The Power Functions of Begg's and Egger's Tests for Publication Bias

Haben Michael[1]

[1]University of Massachusetts, Amherst, MA (hmichael@math.umass.edu)

SUMMARY: Publication bias undermines the results of meta-analyses and other systematic reviews. Applying a formal test for publication bias is therefore a part of many protocols for conducting a meta-analysis. To be effective in this role, the tests must have good power to detect the presence of publication bias. We derive the local limiting power functions and study the Pitman relative efficiency for two popular tests, Begg's and Egger's tests. The analysis suggests that the differences between the two arise out of built-in robustness properties of Begg's test. These work to its detriment in the commonly used Gaussian model, but leave it with better efficiency when study effects are heavy-tailed or skew.

KEYWORDS: Meta-analysis; Publication bias; Local limiting power; Pitman efficiency.

1 Introduction

Meta-analysis is a popular method for combining the conclusions of a body of studies investigating the same question into a single conclusion. A premise of the method is that the studies used in the meta-analysis be representative of all studies conducted on the question under investigation. This premise is violated when publication bias is present. Publication bias is an association between the availability of a study and its conclusions, such as a tendency to publish studies with significant findings.

Protocols for conducting meta-analyses (Guyatt et al., 2011; Cooper, 2015; Page et al., 2021) generally advise the application of a formal test for publication bias before conducting a meta-analysis or before presenting the results of a meta-analysis. While the possibility of an inflated false positive rate of these tests has been studied extensively (Sterne et al., 2000; Peters et al., 2006), it is lower power that leads to poor FPR on the subsequent meta-analysis. As is often the case when hypothesis tests are applied at a preliminary stage, failure to reject the null is taken as grounds to proceed to the main analysis as though the null were in fact true. For example, following the application of a test that fails to reject the null of no publication bias, an analyst may proceed to carry out the meta-analysis without adjusting the body of studies. This procedure is incoherent unless the analyst can be confident in the power of the screening test.

We analyze the power of two of the most popular formal tests for publication bias, Begg's test (Begg and Mazumdar, 1994) and Egger's test (Egger et al., 1997). While there have been many studies of the power of Begg's and Egger's tests (see Sterne and Egger (2005) for an overview), important questions remain. There is little consensus on when one test is to be preferred over the other, and whether or to what extent these tests are redundant. Moreover, many of these studies have been based primarily on simulations and it is not clear how far their conclusions reach. There have also been empirical comparisons of the performance, e.g., Lin et al. (2018), but their conclusions are difficult to interpret since the "ground truth", whether or not publication bias is actually present, is unknown.

Therefore we offer an analytical contribution. We study Begg's and Egger's tests in a simple but prototypical publication bias model: rejecting or "filing away" studies with a large p-value. We find that Egger's test is to be preferred in the commonly assumed Gaussian model, while Begg's test is to be preferred with heavy-tailed or skew data.

1.1 Begg and Egger's tests

The data that goes into a meta-analysis consists of n pairs $(Y_1, \sigma_1^2), \dots, (Y_n, \sigma_n^2)$ representing the estimated effect sizes and sampling variances of n primary studies. In the basic model the study effects have a common mean $\theta \in \mathbb{R}$ and are assumed to be mutually independent conditionally on the sampling variances:

$$\begin{aligned} (Y_1, \sigma_1^2), \dots, (Y_n, \sigma_n^2) &\text{ independent} \\ E(Y_1, \dots, Y_n \mid \sigma_1^2, \dots, \sigma_n^2) &= \theta \mathbb{1}_n \\ \operatorname{Var}(Y_1, \dots, Y_n \mid \sigma_1^2, \dots, \sigma_n^2) &= \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2). \end{aligned} \tag{1}$$

Adding the assumption of Gaussian Y_i, this model is used for carrying out inference in the main meta-analysis.

Begg's and Egger's tests attempt to formalize earlier visual approaches to identifying publication bias. These visual approaches plot the study effects Y_i against their precisions $S_i = 1/\sigma_i$. See Fig. 1 for an illustration. An asymmetry with respect to the precision axis, such as effect sizes increasing with precision, is taken as a sign of possible publication bias. Begg's and Egger's tests formalize this procedure in two seemingly distinct ways.

Begg's test looks for a trend using the correlation between the standardized effect sizes and precisions. The test statistic is Kendall's rank correlation coefficient,

$$\hat{\tau} = \binom{n}{2}^{-1} \sum_{j<k} 2\mathbb{1}\left\{(u_j - u_k)(v_j - v_k) > 0\right\} - 1, \tag{2}$$

applied to measure the correlation between the centered and standardized effects and the study precisions,

$$(u_j, v_j) = \left(\frac{Y_j - \hat{\theta}}{\sqrt{\sigma_j^2 - \sigma_{\hat{\theta}}^2}}, \sigma_j \right), j = 1, \dots, n, \tag{3}$$

where

$$\hat{\theta} = \frac{\sum_{j=1}^n Y_j/\sigma_j^2}{\sum_{j=1}^n 1/\sigma_j^2} \tag{4}$$

is the inverse-variance weighted estimator of θ. The finite-sample corrections $\sigma^2_{\hat{\theta}} = 1/\sum_{j=1}^n (1/\sigma_j^2)$ are usually $O(1/n)$ and negligible to our analyses (Michael and Ghebremichael, 2023, Lemma 1), and will be ignored below.

Egger's test tests for a zero intercept in the the simple linear regression of Y/σ against $1/\sigma$. Under model (1), $E(Y_j/\sigma_j \mid \sigma_j) = \theta/\sigma_j$ and $\mathrm{Var}(Y_j/\sigma_j \mid \sigma_j) = 1$. Therefore,

$$y_j/\sigma_j = \beta_0 + \beta_1/\sigma_j + \epsilon \tag{5}$$

is a correctly specified linear model with independent homoscedastic errors ϵ and $\beta_0 = 0$. The t-statistic for β_0,

$$\hat{t}_0 = \frac{\hat{\beta}_0}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_0)}}, \tag{6}$$

may therefore serve as a consistent test statistic. The null of no publication bias is rejected when $|\hat{t}| > t_{n-2,1-\alpha/2}$.

Fig. 1 illustrates the effect of thresholding on a body of studies, and shows the effects on the trend and intercept of the post-selection study means that Begg's and Egger's tests, respectively, look for.

The model (1) is known as the fixed-effects model since the studies are assumed to have a common fixed mean. A more robust model, the random-effects model, assumes that the mean of each study is drawn from a distribution, usually Gaussian. The marginal variance of study i is then given by $\sigma_i^2 + \tau^2$, where τ^2 is the variance of the means. Tests of publication bias may use this variance to comport with the random-effects model, by plugging in common estimates of the between-study variance, see, e.g., Lin and Chu (2018), though doing so may induce dependence among the observations. Therefore, we ignore the distinction between the fixed-effects and random-effects models in what follows, leaving a more refined analysis to future work.

