Pomona College

Senior Thesis in Mathematics

Multiple Comparison Test

Author:
Helen Lan

Advisor:
Dr. Jo Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 28, 2020

Abstract

Statistical hypothesis testing is important in validating discoveries in real-world data. As data become increasingly easy to collect, there will usually be simultaneous hypothesis tests on different covariates. For example, in considering a new drug's effectiveness as compared with a control group, we might want to look at how patients respond to the new drug in terms of multiple disease symptoms. As the number of symptoms we are looking at increases, it will become more likely that the new drug "improves" patients' health condition in at least one of the many symptoms measured. There will be inevitably some false positives if we apply an individual significance level to each of many tests without any correction. Therefore, multiple comparison tests are critical tools to control Type I Errors.

In my thesis, I will study a number of correction methods that are used to control family-wise error rate (FWER) or false discovery rate (FDR). Specifically, I will illustrate how Sidak, Bonferroni, and Holm correction methods control FWER and how Benjamini-Hochberg controls FDR. I will then conduct simulation studies to test the effectiveness of different types of multiple correction tests.

Contents

Chapter 1

Introduction

In statistical tests, null hypotheses are statements that assume no difference or relationship across two conditions in the variable we are interested in. In order to test the null hypothesis, we use a significance level to measure how strong the evidence in the sample data is to reject the null hypothesis. We compare the significance level to the resulting p-value from the test. If the p-value is smaller than the significance level, 0.05 for example, than it shows that in our sample there is evidence that the null hypothesis is false and therefore we reject the null hypothesis.

Multiple Testing Procedures (MTPs) work to control the error rate of an entire study. There are three types[Emmert-Streib and Dehmer, 2019] of MTPs in general: single-step, step-up (SU), and step-down (SD). Let $H_1, ..., H_m$ be a family of tested null hypotheses and $p_1, ..., p_m$ their corresponding individual p-values rearranged in ascending orders. Suppose that among the m null hypotheses, m_0 of them are true. Each MTP involves a corrected significance level, c_i, to reject the respective null hypothesis, H_i, if $p_i \leq c_i$.

1.1 Single-step

Single-step MTPs set a single correction procedure to all hypothesis tests in one study.[Emmert-Streib and Dehmer, 2019] For instance, the Sidak correction and Bonferroni correction that will be discussed in Section 2.2.1 and 2.2.2 are single-step MTPs.

Formally, in a single-step MTP, reject all H_i such that $p_i \leq c_i$, where

$c_i = c \; \forall \; i$.

1.2 Step-up

Step-up (SU) MTPs start from the least significant, or the largest, p-value and proceed toward the most significant, or the smallest, p-value. [Emmert-Streib and Dehmer, 201 In this sequence, step-up MTPs test successively tests whether $p_i \leq c_i$. At the first index i for which this condition holds the procedure stops and rejects all null hypotheses j with $j \leq i$. If such an index does not exist do not reject any null hypothesis.

Formally, in a step-up MTP, identify the idex

$$i^* = \max_{i \in \{1,...,m | p_i \leq c_i\}}$$

and reject all null hypothesis H_j with $j \leq i^*$.

1.3 Step-down

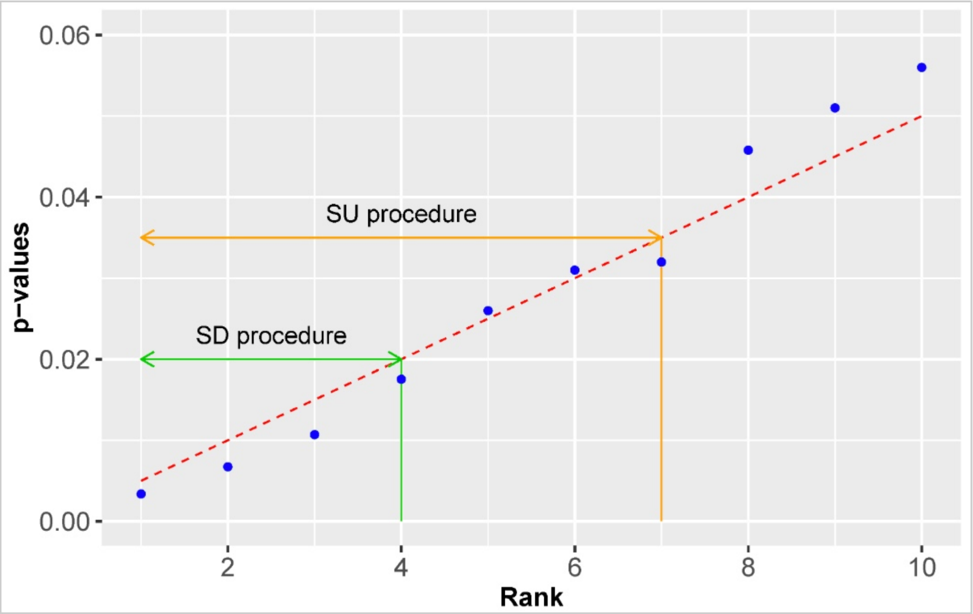
Step-down (SD) MTPs start from the most significant, or the smallest, p-value and proceeds toward the least significant, or the biggest, p-value. [Emmert-Streib and Dehmer, 2019] In this sequence, step-down MTPs successively test whether $p_i \leq c_i$. The first index $i + 1$ for which this condition does not hold induces the procedure to stop and rejects all null hypothesis j with $j \leq i$. If such an index does not exist, reject all null hypotheses.

