Bachelor oppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for matematiske fag

Bård Fineid

Controlling FWER

A comparative study of the procedures of Bonferroni, Holm, Hochberg and Hommel

Bacheloroppgave i Matematiske fag
Veileder: Øyvind Bakke
Juni 2023

NTNU
Kunnskap for en bedre verden

Bård Fineid

Controlling FWER

A comparative study of the procedures of Bonferroni, Holm, Hochberg and Hommel

Bacheloroppgave i Matematiske fag
Veileder: Øyvind Bakke
Juni 2023

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for matematiske fag

NTNU

Kunnskap for en bedre verden

Abstract

The multiple testing procedures of Bonferroni, Holm (1979), Hochberg (1988) and Hommel (1988) are shown to have strong control of the family-wise error rate (FWER) at level α by combining the global tests of Bonferroni and Simes (1986) with the closed testing procedure proposed by Marcus et al. (1976). The procedures of Bonferroni and Holm are valid for all p-value dependency structures and are thus conservative for all instances, except for the unrealistic "worst case" scenario. In contrast, Hochberg and Hommel assume positive dependence through stochastic ordering (PDS), allowing them to ignore the "worst case" dependency structure, resulting in less conservative procedures. Finally, we apply the four procedures to independent, positive dependent and negative dependent tests, and compare the empirical results with the theoretical conclusions.

Additionally, we discuss p-values, with a particular emphasis on the distribution of exact and valid p-values associated with true one-sided hypotheses.

Table of Contents

List of Figures

List of Tables

1 Introduction

In many fields of study, researchers are often interested in testing multiple hypotheses simultaneously, or on the same data over time. A biologist may be interested in testing multiple genes related to a specific phenotype, whereas a finance worker might want to investigate the relationships between different economic variables and stock prices. Multiple testing had computational limits before the computer, however, in the 21st century, with its super-fast computers and access to large amounts of data, these limits are practically non-existent. This enables researchers to conduct hundreds or thousands of tests within minutes. Consequently, the risk of rejecting a true hypothesis, namely a false positive, has become an increasingly frequent problem. To address this issue, multiple testing procedures have been developed to control the rate of false positives, while simultaneously rejecting as many false hypotheses as possible. One such controlling mechanism is the family-wise error rate (FWER), which refers to the probability of at least one false positive result. FWER controlling procedures aim to control the FWER at a predetermined level α, allowing valid inferences to be made regarding the individual hypotheses. The study of FWER controlling methods has been active for decades, and several procedures have been proposed. The most well-known arguably being the procedures of Bonferroni, Holm (1979), Hochberg (1988) and Hommel (1988).

The aim of this thesis is to provide a comparative study of these procedures. Specifically, we demonstrate how the procedures control FWER both theoretically and empirically, and how they differ in power. Relevant mathematical concepts, such as hypothesis testing, p-values, and conservatism, are introduced, together with relevant inequalities and assumptions. In conclusion, we discuss the strengths and weaknesses of FWER control in general and possible alternative methods for controlling the rate of false positives.

2 Theory

This thesis begins by introducing the fundamental theory of hypothesis testing. That involves the mathematical structure of the null and alternative hypotheses, test statistic, and level of significance. We then explore how these concepts can be extended to multiple testing, and the challenges that arise. Additionally, we introduce the concept of global testing, and how it combined with the closed testing procedure by Marcus et al. (1976) build a framework for the procedures of Holm, Hochberg and Hommel.

2.1 Hypothesis testing

To test a hypothesis means to formulate a statement about a particular phenomenon, then collecting and analyzing data in order to evaluate whether that statement is supported by the evidence. Mathematically, we use statistical methods to determine the likelihood of observing a particular result, or set of results, under a specific assumption. A hypothesis test involves a null hypothesis H_0 (as the specific assumption) and a alternative hypothesis H_1, where we look to reject H_0 in favor of H_1. We gather a random sample $\boldsymbol{X}$ from some distribution P_θ, where $\theta \in \Omega$ is the parameter of interest, and Ω is the parameter space. We formulate our null hypothesis as $H_0 \colon \theta \in \omega$ and the alternate hypothesis $H_1 \colon \theta \in \omega^c$, where $\omega \subseteq \Omega$, and ω^c represent the complement of ω. Further, a true hypothesis refers to the null hypothesis being true, while a false hypothesis refers to the null hypothesis being false.

We primarily focus on simple one- and two-sided tests. A one-sided test checks if a parameter is greater or less than a certain value, while a two-sided test checks if the parameter is not equal to a certain value. Such tests can also be referred to as right-, left-, or two-tailed.

Example 1. Testing the average height of a population. We assume the data $\boldsymbol{X}$ is height measured in cm, extracted from a Gaussian distribution where the parameter of interest is the mean μ. We set a null hypothesis $H_0 \colon \mu \leq 180$ against the alternative hypothesis $H_1 \colon \mu > 180$. Using the above notation this corresponds to the parameter of interest θ being μ, our parameter space $\Omega = \mathbb{R}$, and $\omega = (-\infty, 180]$. This is an example of a simple one-sided test. For the test to be two-sided we can reformulate to $H_0 \colon \mu = 180$ against $H_1 \colon \mu \neq 180$, making $\omega = \{180\}$.

To further evaluate a hypothesis, we need some statistical measurement to summarize the difference between the data and what would be expected assuming the null hypothesis is true. The statistic measure is referred to as a test statistic, and one such test statistic is the p-value. The p-value most commonly represents the probability of a test result being at least as extreme as the observed data, assuming the null hypothesis to be true. Note that this is not the same as the probability that the null hypothesis is true, which is a common misinterpretation. "Extreme" is defined in terms of the alternative hypothesis. For our height example, the p-value may represent the probability of our observed average height or higher, assuming that the true average height is 180 cm. However, it is worth noting that the p-value does not necessarily represent a probability. Casella & Berger (2002) define a valid p-value as in Definition 1.

Definition 1. A p-value $p(\boldsymbol{X})$ is a test statistic satisfying $0 \leq p(\boldsymbol{X}) \leq 1$ for every sample point $\boldsymbol{x}$. Small values of $p(\boldsymbol{X})$ give evidence that H_1 is true. A p-value is *valid* if, for every $\theta \in \omega$ and every $0 \leq \alpha \leq 1$,

$$P_\theta(p(\boldsymbol{X}) \leq \alpha) \leq \alpha. \tag{1}$$

When evaluating a hypothesis H_0 there is a possibility of two main errors, type I or type II. A type I error is rejecting a true hypothesis, and a type II error is failing to reject a false hypothesis. Using a valid p-value $p(\boldsymbol{X})$ we can construct a test that controls for the probability of a type I error at a set level α. This test rejects H_0 if and only if $p(\boldsymbol{X}) \leq \alpha$, as the probability of rejecting H_0, assuming H_0 is true, is equal to the probability of $p(\boldsymbol{X}) \leq \alpha$, which from (1) is less than or equal to α. We often refer to α as the level of significance, which represents an upper bound for the probability of rejecting H_0, assuming H_0 is true. This is equivalent to the upper bound for the probability of a type I error, which becomes highly relevant when testing multiple hypotheses.

2.2 Multiple hypothesis testing & FWER

In single hypothesis testing, controlling the probability of a type I error is straightforward. For multiple hypotheses however, a problem quickly occurs. Assume that we simultaneously perform m independent tests with associated valid p-values $p_1, \dots, p_m$ associated with true hypotheses. We set a significance level α, and let V denote the number of false positives. We then calculate the probability of at least one false positive as

$$\begin{aligned} P(V > 0) = 1 - P(V = 0) &= 1 - P(p_1 > \alpha \cap \cdots \cap p_m > \alpha) \\ &= 1 - P(p_1 > \alpha) \cdots P(p_m > \alpha) \\ &\leq 1 - (1 - \alpha)^m. \end{aligned}$$

For $m = 15$ and $\alpha = 0.05$, the probability of at least one false positive is $P(V > 0) \leq 1 - (1 - 0.05)^{15} = 0.537$. Therefore, for only 15 tests, we have a 54% probability of rejecting at least one true hypothesis. By increasing the number of tests to 100, we obtain a 99.4% probability. An interesting question is then, how can we best control the probability of at least one false positive at an assigned level α? This is a known problem in multiple testing, and will be the main objective of the procedures introduced later.

We denote the hypotheses of interest by $\mathcal{H} = \{H_1, \dots, H_m\}$, where $m_0 \leq m$ of the hypotheses in $\mathcal{H}$ are true. We formally define the probability of obtaining at least one false positive as the family-wise error rate (FWER), which we wish to control at a level α, while simultaneously rejecting as many false hypotheses as possible. The simplest example of FWER control is to set $\alpha = 0$. This would not reject any hypotheses, thus not obtaining any false positives. However, allowing a number of false negatives. Therefore, for procedures controlling FWER, a concept called conservatism is frequently discussed. A conservative method controls the FWER, but in a strict sense, which in practice translates to it being harder than necessary to reject a individual hypothesis. This is a problem because we want to reject as many false hypotheses as possible. Moreover, regarding FWER, we discuss weak and strong control. A procedure is said to have weak control if it controls the FWER when all the hypotheses in $\mathcal{H}$ are true, i.e., $m_0 = m$. Meanwhile, a method with strong control, controls the FWER under any combination of true and false hypotheses. In practice, only strong control methods are used (Goeman & Solari, 2014).

Additionally, we separate between raw and adjusted p-values. A raw p-value is a p-value, as stated in Section 2.1, which can be described as the smallest choice α that will reject the hypothesis. Similarly, the adjusted p-value represents the smallest α level from which the multiple testing procedure rejects a hypothesis. The adjusted p-value is thus dependent on the multiple testing procedure used. The reader should be aware that the interpretation of the adjusted p-value can differ between FWER controlling methods, and other methods for controlling the rate of false positives, such as FDR (Goeman & Solari, 2014). Unless stated otherwise, the term "p-value" refers to the raw p-value for further mentions.

2.3 p-values

Firstly, we note that the p-value itself is a random variable as it vary depending on the sample data. Typically, assumptions made regarding p-values only consider p-values of the true hypotheses, where we by Definition 1 have that a valid p-value from a true hypothesis is either uniformly distributed between 0 and 1 or stochastically greater than uniform. Casella & Berger (2002) introduces in Theorem 1, the most common way of defining a valid p-value.

Theorem 1. Let $W(\boldsymbol{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\boldsymbol{x}$, define

$$p(\boldsymbol{x}) = \sup_{\theta \in \omega} P_\theta(W(\boldsymbol{X}) \geq W(\boldsymbol{x})). \tag{2}$$

Then, $p(\boldsymbol{X})$ is a valid p-value.

