THE POWER FUNCTION AND THE "FILE DRAWER PROBLEM"

MARK FISHER

Incomplete

ABSTRACT. The "file drawer problem" arrises when reports of experiments with insignificant results are "left in a file drawer" and the associated datasets remain unobserved. [The phrase was coined by Rosenthal (1979).] This behavior constitutes a selection process that truncates the sample space for datasets. The sampling distribution for observed datasets is restricted to a truncation set implied by the critical region for the test statistic. The truncated sampling distribution is normalized by the probability of the truncation set, and this probability is given by the power function. Because the normalization involves model parameters, the likelihood is affected. Such changes in the likelihood can have dramatic effects on inference regarding the parameters when sample sizes are not sufficiently large.

Date: 10:03 November 30, 2017. *Filename*: `"Publication Selection Process"`. (Original version June 2016.)

1. Introduction

The "file drawer problem" arrises when reports of experiments with insignificant results are left in a file drawer and not published or otherwise observed. Rosenthal (1979) coined the phrase and provides a nice summary of the problem in his abstract:

> For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journal are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results.

Rosenthal addresses the problem by aggregating the p-values from published studies. Rosenthal's approach requires the reviewer to have an opinion about the number of studies in file drawers. If this number is small relative to the number needed to reduce the overall level of significance to an unacceptable value, then the reviewer may suppose the problem is not substantial in this case.

The current paper takes a different approach to addressing the file drawer problem. Unpublished studies constitute missing datasets.[1] As a result of this *selection process*, the sample space for observed datasets is truncated. The sampling distribution for observed datasets is restricted to a truncation set implied by the critical region for the test statistic. The truncated sampling distribution is normalized by the probability of the truncation set, and this probability is given by the power function. Because the normalization involves model parameters, the likelihood is affected. Such changes in the likelihood can have dramatic effects on inference regarding the parameters when sample sizes are not sufficiently large.

The approach taken here has something in common with Iyengar and Greenhouse (1988) who use "weighting functions" which embody the probability of the observations being selected. They discuss multiplying the likelihood function by a weighting function. See also Bayarri and DeGroot (1987) and Bayarri and DeGroot (1991).

Other related literature. Related papers include Gelman and Tuerlinckx (2000), Button et al. (2013), Gelman and Carlin (2014), Gelman and Loken (2014), and the references contained therein.

Outline. Section 2 presents the general idea regarding truncated datasets. Section 3 presents the main illustration. Section 4 presents the truncated sampling distribution for the sufficient statistics. Section 5 introduces multiple datasets. Section 6 introduces some generalizations to the model. Section 7 shows how an optimal Bayesian decision can produce a frequentist selection process.

There are a number of appendices. Appendix A presents addition details about the normal case. Appendix B presents a simple example that illustrates the main point. Appendix C presents a motivating story (for the main case of interest) from an earlier version of this paper that may still be useful. Appendix D presents the Bayesian analysis absent

[1] The missing-data mechanism is not *ignorable* and must be taken into account to obtain valid inference. See Gelman et al. (2014, Chapters 8 and 18). I will not refer to ignorability further since I do not find that framework convenient here.

the selection process. Appendix E begins an investigation into expected power as a guide to sample size.

2. Truncated datasets

When the sample space for a dataset is truncated, only datasets that lie within the truncation set will be observed. The sampling distribution for observed datasets is therefore normalized by the probability of the truncation set. This probability depends on the parameters of the unrestricted distribution. Consequently, the likelihood of the parameters given an observed dataset incorporates the probability of the truncation set, which in turn affects the posterior distribution for the parameters.

The critical region for a test statistic characterizes a truncated sample space for a dataset. The region of truncation will depend on all of the observations jointly. A process that selects only significant results (datasets composed of observations for which the test statistic is in the critical region) produces datasets from a truncated distribution. Such a selection process can be thought of as a form of rejection sampling: Construct a dataset and reject it unless it lies in the truncated set.

When the truncation set is generated by the *critical region* for a *test statistic*, the probability of the truncation set is the *power function*. Consequently, the power function plays a central role in adjusting the posterior distribution to account for the dataset selection process.

Distribution with full sample space. Let the joint density for the dataset (i.e., the observations) y and the parameters θ be given by

$$p(y, \theta) = p(y|\theta)\, p(\theta) \qquad \text{for } (y, \theta) \in \mathcal{Y} \times \Theta. \tag{2.1}$$

Note

$$\int_{\mathcal{Y}} p(y|\theta)\, dy = 1 \qquad \text{for all } \theta \in \Theta. \tag{2.2}$$

The posterior distribution for θ given y is

$$p(\theta|y) = \frac{p(y|\theta)\, p(y)}{p(y)} \propto p(y|\theta)\, p(\theta), \tag{2.3}$$

where

$$p(y) = \int_{\Theta} p(y|\theta)\, p(\theta)\, d\theta. \tag{2.4}$$

Truncation set. Given $\Lambda \subset \mathcal{Y}$, define

$$\mathbb{P}_{\Lambda}(\theta) := \int_{\mathcal{Y}} 1(y \in \Lambda)\, p(y|\theta)\, dy = \int_{\Lambda} p(y|\theta)\, dy, \tag{2.5}$$

where

$$1(x) = \begin{cases} 1 & x \text{ is true} \\ 0 & x \text{ is false} \end{cases}. \tag{2.6}$$

Note that $\mathbb{P}_{\Lambda}(\theta) = \Pr[y \in \Lambda|\theta]$ given (2.1). If $T(y)$ is a *test statistic* and $\Lambda = \{y \in \mathcal{Y} : T(y) \in \Omega\}$ where Ω is the *critical region*, then $\mathbb{P}_{\Lambda}(\theta)$ is the *power function*.

