Multiple Hypothesis Testing - Report

I. Introduction

The idea of multiple hypothesis testing is that a large number of hypotheses are tested simultaneously. However, this raises an issue of having a high proportion of false positives. The more hypotheses are tested, the more likely false rejections will occur. Take, for instance, in the field of genomics, where a researcher could be testing 10,000 hypotheses at the same time.[1] If they are testing at a 0.05 level of significance, they would find 500 significant hypotheses, even if they are not significant. Additionally, the researcher would not be able to distinguish between the true significant hypotheses and the false positives. Moreover, there will likely be hypotheses that do not seem statistically significant when they actually are, hence leading to a proportion of false negatives. Thus, there is an issue of having a high rate of false positives and/or false negatives, which are called Type I and Type II errors. How can we limit the number of false positives that arise in multiple hypothesis testing? Is there a way to increase the power of a test while minimizing its Type I error rate? What is the best way to maximize a test's sensitivity and specificity?

II. Background

There are several methods to control Type I and Type II error rates. The error rates that are approximated in this simulation are defined below.

i. Family-Wise Error Rate (FWER)

The FWER is defined as the probability of exhibiting at least one Type I Error:

$$\text{FWER} = \Pr(N_{1|0} \geq 1)$$

where N1|0 refers to the number of falsely rejected null hypotheses, H_0, when the null is actually true (i.e. a false positive).

Another way to estimate error is by estimating the False Discovery Proportion (FDP), which is defined as the proportion of incorrectly rejected hypotheses:

$$\text{FDP} = \begin{cases} \frac{N_{1|0}}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

where R is the total number of rejected hypotheses. Benjamini and Hochberg suggested a way to control the expectation of the FDP:

ii. False Discovery Rate (FDR)

The FDR is defined as the expected proportion of Type I errors:

$$\text{FDR} = E(\text{FDP})$$

Dudoit et al[2] and Genovese and Wasserman[3] suggested a way to control the tail probability of the FDP:

iii. False Discover Exceedance (FDX):

$$\text{FDX} = \text{tFDP}(c) = \Pr(\text{FDP} > c)$$

In other words, the FDX aims to control the probability of the proportion of false positives being greater than a certain threshold, *c*. In the analysis that follows, this threshold is taken to be 0.1.

Before discussing the different ways to control the aforementioned error rates, the general statistical framework will be defined.

In this model, there are *m* random variables (i.e. hypothesis tests) being replicated *n* times. For each of the *m* variables, there is a null hypothesis, H_0. H_0 is true if the stated model holds. Whether or not H_0 is rejected depends on the p-value that arises from the test statistic. The p-value is defined as:

$$p = \Pr(|T| > |t| \mid H_0 \text{ is true})$$

where *T* is the test statistic and *t* is the observed value. The *m* random variables are used to create the test statistics. A two-sided test is used to determine the p-values that correspond to the test statistics. Note that there are *m* p-values, one for each of the *m* test statistics.

The goal of multiple testing is to accurately estimate the set of true hypotheses while at the same time, minimizing the Type I error rate. One can think of this like a balancing act – performing enough rejections to gain information about statistical significance while also being cautious of having too many false positives. Rejecting more hypotheses gives researchers more statistically significant variables. However, if the method is not conservative enough, then many of the rejections will be false positives, and hence the conclusions cannot be trusted with a high level of confidence.

Many of the methods that minimize $N_{1|0}$ do so by adjusting the p-values. Equivalently, the threshold, or cut-off, that the p-values are compared to when determining rejection are adjusted. There are generally three types of adjustments:

1. **One-step:** This adjustment does not depend on the data. The p-values are compared to a cut-off level that is determined before the data is analyzed. Usually, this cut-off only depends on either the significance level, α, or *m*.
2. **Step-down:** For step-down, the cut-off is adjusted based on the rank of the p-value. The p-values are ranked from smallest to largest. First, the smallest p-value is compared to the threshold. If it is smaller than the threshold, it is rejected, and the next p-value is used for the comparison. This repeats until the p-value is larger than the cut-off. At this point, that p-value, along with the remaining p-values, are not rejected.
3. **Step-up:** Step-up is similar to step-down in the sense that the adjusted cut-off is again based on the rank of the p-value. However, this time, the p-values are ranked from largest to smallest, and the largest p-value is the first to be compared to the threshold. If it is larger than the cut-off, the p-value is not rejected and the next largest p-value is used. This continues until the p-value

is smaller than the cut-off. When this happens, that p-value along with the remaining ones are rejected.

In summary, the threshold is ideally large enough so that more tests are rejected while also ensuring that the error rate (i.e. proportion of false positives) is at most some predetermined level α (hence the test has high power).

This leads to the various methods for controlling the error rate:

Controlling the FWER:

- Bonferroni: this is a one-step method, where the cut-off is predetermined at $\frac{\alpha}{m}$.
- Step-down Holm[4]: Holm aimed to improve the Bonferroni correction (which is very conservative) by adjusting the cut-off with respect to the index of each p-value: $\alpha_j = \frac{\alpha}{m-j+1}$ for the j[th] p-value.
- One-step Sidak[5]: this is another one-step method, but unlike Bonferroni, the cut-off is predetermined at $1 - \sqrt[m]{1-\alpha}$.
- Step-down Sidak[6]: for this method, the cut-off with respect to the index of each p-value is $\alpha_j = 1 - \sqrt[m-j+1]{1-\alpha}$ for the j[th] p-value.
- Step-up Hochberg[7]: this method uses the same cut-off as Step-down Holm ($\alpha_j = \frac{\alpha}{m-j+1}$), but instead, it is used in a step-up approach rather than a step-down approach.

While the FWER is useful when the number of hypotheses is rather small, this error rate has rather lower power when *m* is large. Thus, in cases where *m* is large, one may want to control the FDR instead.

Controlling the FDR:

- Benjamini-Hochberg[8]: this is a step-up procedure, where the cut-off is $\alpha_j = \frac{j\alpha}{m}$ for the j[th] p-value. (Notice how this is very similar to the Bonferroni correction, but now the cut-off depends on the index of the p-value.)
- Step-Down Benjamini-Liu[9]: this step-down procedure fixes the cut-off at $\alpha_j = 1 - [1 - \min\left(1, \frac{m}{m-j+1}\alpha\right)]^{1/(m-j+1)}$ for the j[th] p-value.
- Benjamini-Yekutieli[10]: this is a resampling-based method. Say there are *k* ordered p-values. Then $p_{(k)}$ are the thresholds. When the data is resampled, the resampled p-values that are below each $p_{(k)}$, i.e. the $r_\beta(p_{(k)})$, are then the (1-β) quantile of $r(p_{(k)})$. Then, at each threshold, a $Q^*(p_{(k)})$ is calculated:

$$Q(p_{(k)}) = \begin{cases} \frac{r(p_{(k)})}{r(p_{(k)})+k-p_{(k)}m} & \text{if } p_{(k)}m \le k - r_\beta(p_{(k)}) \\ 1 & \text{otherwise} \end{cases}$$

Then for each *k*, let $k_\alpha = \max_k\{Q^*(p_{(k)}) \le \alpha\}$. Then, the cut-off is set at $p_{(k\alpha)}$ for the k[th] p-value.

For instances when the FDP is not concentrated around its mean, which can happen when the variance is relatively high, controlling the FDX may be more helpful, as it better protects against more extreme situations (hence controlling the tail-probability).

Controlling the FDX:

- Augmentation[2]: first, the FWER is controlled via the Bonferroni approach, which then results in a set of rejected hypotheses. If this approach led to at last one rejection, then define

$$k_n(c,\alpha) = \max\{j \epsilon \{0, \dots, m - S\} \text{ such that } \frac{j}{j+S} \leq c\}$$

  where c is the threshold (taken to be 0.1 in the following analysis) and S is the number of rejected hypotheses after applying the Bonferroni correction. If this new k indicates that more hypotheses should be rejected, then reject those hypotheses.
- Step-down Lehmann-Romano[11]: this step-down method uses a threshold $\alpha_j = \frac{[(cj)+1]\alpha}{m+(cj)+1-j}$ for the jth p-value.

III. Simulation Study

In the Farcomeni paper, n=1000 normal data sets are simulated, with three different numbers of hypotheses: m = {100, 5000, 100000}[12]. The proportion of true hypotheses is held constant at 0.9 (in other words, 10% of the hypotheses in each simulation are false). Consequently, there are three different simulated setups: m = 100 with 90 true hypotheses, m = 5000 with 4500 true hypotheses, and m = 100000 with 90000 true hypotheses. The test statistics are random normal variables with the alternative means sampled from a random uniform (0,5). The variance is known and equal to 1. Then, p-values are generated from these test statistics. The various error-controlling methods are then applied to the raw p-values, resulting in a vector of adjusted p-values. The adjusted p-values are used to determine whether or not to reject the hypothesis. These rejections are compared to the rejections resulting from the raw p-values. The adjusted and raw p-values are compared by calculating the expected number of false positives, the expected number of false negatives, the FWER, the FDR, the FDX, and the FNR (false negative rate, which is calculated as the expected proportion of false negatives).

In this paper, the above procedure is generally followed, but instead of only testing a 0.9 proportion of true hypotheses, a 0.5 proportion is also tested in this simulation. The following packages were used in order to implement the error corrections:

Table 1: R Packages Used to Implement Error Control Procedures

Method	R Package	Function
Control of FWER		
Bonferroni	multtest	mt.rawp2adjp
Step-down Holm	multtest	mt.rawp2adjp
One-step Sidak	multtest	mt.rawp2adjp
Step-down Sidak	multtest	mt.rawp2adjp
Step-up Hochberg	multtest	mt.rawp2adjp
Control of FDR		
Benjamini-Hochberg	multtest	mt.rawp2adjp
Step-down Benjamini-Liu	mutoss	BL
Benjamini-Yekutieli	multtest	mt.rawp2adjp
Control of FDX		
Augmentation	mutoss	augmentation
Step-down Lehmann-Romano	someKFWER	kfwelR

Note that the original code for the function used for the Step-Down Lehmann-Romano correction does not return a vector of adjusted p-values. Instead, it returns the number of rejections. In order to return the vector of adjusted p-values, each raw p-value was multiplied by the inverse of its corrected threshold, α_j, j = 1, …, m. (See Appendix for the edited R code.)

Then, the simsalapar package was used in order to perform the *n* simulations, where the following variables were defined in order to setup the simulation:

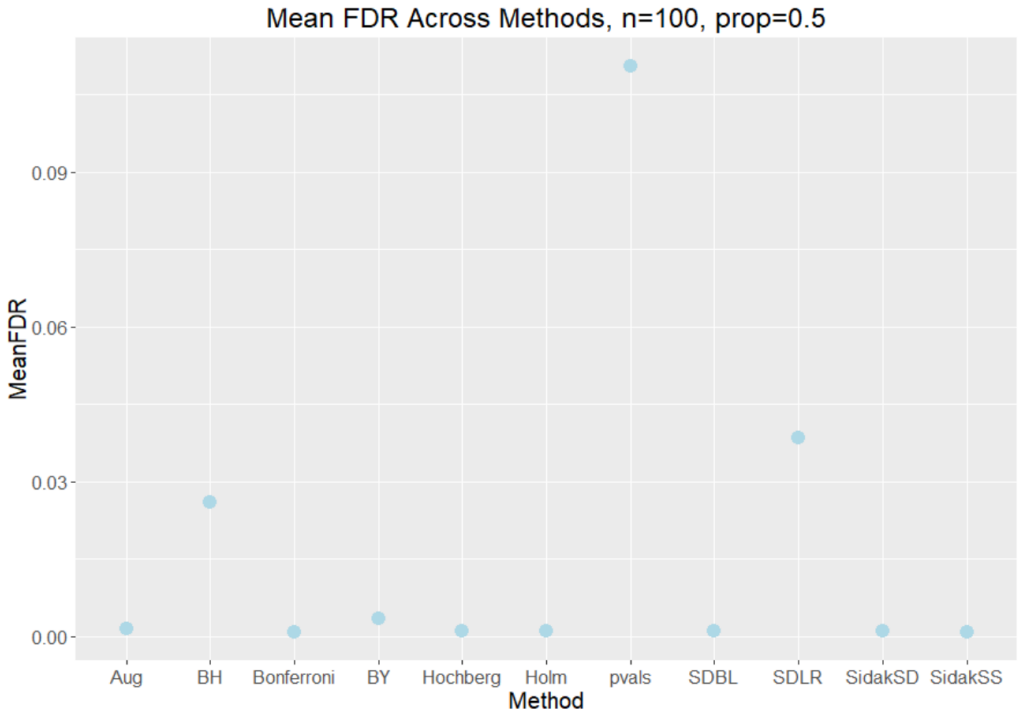
- n.sim = list(type = "N", expr = quote(N[sim]), value = 1000)
- n = list(type = "grid", value = c(100, 5000, 100000))
- prop = list(type = "grid", value = c(0.5,0.9))

IV. Results

First, note that the *n* defined in the simulation setup is equivalent to the *m* defined earlier in the paper (i.e. the number of hypotheses). In the simulation setup when using the simsalapar package, *n.sim* is equivalent to the *n* defined earlier in the paper (i.e. the number of simulations being run). In this paper, 1,000 simulations are performed. Then, the mean error rate values are calculated for each correction (tables of these mean values are included in the appendix). Each time a simulation is run, error rates are calculated, so when the hypotheses are tested 1000 times, there are 1000 values per error rate being calculated. Thus, it is helpful to compare the means of these values in order to better understand the differences between the various methods. Moreover, there are 3 levels of the numbers of hypotheses (*n*) and 2 levels of *prop* (i.e. the proportion of true hypotheses) being tested, so it is also useful to compare the methods across various levels of *n* and *prop*.

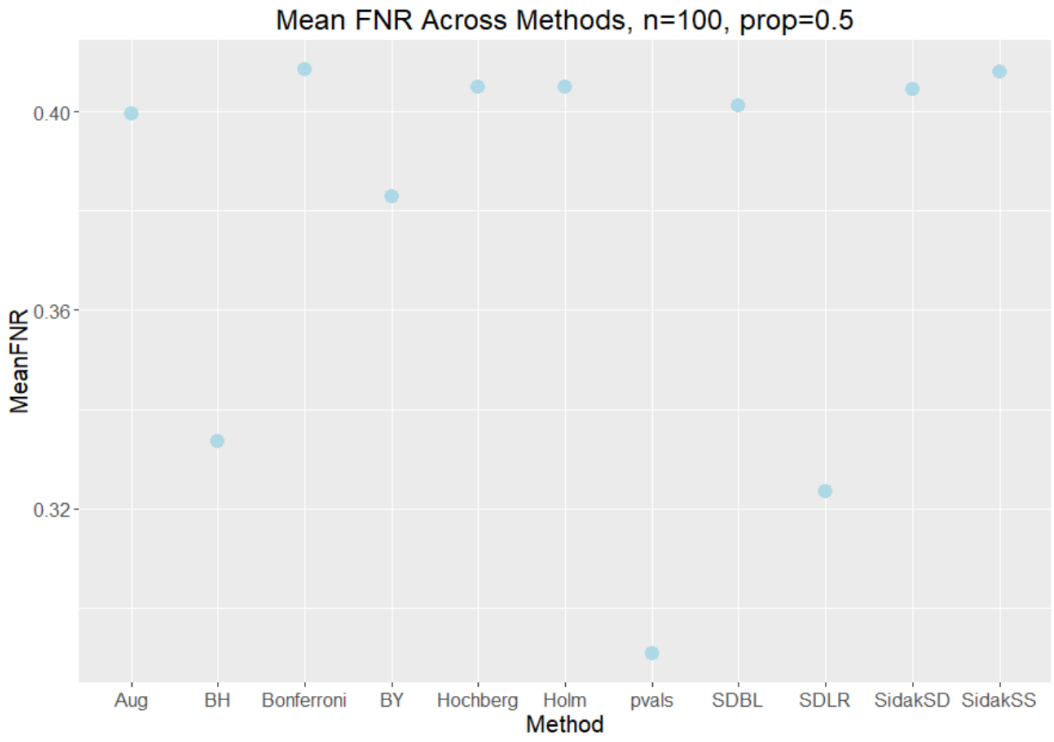
Below are graphs of a few of the mean error values (with the remaining graphs being included in the appendix as well).

