Clara Plaza Alonso

The problem of multiple testing

In the context of genomics, proteomics, and other omics technologies, scientists often simultaneously test numerous hypotheses to identify meaningful signals or associations. Therefore, multiple testing statistics play a crucial role in the realm of modern-day biological technologies, where advancements have enabled researchers to generate massive datasets with high-dimensional measurements.

In hypothesis testing, researchers set a significance level (α) which represents the threshold for statistical significance. The standard for judging α at 0.05 was first suggested by Fisher in 1925, who argued that one should reject a null hypothesis when there is only a 1 in 20 chance (5%) that it is true, which means p values lower than 0.05 should be considered as significant. In multiple testing, the possibility of rejecting the null hypothesis just by chance (not being true) increases due to the huge amount of data. These are called type I errors or false positives.

The contrary would be type II errors, also known as a false negatives, when a test fails to reject a false null hypothesis. The probability of committing a Type II error is called β. So, the power of a test would be defined as 1- β.

In multiple testing we obviously want the least false positives and negatives possible. To avoid type I errors, we can use α values that are much more conservative or involve a smaller chance that the null hypothesis is true, such as 0.01 or 0.001. However, the use of very conservative or stringent significance levels to test hypotheses lead to a loss of power and therefore an increase in the rate of false negatives. On the other hand, the use of significance levels that are too liberal lead to unacceptably high rates of false positives.

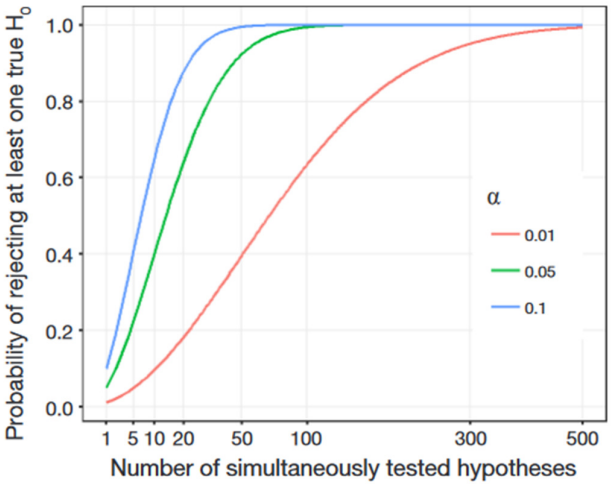

Figure 1. **-** The increased error rate of multiple comparisons. Figure extracted from J. Thorac. Dis. 2017 Jun 1;9(6):1725–9

In the end, the sheer volume of data generated by these technologies needs the application of statistical methods that can effectively control the rate of false positives, maintaining the integrity and reliability of scientific findings. Additionally, they need to take into account unique features of these types of analysis like how they are not independent from each other. Classical statistical models were never designed to handle multiple hypotheses and take all these problems into account. Therefore, numerous corrections have arisen to address this problem.

Two general types of multiple comparison procedures are used, one controlling family-wise error rates (FWERs) and the other controlling for false discovery rates (FDRs). The FWER methods tend to be more conservative and include the classical Bonferroni (and its different adjustments), sequential sampling methods, and resampling methods. The familywise approaches are based on the probability that one, or more of the rejected hypotheses, is true, while the FDRs control the expected number of false rejections from among the rejected hypotheses. Each investigator must determine which method is best for their data in order to strike a balance between type I and type II errors as there is not a consensual resolution.

Clara Plaza Alonso

1. - Family-wise error rates (FWERs)

The **Bonferroni method or correction** is the most common method for dealing with multiple testing problems and involves adjusting the significance level (α) of each test to accommodate the total number of tests performed.

Bonferroni correction can be used to correct 'experiment-wise' and 'family-wise' error rates in multiple comparisons. Experiment-wise error correction is where a large number of independent tests are performed employing basic statistical procedures such as 'Students' or Pearson's correlation coefficient and all tests are included. By contrast, family-wise error correction occurs when a smaller number of related group means are compared following analysis of variance (ANOVA), which is what happens in multiple testing.

