Multiple significance tests: the Bonferroni method

J Martin Bland, Douglas G Altman

This is the tenth in a series of occasional notes on medical statistics.

Many published papers include large numbers of significance tests. These may be difficult to interpret because if we go on testing long enough we will inevitably find something which is "significant." We must beware of attaching too much importance to a lone significant result among a mass of non-significant ones. It may be the one in 20 which we expect by chance alone.

Lee *et al* simulated a clinical trial of the treatment of coronary artery disease by allocating 1073 patient records from past cases into two "treatment" groups at random.[1] They then analysed the outcome as if it were a genuine trial of two treatments. The analysis was quite detailed and thorough. As we would expect, it failed to show any significant difference in survival between those patients allocated to the two treatments. Patients were then subdivided by two variables which affect prognosis, the number of diseased coronary vessels and whether the left ventricular contraction pattern was normal or abnormal. A significant difference in survival between the two "treatment" groups was found in those patients with three diseased vessels (the maximum) and abnormal ventricular contraction. As this would be the subset of patients with the worst prognosis, the finding would be easy to account for by saying that the superior "treatment" had its greatest advantage in the most severely ill patients! This approach to the comparison of subgroups is clearly flawed.

Why does this happen? If we test a null hypothesis which is in fact true, using 0·05 as the critical significance level, we have a probability of 0·95 of coming to a not significant—that is, correct—conclusion. If we test two independent true null hypotheses, the probability that neither test will be significant is $0{\cdot}95 \times 0{\cdot}95 = 0{\cdot}90$. If we test 20 such hypotheses the probability that none will be significant is $0{\cdot}95^{20} = 0{\cdot}36$. This gives a probability of $1 - 0{\cdot}36 = 0{\cdot}64$ of getting at least one significant result—we are more likely to get one than not. The expected number of spurious significant results is $20 \times 0{\cdot}05 = 1$. In general, if we have k independent significant tests at the α level of null hypotheses which are all true, the probability that we will get no significant differences is $(1-\alpha)^k$. If we make α small enough we can make the probability that none of the separate tests is significant equal to 0·95. Then if any of the k tests has a P value less than α we will have a significant difference between the treatments at the 0·05 level. Since α will be very small, it can be shown that $(1-\alpha)^k \approx 1 - k\alpha$. If we put $k\alpha = 0{\cdot}05$, so $\alpha = 0.05/k$, we will have probability 0·05 that one of the k tests will have a P value less than α if the null hypotheses are true. Thus, if in a clinical trial we compare two treatments within five subsets of patients the treatments will be significantly different at the 0·05 level if there is a P value less than 0·01 within any of the subsets. This is the Bonferroni method. Note that they are not significant at the 0·01 level, but at only the 0·05 level.

We can do the same thing by multiplying the observed P value from the significance tests by the number of tests, k, any kP which exceeds one being ignored. Then if any kP is less than 0·05 the two treatments are significant at the 0·05 level.

Williams *et al* randomly allocated elderly patients discharged from hospital to two groups.[2] There were no significant differences overall between the intervention and control groups, but among women aged 75-79 living alone the control group showed significantly greater deterioration in physical score than did the intervention group ($P = 0{\cdot}04$), and among men aged over 80 the control group showed significantly greater deterioration in disability score than did the intervention group ($P = 0{\cdot}03$). Subjects were cross classified by age groups, whether living alone, and sex, so there were at least eight subgroups and three different measurement scales. Even if we considered the scales separately the corrected P values are $8 \times 0{\cdot}04 = 0{\cdot}32$ and $8 \times 0{\cdot}03 = 0{\cdot}24$.

A similar problem arises if we have multiple outcome measurements, where the tests will not in general be independent. Newnham *et al* randomised pregnant women to receive either standard care or a series of Doppler ultrasound blood flow measurements.[3] They found a significantly higher proportion of birth weights in the Doppler group below the 10th and 3rd centiles ($P = 0{\cdot}006$ and $P = 0{\cdot}02$). Birth weight was not the primary outcome variable for the trial. These were only two of many comparisons and one suspects that there might be some spurious significant differences among so many. At least 35 tests were reported in the paper. These tests are not independent because they are all on the same subjects, using variables which may not be independent. The proportions of birth weights below the 10th and 3rd centiles are clearly not independent, for example. The probability that two correlated variables both give non-significant differences when the null hypothesis is true is now greater than $(1-\alpha)^2$, because if the first test is not significant the second has a probability greater than $1-\alpha$ of also being not significant. A P value less than α for any variable, or $kP < 0{\cdot}05$, would still mean that the treatments were significantly different. The overall P value is actually smaller than the nominal 0·05—by an unknown amount which depends on the lack of independence between the tests. The power of the test, its ability to detect true differences in the population, is correspondingly diminished. In statistical terms, the test is conservative.

For the example, we have $\alpha = 0{\cdot}05/35 = 0{\cdot}0014$, and so by the Bonferroni criterion the treatment groups are not significantly different. Alternatively, the P values could be adjusted to give $35 \times 0{\cdot}006 = 0{\cdot}21$ and $35 \times 0{\cdot}02 = 0{\cdot}70$.

Other multiple testing problems arise when we have more than two groups of subjects and wish to compare each pair of groups; when we have a series of observations over time, such as blood pressure every 15 minutes after administration of a drug, where there may be a temptation to test each time point separately; and when we have relations between many variables to examine, as in a survey. For all these problems the multiple tests are highly correlated and the Bonferroni method is inappropriate, as it will be highly conservative and may miss real differences. We shall deal with these types of analysis in separate notes.

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE
J Martin Bland, *reader in medical statistics*

Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX
Douglas G Altman, *head*

BMJ 1995;310:170

1 Lee KL, McNeer JF, Starmer FC, Harris PJ, Rosati RA. Clinical judgements and statistics: lessons from a simulated randomized trial in coronary artery disease. *Circulation* 1980;61:508-15.
2 Williams EI, Greenwell J, Groom LM. The care of people over 75 years old after discharge from hospital: an evaluation of timetabled visiting by health visitor assistants. *J Pub Hlth Med* 1992;14:138-44.
3 Newnham JP, Evans SF, Con AM, Stanley F J, Landau LI. Effects of frequent ultrasound during pregnancy: a randomized controlled trial. *Lancet* 1993;342: 887-91.