Multiple comparisons

Summary

When you perform a large number of statistical tests, some will have *P* values less than 0.05 purely by chance, even if all your null hypotheses are really true. The Bonferroni correction is one simple way to take this into account; adjusting the false discovery rate using the Benjamini-Hochberg procedure is a more powerful method.

The problem with multiple comparisons

Any time you reject a null hypothesis because a *P* value is less than your critical value, it's possible that you're wrong; the null hypothesis might really be true, and your significant result might be due to chance. A *P* value of 0.05 means that there's a 5% chance of getting your observed result, *if* the null hypothesis were true. It does *not* mean that there's a 5% chance that the null hypothesis is true.

For example, if you do 100 statistical tests, and for all of them the null hypothesis is actually true, you'd expect about 5 of the tests to be significant at the $P<0.05$ level, just due to chance. In that case, you'd have about 5 statistically significant results, all of which were false positives. The cost, in time, effort and perhaps money, could be quite high if you based important conclusions on these false positives, and it would at least be embarrassing for you once other people did further research and found that you'd been mistaken.

This problem, that when you do multiple statistical tests, some fraction will be false positives, has received increasing attention in the last few years. This is important for such techniques as the use of microarrays, which make it possible to measure RNA quantities for tens of thousands of genes at once; brain scanning, in which blood flow can be estimated in 100,000 or more three-dimensional bits of brain; and evolutionary genomics, where the sequences of every gene in the genome of two or more species can be compared. There is no universally accepted approach for dealing with the problem of multiple comparisons; it is an area of active research, both in the mathematical details and broader epistomological questions.

Controlling the familywise error rate: Bonferroni correction

The classic approach to the multiple comparison problem is to control the familywise error rate. Instead of setting the critical *P* level for significance, or alpha, to 0.05, you use a lower critical value. If the null hypothesis is true for all of the tests, the probability of getting *one* result that is significant at this new, lower critical value is 0.05. In other words, if all the null hypotheses are true, the probability that the family of tests includes one or more false positives due to chance is 0.05.

The most common way to control the familywise error rate is with the Bonferroni correction. You find the critical value (alpha) for an individual test by dividing the familywise error rate (usually 0.05) by the number of tests. Thus if you are doing 100 statistical tests, the critical value for an individual test would be 0.05/100=0.0005, and you would only consider individual tests with $P<0.0005$ to be significant.

As an example, García-Arenzana et al. (2014) tested associations of 25 dietary variables with mammographic density, an important risk factor for breast cancer, in Spanish women. They found the following results:

Dietary variable	*P* value
Total calories	<0.001
Olive oil	0.008
Whole milk	0.039
White meat	0.041
Proteins	0.042
Nuts	0.06
Cereals and pasta	0.074
White fish	0.205
Butter	0.212

Vegetables	0.216
Skimmed milk	0.222
Red meat	0.251
Fruit	0.269
Eggs	0.275
Blue fish	0.34
Legumes	0.341
Carbohydrates	0.384
Potatoes	0.569
Bread	0.594
Fats	0.696
Sweets	0.762
Dairy products	0.94
Semi-skimmed milk	0.942
Total meat	0.975
Processed meat	0.986

As you can see, five of the variables show a significant ($P<0.05$) P value. However, because García-Arenzana et al. (2014) tested 25 dietary variables, you'd expect one or two variables to show a significant result purely by chance, even if diet had no real effect on mammographic density. Applying the Bonferroni correction, you'd divide $P=0.05$ by the number of tests (25) to get the Bonferroni critical value, so a test would have to have $P<0.002$ to be significant. Under that criterion, only the test for total calories is significant.

The Bonferroni correction is appropriate when a single false positive in a set of tests would be a problem. It is mainly useful when there are a fairly small number of multiple comparisons and you're looking for one or two that might be significant. However, if you have a large number of multiple comparisons and you're looking for many that might be significant, the Bonferroni correction may lead to a very high rate of false negatives. For example, let's say you're comparing the expression level of 20,000 genes between liver cancer tissue and normal liver tissue. Based on previous studies, you are hoping to find dozens or hundreds of genes with different expression levels. If you use the Bonferroni correction, a P value would have to be less than 0.05/20000=0.0000025 to be significant. Only genes with huge differences in expression will have a P value that low, and could miss out on a lot of important differences just because you wanted to be sure that your results did not include a single false positive.

