BRIEF COMMUNICATIONS

Evolution, 59(2), 2005, pp. 464–468

THE FILE-DRAWER PROBLEM REVISITED: A GENERAL WEIGHTED METHOD FOR CALCULATING FAIL-SAFE NUMBERS IN META-ANALYSIS

MICHAEL S. ROSENBERG
Center for Evolutionary Functional Genomics, Biodesign Institute, and School of Life Sciences, Arizona State University, Tempe, Arizona 85287-4501
E-mail: msr@asu.edu

Abstract.—Quantitative literature reviews such as meta-analysis are becoming common in evolutionary biology but may be strongly affected by publication biases. Using fail-safe numbers is a quick way to estimate whether publication bias is likely to be a problem for a specific study. However, previously suggested fail-safe calculations are unweighted and are not based on the framework in which most meta-analyses are performed. A general, weighted fail-safe calculation, grounded in the meta-analysis framework, applicable to both fixed- and random-effects models, is proposed. Recent meta-analyses published in *Evolution* are used for illustration.

Key words.—Fail-safe numbers, file drawer problem, meta-analysis, publication bias, statistical methods.

Received September 30, 2004. Accepted November 24, 2004.

Meta-analysis is becoming a common tool for combining results of independent studies in evolutionary biology (Møller and Thornhill 1997; 1998; Arnqvist and Nilsson 2000; Reed and Frankham 2001; Crnokrak and Barrett 2002; Møller and Jennions 2002; Coltman and Slate 2003; Garamszegi and Møller 2004). Literature reviews, especially quantitative reviews such as meta-analysis, have the potential to be affected by publication bias (Møller and Jennions 2001 Jennions and Møller 2002a,b), the selective publication of articles showing certain types of results over those showing other types of results. The most commonly suspected publication bias is the tendency for journals to only publish studies with statistically significant results; the lack of nonsignificant published studies has been termed the ''file-drawer problem'' (Rosenthal 1979). This bias in publication will lead to an overestimate of the number of significant results on a given topic. A number of methods have been suggested to identify, model, and deal with publication bias (Rosenthal 1979; Orwin 1983; Begg 1985; Hedges and Olkin 1985; Begg and Berlin 1988; Hedges 1992; Begg and Mazumdar 1994; Hedges and Vevea 1996; Wang and Bushman 1998; Palmer 1999; Duval and Tweedie 2000a,b). One of the simplest is the calculation of a fail-safe number. A fail-safe number indicates the number of nonsignificant, unpublished (or missing) studies that would need to be added to a meta-analysis to reduce an overall statistically significant observed result to nonsignificance. If this number is large relative to the number of observed studies, one can feel fairly confident in the summary conclusions. Fail-safe numbers are not necessarily the best way to approach publication bias, but they are a simple first step that can help identify whether more complex approaches are necessary.

The original and most commonly used fail-safe calculation was suggested by Rosenthal (1979). This method calculates the significance of multiple studies by calculating the significance of the mean *Z*-score (the mean of the standard normal deviates of each study). Rosenthal's method calculates the number of additional studies, N_R, with mean null result necessary to reduce the combined significance to a desired α level (usually 0.05). N_R is calculated as

$$N_R = \frac{\left[\sum Z(p_i)\right]^2}{Z_\alpha^2} - n, \quad (1)$$

where n is the number of studies, $Z(p_i)$ are the *Z*-scores for the individual significance values, and Z_α is the one-tailed *Z*-score associated with the desired α.

An alternative method, proposed by Orwin (1983), is based on Cohen's *d* (1969), an effect size estimate that measures the standardized difference between treatment and control means, although it is equally applicable to other similar measures of effect size. It calculates the number of additional studies, N_O, needed to reduce an observed mean effect size to a desired minimal effect size. N_O is calculated as

$$N_O = \frac{n(\bar{E}_o - \bar{E}_m)}{\bar{E}_m - \bar{E}_n}, \quad (2)$$

where n is the number of studies, $\bar{E}_O$ is the mean of the original n studies, $\bar{E}_n$ is the mean of the additional N_O studies, and $\bar{E}_m$ is the desired minimal mean effect size. Cohen (1969) defines a standardized mean difference effect size of 0.2 as ''small,'' 0.5 as ''medium,'' and 0.8 as ''large.'' Generally, the minimal effect size chosen using Orwin's calculation is 0.2 (e.g., VanderWerf 1992).

