Statistical Science
1988, Vol. 3, No. 1, 109–135

Selection Models and the File Drawer Problem

Satish Iyengar and Joel B. Greenhouse

Abstract. Meta-analysis consists of quantitative methods for combining evidence from different studies about a particular issue. A frequent criticism of meta-analysis is that it may be based on a biased sample of all studies that were done. In this paper, we use selection models, or weighted distributions, to deal with one source of bias, namely, the failure to report studies that do not yield statistically significant results. We apply selection models to two approaches that have been suggested for correcting the bias. The fail-safe sample size approach calculates the minimum number of unpublished studies showing nonsignificant results that must have been carried out in order to overturn the conclusion reached from the published studies. The maximum likelihood approach uses a weighted distribution to model the selection bias in the generation of the data and estimates various parameters of interest. We suggest the use of families of weight functions to model plausible biasing mechanisms to study the sensitivity of inferences about effect sizes. By using an example, we show that the maximum likelihood approach has several advantages over the fail-safe sample size approach.

Key words and phrases: Meta-analysis, file drawer problem, selection bias, weighted distributions, maximum likelihood estimation.

1. INTRODUCTION

The application of statistical procedures to collections of results from individual studies for integrating, synthesizing and advancing a research domain is commonly known as meta-analysis. The objective of a meta-analysis is to summarize quantitatively a research literature with respect to a particular question and to examine systematically the manner in which a collection of studies contributes to knowledge about that question. An early and well known example of a method for synthesizing evidence from independent studies is Fisher's method (1932) for combining p-values. Recently, interest in the application of meta-analysis as a primary research tool has grown considerably and with it a corresponding interest in related statistical and methodological problems. Books by Glass, McGaw and Smith (1981), Cooper (1984), Light and Pillemer (1984), Rosenthal (1984) and Wolf (1986) provide extensive discussions on the various aspects of the practice of meta-analysis. In addition, Hedges and Olkin (1985) address the statistical issues involved in integrating independent studies. Applications of meta-analysis as a tool for investigating scientific questions and for guiding public policy decisions are diverse and include, for example, the analysis of the efficacy of psychotherapy (Smith, Glass and Miller, 1980), the assessment of human lung cancer risks from various environmental emissions (DuMouchel and Harris, 1983), and the United States Department of Education study on school desegregation and black achievement (1984).

In practice, the major steps in doing a meta-analysis include identifying, reviewing, abstracting and synthesizing studies from the literature. There are many important issues and difficulties associated with each of these steps, many of which are discussed in the references cited above. In this paper we focus on an issue related to the identification of articles to be included in a meta-analysis. Specifically, a concern in meta-analysis is that studies included in a quantitative literature review may be a biased selection of all studies that were done. There are, of course, many

Satish Iyengar is Postdoctoral Fellow and Joel B. Greenhouse is Assistant Professor of Statistics, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. Satish Iyengar is on leave from the University of Pittsburgh, where he is an Assistant Professor of Statistics.

possible sources of biased reporting. A frequently cited one is that published research is biased in favor of statistically significant findings. Rosenthal (1979) called this publication bias the "file drawer problem," as he imagined these unreported statistically nonsignificant studies sitting in investigators' file drawers. Surveys by Greenwald (1975) and Chase and Chase (1976) amply demonstrate the presence of this publication bias. Moreover, Hedges and Olkin (1985) note that publication bias arises not only out of editorial policy favoring statistically significant results, but also from the reluctance of investigators to report results when p-values do not reach statistical significance.

It is preferable, of course, to eliminate or minimize the file drawer problem from the outset through changes in the attitudes of editors and investigators toward statistical significance. One approach that has been suggested is to establish guidelines for the publication and reporting of the results of studies. Another is to create registries of research studies to use as a sampling frame from which to retrieve studies to include in a meta-analysis (see Chalmers, Hetherington, Newdick, Mutch, Grant, Enkin, Enkin and Dickersin, 1986; Begg and Berlin, 1987). However, such registries are rare, and until better reporting procedures are instituted, it is necessary to consider other approaches.

Our objective in this paper is to provide a framework for modeling the selection of studies for publication by using weighted distributions. We investigate two approaches for dealing with the file drawer problem. The first, suggested by Rosenthal (1979), assesses the magnitude of the file drawer problem by calculating the minimum number of unpublished studies showing null results there must be to overturn the conclusion reached from a meta-analysis based on published studies. The second approach based on selection models explicitly incorporates the reporting process into the likelihood function through the use of a weight function and uses maximum likelihood to estimate the parameters of the model.

This paper is organized as follows. In Section 2, we describe selection models and their application to the file drawer problem. In Section 3, we define Rosenthal's measure for assessing the magnitude of the file drawer problem, known as the fail-safe sample size, and show that Rosenthal's method for calculating the fail-safe sample size is based upon the assumption that the unpublished studies are in fact a random sample of all studies that were done. We suggest a modification of Rosenthal's method and show the effects of this modification upon the estimated fail-safe sample size. A maximum likelihood approach for summarizing the results of a meta-analysis by using weighted distributions is described by Hedges and Olkin (1985). We give extensions of this approach in Section 4 and apply them to a meta-analysis of the effects of open versus traditional education on student creativity. We conclude the paper with a summary and a discussion of some general issues in meta-analysis.