Formally, in a step-down MTP, identify the index

$$i^* = \max_{i \in \{1,...,m | p_j \leq c_j\}} \forall j \in 1, ..., i^*$$

and reject all null hypothesis H_j with $j \leq i^*$.

Note that the SD procedure is more conservative (rejects fewer null hypotheses) than the SU procedure because of the monotonicity requirement.



The graph [Emmert-Streib and Dehmer, 2019] above provides an example which juxtaposes SU and SD procedures. The blue dots represent p-values of individual hypothesis tests and the red dotted line represents the significance level, below which we reject the null hypotheses. An SD procedure starts from the smallest p-value and stops at the first blue dot above the red dotted line, or the first p-value greater than the significance level, rejecting all hypotheses with smaller p-values. In the case illustrated by the graph, the four hypotheses with the smallest four p-values are rejected. The SU procedure starts from the greatest p-value and stops at the first blue dot below the red dotted line, or the first p-value smaller than the corresponding significance level, rejecting all hypotheses with a smaller p-value, including the one that the procedures stops at. In the case illustrated by the graph, the six hypotheses with the smallest six p-values are rejected. A single step procedure would reject every hypothesis with p-values smaller than the cut-off level, or below the red-dotted line. Five hypotheses are rejected in the illustrated example. Therefore, An SD procedure is the most conservative, rejecting fewer hypotheses and therefore declaring fewer significant effects, than an SU or a single-step procedure.

Chapter 2

Controlling Family Wise Error Rate

2.1 Family Wise Error Rate (FWER)

Definition 2.1 *Type I Error happens when a true null hypothesis is falsely rejected.*

Definition 2.2 *Family Wise Error Rate (FWER) is the probability of having at least one Type I Error.*

$$FWER = Pr(\text{at least one Type I Error}) = 1 - Pr(\text{no Type I Error})$$

2.2 Multiple Comparison Procedures Controlling FWER

One goal of multiple comparison correction procedures is to correct the rejection level of each individual test in order to control the overall FWER at a pre-specified value of α for the entire study. Suppose that in a statistical study we have m null hypotheses. Let $H_1, ..., H_m$ be a family of null hypotheses and $p_1, ..., p_m$ their corresponding p-values (rearranged in ascending orders). Among the m null hypotheses, let m_0 of them be true. For all the correction methods that will be introduced, we assume that the goal is to control FWER at α.

2.2.1 Sidak

To control the FWER of the statistical study at α, the Sidak correction method controls the significance level of each individual hypothesis test at an adjusted significance level

$$\alpha_S = 1 - (1-\alpha)^{1/m}$$

assuming all m hypothesis tests are independent in the study.
Sidak Claim:[Šidák, 1967] Let $H_1, ..., H_m$ be a family of independent null hypotheses and $p_1, ..., p_m$ their corresponding p-values (rearranged in ascending orders). Among the m null hypotheses, let m_0 of them be true. The Sidak correction method to adjust the individual significance level to

$$\alpha_S = 1 - (1-\alpha)^{1/m}$$

controls the FWER of the study at α.
Proof By Definition 2.2,

$$\begin{aligned}
FWER &= Pr(\text{at least one Type I Error}) \\
&= 1 - Pr(\text{no Type I Error}) \\
& \text{(Because the null hypotheses are assumed to be independent,} \\
& \text{the probability of making Type I Errors for m hypothesis tests} \\
& \text{is the product of each individual probability)} \\
&= 1 - (Pr(\text{Type I Error for individual hypothesis test}))^m \\
&= 1 - (1-\alpha_S)^m
\end{aligned}$$