Special case. Suppose $y = (y_1, \ldots, y_n) \in \mathcal{Y} = \mathcal{Y}^n$ and $p(y|\theta) = \prod_{j=1}^n p(y_j|\theta)$. In addition suppose $\Lambda = L^n$. Then

$$\mathbb{P}_\Lambda(\theta) = \int_\Lambda p(y|\theta)\, dy = \prod_{j=1}^n \int_L p(y_j|\theta)\, dy_j = \mathbb{P}_L(\theta)^n. \tag{2.7}$$

Truncated distribution. Let $(y, \theta) \in \Lambda \times \Theta$ for some truncation set $\Lambda \subset \mathcal{Y}$ for which $\mathbb{P}_\Lambda(\theta) > 0$ for all $\theta \in \Theta$. Let $\mathcal{S}_\Lambda$ denote the truncation of y to Λ. Then (assuming $y \in \Lambda$)

$$p(y, \theta|\mathcal{S}_\Lambda) = p(y|\theta, \mathcal{S}_\Lambda)\, p(\theta), \tag{2.8}$$

where

$$p(y|\theta, \mathcal{S}_\Lambda) = \frac{p(y|\theta)}{\mathbb{P}_\Lambda(\theta)}. \tag{2.9}$$

Note

$$\int_\Lambda p(y|\theta, \mathcal{S}_\Lambda)\, dy = 1 \qquad \text{for all } \theta \in \Theta. \tag{2.10}$$

The posterior distribution for θ can be expressed as

$$p(\theta|y, \mathcal{S}_\Lambda) \propto p(y|\theta, \mathcal{S}_\Lambda)\, p(\theta) \propto \frac{p(\theta|y)}{\mathbb{P}_\Lambda(\theta)}. \tag{2.11}$$

Thus the probability of the truncation set plays a central role in determining the posterior distribution for the parameter.

Sampling. Equation (2.11) suggests that draws from the target distribution $p(\theta|y, \mathcal{S}_\Lambda)$ can be obtained via *importance sampling* by resampling draws $\{\theta^{(r)}\}_{r=1}^R$ from the proposal distribution $p(\theta|y)$ where the resampling probabilities are proportional to the importance weights $q^{(r)} = 1/\mathbb{P}_\Lambda(\theta^{(r)})$. From this perspective it is evident that if the draws from the unadjusted posterior $p(\theta|y)$ are largely located where the adjustment factor $1/\mathbb{P}_\Lambda(\theta)$ is relatively flat, then the truncation has little effect on inference.

If an analytical expression for $\mathbb{P}_\Lambda(\theta)$ is not available, it can be numerically approximated as follows.[2] Note that if $y' \sim p(y|\theta^{(r)})$, then $1(y' \in \Lambda)$ has a Bernoulli distribution with probability $\mathbb{P}_\Lambda(\theta^{(r)})$. With repeated sampling, one can approximate $q^{(r)} = 1/\mathbb{P}_\Lambda(\theta^{(r)})$ to any desired degree of accuracy. For example, let $s^{(r)} \geq 1$ denote the number of successes in $T^{(r)}$ trials given $\theta^{(r)}$. Then let $\widetilde{q}^{(r)} = (T^{(r)} + 1)/s^{(r)}$.[3]

[2] As an alternative when $\mathbb{P}_\Lambda(\theta)$ is not available, one can adopt approximate Bayesian computation (ABC) to make draws of θ from the posterior. The r-th draw from the posterior for θ is computed as follows. First draw $\theta' \sim p(\theta)$ from the prior. Next draw $y' \sim p(y|\theta')$ from the unrestricted sampling distribution repeatedly until $y' \in \Lambda$. (The expected number of draws equals $1/\mathbb{P}_\Lambda(\theta')$.) If y' is sufficiently "close" to y, then set $\theta^{(r)} = \theta'$. Otherwise discard θ' and start over.

[3] Let $q = 1/\gamma$ where $\gamma \sim \mathsf{Beta}(a, b)$. Then $q - 1 \sim \mathsf{BetaPrime}(b, a)$. Suppose $a = s + 1$ and $b = T - s + 1$. If $s \geq 1$ then the mean is $(T + 1)/s$ and if $s \geq 2$ then the variance is $\frac{(T+1)\,(T-s+1)}{(s-1)\,s^2}$.

3. Main application

The main application involves datasets composed of normally distributed observations with unknown mean and variance. In particular, let $y = (y_1, \ldots, y_n) \in \mathcal{Y} = \mathbb{R}^n$ where

$$p(y|\theta) = \prod_{j=1}^{n} \mathsf{N}(y_j|\mu, \tau^2) \tag{3.1}$$

and $\theta = (\mu, \tau^2) \in \Theta = \mathbb{R} \times \mathbb{R}_{>0}$. It may be useful to think of μ as the unobserved true effect and τ as the standard deviation of the noise inherent in the measurements.

Notation. Some notation is required in order to characterize and measure the truncation set for this application.

Let $\Phi(\,\cdot\,)$ denote the cumulative distribution function (CDF) for the standard normal distribution $\mathsf{N}(0,1)$. Let $\Phi_\nu(\,\cdot\,)$ denote the CDF for $\mathsf{Student\text{-}t}(0,1,\nu)$ and let $\Phi_\nu^{-1}(\,\cdot\,)$ denote the associated quantile function. Further let $\Phi_{\nu,\delta}(\,\cdot\,)$ denote the CDF of the noncentral t distribution with degrees of freedom ν and noncentrality parameter δ. Note that $\Phi_{\nu,0}(x) \equiv \Phi_\nu(x)$. Define

$$f_\nu(x) := 2\,\big(1 - \Phi_\nu(x)\big). \tag{3.2}$$

Note $f_\nu'(x) = -2\,\Phi_\nu'(x) < 0$ for all $x \in \mathbb{R}$. See Figure 1 for an example where $\nu = 19$.

