Multiple Testing

Zitong "Cathy" Qi

Mentor: Anna Neufeld

1. Introduction

1.1. Multiple Testing

Classical statistics focuses on testing a single null hypothesis H_0, where p-value evaluates how extreme is our observed statistic, assuming the null hypothesis is true. Then, we choose a cutoff called alpha, α. If p-value is less than α, we reject the null and we call the result statistically significant. When we conclude that the treatment and the control groups are different, even though in reality they are the same, we make a type I error.

However, things are complicated in the real world. Instead of testing one null hypothesis each time, scientists often perform multiple testing on large data. Multiple testing refers to any instance that involves the simultaneous testing of more than one hypothesis. What if we want to test every week as we recruit new patients to the trial? When scientists want to conduct multiple tests all at once, or when they want to follow an experiment over time, the probability of making a type I error on each individual test is fine. But the probability of making at least one type I error is very high. Notice that α is the probability of making a Type I error when conducting only one single test. Thus, adjustments needed for multiple testing procedure to maintain the overall type I error rate.

1.2. Interim Analysis

When doing hypothesis testing, suppose their is truly no different between the treatment and control group. If we allow ourselves to look at the trial every day to conduct a test for the difference, eventually we will find a day where the treatment and control group are different with a p-value less than α, 0.05. So we should not allow ourselves to look at the data over and over again without penalty.

On the other hand, if a treatment works extremely well and the evidence is clear early on, we want to be able to stop the trial! We should not waste resources, and if the treatment is lifesaving we should not deny treatment to the control group.

1.3. Family-Wise Error Rate

Instead of trying to control the probability of making a false discovery, type I error in a single test, in multiple testing with many null hypotheses, we try to control the probability of making at least one type I error, false positives, among all null hypotheses.

For one single test, there are four different possibilities with the hypothesis being true and false and our decision of rejecting or failing to reject.

	H_0 is True	H_0 is False	Total
Reject H_0	V	S	R
Do Not Reject H_0	U	W	$m - R$
Total	m_0	$m - m_0$	m

When it comes to multiple testing, Family-Wise Error Rate is represented by $Probability(V \geq 1)$. We can control the FWER, making at least one type I error among all tests, when for each hypothesis test, probability of making a type I error is alpha α. Calculated by $FWER(\alpha) = 1 - (1 - \alpha)^m$. While assuming FWER is less than or equals to alpha.

1.4. α-spending functions

If we want to do interim analyses, and we also want to keep the Probability of making at least one type I error rate, the family-wise error rate to α, we need group sequential boundaries where we define a critical value at each interim analysis. To do this, we need to allocate α over k interim analyses. This is exactly the idea of α-spending functions. It treats α, the cutoffs, as an "increasing" function, denoted by $\alpha(t_r)$. Where t_r represents the information fraction, determined by the information observed at t and total information expected at the scheduled terminal, has nothing to do with the actual time. The t_r is calculated by the $\frac{\text{current sample size}}{\text{total sample size}} \cdot$ time(whole trial is 1, half of the trial is 0.5). Notice that t_r is always between 0 and 1.

We are going to analyze three different alpha-spending functions to control the FWER.

The Bonferroni Correction is a very general method that can be used in any multiple testing scenario. which sets the threshold for rejecting each hypothesis to $\frac{\alpha}{m}$, where m represents the total null hypotheses performed and α represents the probability of making at least one false discoveries among all m tests. For example, in order to control the FWER at 0.05, when testing m = 50 null hypotheses, it means to control each null hypothesis at level $\alpha = \frac{0.05}{50} = 0.001$. Such method successfully controls the FWER. However, while making sure we don't falsely reject too many null hypotheses, rejecting few null hypotheses makes more type II errors (when nulls are false, we fail to reject), so power is low. We conclude that the Bonferroni Correction is neither the most powerful nor the best approach, but it is very easy to compute.

To be more specific, let's look at functions for sequential methods that take advantage of the fact that we're using some of the same patients to calculate a p-value at each time point, and we're just adding more patients over time. Therefore, the p-values you compute at each interim analysis are dependent and we can exploit this dependence by using O'Brien and Fleming and Pocock approaches.