Proof. To test $H_0 \colon \theta \in \omega$, we let $\boldsymbol{X}$ be the outcome and $W(\boldsymbol{X})$ be the test statistic such that large values of W provide evidence for rejecting H_0. Fix $\theta \in \omega$, and let $\boldsymbol{x}$ be a realization. Using (2),

$$p_\theta(\boldsymbol{x}) = P_\theta(W(\boldsymbol{X}) \geq W(\boldsymbol{x})).$$

For a second realization $\boldsymbol{x}'$ it follows that for $p_\theta(\boldsymbol{x}') \leq p_\theta(\boldsymbol{x}) \iff W(\boldsymbol{x}') \geq W(\boldsymbol{x})$, which for a random variable X implies that $p_\theta(X) \leq p_\theta(\boldsymbol{x}) \iff W(X) \geq W(\boldsymbol{x})$. Consequently,

$$P_\theta(p_\theta(X) \leq p_\theta(\boldsymbol{x})) = P_\theta(W(X) \geq W(\boldsymbol{x})) = p_\theta(\boldsymbol{x}).$$

Let $0 \leq \alpha \leq 1$, and $\alpha' = \sup\{p_\theta(\boldsymbol{x})\colon p_\theta(\boldsymbol{x}) \leq \alpha\}$ (typically, $\alpha' < \alpha$ for discrete test statistics W). Then,

$$P_\theta(p_\theta(X) \leq \alpha) = P_\theta(p_\theta(X) \leq \alpha') = \alpha' \leq \alpha.$$

As $p(\boldsymbol{x}) = \sup_{\theta'} p_{\theta'}(\boldsymbol{x}) \geq p_\theta(\boldsymbol{x})$ for all sample points $\boldsymbol{x}$,

$$P_\theta(p(X) \leq \alpha) \leq P_\theta(p_\theta(X) \leq \alpha) \leq \alpha.$$

This is true for all $\theta \in \omega$, and every $0 \leq \alpha \leq 1$. Hence, $p(X)$ is a valid p-value. ■

For instances in which (1) is an equality, we call $p(\boldsymbol{X})$ an exact p-value. Note that exact p-values based on a continuous test statistic W, have a uniform distribution between 0 and 1. For strictly valid p-values gathered from true hypotheses, the distribution is stochastically greater than uniform, i.e., skewed towards 1. For p-values gathered from the false hypotheses, we expect a distribution skewed towards 0. These remarks are further emphasized in Section 5.1.

2.4 Dependency structures & probability inequalities

The dependency structure among p-values can significantly impact the performance of multiple testing procedures. Fully utilizing the underlying structure of the tests can lead to significant advantages, and result in less conservative methods. Therefore, FWER controlling procedures usually rely on probability inequalities that make certain assumptions about the distribution of p-values. Generally, stronger assumptions lead to more powerful procedures. Relevant for this thesis are the probability inequalities of Bonferroni and Simes (1986). Let $q_1, \ldots, q_{m_0}$ denote the $m_0 \leq m$ p-values of the true hypotheses. The Bonferroni inequality states that

$$P\left(\bigcup_{i=1}^{m_0} q_i \leq \frac{\alpha}{m_0}\right) \leq \alpha \tag{3}$$

(Goeman & Solari, 2014). Except for (1), the Bonferroni inequality makes no assumptions regarding the p-values. This makes the inequality applicable in most instances. However, this also means that the inequality includes unrealistic "worst case" scenarios, which occur when the inequality is a equality. Often resulting in the inequality being strict. We can note the conservatism of the Bonferroni inequality in its derivation, which is based on Boole's inequality. Boole's inequality states that for events $A_1, \ldots, A_k$,

$$P\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P(A_i). \tag{4}$$

A proof can be found in Appendix A. We then derive the Bonferroni inequality,

$$P\left(\bigcup_{i=1}^{m_0} q_i \leq \frac{\alpha}{m_0}\right) \leq \sum_{i=1}^{m_0} P\left(q_i \leq \frac{\alpha}{m_0}\right) \leq m_0 \frac{\alpha}{m_0} = \alpha.$$

The "worst case" scenario occurs when (4) is an equality, which is when all events A_i are pairwise disjoint. For the Bonferroni inequality, this translates to rejection of the null hypotheses being pairwise disjoint, i.e, at most one null hypothesis can be rejected. This is an example of negative dependency, which is unrealistic for real-life applications. Thus, the inequality is strict in most cases, and methods based on the Bonferroni inequality can therefore be quite conservative.

To avoid considering "worst case", we can make assumptions regarding the dependence structure of the p-values. One such is positive dependence through stochastic ordering (PDS) introduced by Block et al. (1985). In short, we assume positively dependent statistics, which allows us to exclude the unrealistic "worst case" scenario. By assuming PDS and (1) we obtain the Simes inequality, which for ordered p-values associated to true hypotheses, states that

$$P\left(\bigcup_{i=1}^{m_0} q_{(i)} \leq \frac{i\alpha}{m_0}\right) \leq \alpha \tag{5}$$

(Goeman & Solari, 2014). The Simes inequality is also strict for many p-value distributions, but improves on (3). The "worst case" is for exact independent p-values, which a proof can be found in Appendix C. FWER controlling methods exploiting the Simes inequality need thus that the PDS condition hold to validly control FWER at the assigned level α. However, Rødland (2006) shows that the Simes inequality is "valid on average", and instances where it fails are "somewhat bizarre". An example in which the inequality is reversed is presented in Section 5.4.

2.5 The global tests of Bonferroni & Simes

While multiple testing procedures aim to make inferences about the individual hypotheses in $\mathcal{H}$, global tests assess all the hypotheses in $\mathcal{H}$ simultaneously. Using the inequalities of Bonferroni and Simes, we can construct such global hypothesis tests. For the set of hypotheses $\mathcal{H}$ we want to test with FWER control level α, if all individual hypotheses are simultaneously true. That means to test the null hypothesis $\mathcal{H}_\cap = H_1 \cap \ldots \cap H_m$ against the alternative hypothesis stating that at least one of $H_1, \ldots, H_m$ is false. Further, $\mathcal{H}_\cap$ is referred to as a intersection hypothesis.

Using (3) we construct the Bonferroni global test. For the set $\mathcal{H}$, with associated p-values $p_1, p_2, \ldots, p_m$,

$$\text{reject } \mathcal{H}_\cap \text{ if } p_i \leq \frac{\alpha}{m} \text{ for any } i = 1, 2, \ldots, m,$$

where we have that by assuming $\mathcal{H}_\cap$ is true, the probability of rejecting $H_\cap$ is by (3) less than or equal to α. Thus, controlling the FWER at level α. Similarly, we construct the Simes global test using (5). For ordered p-values,

$$\text{reject } \mathcal{H}_\cap \text{ if } p_{(i)} \leq \frac{i\alpha}{m} \text{ for any } i = 1, 2, \ldots, m.$$

Using the same argument as for the Bonferroni global test, Simes control FWER at level α. In Section 4.1, we demonstrate that the Bonferroni inequality allows for strong FWER control when making inferences about individual hypotheses. However, this is not the case for Simes, who only provides weak control of the FWER, i.e., when $m_0 < m$, the probability of a false positive can exceed α (Example 2). Indicating that Simes is only appropriate as a global test.

Example 2. Consider a set of three independent hypotheses H_1, H_2, H_3, of which H_2 and H_3 are true with corresponding exact p-values $p_2 \leq p_3$, based on continuous test statistics. Assume H_1 to be false with corresponding p-value $p_1 = 0$. Set $\alpha = 0.05$. Then,

$$\text{FWER} = P(V > 0) = 1 - P\left(p_2 > \frac{2\alpha}{3} \cap p_3 > \alpha\right) = 1 - P\left(p_2 > \frac{2\alpha}{3}\right) P(p_3 > \alpha).$$

By using the order statistics of a uniform distribution between 0 and 1 we find that $P(p_2 > \frac{2\alpha}{3}) = \frac{841}{900}$ and $P(p_3 > \alpha) = \frac{399}{400}$. Thus $\text{FWER} = 1 - \left(\frac{841}{900}\right)\left(\frac{399}{400}\right) = 0.0679$, which is greater than α. Showing that when using Simes inequality to make inferences on individual hypotheses for m_0 less than m, we are not guaranteed to control the FWER at the assigned level α.

2.6 The closed testing procedure

The closed testing procedure introduced by Marcus et al. (1976) is a method for multiple hypothesis testing that strongly controls FWER. Consider the set of m hypotheses, $\mathcal{H}$. The closed testing procedure states that any $H_i \in \mathcal{H}$ is rejected under valid FWER level α control if all possible intersection hypotheses involving H_i are rejected by a valid level α test, i.e.,

$$\text{reject } H_i \text{ if } \begin{cases} H_i \text{ rejected by valid } \alpha \text{ test,} \\ H_i \cap H_j \text{ rejected for all } j \in \{1, 2, \ldots, m\} \backslash \{i\} \text{ by valid } \alpha \text{ test,} \\ \quad \vdots \\ H_i \cap H_1 \cap \cdots \cap H_{i-1} \cap H_{i+1} \cap \cdots \cap H_m \text{ rejected by valid } \alpha \text{ test.} \end{cases}$$

To illustrate how the closed testing procedure strongly controls FWER, we consider the set $\mathcal{H}^T \subseteq \mathcal{H}$ of true hypotheses. To reject any of the true hypotheses, we must reject the intersection hypotheses $\mathcal{H}^T_\cap$, which depends on a valid α test. Hence,

$$\text{FWER} = P(V > 0) \leq P(\text{reject } \mathcal{H}^T_\cap) \leq \alpha.$$

A valid level α test can for instance be the global tests of Bonferroni or Simes, and we will in Section 4 show how the multiple testing procedures of Holm, Hochberg and Hommel are special cases of the closed testing procedure. In the first place, using the closed testing procedure will be a computationally heavy process. For m hypotheses the closed testing procedure would require 2^{m-1} individual tests to possibly reject one hypotheses. We will however see that the number of tests can be easily reduced in the procedures for the mentioned special cases.

3 Procedures for FWER Control

We present the multiple testing procedures of Bonferroni, Holm (1979), Hochberg (1988) and Hommel (1988). Further derivations, analysis and comparisons will be discussed in Section 4. For all procedures below we consider a set of m hypotheses $\mathcal{H} = \{H_1, \ldots, H_m\}$ with associated p-values $p_1, \ldots, p_m$, where $p_{(i)}$ represent the i'th ordered p-value and $H_{(i)}$ its associated hypothesis.

3.1 The Bonferroni procedure

Reject all individual hypotheses for which

$$p_i \leq \frac{\alpha}{m}.$$

If no such p_i exists for $i = 1, 2, \ldots, m$, no hypotheses are rejected.

