Thailand Statistician
January 2016; 14(1): 15-24
http://statassoc.or.th
Contributed paper

Some Methods for Addressing Publication Bias in Statistical Meta-analysis

April Albertine
Department of Mathematics and Statistics, University of Maryland, Baltimore County, USA.
E-mail: april.albertine@gmail.com

Received: 13 September 2014
Accepted: 3 September 2015

Abstract

In statistical meta-analysis, publication bias may bias the overall conclusion toward a significant result. We examine two approaches for accounting for this bias: the fail safe sample size, for which we provide a correction, and selection models, for which we explore a simplification. In a new approach which combines both of these methods, we use a variation on one type of selection model in order to make better assumptions about the unobserved studies in the fail safe sample size.

Keywords: Fail safe sample size, meta-analysis, publication bias, selection models.

1. Introduction

Meta-analysis, the science of combining the results of many different primary studies in order to arrive at an overall conclusion about the question of interest, consists of several distinct phases, identified by Hartung et al. (2008) as (1) problem formulation, (2) data collection, (3) data evaluation, and (4) data analysis and interpretation. We are presently concerned with the second phase, in which the researcher must make a serious effort to collect all relevant available studies (ranging from peer-reviewed articles to master's theses) for inclusion in the meta-anaysis. Even if the meta-analyst is able to perform an exhaustive literature search and identify all available studies pertaining to the research question, the difficulty may still remain that some studies may be unpublished and unavailable to the researcher. This is the well-known "file drawer problem," in which studies with nonsignificant results languish in the file drawers of authors and journal editors, who are presumably less likely to submit or accept for publication studies that fail to reject the null hypothesis. This publication bias may lead to an overall conclusion of significance when in fact, including the unavailable studies might have yielded an overall nonsignificant result. The objective of this study is to review two categories of methods for addressing publication bias (fail safe sample size and selection models) and then to synthesize ideas from each approach into a new method.

One category of statistical methods for addressing publication bias consists of those based on Rosenthal's (1979) fail safe sample size. Another category includes model-based approaches such as the one developed by Iyengar and Greenhouse (1988). The first approach (outlined in section 2) determines the number of nonsignificant studies that must exist in order for the meta-analysis to

yield a just barely nonsignificant result; we provide a correction. The second approach incorporates a parameter for publication bias into the likelihood of the effect size, and then estimates the publication bias parameter and the effect size using maximum likelihood estimation. We simplify the procedure in section 3 using a normal approximation. Finally, in Section 4, we use a variation of a selection model by Copas (1999) in order to make more informed assumptions about the unpublished studies in the context of the fail safe sample size.

2. Fail Safe Sample Size

2.1. Combined p-value method

One method of assessing publication bias is the "fail safe sample size" approach, an ad hoc method which determines, for a meta-analysis that claims significance of an effect, how many unpublished non-significant results would have to exist in order for the meta-analysis to change from "significant" to "not significant." The subject area scientist can then judge whether the fail safe number of unpublished studies is likely to exist.

Consider a meta-analysis of studies comparing the means of some continuous variable for two treatment groups. Let θ be the standardized mean difference of a treatment group and the mean of a control group. Iyengar and Greenhouse (1988) use as their starting point a method proposed by Rosenthal (1979). Rosenthal uses the inverse normal method for combining (one-tailed) p-values in order to determine what he terms the "tolerance for future null results." Under the null hypothesis (θ=0), the p-values are uniformly distributed, implying that $Z_i = \Phi^{-1}(1 - p_i)$ is standard normal. Then $S_k = Z_1 + ... + Z_k \sim N(0, k)$, so $S_k/k^{1/2} \sim N(0, 1)$. If we suppose that $S_k/k^{1/2} \geq z_a$ (that is, the meta-analysis claims significance without considering publication bias) then a proposed "fail safe sample size" would be the smallest number of additional unpublished studies n_{FS} that would have to be included in this calculation in order to bring this statistic down to less than z_α. Let $\bar{Z}_0$ be the mean Z-value of the unpublished studies. If $\bar{Z}_0$ is zero, this n would be the solution to