Plugging in the significance level, α_S, adjusted by Sidak correction,

$$\begin{aligned}
FWER &= 1 - (1 - 1 + (1-\alpha)^{1/m})^m \\
&= 1 - ((1-\alpha)^{1/m})^m \\
&= 1 - (1-\alpha) \\
&= \alpha
\end{aligned}$$

∎

The problem with Sidak correction method is that its ability to control the

FWER of a study relies on an unrealistic assumption of independent hypothesis tests. In the real world, the independent variables that we are testing tend to correlate with each other to some extent, making independence of hypothesis tests on these variables unrealistic.

2.2.2 Bonferroni

Bonferroni correction method takes the dependence of hypothesis tests in the real world into consideration. In order to control the FWER of a study at α, Bonferroni correction method adjusts the significance level of each individual hypothesis test at

$$\alpha_B = \alpha/m$$

where m is the total number of hypotheses.

The proof of Bonferroni method [Goeman and Solari, 2014] relies on Boole's Inequality, which states that for any finite or countable set of events, the probability that at least one of the events happens is no greater than the sum of the probability of those individual events. Formally, for any countable set of events $A_1, A_2, ..., A_n$,

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

Bonferroni Claim:[Goeman and Solari, 2014] Let $H_1, ..., H_m$ be a family of hypotheses and $p_1, ..., p_m$ their corresponding p-values (rearranged in ascending orders). Among the m null hypotheses, let m_0 of them be true. The Bonferroni correction method to adjust individual significance level to $\alpha_B = \alpha/m$ controls the FWER of the study at α.
Proof [Goeman and Solari, 2014] According to Definition 2.1 and Definition 2.2, FWER measures the probability of having at least one Type I Error, where we falsely rejected the true null hypothesis. There are m_0 true null hypothesis and Type I Error happens when we falsely reject one of those m_0 true null hypotheses. We know that we reject an individual null hypothesis if its corresponding p-value is less than the adjusted significance level.

Therefore, FWER equals the probability that at least one among the m_0 true hypotheses is falsely rejected, or where its corresponding p-value is less than the adjusted significance level.

$$\begin{aligned}
FWER &= Pr\bigg(\bigcup_{i=1}^{m_0}(p \leq \alpha_B)\bigg) \\
&= Pr\bigg(\bigcup_{i=1}^{m_0}(p \leq \alpha/m)\bigg) \\
&\quad \text{(By Boole's Inequality, we know that the probability} \\
&\quad \text{ of having at least one Type I Error is no greater than} \\
&\quad \text{the probability of the sum of having } m_0 \text{ Type I Errors.)} \\
&\leq \sum_{i=1}^{m_0} Pr(p \leq \alpha/m) \\
&\quad \text{(Because when null hypotheses are true, p-values are} \\
&\quad \text{assumed to be uniformly distributed between 0 and 1,} \\
&\quad Pr(p \leq \alpha/m) = \alpha/m.\text{)} \\
&= m_0\alpha/m \\
&\leq m\alpha/m \\
&= \alpha
\end{aligned}$$

$\blacksquare$

2.2.3 Holm

Unlike Sidak and Bonferroni, which are single-step procedures, Holm is a step-down (SD) procedure. Recall from **Section 1.3**, in an SD procedure, we start from the most significant, or the smallest, p-value and proceed towards the least significant, or biggest, p-value. To control the FWER of a statistical study at α, Holm correction method controls the significance level of each individual hypothesis test at

$$\alpha_H = \frac{\alpha}{(m - i + 1)}$$

In Holm's multiple comparison procedure, we successively test whether the individual p-value is less than the adjusted significance level for an individual hypothesis test. We reject all null hypotheses starting from the first hypothesis for which the corresponding p-value is greater than the corresponding adjusted significance level onward.

Holm Claim:[Holm, 1979] Let $H_1, ..., H_m$ be a family of null hypotheses and $p_1, ..., p_m$ their corresponding p-values (rearranged in ascending orders). Among the m null hypotheses, let m_0 of them be true and thus $m - m_0$ of them are false. The Holm procedure adjusts the individual significance level to $\alpha_H = \frac{\alpha}{(m-i+1)}$, where i is the index of the ordered p-values and their corresponding hypotheses. For the minimal index, i, where $p_i > \frac{\alpha}{(m-i+1)}$, stop the procedure and reject all null hypotheses $H_1, H_2, ..., H_{i-1}$, failing to reject $H_i, H_{i+1}, ..., H_m$. The Holm procedure controls the FWER of a statistical study at a given level, α.

