Multiple Hypothesis Tests

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Multiple Hypothesis Tests

1. Introduction
   - Probability of At Least One Type I Error
2. Some Key Concepts
   - Primary Hypotheses
   - Closure
   - Hierarchical Sets and Minimal Hypotheses
   - Families
   - Type I Error Control
   - Power
   - p-Value and Adjusted p-Value
   - Closed Test Procedures
3. Methods Based on Ordered p-Values
   - Methods Based on the First-Order Bonferonni Inequality
   - Methods Based on the Simes Equality
   - Methods Controlling the False Discovery Rate
4. An Example

Introduction

As we saw in Psychology 310, when we test any statistical hypothesis, we realize that our decision may be wrong.

We design a procedure to control the probability of false rejection at α.

Unfortunately, if our data analysis involves many hypothesis tests, the probability of at least one Type I error increases rather sharply with the number of tests.

Introduction

Probability of At Least One Type I Error

For example, if there are m tests and they are independent, and each one is performed with a Type I error rate of α, and all hypotheses are actually true, the probability of at least one Type I error is

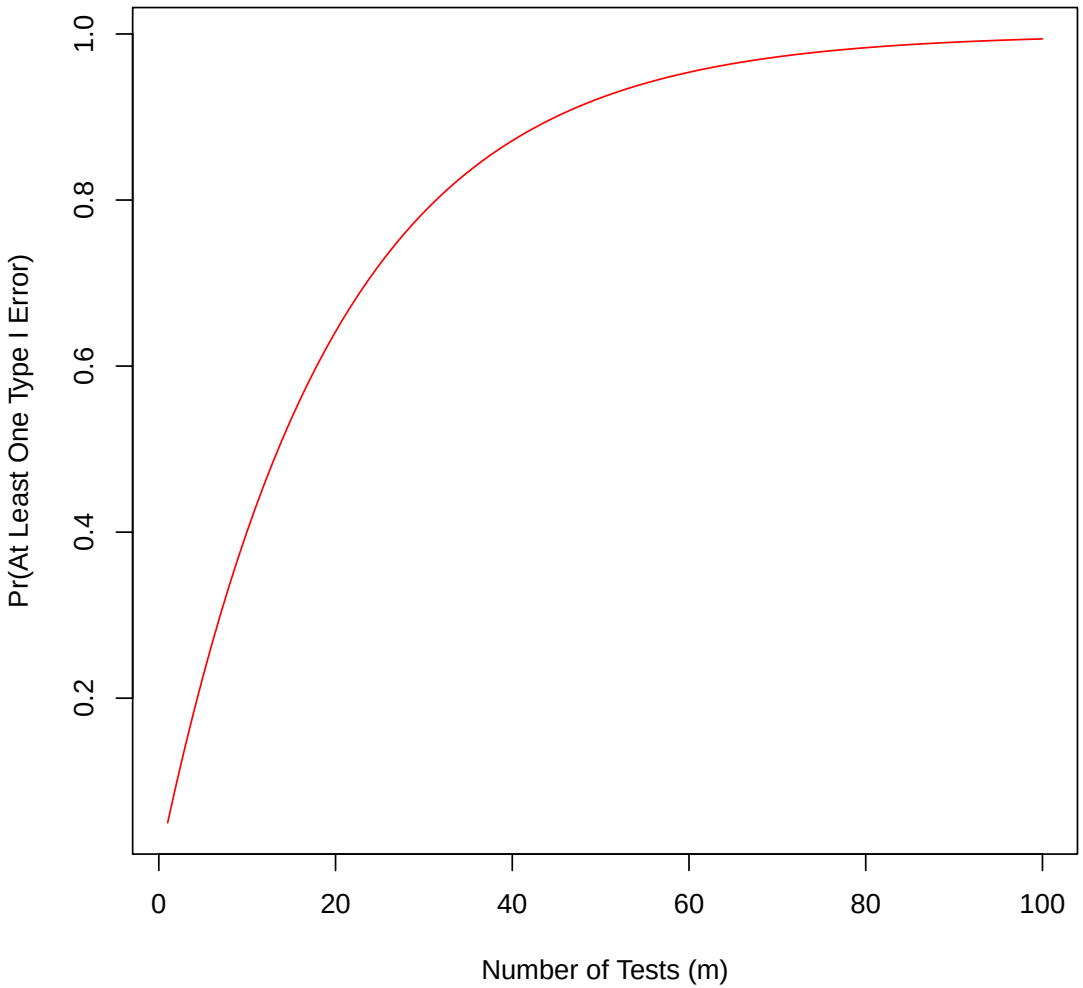
$$\begin{aligned} \Pr(\text{At Least One Error}) &= 1 - \Pr(\text{No Errors}) \\ &= 1 - \Pr(\text{All Decisions Correct}) \\ &= 1 - (1-\alpha)^m \end{aligned} \tag{1}$$