Figure 1: Mean FDR when n=100, prop=0.5



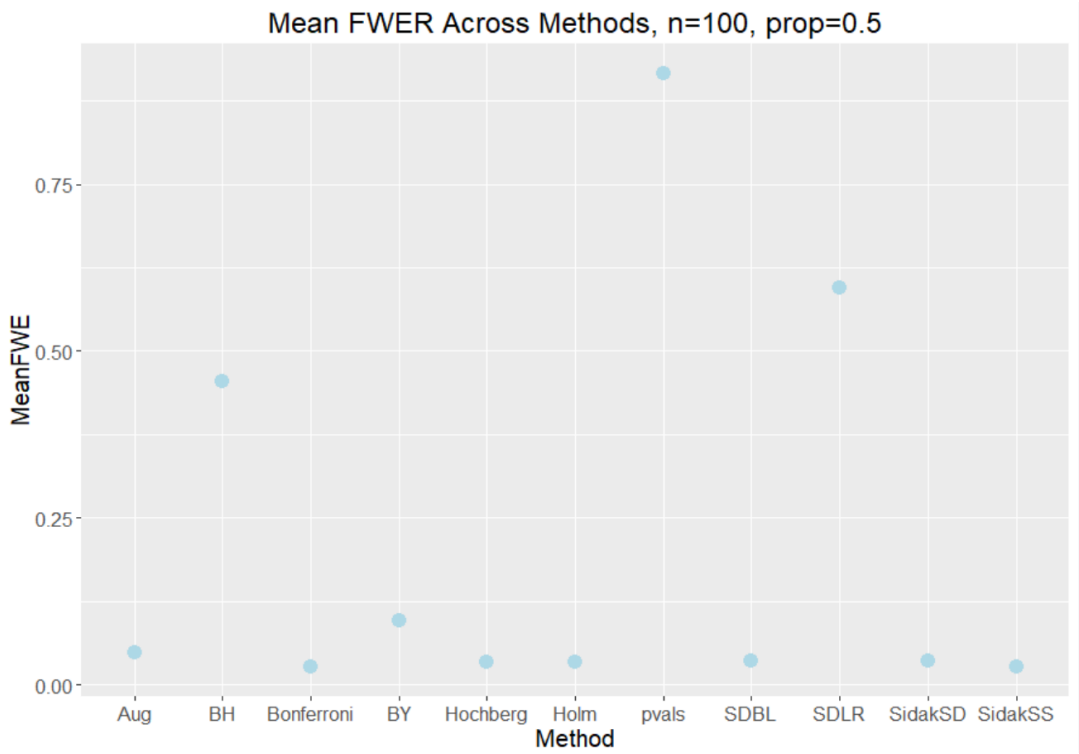
From Figure 1, we see that highest proportion of false positives corresponds to the raw p-values, i.e. when no adjustments are made. When there are no adjustments, the FDR is about 0.11, meaning 11% of the rejections are false positives. The next highest proportion of false positives corresponds to the Step-Down Lehmann-Romano method, where the FDR is about 0.038, meaning 3.8% of the rejections are false positives. The Benjamini-Hochberg method also stands out as one that has a higher proportion of false positives, with an FDR of 0.025, meaning 2.5% or rejections are false positives. The rest of the methods are very close to zero, meaning there are no false positives (likely because these methods rejected very few hypotheses to begin with).

Figure 2: Mean FNR when n=100, prop=0.5



From Figure 2, we see that almost all of the methods that were identified to be conservative from Figure 1, also seem to be conservative here. This is because the methods (Augmentation, Bonferroni, Hochberg, Holm, Step-Down Benjamini-Liu, Step-Down Sidak, and One-Step Sidak) have the highest proportions of false negatives, with all of their FNR values being at or slightly larger than 0.40. However, note that the Benjamini-Yekutieli method has a lower FNR value (0.382) is lower than the other conservative methods, even though from Figure 1, it was deduced that the Benjamini-Yekutieli method is similarly conservative to the others. Perhaps this implies that the Benjamini-Yekutieli method is more effective at rejecting hypotheses that should be rejected, meaning there is a low false positive rate, while also being effective at failing to reject the hypotheses that are, indeed, true. (In other words, the balancing act is in play!) Additionally, the methods with the highest FDR values also have the lowest FNR values, which supports the claim that these methods (Benjamini-Hochberg and Step-Down Lehmann-Romano) are less conservative. Unsurprisingly, the unadjusted p-values have the lowest FNR value.

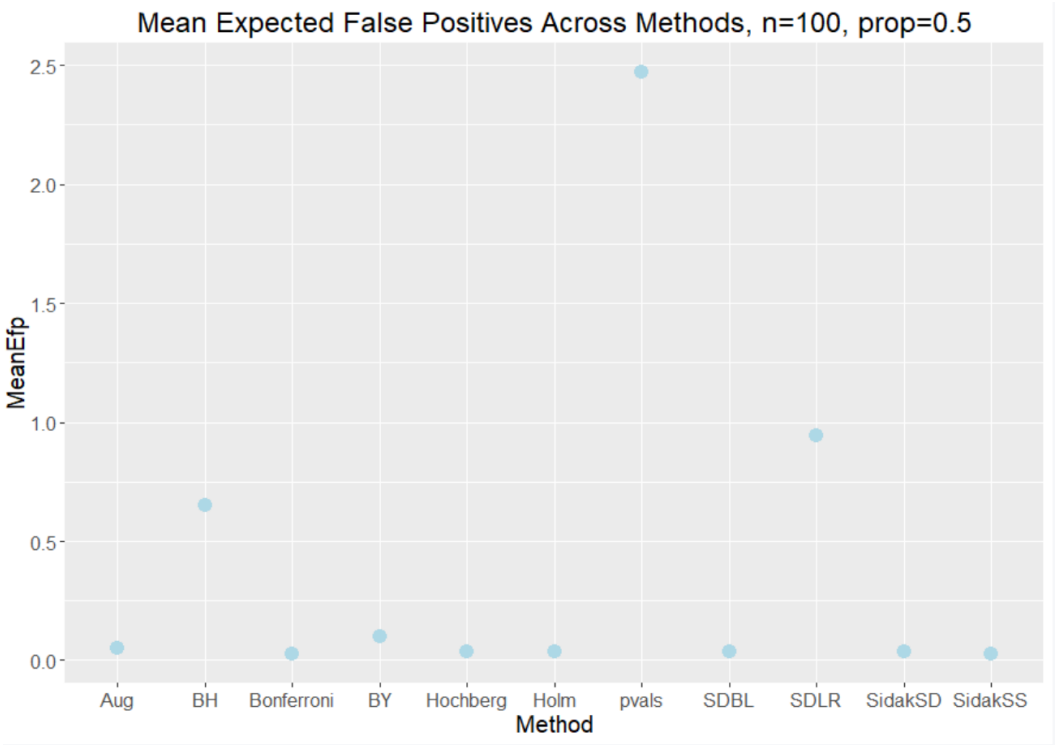
Figure 3: Mean FWER when n=100, prop=0.5



It is also interesting to compare the mean FWER across methods, since the FWER shows which methods are most likely to have at least 1 false positive. Similar to Figure 1, the Benjamini-Hochberg and Step-Down Lehmann-Romano methods have the highest error rate, with FWER values of 0.455 and 0.595, respectively. This means that about half of the time, these methods will result in having at least 1 false positive. Also note that all of the methods, no matter how conservative, have an FWER value greater than 0, meaning at least a small percentage of the time, there will be at least 1 false positive. Moreover, Figure 3 supports the claim that the Benjamini-Yekutieli method is more effective at controlling the Type I error rate than the other, more conservative, methods. This is because its FWER value is only slightly higher than the values for the more conservative methods, meaning the rate at which false positives occur for the BY method is similar to the other methods. However, the difference is that from the FNR comparisons, it was shown that the BY method results in a substantially lower rate of false negatives. This is why the BY method is perhaps more "effective" at controlling the error rate.

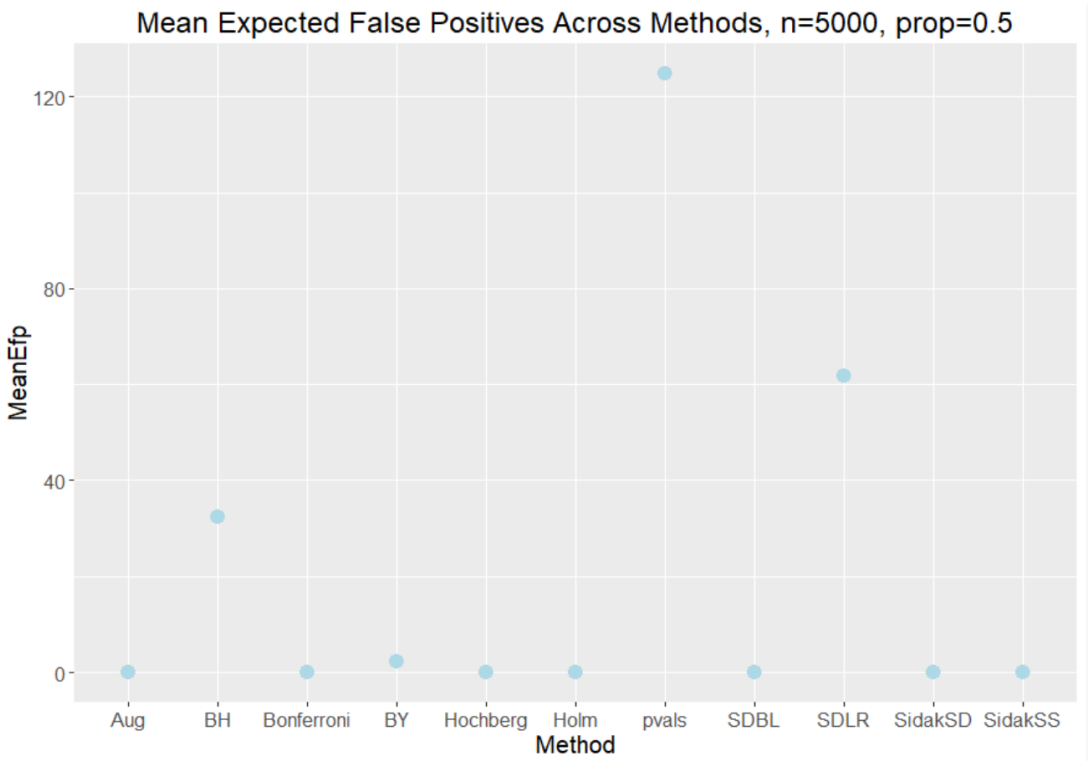
The above analyses are for a fixed level of *n* and *prop*. Now, the differences in error values across various levels of *n* will be discussed by looking at the plots for mean expected false positives (Efp) for a fixed level of *prop*.

Figure 4: Mean Expected False Positives when n=100, prop=0.5



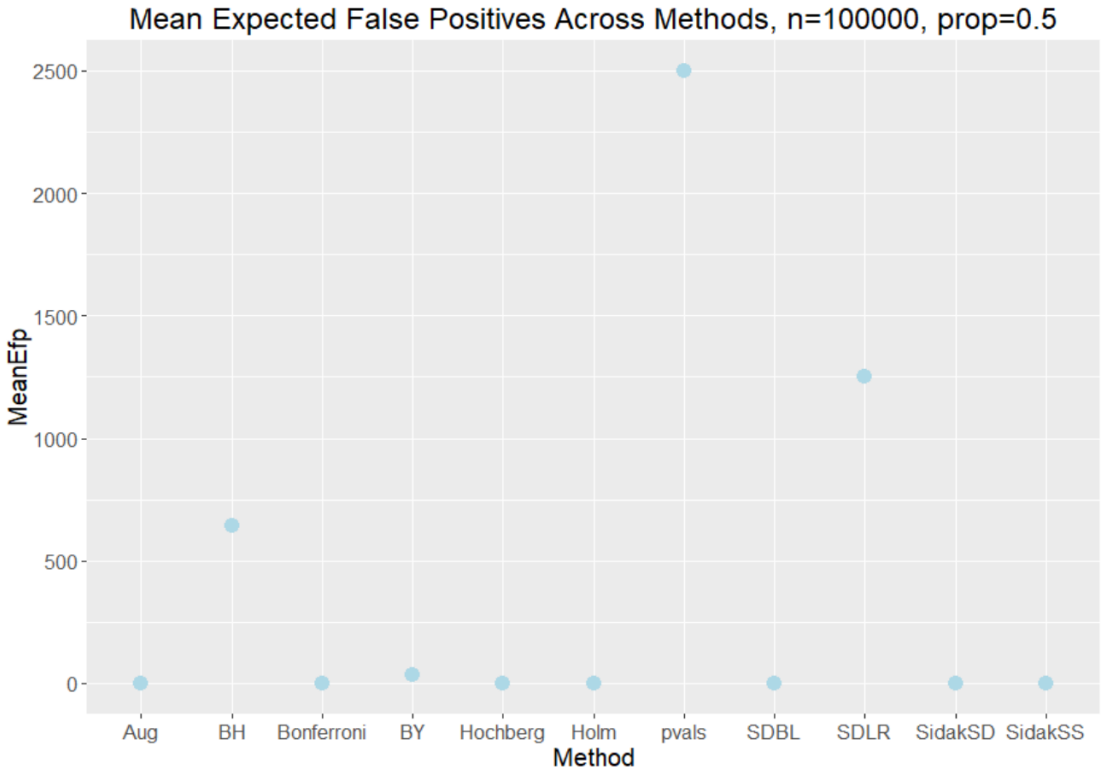
When the number of hypotheses is 100, the uncorrected p-values have almost 2.5 (or about 3) false positives. Step-Down LR has approximately 1 false positive, so it has an improvement of about 66% (as it is about one-third of the Efp of the uncorrected value). This also means that for every 100 hypotheses tested simultaneously, it is expected that there will be one false positive, when the Step-Down LR correction is applied. The Benjamini-Hochberg method has an Efp value of 0.652 which is about a quarter of the uncorrected value, while the remaining methods all have an Efp value of around 0.

Figure 4: Mean Expected False Positives when n=5000, prop=0.5



When the number of hypotheses is 5000, the number of false positives is 124.78 when no correction is applied. The Efp for Step-Down Lehmann-Romano is 61.849, which is about one-half of the number of false positives without any corrections. This is significantly higher than when *n* was 100 (recall, the SDLR Efp value was one-third of the uncorrected value). Additionally, the Benjamini-Hochberg Efp value is 32.381, which is about a quarter of uncorrected value. This is about the same as when *n* was 100. Again, the rest of the corrections all have an Efp very close to zero.

Figure 5: Mean Expected False Positives when n=100000, prop=0.5



When the number of hypotheses is 100000, the uncorrected p-values result in about 2500 false positives. The SDLR method has an Efp value of 1250.282, which is one-half of the uncorrected value, just like in the case when *n* was 5000. The Benjamini-Hochberg method has an Efp value of 645.835, which is, again, a quarter of the uncorrected value. Just like the previous two cases, in this case when *n* is 100000, the rest of the methods have practically zero false positives.

Hence, the Benjamini-Hochberg method is rather consistent in its improvement of the number of false positives, no matter how many hypotheses are tested. As the number of hypotheses grows, the more consistent the improvement is for the SDLR method. However, when the number of hypotheses is small, the SDLR method gives fewer false positives than when the number of hypotheses is large in comparison to the unadjusted number of false positives. This is interesting because as *n* grows, the Efp values for the other, more conservative, methods get closer and closer to zero. Hence, just because the number of hypotheses grows, it does not generally mean that the number of false positives will also grow when using many of these error-control methods.

Finally, the difference in error proportions across methods as the proportion of true hypotheses changes will be analyzed, for a fixed level of *n*. The following comparisons will be made by investigating the differences in the false negative rate (FNR) when *n* is 100000.