Proof [Holm, 1979] Let I_0 be the set of indices corresponding to the unknown true null hypotheses. Since we assumed that m_0 out of m hypotheses are truly null, the size of I_0 is m_0. Recall that FWER is defined to be the probability of having at least one Type I Error, or having at least one true hypothesis to be falsely rejected. Given all hypotheses are ranked by ascending p-values, let i^* be the index of the first falsely rejected null hypothesis. Then $H_1, H_2, ..., H_{i^*-1}$ are all correctly rejected false hypotheses and $i^* - 1 \leq m - m_0$ since there are $m - m_0$ false hypotheses in total. It follows that $m - i^* + 1 \geq m_0$. Therefore, $\frac{1}{m-i^*+1} \leq \frac{1}{m_0}$. Since H_{i^*} is assumed to be rejected, according to the definition of Holm procedure, we know that $p_{i^*} \leq \frac{\alpha}{m-i^*+1}$. Since $\frac{1}{m-i^*+1} \leq \frac{1}{m_0}$, we also know that $p_{i^*} \leq \frac{\alpha}{m-i^*+1} \leq \frac{\alpha}{m_0}$. Thus, if we falsely reject a true null hypothesis, there has to be a true null hypothesis with p-value no greater than $\frac{\alpha}{m_0}$.

Let $A_i = \{p_i \leq \frac{\alpha}{m_0}\}$ for $i \in I_0$. Since I_0 is assumed to be the set of indices corresponding to the unknown true null hypotheses, i represent the index of a true hypothesis. If the i^{th} null hypothesis is rejected, then $i \in A_i$. In other words, A_i includes the event in which a Type I Error occurs. Given p-values are assumed to be uniformly distributed when the null hypothesis is true,

$Pr(A_i) = \frac{\alpha}{m_0}$.

$$\begin{aligned}
FWER &= Pr(\text{at least one Type I Error}) \\
&= Pr(\text{at least one true null hypothesis is falsely rejected}) \\
&= Pr\left(p_i \leq \frac{\alpha}{m-i+1}\right) \text{ for } i \in I_0 \\
&\quad \text{(a true null hypothesis can only be falsely rejected} \\
&\quad \text{if its corresponding p-value, } p_i\text{, is smaller than} \\
&\quad \text{the adjusted Holm cutoff)} \\
&\leq Pr\left(\bigcup_{i \in I_0} (A_i)\right) \\
&\quad \text{(since } A_i \text{ includes the event in which a Type I Error} \\
&\quad \text{occurs, the probability of at least one Type I Error} \\
&\quad \text{is smaller or equal to the probability of the union of } A_i\text{)} \\
&\leq \sum_{i \in I_0} Pr(A_i) \\
&\quad \text{(by Boole's Inequality)} \\
&= m_0 \times Pr(A_i) \\
&= m_0 \times \frac{\alpha}{m_0} \\
&= \alpha
\end{aligned}$$

∎

Chapter 3

Controlling False Discovery Rate

3.1 False Discovery Rate

Definition 3.1 *False Discovery Rate (FDR) measures the expected proportion of falsely rejected hypotheses to all rejections or all tests declared significant. [Benjamini and Hochberg, 1995]*

$$FDR = \mathbb{E}\left(\frac{V_m}{R_m}\right)$$

	Fail to reject null	Reject null	Total
Null true	U$_m$	V$_m$	m_0
Null false	T$_m$	S$_m$	m_1
	W$_m$	R$_m$	m

Table 3.1: Possible outcomes from m hypothesis tests

3.2 Multiple Comparison Procedures Controlling FDR

As we saw in Chapter 2, multiple comparison procedures designed to control the FWER are one category of methods to avoid too many Type I Errors in multiple comparison studies. Another category of methods aims to control False Discovery Rate.

In all of the procedures below, our goal is to control FDR at α. Let $H_1, ..., H_m$ be a family of hypotheses and $p_1, ..., p_m$ their corresponding p-values (rearranged in ascending orders). Let m be the total number of hypotheses and m_0 the number of true null hypotheses.

3.2.1 Benjamini-Hochberg

The Benjamini-Hochberg procedure assumes ordered p-values as introduced in Section 3.2. Then by a step-up procedure it identifies the largest index k for which

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

holds and rejects the null hypotheses $H_1, ..., H_k$. In the original paper by Benjamini and Hochberg [Benjamini and Hochberg, 1995], they proved by induction over m that the step-up procedure controls FDR within $\frac{m_0\alpha}{m}$. In this paper, an interpreted proof of a theorem from Finner and Roters [Finner and Roters, 2001] will show that the Benjamini-Hochberg procedure controls FDR at α.