The study variances σ^2 are often treated as fixed but in most of what follows we allow them to be random. However, in assuming that $\sigma_i^2 + \tau^2, i = 1, \dots, n$, in the random-effects model, or $\sigma_i^2, i = 1, \dots, n$, in the fixed effects model (1), are the true marginal variances, we are ignoring the estimation error in these reported and computed variances.

1.2 A heuristic argument

Previous studies have shown that Begg's and Egger's tests frequently disagree in practice (Lin et al., 2018). What is the difference between testing for publication bias using a correlation as in Begg's test or the intercept in a related regression as in Egger's test? The purpose of this section is to give intuition for the formal results and simulations that follow. We relate Begg's and Egger's tests in a way that better facilitates comparison.

A variant of Begg's test is formed by using Pearson's correlation rather than Kendall's rank correlation in (2),

$$\hat{r} = \frac{\sum((u_j - \overline{u})(v_j - \overline{v}))}{\sqrt{\sum(u_j - \overline{u})^2 \sum(v_j - \overline{v})^2}}. \tag{7}$$

(Cf. Gjerdevik and Heuch (2014), suggesting Spearman's rank correlation instead.) A test of publication bias rejects if (7) is large. Given IID normal or large-sample data, which the above is not since $\hat{\theta}$ is common to the u_j, such a test would usually refer the standardized statistic

$$\hat{t}_1 = \hat{r}\sqrt{\frac{n-2}{1-\hat{r}^2}} \tag{8}$$

to a Student's t distribution with $n-2$ degrees of freedom. Doing so is equivalent to testing for a slope of 0 in the formal simple linear regression

$$\frac{Y_i - \hat{\theta}}{\sigma_i} \text{ vs. intercept and } \frac{1}{\sigma_i},\ i = 1, \ldots, n. \tag{9}$$

The fixed effects estimator $\hat{\theta}$ (4) is the projection of Y/σ onto the precisions $1/\sigma$, and the regressand in (9), $(Y - \hat{\theta})S$, is the residual. Therefore the intercept in (9) is the same as the intercept in the Egger regression (5). The residual sum of squares is also the same. Therefore testing the intercept being 0 is the same as carrying out Egger's regression, and testing the slope being 0 carries out the Pearson variant of Begg's test.

This raises the possibility that Egger's test and the Pearson variant of Begg's test can be carried out simultaneously with an F-test of the null $\beta_0 = \beta_1 = 0$ in model (9). Such a test would enjoy the benefits of both Egger's test and Begg's test, to the extent that the benefits of the latter carry over to the Pearson variant. However, it follows from Prop. 1 that this combined test, at least from the standpoint of normal OLS theory, is redundant to Egger's test.

Proposition 1. *Let $\hat{\beta}_0/\hat{sd}(\hat{\beta}_0), \beta_1/\hat{sd}(\hat{\beta}_1)$, and $\hat{F}$ be the test statistics of normal OLS theory for the nulls $\beta_0 = 0, \beta_1 = 0$, and $\beta_0 = \beta_1 = 0$ in the formal regression (9). Let $S_i = 1/\sigma_i$ denote the precisions of the studies and $Z_i = Y_i/\sigma_i$ the standardized effect sizes.*

1. *$\hat{\beta}_0/\hat{sd}(\hat{\beta}_0) = \hat{t}_0$ and $\hat{\beta}_1/\hat{sd}(\hat{\beta}_1) = \hat{r}/\hat{sd}(\hat{r}) = \hat{t}_1$, where $\hat{t}_0$ is defined in (6) and $\hat{t}_1$ in (8)*
2. *$\overline{S}\hat{\beta}_0 = -\overline{S^2}\hat{\beta}_1$ and $\hat{t}_0^2 - \hat{t}_1^2 = \frac{n}{(\overline{S^2})^2 RSS}(\overline{ZS}\,\overline{S} - \overline{Z}\,\overline{S^2})^2 > 0$*
3. *For Gaussian Y,*

$$\begin{aligned} \hat{t}_0 &\sim t_{n-2} \\ \overline{S^2}/\overline{S}\hat{t}_1 &\sim t_{n-2} \\ 2\hat{F} = \hat{t}_0^2 &\sim F_{1,n-2} \end{aligned}$$

Prop. 1(2) shows that the power function of the Pearson variant of Begg's test cannot exceed that of Egger's test. Furthermore, since the Egger statistic is asymptotically level α under model (1) (see Sec. 1.1), the Pearson variant of Begg's statistic cannot be unless

$$n(\overline{ZS}\,\overline{S} - \overline{Z}\,\overline{S^2})^2 \to 0,$$

which is ruled out by non-degenerate data. The observations are not IID due to the common term $\hat{\theta}$, so that the usual consistency assumptions for OLS are unmet. This problem is not specific to the Pearson variant of Begg's test. The original Begg's test also suffers from bias in the same direction due to $\hat{\theta}$ (Michael and Ghebremichael, 2023).

The bias of the F test can be corrected to derive a valid test. The effect of regressing the residuals is that the coefficients are linearly related, Prop. 1(2), and a degree of freedom is lost for the regression sum of squares, Prop. 1(3). The correction results in a test identical to Egger's test, Prop. 1(3). Therefore the difference between the Pearson variant of Begg's test and Egger's test is due entirely to the uncorrected bias. As the Pearson variant and original Begg's test differ only in the choice of correlation measure, we surmise that differences between Egger's test and Begg's test arise due to 1) the robustness and efficiency properties of using a rank correlation measure rather than the Pearson correlation, and 2) an unaccounted-for bias in Begg's test.

This reasoning is only suggestive because there are other tests of the slope being 0 besides the F-test, and these may contain different information from Egger's test. However the following theoretical analysis appears to confirm this conclusion.

2 Limiting power analysis

We investigate the power of Begg's and Egger's tests using their asymptotic power functions. Let $T^{(n)}$ and $c_\alpha^{(n)}$ be test statistics and critical values for a sequence of level α hypothesis tests, such as might come about by applying Begg's or Egger's test to a sequence of data samples of increasing sample size. Let a sequence of alternative hypotheses parametrized by reals $\theta^{(n)}$ converge to the null hypothesis $\theta^{(\infty)} = \theta_0$ at a $n^{-1/2}$ rate. The local limiting power at θ_0 is defined as

$$\lim_{n\to\infty} P_{\theta^{(n)}}(T^{(n)} > c_\alpha^{(n)}) \tag{10}$$

provided the limit exists. The motivation for definition (10) is that tests in common use are expected to be perfectly discriminating in the limit, i.e., for any α and associated critical values $c_\alpha^{(n)}$, $\limsup_{n\to\infty} P_{\theta_0}(T^{(n)} > c_\alpha^{(n)}) \leq \alpha$ while for any $\theta \neq \theta_0$, $\lim P_\theta(T^{(n)} > c_\alpha^{(n)}) = 1$. Therefore tests are often incomparable on this basis, and the asymptotic power function allows a more refined analysis. As it turns out, the standard Begg's test (Begg and Mazumdar, 1994) isn't consistent in the sense just given, being under-powered (Lin et al., 2018; Sterne et al., 2000; Begg and Mazumdar, 1994; Macaskill et al., 2001) due to a faulty asymptotic variance, and we will need to derive and use the correct asymptotic variance below.

