Asymptotic Approximations for the Radial Plot in Meta Analysis, and a Bias Correction to the Egger Test.

John Copas *and Claudia Lozada
University of Warwick, UK

Summary

Fixed effects meta analysis can be thought of as least squares analysis of the radial plot, the plot of standardized treatment effect against precision (reciprocal of the standard deviation) for the studies in a systematic review. For example, the Egger test for publication bias is equivalent to testing the significance of the intercept in linear regression. In practice, both x- and y-coordinates of the points in a radial plot are subject to sampling error, which may be correlated, and so the standard theory of least squares does not apply. For the Egger test, the actual significance levels are inflated, sometimes substantially so. We derive approximations to the sampling properties of the radial plot, assuming that the within-study sample sizes are large. This leads to an asymptotic bias correction for the Egger test. A simulation study suggests that the bias correction controls the significance level of the Egger test without any appreciable loss of power in detecting non-random study selection. A clinical trials example is used as an illustration.

Key words: Radial plot; Publication bias; Small study effects; Egger test; Asymptotic bias correction.

*Corresponding author: jbc@stats.warwick.ac.uk; phone 02476523370; fax 02476524532

1 Introduction: radial plots and small sample effects

The standard fixed effects model in meta analysis is that we have k separate studies, each reporting an estimate $\hat{\theta}$ of a common parameter θ. Each estimate (typically a treatment effect) is assumed to be independent and normally distributed

$$\hat{\theta} \sim N(\theta, \frac{\sigma^2}{n}), \tag{1.1}$$

where n is the sample size and σ^2 an underlying variance parameter. Note that θ is common (the fixed effects assumption) but n and σ^2 will usually vary across the studies. An immediate consequence of (1.1) is that the standardized treatment effects can be written as a linear regression, the so-called *Radial Plot* (Galbraith, 1988)

$$\frac{\hat{\theta}\sqrt{n}}{\sigma} = \alpha + \theta\frac{\sqrt{n}}{\sigma} + \epsilon, \tag{1.2}$$

where $\alpha = 0$ and ϵ is a standard normal residual. Ideally, the plot of standardized effects against study precision, measured by $\sqrt{n}/\sigma$, should be a straight line radiating out from the origin, with slope equal to the true value of θ. See Sutton *et al.* (2000) for a good introduction to meta analysis and radial plots.

Small studies (small values of $\sqrt{n}/\sigma$) will give the points on the radial plot closest to the origin, large studies (large $\sqrt{n}/\sigma$) will be the points furthest away from the origin. Thus, if there is a tendency for smaller studies to give larger estimates of θ than the larger studies, then the left hand side of the line will be pulled upwards resulting in a positive value of the intercept α. Conversely, if smaller studies tend to give smaller estimates of θ then the intercept will become negative. This is the motivation behind the *Egger test* for a small study effect: fit a line to the radial plot by ordinary least squares and test the fitted intercept $\tilde{\alpha}$ for a significant departure from zero (Egger *et al.*, 1997). Although the Egger test could use the fact that the residual variance in (1.1) is known to equal one, the sample residual variance is usually used in calculating the variance of $\tilde{\alpha}$. The test is then reasonably robust to modest heterogeneity between the studies. The Egger test is usually thought of as test for publication bias, as a tendency for publication to favour studies with significant results would typically result in non-zero value of α in (1.2).

In practice, the study-specific variances σ^2 will never be known, and so we have to use estimates $\hat{\sigma}^2$ when plotting radial plots and testing for small study effects. We thus plot y against x where

$$y = \frac{\hat{\theta}\sqrt{n}}{\hat{\sigma}} \quad \text{and} \quad x = \frac{\sqrt{n}}{\hat{\sigma}}. \tag{1.3}$$

The least squares estimates of θ and α in (1.2), and the test statistic for the Egger test, are respectively

$$\tilde{\theta} = \frac{c_{x,y}}{c_{x,x}} \quad , \quad \tilde{\alpha} = \bar{y} - \tilde{\theta}\bar{x} \quad \text{and} \quad t = \tilde{\alpha}\left\{\frac{s^2}{k}\left(1 + \frac{\bar{x}^2}{c_{x,x}}\right)\right\}^{-\frac{1}{2}}, \tag{1.4}$$

where $s^2 = k(c_{y,y} - c_{x,x}^{-1}c_{y,x}^2)/(k-2)$ is the usual unbiased estimate of the residual variance.

In this and in all following expressions we use the generic covariance notation

$$c_{x,y} = \frac{1}{k}\sum_{i=1}^{k}(y_i - \bar{y})(x_i - \bar{x}).$$

The Egger test refers the statistic t in (1.4) to the t-distribution on $(k-2)$ degrees of freedom. Notice that, throughout, our sample mean and covariance notation refers to variation *between* the k studies.

Replacing σ by $\hat{\sigma}$ in x and y upsets the assumptions necessary for the linear regression (1.2). The stochastic error in x may lead to bias in the parameter estimates, and the variability in $\hat{\sigma}$ may also induce a within-study correlation between x and y. Both effects mean that t no longer has a t-distribution. Several recent simulation studies (Macaskill *et al.*, 2001; Schwarzer *et al.*, 2002; Peters *et al.*, 2006; Harbord *et al.*, 2006) have shown that the true significance level of the Egger test can be substantially inflated, meaning that it rejects the null hypothesis that $\alpha = 0$ more often than it should, even when the study sample sizes are quite large.

The aim of our paper is to study the distribution of x and y, and to derive asymptotic approximations to sampling properties of the radial plot, asymptotic in the sense that the study-specific sample sizes are large. This leads to a relatively simple bias correction to the Egger test, which is easy to use in practice and which in most applications gives an effective correction for the inflation in significance level. As we shall see, the extent of these biases depends principally on the spread in the values of x across the studies, and on the sampling properties of the particular within-study estimates $\hat{\theta}$ and $\hat{\sigma}^2$. We discuss two contrasting examples:

Example 1. Here $\hat{\theta}$ is the sample mean of a random sample of size n from $N(\theta, \sigma^2)$ and $\hat{\sigma}^2$ is their sample variance. In this case, $\hat{\theta}$ and $\hat{\sigma}^2$ are independent which simplifies much of the theory developed below.

Example 2. This is the most common problem in meta analysis, when we have two treatments and a binary outcome, and hence two binomial distributions $\text{bin}(m_1, p_1)$ and $\text{bin}(m_2, p_2)$. Each study gives an estimated log-odds ratio $\hat{\theta}$ with variance σ^2/n based on a total sample size $n = m_1 + m_2$, where

$$\theta = \log\frac{p_1(1-p_2)}{(1-p_1)p_2}, \quad \sigma^2 = \gamma_1 + \gamma_2, \tag{1.5}$$

and

$$\gamma_j = \frac{1}{\alpha_j p_j(1-p_j)} \quad , \quad \alpha_j = \frac{m_j}{n};\ j = 1, 2. \tag{1.6}$$

In this case, the usual estimates of θ and σ^2 can be substantially correlated, leading to a serious inflation in the significance level of the Egger test.

