

Monte Carlo study of multiple comparisons corrections in t-test

André Felipe B. Menezes
Vinícius Basseto Félix

State University of Maringá
Department of Statistics
5th Workshop on Probabilistic and Statistical Methods

February 7, 2017

Outline

- 1 Introduction
- 2 The Student's t-test and its corrections
- 3 Simulation study
- 4 Results
- 5 Conclusion
- 6 References

Motivation and Purposes

- ▶ Pairwise multiple comparisons of treatments means are common in several fields of knowledge;
- ▶ The Student's t-test is one of the first procedures developed;
- ▶ However p-values associated with the t-test are inaccurate, since there is no control on the familywise Type I error;
- ▶ Show the several corrections developed to solve this;
- ▶ Comparing all of these corrections in different scenarios, and find out which one is the best.

Hypothesis

The comparisons were made between all pairs of means. Therefore we have the following hypothesis

$$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases} \quad (1)$$

with $(i \neq j)$. This procedure is repeated for all $m = \frac{k(k-1)}{2}$ comparisons, where k is the number of treatments.

Type I Error

- ▶ **Comparisonwise error rate:** This is the probability of a Type I error for a particular test. We will denote this error rate α .
- ▶ **Familywise error rate:** This is the probability of making one or more type-I errors in the set (family) of comparisons. We will denote this error rate α_{FWE} .
- ▶ The relationship between these two error rates when the tests are independent is given by:

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

where m is the number of hypothesis tested.

The Student's t-test and its corrections

The Student's t-test

Proposed in 1908 the t-test can be used for different purposes. In multiple comparison the statistic is defined by

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

where T has t-Student distribution with $(N - 2)$ degrees of freedom, been $N = n_i + n_j$, \bar{x}_i and \bar{x}_j the means of treatments i and j , respectively and MSE the mean squared due to error provided by analysis of variance.

The Student's t-test and its corrections

Bonferroni correction

The Bonferroni correction consists of calculate a new significance level to keep the familywise Type I error at α , given by

$$\alpha_B = \frac{\alpha}{m}, \quad (4)$$

where m is the number of hypothesis tested.

The Student's t-test and its corrections

Holm correction

The Holm correction is intended to control the familywise error rate and offers a simple test uniformly more powerful than the Bonferroni correction.

Let H_1, \dots, H_m be a family of hypothesis and P_1, \dots, P_m the respective p-values, we do the following

- 1 Order the p-values from lowest to highest, been $P_{(1)}, \dots, P_{(m)}$ and consequently $H_{(1)}, \dots, H_{(m)}$;
- 2 Let k be the index such that $P_k > \frac{\alpha}{m+1-k}$;
- 3 Then, we reject the null hypothesis $H_{(1)}, \dots, H_{(k-1)}$.

The Student's t-test and its corrections

Hochberg correction

The Hochberg correction follows the Holm correction idea, considering now a index that $P_{(k)}$ is less or equal the corrected significance level:

- 1 Order the p-values from lowest to highest, been $P_{(1)}, \dots, P_{(m)}$ and consequently $H_{(1)}, \dots, H_{(m)}$;
- 2 Let k be the index such that $P_k \leq \frac{\alpha}{m+1-k}$;
- 3 Then, we reject the null hypothesis $H_{(1)}, \dots, H_{(k)}$.

The Student's t-test and its corrections

Hommel correction

The Hommel correction is more powerful than Hochberg but is more complex. Let j be the largest integer for which

$$P_{(m-j+k)} > \frac{k\alpha}{j}. \quad (5)$$

for all $k = 1, \dots, j$. If no such j exists, reject all hypotheses; otherwise, reject all H_i whenever $P_i \leq \frac{\alpha}{j}$.

The Student's t-test and its corrections

Benjamini–Hochberg correction

The Benjamini–Hochberg correction (“BH” for short) work as follows

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

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

The Student's t-test and its corrections

Benjamini–Hochberg–Yekutieli correction

The Benjamini–Hochberg–Yekutieli correction (also called the “BY”) is a refinement to work under positive dependence assumptions.

- ❶ Be k the largest number such that $P_k \leq \frac{k}{mc(m)}\alpha$;
- ▶ If the tests are independent or positively correlated, $c(m) = 1$;
 - ▶ Under arbitrary dependence, $c(m) = \sum_{i=1}^m \frac{1}{i}$;
 - ▶ In the case of negative correlation, $c(m)$ can be approximated by using the Euler's constant (γ).

$$\ln(m) + \gamma + \frac{1}{2m} \quad (6)$$

- ❷ Then, we reject the null hypothesis $H_{(1)}, \dots, H_{(k)}$.

Simulation Study

We evaluate the t-test and its corrections with respect to the power and Type I error rate through Monte Carlo simulation in different scenarios.

