What's wrong with Bonferroni adjustments

Thomas V Perneger

Institute of Social and Preventive Medicine, University of Geneva, CH-1211 Geneva 4, Switzerland

Thomas V Perneger, *medical epidemiologist*

Correspondence to: Dr Perneger perneger@cmu.unige.ch

BMJ 1998;316:1236–8

When more than one statistical test is performed in analysing the data from a clinical study, some statisticians and journal editors demand that a more stringent criterion be used for "statistical significance" than the conventional $P<0.05$.[1] Many well meaning researchers, eager for methodological rigour, comply without fully grasping what is at stake. Recently, adjustments for multiple tests (or Bonferroni adjustments) have found their way into introductory texts on medical statistics, which has increased their apparent legitimacy.[2,3] This paper advances the view, widely held by epidemiologists, that Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference.[4,5]

Summary points

Adjusting statistical significance for the number of tests that have been performed on study data—the Bonferroni method—creates more problems than it solves

The Bonferroni method is concerned with the general null hypothesis (that all null hypotheses are true simultaneously), which is rarely of interest or use to researchers

The main weakness is that the interpretation of a finding depends on the number of other tests performed

The likelihood of type II errors is also increased, so that truly important differences are deemed non-significant

Simply describing what tests of significance have been performed, and why, is generally the best way of dealing with multiple comparisons

Adjustment for multiple tests

Bonferroni adjustments are based on the following reasoning.[1–3] If a null hypothesis is true (for instance, two treatment groups in a randomised trial do not differ in terms of cure rates), a significant difference ($P<0.05$) will be observed by chance once in 20 trials. This is the type I error, or α. When 20 independent tests are performed (for example, study groups are compared with regard to 20 unrelated variables) and the null hypothesis holds for all 20 comparisons, the chance of at least one test being significant is no longer 0.05, but 0.64. The formula for the error rate across the study is $1-(1-\alpha)^n$, where n is the number of tests performed. However, the Bonferroni adjustment deflates the α applied to each, so the study-wide error rate remains at 0.05. The adjusted significance level is $1-(1-\alpha)^{1/n}$ (in this case 0.00256), often approximated by α/n (here 0.0025). What is wrong with this statistical approach?

Problems

Irrelevant null hypothesis

The first problem is that Bonferroni adjustments are concerned with the wrong hypothesis.[4–6] The study-wide error rate applies only to the hypothesis that the two groups are identical on all 20 variables (the universal null hypothesis). If one or more of the 20 P values is less than 0.00256, the universal null hypothesis is rejected. We can say that the two groups are not equal for all 20 variables, but we cannot say which, or even how many, variables differ. Such information is usually of no interest to the researcher, who wants to assess each variable in its own right. A clinical equivalent would be the case of a doctor who orders 20 different laboratory tests for a patient, only to be told that some are abnormal, without further detail. Thus, Bonferroni adjustments provide a correct answer to a largely irrelevant question.

Inference defies common sense

Bonferroni adjustments imply that a given comparison will be interpreted differently according to how many other tests were performed. For example, the difference in remission rates between two chemotherapeutic treatments could be interpreted as statistically significant or not depending on whether or not survival rates, quality of life scores, and complication rates were also tested. In a clinical setting, a patient's packed cell volume might be abnormally low, except if the doctor also ordered a platelet count, in which case it could be deemed normal. Surely this is absurd, at least within the current scientific paradigm. Evidence in data is what the data say—other considerations, such as how many other tests were performed, are irrelevant.

Increase in type II errors

Type I errors cannot decrease (the whole point of Bonferroni adjustments) without inflating type II errors (the probability of accepting the null hypothesis when the alternative is true).[4] And type II errors are no less false than type I errors. In clinical practice, if a high concentration of creatine kinase were considered compatible with "no myocardial infarction" by virtue of a Bonferroni adjustment, the patient would be denied appropriate care. In research, an effective treatment may be deemed no better than placebo. Thus, contrary to what some researchers believe, Bonferroni adjustments do not guarantee a "prudent" interpretation of results.

What tests should be included?