Finner and Roters Claim:[Finner and Roters, 2001] Let m be the total number of independent null hypotheses. Let m_0 be the number of true hypotheses and $m-m_0$ be the number of false null hypotheses. If the p-values are uniformly distributed under the corresponding null hypotheses, then

$$FDR = \frac{m_0\alpha}{m}$$

Proof [Finner and Roters, 2001] Without the loss of generality, let $J_n = \{1, ..., m - m_0\}$ denote the index set of false null hypotheses and $I_n = \{m - m_0 + 1, ..., m\}$ denote the index set of true null hypotheses, and let $r = m - m_0$. As in Table 3.1, let V_m represent the number of falsely rejected null hypotheses and let S_m denote the number of correctly rejected null hypotheses. Let $R_m = V_m + S_m$ denote the total number of null hypotheses rejected in the tests. Let α_i be the adjusted significance level of individual hypothesis test under the Benjamini-Hochberg procedure. Then, according to the definition of FDR, the actual FDR in this study can be calculated as

[Finner and Roters, 2001]

$$FDR = \sum_{v=1}^{m_0} \sum_{s=0}^{r} \frac{v}{v+s} \Pr(V_m = v, S_m = s, |I_m, J_m, (\alpha_1, ..., \alpha_m)), \qquad (3.1)$$

which is the expected proportion of false rejections to all rejections in the event where v hypotheses are falsely rejected and s hypotheses are correctly rejected dependent on I_m, J_m, and the step-up Benjamini-Hochberg procedure with critical values $\alpha_1, ..., \alpha_m$.

Since the p-values are assumed to be independent and uniformly distributed over [0,1] when the null hypothesis is true, we obtain [Finner and Roters, 2001]

$$\Pr\Bigg(V_m = v, S_m = s, |I_m, J_n, (\alpha_1, ..., \alpha_m)\Bigg) =$$

$$\tbinom{m_0}{v}(\alpha_{v+s})^v \times Pr\Bigg(V_{m-v} = 0, S_{m-v} = s, |I_m \backslash m - v + 1, ..., m, J_m, (\alpha_{v+1}, ..., \alpha_m)\Bigg)$$

(3.2)

This equation says that we allow v out of m_0 hypotheses to be falsely rejected at significance level α and none of the rest $m - v$ hypotheses are falsely rejected.

According to the Benjamini-Hochberg procedure, each significance level is adjusted to be $\alpha_i = \frac{i}{m}\alpha$, therefore

$$
\begin{aligned}
\frac{v}{v+s}\binom{m_0}{v}(\alpha_{v+s})^v &= \frac{v}{v+s}\binom{m_0}{v}\left(\frac{v+s}{m}\alpha\right)^v \\
&= \frac{v}{v+s}\left(\frac{v+s}{m}\alpha\right)\binom{m_0}{v}\left(\frac{v+s}{m}\alpha\right)^{v-1} \\
&= \alpha \times \frac{v}{m}\binom{m_0}{v}\left(\frac{v+s}{m}\alpha\right)^{v-1} \\
&= \alpha\frac{v}{m}\left(\frac{m_0!}{v!(m_0-v)!}\right)\left(\frac{v+s}{m}\alpha\right)^{v-1} \\
&= \alpha\frac{v}{m}\frac{m_0}{v}\left(\frac{(m_0-1)!}{(v-1)!(m_0-v)!}\right)\left(\frac{v+s}{m}\alpha\right)^{v-1} \\
&= \alpha\frac{m_0}{m}\binom{m_0-1}{v-1}\left(\frac{v+s}{m}\alpha\right)^{v-1} \\
&= \frac{m_0}{m}\alpha\binom{m_0-1}{v-1}(\alpha_{v+s})^{v-1}
\end{aligned}
$$

(3.3)

