Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
`http://siba-ese.unisalento.it/index.php/ejasa/index`
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v11n1p74

Comparisons of ten corrections methods for t-test in multiple comparisons via Monte Carlo study

By Félix, Menezes

Published: 26 April 2018

Electronic Journal of Applied Statistical Analysis
Vol. 11, Issue 01, April 2018, 74-91
DOI: 10.1285/i20705948v11n1p74

Comparisons of ten corrections methods for t-test in multiple comparisons via Monte Carlo study

Vinícius Basseto Félix*and André Felipe Berdusco Menezes

Department of Statistics, Universidade Estadual de Maringá, Maringá, PR, Brasil

Published: 26 April 2018

Multiple comparisons of treatments means are common in several fields of knowledge. The Student's t-test is one of the first procedures to be used in multiple comparisons, however the p-values associated with it are inaccurate, since there is no control on the family-wise Type I error. To solve this problem, several corrections were developed. In this work, based on Monte Carlo simulations, we evaluated the t-test and the following corrections: Bonferroni, Holm, Hochberg, Hommel, Holland, Rom, Finner, Benjamini–Hochberg, Benjamini–Yekutieli and Li with respect to their power and Type I error rate. The study was lead varying the sample size, the sample distribution and the degree of variability. For all instances we regarded three balanced treatments and the probability distributions considered were: Gumbel, Logistic and Normal. Although the corrections were approaching when the sample size increased, our study reveals that the BH correction provides the best family-wise Type I error rate and the second overall most powerful correction.

keywords: t-test, Monte Carlo simulation, Multiple comparison, Type I error rate, Power.

1 Introduction

A common issue in several fields of knowledge is to compare the effects of treatments in order to evaluate if there is difference between them. The researchers can use different statistical methodologies according to the problem. For example, comparisons of

*Corresponding author:felix_prot@hotmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

means, factorial ANOVA, factorial ANOVA with multiple outcome variables, multiple chi-squared tests, multiple log-rank tests, etc. In fact, the problem of testing simultaneously hypotheses is very common in applied statistics. Hence, it is important that these procedures control the family-wise Type I error and keep a higher power.

Regarding the interest of comparing r treatment means, thus the following $m = r(r-1)/2$ hypotheses are tested:

$$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases} \tag{1}$$

with $i \neq j$. Moreover, when several simultaneous hypotheses statistical tests are conducted, both comparison-wise and family-wise Type I error should be considered. As seen in van der Laan et al. (2004) the family-wise Type I error (α_{FWE}) is the main problem, since each comparison inflate the Type I error.

So, let α_{FWE} be the family-wise Type I error rate, that is, the probability of making one or more Type I errors in the set (family) of comparisons, and the comparison-wise error rate the probability of a Type I error for a particular test, denoted by α (see Sheskin, 2000, pp. 533–535). The relationship between these two error rates, when the tests are independent is given by:

$$\alpha_{FWE} = 1 - (1 - \alpha)^m, \tag{2}$$

where m is the number of hypotheses tested.

Full statistical procedures and extensive literature are available to deal with multiple comparisons, and interested readers can refer to Westfall et al. (2011), Howell (2010), Tamhane (2009), Sheskin (2000), Hsu (1996), Shaffer (1995), Westfall and Young (1993), Carmer and Walker (1985), Miller Jr. (1981), Hochberg and Tamhane (1978) and Kemp (1975). On the other hand, several Monte Carlo studies have been conducted to provide the guidance to researchers in selecting a multiple comparisons procedures, interested readers may refer to Dolgun and Demirhan (2016), Li (2015), Demirhan et al. (2010), García et al. (2010), Girardi et al. (2009), Matsunaga (2007), Wu and Liao (2004), Seco et al. (2001), Olejnik et al. (1997), Klockars et al. (1995), Kromrey and La Rocca (1995), Seaman et al. (1991), Einot and Gabriel (1975) and Carmer and Swanson (1973).

The main aim of this paper is to compare the performance of the t-test and the ten corrections in the context of multiple comparisons of means. It is well known that t-test require the assumption of normality in the data, however in practical applications the data not necessarily follows the Normal distribution. Hence, it is crucial to know the properties of t-test and the considered correction when these assumption is not attended. In this view, the simulation study was conducted using different groups sizes, variances and data from different distributions, besides Normal.

Following García et al. (2010), the corrections considered can be classified into:

- one-step: Bonferroni-Dunn (Bonferroni, 1936; Dunn, 1961);
- step-down: Holm (Holm, 1979), Holland (Holland and Copenhaver, 1987) and Finner (Finner, 1993);

- step-up: Hochberg (Hochberg, 1988), Hommel (Hommel, 1988), Rom (Rom, 1990), Benjamini–Hochberg (Benjamini and Hochberg, 1995) and Benjamini–Yekutieli (Benjamini and Yekutieli, 2001);
- two-step: Li (Li, 2008).

It should be point out that although this corrections are used here to correct the p-values of t-test they can be use for other procedures and situations. For instance, in chi-squared test to compare the frequency distributions of a multinomial outcome of interest for two or more groups (Jin and Wang, 2014) or nonparametric tests (Log-rank, Wilcoxon, Peto-Peto, etc.) used to compare different groups in survival analysis (SAS Institute Inc., 2011).

The remainder of the paper is organized as follows. Section 2.2 discuss the Student's t-test and the ten corrections considered in this paper. The specifications of the Monte Carlo simulations are introduced in Section 2.3. The results are presented and discussed in Section 3. Some concluding remarks in Section 4 finalize this paper.

The contribution of this work comes from the fact that there has been no previous work comparing all of these corrections in different scenarios, especially on the breakdown of assumptions. In this sense, we believe that this work, although not presenting methodological innovations, is useful as an effort to provide guidance to researchers in selecting among these corrections the best, mainly in many real situations where small samples are available.

2 Material and methods

2.1 The Student's t-test

Proposed by Student (1908) the t-test has applications in several areas, been one of the most frequently used procedures in statistics, it can be used to test if the population mean is equal to a specified value (μ_0), if the slope of a linear regression is equal to zero or in certain situations where we can compare the mean of, independent or dependent, samples.

Regard the interest in testing hypothesis, defined in Equation (1), it is well known that the test statistic defined by:

$$T = \frac{\overline{\mathrm{x}}_i - \overline{\mathrm{x}}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}, \tag{3}$$

where $\overline{\mathrm{x}}_i$ and $\overline{\mathrm{x}}_j$ are the means of treatments i and j, respectively, n_i and n_j are the sample sizes of the treatments i and j, respectively and MSE is the mean squared due to error provided by analysis of variance (ANOVA).

Under the null hypothesis, T has t-Student distribution with $(N-2)$ degrees of freedom, where $N = n_i + n_j$, and we reject the null hypothesis, at significance level α, if $\mid T \mid > t(\frac{\alpha}{2}, N-2)$, where $t(\frac{\alpha}{2}, N-2)$ is the quantile $(100 \times \alpha)\%$ from the t-Student distribution with $(N-2)$ degrees of freedom.

2.2 The corrections

This subsection is devoted to describe the corrections considered in this paper. Initially, we introduce the notation used in the paper.

- P_i: the i^{th} p-value;
- $P_{(i)}$: the i^{th} order p-value, after ordering the p-values from lowest to highest;
- H_i: the i^{th} null hypothesis;
- $H_{(i)}$: the i^{th} null hypothesis concerning the respective $P_{(i)}$;
- m: the number of hypotheses tested.

