MODELS FOR ESTIMATING THE
NUMBER OF UNPUBLISHED STUDIES

by

L.J. Gleser
I. Olkin

Technical Report No. 313
September 1995

Prepared Under the Auspices
of
National Science Foundation Grant
DMS 93-01366
Ingram Olkin, Project Director

Department of Statistics
Stanford University
Stanford, California
MODELS FOR ESTIMATING THE
NUMBER OF UNPUBLISHED STUDIES

by

L.J. Gleser
I. Olkin

Technical Report No. 313
September 1995

Prepared Under the Auspices
of
National Science Foundation Grant
DMS 93-01366
Ingram Olkin, Project Director

Department of Statistics
Stanford University
Stanford, California
September 10, 1995

Models for Estimating the Number of Unpublished Studies

Leon J. Gleser$^1$ and Ingram Olkin$^2$
University of Pittsburgh and Stanford University

Abstract

The possible existence of unreported studies can cast doubt on the conclusions of a meta-analytic summary of the literature, particularly if there is reason to believe that there is a publication bias against nonsignificant results. The present article proposes two general models that describe how the preponderance of published studies could report significant $p$-values even when testing a null hypothesis that is, in fact, true. Each such model allows one to estimate the number, $N$, of unpublished studies using the $p$-values reported in the published studies; this estimated value of $N$, or related confidence bounds, can then be evaluated for plausibility by the meta-analyst. Use of models of the kind suggested here allows meta-analysts to assess the problem of unpublished studies from various perspectives and thus can lead to greater understanding of, and confidence in, meta-analytic conclusions.

Key words and phrases: meta-analysis; publication bias; selection models; maximum likelihood estimation; one-sided confidence intervals.

$^1$Supported in part by National Science Foundation Grant DMS-9002847 at the University of Pittsburgh.

$^2$Supported in part by National Science Foundation Grant DMS-9301366 at Stanford University, and the Centers for Disease Control and Prevention.

Requests for reprints should be sent to Leon J. Gleser, Department of Mathematics and Statistics, Thackeray Hall, University of Pittsburgh, Pittsburgh, PA 15260.
August 28, 1995

Models for Estimating the Number of Unpublished Studies

Leon J. Gleser and Ingram Olkin
University of Pittsburgh and Stanford University

1 Introduction

A problem common to all meta-analyses is the possible existence of studies bearing on
the question under investigation that were never published nor reported, and thus are
unavailable. Rosenthal (1979) visualized the data from such unreported studies as being
"tucked away in file drawers."

The concern about such unreported studies is that they may contain valid results
which conflict with the evidence summarized in the meta-analysis. If the meta-analysis
deals with the statistical test of a null hypothesis $H$, the reported data may appear to
strongly favor rejection of that hypothesis only because of a "publication bias" against
reporting non-significant results. An extreme view of this possibility, assuming that
"significance" is defined by $p$-values less than or equal to 0.05 and that $H$ is true, is that
the literature contains the 5% of all studies (reported and non-reported) that falsely
rejected $H$ and the file drawers contain the remaining 95% of the studies. If so, one
expects the file drawers to contain $(0.95/0.05) = 19$ times as many studies as have been
reported.

That some publication bias favoring statistically significant results exists has been
extensively documented; see Beggs (1994) and Sterling, Rosenbaum and Weinkam (1995)
and the references cited in these papers for a review of the relevant literature. Sterling,
Rosenbaum and Weinkam find little changed from Sterling's (1959) damning findings
about publication bias in the social sciences. These authors admit that the situation
is somewhat better in the health and physical sciences, although even Sommer's (1987)
relatively optimistic findings regarding studies conducted by the Society of Menstrual
Research revealed that a higher proportion of published studies reported statistically
significant results than did unpublished studies. The possibility of publication bias can
cast serious doubt on the validity of any meta-analytic study. This problem has led
Rosenthal (1979), Orwin (1983), Iyengar and Greenhouse (1988) and others to seek
indices that can measure the extent to which publication bias exists in a given area of
research.

One index of publication bias is the number, $N$, of unreported studies. If this num-
ber can be estimated or determined in some way from the available data, it can then be
compared to one's knowledge about the field. For example, in the extreme case mentioned above where the reported studies contain all of the significant studies and the unreported studies contain all of the nonsignificant studies, $N$ would equal $19k$, where $k$ is the number of reported studies. If $k = 20$, then $N = 380$ and it would be surprising to have this large number of unreported studies. This finding, by itself, would cast doubt on the existence of such an extreme version of publication bias.

Notice that the formula $N = 19k$ arises from a model of how published (reported) studies are selected when the null hypothesis is true. The large value of $N$ obtained when $k = 20$ casts doubt on this model without the need for a detailed look at the $p$-values in the reported studies. A reported $p$-value greater than 0.05 would also cast doubt on this model. The approach advocated in the present paper makes use of models and reported $p$-values to make inferences about the actual number, $N$, of unpublished (unreported) studies. This allows us to use accepted methods of statistical estimation (e.g., maximum likelihood estimation) and to evaluate the accuracy of our estimates using either mean square error or confidence bounds. Equally important, models such as those proposed here can be tested against the data.

2 The Fail-Safe $N$ Approach

Rosenthal (1979) suggested computing the minimum number of (mostly nonsignificant) unreported studies that would be required to overturn rejection of $H$ based on a certain meta-analytic combination of $p$-values from the reported studies. For any number $p$ between 0 and 1, let $z(p)$ be the $100(1 - p)$th percentile of the standard normal distribution.