2. SELECTION MODELS

For ease of exposition, we assume henceforth that we have retrieved from the literature several studies that were done to compare the efficacy of a treatment relative to a control, and that a meta-analysis is based on those studies. A data point in the meta-analysis consists of the results from a single study. We usually work with a summary measure from each study, say x, such as a p-value or an estimate of an effect size that, for instance, may be defined as the standardized difference between a treatment group and a control group. We assume further that if there were no bias in the selection of studies included in the meta-analysis, x would be governed by the probability density function $f(x;\, \theta)$, where θ is some unknown parameter (for example, θ may be the effect size). One objective of a meta-analysis is to obtain an estimate of θ. For example, the usual estimate of the effect size defined above is proportional to a noncentral t variate with noncentrality parameter proportional to θ.

When there is selection bias, however, a more appropriate specification of the probabilistic model is needed. One way to model the bias is through the use of weighted distributions; that is, by using the density proportional to $f(x;\, \theta)w(x)$, where $w(x)$ is a non-negative weight function. We assume that $f(x;\, \theta)w(x)$ is integrable, so that it can be normalized to yield a proper density. For example, if we only obtain data from a subset, A, of the sample space, w would be the indicator function of that set: $w(x) = I_A(x)$. Another important example arises in renewal theory in connection with the waiting time paradox, where $w(x) = x$. Weighted distributions also originate with the work of Fisher (1934), who recognized the importance of ascertainment bias in his genetic investigations. However, it was not until Rao (1965) formalized this idea that weighted distributions, or selection models, came into common use. (Some authors reserve the term *selection model* for the special case of indicator weight functions; we use the term in the broader sense of any appropriate non-negative $w(x)$.) Rao also allowed the weight function to be only partially specified, so that it was parametrized by an unknown, say β. He then suggested the joint estimation of β and θ from the data.

More recently, weighted distributions have been used in many other problems, such as in sample surveys, family studies and geology. A principal contributor to this field is G. P. Patil, who has applied weighted distributions to problems in ecology. He has also derived characterizations of certain classes of weighted distributions that have important implications for other practical applications (see, for instance, Mahfoud and Patil, 1982). Good overviews of this topic are given in Patil and Rao (1977), Rao (1985) and Bayarri and DeGroot (1986c).

One way to characterize the file drawer problem is by the following simple weight function $w(x; \alpha)$, where x is an observed p-value:

$$w(x; \alpha) = I_{[0,\alpha]}(x), \tag{1}$$

for some $\alpha \in [0, 1]$. A random sample from the corresponding weighted distribution is obtained if an investigator does independent replications of an experiment, and only reports the ones that yield results significant at the level α. As a model for the selection of studies for publication, the weight function in (1) is unrealistic because studies showing nonsignificant results do indeed get published. We discuss other choices of weight functions in Section 4. In this paper, we interpret $w(x)$ as the probability of reporting a study when the data takes on the value x. Thus, all our weight functions take values between 0 and 1, so that $f(x; \theta)w(x)$ can always be renormalized to yield a proper density. We also consider families of weight functions: in particular, if there are k studies, we may have weight functions $\{w_j(x; \beta): j = 1, \ldots, k\}$, where w_j applies to the jth study and β is an unknown parameter. Two examples of just such a family are given in Section 4 below. The flexibility afforded by such families of weight functions is important, for it allows us to investigate the sensitivity of the inferences from a meta-analysis when various weight functions are used to model plausible biasing mechanisms (see, for example, Greenhouse, Fromm, Iyengar, Dew, Holland and Kass, 1986).

3. FAIL-SAFE SAMPLE SIZE APPROACH

We start with Rosenthal's approach to assessing the magnitude of the file drawer problem. He computes the number of unpublished studies needed to offset the conclusion reached on the basis of the observed studies. Cooper (1979) called this number the fail-safe sample size. Rosenthal uses the inverse normal method for combining one-tailed significance levels, which is generally attributed to Stouffer (see Mosteller and Bush, 1954). Let $p_1, \ldots, p_k$ be the significance levels from k observed studies. Under the null hypothesis that the mean effect size is zero, each p_i is uniformly distributed on $[0, 1]$. Let Z_i be the standard normal deviate associated with p_i; that is, $Z_i = \Phi^{-1}(1 - p_i)$, where Φ is the standard normal cumulative distribution function. The inverse normal method then computes the overall significance by using $S_k/k^{1/2}$, where $S_k = Z_1 + \cdots + Z_k$. Suppose now that $S_k/k^{1/2} \geq z_\alpha$, where z_α is the critical value for a one-sided level α test of the normal mean. Rosenthal (1979) assumes that the mean Z value for the unpublished studies is zero, and proposes that the fail-safe sample size, $n(0)$, is the solution to

$$S_k/(k + n(0))^{1/2} = z_\alpha. \tag{2}$$

In practice, $n(0)$ is assessed informally by using one's knowledge of the particular field of study. If $n(0)$ is large, then one may argue that it is unlikely that so many unpublished studies exist. Alternatively, if $n(0)$ is small, the result of the combined significance test may well be due to the biased ascertainment of the studies. Now, if there were publication bias in favor of studies with statistically significant findings, then the Z values for the unpublished studies would not be a sample from the standard normal distribution. Instead, they would be selected from the part of the population of studies whose significance levels exceed α, and hence, whose Z values are less than z_α. If ϕ denotes the standard normal density, then this selected variate has a truncated normal density

$$g(x; z_\alpha) = \begin{cases} \dfrac{\phi(x)}{\Phi(z_\alpha)}, & \text{for } x \leq z_\alpha, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The mean value under the density $g(x; z_\alpha)$ is $M(\alpha) = -\phi(z_\alpha)/\Phi(z_\alpha)$. For commonly used values of α, Table 1 gives values of $M(\alpha)$.