$$S_k/(k + n_{\text{FS}})^{1/2} = z_\alpha. \tag{1}$$

Iyengar and Greenhouse (1988) build on this approach by noting that if publication bias is a factor in the unpublished studies being unpublished, then the Z values corresponding to those studies would not be distributed as standard normal. Rather, since the p-values of the unpublished studies would tend to be higher than usual, the authors propose a truncated normal distribution for the corresponding Z values, with the cutoff located at z_α. This is making the assumption that there are no studies with significant results among the unpublished studies. The mean of this truncated distribution is $M(\alpha) = -\frac{\phi(z_\alpha)}{\Phi(z_\alpha)}$. Letting $\bar{Z}_0 = M(\alpha)$, the numerator of the LHS above would become $S_k + n_{\text{FS}}(\alpha)M(\alpha)$. Note that the calculated fail safe n_{FS} depends on α since the distribution of unobserved Z values is truncated at z_α.

For example, consider Saavedra and Garcia's (2013) meta-analysis of the impact of Conditional Cash Transfer (CCT) programs on school dropout rates in developing countries, reported in Table1. CCT programs are designed to reduce extreme poverty in developing nations by incentivizing desired health and education behaviors among the very poor. In educational CCT programs, for example, students or families receive cash payments conditional on school enrollment, attendance, or performance. In each of the primary studies in Saavedra and Garcia's (2013) meta-analysis, the authors compared the difference in the secondary school dropout rate between students who received a CCT

Table 1 Saavedra and Garcia's conditional cash transfer data

Author	$\hat{\theta}$	$\hat{\sigma}(\hat{\theta})$
De Janvry (2006)	7.4	0.200
Glewwe (2008)	0.27	0.051
Cameron (2009)	1.42	0.395
Todd (2005)	2.89	0.992
Raymond (2003)	1.94	0.418
Behrman (2005)	8.36	1.283

scholarship (π_t), and those who did not (π_c). Each study reported a standard error $\hat{\sigma}(\hat{\theta})$ for its estimate $\hat{\theta}$ of $\theta = \pi_c - \pi_t$. Using Rosenthal's approach, the number of unpublished studies necessary to bring the p-value up to .05 is 1,331. On the other hand, incorporating the mean of the truncated distribution (with $\alpha = .05$) for the missing Z values lowers the fail safe number to 292.

2.2. Effect size method

Another way to determine an appropriate fail safe sample size is due to Orwin (1983) who suggested choosing the number of unpublished studies that would push the effect size below some critical threshold d_c, which would be determined by the subject area experts. Consider, for example, Cohen's d (1988) as an estimate of the effect size δ, the standardized mean difference between two groups. Let $\bar{X}_j$, s_j^2, n_j be the sample mean, sample variance, and sample size for the jth group, $j = 1, 2$. Cohen's d is defined for the ith study as $d_i = \frac{\bar{X}_1 - \bar{X}_2}{S}$ with $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$ In the case where all of the studies have identical designs, and are based on similar populations, a reasonable combined effect size would be a simple average, $\bar{d}$. Assuming that the unpublished studies have a mean effect size of μ, the average of all of the effect sizes, published and unpublished, would be

$$\frac{k\bar{d} + n_{\text{FS}}\bar{d}_0}{k + n_{\text{FS}}} = d_c \tag{2}$$

Choosing a tolerance d_c and solving for n_{FS} yields the fail safe sample size. The only unknown in the equation above is the sample mean $\bar{d}_0$ of the unobserved studies. Iyengar and Greenhouse (1988) suggest specifying $\bar{d}_0 = M(\alpha)$. However, the mean of the truncated distribution in this case is based on the unobserved d's instead of the Z-scores in the inverse normal method above, so the mean will be slightly different. Under the null hypothesis $H_0 : \delta = 0$, $d \approx N(0, a)$ where $a = \frac{n_1 + n_2}{n_1 n_2}$. Again assuming that there are no significant results among the unobserved studies, we truncate d such that $\frac{d}{\sqrt{a}} < z_\alpha$. Hence,