Let $p_1, \ldots, p_k$ be the $p$-values in the reported studies and $p_{k+1}, \ldots, p_{k+N}$ be the $p$-values in the unreported studies. Let $z_i = z(p_i)$, $i = 1, \ldots, k + N$, $S = \sum_{i=1}^{k} z_i$, $U = \sum_{i=k+1}^{k+N} z_i$. If all $N + k$ $p$-values were available, the method of combining $p$-values assumed by Rosenthal rejects the null hypothesis $H$ at, say, the 5% level of significance if

$$(k + N)^{-1/2}(S + U) \geq z(0.05) = 1.645.$$ 

Rosenthal assumes that the $z_i$'s from the unreported studies sum to 0; that is, $U = 0$. (Actually, it is only necessary to assume that $U \leq 0$.) If the combined test based only on the reported studies rejects the null hypothesis (that is, $k^{-1/2}S \geq 1.645$), then the number $N$ of unreported studies must be at least

$$N_{FS} = (S/1.645)^2 - k$$

in order to overturn this decision. In the meta-analysis literature the quantity $N_{FS}$ is often called the "fail-safe $N".\footnote{In general, if it is assumed that $U \leq u$, then $N_{FS}(u) = ((S + u)/1.645)^2 - k$.}

Rosenthal's (1979) original article made no attempt to specify the mechanisms that might lead to nonreporting of studies, nor to justify his assumption that the sum of the
unreported z-values is (no greater than) zero. His later comments and those of critics of his approach (see, for example, Iyengar and Greenhouse (1988), Begg and Berlin (1988), and the discussions following these articles) have emphasized that the fail-safe N approach cannot be universally applied. At least some specification of the mechanism that consigns studies to file drawers is necessary to justify this method. In any case, the goal of the “fail-safe N” approach differs from the goals of this paper; the “fail-safe N” has no necessary relation to the actual number, N, of unreported studies. It is thus not surprising that values of \(N_{FS}\) usually differ considerably from estimates of \(N\).

3 Outline

In this article, we discuss estimation of \(N\) under two models. Both models make the following assumptions:

\[(3.1)\]

(i) that the \(N + k\) studies, both reported and unreported, are mutually statistically independent;

(ii) that the \(p\)-value in each study is based on a continuous test statistic;

(iii) the null hypothesis \(H\) is true.\(^4\)

Under these assumptions, if there were no bias in the selection of those studies that are reported, the \(N + k\) \(p\)-values would be a random sample from the uniform distribution on the interval [0,1].

We first consider the possibility that the \(p\)-values observed are the \(k\) smallest \(p\)-values among the \(N + k\) reported and unreported studies. This is not a selection model (Iyengar and Greenhouse, 1988); rather, it resembles observational models used in accelerated life-testing, where sampling ceases once \(k\) lifetimes (which are necessarily the smallest lifetimes) are observed. We show how to find the maximum likelihood and best unbiased point estimators of \(N\), and construct a lower 100(1 - \(\alpha\))\% confidence bound for \(N\) assuming that \(H\) is true. A more realistic modification of this model is also treated, in which we observe the \(m\) smallest \(p\)-values plus a random sample of \(k - m\) of the \(N + k - m\) remaining \(p\)-values.

The second type of model considered is a true selection model in which, following Iyengar and Greenhouse (1988), Bayarri and DeGroot (1986) and Dear and Begg (1992), the probability that a study is reported is a function \(g(p)\), of the attained \(p\)-value. If this function is totally unknown, we show that \(N\) is not identifiable. It is often the case, however, that we know the probability that studies with \(p\)-values in some interval \([a,b]\) will be reported. For example, it might be plausible to assume that studies with \(p\)-values between 0.00 and 0.05 are certain to be reported (Rosenthal, 1974). In this case, we show that \(N\), and also \(g(p)\), can be estimated.

\(^4\)If \(H\) is compound (contains more than one model), then it must also be assumed that in each study the \(p\)-value is calculated using the specific model in \(H\) that is true. For this assumption to be satisfied, it is sufficient that the test statistic leading to the \(p\)-value is similar (has the same distribution for all models in \(H\)).
4 A Simple Model for $p$-Values

Assume that the studies we observe are the ones having the smallest $p$-values. That is, if $p_{(i)}$ denotes the $i$th largest $p$-value from among all $N + k$ studies, both reported and unreported, so that

$$0 \leq p_{(1)} \leq \cdots \leq p_{(k)} \leq p_{(k+1)} \leq \cdots \leq p_{(k+N)} \leq 1,$$

the $p$-values that we observe are $p_{(1)}, \ldots, p_{(k)}$. Under the assumptions (3.1), the joint distribution of the observed $p$-values is

$$f(p_{(1)}, \ldots, p_{(k)}|N) = \frac{(N+k)!}{N!}(1-p_{(k)})^N, \quad 0 \leq p_{(1)} \leq \cdots \leq p_{(k)} \leq 1,$$

from which it follows that $p_{(k)}$ is a (minimal) sufficient statistic for $N$. It is shown in Appendix A.1 that the maximum likelihood estimator (MLE) of the unknown $N$ is

$$N_{MLE} = \lfloor k(1-p_{(k)})/p_{(k)} \rfloor,$$

where $\lfloor x \rfloor$ is the integer part of $x$. The estimator $N_{MLE}$, however, is positively biased; the unique unbiased estimator of $N$ based on $p_{(k)}$ is