Studying the power of Begg's and Egger's tests requires a model for publishing bias. A simple but prototypical model, thresholding on p-values, is presented in Section 2.1. Formulas for the local limiting power are given in Section 2.2. These are obtained by representing the test statistics as asymptotically equivalent IID averages and applying the CLT.

There are many other applications of the local limiting power besides comparing tests. These include determining the minimum sample size to detect a given effect size with a given probability, and determining the minimum effect size to attain a given power at a given sample size.

2.1 Selection model

We model publication bias by supposing that insufficiently significant findings are suppressed. Specifically, where (1) models studies in the absence of publication bias, a study is published only when its p-value is sufficiently small, or equivalently,

$$\frac{Y}{\sigma} > c \tag{11}$$

for some $c \in \mathbb{R}$. Greater publication bias corresponds to greater values of the cutoff c. Thresholding has the effect of shifting up the mean of the standardized study effects $Z = Y/\sigma$,

$$\theta(c) = \frac{\int_c^\infty z f_Z(z) dz}{1 - F_Z(c)} \tag{12}$$

See Fig. 1. With the further assumption that $Z = Y/\sigma$ has a density with convex support, we can use the post-selection means (12) to parametrize the alternatives, since the former are strictly increasing in the threshold:

$$\begin{aligned} \frac{d}{dc}\theta(c) &= \frac{-c f_Z(c)}{1 - F_Z(c)} + \frac{\left(\int_c^\infty z f_Z(z) dz\right) f_Z(c)}{(1 - F_Z(c))^2} \\ &= \frac{f_Z(c)}{(1 - F_Z(c))^2} \left(\int_c^\infty (z - c) f_Z(z) dz \right) > 0. \end{aligned}$$

The p-value thresholding model (11) has convenient technical properties. If we assume that Y belongs to a scale family, that is, Y/σ follows a fixed distribution whatever the value $\sigma = 1/S$, say

$$\frac{Y}{\sigma} = Z \sim F_Z, \tag{13}$$

then $Z \perp\!\!\!\perp S$. It follows that under the the p-value thresholding model (11), or any other model where selection depends only on $Z = Y/\sigma$, the standardized effect sizes and precisions remain independent after thresholding,

$$Z \perp\!\!\!\perp S \mid Z > c.$$

Moreover, the distribution of the precisions after thresholding is unchanged. While more general than the Gaussian model common in meta-analysis, the scale family assumption (13) is stronger than the basic meta-analysis model (1) by requiring the entire distribution of Z be specified by σ instead of just the first 2 moments.

Begg's and Egger's tests ought to be consistent against p-value thresholding. Section 1.2 shows that the two tests can be approximated by testing for zero slope and zero intercept, respectively, in the regression line fitted to $(S_i, Z_i - \hat{\theta} S_i)$. By the above remarks both conditional post-thresholding means $E(Z|Z > c, S)$ and $E(\hat{\theta} \mid Z > c, S)$ equal $E(Z \mid Z > c)$. If $E(Z) = 0$ under the null and the thresholding has any force, then $E(Z \mid Z > c) > 0$, i.e., both slope and intercept are non-zero.

Thresholding on the p-value is perhaps the single most widely accepted model for publication bias (Dickersin et al., 1992; Dickersin, 1997; Easterbrook et al., 1991; Ioannidis, 1998; Stern and Simes, 1997), showing up in the earliest studies of publication bias, such as Lane and Dunlap (1978) and Hedges (1984), as well as many subsequent, e.g., Givens et al. (1997); Copas (1999). An overview of models for publication bias is given in Hedges and Vevea (2005). Variants of p-value thresholding (11) include publishing studies with a probability related to the size of the p-value, as in, e.g., Begg and Mazumdar (1994). Many of the convenient technical properties of (11) carry over to weighted p-value thresholding and we expect the conclusions of this study do as well. Qualitatively different models include thresholding on the effect sizes Y. More realistically, there is no reason to suppose study authors obey the same simple selection mechanism and there also appear to be vastly different selection effects depending on the discipline from which the meta-analysis draws its studies. Due to the manifold and complicated sources of publication bias, models such as (11) can only be regarded as very stylized.

2.2 Local limiting power function

Let $P^{(n)}, n \geq 1$, be a sequence of continuous distributions for (Z, S) in the upper half-plane with a limit law $P^{(\infty)} = P$. Suppose that the distribution of S is the same for all $P^{(n)}$ and it has 2 moments. Suppose further that $E^{(n)}Z^2 \to E^{(\infty)}Z^2$ and that $Z \perp\!\!\!\perp S$ under $P^{(n)}$. These assumptions on Z are met under selection by p-value thresholding (11) with a square-integrable Z-statistic distribution. In summary,

$$\begin{aligned} P^{(n)} &\Rightarrow P^{(\infty)} \\ S &\sim P^{(\infty)} \\ \mu_p = ES^p &< \infty \text{ for } p = 1, 2 \\ Z &\perp\!\!\!\perp S \text{ under } P^{(n)}, n \geq 0 \\ E^{(n)}Z^2 &\to E^{(\infty)}Z^2 < \infty \end{aligned} \tag{14}$$

Theorem 2 gives the local limiting power for Egger's test along the alternatives $\theta^{(n)}$ given by p-value thresholding.

Theorem 2. *In the model* (14)*, assume further that*

1. $\theta^{(n)} = E^{(n)}Z \to E^{(\infty)}Z = \theta^{(\infty)} = 0$.

Then,

1. $$\frac{\hat{\beta}_0 - \theta^{(n)}}{\sqrt{\hat{\operatorname{Var}}(\hat{\beta}_0)}} = n^{-1/2} \frac{\sum_{j=1}^n \left(\operatorname{Var}(S)^{-1}(E(S^2) - E(S)S_j)Z_j - \theta^{(n)} \right)}{\sqrt{\operatorname{Var}(Z)E(S^2)/\operatorname{Var}(S)}} + o_{P_n}(1).$$

2. *The local limiting power function of Egger's test along alternatives $\theta^{(n)} = h/\sqrt{n}$ is*

$$\lim_n P^{(n)} \left(\frac{|\hat{\beta}_0|}{\sqrt{\hat{\operatorname{Var}}(\hat{\beta}_0)}} > t_{n-1,1-\alpha/2} \right) = 1 - \Phi(z_{1-\alpha/2} - m_{\hat{\beta}_0} h) + \Phi(-z_{1-\alpha/2} - m_{\hat{\beta}_0} h),$$

where

$$m_{\hat{\beta}_0} = \sqrt{\frac{\operatorname{Var}(S)}{\operatorname{Var}(Z)E(S^2)}}.$$

Begg's theorem "out of the box" is not consistent, as mentioned earlier. The reason is that the asymptotic variance used for the test statistic is derived under the assumption the observations (3) are IID, which does not hold due to the common $\hat{\theta}$ term. Michael and Ghebremichael (2023) gives the correct bias in the case of Gaussian study effects Y while the general case is treated below. Theorem 3 gives an IID representation of the Begg statistic, Corollary 4 gives the correct asymptotic variance, and Corollary 5 gives the resulting local power function.