An important issue with the Bonferroni correction is deciding what a "family" of statistical tests is. García-Arenzana et al. (2014) tested 25 dietary variables, so are these tests one "family," making the critical P value 0.05/25? But they also measured 13 non-dietary variables such as age, education, and socioeconomic status; should they be included in the family of tests, making the critical P value 0.05/38? And what if in 2015, García-Arenzana et al. write another paper in which they compare 30 dietary variables between breast cancer and non-breast cancer patients; should they include those in their family of tests, and go back and reanalyze the data in their 2014 paper using a critical P value of 0.05/55? There is no firm rule on this; you'll have to use your judgment, based on just how bad a false positive would be. Obviously, you should make this decision before you look at the results, otherwise it would be too easy to subconsciously rationalize a family size that gives you the results you want.

Controlling the false discovery rate: Benjamini–Hochberg procedure

An alternative approach is to control the false discovery rate. This is the proportion of "discoveries" (significant results) that are actually false positives. For example, let's say you're using microarrays to compare expression levels for 20,000 genes between liver tumors and normal liver cells. You're going to do additional experiments on any genes that show a significant difference between the normal and tumor cells, and you're willing to accept up to 10% of the genes with significant results being false positives; you'll find out they're false positives when you do the followup experiments. In this case, you would set your false discovery rate to 10%.

One good technique for controlling the false discovery rate was briefly mentioned by Simes (1986) and developed in detail by Benjamini and Hochberg (1995). Put the individual P values in order, from smallest to largest. The smallest P value has a rank of i=1, then next smallest has i=2, etc. Compare each individual P value to its

Benjamini-Hochberg critical value, (*i*/*m*)*Q*, where *i* is the rank, *m* is the total number of tests, and *Q* is the false discovery rate you choose. The largest *P* value that has *P*<(i/m)Q is significant, and *all* of the *P* values smaller than it are also significant, even the ones that aren't less than their Benjamini-Hochberg critical value.

To illustrate this, here are the data from García-Arenzana et al. (2014) again, with the Benjamini-Hochberg critical value for a false discovery rate of 0.25.

Dietary variable	*P* value	Rank	(i/m)Q
Total calories	<0.001	1	0.010
Olive oil	0.008	2	0.020
Whole milk	0.039	3	0.030
White meat	0.041	4	0.040
Proteins	0.042	5	0.050
Nuts	0.060	6	0.060
Cereals and pasta	0.074	7	0.070
White fish	0.205	8	0.080
Butter	0.212	9	0.090
Vegetables	0.216	10	0.100
Skimmed milk	0.222	11	0.110
Red meat	0.251	12	0.120
Fruit	0.269	13	0.130
Eggs	0.275	14	0.140
Blue fish	0.34	15	0.150
Legumes	0.341	16	0.160
Carbohydrates	0.384	17	0.170
Potatoes	0.569	18	0.180
Bread	0.594	19	0.190
Fats	0.696	20	0.200
Sweets	0.762	21	0.210
Dairy products	0.94	22	0.220
Semi-skimmed milk	0.942	23	0.230
Total meat	0.975	24	0.240
Processed meat	0.986	25	0.250

Reading down the column of *P* values, the largest one with *P*<(*i*/*m*)*Q* is proteins, where the individual *P* value (0.042) is less than the (*i*/*m*)*Q* value of 0.050. Thus the first five tests would be significant. Note that whole milk and white meat are significant, even though their *P* values are not less than their Benjamini-Hochberg critical values; they are significant because they have *P* values less than that of proteins.

When you use the Benjamini-Hochberg procedure with a false discovery rate greater than 0.05, it is quite possible for individual tests to be significant even though their *P* value is greater than 0.05. Imagine that all of the *P* values in the García-Arenzana et al. (2014) study were between 0.10 and 0.24. Then with a false discovery rate of 0.25, all of the tests would be significant, even the one with *P*=0.24. This may seem wrong, but if all 25 null hypotheses were true, you'd expect the largest *P* value to be well over 0.90; it would be extremely unlikely that the largest *P* value would be less than 0.25. You would only expect the largest *P* value to be less than 0.25 if most of the null hypotheses were false, and since a false discovery rate of 0.25 means you're willing to reject a few true null hypotheses, you would reject them all.