Following Rosenthal, if the mean Z value for the unpublished studies is $M(\alpha)$, then the equation for the fail-safe sample size, $n(\alpha)$, becomes

$$(S_k + n(\alpha)M(\alpha))/(k + n(\alpha))^{1/2} = z_\alpha. \tag{4}$$

It is easy to see that equation (2) is a special case of equation (4), indicating that Rosenthal's calculation of the fail-safe sample size assumes that there is no bias among the unpublished studies. Also, equation (4) always yields a smaller estimate of the fail-safe sample size than does equation (2). Notice that for the given values of α, the values of $M(\alpha)$ in Table 1 are not very different from zero. However, their effect upon the estimate of the fail-safe sample size can be quite large. This can be seen by considering the dependence of the ratio $n(\alpha)/k$ upon the overall standard normal deviate $S_k/k^{1/2}$. Simple algebra shows that for

TABLE 1
Mean value for normal density truncated at z_α

a	z_a	$M(a)$
0.01	2.326	−0.0269
0.05	1.645	−0.1085
0.10	1.282	−0.1949

Rosenthal's method, $n(\alpha)/k$ is a quadratic function of $S_k/k^{1/2}$, whereas for the modification given in equation (4), $n(\alpha)/k$ is approximately linear.

Table 2 provides a comparison of the fail-safe n values calculated from Rosenthal's method and from equation (4). Both examples in Table 2 are taken from Rosenthal (1979) and come from his studies of the effects of interpersonal expectations. Both examples show that this method is quite unstable with respect to the choice of the mean Z value of the unpublished studies: the two estimates can differ by an order of magnitude. In many applications, (4) is probably a better approximation than (2), because the unpublished studies showing significant results probably number far less than the nominal 5% of all unpublished studies.

Hedges and Olkin (1985) list other methods of combining significance levels. A well known one is Fisher's method, which computes $S_k = -\sum \log(p_i)$, and refers S_k to a χ^2 distribution with $2k$ degrees of freedom (d.f.). Our modification described above can also be applied to these methods. For instance, for Fisher's method (3) would be replaced by a truncated exponential distribution, which is the null distribution of $-\log(p)$.

Orwin (1983) proposed a procedure analogous to Rosenthal's, but which is based upon effect size estimates. Recall that an effect size may be defined as the standardized difference between a treatment group and a control group. His criterion is the number of unpublished studies showing null results needed to bring an observed effect size estimate, $\bar{d}$, down to some negligible level, d_c. This critical level d_c is presumably determined by subject matter considerations. Denoting the mean effect size of the unpublished studies by μ, his fail-safe sample size, $m(\mu)$, is the solution to

$$(5) \qquad (k\bar{d} + \mu m(\mu))/(k + m(\mu)) = d_c,$$

where k is once again the number of observed studies.

Orwin did not address the problem of how to specify μ. He only said that "a researcher may have reason to believe that the file drawer studies have a non-zero mean effect size" (page 158). An argument similar to the one leading to (4) implies that $\mu = M(\alpha)$ is an appropriate choice when, once again, the statistical significance criterion is the main source of biased reporting. Because the ratio $m(\mu)/k$ depends linearly upon $\bar{d}$ for all μ, the choice of μ does not dramatically affect the estimated fail-safe sample size, as it did in (4). This is seen in Table 3, which compares the fail-safe sample size estimates for $\mu = 0$ and $\mu = M(.05)$. The first two rows of Table 3 come from the meta-analysis of psychotherapy outcomes by Smith, Glass and Miller (1980), whereas the last one comes from Cooper's (1979) meta-analysis of research on gender differences in conforming in face to face situations. It is clear that Orwin's scheme is more stable with respect to the choice of the mean of the unreported studies than Rosenthal's scheme.

TABLE 2
Comparison of fail-safe sample sizes from (2) and (4)

k	S_k	$n(0)$	$n(.05)$
94	95.32	3263	507
345	420.90	65123	3002

TABLE 3
Fail-safe sample sizes from (5)

k	$\bar{d}$	d_c	$m(0)$	$m(M(.05))$
1766	0.85	0.5	1236	1015
1766	0.85	0.2	5740	3720
16	0.28	0.2	6	2

Rosenthal's clever formulation of the file drawer problem, and Orwin's modification of it yield easily computed indices of its magnitude. Rosenthal (1984) has provided a rough guide to help decide what is an unlikely number of studies in the file drawers, but this guide does not seem to be used, due to its ad hoc nature. Instead, in practice the fail-safe sample size is often used subjectively, relying upon one's knowledge of the field in question to assess the magnitude of the file drawer problem. See, for example, Rosenthal and Rubin (1978), Booth-Kewley and Friedman (1987) and Hazelrigg, Cooper and Borduin (1987).

This approach, however, has certain drawbacks that render it less useful for the broader purposes of meta-analysis. First, we have shown above that Rosenthal's original solution is very sensitive to the choice of weight function. Also, this approach relies on a combined statistic (S_k or $\bar{d}$), which assumes that the k studies are replications of the same experiment. Thus, important heterogeneities (due to variations in experimental design, subject pool, etc.) among the studies are ignored. In the next section, we turn to the maximum likelihood approach, which is more flexible, because it can be used to model such heterogeneities in order to provide a better summary of the data from a meta-analysis.