These methods have a number of disadvantages. The first is that they are both explicitly unweighted. One of the primary attributes of contemporary meta-analysis is weighting; studies with large sample size or small variance are given higher weight than those with small sample sizes or large variance. Neither method accounts for the weight of the observed or the hypothesized unpublished studies. A second problem with Rosenthal's method is that the method of adding *Z*-scores is not normally the method by which one combines studies in a meta-analysis; most modern meta-analyses are based on the combination of effect sizes, not simply significance values (Rosenberg et al. 2000). Rosenthal's calculation is therefore not precisely applicable to the actual significance obtained

from a meta-analysis. Orwin's method is not based on significance testing; the choice of a desired minimal effect size to test the observed mean against seems unstable without a corresponding measure of variance.

Because the above methods do not contain any specification of the sample size of the studies involved, L'Abbé et al. (1987) suggested simulating a single study of large negative effect and determining the sample size of that study necessary to raise an observed significance above 0.05. Similarly, they suggested simulating several relatively small studies of no effect and calculating how many of these studies it would take to raise an observed significance above 0.05.

Here I illustrate an approach for directly calculating a fail-safe number that is explicitly grounded in the way meta-analyses are normally conducted, including the methods by which we calculate mean effect sizes, weighting, and significance testing (software for performing these analyses is available at the author's website: http://lsweb.la.asu.edu/rosenberg). In general, we begin a meta-analysis with n independent studies, each with an observed effect size E_i and variance s_i^2. The nature of the effect size depends on the parameters being studied but is commonly a Z-transformed correlation coefficient, a standardized difference between means from a control and experiment, or an estimate of risk from a 2×2 table, such as the odds ratio, risk difference, or relative risk (Rosenberg et al. 2000). Each study will be weighted by the inverse of its variance, $w_i = 1/s_i^2$. The mean effect size is calculated as a weighted average: the sum of the product between individual effects and their weights divided by the sum of the weights,

$$\bar{E} = \frac{\sum w_i E_i}{\sum w_i}, \tag{3}$$

with variance

$$s_{\bar{E}}^2 = \frac{1}{\sum w_i}. \tag{4}$$

Note that this generalization is true regardless of the effect size being used (Rosenberg et al. 2000). The typical way of testing whether this mean differs from some null value O would be via a t-test:

$$t_s = \frac{\bar{E} - O}{s_{\bar{E}}}. \tag{5}$$

For most metrics, the null O equals zero, leading to the simplification:

$$t_s = \frac{\bar{E}}{s_{\bar{E}}}. \tag{6}$$

Substituting equations (3) and (4), we find:

$$t_s = \frac{\sum w_i E_i \big/ \sum w_i}{\sqrt{1 \big/ \sum w_i}} \quad \text{or}$$

$$t_s^2 = \frac{\left(\sum w_i E_i\right)^2 \big/ \left(\sum w_i\right)^2}{1 \big/ \sum w_i} = \frac{\left(\sum w_i E_i\right)^2}{\sum w_i}. \tag{7}$$

When calculating a fail-safe number, the question being asked is how many new studies of mean null effect (zero) would need to be added to the analysis to produce a t-score at a desired significance level α (e.g., 0.05)? If the mean effect of the new studies is zero, and W' is the amount of additional weight needed to produce the desired α level, equation (7) would become:

$$t_{\alpha(v)}^2 = \frac{\left(\sum w_i E_i\right)^2}{W' + \sum w_i}. \tag{8}$$

Solving for W',

$$W' = \frac{\left(\sum w_i E_i\right)^2}{t_{\alpha(v)}^2} - \sum w_i, \tag{9}$$

where v represents the degrees of freedom (see below). Dividing W' by the mean weight $[(\Sigma w_i)/n]$ yields

$$N = \frac{nW'}{\sum w_i}. \tag{10}$$

N is equivalent to the number of studies of null effect and mean weight necessary to reduce the observed significance level to α. N could also be interpreted as the relative size of a single study of no effect needed to reduce the significance level to α, where relative size means that the single study would need to be weighted N times the mean weight. Because we expect weights to be roughly proportional to the sample size of each study (for some metrics this is exactly the case), N could also be thought of as an estimate of how many times larger the sample size of a single study (compared to the mean sample size) would need to be to reduce the significance of the mean effect to α.

