Multiple Testing

Tim Hanson

Department of Statistics
University of South Carolina

January, 2017

Modified from originals by Gary W. Oehlert

Type I error

A Type I error is to wrongly reject the null hypothesis of what you are testing. When we say "reject H_0 at 5% level" we mean we reject the null hypothesis H_0 but are willing to make a mistake – reject when the null is true – 5% of the time, i.e. one in twenty tests.

When we have multiple hypotheses, if we reject each at 5% there is a greater chance than 5% of wrongly rejecting one or more nulls.

Background

Suppose that you had a 20-sided die. Nineteen of the sides are labeled 0 and one of the sides is labeled 1.

You roll the die once. What is the chance of getting a 1? Easy, 5%.

Now roll the die 20 times. What is the chance of getting at least one 1?
$1 - .95^{20} = .642$

Roll it 100 times, and the probability of at least one 1 is now $1 - .95^{100} = .994$

Background

Doing a 5% level test when the null is true is like rolling the die. You have a 5% chance of rejecting that true null, just like one roll of the die.

Now do 20 tests at the 5% level, with the null true every time. The chance of one or more nulls being rejected is .642. With 100 tests of true nulls, the chance of making at least one false rejection is virtual certainty.

That is the essence of the multiple testing problem: how do you control error rates when you do lots of tests?

Data snooping

Things are even worse if you don't just do lots of tests but instead snoop in the data to find something that looks interesting, and then test that interesting looking thing.

Example: when we looked at the fruit fly mating data, side-by-side boxplots indicated than only the last group (8 virgin females) looked to have a mean different from the rest. Looking at $\frac{1}{4}(\mu_1 + \mu_2 + \mu_3 + \mu_4) - \mu_5$ is snooping!

In this case, your chance of rejecting the null in that single test is very high, even if null is true and what you detected is just random variation.

It takes a ~~heavy, blunt instrument~~ powerful procedure to keep error rates under control in that situation.

Notation

We have several null hypotheses $H_{01}, H_{02}, \ldots, H_{0k}$.

H_0 is the overall or combined null hypothesis that all of the other nulls are true

$$H_0 = H_{01} \cap H_{02} \cap \cdots \cap H_{0k}$$

H_0 is true only if all k of $H_{01}, \ldots, H_{0k}$ are true. If any of them are false then H_0 is false.

$\mathcal{E}_i$ is the Type I error rate for the ith test; $\mathcal{E}$ is the Type I error rate for the combined null.

Hypothesis test vs. CI

There is a useful and widely-used relationship between hypothesis tests and CI's:

Let θ be a parameter to be estimated, e.g. $\theta = \mu_1$, $\theta = \mu_3 - \mu_7$, $\theta = w_1\mu_1 + \cdots + w_g\mu_g$, or in quadratic regression $y_{ij} = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \epsilon_{ij}$ maybe $\theta = -0.5\beta_1/\beta_2$, the maximum or minimum of the quadratic.

A 95% CI for a parameter θ includes all values where we would accept $H_0 : \theta = \theta_0$ (vs. $H_1 : \theta \neq \theta_0$) at the 5% level.

A 99% CI for a parameter θ includes all values where we would accept $H_0 : \theta = \theta_0$ at the 1% level.

To test $H_0 : \theta = \theta_0$ we accept if θ_0 is in the CI and reject otherwise.

Errors

This is errors as in mistakes.

Declaring a true null to be false is a Type I error. This is a false positive, declaring something to be happening when it is not.

Accepting a false null is a Type II error. This is a false negative, saying something is not happening when, in fact, something is happening.

Truth table for one hypothesis

	Reality/State of nature	
Decision	Null correct	Null false
Fail to reject	True negative	False negative
Reject	False positive	True positive

	Reality/State of nature	
Decision	Null correct	Null false
Fail to reject	☺	Type II error
Reject	Type I error	☺

The general approach in classical statistics is to control the probability of a Type I error ($\mathcal{E}$), and among procedures that control that error choose one that makes the Type II error rate low.