Theorem 3. *For a given n and an IID sample $(Z_1, S_1), \ldots, (Z_n, S_n)$, under $P^{(n)}$, define*

$$\begin{aligned}
\hat{\theta} &= \left(\sum_{j=1}^{n} Z_j S_j\right) / \sum_{j=1}^{n} S_j^2 \\
\hat{\tau}(\theta) &= 2\binom{n}{2}^{-1} \sum_{1 \le j < k \le n} \left\{ \frac{Z_j - Z_k}{S_j - S_k} < \theta \right\} - 1 \\
\tau(\theta) &= E^{(n)}\hat{\tau}(\theta) = 2\mathbb{P}^{(n)}\left(\frac{z - z'}{s - s'} < \theta\right) - 1 \\
\Pi^{(n)}\hat{\tau}(\theta) &= \frac{2}{n}\sum_{j=1}^{n}\left(2P^{(n)}\left(\frac{Z_j - Z}{S_j - S} < \theta \,\middle|\, Z_j, S_j\right) - 1\right) - \left(2\mathbb{P}^{(n)}\left(\frac{z - z'}{s - s'} < \theta\right) - 1\right)
\end{aligned}$$

Assuming:

1. $\theta^{(n)} = E^{(n)}Z \to E^{(\infty)}Z = \theta^{(\infty)} = 0$
2. $E^{(\infty)} f_Z^{(\infty)}(Z) = \int \left(f_Z^{(\infty)}\right)^2 < \infty$
3. $f_{Z-Z'}^{(n)} \to f_{Z-Z'}^{(\infty)}$ *uniformly*

Then:

$$\begin{aligned}
&\sqrt{n}\left(\hat{\tau}(\hat{\theta}) - (E^{(n)}\hat{\tau})\left(\frac{\mu_1}{\mu_2}\theta^{(n)}\right)\right) \\
&\qquad = \sqrt{n}(\Pi^{(n)}\hat{\tau})(0) + \sqrt{n}\left(\frac{\overline{ZS}}{\mu_2} - \frac{\mu_1}{\mu_2}\theta^{(n)}\right) \cdot 2E|S_1 - S_2| E^{(\infty)} f_Z^{(\infty)}(Z) + o_{\mathbb{P}^{(n)}}(1).
\end{aligned}$$

Corollary 4. *Under the assumptions of Theorem 3,*

$$\operatorname{Var}^{(n)}(\sqrt{n}\hat{\tau}(\hat{\theta})) \to \frac{4}{9} + 4\frac{(E|S_1 - S_2|)^2}{ES^2} Ef_Z(Z)(Ef_Z(Z)\operatorname{Var} Z - 2E(ZF_Z(Z))).$$

Corollary 5. *Under the assumptions of Theorem 3, the local limiting power function of the debiased Begg test along alternatives $\theta^{(n)} = h/\sqrt{n}$ is*

$$\lim_n P^{(n)}\left(\frac{\sqrt{n}|\hat{\tau}(\hat{\theta})|}{\sqrt{\operatorname{Var}^{(n)}(\hat{\tau}(\hat{\theta}))}} > z_{1-\alpha}\right) = 1 - \Phi(z_{1-\alpha/2} - m_{\hat{\tau}}h) + \Phi(-z_{1-\alpha/2} - m_{\hat{\tau}}h),$$

where

$$m_{\hat{\tau}} = \frac{2Ef_Z(Z)E|S_1 - S_2|ES/ES^2}{\sqrt{\frac{4}{9} + 4Ef_Z(Z)(Ef_Z(Z)\operatorname{Var}Z - 2E(ZF_Z(Z)))(E|S_1 - S_2|)^2/ES^2}}.$$

The limiting power functions will be used to contrast the two tests in Section 3, but two common properties can be directly read off:

1. Loss of power when the primary study variances are nearly constant. Both test slopes $m_{\hat{\beta}_0}$ and $m_{\hat{\tau}}$ drop to 0 as the primary study variances, or equivalently the precisions, approach constancy. The terms $\operatorname{Var}(S)$ and $E|S_1 - S_2|$ in the numerators then approach 0, while the denominators remain > 0 provided the study variances are finite. Insufficient variation in the study variances has been previously associated with inflated Type I error rates in Begg's and Egger's test (Sterne et al., 2000), and the foregoing shows that it is also a problem for their power.

2. Test equivariance. The local limiting power functions of Begg's and Egger's tests are unaffected by reflecting the mean-zero standardized study effect distribution, $Z \mapsto -Z$, just as a funnel plot would not lean more or less toward publication bias if flipped about the precision axis. Likewise, the power functions are unaffected by translations of Z, just as applying a common shift to all the study effects would not usually affect the interpretation of a funnel plot. This invariance property actually holds in finite samples, as well; see Michael (2024, Sec. 2.3).

   It is thus odd that, through the ES and ES^2 terms, both Begg's and Egger's local power functions are sensitive to the location of the study variances. Were the precisions of all studies being conducted to increase, other factors being held constant, Begg's and Egger's tests would be less powerful. See Fig 6. The funnel plots would look exactly the same, other than the tick labels on the precision axis. The formal tests fail to capture this invariance property of the funnel plots that inspired them.

3 Asymptotic relative efficiency

The asymptotic relative efficiency, or Pitman relative efficiency, of Egger's test relative to Begg's is defined as

$$\frac{m^2_{\hat{\beta}_0}}{m^2_{\hat{\tau}}} = \frac{\operatorname{Var}(S)}{(Ef_Z(Z))^2(ES)^2}\left(\frac{ES^2}{(3E|S_1 - S_2|)^2} + Ef_Z(Z)\left(Ef_Z(Z) - 2E(ZF_Z(Z))\right)\right). \quad (15)$$

Expression (15) is the square of ratio of the slopes $m_{\hat{\beta}_0}$ and $m_{\hat{\tau}}$ defined in Theorem 2 and Corollary 5 respectively. The ARE has the following practical application to study design . Let $N_{\hat{\tau}}(\theta)$ and $N_{\hat{\beta}_0}(\theta)$ be the minimal sample size for Begg's and Egger's tests, respectively, to attain a pre-specified power $\beta \in (0,1)$ at alternative θ. Then the ARE is

$$\lim_{n\to\infty} \frac{N_{\hat{\tau}}(\theta^{(n)})}{N_{\hat{\beta}_0}(\theta^{(n)})}.$$

The ratio (15) depends on nuisance parameters, namely, $ES, Ef_Z(Z), E(ZF_Z(Z)), ES^2$, and $E|S_1 - S_2|$. Whether the ARE exceeds or falls below 1, i.e., whether or not Egger's test outperforms Begg's, and to what extent, depends on the distributions of the precisions S and standardized study effects Z giving rise to these nuisance parameters. Section 3.1 considers Gaussian Z, where Egger's test will have the advantage. Section 3.2 considers Student's t and beta distributions for Z, where Begg's test can outperform Egger's.

