3. OTHER TYPES OF SELECTION

Sometimes, usually in studies testing a small number of hypotheses, there is an interest in CIs for the accepted hypotheses, or separately for the accepted and rejected hypotheses. There has been a concern, especially in the social sciences, with the inadequate power of most research, stemming from the work of Cohen (1962), who was the first to estimate the typical power of psychological research studies. In these cases, often the magnitude of a departure from the null value is not of special interest, as long as the null is rejected, and CIs are calculated for accepted hypotheses to give an indication of the range of plausible parameter values, in view of the presumed low power due to practical constraints. Given acceptance, parameter values close to the null are more likely to be included in the interval than when the null hypothesis is rejected, and the CIs have conditional coverage probabilities greater than the nominal probability $1 - \alpha$ or noncoverage probabilities less than α, whereas for parameter values far from the null, the conditional confidence coverage probabilities approach 0 and noncoverage probabilities approach 1. For $m = 1$, with the test statistic distributed $N(\theta, 1)$ and for $\theta = 0$, .5, 1, 2, 4, 5, and 6, when CIs are calculated only if the hypothesis is accepted, the conditional noncoverage probabilities are 0, .02, .03, .05, 1, 1, and 1, whereas the marginal noncoverage probabilities are 0, .02, .02, .02, .02, 0, and 0. These values suggest that the FCR approach may be less useful for parameters selected when hypotheses are not rejected, because the CIs may be too wide for useful inferences.

4. CONCLUSIONS

Although it is a good idea for researchers to be aware of problems with conditional coverage of CIs, there is not much that can be done to address them. Because the true values of the relevant parameters are unknown, there is no way of adjusting for the conditional coverage probabilities of the associated CIs, given parameters selected on the basis of the data. What BY show is that the joint probability of some parameters being selected and parameter noncoverage rates can be controlled at a level smaller than a specified α, for independent tests and some types of positively dependent tests, regardless of the selection method used. A simple method of guaranteeing this maximum noncoverage probability, the FCR, is to test the selected hypotheses at level $R\alpha/m$, where R is the number selected and m is the total number, although improvements are possible using adaptive methods. This approach is useful when there are large numbers of hypotheses and many hypotheses are expected to be false, with CIs desired for rejected hypotheses. The article makes a valuable contribution to analysis in such situations. The methodology appears to be less useful with small numbers of hypotheses and in studies with low power to reject any hypotheses.

ADDITIONAL REFERENCES

Black, M. A. (2004), "A Note on the Adaptive Control of False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 66, 297–304.

Cohen, J. (1962), "The Statistical Power of Abnormal Social Psychological Research: A Review," *Journal of Abnormal and Social Psychology*, 65, 145–153.

Cox, D. R., and Wong, M. Y. (2004), "A Simple Procedure for the Selection of Significant Effects," *Journal of the Royal Statistical Society*, Ser. B, 66, 395–400.

Olshen, R. A. (1973), "The Conditional Level of the *F*-Test," *Journal of the American Statistical Association*, 68, 692–698.

Scheffé, H. (1977), "A Note on a Reformulation of the *S*-Method of Multiple Comparison," *Journal of the American Statistical Association*, 72, 143–146.

Comment

Ajit C. TAMHANE

I congratulate the authors for providing a solution to the vexing problem of constructing multiple confidence intervals (CIs) with controlled error rate for parameters selected by a multiple-testing procedure. There are a number of new important ideas in the article, a thorough discussion of which would require much additional work. I am sure that there will be many follow-up articles that will explore these ideas in detail; here I restrict my comments to only a few basic points.

The authors begin by demonstrating that unadjusted and Bonferroni-adjusted procedures do not ensure prescribed conditional coverage probability if CIs are computed only for those means for which the null hypothesis that the mean equals 0 is rejected (the so-called "discoveries"). For each such discovery, the set of "acceptable" values of the mean is used as its CI, which is therefore dual to the corresponding significance test; in particular, it excludes 0. This obviously makes the conditional coverage probability equal to 0 when the null hypothesis holds. For small nonzero means, the conditional coverage probability still falls below the nominal confidence level. One reason for this phenomenon is that the estimates of the selected means are highly biased (except when the true mean is 0, in which case the estimate is unbiased). As a result, the intervals are incorrectly centered at these biased estimates. Would it be possible to use shrinkage estimates instead, although the resulting intervals will not be duals of the corresponding significance tests?

To give an idea of the bias involved in selected means, consider independent $T_j \sim N(\theta_j, 1)$, $j = 1, 2, \ldots, m$. A "nominal" $(1 - \alpha)$ marginal or simultaneous CI, $T_j \pm c$, for θ_j is computed conditional on an α-level test of $\theta_j = 0$ rejecting when

Ajit C. Tamhane is Professor and Chairman, Department of Industrial Engineering and Management Sciences (IE/MS) and Professor of Statistics, Northwestern University, Evanston, IL 60208 (E-mail: *ajit@iems.northwestern.edu*). The author thanks Dingxi Qiu, a graduate student in the IE/MS Department, for providing computational help and useful comments.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001925

Table 1. Bias in T_j Conditional on $|T_j| > c$ for $\alpha = .05$

θ	Unadjusted test	Bonferroni-adjusted test
.5	1.4927	3.2798
1.0	1.4503	2.9680
2.0	.7722	2.0778
4.0	.0509	.5961

NOTE: The unadjusted test uses $c = Z_{.975} = 1.96$, whereas the Bonferroni-adjusted test uses $c = Z_{.999875} = 3.6623$.