In Section 2 we discuss some asymptotic properties of x and y for a single study, and then use these results to study asymptotic properties of the radial plot in Section 3. This leads to the proposed bias-corrected version of the Egger test in Section 4. In each of

these sections we illustrate the general set-up by looking at Examples 1 and 2. Example 2 (the meta analysis of 2×2 tables) is the main case of practical interest, and so in Section 5 we discuss this example in more detail, setting out the steps needed to calculate our bias correction. A case study from the Cochrane database is used as an illustration, with bootstrap estimates of the actual significance levels of the corrected and uncorrected Egger tests. Section 6 reports a more general simulation study. Some concluding comments are given in Section 7.

2 Asymptotic approximations for a single study

We now consider the joint distribution of $\hat{\theta}$ and $\hat{\sigma}^2$ for a single study. Model (1.1) states that $\hat{\theta}$ is asymptotically $N(\theta, \sigma^2/n)$; first we extend this to suppose that $(\hat{\theta}, \hat{\sigma}^2)$ is asymptotically jointly normal with mean (θ, σ^2). We assume that the biases in both estimates are of order $O(n^{-1})$, that $\mathrm{Var}(\hat{\theta}) = \sigma^2/n + O(n^{-2})$, and that $\mathrm{Var}(\hat{\sigma}^2) = O(n^{-1})$. These are standard properties of maximum likelihood estimates; in practice they will hold for any 'sensible' estimates of θ and σ^2.

We will be interested in the following three functions of x and y:

$$u = x - \frac{\sqrt{n}}{\sigma} = \sqrt{n}\left(\frac{1}{\hat{\sigma}} - \frac{1}{\sigma}\right), \quad v = y - \frac{\theta\sqrt{n}}{\sigma} = \sqrt{n}\left(\frac{\hat{\theta}}{\hat{\sigma}} - \frac{\theta}{\sigma}\right), \quad w = v - \theta u = \frac{(\hat{\theta} - \theta)\sqrt{n}}{\hat{\sigma}}. \tag{2.1}$$

Our approximations will involve two moments in particular, $\mathrm{E}(w)$ and $\mathrm{E}(uw)$, which from the above assumptions have orders of magnitude

$$\mathrm{E}(w) = O(n^{-\frac{1}{2}}) \quad \text{and} \quad \mathrm{E}(uw) = O(1). \tag{2.2}$$

We will also need to note that, again as a consequence of the above assumptions,

$$\mathrm{E}(u) = O(n^{-\frac{1}{2}}), \quad \mathrm{Var}(w) = 1 + O(n^{-1}) \quad \text{and} \quad \mathrm{Cov}(w, uw) = O(n^{-\frac{1}{2}}). \tag{2.3}$$

Example 1 (cont.) For the normal mean example, $\hat{\theta}$ and $\hat{\sigma}^2$ are independent, and $E(\hat{\theta} - \theta) = 0$. It follows immediately that $\mathrm{E}(w) = \mathrm{E}(uw) = 0$.

Example 2 (cont.) For the 2×2 table, first consider the case of a single binomial distribution, with frequency $f \sim \mathrm{bin}(m, p)$, say. Let z be the asymptotic (large m) standard normal deviate corresponding to f, with

$$z = \frac{f - np}{\sqrt{n/\gamma}} \quad \text{and} \quad \gamma = \frac{1}{p(1-p)}.$$

In estimating the log odds, $\log\{p/(1-p)\}$, we follow the standard practice of adding $\frac{1}{2}$ to each of the observed frequencies f and $m - f$, making the sample size effectively $(m + 1)$. Then after some tedious but straightforward algebra we find

$$\log\frac{f + \frac{1}{2}}{m - f + \frac{1}{2}} = \log\frac{p}{1-p} + \gamma^{\frac{1}{2}} z m^{-\frac{1}{2}} + \frac{1}{2}\gamma(1-2p)(1-z^2)m^{-1} + O_p(m^{-3/2}) \tag{2.4}$$

$$\frac{m+1}{(f + \frac{1}{2})(m - x + \frac{1}{2})} = \gamma m^{-1} - \gamma^{3/2}(1-2p)z m^{-3/2} + O_p(m^{-2}). \tag{2.5}$$

Extending this to two binomial distributions $f_1 \sim \text{bin}(m_1, p_1)$ and $f_2 \sim \text{bin}(m_2, p_2)$, with $n = m_1 + m_2$, our estimates of θ and σ^2 in (1.5) are

$$\hat{\theta} = \log \frac{f_1 + \frac{1}{2}}{m_1 - f_1 + \frac{1}{2}} - \log \frac{f_2 + \frac{1}{2}}{m_2 - f_2 + \frac{1}{2}} \tag{2.6}$$

$$\hat{\sigma}^2 = n \left\{ \frac{m_1 + 1}{(f_1 + \frac{1}{2})(m_1 - f_1 + \frac{1}{2})} + \frac{m_2 + 1}{(f_2 + \frac{1}{2})(m_2 - f_2 + \frac{1}{2})} \right\}. \tag{2.7}$$

We can now use (2.4) and (2.5) to expand $\hat{\theta}$ and $\hat{\sigma}^2$ in powers of $n^{-\frac{1}{2}}$ and in terms of two independent standard normal deviates z_1 and z_2. Recalling the notation γ_j in (1.6), this leads to the following approximations to the quantities u and w in (2.1),

$$u = \frac{1}{2}(\gamma_1 + \gamma_2)^{-3/2}\{\gamma_1^{3/2}(1 - 2p_1)z_1 + \gamma_2^{3/2}(1 - 2p_2)z_2\} + O_p(n^{-\frac{1}{2}}),$$

and

$$\begin{aligned} w &= (\gamma_1 + \gamma_2)^{-\frac{1}{2}}(\gamma_1^{\frac{1}{2}} z_1 + \gamma_2^{\frac{1}{2}} z_2) + \frac{1}{2} n^{-\frac{1}{2}} \left[2u(\gamma_1^{\frac{1}{2}} z_1 - \gamma_2^{\frac{1}{2}} z_2) \right. \\ &+ \left. (\gamma_1 + \gamma_2)^{-\frac{1}{2}}\{\gamma_1(1 - 2p_1)(1 - z_1^2) - \gamma_2(1 - 2p_2)(1 - z_2^2)\} \right] + O_p(n^{-1}). \end{aligned}$$

These give

$$\text{E}(w) = \frac{1}{2}(\gamma_1 + \gamma_2)^{-3/2}\{\gamma_1^2(1 - 2p_1) - \gamma_2^2(1 - 2p_2)\} n^{-\frac{1}{2}} + O(n^{-1}), \tag{2.8}$$

$$\text{E}(uw) = (\gamma_1 + \gamma_2)^{-\frac{1}{2}}\{n^{\frac{1}{2}}\text{E}(w)\} + O(n^{-\frac{1}{2}}). \tag{2.9}$$

3 Asymptotic properties of the radial plot

Now consider the distribution of the radial plot least squares estimates (1.4). Our aim is to derive asymptotic approximations which are valid on the assumption that all the study sample sizes $n_1, n_2, \cdots, n_k$ are large. To make this clear, let $N = \sum_1^k n_i$ and $\lambda_i = n_i/N$. Then we imagine that $N \to \infty$ while the λ_is remain fixed.