4. MAXIMUM LIKELIHOOD APPROACH

As indicated above, one major aim of meta-analysis is to estimate a treatment or intervention effect from studies that were done. If the collection of studies obtained from the literature represents a selected sample, the estimate of the average effect size will be biased. Instead of dealing with the problem of publication bias by using the fail-safe sample size to assess the magnitude of the effect of the bias, an alternative approach is to model the selection process explicitly by incorporating a weight function into the likelihood function. The maximum likelihood estimate (MLE) of the effect size can then be obtained, as well as its approximate variance by using the observed information.

For the problem of estimating the mean of a normal distribution with known variance, Bayarri and DeGroot (1986a) studied the MLE for $w(x) = I_{(z_\alpha, \infty)}(x)$. For the same problem with unknown variance, Hedges and Olkin (1985) studied the MLE for $w(t) = I_B(t)$, where $B = \{t: |t| \geq t(q, \alpha)\}$, $t(q, \alpha)$ is the critical value for a size α two-sided t test with q d.f. In this case, the observation is, of course, a t statistic, and the weight function acts upon it. Hedges and Olkin give a detailed discussion of the likelihood function and the MLE for the latter case, when a single study is observed. They note that "More complicated censoring schemes are possible. In many practical situations the censoring rule is unknown, so that the model and results described (above) may not be applicable. However, it does provide a framework from which to make modifications" (page 288). Below, we give examples of such modifications.

To fix ideas, we consider an example given in Hedges and Olkin (1985, page 303). The data come from a meta-analysis of ten studies comparing the effects of experimental open classroom education with traditional education on student creativity and are reproduced in Table 4. For illustrative purposes (and following Hedges and Olkin), we assume that all studies are estimating the same effect size, denoted by θ. For the ith study, the second column gives the sample size, N_i in each of the two samples. The third column gives the effect size estimate, $\hat{\theta}_i$, where $\hat{\theta}_i$ is the difference between the means for each education type divided by a pooled estimate of the standard deviation. The fourth and fifth columns give, respectively, the t statistic, t_i, and the corresponding d.f., $q_i = 2N_i - 2$ for each study.

Denote the density of a noncentral t distribution with noncentrality η and q d.f. by $f(t; \eta, q)$. If there were no selection bias in the reporting of the studies in Table 4, then t_i would have density $f(t; (N_i/2)^{1/2}\theta, q_i)$. Assuming that the observed studies are independent, and that the weight function $w(t)$ models the selection bias in reporting a result for which the value, t, of the t statistic is observed, the likelihood function for θ is given by

$$(6) \quad L(\theta, w) = \frac{\prod_{i=1}^{10} f(t_i; (N_i/2)^{1/2}\theta, q_i) w(t_i)}{\prod_{i=1}^{10} A((N_i/2)^{1/2}\theta, w, q_i)},$$

where

$$(7) \quad A(\eta, w, q) = \int_{-\infty}^{\infty} f(t; \eta, q) w(t)\, dt.$$

TABLE 4
Studies of effects of open vs. traditional education on creativity

i	N_i	$\hat{\theta}_i$	t_i	q_i
1	90	−0.583	−3.91	178
2	40	0.535	2.39	78
3	36	0.779	3.31	70
4	20	1.052	3.33	38
5	22	0.563	1.87	42
6	10	0.308	0.69	18
7	10	0.081	0.18	18
8	10	0.598	1.34	18
9	39	−0.178	−0.79	76
10	50	−0.234	−1.17	98

We note that the independence assumption may be violated if, for example, several studies came from the same investigator or laboratory; in such cases we can make appropriate modifications by modeling the dependence explicitly. Four of the ten studies in Table 4 yielded results that were significant at the .05 level, so an appropriate weight function for this data should be non-zero everywhere. To examine different possible selection schemes and to examine the effect of the choice of the weight function upon our inferences about θ, we consider the following two parametric families of weight functions:

$$(8) \quad w_1(x; \beta, q) = \begin{cases} \dfrac{|x|^\beta}{t(q, .05)^\beta}, & \text{if} \quad |x| \leq t(q, .05), \\ 1, & \text{otherwise,} \end{cases}$$

and

$$(9) \quad w_2(x; \gamma, q) = \begin{cases} e^{-\gamma}, & \text{if} \quad |x| \leq t(q, .05), \\ 1, & \text{otherwise.} \end{cases}$$

In both cases, $P(|T_0| \geq t(q, .05)) = 0.05$ where T_0 has a central t distribution with q d.f. For w_1, $\beta \geq 0$, and for w_2, $\gamma \geq 0$. Henceforth, we write $L(\theta, \beta)$ and $L(\theta, \gamma)$ for $L(\theta, w_1)$ and $L(\theta, w_2)$, respectively.

Both families of weight functions imply that all studies showing statistically significant results will be reported. In addition, these families have the following features: when β and γ are zero, the weight functions indicate no selection bias; when β and γ are infinite, the weight functions become the Hedges-Olkin scheme described above, in which only statistically significant results were reported. The two weight function families differ only for nonsignificant results: w_1 says that the reporting probability increases as the outcome approaches statistical significance, whereas w_2 says that the reporting probability is constant for all nonsignificant results.

The maximum likelihood estimates, $(\hat{\theta}, \hat{\beta})$ and $(\hat{\theta}, \hat{\gamma})$, of the vector parameters (θ, β) and (θ, γ), respectively, along with their estimated standard deviations and their estimated covariance matrices (V) are obtained from equations (6) and (7) and presented in Table 5. Computational details are given in the appendix. For (θ, β), we define the relative log likelihood function $r(\theta, \beta) = \log L(\theta, \beta) - \log L(\hat{\theta}, \hat{\beta})$ and similarly define $r(\theta, \gamma)$. Contour plots for $r(\theta, \beta)$ and $r(\theta, \gamma)$ near the MLEs are given in Figures 1 and 2, respectively.