$|T_j| > c$. Here $c = Z_{1-\alpha/2}$ for an unadjusted test coupled with a marginal CI and $c = Z_{1-\alpha/2m}$ for the Bonferroni-adjusted test coupled with a simultaneous CI. Assume that $\theta_j = \theta$ for all $j = 1, 2, \ldots, m$. It is easily shown that the conditional expectation of T_j, conditioned on $|T_j| > c$, is given by

$$E\big(T_j \big| |T_j| > c\big) = \frac{\theta - \int_{-c}^{c} t\phi(t-\theta)\,dt}{\Phi(\theta - c) - \Phi(-\theta - c)}$$
$$= \theta + \frac{\phi(\theta - c) - \phi(-\theta - c)}{\Phi(\theta - c) + \Phi(-\theta - c)},$$

where ϕ and Φ are the pdf and cdf of the standard normal distribution. The second term gives the bias, which has the same sign as θ. Table 1 gives the bias values for selected θ for both unadjusted and Bonferroni-adjusted procedures when $\alpha = .05$ and $m = 200$. We see that the bias is quite large for small values of θ and decreases with θ.

Some readers may be confused, as indeed I was, by the fact that the estimated FCRs for the unadjusted procedure in example 3 equal exactly 1 minus the corresponding conditional coverage probabilities from example 1 (in particular, the FCR equals 1 when $\theta = 0$), whereas this relation does not hold (in particular, the FCR does not equal 1, but equals .05 when $\theta = 0$) for the Bonferroni-adjusted procedure in example 4. The reason for this is that the ratio V_{CI}/R_{CI} is defined as 0 when $R_{CI} = 0$; hence the FCR can be expressed as

$$FCR = E\left(\frac{V_{CI}}{R_{CI}}\middle| R_{CI} > 0\right) P(R_{CI} > 0).$$

If $R_{CI} > 0$ when $\theta = 0$, then $V_{CI}/R_{CI} \equiv 1$ for both the unadjusted and Bonferroni-adjusted procedures. Therefore, $FCR = P(R_{CI} > 0)$. For the unadjusted procedure,

$$P(R_{CI} > 0) = 1 - (.95)^{200} \approx 1,$$

and hence $FCR \approx 1$. In contrast, for the Bonferroni-adjusted procedure,

$$P(R_{CI} > 0) = 1 - (.99975)^{200} \approx .05,$$

and hence $FCR \approx .05$.

The foregoing explanation demonstrates that the FCR is controlled for the Bonferroni-adjusted procedure at the .05 level even for $\theta = 0$, because CIs are computed in only 5% of the cases, although all of them miss the true means. To me, this does not provide the necessary security about the accuracy of the CIs, and suggests that the positive FCR,

$$pFCR = E\left(\frac{V_{CI}}{R_{CI}}\middle| R_{CI} > 0\right),$$

may be a more appropriate criterion. I recognize, as the authors note, that the pFCR is equivalent to the conditional coverage probability and cannot be controlled for all parameter values. However, there are other criteria that could be used instead. In summary, I think that the debate on the choice between

$$\frac{FDR}{FCR} \quad \text{versus} \quad \frac{pFDR}{pFCR}$$

is far from over.

As an aside, I note that it is not necessary to estimate the quantities in examples 1–4 by simulation, because the following exact expressions for them can be readily derived. First, the conditional coverage probability is given by

$$P\big(\theta \in [T_j - c, T_j + c] \big| |T_j| > c\big) = \frac{\Phi[\min(c, \theta - c)] - \Phi(-c)}{\Phi(\theta - c) + \Phi(-\theta - c)}.$$

Next, the FCR is given by

$$FCR = E\left(\frac{V_{CI}}{R_{CI}}\middle| R_{CI} > 0\right) P(R_{CI} > 0)$$
$$= P\big(\theta \notin [T_j - c, T_j + c] \big| |T_j| > c\big) \times \left\{1 - [P\{-c \le T_j \le c\}]^m\right\}$$
$$= \left\{1 - \frac{\Phi[\min(c, \theta - c)] - \Phi(-c)}{\Phi(\theta - c) + \Phi(-\theta - c)}\right\} \times \left\{1 - [\Phi(-\theta + c) - \Phi(-\theta - c)]^m\right\}.$$

This last expression holds only when $\theta_j = \theta$ for all $j = 1, 2, \ldots, m$.

In closing, I congratulate the authors once again for a thought-provoking article, and I thank the editor for giving me an opportunity for contributing to its discussion.

Comment

Peter H. WESTFALL

1. INTRODUCTION

Benjamini and Yekutieli (BY) solve important problems in false discovery rate–controlling multiple-comparison procedures (FDRMCPs), thus increasing their utility and applicability. Familywise error rate–controlling multiple-comparison procedures (FWEMCPs) have historically been interval-based

Peter H. Westfall is Horn Professor of Statistics, Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, TX 79409 (E-mail: *peter.westfall@ttu.edu*).

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000001916