Multiple Comparison Methods Simulation

Group 5, Project 2: Zelos Zhu, Quinton Neville, Soohyun Kim

2/14/2019

Objectives

In disease etiology research, patients are often assessed by multiple outcomes/biomarkers. In a two-arm randomized clinical trial with K outcomes, conducting multiple two sample t-tests to determine treatment effects on individual outcomes can result in Type I error inflation. To control this inflation and attempt to preserve Power of the individual tests, we set a desired level $\alpha_{0.05}$ and chose between a variety of multiple-comparison adjustment methods. The objective of this simulation was to compare the control of family-wise error rate (FWER) and preservation of Power for the following four methods:

- Bonferroni method
- Holm's step-down approach
- Hochberg's step-up approach
- Resampling methods

Exploration of FWER and Power dynamics included the variation of these key parameters:

- The number of outcomes (K)
- Between-outcome correlations (ρ)
- Effect size (difference in $\mu_t - \mu_c$)

Methods Studied:

Bonferroni method: Reduce the significance level α (for individual tests) by $\alpha_{adj.} = \alpha/M$

Holm's step-down approach: Sort the M p-values such that $p_1 > p_2 > \cdots > p_M$. Begins by adjusting the smallest p-value p_M by $P_{M,adj} = \min\{M * p_M, 1\}$, and then adjust the subsequent p_j with $j = M-1, M-2, \ldots, 1$ by $P_{j,adj} = \min\{\max(j * p_j, (j+1) * p_{j+1}, ..., M * p_M), 1\}$. Reject the null hypothesis sequentially until a $P_{j,adj} > \alpha$.

Hochberg's step-up approach: Adjustment begins with the largest p-value, $p_{1,adj} = 1 * p_1$, and adjust the subsequent p-values by $P_{j,adj} = \{\max(j * p_j, (j-1) * p_{j-1}, ..., 1 * p_1)\}$. Reject the null hypothesis when a $P_{j,adj} < \alpha$ and reject the subsequent hypothesis.

Holm and Hochberg are step-wise Bonferroni methods.

Resampling methods Define $p_{\min} = \min\{p_1, p_2, ..., p_M\}$ be the minimum p-values from the M tests, and then generate the distribution of $p_{\min}$ using re-sampling data under the Null distribution.

Suppose $(x_{i,1}, ..., x_{i,k}, y_{i,1}, ..., y_{i,k})$ is a random sample from a two-arm clinical trial with k outcomes, where $(x_{i,1}, ..., x_{i,k})$ is the k outcomes from the treatment group, while $(y_{i,1}, ..., y_{i,k})$ is that from the control group. One can generate the distribution of $p_{\min}$ under the Null following a resampling procedure.

1. For each outcome k, we center the responses by their group specific means. i.e. $y^*_{i,k} = y_{i,k} - \bar{y}_k$ and $x^*_{i,k} = x_{i,k} - \bar{x}_k$, where $\bar{y}_k$ and $\bar{x}_k$ are sample means of the kth outcome in treatment and control groups, respectively.
2. Resample $y^*_{i,k}$ and $x^*_{i,k}$ separately, and apply the two-sample t-test to the resampled data, and calculate the $p_{\min}$ accordingly. Repeat this procedure ℓ times, and you obtain ℓ resampled $p_{\min}$'s.
3. Let $q^1_{\min}, q^1_{\min}, ..., q^\ell_{\min}$ be resampled $p_{\min}$s. The P-value of $p_{\min}$ can be calcuated by the proportion of resampled $p_{\min}$s that are smaller than the observed $p_{\min}$. i.e.

$$p - value = \frac{1}{\ell} \sum_{j=1}^{\ell} I\{q^j_{\min} < p_{\min}\}$$

Scenarios Investigated:

At baseline, we assume that for each K outcome, the treatment and control arms follow an approximately standard multivariate normal distribution with equal variance and a fixed sample size of $N = 30$, respectively. Additionally, as we vary the parameters of interest, we further fix $K = 20$, $\rho = 0$, and $\mu_t - \mu_c = (0, 1)$ corresponding to the null (H_0) and alternative (H_A) hypotheses, respectively. In order to study how paramter variation affects FWER and Power:

- (K) – The number of outcomes were varied on a 10 to 30 grid, under the null and alternative hypotheses for FWER and Power, respectively.
- (ρ) – The between-outcome correlations were varied on a 0 to 0.9 grid, and assumed equal and fixed between all pairs of treatment and control outcomes, respectively. Additionally, simulations were performed under the null and alternative hypotheses for FWER and Power, respectively.
- $(\mu_t - \mu_c)$ – The effect size was varied on a 0.5 to 1 grid, approximately corresponding to the number of standard deviations from the mean (under approximately standard normal assumptions). In this case, FWER is incalculable as the null hypothesis is synthetically not true; instead Power is calculated under the alternative hypothesis.

Simulation Generation:

Utilizing the `mvtnorm` library, approximately standard multivariate normal samples of size $N = 30$ were generated for K outcomes, where $\mu_{i,j}$ were constant and $\sigma^2_{i,j} \sim U(0.8, 1.2)$ using Monte Carlo methods. These data were repeatedly randomly simulated over the course of 1000 iterations. Covariance matrices for multivariate normal sample generation were manipulated using the `MATRIX` library, ensuring symmetry and utilizing an algorithm to find the closest positive definite matrix with respect to the random variance. Lastly, the resampling method was performed using 100 bootstrap iterations for simulation speed, although 1000 would have been optimal.

Performance Measures:

FWER for each unadjusted and adjusted method was calculated by counting the number of simulations where *at least one* of the p-values for the K-outcomes was significant (thus rejected under the null hypothesis) and dividing it by the total number of simulations run (1000). Conversely, Power was calculated by counting the number of K individual t-tests with significant p-values, under the alternative hypothesis, and dividing by the total number of individual tests and simulations run $(K \cdot 1000)$.

Simulation Results:

Varying Number of Outcomes

Under the null hypothesis, it was expected to see FWER be the same for both the Bonferroni and Holm adjustment methods while Hochberg's would be very similar. This was shown to be true according to our simulation. However, the resampling method consistently had a higher FWER, especially at low number of outcomes (K). Interestingly, we observed a "dip" trend among all adjustment methods as K increased from 10 to 20, and return to near original FWER state as K went from 20 to 30. As K increased, the *difference* in FWER between the resampling and other methods decreased. (See Figure 1)

As K increased, we saw a decrease with regards to power for the Bonferroni and Holm methods while the Hochberg and resampling methods were robust to changes to K. Overall, Bonferroni had the worst power preservation while the other three methods exhibited much better preservation with the resampling method performing the best. (See Figure 2)

Varying Correlation

Considering the effect that correlation (ρ) demonstrates with respect to FWER control, under the Null Hypothesis ($\mu_t = \mu_c$), we fixed the correlation between all $K = 20$ observations, and varied ρ from 0 (uncorrelated) to 0.9 (highly correlated). Simulation results indicate that all p-value adjustment methods remain near or below 0.05 even as ρ increases. Furthermore, as ρ increases, we see a reduction in the FWER for all adjustment methods, with the resample method eliciting the greatest decrease in magnitude. It was also interesting to note that the FWER for performing multiple tests with unadjusted p-values decreased from more than 60% to less than 20% as ρ increased. (See Figure 3)

Next, the same simulation as performed under the Alternative Hypothesis ($\mu_t \neq \mu_c$) to observe the effect of correlation on the Power of each method, again with $K = 20$, $rho = (0 - 0.9)$, and an effect size of 0.5 and 1 for visualization. At low effect size ($\mu_t - \mu_c = 0.5$), the Bonferroni, Hochberg, and Holm demonstrated the lowest Power to detect this difference in means; although Hochberg and Holm methods showed an approximate additive 10% increase in Power as ρ increased. Conversely, the resampled adjustment and unadjusted methods demonstrated little change in Power as correlation increased, reamining relatively stagnant at ~50%. Considering a larger effect size ($\mu_t - \mu_c = 0.5$), Power for all methods increased significantly however the trend in Power with respect to correlation was less clear. Speculatively, it appeared as though Power increased for all methods as ρ initially increased from 0 to 0.4, then decreased or remained stagnant as correlation increased further. This may imply that for larger effect size, Power increases for low to moderately correlated data, but is reduced or unnafected for highly correlated data. (See Figure 4)