The Bonferroni procedure has strong FWER control and is valid for all dependency structures, with (1) as the only assumption. It is however conservative for most cases, especially for a large proportion of false hypotheses, or when the p-values have positive associations.

3.2 The Holm procedure

For $j = 1, 2, \ldots, m$, if

$$p_{(j)} \leq \frac{\alpha}{m - (j-1)},$$

reject $H_{(j)}$ and continue with $j + 1$, otherwise stop. If $p_{(1)}$ is greater than $\frac{\alpha}{m}$, no hypotheses are rejected.

The Holm procedure has strong FWER control, and is valid for all dependency structures, with (1) as the only assumption. The procedure rejects at least as many hypotheses as the Bonferroni procedure and one expect the largest gain in power for a large proportion of false hypotheses.

3.3 The Hochberg procedure

Find the largest integer $j \in \{1, 2, \ldots, m\}$ such that

$$p_{(j)} \leq \frac{\alpha}{m - (j-1)}.$$

Reject $H_{(1)}, \ldots, H_{(j)}$. If no such j exists, no hypotheses are rejected.

The Hochberg procedure has strong FWER control, and is valid under the PDS assumption and (1). It rejects at least as many hypotheses as Holm, possibly more.

3.4 The Hommel procedure

Find the smallest integer $j \in \{1, 2, \ldots, m\}$ such that

$$p_{(k)} > \frac{(k - (j-1))\alpha}{m - (j-1)} \text{ for all } k = j, \ldots, m.$$

Reject all hypotheses with corresponding p-value less than or equal to $\frac{\alpha}{m-(j-1)}$. If no such j exists all hypotheses are rejected, if $j = 1$ no hypotheses are rejected.

The Hommel procedure has strong FWER control, and is valid under the PDS assumption and (1). The procedure is more powerful than Hochberg, but requires more computations.

4 Derivation and Comparison of the Procedures

The aim of the procedures in Section 3 is to control the FWER at level α, while simultaneously rejecting as many false hypotheses as possible. Combining the closed testing procedure with the global tests of Bonferroni and Simes presented in Section 2, we derive the procedures of Holm, Hochberg and Hommel. The Bonferroni procedure can be derived directly from the Bonferroni inequality. From its derivations, we can further compare and examine the procedures in terms of their power and practical use.

Firstly, we introduce a "loop-hole" in the closed testing procedure when combined with the global tests of Bonferroni or Simes. Let $\mathcal{H} = \{H_1, H_2, \ldots, H_m\}$ be a set of m hypotheses, where $p_{(i)}$ is the i'th ordered p-value and $H_{(i)}$ its associated hypothesis. By the closed testing procedure we reject a hypothesis H_i if all possible intersection hypotheses involving H_i are rejected. Using the Bonferroni or Simes global tests it is however enough to reject intersection combinations involving only the largest p-values. Hence,

$$\text{reject } H_i \text{ if } \begin{cases} H_i \text{ rejected by valid } \alpha \text{ test,} \\ H_i \cap H_{(m)} \text{ rejected by valid } \alpha \text{ test,} \\ H_i \cap H_{(m-1)} \cap H_{(m)} \text{ rejected by valid } \alpha \text{ test,} \\ \quad \vdots \\ H_{(1)} \cap \cdots \cap H_i \cap \cdots \cap H_{(m)} \text{ rejected by valid } \alpha \text{ test.} \end{cases}$$

This is due to the structure of the global tests of Bonferroni and Simes. Having rejected the intersection combination with the largest p-values, rejection of other subsets of equally many hypotheses automatically follow.

4.1 Bonferroni: The Bonferroni inequality

The Bonferroni procedure reject an individual hypothesis if the associated p-value is less than or equal to $\frac{\alpha}{m}$. To prove strong FWER control we use the same arguments as for when deriving the Bonferroni inequality in Section 2.4. Let $m_0 \leq m$ be the number of true hypotheses, and $q_1, q_2, \ldots, q_{m_0}$ their associated p-values. Then,

$$\text{FWER} = P(V > 0) = P\left(\bigcup_{i=1}^{m_0} q_i \leq \frac{\alpha}{m}\right) \leq \sum_{i=1}^{m_0} P\left(q_i \leq \frac{\alpha}{m}\right) \leq m_0 \frac{\alpha}{m} \leq \alpha.$$

The above proof highlights the Bonferroni procedures conservatism. The first inequality (from the left) is strict for all cases where the events $q_i \leq \frac{\alpha}{m}$ are not pairwise disjoint. The second inequality, due to assuming (1), shows conservatism for strictly valid p-values, and the third inequality shows conservatism when the proportion of false hypotheses is large, i.e., $m_0 \ll m$. Hence, the (unrealistic) scenario for which the procedure is not conservative is for pairwise disjoint events, exact p-values, and all hypotheses in $\mathcal{H}$ are true. An example of this is shown in Section 5.4. One can easily find the adjusted p-values by $\min(mp_i, 1)$, where p_i represent the raw p-value.

4.2 Holm: closed testing & the Bonferroni global test

By combining the closed testing procedure with the Bonferroni global test we can improve upon the Bonferroni procedure. Considering $H_{(1)} \in \mathcal{H}$, we reject $H_{(1)}$ if we can reject all intersection hypotheses involving $H_{(1)}$ using a valid level α test, i.e.,

$$\text{reject } H_{(1)} \text{ if } \begin{cases} H_{(1)} \text{ rejected by valid } \alpha \text{ test,} \\ H_{(1)} \cap H_{(m)} \text{ rejected by valid } \alpha \text{ test,} \\ \quad \vdots \\ H_{(1)} \cap \cdots \cap H_{(m)} \text{ rejected by valid } \alpha \text{ test.} \end{cases}$$

By using the Bonferroni global test as our valid α test, this translates to

$$\text{reject } H_{(1)} \text{ if } \begin{cases} p_{(1)} \leq \alpha, \\ p_{(1)} \leq \frac{\alpha}{2} \text{ or } p_{(m)} \leq \frac{\alpha}{2}, \\ \quad \vdots \\ p_{(1)} \leq \frac{\alpha}{m} \text{ or } p_{(2)} \leq \frac{\alpha}{m} \text{ or } \ldots \text{or } p_{(m)} \leq \frac{\alpha}{m}. \end{cases}$$

Thus, for $p_{(1)}$ less than or equal $\frac{\alpha}{m}$, $H_{(1)}$ is surely rejected. If $p_{(1)}$ is greater than $\frac{\alpha}{m}$ no hypotheses are rejected as the p-values are ordered. Assume we reject $H_{(1)}$, evaluating $H_{(2)}$, we use the same arguments as for $H_{(1)}$. However, we do not need to assess intersection hypotheses involving $H_{(1)}$, as these intersections are already rejected. Hence,

$$\text{reject } H_{(2)} \text{ if } \begin{cases} p_{(2)} \leq \alpha, \\ p_{(2)} \leq \frac{\alpha}{2} \text{ or } p_{(m)} \leq \frac{\alpha}{2}, \\ \quad \vdots \\ p_{(2)} \leq \frac{\alpha}{m-1} \text{ or } p_{(3)} \leq \frac{\alpha}{m-1} \text{ or } \ldots \text{or } p_{(m)} \leq \frac{\alpha}{m-1}. \end{cases}$$

Given that $H_{(1)}$ is rejected, having that $p_{(2)}$ is less than or equal to $\frac{\alpha}{m-1}$ ensures rejection of $H_{(2)}$. Continuing to $H_{(3)}$, and so on until the first instance where $p_{(j)}$ is greater than $\frac{\alpha}{m-(j-1)}$, for $j = 1, 2, \ldots, m$.

This results in the Holm procedure as stated in Section 3.2. The gain in power compared to the Bonferroni procedure is clear. Bonferroni compare every p-value to $\frac{\alpha}{m}$, while Holm compare the j'th ordered p-value to $\frac{\alpha}{m-(j-1)}$, ensuring the Holm procedure to reject at least as many hypotheses as Bonferroni. We expect the greatest increase in power when the proportion of false hypotheses is large. The procedure is based on the closed testing procedure using the Bonferroni global test, and has thus strong FWER control for all p-value dependency structures, only assuming (1). A alternative proof of the Holm procedures' strong control of FWER can be found in Appendix B.

4.3 Hochberg: closed testing & the Simes global test

To obtain the Hochberg procedure we use the same idea as for Holm, but instead of using the Bonferroni global test, we use Simes. We consider a hypothesis $H_{(i)} \in \mathcal{H}$,

$$\text{reject } H_{(i)} \text{ if } \begin{cases} p_{(i)} \leq \alpha, \\ p_{(i)} \leq \frac{\alpha}{2} \text{ or } p_{(m)} \leq \alpha, \\ \quad \vdots \\ p_{(1)} \leq \frac{\alpha}{m} \text{ or } p_{(2)} \leq \frac{2\alpha}{m} \text{ or } \ldots \text{or } p_{(i)} \leq \frac{i\alpha}{m} \text{ or } \ldots \text{or } p_{(m)} \leq \alpha. \end{cases}$$

We observe directly that if $p_{(i)}$ is less than or equal to $\frac{\alpha}{m-(i-1)}$ intersection hypotheses of up to $m-(i-1)$ hypotheses are rejected, since

$$p_{(i)} \leq \frac{\alpha}{m-(i-1)} < \cdots < \frac{\alpha}{3} < \frac{\alpha}{2} < \alpha.$$

For intersection hypotheses of more than $m-(i-1)$ hypotheses, $p_{(i)}$ will be compared to $\frac{k\alpha}{m-(i-k)}$, for $k = 1, \ldots, i$, where we have that

$$p_{(i)} \leq \frac{\alpha}{m-(i-1)} = \frac{2\alpha}{2(m-(i-1))} \leq \frac{2\alpha}{m-(i-2)} \leq \frac{k\alpha}{m-(i-k)}.$$

Hence, if $p_{(i)}$ is less than or equal to $\frac{\alpha}{m-(i-1)}$ we reject $H_{(i)}$. It follows that hypotheses $H_{(1)}, \ldots, H_{(i-1)}$ are rejected. This is because $p_{(1)} \leq \cdots \leq p_{(i)} \leq \frac{\alpha}{m-(i-1)}$, rejecting intersection hypotheses of up to $m-(i-1)$ hypotheses. And intersection hypotheses of more than $m-(i-1)$ hypotheses are rejected as $p_{(i)}$ is less than or equal to $\frac{\alpha}{m-(i-1)}$. Hence, by finding the largest integer $j \in \{1, 2, \ldots, m\}$ such that

$$p_{(j)} \leq \frac{\alpha}{m-(j-1)},$$

we reject hypotheses $H_{(1)}, \ldots, H_{(j)}$, which is the Hochberg procedure as stated in Section 3.3.