Below is code to create a plot of the probability of at least one error.

```
> curve(1 - (1 - 0.05)^x, 1, 100, xlab = "Number of Tests (m)", ylab = "Pr(
+     col = "red")
```

Introduction

Probability of At Least One Type I Error



Some Key Concepts

In this section, we discuss some key "Organizing Concepts" useful for discussing the problem of multiple hypothesis testing

We begin by assuming that some set of null hypotheses is of primary interest, and that we have a set of observations with a joint distribution depending on parameters relevant to the hypothesis set, and that the hypotheses limit the values of the parameters in some way.

For example, suppose we have the means of J populations, i.e., $\mu_1, \mu_2, \ldots, \mu_J$.

Let δ_{ij} stand for the difference between μ_i and μ_j. Let δ_{ijk} stand for the set of differences among δ_i, δ_j, and δ_k.

Suppose the hypotheses are written $H_{ijk\ldots} : \delta_{ijk\ldots} = 0$, indicating that all subscripted means are equal.

For example, H_{1234} is shorthand for the hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

Some Key Concepts

Primary Hypotheses

The *primary hypotheses* in a testing situation are the elements of the universal set of all hypotheses of interest.

Some Key Concepts

Closure

The *closure* of a set of hypotheses is the collection of the original set plus all distinct hypotheses formed by intersections of the hypotheses in the original set.

For example, if the original set is $A = \{H_{12}, H_{13}\}$, the *closure* of A is H_{123}, since if $\mu_1 = \mu_3$ and $\mu_2 = \mu_3$, then $\mu_1 = \mu_2 = \mu_3$.

The hypotheses included in an intersection are called the *components* of the intersection hypothesis.

Note that technically, an intersection is a component of itself. So we introduce the notion of *proper component*, representing any other component of an intersection.

In the preceding example, the proper components (so long as they are included in the primary hypotheses) of H_{123} are H_{12},H_{13},H_{23}.

Note that the truth of the closure of a set of hypotheses implies the truth of all its proper components.

Some Key Concepts

Hierarchical Sets and Minimal Hypotheses

Any set of hypotheses in which some are proper components of others will be called a *hierarchical set*.

A closed set is therefore hierarchical. The top of the hierarchy is the intersection of all the hypotheses.

The bottom of the hierarchy consists of the sets that have no proper components. These are called the *minimal hypotheses*.

A mimimal hypothesis is also one that does not imply the truth of any other hypotheses in the set.

Some Key Concepts

Families

A key decision in analyzing data is to decide on the set of hypotheses to consider as a *family*.

A family is a set for which significance statements and related error rates will be controlled jointly.

Note: In the early multiple comparisons literature (e.g., Ryan, 1959, 1960), the term "experiment" was used instead of "family."

As research grew more complex, the use of the term "experiment" was found to be limiting. Consider, for example, factorial experiment or a large survey.

Because of the inverse relationship between control of Type I errors and power, it would be unreasonable to expect the probability of a Type I error to be controlled over the entire experiment at conventional levels like 0.05.

Some Key Concepts

Families

Even within the same data set, different families may be analyzed for different reasons. For example, suppose you have data for 50 schools. You may be interested in all the pairwise comparisons among the schools. On the other hand, the principal of school A may only be interested in the family of pairwise comparison of her school with the other 49.

Some Key Concepts

Type I Error Control

Strong error rate control methods control the Type I error rates (of various kinds) for any combination of true and false null hypotheses in a family.

Weak error rate control methods control the various Type I error rates only when all the null hypotheses in a family are simultaneously true.

We will concentrate on methods with strong control.