Setup

- ▶ Three treatments (groups)
- ▶ Group size: $n = 2, 3, 5, 10, 15, 20$
- ▶ Location parameter: $\mu = 0$
- ▶ Alternative hypothesis: $\mu_a = -6, \dots, -1, 1, \dots, 6$
- ▶ Scale parameter: $\sigma = 1, 2$
- ▶ Number of replications: $B = 5000$
- ▶ Sampling distribution: Normal, Logistic and Gumbel
- ▶ Significance level: $\alpha = 0.05$
- ▶ Programming language: R 3.3.2

Experimentwise type I error rate

We can estimate the experimentwise type I error rate ($\hat{\alpha}_{FWE}$) generating independent sample from null hypothesis and calculating the proportion of times that H_0 was rejected wrongly, that is,

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

Power

In this study we consider one treatment with different means and the others with the same, therefore, the empirical power of the test ($\hat{\tau}$) were obtained by

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

Ranking

For the experimentwise type I error rate ranking, was taking into account the

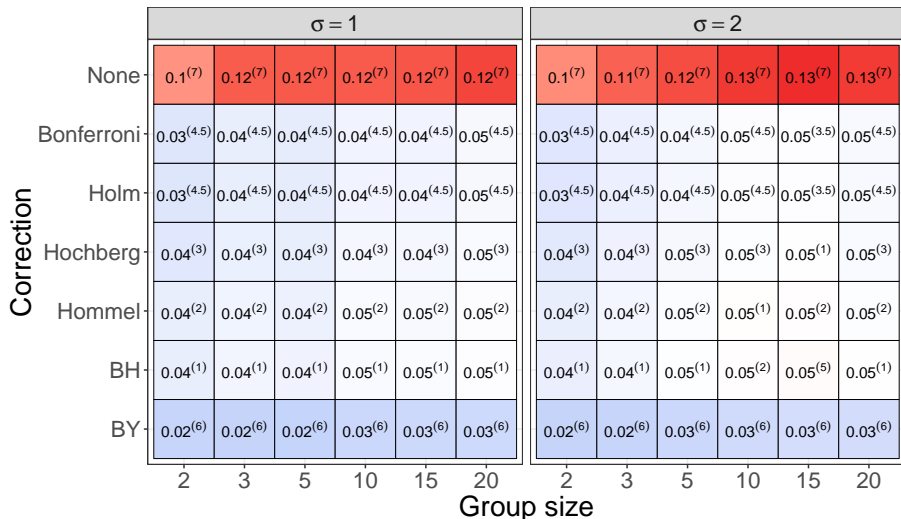
$$|\alpha - \hat{\alpha}_{FWE}|, \quad (7)$$

So the lowest ranking means a better correction.

For the power ranking, it was used the power itself, therefore the lowest ranking means a powerful test.

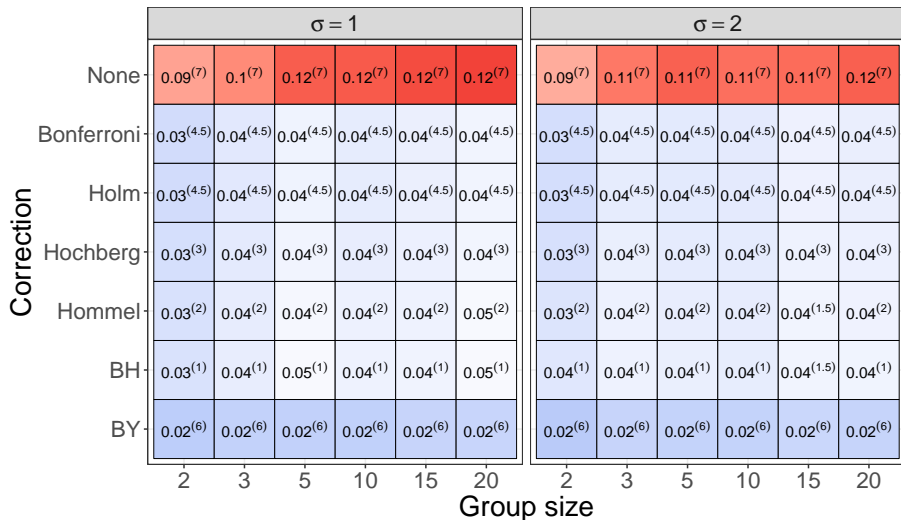
Results - Normal

Experimentwise Type I Error Rate



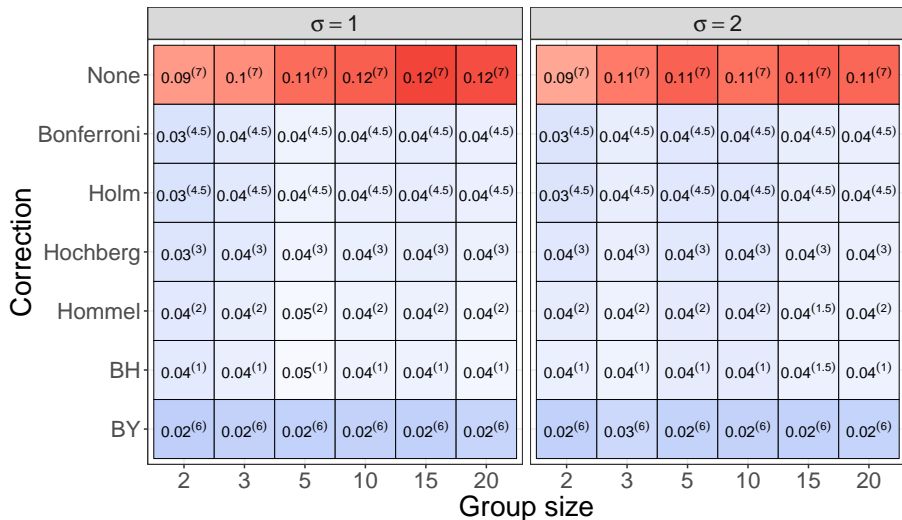
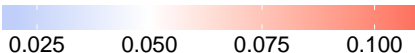
Results - Logistic