Figure 6: Mean FNR when n=100000, prop=0.5

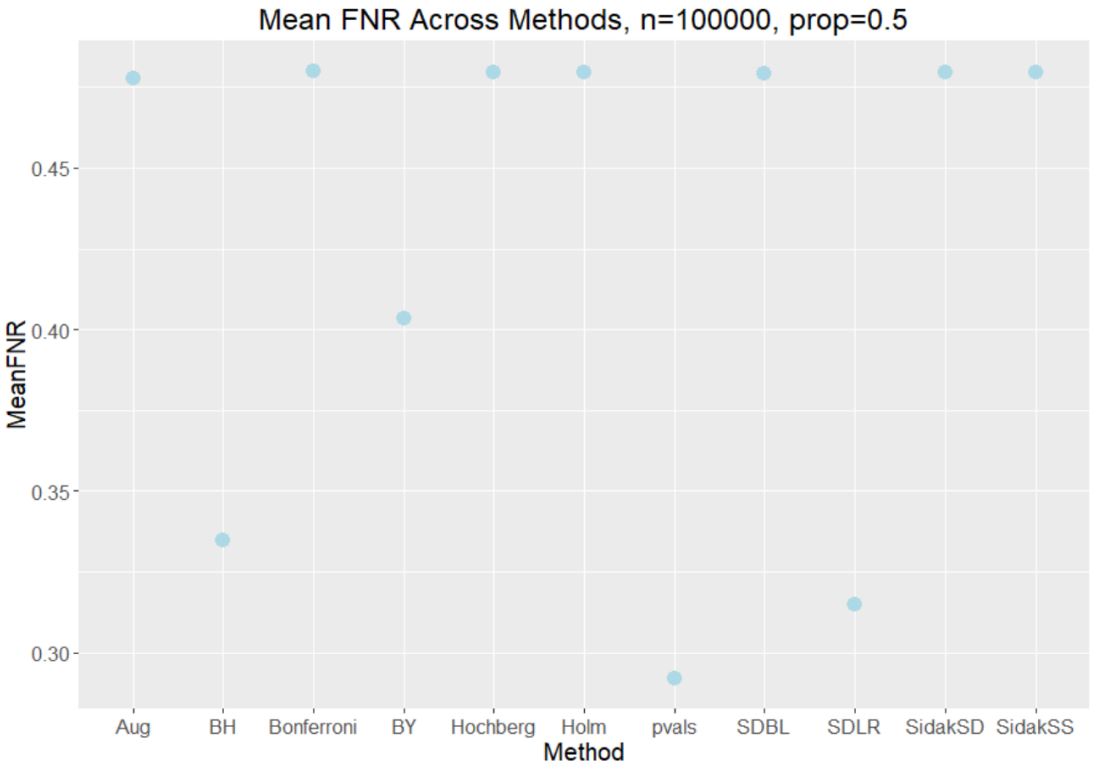
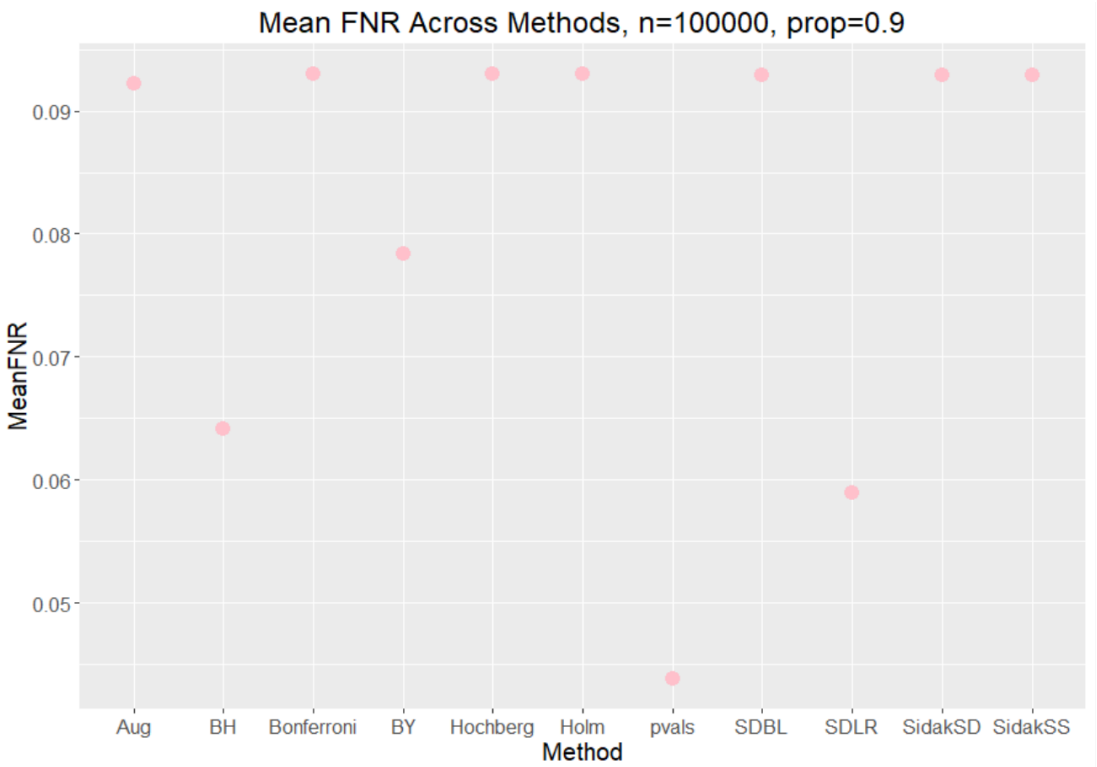


Figure 6: Mean FNR when n=100000, prop=0.9



When *prop* is 0.5, all of the FNR values are more than four times as large as the FNR values when *prop* is 0.9. The relationships between FNR values for various methods stay about the same (the graphs look very similar), but the values are much smaller when the proportion of false hypotheses is small. In other words, when the true proportion of false hypotheses is small, fewer of the total not-rejected hypotheses should have been rejected.

V. Conclusion

From the results discussed above, it can be concluded that most of the error-correction methods are very conservative, to a point where very few hypotheses are rejected at all. This causes the false negative rate to be much higher than the false positive rate, no matter which method is applied. Additionally, every method makes a noticeable improvement on the false positive and false negative error approximations. When the p-values are not corrected, the number of, and proportion of, false positives and negatives is a lot higher than when the p-values are adjusted.

Depending on how conservative a researcher wants to be, certain methods are likely better to use than others. Similarly, depending on if the researcher wants to be able to reject a fair number of hypotheses (and thus gain more information), certain methods are better suited to the goal than others. Specifically, the more conservative methods seem to be Augmentation, Bonferroni, Hochberg, Holm, Step-Down Benjamini-Liu, Step-down Sidak, and One-Step Sidak. Hence, if a researcher wants to be more cautious with their rejections, or perhaps wants to ensure a very low false positive rate, these methods would be the most appropriate. As discussed in the Results section, the Benjamini-Yekutieli method is one of the most effective methods when it comes to balancing precaution and power. The Step-Down Lehmann-Romano method tends to reject a greater proportion of hypotheses than the Benjamini-Yekutieli method, but at the same time, it also tends to have a greater proportion of false positives.

VI. Discussion

It may be enticing to choose a “best” method, one that clearly stands out from the rest. However, the parameters and context, such as number of hypotheses being tested and the proportion of true hypotheses, play a roll in determining which methods are better than others. There are so many academic fields in which research is carried out, and each has its own level of precision and its magnitude of hypotheses being tested. Thus, choosing one, best, method is likely not possible. However, knowing the context in which the multiple hypothesis testing is being carried out in can help one decide which error-controlling method could be more useful.

There are further additions that can be made to this simulation study. First, it would be interesting to investigate the effect that small sample sizes have on how well these methods control Type I error. As the test statistics move further away from normality, how well do

these methods control the error rate? Additionally, there are newer methods, such as the sign error rate. Comparing this method to some of the older ones (such as Bonferroni) may be interesting to see how the more classical approaches are being updated and improved upon.

VII. Appendix

Step-Down Lehmann-Romano R Code:

```
kfweLR3=function (p, k = 1, alpha = 0.01, disp = TRUE)
{
 s <- length(p)
 sdconst <- rep(1, s)
 sdconst[1:min(k, s)] <- k * alpha/s
 if (s > k)
  sdconst[(k + 1):s] <- k * alpha/(s + k - ((k + 1):s))
 ps <- sort(p)
 u <- ps < sdconst
 res <- 0
 if (any(u)) {
  w <- min(which(!u)) - 1
  res <- ps[w]
 }

 adjp <- p*alpha/res
 p[which(p > res)] <- 1
 p[p <= alpha] <- 0
 h = (!p)
 if (disp) {
  #cat(paste("Lehmann e Romano k-FWER Step Down procedure\n ",
  #         length(p), " tests, k=", k, ", alpha=", alpha, "\n ",
  #         sum(h), " rejections\n\n", sep = ""))
 }
 adjp[adjp >1] <- 1
 return(adjp)
}
```

Table 2: Mean Efp Values for all levels of n, prop

Method	n	prop	MeanEfp	Method	n	prop	MeanEfp
Aug	1e+02	0.5	0.05	Label	1e+02	0.5	50
Aug	1e+02	0.9	0.049	Label	1e+02	0.9	90
Aug	5e+03	0.5	0.036	Label	5e+03	0.5	2500
Aug	5e+03	0.9	0.061	Label	5e+03	0.9	4500
Aug	1e+05	0.5	0.036	Label	1e+05	0.5	50000
Aug	1e+05	0.9	0.063	Label	1e+05	0.9	90000
BH	1e+02	0.5	0.652	pvals	1e+02	0.5	2.471
BH	1e+02	0.9	0.241	pvals	1e+02	0.9	4.48
BH	5e+03	0.5	32.381	pvals	5e+03	0.5	124.78
BH	5e+03	0.9	9.078	pvals	5e+03	0.9	224.901
BH	1e+05	0.5	645.835	pvals	1e+05	0.5	2499.21
BH	1e+05	0.9	180.588	pvals	1e+05	0.9	4499.74
Bonferroni	1e+02	0.5	0.027	SDBL	1e+02	0.5	0.036
Bonferroni	1e+02	0.9	0.049	SDBL	1e+02	0.9	0.052
Bonferroni	5e+03	0.5	0.022	SDBL	5e+03	0.5	0.026
Bonferroni	5e+03	0.9	0.04	SDBL	5e+03	0.9	0.041
Bonferroni	1e+05	0.5	0.023	SDBL	1e+05	0.5	0.026
Bonferroni	1e+05	0.9	0.043	SDBL	1e+05	0.9	0.044
BY	1e+02	0.5	0.101	SDLR	1e+02	0.5	0.946
BY	1e+02	0.9	0.039	SDLR	1e+02	0.9	0.192
BY	5e+03	0.5	2.338	SDLR	5e+03	0.5	61.849
BY	5e+03	0.9	0.59	SDLR	5e+03	0.9	21.861
BY	1e+05	0.5	33.525	SDLR	1e+05	0.5	1250.28
BY	1e+05	0.9	8.647	SDLR	1e+05	0.9	448.334
Hochberg	1e+02	0.5	0.034	SidakSD	1e+02	0.5	0.035
Hochberg	1e+02	0.9	0.049	SidakSD	1e+02	0.9	0.051
Hochberg	5e+03	0.5	0.023	SidakSD	5e+03	0.5	0.024
Hochberg	5e+03	0.9	0.04	SidakSD	5e+03	0.9	0.041
Hochberg	1e+05	0.5	0.024	SidakSD	1e+05	0.5	0.024
Hochberg	1e+05	0.9	0.043	SidakSD	1e+05	0.9	0.044
Holm	1e+02	0.5	0.034	SidakSS	1e+02	0.5	0.027
Holm	1e+02	0.9	0.049	SidakSS	1e+02	0.9	0.049
Holm	5e+03	0.5	0.023	SidakSS	5e+03	0.5	0.022
Holm	5e+03	0.9	0.04	SidakSS	5e+03	0.9	0.04
Holm	1e+05	0.5	0.024	SidakSS	1e+05	0.5	0.023
Holm	1e+05	0.9	0.043	SidakSS	1e+05	0.9	0.043

Table 3: Mean Efn Values for all levels of n, prop

Method	n	prop	MeanEfn	Method	n	prop	MeanEfn
Aug	1e+02	0.5	33.408	Label	1e+02	0.5	50
Aug	1e+02	0.9	6.907	Label	1e+02	0.9	10
Aug	5e+03	0.5	2080.55	Label	5e+03	0.5	2500.02
Aug	5e+03	0.9	416.66	Label	5e+03	0.9	500.002
Aug	1e+05	0.5	45711.1	Label	1e+05	0.5	50005.6
Aug	1e+05	0.9	9141.66	Label	1e+05	0.9	10000.2
BH	1e+02	0.5	24.949	pvals	1e+02	0.5	19.672
BH	1e+02	0.9	6.126	pvals	1e+02	0.9	3.954
BH	5e+03	0.5	1240.53	pvals	5e+03	0.5	978.674
BH	5e+03	0.9	307.787	pvals	5e+03	0.9	195.857
BH	1e+05	0.5	24844.4	pvals	1e+05	0.5	19593.9
BH	1e+05	0.9	6155.16	pvals	1e+05	0.9	3920.07
Bonferroni	1e+02	0.5	34.624	SDBL	1e+02	0.5	33.637
Bonferroni	1e+02	0.9	6.907	SDBL	1e+02	0.9	6.873
Bonferroni	5e+03	0.5	2122.09	SDBL	5e+03	0.5	2106.99
Bonferroni	5e+03	0.9	424.593	SDBL	5e+03	0.9	423.71
Bonferroni	1e+05	0.5	46140.2	SDBL	1e+05	0.5	46038
Bonferroni	1e+05	0.9	9227.1	SDBL	1e+05	0.9	9219.21
BY	1e+02	0.5	31.165	SDLR	1e+02	0.5	23.667
BY	1e+02	0.9	7.121	SDLR	1e+02	0.9	6.269
BY	5e+03	0.5	1643.08	SDLR	5e+03	0.5	1120.47
BY	5e+03	0.9	375.283	SDLR	5e+03	0.9	281.375
BY	1e+05	0.5	33772.3	SDLR	1e+05	0.5	22404.1
BY	1e+05	0.9	7658.26	SDLR	1e+05	0.9	5606.25
Hochberg	1e+02	0.5	34.148	SidakSD	1e+02	0.5	34.083
Hochberg	1e+02	0.9	6.896	SidakSD	1e+02	0.9	6.883
Hochberg	5e+03	0.5	2115.91	SidakSD	5e+03	0.5	2113.83
Hochberg	5e+03	0.9	424.353	SidakSD	5e+03	0.9	423.957
Hochberg	1e+05	0.5	46102.3	SidakSD	1e+05	0.5	46077.8
Hochberg	1e+05	0.9	9225.61	SidakSD	1e+05	0.9	9220.76
Holm	1e+02	0.5	34.15	SidakSS	1e+02	0.5	34.547
Holm	1e+02	0.9	6.896	SidakSS	1e+02	0.9	6.896
Holm	5e+03	0.5	2115.91	SidakSS	5e+03	0.5	2120.09
Holm	5e+03	0.9	424.353	SidakSS	5e+03	0.9	424.203
Holm	1e+05	0.5	46102.3	SidakSS	1e+05	0.5	46115.9
Holm	1e+05	0.9	9225.61	SidakSS	1e+05	0.9	9222.25

Table 4: Mean FWER Values for all levels of n, prop

Method	n	prop	MeanFWE	Method	n	prop	MeanFWE
Aug	1e+02	0.5	0.049	Label	1e+02	0.5	1
Aug	1e+02	0.9	0.048	Label	1e+02	0.9	1
Aug	5e+03	0.5	0.036	Label	5e+03	0.5	1
Aug	5e+03	0.9	0.06	Label	5e+03	0.9	1
Aug	1e+05	0.5	0.036	Label	1e+05	0.5	1
Aug	1e+05	0.9	0.063	Label	1e+05	0.9	1
BH	1e+02	0.5	0.455	pvals	1e+02	0.5	0.916
BH	1e+02	0.9	0.201	pvals	1e+02	0.9	0.994
BH	5e+03	0.5	1	pvals	5e+03	0.5	1
BH	5e+03	0.9	0.999	pvals	5e+03	0.9	1
BH	1e+05	0.5	1	pvals	1e+05	0.5	1
BH	1e+05	0.9	1	pvals	1e+05	0.9	1
Bonferroni	1e+02	0.5	0.027	SDBL	1e+02	0.5	0.036
Bonferroni	1e+02	0.9	0.048	SDBL	1e+02	0.9	0.051
Bonferroni	5e+03	0.5	0.022	SDBL	5e+03	0.5	0.026
Bonferroni	5e+03	0.9	0.04	SDBL	5e+03	0.9	0.041
Bonferroni	1e+05	0.5	0.023	SDBL	1e+05	0.5	0.026
Bonferroni	1e+05	0.9	0.043	SDBL	1e+05	0.9	0.044
BY	1e+02	0.5	0.096	SDLR	1e+02	0.5	0.595
BY	1e+02	0.9	0.038	SDLR	1e+02	0.9	0.164
BY	5e+03	0.5	0.903	SDLR	5e+03	0.5	1
BY	5e+03	0.9	0.433	SDLR	5e+03	0.9	1
BY	1e+05	0.5	1	SDLR	1e+05	0.5	1
BY	1e+05	0.9	1	SDLR	1e+05	0.9	1
Hochberg	1e+02	0.5	0.034	SidakSD	1e+02	0.5	0.035
Hochberg	1e+02	0.9	0.048	SidakSD	1e+02	0.9	0.05
Hochberg	5e+03	0.5	0.023	SidakSD	5e+03	0.5	0.024
Hochberg	5e+03	0.9	0.04	SidakSD	5e+03	0.9	0.041
Hochberg	1e+05	0.5	0.024	SidakSD	1e+05	0.5	0.024
Hochberg	1e+05	0.9	0.043	SidakSD	1e+05	0.9	0.044
Holm	1e+02	0.5	0.034	SidakSS	1e+02	0.5	0.027
Holm	1e+02	0.9	0.048	SidakSS	1e+02	0.9	0.048
Holm	5e+03	0.5	0.023	SidakSS	5e+03	0.5	0.022
Holm	5e+03	0.9	0.04	SidakSS	5e+03	0.9	0.04
Holm	1e+05	0.5	0.024	SidakSS	1e+05	0.5	0.023
Holm	1e+05	0.9	0.043	SidakSS	1e+05	0.9	0.043