$$E\left(\frac{d_{tr}}{\sqrt{a}}\right) = -\frac{\phi(z_\alpha)}{\Phi(z_\alpha)} \tag{3}$$

where d_{tr} is the truncated (unobserved) version of d. Then the mean $\bar{d}_0$ of the unobserved ds is $\sqrt{a}M(\alpha)$. The multiplier $\sqrt{a}$, being less than 1 for all sample sizes $n_1, n_2 > 1$, has the effect of shrinking $\bar{d}_0$ closer to zero, and therefore of bringing the fail safe sample size closer to the one corresponding to the non-truncated estimate, 0. On identical, known study designs, a is a fixed and known quantity. However, in a more realistic scenario where the sample sizes vary across studies, this approach may be modified by estimating the average sample sizes of the unknown studies. Letting N be the average study size over all studies and treatment groups, $\hat{a} = 2/N$ for each of the unobserved studies. Therefore, our suggested correction is to estimate the mean $\bar{d}_0$ of the unpublished studies

as $\sqrt{\frac{2}{N}}M(\alpha)$. A different context for applying Orwin's fail safe sample size is the meta-analysis by Andrews (1990) which evaluates the evidence that "appropriate correctional services" reduces prisoner recidivism. "Appropriate correctional services" were those programs for rehabilitating prisoners which met certain criteria, such as adapting to the learning needs of the prisoners. Here, for each primary case-control study, we compute the ϕ-coefficient for the 2x2 contingency table relating recidivism and receipt of these correctional services. Let n be the total number of subjects in a primary study. Using $1/n$ (the known asymptotic variance of ϕ) for the sampling variance, we approximate $\phi \approx N(0, 1/n)$ under the null hypothesis that the services have no effect on recidivism. Then, we can estimate the mean of the unobserved studies as $\sqrt{\frac{1}{n}}M(\alpha)$. Let $\bar{\phi}$ be the mean of the observed effect sizes, and let $\bar{\phi}_0$ be the mean of the unobserved effect sizes. We can select the fail safe n_{FS} such that

$$\frac{k\bar{\phi} + n_{\text{FS}}\bar{\phi}_0}{k + n_{\text{FS}}} = \phi_c, \tag{4}$$

where ϕ_c is the designated fail safe tolerance.

Table 2 Fail safe sample size for recidivism data using effect size tolerance approach

	$\bar{\phi}_0 = 0$	$\bar{\phi}_0 = \sqrt{\frac{1}{n}}M(\alpha)$
$\phi_c = .20$	14	12
$\phi_c = .15$	35	33
$\phi_c = .10$	80	73
$\phi_c = .05$	211	179
$\phi_c = .01$	1267	665

Table 2 shows the fail safe sample size for the recidivism data for different values of ϕ_c, assuming $\bar{\phi}_0 = 0$ and also estimating $\bar{\phi}_0$ as $\sqrt{\frac{1}{n}}M(\alpha)$.

3. Selection Model Approach

Although the fail safe sample size approach is attractive due to its simplicity and appeal to intuition, it is not without its drawbacks. One issue is that among the unpublished studies, those with large sample sizes are treated no differently that those with small sample sizes. Furthermore, although we can make statements about the presence or extent of publication bias using this approach, it provides no information about the impact on the effect size of interest. Selection models such as the ones outlined in this section are one possibility for addressing these issues.

3.1. Weighted non-central t model

When there is no publication bias, the summary measure x follows some density function $f(x, \theta)$. One method of accounting for the bias, due to Iyengar and Greenhouse (1988), is to explicitly model it by choosing a nonnegative weight function $w(x)$ which assigns greater weight to values of x that are more likely to be published.