Comparing Hochberg to Holm it is clear that Hochberg has greater power. Holm finds the first instance where $p_{(j)}$ is greater than $\frac{\alpha}{m-(j-1)}$ and rejects hypotheses $H_{(1)}, \dots, H_{(j-1)}$, while Hochberg find the largest j' such that $p_{(j')}$ is less than or equal to $\frac{\alpha}{m-(j'-1)}$, rejecting $H_{(1)}, \dots, H_{(j')}$, where $j' \geq j-1$. The Hochberg procedure is valid by closed testing and Simes global test, thus ensuring strong FWER control assuming PDS and (1).

Example 3. Let $\mathcal{H} = \{H_1, H_2, H_3, H_4, H_5\}$ with associated ordered p-values. Using Hochberg's definition of j, assume $j = 3$. This gives that $p_3 \leq \frac{\alpha}{3}$, while $p_4 > \frac{\alpha}{2}$ and $p_5 > \alpha$. For the closed testing procedure combined with Simes' global test,

$$\text{reject } H_3 \text{ if } \begin{cases} p_3 \leq \alpha, \\ p_3 \leq \frac{\alpha}{2} \text{ or } p_5 \leq \alpha, \\ p_3 \leq \frac{\alpha}{3} \text{ or } p_4 \leq \frac{2\alpha}{3} \text{ or } p_5 \leq \alpha, \\ p_2 \leq \frac{\alpha}{4} \text{ or } p_3 \leq \frac{2\alpha}{4} \text{ or } p_4 \leq \frac{3\alpha}{4} \text{ or } p_5 \leq \alpha, \\ p_1 \leq \frac{\alpha}{5} \text{ or } p_2 \leq \frac{2\alpha}{5} \text{ or } p_3 \leq \frac{3\alpha}{5} \text{ or } p_4 \leq \frac{4\alpha}{5} \text{ or } p_5 \leq \alpha. \end{cases}$$

Observe directly that H_3 will be rejected, as

$$p_3 \leq \frac{\alpha}{3} \leq \frac{\alpha}{2} = \frac{2\alpha}{4} \leq \frac{3\alpha}{5} \leq \alpha.$$

Moreover, H_1 and H_2 will be rejected since $p_1 \leq p_2 \leq p_3$. Intersection hypotheses of up to three hypotheses will obviously be rejected, while intersections of four and five hypotheses will compare p_3 to the same values as for when rejecting H_3. This shows that by finding $j = 3$, we can with FWER control level α reject H_1, H_2 and H_3 by the Hochberg procedure. To reject the same number of hypotheses using the Bonferroni or Holm procedures, we need additional requirements for the p-values, namely that $p_1 \leq p_2 \leq p_3 \leq \frac{\alpha}{5}$ for Bonferroni, and $p_1 \leq \frac{\alpha}{5}$ and $p_2 \leq \frac{\alpha}{4}$ for Holm.

4.4 Hommel: closed testing & the Simes global test

As for Hochberg, we use closed testing and Simes global test as our framework. For a hypothesis $H_{(i)} \in \mathcal{H}$,

$$\text{reject } H_{(i)} \text{ if } \begin{cases} p_{(i)} \leq \alpha, \\ p_{(i)} \leq \frac{\alpha}{2} \text{ or } p_{(m)} \leq \alpha, \\ \quad \vdots \\ p_{(1)} \leq \frac{\alpha}{m} \text{ or } p_{(2)} \leq \frac{2\alpha}{m} \text{ or } \dots \text{or } p_{(i)} \leq \frac{i\alpha}{m} \text{ or } \dots \text{or } p_{(m)} \leq \alpha. \end{cases}$$

For obtaining the Hochberg procedure in Section 4.3 we focused on how we directly can reject a hypothesis. For Hommel, our focus shifts to identifying the hypotheses that cannot be rejected. We do this be finding the smallest integer $j \in \{1, 2, \dots, m\}$ such that

$$p_{(k)} > \frac{(k-(j-1))\alpha}{m-(j-1)} \text{ for all } k = j, \dots, m.$$

By finding this value j, we will not be able to reject the intersection hypothesis $H_{(j)} \cap H_{(j+1)} \cap \dots \cap H_{(m)}$. Consequently, by the closed testing procedure, we are not able to reject any of $H_{(j)}, H_{(j+1)}, \dots, H_{(m)}$. Moreover, we know from the definition of j that the largest indexed intersection hypotheses of more than $m-(j-1)$ hypotheses will be rejected, since

$$p_{(k)} < \frac{(k-(j'-1))\alpha}{m-(j'-1)} \text{ for at least one } k = j', \dots, m$$

must be true for $j' = 1, 2, \dots, j-1$. Therefore, the remaining hypotheses for possible rejection are $H_{(1)}, \dots, H_{(j-1)}$, where we only need to evaluate intersection hypotheses of up to $m-(j-1)$ hypotheses. Thus, for $i \in \{1, 2, \dots, j-1\}$,

$$\text{reject } H_{(i)} \text{ if } \begin{cases} p_{(i)} \leq \alpha, \\ p_{(i)} \leq \frac{\alpha}{2} \text{ or } p_{(m)} \leq \alpha, \\ \quad \vdots \\ p_{(i)} \leq \frac{\alpha}{m-(j-1)} \text{ or } p_{(j+1)} \leq \frac{2\alpha}{m-(j-1)} \text{ or } \dots \text{or } p_{(m)} \leq \alpha. \end{cases}$$

We then reject all hypotheses with corresponding p-value less than or equal to $\frac{\alpha}{m-(j-1)}$, which results in the Hommel procedure as stated in Section 3.4; find the smallest integer $j \in \{1, 2, \ldots, m\}$ such that

$$p_{(k)} > \frac{(k-(j-1))\alpha}{m-(j-1)} \text{ for all } k = j, \ldots, m.$$

Then, all hypotheses with a corresponding p-value less than or equal to $\frac{\alpha}{m-(j-1)}$ are rejected.

The Hommel procedure automatically rejects all hypotheses that would have been rejected using closed testing combined with Simes, which is not the case for Hochberg. As a result, the Hommel procedure rejects at least as many hypotheses as Hochberg and should be the preferred option. The largest gain in power is expected for a large proportion of false hypotheses or for positive associations between the p-values. In Section 4.5, the gain in power compared with Hochberg become more apparent. The procedure is based on closed testing and Simes, which provide strong FWER control, assuming PDS and (1).

Example 4. Let $\mathcal{H} = \{H_1, H_2, H_3, H_4, H_5\}$ with associated ordered p-values. Using Hommel's definition of j, assume $j = 3$. This gives us that $p_3 > \frac{\alpha}{3}$, $p_4 > \frac{2\alpha}{3}$ and $p_5 > \alpha$, which is equivalent to us not being able to reject $H_3 \cap H_4 \cap H_5$ by Simes. We are thus not able to reject any of H_3, H_4 or H_5 by closed testing. Moreover, from $j = 3$ we know that $H_2 \cap H_3 \cap H_4 \cap H_5$ and $H_1 \cap H_2 \cap H_3 \cap H_4 \cap H_5$ will be rejected by Simes. Hence, when evaluating H_1 and H_2 with closed testing, we only need to consider intersection hypotheses of up to three hypotheses. In fact, we only need to reject the largest indexed intersection hypothesis that involves three hypotheses, i.e.,

$$H_i \cap H_4 \cap H_5, \text{ where } i = 1, 2.$$

As we know that $p_4 > \frac{2\alpha}{3}$ and $p_5 > \alpha$, the intersection hypotheses is rejected if and only if p_i is less than or equal to $\frac{\alpha}{3}$, which consequently ensures rejection of intersection hypotheses of one and two hypotheses. Using Hommel it is enough for p_1 and p_2 to be less than $\frac{\alpha}{3}$ to reject H_1 and H_2. For Hochberg however, we additionally demand that p_2 is less than or equal to $\frac{\alpha}{4}$.

4.5 Comparing Hochberg & Hommel

The improvements of Holm over Bonferroni, and Hochberg over Holm, are relatively straightforward, but comparing Hochberg and Hommel requires a more in-depth analysis. We write the Hommel procedure as finding the smallest $j \in \{1, 2, \ldots, m\}$ such that

$$p_{(k)} > \frac{(k-(j-1))\alpha}{m-(j-1)} \text{ for all } k = j, \ldots, m,$$

where we reject all hypotheses with a corresponding p-value less than or equal to $\frac{\alpha}{m-(j-1)}$. Hochberg finds the largest $j' \in \{1, 2, \ldots, m\}$ such that

$$p_{(j')} \leq \frac{\alpha}{m-(j'-1)},$$

rejecting all hypotheses with corresponding p-value less than or equal to $\frac{\alpha}{m-(j'-1)}$. The more powerful procedure will thus have the largest value of j (or j'). Assuming both j and j' exists, it is clear that j' can be less than j, and that they can not be equal. For which Hommel will be the more powerful procedure. Thus, what is left to check for is j' to be larger than j. Equivalent to Hochberg being the more powerful procedure. Let $j' = j + r$ for any $r = 1, 2, \ldots, m - j$. From the definition of j we know that

$$p_{(j)} > \frac{\alpha}{m-(j-1)} \cap p_{(j+1)} > \frac{2\alpha}{m-(j-1)} \cap \cdots \cap p_{(j+r)} > \frac{(r+1)\alpha}{m-(j-1)} \cap \cdots \cap p_{(m)} > \alpha.$$

And from the definition of j' we have that

$$p_{(j')} = p_{(j+r)} \leq \frac{\alpha}{m-((j+r)-1)} = \frac{(r+1)\alpha}{(r+1)(m-((j+r)-1)} \leq \frac{(r+1)\alpha}{m-(j-1)} < p_{(j+r)},$$

which is a contradiction. Hence, instances where $j' \geq j$ do not exist, making the Hommel procedure the more powerful procedure. Special cases for instances where $j' = m$, then j do not exist and both

procedures reject all hypotheses in $\mathcal{H}$. If j' do not exist, Hochberg will not reject any hypotheses, while Hommel will have j less than or equal to $m-1$, i.e., still capable of rejecting hypotheses.

Since Hochberg is more powerful than both the Bonferroni and Holm procedures, the above result shows that Hommel is also more powerful than Bonferroni and Holm. Note however that even though the procedures of Hochberg and Hommel have successfully extended Simes global test to make inferences on individual hypotheses, the procedures are conservative. That means there are instances where the Simes global test is rejected but the procedures of Hochberg and Hommel do not reject any individual hypotheses.