2.2.1 Bonferroni correction

The first of them was the Bonferroni correction, named after Carlo Emilio Bonferroni, since its use of the Bonferroni inequalities, as seen in Bonferroni (1936), it is also called the Bonferroni-Dunn correction, since the modern approach given by Dunn (1961). This method consists of calculating a new significance level to keep the family-wise Type I error at α, it is given by

$$\alpha_B = \frac{\alpha}{m}, \tag{4}$$

where m is the number of hypotheses tested.

The Bonferroni is probably the most commonly used correction, because it is highly flexible, very simple to compute, and can be used with any type of statistical test, not just post hoc tests after an omnibus test like ANOVA, since corrects the experiment-wise error when a large number of independent tests are performed Armstrong (2014). However, it tends to lack power, some of the reasons are given by Perneger (1998) and Keppel and Wickens (2004):

1. The α_{FWE} calculation depends on the assumption that, for all tests, the null hypothesis is true, this is unlikely to be the case, especially after a significant omnibus test;
2. All tests are assumed to be orthogonal (i.e., independent or nonoverlapping) when calculating the α_{FWE} test, and this is usually not the case when all pairwise comparisons are made;
3. It does not take into account whether the findings are consistent with theory and past research. If consistent with previous findings and theory, an individual result should be less likely to be a Type I error;
4. Type II error rates are too high for individual tests. In other words, the Bonferroni over corrects for Type I error.

2.2.2 Holm correction

The Holm correction (also called the Holm-Bonferroni or Bonferroni-Holm correction), Holm (1979), is intended to control the family-wise error rate and offer a simple test uniformly more powerful than the Bonferroni correction, the procedure consists of

1. Let k be the minimal index such that $P_k > \frac{\alpha}{m+1-k}$;
2. Then, we reject the null hypothesis $H_{(1)}, \ldots, H_{(k-1)}$.

2.2.3 Hochberg correction

The Hochberg correction, Hochberg (1988), follows the Holm correction idea, but with a chance in the second step, considering now a index that $P_{(k)}$ is less or equal the corrected significance level

1. For any $k = m,\ m-1, \ldots, 1$, if $P_k \leq \frac{\alpha}{m+1-k}$;
2. Then, we reject the null hypothesis $H_{(1)}, \ldots, H_{(k)}$.

2.2.4 Holland correction

The Holland correction, Holland and Copenhaver (1987), is a improve to the Holm correction.

1. Let k the smallest integer such that $P_k > 1 - (1-\alpha)^{1/k}$;
2. Then, we reject the null hypothesis $H_{(1)}, \ldots, H_{(k-1)}$ and accept $H_{(k)}, \ldots, H_{(m)}$;
3. If no k satisfy the condition we reject all hyphotheses.

2.2.5 Hommel correction

The Hommel correction, Hommel (1988), is more powerful than the Hochberg correction, but is also more complex.

1. Let j be the largest index such that $P_{n-j+k} > \frac{k\alpha}{j}, \quad k = 1, \ldots, j.$;
2. Then, if no such j exists, reject all hypothesis; otherwise, reject all H_i with $P_i \leq \frac{\alpha}{j}$.

2.2.6 Rom correction

The Rom correction, Rom (1990), is a modification of Hochberg procedure to increase its power. The difference occurs in how the adjusted significance level (α) is obtained, i.e, the α values are determined by the solution of the following equation:

$$\sum_{i=1}^{n-1} c_{m_m}^{i} - \binom{n}{i} c_{(m-i)_m}^{m-i} = 0 \tag{5}$$

where $i = 1, 2, \ldots, m$ and c is a replacement of the critical points given by $A(\alpha)_m = pr(p_{(1)} \geq c_{1_m}, \ldots, p_{(m)} \geq c_{m_m}) = 1 - \alpha$, where under the global null hypothesis, $p_{(1)}, \ldots, p_{(m)}$ are the order statistics of n independent uniform $(0, 1)$ random variables with joint density $n!$ $(0 \leq p_1 \leq ... \leq p_n \leq 1)$; 0, otherwise (Rom, 1990).

2.2.7 Finner correction

The Finner correction, Finner (1993), works in a quite similar way to Holm and Holland corrections.

1. Let k the smallest integer such that $P_k > 1 - (1 - \alpha)^{(m-1)/k}$;
2. Then, we reject the null hypothesis $H_{(1)}, \ldots, H_{(k-1)}$.

2.2.8 Benjamini–Hochberg correction

The Benjamini–Hochberg correction (also called the BH step-up or BH correction), Benjamini and Hochberg (1995), which was developed to control the false discovery rate, work as follows

1. Let k the largest number such that $P_k \leq \frac{k}{m}\alpha$;
2. Then, we reject the null hypothesis $H_{(1)}, \ldots, H_{(k)}$.

However this method is valid if only the m tests are independent.

2.2.9 Benjamini–Hochberg–Yekutieli correction

The Benjamini–Hochberg–Yekutieli correction (also called the BY correction), Benjamini and Yekutieli (2001), is a refinement of the BH correction to work under positive dependence assumptions, so

1. Let k the largest number such that $P_k \leq \frac{k}{mc(m)}\alpha$, where $c(m) = \sum_{i=1}^{m} \frac{1}{i}$;
2. Then, we reject the null hypothesis $H_{(1)}, \ldots, H_{(k)}$.

And, as seen in Li (2015) this procedure is more conservative than the procedure proposed by Benjamini and Hochberg (1995) due to its smaller critical value, then it has strong control of the false discovery rate for both independent and dependent test statistics.

2.2.10 Li correction

The Li correction, Li (2008), is a two-step procedure and work as follows

1. Reject all H_i, if $P_{m-1} \leq \alpha$. Otherwise accept the hypothesis associated with P_m;
2. Reject any remaining H_i whenever $P_i \leq \frac{1-P_m}{(1-\alpha)\,\alpha}$

2.3 Simulation Study

We evaluated the corrections described with respect to their power and Type I error rate, via Monte Carlo simulation. The study was lead in software `R`, R Core Team (2016), and the package `scmamp` (Statistical Comparison of Multiple Algorithms in Multiple Problems) developed by Calvo and Santafe (2015) that computes the corrections Holland, Rom, Finner and Li.

The scenarios consisted of three balanced treatments, taking as group sample sizes (n_i), $n_i = 2, 3, 5, 10, 15$ and 20 and scale parameter $\sigma = 1$ and 2. Without loss of generality we considered the location parameter $\mu = 0$. For each combination (n, σ) we have generated $B = 10000$ pseudo-random samples from Normal, Logistic and Gumbel distributions (see Table 1).

As mentioned in the previous section it is very crucial to know the behaviour of the t-test and corrections when the assumption of normality is violated. This motivate us to generate data from other distributions. The choice of Logistic comes from the fact that is a symmetric distribution as Normal, but with heavier tails. Further the Gumbel distribution was chosen because their asymmetric behaviour.

Table 1: Distributions considered on the study.