Decision counts for k hypotheses

That's pretty well defined for a single hypothesis, but working with multiple hypotheses requires a bit more. Consider this table.

Numbers of decisions

	Reality/State of nature	
Decision	Null correct	Null false
Fail to reject	A	B
Reject	C	D

For k hypotheses, we have $A + B + C + D = k$.

In practice, we will never know these counts, but we can work with them theoretically.

Error rates: per comparison

The **per comparison** error rate ignores the multiple testing issue.

Here you just do a separate test for each null hypothesis ignoring all of the other tests. Per comparison error control is

$$P[\text{reject } H_{0i} | H_{0i} \text{ true}] \leq \mathcal{E}$$

In effect, we have k different tables with A_i, B_i, C_i, and D_i. Because we assume that all nulls are true, $B_i = D_i = 0$ for all tables (sub-hypotheses). Or,

$$P[C_i = 1 | H_{0i} \text{ true}] \leq \mathcal{E}$$

Error rates: per experiment

The **per experiment** error rate or **experimentwise** error rate controls the probability that any H_{0i} is rejected (thus rejecting H_0) when all H_{0i} (and H_0) are true. Per experiment error control is

$$P[\text{reject any } H_{0i} | H_0 \text{ true}] \leq \mathcal{E}$$

Again, because we have all nulls true, $B = D = 0$ and per experiment control can be written as

$$P[C > 0 | H_0 \text{ true}] \leq \mathcal{E}$$

Protected by overall F-test of $H_0 : \mu_1 = \cdots = \mu_g$, or joint F-test for several contrasts, i.e. testing a contrast matrix.

Error rates: false discovery rate (FDR)

The **False Discovery Rate** allows for the possibility that some of the H_{0i} are false.

Let F = C/(C+D) (or zero when C+D=0). This is the false discovery fraction—the fraction of rejections that are incorrect.

Controlling the FDR is making sure

$$E\left[\frac{C}{C+D}\right] \leq \mathcal{E}$$

so the expected fraction of false rejections is at most $\mathcal{E}$. Note that the more correct rejections you make, the more false rejections FDR lets you make.

Error rates: strong familywise (SFER)

The **strong familywise error rate** also allows for the possibility that some of the H_{0i} are false, but unlike the FDR it cuts you no slack for making correct rejections. SFER control is

$$P[\text{reject any } H_{0i} | H_{0i} \text{ true}] \leq \mathcal{E}$$

Controlling the SFER is

$$P[C > 0] \leq \mathcal{E}$$

Compare this carefully with the experimentwise error rate.

Simultaneous confidence intervals

If we are forming multiple confidence intervals instead of just testing, then **simultaneous confidence intervals** satisfy

$$P[\text{One or more of the CIs fails to cover its parameter}] \leq \mathcal{E}$$

or

$$P[\text{All CIs simultaneously cover their parameters}] \geq 1 - \mathcal{E}$$

The coverage rate of individual intervals within a simultaneous confidence interval procedure will typically be larger than 1-$\mathcal{E}$.

(In effect, SFER only requires simultaneous confidence intervals for null values, so this requires more than SFER.)

Type I error control

Error rates were presented from weakest (per comparison) to strongest (simultaneous CIs). If a procedure controls one rate, it will also control the weaker rates.

If a procedure controls an error rate at $\mathcal{E}$, it controls the weaker error rates at (something usually less than) $\mathcal{E}$.

The stronger the Type I error rate control, the harder it is to see differences that are really there.

Type I error control

As you make Type I control stronger, you make more and more Type II errors.

Review:

- Per comparison hardly cares how many incorrect rejections in total.
- Per experiment doesn't want you to make an incorrect rejection, but if you make one correct rejection, then it doesn't care how many incorrect ones you make.
- FDR gives you some slack; for example, for every 19 correct rejection it gives you a pass on one incorrect rejection.
- SFER doesn't care how many correct rejections you make, it still doesn't want you to make an incorrect rejection.
- Simultaneous confidence intervals not only requires you to get the nulls right and the non-nulls right, and also need to say where all the parameter values are.