Table 5: Mean FDR Values for all levels of n, prop

Method	n	prop	MeanFDR	Method	n	prop	MeanFDR
Aug	1e+02	0.5	0.00151	Label	1e+02	0.5	0.5
Aug	1e+02	0.9	0.006528	Label	1e+02	0.9	0.9
Aug	5e+03	0.5	1.73E-05	Label	5e+03	0.5	0.499998
Aug	5e+03	0.9	0.000146	Label	5e+03	0.9	0.9
Aug	1e+05	0.5	7.88E-07	Label	1e+05	0.5	0.499972
Aug	1e+05	0.9	6.89E-06	Label	1e+05	0.9	0.899998
BH	1e+02	0.5	0.025965	pvals	1e+02	0.5	0.110554
BH	1e+02	0.9	0.038858	pvals	1e+02	0.9	0.52073
BH	5e+03	0.5	0.025448	pvals	5e+03	0.5	0.113047
BH	5e+03	0.9	0.028673	pvals	5e+03	0.9	0.534317
BH	1e+05	0.5	0.025337	pvals	1e+05	0.5	0.11312
BH	1e+05	0.9	0.028505	pvals	1e+05	0.9	0.534411
Bonferroni	1e+02	0.5	0.000787	SDBL	1e+02	0.5	0.001081
Bonferroni	1e+02	0.9	0.006528	SDBL	1e+02	0.9	0.006933
Bonferroni	5e+03	0.5	1.03E-05	SDBL	5e+03	0.5	1.23E-05
Bonferroni	5e+03	0.9	9.4E-05	SDBL	5e+03	0.9	9.65E-05
Bonferroni	1e+05	0.5	4.99E-07	SDBL	1e+05	0.5	5.65E-07
Bonferroni	1e+05	0.9	4.66E-06	SDBL	1e+05	0.9	4.77E-06
BY	1e+02	0.5	0.003336	SDLR	1e+02	0.5	0.038572
BY	1e+02	0.9	0.005514	SDLR	1e+02	0.9	0.028952
BY	5e+03	0.5	0.001422	SDLR	5e+03	0.5	0.052294
BY	5e+03	0.9	0.00157	SDLR	5e+03	0.9	0.071952
BY	1e+05	0.5	0.000992	SDLR	1e+05	0.5	0.052855
BY	1e+05	0.9	0.001128	SDLR	1e+05	0.9	0.074043
Hochberg	1e+02	0.5	0.001001	SidakSD	1e+02	0.5	0.001031
Hochberg	1e+02	0.9	0.006552	SidakSD	1e+02	0.9	0.006821
Hochberg	5e+03	0.5	1.09E-05	SidakSD	5e+03	0.5	1.13E-05
Hochberg	5e+03	0.9	9.41E-05	SidakSD	5e+03	0.9	9.65E-05
Hochberg	1e+05	0.5	5.21E-07	SidakSD	1e+05	0.5	5.21E-07
Hochberg	1e+05	0.9	4.66E-06	SidakSD	1e+05	0.9	4.77E-06
Holm	1e+02	0.5	0.001001	SidakSS	1e+02	0.5	0.00079
Holm	1e+02	0.9	0.006552	SidakSS	1e+02	0.9	0.006552
Holm	5e+03	0.5	1.09E-05	SidakSS	5e+03	0.5	1.03E-05
Holm	5e+03	0.9	9.41E-05	SidakSS	5e+03	0.9	9.41E-05
Holm	1e+05	0.5	5.21E-07	SidakSS	1e+05	0.5	4.99E-07
Holm	1e+05	0.9	4.66E-06	SidakSS	1e+05	0.9	4.66E-06

Table 6: Mean FDX Values for all levels of n, prop

Method	n	prop	MeanFDX	Method	n	prop	MeanFDX
Aug	1e+02	0.5	0	Label	1e+02	0.5	1
Aug	1e+02	0.9	0.044	Label	1e+02	0.9	1
Aug	5e+03	0.5	0	Label	5e+03	0.5	1
Aug	5e+03	0.9	0	Label	5e+03	0.9	1
Aug	1e+05	0.5	0	Label	1e+05	0.5	1
Aug	1e+05	0.9	0	Label	1e+05	0.9	1
BH	1e+02	0.5	0.04	pvals	1e+02	0.5	0.504
BH	1e+02	0.9	0.197	pvals	1e+02	0.9	0.994
BH	5e+03	0.5	0	pvals	5e+03	0.5	0.925
BH	5e+03	0.9	0	pvals	5e+03	0.9	1
BH	1e+05	0.5	0	pvals	1e+05	0.5	1
BH	1e+05	0.9	0	pvals	1e+05	0.9	1
Bonferroni	1e+02	0.5	0	SDBL	1e+02	0.5	0
Bonferroni	1e+02	0.9	0.044	SDBL	1e+02	0.9	0.047
Bonferroni	5e+03	0.5	0	SDBL	5e+03	0.5	0
Bonferroni	5e+03	0.9	0	SDBL	5e+03	0.9	0
Bonferroni	1e+05	0.5	0	SDBL	1e+05	0.5	0
Bonferroni	1e+05	0.9	0	SDBL	1e+05	0.9	0
BY	1e+02	0.5	0	SDLR	1e+02	0.5	0.082
BY	1e+02	0.9	0.038	SDLR	1e+02	0.9	0.156
BY	5e+03	0.5	0	SDLR	5e+03	0.5	0
BY	5e+03	0.9	0	SDLR	5e+03	0.9	0.032
BY	1e+05	0.5	0	SDLR	1e+05	0.5	0
BY	1e+05	0.9	0	SDLR	1e+05	0.9	0
Hochberg	1e+02	0.5	0	SidakSD	1e+02	0.5	0
Hochberg	1e+02	0.9	0.044	SidakSD	1e+02	0.9	0.046
Hochberg	5e+03	0.5	0	SidakSD	5e+03	0.5	0
Hochberg	5e+03	0.9	0	SidakSD	5e+03	0.9	0
Hochberg	1e+05	0.5	0	SidakSD	1e+05	0.5	0
Hochberg	1e+05	0.9	0	SidakSD	1e+05	0.9	0
Holm	1e+02	0.5	0	SidakSS	1e+02	0.5	0
Holm	1e+02	0.9	0.044	SidakSS	1e+02	0.9	0.044
Holm	5e+03	0.5	0	SidakSS	5e+03	0.5	0
Holm	5e+03	0.9	0	SidakSS	5e+03	0.9	0
Holm	1e+05	0.5	0	SidakSS	1e+05	0.5	0
Holm	1e+05	0.9	0	SidakSS	1e+05	0.9	0

Table 7: Mean FNR Values for all levels of n, prop

Method	n	prop	MeanFNR	Method	n	prop	MeanFNR
Aug	1e+02	0.5	0.399653	Label	1e+02	0.5	1
Aug	1e+02	0.9	0.071089	Label	1e+02	0.9	1
Aug	5e+03	0.5	0.454208	Label	5e+03	0.5	1
Aug	5e+03	0.9	0.084743	Label	5e+03	0.9	1
Aug	1e+05	0.5	0.477595	Label	1e+05	0.5	1
Aug	1e+05	0.9	0.092208	Label	1e+05	0.9	1
BH	1e+02	0.5	0.333661	pvals	1e+02	0.5	0.290886
BH	1e+02	0.9	0.063553	pvals	1e+02	0.9	0.043927
BH	5e+03	0.5	0.334501	pvals	5e+03	0.5	0.291768
BH	5e+03	0.9	0.064133	pvals	5e+03	0.9	0.043801
BH	1e+05	0.5	0.334835	pvals	1e+05	0.5	0.292032
BH	1e+05	0.9	0.064133	pvals	1e+05	0.9	0.043838
Bonferroni	1e+02	0.5	0.40842	SDBL	1e+02	0.5	0.401284
Bonferroni	1e+02	0.9	0.071089	SDBL	1e+02	0.9	0.070759
Bonferroni	5e+03	0.5	0.459113	SDBL	5e+03	0.5	0.45734
Bonferroni	5e+03	0.9	0.086217	SDBL	5e+03	0.9	0.086053
Bonferroni	1e+05	0.5	0.479926	SDBL	1e+05	0.5	0.479373
Bonferroni	1e+05	0.9	0.09299	SDBL	1e+05	0.9	0.092918
BY	1e+02	0.5	0.382887	SDLR	1e+02	0.5	0.323469
BY	1e+02	0.9	0.073077	SDLR	1e+02	0.9	0.064963
BY	5e+03	0.5	0.396782	SDLR	5e+03	0.5	0.314828
BY	5e+03	0.9	0.076981	SDLR	5e+03	0.9	0.059114
BY	1e+05	0.5	0.403304	SDLR	1e+05	0.5	0.314867
BY	1e+05	0.9	0.078426	SDLR	1e+05	0.9	0.058915
Hochberg	1e+02	0.5	0.405017	SidakSD	1e+02	0.5	0.404548
Hochberg	1e+02	0.9	0.070981	SidakSD	1e+02	0.9	0.070857
Hochberg	5e+03	0.5	0.458388	SidakSD	5e+03	0.5	0.458144
Hochberg	5e+03	0.9	0.086173	SidakSD	5e+03	0.9	0.086099
Hochberg	1e+05	0.5	0.479721	SidakSD	1e+05	0.5	0.479588
Hochberg	1e+05	0.9	0.092976	SidakSD	1e+05	0.9	0.092932
Holm	1e+02	0.5	0.405033	SidakSS	1e+02	0.5	0.407877
Holm	1e+02	0.9	0.070981	SidakSS	1e+02	0.9	0.070981
Holm	5e+03	0.5	0.458388	SidakSS	5e+03	0.5	0.458879
Holm	5e+03	0.9	0.086173	SidakSS	5e+03	0.9	0.086145
Holm	1e+05	0.5	0.479721	SidakSS	1e+05	0.5	0.479794
Holm	1e+05	0.9	0.092976	SidakSS	1e+05	0.9	0.092945

Below are the remaining tables that were not included prior in the paper.