3.1 $ARE > 1$

The assumption of standard normal $Z = Y/\sigma = YS$ is of special importance in meta-analysis (Begg and Mazumdar, 1994) where Y represents study effects thought to be subject to the CLT. For standard normal Z, $Ef_Z(Z) = E(ZF_Z(Z)) = 1/(2\sqrt{\pi})$, and the ARE (15) becomes

$$\frac{\mathrm{Var}(S)\left(\frac{4}{9}\pi E(S^2) - (E|S_1 - S_2|)^2\right)}{(E(S)E|S_1 - S_2|)^2}. \tag{16}$$

The ARE (16) reveals precision distributions where Egger's test will out-perform Begg's, two categories of which are discussed below. A common cause appears to be the relationship between the two measures of dispersion used by the two tests, the usual variance that appears in the numerator of (16) and comes from Egger's test and the term $E|S_1 - S_2|$ in the denominator, which comes from Begg's test. Given an RV X, and X_1, X_2 IID as X, define the statistical functional

$$D(X) = (E|X_1 - X_2|)^2.$$

Like the variance, $D(X)$ is a measure of statistical dispersion, in that if $X' = a + bX, a, b \in \mathbb{R}$, then $D(X') = b^2 D(X)$. A formula for $D(X)$ is given by Michael and Ghebremichael (2023, Proof of Theorem 4),

$$\sqrt{D(X)} = 2\int_{-\infty}^{\infty} F_X(x)(1 - F_X(x))dx.$$

The parameter $D(X)$ has certain robustness properties that appear to work against the power of Begg's test.

1. **Heavy-tailed distributions.** The parameter $D(X)$ does not have the quadratic character of the variance. For example, it is finite as long as X is integrable, $D(X) = (E|X_1 - X_2|)^2 \leq 4(E|X|)^2$. Therefore, for any family of distributions where the variance can diverge but the mean remains finite, there will be choices of the precision distribution F_S such that the ARE (16) is arbitrarily large. For example, the

Type I Pareto distribution parametrized by location $s_0 > 0$ and shape α has density $f_S(s) = \alpha s_0^\alpha / s^{\alpha-1}$ on $\{s \geq s_0\}$. The mean is $\alpha s_0/(\alpha - 1), \alpha > 1$, and the variance is $s_0^2 \alpha/((\alpha-1)^2(\alpha-2)), \alpha > 2$. As $\alpha \downarrow 2$ with s_0 fixed, the variance goes to infinity while the mean is $< 2s_0$, and as a result the ARE (16) diverges. See Fig. 2 for a simulation.

2. **Skew distributions.** It is not essential that the precision S distribution have very heavy tails for the ARE to be large. On the contrary, even supposing that S is bounded, $a < S < b$, left or right skew can push the ARE up. As in the case of heavy tails, this difference is connected to the different behavior of the dispersion measure D and the usual variance. Relaxing for the moment the assumption (14) that S be continuous, suppose S has the two-point distribution $p\delta_a + (1-p)\delta_b,\ a, b > 0$. Then

$$\frac{\operatorname{Var} S}{D(S)} = \frac{1}{4p(1-p)}$$

is unbounded as $p \to 0$ or 1, and the same holds of the ARE,

$$\text{ARE } = \frac{1}{4p(1-p)} \frac{\frac{4}{9}\pi(a^2 + b^2(1-p)) - (2p(1-p))^2(a-b)^2}{(ap + b(1-p))^2} \geq \frac{1}{4p(1-p)} \frac{4}{9}\pi.$$

To show that this behavior holds more generally than the two-point distribution, suppose again $a \leq S \leq b$. With F_S denoting the CDF of S, introduce the family of CDFs indexed by $\alpha, 0 \leq \alpha \leq 1$,

$$F_S^{(\alpha)}(s) = \begin{cases} 0, & s < a \\ 1 - \alpha(1 - F_S(s)), & a \leq s \leq b \\ 1, & s > b \end{cases}$$

The original CDF F_S corresponds to $\alpha = 1$ and $\alpha = 0$ corresponds to an atom at a. As $\alpha \downarrow 0$, mass is moved toward the left support boundary a, i.e., the distributions become more right-skew. Since $1 - F^{(\alpha)} = \alpha(1 - F)$ on $[a, b]$,

$$\begin{aligned} E^{(\alpha)} S^k &= k \int s^{k-1}(1 - F_S^{(\alpha)}(s)) = \alpha E S^k, k \geq 1 \\ \operatorname{Var}^{(\alpha)} S &= \alpha \operatorname{Var} S - \alpha(\alpha - 1)(ES)^2 \\ E^{(\alpha)}|S_1 - S_2| &= 2 \int F_S^{(\alpha)}(s)(1 - F_S^{(\alpha)}(s)) = 2\alpha(1-\alpha)ES + \alpha^2 E|S_1 - S_2|. \end{aligned}$$

Substituting these values in (16), the ARE at α is

$$\begin{aligned} \text{ARE}^{(\alpha)} &= \frac{\operatorname{Var}(S) - (\alpha - 1)(ES)^2}{\alpha(2(1-\alpha)ES + \alpha E|S_1 - S_2|)^2} \cdot \frac{\frac{4}{9}\pi ES^2 - \alpha(2(1-\alpha)ES + \alpha E|S_1 - S_2|)^2}{\alpha(ES)^2} \\ &\geq \text{constant}/\alpha^2 \end{aligned}$$

as $\alpha \downarrow 0$.

See Fig. 3 for a simulation using asymmetric beta distributions to model the precisions. A similar family of CDFs shifting mass toward b shows that the ARE also diverges with increasing left skew.

The examples above show that the relative efficiency of Egger's test relative to Begg's may be arbitrarily large. On the other hand, with Gaussian study effects, Begg's test cannot outperform Egger's in detecting the presence of p-value thresholding, at least by the criterion of Pitman efficiency.

Proposition 6. *Given a random variable S with $ES^2 < \infty$ and $ES \neq 0$, the ARE* (16) *satisfies the inequality*

$$\frac{\operatorname{Var}(S)\left(\frac{4}{9}\pi E(S^2) - (E|S_1 - S_2|)^2\right)}{(E(S)E|S_1 - S_2|)^2} > 1.$$

Proof. The ARE is greater than 1 if and only if

$$\frac{(E|S_1 - S_2|)^2}{\operatorname{Var}(S)} < \frac{4}{9}\pi.$$

Lemma 7 shows that, whatever the precision distribution F_S, as long as the moment assumptions are satisfied the left-hand side is bounded by $4/3$. The required inequality therefore holds strictly, by a factor of $\pi/3$. □

Lemma 7. *Given independent and identically distributed, square-integrable random variables X,*

$$\frac{D(X)}{\operatorname{Var}(X)} \leq \frac{4}{3}.$$

The bound occurs for a uniformly distributed RV.

3.2 *ARE* < 1

Section 3.1 focused on the case of Gaussian study effects where the the robustness property of Begg's test is perhaps more likely to be an encumbrance than an asset. We next consider non-Gaussian study effects. There does not appear to be a standard or default choice of precision distribution that we can fix corresponding to the role of the Gaussian distribution for study effects. Therefore we consider two representative families of study effect distributions, Student's t for heavy-tailed distributions and beta for skew distributions. Skew distributions as a model of the study effects is of particular interest in meta-analysis due to the prevalence of odds ratios and similar statistics.