Gene expression data

Suppose that we have done a genomic assay on 30 women, 15 with breast cancer and 15 without. We have gene expression data on 5,000 genes.

To be concrete, let's introduce another index $k = 1, \ldots, 5000$ for gene. Then a typical model is

$$Y_{ijk} = \mu_{ik} + \epsilon_{ijk},$$

where $i = 1, 2$ denotes case or control, $j = 1, \ldots, 15$ women in each group (case or control), and $k = 1, \ldots, 5000$ is the gene being tested. There are 5000 potential hypotheses to think about $H_{0(k)} : \mu_{1k} - \mu_{2k} = 0$. And so 5000 "usual" per experiment p-values to compute, e.g. using t.test in R.

To use Benjamini-Hochberg, which controls FDR (later) the tests must be independent; is this reasonable?

Gene expression data

If we just had three genes in mind and didn't care about the others, we might use a per comparison error rate.

If we were primarily interested in whether there is some genetic influence, but want to cast a wide net for potential genetic markers if there is a genetic component, then we might use an experimentwise method.

Gene expression data

If we don't want to be bombarded with a lot of genes incorrectly identified as active but can work with a limited percentage of false positives, then FDR would do the trick.

If we want to have a controlled probability of making any false statement that a gene is involved in breast cancer, then we control the SFER.

If we want to be able to estimate expression on all of the genes with simultaneous coverage, then we need a simultaneous confidence interval method.

Approaches for controlling Type I error

Find the weakest Type I error rate control that is compatible with the kind of inference you wish to make. Then choose a procedure that controls that error rate.

We'll examine several approaches in common use; many are pairwise approaches.

Scheffé

Let's begin with the heaviest, bluntest instrument of them all: the Scheffé adjustment for contrasts.

The Scheffé procedure will control the strong familywise error rate for arbitrarily many contrasts, including contrasts suggested by the data.

The price you pay for this amazing Type I control is lots of Type II errors; differences have to be pretty big before Scheffé will reject the null.

The underlying idea of this procedure is to treat the SS from any contrast as if it had $g - 1$ degrees of freedom (instead of 1).

Scheffé

To test $H_0 : \sum_i w_i \mu_i = 0$, use

$$F = \frac{(\sum_i w_i \bar{y}_{i\bullet})^2}{(g-1) MS_E \sum_i w_i^2 / n_i}$$

and compute the p-value from a F distribution with $g-1$ and $N-g$ df. (This "F" is the square of the t-test for the contrast divided by $g-1$.) Note that the denominator uses $g-1$ instead of 1, thus making the F-statistic smaller $\Rightarrow$ p-value bigger $\Rightarrow$ harder to reject H_0.

For a confidence interval use

$$\sum_i w_i \bar{y}_{i\bullet} \pm \sqrt{(g-1) F_{\mathcal{E},g-1,N-g} MS_E \sum_i w_i^2 / n_i}$$

For example, if $g = 5$, $N - g = 20$, and $\mathcal{E} = .05$, then the usual t-based multiplier for the interval would be 2.08, but the Scheffé-based multiplier is 3.386 (equivalent to a t with $\mathcal{E}$=.0029).

Scheffé example in R

Recall w/ fruity fly data that we "snooped" and saw that group 5 looked different from the other 4.

```
library(cfcdae) # contains linear.contrast function, otherwise...
source("http://people.stat.sc.edu/hansont/stat506/cfcdae.R")
d=read.table("http://users.stat.umn.edu/~gary/book/fcdae.data/pr3.2",header=T)
attach(d)
ftrt=factor(trt)
f=lm(days~ftrt) # parameterization doesn't matter, here alpha1=0
linear.contrast(f,ftrt,c(.25,.25,.25,.25,-1)) # 8 virgins vs. the rest
linear.contrast(f,ftrt,c(.25,.25,.25,.25,-1),scheffe=T) # Scheffe
```

Intepretation? Note that p-value does not change; this function does not correctly modify the p-value, only the CI. However, the Scheffé CI does not include zero, so we reject $H_0 : \frac{1}{4}(\mu_1 + \mu_2 + \mu_3 + \mu_4) - \mu_5$ at the 5% level, correctly correcting for data snooping. Note we can examine as many "snooped" contrasts as we want using Scheffé. We simply reject those with CI's that do not include zero.