Some Key Concepts

Type I Error Control

The *error rate per hypothesis*, often called the error rate per comparison or PCER, is the Type I error rate for each individual hypothesis test.

Some Key Concepts

Type I Error Control

The *error rate per family*, or PFER, is the expected number of false rejections in the family.

Some Key Concepts

Type I Error Control

The *familywise error rate (FWER)* is the probability of at least one Type I error in the family of tests.

Some Key Concepts

Type I Error Control

Let V_m stand for the number of Type I errors committed in a family of tests, and R_m be the number of rejected hypotheses. The *generalized familywise error rate* $gFWER(k) = Pr(V_m > k)$, or chance of at least $(k + 1)$ false positives. The special case $k = 0$ corresponds to the usual family-wise error rate, FWER.

Some Key Concepts

Type I Error Control

The *False Discovery Rate (FDR)* is (V_m/R_m), the long run proportion of rejections that are Type I errors.

Some Key Concepts

Power

Just as there are multiple ways of looking at Type I error rates, there are several conceptualizations of the notion of power in multiple hypothesis testing. In the context of pairwise mean comparisons, these have been referred to as

1. *Any-pair power*. The probability of rejecting at least one false null hypothesis.
2. *Per-pair power*. The average probability of rejecting a false null hypothesis.
3. *All-pairs power*. The probability of rejecting all false null hypotheses in the set.

Some Key Concepts

p-Value and Adjusted p-Value

Many scientists now report p-values rather than simply giving a test statistic and the result of the hypothesis test.

Extension of the idea of a p-value to multiple testing is not straightforward.

Some authors have championed the use of the *adjusted p-value*, which is the value of the error rate for the entire procedure that, if it had been employed on the entire set of test statistics under consideration, would have resulted in the null hypothesis for a particular hypothesis test barely rejecting.

Some Key Concepts

Closed Test Procedures

The most powerful procedures designed to control FWER are in the class of *closed test procedures*.

Assume a set of hypotheses of primary interest, add hypotheses as necessary to form the closure of this set, and recall that the closed set consists of a hierarchy of hypotheses.

The *closure principle* is as follows: A hypothesis is rejected at level α if and only if it and every hypothesis directly above it in the hierarchy (i.e. every hypothesis that includes it in an intersection and thus implies it) is rejected at level α.

Some Key Concepts

Closed Test Procedures

Consider a hypothesis set involving 4 means, with the highest hypothesis in the hierarchy H_{1234} and the six hypotheses $H_{ij}, i \neq j = 1, 2, 3, 4$ as the minimal hypotheses.

No hypotheses below H_{1234} can be rejected unless H_{1234} is rejected. Suppose H_{1234} is rejected. Then H_{12}, for example, cannot be rejected unless H_{124}, H_{123} are rejected.

But since the intersection hypothesis $H_{12} \cap H_{34}$ is also implied by H_{1234} yet is formally distinct from it, this intersection hypothesis ranks below H_{1234} but above H_{12}.

So the intersection hypothesis $H_{12} \cap H_{34}$ must be tested and rejected at the α level before H_{12} is tested by itself at the α level. It is only when the final test is rejected that one declares μ_1 and μ_2 to be significantly different.

Some Key Concepts

Closed Test Procedures

The proof that Closed Test Procedures control FWER is straightforward, and is given in a *Biometrika* article by Marcus et al.(1976). Let's consider the proof in connection with the following situation. There are 4 means, and $\mu_1 = \mu_2 = \mu_3 \neq \mu_4$. In this case, $H_P = H_{123}$ is the closure of all true hypotheses. the intersection of all H_{ij} that are true.

- Consider every possible true situation, each of which can be represented as the intersection of null hypotheses and their alternatives. *Only one of these can be the true one.* In our current example, this is

$$H_Q = H_{12} \cap H_{13} \cap H_{23} \cap \overline{H}_{14} \cap \overline{H}_{24} \cap \overline{H}_{34}$$

- Now, consider H_T, the closure of the H_{ij} that are true.

$$H_T = H_{12} \cap H_{13} \cap H_{23}$$

.