Inspection of Figures 1 and 2 suggests that the log likelihood contours are approximately elliptical in the neighborhoods of the MLEs, supporting the use of the normal approximation for making inferences about the parameters. Hodges (1987) discusses several methods for assessing the adequacy of this approximation. Application of his method further supports the normal approximation; we omit the details here. Notice that the estimates of θ and its standard deviation and the log likelihood contours are virtually the same for the two weight functions. Thus, for these data the inferences about the effect size seem to be insensitive to the choice of weight function. Also, the intervals $\hat{\beta} \pm 2(SD(\hat{\beta}))$ and $\hat{\gamma} \pm 2(SD(\hat{\gamma}))$ do not include zero, suggesting the presence of some selection bias. Perhaps the most interesting feature of the log likelihood contours is their width. The log likelihood surface is not very sharp in the neighborhoods of $(\hat{\theta}, \hat{\beta})$ and $(\hat{\theta}, \hat{\gamma})$, indicating that this meta-analysis based on ten studies is not very informative for the parameters and in particular θ.

We now compare the maximum likelihood estimate of θ with other estimates that are often used in meta-analysis. The unweighted average of all ten effect size estimates is 0.291, with a standard error of 0.162. The sample size weighted average of all effect size estimates is 0.057, with a standard error of 0.080, which is approximately the MLE if we assume no publication bias ($\beta = \gamma = 0$). At the other extreme, there is the estimate based only upon the four studies showing statistically significant results. Hedges and Olkin (1985) provide tables to compute an MLE of θ approximately the sample size weighted average of the individual MLEs. They report that this weighted average is 0.01, which is very close to the actual MLE, 0.011. They do not, however, provide a standard error for their approximate MLE; the details described in the Appendix give a standard error of 0.035 for this case. Large sample 95% confidence intervals for all the estimators of θ include zero.

TABLE 5

MLEs of effect size and weight function parameters from (6) *to* (9)

$(\hat{\theta}, \hat{\beta}) = (0.026, 1.33)$	$(\hat{\theta}, \hat{\gamma}) = (0.022, 2.53)$
$(SD(\hat{\theta}), SD(\hat{\beta})) = (0.052, 0.59)$	$(SD(\hat{\theta}), SD(\hat{\gamma})) = (0.049, 0.65)$
$V(\hat{\theta}, \hat{\beta}) = \begin{bmatrix} 0.003 & -0.003 \\ -0.003 & 0.348 \end{bmatrix}$	$V(\hat{\theta}, \hat{\gamma}) = \begin{bmatrix} 0.002 & 0.000 \\ 0.000 & 0.417 \end{bmatrix}$

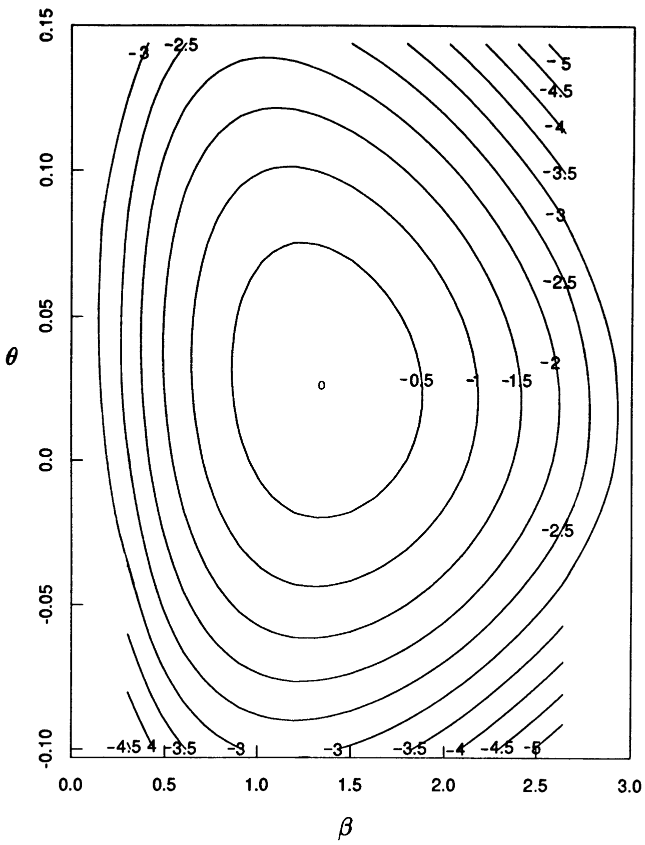

FIG. 1. *Contours of the relative log likelihood function, $r(\theta, \beta)$, for weight function w_1.*

Two features of the comparison above are interesting. First, the MLE for θ based upon all the studies ($\hat{\theta} = 0.022$ or 0.026) is further away from the null value ($\theta = 0$) than the MLE ($\hat{\theta} = 0.011$) derived from just the statistically significant results. In this case, the MLEs come from different models: one model assumes that the weight function is known (that is, either β or γ is known), whereas the other estimates all the parameters. However, this can also happen under the same model when we have a two-sided weight function. For example, if the statistically significant results include both positive and negative values so that the MLE based upon them is near 0, and if the statistically nonsignificant results are predominantly positive, say, then the MLE based upon all the data will be away from zero. An analytic demonstration of this for the noncentral t distribution seems difficult; however, a careful examination of (18) to (21) in the Appendix shows that this can happen for the normal distribution. The other interesting comparison is that the accuracy of the MLE is greater in the presence of publication bias than otherwise. This empirical result is consistent with the analytical work of Bayarri and DeGroot (1986b), who show that in the normal case a selected sample can have greater Fisher information