Distribution	Density	Mean	Std. deviation
Normal	$\frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	μ	σ
Logistic	$\frac{1}{\sigma} \frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{\left[1+\exp\left(\frac{x-\mu}{\sigma}\right)\right]^2}$	μ	$\frac{\pi\,\sigma}{\sqrt{3}}$
Gumbel	$\frac{1}{\sigma} \exp\left[\left(\frac{x-\mu}{\sigma}\right) - \exp\left(\frac{x-\mu}{\sigma}\right)\right]$	$\mu + \sigma\,\gamma$	$\frac{\pi\,\sigma}{\sqrt{6}}$

One of the main problem in a study on the power of a test is the unlimited number of ways that the alternative hypothesis can be formulated. So, in this study we consider one of the treatments with different means (from -6 to 6). Therefore, the empirical power of the test $(\widehat{\tau})$ was obtained by:

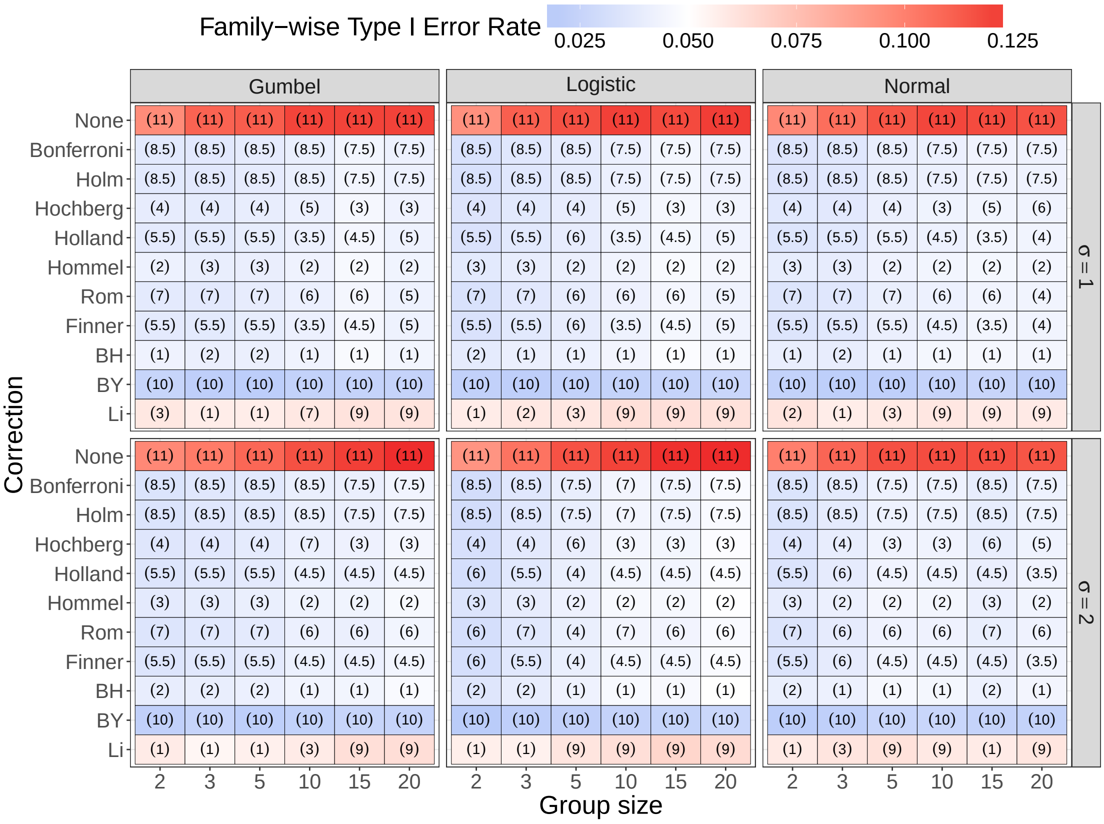
$$\widehat{\tau} = \frac{\text{Number of times that } H_0 \text{ is rejected in the specifics hypothesis } \mid \ H_0 \text{ is false}}{B}.$$

On the other hand, we can estimate the family-wise Type I error rate $(\widehat{\alpha}_{FWE})$ generating independents samples from the null hypothesis and calculating the proportion of times that H_0 was rejected wrongly, considering the nominal significance level 5%. Formally, we have:

$$\widehat{\alpha}_{FWE} = \frac{\text{Number of times that } H_0 \text{ is rejected in at least one hypothesis} \mid H_0 \text{ is true}}{B}.$$

3 Results and Discussion

To make easier to identify the best corrections a ranking was made for each group size, for the family-wise Type I error rate we took into account the absolute difference between $\widehat{\alpha}_{FWE}$ and α, so the lower the ranking the better the correction, as we can see in Figure 1. And for the power ranking, the value of the empirical power itself was used, so a higher power means a lower ranking and consequently a better correction.



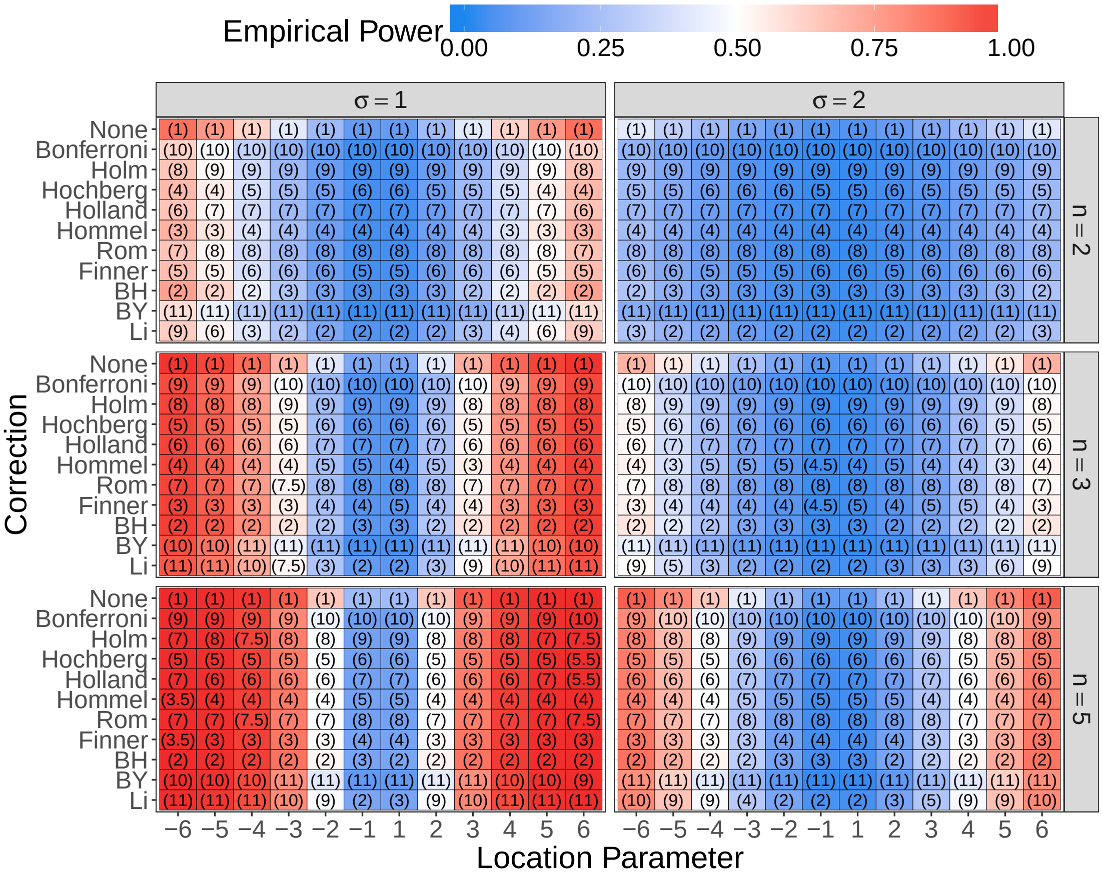
σ = 1	Gumbel						Logistic						Normal					
Correction	2	3	5	10	15	20	2	3	5	10	15	20	2	3	5	10	15	20
None	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)
Bonferroni	(8.5)	(8.5)	(8.5)	(8.5)	(7.5)	(7.5)	(8.5)	(8.5)	(8.5)	(7.5)	(7.5)	(7.5)	(8.5)	(8.5)	(8.5)	(7.5)	(7.5)	(7.5)
Holm	(8.5)	(8.5)	(8.5)	(8.5)	(7.5)	(7.5)	(8.5)	(8.5)	(8.5)	(7.5)	(7.5)	(7.5)	(8.5)	(8.5)	(8.5)	(7.5)	(7.5)	(7.5)
Hochberg	(4)	(4)	(4)	(5)	(3)	(3)	(4)	(4)	(4)	(5)	(3)	(3)	(4)	(4)	(4)	(3)	(5)	(6)
Holland	(5.5)	(5.5)	(5.5)	(3.5)	(4.5)	(5)	(5.5)	(5.5)	(6)	(3.5)	(4.5)	(5)	(5.5)	(5.5)	(5.5)	(4.5)	(3.5)	(4)
Hommel	(2)	(3)	(3)	(2)	(2)	(2)	(3)	(3)	(2)	(2)	(2)	(2)	(3)	(3)	(2)	(2)	(2)	(2)
Rom	(7)	(7)	(7)	(6)	(6)	(5)	(7)	(7)	(6)	(6)	(6)	(5)	(7)	(7)	(7)	(6)	(6)	(4)
Finner	(5.5)	(5.5)	(5.5)	(3.5)	(4.5)	(5)	(5.5)	(5.5)	(6)	(3.5)	(4.5)	(5)	(5.5)	(5.5)	(5.5)	(4.5)	(3.5)	(4)
BH	(1)	(2)	(2)	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(1)
BY	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)
Li	(3)	(1)	(1)	(7)	(9)	(9)	(1)	(2)	(3)	(9)	(9)	(9)	(2)	(1)	(3)	(9)	(9)	(9)

