J. R. Statist. Soc. A (1995)
158, *Part* 1, *pp.* 175–177

The Graphical Presentation of a Collection of Means

By HARVEY GOLDSTEIN† and MICHAEL J. R. HEALY

Institute of Education, London, UK

[Received July 1993]

SUMMARY

When a study produces estimates for many units or categories a common problem is that end-users will wish to make their own comparisons among a subset of these units. This problem will occur, for example, when estimates of school performance are produced for all schools. The paper proposes a procedure, based on the graphical presentation of confidence intervals, which enables such comparisons to be carried out while maintaining an average required type I error rate.

Keywords: CONFIDENCE INTERVAL; SCHOOL EFFECTIVENESS; SIGNIFICANCE TEST; TYPE I ERROR AVERAGED CONFIDENCE INTERVALS

1. INTRODUCTION

When the means of two independent samples are to be presented graphically, it is a common practice to accompany the two points by error bars giving the 95% confidence intervals for each mean. As a visual aid, these bars are not very effective in assessing the statistical significance of the quantity of interest, which is the difference between the means. It is a common statistical misconception to suppose that two quantities whose 95% confidence intervals just fail to overlap are significantly different at the 5% level. Clearly, however, it is possible to adjust the confidence level so that the required significance level is achieved by the non-overlap criterion. With equal known standard errors, and assuming normality, the width of the intervals to achieve a 5% significance level should be $\pm 1.39\sigma$.

The problem is more acute and difficult when several means are to be presented from a large study which is of interest to a variety of consumers. The results reported are likely to be used by different individuals for their own purposes and any two out of the set of means may need to be compared. This can occur, for instance, in the publication of results from population surveys, where estimates of a characteristic for each geographical unit are available. In the simplest case each individual will be interested only in a single comparison: in this situation multiple-comparison considerations do not arise. We are concerned to provide a simple presentation which will allow the results of a statistical analysis to be properly appreciated by a reader with little statistical sophistication.

Our proposal is that the means presented graphically should be accompanied by error bars corresponding to confidence intervals at a level β, drawn so that the non-overlap significance level averaged over all possible pairs is equal to the required value.

†*Address for correspondence*: Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, UK.
E-mail: hgoldstn@ioe.ac.uk

 0035-9238/95/158175

2. PROCEDURE

Suppose that there are n independently normally distributed estimates, one for each sample or category, with known standard error. The procedure that we propose is to present the estimates together with confidence intervals and to recommend a judgment of statistical significance when the relevant confidence intervals do not overlap and we need to know the level of significance which this procedure implies. We consider first the basic case where only two categories are to be compared.

Suppose that the sample means of categories i and j are m_i and m_j, independently distributed with standard errors σ_i and σ_j supposed known. Assuming normality, the confidence intervals at level β do not overlap if

$$|m_i - m_j| > z_\beta(\sigma_i + \sigma_j) \tag{1}$$

where z_β is the (positive) normal deviate with a two-tailed probability β. If we write

$$\text{var}(m_i - m_j) = \sigma_i^2 + \sigma_j^2 = \sigma_{ij}^2$$

say, then the probability that inequality (1) occurs is given by

$$\gamma_{ij} = 2[1 - \Phi\{z_\beta(\sigma_i + \sigma_j)/\sigma_{ij}\}] \tag{2}$$

where Φ is the normal integral. This varies between $2(1 - \Phi\{z_\beta\})$ and $2(1 - \Phi\{z_\beta\sqrt{2}\})$ according to the ratio σ_i/σ_j and is a minimum when this ratio is 1.

Where there are more than two categories we propose that β should be selected so that the average value of γ_{ij} over all (i, j) is a predetermined value, say α, typically 0.05 or 0.01. For a given data set this can be determined by a straightforward search procedure. A starting point for z_β is the average of $z_\alpha\sigma_{ij}/(\sigma_i + \sigma_j)$ taken over all the pairs (i, j). The confidence interval for the ith category is then given by $m_i \pm z_\beta\sigma_i$.

3. EXAMPLE

We use as an illustration some public examination results for 16-year-olds from an analysis described in Goldstein *et al.* (1993). The analysis used a two-level variance components model to derive 'value-added' estimates of 'school effectiveness' after adjusting for intake achievements of the examination students at the age of 11 years just before entry to secondary school. The standard errors for the effectiveness estimates, i.e. level 2 residuals for the 64 schools in the sample, were estimated and inputted to the procedure for a 5% average type I error for pairwise comparisons. The calculated value of z_β was 1.396. This is close to the minimum when all the standard errors are equal and reflects the fact that the range of standard errors is only from 0.05 to 0.10.

Fig. 1 shows the set of 64 confidence intervals ordered on the m_i.

It is clear from Fig. 1 that there is considerable overlap of intervals, so that only relatively widely separated schools can be judged as having significantly different means. It should also be remembered that in practice many individuals will wish to make comparisons between three or more schools, leading to substantially wider intervals. In addition, no account has been taken here of the imprecision of the standard error estimates. Thus the set of intervals in Fig. 1 represents the most optimistic picture in terms of the type I error.

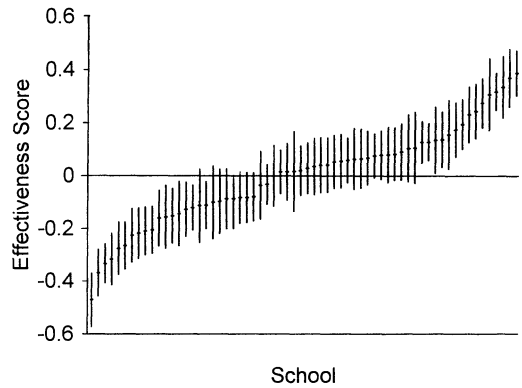

Fig. 1. Effectiveness scores for 64 schools after adjusting for intake achievement

4. GENERALIZATIONS AND CONCLUSIONS

The procedure can be generalized in several ways.

First we can attach weights to each pairwise comparison, e.g. to reflect the probability of the comparison being used. In this case we require the average of $w_{ij}\gamma_{ij}$ to be α, with

$$\sum_{i<j} w_{ij} = 1 \tag{3}$$

where the w_{ij} are the chosen weights.

Secondly, individual users may wish to make several comparisons at a time. For example, a particular school may serve as a 'control' and others compared with it by using an appropriate multiple-comparisons procedure. Or we may wish to compare all pairs of a set of schools, chosen for example within a well-defined locality. For these situations a suitable multiple-comparisons procedure will be required.

If we can anticipate where such uses will occur, or at least can obtain a reasonable estimate of the relevant probabilities of occurrence, then the above procedure can be modified readily. For any particular set of comparisons, the confidence intervals can be constructed for a chosen level β. We then carry out the weighted version of the procedure where the weights are chosen over the set of all defined comparisons. A somewhat more complex search procedure can now be implemented.

Although our discussion has been in terms of normality assumptions, it is readily adapted to other distributional assumptions, such as that of a t-distribution and to statistics other than the mean, e.g. to odds ratios resulting from linear-logistic models.

REFERENCES

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. L. and Thomas, S. (1993) A multilevel analysis of school examination results. *Oxf. Rev. Educ.*, **19**, 425–433.