Experimentwise Type I Error Rate



Results - Gumbel


Experimentwise Type I Error Rate



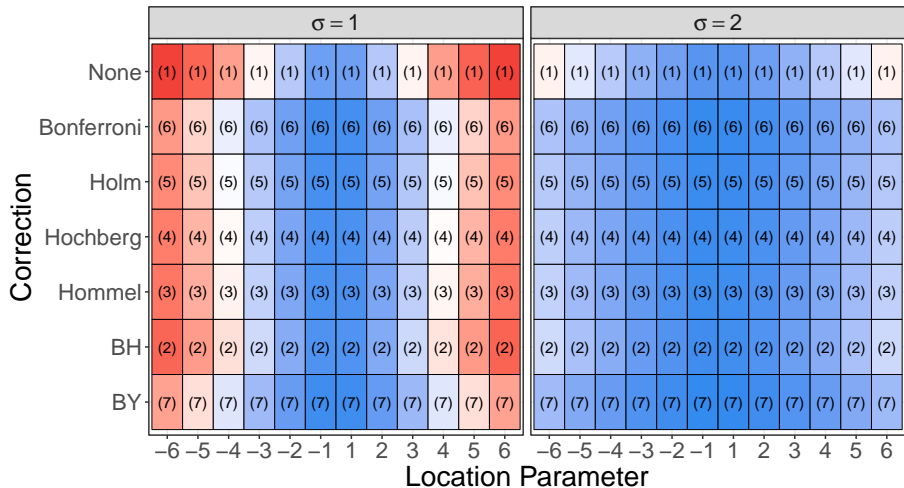
Results - Normal

$n = 2$

Empirical Power



0.00 0.25 0.50 0.75 1.00

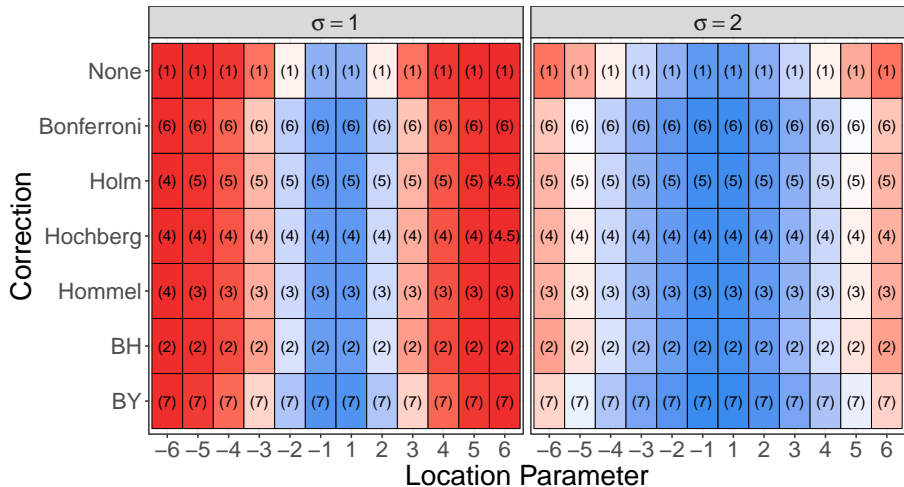


Results - Normal

$n = 3$

Empirical Power


0.00 0.25 0.50 0.75 1.00



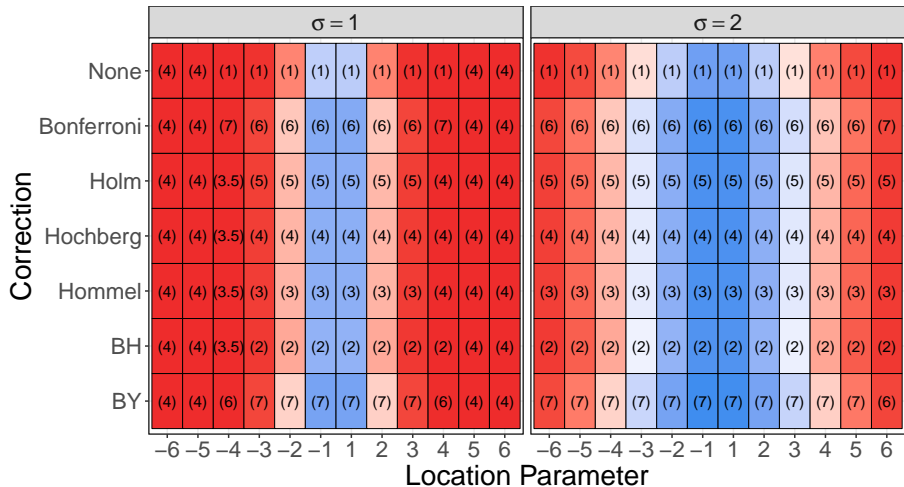
Results - Normal

$n = 5$

Empirical Power



0.00 0.25 0.50 0.75 1.00

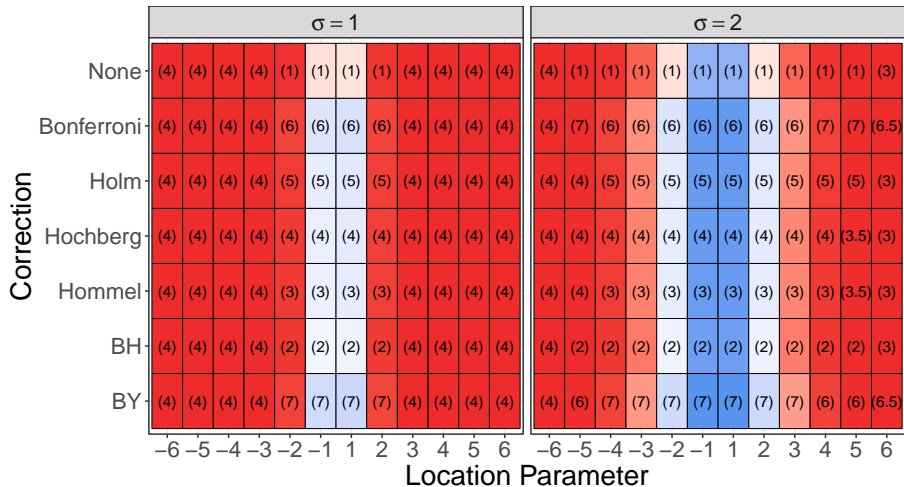


Results - Normal

$n = 10$

Empirical Power


0.00 0.25 0.50 0.75 1.00



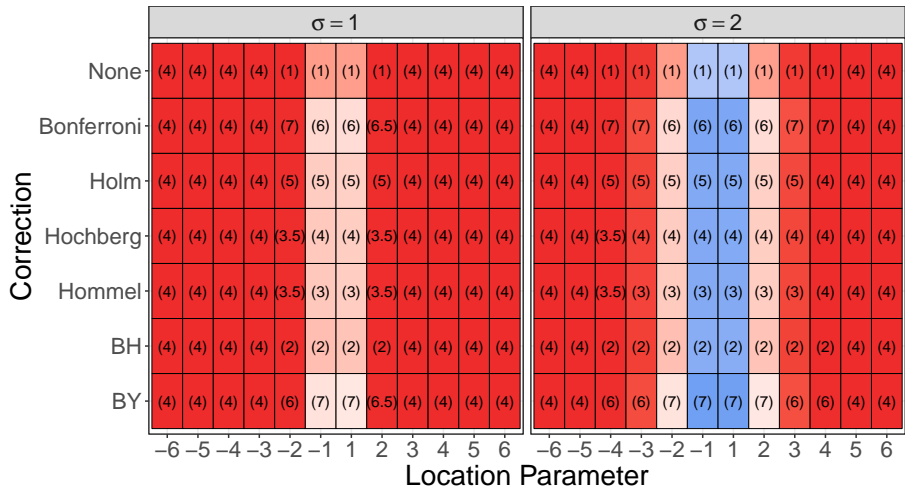
Results - Normal

n = 15

Empirical Power