Proposition 8. *1. Let the RV Y follow Student's t distribution with degrees of freedom ν. Then*

$$Ef_Y(Y) = 2\sqrt{\nu}(\nu + 1)c(\nu)^2 B(3/2, \nu + 1/2)$$
$$E(YF_Y(Y)) = 2\nu^{5/2}/(\nu - 1)c(\nu)^2 B(3/2, \nu - 1/2),$$

where $c(\nu) = \Gamma((\nu + 1)/2)/(\sqrt{\pi\nu}\Gamma(\nu/2)$ denotes the integrating factor in the Student's t density and B denotes the beta function. If $Z \sim \sqrt{(\nu - 2)/\nu}Y$ is a standardized Student's t with degrees of freedom ν, then as $\nu \downarrow 2$,

$$ARE \to \frac{\operatorname{Var} S}{(ES)^2}. \tag{17}$$

2. *Let Y follow a beta distribution with shape parameters a and b. Then*

$$\begin{aligned} Ef_Y(Y) &= B(2a-1,2b-1)/B(a,b)^2 \\ E(YF_Y(Y)) &= \frac{2^{2(a+b)-1}\Gamma(a+\frac{1}{2})\Gamma(b+\frac{1}{2})\Gamma(a+b)^2}{\pi\Gamma(a)\Gamma(b)\Gamma(2(a+b)+1)} + \frac{1}{2}\frac{a}{a+b}. \end{aligned}$$

Let $Z = (Y-\mu)/\sigma, \mu = EY, \sigma^2 = \operatorname{Var} Y$, follow the distribution of a beta RV with shape parameters a and b after centering and standardizing. Then (17) *holds for fixed $b > .5$, as $a \downarrow .5$.*

For example, when the precisions S are uniformly distributed on $(0,1)$, the limiting ARE is $\operatorname{Var} S/(ES)^2 = 1/3$. The ratio $\operatorname{Var} S/(ES)^2 = E(S^2)/(ES)^2 - 1$, and therefore the ARE, approaches 0 as the distribution of precisions S approaches 0.

The mechanism is similar in both cases. Writing the ARE (15) as

$$\frac{\operatorname{Var}(S)}{(ES)^2}\left(1 + \frac{ES^2}{(3E|S_1-S_2|)^2}\frac{1}{(Ef_Z(Z))^2} - 2\frac{E(ZF_Z(Z))}{Ef_Z(Z)}\right),$$

we find that there are boundary points of the model (14) near which $Ef_Z(Z)$ diverges while $EZF_Z(Z)$ remains bounded. These boundary points are $\nu = 2$ for the Students t, where the variance is infinite, and $a = .5$ for the beta, where $Ef_Z(Z)$ is infinite. See Figs. 4 and 5.

4 Simulations

There are two sources of approximation error in the definition of limiting local power (10). Besides the asymptotic approximation in matching the finite sample power to the limiting power, there is also the error in matching an alternative near the null to the null itself. The first error is exacerbated in the setting of meta-analysis, where sample sizes are frequently small. Therefore we check the theoretical conclusions of Sections 2 and 3 against simulations. The simulations confirm the qualitative aspects of the conclusions though the effects are attenuated.

In all simulations, we take the number of primary studies to be $n = 20$, which was informed by Lin and Chu (2018), and the significance level of the publishing bias tests to be .1, following Macaskill et al. (2001), Sterne et al. (2000), and others. These settings may be easily modified in the provided software. The simulations are summarized by figures. Each figure consists of the observed power curves for Egger's and Begg's test at thresholding alternatives and the ratio of the power curves. The first two simulations correspond to Section 3.1, where heavy-tailed and skew precision distributions are shown to benefit Egger's test over Begg's. The next two correspond to Section 3.2, where heavy-tailed and skew study effect distributions are shown to benefit Begg's test over Egger's.

1. Fig. 2. Precisions: Pareto; study effects: Gaussian. Heavier tails for the precisions, i.e., increasing the probability of very precise studies, improves the power of both tests. However it improves the power of Egger's test more so.

2. Fig. 3. Precisions: beta; study effects: Gaussian. Left or right skew in the precisions, i.e., keeping the range of precisions fixed but increasing the imbalance between the proportion of precise and imprecise studies, affects the power of both tests adversely. But the effect on Begg's test is more severe.

3. Fig. 4. Precisions: uniform; study effects: Student's t, scaled. As the tails of the study effects become heavier, the power of each test improves. However Begg's test benefits more than Egger's.

4. Fig. 5. Precisions: uniform; study effects: beta, centered and scaled. Letting the first shape parameter a approach .5 and the second grow as $1/(a-.5)$, Begg's test becomes more efficient than Egger's.

Lastly, we examine the effect of shifting the precision distribution.

5. Fig. 6. Precisions: uniform; study effects: Gaussian. A location shift of the precisions to the right, i.e., keeping the spread and range of the precisions fixed but increasing all precisions by the same fixed amount amount, affects the power of both tests adversely. Each appears affected to about the same degree

The software used to carry out the simulations and generate the figures presented is publicly available at `https://github.com/haben-michael/pubbias-power.`

5 Discussion

The principal conclusions of the local power analysis are:

1. The differences between Begg's and Egger's tests are due mainly to the robustness of the former and a bias in Begg's test.

2. After the bias in Begg's test is repaired, the robustness of Begg's test has two tendencies.

   (a) With respect to the precision distribution, the robustness is a detriment. As a result, in the popular Gaussian study effect model, the performance of Egger's test is superior.

   (b) With respect to the standardized study effect distribution, the robustness is an asset. As a result, Begg's test may be preferable to Egger's in the case of odds ratios and other statistics slow to converge to normality.

While the sample sizes involved in typical meta-analyses, i.e., the number of primary studies, undercut any asymptotic analyses, simulations confirm the conclusions reached by the local power analysis.

An area of further work is to incorporate the measurement error in the reported study variances. Ignoring this measurement error appears widespread in the methodological meta-analysis literature, including in the original formulations of the publication bias tests considered here. If measurement error were incorporated, the variances of the standardized effect sizes, rather than being exactly 1, would deviate from 1 by some stochastic error. This error is likely non-ignorable when analyzing publication bias (Macaskill et al., 2001; Schwarzer et al., 2002): Smaller meta-analyses are likely to have larger errors. There is greater variability at the lower end of the precision axis of the funnel plot that is unaccounted for.

References

Begg, C. and M. Mazumdar (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088–1101.

Cooper, H. (2015). *Research synthesis and meta-analysis: A step-by-step approach*, Volume 2.

Copas, J. (1999). What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 162*(1), 95–109.

Dickersin, K. (1997). How important is publication bias? a synthesis of available data. *AIDS education and prevention 9*, 15–21.

Dickersin, K., Y.-I. Min, and C. L. Meinert (1992). Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *Jama 267*(3), 374–378.