Truncation set. Define the following statistics:

$$\widehat{\mu} := \frac{1}{n}\sum_{j=1}^{n} y_j \qquad\qquad \widehat{\sigma} := \sqrt{\frac{\sum_{j=1}^{n}(y_j - \widehat{\mu})^2}{n\,(n-1)}} \tag{3.3a}$$

$$\widehat{t} := \widehat{\mu}/\widehat{\sigma} \qquad\qquad \widehat{\pi} := f_{n-1}(|\widehat{t}|). \tag{3.3b}$$

Note that $\widehat{\pi} \in (0,1]$. Given $\alpha \in (0,1]$, the truncation set is

$$\Lambda = \{y \in \mathcal{Y} : \widehat{\pi} < \alpha\}. \tag{3.4}$$

The probability of the truncation set is

$$\mathbb{P}_\Lambda(\mu,\tau) = 1 - \Phi_{n-1,\sqrt{n}\,\mu/\tau}(c_\alpha) + \Phi_{n-1,\sqrt{n}\,\mu/\tau}(-c_\alpha) \qquad \text{where } c_\alpha = f_{n-1}^{-1}(\alpha). \tag{3.5}$$

Note that $\mathbb{P}_\Lambda(\mu,\tau)$ depends on μ/τ, n, and α. It is convenient to adopt notation that expresses this dependence:

$$\mathcal{P}(\mu/\tau, n, \alpha) := \mathbb{P}_\Lambda(\mu/\tau, 1). \tag{3.6}$$

Note

$$\mathcal{P}(0,n,\alpha) = 1 - \Phi_{n-1}(c_\alpha) + \Phi_{n-1}(-c_\alpha) = f_{n-1}(c_\alpha) = \alpha. \tag{3.7}$$

Also note, $\mathcal{P}(\mu/\tau, n, 1) = 1$.[4] See Figure 2 for a plot of $\mathcal{P}(\lambda, n, \alpha)$ and Figure 3 for a plot of $1/\mathcal{P}(\lambda, n, \alpha)$.

In passing note that if $n = 2$ and $\alpha = 1/2$, then $\Lambda = \{(y_1, y_2) \in \mathbb{R}^2 : y_1\,y_2 > 0\}$ and $\mathcal{P}(\mu/\tau, 2, 1/2) = \Phi(-\mu/\tau)^2 + \Phi(\mu/\tau)^2$.

[4] If $\alpha = 1$ then $\Lambda = \mathcal{Y} \setminus B$ where $B = \{y \in \mathcal{Y} : \widehat{\mu} = 0\}$ and $\int_B p(y|\theta)\,dy = 0$.

Null hypothesis and significance level. Consider the null hypothesis $H_0 : \mu = 0$ and the alternative hypothesis $H_1 : \mu \neq 0$. Let the test statistic be the t-statistic $\widehat{t}$ and let $1(|\widehat{t}| > c_\alpha)$ characterize the critical region where c_α is the critical value. Note that $1(|\widehat{t}| > c_\alpha)$ is equivalent to $1(\widehat{\pi} < \alpha)$ where $\widehat{\pi}$ is the p-value. Consequently, $\mathcal{P}(\mu/\tau, n, \alpha)$ is the power function and α is the significance level (the power function evaluated at the null hypothesis).[5]

Likelihood and posterior. Is is convenient to express the likelihood for the truncated dataset as

$$p(y|\mu, \tau^2, \alpha, \mathcal{S}) = \frac{1(\widehat{\pi} < \alpha)\, p(y|\mu, \tau^2)}{\mathcal{P}(\mu/\tau, n, \alpha)}. \tag{3.8}$$

Note that it is possible to learn about α as well as μ and τ^2.

The joint posterior for (μ, τ, α) can be expressed as

$$p(\mu, \tau^2, \alpha|y, \mathcal{S}) \propto p(y|\mu, \tau^2, \alpha, \mathcal{S})\, p(\mu, \tau^2)\, p(\alpha) \propto p(\mu, \tau^2|y)\, \frac{1(\widehat{\pi} < \alpha)\, p(\alpha)}{\mathcal{P}(\mu/\tau, n, \alpha)}. \tag{3.9}$$

The marginal posterior for (μ, τ) is given by

$$p(\mu, \tau|y, \mathcal{S}) \propto p(\mu, \tau|y)\, \mathcal{W}(\mu/\tau, n, \widehat{\pi}), \tag{3.10}$$

where the weighted inverse power is

$$\mathcal{W}(\mu/\tau, n, \widehat{\pi}) = \int_{\widehat{\pi}}^{1} \frac{p(\alpha)}{\mathcal{P}(\mu/\tau, n, \alpha)}\, d\alpha. \tag{3.11}$$

Prior. As an example, consider a discrete prior for α:

$$p(\alpha) = \begin{cases} 1/4 & \alpha \in A = \{1/100, 1/20, 1/10, 1\} \\ 0 & \text{otherwise} \end{cases}. \tag{3.12}$$

This prior includes the possibility of no truncation ($\alpha = 1$). With this prior,

$$\mathcal{W}(\mu/\tau, n, \widehat{\pi}) = \frac{1}{4} \sum_{\alpha \in A} \frac{1(\widehat{\pi} < \alpha)}{\mathcal{P}(\mu/\tau, n, \alpha)}. \tag{3.13}$$

Prior for (μ, τ^2). Let the prior for (μ, τ^2) be given by[6,7]

$$p(\mu, \tau^2) = \mathsf{N}(\mu|m_0, \tau^2/\kappa_0)\, \textsf{Inv-Gamma}(\tau|a_0/2, b_0/2). \tag{3.14}$$

Then

$$p(\mu, \tau^2|y) = \mathsf{N}(\mu|m_1, \tau^2/\kappa_1)\, \textsf{Inv-Gamma}(\tau^2|a_1/2, b_1/2), \tag{3.15}$$

[5] A *type I error* is committed when one rejects a true null hypothesis. The probability of a type I error is given by the significance level of the test, α. A *type II error* is committed when one fails to reject a false null hypothesis. The probability of a type II error (which is denoted β) equals $1 - \mathcal{P}(\mu/\tau, n, \alpha)$.

[6] If $p(\tau^2) = \textsf{Inv-Gamma}(\tau^2|a, b) \propto (1/\tau^2)^{1+a} e^{-b/\tau^2}$, then $p(\tau) = 2\,\tau\, \textsf{Inv-Gamma}(\tau^2|a, b)$.