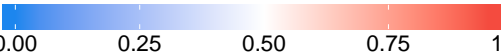
0.00 0.25 0.50 0.75 1.00



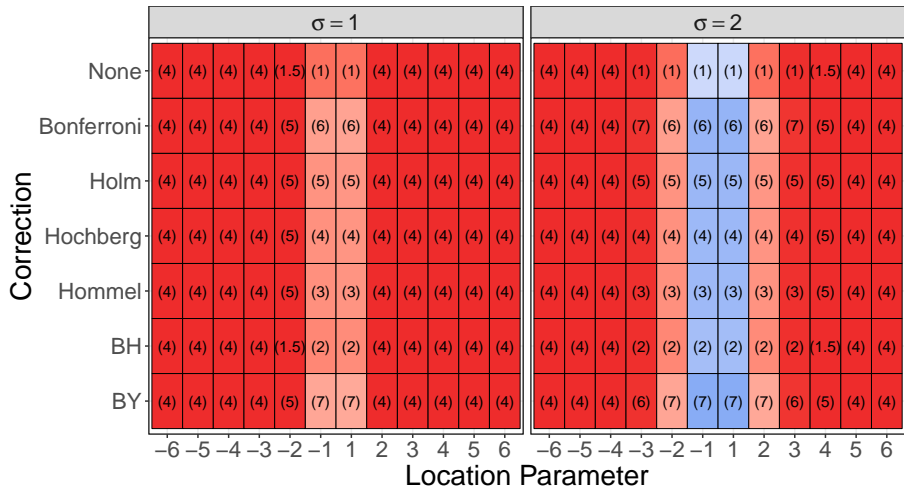
Results - Normal

n = 20

Empirical Power




0.00 0.25 0.50 0.75 1.00



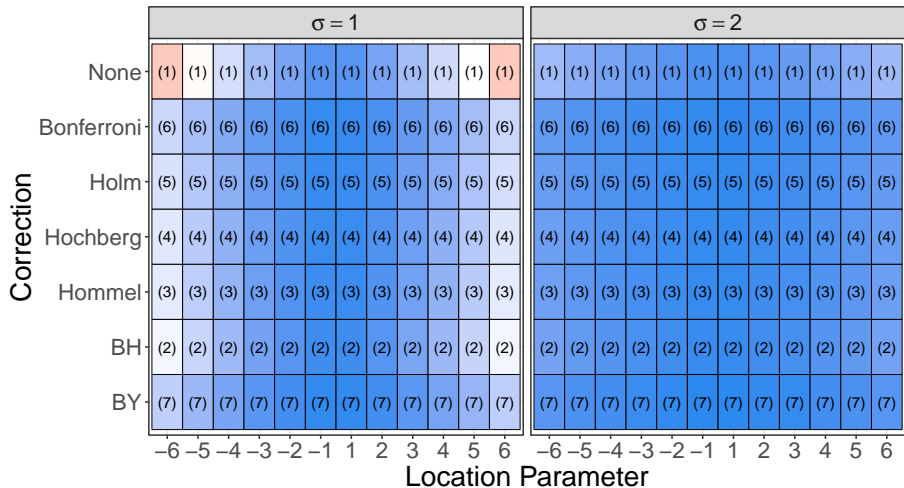
Results - Logistic

$n = 2$

Empirical Power




0.00 0.25 0.50 0.75 1.00



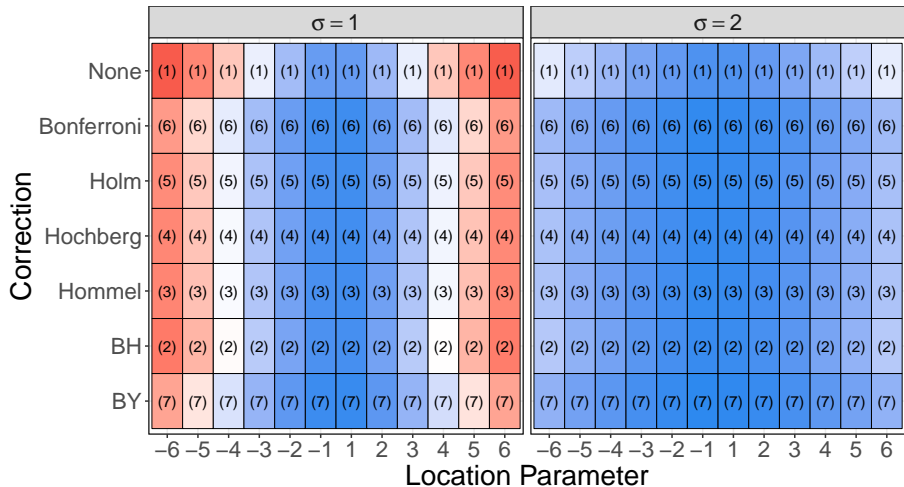
Results - Logistic

$n = 3$

Empirical Power




0.00 0.25 0.50 0.75 1.00



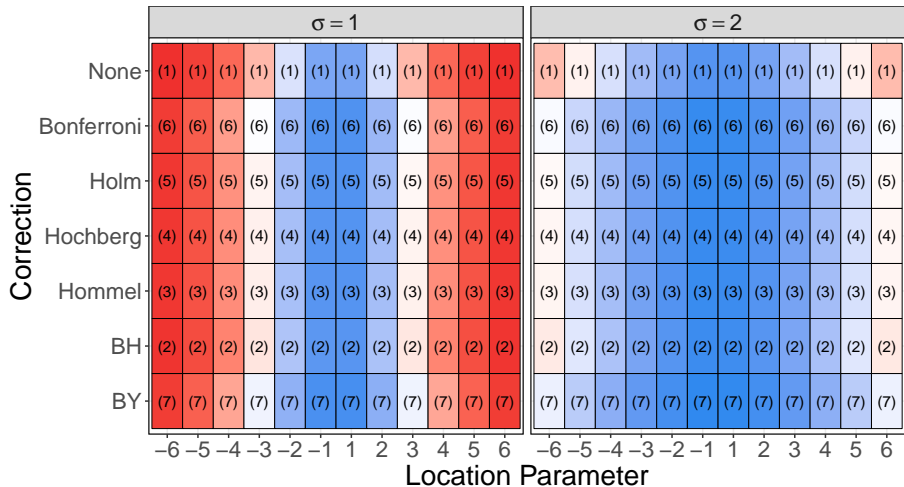
Results - Logistic

$n = 5$

Empirical Power



0.00 0.25 0.50 0.75 1.00

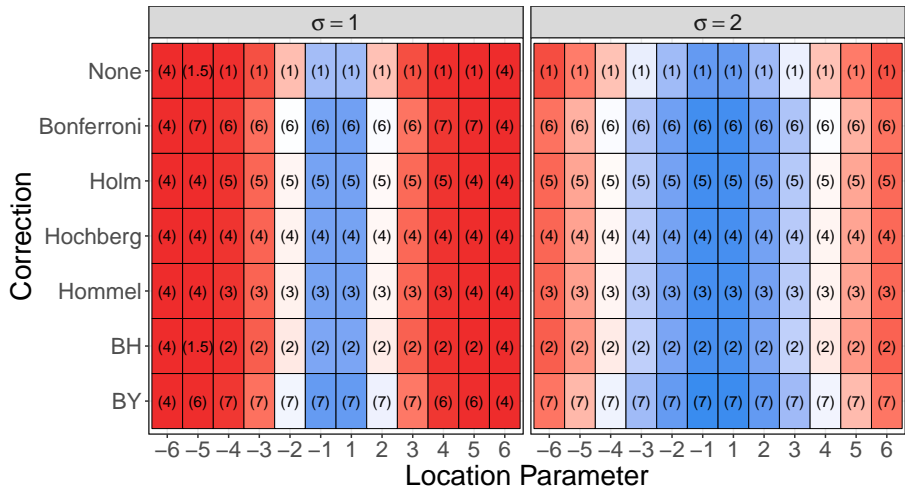


Results - Logistic

n = 10

Empirical Power