$$(4.1) \quad \hat{N} = \frac{k(1-p_{(k)}) - 1}{p_{(k)}}$$

It can be shown that

$$N_{MLE} \geq (k\hat{N} + 1)/(k - 1) > \hat{N},$$

from which the positive bias of the MLE is immediate. As can be seen from Table 1, the two estimators are in relatively close agreement only when $p_{(k)}$ is large. Although $\hat{N}$ is the minimum variance unbiased estimator of $N$, it does not necessarily take on integer values. An improvement to this estimator would be the integer closest to $\hat{N}$. Because in the present context it is preferable to underestimate rather than overestimate $N$, we recommend using the integer part $\lfloor \hat{N} \rfloor$ of $\hat{N}$ as an estimator of $N$. Thus, if the largest reported $p$-value in $k = 10$ studies is $p_{(10)} = 0.40$, we would estimate the number of unpublished studies to be $\lfloor \hat{N} \rfloor = \lfloor 12.5 \rfloor = 12$.

[TABLE 1 ABOUT HERE]

Under assumptions (3.1), the largest reported $p$-value $p_{(k)}$ has a beta distribution with parameters $k$ and $N + 1$. From this fact, it is possible to obtain the standard deviation of $\lfloor \hat{N} \rfloor$. Of greater interest, however, in meta-analytic contexts are confidence bounds for $N$, particularly a lower confidence bound $N_L$.

As shown in Appendix A.2, a prescription for obtaining a 100$(1 - \alpha)$% lower confidence bound $N_L$ for $N$ is the following: $N_L$ is the $100\alpha$-th percentile of the negative binomial distribution with parameters $k$ and $p_{(k)}$. 
It is worth noting that the lower confidence bound becomes uninformative ($N_L = 0$) if $p(k) \geq \alpha^{1/k}$. Thus, for example, for $k = 5$ and $\alpha = 0.05$, $N_L$ would be 0 whenever $p(5) \geq 0.55$; for $k = 10$ and the same $\alpha$, $N_L$ is 0 whenever $p(10) \geq 0.74$. When $p(k) < \alpha^{1/k}$, we can find $N_L$ from the prescription above:

$$N_L = \min_{n \geq 0} \left\{ n : \sum_{i=0}^{n} \binom{k+i-1}{k-1} p(k)^i (1 - p(k))^i \geq \alpha \right\}.$$

The search for $N_L$ is easily done by calculator except when $p(k)$ is small or $k$ is large; in such cases, use of a computer is recommended. Alternatively, if tables of percentiles of the beta- or $F$-distributions are available (e.g., Beyer, 1968), one of the following equivalent descriptions of $N_L$ may prove more convenient to apply:

$$N_L = \min_{n \geq 0} \left\{ n : B_{k,n+1;\alpha} < p(k) \right\};$$

where $B_{s,t;\alpha}$ is the $100\alpha$-th percentile of the beta distribution with parameters $s$ and $t$, or

$$N_L = \min_{n \geq 0} \left\{ n F_{2k,2(n+1);\alpha} < (n + 1)p(k)/k(1 - p(k)) \right\},$$

where $F_{2k,2(n+1);\alpha}$ is the $100\alpha$-th percentile of the $F$-distribution with $2k$ and $2(n+1)$ degrees of freedom.

For example, if $k = 7$ and $p(7) = 0.095$, then the 95\% lower confidence bound for $N$ is $N_L = 30$; if $p(7) = 0.05$, then the 95\% lower confidence bound is $N_L = 60$. (Compare these values to $N_{MLE} = 133$ and $|N| = 113$.)

Because the model implies that the mean of $p(i)$ is a linear function of $i$, an approximate graphical check of the model is to plot the pairs $(i, p(i))$, $i = 1, \ldots, k$, and see whether they deviate excessively from a straight line.

5 A Generalization of the Simple Model

It may be unrealistic to believe that we only observe the $k$ smallest $p$-values among those obtained in all $N+k$ studies (reported and unreported). A more realistic scenario is that we have obtained the $m(1 \leq m < k)$ smallest $p$-values, and that the remaining $k-m$ observed $p$-values have been sampled at random from the $N+k-m$ largest $p$-values. In this case the joint density of the observed $p$-values is

$$\frac{(N+k)!}{(N+k-m)!} (1 - p(m))^{N+k-m}, \quad 0 \leq p(1) \leq \cdots \leq p(m) \leq p(m+1) \leq \cdots \leq p(k) \leq 1.$$

Hence, in parallel to our earlier results, with $m$ replacing $k$ and $N+k-m$ replacing $N$, the maximum likelihood estimator of $N$ is

$$N_{MLE} = |m(1 - p(m))p(m)| + m - k = |mp(m)| - k,$$

(5.1)
if the right-hand side of (5.1) is nonnegative, and \( N_{MLE} = 0 \) if the right-hand side of (5.1) is negative. If the right-hand side of (5.1) is negative, however, this casts doubt on the applicability of the model to the data at hand.

Similarly, an unbiased estimator of \( \hat{N} \) is given by

\[
\hat{N} = (m - 1)p_{(m)}^{-1} - k.
\]

As in the previous model, the inequality \( N_{MLE} > \hat{N} \) holds with probability 1.