Most proponents of the Bonferroni method would count at least all the statistical tests in a given report as a basis for adjusting P values. But how about tests that were performed, but not published, or tests published in other papers based on the same study? If several

papers are planned, should future ones be accounted for in the first publication? Should we worry about error rates related to an investigator—taking the number of tests he or she has done in their lifetime into consideration[6]—or error rates related to journals? Should confidence intervals, which are not statistical tests, but are often interpreted as such (the confidence interval includes 0, hence the groups do not differ) be counted? No statistical theory provides answers for these practical issues.

A futuristic scenario

What would happen to biomedical research if Bonferroni adjustments became routine? Cynical researchers would slice their results like salami, publishing one P value at a time to escape the wrath of the statistical reviewer. Idealists would conduct studies to examine only one association at a time—wasting time, energy, and public money. Meta-analysts would go out of business, since a pooled analysis would invalidate retrospectively all original findings by adding more tests to be adjusted for. Journals would have to create a new section entitled "P value updates," in which P values of previously published papers would be corrected for newly published tests based on the same study. And so on ....

Back to the Neyman-Pearson theory

These objections seem so compelling that the reader may wonder why adjustments for multiple tests were developed at all. The answer is that such adjustments are correct in the original framework of statistical test theory, proposed by Neyman and Pearson in the 1920s.[7] This theory was intended to aid decisions in repetitive situations. Imagine that your factory produces light bulbs in lots of 1000, and that testing each bulb before shipment would be impractical. You can decide to test only a sample in each lot, and to reject (literally) any lots in which more than a predefined number (x) of bulbs in the sample are defective. Of course, your decision might be wrong for any particular lot, but the Neyman-Pearson theory provides a decision rule (the number x), so that over many trials your error rates (type I and type II) will be minimised. Now, if for some reason you took 20 samples out of a given lot instead of one, and decided that you would reject the lot if the number of defective bulbs exceeded x in only one sample, you would be much too likely to reject a good lot in error, and a Bonferroni adjustment would restore the original optimal error rates.

DAVID HITCH

The catch is that Neyman and Pearson developed their statistical tests to aid decision making, not to assess evidence in data. The latter practice may be objected to for several reasons (this topic would deserve a discussion of its own), and alternative approaches to statistical inference, such as estimation procedures, use of likelihood ratios, and Bayesian methods, have been proposed.[8–11] Bonferroni adjustments follow the original logic of statistical tests as supports of repeated decisions, but they are of little help in determining what the data say in one particular study.

Should Bonferroni adjustments ever be used?

Statistical adjustment for multiple tests make sense in a few situations. Firstly, the universal null hypothesis is occasionally of interest. For instance, to verify that a disease is not associated with an HLA phenotype, we may compare available HLA antigens (perhaps 40) in a group of cases and controls. If no association existed, at least one test would be significant with a probability of 0.87, and Bonferroni adjustments would protect against making excessive claims. A clinical equivalent is the case of a healthy person undergoing several laboratory tests as part of a general health check. Secondly, adjustments are appropriate when the same test is repeated in many subsamples, such as when stratified analyses (by age group, sex, income status, etc) are conducted without an a priori hypothesis that the primary association should differ between these subgroups. Note that this is the scenario, reminiscent of repeated sampling of the same lot, that Tukey and Bland and Altman use in their justifications of multiple test adjustments.[1 3] Sequential testing of trial results also falls in this category. A final situation in which Bonferroni adjustments may be acceptable is when searching for significant associations without pre-established hypotheses.

The best approach

However, even in these situations, simply describing what was done and why, and discussing the possible interpretations of each result, should enable the reader to reach a reasonable conclusion without the help of Bonferroni adjustments.[5 12] There is an important difference between what the data say and what the researcher (or the reader) believes to be true.[8] The latter depends not only on the data at hand but also on considerations such as whether a finding is biologically plausible or whether the significant test was a serendipitous finding in a fishing expedition. The

integration of prior beliefs with evidence is best achieved by Bayesian methods, not by Bonferroni adjustments. In summary, Bonferroni adjustments have, at best, limited applications in biomedical research, and should not be used when assessing evidence about specific hypotheses.

I thank Dr Richard M Royall, Department of Biostatistics, Johns Hopkins University, for helpful comments on the manuscript.