0.00 0.25 0.50 0.75 1.00



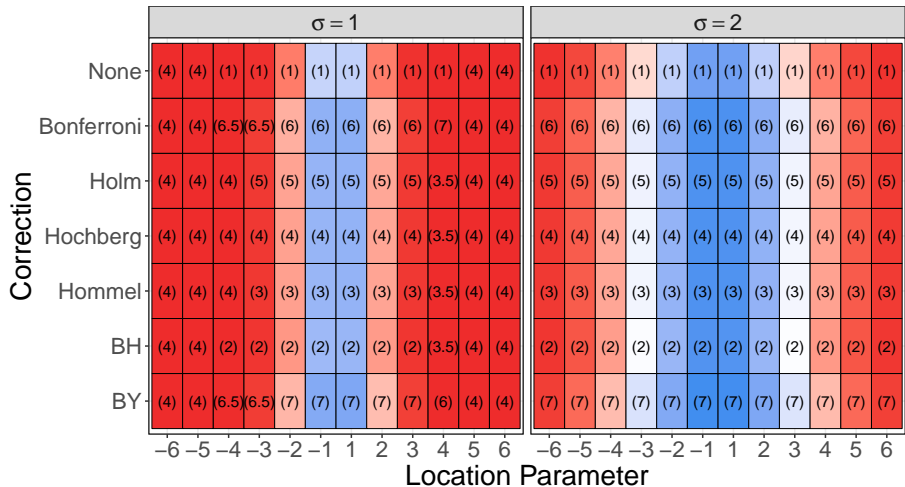
Results - Logistic

n = 15

Empirical Power




0.00 0.25 0.50 0.75 1.00



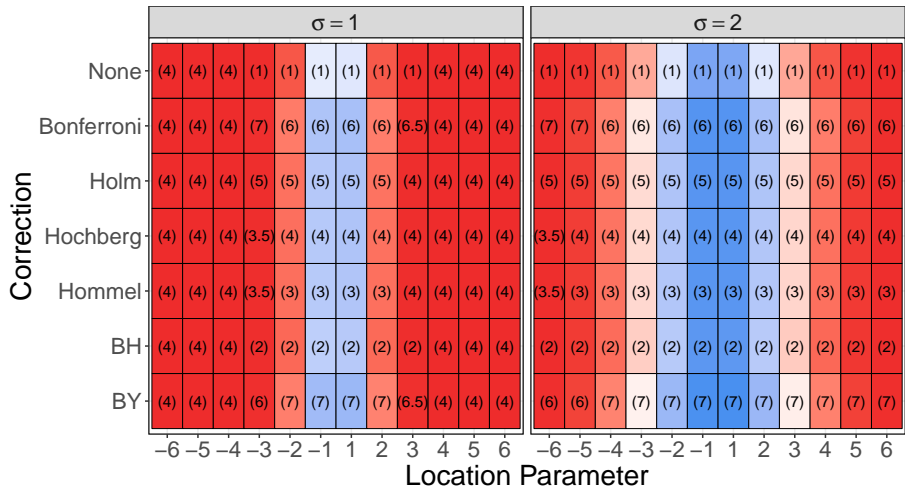
Results - Logistic

n = 20

Empirical Power




0.00 0.25 0.50 0.75 1.00



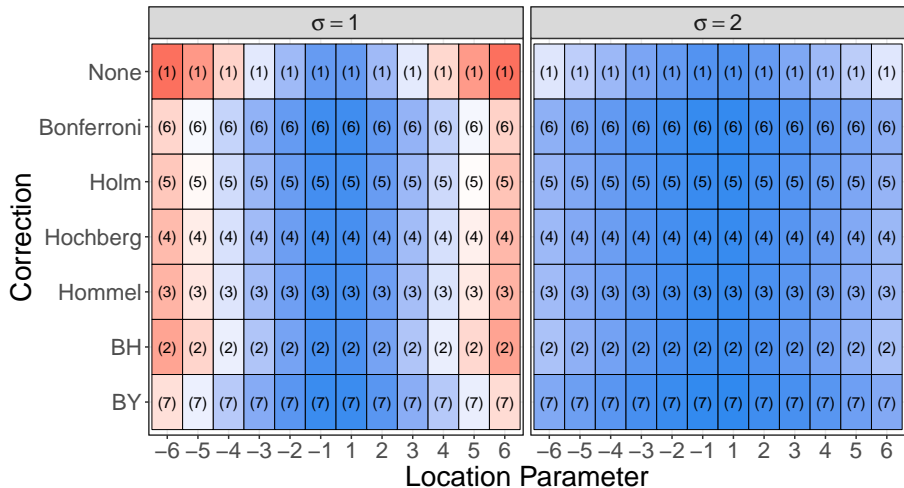
Results - Gumbel

$n = 2$

Empirical Power



0.00 0.25 0.50 0.75 1.00

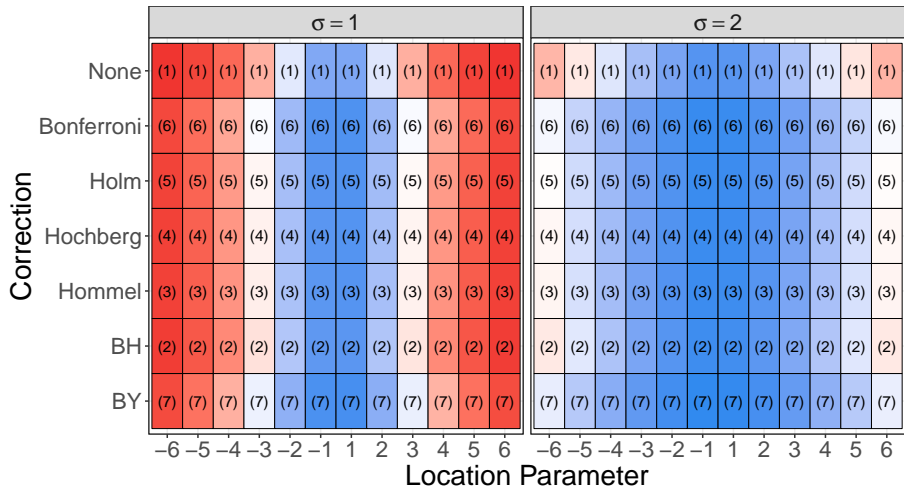


Results - Gumbel

$n = 3$

Empirical Power


0.00 0.25 0.50 0.75 1.00



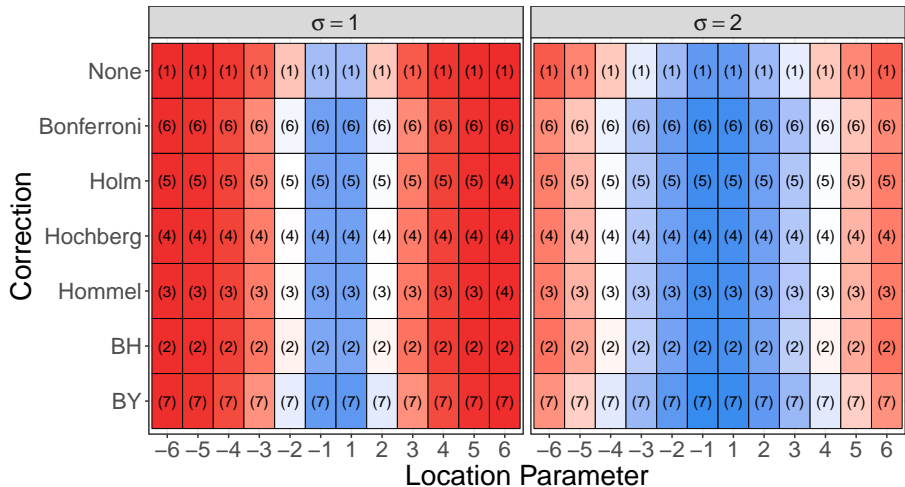
Results - Gumbel

$n = 5$

Empirical Power




0.00 0.25 0.50 0.75 1.00



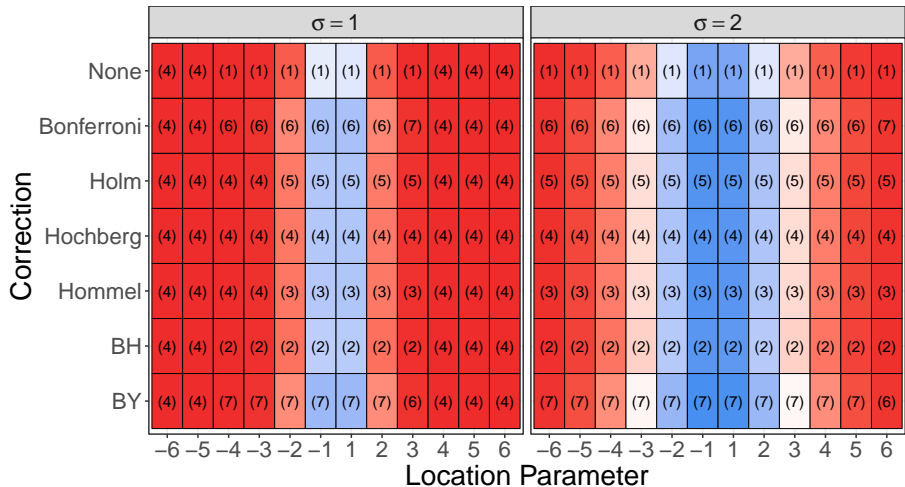