The Bonferroni method controls for the overall type I error rate in a study (e.g., α = 0.05), such that the total of all tests conducted will together produce no more than a probability of for false positive results. In general the error rate will be:

$$1 - (1 - \alpha)/T$$

where 'α' is the significance and 'T' is the number of tests performed. In practice, an adjusted significance level of α /T is used as an approximation to. There if 'T' tests are performed, then the Bonferroni-corrected α for each individual test is α /T. Thus, to attain an overall significance level of 0.05 for 10 tests, each individual test must reach an α level of (0.05/10)=0.005.

Despite the widespread use of the Bonferroni method, there has been continuing controversy regarding its use. Different drawbacks are suggested by experts:

1. **Focus on the Wrong Hypothesis:** the method often tests the 'universal' null hypothesis, checking if groups are identical in all comparisons, which may be of little relevance to researchers interested in the significance of individual tests.
2. **Interpretation Dependency on Other Tests:** the interpretation of a single test is argued to depend on the number of other tests performed, suggesting that conclusions drawn should not be altered based solely on the number of additional tests.
3. **Trade-off between Type I and Type II Errors:** the probability of a Type I error cannot be decreased without increasing the risk of a Type II error, potentially leading to the failure to detect real differences. As the number of tests increases, the adjusted p-value for statistical significance decreases, lowering the test's power. For example, a Bonferroni correction for a GWAS study involving around 500,000 tests corresponds to a p value of 10^{-7}. While these α -levels certainly will provide a safeguard against type I errors, they also will lead to an unacceptably type II error rate. Therefore this method works well when there are only a few tests are being performed (10-20 tests) but not for those involved a lot.
4. **Unclear Scope for Correction Application:** questioning what constitutes the population of tests for which the correction should be applied, such as all tests in a report, a subset of them, tests performed but not included in the report, or tests from the same data included in other reports.

Different adjustments to the Bonferroni correction raised due to the need to address these and other limitations and challenges associated with this correction method.

The **Holm Adjustment** is an extension of the classical Bonferroni correction. It is a procedure that adjusts the significance level for each individual test based on its rank or order of significance. The most significant test receives the strictest adjustment, while subsequent tests receive less stringent adjustments. This method provides a balance between controlling the family-wise error rate and maintaining power in the analysis. It works as follows:

1. All p-values are sorted from smallest to largest, with K as the number of p-values.
2. If the first p-value is greater than or equal to α/K, the procedure is stopped and no p-values are significant.
3. If the first p-value is declared significant, the second p-value is compared to α/(K-1). If the second p-value is greater than or equal to α/(K-1), the procedure is stopped and no further p-values are significant. Otherwise, we go on until the i-th ordered

   p-value is such that:

$$p_{(i)} \geq \alpha/(K - i + 1)$$

Bonferroni-Holm's great advantage is that it is sequentially rejective.

Similar to Holm's, the **Hochberg Adjustment** computes significance levels in a stepwise manner. It adjusts the significance level for each test, but in a slightly different manner. The Hochberg method begins by sorting the p-values in ascending order and it conducts statistical inference starting with the largest p-value, stopping when the adjusted level is exceeded. Therefore when we first observe $p_{(i)} < \alpha/(K - i + 1)$ the comparison stops. Generally is considered more powerful than Holm's.

Another one would be the **Hommel Adjustment.** Originally it was an extension of the Simes' Adjustment to Bonferroni's Correction. The Simes' procedure, originally developed in 1986, was designed for global tests, and Hommel adapted it for individual tests. For each hypothesis the p-value an index is calculated. All hypotheses are rejected if the index does not exist. If it does exist, all the hypotheses with p-values lower than the index are rejected.

Simulation-based techniques

This type of techniques compare observed p-values with p-values calculated from simple repeated perturbations of the data. There are three main methods; jackknife, bootstrap, and permutation tests.