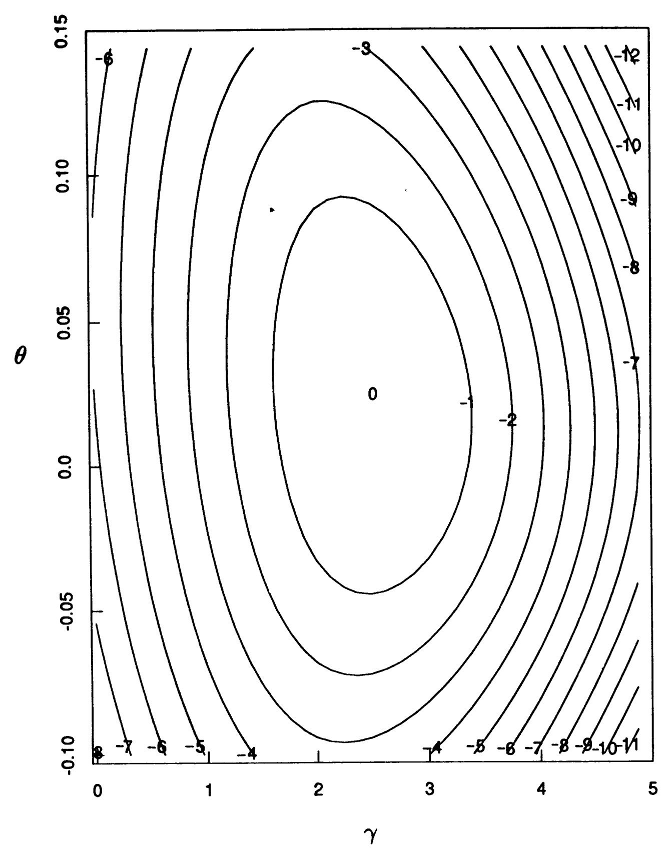

FIG. 2. *Contours of the relative log likelihood function, $r(\theta, \gamma)$, for weight function w_2.*

for certain parameter values than an unselected sample.

5. DISCUSSION

As discussed earlier, the major steps in doing a meta-analysis include identifying, reviewing, abstracting and synthesizing studies from the literature. To date, much of the statistical research in meta-analysis has focused on methods for synthesizing information from studies retrieved from the literature. In this paper, we have considered the problem of the selection of studies to be included in a meta-analysis and the concern that the published research may be biased in favor of statistically significant findings. If we think of the retrieval of studies from the literature as a sampling experiment, then the problem of selection bias in meta-analysis arises because the population of relevant studies is often not well defined. We have suggested an approach for modeling the selection of studies for publication using weighted distributions and maximum likelihood estimation. Our model for selective reporting is necessarily simple, but can be modified to include additional information about the reporting process. We note, however, that the methodology used here is more general and that whenever the various sources of bias can be assessed, and can be modeled by a weight function, or a parametric family of weight functions, the techniques described above can be applied.

Rosenthal's fail-safe sample size approach is an elegant formulation, and it is computationally simple; yet it has several drawbacks that limit its usefulness. This approach initially combines the results of the observed studies as if they constituted an unselected sample, and then provides ad hoc guides for assessing the potential effects of selection bias. This second step becomes unnecessary if we model the selection and appropriately combine the results from the observed studies; the maximum likelihood approach is one way of doing this.

The maximum likelihood approach based on selection models does involve much more computation, but that is justified, we believe, by its many advantages. This approach is flexible: it allows us to see how the parameter estimates and our inferences change as we change our assumptions about the selection model. The accuracy of the parameter estimate is also available, in contrast to the fail-safe sample size method. Also, it is possible to examine the log likelihood surface to see how informative the data are about the parameters of the model. In the example considered, we found that a meta-analysis based on ten studies was not very informative for the population effect size. This raises important questions about design issues in meta-analysis, a topic that has not received much attention.

It is clear that meta-analysis, like rock and roll, is here to stay and that statisticians have made and will continue to make significant contributions to the theory and practice of meta-analysis. Although here we have focused on methods for dealing with the bias due to unpublished studies, others have suggested that "more research on methods for improving the yield of published studies which are actually found would be useful" (Laird, 1986). Presently, this yield typically includes an estimate of effect size. We have used one definition of an effect size, but many other measures have been proposed, and there is much debate about the use and interpretation of the different measures. (Rosenthal, 1984; Laird, 1986). The problem of choosing an appropriate effect size measure is yet to be resolved. Another issue about which we are especially concerned is the problem of using results from studies that differ in quality. As Rosenthal (1984) points out, differing quality may be accommodated within the formalism of meta-analysis by assigning an appropriate weight to each study. Summaries such as an overall effect size, may then be computed. The process of assessing the quality of studies, however, deserves as much attention as the final result. Efforts should be made to articulate the criteria used to distinguish the good quality studies from the bad. Then, methods for integrating the results from heterogeneous studies based on hierarchical models can be applied (see for example, DuMouchel and Harris, 1983; DerSimonian and Laird, 1986; Greenhouse, Fromm, Iyengar, Dew, Holland and Kass, 1986).

