EBMH Notebook

p Value fetishism and use of the Bonferroni adjustment

The p value guides us in judgements of statistical significance, accepting or rejecting the "null hypothesis", but can exert excessive influence over judgements of clinical significance. Karl Marx wrote of the "fetishism of commodities" in criticising the undue influence of property on capitalism's social processes. In much the same way, evidence-based medicine has created a "fetishism of p values". Where used to make an appraisal of the significance of difference between two groups, the p value is termed the "significance level". The convention of relying on $p<0.05$ to indicate clinical significance is now deeply embedded in our critical appraisal of research papers, to such an extent that we can sometimes forget its true meaning.

A p value less than 0.05 indicates that the differences between, say, two treatment groups will result from "chance" one time in 20, as a Type I error (signified α). This is conventionally interpreted as indicating clinical significance, useful as an aide memoire but sometimes given unwarranted influence when conceptually detached from the clinical situation in which the test has been conducted. p values must be interpreted in light of the clinical scenario. We misuse p values when treating them in a purely dichotomous manner, as "significant" or "not significant". The difference between $p = 0.051$ and $p = 0.049$ may be actuarially meaningless, and can only be understood with reference to the circumstances and power of the test.

BONFERRONI ADJUSTMENT

When we carry out repeated independent tests on a single study group, the probability of finding a significant difference is artificially inflated, and the error rate can be calculated on the basis of the number, n, of these independent tests, using the formula:

$$1-(1-\alpha)^n$$

It has therefore become standard practice to employ a statistical adjustment, known as the Bonferroni adjustment, to counteract the effect of multiple tests.[1] This employs the reverse formula to adjust the significance level and maintain an error rate of 0.05:

$$1-(1-\alpha)^{1/n}$$

Bonferroni adjustments are one of many different types of statistical adjustment.[2] In the hands of well-informed statisticians, making such adjustments can sometimes be useful. But the trend has been to employ adjustments injudiciously. There are several arguments against Bonferroni adjustments in particular.[3]

TRADING TYPE I FOR TYPE II ERRORS

The Bonferroni adjustment is employed to minimise Type I errors, but will only do so by increasing the probability of accepting the null hypothesis when the alternative is true, or Type II error. Sins of omission are no less than sins of commission. In a psychiatric setting if, say, we fail to accept the hypothesis that "domestic violence increases the probability of hospital admission", the Type II error resulting from excessive statistical conservatism is no better than its opposite. Bonferroni adjustments are not a sign of judicious statistical caution, but simply a method of reducing Type I errors and increasing Type II errors.

WHAT GOES INTO THE POT?

In determining the number ("n") of independent tests, it is common to ascertain the number of independent tests in the published paper. But this neglects the hidden layers of tests employed in addition or in preparation for that paper. Statistical tests may have been carried out, but not published. Tests may have been conducted in studies before the publication in question, but essential to that publication. To be theoretically "pure", we should consider error rates beyond the immediate study. The Bonferroni adjustment applies a veneer of authenticity.

WHAT IS THE CORRECT "NULL HYPOTHESIS"?

The Bonferroni adjustment attempts to remove the study-wide error rate across a wide range of independent tests. If significance is detected, and the "null hypothesis" seemingly rejected, this leaves us ignorant as to which of the individual tests are significant and which are not. We can only conclude that the "universal" null hypothesis is rejected. But the universal null hypothesis—that some of the individual tests are significant in some way—is of no clinical interest. Clinical interest is served by knowing which test is significant and in what way.

ARE BONFERRONI ADJUSTMENTS EVER JUSTIFIED?

Bonferroni adjustments were developed by Neyman and Pearson[4] in the 1920s as a means of enhancing decisions in recurring and repetitive circumstances. As a means of improving decision-making through statistical inference, Bonferroni adjustments have a role to play. They can be misapplied in biomedical research, which is dealing with a distinct paradigm. Bonferroni adjustments only have relevance to us where the universal null hypothesis is of greater interest than individual hypotheses concerning individual independent tests. This is particularly true to hypothesis-generating, rather than hypothesis-testing, research. Even in that instance, there is a need to tether interpretation to clinically relevant information, and to consider the implication for Type I as much as Type II error rates.

There are alternative multiple test procedures to the Bonferroni method[2] which overcome some, but not all, of the difficulties described above. These include the Holm method[5] which, however, is inaccessible to most biomedical researchers. Judgements on which method, if any, should be employed to adjust for multiple statistical tests need to be made judiciously, with knowledge of what p values actually represent.

CONCLUSION

p values communicate a specific meaning about probability that should be appraised clinically. p value fetishism can distract from that underlying meaning. $p<0.05$ is a useful aide memoire, but we should know its limitations. It is preferable to report estimated difference with a confidence interval, and not just as a p value. This fetishism finds its greatest expression in the Bonferroni adjustment for repeated tests. Bonferroni adjustments were created to inform decision

making, and have sometimes been misapplied to biomedical research. Alternatives to Bonferroni adjustments are available but, when adjustments have been made, it is desirable to report both adjusted and non-adjusted analyses.