Ultimately, we observed that the resampling adjustment method led to relatively good FWER control as correlation increases, while maintaining much higher Power than the other adjustment methods (as high, or higher than the unadjusted which has poor FWER control), and was robust to correlated data. However, the Hochberg and Holm methods ellicited improved Power for increasingly correlated data, while Bonferroni was likewise robust to correlation change at smaller effect sizes. Additionally, a similar simulation with Monte Carlo generated random correlations between the K observations was considered but results are omitted here, as there were not significantly different results or noteworthy trends compared to the fixed ρ simulations.

Varying Effect Size

Unsurprisingly, Power of the study increases as effect size increases. By definition Power is the probability to reject the null when the alternative is true. If effect size increases, difference in means in this case, it should generally easier to detect these differences and that reflects in our data. (See Figure 5)

Conclusion

Overall the Bonferroni, Holm, and Hochberg readjustment methods performed the best at controlling FWER, robust to changes in number of outcomes, K, and between outcome correlations ρ. The resampling method consistently performed the best at preserving power robust to outcomes, correlation, and effect size, while having only a marginally worse control on FWER. Although each adjustment method may be advantageous in a particular scenario, with respect to FWER and power we would recommend the resampling method as the most generally optimal adjustment method.

References

1 Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6:65–70.

2 Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75:800–802.

3 Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Statistics in Medicine. 1990;9:811–818

4 Westfall PH, Young SS. Resampling-based multiple testing: Examples and methods for p-value adjustment. New York: Wiley; 1993.