The **jackknife method** constructs an empirical distribution of a relevant test statistic computed with the actual data though the use of subsets of the data. In the context of multiple testing, the jackknife method can be applied to assess the stability and variability of p-values.

After performing the multiple testing, we obtain a set of p-values. To assess the stability of our findings, we create jackknife samples by excluding one observation at a time from the dataset. In other words, the method leaves out the first sample observation to create the first jackknife sample, then leaves out the second observation to create the second jackknife sample and so on to finally leave out the last observation to create the n-th jackknife sample. We then rerun the hypothesis tests for each jackknife sample, resulting in a set of modified p-values. If a result consistently persists in appearing across most jackknife samples this suggests robust findings. However, significant variation when certain observations are excluded indicates potential instability.

The **bootstrap method** draws randomly with replacement from a set of data points to estimate the empirical distribution of a test statistic. It is very similar to jackknife, but in this case bootstrap involves randomness due to the sampling with replacement, while jackknife does not involve randomness as it systematically creates subsamples (as previously described, first we create a subsample without the first data and so on). As bootstrap samples involve replacement, it allows repeated inclusion of data points, while the jackknife systematically creates subsamples by omitting one observation at a time.

Although the jackknife and bootstrap often produce similar outcomes, jackknife is preferable for handling intricate sampling designs whereas bootstrap is generally more computationally efficient when dealing with large samples. The jackknife and bootstrap may often yield similar results, but bootstrap gives slightly different results when repeated on the same data, whereas jackknife gives exactly the same result each time. This is because bootstrap involves drawing random samples from the sample observations while jackknife does not involve any randomness.

Permutation tests estimate the distribution of a test statistic by exchanging data points among the units of observation. In genetic applications, the permutation test is the most widely used resampling method. Initially, the multiple testing is performed and its corresponding p-values are computed using the original dataset. Subsequently, the dataset undergoes random permutation n times, and test statistics are recalculated for each permuted dataset. Next, a permuted distribution is constructed by counting whenever the p-value in the original dataset is smaller than the p-value obtained from the permuted datasets. This count is divided by the total number of random permutations, generating the permutation probability.

This kind of testing, while computationally demanding and requiring advanced programming has an incredible robustness, effectiveness in maintaining good type I error rates and power. Importantly, the significance threshold is directly derived from the data, ensuring that the null distribution aligns with the original data's characteristics, including normality, allele frequencies, and missing data patterns.

2. - False Discovery Rate (FDR)

All the adjustments before were focused in controlling the FWER but other approaches raised where they focused in controlling FDR. The FDR-based control is less stringent with the increased gain in power and has been widely used in cases where a large number of hypotheses are simultaneously tests.

The **Benjamini-Hochberg (BH) Adjustment** determines a pre-specified false discovery rate (Q) and computes the Benjamini-Hochberg's critical value. The process is as follows:

1. Put the individual p-values in ascending order.
2. Assign ranks to the p-values. For example, the smallest has a rank of 1, the second smallest has a rank of 2.
3. Calculate each individual p-value's Benjamini-Hochberg critical value, using the formula (i/K)Q, where:
    - i = the individual p-value's rank
    - K = total number of tests
    - Q = the false discovery rate (a chosen percentage)
4. Compare your original p-values to the BH critical value
5. Find the largest p-value that is smaller than the critical value.

6. **All** values above it (i.e. those with lower p-values) are considered significant, even if those p-values are lower than the critical values.

The **Benjamini and Yekutieli (BY) Adjustment** is an extension of the Benjamini-Hochberg's (BH) method and is particularly useful when there may be dependencies or correlations among the tests. The main difference between BH and BY is that we calculate the false discovery rate, considering the number of tests, the distribution of null p-values, and potential dependencies among the tests.