Bonferroni

Our second general procedure is Bonferroni. Bonferroni works for k pre-planned tests, so it does not work for data snooping.

The tests can be of any type, of mixed type, independent or dependent, they just have to be tests.

Bonferroni says divide your overall error $\mathcal{E}$ into k parts: $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k$ with $\sum_i \mathcal{E}_i = \mathcal{E}$ (usually $\mathcal{E}_i = \mathcal{E}/k$). Run test i of H_{0i} at the $\mathcal{E}_i$ error level. This will control the strong familywise error rate.

If you are doing confidence intervals, compute the ith interval with coverage $1 - \mathcal{E}_i$. Then you will have simultaneous confidence intervals with coverage $1 - \mathcal{E}$.

Bonferroni

Another way to think of this is do your tests and multiply the p-values by k. If any of them still look small, then reject.

The advantage of Bonferroni is that it is easy and widely applicable.

The disadvantage of Bonferroni is that in many special cases there are better procedures that control the same error rate.

Better in this case means fewer Type II errors or shorter confidence intervals, all while still controlling the error of interest.

Either Scheffé or Bonferroni can be optimal, it depends on the number of tests to be carried out k. I often fit both and use the one that produces smaller CI's...as long as I'm not data snooping.

Multiple contrasts in R

Let's look again at the three contrasts examined for the fruit fly data in Chapter 4.

```
cm=matrix(c(-1,.25,.25,.25,.25,0,.5,-.5,.5,-.5,0,-.5,-.5,.5,.5),5,3)
cm # matrix of contrasts (each column are contrast coefficients)
linear.contrast(f,ftrt,cm,bonferroni=T)
linear.contrast(f,ftrt,cm,scheffe=T)
```

Bonferroni finds two significant differences whereas Scheffé only finds one! Here, Bonferroni has more power. Both control the SFER, so we prefer Bonferroni. We are able to use Bonferroni because we thought about and constructed the contrasts ahead of time.

Again, note that only the CI's are correctly adjusted, not the p-values. Be careful.

Pairwise comparisons vs. general contrasts

One special contrast has $w_i = 1$ for group i and a $w_j = -1$ for group j, the rest being zero. Then $w_1\mu_1 + w_2\mu_2 + \cdots + w_g\mu_g = \mu_i - \mu_j$, a **pairwise comparison**.

Bonferroni and Scheffé work for tests of any contrasts, including pairwise comparisons. However, they are blunt instruments.

There are several more refined procedures that look at (a) all pairwise comparisons, or (b) all pairwise comparisons with a control, while fixing certain Type I errors. First let's look at the Studentized range distribution.

Studentized range

Suppose $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$ (the single mean model) is true. Look at the distribution of

$$\max_{i,j} \frac{\bar{y}_{i\bullet} - \bar{y}_{j\bullet}}{\sqrt{MS_E/n}}$$

This distribution is called the Studentized range. Its upper $\mathcal{E}$ percent point is denoted $q_{\mathcal{E}}(g, \nu)$ where there are g groups and ν is the df for the MS_E.

The Studentized range works by noting that $\bar{y}_{(g)\bullet} - \bar{y}_{(1)\bullet} \geq |\bar{y}_{i\bullet} - \bar{y}_{j\bullet}|$ for all i and j.

Studentized range

It's not obvious, but $q_{\mathcal{E}}(2, \nu) = \sqrt{2} t_{\mathcal{E}/2, \nu}$. That is, with two groups you can link the Studentized range to t.

It is possible to replace the F test comparing the separate means model to the single mean model with a test based on the Studentized range. They usually, but not always, agree.

Pairwise comparisons

Pairwise comparisons are simple comparisons of the mean of one treatment group to the mean of another treatment group, estmiated by

$$\hat{\mu}_i - \hat{\mu}_j = \bar{y}_{i\bullet} - \bar{y}_{j\bullet}$$

Lets examine procedures according to the error rate that they control.