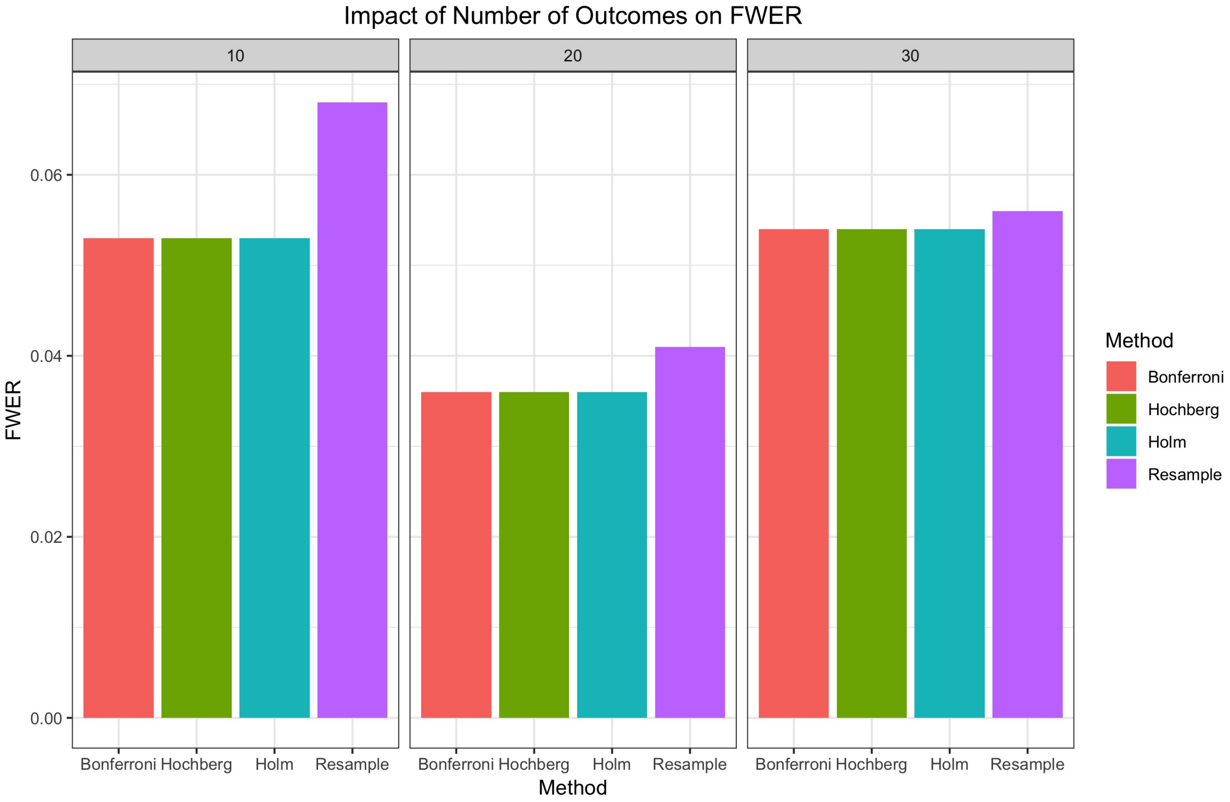

Figure 1: Impact of Number of Outcomes on FWER