- The probability under a closed testing procedure of rejecting H_T is $\leq \alpha$. Why? (continued on next slide)

Some Key Concepts

Closed Test Procedures

- All true null hypotheses in the primary set are contained in H_Q, and none of them can be rejected unless that configuration is rejected. Let A be the event that all hypotheses in H_Q ranking above H_T and including elements of H_T are rejected. Clearly $\Pr(A) \leq 1$. Let B be the event that H_T is rejected. Since B can only occur when A has already occurred, $B = A \cap B$, and so $\Pr(B) = \Pr(A \cap B) = \Pr(A)\Pr(B|A)$. But $\Pr(B|A) = \alpha$, since once one arrives at the point of testing H_T, that test is performed at the α level.
- Consequently $\Pr(B) \leq \alpha$. And since rejection of any primary hypothesis requires event B, the probability of one or more such rejections must be less than or equal to $\Pr(B)$, and so must also be less than or equal to α.

In other words, when working through the hierarchy, when one encounters the first hypothesis at the top of the hierarchy of true hypotheses, the probability of rejecting it is $\leq \alpha$. If it is rejected, then a Type I error has occurred. If not, no more tests below that point in the hierarchy can be done. So the Type I error rate at the head of the hierarchy is also the FWER.

Methods Based on Ordered p-Values

A finite set of minimal hypotheses $H_i, i = 1, \ldots, m$ is to be tested. Corresponding to the H_i are test statistics T_i (or their absolute values) such that p_i corresponding to each hypothesis may be computed.

Assume that the p_i are ordered such that $p_1 \leq p_2 \leq \ldots \leq p_m$. With the exception of the methods in the the subsection on False Discovery Rate, these methods provide strong FWER control.

Methods Based on Ordered p-Values

Methods Based on the First-Order Bonferonni Inequality

The Simple Bonferonni Method.

The first-order Bonferroni inequality states that, for events $A_i, i = 1, \ldots, n$,

$$\Pr\left(\bigcup_{i=1}^{m} A_i\right) \leq \sum_{i=1}^{m} \Pr(A_i) \tag{2}$$

This inequality is the basis for several general methods.

Methods Based on Ordered p-Values

Methods Based on the First-Order Bonferonni Inequality

The simple method is, reject H_i if $p_i \leq \alpha_i$, where $\sum_{i=1}^{m} \alpha_i = \alpha$.

This method controls FWER at or below α.

Usually, all α_i are set equal to α/m, a procedure sometimes called the *unweighted simple Bonferroni method*.

Of course, with this method, power suffers increasingly as m becomes large.

Methods Based on Ordered p-Values

Methods Based on the First-Order Bonferonni Inequality

Holm's Sequentially Rejective Bonferonni Method

The unweighted method is as follows. At the first stage, H_1 is rejected if $p_1 \leq \alpha/m$.

If H_1 is not rejected, all subsequent hypotheses are accepted without further testing.

If H_1 is rejected, H_2 is tested at the $\alpha/(m-1)$ level. If H_2 is not rejected, all subsequent hypotheses are accepted without further testing.

The procedure continues, with the ith test performed at the $\alpha/(m-i+1)$ level, until the first non-rejection occurs.

Methods Based on Ordered p-Values

Methods Based on the First-Order Bonferonni Inequality

Proof.

First imagine all m hypotheses are true, and remember that an acceptance can never be followed by a rejection with this procedure. So, if there is going to be any incorrect rejection, it has to occur on the first test of a true hypothesis, because otherwise there will be no further tests.

So if all hypotheses are true, the probability of at least one rejection is the probability of getting a rejection on the first test, which is α/m. What happens after that is irrelevant to the FWER, because all patterns of subsequent results will fit the definition of a Familywise Error having occurred. (continued on next slide)

Methods Based on Ordered p-Values

Methods Based on the First-Order Bonferonni Inequality

Now imagine that there are $k \leq m$ true null hypotheses in the collection of m hypotheses to be tested. Suppose $k = m - 1$. Then the first true hypothesis will be tested in position 1 or 2, and so the probability of a Familywise Error can be no more than $\alpha/(m-1)$.

If $k = m - 2$, the first true hypothesis will be tested in position 1, 2, or 3 and so the probability of a Familywise Error can be no more than $\alpha/(m-2)$, etc.

If there is only one true null hypothesis, and it is tested last, the probability of a rejection is α.

This completes the proof.