Figure 7: Mean Efn when n=100, prop=0.5

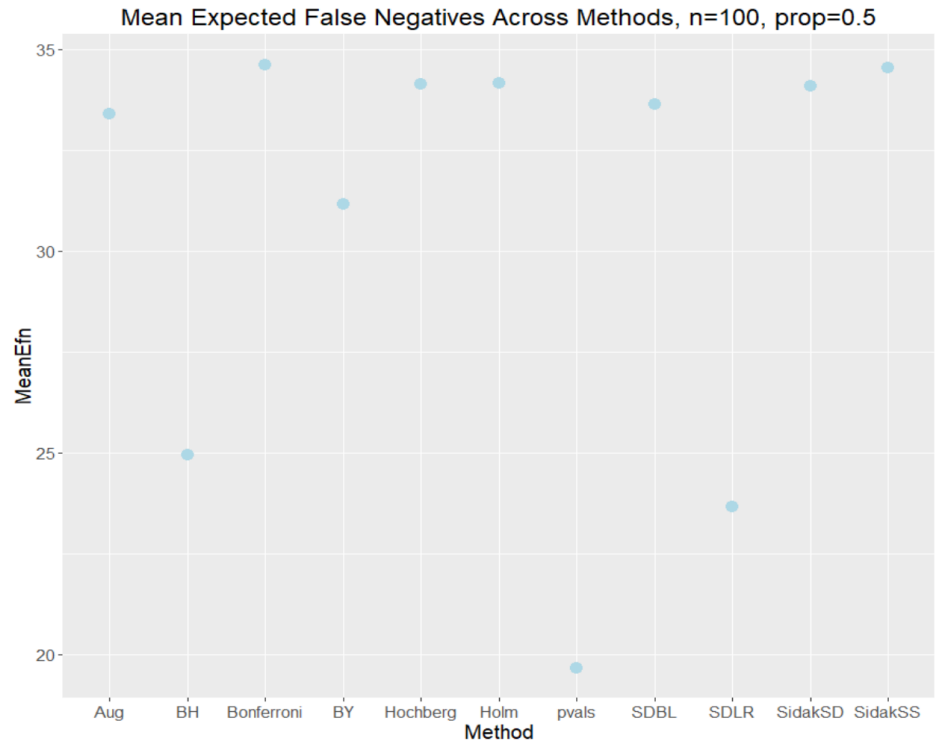


Figure 8: Mean Efn when n=5000, prop=0.5

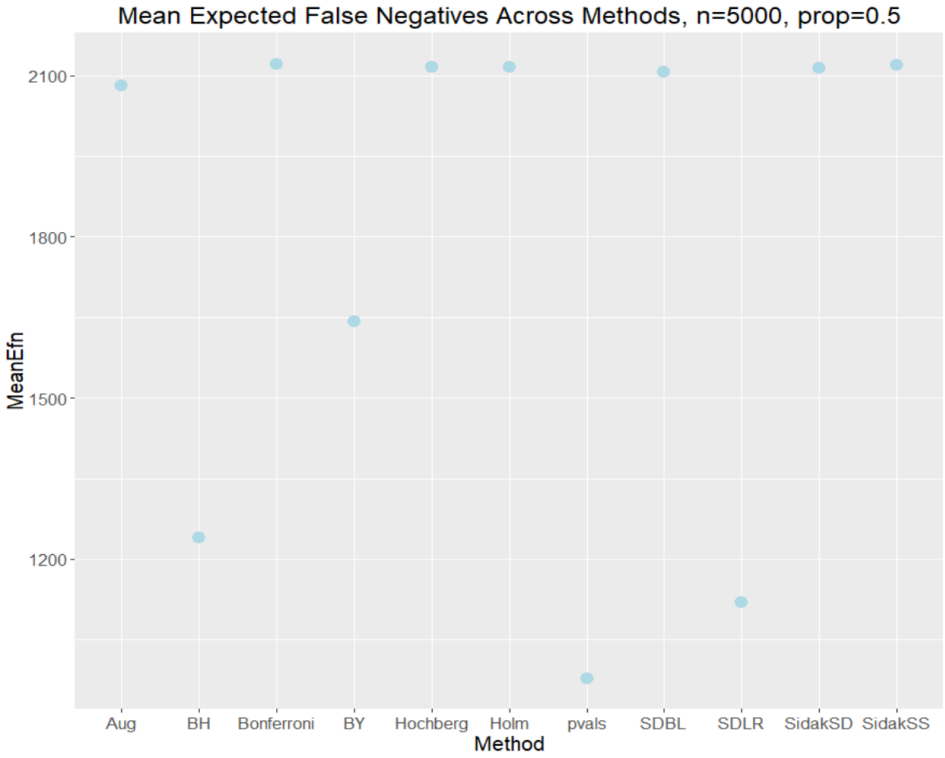


Figure 9: Mean Efn when n=100000, prop=0.5

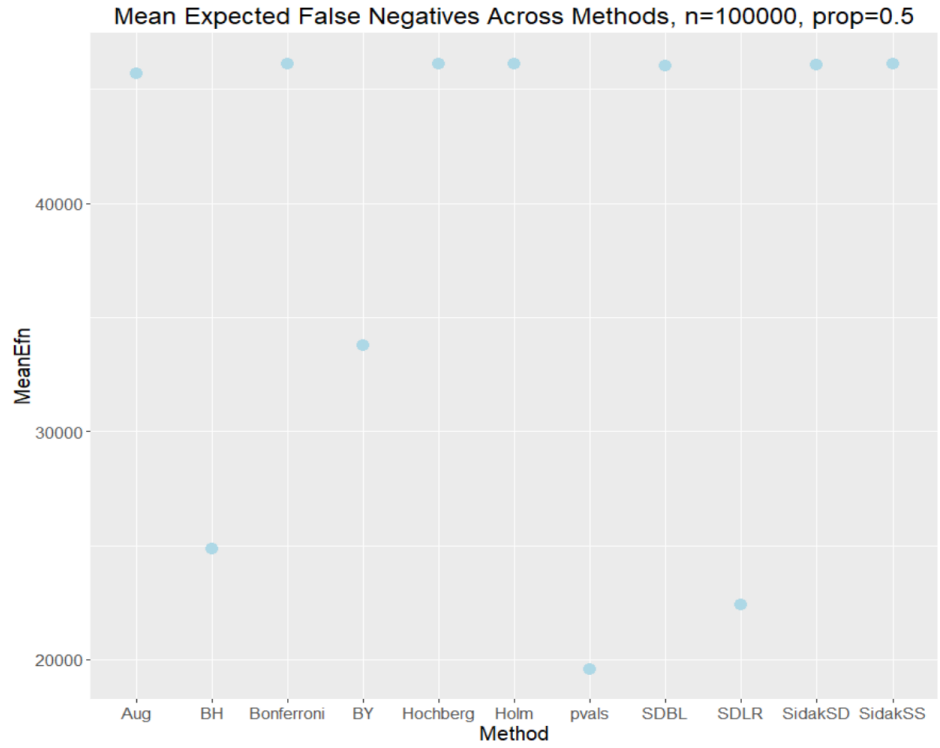


Figure 10: Mean FWER when n=5000, prop=0.5

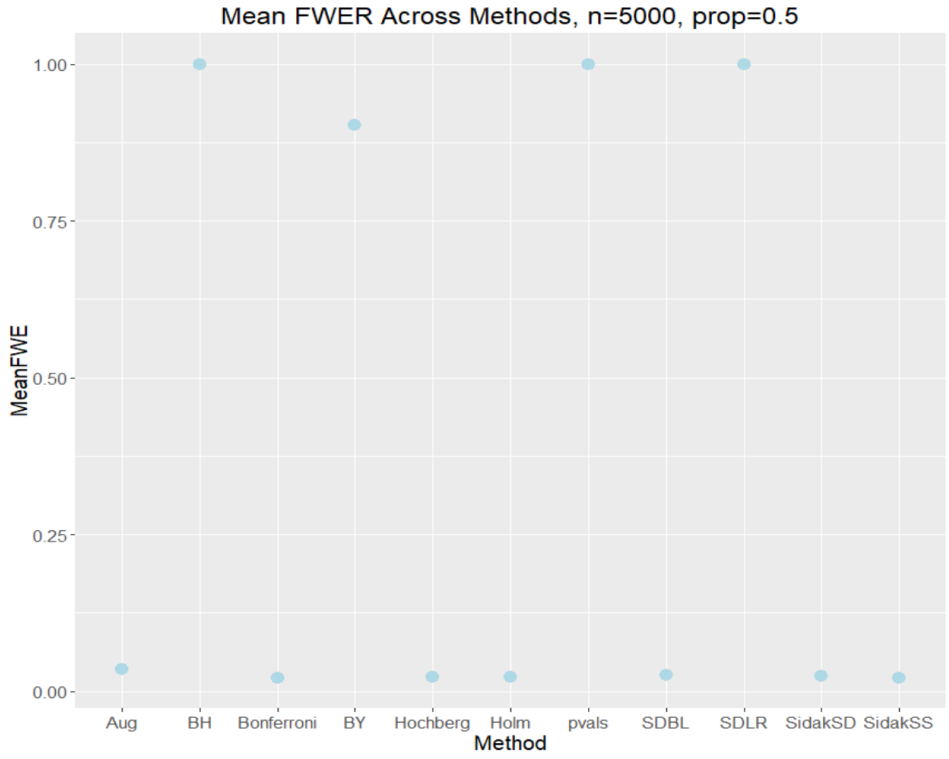


Figure 11: Mean FWER when n=100000, prop=0.5

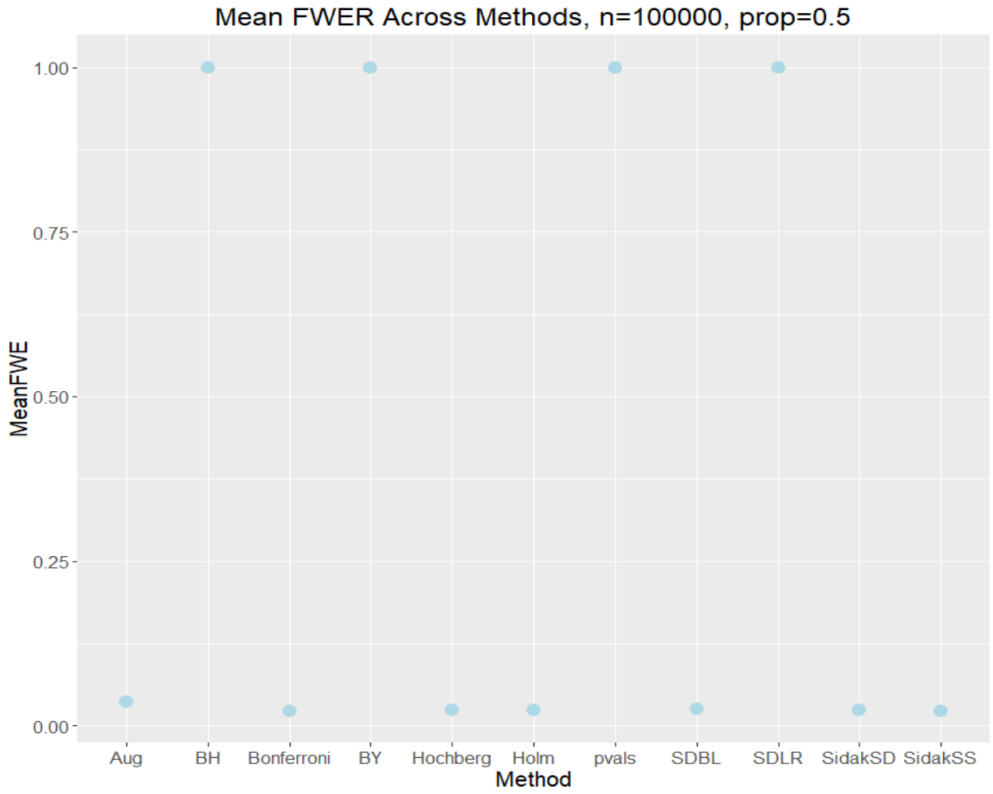


Figure 12: Mean FDR when n=5000, prop=0.5

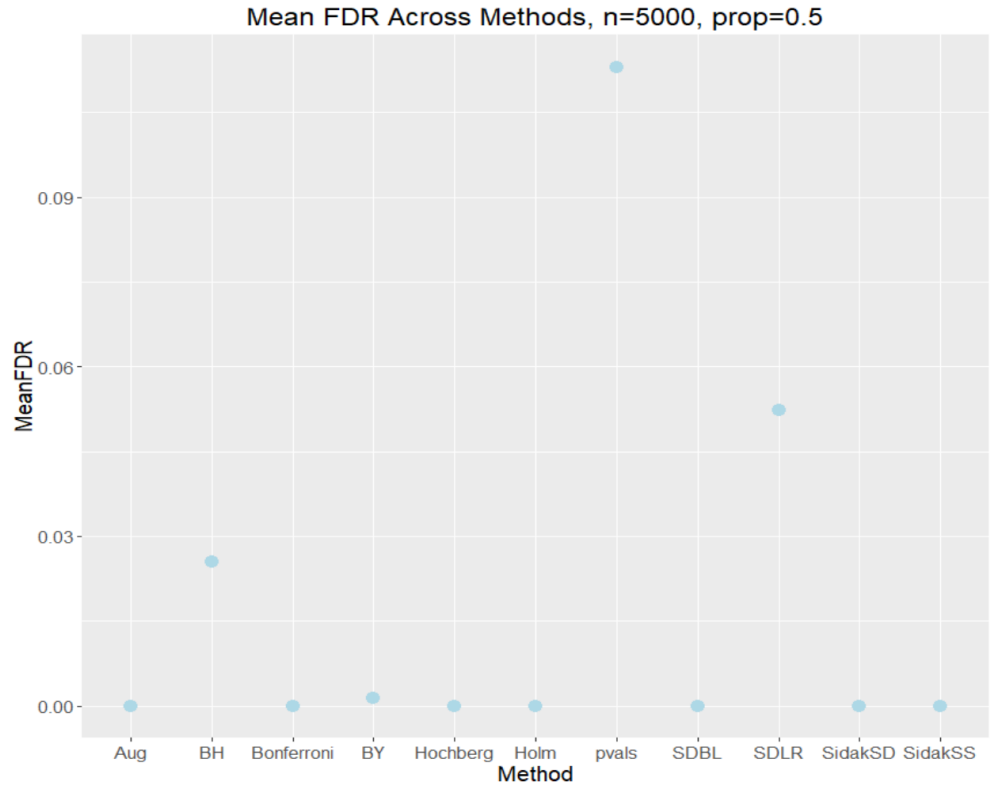


Figure 13: Mean FDR when n=100000, prop=0.5

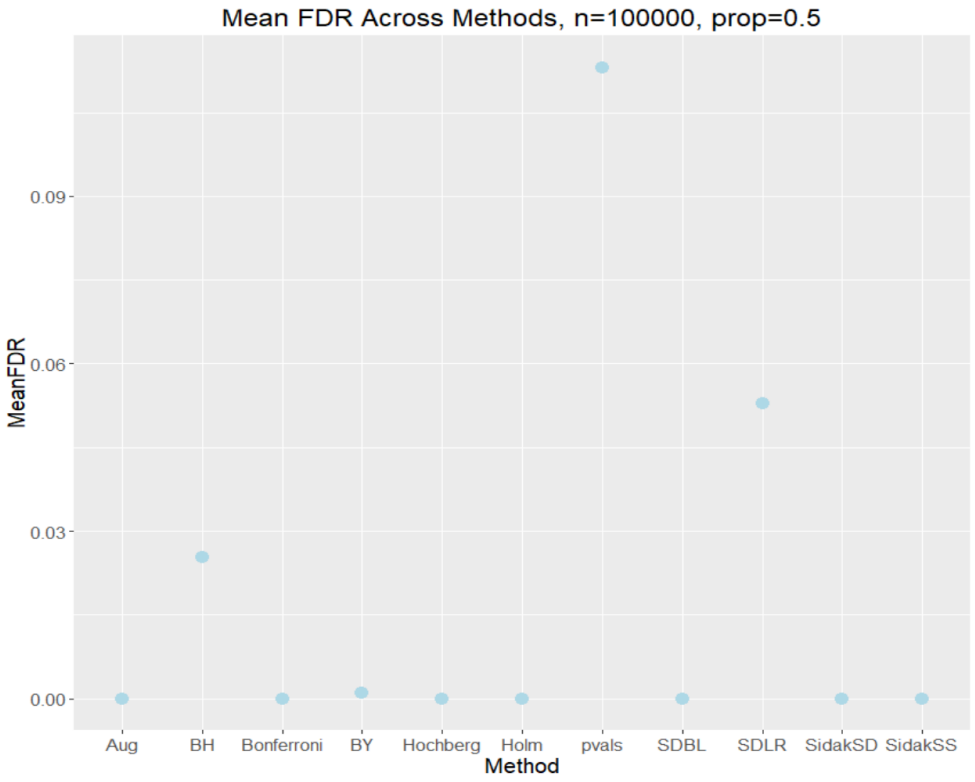


Figure 14: Mean FDX when n=100, prop=0.5

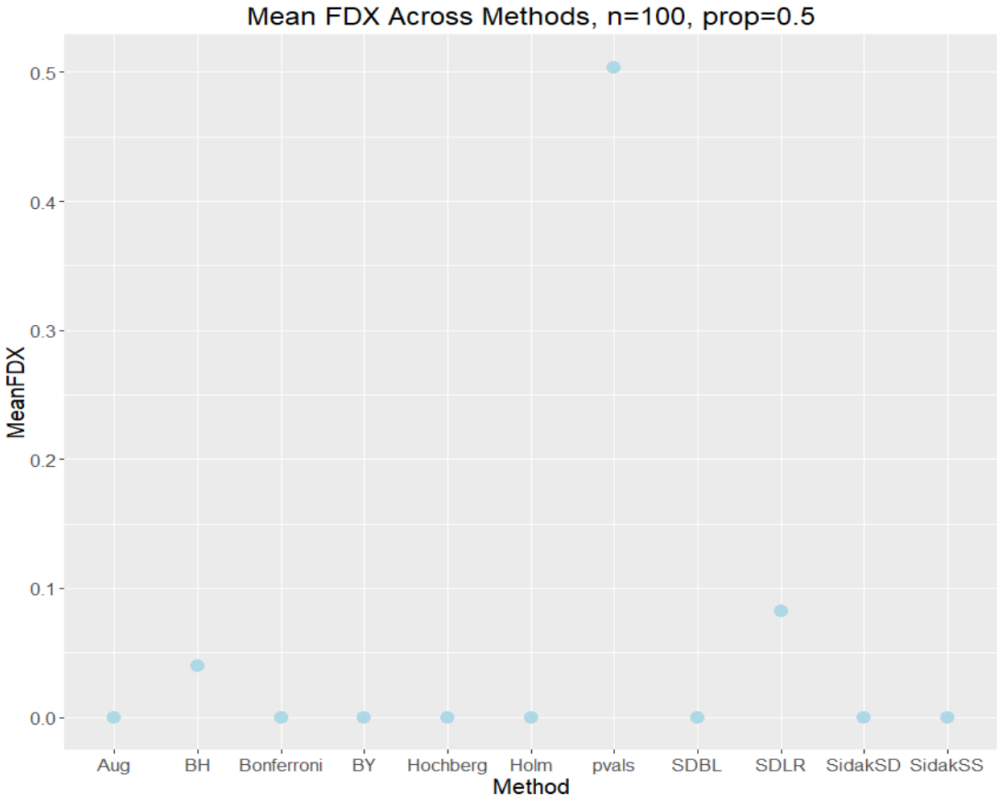


Figure 15: Mean FDX when n=5000, prop=0.5

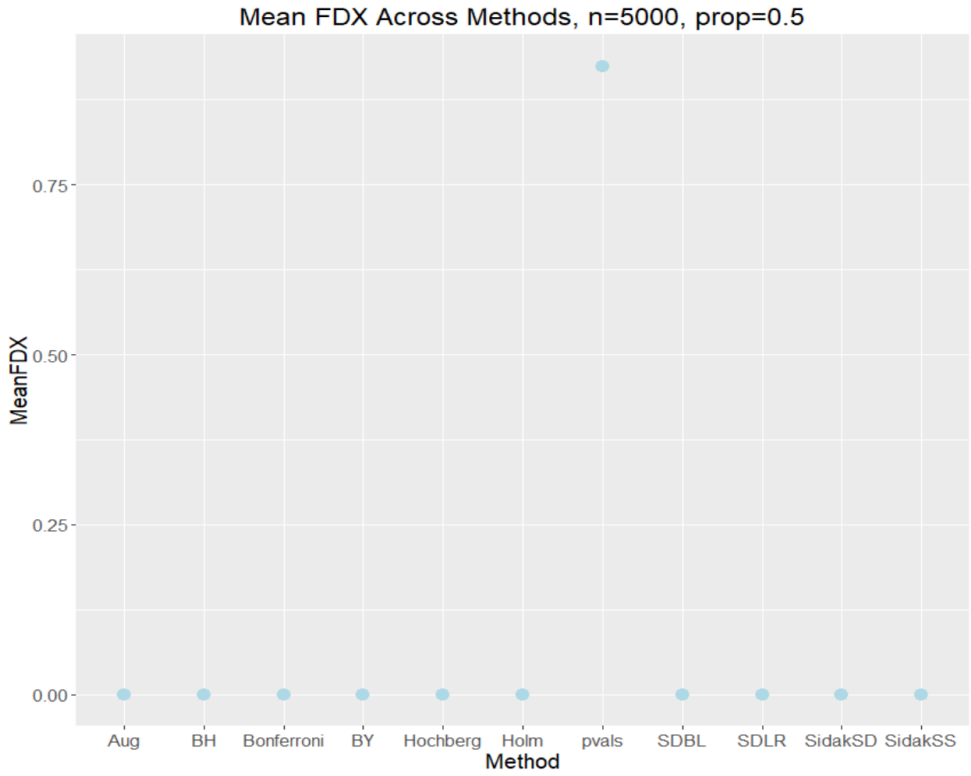


Figure 16: Mean FDX when n=100000, prop=0.5

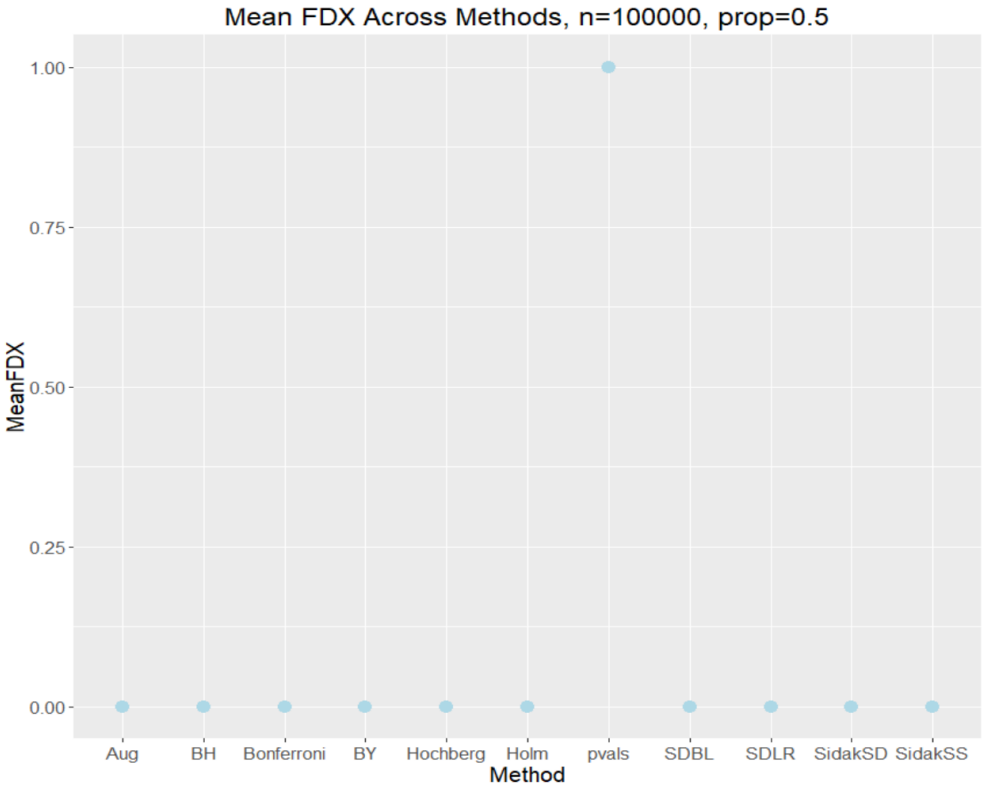


Figure 17: Mean FNR when n=100, prop=0.5

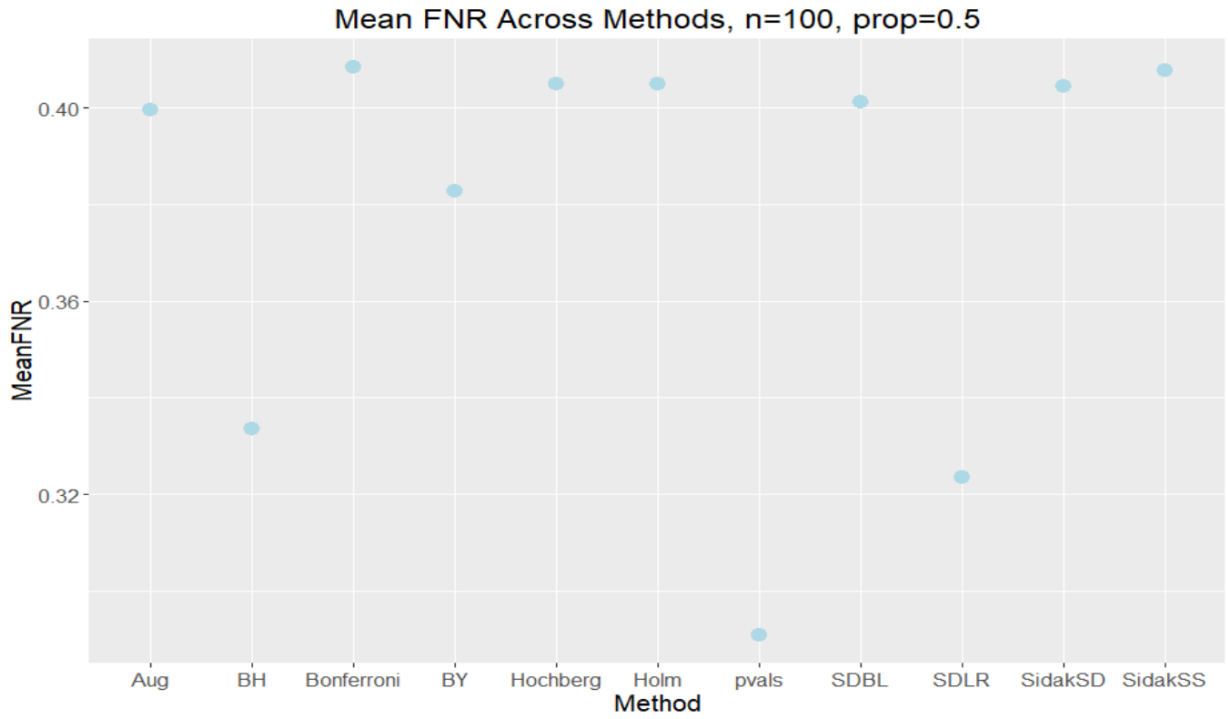