One immediate complication is that the degrees of freedom of the t-test (v) is based on the number of studies used to construct the mean. If N is being interpreted as the relative size of a single study $v = n - 1 + 1 = n$. However, if N is being interpreted as multiple studies of mean weight, $v = n + N - 1$. In this case, N must be solved for iteratively (estimate N with the original degrees of freedom, then recalculate with the adjusted degrees of freedom, and repeat until N stabilizes). Because variation in t is quite small with even moderate degrees of freedom, only a few iterations are required for convergence. To differentiate between the interpretations of N, the single study value will be designated N_1; the multiple study value N_+. Although one could avoid the degrees-of-freedom issue by using the standard normal distribution (as is usually suggested in the older meta-analysis literature), the sample size in many meta-analyses can be quite small and it seems wise to err conservatively.

A primary assumption of N_+ is that the missing studies have similar sample sizes to those that were included in the original analysis. This may not be true; in fact, it is generally assumed that the majority of unpublished studies have smaller sample sizes than those found in the literature. This would make N_+ a minimal estimate because smaller studies would naturally lead to even larger fail-safe estimates. However, one can easily adjust the iterative procedure to find the number of necessary unpublished studies with any desired fraction of the mean weight.

TABLE 1. Fail-safe numbers calculated by the described methods for data on microsatellite measures of inbreeding from Coltman and Slate (2003). Only a subset of their significant results are presented. Mean effect sizes are estimated as *Z*-transformed correlation coefficients under a fixed-effects model (mean effects sizes under the random-effects model are not shown); *n* indicates the number of studies; N_R is Rosenthal's fail-safe number; N_1 was calculated for a single additional study; N_+ was calculated for many studies of mean weight. The desired $\alpha = 0.05$ for all methods. An asterisk indicates the fail-safe number is robust ($> 5n + 10$); a dash indicates the random-effects model collapses to the fixed model; n.s. indicates the effect size estimate was not significant. Most discrepancies between the numbers below and those in the original publication are due to rounding.

				Fixed		Random	
Variable	*n*	Fixed effect	N_R	N_1	N_+	N_1	N_+
Multilocus heterozygosity							
All traits	115	0.0274	2273*	530	541	323	—
Life history	47	0.0858	1274*	637*	671*	184	—
Life history (mammal)	24	0.0982	383*	198*	220*	48	—
Published studies	50	0.1088	1916*	999*	1049*	219	—
Mean d^2							
All traits	110	0.0156	779*	77	79	67	85
Life history	54	0.0479	472*	171	179	55	—
Life history (mammal)	21	0.0453	34	11	13	n.s.	n.s.
Published studies	38	0.0816	696*	291*	311*	55	—

Another assumption of these methods is the use of a fixed-effects model (Gurevitch and Hedges 1993; Hedges and Vevea 1998; Overton 1998; Rosenberg et al. 2000). Use of the alternative, the random-effects model, can be quite controversial (Gurevitch and Hedges 1993; Greenland 1994; Raudenbush 1994). In a random-effects model, individual studies are not weighted by simply the inverse of the variance, but rather as

$$w_i' = \frac{1}{s_i^2 + \sigma^2_{\text{pooled}}}, \tag{11}$$

where σ^2_{pooled} is an estimate of the pooled variance. All other calculations (e.g., mean effect and variance) proceed identically using this new weight. For a general meta-analysis with no internal data structure, the pooled variance is estimated as

$$\sigma^2_{\text{pooled}} = \frac{Q_T - (n - 1)}{\sum w_i - \dfrac{\sum w_i^2}{\sum w_i}}, \tag{12}$$

where Q_T is the total heterogeneity, measured as

$$Q_T = \sum w_i(E_i - \bar{E})^2. \tag{13}$$

The original fail-safe calculations are based on the fixed-effect model; using a similar approach we can estimate a fail-safe number for random-effects model meta-analysis. However, the necessity of estimating the pooled variance complicates the determination of a fail-safe number. Although *N* is generally thought to be the number of studies of null effect necessary to change a significant outcome, in the fixed-effects calculations described above, these studies need only have an average effect of zero. Because the random-effects model involves a sum-of-squares calculation (in the determination of Q_T), we need to explicitly assume all of the missing studies have effects that are precisely zero. This assumption could be avoided in part by simulating missing studies with a desired variance, although this raises additional complications and will not be explored further at this time.