σ = 2	Gumbel						Logistic						Normal					
Correction	2	3	5	10	15	20	2	3	5	10	15	20	2	3	5	10	15	20
None	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)
Bonferroni	(8.5)	(8.5)	(8.5)	(8.5)	(7.5)	(7.5)	(8.5)	(8.5)	(7.5)	(7)	(7.5)	(7.5)	(8.5)	(8.5)	(7.5)	(7.5)	(8.5)	(7.5)
Holm	(8.5)	(8.5)	(8.5)	(8.5)	(7.5)	(7.5)	(8.5)	(8.5)	(7.5)	(7)	(7.5)	(7.5)	(8.5)	(8.5)	(7.5)	(7.5)	(8.5)	(7.5)
Hochberg	(4)	(4)	(4)	(7)	(3)	(3)	(4)	(4)	(6)	(3)	(3)	(3)	(4)	(4)	(3)	(3)	(6)	(5)
Holland	(5.5)	(5.5)	(5.5)	(4.5)	(4.5)	(4.5)	(6)	(5.5)	(4)	(4.5)	(4.5)	(4.5)	(5.5)	(6)	(4.5)	(4.5)	(4.5)	(3.5)
Hommel	(3)	(3)	(3)	(2)	(2)	(2)	(3)	(3)	(2)	(2)	(2)	(2)	(3)	(2)	(2)	(2)	(3)	(2)
Rom	(7)	(7)	(7)	(6)	(6)	(6)	(6)	(7)	(4)	(7)	(6)	(6)	(7)	(6)	(6)	(6)	(7)	(6)
Finner	(5.5)	(5.5)	(5.5)	(4.5)	(4.5)	(4.5)	(6)	(5.5)	(4)	(4.5)	(4.5)	(4.5)	(5.5)	(6)	(4.5)	(4.5)	(4.5)	(3.5)
BH	(2)	(2)	(2)	(1)	(1)	(1)	(2)	(2)	(1)	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(2)	(1)
BY	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)
Li	(1)	(1)	(1)	(3)	(9)	(9)	(1)	(1)	(9)	(9)	(9)	(9)	(1)	(3)	(9)	(9)	(1)	(9)

Figure 1: Family-wise Type I error rate of the multiple comparisons corrections in all scenarios, with their respective rankings.

In Figure 1 we see that in almost all scenarios the BH correction is the best, with just a few exceptions. As expected, with no correction the $\widehat{\tau}$ is the worst in all scenarios. Besides that, the BY correction produces lower rates than the rest and the Li correction produces one of the best rankings when the group size is lower and it became worse as the group size grows.

We also see the power of those corrections, so we vary the location parameter from −6 to 6, as we can see in Figures 2,3, 4, 5, 6 and 7.



Correction	−6	−5	−4	−3	−2	−1	1	2	3	4	5	6
None	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
Bonferroni	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)
Holm	(8)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(8)
Hochberg	(4)	(4)	(5)	(5)	(5)	(6)	(6)	(5)	(5)	(5)	(4)	(4)
Holland	(6)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(6)
Hommel	(3)	(3)	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(3)	(3)	(3)
Rom	(7)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(7)
Finner	(5)	(5)	(6)	(6)	(6)	(5)	(5)	(6)	(6)	(6)	(5)	(5)
BH	(2)	(2)	(2)	(3)	(3)	(3)	(3)	(3)	(2)	(2)	(2)	(2)
BY	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)
Li	(9)	(6)	(3)	(2)	(2)	(2)	(2)	(2)	(3)	(4)	(6)	(9)

Correction	−6	−5	−4	−3	−2	−1	1	2	3	4	5	6
None	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
Bonferroni	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)
Holm	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)
Hochberg	(5)	(5)	(6)	(6)	(6)	(5)	(5)	(6)	(5)	(5)	(5)	(5)
Holland	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)
Hommel	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)
Rom	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)
Finner	(6)	(6)	(5)	(5)	(5)	(6)	(6)	(5)	(6)	(6)	(6)	(6)
BH	(2)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(2)
BY	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)
Li	(3)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(3)

Correction	−6	−5	−4	−3	−2	−1	1	2	3	4	5	6
None	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
Bonferroni	(9)	(9)	(9)	(10)	(10)	(10)	(10)	(10)	(10)	(9)	(9)	(9)
Holm	(8)	(8)	(8)	(9)	(9)	(9)	(9)	(9)	(8)	(8)	(8)	(8)
Hochberg	(5)	(5)	(5)	(5)	(6)	(6)	(6)	(6)	(5)	(5)	(5)	(5)
Holland	(6)	(6)	(6)	(6)	(7)	(7)	(7)	(7)	(6)	(6)	(6)	(6)
Hommel	(4)	(4)	(4)	(4)	(5)	(5)	(4)	(5)	(3)	(4)	(4)	(4)
Rom	(7)	(7)	(7)	(7.5)	(8)	(8)	(8)	(8)	(7)	(7)	(7)	(7)
Finner	(3)	(3)	(3)	(3)	(4)	(4)	(5)	(4)	(4)	(3)	(3)	(3)
BH	(2)	(2)	(2)	(2)	(2)	(3)	(3)	(2)	(2)	(2)	(2)	(2)
BY	(10)	(10)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(10)	(10)
Li	(11)	(11)	(10)	(7.5)	(3)	(2)	(2)	(3)	(9)	(10)	(11)	(11)