You should carefully choose your false discovery rate before collecting your data. Usually, when you're doing a large number of statistical tests, your experiment is just the first, exploratory step, and you're going to follow up with more experiments on the interesting individual results. If the cost of additional experiments is low and the cost of a false negative (missing a potentially important discovery) is high, you should probably use a fairly high false discovery rate, like 0.10 or 0.20, so that you don't miss anything important. Sometimes people use a false discovery

rate of 0.05, probably because of confusion about the difference between false discovery rate and probability of a false positive when the null is true; a false discovery rate of 0.05 is probably too low for many experiments.

The Benjamini-Hochberg procedure is less sensitive than the Bonferroni procedure to your decision about what is a "family" of tests. If you increase the number of tests, and the distribution of *P* values is the same in the newly added tests as in the original tests, the Benjamini-Hochberg procedure will yield the same proportion of significant results. For example, if García-Arenzana et al. (2014) had looked at 50 variables instead of 25 and the new 25 tests had the same set of *P* values as the original 25, they would have 10 significant results under Benjamini-Hochberg with a false discovery rate of 0.25. This doesn't mean you can completely ignore the question of what constitutes a family; if you mix two sets of tests, one with some low *P* values and a second set without low *P* values, you will reduce the number of significant results compared to just analyzing the first set by itself.

Sometimes you will see a "Benjamini-Hochberg adjusted *P* value." The adjusted *P* value for a test is either the raw *P* value times *m/i* or the adjusted *P* value for the next higher raw *P* value, whichever is smaller (remember that *m* is the number of tests and *i* is the rank of each test, with 1 the rank of the smallest *P* value). If the adjusted *P* value is smaller than the false discovery rate, the test is significant. For example, the adjusted *P* value for proteins in the example data set is 0.042×(25/5)=0.210; the adjusted *P* value for white meat is the smaller of 0.041×(25/4)=0.256 or 0.210, so it is 0.210. In my opinion "adjusted *P* values" are a little confusing, since they're not really estimates of the probability (*P*) of anything. I think it's better to give the raw *P* values and say which are significant using the Benjamini-Hochberg procedure with your false discovery rate, but if Benjamini-Hochberg adjusted *P* values are common in the literature of your field, you might have to use them.

Assumption

The Bonferroni correction and Benjamini-Hochberg procedure assume that the individual tests are independent of each other, as when you are comparing sample A vs. sample B, C vs. D, E vs. F, etc. If you are comparing sample A vs. sample B, A vs. C, A vs. D, etc., the comparisons are not independent; if A is higher than B, there's a good chance that A will be higher than C as well. One place this occurs is when you're doing unplanned comparisons of means in anova, for which a variety of other techniques have been developed, such as the Tukey-Kramer test. Another experimental design with multiple, non-independent comparisons is when you compare multiple variables between groups, and the variables are correlated with each other within groups. An example would be knocking out your favorite gene in mice and comparing everything you can think of on knockout vs. control mice: length, weight, strength, running speed, food consumption, feces production, etc. All of these variables are likely to be correlated within groups; mice that are longer will probably also weigh more, would be stronger, run faster, eat more food, and poop more. To analyze this kind of experiment, you can use multivariate analysis of variance, or manova, which I'm not covering in this textbook.

Other, more complicated techniques, such as Reiner et al. (2003), have been developed for controlling false discovery rate that may be more appropriate when there is lack of independence in the data. If you're using microarrays, in particular, you need to become familiar with this topic.

When not to correct for multiple comparisons

The goal of multiple comparisons corrections is to reduce the number of false positives, because false positives can be embarrassing, confusing, and cause you and other people to waste your time. An unfortunate byproduct of correcting for multiple comparisons is that you may increase the number of false negatives, where there really is an effect but you don't detect it as statistically significant. If false negatives are very costly, you may not want to correct for multiple comparisons at all. For example, let's say you've gone to a lot of trouble and expense to knock out your favorite gene, mannose-6-phosphate isomerase (*Mpi*), in a strain of mice that spontaneously develop lots of tumors. Hands trembling with excitement, you get the first $Mpi^{-/-}$ mice and start measuring things: blood pressure, growth rate, maze-learning speed, bone density, coat glossiness, everything you can think of to measure on a mouse. You measure 50 things on $Mpi^{-/-}$ mice and normal mice, run the approriate statistical tests, and the smallest *P* value is 0.013 for a difference in tumor size. If you use a Bonferroni correction, that *P*=0.013 won't be close to significant; it might not be significant with the Benjamini-Hochberg procedure, either. Should you conclude that there's no significant difference between the $Mpi^{-/-}$ and $Mpi^{+/+}$ mice, write a boring little paper titled "Lack of anything interesting in $Mpi^{-/-}$ mice," and look for another project? No, your paper should be "Possible effect of *Mpi* on cancer." You should be suitably cautious, of course, and emphasize in the paper that there's a good chance that your result is a false positive; but the cost of a false positive—if further experiments show that *Mpi* really has no effect on tumors—

is just a few more experiments. The cost of a false negative, on the other hand, could be that you've missed out on a hugely important discovery.

