**In Neil Salkind (Ed.), *Encyclopedia of Research Design.***
**Thousand Oaks, CA: Sage. 2010**

# Holm's Sequential Bonferroni Procedure

**Hervé Abdi**

## 1 Overview

The more statistical tests we perform, the more likely we are to reject the null hypothesis when it is true (*i.e.,* a "false alarm," also called a "Type 1" error). This is a consequence of the logic of hypothesis testing: We reject the null hypothesis for *rare* events, and the larger the number of tests, the easier it is to find rare events which are false alarms. 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. The most well known correction is called the Bonferroni correction, it consists in multiplying each probability by the total number of tests performed. A more powerful (*i.e.,* more likely to detect an effect it it exists) sequential version has been proposed by Holm in 1979. In Holm's sequential version, the tests need first to be performed in order to obtain their "$p$-values." The tests are then ordered from the one with the smallest $p$-value to the

Hervé Abdi
The University of Texas at Dallas
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

one with the largest $p$-value. The test with the lowest probability is tested first with a Bonferroni correction involving all tests. The second test is tested with a Bonferroni correction involving one less test and so on for the remaining tests. Holm's approach is more powerful than the Bonferroni approach but it still keeps under control the inflation of the Type 1 error.

## 2 Preliminary: The different meanings of alpha

When we perform more than one statistical test, we need to distinguish between two interpretations of the $\alpha$ level which represents the probability of a Type 1 error. The first interpretation evaluates the probability of a Type 1 error for the whole set of tests whereas the second evaluates the probability for only one test at a time.

### 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 $\alpha$=.05. For each test 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 $\alpha$ level of .05 for *each* of the 10 tests, the probability of incorrectly rejecting the null hypothesis is .401.

This example makes clear the need to distinguish between two meanings of $\alpha$ 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.

## 2.2 How to correct for multiple tests: Šidàk, Bonferroni, Boole, Dunn

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

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

This equation can be rewritten as

$$\alpha[PT] = 1-(1-\alpha[PF])^{1/C} \ . \tag{2}$$

This formula—derived assuming *independence* of the tests—is sometimes called the Šidàk equation. It shows that in order to maintain 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 *Bonferroni* (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

$$\alpha[PF] \approx C \times \alpha[PT] \quad , \tag{3}$$

and

$$\alpha[PT] \approx \frac{\alpha[PF]}{C} \ . \tag{4}$$

Šidàk and Bonferroni are linked to each other by the inequality

$$\alpha[PT] = 1 - (1 - \alpha[PF])^{1/C} \geq \frac{\alpha[PF]}{C} \ . \tag{5}$$

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

The Šidàk-Bonferroni 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 Bonferroni 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.

### 2.3 Bonferroni and Šidàk correction for a $p$ value

When a test has been performed as part of a family comprising $C$ tests, the $p$ value of this test can be corrected with the Šidàk or Bonferroni approaches by replacing $\alpha[PF]$ by $p$ in Equations 1 or 3. Specifically, the Šidàk corrected $p$-value for $C$ comparisons, denoted $p_{\text{Šidàk, } C}$ becomes

$$p_{\text{Šidàk, } C} = 1 - (1-p)^C \ , \tag{6}$$

and the Bonferroni corrected $p$-value for $C$ comparisons, denoted $p_{\text{Bonferroni, } C}$ becomes

$$p_{\text{Bonferroni, } C} = C \times p \ . \tag{7}$$

Note that the Bonferroni correction can give a value of $p_{\text{Bonferroni, } C}$ larger than 1. In such cases, $p_{\text{Bonferroni, } C}$ is set to 1.

## 3 Sequential Holm-Šidàk and Holm-Bonferroni

Holm's procedure is a sequential approach whose goal is to increase the power of the statistical tests while keeping under control the familywise Type I error. As previously, suppose that we want to evaluate a family comprising $C$ tests. The first step in Holm's procedure is to perform the tests in order to obtain their $p$-values, Then we order the tests from the one with the smallest $p$-value to the one with the largest $p$-value. The test with the smallest probability will be tested with a Bonferroni or a Šidàk correction for a family of $C$ tests (Holm used a Bonferroni correction, but Šidàk gives an accurate value and should be preferred to Bonferroni which is an approximation). If the test is not significant, then the procedure stops. If the first test is significant, the test with the second smallest $p$-value is then corrected with a Bonferroni or a Šidàk approach for a family of $(C-1)$ tests. The procedure stops when the first non-significant test is obtained or when all the tests have been performed. Formally, assume that the tests are ordered (according to their $p$-values) from 1 to $C$, and that the procedure stops at the first non-significant test. When using the Šidàk correction with Holm's approach, the corrected $p$-value for

the $i$th-test, denoted $p_{\text{Šidàk, } i|C}$ is computed as:

$$p_{\text{Šidàk, } i|C} = 1 - (1-p)^{C-i+1} \ . \tag{8}$$

When using the Bonferroni correction with Holm's approach, the corrected $p$-value for the $i$th-test, denoted $p_{\text{Bonferroni, } i|C}$ is computed as:

$$p_{\text{Bonferroni, } i|C} = (C - i + 1) \times p \ . \tag{9}$$

Just like the standard Bonferroni procedure, corrected $p$-values larger than 1 are set equal to 1.

