The Bonferonni and Šidák Corrections for Multiple Comparisons

Hervé Abdi[1]

1 Overview

The more tests we perform on a set of data, the more likely we are to reject the null hypothesis when it is true (*i.e.,* a "Type I" error). This is a consequence of the logic of hypothesis testing: We reject the null hypothesis if we witness a *rare* event. But the larger the number of tests, the easier it is to find rare events and therefore the easier it is to make the mistake of thinking that there is an effect when there is none. This problem is called the *inflation* of the alpha level. In order to be protected from it, one strategy is to correct the alpha level when performing multiple tests. Making the alpha level more stringent (*i.e.,* smaller) will create less errors, but it may also make it harder to detect real effects.

2 The different meanings of alpha

Maybe it is because computers make it easier to run statistical analyses that researchers perform more and more statistical tests on a

[1]In: Neil Salkind (Ed.) (2007). *Encyclopedia of Measurement and Statistics.* Thousand Oaks (CA): Sage.
Address correspondence to: Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utd.edu/~herve

same set of data. For example, brain imaging researchers will routinely run millions of tests to analyze an experiment. Running so many tests increases the risk of false alarms. To illustrate, imagine the following "pseudo-experiment":

> I toss 20 coins, and I try to force the coins to fall on the heads. I know that, from the "binomial test," the null hypothesis is rejected at the $\alpha = .05$ level if the number of heads is greater than 14. I repeat this experiment 10 times.

Suppose that one trial gives the "significant" result of 16 heads *versus* 4 tails. Did I influence the coins on that occasion? Of course not, because the larger the number of experiments, the greater the probability of detecting a low-probability event (like 16 *versus* 4). In fact, waiting long enough is a sure way of detecting rare events!

2.1 Probability in the family

A *family of tests* is the technical term for a series of tests performed on a set of data. In this section we show how to compute the probability of rejecting the null hypothesis at least once in a family of tests when the null hypothesis is true.

For convenience, suppose that we set the significance level at α=.05. For each test (*i.e.*, one trial in the example of the coins) the probability of making a *Type I error* is equal to $\alpha = .05$. The events "making a Type I error" and "not making a Type I error" are *complementary events* (they cannot occur simultaneously). Therefore the probability of *not making a Type I error* on one trial is equal to

$$1 - \alpha = 1 - .05 = .95 \, .$$

Recall that when two events are *independent*, the probability of observing these two events together is the *product* of their probabilities. Thus, if the tests are independent, the probability of not making a Type I error on the first *and* the second tests is

$$.95 \times .95 = (1 - .05)^2 = (1 - \alpha)^2 \, .$$

With 3 tests, we find that the probability of not making a Type I error on all tests is:

$$.95 \times .95 \times .95 = (1 - .05)^3 = (1 - \alpha)^3 .$$

For a family of C tests, the probability of *not* making a Type I error for the *whole family* is:

$$(1 - \alpha)^C .$$

For our example, the probability of not making a Type I error on the family is

$$(1 - \alpha)^C = (1 - .05)^{10} = .599 .$$

Now, what we are looking for is the probability of making one or more Type I errors on the family of tests. This event is the complement of the event *not making a Type I error on the family* and therefore it is equal to

$$1 - (1 - \alpha)^C .$$

For our example, we find

$$1 - (1 - .05)^{10} = .401 .$$

So, with an α level of .05 for *each* of the 10 tests, the probability of wrongly rejecting the null hypothesis is .401.

This example makes clear the need to distinguish between two meanings of α when performing multiple tests:

- The probability of making a Type I error when dealing only with a specific test. This probability is denoted $\alpha[PT]$ (pronounced “alpha *per* test"). It is also called the *testwise* alpha.
- The probability of making at least one Type I error for the whole family of tests. This probability is denoted $\alpha[PF]$ (pronounced “alpha *per* family of tests”). It is also called the *familywise* or the *experimentwise* alpha.