First order the $\hat{\mu}_1, \ldots, \hat{\mu}_g$...

Introduce new labels on the sample means so that $\bar{y}_{(1)\bullet}$ is the smallest and $\bar{y}_{(g)\bullet}$ is the largest.

From $\bar{y}_{(1)\bullet}$ to $\bar{y}_{(g)\bullet}$ is a stretch of g means.
From $\bar{y}_{(2)\bullet}$ to $\bar{y}_{(g)\bullet}$ is a stretch of $g-1$ means.
From $\bar{y}_{(2)\bullet}$ to $\bar{y}_{(4)\bullet}$ is a stretch of 3 means.

Step-down methods

Step-down methods look at pairwise comparisons starting with the most extreme pair and working in. When you get to a pair whose equality of means cannot be rejected, then you do not reject equality for every pair of means included in the stretch.

Step-down methods can only declare a stretch of means significantly different (i.e., the ends are different) if the stretch exceeds its critical minimum and every stretch containing the stretch also exceeds its critical minimum.

So failure to reject the null that the treatments corresponding to $\bar{y}_{(2)\bullet}$ and $\bar{y}_{(4)\bullet}$ have equal means implies that we must fail to reject the comparisons between (2) and (3) as well as (3) and (4).

Stopping rule

The step-down stopping rule is only needed if the critical minimum difference for rejecting the null gets smaller as the stretches get shorter. If they all stay the same, then failure to reject the endpoints of a stretch of means implies that you will not reject any stretch within.

A couple of the forthcoming methods are real, genuine step-down methods (SNK and REGWR). A couple have constant sized critical minima (LSD and HSD). However, we will talk about them all as step-down because we can frame them together that way.

All pairwise methods work the same

Consider the difference

$$\bar{y}_{(j)\bullet} - \bar{y}_{(i)\bullet}$$

The critical value, often called the "significant difference," for a comparison is

$$|\bar{y}_{(j)\bullet} - \bar{y}_{(i)\bullet}| > \frac{X}{\sqrt{2}} \sqrt{MS_E} \sqrt{\frac{1}{n_{(i)}} + \frac{1}{n_{(j)}}}$$

We say treatment means (i) and (j) differ if the observed difference in means exceeds this significant difference.

All we need to do is set the mysterious X.

Several pairwise methods

Method	X
LSD	$q_{\mathcal{E}}(2, N-g) = \sqrt{2}t_{\mathcal{E}/2,\nu}$
PLSD	$q_{\mathcal{E}}(2, N-g)$ but F test must reject
SNK	$q_{\mathcal{E}}(k, N-g)$
REGWR	$q_{\mathcal{E}_k}(k, N-g)$
HSD	$q_{\mathcal{E}}(g, N-g)$

The mysterious $\mathcal{E}_k$ in REGWR is $\mathcal{E}_k = \mathcal{E}$ for $k = g, g-1$ and $\mathcal{E}_k = k\mathcal{E}/g$ for $k < g-1$.

In general, $N-g$ is replaced by df in the MS_E.

LSD and PLSD are usually formulated using t distributions (i.e., use t and get rid of the $\sqrt{2}$).

Approaches and error control

LSD is *least significant difference*. It protects the per comparison error rate.

PLSD is *Protected LSD*. Do the ANOVA F test first. If it rejects, then proceed with LSD. If it fails to reject, then say no differences. The F-test protects experimentwise error rate.

SNK is Student-Neuman-Keuls. Protects FDR.

REGWR is Ryan-Einot-Gabriel-Welsch range test. Protects SFER.

HSD is the Honest significant difference (also called the Studentized range procedure or the Tukey W). It produces simultaneous confidence intervals (as difference plus or minus significant difference).

Visualization

Write treatment labels so means are in increasing order, then draw a line under treatments that are not significantly different.