How to do the tests

Spreadsheet

I have written a spreadsheet to do the Benjamini-Hochberg procedure on up to 1000 *P* values. It will tell you which *P* values are significant after controlling for the false discovery rate you choose. It will also give the Benjamini-Hochberg adjusted *P* values, even though I think they're kind of stupid.

I have also written a spreadsheet to do the Bonferroni correction on up to 1000 *P* values.

Web pages

I'm not aware of any web pages that will perform the Benjamini-Hochberg procedure.

R

Salvatore Mangiafico's *R Companion* has a sample R programs for the Bonferroni, Benjamini-Hochberg, and several other methods for correcting for multiple comparisons.

SAS

There is a PROC MULTTEST that will perform the Benjamini-Hochberg procedure, as well as many other multiple-comparison corrections. Here's an example using the diet and mammographic density data from García-Arenzana et al. (2014).

```
DATA mammodiet;
   INPUT food $ Raw_P;
   cards;
Blue_fish .34
Bread .594
Butter .212
Carbohydrates .384
Cereals_and_pasta .074
Dairy_products .94
Eggs .275
Fats .696
Fruit .269
Legumes .341
Nuts .06
Olive_oil .008
Potatoes .569
Processed_meat .986
Proteins .042
Red_meat .251
Semi-skimmed_milk .942
Skimmed_milk .222
Sweets .762
Total_calories .001
Total_meat .975
Vegetables .216
White_fish .205
White_meat .041
Whole_milk .039
;
PROC SORT DATA=mammodiet OUT=sorted_p;
   BY Raw_P;
PROC MULTTEST INPVALUES=sorted_p FDR;
RUN;
```

Note that the *P* value variable *must* be named "Raw_P". I sorted the data by "Raw_P" before doing the multiple comparisons test, to make the final output easier to read. In the PROC MULTTEST statement, INPVALUES tells you what file contains the Raw_P variable, and FDR tells SAS to run the Benjamini-Hochberg procedure.

The output is the original list of *P* values and a column labeled "False Discovery Rate." If the number in this column is less than the false discovery rate you chose before doing the experiment, the original ("raw") *P* value is

significant.

Test	Raw	False Discovery Rate
1	0.0010	0.0250
2	0.0080	0.1000
3	0.0390	0.2100
4	0.0410	0.2100
5	0.0420	0.2100
6	0.0600	0.2500
7	0.0740	0.2643
8	0.2050	0.4911
9	0.2120	0.4911
10	0.2160	0.4911
11	0.2220	0.4911
12	0.2510	0.4911
13	0.2690	0.4911
14	0.2750	0.4911
15	0.3400	0.5328
16	0.3410	0.5328
17	0.3840	0.5647
18	0.5690	0.7816
19	0.5940	0.7816
20	0.6960	0.8700
21	0.7620	0.9071
22	0.9400	0.9860
23	0.9420	0.9860
24	0.9750	0.9860
25	0.9860	0.9860

So if you had chosen a false discovery rate of 0.25, the first 6 would be significant; if you'd chosen a false discovery rate of 0.15, only the first two would be significant.

References

García-Arenzana, N., E.M. Navarrete-Muñoz, V. Lope, P. Moreo, S. Laso-Pablos, N. Ascunce, F. Casanova-Gómez, C. Sánchez-Contador, C. Santamariña, N. Aragonés, B.P. Gómez, J. Vioque, and M. Pollán. 2014. Calorie intake, olive oil consumption and mammographic density among Spanish women. International journal of cancer 134: 1916-1925.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57: 289-300.

Reiner, A., D. Yekutieli and Y. Benjamini. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19: 368-375.

Simes, R.J. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 751-754.

This page was last revised July 20, 2015. Its address is http://www.biostathandbook.com/multiplecomparisons.html. It may be cited as:
McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland. This web page contains the content of pages 254-260 in the printed version.

©2014 by John H. McDonald. You can probably do what you want with this content; see the permissions page (http://www.biostathandbook.com/permissions.html) for details.