Iyengar and Greenhouse (1988) consider ten experiments comparing the effects of open classroom education versus traditional classroom education on a measure of student creativity (Table 3). Each study has balanced data, so N_i is the sample size of both the control and the treatment group in the ith study. g_i is the difference in group means for the ith study divided by the estimate of the pooled standard deviation for that group. t_i represents the corresponding t statistic with degrees of

Table 3 Effect of open classroom on student creativity

i	N_i	g_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

freedom q_i. Without publication bias, t_i is distributed as noncentral t with noncentrality parameter $\eta_i = (N_i/2)^{1/2}\theta$ and degrees of freedom $q_i = 2N_i - 2$. Calling this density $f(t; \eta_i, q_i)$, the authors suggest weighting it with some nonnegative weight function $w(t)$ and normalizing so that the new weighted density is

$$\frac{f(t_i; \eta_i, q_i)w(t_i)}{\int_{-\infty}^{\infty} f(t, \eta_i, q_i)w(t)dt}.$$

By independence of the studies, the joint likelihood in the presence of publication bias is

$$\prod_{i=1}^{10} \frac{f(t_i; \eta_i, q_i)w(t_i)}{\int_{-\infty}^{\infty} f(t, \eta_i, q_i)w(t)dt}.$$

From here, we may simply calculate the MLE of θ, for our choice of weight function for the publication bias. However, to allow for greater flexibility and also to inform our choice of weight function using the data itself, we can also incorporate a parameter, say β, in the weight function that will allow us to choose the most likely function from a class of functions parametrized by β. Let $t(q, \alpha)$ be the critical value for a two-sided t-test with q degrees of freedom and size α. The classes of weight functions suggested by Iyengar and Greenhouse (1988) are

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

and

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

where $q = N - 2$, $\beta \geq 0$, and $\gamma \geq 0$.

3.2. Normal approximation

In order to simplify the noncentral t model, we suggest a normal approximation of Hedges' g (1981) to estimate the standardized mean difference. Let $\bar{X}_j$, s_j^2, n_j be the sample mean, sample variance, and sample size for the jth group, $j = 1, 2$. Hedges' g is defined as $g = \frac{\bar{X}_1 - \bar{X}_2}{S^*}$ with $S^{*2} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$. We approximate the density of g:

$$f(g; \theta) \approx n\left(\theta, \frac{2}{N} + \frac{\theta^2}{2(2N-2)}\right)$$

The weight functions are based on the acceptance region of hypothesis $H_0 : \theta = 0$:

$$w_1(g;\beta) = \begin{cases} \left(\frac{|g|}{\sqrt{\frac{2}{N}}z_{\alpha/2}}\right)^{\beta}, & \text{if } |g| \leq \sqrt{\frac{2}{N}}z_{\alpha/2} \\ 1, & \text{otherwise} \end{cases}$$

and

$$w_2(g;\gamma) = \begin{cases} e^{-\gamma}, & \text{if } |g| \leq \sqrt{\frac{2}{N}}z_{\alpha/2} \\ 1, & \text{otherwise} \end{cases}$$

where $N = N_1 = N_2$, $\beta \geq 0$, and $\gamma \geq 0$.