Example 5. We illustrate how the procedures of Hochberg and Hommel can be conservative by assuming a set of three hypotheses H_1, H_2 and H_3 with associated ordered p-values. We assume that for a significance level α,

$$\frac{\alpha}{2} < p_1 \leq p_2 < \frac{2\alpha}{3} \text{ and } p_3 > \alpha.$$

The Simes global test is rejected as p_2 is less than $\frac{2\alpha}{3}$. For Hochberg, no hypotheses are rejected since $p_1 > \frac{\alpha}{3}$, $p_2 > \frac{\alpha}{2}$ and $p_3 > \alpha$. And for Hommel we get $j = 2$, thus no rejections since p_1 is greater than $\frac{\alpha}{2}$. Showing that the Simes global test is rejected, but Hochberg and Hommel are not able to reject any individual hypotheses.

Calculating the adjusted p-values for the Holm and Hochberg procedures can be performed using a simple algorithm, however, not shown here. Hommel encounter greater complexity when calculating this value. The reader is advised to refer to Goeman & Solari (2014). However, an interesting note regarding the adjusted p-values is their possibility of being equal, even though the raw p-values were not, often resulting in long lists of identical values. Most often from true hypotheses, being equal to one.

5 Simulations & Examples

To demonstrate the differences in behavior between the presented procedures, we apply them to data and compare the empirical results to the theoretical conclusions made earlier. Specifically, we test the procedures using high and low proportions of false hypotheses, as well as independent, positively dependent, and negatively dependent tests. All of the data are simulated using `R` and code can be found in Appendix D.

5.1 Distribution of exact & valid p-values

Continuing on Example 1, we have the null hypothesis $H_0 \colon \mu \leq 180$ and the alternative hypothesis $H_1 \colon \mu > 180$. We perform the hypothesis test 10 000 times for independent random samples, calculating the 10 000 associated p-values. We simulate for two instances. First, we simulate for true values of $\mu < 180$. Second, we simulate for $\mu = 180$. Both instances will thus produce p-values associated to true hypotheses. We let $\sigma = 3$ (known) and $n = 100$ be the number of observations in each sample.

For each sample we collect height observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_{100})$, where $\bar{x}$ denote the mean of the 100 observations. Using (2) to define a valid p-value we set the mean as our continuous test statistic W. Hence, $p(\boldsymbol{x}) = \sup_{\mu \in \omega} P_\mu(\bar{X} \geq \bar{x})$. The supremum argument for $p(\boldsymbol{x})$ will clearly be for $\mu = 180$ for all sample points $\boldsymbol{x}$, as this gives the highest expected value for $\bar{X}$. From the proof in Section 2.3 we know that $P_\mu(p_\mu(\boldsymbol{X}) \leq \alpha) = \alpha$ for all $\mu \in \omega$. Therefore $p(\boldsymbol{x}) = p_\mu(\boldsymbol{x})$ give exact p-values, which for our case happens when $\mu = 180$, i.e., $H_0 \colon \mu = 180$ is true. Consequently, for instances where $\mu < 180$, we have $p(\boldsymbol{x}) > p_\mu(\boldsymbol{x})$, giving strictly valid p-values, as $P_\mu(p(\boldsymbol{X}) \leq \alpha) < P_\mu(p_\mu(\boldsymbol{X}) \leq \alpha) = \alpha$.

For the first instance, we simulated the samples using true mean value $\mu = 179$. We obtain 10000 p-values from true hypotheses and expect a skewed distribution towards 1, as these correspond to strictly valid p-values. The resulting p-value distribution is illustrated in Figure 1. As expected,

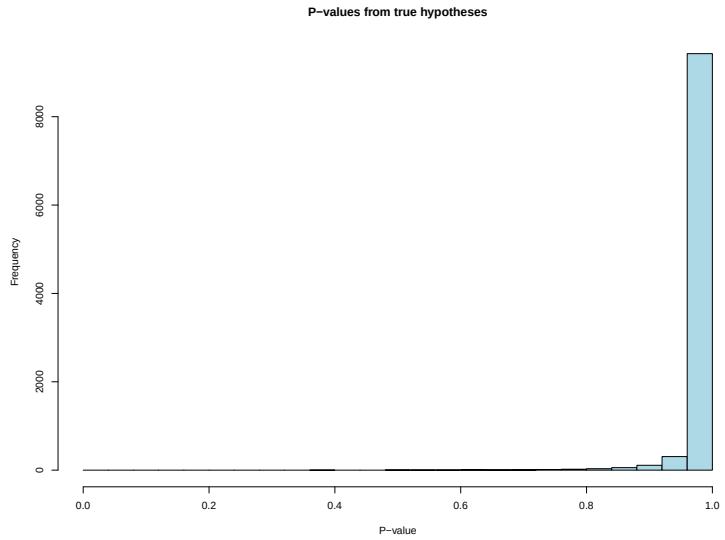


Figure 1: Histogram of p-values from testing $H_0 \colon \mu \leq 180$ for 10 000 independent samples. Each sample consist of 100 height values simulated using true $\mu = 179$ and constant (known) variance $\sigma^2 = 3^2$.

the p-values are skewed towards 1, empirically showing that $P(p(\boldsymbol{X}) \leq \alpha) < \alpha$.

For the second case we simulated observations using true mean $\mu = 180$. This gives 10 000 p-values from true null hypotheses, expected to be uniformly distributed between 0 and 1. The resulting p-value distribution is illustrated in Figure 2. From the histogram, we observe an approximately uniform distribution, which agrees with the theory of exact p-values presented previously.

Hence, in addition to theory we have empirical results showing that, if $H_0 \colon \mu \in \omega$ is true, where μ corresponds to the supremum argument in (2), we expect a uniform distribution of p-values

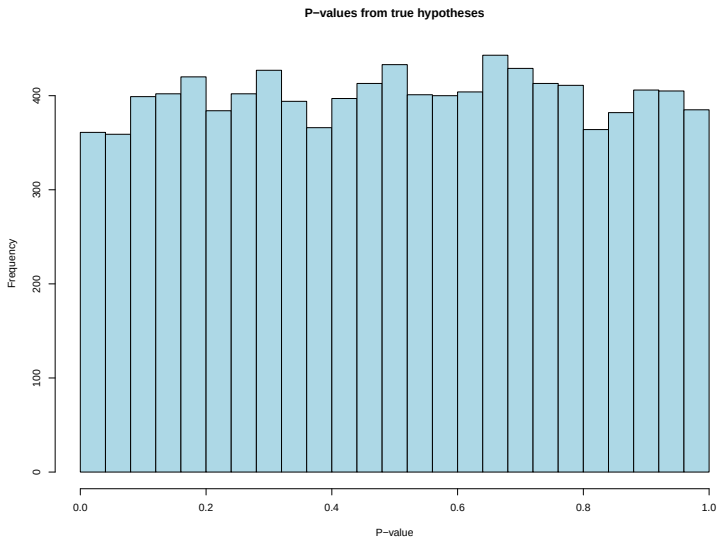


Figure 2: Histogram of p-values from testing $H_0 \colon \mu \leq 180$ for 10 000 independent samples. Each sample consist of 100 height values simulated using true $\mu = 180$ and constant (known) variance $\sigma^2 = 3^2$.

between 0 and 1. Otherwise, if $H_0 \colon \mu' \in \omega \setminus \{\mu\}$ is true, we expect a distribution skewed towards 1. Note that for a simple two-sided test ($|\omega| = 1$) with a continuous test statistic, we expect a uniform distribution of p-values for all the true hypotheses. Whether or not a p-value is exact or valid can influence the conservatism of a multiple testing method. E.g., the second inequality in the Bonferroni derivation in Section 4.1, will be an equality for exact p-values.

Additionally, we show the distribution of p-values corresponding to false hypotheses. We simulated samples for different true mean values $\mu \in (180, 182]$, giving 10 000 p-values from false hypotheses. The resulting p-value distribution is presented in Figure 3. We observe a p-value distribution skewed towards 0, which corresponds to $P(p(\boldsymbol{X}) \leq \alpha) > \alpha$. This is consistent with theory. Let $\boldsymbol{X}$ be a random variable corresponding to true $\mu = 180$, and $\boldsymbol{X}'$ is a random variable corresponding to true $\mu > 180$. Testing the null hypotheses $H_0 \colon \mu \leq 180$ for both variables we expect $W(\boldsymbol{x}) < W(\boldsymbol{x}')$ for observations $\boldsymbol{x}$ and $\boldsymbol{x}'$. This in turn give that $p(\boldsymbol{x}) > p(\boldsymbol{x}')$. Since $p(\boldsymbol{X})$ is exact, then $P_\mu(p(\boldsymbol{X}') \leq \alpha) > P_\mu(p(\boldsymbol{X}) \leq \alpha) = \alpha$.

5.2 FWER control of independent tests

The presented procedures are all valid for independent tests, hence, we construct 10 000 independent tests to compare their performances. We start by considering a scenario with a high proportion of true hypotheses, 7 000 true and 3 000 false. We assume the data is Gaussian with a known variance $\sigma^2 = 1$. Testing the null hypothesis

$$H_0 \colon \mu = 0 \text{ against the alternative hypothesis } H_1 \colon \mu \neq 0,$$

for 10 000 independent random samples. For the 7 000 true hypotheses, we simulated observations from $\mathcal{N}(0, 1)$. For the 3 000 false hypotheses, 1 500 observations are simulated from $\mathcal{N}(4, 1)$ and 1 500 from $\mathcal{N}(-4, 1)$. We find the p-value for each test using (2) and set $\alpha = 0.05$. We perform the Bonferroni, Holm, Hochberg, and Hommel procedures on the 10 000 p-values and report the number of rejected hypotheses, as well as the number of false positives. We also compare to not performing any multiple testing correction at all, i.e., simply rejecting a hypothesis if the corresponding p-value is less than or equal to α. We repeated this task 100 times, and Table 1 shows the mean number of rejected hypotheses for the procedures.

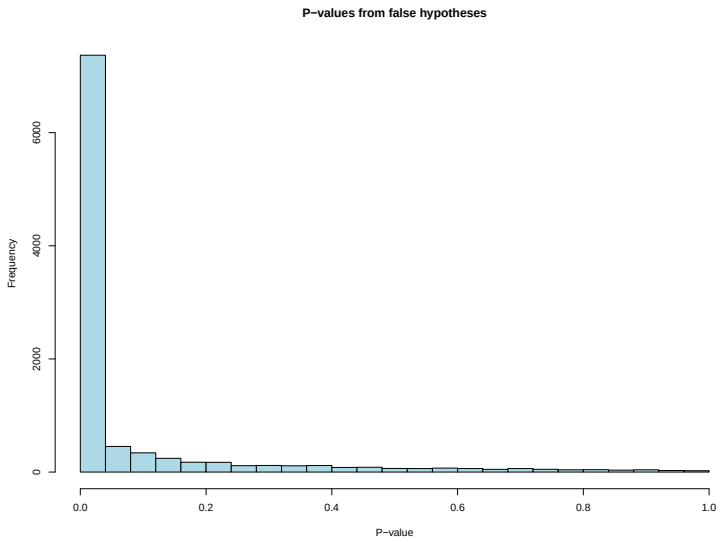


Figure 3: Histogram of p-values from testing $H_0 : \mu \leq 180$ for 10 000 independent samples. Each sample consist of 100 height values simulated using true $\mu \in (180, 182]$ and constant (known) variance $\sigma^2 = 3^2$.