Results - Gumbel

$n = 10$

Empirical Power




0.00 0.25 0.50 0.75 1.00



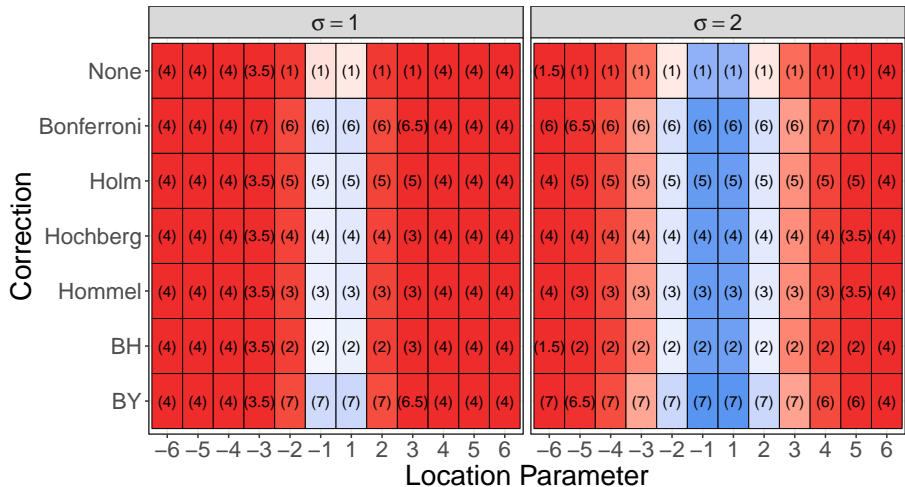
Results - Gumbel

n = 15

Empirical Power




0.00 0.25 0.50 0.75 1.00



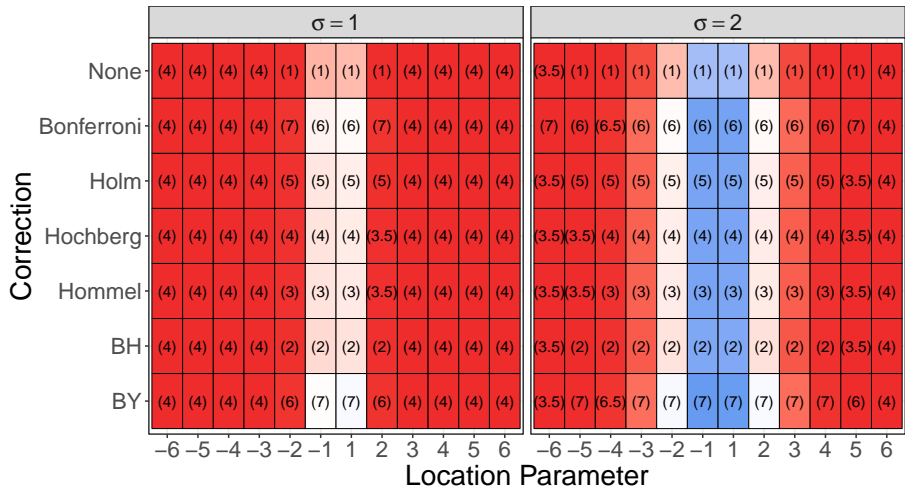
Results - Gumbel

n = 20

Empirical Power



0.00 0.25 0.50 0.75 1.00



Conclusion

- ▶ It was verified that when the sample size increases, the experimentwise Type I error rate of the corrections approaches the nominal level reasonably, even when the assumptions are not attended.
- ▶ As expected the t-test without correction was better than the others with respect to the power.