Table 4 Comparison of MLEs

		Non-central t MLE	Normal approximation MLE
w1	$(\hat{\theta}, \hat{\beta})$	$(0.026, 1.33)$	$(0.031, 1.393)$
	$SD(\hat{\theta}), SD(\hat{\beta})$	$(0.052, 0.59)$	$(0.054, 0.61)$
	$Var(\hat{\theta}, \hat{\beta})$	$\begin{bmatrix} 0.003 & -0.003 \\ -0.003 & 0.348 \end{bmatrix}$	$\begin{bmatrix} 0.003 & 0.278 \\ 0.278 & 0.374 \end{bmatrix}$
	$\hat{\theta} \pm 2SD(\hat{\theta})$	$(-0.078, 0.130)$	$(-0.077, 0.139)$
	$\hat{\beta} \pm 2SD(\hat{\beta})$	$(0.15, 2.51)$	$(0.17, 2.62)$
w2	$(\hat{\theta}, \hat{\gamma})$	$(0.022, 2.53)$	$(0.025, 2.51)$
	$SD(\hat{\theta}), SD(\hat{\gamma})$	$(0.049, 0.65)$	$(0.049, 0.65)$
	$Var(\hat{\theta}, \hat{\gamma})$	$\begin{bmatrix} 0.002 & 0.000 \\ 0.000 & 0.417 \end{bmatrix}$	$\begin{bmatrix} 0.002 & 0.214 \\ 0.214 & 0.417 \end{bmatrix}$
	$\hat{\theta} \pm 2SD(\hat{\theta})$	$(-0.076, 0.120)$	$(-0.073, 0.123)$
	$\hat{\gamma} \pm 2SD(\hat{\gamma})$	$(1.23, 3.83)$	$(1.22, 3.80)$

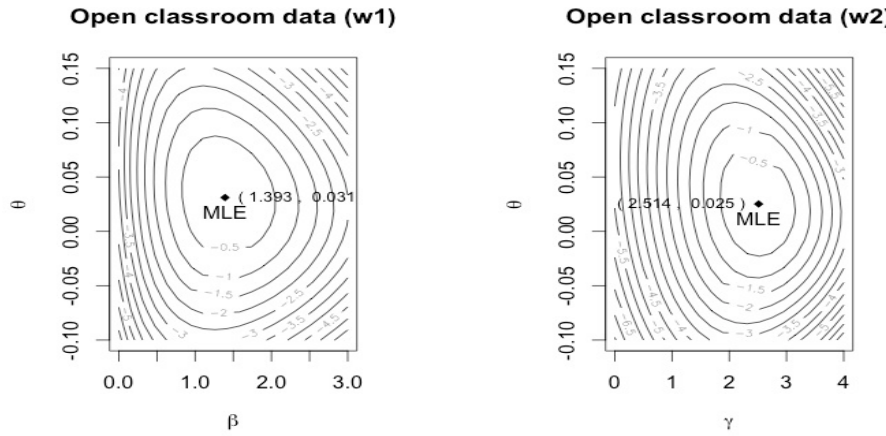
The contours of the Normal likelihood surface are shown in Figure 1. The MLE and related statistics using the Normal approximation are very close to the results reported by Iyengar and Greenhouse (1988) for the non-central t approach. Table 4 shows the results of the Normal approximation side by side with the results of noncentral t approach. $\beta = 0$ or $\gamma = 0$ would correspond to constant weight functions (and hence, no publication bias). Hence, there is evidence that publication bias played a role in the availability of the studies since the confidence intervals for these parameters do not cover zero. On the other hand, since the estimate for theta is so small compared to its standard error, we cannot draw helpful conclusions about the true mean θ.

The normal approximation allows straightforward application of the model in a generic meta-analysis. Assume that all reported effect sizes estimate the same effect θ. Let $\hat{\theta}$ be and estimate of θ from a study, and let $\hat{\sigma}^2$ be its estimated standard error, which we treat as fixed. Then a normal approximation for the the estimated effect size is $\hat{\theta} \sim N(\theta, \hat{\sigma}^2)$, and we can write the generic weight functions as

$$w_1(\hat{\theta};\beta) = \begin{cases} \left(\frac{\hat{\theta}}{\hat{\sigma}z_{\alpha/2}}\right)^{\beta}, & \text{if } |\hat{\theta}| \leq \hat{\sigma}z_{\alpha/2} \\ 1, & \text{otherwise} \end{cases}$$

and

$$w_2(\hat{\theta};\gamma) = \begin{cases} e^{-\gamma}, & \text{if } |\hat{\theta}| \leq \hat{\sigma}z_{\alpha/2} \\ 1, & \text{otherwise} \end{cases}$$



(a) Weight function w1 (b) Weight function w2

Figure 1 Contour plots of the joint likelihood of θ and the publication bias parameter for each of the two weight functions (Open classroom data)

where $N = N_1 = N_2$, $\beta \geq 0$, and $\gamma \geq 0$.