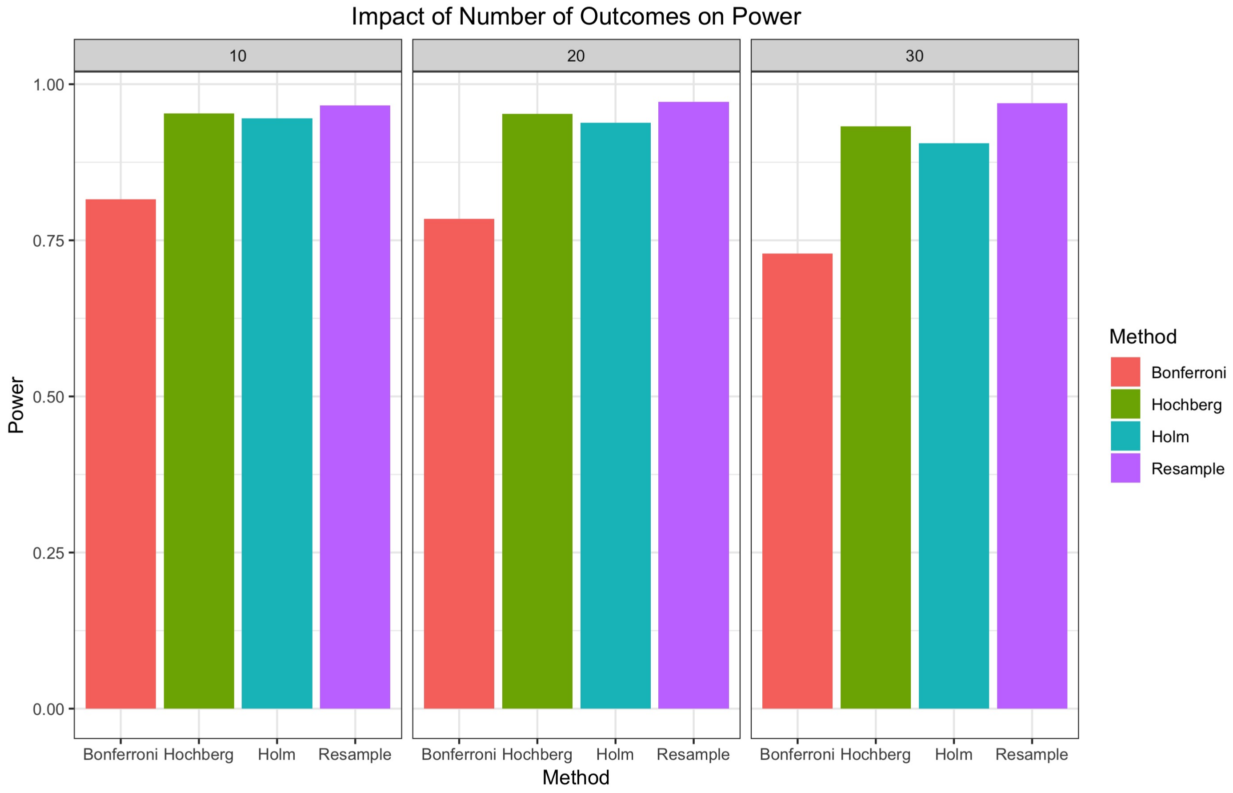

Figure 2: Impact of Number of Outcomes on FWER

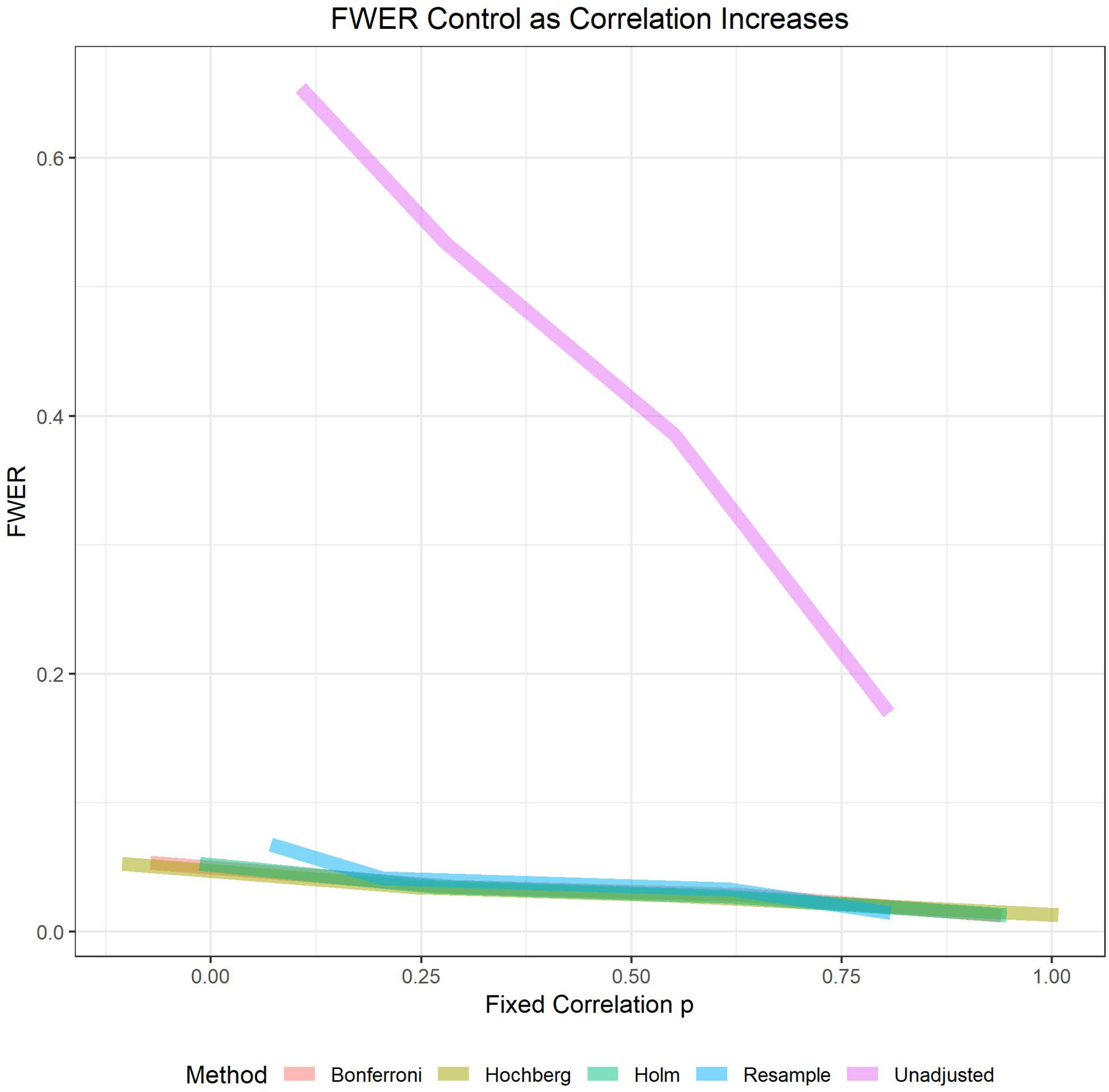