	No correction	Bonferroni	Holm	Hochberg	Hommel
True hypotheses rejected	347.4	0.03	0.03	0.03	0.03
False hypotheses rejected	2937.56	858.09	878.21	878.21	911.42
Total	3284.96	858.12	878.24	878.24	911.45

Table 1: The mean number of rejected hypotheses among 10 000 independent tests, of which 7 000 true, and 3 000 false.

From Table 1 we read that the multiple testing procedures gave a false positive finding on three different occasions, which corresponds to FWER $= P(V > 0) = 0.03$, hence, controlling the FWER (strictly) at level α. With no multiple testing correction at all, there were false positives in each of the 100 repetitions, giving FWER $= P(V > 0) = 1$. The difference in power between the procedures is however minimal. The least conservative option (Hommel) rejected on average 53.33 more hypotheses than the most conservative option (Bonferroni). However, Hommel still failed to reject 2088.58 false hypotheses on average.

Repeating the simulations, only now for a larger proportion of false hypotheses, namely 7 000 false and 3 000 true. Simulated from the same Gaussian distributions. The results are shown in Table 2.

	No correction	Bonferroni	Holm	Hochberg	Hommel
True hypotheses rejected	150.24	0.01	0.02	0.02	0.05
False hypotheses rejected	6855.85	1997.68	2117.4	2117.42	2460.57
Total	7006.09	1997.69	2117.42	2117.44	2460.62

Table 2: The mean number of rejected hypotheses among 10 000 independent tests, of which 3 000 true, and 7 000 false.

The results in Table 2 illustrate that Bonferroni had one instance of a false positive, Holm and Hochberg had two instances, while Hommel had five. Each false positive in a different simulation. Hence, all procedures control FWER at level α, however Bonferroni conservatively compared to

Holm and Hochberg, and Holm and Hochberg conservatively compared to Hommel. This is coherent with the theoretical conclusions when having a large proportion of false hypothesis. The least conservative option (Hommel) rejected on average 462.93 more hypotheses than the most conservative option (Bonferroni), which is a great increase in power compared to when having a large proportion of true hypotheses. Hommel still failed to reject about 4 540 false hypotheses, but rejected a higher proportion; 0.35 compared to 0.3 in Table 1.

5.3 FWER control of positive dependent tests

For positive dependent tests we expect especially Bonferroni to be conservative compared to the other procedures. We construct 10 011 tests by comparing 142 groups from Gaussian distributions with mean values $\mu_1, \mu_2, \ldots, \mu_{142}$ and common (known) variance $\sigma^2 = 3^2$. Our null hypotheses are

$$H_0\colon \mu_i = \mu_j \text{ against the alternative hypotheses } H_1\colon \mu_i \neq \mu_j,$$

for $i, j = 1, 2, \ldots, 142$ and $i \neq j$. These tests are positively dependent as the p-values will exhibit a positive correlation between each other. We constructed the 142 groups by having 30 groups with true mean $\mu = 15$, 50 groups with true mean $\mu = 40$, 25 groups with true mean $\mu = 60$, 30 groups with true mean $\mu = -5$ and seven groups with true mean $\mu = 30$. This give in total 7 595 false hypotheses. As in Section 5.2, we simulated the process 100 times, looking at the mean number of rejected hypotheses for each procedure. The results are shown in Table 3.

	No correction	Bonferroni	Holm	Hochberg	Hommel
True hypotheses rejected	121.08	0.01	0.04	0.04	0.04
False hypotheses rejected	7449.67	5994.45	6227.37	6227.37	6282.34
Total	7570.75	5994.46	6227.41	6227.41	6282.38

Table 3: The mean number of rejected hypotheses among 10 011 positive dependent tests, of which 2 416 true, and 7 595 false.

Firstly, we observe from Table 3 that a greater proportion of false hypotheses are rejected by the procedures. This is however influenced by a number of factors, most notably the power of each individual test. In terms of power between the procedures, there are not much separating them. Holm, Hochberg and Hommel performed relatively similar (Hommel rejected more but not by much), while Bonferroni performed the worst. Bonferroni have allowed one false positive finding, while the other procedures have allowed a total of four. Empirically showing again how Bonferroni controls FWER at a stricter level than α. One should note that two of the four false positives found in Holm, Hochberg and Hommel happened simultaneously. This is however according to the definition of FWER, as we control $P(V > 0)$, not $P(V = 1)$. Consequently, giving an empirical estimate of FWER at 0.03.

5.4 The Simes global tests for negatively dependent tests

As mentioned in Section 2.3, instances where the Simes inequality fail are "somewhat bizarre". Nonetheless, they exist. Block et al. (2008) show a negative dependence concept where the Simes inequality is reversed. Let $W_1, W_2, \ldots, W_m$ have a multivariate normal distribution with means 0, variances 1, and negative correlations. Assume $W_1, W_2, \ldots, W_m$ to be the m test statistics for the one-sided hypotheses ($H_0\colon \mu \leq 0$ against the alternative hypothesis $H_1\colon \mu > 0$) $H_1, H_2, \ldots, H_m$, with associated p-values $p_1, \ldots, p_m$. Then,

$$P\left(\bigcup_{i=1}^{m} p_{(i)} \leq \frac{i\alpha}{m}\right) \geq \alpha.$$

Note that all the null hypotheses $H_1, H_2, \ldots, H_m$ are true. To simulate an example we set $m = 6$, with correlation $\rho = -0.16$ between all test statistics $W_1, W_2, \ldots, W_6$. We performed the Simes- and Bonferroni global tests on the resulting p-values. Repeating this procedure multiple times we

calculate an empirical estimate of the probability of a false positive for the two tests. We tested for different significance levels α. The results are shown in Table 4.

α	α_B	α_S
0.01	0.0098	0.0098
0.02	0.0203	0.0203
0.05	0.0497	0.0501
0.10	0.098	0.10
0.15	0.1476	0.1526
0.20	0.1936	0.204

Table 4: Estimated probabilities of a false positive for the Bonferroni global test (α_B), and the Simes global test (α_S). For negatively correlated test statistics.

Table 4 illustrate that the probability of a false positive is nearly the same for Bonferroni and Simes. The simulation of α_S indicates that the Simes inequality (5) is reversed, but the effect is minimal and hardly noticeable. As we have pairwise disjoint events, exact p-values and all null hypotheses are true, the Bonferroni inequality (3), as mentioned in Section 4.1, is an equality. This is supported by the estimated probabilities α_B which lies very close to α.

6 Discussion

In this thesis, we illustrated how the FWER controlling procedures of Bonferroni, Holm, Hochberg and Hommel are powerful tools for controlling the rate of false positives for multiple hypotheses testing. Based on the derivations of the procedures, we concluded that Hommel is the least conservative of the four and should be the preferred option, as instances where its assumptions fail are fairly unrealistic. Further, we have identified the potential of the general framework upon which the procedures are based, namely closed testing with local level α tests. The framework can derive less conservative procedures by having more powerful global tests (likely involving stronger assumptions) or equally powerful procedures, but with fewer assumptions. One would imagine that the former is of greater practical interest.

Goeman & Solari (2014) mention a few additional ways of enhancing the performance of FWER controlling procedures, namely by either selection or aggregation (or both). FWER controlling procedures are more powerful for a small number of hypotheses. Therefore, discarding hypotheses prior to testing can potentially lead to more rejections amongst the remaining hypotheses. Naturally, one should be considerate before discarding, as there is a risk of discarding hypotheses that may be significant. Therefore, discarded hypotheses should be selected based on either their lack of interest or their low power. An alternative method to lower the number of test is prior aggregation. Either on the basis of the data, or field knowledge. A researcher can cluster hypotheses prior to testing, the global hypothesis for a cluster is then rejected before testing the individual hypotheses using an FWER controlling procedure. Permutation-based FWER controlling procedures also exist, such as Westfall & Young (1993), which do not make any assumptions regarding the dependency structure of the p-values, thus removing the associated conservatism found in the probability inequalities of Bonferroni and Simes.

However, FWER controlling procedures are generally quite conservative for many data sets. The procedures focuses primarily on avoiding false positives and consequently allow for a large number of false negatives. All hypotheses rejected by an FWER controlling procedure is however individually reliable, i.e., a $100(1-\alpha)\%$ confidence that every rejection is correct, which is sufficient for many experiments. The rate of false negatives can however not be overlooked. Missing out on potentially significant discoveries due to a conservative procedure can result in a waste of resources and frustration when another researcher later receives credit for the findings.

One alternative for controlling the rate of false positives is FDR-based methods. These methods aim to find a trade-off between false positives and false negatives. Let R be the number of rejected hypotheses and V be the number of false positives from a multiple testing procedure. We define

$$Q = \begin{cases} V/R \text{ , if } R > 0, \\ 0 \quad\quad \text{, otherwise} \end{cases}$$

as the false discovery proportion (FDP), and $E(Q)$ as the false discovery rate (FDR). Therefore, while FWER focuses on the probability of any error among the rejections, FDR focuses on the expected proportion of error. Further, we have that $E(Q) \leq P(Q > 0) = P(V > 0) = \text{FWER}$, which implies that any FWER controlling procedure is naturally an FDR controlling procedure, but easier to control at a set level α. Consequently, one expects FDR-based procedures to be more powerful than FWER-based, particularly for a large proportion of false hypotheses (Goeman & Solari, 2014).

Bibliography

Block, H.W., T.H. Savits and M. Shaked (1985). 'A concept of negative dependence using stochastic ordering'. In: *Statistics & Probability Letters* 3, pp. 81–86.

Block, H.W., T.H. Savits and J. Wang (2008). 'Negative dependence and the Simes inequality'. In: *Journal of Statistics Planning and Inference* 138, pp. 4107–4110.

Casella, G. and R.L. Berger (2002). *Statistical Inference. 2 ed.* Duxbury Pacific Grove.

Goemann, J.J. and A. Solari (2014). 'Multiple hypothesis testing in genomics'. In: *Statistics in medicine* 33, pp. 1946–1978.