Hence we obtain by substituting $v^* = v-1$ for v [Finner and Roters, 2001]

$$
\begin{aligned}
FDR &= \frac{m_0}{m}\alpha\sum_{v^*=0}^{m_0-1}\sum_{s=0}^{r}\binom{m_0-1}{v^*}\alpha^v_{v^*+s+1} \\
&\quad \times \Pr(V_{m-v^*-1} = 0, S_{m-v^*-1} = s, |I_m \backslash m - v^* + 1, ..., m, J_m, (\alpha_{v^*+2}, ..., \alpha_m)) \\
&= \frac{m_0}{m}\alpha\sum_{v^*=0}^{m_0-1}\sum_{s=0}^{r} \times \Pr(V_{m-1} = v^*, S_{m-1} = s, |I_m \backslash m, J_m, (\alpha_2, ..., \alpha_m)) \\
&= \frac{m_0}{m}\alpha \times \Pr(V_{m-1} \in 0, ..., m_0 - 1, S_{m-1} \in 0, ..., r | I_m \backslash m, J_m, (\alpha_2, ..., \alpha_m)) \\
&= \frac{m_0}{m}\alpha
\end{aligned}
$$

The last step follow from the Law of Total Probability because those sets contain the only values that V_{m-1} and S_{m-1} can take. Therefore, the

Benjamini-Hochberg procedure makes the average FDR equal to $\frac{m_0}{m}\alpha$. Since m_0 is always smaller or equal to m, $\frac{m_0}{m} \leq 1$ and $\frac{m_0}{m}\alpha \leq \alpha$. Therefore, the Benjamini-Hochberg procedure controls FDR at α.

$\blacksquare$

Chapter 4

Application and Simulation

4.1 Recipe Data Applications

The purpose of multiple comparison test is to improve conclusions made in real applications. Although we are not able to distinguish which hypotheses are null and which are alternative in the real world, it is proved in the previous two chapters that different multiple comparison procedures can control the FWER and the FDR. To illustrate how multiple comparison tests work in practice, I used a recipe dataset on Kaggle scraped from Epicurious.com [1]. This dataset contains 679 variables and 14,429 recipes. There are two types of variables in the dataset, numeric and binary. Numeric variables represent descriptive and nutritional acts about the recipe, namely ratings, calories, protein, fat, and sodium standardized in one serving portion. Binary variables are tags of recipes. A tag can be an incredient, such as broccoli, a type of food, such as alcoholic, a situation associated with the recipe, such as birthday, etc.

To explore the research question of whether certain tags significantly differentiates calories of the recipes, we conduct t-tests to test for significant difference in means among two groups: the group of recipes that contains a certain tag and the group that does not. Let us consider two tags, for example, alcoholic and almond. The mean calories of the groups of recipes that contain alcoholic and almond, respectively, or not are shown in Table 4.1.

[1] https://www.kaggle.com/hugodarwood/epirecipes

	tag	sample size	mean	sd	p-value from t-test
alcoholic	1	581	514.8	624.6	$5.51 * 10^{-118}$
non-alcoholic	0	13,848	225.1	161.4	
	tag	sample size	mean	sd	p-value from t-test
almond	1	448	503.7	618.2	0.515
no almond	0	13,981	520.5	487. 3	

Table 4.1: Mean calories

Suppose that we set the significance level at 0.05 for the tests conducted in this paper. The resulting p-values from t-tests suggest that the tag alcoholic significantly differentiates calories whereas the tag almond does not. Such results are consistent with what the mean values demonstrate. Alcoholic and non-alcoholic groups have vastly different means whereas almond and non-almond groups have a much smaller difference in means. For the 380 tags with at least 30 observations, t-tests were run. The overall study contains m=380 hypothesis tests.

Correction Procedure	Adjusted Significance Level	Rejected Hypotheses
Sidak	$\alpha_S = 1 - (1-\alpha)^{1/m}$	145
Bonferroni	$\alpha_B = \alpha/m$	145
Holm	$\alpha_H = \alpha/(m-i+1)$	150
Benjamini-Hochberg	$\alpha_{BH} = i\alpha/m$	156
Unadjusted	$\alpha = 0.05$	238

Table 4.2: Results of Multiple Comparison Procedure

From Table 4.2, we can see that if we apply the 0.05 unadjsuted significance level to the 380 t-tests, we end up rejecting 238 hypotheses. However, if we apply multiple comparison tests, we end up with a much smaller number of rejections. The results show that single-step procedures, Sidak and Bonferroni, seem most conservative because they declared the fewest significant results. The step-up procedure, Holm, is less conservative than single-step procedures. The least conservative procedure is Benjamini-Hochberg, which is a step-down procedure. All of these results align with what was introduced in the theoretical sections.