Methods Based on Ordered p-Values

Methods Based on the First-Order Bonferonni Inequality

An Enhancement for Independent (and some Dependent) Tests.

If tests are independent, α/m may be replaced by $1-(1-\alpha)^{1/m}$, which is always greater than α/m.

For certain other classes of tests that are *positive orthant dependent*, this enhancement may also be applied. This includes the set of pairwise two-sided t-tests in a 1-way ANOVA layout.

Methods Based on Ordered p-Values

Methods Based on the Simes Equality

If X is a continuous test statistic based on an assumed null distribution, having a continuous cumulative distribution F, then F has a Uniform(0,1) distribution if the null hypothesis is true. If you order m observed test statistics $x_{1:m}$, you can order their corresponding cumulative probabilities $u_{i:m}$. Simes proved that if the Xs are independent then for a value α between 0 and 1,

$$\Pr\left(u_{i:m} \geq i\alpha/m, i = 1, \ldots, m\right) = 1 - \alpha \tag{3}$$

Methods Based on Ordered p-Values

Methods Based on the Simes Equality

Hochberg's Sequential Step-Up Procedure.

Order the m tests in terms of their p values, with p_1 the smallest and p_m the largest.

Choose a FWER α. If $p_m \leq \alpha$, reject *all* hypotheses.

If $p_m > \alpha$, compare p_{m-1} to $\alpha/2$, and if $p_{m-1} \leq \alpha/2$ reject all $m-1$ remaining hypotheses.

If $p_{n-1} > \alpha/2$, compare p_{n-2} to $\alpha/3$, etc.

An alternative way of viewing this process is that one rejects the subset of the (ordered) hypotheses $H_1, H_2, \ldots, H_k$, where

$$k = \max\left\{ i : p_i \leq \frac{\alpha}{m - i + 1} \right\} \tag{4}$$

Methods Based on Ordered p-Values

Methods Controlling the False Discovery Rate

As articulated by Benjamini and Hochberg (1995) in the quote below, if there are a lot of rejections expected in a set of m tests, control of FWER may not be feasible, because of its damaging effect on power.

> *(b) Classical procedures that control the FWER in the strong sense, at levels conventional in single-comparison problems, tend to have substantially less power than the per comparison procedure of the same levels.*
>
> *(c) Often the control of the FWER is not quite needed. The control of the FWER is important when a conclusion from the various individual inferences is likely to be erroneous when at least one of them is. This may be the case, for example, when several new treatments are competing against a standard, and a single treatment is chosen from the set of treatments which are declared significantly better than the standard. However, a treatment group and a control group are often compared by testing various aspects of the effect (different end points in clinical trials terminology). The overall conclusion that the treatment is superior need not be erroneous even if some of the null hypotheses are falsely rejected.*

Methods Based on Ordered p-Values

Methods Controlling the False Discovery Rate

The Benjamini-Hochberg method is as follows. Suppose there are m null hypotheses, and, unknown to the experimenter, m_0 are true. The following method controls FDR at or below $\alpha m_0/m$ (which of course is less than or equal to α).

Consider again the ordered p values $p_1 \leq p_2 \leq \ldots \leq p_m$. Reject the set of hypotheses $H_1, H_2, \ldots H_k$ for which

$$k = \max \left\{ i : p_i \leq \frac{i}{m} \alpha \right\} \tag{5}$$

An Example

As discussed in Benjamini and Yekutieli (2001), Needleman et al (*New England Journal of Medicine 300* 689–695) studied the neuropsychologic effects of unidentified childhood exposure to lead by comparing various psychological and classroom performances between two groups of children differing in the lead level observed in their shed teeth. While there is no doubt that high levels of lead are harmful, Needleman's findings regarding exposure to low lead levels, especially because of their contribution to the Environmental Protection Agencys review of lead exposure standards, are controversial. The study was attacked on the ground of methodological flaws, because Needleman et al. analyzed three separate families of "endpoints" in their study (and the p-values observed):

1. Teacher's Behavioral Ratings (0.003,0.05,0.05,0.14, 0.08,0.01,0.04,0.01,.050,0.003,0.003)
2. WISC scores (0.04,0.05,0.02,0.49,0.08,0.36,0.03, 0.38,0.15,0.90,0.37,0.54)
3. Verbal Processing and Reaction Time scores. (0.002,0.03,0.07,0.37,0.90,0.42,0.05,0.04, 0.32,0.001,0.001,0.01)