[7] If $p(\mu, \tau^2) = \mathsf{N}(\mu|m, \tau^2/\kappa)\, \textsf{Inv-Gamma}(\tau^2|a/2, b/2)$, then $p(\mu) = \textsf{Student-t}(\mu|m, b/(a\,\kappa), a)$.

where

$$\kappa_1 = \kappa_0 + n \tag{3.16a}$$

$$m_1 = \frac{\kappa_0}{\kappa_0 + n}\, m_0 + \frac{n}{\kappa_0 + n}\, \widehat{\mu} \tag{3.16b}$$

$$a_1 = a_0 + n \tag{3.16c}$$

$$b_1 = b_0 + \frac{\kappa_0\, n}{\kappa_0 + n}\, (\widehat{\mu} - m_0)^2 + n\,(n-1)\,\widehat{\sigma}^2. \tag{3.16d}$$

Notice that the observations enter the posterior distribution solely via the sufficient statistic $(n, \widehat{\mu}, \widehat{\sigma})$.[8]

4. Truncated sampling distribution for the measured effect

In this section we examine the marginal sampling distribution for the *measured effect* $\widehat{\mu}$ (i.e., the sample mean) given the truncated sample space. This distribution allows us to characterize the Type M (magnitude) and Type S (sign) errors described by Gelman and Carlin (2014).

The sampling distribution for $(\widehat{\mu}, \widehat{\sigma})$ is given by[9]

$$p(\widehat{\mu}, \widehat{\sigma}|\mu, \tau^2, n) = p(\widehat{\mu}|\mu, \tau^2, n)\, p(\widehat{\sigma}|\tau^2, n), \tag{4.1}$$

where[10]

$$p(\widehat{\mu}|\mu, \tau^2, n) = \mathsf{N}\Big(\widehat{\mu}\,\Big|\,\mu, \frac{\tau^2}{n}\Big) \quad \text{and} \quad p(\widehat{\sigma}|\tau^2, n) = \mathsf{Nakagami}\Big(\widehat{\sigma}\,\Big|\,\frac{n-1}{2}, \frac{\tau^2}{n}\Big). \tag{4.2}$$

The truncation set may be characterized by $1(|\widehat{\mu}| > c_\alpha\, \widehat{\sigma})$ where $c_\alpha = f_{n-1}^{-1}(\alpha)$. The selection process produces a truncated sampling distribution for the sufficient statistic:

$$p(\widehat{\mu}, \widehat{\sigma}|\mu, \tau^2, n, \alpha, \mathcal{S}) = \frac{1(|\widehat{\mu}| > c_\alpha\, \widehat{\sigma})\, p(\widehat{\mu}, \widehat{\sigma}|\mu, \tau^2, n)}{\mathcal{P}(\mu/\tau, n, \alpha)}. \tag{4.3}$$

The truncated distribution is normalized by the probability of selection (i.e., the power).

The truncated sample space produces dependence between $\widehat{\mu}$ and $\widehat{\sigma}$. The marginal sampling distribution for $\widehat{\mu}$ given α is obtained by integrating $\widehat{\sigma}$ out over the selection region (i.e., $\widehat{\sigma} < |\widehat{\mu}|/c_\alpha$):

$$\begin{aligned} p(\widehat{\mu}|\mu, \tau^2, n, \alpha, \mathcal{S}) &= \int_0^{|\widehat{\mu}|/c_\alpha} p(\widehat{\mu}, \widehat{\sigma}|\mu, \tau^2, n, \alpha, \mathcal{S})\, d\widehat{\sigma} \\ &= \frac{p(\widehat{\mu}|\mu, \tau^2, n)}{\mathcal{P}(\mu/\tau, n, \alpha)}\, \mathcal{C}(|\widehat{\mu}|/c_\alpha, \tau^2, n), \end{aligned} \tag{4.4}$$

[8] The Jeffreys prior of $p(\mu, \tau^2) \propto 1/\tau^2$ produces the posterior (3.15)–(3.16) where $\kappa_0 = b_0 = 0$, $a_0 = -1$, and m_0 is unspecified. Consequently, $p(\mu|y) = \mathsf{Student\text{-}t}(\mu|\widehat{\mu}, \widehat{\sigma}^2, n-1)$.

[9] See Appendix A for omitted details.

[10] If $\widehat{\sigma} \sim \mathsf{Nakagami}(\frac{n-1}{2}, \frac{\tau^2}{n})$, then $\widehat{\sigma}^2 \sim \mathsf{Gamma}(\frac{n-1}{2}, \frac{2\,\tau^2}{n\,(n-1)})$.

where the "correction factor" $\mathcal{C}$ is the CDF of the Nakagami distribution evaluated at the boundary [see (A.7) and (A.8)]:

$$\mathcal{C}(x,\tau^2,n) = \int_0^x p(\widehat{\sigma}|\tau^2,n)\,d\widehat{\sigma} = 1 - \frac{\Gamma\left(\frac{n-1}{2}, x^2\left(\frac{n-1}{2}/\frac{\tau^2}{n}\right)\right)}{\Gamma(\frac{n-1}{2})}. \tag{4.5}$$

To illustrate (4.3) and (4.4), let $\mu = 0.1$, $\tau = 1$, and $\alpha = 5\%$. See Figure 4 for the sampling distribution for $(\widehat{\mu},\widehat{\sigma})$ subject to the restricted sample space. See Figure 5 for the marginal sampling distribution for $\widehat{\mu}_i$ computed from the distribution shown in Figure 4. See Figure 6 for the marginal distribution for $\widehat{\mu}$ for $n \in \{10, 20, 50\}$. Note that smaller studies are associated with larger measured effects in absolute value and larger probabilities of the incorrect sign. The correction factor is plotted in Figure 7.