4.2 Simulations

Since the truth is unknown, a simulation is useful so that the proportion of true null hypotheses to investigate is known. The effectiveness of multiple comparison tests can then be assessed given the truth. The purpose of the simulation is to demonstrate multiple comparison tests' ability to control error rates and reduce the number of false positives, or Type I Errors, in real-world data. Although independence is required for proofs of multiple comparison procedures in Chapters 2 and 3 except for the Bonferroni method, the underlying dependence structure in real-world data is unknown. Thus, multiple comparison tests may not control the error rate at exactly the given level. Because it is difficult to define correlation between independent variables, the simulation described in this paper is built upon the nutrition data set described in Section 4.1. No matter what the relationship between variables is, it is preserved in the simulation.

4.2.1 Method

First, the nutrition data set is randomly shuffled, i.e., calories are randomly re-assigned to each recipe. This randomness simulates a dataset with entirely true null hypotheses, where tags do not differentiate calories, while keeping the correlation strucutre from recipe to recipe. The re-sampled data set is then filtered to only keep tags with more than 30 observations for the validity of t-tests later. A proportion of the remaining null hypotheses on 380 tags is randomly selected to be false This is done through an addition of 200 calories on the recipes where those selected tags are present. Given that the calories in this recipe data set have a mean of 503 and a standard deviation of 615, a slight variation in calories should be tolerated for insignificant differentiation. An addition of 200 calories can reasonably set a tag's null hypothesis to be false. Whether a tag significantly differentiates calories, or whether it is associated with a true or false null hypothesis, is kept track of by a binary variable.

Next, 380 t-tests are run and the resulting p-values are stored. The null hypotheses are rejected the 5% significance level according to unadjusted p-values. Then four multiple comparison tests discussed in Chapters 2 and 3, namely Sidak, Bonferroni, Holm, and Benjamini-Hochberg, are performed and the null hypotheses are rejected at an overall error rate of 5%. Each tag's rejection status is indicated by a binary variable. The False Discovery

Proportion, the proportion of false rejections to all rejections, is calculated in each of the five procedures, one unadjusted and four adjusted by multiple comparison tests.

The above procedure is iterated for 100 times. Family Wise Error rate is calculated by the number of times where at least one Type I Error, or false rejection, occurred over 100, the number of iterations. The average value of the 100 False Discovery Proportions is calculated to be the False Discovery Rate of the 100 iterations.

4.2.2 Results

The proportion of false null hypotheses is set to be 0.5 for the first 100 iterations, which means that half of the hypotheses are set so that the null is true and the other half are set so that the null is false. The results are presented in Table 4.3.

Multiple comparison test	FDR on 100 iterations	FWER on 100 iterations
Unadjusted	0.088 (0.022)	1.00
Sidak		0.72
Bonferroni		0.72
Holm		0.77
Benjamini-Hochberg	0.063 (0.019)	
	(sd in parentheses)	

Table 4.3: FDR and FWERx with 50% False Null Hypothesis

As shown in Table 4.3, without multiple comparison tests, the FDR is 0.088 and the FWER is 1, both of which are greater than 0.05, the level we want to control. FWERs after the Sidak, Bonferroni, and Holm corrections are still very high, above 70%. Since Sidak and Bonferroni are the most conservative procedures, they yield the smallest FWERs. The high resulting FWERs demonstrate that the independence assumption required by the Sidak and Holm proofs are unrealistic. Yet the Bonferroni procedure, which does not require independence, also failed to control the FWER at 0.05. Part of this result may be attributable to the small number of iterations (only 100) run due to computational burdens in R. The high FWERs might also speak to the conservative nature of FWER, which looks at the events where at least one Type I Error occurred. When it comes to real-world data, it is highly likely that a Type I Error will occur without knowing the truth underlying the data. Therefore, the resulting FWERs from the simulations speaks to why FWER is not the best measure of overall error rate and why FDR was proposed as an alternative by Benjamini and Hochberg. [Benjamini and Hochberg, 1995]

The Benejamini-Hochberg procedure yields an FDR of 0.063, lower than the unadjusted FDR, yet failing to control the FDR at 0.05. This happens because of a few reasons. First, the proof in Section 3.2.1 requires independence, which might not exist in the data set. For instance, recipes tagged "birthday" are more likely to be tagged "cake" as well. Although Benjamini and Yekutieli proved that the Benjamini-Hochberg procedure also works when the test statistics have positive regression dependency on each of the test statistics corresponding to the true null hypotheses [Benjamini and Yekutieli, 2001], we do not know if the condition holds in the recipe data. For example, recipes tagged "broccoli" are less likely to be tagged "banana". Second, the FDR calculated is the expected value of FDP from 100 iterations. A larger number of iterations is likely to yield different results. Third, the proportion of true vs. false null hypotheses is crucial in determining the probability of false rejections. Recall that the Benjamini-Hochberg procedure is proved to set FDR $= \frac{m_0\alpha}{m}$ under independence. The greater the proportion of true null hypotheses, $\frac{m_0}{m}$, the larger the FDR. The greater the proportion of false null hypotheses, the smaller the expected proportion of true null hypotheses in rejected hypotheses, i.e., false rejections.