Easterbrook, P. J., R. Gopalan, J. Berlin, and D. R. Matthews (1991). Publication bias in clinical research. *The Lancet 337*(8746), 867–872.

Egger, M., G. D. Smith, M. Schneider, and C. Minder (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ 315*(7109), 629–634.

Givens, G. H., D. Smith, and R. Tweedie (1997). Publication bias in meta-analysis: a bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science 12*(4), 221–250.

Gjerdevik, M. and I. Heuch (2014). Improving the error rates of the begg and mazumdar test for publication bias in fixed effects meta-analysis. *BMC Medical Research Methodology 14*(1), 1–16.

Guyatt, G. H., A. D. Oxman, V. Montori, G. Vist, R. Kunz, J. Brozek, P. Alonso-Coello, B. Djulbegovic, D. Atkins, Y. Falck-Ytter, et al. (2011). Grade guidelines: 5. rating the quality of evidence—publication bias. *Journal of clinical epidemiology 64*(12), 1277–1282.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics 9*(1), 61–85.

Hedges, L. V. and J. Vevea (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, and M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, pp. 145–174.

Ioannidis, J. P. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Jama 279*(4), 281–286.

Lane, D. M. and W. P. Dunlap (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology 31*(2), 107–112.

Lin, L. and H. Chu (2018). Quantifying publication bias in meta-analysis. *Biometrics 74*(3), 785–794.

Lin, L., H. Chu, M. H. Murad, C. Hong, Z. Qu, S. R. Cole, and Y. Chen (2018). Empirical comparison of publication bias tests in meta-analysis. *Journal of general internal medicine 33*, 1260–1267.

Macaskill, P., S. D. Walter, and L. Irwig (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine 20*(4), 641–654.

Michael, H. (2024). The effect of screening for publication bias on the outcomes of meta-analyses. *Forthcoming in Scandinavian Journal of Statistics*.

Michael, H. and M. Ghebremichael (2023). A correction to Begg's test for publication bias. *Communications in Statistics - Theory and Methods 0*(0), 1–21.

Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj 372*.

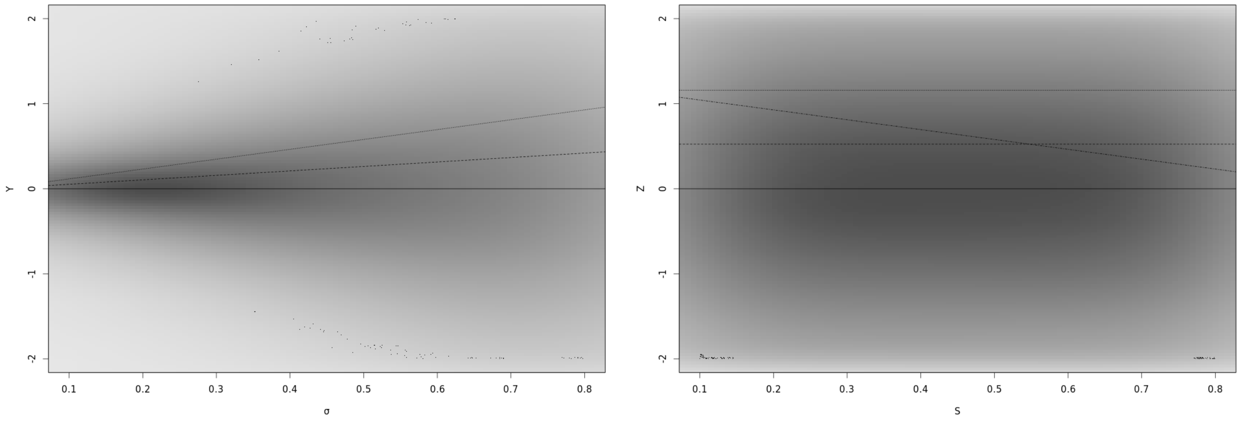
Peters, J. L., A. J. Sutton, D. R. Jones, K. R. Abrams, and L. Rushton (2006). Comparison of two methods to detect publication bias in meta-analysis. *Jama 295*(6), 676–680.

Schwarzer, G., G. Antes, and M. Schumacher (2002). Inflation of type i error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Statistics in medicine 21*(17), 2465–2477.

Stern, J. M. and R. J. Simes (1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *Bmj 315*(7109), 640–645.

Sterne, J. A. and M. Egger (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, and M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, pp. 99–110.

Sterne, J. A., D. Gavaghan, and M. Egger (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology 53*(11), 1119–1129.



(a) (Y, σ) axes

(b) (Z, S) axes

Figure 1: The effect of screening out studies with large p-values. The left plot uses the study effects and variances as coordinates, and the right plot uses the centered and standardized study effects and study precisions. The grayscale background indicates the null distribution of the data, and the expected funnel plot shape, oriented horizontally, is seen in the left plot. In both plots the horizontal solid line through 0 gives the mean of the studies under the null, the dashed line gives a cutoff p-value for suppressing or publishing a study, and the dotted line gives the mean of the studies after thresholding. On the right, the broken line is the line the intercept and slope of which is targeted by Egger's and Begg's test, respectively, in detecting departures from the null (Sec. 2.1).

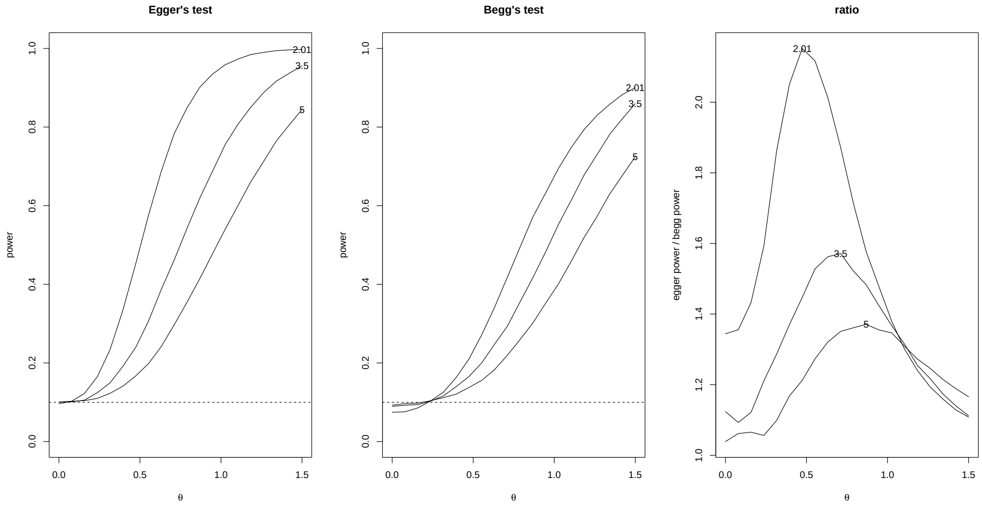


Figure 2: Pareto precision distribution (varying shape parameter), Gaussian study effect distribution.