A 100(1 - \( \alpha \))% lower confidence bound for \( N \) is provided by \( N_L \) where \( N_L + k - m \) is the \( 100\alpha \)-th percentile of the negative binomial distribution with parameters \( m \) and \( p(\theta) \). Alternatively,

\[
N_L + k - m = \min_{q \geq 0} \{ n : B_{m,q+1,\alpha} < p(\theta) \} = \min_{q \geq 0} \{ n : F_{q,m,2(q+1),\alpha} < (q + 1)(1 - p(\theta))/mp(\theta) \}.
\]

If the generalized model holds, the graph of \((i, p(i))\), \( i = 1, \ldots, k \), will be approximately fit by two straight lines, one line passing through the points \((i, p(i))\), \( i = 1, \ldots, m \), and the other line through the remaining \( k - m \) points. The slope of the second line should be approximately \((N + k + 1)(1 - p(\theta))/(k - m + 1)\) times as large as the slope of the second line; the second line’s slope will usually be larger than that of the first line. If \( m \) is not known, a tentative value can be obtained for this parameter by inspecting the scatterplot of the points \((i, p(i))\) and looking for the value of \( i \) where the second line seems to begin. Be aware, however, that if \( m \) is selected in this way, then the estimators \( N_{MLE} \) and \( \hat{N} \) do not necessarily have the maximum likelihood and unbiasedness properties respectively claimed for them, and \( N_L \) is only approximately a 100(1 - \( \alpha \))% lower confidence bound for \( N \).

6 An Alternative Model

Assume that the decision whether or not to publish a study is based solely on the observed \( p \)-value. This is not necessarily realistic because, as noted by the discussants of Iyengar and Greenhouse (1988) and also Begg and Berlin (1988), other factors (such as sample size) may enter into such a decision. We adopt this assumption, however, as a reasonable first approximation. Let \( g(p) \) denote the conditional probability that a study with \( p \)-value equal to \( p \) is reported:

\[
g(p) = P\{\text{study is reported}|p\text{-value} = p\}, \quad 0 \leq p \leq 1.
\]

We call this conditional probability the selection function. Under the above assumption and the assumptions in (3.1), the proportion \( \theta \) of all studies that are reported is given by the mean of the selection function \( g(p) \); that is

\[
\theta = \int_0^1 g(p) \, dp,
\]

7
because under the assumptions in (3.1) the p-value $p$ has a uniform distribution on the interval $[0,1]$. Therefore, $g^*(p) = \theta^{-1} g(p)$ is the (conditional) density of the p-value given that the study producing this p-value is reported. The joint likelihood of the reported p-values $p_1, \ldots, p_k$ when a total of $N + k$ studies were run is

$$
(6.3) \quad \theta^k (1 - \theta)^N g^*(p_1) g^*(p_2) \cdots g^*(p_k).
$$

When $\theta$ is unknown, then $\theta$ and $N$ are not identifiable even if $g^*(p)$ is known. (This is intuitively reasonable because $N$ is involved in the likelihood only with $\theta$, so that $\theta$ and $N$ are confounded.) On the other hand, standard density estimation methods can be used to estimate the density $g^*(p)$ from the observed p-values $p_1, \ldots, p_k$, so that the shape of the selection function $g(p)$ can be estimated.

To identify $N$, we either need to know $\theta$ or have some information about $g(p)$. It is sufficient to know the area $\tau$ of $g(p)$ over any interval $[a, b]$, $0 \leq a < b \leq 1$. (Note, however, that to obtain a reasonable estimate of $N$, such an interval should contain at least one observed p-value.) It is straightforward to show that $\tau(b - a)^{-1}$ is the conditional probability that a p-value will be reported, given that the p-value lies in the interval $[a, b]$; consequently, $\tau$ can be obtained from this conditional probability. For example, we might be willing to assert (in the spirit of the extreme model of publication bias mentioned in the introduction) that all p-values in the interval $[0.00, 0.05]$ will be reported; here, $a = 0.00, b = 0.05, \tau(0.05 - 0.00)^{-1} = 1.00$, so that $\tau = 0.05$. For $\tau$ known, let

$$
h(p) = \begin{cases} 
g(p)/\tau, & \text{if } a \leq p \leq b, 
g(p)/((\theta - \tau)), & \text{if } 0 \leq p < a \text{ or } b < p \leq 1.
\end{cases}
$$

The likelihood of the observed p-values is then

$$
(6.4) \quad \tau^d (\theta - \tau)^{k-d} (1 - \theta)^N h(p_1) h(p_2) \cdots h(p_k),
$$

where $d$ is the number of $p_1$'s in the interval $[a, b]$, $0 \leq d \leq k$. Note that $h(p)$ restricted to the interval $[a, b]$ is a probability density function and can be estimated (if it is not known) using the $d$ p-values observed in the interval $[a, b]$. Similarly, $h(p)$ also defines a density over the union of the intervals $[0, a)$ and $(b, 1]$, and can be estimated from the $k - d$ reported p-values in that union of intervals.

An unbiased estimator of $N$ is given by

$$
(6.5) \quad \hat{N} = d\tau^{-1} - k,
$$

which follows from the fact that $d$ has a binomial distribution with parameters $(N + k, \tau)$. A corresponding estimator of $\theta$ is then

$$
(6.6) \quad \hat{\theta} = k(\hat{N} + k)^{-1} = k\tau d^{-1}.
$$

Unfortunately, these estimators do not always assume values in the parameter space, and thus it should be possible to improve upon them. For example, although it is
negatively biased, the integer part $|\tilde{N}|$ of $\tilde{N}$ seems preferable to $\tilde{N}$ as an estimator of $N$. We mention that this model provides an illustration of a situation in which maximum likelihood estimation (of $N$) does not yield a good estimator. (See Appendix A.3.)