Hochberg, Y. (1988). 'A sharper Bonferroni procedure for multiple tests of significance'. In: *Biometrika* 75, pp. 800–802.

Hochberg, Y. and D. Rom (1995). 'Extensions of multiple testing procedures based on Simes' test'. In: *Journal of Statistics Planning and Inference* 48, pp. 141–152.

Holm, S. (1979). 'A simple sequentially rejective multiple test procedure'. In: *Scandinavian journal of statistics*, pp. 65–70.

Hommel, G. (1988). 'A stagewise rejective multiple test procedure based on a modified Bonferroni test'. In: *Biometrika* 75, pp. 383–386.

Marcus, R., P. Eric and K.R. Gabriel (1976). 'On closed testing procedures with special reference to ordered analysis of variance'. In: *Biometrika* 63, pp. 655–660.

Rødland, E. (2006). 'Simes' procedure is 'valid on average''. In: *Biometrika* 93, pp. 742–746.

Sarkar, S.K. (1998). 'Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture'. In: *The Annals of Statistics* 26, pp. 494–504.

Simes, R.J. (1986). 'An improved Bonferroni procedure for multiple tests of significance'. In: *Biometrika* 73, pp. 751–754.

Appendix

A Proof: Boole's inequality

For events $A_1, \ldots, A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

Proof. One can prove Boole's inequality for a finite collection of events $A_1, \ldots, A_n$ by induction. For the simplest case, when $n = 1$, it is trivially true that

$$P(A_1) \leq P(A_1).$$

Assuming the inequality hold for $n = k$,

$$P\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P(A_i),$$

we will show that it holds for $n = k+1$. Further, we know that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, where $P(A \cap B) \geq 0$. As the union operation is associative

$$P\left(\bigcup_{i=1}^{k} A_i \cup A_{k+1}\right) = P\left(\bigcup_{i=1}^{k} A_i\right) + P(A_{k+1}) - P\left(\bigcup_{i=1}^{k} A_i \cap A_{k+1}\right) \leq P\left(\bigcup_{i=1}^{k} A_i\right) + P(A_{k+1}),$$

which give that

$$P\left(\bigcup_{i=1}^{k+1} A_i\right) = P\left(\bigcup_{i=1}^{k} A_i \cup A_{k+1}\right) \leq \sum_{i=1}^{k} P(A_i) + P(A_{k+1}) = \sum_{i=1}^{k+1} P(A_i).$$

■

B Proof: The Holm procedure control FWER at level α

For a set of hypotheses $\mathcal{H} = \{H_1, H_2, \ldots, H_m\}$ with associated p-values $p_1, \ldots, p_m$, the Holm procedure has strong FWER control at level α.

Proof. By the Holm procedure as stated in Section 3.2, let k be minimal such that $H_{(k)}$ is a false positive, and let $m_0 \leq m$ be the number of true hypotheses in $\mathcal{H}$. This in turn gives that $H_{(1)}, H_{(2)}, \ldots, H_{(k-1)}$ are rejected false hypotheses, and

$$k - 1 \leq m - m_0,$$

since $m - m_0$ is the true number of false hypotheses. Further,

$$m_0 \leq m - (k-1) \iff \frac{1}{m-(k-1)} \leq \frac{1}{m_0}.$$

As $H_{(k)}$ was rejected by Holm, $p_{(k)} \leq \frac{\alpha}{m-(k-1)} \leq \frac{\alpha}{m_0}$. Therefore, to obtain at least one false positive, the first instance must have a associated p-value less than or equal to $\frac{\alpha}{m_0}$. Let $q_1, q_2, \ldots, q_{m_0}$ denote the p-values of the true hypotheses. Thus by Boole's inequality and (1),

$$\text{FWER} = P(V > 0) = P\left(\bigcup_{i=1}^{m_0} q_i \leq \frac{\alpha}{m_0}\right) \leq \sum_{i=1}^{m_0} P(q_i \leq \frac{\alpha}{m_0}) \leq m_0 \frac{\alpha}{m_0} = \alpha.$$

■

C Proof: The Simes inequality

For m independent hypotheses with associated exact p-values $p_1, \ldots, p_m$ based on continuous test statistics, we have for a significance level $0 \leq \alpha \leq 1$,

$$P\left(\bigcup_{i=1}^{m} p_{(i)} \leq \frac{i\alpha}{m}\right) = \alpha,$$

where $p_{(i)}$ represent the i'th ordered p-value.

Proof. Note that $p_{(i)}$ is the i'th order statistic of m iid $\mathcal{U}(0,1)$. Let $A_m(\alpha) = P\left(\bigcup_{i=1}^{m} p_{(i)} \leq \frac{i\alpha}{m}\right)$. We will by induction show that $A_m(\alpha) = \alpha$ for all $m = 1, 2, \ldots$. For $m = 1$ it is trivially true that $A_1(\alpha) = P(p_{(1)} \leq \alpha) = \alpha$. Further, for $m > 1$ we divide $p_{(1)}, p_{(2)}, \ldots, p_{(m-1)}$ by $p_{(m)}$. This gives the order statistics of $(m-1)$ iid $\mathcal{U}(0,1)$ random variables $\frac{p_{(1)}}{p_{(m)}} \leq \frac{p_{(2)}}{p_{(m)}} \leq \cdots \leq \frac{p_{(m-1)}}{p_{(m)}}$, independent of $p_{(m)}$. Let $\alpha < p \leq 1$, then

$$\begin{aligned}
P\left(\bigcup_{i=1}^{m-1} p_{(i)} \leq \frac{i\alpha}{m} \mid p_{(m)} = p\right) &= P\left(\bigcup_{i=1}^{m-1} \frac{p_{(i)}}{p} \leq \frac{i\alpha}{pm}\left(\frac{m-1}{m-1}\right) \mid p_{(m)} = p\right) \\
&= P\left(\bigcup_{i=1}^{m-1} p_{(i)} \leq \frac{i\frac{\alpha(m-1)}{pm}}{m-1}\right) \\
&= A_{m-1}\left(\frac{\alpha(m-1)}{pm}\right).
\end{aligned}$$

If $p_{(m)} \leq \alpha$, then clearly $A_m(\alpha) = 1$ for all m. We assume $A_m(\alpha) = \alpha$ for $m = k-1$, so we want to check for $m = k$. We write

$$A_k(\alpha) = \int_\alpha^1 P\left(\bigcup_{i=1}^{k-1} p_{(i)} \leq \frac{i\alpha}{k} \mid p_{(k)} = p\right) P(p_{(k)} = p)\, dp + \int_0^\alpha P(p_{(k)} = p)\, dp,$$

by the law of total probability. As $p_{(k)}$ is the k'th order statistic of k iid $\mathcal{U}(0,1)$ random variables, it's density function is $f(p) = kp^{k-1}$ for $0 \leq p \leq 1$. This gives

$$\begin{aligned}
A_k(\alpha) &= \int_\alpha^1 \frac{\alpha(k-1)}{pk} kp^{k-1}\, dp + \int_0^\alpha kp^{k-1}\, dp \\
&= \alpha(k-1)\int_\alpha^1 p^{k-2}\, dp + k\int_0^\alpha p^{k-1}\, dp \\
&= \alpha - \alpha^k + \alpha^k \\
&= \alpha.
\end{aligned}$$

Thus, proven by induction, $A_m(\alpha) = \alpha$ for all $m = 1, 2, \ldots$.

■

The above proof hold for Simes' original assumptions, namely independent tests with uniformly distributed p-values. These assumptions have however been proven more tolerant since Simes original paper in 1986. Later publications by Hochberg & Rom (1995), Sarkar (1998) and Block et al. (2013) have generalized Simes inequality to hold for positively dependent statistics, but it may be reversed for negatively dependent statistics.

D R code used in simulations

D.1 Distribution of p-values (Section 5.1)

```
set.seed(100)
p.val_case1 <- c()
p.val_case2 <- c()
p.val_case3 <- c()
for (i in 1:10000) {
  #First instance
  pop.heights <- rnorm(100,179,3)
  test.stat <- (mean(pop.heights)-180)/(3)*sqrt(100)
  p.val_case1 <- c(p.val_case1,pnorm(test.stat,0,1, F))

  #Second instance
  pop.heights <- rnorm(100,180,3)
  test.stat <- (mean(pop.heights)-180)/(3)*sqrt(100)
  p.val_case2 <- c(p.val_case2,pnorm(test.stat,0,1, F))

  #Third instance
  mean <- runif(1,180.0001,182)
  pop.heights <- rnorm(100,mean,3)
  test.stat <- (mean(pop.heights)-180)/(3)*sqrt(100)
  p.val_case3 <- c(p.val_case3,pnorm(test.stat,0,1, F))
}

breaks <- seq(0,1, by = 0.04)
hist(p.val_case1, breaks, col = "lightblue", main = "P-values from true hypotheses",
     xlab = "P-value")
hist(p.val_case2, breaks, col = "lightblue", main = "P-values from true hypotheses",
     xlab = "P-value")
hist(p.val_case3, breaks, col = "lightblue", main = "P-values from false hypotheses",
     xlab = "P-value")
```

