Multiple Testing

Gary W. Oehlert

School of Statistics
University of Minnesota

January 28, 2016

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%.

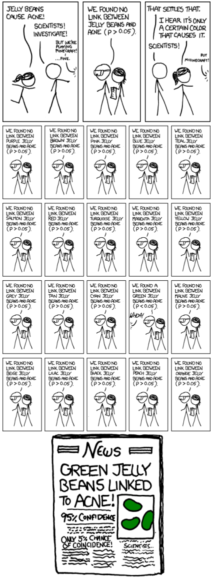
Now roll the die 20 times. What is the chance of getting at least one 1?
$1 - .95^{20} = .642$

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

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?


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

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.

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

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

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

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.

Failing to reject a false null[1] is a Type II error. This is a false negative, saying something is not happening when, in fact, something is happening.

[1] In ye olde days one would say "accept the null," but I prefer "fail to reject."

	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.

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

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,

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

The **per experiment** 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

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

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.

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.

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.)

I have described the error rates 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, the harder it is to see differences that are really there.

As you control stronger and stronger type I error rates, 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 pushes you to get the nulls right and the non-nulls right, you also have to be able to say where all the parameter values are.

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.

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.

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.

Search your soul to find the weakest type I error rate that is compatible with the kind of inference you wish to make. Then choose a procedure that controls that error rate. It's a Goldilocks problem where you need to balance the types of errors.

There are many different procedures, particularly pairwise comparison procedures, and people argue for their favorites. My philosophy is to argue of the type I error rate to be controlled, and then choose the corresponding procedure.

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.

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

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

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.)

For a confidence interval use

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

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).

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 $1 - \mathcal{E}$.

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 dead 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.

Fiber percent example.

Studentized range

Before moving on, we need a new distribution called the Studentized range. Suppose $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$ (the single mean model) is true. Look at the distribution of

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

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.

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:

$$\overline{y}_{i\bullet} - \overline{y}_{j\bullet}$$

These comparisons are an obvious thing to do, and there are lots of procedures out there to do them. We will work on them according to the error rate that they control.

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

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

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

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.

Consider the difference

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

This is a stretch of $i - j + 1$ means. (Let $i - j + 1 = k$, i.e., k is the stretch length.)

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

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

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.

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 $\sqrt{2}$).

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. I am pretty sure that it protects FDR, but I have failed to prove it.

REGWR is Ryan-Einot-Gabriel-Welsch range test. It 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

Or use letters or numbers; treatments sharing a letter or number are not significantly different.

C (1) A (12) B (2)

C (a) A (ab) B (b)

C^1 A^{12} B^2

C^a A^{ab} B^b

Danger!

There are many, many other procedures, but beware.

There is a procedure called Duncan's New Multiple Range test. Some people like it because it finds lots of differences.

It finds lots of differences because it does not control any of our type I error rates including, believe or not, the per comparison error rate.

I keep away.

Cheese inoculants example.

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.[2]

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

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

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

[2]Actually, I almost always want to compare the new treatments with each other as well, so I don't wind up doing this very often.

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

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

If you are really wedded to just comparing new to control, design with $n_g/n_i \approx \sqrt{g-1}$. This gives best overall results.

Compare to best

Here is something that I think is very useful. We 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 $j \neq i$:

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

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-$\mathcal{E}$.