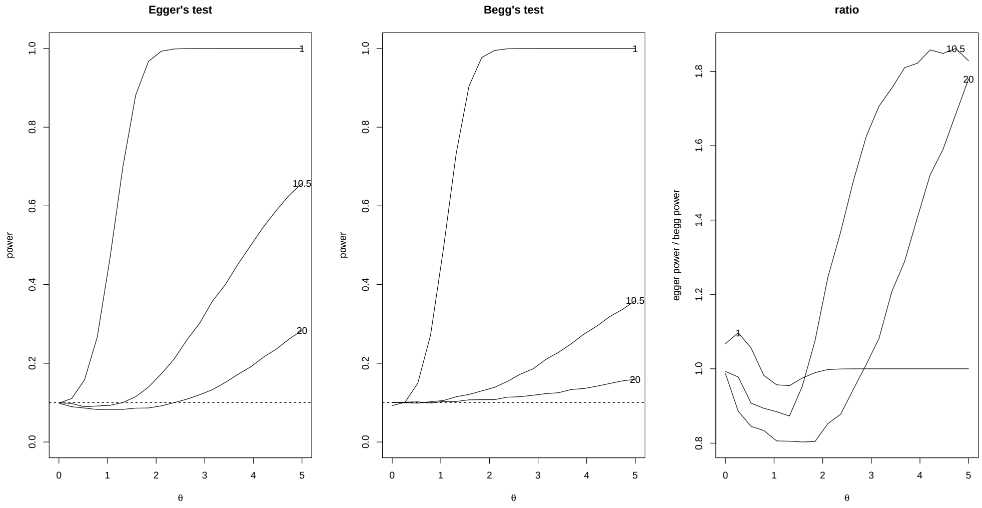


Figure 3: Beta precision distribution (varying shape parameter), Gaussian study effect distribution.

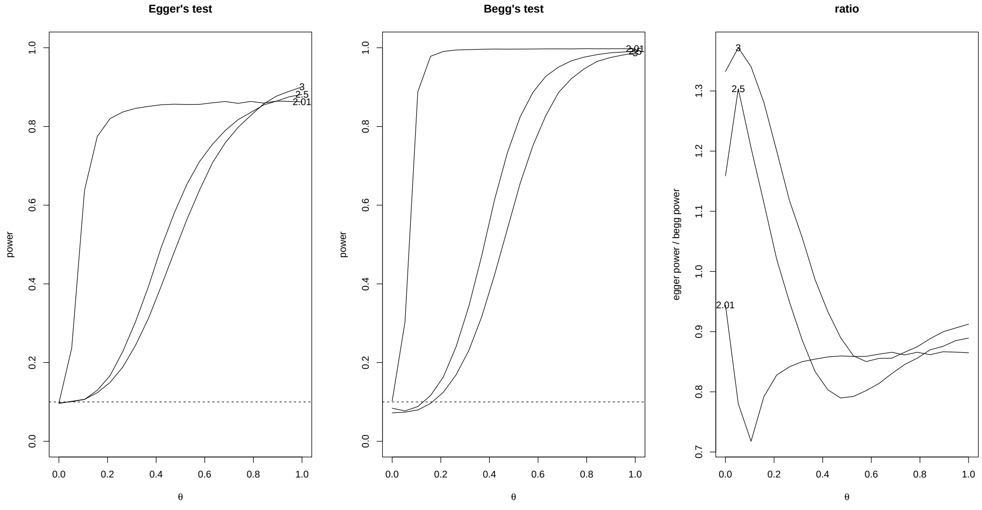


Figure 4: Uniformly distributed precisions, Student's t study effect distribution (varying degrees of freedom).

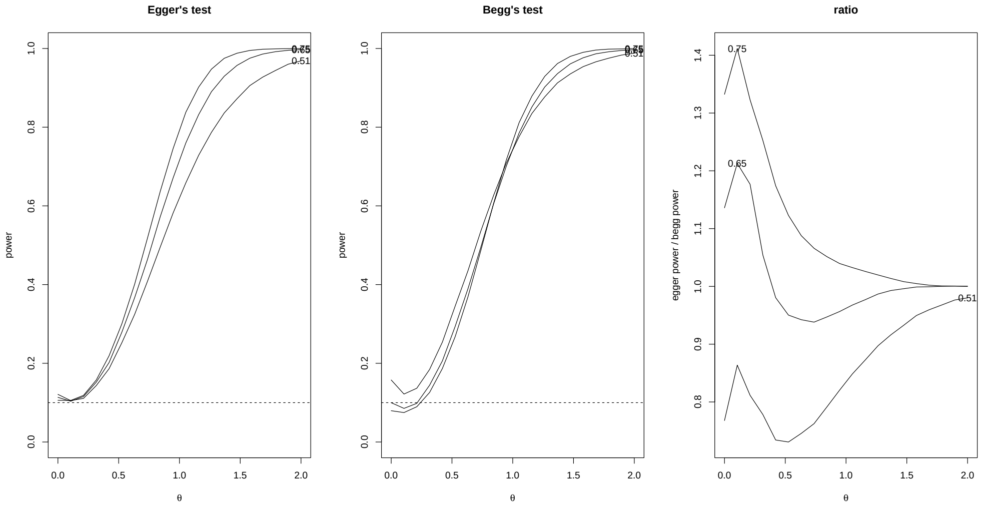


Figure 5: Uniformly distributed precisions, beta study effect distribution (varying shape parameter).

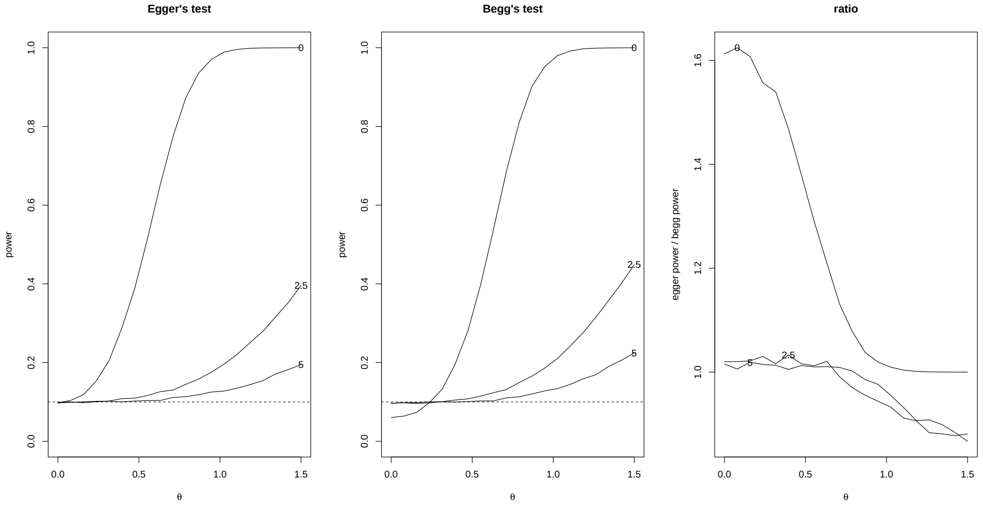


Figure 6: Uniformly distributed precisions (varying location), Gaussian study effect distribution.