Figure 3: Impact of Correlation on FWER

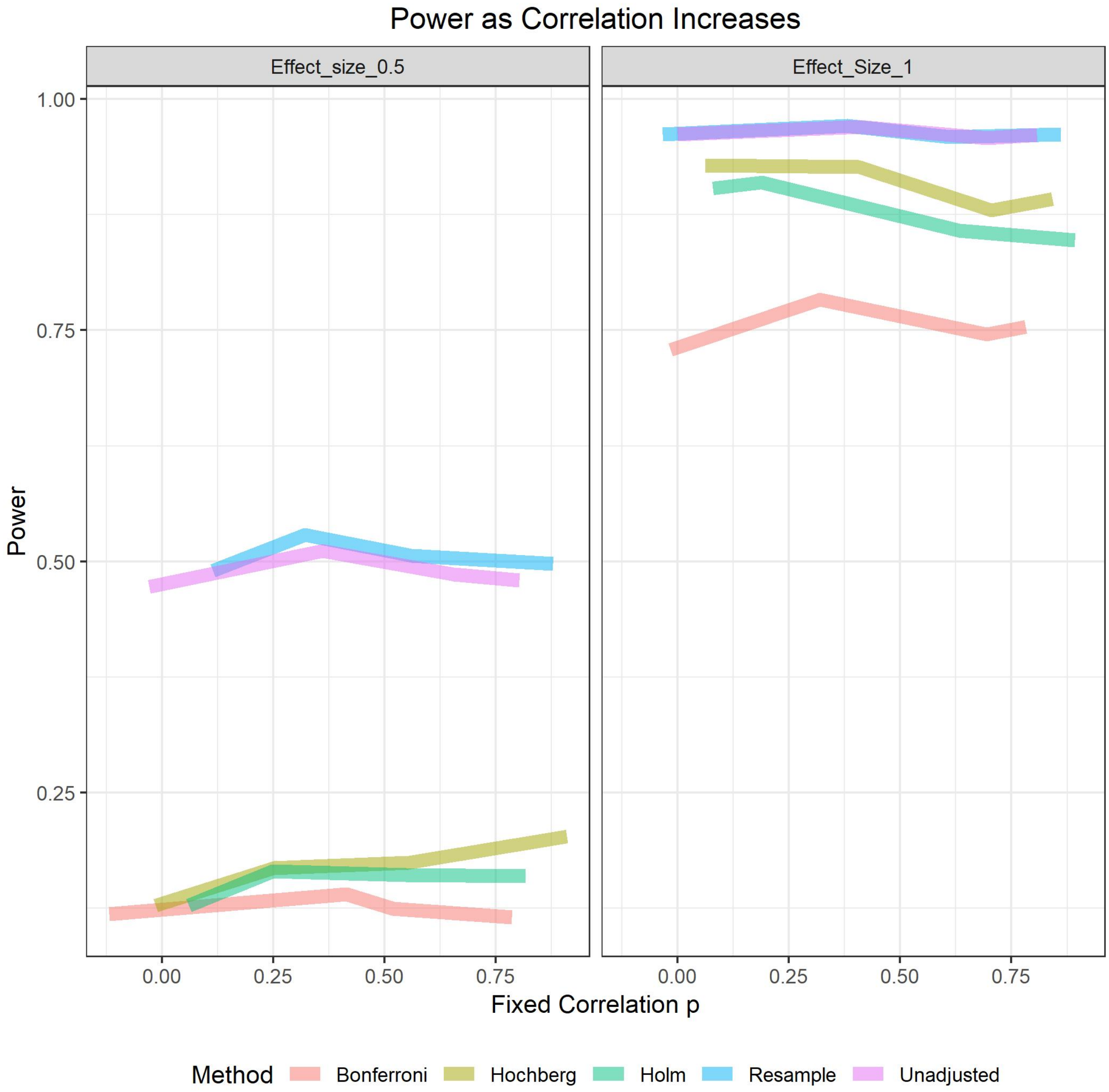

Figure 4: Impact