Using the fact that $d$ has a binomial distribution with sample size $N + k$ and known probability $\tau$ of success, we can construct a $100(1 - \alpha)\%$ lower confidence bound $N_L^*$ for $N$; namely,

\begin{equation}
N_L^* = \min_{n \geq 0} \left\{ n : \sum_{i=d}^{n+k} \binom{n+k}{i} \tau^i (1 - \tau)^{n+k-i} \geq \alpha \right\}.
\end{equation}

The search for $N_L^*$ can be programmed on a computer, or can be done with the aid of detailed tables of binomial probabilities. Alternatively (Appendix A.2),

\begin{equation}
N_L^* + k - d = \min_{q \geq 0} \left\{ n : B_{d,q+1;\alpha} \leq \tau \right\}
\end{equation}

is the $100\alpha$-th percentile of the negative binomial distribution with parameters $d$ and $\tau$.

Unless the shape of $g(p)$ is specified, the model is sufficiently general that it cannot be tested using the reported $p$-values. If the shape of $g(p)$ is not specified, the reported $p$-values can have any distribution whatsoever even if the null hypothesis $H$ is true.

7 Illustrations

Example 1. Chalmers and Buyse (1988) discuss the analysis of several data sets relating to patient care in gastroenterology. Table 2 shows the results of $k = 8$ studies conducted between 1981 and 1984. These studies compare "an experimental surgical intervention (proximal gastric vagotomy) to an established intervention (truncal vagotomy plus drainage). Here the event of interest is recurrence, the 'control' group is truncal vagotomy plus drainage, and the treatment group is the experimental intervention." The original table provides two analyses: one based on normalized observed minus expected frequencies yielding an odds ratio, whereas the other uses risk differences. The two methods yield comparable results; we have used the risk difference data in Table 2. (Note that the test statistics are discrete. However, continuous approximations are used to obtain the $p$-values and the sample sizes in these studies appear large enough to sustain such approximations. Thus, although (ii) of assumptions (3.1) may be violated, the violation is not likely to have serious consequences.)

[TABLE 2 ABOUT HERE]

Our first simple model is compatible with the smallest six $p$-values, but not with all eight. For the simple model, $p(8) = 0.43$, $N_{MLE} = 10$, $|\tilde{N}| = |8.279| = 8$ and a 95% lower confidence bound is $N_L = 4$. The generalization of the simple model with $m = 6$ seems to adequately fit the data. With $m = 6$ and $p(6) = 0.18$, we find that $N_{MLE} = 25$ and $|\tilde{N}| = |19.78| = 19$, whereas the 95% lower confidence bound is $N_L = 8$. 9
Our alternative model requires specification of an interval \([a, b]\), \(0 \leq a < b \leq 1\), and the value \(\tau\) of the area under the curve of the selection function \(g(p)\) over this interval. Suppose that we are willing to believe that 90% of the studies having \(p\)-values between 0.00 and 0.05 will be published. Then

\[\tau = (0.90)(0.05 - 0.00) = 0.045.\]

Because there are \(d = 3\) reported \(p\)-values in the interval \([0.00, 0.05]\), our unbiased estimator of \(N\) is \(\bar{N} = 58.67\), which we have suggested rounding down to 58. This value of \(N\) paints a much different picture of the number of unreported studies than do the point estimates of \(N\) from our first model. (On the other hand, the 95% lower confidence bound \(N_L^* = 11\), is not greatly different from the two 95% lower confidence bounds obtained from the first model.)

If we assume that 100% of all studies with \(p\)-values between 0.00 and 0.05 are reported, then \(\tau = 0.05\) and \(\bar{N} = 52\). Had we assumed that 100% of the \(p\)-values in the interval \([0.00, 0.10]\) are reported, then \(\tau = 0.100\), \(d = 4\) and \(\bar{N} = 32\). This example shows the sensitivity of the unbiased point estimate of \(N\) to the choice of the interval \([a, b]\).

**Example 2.** Raudenbush (1984) and Raudenbush and Bryk (1985) analyze \(k = 19\) studies consisting of randomized experiments of the effects of teacher expectancy on later pupil performance on an IQ test. Table 3 gives the resulting \(p\)-values obtained from tests of the null hypothesis that teacher expectancy has no effect on mean IQs versus a one-sided alternative. (This data is taken from Becker, 1994, p. 224.)

**[TABLE 3 AND FIGURE 1 ABOUT HERE]**

Figure 1 gives a plot of the pairs \((i, p_{(i)})\), \(i = 1, 2, \ldots, 19\). The fit of the data to our first simple model is not particularly good. The absence of reported \(p\)-values between 0.040 and 0.208 is particularly surprising, considering how many larger (and insignificant) \(p\)-values are reported. If it were not for the 5 \(p\)-values between 0.211 and 0.243, the generalization of the first model with \(m = 5\) would fit the data fairly closely. Using the simple model, the point estimates of \(N\) are \(N_{MLE} = 1\) and \(|\bar{N}| = 0\) and the 95% lower confidence bound for \(N\) is \(N_L = 0\) (which is uninformative). Using the much better fitting generalization of this model with \(m = 5\), \(p_{(5)} = 0.051\), we find that \(N_{MLE} = 79\), \(|\bar{N}| = 59\) and the 95% lower confidence bound for \(N\) is \(N_L = 21\). The second model with \([a, b] = (0.00, 0.055)\) and assuming that 90% of the \(p\)-values in this interval will be reported, yields very similar results: \(\tau = 0.90(0.055) = 0.0495\), \(d = 5\), \(\bar{N} = 82.01\) and \(N_L^* = 22\). Although as noted above, one can always find a selection function \(g(p)\) that will fit the data, the shape of the selection function required to fit this data seems somewhat nonintuitive.