Table 1: Results of a Monte Carlo simulation. Numbers of Type 1 errors when performing $C = 5$ tests for 10,000 families when H_0 is true. How to read the table? For example, 192 families over 10,000 have 2 Type 1 errors, this gives $2 \times 192 = 384$ Type 1 errors.

Number of families with X Type I errors	X: Number of Type I errors *per* family	Number of Type I errors
7,868	0	0
1,907	1	1,907
192	2	384
20	3	60
13	4	52
0	5	0
10,000		2,403

2.2 A Monte Carlo illustration

A “Monte Carlo" simulation can illustrate the difference between $\alpha[PT]$ and $\alpha[PF]$. The Monte Carlo technique consists of running a simulated experiment many times using random data. This gives the pattern of results that happens on the basis of chance.

Here 6 groups with 100 observations *per* group were created with data randomly sampled from the same normal population. By construction, H_0 is true (*i.e.*, all population means are equal). Call that procedure an *experiment*. We performed 5 *independent tests* from these 6 groups. For each test, we computed an F-test. If its probability was smaller than $\alpha = .05$, the test was declared significant (*i.e.*, $\alpha[PT]$ is used). We performed this experiment 10,000 times. Therefore, there were 10,000 experiments, 10,000 families, and $5 \times 10{,}000 = 50{,}000$ tests. The results of this simulation are given in Table 1.

Table 1 shows that H_0 is rejected for 2,403 tests over 50,000 tests performed. From these data, an estimation of $\alpha[PT]$ is computed as:

$$\alpha[PT] = \frac{\text{number of significant tests}}{\text{total number of tests}}$$

$$= \frac{2,403}{50,000} = .0479 \,. \tag{1}$$

This value falls close to the theoretical value of $\alpha = .05$.

For 7,868 families, no test reaches significance. Equivalently for 2,132 families (10,000 − 7,868) at least one Type I error is made. From these data, $\alpha[PF]$ can be estimated as:

$$\alpha[PF] = \frac{\text{number of families with at least 1 Type I error}}{\text{total number of families}}$$

$$= \frac{2,132}{10,000} = .2132 \,. \tag{2}$$

This value falls close to the theoretical value of

$$\alpha[PF] = 1 - (1 - \alpha[PT])^C = 1 - (1 - .05)^5 = .226 \,.$$

2.3 How to correct for multiple tests: Šidàk, Bonferonni, Boole, Dunn

Recall that the probability of making *as least one* Type I error for a family of C tests is

$$\alpha[PF] = 1 - (1 - \alpha[PT])^C \,.$$

This equation can be rewritten as

$$\alpha[PT] = 1 - (1 - \alpha[PF])^{1/C} \,.$$

This formula—derived assuming *independence* of the tests—is sometimes called the Šidàk equation. It shows that in order to reach a given $\alpha[PF]$ level, we need to adapt the $\alpha[PT]$ values used for each test.

Because the Šidàk equation involves a fractional power, it is difficult to compute by hand and therefore several authors derived

a simpler approximation which is known as the *Bonferonni* (the most popular name), or *Boole*, or even *Dunn* approximation. Technically, it is the first (linear) term of a Taylor expansion of the Šidàk equation. This approximation gives

$$\alpha[PT] \approx \frac{\alpha[PF]}{C} .$$

Šidàk and Bonferonni are linked to each other by the inequality

$$\alpha[PT] = 1 - (1 - \alpha[PF])^{1/C} \geq \frac{\alpha[PF]}{C} .$$

They are, in general, very close to each other but the Bonferonni approximation is pessimistic (it always does worse than Šidàk equation). Probably because it is easier to compute, the Bonferonni approximation is more well known (and cited more often) than the exact Šidàk equation.

The Šidàk-Bonferonni equations can be used to find the value of $\alpha[PT]$ when $\alpha[PF]$ is fixed. For example, suppose that you want to perform 4 *independent* tests, and you want to limit the risk of making at least one Type I error to an overall value of $\alpha[PF] = .05$, you will consider a test significant if its associated probability is smaller than

$$\alpha[PT] = 1 - (1 - \alpha[PF])^{1/C} = 1 - (1 - .05)^{1/4} = .0127 .$$

With the Bonferonni approximation, a test reaches significance if its associated probability is smaller than

$$\alpha[PT] = \frac{\alpha[PF]}{C} = \frac{.05}{4} = .0125 ,$$

which is very close to the exact value of .0127.