5. Multiple datasets

Suppose there were multiple sets of observations, $y_{1:N} = (y_1, \ldots y_N) \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_N$, where $y_i = (y_{i1}, \ldots, y_{in_i}) \in \mathcal{Y}_i = \mathbb{R}^{n_i}$. Absent truncation, let

$$p(y_{1:N}|\mu,\tau^2) = \prod_{i=1}^N p(y_i|\mu,\tau^2) = \prod_{i=1}^N \prod_{j=1}^{n_i} \mathsf{N}(y_{ij}|\mu,\tau^2). \tag{5.1}$$

The posterior distribution for (μ,τ^2) is given by

$$p(\mu,\tau^2|y_{1:N}) \propto p(y_{1:N}|\mu,\tau^2)\,p(\mu,\tau^2). \tag{5.2}$$

The sufficient statistic for each dataset is $(n_i,\widehat{\mu}_i,\widehat{\sigma}_i)$. The sufficient statistic for the collection of datasets is given by $(n,\widehat{\mu},\widehat{\sigma})$, where $n = \sum_{i=1}^N n_i$ and

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^N \sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^N \left(\frac{n_i}{n}\right)\widehat{\mu}_i \tag{5.3a}$$

$$\widehat{\sigma}^2 = \frac{1}{n\,(n-1)}\sum_{i=1}^N \sum_{j=1}^{n_i} (y_{ij}-\widehat{\mu})^2 = \frac{1}{n-1}\sum_{i=1}^N \left(\frac{n_i}{n}\right)\Big((n_i-1)\,\widehat{\sigma}_i^2 + (\widehat{\mu}_i-\widehat{\mu})^2\Big). \tag{5.3b}$$

Now let us take the possibility of truncation into account. Let $\Lambda_i = \{y_i \in \mathcal{Y}_i : \widehat{\pi}_i < \alpha_i\}$. We have the likelihood

$$\begin{aligned} p(y_{1:N}|\mu,\tau^2,\alpha_{1:N},\mathcal{S}) &= \prod_{i=1}^N p(y_i|\mu,\tau^2,\alpha_i,\mathcal{S}) \\ &= \prod_{i=1}^N \frac{1(\widehat{\pi}_i < \alpha_i)\,p(y_i|\mu,\tau^2)}{\mathcal{P}(\mu/\tau,n_i,\alpha_i)} = p(y|\mu,\tau^2)\prod_{i=1}^N \frac{1(\widehat{\pi}_i<\alpha_i)}{\mathcal{P}(\mu/\tau,n_i,\alpha_i)}. \end{aligned} \tag{5.4}$$

Assuming $p(\alpha_{1:N}) = \prod_{i=1}^N p(\alpha_i)$, the posterior is given by

$$p(\mu,\tau^2,\alpha_{1:N}|y_{1:N},\mathcal{S}) \propto p(\mu,\tau^2|y_{1:N})\prod_{i=1}^N \frac{1(\widehat{\pi}_i<\alpha_i)\,p(\alpha_i)}{\mathcal{P}(\mu/\tau,n_i,\alpha_i)}, \tag{5.5}$$

If all datasets are subject to truncation by the same significance level, the final factor in (5.5) is replaced by

$$p(\alpha)\prod_{i=1}^{N}\frac{1(\widehat{\pi}_i<\alpha)}{\mathcal{P}(\mu/\tau,n_i,\alpha)}=p(\alpha)\,1(\widehat{\pi}_{\max}<\alpha)\prod_{i=1}^{N}\frac{1}{\mathcal{P}(\mu/\tau,n_i,\alpha)},\tag{5.6}$$

where $\widehat{\pi}_{\max}=\max(\widehat{\pi}_{1:N})$.

Numerical illustration. In planning an experiment, the researchers decided that the study should be published in the *Journal of Correct Signs*. This journal only publishes studies of experiments for which $\widehat{\pi}_i<\alpha_i$ for some α_i. (A rationale for this decision rule is given in Section 7.) If the study is not accepted for publication, then the result of the experiment will not be observed (by the public).

Let $\mu=0.1$, $\tau=1$, and $\alpha=5\%$.

From a total of $N^*=200$ studies conducted (with sample sizes n_i ranging from 5 to 50), $N=20$ studies (10%) satisfied the acceptance criterion for publication. See Figure 9. The acceptance rate for publication is in line with the rejection rate of the null hypothesis given the average power of about 8%. (The power ranges from about 5.5% for $n_i=5$ to about 10.5% for $n_i=50$.)

Let the prior be $p(\mu,\tau)\propto 1/\tau$. Figure 10 shows the joint posterior distribution for (μ,τ) given y that takes the selection process into account. The associated marginal posterior distributions for μ and τ^2 are shown in Figures 11 and 12. Figure 13 compares $p(\mu|y,\alpha,\mathcal{S},\mathcal{M},\mathcal{J})$ with $p(\mu|y,\mathcal{M},\mathcal{J})$.

Up to this point we have assumed the value for was known to be 5%. Now let the prior for α be given by (3.12). The marginal likelihoods $p(y|\alpha_j,\mathcal{S},\mathcal{M},\mathcal{J})$ can be calculated following the procedure described in Appendix A. The posterior odds ratios are equal to the Bayes factors since the prior probabilities are all equal. For comparison purposes, let us take $\alpha=5\%$ as the base model. Note that $\alpha=1\%$ is impossible because $1\%<\pi_{\max}=0.049$. The Bayes factor in favor of $\alpha=10\%$ is on the order of 4×10^{-5} (not very likely) and the Bayes factor in favor of $\alpha=100\%$ (no selection process) is on the order of 6×10^{-12} (extremely unlikely). Thus the evidence overwhelmingly favors $\alpha=5\%$ to the alternatives.

6. A more general model: Some considerations

In this section I introduce some generalizations. Thus far, the parameters (μ,τ) have been shared across all studies. We can generalized the likelihood to allow for study-specific values of these parameters: $p(y_i|\mu_i,\tau_i,\alpha_i,\mathcal{S})$. Having a well-structured prior for $\{(\mu_i,\tau_i)\}_{i=1}^N$ becomes important.

Different types of experiments involving the same effect may exhibit heterogeneity across the amount of experimental noise. For this case, one can model the individual noises independently or one can tie them together via prior jointness.

The point of a meta-analysis is to combine the results of different studies of the same effect. Nevertheless, it may be the case that different experiments measure the same effect in different ways. In this case, the goal is to compute the posterior distribution for the *generic* effect. We can think of this a density estimation for latent variables.