For each study, let

$$a = \frac{\lambda^{\frac{1}{2}}}{\sigma},$$

so that

$$x = \frac{\sqrt{n}}{\hat{\sigma}} = \frac{\lambda^{\frac{1}{2}} N^{\frac{1}{2}}}{\sigma} + u = aN^{\frac{1}{2}} + u, \tag{3.1}$$

and

$$y = \frac{\hat{\theta}\sqrt{n}}{\hat{\sigma}} = \frac{\theta\lambda^{\frac{1}{2}} N^{\frac{1}{2}}}{\sigma} + v = a\theta N^{\frac{1}{2}} + v. \tag{3.2}$$

Then

$$c_{x,x} = N c_{a,a} + 2N^{\frac{1}{2}} c_{a,u} + c_{u,u} \ ,$$

and

$$\begin{aligned} c_{x,y} &= \theta N c_{a,a} + N^{\frac{1}{2}}(c_{a,v} + \theta c_{a,u}) + c_{u,v} \\ &= \theta c_{x,x} + N^{\frac{1}{2}} c_{a,w} + c_{u,w}, \end{aligned}$$

where, as before, $w = v - \theta u$. Thus the least squares slope is

$$\tilde{\theta} = \frac{c_{x,y}}{c_{x,x}} = \theta + \frac{c_{a,w}}{c_{a,a}} N^{-\frac{1}{2}} + \frac{c_{u,w}c_{a,a} - 2c_{a,w}c_{a,u}}{c_{a,a}^2} N^{-1} + O_p(N^{-3/2}). \tag{3.3}$$

From (3.1) and (3.2), $\bar{x} = \bar{a}N^{\frac{1}{2}} + \bar{u}$ and $\bar{y} = \theta\bar{a}N^{\frac{1}{2}} + \bar{v}$, and so the least squares intercept becomes

$$\begin{aligned} \tilde{\alpha} &= \bar{y} - \tilde{\theta}\bar{x} \\ &= \bar{w} - \frac{c_{a,w}\bar{a}}{c_{a,a}} - \left\{ \frac{c_{a,w}\bar{u}}{c_{a,a}} + \frac{\bar{a}(c_{u,w}c_{a,a} - 2c_{a,w}c_{a,u})}{c_{a,a}^2} \right\} N^{-\frac{1}{2}} + O_p(N^{-1}). \end{aligned} \tag{3.4}$$

Now, for each study, let

$$b = \lim_{N\to\infty} N^{\frac{1}{2}}\mathrm{E}(w) \quad \text{and} \quad c = \lim_{N\to\infty} \mathrm{E}(uw).$$

From(2.2), all three study-specific quantities (a, b, c) are of order $O(1)$ in our approximations as $N \to \infty$. Then the expectations of the random quantities appearing in (3.3) and (3.4) can be written

$$\begin{aligned} \mathrm{E}(\bar{w}) &= \bar{b}N^{-\frac{1}{2}} + O(N^{-1}) \\ \mathrm{E}(c_{a,w}) &= k^{-1}\sum(a_i - \bar{a})\mathrm{E}(w) = c_{a,b}N^{-\frac{1}{2}} + O(N^{-1}) \\ \mathrm{E}(c_{a,w}\bar{u}) &= k^{-2}\sum(a_i - \bar{a})\mathrm{E}(u_i w_i) + O(N^{-1}) \\ &= k^{-1}c_{a,c} + O(N^{-1}) \\ \mathrm{E}(c_{u,w}) &= k^{-1}\sum \mathrm{E}(u_i w_i) - \mathrm{E}(\bar{u}\bar{w}) \\ &= (1 - k^{-1})\bar{c} + O(N^{-\frac{1}{2}}) \\ \mathrm{E}(c_{a,u}c_{a,w}) &= k^{-2}\sum(a_i - \bar{a})\mathrm{E}(u_i w_i) + O(N^{-\frac{1}{2}}) \\ &= k^{-1}(c_{(a-\bar{a})^2,c} + \bar{c}c_{a,a}) + O(N^{-\frac{1}{2}}). \end{aligned}$$

Substituting these into (3.3) and (3.4) gives

$$\mathrm{E}(\tilde{\theta}) = \theta + \frac{N^{-1}}{c_{a,a}} \left\{ c_{a,b} + \left(1 - \frac{3}{k}\right)\bar{c} - \frac{2c_{(a-\bar{a})^2,c}}{kc_{a,a}} \right\} + O(N^{-3/2}), \tag{3.5}$$

and

$$\begin{aligned} \mathrm{E}(\tilde{\alpha}) &= \frac{N^{-\frac{1}{2}}}{c_{a,a}} \left\{ c_{a,a}\bar{b} - c_{a,b}\bar{a} - \frac{c_{a,c}}{k} - \left(1 - \frac{3}{k}\right)\bar{a}\bar{c} + \frac{2\bar{a}c_{(a-\bar{a})^2,c}}{kc_{a,a}} \right\} + O(N^{-1}) \\ &= \alpha + O(N^{-1}) \ , \ \text{say.} \end{aligned} \tag{3.6}$$

Notice the different orders of magnitude of the biases in $\tilde{\theta}$ and $\tilde{\alpha}$. We can think of estimating the intercept of a linear regression as extrapolating from the observed values of x back to $x = 0$, a distance which here increases at the rate $N^{\frac{1}{2}}$ because of the factor $\sqrt{n}$ in the definitions of x and y. Hence the bias of order $O(N^{-1})$ in the slope, which is negligibly small if the study sample sizes are sufficiently large, is magnified into the much bigger bias of order $O(N^{-\frac{1}{2}})$ in the intercept. We see this in approximations (3.5) and (3.6).

Example 1 (cont.) Here $b = c = 0$ and so the leading bias terms in (3.5) and (3.6) are both zero. Although the estimates are not exactly unbiased, their biases are an order of magnitude smaller than in the general case.

Example 2 (cont.) From (2.8) and (2.9),

$$c = \mathrm{E}(uw) = \frac{\gamma_1^2(1-2p_1) - \gamma_2^2(1-2p_2)}{2(\gamma_1+\gamma_2)^2}. \tag{3.7}$$

Hence

$$b = N^{\frac{1}{2}}\mathrm{E}(w) = \frac{n^{\frac{1}{2}}\mathrm{E}(w)}{\lambda^{\frac{1}{2}}} = \frac{n^{\frac{1}{2}}\mathrm{E}(w)}{a\sigma} = \frac{c}{a}, \tag{3.8}$$

this last step following from (2.9) as $\sigma^2 = \gamma_1 + \gamma_2$.