D.2 Independent tests (Section 5.2)

```
#Low proportion of false hypotheses
numb_reject_nc = c()
numb_reject_bonf = c()
numb_reject_holm = c()
numb_reject_hochberg = c()
numb_reject_hommel = c()
nc_fp <- c()
bonf_fp <- c()
holm_fp <- c()
hochberg_fp <- c()
hommel_fp <- c()

for (i in 1:100) {
  #Constructing the p-values
  obs = c(rnorm(7000,0,1),rnorm(1500,4,1),rnorm(1500,-4,1))
  p_val = 2 * pnorm(-abs(obs))

  #Performing the procedures
  bonf = p.adjust(p_val, method = "bonferroni", n = length(p_val))
  holm = p.adjust(p_val, method = "holm", n = length(p_val))
  hochberg = p.adjust(p_val, method = "hochberg", n = length(p_val))
  hommel = p.adjust(p_val, method = "hommel", n = length(p_val))
```

```
  #Finding the number of rejections
  numb_reject_nc <- c(numb_reject_nc, length(which(p_val < 0.05)))
  numb_reject_bonf = c(numb_reject_bonf, length(which(bonf < 0.05)))
  numb_reject_holm = c(numb_reject_holm, length(which(holm < 0.05)))
  numb_reject_hochberg = c(numb_reject_hochberg, length(which(hochberg < 0.05)))
  numb_reject_hommel = c(numb_reject_hommel, length(which(hommel < 0.05)))

  nc_fp <- c(nc_fp, length(which(p_val[1:7000] < 0.05)))
  bonf_fp <- c(bonf_fp,length(which(bonf[1:7000] < 0.05)))
  holm_fp <- c(holm_fp,length(which(holm[1:7000] < 0.05)))
  hochberg_fp <- c(hochberg_fp,length(which(hochberg[1:7000] < 0.05)))
  hommel_fp <- c(hommel_fp,length(which(hommel[1:7000] < 0.05)))

}

mean(numb_reject_nc)
mean(numb_reject_bonf)
mean(numb_reject_holm)
mean(numb_reject_hochberg)
mean(numb_reject_hommel)

mean(nc_fp)
mean(bonf_fp)
mean(holm_fp)
mean(hochberg_fp)
mean(hommel_fp)

#Large proportion of false hypotheses
numb_reject_nc = c()
numb_reject_bonf = c()
numb_reject_holm = c()
numb_reject_hochberg = c()
numb_reject_hommel = c()
nc_fp <- c()
bonf_fp <- c()
holm_fp <- c()
hochberg_fp <- c()
hommel_fp <- c()

for (i in 1:100) {

  #Constructing the p-values
  obs = c(rnorm(3000,0,1),rnorm(3500,4,1),rnorm(3500,-4,1))
  p_val = 2 * pnorm(-abs(obs))

  #Performing the procedures
  bonf = p.adjust(p_val, method = "bonferroni", n = length(p_val))
  holm = p.adjust(p_val, method = "holm", n = length(p_val))
  hochberg = p.adjust(p_val, method = "hochberg", n = length(p_val))
  hommel = p.adjust(p_val, method = "hommel", n = length(p_val))

  #Finding the number of rejections
  numb_reject_nc <- c(numb_reject_nc, length(which(p_val < 0.05)))
  numb_reject_bonf = c(numb_reject_bonf, length(which(bonf < 0.05)))
  numb_reject_holm = c(numb_reject_holm, length(which(holm < 0.05)))
  numb_reject_hochberg = c(numb_reject_hochberg, length(which(hochberg < 0.05)))
  numb_reject_hommel = c(numb_reject_hommel, length(which(hommel < 0.05)))
```

```
  nc_fp <- c(nc_fp, length(which(p_val[1:3000] < 0.05)))
  bonf_fp <- c(bonf_fp,length(which(bonf[1:3000] < 0.05)))
  holm_fp <- c(holm_fp,length(which(holm[1:3000] < 0.05)))
  hochberg_fp <- c(hochberg_fp,length(which(hochberg[1:3000] < 0.05)))
  hommel_fp <- c(hommel_fp,length(which(hommel[1:3000] < 0.05)))

}

mean(numb_reject_nc)
mean(numb_reject_bonf)
mean(numb_reject_holm)
mean(numb_reject_hochberg)
mean(numb_reject_hommel)

mean(nc_fp)
mean(bonf_fp)
mean(holm_fp)
mean(hochberg_fp)
mean(hommel_fp)
```

D.3 Positive dependent tests (Section 5.3)

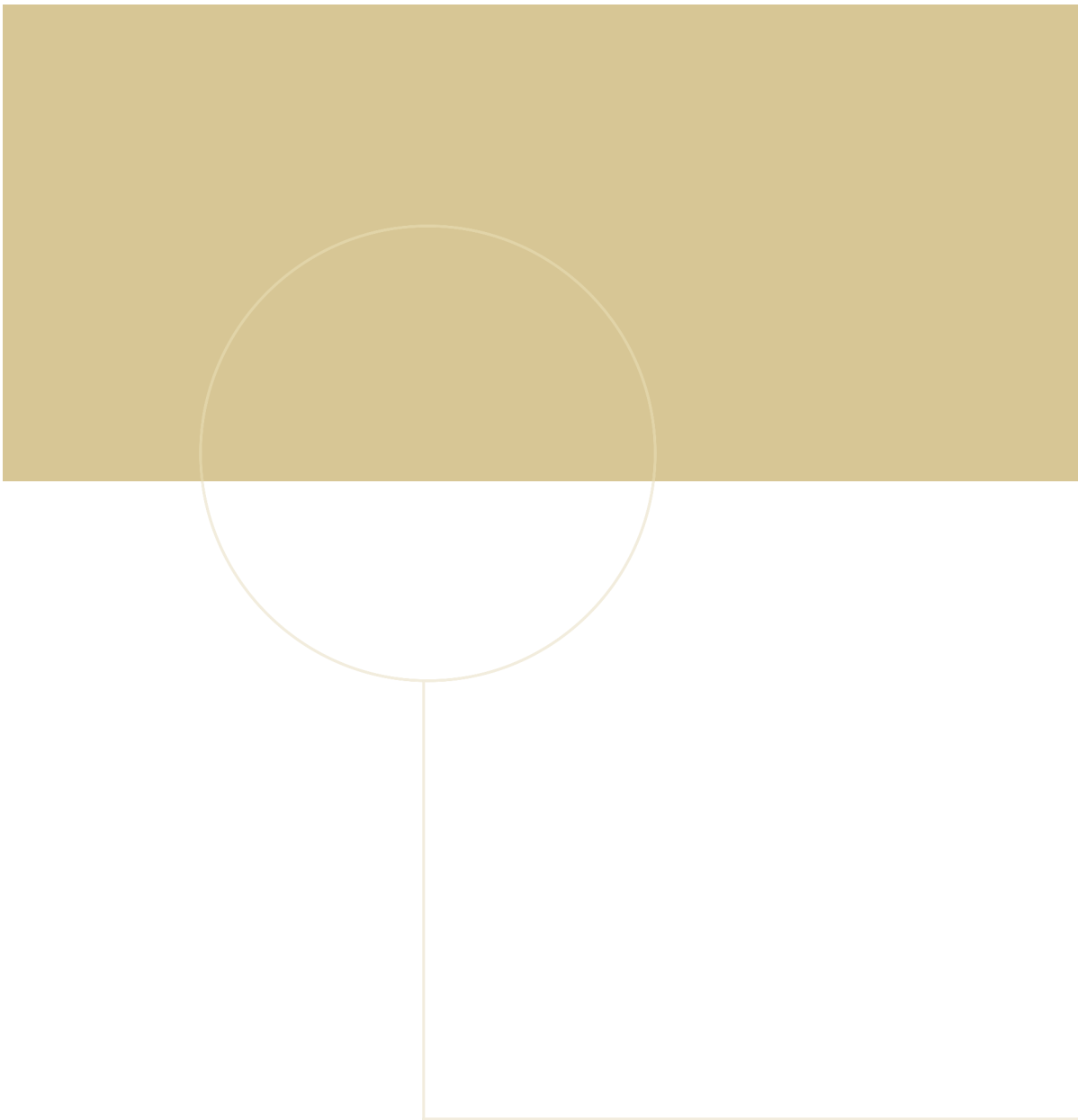
```
numb_reject_nc = c()
numb_reject_bonf = c()
numb_reject_holm = c()
numb_reject_hochberg = c()
numb_reject_hommel = c()
nc_fp <- c()
bonf_fp <- c()
holm_fp <- c()
hochberg_fp <- c()
hommel_fp <- c()

for (i in 1:100) {

  #Constructing the p-values
  g1 <- rnorm(30,15,3)
  g2 <- rnorm(50,40,3)
  g3 <- rnorm(25,60,3)
  g4 <- rnorm(30,-5,3)
  g5 <- rnorm(7,30,3)
  obs <- c(g1,g2,g3,g4,g5)

  p_val <- c()
  for (i in 1:(length(obs)-1)) {
    for (j in (i+1):length(obs)) {
      Z = (obs[i]-obs[j])/(sqrt(18))
      p_val <- c(p_val, 2*pnorm(-abs(Z)))
    }
  }

  #Performing the procedures
  bonf = p.adjust(p_val, method = "bonferroni", n = length(p_val))
  holm = p.adjust(p_val, method = "holm", n = length(p_val))
  hochberg = p.adjust(p_val, method = "hochberg", n = length(p_val))
  hommel = p.adjust(p_val, method = "hommel", n = length(p_val))

  #Finding the number of rejections
```

```
  numb_reject_nc <- c(numb_reject_nc, length(which(p_val < 0.05)))
  numb_reject_bonf = c(numb_reject_bonf, length(which(bonf < 0.05)))
  numb_reject_holm = c(numb_reject_holm, length(which(holm < 0.05)))
  numb_reject_hochberg = c(numb_reject_hochberg, length(which(hochberg < 0.05)))
  numb_reject_hommel = c(numb_reject_hommel, length(which(hommel < 0.05)))

  nc_fp <- c(nc_fp, length(which(p_val[true] < 0.05)))
  bonf_fp <- c(bonf_fp,length(which(bonf[true] < 0.05)))
  holm_fp <- c(holm_fp,length(which(holm[true] < 0.05)))
  hochberg_fp <- c(hochberg_fp,length(which(hochberg[true] < 0.05)))
  hommel_fp <- c(hommel_fp,length(which(hommel[true] < 0.05)))

}

mean(numb_reject_nc)
mean(numb_reject_bonf)
mean(numb_reject_holm)
mean(numb_reject_hochberg)
mean(numb_reject_hommel)

mean(nc_fp)
mean(bonf_fp)
mean(holm_fp)
mean(hochberg_fp)
mean(hommel_fp)
```

D.4 Negatively dependent tests (Section 5.4)

```
#Creating mean vector and covariance matrix.
n <- 6
var_matrix <- diag(1, n, n)
corr <- -0.16
var_matrix[lower.tri(var_matrix)] <- corr
var_matrix[upper.tri(var_matrix)] <- corr
Sigma <- var_matrix
mu <- rep(0, times = n)

#Function for calculating the average proportion of false positives using
  #Bonferroni and Simes global test.
falsePositive <- function(Sigma, mu, alpha) {
  false_pos.s <- c()
  false_pos.b <- c()
  for (i in 1:500) {
    false.s <- c()
    false.b <- c()
    for (i in 1:1000) {
      fp.s = 0
      fp.b = 0
      tests <- mvrnorm(1,mu,Sigma)
      p.val <- pnorm(tests,0,1, F)
      p.val <- sort(p.val)
      for (i in 1:6) {
        if (p.val[i] < (i*alpha)/6) {
          fp.s = 1
          break
        }
      }
      if (p.val[1] < alpha/6) {
```

```
        fp.b = 1
      }

      false.s <- c(false.s, fp.s)
      false.b <- c(false.b, fp.b)

    }

    false_pos.s <- c(false_pos.s,length(which(false.s == 1)))
    false_pos.b <- c(false_pos.b, length(which(false.b == 1)))

  }

  c(mean(false_pos.b)/1000, mean(false_pos.s)/1000)

}

#Use for different levels of alpha (here alpha = 0.05).
falsePositive(Sigma, mu, 0.05)
```



NTNU

Kunnskap for en bedre verden