Figure 18: Mean FNR when n=5000, prop=0.5

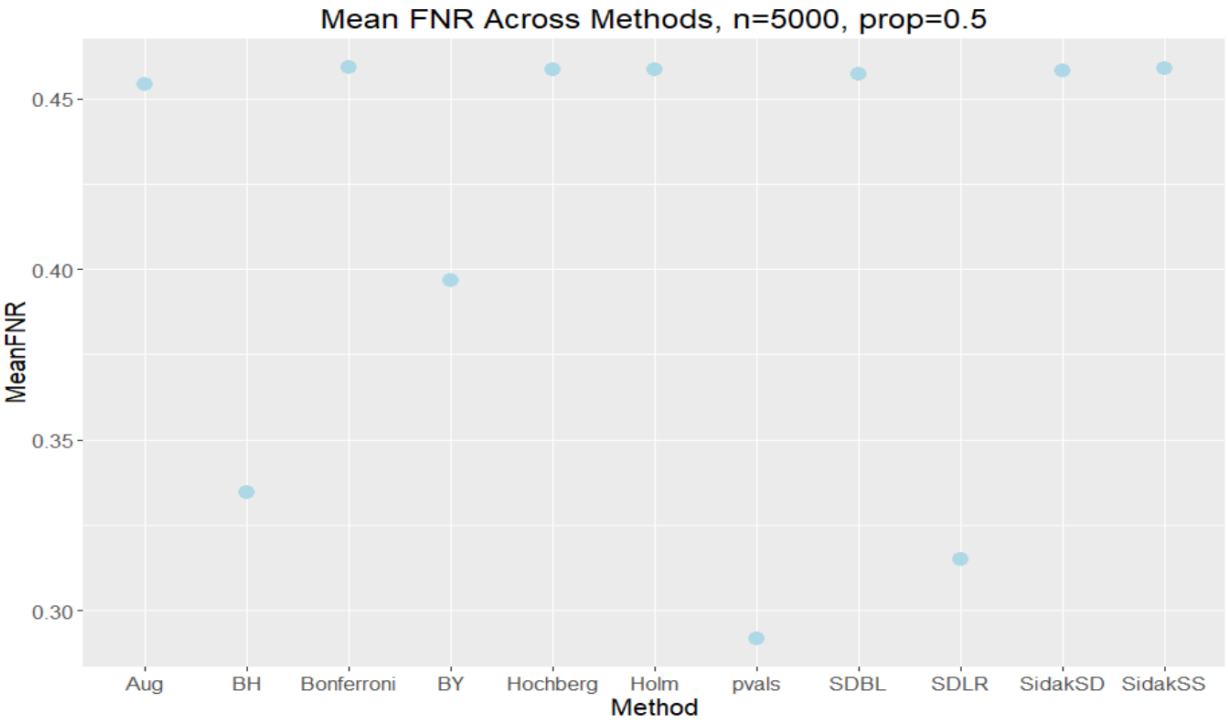


Figure 19: Mean Efp when n=100, prop=0.9

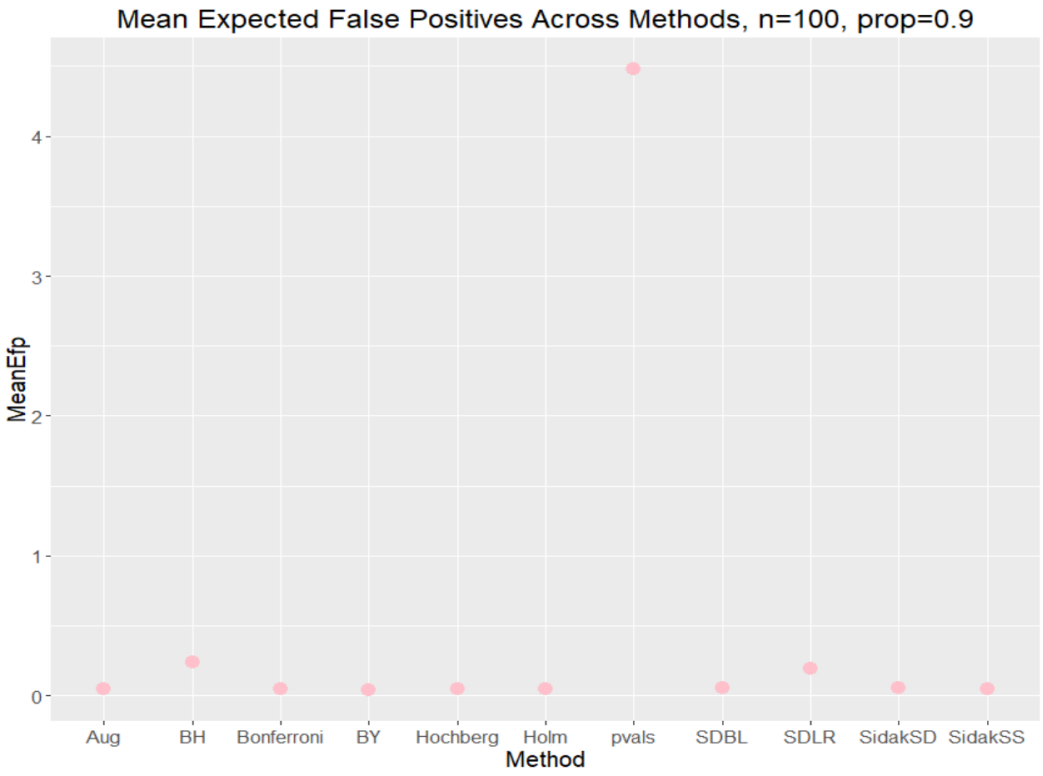


Figure 20: Mean Efp when n=5000, prop=0.9

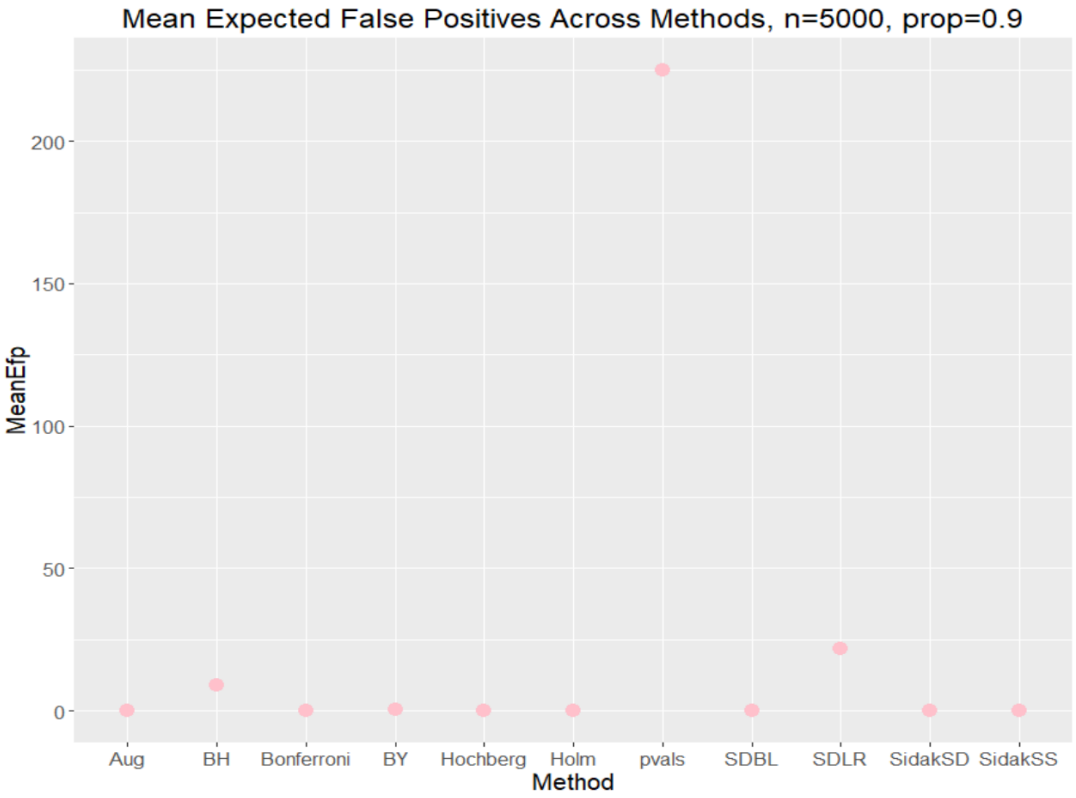


Figure 21: Mean Efp when n=100000, prop=0.9

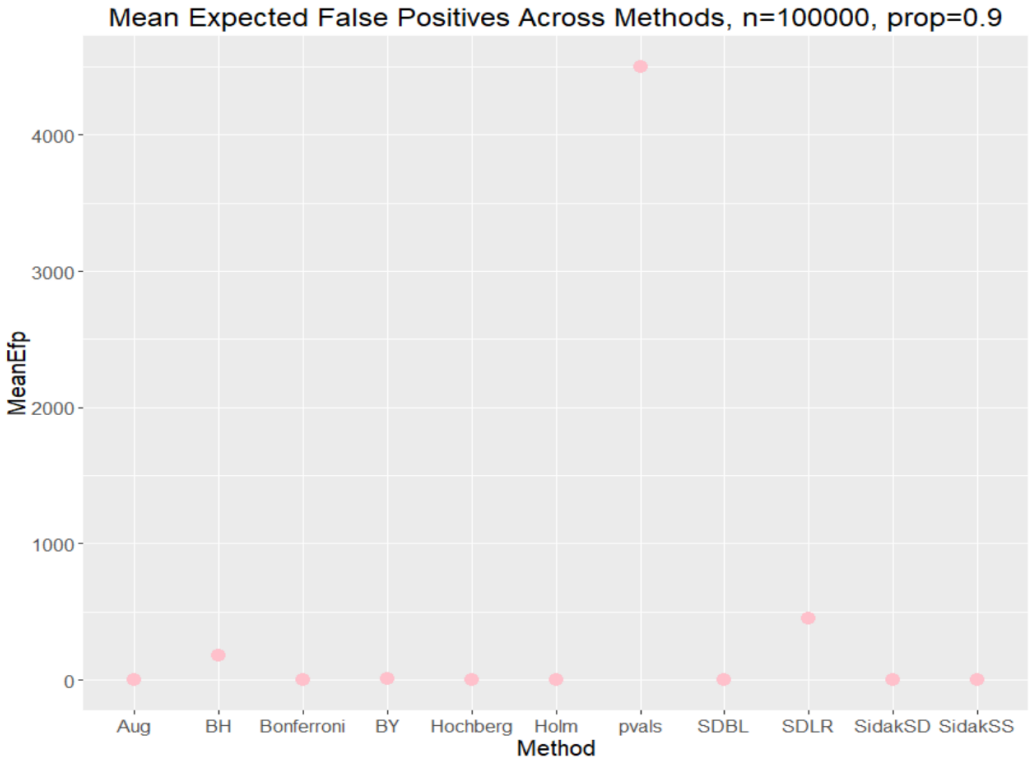


Figure 22: Mean Efn when n=100, prop=0.9

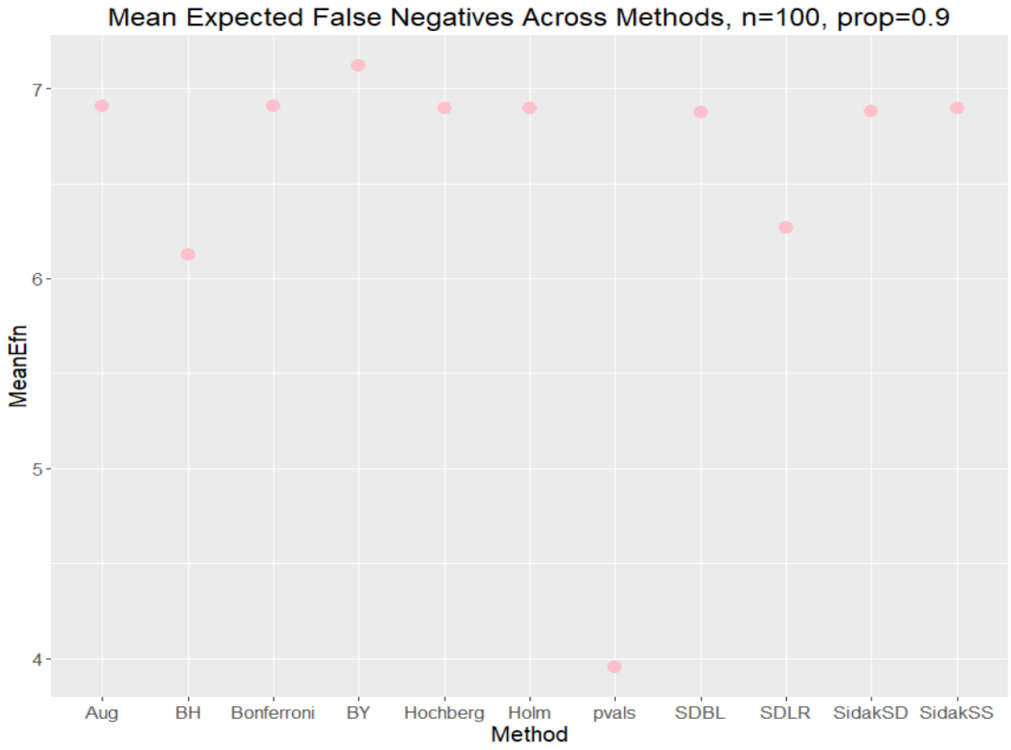


Figure 23: Mean Efn when n=5000, prop=0.9

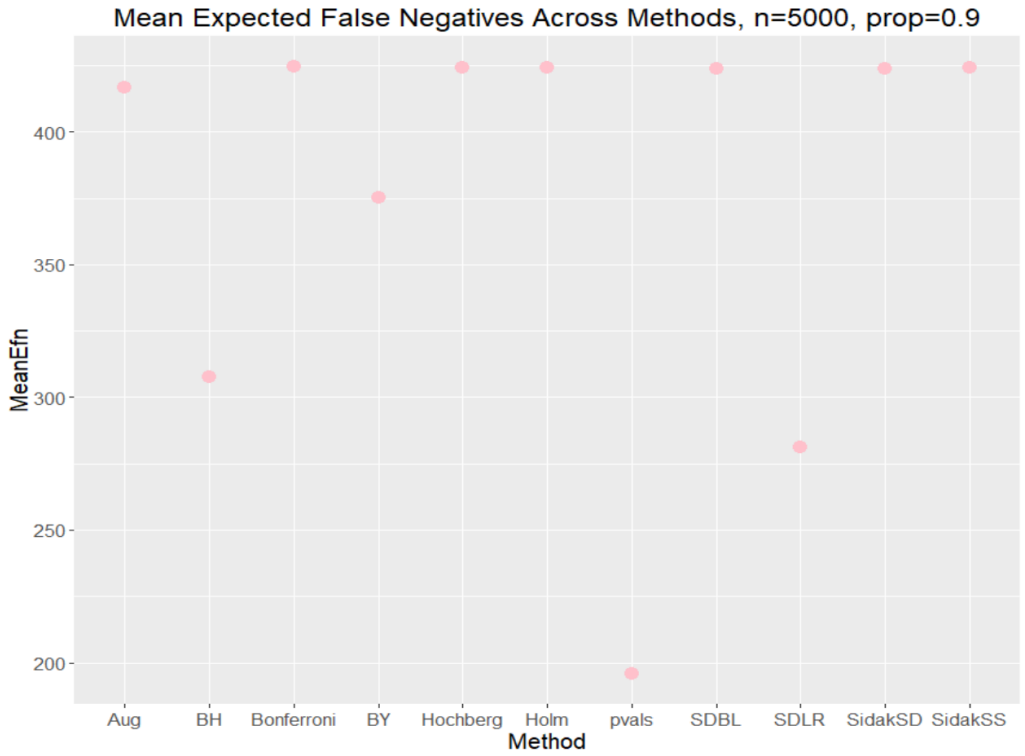


Figure 24: Mean Efn when n=100000, prop=0.9

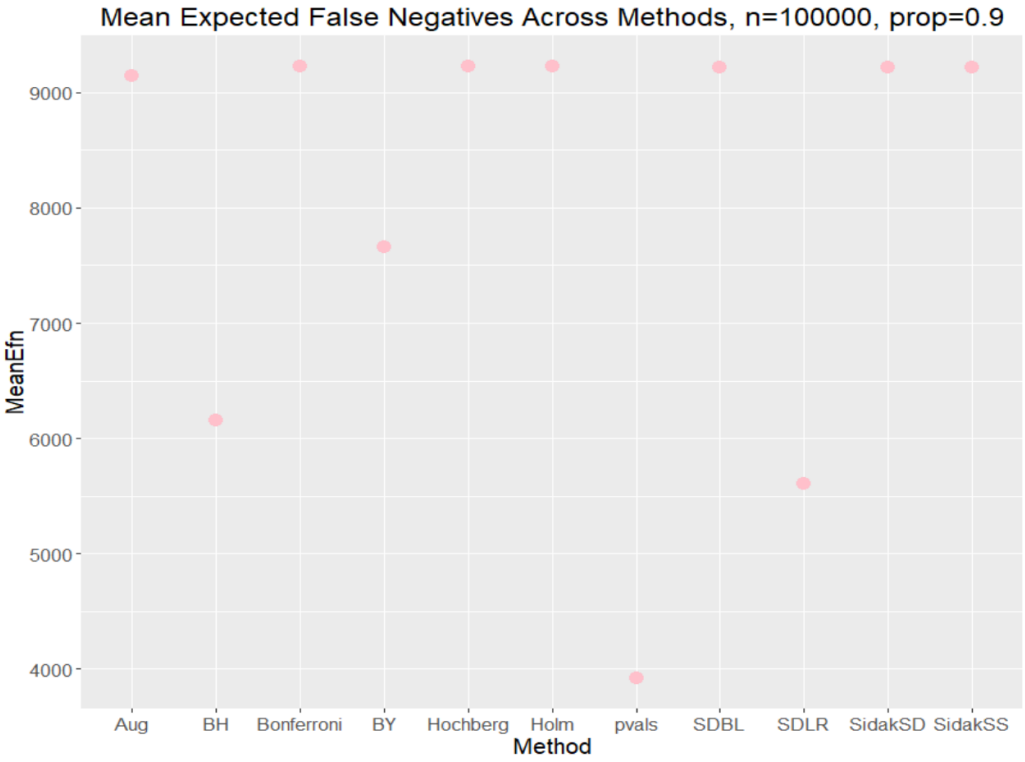


Figure 25: Mean FWER when n=100, prop=0.9

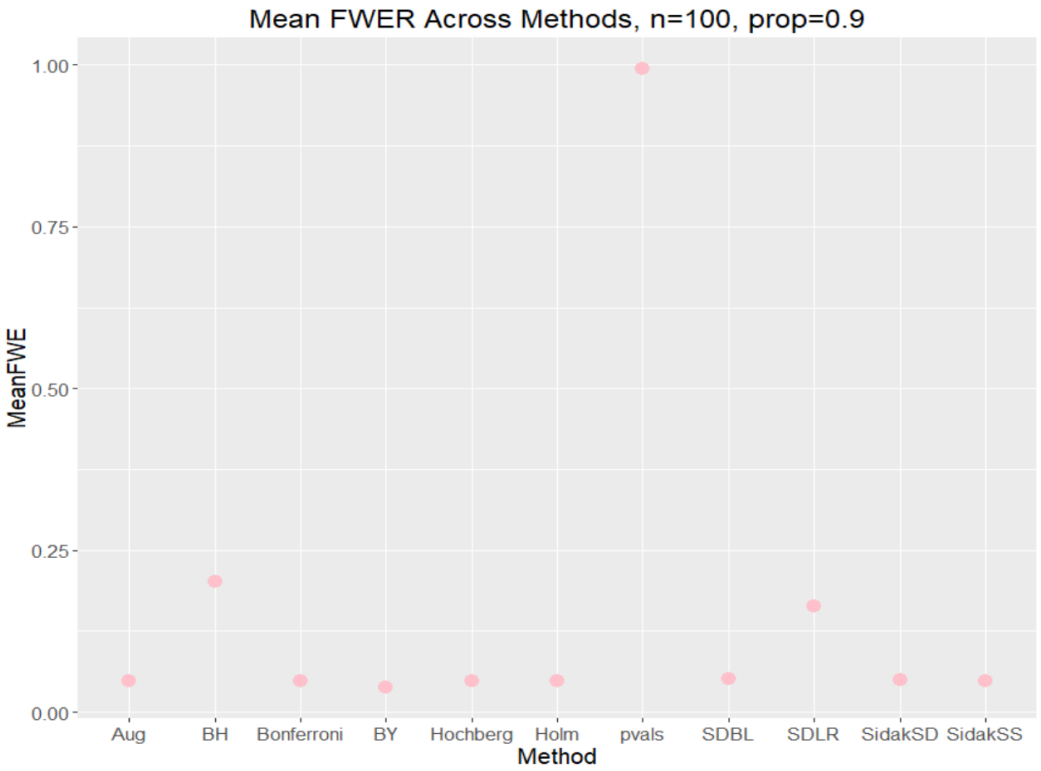


Figure 26: Mean FWER when n=5000, prop=0.9

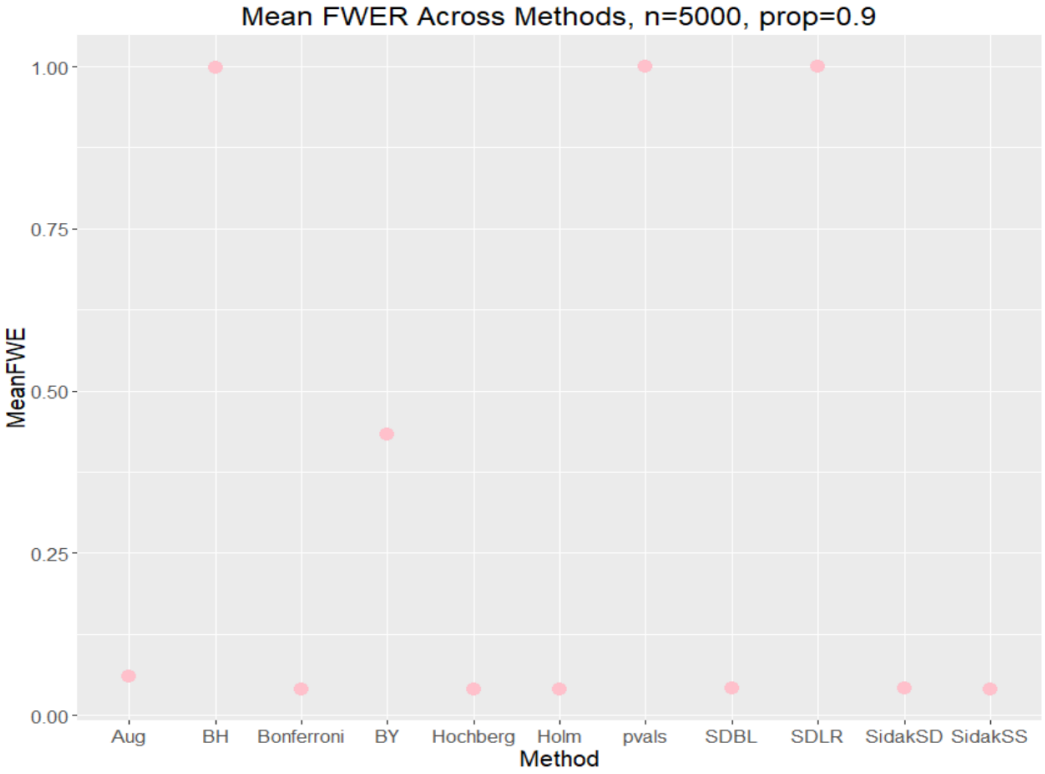


Figure 27: Mean FWER when n=100000, prop=0.9

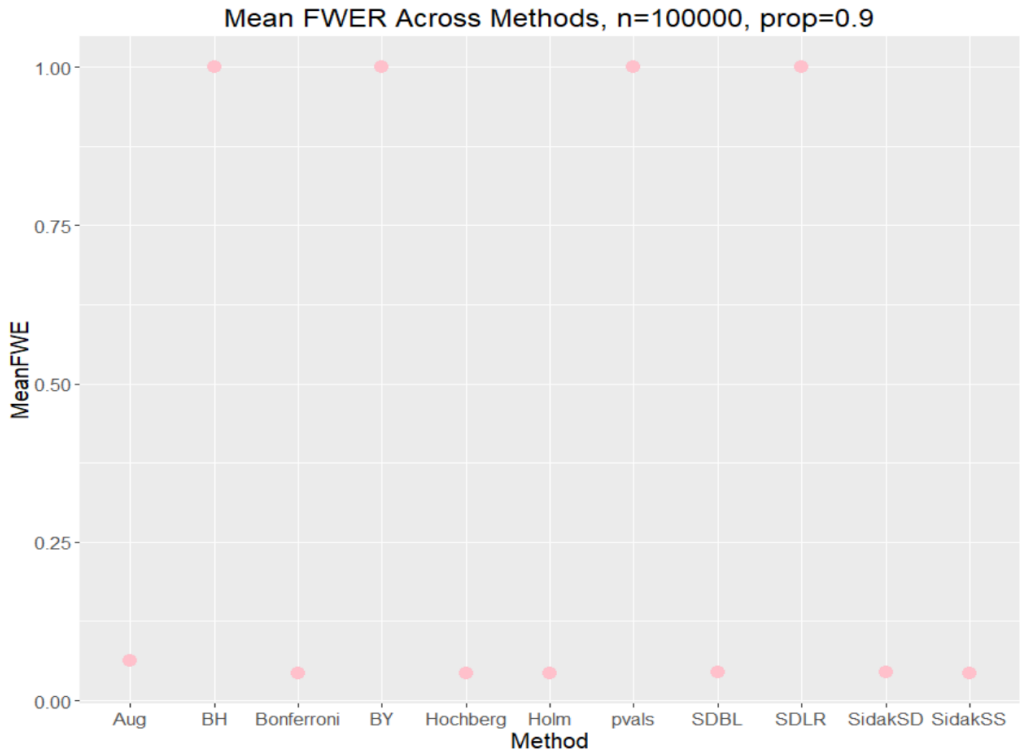


Figure 28: Mean FDR when n=100, prop=0.9

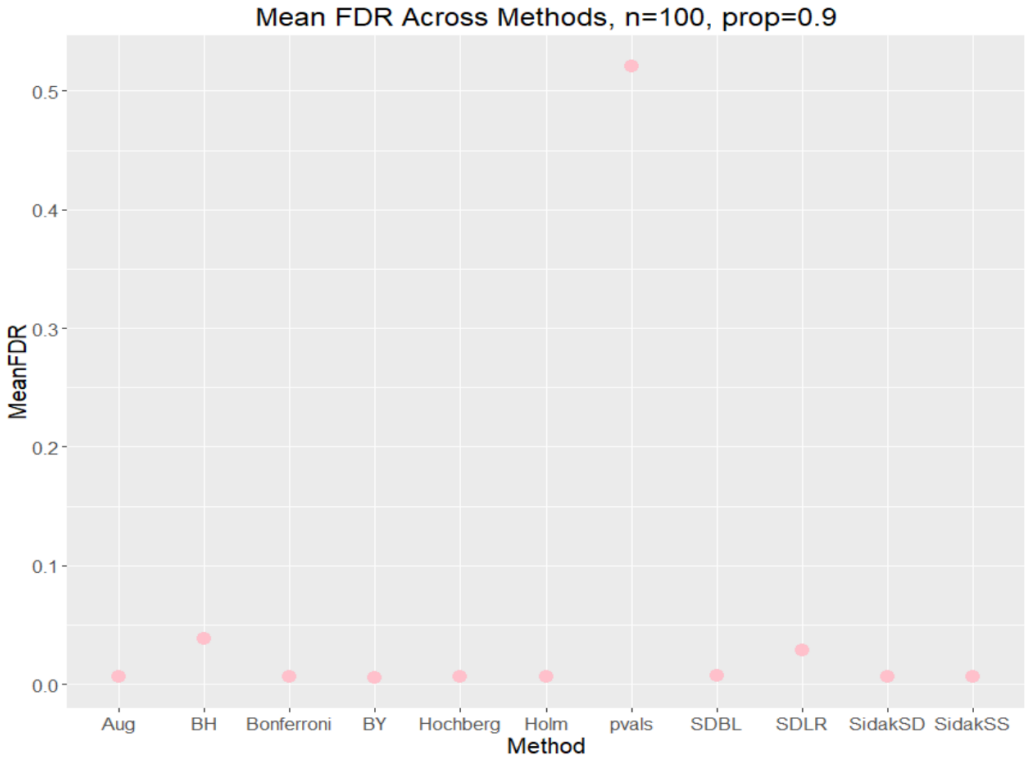