To illustrate the effect of the proportion of true versus false null hypotheses, two simulations are designed with 80% false null hypotheses and 20% false null hypotheses, respectively. Table 4.4 and 4.5 show the result out of 100 iterations where only the proportion of false null hypotheses changed. The resulting FDRs and FWERs in Table 4.4 with 80% false null hypotheses are lower than the corresponding FDRs and FWERs in Table 4.5 with 20% false null hypotheses.

Multiple comparison test	FDR on 100 iterations	FWER on 100 iterations
Unadjusted	0.025 (0.009)	1.00
Sidak		0.41
Bonferroni		0.41
Holm		0.41
Benjamini-Hochberg	0.021 (0.008)	
	(sd in parentheses)	

Table 4.4: FDR and FWER with 80% False Null Hypothesis

Multiple comparison test	FDR on 100 iterations	FWER on 100 iterations
Unadjusted	0.279 (0.051)	1
Sidak		0.87
Bonferroni		0.87
Holm		0.87
Benjamini-Hochberg	0.156 (0.050)	
	(sd in parentheses)	

Table 4.5: FDR and FWER with 20% False Null Hypothesis

Chapter 5

Discussion and Conclusion

Multiple comparison tests are tools to control Type I Errors in statistical studies with large numbers of variables. This paper introduces single step (Sidak and Bonferroni), step-up (Holm), and step-down (Benjamini-Hochberg) multiple comparison procedures to control Family Wise Error Rate (FWER) and False Discovery Rate (FDR), which are overall measures of error rate of a statistical study. Proofs of these multiple comparison procedures are rewritten from past literature to incorporate intuitions and contexts.

Four multiple comparison tests are applied on a recipe data set. The results align with the theories. Single step multiple comparison tests, Sidak and Bonferroni yield the most conservative results with the least number of rejections; step-up procedure, Holm, is less conservative; step-down procedure, Benjamini-Hochberg, is the least conservative, giving the largest number of rejections.

To address the problem of unknown truth, simulations are designed using the recipe data set to force the proportion of true versus false null hypotheses. The FWERs and FDRs after applications of multiple comparison tests vary with the proportion of false null hypotheses. A larger proportion of false null hypotheses result in smaller FWERs and FDR when everything else holds equal. Yet the expected FWERs and FDR are not controlled at 0.05 with 100 iterations where half of the hypotheses are null. This result is reasonable. For the Sidak and Holm procedures controlling the FWER, the independence assumption is too strong to hold in real data. FWER also has a very conservative definition measuring how often we see at least one Type I Error, which makes the expected FWER high. The FDR is about 1%

above the designated level of 5% under the Benjamini-Hochberg procedure because FDR is an expected value of the FDP (False Discovery Proportion). The proof of the Benjamini-Hochberg procedure shows that the multiple comparison test controls the FDR at the desired level on average, not every time. Moreover, the Benjamini-Hochberg procedure requires independence or positive regression dependence [Benjamini and Yekutieli, 2001], which might not hold true in the recipe data the simulations are built upon.

Despite the fact that the multiple comparison procedures discussed in this paper do not always control the FWERs and FDRs at the desired level, the simulation results show that all procedures yield lower FWERs and FDRs than the unadjusted results. Even thought we do not know the true proportion of false null hypotheses in real-world data, we know that applying multiple comparison tests can at least reduce the overall error rate of the study and therefore improve the validity of the statistical reports.

Bibliography

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

[Benjamini and Yekutieli, 2001] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

[Emmert-Streib and Dehmer, 2019] Emmert-Streib, F. and Dehmer, M. (2019). Large-scale simultaneous inference with hypothesis testing: Multiple testing procedures in practice. *Machine Learning and Knowledge Extraction*, 1(2):653–683.

[Finner and Roters, 2001] Finner, H. and Roters, M. (2001). On the false discovery rate and expected type i errors. *Biometrical Journal*, 43(8):985–1005.

[Goeman and Solari, 2014] Goeman, J. J. and Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978.

[Holm, 1979] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

[Šidák, 1967] Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.