Start here:

$$p(y_{1:N}|\mu_{1:N},\tau^2_{1:N},\mathcal{S}) = \prod_{i=1}^{N} p(y_i|\mu_i,\tau_i^2,\alpha_i,\mathcal{S}), \tag{6.1}$$

where

$$p(y_i|\mu_i,\tau_i^2,\alpha_i,\mathcal{S}) = \frac{1(\widehat{\pi}_i < \alpha_i)\, p(y_i|\mu_i,\tau_i^2)}{\mathcal{P}(\mu_i/\tau_i, n_i, \alpha_i)} \propto \frac{1(\widehat{\pi}_i < \alpha_i)\, p(\widehat{\mu}_i,\widehat{\sigma}_i|\mu_i,\tau_i,n_i)}{\mathcal{P}(\mu_i/\tau_i, n_i, \alpha_i)}. \tag{6.2}$$

Let

$$p(\mu_{1:N},\tau_{1:N},\alpha_{1:N}|\psi_\mu,\psi_\tau,\psi_\alpha) = p(\mu_{1:N}|\psi_\mu)\, p(\tau_{1:N}|\psi_\tau)\, p(\alpha_{1:N}|\psi_\alpha), \tag{6.3}$$

and where (for example)

$$p(\mu_{1:N}|\psi_\mu) = \prod_{i=1}^{N} p(\mu_i|\psi_\mu) \tag{6.4}$$

where (for example)

$$p(\mu_i|\psi_\mu) = \sum_{c=1}^{\infty} w_{c\mu}\, f_\mu(\mu|\theta_\mu). \tag{6.5}$$

7. Deciding whether to publish a study

This section deals with the editor of the journal. The editor's decision problem is described and then it is shown how the editor can use prior knowledge in the decision process.

The decision problem. An editor must make a decision as to whether to publish a study or not. Let δ denote the decision, where $\delta \in \{\text{accept}, \text{reject}\}$. (The terms accept and reject refer to the publication decision and not to the null hypothesis. Indeed, if the null hypothesis is rejected then the paper will be accepted (and vice-versa).)

Consider the decision problem facing the editor of the *Journal of Correct Signs*. Papers published in this journal claim either a positive effect or a negative effect. An author's claim, denoted γ, is determined by the sign of $\widehat{\mu}$:

$$\gamma = \big(\text{sgn}(\mu) = \text{sgn}(\widehat{\mu}_i)\big). \tag{7.1}$$

The editor wishes to publish papers for which the claim about the sign is correct — i.e., for which

$$1(\gamma) = 1. \tag{7.2}$$

The probability that the claim is correct can be expressed as

$$\xi = \Pr[\,1(\gamma) = 1 \,|\, y, \mathcal{E}] = \begin{cases} \Pr[\mu > 0 | y, \mathcal{E}] & \widehat{\mu} > 0 \\ \Pr[\mu < 0 | y, \mathcal{E}] & \widehat{\mu} < 0 \end{cases}, \tag{7.3}$$

where $\mathcal{E}$ denotes the editor's prior information. (Two forms of such prior information are described below.)

The optimal decision can be characterized as minimizing the expected loss given the information the editor has. Suppose the loss function has the following form:

$$L(\gamma, \delta) = \begin{cases} \ell_{0r} & 1(\gamma) = 0 \text{ and } \delta = \text{reject} \\ \ell_{0a} & 1(\gamma) = 0 \text{ and } \delta = \text{accept} \\ \ell_{1r} & 1(\gamma) = 1 \text{ and } \delta = \text{reject} \\ \ell_{1a} & 1(\gamma) = 1 \text{ and } \delta = \text{accept} \end{cases}. \tag{7.4}$$

The expected losses from the two possible decisions are

$$E[L(\gamma, \text{reject})|y, \mathcal{E}] = \ell_{0r}\,(1 - \xi) + \ell_{1r}\,\xi \tag{7.5}$$

$$E[L(\gamma, \text{accept})|y, \mathcal{E}] = \ell_{0a}\,(1 - \xi) + \ell_{1a}\,\xi. \tag{7.6}$$

Suppose there is no loss for making the correct decision: $\ell_{0r} = \ell_{1a} = 0$. Then the editor will accept the study for publication if $\ell_{0a}\,(1 - \xi) < \ell_{1r}\,\xi$, which can be expressed as

$$\frac{\xi}{1 - \xi} > \frac{\ell_{0a}}{\ell_{1r}}. \tag{7.7}$$

The expression on the left-hand side is the posterior odds ratio in favor of the claim being true, while the expression on the right-hand side is the ratio of the loss of publishing a paper with a false claim to the loss of rejecting a paper with a true claim. Further suppose

$$\frac{\ell_{0a}}{\ell_{1r}} = \frac{1 - \alpha/2}{\alpha/2}. \tag{7.8}$$

(For example, if $\alpha = 1/20$ then $\ell_{0a}/\ell_{1r} = 39$.) In this case the editor will accept the paper for publication if

$$2\,(1 - \xi) < \alpha. \tag{7.9}$$

We now turn to how an editor may determine ξ_i. We consider two types of editor: the ignorant/objective editor and the informed editor.

An ignorant/objective editor. We characterize the ignorant/objective editor in terms of the Jeffreys prior. The marginal posterior distribution for μ given the Jeffreys prior is $p(\mu|y, \mathcal{J}) = \mathsf{Student\text{-}t}(\mu|\widehat{\mu}, \widehat{\sigma}^2, n - 1)$ [see (D.5)]. Using this distribution, the probability that μ has the same sign as $\widehat{\mu}$ is

$$\begin{aligned} \xi = \Pr[\,1(\gamma) = 1\,|y, \mathcal{J}] &= \begin{cases} \int_0^\infty p(\mu|y, \mathcal{J})\,d\mu & \widehat{\mu} > 0 \\ \int_{-\infty}^0 p(\mu|y, \mathcal{J})\,d\mu & \widehat{\mu} < 0 \end{cases} \\ &= \Phi_{n-1}(|\widehat{t}|) \\ &= 1 - \widehat{\pi}/2. \end{aligned} \tag{7.10}$$

Thus, the p-value equals twice the probability that μ has the *opposite* sign of $\widehat{\mu}$:

$$\widehat{\pi} = 2\,(1 - \xi). \tag{7.11}$$

Therefore, an ignorant/objective editor would publish the paper if $\widehat{\pi} < \alpha$.

(We can visualize (7.10) by considering the posterior probability of the equivalent condition $\text{sgn}(\mu/\widehat{\sigma}) = \text{sgn}\big(\widehat{t}\big)$. The visualization is presented in Figure 14 for $\widehat{\mu} > 0$.)