C A B

The `pairwise` function gives

(a) The estimated pairwise differences $\hat{\mu}_i - \hat{\mu}_j = \bar{y}_{i\bullet} - \bar{y}_{j\bullet}$

(b) The significant difference(s) $\frac{X}{\sqrt{2}}\sqrt{MS_E}\sqrt{\frac{1}{n_{(i)}} + \frac{1}{n_{(j)}}}$

(c) Confidence intervals $\bar{y}_{i\bullet} - \bar{y}_{j\bullet} \pm \frac{X}{\sqrt{2}}\sqrt{MS_E}\sqrt{\frac{1}{n_{(i)}} + \frac{1}{n_{(j)}}}$

Using the `lines` function with `pairwise` gives the visualization above.

Cheese inoculants

Total free amino acids in cheese after 168 days of ripening when subjected to four different adjunct (nonstarter) bacterial treatments. Treatments are control, add strain A, add strain B, add strains A and B.

```
cheese=read.table("http://users.stat.umn.edu/~gary//book/fcdae.data/exmpl5.5",header=T)
names(cheese)
cheese$trt=factor(cheese$trt)
f=lm(y~trt,data=cheese)
anova(f) # are there treatment differences at 5%?
?pairwise # look at available options, note e.g. confidence=0.9 changes CI level
# if you got the cfcdae package to work do not need the "print" wrapper
print(pairwise(f,trt)) # default is HSD
print(pairwise(f,trt,type="regwr")) # REGWR
print(pairwise(f,trt,type="snk")) # SNK
lines(pairwise(f,trt)) # default is HSD
lines(pairwise(f,trt,type="regwr")) # REGWR
lines(pairwise(f,trt,type="snk")) # SNK
```

A & B significantly different from control and A alone w/ SFER capped at 5%.

Compare to control

Sometimes we have a control treatment, and all we really want to do is compare each treatment to control, but not the non-control treatments to each other.

Should you want to do this, there is a procedure called Dunnett's Significant Difference that will give you simultaneous confidence intervals or control SFER. Comparing treatment g to the other treatments, use

$$\bar{y}_{i\bullet} - \bar{y}_{g\bullet} \pm d_{\mathcal{E}}(g-1,\nu)\sqrt{MS_E}\sqrt{1/n_i + 1/n_j}$$

You get $d_{\mathcal{E}}(g-1,\nu)$ from the two-sided Dunnett's table.

Dunnett's Significant Difference

For one sided test, say with new yielding higher than control as the alternative, use

$$\bar{y}_{i\bullet} - \bar{y}_{g\bullet} > d'_{\mathcal{E}}(g-1,\nu)\sqrt{MS_E}\sqrt{1/n_i + 1/n_j}$$

If you only want to compare new to control, design with $n_g/n_i \approx \sqrt{g-1}$. This gives best overall results.

```
compare.to.control(f,trt,control=1)
```

Compare to best

Very useful: can use Dunnett to identify the group of treatments that distinguishes itself as best.

Best subset (assuming bigger is better) is all i such that for any $j \neq i$:

$$\bar{y}_{i\bullet} > \bar{y}_{j\bullet} - d'_{\mathcal{E}}(g-1,\nu)\sqrt{MS_E}\sqrt{1/n_i + 1/n_j}$$

Best subset is all treatments not significantly less than the highest mean using a one-sided Dunnett allowance.

The probability of truly best treatment being in this group is 1-$\mathcal{E}$.

Weed control

Percent weed control in soybeans under 14 different treatments. Columns are treatment number and percent control.

```
soybeans=read.table("http://users.stat.umn.edu/~gary/book/fcdae.data/exmpl5.10",header=TRUE)
names(soybeans)
f=lm(sqrt(100-y)~as.factor(trt),data=soybeans)
anova(f)
pairwise(f,as.factor(trt))
lines(pairwise(f,as.factor(trt)))
compare.to.best(f,as.factor(trt)) # one option: add conf=0.99
# not quite right, original response % weed control (high is good)
# transformed response is sqrt(100-y) so low is now good
compare.to.best(f,as.factor(trt),lowisbest=TRUE)
```

Treatments 1, 6, 7, 10, and 11 are the best and not significantly different according to our model. Any problems here? Look at side-by-side boxplots...

