Comment

William DuMouchel

1. THE SELECTION MODEL APPROACH TO THE FILE DRAWER PROBLEM

Professors Iyengar and Greenhouse provide a nice introduction to the file drawer problem and then criticize and improve the so-called "fail-safe" method of Rosenthal (1979), as well as its modification in Orwin (1983). In doing so, they provide a strong argument that Orwin's modification, based on averaging effect sizes, is to be preferred to the Rosenthal version, based on computing a joint p-value for a set of studies.

But the major purpose of their paper is to present and advocate another approach, an extension of that of Hedges and Olkin (1985), in which the probability of selection bias for each study is explicitly incorporated into the likelihood function. Iyengar and Greenhouse provide an example meta-analysis of the ten studies summarized in Table 4, to illustrate the use of their method. Using either of two proposed parametric models of selection bias, they jointly estimate the degree of selection bias and the common effect size in the ten studies. The resulting inferences are claimed to be more complete, useful and robust than those from the fail-safe method.

Although I am inclined to grant them those modest claims, I would have liked to have seen a deeper discussion of their own example. On page 13 the authors state "Perhaps the most interesting feature of the log likelihood contours is their width . . . this meta-analysis is not very informative for θ." I disagree with both clauses. To take the latter clause first, a standard error of .05 for θ seems quite an achievement for estimation of any effect size, and I suspect that most social scientists reviewing the data in Table 4, where estimates range from $-.58$ to 1.05, would call it unduly optimistic. Second, that feature of the figures which startled me the most, and which is also reflected in the covariance matrices of Table 5, is that there is hardly any correlation between the estimates of θ and β (or of θ and γ). If we work with the normal approximation to the joint posterior distribution of (θ, β) as shown in Table 5, the correlation between the two parameters is about -0.1. Thus, knowledge that $\beta = .1$ would lead one to predict $\theta = 0.036$, although increasing β by four standard deviations to 2.5 leads to the prediction of $\theta = 0.016$. Large changes in the assumed size of the selection bias lead to trivial changes in the estimate of the parameter of interest. Does this mean that the file drawer problem is not a problem at all?

Looking at their Figure 1 more closely, it can be seen that there is a lot more information in the likelihood function for θ (the contours are more closely spaced in the vertical direction) when $\beta = 2.5$ than when $\beta = 0.1$; does this mean that selection bias helps us estimate θ? One can get some insight into this puzzling behavior by studying equation (6). Because the numerator of the expression for $L(\theta, w)$ factors into a function of θ alone and a function of w (i.e., β) alone, all the dependence between the two parameters in the joint likelihood must come from the term in the denominator of (6). Now this term does not even depend on the data $(t_1, \ldots, t_{10})$, but only on the sample sizes N_i and on the assumed forms of $f(t; \theta)$ and of $w(t)$. Thus, inferences drawn about the relationship between θ and β, based on that likeliood function, may depend more on prior assumptions than on the data.

2. THE ASSUMPTION OF HOMOGENEOUS EFFECT SIZES

The authors state that "for illustrative purposes (and following Hedges and Olkin) we assume that all studies are estimating the same effect size." I hereby propose that statisticians never recommend for general use any method of meta-analysis which does not include, somewhere in the model, a parameter or component of variance for between study variability. This admittedly extreme proposal would rule out recommendation of the inverse normal method of combining one-tailed significance tests, and thus the Rosenthal fail-safe sample size procedure, because, contrary to the implication of the authors in Section 3, one needs to assume that *every* effect size is zero, not just that the mean effect size is zero, in order to have each p_i uniformly distributed on [0, 1]. The Orwin procedure based on average effect sizes does not seem to need this assumption, providing another reason to prefer it to Rosenthal's version.

Figure 1 graphs the ten estimates from Table 4 with approximate 95% error bars based on plus or minus two standard errors. A glance at Figure 1 makes it

William DuMouchel is Chief Statistician at BBN Software Products Corporation, 10 Fawcett Street, Cambridge, Massachusetts 02238.