O'Brien and Fleming approach uses more conservative stopping boundaries at early stages. Puts more priority on power at the end of the study. Its cutoff over time can be represented by an increasing model, which has larger thresholds as we perform more analyses and make it hard to reject the null at the beginning of the test.

Pocock approach uses the same significance level at each interim analysis. Puts more priority on being able to stop early. Its cutoff over time can be represented by a flat model.

2. R simulation

2.1. Setting

We are going to use R simulation to verify power and interim analyses properties of Alpha-spending functions and compare those properties between no correction, Bonferroni's Correction, O'Brien & Fleming method, and Pocock method. We are going to sequentially monitor trials both under the null(same mean for treatment and control group) and under the alternative hypotheses(there is some treatment effect between the drug and the placebo).

Assuming we have 100 people in treatment and 100 people in control group. The scheduled terminal is 10 analyses. The cutoff is computed automatically from a shiny app:

No correction's cutoff: (0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05)

Bonferroni's cutoff: (0.05/10, 0.05/10, 0.05/10, 0.05/10, 0.05/10, 0.05/10, 0.05/10, 0.05/10, 0.05/10, 0.05/10)

O'Brien and Fleming's cutoff: (0.0001, 0.0001, 0.0001, 0.0010, 0.0032, 0.0071, 0.0126, 0.0197, 0.0279, 0.0369)

Pocock's cutoff:(0.0106, 0.0106, 0.0106, 0.0106, 0.0106, 0.0106, 0.0106, 0.0106, 0.0106, 0.0106)

2.2. Null hypothesis is true

When the null hypothesis is true, if we reject the null, we make a type I error and stop the experiment. We repeat this process 5000 times. The mean of making a type I error among 5000 trials is the Family-Wise Error Rate and we are also going to look at the average stopping point of interim analyses among trials where we stopped.

	FWER(probability of stopping the trial and concluding treatment and control are different)	K(average stopping time, among trials where we stopped)
No correction	0.20	3.79
Bonferroni	0.02	4.20
O'Brien & Felming	0.05	8.33
Pocock	0.05	4.24

When null is true, our goal is to control the FWER to be below alpha which is 0.05. Alpha-spending functions: Bonferroni, O'Brien & Fleming, Pocock successfully did that. Bonferroni's FWER is much lower than alpha, which is good as it means fewer type I mistakes, but it will hurt the power as it is too conservative and makes it almost impossible to reject any of the false null hypotheses. Bonferroni and Pocock has similar average stopping time since they have similar constant thresholds. The average stopping time for O'Brien & Fleming is much larger, it makes sense since it has increasing thresholds that make it hard to reject at the beginning.

2.3. Null hypothesis is false

When null is false, if we reject the null, we stop the experiment. We repeat this process for 500 times. The mean of rejecting the null is the power of the method.

Figure 1:

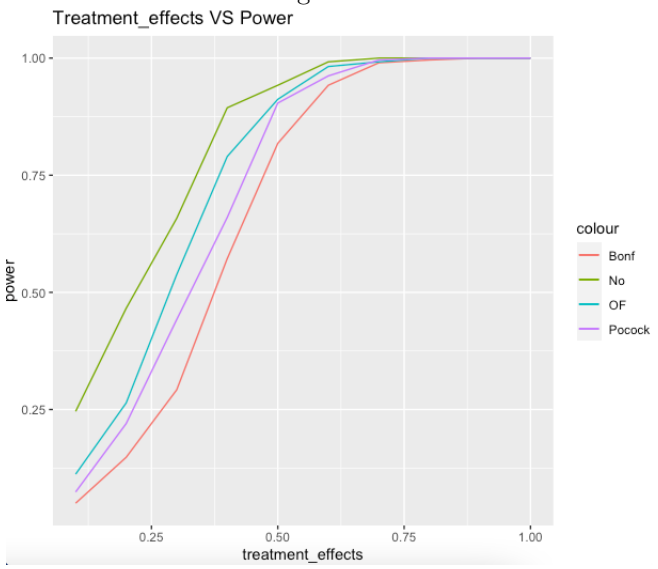
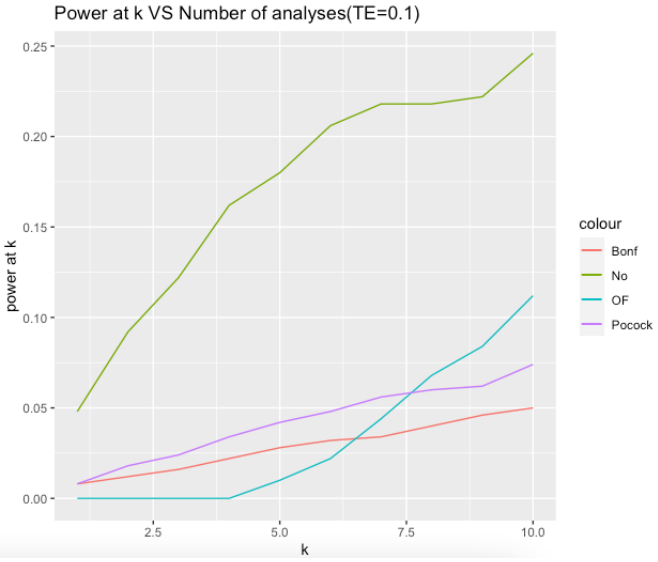


Figure 2:



From the above figures, we can conclude that no correction has the highest power but we don't care because its trade-off is that it does not control the type I error rate at all. Pocock's power is higher than Bonferroni because it exploits dependence. O'Brien & Fleming has lower power than Pocock at the beginning than higher later, this makes sense because the cutoff for O'Brien & Fleming is pretty low at first as an increasing function, so it makes it hard to reject the null, thus resulting in lower power.

3. Extensions

3.1. The Marginal False Discovery Rate

In the R simulation, we see that the Pocock procedure is more powerful than the Bonferroni Correction as it exploits dependence. Can we do even better by exploiting the dependence more?

While controlling the FWER is too conservative and makes it almost impossible to reject any of the false null hypotheses, it has extremely low power. Instead we can tolerate a few type I errors while making sure most of the rejected null are not false positives, type I errors. In other words, reject as many null hypotheses as possible while guaranteeing no more than α percent of those rejected null are false positives. This is the process of controlling marginal false discovery rate.

3.2. Alpha-investing functions

When we sequentially monitor trials with multiple testing, in order to control the marginal false discovery rate, we need alpha-investing functions. We are going to compare their properties with the alpha-spending functions.

Alpha-spending functions have fixed boundaries that depend on number of planned analyses and the initial alpha, so you can tell all the cutoffs before actually performing the test. However, alpha-investing functions have more advanced boundaries that can be changed based on results of previous tests, so you can't tell all cutoffs in advance until you actually start the test.

The goal for alpha-spending functions is to control the probability of making at least one type I error, the family-wise error rate, while the goal for alpha-investing functions is to control a rate that depends on number of all rejected null, and number of rejected true nulls, the marginal false discovery rate.

References

[1] James, Gareth, et al. An Introduction to Statistical Learning: With Applications in R.

[2] Multiple Testing - University of Chicago.
https://home.uchicago.edu/amshaikh/webfiles/palgrave.pdf.

[3] Foster, Dean P., and Robert A. Stine. "A-Investing: A Procedure for Sequential Control of Expected False Discoveries." Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 70, no. 2, 2008, pp. 429–444.

[4] Demets, David L., and K. K. Lan. "Interim Analysis: The Alpha Spending Function Approach." Statistics in Medicine, vol. 13, no. 13-14, 1994, pp. 1341–1352.

[5] Lark, R. M. "Controlling the Marginal False Discovery Rate in Inferences from a Soil Dataset With α-Investment." European Journal of Soil Science, vol. 68, no. 2, 2017, pp. 221–234.