Correction	−6	−5	−4	−3	−2	−1	1	2	3	4	5	6
None	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
Bonferroni	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)
Holm	(8)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(9)	(8)
Hochberg	(5)	(6)	(6)	(6)	(6)	(6)	(6)	(6)	(6)	(6)	(5)	(5)
Holland	(6)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(7)	(6)
Hommel	(4)	(3)	(5)	(5)	(5)	(4.5)	(4)	(5)	(4)	(4)	(3)	(4)
Rom	(7)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(7)
Finner	(3)	(4)	(4)	(4)	(4)	(4.5)	(5)	(4)	(5)	(5)	(4)	(3)
BH	(2)	(2)	(2)	(3)	(3)	(3)	(3)	(2)	(2)	(2)	(2)	(2)
BY	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)
Li	(9)	(5)	(3)	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(6)	(9)

Correction	−6	−5	−4	−3	−2	−1	1	2	3	4	5	6
None	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
Bonferroni	(9)	(9)	(9)	(9)	(10)	(10)	(10)	(10)	(9)	(9)	(9)	(10)
Holm	(7)	(8)	(7.5)	(8)	(8)	(9)	(9)	(8)	(8)	(8)	(7)	(7.5)
Hochberg	(5)	(5)	(5)	(5)	(5)	(6)	(6)	(5)	(5)	(5)	(5)	(5.5)
Holland	(7)	(6)	(6)	(6)	(6)	(7)	(7)	(6)	(6)	(6)	(7)	(5.5)
Hommel	(3.5)	(4)	(4)	(4)	(4)	(5)	(5)	(4)	(4)	(4)	(4)	(4)
Rom	(7)	(7)	(7.5)	(7)	(7)	(8)	(8)	(7)	(7)	(7)	(7)	(7.5)
Finner	(3.5)	(3)	(3)	(3)	(3)	(4)	(4)	(3)	(3)	(3)	(3)	(3)
BH	(2)	(2)	(2)	(2)	(2)	(3)	(2)	(2)	(2)	(2)	(2)	(2)
BY	(10)	(10)	(10)	(11)	(11)	(11)	(11)	(11)	(11)	(10)	(10)	(9)
Li	(11)	(11)	(11)	(10)	(9)	(2)	(3)	(9)	(10)	(11)	(11)	(11)

Correction	−6	−5	−4	−3	−2	−1	1	2	3	4	5	6
None	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
Bonferroni	(9)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(10)	(9)
Holm	(8)	(8)	(8)	(9)	(9)	(9)	(9)	(9)	(9)	(8)	(8)	(8)
Hochberg	(5)	(5)	(5)	(6)	(6)	(6)	(6)	(6)	(6)	(5)	(5)	(5)
Holland	(6)	(6)	(6)	(7)	(7)	(7)	(7)	(7)	(7)	(6)	(6)	(6)
Hommel	(4)	(4)	(4)	(5)	(5)	(5)	(5)	(5)	(4)	(4)	(4)	(4)
Rom	(7)	(7)	(7)	(8)	(8)	(8)	(8)	(8)	(8)	(7)	(7)	(7)
Finner	(3)	(3)	(3)	(3)	(4)	(4)	(4)	(4)	(3)	(3)	(3)	(3)
BH	(2)	(2)	(2)	(2)	(3)	(3)	(3)	(2)	(2)	(2)	(2)	(2)
BY	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)	(11)
Li	(10)	(9)	(9)	(4)	(2)	(2)	(2)	(3)	(5)	(9)	(9)	(10)

Figure 2: Empirical power of the multiple comparisons corrections for the group sizes of 2, 3 and 5 of the Gumbel distribution, with their respective rankings.

In Figure 2 first we observe the scenarios of the smaller groups in the Gumbel distribution, is visible the difference between the scales parameters, since when $\sigma = 2$ the empirical power is lower, even for the cases when μ is distant from 0. Also, unlike in the family-wise Type I error rate, now the absent of correction provide the higher empirical powers in all cases, however, the BH correction still maintain a good ranking been the second best overall, as the Li correction been good when the group size is small. After, we seek to observe the empirical power when the group size is bigger (Figure 3).

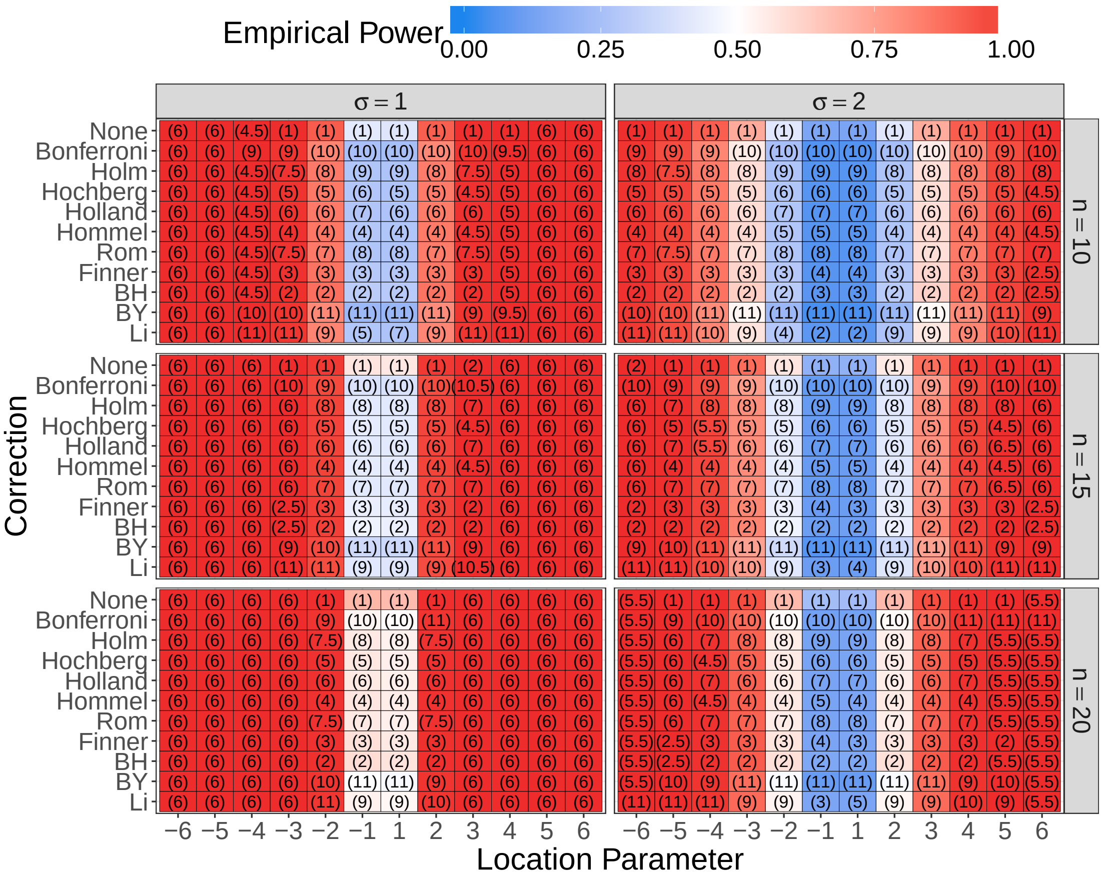


Figure 3: Empirical power of the multiple comparisons corrections for the group sizes of 10, 15 and 20 of the Gumbel distribution, with their respective rankings.