An Example

We abbreviate the 3 families as TBR, WISC, and RT R can sort the values:

```
> TBR <- sort(c(0.003, 0.05, 0.05, 0.14, 0.08, 0.01, 0.04, 0.01, 0.05, 0.00
+     0.003))
> WISC <- sort(c(0.04, 0.05, 0.02, 0.49, 0.08, 0.36, 0.03, 0.38, 0.15, 0.9,
+     0.54))
> RT <- sort(c(0.002, 0.03, 0.07, 0.37, 0.9, 0.42, 0.05, 0.04, 0.32, 0.001,
+     0.01))
> TBR

 [1] 0.003 0.003 0.003 0.010 0.010 0.040 0.050 0.050 0.050 0.080 0.140

> WISC

 [1] 0.02 0.03 0.04 0.05 0.08 0.15 0.36 0.37 0.38 0.49 0.54 0.90

> RT

 [1] 0.001 0.001 0.002 0.010 0.030 0.040 0.050 0.070 0.320 0.370 0.420
[12] 0.900
```

An Example

Suppose we process the 3 families separately, and set FWER to $\alpha = 0.05$ for each of the 3 families. If we use the simple Bonferroni procedure, how many rejections do we get?

```
> Bonf.reject <- function(pvalues, alpha) {
+     return(sort(pvalues) <= alpha/length(pvalues))
+ }
> Bonf.reject(TBR, 0.05)

 [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

> Bonf.reject(WISC, 0.05)

 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[12] FALSE

> Bonf.reject(RT, 0.05)

 [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[12] FALSE
```

An Example

The Holm step-down and the Hochberg step-up procedures that control FWER help a bit in this case.

```
> Holm.reject <- function(pvalues, alpha) {
+     pvalues <- sort(pvalues)
+     m <- length(pvalues)
+     results <- rep(FALSE, m)
+     crits <- alpha/(m:1)
+     for (i in 1:m) if (pvalues[i] <= crits[i])
+         results[i] <- TRUE else break
+     return(results)
+ }
> Hochberg.reject <- function(pvalues, alpha) {
+     pvalues <- sort(pvalues)
+     m <- length(pvalues)
+     results <- rep(TRUE, m)
+     for (i in m:1) if (pvalues[i] <= alpha/(m - i + 1))
+         break else results[i] <- FALSE
+     return(results)
+ }
```

An Example

Holm's method the same pattern of rejections as Hochberg's, in this case.

```
> Holm.reject(TBR, 0.05)

 [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

> Holm.reject(WISC, 0.05)

 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[12] FALSE

> Holm.reject(RT, 0.05)

 [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[12] FALSE

> Hochberg.reject(TBR, 0.05)

 [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

> Hochberg.reject(WISC, 0.05)

 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[12] FALSE

> Hochberg.reject(RT, 0.05)

 [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[12] FALSE
```

An Example

A point of contention with respect to the article was the choice by the authors to analyze 3 separate families.

Some authors argued that all tests should have been combined and analyzed as one family.

With that approach, only two hypotheses would have been rejected.

```
> ALL.FAMILIES <- c(TBR, WISC, RT)
> Hochberg.reject(ALL.FAMILIES, 0.05)

 [1]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[34] FALSE FALSE
```

An Example

On the other hand, since the family is now quite large, encompassing 35 tests, it may make more sense at this point to control the FDR rather than the FWER. We write a function to implement the method.

```
> FDR.reject <- function(pvalues, alpha) {
+     pvalues <- sort(pvalues)
+     m <- length(pvalues)
+     results <- rep(TRUE, m)
+     for (i in m:1) if (pvalues[i] <= (i * alpha/m))
+         break else results[i] <- FALSE
+     return(results)
+ }
```

Let's try it out!

An Example

```
> FDR.reject(ALL.FAMILIES, 0.05)

 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
[12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[34] FALSE FALSE
```

Now 9 hypotheses are rejected when the combined families are processed as a unit. If we go back and reanalyze the individual families while controlling the false discovery rate, we find that there are 5 significant differences in the TBR family and 4 in the RT family.