3. – FWER and FDR

FWER management aims to minimize the probability of making any Type I errors across all comparisons, resulting in a more conservative approach that adjusts p-values to maintain a stringent significance threshold. On the other hand, FDR control is more permissive, allowing for a higher proportion of false positives among rejected hypotheses. Consequently, FDR-controlling procedures adjust p-values in a way that balances the identification of true signals with the acceptance of some false positives, offering a more flexible framework for hypothesis testing in large-scale analyses. The choice between FWER and FDR adjustment depends on the research objectives, emphasizing the trade-off between the stringent control of overall error rates and the potential discovery of true associations amidst multiple comparisons.

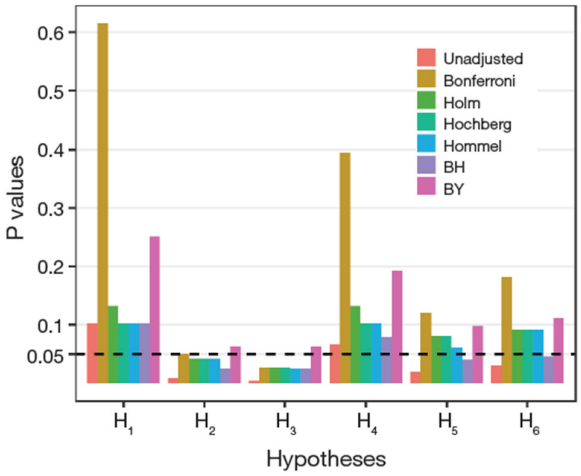

Figure 2. **-** Differences of the adjusted P values among various methods. The dashed horizontal line denotes the pre-specified significance level. Figure extracted from J. Thorac. Dis. 2017 Jun 1;9(6):1725–9

4. – Other methods

Other approaches have surfaced to tackle the multiple testing issue. One example is the Non-Discovery Rate (NDR), which tells us the proportion of missed true positives. Another approach involves using sequential sampling methods, where data is collected and analysed in stages. The idea here is to split the sample into two groups, one for exploring the data and the other for confirming findings. This helps verify signals and keep the number of identified signals under control

5. – Conclusions

The progress in molecular genetic technologies has led to an abundance of data that frequently surpasses researchers' capacity to draw meaningful conclusions. To navigate this challenge, there is a demand for innovative statistical methods and perspectives to effectively interpret all of this data.

6. – Bibliography

1. Streiner DL, Norman GR. Correction for multiple testing. Chest [Internet]. 2011 Jul 1;140(1):16–8. Available from: https://doi.org/10.1378/chest.11-0523
2. Rice T, Schork NJ, Rao DC. Methods for handling multiple testing. In: Advances in Genetics [Internet]. 2008. p. 293–308. Available from: https://doi.org/10.1016/s0065-2660(07)00412-9
3. Armstrong RA. When to use the Bonferroni correction. Ophthalmic and Physiological Optics [Internet]. 2014 Apr 2;34(5):502–8. Available from: https://doi.org/10.1111/opo.12131
4. Giacalone M, Zirilli A, Cozzucoli PC, Alibrandi A. Bonferroni-Holm and permutation tests to compare health data: methodological and applicative issues. BMC Medical Research Methodology [Internet]. 2018 Jul 20;18(1). Available from: https://doi.org/10.1186/s12874-018-0540-8
5. Chen S, Feng Z, Xiaolian Y. A general introduction to adjustment for multiple comparisons. Journal of Thoracic Disease [Internet]. 2017 Jun 1;9(6):1725–9. Available from: https://doi.org/10.21037/jtd.2017.05.34
6. Stephanie. Benjamini-Hochberg procedure - statistics how to [Internet]. Statistics How To. 2021. Available from: https://www.statisticshowto.com/benjamini-hochberg-procedure/
7. Sinharay S. Jackknife methods. In: Elsevier eBooks [Internet]. 2010. p. 229–31. Available from: https://doi.org/10.1016/b978-0-08-044894-7.01338-5
8. Camargo A, Azuaje F, Wang H, Zheng H. Permutation – based statistical tests for multiple hypotheses. Source Code for Biology and Medicine [Internet]. 2008 Oct 21;3(1). Available from: https://doi.org/10.1186/1751-0473-3-15