**Example 3.** Cohen (1983) reviewed studies of the validity of student ratings of instructors. The studies were usually conducted in multisection courses (different instructors),
but had a common final exam. In each study the partial correlation, over sections, of mean instructor ratings and mean final examination scores, conditional on a measure of student ability, was used as an index of validity for the ratings. [Only studies with 10 or more sections were selected by Cohen; there are k = 20 such studies.] Instead of the partial correlations, in Table 4 we give the p-values of standard one-sided tests of the null hypothesis that the population partial correlation is zero (versus alternative of positive correlation). This data is discussed by Becker (1994, p. 226).

[TABLE 4 ABOUT HERE]

The fact that these studies are spread over 20 years, but that 7 of the 20 studies come from one year (and all have small p-values), suggests that the assumption of independence of studies may not hold. The granularity observable in the p-values of Example 2 also can be observed here: there is one cluster of 7 p-values between 0.000 and 0.005, and a smaller cluster of 3 p-values between 0.225 and 0.250. Further the pairs \((i, p(i))\), \(i \geq 9\), trace a highly curvilinear path. (Note that the gaps between successive \(p(i)\)'s increase rapidly with \(i\).) Consequently, neither our first simple model nor its generalization are a very good fit to the data. For this reason, we confine our estimates of \(N\) to those obtained from our second model. Because p-values less than or equal to 0.01 are considered to be “highly significant” in the behavioral sciences, it is intuitively reasonable to believe that 100% of such p-values will be reported. Note that in this case the interval \([a, b] = [0.000, 0.010]\) and \(\tau = b = 0.010\). Observing that \(d = 8\), our point estimate of \(\bar{N}\) is \(\bar{N} = 780\) and the 95% lower confidence bound for \(N\) is \(N_{L}^{*} = 380\). These extremely large values for the number of unpublished studies casts doubt on our model. (Also see the discussion below.)

8 Discussion

Even though the point estimators (5) and (11) of \(N\) are based on different models, there is a formal resemblance between them, and also between the corresponding 100(1 - \(\alpha\))% lower confidence bounds for \(N\). Taking the interval \([a, b]\) in our second model to be of the form \([0, b]\), and assuming that 100% of the p-values in this interval are reported (in which case \(\tau = b\)), the difference in the two models is whether we fix the number, \(m\), of p-values in this interval and let the endpoint \(b = p(m)\) be random as in our first model, or fix \(b\) and let the number \(d\) of points in the interval be random, as in our second model. For our first model this suggests choosing \(m\) as the rank of the largest observed p-value for which we are certain that no smaller p-value occurs in the unreported studies, even though this is a data dependent choice.

It is important to note that our approach to estimating the number \(N\) of unreported studies takes advantage of the fact that p-values have, under the null hypothesis, a common known distribution that does not depend upon the designs, sample sizes or concomitant variables used in the studies (reported or unreported). The individual
distributions of the $p$-values under the alternative hypothesis, on the other hand, can depend heavily on these factors. If no publication bias exists, we would expect the $p$-values to be stochastically smaller under the alternative hypothesis than under the null hypothesis, so that observed $p$-values would be more crowded at the low (near 0) end of the $p$-scale, as in Example 3. Selection of studies on the basis of their $p$-values can also act to make the observed $p$-values appear to come from a distribution stochastically smaller than the uniform distribution on $[0, 1]$, even when the null hypothesis is true. That is, publication bias can be confounded with the truth of the alternative hypothesis being tested. It is for this reason that we require some knowledge of the selection function $g(p)$ in our second model.

We have concentrated in this article on $p$-values rather than on $z$-scores because the influence of selection on the distribution of the $p$-values is more transparent. The question of how best to transform and combine $p$-values in order to test a null hypothesis depends on the nature of the distributions of the $p$-values under the alternative hypothesis, and thus in turn on the sample sizes and designs of the reported studies. Although this issue is important, it is tangential to the question of estimating the number, $N$, of unreported studies (which, here, is done assuming that the null hypothesis is true). It is true that smaller observed $p$-values lead by our methods to larger estimates of $N$. In this sense, the estimate of $N$ serves as a test statistic for the null hypothesis; however, no appropriate critical value (other than that provided by subjective opinion) exists for this statistic. Because estimated $N$'s under our modeling approach do not involve the designs and sample sizes of the reported studies, it is clear that such estimates can at best only provide inefficient "tests" (see above) of the truth of a null hypothesis.