We can easily apply this approach to the Conditional Cash Transfer meta-analysis data, which reports an estimated effect $\hat{\theta}_i$ and estimated standard error $\hat{\sigma}_i$ for studies $i = 1, \ldots, 6$, where $\theta = \pi_1 - \pi_2$, the difference in population proportions. This is a case where very little can be determined about the presence of publication bias. The contour maps show horizontal, nearly parallel lines, indicating that the likelihood is very flat in the direction of the publication bias parameter. θ is restricted to a relatively small range centered near -.75, regardless of the extent of the publication bias. The MLE for the weighted density is very close to MLE computed without weighting (-.753), which is equivalent to the estimate of θ using the fixed effects model. This is unsurprising for a data set like this one where there are a small number of studies, all of which report large standardized differences in the sample proportions. The weighting function impacts the density only for values of $\hat{\theta}$ that are very close to 0, so it has very little impact on densities that are centered far from zero, such as the ones in this data set.

Applying this approach to the recidivism data tells a different story. Here, there is reason to believe that publication bias plays a role, albeit small, in inflating the overall effect size estimate. For example, under w2, γ has a 95% confidence interval of (0.326,1.662), which does not cover zero. Furthermore, the MLE of θ is .234, somewhat lower than the unweighted MLE of .249. Although the point estimate is lower, the 95% confidence interval for θ still does not include zero, even accounting for publication bias in this way.

4. Selection Model for Informing Fail Safe Sample Size

One way we might choose to combine these two approaches is to use a selection model to estimate the reported effect sizes of the missing studies. Copas (1997 and 1999) suggests a selection model which accounts for sample size and effect size separately, using the Andrews (1990) recidivism data as an example. Using the same notation, the model for $y = \phi$ is:

$$y = \mu + \left(\tau^2 + n^{-1}\right)^{1/2} \epsilon,$$

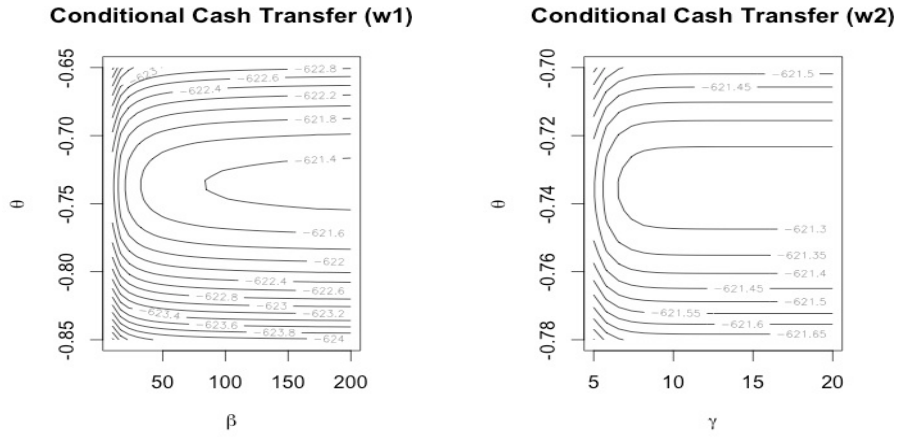


Figure 2 Conditional cash transfer data

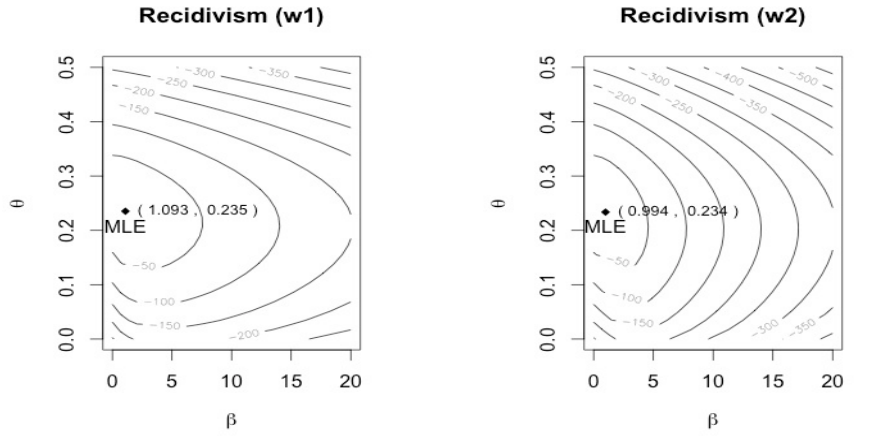


Figure 3 Recidivism data

where $\epsilon \sim N(0,1)$. This represents the random effects model of meta-analysis with no covariates. A separate correlated variable z is written as

$$z = \gamma_0 + \gamma_1 n^{1/2} + \delta,$$

where $\delta \sim N(0,1)$ and $\text{cov}(\epsilon, \delta) = \rho$. When $z > 0$, the study is published; when $z \leq 0$, the study is not published. Intuitively this model means that a study is more likely to be published if it has a large sample size (small standard error), or if it has large δ, which would tend to mean a large estimated effect size. All observed studies come from the conditional distribution of y given $z > 0$. Copas examined this conditional likelihood, maximizing it over μ, τ and ρ for different fixed values of γ_0 and γ_1.

We use this model (modified to include one covariate) to estimate the mean effect size of the unobserved studies. Let the model for y be defined as

$$y = \mu + x\beta + \left(\tau^2 + n^{-1}\right)^{1/2} \epsilon,$$

where $\epsilon \sim N(0,1)$ and x is a covariate of y. Then the density for y given $z < 0$ (that is, the study was NOT selected) is

$$f(y|z<0) = -\frac{1}{2}\log\left(\tau^2 + n^{-1}\right) - \frac{1}{2}\frac{(y-\mu)^2}{\tau^2 + n^{-1}} - \log\left(\Phi(-u)\right) + \log\left(\Phi(-v)\right)$$

$$u = \gamma_0 + n^{1/2}\gamma_1$$

and

$$v = \frac{u + \rho(y-\mu)(\tau^2 + n^{-1})^{-1/2}}{(1-\rho^2)^{1/2}}$$

Using the estimated parameters for the model for the observed studies, we can estimate the expected value of y for an unobserved study of sample size n_0 and covariate x_0. We estimate the fail safe sample size by making draws from the distribution of unobserved effect sizes until the average effect size of the observed studies and the "new" unobserved studies becomes less than or equal to some tolerance level ϕ_c. To get the covariate x_0 of a "new" unobserved study, we resample x_0 from the observed values of the covariate x. For choosing the sample size n_0, we restrict ourselves to resampling from the lower half of the observed sample sizes, since studies with very large sample sizes are unlikely to go unpublished. For various fixed values of γ_0 and γ_1, we repeat this procedure for estimating the fail safe sample size 100 times in order to get a Monte Carlo estimate for the true fail safe sample size. We apply this procedure for the recidivism data, using as x the categorical variable representing the quality of the study (high/low).