In Figure 3 we can observe that for larger group sizes the empirical power is higher, and pretty much the same for all corrections, still for locations parameters near of 0 the corrections rankings are the same that for the smaller group sizes (Figure 2).

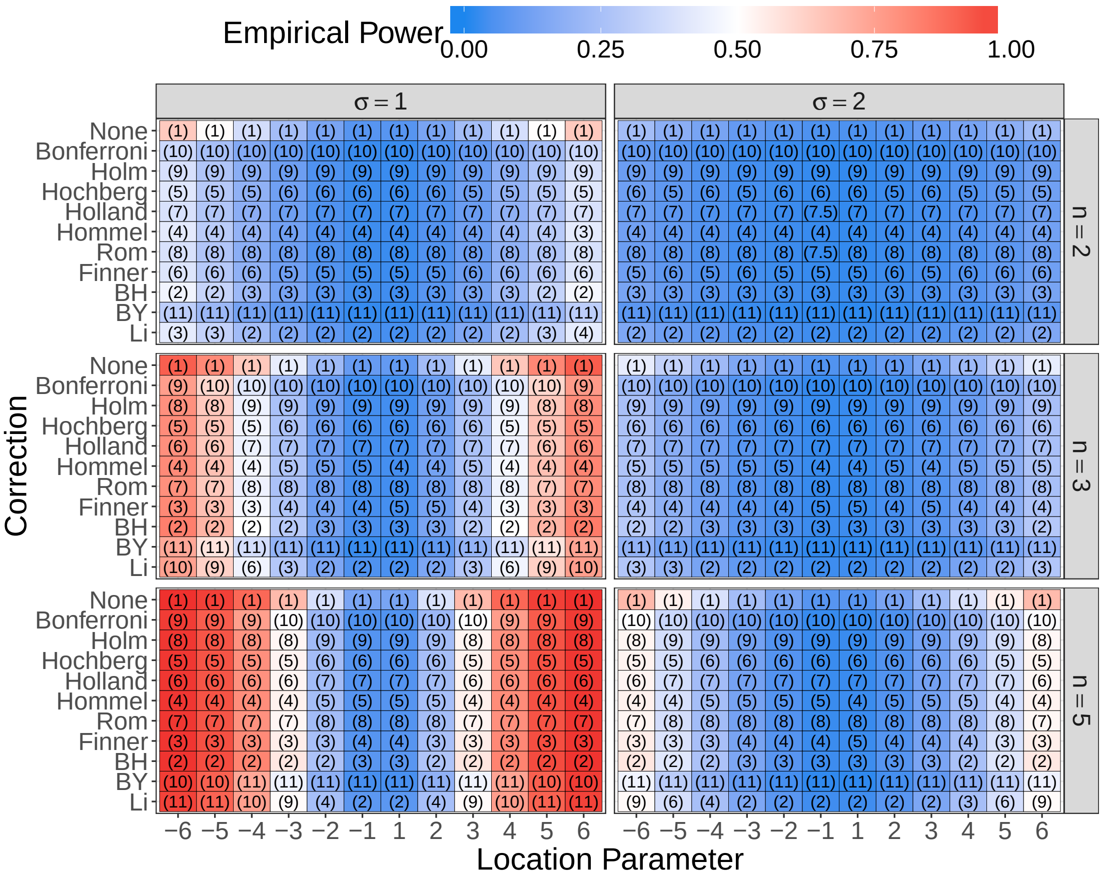


Figure 4: Empirical power of the multiple comparisons corrections for the group sizes of 2, 3 and 5 of the Logistic distribution, with their respective rankings.

In Figure 4 we see the scenarios in the Logistic distribution, and if compared with Gumbel distribution (Figure 2) the empirical power is a lot smaller, since this distribution has heavy tails. Still, the absence of correction continue to be the most powerful for all scenarios, and with the exception of Li correction the others keeps their ranking as the location parameter vary.

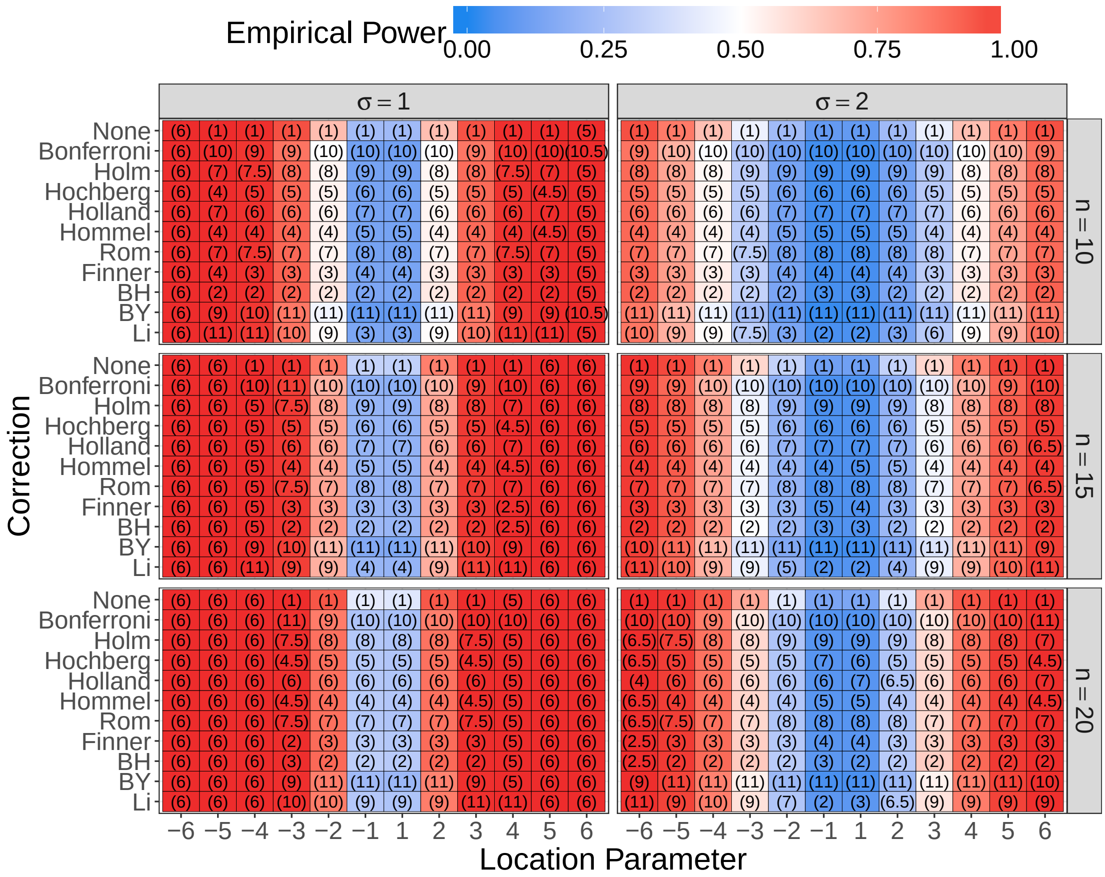


Figure 5: Empirical power of the multiple comparisons corrections for the group sizes of 10, 15 and 20 of the Logistic distribution, with their respective rankings.

