

Learning Curve

Multiple Testing and Protection Against a Type 1 (False Positive) Error Using the Bonferroni and Hochberg Corrections

Chittaranjan Andrade

ABSTRACT

In a given study, if many related outcomes are tested for statistical significance, one or more outcomes may emerge significant at the $P < 0.05$ level not because they are truly significant in the population but because of chance. The larger the number of statistical tests performed, the greater the risk that some of the significant findings are significant because of chance. There are many ways to protect against such false positive or Type 1 errors. The simplest way is to set a more stringent threshold for statistical significance than $P < 0.05$. This can be done using either the Bonferroni or the Hochberg correction. Using the Bonferroni correction, 0.05 is divided by the number of statistical tests being performed and the result is set as the critical P value for statistical significance. Using the Hochberg correction, the P values obtained from the different statistical tests are arranged in descending order of magnitude, and each P value is assessed for significance against progressively more stringent levels for significance. The Bonferroni and Hochberg procedures are explained with the help of examples.

Key words: *Bonferroni correction, false positive error, Hochberg correction, multiple testing, P value, type 1 error*

Imagine that you conduct a 3-month trial in which patients randomized to receive risperidone or haloperidol are examined to determine which antipsychotic is associated with better outcomes for negative symptoms and cognitive functioning. In this trial, negative symptoms are assessed using the Positive and Negative Syndrome Scale-Negative Syndrome subscale (PANSS-N) and the Scale for Assessment of Negative Symptoms (SANS); the total score on each scale is the outcome of interest. Cognitive functioning is assessed using tests of attention and concentration,


visual memory, verbal memory, working memory, and ideational fluency; each test yields a single score. Thus, there are two negative symptom outcomes and five cognitive outcomes, making a total of seven outcomes to be compared between groups.

You know that if you compare just one outcome between the two groups, and if the two groups actually (in the population) do not differ on this outcome, there is only a

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

How to cite this article: Andrade C. Multiple testing and protection against a type 1 (false positive) error using the Bonferroni and Hochberg corrections. *Indian J Psychol Med* 2019;41:99-100.

Access this article online	
Website: www.ijpm.info	Quick Response Code 
DOI: 10.4103/IJPSYM.IJPSYM_499_18	

Department of Psychopharmacology, National Institute of Mental Health and Neurosciences, Bangalore, Karnataka, India

Address for correspondence: Dr. Chittaranjan Andrade

Department of Psychopharmacology, National Institute of Mental Health and Neurosciences, Bangalore - 560 029, Karnataka, India.

E-mail: andradec@gmail.com

5% probability that the result will be statistically significant because of chance; this is what $P < 0.05$ means.^[1] You also know that the larger the number of outcomes compared between the two groups, the greater the likelihood that one or more outcomes will be significant by chance alone. In fact, if five related outcomes are tested, there is a 23% probability that one of the outcomes will be significant by chance.^[1] This is known as a false positive error or a Type 1 statistical error.^[1] So how would you protect against an inflated Type 1 error when you compare the risperidone and haloperidol groups?

Although negative symptom burden and cognitive impairment are correlated, because they represent different conceptual entities it would be reasonable to protect against a Type 1 error separately for the two negative symptom outcomes and for the five cognitive outcomes. Protection against a Type 1 error can be done in many ways. One method sets a more stringent value of P for statistical significance. This can be done using the Bonferroni correction or the Hochberg correction.^[2]

THE BONFERRONI CORRECTION

With this method, the value of P for statistical significance (conventionally, 0.05) is divided by the number of statistical tests performed. So, for the negative symptom outcomes, because there are two tests (one for PANSS-N and one for SANS), P for statistical significance is set at $0.05/2$ or 0.025. This means that the outcomes for PANSS-N and SANS will be considered significant only if the P values associated with these tests are <0.025 instead of <0.05 , as conventional. With regard to the cognitive outcomes, because there are five tests, for any of the five outcomes to be considered statistically significant, it should result in a P value that is $<0.05/5$; that is, <0.01 .

The Bonferroni correction is considered *conservative*; that is, it makes it quite difficult to obtain statistically significant results. This is because when the number of tests performed is large, the P value required for statistical significance becomes quite small and is hard to achieve. In other words, the Bonferroni correction magnifies the risk of a false negative or Type 2 statistical error.^[1] The Hochberg sequential procedure offers a better balance between the Type 1 and Type 2 error risks.

THE HOCHBERG SEQUENTIAL PROCEDURE

With this method, after the groups are compared on each of the five cognitive outcomes, the P values obtained are arranged in descending order of magnitude. If the outcome with the largest P value is significant at the 0.05 level (i.e., $P < 0.05$), then all the outcomes are considered

significant. If the first P value is >0.05 , then the second P value is examined; if the second P value is $<0.05/2$ (that is, 0.025), then this outcome and all the outcomes with smaller P values are considered significant. If the second P value is >0.025 , then the third P value is examined; if the third P value is $<0.05/3$ (that is, 0.017), then this outcome and all the outcomes with smaller P values are considered significant; and so on.

For the negative symptom outcomes, if the larger of the two P values is <0.05 , then both outcomes are considered significant. If the larger value is >0.05 , the second P value will be considered significant only if it is $<0.05/2$; that is, 0.025.

Effectively, the Hochberg sequential procedure applies progressively more stringent criteria for statistical significance, and the last P value is examined at the Bonferroni correction level if the previous P values were not significant on Hochberg testing.

NOTES

1. Corrections for a Type 1 statistical error are necessary only when many tests of the same construct (e.g., cognition) are conducted. Correction is generally considered unnecessary if different tests examine different constructs (e.g., psychosis, memory, and extrapyramidal symptoms). However, in such a context, the issue of primary outcome vs secondary outcomes must be considered.^[3]
2. Avoidance of a Type 1 error is desirable in confirmatory studies but may be dispensed with in exploratory studies where authors do not wish to miss a potentially significant outcome
3. Sometimes, authors may set an arbitrarily conservative P value (e.g., $P < 0.01$) for all tests to modestly protect against a Type 1 error.^[4]

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCE

1. Norman GR, Streiner DL. Biostatistics: The Bare Essentials, 4th ed. Shelton, Connecticut, USA: People's Medical Publishing House; 2014.
2. Cao J, Zhang S. Multiple comparison procedures. JAMA 2014;312:543-4.
3. Andrade C. The primary outcome measure and its importance in clinical trials. J Clin Psychiatry 2015;76:e1320-3.
4. Singh NM, Sathyaprabha TN, Thirthalli J, Andrade C. Effects of electrical stimulus composition on cardiac electrophysiology in a rodent model of electroconvulsive therapy. Indian J Psychiatry 2018;60:17-23.