Funding: Swiss National Science Foundation (PROSPER 3233-32609.91).

Conflict of interest: None.

1 Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 1977;198:679-84.
2 Greenhalgh T. Statistics for the non-statistician. I. Different types of data need different statistical tests. *BMJ* 1997;315:364-6.
3 Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170.
4 Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;1:43-6.
5 Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am J Epidemiol* 1995;142:904-8.
6 Thomas DC, Siemiatycki J, Dewar R, Robins J, Goldberg M, Armstrong RG. The problem of multiple inference in studies designed to generate hypotheses. *Am J Epidemiol* 1985;122:1080-95.
7 Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* 1928;20A:175-240, 263-97.
8 Royall RM. *Statistical inference: a likelihood paradigm.* New York: Chapman and Hall, 1997.
9 Berkson J. Tests of significance considered as evidence. *J Am Stat Assoc* 1942;37:325-35.
10 Goodman SN, Royall R. Evidence and scientific research. *Am J Public Health* 1988;78:1568-74.
11 Oakes M. *Statistical inference.* Boston: Epidemiologic Resources, 1990.
12 Jones DR, Rushton L. Simultaneous inference in epidemiologic studies. *Int J Epidemiol* 1982;11:276-82.

(Accepted 16 January 1998)

Coping with loss

The doctor's losses: ideals versus realities

Glin Bennet

This is the sixth in a series of 10 articles dealing with the different types of loss that doctors will meet in their practice

44 Wellington Park, Bristol BS8 2UW
Glin Bennet, *formerly consultant psychiatrist, United Bristol Healthcare NHS Trust*
Glin.Bennet@quadbtinternet.com

Series editors: Colin Murray Parkes and Andrew Markus

BMJ 1998;316:1238–40

After five years of study, newly qualified doctors may find it hard to realise that much of their future development will involve loss. They will go on gathering information and acquiring skills, but if they are to retain their enthusiasm and to mature as people, they will be learning to live with various losses.

Summary points

Reality often disappoints the expectations of young doctors, who become tired and disillusioned with themselves and with the health care system

A plateau in middle life is often associated with loss of further opportunities, and high achievers may interpret this as failure

To enjoy medicine we must achieve a balance between meeting the needs of our patients and maintaining our own resources of strength, energy, and commitment

Doctors who can acknowledge their own fallibility, accept their own wounds, and accept help from colleagues or others may emerge warmer and more humane

Tiredness

New doctors should enjoy the initial enthusiasm, the ideals and the sense of omnipotence and invulnerability, the buoyant feeling of being able to contribute to the general good, because it may not last for long. Very likely a few months of broken nights will blur the ideals and push the ambitions into the distance. The immediate objective becomes to get through the job.

The grinding tiredness teaches them a lot: about their limitations, that sleep matters, and that it is difficult to be a good doctor when their eyes will not stay open. They become impatient over explanations, and tiredness comes up like a barrier so that they can no longer reach out to anxious and grieving patients.

They are learning that they cannot meet the ideals they set for themselves or the expectations of others. But tiredness is cured by a good sleep and enthusiasm is restored by a relaxing weekend. They can be admired for the long hours they work. They work harder than other people, they work amid the basic crises of living, they know about suffering, they see that people get better through their individual efforts, though they are not successful all the time. The death of a patient is a loss that reminds doctors of their limitations and the limitations of medical science, in which they had been taught to have so much faith. The first time it happens, the doctor is sad, shocked, perhaps angry that the patient could have done that to them.

Loss of unreality

Most doctors have relatively simple lives in these early years, so it is possible, if they want, to give all their waking hours to the work in hand. Then there comes a time when the work is not sufficiently sustaining on its own—at least it ceases to be for most people, especially when the needs of others have to be considered. Now the people with the idealism and enthusiasm are confronted with a fresh reality, and much of a doctor's subsequent life and career will depend on how this matter is addressed.

This is a further lesson in the loss of omnipotence, but in no way is it the beginning of a decline. It is a time for redirecting energy. Doctors who accomplish this and can control the circumstances of their work can have a satisfying life, because medicine offers such abundant opportunities.