4 Bias correction for the Egger test

From (3.4) we can write

$$\tilde{\alpha} = \frac{1}{k}\sum_1^k w_i\left(1 - \frac{\bar{a}(a_i - \bar{a})}{c_{a,a}}\right) + AN^{-\frac{1}{2}} + O_p(N^{-1}),$$

where A is the factor multiplying $N^{-\frac{1}{2}}$ in expression (3.4). Now A can be written as a linear combination of products of the form $u_i w_j$, and so from (2.3) we have

$$\mathrm{Cov}(w_i, A) = O(N^{-\frac{1}{2}}).$$

Hence, again using (2.3), we get

$$\mathrm{Var}(\tilde{\alpha}) = \sigma_\alpha^2 + O(N^{-1}),$$

where

$$\sigma_\alpha^2 = \frac{1}{k}\left(1 + \frac{\bar{a}^2}{c_{a,a}}\right). \tag{4.1}$$

Then the standardized value of $\tilde{\alpha}$ is

$$\frac{\tilde{\alpha} - \mathrm{E}(\tilde{\alpha})}{\sqrt{\mathrm{Var}(\tilde{\alpha})}} = \frac{\tilde{\alpha} - \alpha}{\sigma_\alpha} + O(N^{-1}). \tag{4.2}$$

We can think of (4.2) as a bias-corrected version of the Egger test statistic. To implement this in practice, we need to estimate the parameters α and σ_α, by $\hat{\alpha}$ and $\hat{\sigma}_\alpha$, say. We do

this by estimating the values of (a, b, c) for each study and then calculating the required averages and sums of squares and products.

Notice that if we replace a by $aN^{\frac{1}{2}} = \sqrt{n}/\sigma$ in (4.1) then σ_α is unchanged, and if we replace a by $aN^{\frac{1}{2}}$ in (3.6), we just remove the factor $N^{-\frac{1}{2}}$. But $\mathrm{E}(x) = aN^{\frac{1}{2}} + O(N^{-\frac{1}{2}})$, so one way of calculating these quantities is to simply replace a by x in the above formulae. If we do this, σ_α becomes exactly the same as the denominator of the original test statistic t in (1.4), except that the theoretical variance of one in (1.2) is replaced by the estimated residual variance s^2. We suggest that the sample residual variance of the radial plot is again used in the denominator, giving the bias-corrected Egger test statistic as

$$t^* = \frac{\tilde{\alpha} - \hat{\alpha}}{s\hat{\sigma}_\alpha}. \tag{4.3}$$

Example 1 (cont.) Here $b = c = 0$ in each study and so $\alpha = 0$. Thus no bias correction is needed in this case.

Example 2 (cont.) From (1.5),

$$aN^{\frac{1}{2}} = \frac{\sqrt{n}}{\sigma} = \sqrt{\frac{n}{\gamma_1 + \gamma_2}}.$$

This, together with (3.7) and (3.8), shows that $(aN^{\frac{1}{2}}, b, c)$ can all be written as functions of the values of (γ_1, γ_2) and hence of (p_1, p_2) in the k studies. Taking p_1 and p_2 equal to the corresponding observed relative frequencies therefore gives the simplest way of estimating these quantities. This means we would simply replace a by x and estimate c (and hence b) in the obvious way from the earlier formulae.

However, we get more stable estimates of (p_1, p_2) if we exploit the fixed effects assumption, that the pairs (p_1, p_2) are related through the common log-odds ratio θ. The constrained maximum likelihood estimates of p_1 and p_2 given that

$$\Theta(p_1, p_2) = \log \frac{p_1(1 - p_2)}{(1 - p_1)p_2} = \theta$$

are

$$p_1(\lambda) = \frac{f_1 + \lambda + \frac{1}{2}}{m_1 + 1} \quad \text{and} \quad p_2(\lambda) = \frac{f_2 - \lambda + \frac{1}{2}}{m_2 + 1}, \tag{4.4}$$

where the Lagrange multiplier λ is a solution of the quadratic equation $\Theta\{p_1(\lambda), p_2(\lambda)\} = \theta$. For consistency with the earlier definitions of x and y, we have again added one half onto all of the observed frequencies in these calculations. By examining the function $\Theta(p_1, p_2)$ it is easy to check that the quadratic equation for λ has two real solutions, and that it is uniquely the larger solution which gives values of p_1 and p_2 in $[0, 1]$. We suggest taking $\theta = \tilde{\theta} = \sum xy / \sum x^2$, the standard fixed effects estimate of θ, finding the corresponding pairs (p_1, p_2), and then using these to calculate $(aN^{\frac{1}{2}}, b, c)$.

5 Bias correction for 2×2 tables

As Example 2, the meta analysis of 2×2 tables, is the main case of practical interest (in fact tacitly assumed in almost all of the literature on the Egger test), we set out in more detail how our correction to the Egger test is calculated.

5.1 Calculation

As before, we assume we have k studies, each with data $f_1 \sim \text{bin}(m_1, p_1)$ and $f_2 \sim \text{bin}(m_2, p_2)$.

- First calculate the study-specific estimates $(\hat{\theta}, \hat{\sigma}^2)$ in (2.6) and (2.7), and hence the usual radial plot co-ordinates (x, y) in (1.3) and the least squares estimates $\tilde{\alpha}$ and s in (1.4).

- Let $\eta = \exp(\sum xy / \sum x^2)$. If $\eta \neq 1$, calculate the Lagrange multipliers

$$\begin{aligned} \lambda &= \{2(1-\eta)\}^{-1} \left[-\{f_1 + m_2 - f_2 + 1 + \eta(f_2 + m_1 - f_1 + 1)\} \right. \\ &+ \left. \left\{ \{f_1 + f_2 - m_2 + \eta(m_1 - f_1 - f_2)\}^2 + 4\eta(m_1+1)(m_2+1) \right\}^{\frac{1}{2}} \right], \end{aligned}$$

  and hence each study's estimate of (p_1, p_2) in (4.4). If $\eta = 1$ then we simply take $p_1 = p_2 = (f_1 + f_2 + 1)/(m_1 + m_2 + 2)$. We can now calculate the estimates $(\hat{\gamma}_1, \hat{\gamma}_2)$ from (1.6) and hence the following four quantities for each study:

$$\begin{aligned} e &= \sqrt{\frac{n}{\hat{\gamma}_1 + \hat{\gamma}_2}}, \\ c &= \frac{\hat{\gamma}_1^2(1 - 2\hat{p}_1) - \hat{\gamma}_2^2(1 - 2\hat{p}_2)}{2(\hat{\gamma}_1 + \hat{\gamma}_2)^2}, \\ b &= \frac{c}{e}, \\ d &= (e - \bar{e})^2. \end{aligned}$$

- Calculate the average values across the k studies of the quantities (b, c, d, e), and their empirical covariances $c_{b,e}$, $c_{c,e}$ and $c_{c,d}$. Then, from (3.6) and (4.1),

$$\begin{aligned} \hat{\alpha} &= \bar{b} - \frac{\bar{e}c_{b,e}}{\bar{d}} - \frac{c_{c,e}}{k\bar{d}} - \frac{(k-3)\bar{e}\bar{c}}{k\bar{d}} + \frac{2\bar{e}c_{c,d}}{k\bar{d}^2} \\ \hat{\sigma}_\alpha^2 &= \frac{1}{k}\left(1 + \frac{\bar{e}^2}{\bar{d}}\right). \end{aligned}$$

- The corrected Egger test statistic is now t^* in (4.3), namely

$$t^* = \frac{\tilde{\alpha} - \hat{\alpha}}{s\hat{\sigma}_\alpha}.$$

  This is the same as the usual Egger test statistic t in (1.4), except for the subtraction of the bias correction in the numerator and the use of e instead of x in the denominator. As before, we refer t^* to the t-distribution on $(k-2)$ degrees of freedom.