An informed editor. Suppose the editor has the conjugate prior for (μ, τ^2) given by (3.14). The marginal posterior for μ is given by $\mathsf{Student\text{-}t}(\mu|m_1, s_1^2, a_1)$, where $s_1^2 = b_1/(a_1\,\kappa_1)$. Define $t_1 = m_1/s_1$ and $\pi_1 = f_{a_1}(|t_1|)$. Following (7.10), $\xi = 1 - \pi_1/2$. Given the loss function described above, the informed editor would publish the paper if $\pi_1 < \alpha$. To the outside observer who does not know the editor's prior, the acceptance criterion will appear to be $\widehat{\pi} < \widetilde{\alpha}$, where the apparent significance level, $\widetilde{\alpha} = (\widehat{\pi}/\pi_1)\,\alpha$, depends on the prejudices of the editor.

Appendix A. Additional material

Sampling theory. We know from from sampling theory

$$z := \frac{\widehat{\mu}_i - \mu}{\tau/\sqrt{n_i}} \,\Big|\, \mu, \tau, n_i \sim \mathsf{N}(0,1) \tag{A.1}$$

and

$$v := \frac{n_i\,(n_i-1)\,\widehat{\sigma}_i^2}{\tau^2} \,\Big|\, \tau, n_i \sim \chi^2_{n_i-1}. \tag{A.2}$$

We also know z and v are (conditionally) independent. Therefore,

$$\widehat{\mu}_i = \frac{\tau\,z}{\sqrt{n_i}} + \mu \,\Big|\, \mu, \tau, n_i \sim \mathsf{N}(\mu, \tau^2/n_i) \tag{A.3}$$

and

$$\widehat{\sigma}_i = \sqrt{\frac{\tau^2\,v}{n_i\,(n_i-1)}} \,\Big|\, \tau, n_i \sim \mathsf{Nakagami}\Big(\frac{n_i-1}{2}, \frac{\tau^2}{n_i}\Big), \tag{A.4}$$

where $\widehat{\mu}_i$ and $\widehat{\sigma}_i$ are (conditionally) independent. The mean and variance of $\widehat{\sigma}_i$ are $\zeta_i\,(\tau/\sqrt{n_i})$ and $(1-\zeta_i^2)\,(\tau^2/n_i)$, where

$$\zeta_i = \frac{\Gamma(n_i/2)}{\Gamma\big((n_i-1)/2\big)}\,\sqrt{\frac{2}{n_i-1}}. \tag{A.5}$$

Note $\zeta_i < 1$ and $\lim_{n_i\to\infty}\ \zeta_i = 1$.

Nakagami distribution. If $x \sim \mathsf{Gamma}(a,c)$, then $\sqrt{x} \sim \mathsf{Nakagami}(a,b)$ where $b = a\,c$. The PDF for the Nakagami distribution is

$$\mathsf{Nakagami}(x|a,b) = \frac{2\,\big(\frac{a}{b}\big)^a\ x^{2\,a-1} e^{-\frac{a\,x^2}{b}}}{\Gamma(a)}. \tag{A.6}$$

The CDF for the Nakagami distribution is

$$\int_0^x \mathsf{Nakagami}(s|a,b)\,ds = 1 - \frac{\Gamma\big(a, x^2\,(a/b)\big)}{\Gamma(a)}, \tag{A.7}$$

where $\Gamma(a,c) = \int_c^\infty t^{a-1}\,e^{-t}\,dt$ is the incomplete gamma function. Thus

$$\int_0^{|\widehat{\mu}_i|/c_i} p(\widehat{\sigma}_i|\tau,n_i)\,d\widehat{\sigma}_i = 1 - \frac{\Gamma\big(\frac{n_i-1}{2}, (\widehat{\mu}_i/c_i)^2\,(\frac{n_i-1}{2}/\frac{\tau^2}{n_i})\big)}{\Gamma(\frac{n_i-1}{2})}. \tag{A.8}$$

Sampling distribution for the t statistic. The sampling distribution for the t statistic can be derived from the distribution for $(\widehat{\mu}_i, \widehat{\sigma}_i)$. By the change of variables formula the sampling distribution for $(\widehat{t}_i, \widehat{\sigma}_i)$ is

$$\begin{aligned} p(\widehat{t}_i, \widehat{\sigma}_i|\mu, \tau, n_i) &= \widehat{\sigma}_i \, p(\widehat{\mu}_i, \widehat{\sigma}_i|\mu, \tau, n_i)|_{\widehat{\mu}_i = \widehat{t}_i \, \widehat{\sigma}_i} \\ &= \mathsf{N}(\widehat{t}_i|\mu/\widehat{\sigma}_i, (\tau^2/n_i)/\widehat{\sigma}_i) \, \mathsf{Nakagami}(\widehat{\sigma}_i|(n_i - 1)/2, \tau^2/n_i). \end{aligned} \tag{A.9}$$

We can integrate out $\widehat{\sigma}_i$ thereby obtaining the sampling distribution for the t statistic:

$$\begin{aligned} p(\widehat{t}_i|\mu, \tau, n_i) &= \int_0^\infty p(\widehat{t}_i, \widehat{\sigma}_i|\mu, \tau, n_i) \, d\widehat{\sigma}_i \\ &= \mathsf{Noncentral\text{-}t}\big(\widehat{t}_i|n_i - 1, \sqrt{n_i}\,(\mu/\tau)\big). \end{aligned} \tag{A.10}$$

The power function. We can express the likelihood in terms of these sample statistics as

$$p(y_i|\mu, \tau) = \prod_{j=1}^{n_i} \mathsf{N}(y_{ij}|\mu, \tau^2) = p(\widehat{\mu}_i, \widehat{\sigma}_i|n_i, \mu, \tau) \, h(y_i), \tag{A.11}$$

where $z_i = (\widehat{\mu}_i, \widehat{\sigma}_i)$ is a sufficient statistic for y_i and where $\mathcal{Z}_i = \mathbb{R} \times \mathbb{R}^+$.