Working with p-values directly

`linear.contrast` produces properly adjusted CIs but not p-values. We reject if the adjusted CIs do not include zero (for multiple contrasts).

One approach works with the k p-values directly. There are three approaches described in your book: Bonferroni, Holm, and Benjamini-Hochberg. Bonferroni controls SCI, Holm the SFER, and Benjamini-Hochberg the FDR. See p. 82.

Say we have $k = 10$ hypotheses to test; fix the error at $\mathcal{E} = 0.05$. Our per comparison, i.e. usual p-values for the 10 tests are:

Hyp.	$H_{0(1)}$	$H_{0(2)}$	$H_{0(3)}$	$H_{0(4)}$	$H_{0(5)}$
p-value	0.0052	0.0041	0.0395	0.4533	0.0070
Hyp.	$H_{0(6)}$	$H_{0(7)}$	$H_{0(8)}$	$H_{0(9)}$	$H_{0(10)}$
p-value	0.1568	0.0032	0.0919	0.0094	0.0149

Bonferroni, Holm, and Benjamini-Hochberg

First we order the 10 p-values and then consider the criteria for Bonferroni, Holm, and Benjamini-Hochberg.

i	$p_{(i)}$	$\frac{\mathcal{E}}{k}$	$\frac{\mathcal{E}}{(k-i+1)}$	$\frac{i\mathcal{E}}{k}$	Hyp.
1	0.0032	0.0050	0.0050	0.0050	$H_{0(7)}$
2	0.0041	0.0050	0.0056	0.0100	$H_{0(2)}$
3	0.0052	0.0050	0.0063	0.0150	$H_{0(1)}$
4	0.0070	0.0050	0.0071	0.0200	$H_{0(5)}$
5	0.0094	0.0050	0.0083	0.0250	$H_{0(9)}$
6	0.0149	0.0050	0.0100	0.0300	$H_{0(10)}$
7	0.0395	0.0050	0.0125	0.0350	$H_{0(3)}$
8	0.0919	0.0050	0.0167	0.0400	$H_{0(8)}$
9	0.1568	0.0050	0.0250	0.0450	$H_{0(6)}$
10	0.4533	0.0050	0.0500	0.0500	$H_{0(4)}$

Reject $H_{0(7)}$ and $H_{0(2)}$ using Bonferroni (SCI). Reject $H_{0(7)}$, $H_{0(2)}$, $H_{0(1)}$, $H_{0(5)}$ using Holm (SFER). Reject $H_{0(7)}$, $H_{0(2)}$, $H_{0(1)}$, $H_{0(5)}$, $H_{0(9)}$, $H_{0(10)}$ using Benjamini-Hochberg (FDR). Simply add $H_{0(3)}$ to the rejection list for per comparison error rate.

Review

- If we perform multiple, say k hypotheses there are different ways to control Type I error from weakest to strongest: (1) per comparison, (2) experimentwise, (3) false discovery rate, (4) strong familywise, and (5) simultaneous CIs.
- `linear.contrast` allows experimentwise control (`jointF=T`), SFER (`bonferroni=T`) for contrasts considered before looking at the data, SFER (`scheffe=T`) for any contrasts.
- If we only care about pairwise comparisons, `pairwise` will give all of them; only the CIs are properly adjusted. We reject $H_0 : \mu_i - \mu_j = 0 \Leftrightarrow$ the CI does not include zero. FDR protected by `type="snk"`, SFER protected by `type="regwr"`, SCI given by default, Tukey. Use `lines` to visualize treatments that are not sig. different.

Review

- `compare.to.control` does exactly that; provides SCI (which also protects SFER).
- `compare.to.best` finds the best treatments – must specify if low or high is "good" – and whether there are significant differences among them, via SCI (protects SFER).
- The "Bonferroni" style methods on the couple slides before the review work great if you only have a bunch of p-values for your k hypotheses. Just order them and compare to the cutoffs.