5.2 Example

To illustrate these calculations, we revisit the example used in Copas and Jackson (2004), a systematic review of $k = 14$ clinical trials into the effectiveness of corticosteroids in cases of premature birth (see the cited paper for further details, including a table giving the raw data). Here, θ is the log-odds ratio measuring the reduction in infant deaths for active treatment versus placebo. As there is no evidence of between-study heterogeneity, the standard fixed effects analysis is used to give an estimated log-odds ratio of -0.477 with standard error 0.118, a strongly significant treatment effect.

However the radial plot in Figure 1 shows evidence of a small study effect, with the smaller studies (points to the left of the graph) tending to report more negative estimates of θ (i.e. greater improvements in mortality). Two regression lines are shown on the plot. The solid line is constrained to pass through the origin, and has slope -0.477, the regression we would expect if the model (1.1) is correct. The dashed line is the usual (unconstrained) least squares linear regression. The substantial difference between these lines is again indicative of a small study effect. The dashed line has intercept $\tilde{\alpha} = -1.217$ and residual standard deviation $s = 0.872$. The conventional Egger test statistic in (1.4) is $t = -2.213$, giving a (2-sided) P-value of 0.047 on $k - 2 = 12$ degrees of freedom. Thus, taking the conventional significance level as 5%, the Egger test indicates a significant small study effect.

Estimating the quantities required for our asymptotic approximations, we find $\hat{\alpha} = -0.043$ and $\hat{\sigma}_\alpha = .622$, giving $t^* = -1.990$, a (2-sided) P-value of 0.070. The test statistic has shrunk by 10%. In this case, the difference between t and t^* is quite small, but enough to cast doubt on the alleged significance of a small study effect.

However, the sample sizes in some of these trials are small (only 23 and 22 in one study) and the probabilities p_1 and p_2 are also small (the numbers of infant deaths observed in this small trial are 0 and 1). Such sparse data raise doubts about the accuracy of our asymptotic approximations. As a check, Table 1 reports the results of a limited parametric bootstrap simulation designed to assess the actual significance levels of these tests. We have generated independent samples from 14 pairs of binomial distributions, taking the sample sizes as the actual values of m_1 and m_2 and the probabilities as $p_1 = 0.079$ and $p_2 = 0.135$, these two numbers being the averages across the studies of the empirical estimates of p_1 and p_2. For each simulated data set we calculate t and t^* and compare these with the 95th and $97\frac{1}{2}$th percentiles of the t distribution on 12 degrees of freedom, namely 1.783 and 2.179. For example, the first numerical row of the table records the percentage of simulated data sets with $t \leq -1.783$ (left tail), with $t \geq +1.783$ (right tail), and with $|t| \geq 1.783$ (actual 2-tailed significance level), and similarly for the other rows in the table. These figures are based on $100,000$ simulations, giving Monte Carlo standard errors of 0.07 and 0.05 on the percentages shown for the 5% and $2\frac{1}{2}$% tests respectively. This suggests that most of the percentages in the table are correct to the accuracy given, with errors at most one in the first decimal place.

Comparing the first and last columns of the table, the 2-tailed significance levels of t are about 1% higher than nominal, whereas those of t^* match the nominal levels more closely. However, the major effect is the bias in $\tilde{\alpha}$, which shifts the intercept downwards leading to more false rejections in the left tail than the right. The bias correction in t^*

over-corrects for this, leading to an imbalance the other way round, although the two tails for t^* are more nearly equal than for t. The table suggests that even for relatively sparse data, the asymptotic corrections are reasonably effective in achieving nominal significance levels, equally divided between the left and right tails.

nominal (2-sided) significance level %	test	actual left tail %	actual right tail %	actual (2-sided) significance level %
10	t	6.8	4.4	11.2
	t^*	4.7	5.6	10.3
5	t	3.7	2.1	5.8
	t^*	2.3	2.9	5.2

Table 1. Bootstrap significance levels for the corticosteroids meta analysis.

The size of the bias correction depends on the characteristics of the studies in the review, through the quantities b and c in (3.7) and (3.8). Sometimes the bias correction is zero, for example when the trials are balanced ($m_1 = m_2$) and there is no treatment effect ($p_1 = p_2$), for in this case $b = c = 0$ and so $\mathrm{E}(\tilde{\alpha}) = 0$. In Section 6 we go on to consider a wider range of possibilities, some of which produce a much more marked distortion of significance levels than we see in Table 1.

6 Simulation study

In this section we present some further simulation results for 2×2 tables. Clearly it is impossible to cover all possibilities, but for Figs 2-4 below we have made the following choices:

- k, the number of studies, is between 10 and 100
- all studies are balanced with $m_1 = m_2 = m$, where the values of m are chosen randomly from the uniform distribution between 30 and 150
- $\theta = \log(0.67)$, an odds ratio of $2/3$ (or equivalently 1.5 if the labels on the treatments are reversed)
- p_E, defined to be the average of p_1 and p_2 on the logit scale, is 0.3 or 0.1. With $\theta = \log(.67)$ this means that (p_1, p_2) is either $(0.26, 0.34)$ (with $p_E = 0.3$) or $(0.08, 0.12)$ (with $p_E = 0.1$). In the second case the data can be quite sparse.

For each randomly generated set of 2×2 tables, the values of t in (1.4) and t^* in (4.3) are calculated as set out in Section 5. The simulations are repeated $10,000$ times.

Firstly, we illustrate the bias in t by looking at its distribution when $k = 50$. Figure 2 shows kernel density estimates for t and t^*, compared with the density of the t-distribution on $k - 2 = 48$ degrees of freedom. For both values of p_E the densities of t^* are virtually

indistinguishable from the t-distribution, but the densities of t are noticeably shifted to the left. The vertical lines represent the quantiles of the t-distribution for a two-sided test at the 10% level. At the smaller value of p_E nearly all the rejections for the test based on t will be in the left tail. Essentially, the quantity $\hat{\alpha}$ defined in Section 5 is estimating the degree of shift shown in Figure 2.

Figure 3 shows the actual Type I error rates for t and t^* for nominal 10% (two-sided) and 5% (one-sided) tests. In all four cases shown, the inflation in significance level of t increases as k increases, and can be very substantial. This is in line with the findings of Macaskill *et al.* (2001), Schwarzer *et al.* (2002), Peters *et al.* (2006) and Harbord *et al.* (2006). With $p_E = 0.3$ (when there are no small frequencies) the Type I error rates of t^* keep very close to the nominal levels. For more sparse data ($p_E = 0.1$), t^* overcompensates for the imbalance in the tails, as already noted in Table 1 above. In this case t^* gives an overly conservative one-sided test, but then the Egger test shows an even greater distortion in the opposite direction *i.e.* rejects the null hypothesis too often.