The main difference observed in Figure 5 is the behavior of the Li correction if compared to smaller group sizes (Figure 4), since it was one of the best corrections for the nearest locations parameters to 0, and now is one the worst.

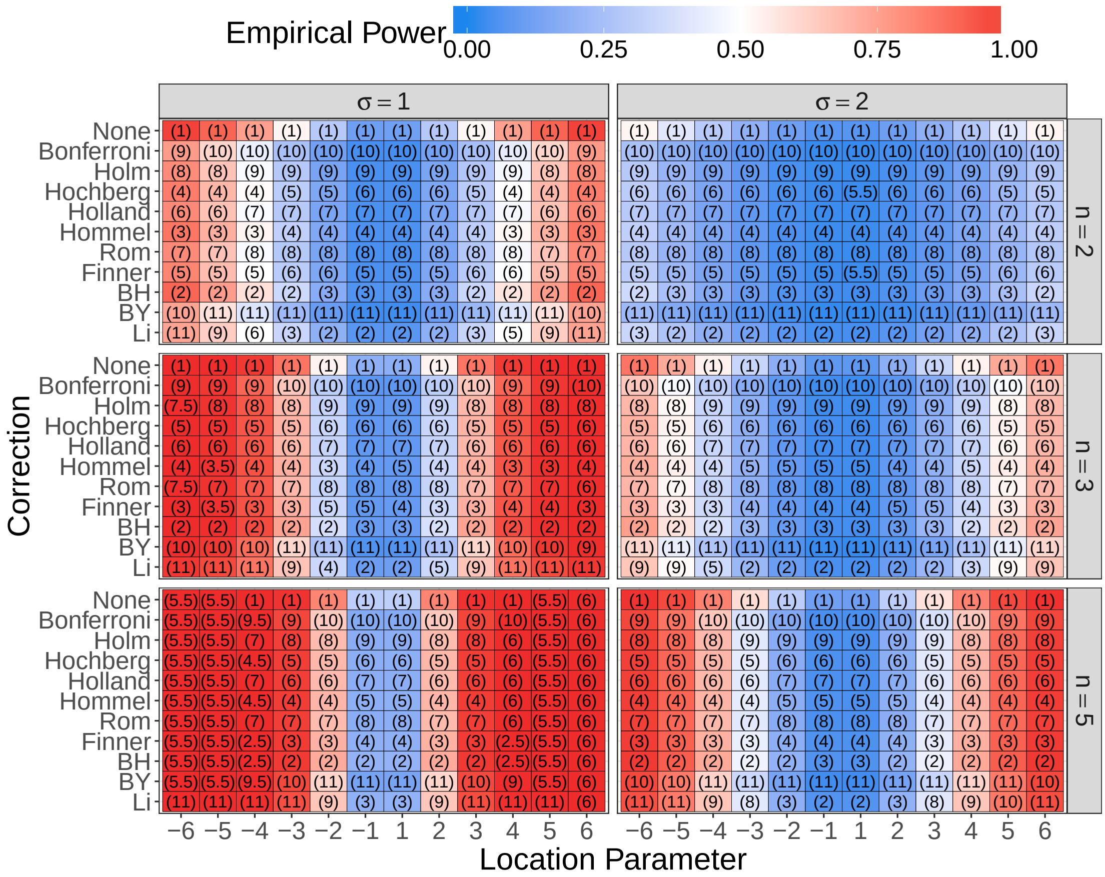


Figure 6: Empirical power of the multiple comparisons corrections for the group sizes of 2, 3 and 5 of the Normal distribution, with their respective rankings.

Finally, we see the Normal distribution in Figure 6, and the results are pretty similar to Gumbel (Figure 2), having pretty much the same behavior and interpretations.

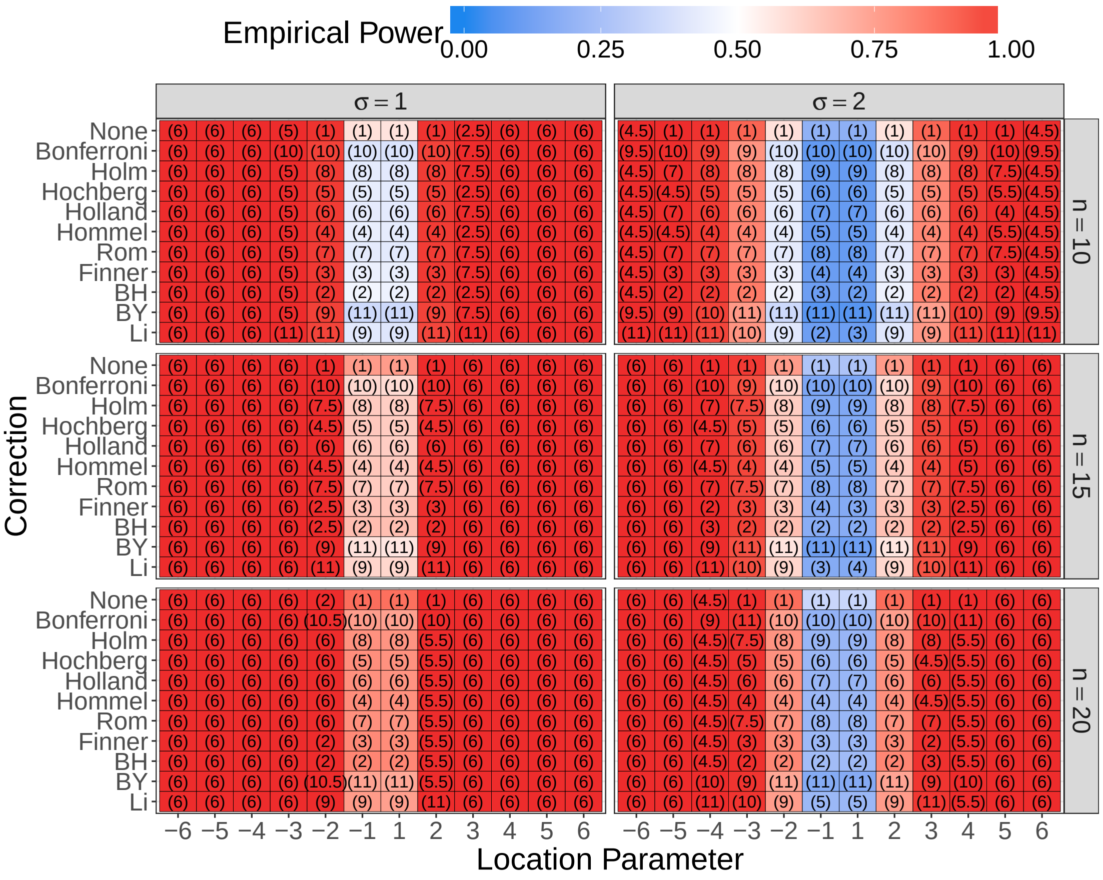


Figure 7: Empirical power of the multiple comparisons corrections for the group sizes of 10, 15 and 20 of the Normal distribution, with their respective rankings.