APPENDIX

Let T_η have a noncentral t distribution with q d.f. and noncentrality parameter η. Denote the density and distribution functions of T_η by $f(t;\ \eta,\ q)$ and $F(t;\ \eta, q)$, respectively. Then

$$\begin{aligned} &f(t;\eta, q)\\ (10)\quad &= \exp(-\eta^2/2)(\pi)^{-1/2}(1 + t^2/q)^{-(q+1)/2}\\ &\quad\times \sum_{n=1}^{\infty} \frac{b^n\Gamma((n+q+1)/2)}{n!\, q^{1/2}\Gamma(q/2)}, \end{aligned}$$

where

$$(11)\qquad b = \eta t 2^{1/2}/(q + t^2)^{1/2}.$$

We need the noncentral t density to perform the likelihood calculations in Section 5. To do this, we summed the series in (10) when $q \le 80$; and for larger q, we used the following saddle-point approximation due to Resnikoff and Lieberman (1957):

$$\begin{aligned} &f(t;\eta, q)\\ (12)\quad &= C_q\exp(-q\eta^2/2(q + \eta^2))\\ &\quad\times (1 + t^2/q)^{-(q+1)/2}H_q(-t\eta/(q + t^2)^{1/2}), \end{aligned}$$

where

$$(13)\qquad C_q = 2^{-(q-1)/2}\Gamma(q/2)^{-1}(\pi q)^{-1/2},$$

$$\begin{aligned} &H_q(x) = u^q\exp(-(x + u)^2/2)\left(\frac{2\pi u^2}{q + u^2}\right)^{1/2}\\ (14)\quad &\quad\times\left[1 - \frac{3q}{4(q + u)^2} + \frac{5q^2}{6(q + u)^3}\right], \end{aligned}$$

and

$$(15)\qquad u = \frac{-x + (x^2 + 4q)^{1/2}}{2}.$$

The normalizing constants (7) for the weight functions given in (8) and (9) are

$$\begin{aligned} (16)\quad A(\eta;\ w_1, q) &= 1 - (F(c;\ \eta, q) - F(-c;\ \eta, q))\\ &\quad + c\int_{-1}^{1} |x|^\beta f(cx;\ \eta, q)\, dx \end{aligned}$$

and

$$\begin{aligned} &A(\eta;\ w_2, q)\\ (17)\quad &= 1 - (1 - e^{-\gamma})(F(c;\ \eta, q) - F(-c;\ \eta, q)), \end{aligned}$$

respectively, where $P(|T_0| \ge c) = 0.05$. IMSL's routine MDTN gave $F(x;\ \eta, q)$, and the integral in (16) was evaluated by Gaussian quadrature.

To get the covariance matrices in Table 5, we first compute the observed information matrices, $I(\hat\theta,\ \hat\beta)$ and $I(\hat\theta,\ \hat\gamma)$, which are the Hessians of minus the log likelihood evaluated at the MLEs. We then invert them, so that $V = I^{-1}$. We use second central differences to evaluate the second derivatives involved here. The variance of $\hat\theta$ for fixed β or γ is similarly derived.

We now provide some details for the claims made at the end of Section 4. When the underlying density is $\phi(x - \theta)$, and the weight function is

$$(18)\qquad w(x;\ \gamma, a) = e^{-\gamma}I(|x| \le a) + I(|x| \ge a),$$

the normalizing constant is

$$\begin{aligned} &A(\theta, \gamma, a)\\ (19)\quad &= e^{-\gamma} + (1 - e^{-\gamma})(\Phi(-a - \theta) + \Phi(-a + \theta)). \end{aligned}$$

For a sample $(X_1, \cdots, X_k)$ from the weighted distribution $w(x;\ \gamma, a)\phi(x - \theta)/A(\theta, \gamma, a)$, the MLE $\hat\theta(X_1, \cdots, X_k)$, is the unique solution to

$$(20)\qquad \bar X = \hat\theta + B(\hat\theta, \gamma, a),$$

where

$$(21)\qquad B(\theta, \gamma, a) = \frac{d}{d\theta}\log A(\theta, \gamma, a).$$

Examination of $A(\theta, \gamma, a)$ shows that $\hat\theta(X_1, \cdots, X_k)$ is closer to the origin than $\bar X$. Now suppose that $X_1, \cdots, X_j$ are all statistically significant (that is, $|X_i| \ge a$ for $i \le j$), and denote the MLE based upon them by $\hat\theta(X_1, \cdots, X_j)$. Suppose also that $X_{j+1}, \cdots, X_k$ are not significant. It is easy to see that if the mean of the nonsignificant observations is larger in absolute value than the mean of the significant ones, then $\hat\theta(X_1, \cdots, X_k)$ can be farther from zero than $\hat\theta(X_1, \cdots, X_j)$.

ACKNOWLEDGMENTS

We would like to acknowledge our gratitude to Professors Stephen Fienberg and Kenneth Wachter for their gracious invitation to participate in the "Workshop on the Future of Meta-Analysis," October 1986, sponsored by the Committee on National Statistics of the National Academy of Sciences. This work grew out of our participation in that meeting. We would also like to thank Professor Morris DeGroot for many stimulating discussions on weighted distributions and for his encouragement and guidance in completing this paper. This work was supported in part by National Institute of Mental Health Grant MH 15758 at Carnegie-Mellon University, National Institute of Mental Health Clinical Research Center Grant MH

30915 and the John D. and Catherine T. MacArthur Research Network on the Psychobiology of Depression.