Accepting the additional assumption that all missing studies have an effect of exactly zero, we can solve for the random-effects fail-safe number iteratively, using either the N_1 or N_+ approach described above, recalculating Q_T and σ^2_{pooled} each step (again, convergence usually occurs after only a few iterations). In practice, however, an additional complication quickly develops. The random-effects model operates under the assumption that there is true variance around the grand mean effect size, thus the incorporation of σ^2_{pooled} into the weights. The incorporation of this value in the determination of the weights can drastically increase the variance of the mean effect size; thus, one usually expects a random-effects model to have a much smaller fail-safe number than the same data in a fixed-effects model. However, when the estimate of σ^2_{pooled} equals zero (or is negative), the analysis collapses to a fixed-effects model (Hedges and Olkin 1985). Examination of the numerator of equation (12) shows that σ^2_{pooled} will become negative as the increase in the number of studies outpaces the increase in Q_T. Because the hypothesized studies being added to the analysis all have an effect of zero, as studies are added the grand mean will approach zero, rapidly reducing the increase in Q_T relative to the increase in *n*. In practice, the random-effects fail-safe number calculation of N_+ will often collapse to a fixed-effects model (this is less of a problem with N_1).

Because they are directly derived from the methods used in modern meta-analysis, the calculations described above are necessarily more reasonable predictors of the fail-safe number than the original approaches of Rosenthal (1979) and Orwin (1983). In practice, Rosenthal's fail-safe number overestimates the number of studies needed to reduce a meta-analysis to nonsignificance. Table 1 illustrates fail-safe numbers for a set of meta-analyses on microsatellites and inbreeding recently published in *Evolution* (Coltman and Slate 2003). Only a subset of the analyses for which the mean effect was significantly different from zero are included. The fail-safe numbers calculated with the new formula are often significantly lower than those reported in Coltman and Slate (2003) using Rosenthal's method. A fail-safe number is often considered robust if it is greater than $5n + 10$, where *n* is

the original number of studies (Rosenthal 1991). Many of the fail-safe numbers considered robust using Rosenthal's method in the original paper fail this criterion using the new calculations (the criterion is fairly arbitrary; it is difficult to consider a fail-safe number requiring more than 500 missing studies unrobust). It must be noted that the smaller fail-safe numbers do not imply that the results or conclusions of Colman and Slate (2003) are incorrect; fail-safe numbers are an attempt to judge the robustness of results against publication bias. Coleman and Slate account for publication bias to a certain extent by including unpublished results in their study (which did, in fact, have a much smaller average effect size than published studies). As expected, random-effects model fail-safe numbers (when they can be calculated) are usually quite a bit smaller than their fixed-effects model equivalents. Another recent meta-analysis in *Evolution* examined the correlation between quantitative and molecular measures of genetic variation (Reed and Frankham 2001). From 71 studies, their average effect was 0.279 (this differs slightly from the reported number because the authors average correlation coefficients directly rather than through Fisher's *Z*-transformation). The fail-safe numbers for this study (not reported by the authors) are $N_R = 8473$, $N_1 = 7586$, and $N_+ = 7851$, all quite high. In contrast, if we were to use a random-effects model rather than the fixed-effects model, both N_1 and N_+ become a rather paltry 8.

One needs to remember that a fail-safe calculation is neither a method of identifying publication bias nor a method of accounting for publication bias that does exist. It is simply a procedure by which one can estimate whether publication biases (if they exist) may be safely ignored. Many approaches to modeling and identifying publication bias have been and continue to be developed (Begg 1985, 1994; Hedges and Olkin 1985; Begg and Berlin 1988; Hedges 1992; Begg and Mazumdar 1994; Hedges and Vevea 1996; Wang and Bushman 1998; Duval and Tweedie 2000a,b). While fail-safe numbers are not the best approach to dealing with publication bias, they have certain advantages, including simplicity and intuitive appeal. Many of the other methods require one to estimate the potential bias by modeling the probability of publication as a function of significance or sample size. Specification of this function gives an estimate of the degree to which publication bias may exist in a given dataset; this bias may then be accounted for by complex factoring of the modeled missing publications. A more recent approach, the trim-and-fill method (Duval and Tweedie 2000a,b), assumes the data is symmetrically distributed around the mean in a funnel plot and models missing studies to symmetrize observed asymmetric distributions. While perhaps not as elegant as some of these methods, a fail-safe number is much simpler to calculate. Hopefully, the approach presented here will allow us to better estimate the potential for unpublished or missing studies to alter our conclusions; a low fail-safe number should certainly encourage researchers to pursue the more complicated publication bias methodologies.

Acknowledgments

Thanks to L. Hedges, J. Gurevitch, and anonymous reviewers for comments and suggestions on early versions of this manuscript, and to D. Coltman and J. Slate for providing their data for reanalysis.