Figure 29: Mean FDR when n=5000, prop=0.9

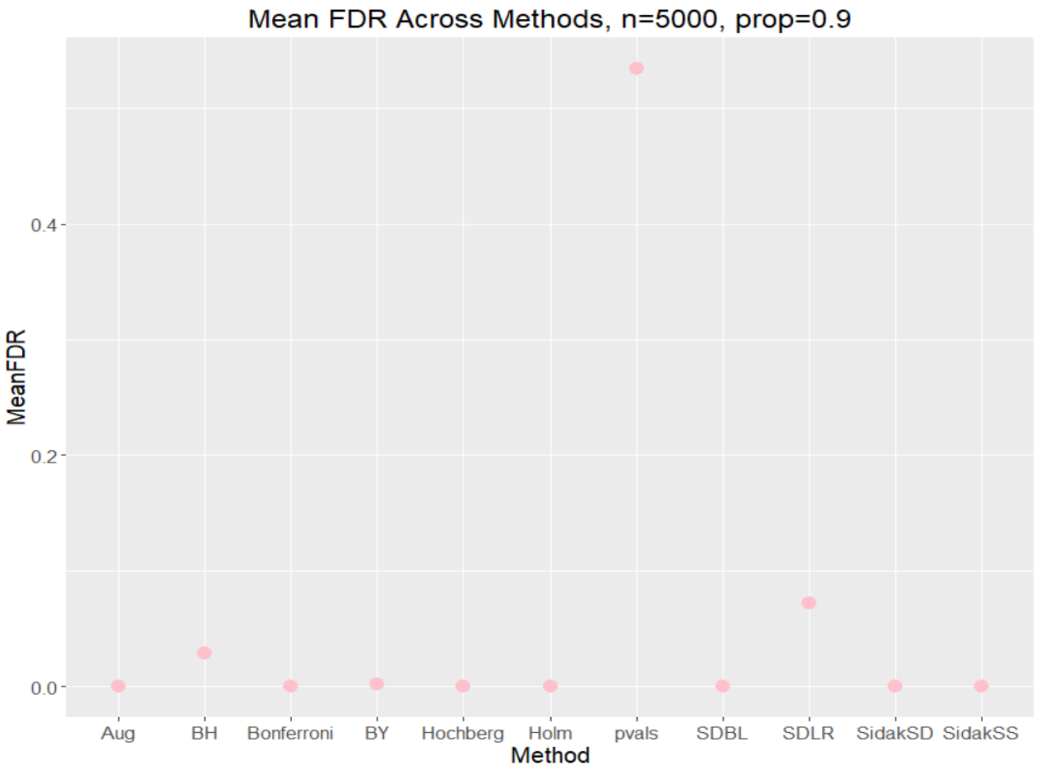


Figure 30: Mean FDR when n=100000, prop=0.9

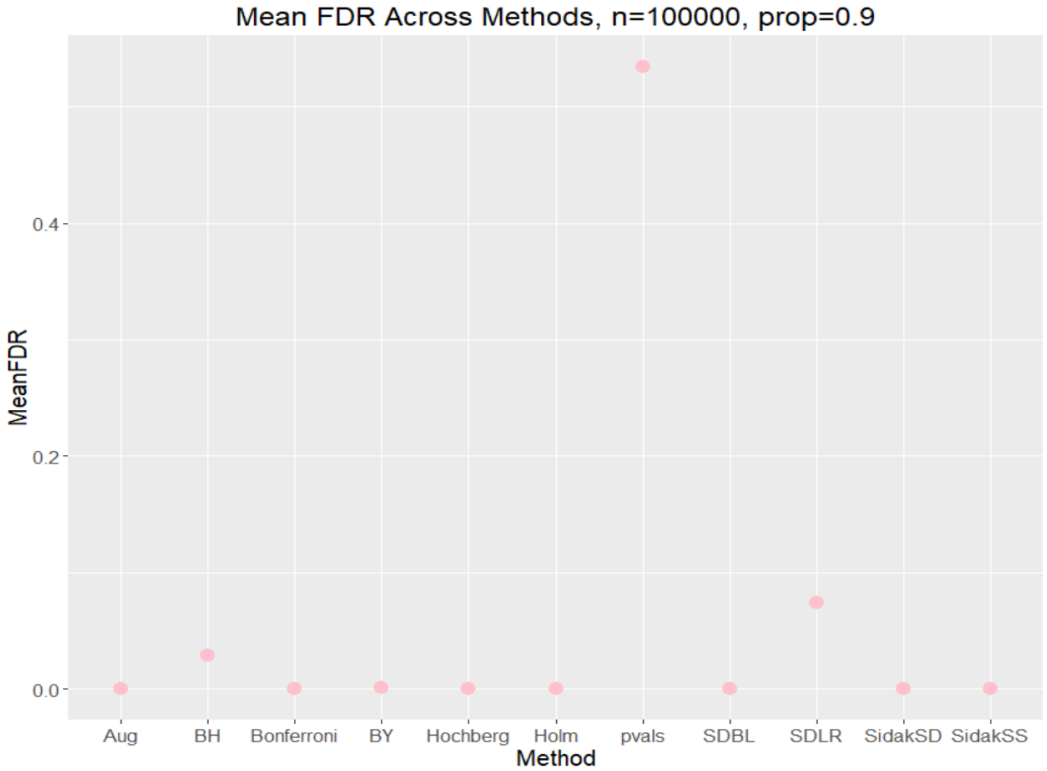


Figure 31: Mean FDX when n=100, prop=0.9

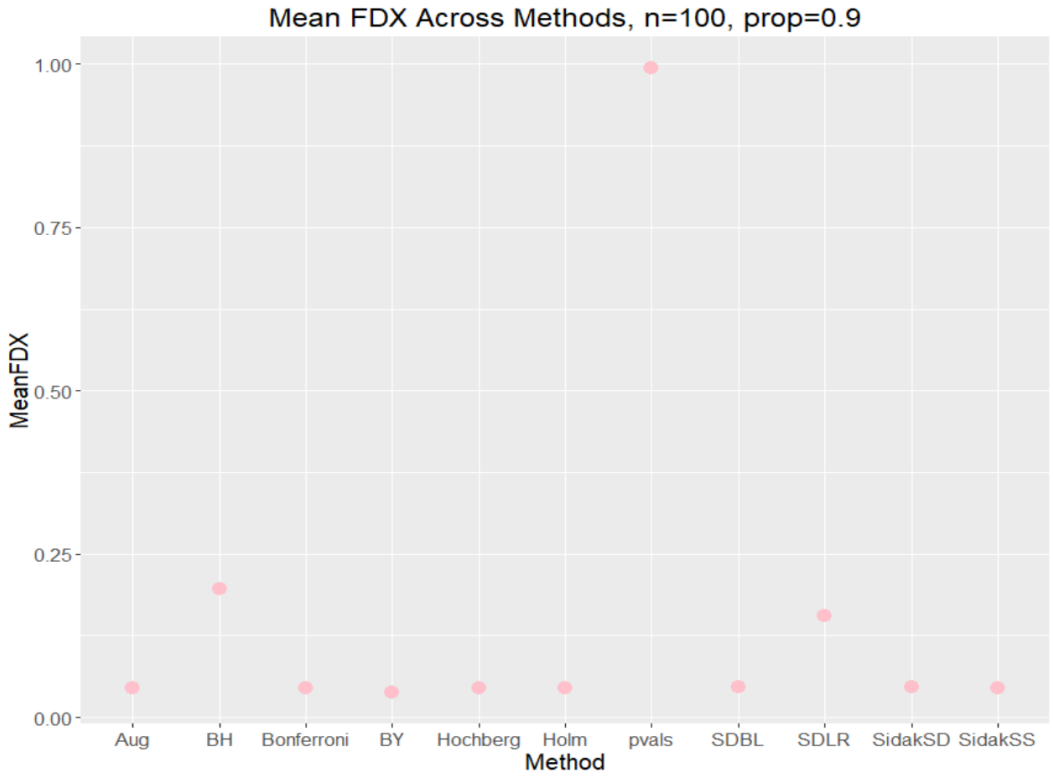


Figure 32: Mean FDX when n=5000, prop=0.9

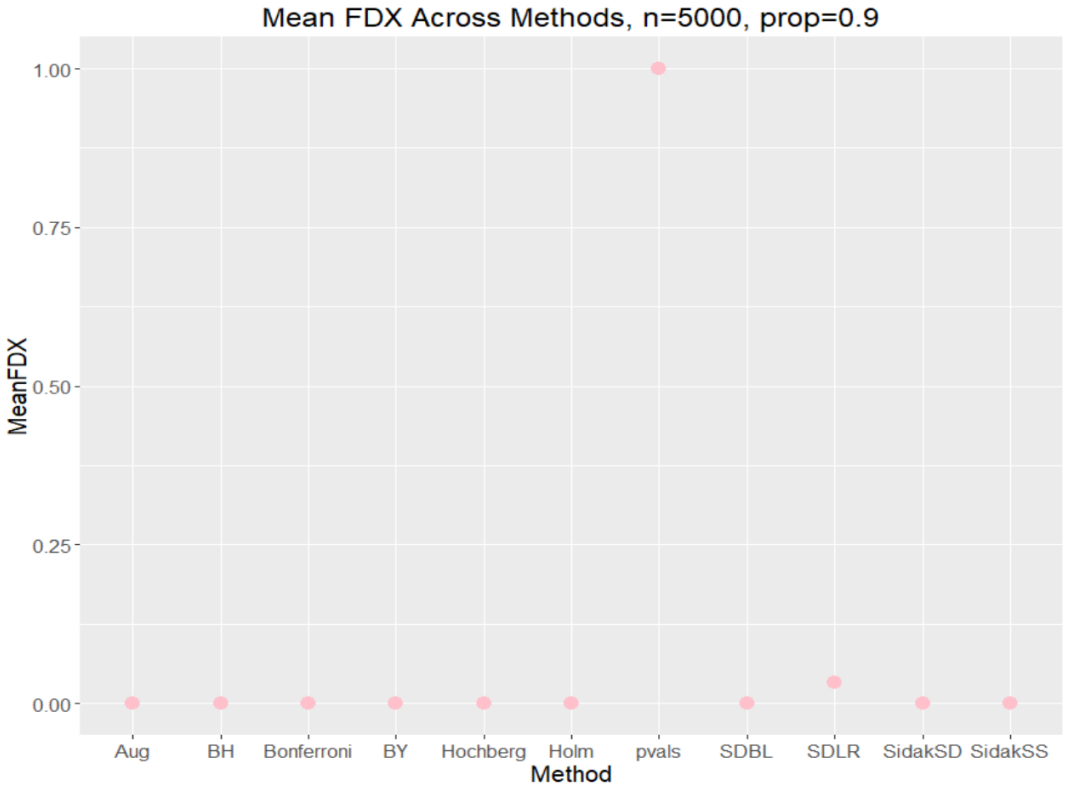


Figure 33: Mean FDX when n=100000, prop=0.9

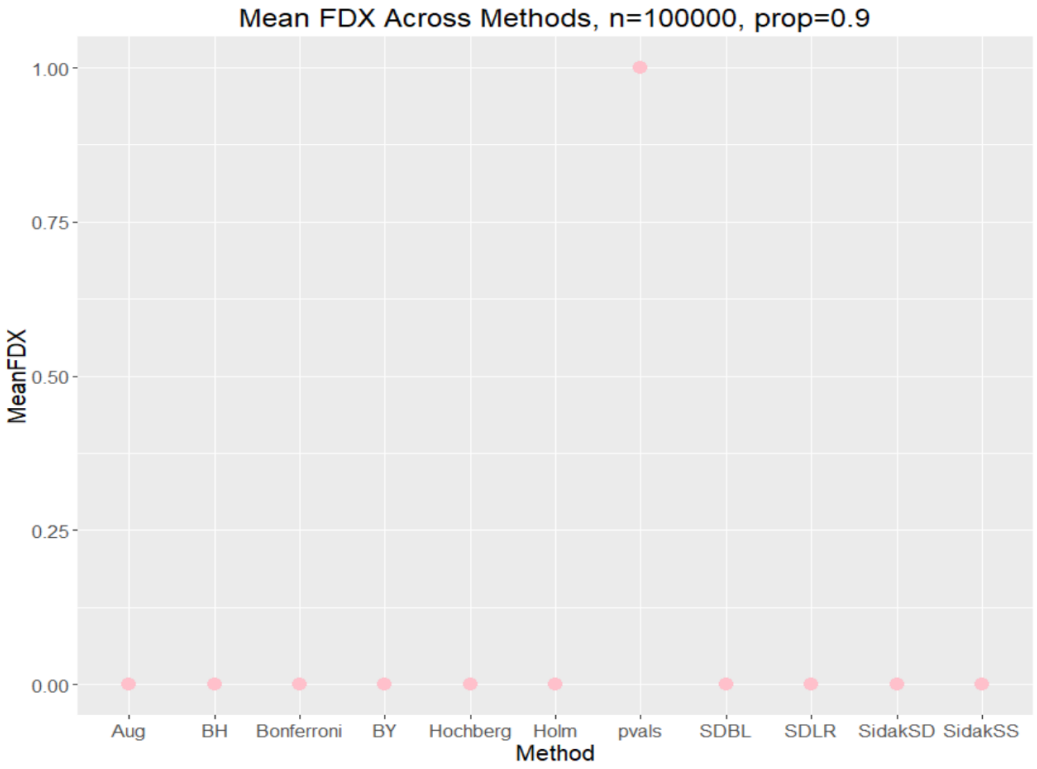


Figure 33: Mean FNR when n=100, prop=0.9

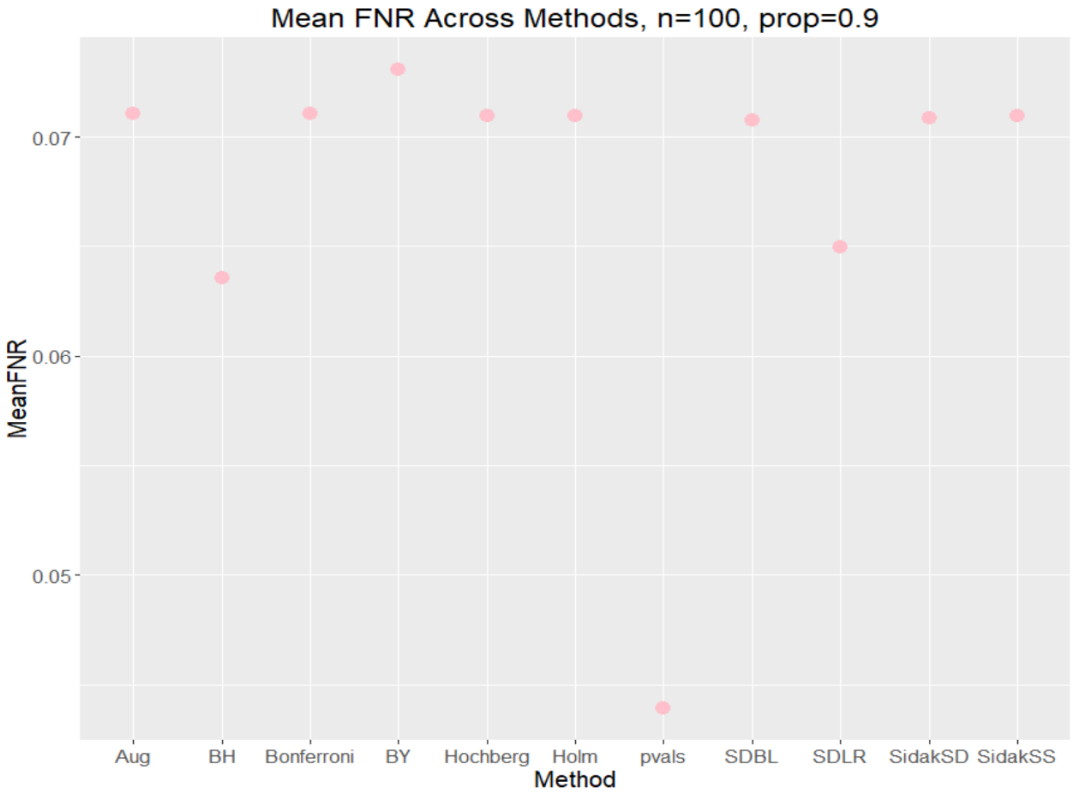
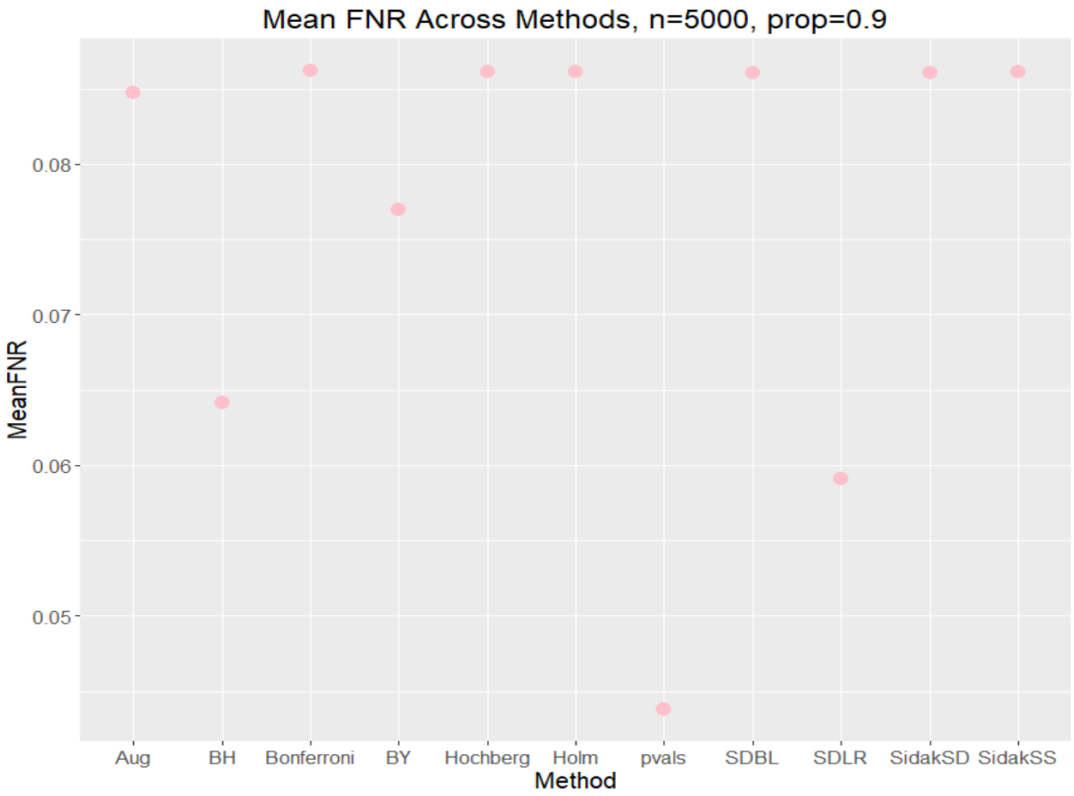


Figure 34: Mean FNR when n=5000, prop=0.9



VIII. References

1 Dudoit S, Shaffer PJ, Boldrick JC. Multiple hypothesis testing in microarray experiments. Statistical Science 2003; 18: 71–103.

2 van der Laan MJ, Dudoit S, Pollard KS. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Statistical Applications in Genetics and Molecular Biology 2004; 3(1).

3 Genovese CR, Wasserman L. Exceedance control of the false discovery proportion. Journal of the American Statistical Association 2006; 101: 1408–17.

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

5 Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Test 2003; 12: 1–77.

6 Pesarin F.Multivariate permutation tests with applications to biostatistics. Wiley, 2001.

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

8 Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986; 73: 751–54.

9 Benjamini Y, Liu W. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. Journal of Statistical Planning and Inference 1999; 82: 163–70.

10 Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference 1999; 82: 171–96.

11 Lehmann EL, Romano JP. Generalizations of the familywise error rate. Annals of Statistics 2005; 33: 1138–54.

12 Farcomeni, Alessio. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Statistical Methods in Medical Research 2008; 17: 347-388.