REFERENCES

BAYARRI, M. J. and DEGROOT, M. (1986a). Bayesian analysis of selection models. Technical Report 365, Dept. Statistics, Carnegie Mellon Univ.

BAYARRI, M. J. and DEGROOT, M. (1986b). Information in selection models. Technical Report 368, Dept. Statistics, Carnegie Mellon Univ.

BAYARRI, M. J. and DEGROOT, M. (1986c). A Bayesian view of weighted distributions and selection models. Technical Report 375, Dept. Statistics, Carnegie Mellon Univ.

BEGG, C. and BERLIN, J. (1987). Publication bias: A problem in interpreting medical data. Technical Report 490Z, Dept. Biostatistics, Dana-Farber Cancer Institute.

BOOTH-KEWLEY, S. and FRIEDMAN, H. (1987). Psychological predictors of heart disease: A quantitative review. *Psychol. Bull.* **101** 343–362.

CHALMERS, I., HETHERINGTON, J., NEWDICK, M., MUTCH, L., GRANT, A., ENKIN, M., ENKIN, E. and DICKERSIN, K. (1986). The Oxford database of perinatal trials: Developing a register of published reports of controlled trials. *Controlled Clin. Trials* **7** 306–324.

CHASE, L. and CHASE, R. (1976). Statistical power analysis of applied psychological research. *J. Appl. Psychology* **61** 234–237.

COOPER, H. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *J. Personality Social Psych.* **37** 131–146.

COOPER, H. (1984). *The Integrative Research Review: A Systematic Approach.* Sage Press, Beverly Hills, Calif.

DERSIMONIAN, R. and LAIRD, N. (1986). Meta-analysis in clinical trials. *Controlled Clin. Trials* **7** 177–188.

DUMOUCHEL, W. and HARRIS, J. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *J. Amer. Statist. Assoc.* **78** 293–315.

FISHER, R. A. (1932). *Statistical Methods for Research Workers*, 4th ed. Oliver and Boyd, London.

FISHER, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Ann. Eugen.* **6** 13–25.

GLASS, G. V. (1976). Primary, secondary, and meta-analysis of research. *Ed. Res.* **5** 3–8.

GLASS, G. V., MCGAW, B. and SMITH, M. L. (1981) *Meta-Analysis in Social Research.* Sage Press, Beverly Hills, Calif.

GREENHOUSE, J. B., FROMM, D., IYENGAR, S., DEW, M. A., HOLLAND, A. and KASS, R. (1986). The making of a meta-analysis: A case study of a quantitative review of the aphasia treatment literature. Technical Report 379, Dept. Statistics, Carnegie Mellon Univ.

GREENWALD, A. (1975). Consequences of prejudice against the null hypothesis. *Psychol. Bull.* **82** 1–20.

HAZELRIGG, M., COOPER, H. and BORDUIN, C. (1987). Evaluating the effectiveness of family therapies: An integrative review and analysis. *Psychol. Bull.* **101** 428–442.

HEDGES, L. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis.* Academic, Orlando, Fla.

HODGES, J. (1987). Assessing the accuracy of normal approximations. *J. Amer. Statist. Assoc.* **82** 149–154.

LAIRD, N. (1986). Discussion on "The making of a meta-analysis: A case study of a quantitative review of the aphasia treatment literature." Presented at CNSTAT Conference on Meta-Analysis, October 1986.

LIGHT, R. and PILLEMER, D. (1984). *Summing Up: The Science of Reviewing Research.* Harvard Univ. Press, Cambridge, Mass.

MAHFOUD, M. and PATIL, G. P. (1982). On weighted distributions. In *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh, eds.) 479–492. North-Holland, Amsterdam.

MOSTELLER, F. and BUSH, R. (1954). Selected quantitative techniques. In *Handbook of Social Psychology* (G. Lindzey, ed.) **1** 289–334. Addison-Wesley, Cambridge, Mass.

ORWIN, R. (1983). A fail-safe *N* for effect size in meta-analysis. *J. Ed. Statist.* **8** 157–159.

PATIL, G. P. and RAO, C. R. (1977). The weighted distributions: A survey of their applications. In *Applications of Statistics* (P. R. Krishnaiah, ed.) 383–405. North-Holland, Amsterdam.

RAO, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyā Ser. A* **27** 311–324.

RAO, C. R. (1985). Weighted distributions arising out of methods of ascertainment: What population does a sample represent? In *A Celebration of Statistics: The ISI Centenary Volume* (A. C. Atkinson and S. E. Fienberg, eds.) 543–569. Springer, New York.

RESNIKOFF, G. and LIEBERMAN, G. (1957). *Tables of the Non-Central t-Distribution.* Stanford Univ. Press, Stanford, Calif.

ROSENTHAL, R. (1979). The "file drawer problem" and tolerance for null results. *Psychol. Bull.* **86** 638–641.

ROSENTHAL, R. (1984). *Meta-Analytic Procedures for Social Research.* Sage Press, Beverly Hills, Calif.

ROSENTHAL, R. and RUBIN, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral Brain Sci.* **3** 377–415.

SMITH, M., GLASS, G. and MILLER, T. (1980). *The Benefits of Psychotherapy.* Johns Hopkins Univ. Press, Baltimore, Md.

U.S. DEPARTMENT OF EDUCATION (1984). *School Desegregation and Black Achievement.* National Institute of Education, Washington.

WOLF, F. (1986). *Meta-Analysis: Quantitative Methods for Research Synthesis.* Sage Press, Beverly Hills, Calif.