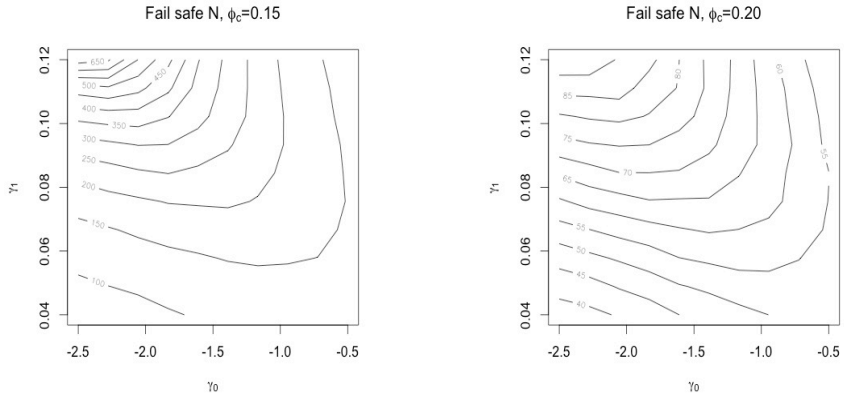


Figure 4 Recidivism data

The contours of the estimated fail safe sample size are shown in Figure 4 for a range of values of γ_0 and γ_1. The probability of selection (that is, that $p(z > 0|n)$) for a study of size n is $\Phi(\gamma_0 + n^{1/2}\gamma_1)$, so the range of values for γ_0 and γ_1 cover a wide range of probabilities of publication. For example, for a study of size 150, the probability of publication would be .02 in the lower left corner of the plot and .98 in the upper right corner. The contours are shown for the tolerance level $\phi_c = .15$ and $\phi_c = .20$. We see that only a small portion of the contour plot for $\phi = .15$ is less than 100, indicating that no matter what the true values of γ_0 and γ_1 are, publication bias could not account for an overall effect size less than 0.15. For $\phi_c = .20$, the fail safe sample size ranges from about 40 to

about 90. The contours indicate that publication bias is less of a problem than is implied by the fail safe sample sizes computed using Orwin's method in Table 2.

5. Conclusions

In this paper we have examined two very different methods for approaching the problem of publication bias in meta-analysis. While the fail safe sample size (Section 2) is both simple and intuitive, the researcher must make assumptions about the p-values or estimated effect sizes of the unobserved studies. These assumptions (for example, that the mean of the unobserved studies come from a truncated distribution) impact the estimate of the fail safe sample size. Selection models such as Iyengar and Greenhouse's (1988) model in Section 3.1 have been used to directly estimate parameters of interest in the presence of publication bias, and certainly the normal approximation simplifies this approach for use in a variety of contexts. However, as shown in Section 4, selection models may also be used to make an informed specification of the unpublished studies in calculating the fail safe sample size.

Acknowledgements

I would like to thank my advisor, Dr. Bimal Sinha, for his helpful guidance and corrections. Thanks are also due to the reviewers, particularly for referring me to Copas' work on selection models. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144243.

References

Andrews DA, Zinger I, Hoge R, Bonta J, Gendreau P, Cullen F. Does correctional treatment work?: a clinically relevant and psychologically informed meta-analysis. Criminology. 1990; 28: 369-429.

Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2 ed. Erlbaum; 1988.

Copas JB, LHG. Inference for non-random samples. J. Roy. Statist. Soc. Ser. B 1997; 59: 55-95.

Copas JB. What works?: selectivity models and meta-analysis. J. Roy. Statist. Soc. Ser. A 1999; 162: 95-109.

Hartung J, Knapp G, Sinha BK. Statistical meta-analysis with applications. Hoboken, N.J.:Wiley; 2008.

Hedges LV. Distribution theory for glass's estimator of effect size and related estimators. J. Educ. Stat. 1981; 6: 107-128.

Iyengar S, Greenhouse JB. Selection models and the file drawer problem. Stat. Sci. 1988; 3: 109-135.

Orwin R. A fail-safe n for effect size in meta-analysis. J. Educ. Stat. 1983; 8: 157-159.

Rosenthal R. The "file drawer problem" and the tolerance for null results. Psycho. Bull. 1979; 86: 638-641.

Saavedra JG, Garcia S. Educational impacts and cost-effectiveness of conditional cash transfer programs in developing countries: A meta-analysis. University of Southern California Center for Economic and Social Research Working Paper Series. 2013; 2013-007.