Although the fail-safe $N$ suggested by Rosenthal (1979) has some similarities to the approach described here, the resemblance is superficial. Our approach assumes that the null hypothesis being tested is true, and provides a model (or models) to explain how reported $p$-values could be predominantly small (significant) in this situation. In our first model, it is possible to graphically check the model against the data. Both of our models also produce good point estimates of, or lower confidence bounds for, the number, $N$, of unreported studies. If these estimates (or lower bounds) are excessively large, this fact can be viewed as either discrediting the model or else providing evidence against the null hypothesis. In contrast, the fail-safe $N$ approach tries to buttress a decision against the null hypothesis by showing, under unverifiable assumptions about the $p$-values in unreported studies, that the smallest number of unreported studies that would be needed to overturn that decision is impossibly large. Note that no claim is made that this smallest number of unreported studies is actually equal to the value of $N$. Because the two approaches address the problem of combining the results of published studies from different perspectives, they will not always produce similar results. Thus,

---

5Because the null hypothesis is assumed to be true, the present approach is inappropriate for studying the effects of publication bias and selection on effect sizes.
in Example 3, the estimate of the true number, \( N \), of unpublished studies obtained by our second model is \( \hat{N} = 780 \), but \( N_{FS} \) is equal to 459.80.\(^6\)

We note that either one of our models can be used to calculate a model-based expected value (or upper confidence bound) for the sum \( U \) of the z-scores of the unpublished studies. The value of \( N_{FS} \) calculated using this value of \( U \) (rather than the arbitrary \( U = 0 \)) can then be tested against information contained in the data about the true number, \( N \), of unpublished studies contained in the data, assuming that the model and the null hypothesis are both true. That is, we can test \( H^*: N \geq N_{FS} \) versus the alternative, \( A: N < N_{FS} \), at some suitable level of significance. Rejection of \( H^* \) would indicate that there are an insufficient number of unpublished studies to overturn the decision to reject the original null hypothesis \( H \), even assuming that \( H \) is true. This approach, perhaps applied using several different possible models, can help to remove some of the arbitrariness in the fail-safe \( N \) method by basing it on a specific statistical model of publication bias.

The main concern with unreported studies is that we suspect that these studies, if published, might lead to changes in the conclusions that have been reached from the published studies. Different assumptions about the contents of unreported studies and the mechanisms by which these studies came to be unreported lead to different interpretations and conclusions. It is probably impossible to resolve such differences of interpretation using data from a single meta-analytic problem. However, in disciplines that have commonly accepted paradigms for what constitutes a publishable (or reportable) study, a single model of publication bias or selection should be valid over a wide range of hypothesis testing problems. The meta-analytic studies in a field, particularly those of null hypotheses that current wisdom now considers to be true (or at least not rejected), can then be used to suggest, fit and test such a model; such a statistical analysis of meta-analyses might be called a global meta-analysis. Determining a model, or even a class of models, that successfully predicts the effect of publication bias or selection can help clarify the problems of interpretation of meta-analytic data posed by the possible existence of unpublished studies.

9 Acknowledgement

The authors are grateful to Larry Hedges for his comments and suggestions.

\(^6\)In Example 1, \( N_{FS} = 26.627 \). In Example 2, \( N_{FS} = 17.5828 \).
A Appendix

A.1 Maximum likelihood estimation of \( N \) in the “Simple Model”

The likelihood function of \( N \) under the simple model is

\[
L(N|p(k)) = f(p(1), \ldots, p(k)|N) = \frac{(N + k)!}{N!} (1 - p(k))^N,
\]

for \( 0 \leq p(1) \leq \cdots \leq p(k) \leq 1 \). This function of the parameter \( N \) depends on the data only through \( p(k) \); further, for any two distinct values \( p \) and \( p^* \) of \( p(k) \), \( L(N|p) \) is not the same function of \( N \) as is \( L(N|p^*) \). Hence, \( p(k) \) is a (minimal) sufficient statistic for \( N \). Further,

\[
\frac{L(N + 1|p(k))}{L(N|p(k))} = \frac{(N + k + 1)}{(N + 1)} (1 - p(k))
\]

strictly decreases from \((k + 1)(1 - p(k)) \) at \( N = 0 \) to \((1 - p(k)) \) as \( N \to \infty \). If \( p(k) > k/(k+1) \), then the ratio (A.1) is always less than 1, and it follows that \( L(N|p(k)) \) is maximized for \( N = 0 \). If \( p(k) = k/(k+1) \), then both \( N = 0 \) and \( N = 1 \) maximize the likelihood. Otherwise, the ratio (A.1) changes direction from being greater than 1 to being less than or equal to 1 at

\[
(A.2) \quad N_{MLE} = \lfloor k(1 - p(k))/p(k) \rfloor,
\]

where \( \lfloor x \rfloor \) denotes the integer part of \( x \). If \( k(1 - p(k))/p(k) \) is an integer, then both \( N_{MLE} \) and \( N_{MLE} + 1 \) maximize the likelihood. Note that the results for the case \( p(k) > k/(k+1) \) are consistent with (A.2) and the sentence following that equation, so that (A.2) gives the maximum likelihood estimator of \( N \) for all values of \( p(k), 0 \leq p(k) \leq 1 \).

In the generalization of the simple model, the likelihood for \( N' \) has the same form as (A.1) with \( m \) replacing \( k \), \( p(m) \) replacing \( p(k) \), and \( N + k - m \) replacing \( N \). This yields \( \lfloor m(1 - p(k))/p(k) \rfloor \) as the maximum likelihood estimator of \( N' = N + k - m \) and (5.1) as the maximum likelihood estimator of \( N \). Here, however, it is possible that the likelihood is maximized at a value of \( N' \) that is less than \( k - m \) (for example, when \( p(m) > m/(m+1) \), then \( N' = 0 \) maximizes the likelihood). However, the derivation of the maximum likelihood estimator in the simple model shows that in this case the likelihood is decreasing in \( N' \), so that the likelihood is maximized over the parameter space at the smallest acceptable value of \( N' \), namely, \( N' = k - m \). Consequently, in this case \( N_{MLE} = 0 \), as asserted in the sentence following (5.1).