of Correlation on Power

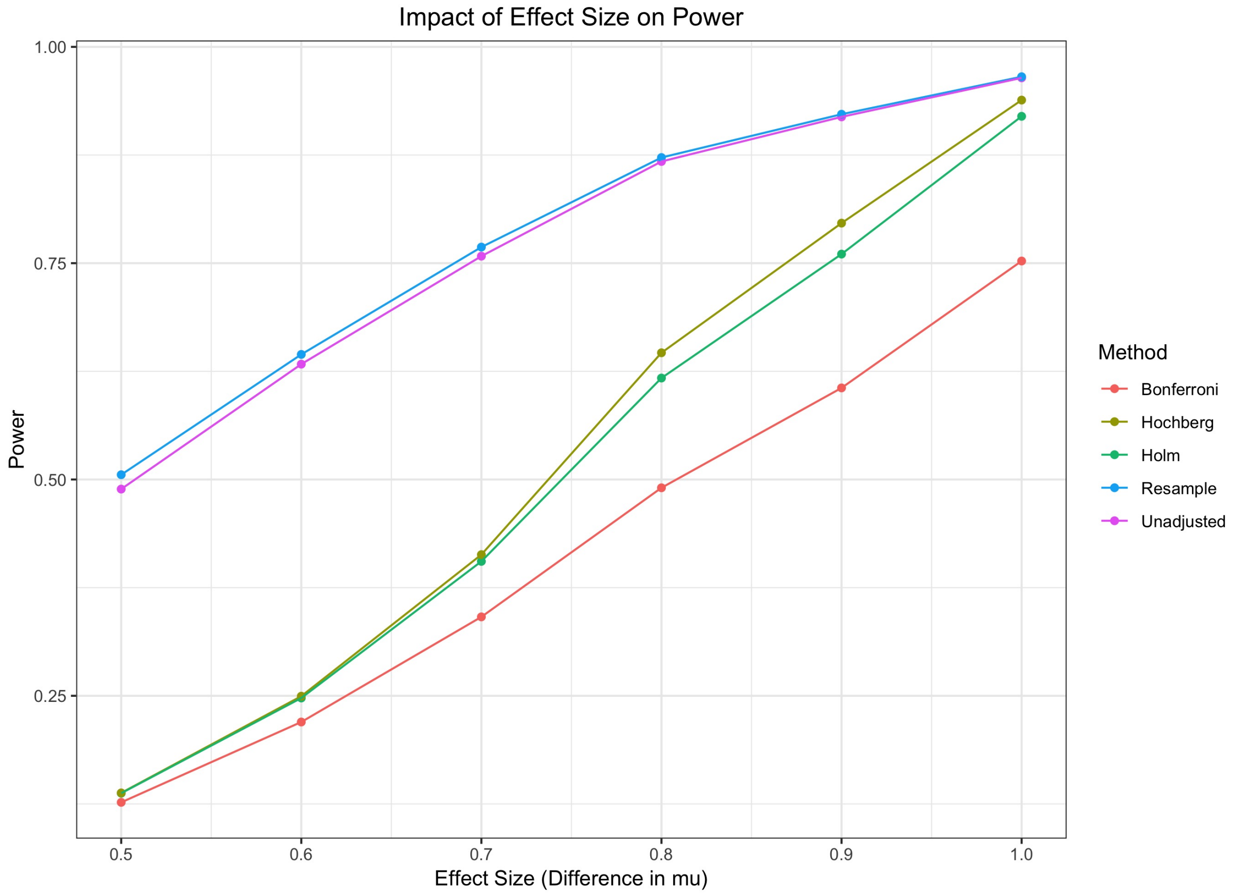

Figure 5: Impact of Effect Size on Power