2.4 Correction for non-independent tests

The Šidàk equation is derived assuming *independence* of the tests. When they are not independent, it gives a lower bound (*cf.* Šidàk, 1967; Games, 1977), and then:

$$\alpha[PF] \leq 1 - (1 - \alpha[PT])^{C} .$$

As previously, we can use a *Bonferonni* approximation because:

$$\alpha[PF] < C\alpha[PT] \; .$$

Šidàk and Bonferonni are related by the inequality

$$\alpha[PF] \leq 1-(1-\alpha[PT])^{C} < C\alpha[PT] \; .$$

The Šidàk and Bonferonni inequalities can also be used to find a correction on $\alpha[PT]$ in order to keep $\alpha[PF]$ fixed. the Šidàk inequality gives

$$\alpha[PT] \approx 1-(1-\alpha[PF])^{1/C} \; .$$

This is a conservative approximation, because the following inequality holds:

$$\alpha[PT] \geq 1-(1-\alpha[PF])^{1/C} \; .$$

The Bonferonni approximation gives

$$\alpha[PT] \approx \frac{\alpha[PF]}{C} \; .$$

2.5 Splitting up $\alpha[PF]$ with unequal slices

With the Bonferonni approximation we can make an unequal allocation of $\alpha[PF]$. This works because with the Bonferonni approximation, $\alpha[PF]$ is the sum of the individual $\alpha[PT]$:

$$\alpha[PF] \approx C\alpha[PT] = \underbrace{\alpha[PT]+\alpha[PT]+\cdots+\alpha[PT]}_{C \text{ times}} \; .$$

If some tests are judged more important *a priori* than some others, it is possible to allocate unequally $\alpha[PF]$ (*cf.* Rosenthal & Rosnow, 1985). For example, suppose we have 3 tests that we want to test with an overall $\alpha[PF] = .05$, and we think that the first test is the most important of the set. Then we can decide to test it with $\alpha[PT] = .04$, and share the remaining value $.01 = .05 - .04$ between the last 2 tests, which will be evaluated each with a value of $\alpha[PT] = .005$. The overall Type I error for the family is equal to $\alpha[PF] = .04 + .005 + .005 = .05$ which was indeed the value we set

beforehand. It should be emphasized, however, that the (subjective) importance of the tests and the unequal allocation of the individual $\alpha[PT]$ should be decided *a priori* for this approach to be statistically valid. An unequal allocation of the $\alpha[PT]$ can also be achieved using the Šidàk inequality, but it is more computationally involved.

3 Alternatives to Bonferonni

The Šidàk-Bonferonni approach becomes very conservative when the number of comparisons becomes large and when the tests are not independent (*e.g.,* as in brain imaging). Recently, some alternative approaches have been proposed (see Shaffer, 1995, for a review) to make the correction less stringent (*e.g.,* Holm 1979, Hochberg, 1988). A more recent approach redefines the problem by replacing the notion of $\alpha[PF]$ by the false discovery rate (FDR) which is defined as the ratio of the number of Type I errors by the number of significant tests (Benjamini & Hochberg, 1995).

References

[1] Benjamini & Hochberg, (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Serie B*, **57**, 289–300.

[2] Games, P.A. (1977). An improved *t* table for simultaneous control on *g* contrasts. *Journal of the American Statistical Association*, **72**, 531–534.

[3] Hochberg Y. (1988). A sharper Bonferonni procedure for multiple tests of significance. *Biometrika*, **75**, 800–803.

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

[5] Rosenthal, R. & Rosnow, R.L. (1985). *Contrast analysis: focused comparisons*. Boston: Cambridge University Press.

[6] Shaffer, J.P. (1995). Multiple Hypothesis Testing *Annual Review of Psychology*, **46**, 561–584.

[7] Šidàk, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association*, **62**, 626–633.