To investigate the power of the tests based on t and t^* we need to simulate data from models representing varying degrees of publication bias. Since, with $\theta < 0$, we would expect publication bias to result in a negative intercept for the radial plot, working with one-sided tests seems more meaningful. For each study, we simulate its 2×2 table and test to see if the treatment effect is significant (testing $H_0 : \theta = 0$ against $H_1 : \theta < 0$), using significance level α_p. We continue simulating 2×2 tables until H_0 is rejected. All the tables obtained in this way therefore show a significant treatment effect at the α_p level. If $\alpha_p = 1$ all tables are accepted and there is no publication bias. The smaller is α_p the more severe is the selection and hence the greater is the potential for publication bias.

It would clearly be misleading simply to compare the proportion of simulations in which the tests based on t and t^* reject the null hypothesis of no publication bias, because of the great imbalance in the Type I error rates. For a fair comparison of power, we need to be sure that they have the same probability of rejection when the null hypothesis is true. We define the "shifted Egger test" to be the one-tail test using t, but using a nominal significance α_s which may be different from the target significance level. From the above simulations we estimate α_s so that the *actual* Type I error rate of the Egger test is 5%. This means that we compare t against a percentile of the t-distribution which is further into the left tail than the 5th percentile.

Figure 4 compares the power functions of t^* and the shifted version of t, for $p_E = 0.3$ with $k = 10$ and $k = 50$. These graphs plot the probability of rejecting the hull hypothesis of no publication bias against the significance level α_p. The t^* test is less powerful, as expected because of the added variability from estimating the intermediate quantities needed in Section 5.1. However, the power functions shown in Figure 4 are quite similar, especially at $k = 50$, suggesting that the power of the Egger test is not substantially compromised by the proposed bias correction.

More generally, power plots like Figure 4 illustrate the difficulty of the whole idea of trying to test for publication bias. When $\alpha_p = 0.5$, which means that all studies with $\hat{\theta} > 0$ are filtered out, the probability of being able to detect this from the data is disappointingly low, even with 50 studies. When $\alpha_p = 0.1$, we would need at least 20 studies for there to be any reasonable chance of being able to detect that selection of this severity was taking place.

A strong assumption in these simulations so far is that p_1 and p_2 remain the same for all k studies. This is unrealistic: even the fixed effects model allows these probabilities to vary between studies provided θ is constant. We have rerun the above simulations with p_E generated randomly, with medians 0.3 and 0.1 and normal distributions for logit(p_E). With median 0.3 and coefficient of variation of logit(p_E) up to 50%, the results are almost indistinguishable from those already shown for fixed $p_E = 0.3$. For median 0.1 this also holds for coefficient of variation up to 20% — a higher variability for logit(p_E) in this case is unrealistic in that p_E can be very close to zero in which case both arms of the trial will have no observed events.

The above comparisons depend strongly on the value of θ. Our choice $\theta = \log(0.67)$ represents the kind of moderate treatment effect that one might hope to find in clinical trials or epidemiological studies. When $\theta = 0$, there is no bias in the radial plot and so t and t^* are essentially the same. We have rerun the simulations for the stronger treatment effect $\theta = \log(0.5)$, or equivalently $\theta = \log(2)$. The inflation in the significance level of t is then much more marked, for example with $k = 100$ and $p_E = 0.1$ the actual significance of the two-sided Egger test at the nominal 10% level is in fact 70%. However, the shape (but not the magnitude) of the various plots in Figs 2-4 remains broadly similar.

7 Discussion

1. Each point (1.3) of the radial plot can be thought of as an estimate of the true point $(\theta\sqrt{n}/\sigma, \sqrt{n}/\sigma)$ for that study. These true points become unbounded as $n \to \infty$. Thus the asymptotics in this paper are somewhat non-standard, unlike more usual asymptotic discussions where we are studying properties of estimates of fixed parameters. Note in particular that the covariance notation used in Sections 2-5 refers to variation across rather than within studies. For example, the variance $c_{a,a}$ in (3.3) is a positive quantity even through the study specific a's are defined as fixed parameters.

2. The radial plot is a key to several aspects of meta analysis, not just testing for publication bias. For example, the constrained least square slope is just the usual fixed effects estimate of θ, and the residual sum of squares gives the usual chi-squared test for heterogeneity. Thus, at least in principle, the asymptotic theory of Sections 2 and 3 allows us to address wider aspects of the bias inherent in nearly all of the literature in meta analysis, caused by the assumption that within study variances are known when in practice they are estimated.

3. The Egger test was proposed by Egger (1997) in a *British Medical Journal* article, and has been very widely used. However there have been many papers raising doubts about its statistical validity, including letters published in the same journal soon after the original paper (*e.g.* Irwig, *et al.*, 1998). We have retained the Egger test and its simple graphical motivation, whilst addressing its statistical problems with the proposed bias correction. An alternative approach, suggested in several recent papers, is to abandon the Egger test altogether and develop better tests for publication bias (for example Harbord *et al.*, 2005; Macaskill *et al.*, 2001; Peters *et al.*, 2006). All these tests are claimed to have a more stable Type I error rate than the Egger test.

4. The fixed effects assumption lies behind the motivation of the Egger test, and is

assumed in the theory of Section 3. But in practice the Egger test seems to be used with scant regard to whether a fixed or random effects model would be more appropriate. Using the estimated residual variance rather than the true variance in the denominator of t gives some robustness to heterogeneity, as confirmed by repeating the simulations of Section 6 with θ varying randomly with variance τ^2. We find that when τ^2 is small the sampling properties of t and t^* are very similar to those shown above for $\tau^2 = 0$. When τ^2 is large, unweighted least squares on the radial plot is no longer appropriate and we would need to rework the theory of Section 3. This would be much more challenging than the development here, particularly as estimates of τ^2 would depend on all the data and not just on the individual points in the radial plot.

5. An emphasis on testing for publication bias runs the risk of encouraging the belief that if there is no significant evidence of publication bias then no bias exists. This can be grossly misleading, particularly when k is small, as is often the case in practice. Study selection can have a marked effect on the estimate of θ without being detectable from the data, as evident from the power functions shown in Figure 4. Arguably, sensitivity analysis is a better approach, interpreted in the scientific context of each particular systematic review. See Rothstein *et al.* (2005) for a general discussion of publication bias and its importance in meta analysis.

References

Copas, J. B. and Jackson, D. (2004). A bound for publication bias based on the fraction of unpublished studies. *Biometrics*, **60,** 146-153.

Egger, M., Smith, G. D., Schneider, M. and Minder, C. (1997). Bias in meta analysis detected by a simple graphical test. *Brit. Med. J.*, **315**, 629-634.

Galbraith, R. F. (1988). A note on graphical representation of estimated odds ratios from several clinical trials. *Statist. in Med.*, **7**, 889-894.

Harbord, R. M., Egger, M. and Sterne, J. A. C. (2005). A modified test for small study effects in meta analysis of controlled trials with binary endpoints. *Statist. in Med.*, **25**, 3443-3457.

Irwig, L., Macaskill, P. and Berry, G. (1998). Bias in meta analysis detected by a simple graphical test. Graphical test is itself biased (letter to the editor). *Brit. Med. J.*, **316**, 469.