In Figure 7 we observe that for the case when the group sample size is equal to 10 and nearest location parameters to 0, the Li correction is the $9^t h$ correction for a scale parameter of 1, but in the case of $\sigma = 2$ it became one of the best.
Based on the results from the charts, we observed that all corrections works, in the sense that all of them minimize the Type I error rate, since is smaller than in the absence of a correction, and somewhat near of the nominal level. Therefore, we can summary the results as:

- Absence of correction: As expected the worst family-wise Type I error rate, and the most powerful one in all scenarios;

- Bonferroni correction: The most popular correction, it was actually the second overall worst correction in both criteria;

- Holm correction: In the Type I measure provided the same results as the Bonferroni but it was slightly more powerful;

- Hochberg correction: This correction was pretty average, since in general it varied its ranking between 4 and 6;
- Holland correction: The results agreed with theory, because even overall the correction was average it was indeed better than Holm, as it was already its purpose;
- Hommel correction: It was as stated in literature slightly more powerful than Hochberg;
- Rom correction: On the contrary of Holland and Hommel, the correction was less powerful than Hochberg, been one of the worst overall;
- Finner correction: For the family-wise Type I error rate it was pretty average, but one of the most powerful when the group size was bigger than 2;
- BH correction: Not only provided the best family-wise Type I error rate, but also the second overall most powerful correction;
- BY correction: Analyzing both criteria it was the worst, as expected since it was a refinement of the BH correction to work under positive dependence assumptions, and that was not the case in those simulations;
- Li correction: Provided interesting results, such as been better minimizing the α_{FWE} in smaller groups and been more powerful when the scale parameter is higher.

In terms of distribution, when the normality assumption is not broken the power is higher in comparison to Gumbel and Logistic.

4 Conclusion

In this paper, we compared by intensive simulation experiments ten corrections applied in the t-test, most of them were made to minimize the family-wise Type I error, ensuring that the inflation effect was minimal, providing a test with the optimal Type I error.

As we expected the t-test with no correction was the best in respect to the power, however it was the worst with respect to family-wise Type I error rate. Interestingly, the most popular correction, Bonferroni, was one of the worst. On the other side the BH correction was the best overall, that is, it was good in both criteria

References

Armstrong, R. A. (2014). When to use the bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber.

Calvo, B. and Santafe, G. (2015). scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal*, Accepted for publication.

Carmer, S. G. and Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by monte carlo methods. *Journal of the American Statistical Association*, 68(341):66–74.

Carmer, S. G. and Walker, W. M. (1985). Pairwise multiple comparisons of treatment means in agronomic research. *Journal of Agronomic Education*.

Demirhan, H., Dolgun, N. A., Demirhan, Y. P., and Özgür Dolgun, M. (2010). Performance of some multiple comparison tests under heteroscedasticity and dependency. *Journal of Statistical Computation and Simulation*, 80(10):1083–1100.

Dolgun, A. and Demirhan, H. (2016). Performance of nonparametric multiple comparison tests under heteroscedasticity, dependency, and skewed error distribution. *Communications in Statistics - Simulation and Computation*, pages 1–18.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Einot, I. and Gabriel, K. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, 70(351a):574–583.

Finner, H. (1993). On a monotonicity problem in step-down multiple test procedures. *Journal of the American Statistical Association*, 88(423):920–923.

García, S., Fernández, A., Luengo, J., and 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(10):2044–2064.

Girardi, L. H., Cargnelutti Filho, A., and Storck, L. (2009). Type i error and power of five multiple comparison procedures for means. *Rev. Bras. Biom.*, 27(1):23–36.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Hochberg, Y. and Tamhane, A. C. (1978). *Multiple Comparison Procedures*. John Wiley & Sons.

Holland, B. S. and Copenhaver, M. D. (1987). An improved sequentially rejective bonferroni test procedure. *Biometrics*, pages 417–423.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386.

Howell, D. C. (2010). *Statistical Methods for Psychology*. Wadsworth, Cengage Learning, Seventh edition.

Hsu, J. C. (1996). *Multiple Comparison Theory and Methods.* Chapman & Hall, Second edition.

Jin, M. and Wang, B. (2014). Implementing multiple comparisons on pearson chi-square test for an rxc contingency table in sas®.

Kemp, K. (1975). Multiple comparisons: comparisonwise versus experimentwise type i error rates and their relationship to power. *Journal of dairy science*, 58(9):1374–1378.

Keppel, G. and Wickens, T. (2004). Simultaneous comparisons and the control of type i errors. *Design and analysis: A researcher's handbook. 4th ed. Upper Saddle River (NJ): Pearson Prentice Hall. p*, pages 111–130.

Klockars, A. J., Hancock, G. R., and McAweeney, M. J. (1995). Power of unweighted and weighted versions of simultaneous and sequential multiple-comparison procedures. *Psychological Bulletin*, 118(2):300.

Kromrey, J. D. and La Rocca, M. A. (1995). Power and type i error rates of new pairwise multiple comparison procedures under heterogeneous variances. *The Journal of Experimental Education*, 63(4):343–362.

Li, D. (2015). Power and stability comparisons of multiple testing procedures with false discovery rate control. *Journal of Statistical Computation and Simulation*, 85(14):2808–2822.

Li, J. D. (2008). A two-step rejection procedure for testing multiple hypotheses. *Journal of Statistical Planning and Inference*, 138(6):1521–1527.

Matsunaga, M. (2007). Familywise error in multiple comparisons: Disentangling a knot through a critique of o'keefe's arguments against alpha adjustment. *Communication Methods and Measures*, 1(4):243–265.

Miller Jr., R. G. (1981). *Simultaneous Statistical Inference.* Springer, First edition.

Olejnik, S., Li, J., Supattathum, S., and Huberty, C. J. (1997). Multiple testing and statistical power with modified bonferroni procedures. *Journal of educational and behavioral statistics*, 22(4):389–406.

Perneger, T. V. (1998). What's wrong with bonferroni adjustments. *BMJ: British Medical Journal*, 316(7139):1236.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified bonferroni inequality. *Biometrika*, pages 663–665.

SAS Institute Inc. (2011). *The LIFETEST Procedure, SAS®/STAT User's Guide, Version 9.3.* Cary, NC: SAS Institute Inc. 4322–4412.

Seaman, M. A., Levin, J. R., and Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin*, 110(3):577.

Seco, G. V., Menéndez de la Fuente, I. A., and Escudero, J. R. (2001). Pairwise multiple comparisons under violation of the independence assumption. *Quality and Quantity*, 35(1):61–76.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584.

Sheskin, D. J. (2000). *Parametric and nonparametric statistical procedures.* Chapman & Hall.

Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

Tamhane, A. C. (2009). *Statistical Analysis of Designed Experiments Theory and Applications.* John Wiley & Sons, Inc.

van der Laan, M. J., Dudoit, S., Pollard, K. S., et al. (2004). Multiple testing. part ii. step-down procedures for control of the family-wise error rate. *Statistical applications in genetics and molecular biology*, 3(1):1041.

Westfall, P. H., Tobias, R. D., and Wolfinger, R. D. (2011). *Multiple comparisons and multiple tests using SAS.* SAS Institute.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* Joh.

Wu, S.-F. and Liao, B.-X. (2004). A simulation study of multiple comparisons with the average under heteroscedasticity. *Communications in Statistics - Simulation and Computation*, 33(3):639–659.