## 3.1 Example Holm-Šidàk and Holm-Bonferroni

Suppose that we have designed a study involving analysis of variance and we want to perform three tests (see Contrast entry for more details). The $p$ values for these three tests are equal to .0000040, .016100, and .0612300 (we have ordered them from the smallest to the largest). So, here we have $C = 3$. The first test has an original $p$ value of $p = .0000040$. Because it is the first of the series, we have $i = 1$, and its corrected $p$-value using the Holm-Šidàk approach (*cf.* Equation 8) is equal to:

$$\begin{aligned} p_{\text{Šidàk, } i|C} &= 1 - (1-p)^{C-i+1} = p_{\text{Šidàk, } 3|3} = 1 - (1 - .0000040)^{3-1+1} \\ &= p_{\text{Šidàk, } 1|3} = 1 - (1 - .0000040)^3 = 1 - .999960^3 \\ &= 000119 \ . \end{aligned} \tag{10}$$

Using the Bonferroni approximation (*cf.* Equation 9) will give a corrected $p$ value of $p_{\text{Bonferroni, } 1|3} = .000120$. Because the corrected $p$ value for the first test is significant, we proceed to the second test for which $i = 2$ and $p = .016100$. Using Equations 8 and 9 we find the corrected $p$ values of $p_{\text{Šidàk, } 2|3} = .031941$, and $p_{\text{Bonferroni, } 2|3} = .032200$. The corrected $p$ values are significant and, so, we proceed to evaluating the last lest for which $i = 3$. Because this is the last of the series, the corrected $p$ values are now equal to the uncorrected $p$ value of $p = p_{\text{Šidàk, } 3|3} = p_{\text{Bonferroni, } 3|3} = .612300$, which is clearly not significant. Table 1 gives the results of the Holm's sequential procedure along with the values of the standard Šidàk and Bonferroni corrections.

**Table 1:** Šidàk, Bonferroni, Holm-Šidàk, and Holm-Bonferroni corrections for multiple comparisons for a set of $C = 3$ tests with $p$-values of .0000040, .016100, and .0612300.

| | $\alpha[PT]$ | Šidàk $p_{\text{Šidàk, } C}$ | Bonferroni $p_{\text{Bonferroni, } C}$ | Holm-Šidàk $p_{\text{Šidàk, } i\|C}$ | Holm-Bonferroni $p_{\text{Bonferroni, } i\|C}$ |
|---|---|---|---|---|---|
| $i$ | $p$ | $1-(1-p)^{C-i+1}$ | $C \times p$ | $1-(1-p)^{C-i+1}$ | $(C-i+1) \times p$ |
| 1 | 0.000040 | 0.000119 | 0.000120 | 0.000119 | 0.000120 |
| 2 | 0.016100 | 0.047526 | 0.048300 | 0.031941 | 0.032200 |
| 3 | 0.612300 | 0.941724 | 1.000000 | 0.612300 | 0.612300 |

## 4 Correction for non-independent tests

The Šidàk equation is derived assuming *independence* of the tests. When they are not independent, it gives a conservative estimate (*cf.* Šidàk, 1967; Games, 1977). The Bonferonni being a conservative estimation of Šidàk will also give a conservative estimate. Similarly, the sequential Holm's approach is conservative when the tests are not independent. Holm's approach is obviously more powerful than Šidàk (because the $p_{\text{Šidàk, } i|C}$ values are always smaller than or equal to the $p_{\text{Šidàk, } C}$ values), but it still controls the overall familywise error rate. The larger the number of tests, the larger the increase in power with Holm's procedure compared to the standard Šidàk (or Bonferroni) correction.

## 5 Alternatives to Holm-Bonferroni

The Šidàk-Bonferroni as well as Holm's approaches become 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.,* 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).

## Further readings

Abdi, H., Edelman, B., Valentin, D., & Dowling, W.J. (2009). *Experimental design and analysis for psychology*. Oxford: Oxford University Press.

Aickin, M. & Gensler, H. (1996) Adjusting for multiple testing when reporting research results: The bonferroni *vs.* Holm methods. *American Journal of Public Health*, **86**, 726–728.

Benjamini, Y. & Hochberg, T. (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.

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

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

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

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

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