Macaskill, P., Walter, S. D. and Irwig, L. (2001). A comparison of methods to detect publication bias in meta analysis. *Statist. in Med.*, **20**, 641-654.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R. and Rushton, L. (2006). Comparison of two methods to detect publication bias in meta analysis. *J. Am. Med. Assoc.*, **295**, 676-680.

Rothstein, H. R., Sutton, A. J. and Borenstein, M. (eds) (2005). *Publication bias in meta-analysis*. Chichester: Wiley.

Shwarzer, G., Antes, G. and Shumacher, M. (2002). Inflation in Type I error rate in two statistical tests for the detection of publication bias in meta analysis with binary outcomes. *Statist. in Med.*, **21**, 2465-2477.

Sutton, A. J., Abrams, K. R., Jones, D. R. and Sheldon, T. A. (2000). *Methods for meta-analysis in medical research.* Chichester: Wiley.

List of Table and Figures

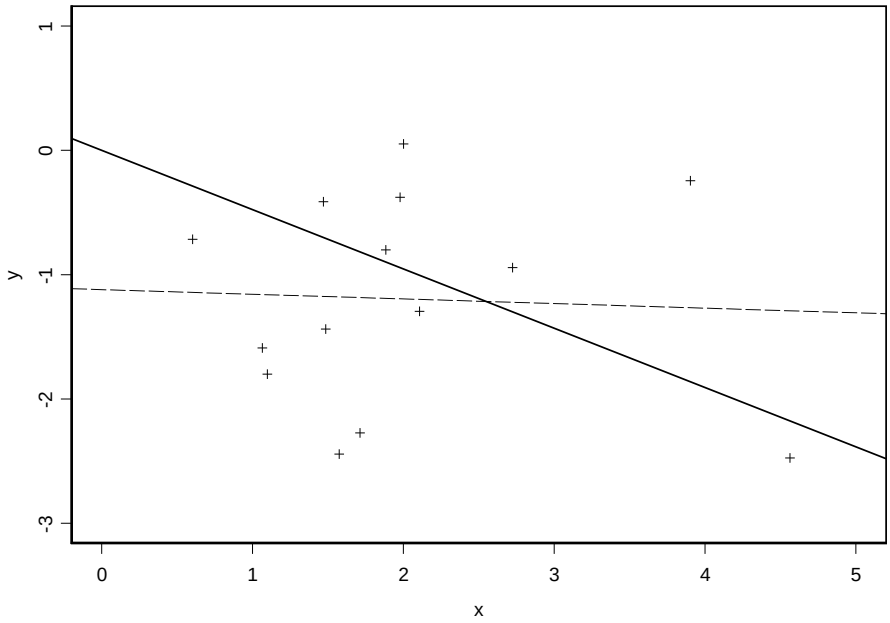

Figure 1: Radial plot for corticosteroids example

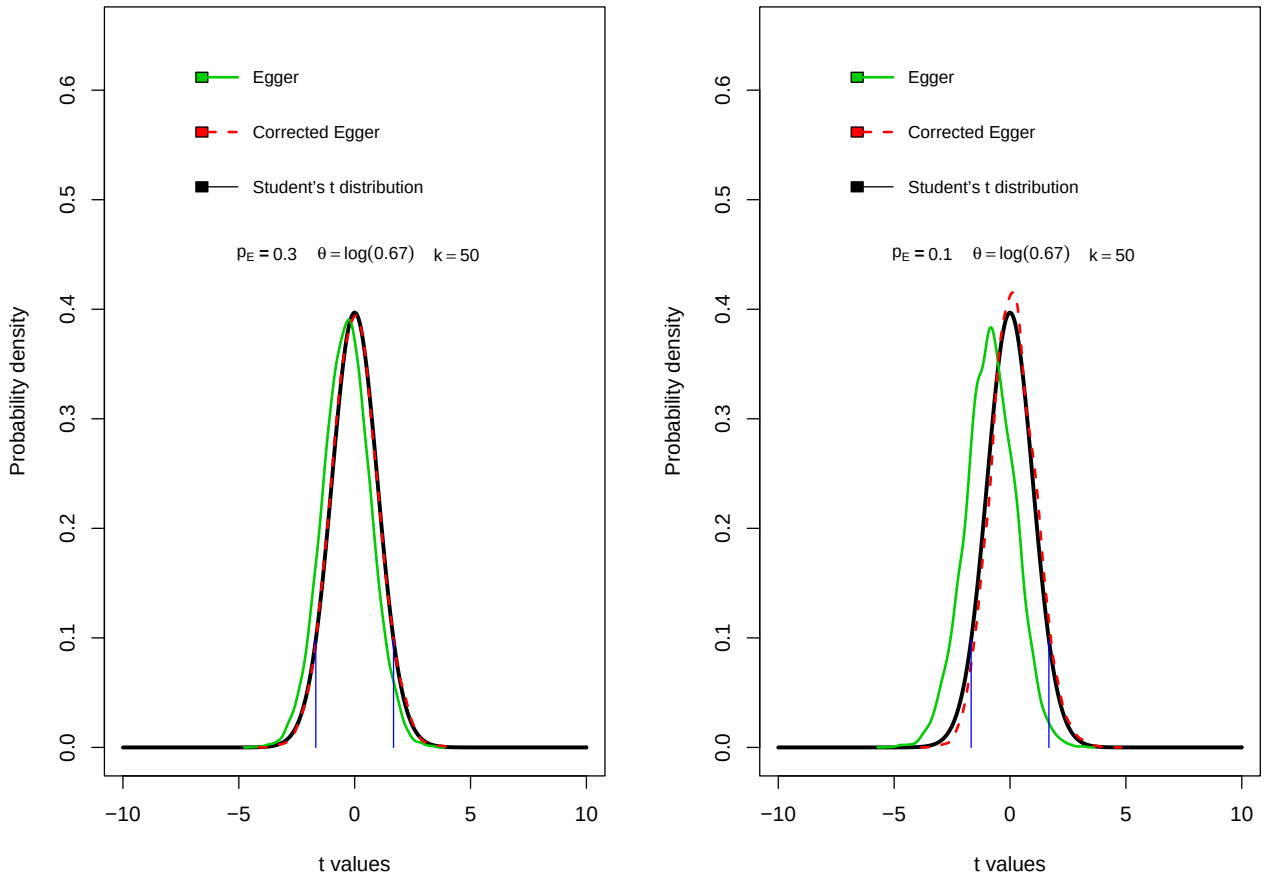

Figure 2: Distributions of t and t^*

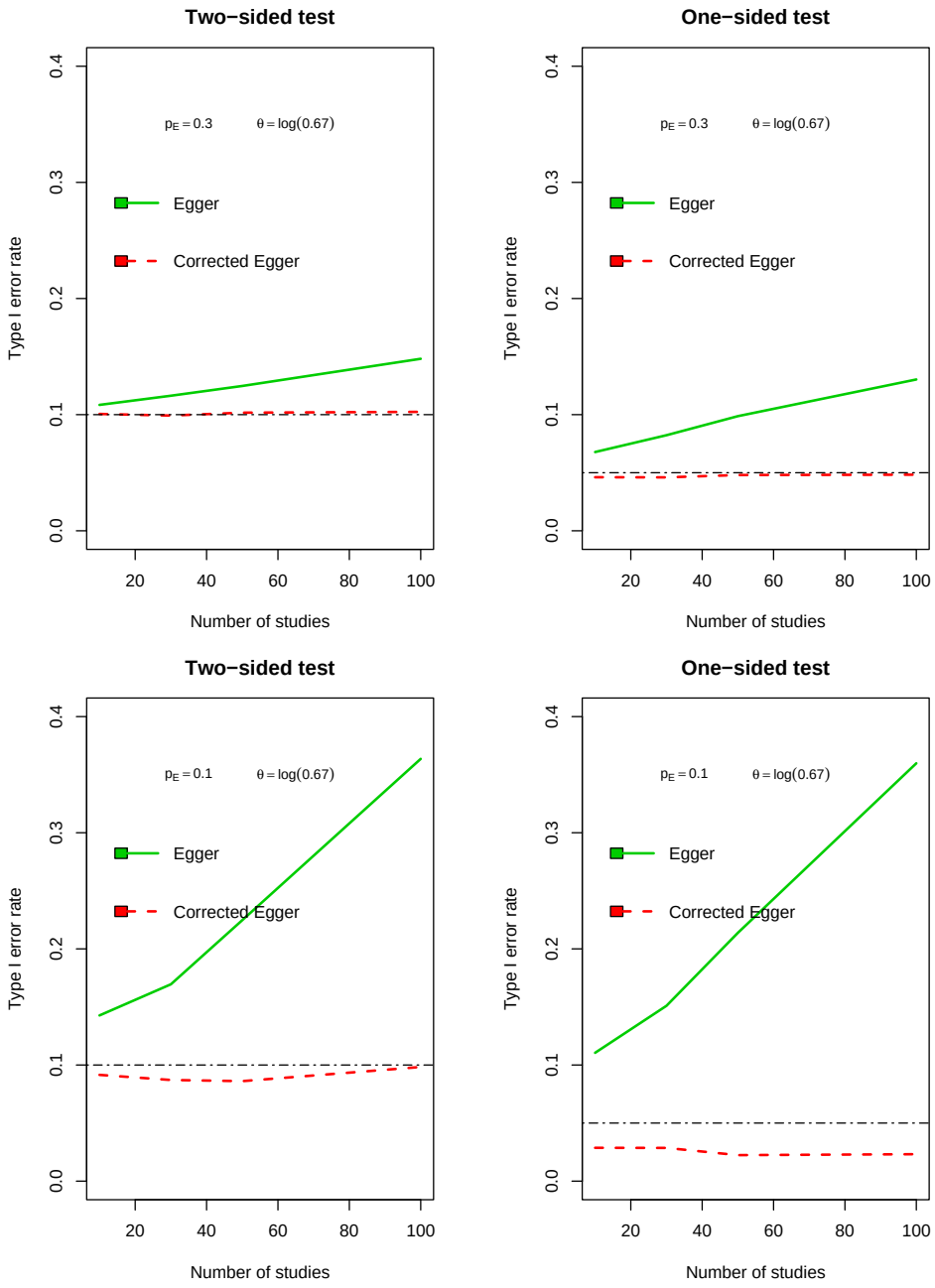

Figure 3: Actual Type I error rates for tests based on t and t^*

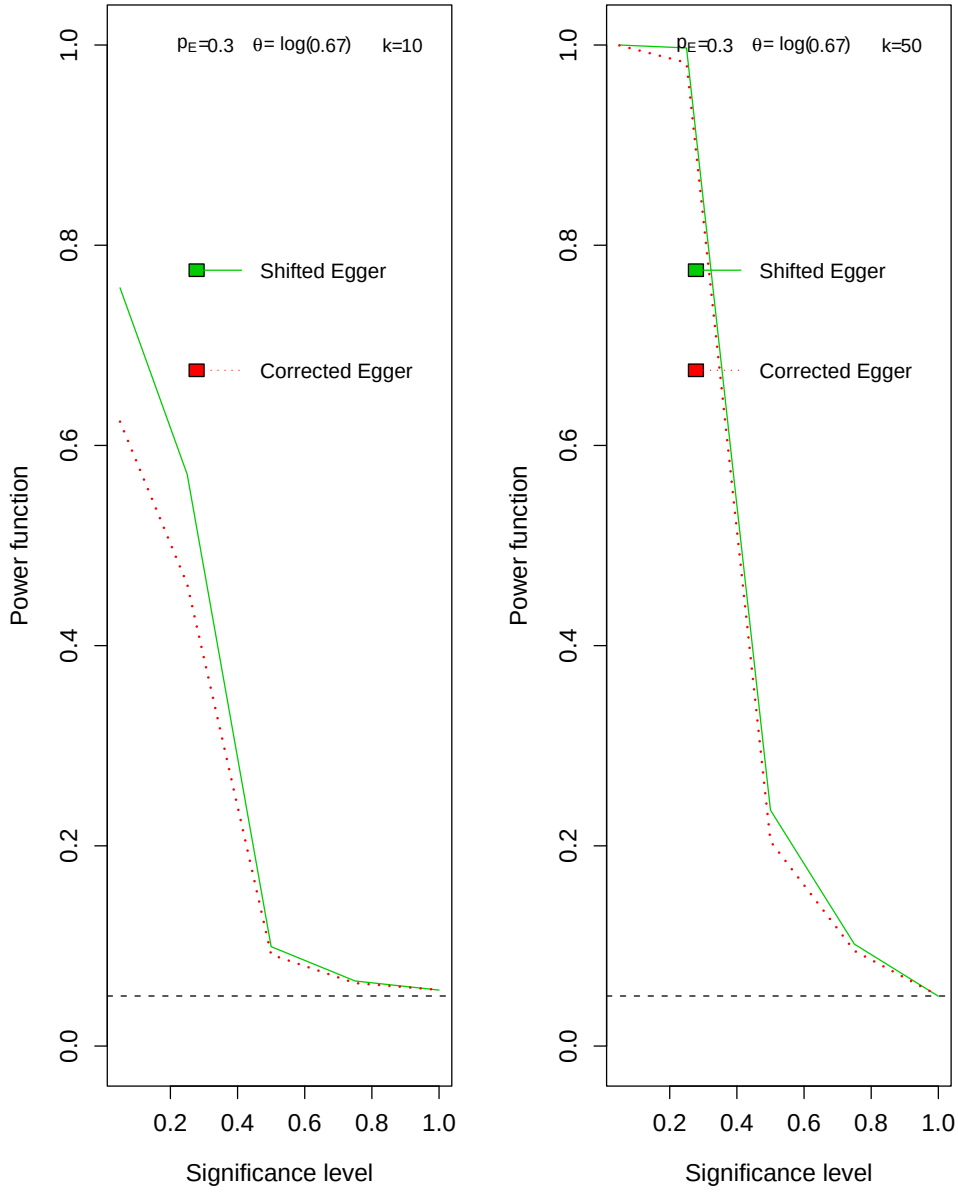

Figure 4: Power functions for t^* and shifted-t