JOHN F MORGAN
Consultant Psychiatrist, Yorkshire Centre for Eating Disorders, Newsam Centre, Seacroft Hospital, Leeds LS14 6WB, UK; john.morgan@leedsmh.nhs.uk

REFERENCES

1 Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 1977;**198**:679–84.
2 Bauer P. Multiple testing in clinical trials. *Stat Med* 1991;**10**:871–90.
3 Thomas V Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998;**316**:1236–8.
4 Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* 1928;**20A**:263–97.
5 Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am J Public Health* 1996;**86**:726–8.

Experimental interventions with sex offenders: a brief review of their efficacy

Sexual offending is an area which generates considerable public concern and which has received much attention, with a marked growth in a range of experimental interventions designed to reduce the risk of re-offending among participants. A number of significant challenges arise from this including the polythetic nature of the category ''sex offender''. Any two people identified as sex offenders may have different and non-overlapping patterns of behaviour. Additionally most research and practice has focussed on those detected and convicted of sexual offences. Rates of reporting, detection and conviction in this area are generally very low, which suggests the existence of a large pool of undetected offenders. The extent to which this group differs from convicted groups is rarely acknowledged, yet an understanding of this is likely to be fundamental to efforts to prevent such offending and improve public protection.[1 2]

There have been two Cochrane Collaboration reviews of interventions with known sex offenders.[3 4] Strikingly both reviews found no high quality randomised studies. The great majority of the research was excluded from review and was characterised as small scale and non-randomised, having high exclusion rates and being of relatively poor standard. Both reviews found only a small number of randomised studies, all of which had important weaknesses in terms of the description of randomisation and efforts to ensure blinding in assessment.

Currently, cognitive behavioural group work is fashionable in the treatment of convicted sex offenders. The use of this approach, based on relapse prevention principles, was compared with a no treatment group by Marques *et al*.[5] The mean duration of follow-up for this study was three years. No difference was found between the two groups in terms of rates of sexual offending (OR 0.76, 95% CI 0.26 to 2.28). The treatment group showed lower rates for non-sexual violent offences (OR 0.3, 95% CI 0.1 to 0.89; NNT 10, 95% CI 5 to 85) and also showed lower rates of recidivism when violent and sexual offences were combined (OR 0.14, 95% CI 0.02 to 0.98; NNT 20, 95% CI 10 to 437).

A later and more methodologically sophisticated study was reported by Marques *et al*.[6] A mixed group of 704 rapists and child molesters were randomly allocated to cognitive behavioural relapse prevention treatment, a volunteer control group and a non-volunteer control group, with five-year follow-up. The treatment group underwent an intensive two-year inpatient treatment programme which was manualised to ensure treatment consistency, followed by one year's community support.

The treatment group showed no reduction in sexual offending (χ^2 (2, n = 704) = 0.28 p = 0.870) or non-sexual violent offending (χ^2 (2, n = 704) = 0.66 p = 0.719). Survival analyses were conducted and similar patterns were reported across all three groups.[6] The survival analysis for early treatment dropouts was distinct and this subgroup showed markedly poorer outcomes; a finding of particular significance given the propensity of earlier studies to use such groups for comparison purposes.

Going on to look at severity of re-offending a trend for the volunteer control group to commit more severe offences was noted in terms of sexual penetration (χ^2 (2, n = 178) = 6.48 p = 0.039) and victim injury (χ^2 (2, n = 155) = 7.51 p = 0.023). However, this effect disappeared when statistically corrected for repeated comparisons. Similar findings were reported when the groups were retrospectively banded, on the basis of a number of static risk factors, into high- medium- and low-risk offenders.

In summary Marques *et al*[6] reported no treatment effect and went on to explore a number of explanations for this finding, concluding that in the case of sex offenders understanding how and when treatment is effective is still lacking. Curiously they did not directly address the question of whether psychological interventions have a positive treatment effect.

More encouraging conclusions have derived from meta-analytic studies based on the combination of random allocation and non-randomised studies. Small but statistically significant positive treatment effects have been reported.[7] Here overall re-offending rates based on pooled data were reported. For treated groups this was 12.3% compared to 16.8% for untreated sex offenders. It was noted that cognitive behavioural therapy or systemic approaches produced reductions in recidivism of between 17.4% and 9.9%. A review of this study by Rice and Harris[8] criticised a number of aspects of the methodology. Based on a detailed analysis, Rice and Harris[8] concluded that the methodologies in most of the studies included by Hanson *et al*[7] were inadequate and, as such, were too weak to support the conclusions drawn.

A later and more methodologically sophisticated meta-analysis by Lösel and Schmucker[9] noted the problems with and varied results obtained from earlier reviews in this area. In reviewing 2039 citations they identified 66 studies that met basic scientific criteria for inclusion, which allowed for 80 comparisons. These were rated on an adaptation of the Maryland Scale of Scientific Rigour.[10] Sample sizes varied from 15 to 2557 (median 118), with around a third of studies having samples of 50 or less. Of these, seven studies used randomised designs with six meeting the requirements for level 5 on the Maryland scale. It was notable that 60% of studies only reached level 2 on this scale.