A.2 Lower confidence bound for \( N \) in the “Simple Model”

Under the assumptions (3.1), the \( k \)-th largest \( p \)-value \( p(k) \) in all \( N + k \) studies, reported and unreported, has a beta distribution with parameters \( k \) and \( N + 1 \). (See Lehmann, 1986, p. 345.) We have already noted that \( p(k) \) is minimal sufficient for \( N \). Recall that
$B_{s,t;\alpha}$ denotes the 100\(\alpha\)-th percentile of the beta distribution with parameters \(s\) and \(t\). It can be shown (see Lehmann, 1986, pp. 78-83) that

\[
C = \{n : n \geq 0 \text{ and } B_{k,n+1;1-\alpha} < p(k)\}
\]

is the uniformly most accurate 100\((1 - \alpha)\)% lower confidence region for \(N\); that is, \(C\) has minimal probability of including any value of \(N\) smaller than the true value among all 100\((1 - \alpha)\)% confidence regions for \(N\). Because a random variable having a beta distribution with parameters \(s\) and \(t\) decreases stochastically with \(t\) for fixed \(s\), it follows that \(B_{k,n+1;\alpha}\) is a decreasing function of \(n\). Hence, the region \(C\) has the form of a one-sided interval

\[
C = [N_L, \infty),
\]

where \(N_L\) is the smallest nonnegative integer \(n\) satisfying \(B_{k,n+1;\alpha} \leq p(k)\). This verifies one of the equivalent definitions of \(N_L\). The other definitions of \(N_L\) are consequences of the following known relationships between the beta, \(F\), binomial and negative binomial distributions:

(A.3) \[
P\{X \leq x\} = P\{F \leq tx/s(1-x)\} = P\{U \geq s\} = P\{V \leq t-1\},
\]

where \(0 < x < 1\), \(s\) and \(t\) are positive integers, \(X\) has a beta distribution with parameters \(s\) and \(t\), \(F\) has an \(F\)-distribution with parameters \(2s\) and \(2t\), \(U\) has a binomial distribution with parameters \(s + t - 1\) and \(x\), and \(V\) has a negative binomial distribution with parameters \(s\) and \(x\). (See Olkin, Gleser and Derman, 1995, pages 344 and 461 and Lehmann, 1986, p. 200.)

The beta distribution with parameters \(k\) and \(N + 1\), where \(N\) is an unspecified nonnegative integer, is a complete family of distributions. Because \(p(k)\) has this family of distributions, it follows that there is at most one unbiased estimator of \(N\) based on \(p(k)\). It can be straightforwardly derived from the distribution of \(p(k)\) that \(\hat{N}\) defined by (5) is an unbiased estimator of \(N\); because \(p(k)\) is sufficient, it follows that \(\hat{N}\) is the uniformly minimum variance unbiased estimator of \(N\).

In the generalization of the simple model, once we note that \(p(m)\) has a beta distribution with parameters \(m\) and \(N + k - m + 1\), we see that the 100\((1 - \alpha)\)% lower confidence bound \(N_L\) for \(N\), the uniformly most accurate property of this confidence bound, and the uniformly minimum variance unbiased property of the estimator \(\hat{N}\) can be derived from results for the simple model by making the substitutions \(m\) for \(k\), \(p(m)\) for \(p(k)\), etc. already mentioned.

A.3 Maximum likelihood estimators of \(\theta\) and \(N\) in the second model.

The maximum likelihood estimators of \(\theta\) and \(N\) are obtained from (10) by maximizing

(A.4) \[
\tau^d(\theta - \tau)^{k-d}(1 - \theta)^N
\]
assuming that \((\theta, N)\) is functionally unrelated to the function \(h(p)\). For fixed \(N\), the expression (13) is maximized over \(\tau \leq \theta \leq 1\) by

\[
\theta(N) = \frac{k - d + N\tau}{N + k - d},
\]

as is easily shown by inspecting the sign of the first derivative of (A.4) with respect to \(\theta\). Substituting (A.5) for \(\theta\) in (A.4) yields

\[
Q(N) = \tau^d \left(\frac{(k - d)(1 - \tau)^{k-d}}{N + k - d}\right) \left(\frac{N(1 - \tau)}{N + k - d}\right)^N.
\]

From the fact that \(r(z) = (z + 1)^{z+1}/z^z\) is monotone increasing in \(z\), it follows that the ratio

\[
Q(N+1)/Q(N) = (1 - \tau)r(N)/r(N + k - d)
\]

is less than 1. Consequently, \(Q(N)\) is decreasing in \(N\) and thus is maximized for \(N = 0\). Thus, \((\theta, N) = (\theta(0), 0) = (1, 0)\) maximizes the likelihood regardless of the data, showing that maximum likelihood estimation does not provide us with much information. Indeed, we expect to be able to obtain a better estimate of \((\theta, N)\) because intuitively \(d\) gives us information about \(N\) when \(\tau\) is known.

If \(\tau\) and the function \(h(p)\) are known, then \(d\) is a sufficient statistic for \((\theta, N)\), as can be seen from (10). It is easily seen that \(d\) has a binomial distribution with parameters \(N + k\) and \(\tau\). Consequently, the mean of \(d\) is \((N + k)\tau\) and it immediately follows that \(\bar{N} = d\tau^{-1} - k\) is an unbiased estimator of \(N\).

**References**