- ▶ In general, a bigger scale parameter provides least powerful test.
- ▶ The BH correction provides the best experimentwise type I error rate and also the second overall most powerful correction.

- ▶ BENJAMINI, Y.; HOCHBERG, Y. **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** Journal of the royal statistical society. Series B (Methodological), JSTOR, p. 289–300, 1995.
- ▶ BENJAMINI, Y.; YEKUTIELI, D. **The control of the false discovery rate in multiple testing under dependency.** Annals of statistics, JSTOR, p. 1165–1188, 2001.
- ▶ HOCHBERG, Y. **A sharper bonferroni procedure for multiple tests of significance.** Biometrika, Biometrika Trust, v. 75, n. 4, p. 800–802, 1988.
- ▶ HOLM, S. **A simple sequentially rejective multiple test procedure.** Scandinavian journal of statistics, JSTOR, p. 65–70, 1979.

- ▶ HOMMEL, G. **A stagewise rejective multiple test procedure based on a modified bonferroni test.** Biometrika, Biometrika Trust, v. 75, n. 2, p. 383–386, 1988.
- ▶ R Core Team. **R: A Language and Environment for Statistical Computing.** Vienna, Austria, 2016. Disponível em: <<https://www.R-project.org/>>.
- ▶ STUDENT. **The probable error of a mean.** Biometrika, JSTOR, p. 1–25, 1908.
- ▶ WEISSTEIN, E. W. **Bonferroni correction.** Wolfram Research, Inc., 2004.