Literature Cited

Arnqvist, G., and T. Nilsson. 2000. The evolution of polyandry: multiple mating and female fitness in insects. Anim. Behav. 60: 145–164.

Begg, C. B. 1985. A measure to aid in the interpretation of published clinical trials. Stat. Med. 4:1–9.

———. 1994. Publication bias. Pp. 399–409 *in* H. Cooper and L. V. Hedges, eds. The handbook of research synthesis. Russell Sage Foundation, New York.

Begg, C. B., and J. A. Berlin. 1988. Publication bias: a problem in interpreting medical data. J. R. Stat. Soc. A 151:419–463.

Begg, C. B., and M. Mazumdar. 1994. Operating characteristics of a rank correlation test for publication bias. Biometrics 50: 1088–1101.

Cohen, J. 1969. Statistical power analysis for the behavioral sciences. Academic Press, New York.

Coltman, D. W., and J. Slate. 2003. Microsatellite measures of inbreeding: a meta-analysis. Evolution 57:971–983.

Crnokrak, P., and S. C. H. Barrett. 2002. Perspective: Purging the genetic load: a review of the experimental evidence. Evolution 56:2347–2358.

Duval, S., and R. Tweedie. 2000a. A non-parametric ''trim and fill'' method of assessing publication bias in meta-analysis. J. Am. Stat. Assoc. 95:89–98.

———. 2000b. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 56:455–463.

Garamszegi, L. Z., and A. P. Møller. 2004. Extrapair paternity and the evolution of bird song. Beha. Ecol. 15:508–519.

Greenland, S. 1994. A critical look at some popular meta-analytic methods. Am. J. Epidemiol. 140:290–296.

Gurevitch, J., and L. V. Hedges. 1993. Meta-analysis: combining the results of independent experiments. Pp. 378–398 *in* S. M. Scheiner and J. Gurevitch, eds. Design and analysis of ecological experiments. Chapman and Hall, New York.

Hedges, L. V. 1992. Modeling publication selection effects in meta-analysis. Stat. Sci. 7:246–255.

Hedges, L. V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, San Diego, CA.

Hedges, L. V., and J. L. Vevea. 1996. Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. J. Educ. Behav. Stat. 21: 299–332.

———. 1998. Fixed- and random-effects models in meta-analysis. Psychol. Methods 3:486–504.

Jennions, M. D., and A. P. Møller. 2002a. Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. Biol. Rev. 77:211–222.

———. 2002b. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. Proc. R. Soc. Lond. B 269:43–48.

L'Abbé, K. A., A. S. Detsky, and K. O'Rourke. 1987. Meta-analysis in clinical research. Ann. Internal Med. 107:224–233.

Møller, A. P., and M. D. Jennions. 2001. Testing and adjusting for publication bias. Trends Ecol. Evol. 16:580–586.

———. 2002. How much variance can be explained by ecologists and evolutionary biologists? Oecologia 132:492–500.

Møller, A. P., and R. Thornhill. 1997. A meta-analysis of the heritability of developmental stability. J. Evol. Biol. 10:1–16.

———. 1998. Bilateral symmetry and sexual-selection: a meta-analysis. Am. Nat. 151:174–192.

Orwin, R. G. 1983. A fail-safe *N* for effect size in meta-analysis. J. Educ. Stat. 8:157–159.

Overton, R. C. 1998. A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. Psychol. Methods 3:354–379.

Palmer, A. R. 1999. Detecting publication bias in meta-analysyes:

a case study of fluctuating asymmetry and sexual selection. Am. Nat. 154:220–233.
Raudenbush, S. W. 1994. Random effects models. Pp. 301–321 *in* H. Cooper and L. V. Hedges, eds. The handbook of research synthesis. Russell Sage Foundation, New York.
Reed, D. H., and R. Frankham. 2001. How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. Evolution 55:1095–1103.
Rosenberg, M. S., D. C. Adams, and J. Gurevitch. 2000. MetaWin: statistical software for meta-analysis. Sinauer Associates, Sunderland, MA.
Rosenthal, R. 1979. The ''file drawer problem'' and tolerance for null results. Psychol. Bull. 86:638–641.
———. 1991. Meta-analytic procedures for social research. Sage, Newbury Park, CA.
Vander Werf, E. 1992. Lack's clutch size hypothesis: an examination of the evidence using meta-analysis. Ecology 73: 1699–1705.
Wang, M. C., and B. J. Bushman. 1998. Using the normal quantile plot to explore meta-analytic data sets. Psychol. Methods 3: 46–54.

Corresponding Editor: C. Goodnight