Given this criterion, the restricted space for the sufficient statistic is

$$\Omega_{c_i} = \{(\widehat{\mu}_i, \widehat{\sigma}_i) \in \mathcal{Z}_i : |\widehat{\mu}_i| > c_i \, \widehat{\sigma}_i\}, \tag{A.12}$$

where the selection criterion has been expressed explicitly in terms of the sufficient statistic.

The power function can be computed using the sampling distribution for the t statistic, which itself is computed from the sampling distribution for the sufficient statistic:

$$p(\widehat{t}_i|\mu, \tau, n_i) = \mathsf{Noncentral\text{-}t}\big(\widehat{t}_i|n_i - 1, \sqrt{n_i}\,(\mu/\tau)\big), \tag{A.13}$$

with $n_i - 1$ degrees of freedom and noncentrality parameter $\sqrt{n_i}\,(\mu/\tau)$. Note the sampling distribution for the t statistic depends on μ and τ only via their ratio μ/τ. Also note

$$\mathsf{Noncentral\text{-}t}(\nu, 0) \equiv \mathsf{Student\text{-}t}(0, 1, \nu). \tag{A.14}$$

The power function is given by

$$\begin{aligned} \mathcal{P}_{\Omega_{c_i}}(\mu, \tau) &= 1 - \int_{-c_i}^{c_i} p(\widehat{t}_i|\mu, \tau, n_i) \, d\widehat{t}_i \\ &= 1 - \int_{-c_i}^{c_i} \mathsf{Noncentral\text{-}t}\big(\widehat{t}_i|n_i - 1, \sqrt{n_i}\,(\mu/\tau)\big) \, d\widehat{t}_i \\ &= 1 - \Phi_{n_i - 1, \sqrt{n_i}\,(\mu/\tau)}(c_i) + \Phi_{n_i - 1, \sqrt{n_i}\,(\mu/\tau)}(-c_i). \end{aligned} \tag{A.15}$$

Figure 15 illustrates (A.15).

Numerical evaluation. Consider the case of a meta-analysis. (The case of a single study is a simple specialization.) Most of the computation is involved in evaluating the power function. For each required value of n_i and α_j, one can precompute $\mathcal{P}(\lambda_k, n_i, \alpha_j)$ over a grid $\{\lambda_k\}$ from 0 to $\ell_{\alpha_j}^{n_i}$ where $\mathcal{P}(\ell_{\alpha_j}^{n_i}, n_i, \alpha_j) \approx 1$. Let

$$K_j(\mu, \tau) := \frac{p(\mu, \tau|y, \mathcal{M}, \mathcal{I})}{\prod_{i=1}^N \mathcal{P}(\mu/\tau, n_i, \alpha_j)}. \tag{A.16}$$

One can numerically integrate $K_j(\mu, \tau)$ to obtain

$$p(y|\alpha_j, \mathcal{S}, \mathcal{M}, \mathcal{I}) = \int_0^\infty \int_{-\infty}^\infty K_j(\mu, \tau)\, d\mu\, d\tau \tag{A.17}$$

and compute

$$p(y|\mathcal{S}, \mathcal{M}, \mathcal{I}) = \sum_{j=1}^{J} p(\alpha_j)\, p(y|\alpha_j, \mathcal{S}, \mathcal{M}, \mathcal{I}). \tag{A.18}$$

Posterior probabilities for α_j are given by

$$p(\alpha_j|y, \mathcal{S}, \mathcal{M}, \mathcal{I}) = \frac{p(\alpha_j)\, p(y|\alpha_j, \mathcal{S}, \mathcal{M}, \mathcal{I})}{p(y|\mathcal{S}, \mathcal{M}, \mathcal{I})}. \tag{A.19}$$

Implied prior for λ. Given the informed editor's prior for (μ, τ^2) [see (**??**)], the general expression for the implied prior for λ is

$$p(\lambda|\mathcal{E}) = \mathsf{N}(\lambda|0, 1/\kappa) \times \left(\frac{\nu\, s^2}{\kappa\, m^2 + \nu\, s^2}\right)^{\nu/2} \times \left\{ {}_1F_1\left(\frac{\nu}{2}; \frac{1}{2}; \frac{m^2\kappa^2\lambda^2}{2\,(\kappa\, m^2 + \nu\, s^2)}\right) + \frac{\sqrt{2}\,\kappa\,\lambda\, m\, \Gamma\left(\frac{\nu+1}{2}\right)\, {}_1F_1\left(\frac{\nu+1}{2}; \frac{3}{2}; \frac{m^2\kappa^2\lambda^2}{2\,(\kappa\, m^2+\nu\, s^2)}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\kappa\, m^2 + \nu\, s^2}} \right\}, \tag{A.20}$$

where ${}_1F_1(\,\cdot\,, \cdot\,, \cdot\,)$ is the Kummer confluent hypergeometric function.

Appendix B. A simple example: Widgets in a box

This section presents a simple example that illustrates many of the points discussed in the main text. Some readers may find it useful.

A machine produces widgets in batches and deposits them in a box. Each time the machine is activated it is supposed to produce a single widget, but sometimes it fails and produces nothing. The probability of success for a given activation does not depend on the success of other activations.

Box i has n_i slots, each of which can hold one widget. The machine is activated once for each slot. Let y_{ij} indicate whether slot j in box i holds a widget after the machine is finished with its activations:

$$y_{ij} = \begin{cases} 1 & \text{success: slot } j \text{ in box } i \text{ is occupied} \\ 0 & \text{failure: slot } j \text{ in box } i \text{ is empty} \end{cases}. \tag{B.1}$$

We may write $y_{ij} \in \mathcal{Y} = \{0, 1\}$ and $y_i = (y_{i1}, \ldots, y_{in_i}) \in \mathcal{Y}^{n_i} = \{0, 1\}^{n_i}$. The sample space $\mathcal{Y}^{n_i}$ contains 2^{n_i} elements. Let θ denote the probability of success. (I will also refer to θ as the efficiency of the machine). The sampling distribution for a single activation can be expressed as

$$p(y_{ij}|\theta) = \mathsf{Bernoulli}(y_{ij}|\theta) = (1-\theta)^{1-y_{ij}}\, \theta^{y_{ij}}. \tag{B.2}$$

The sampling distribution for the sequence of slots in a box is given by

$$p(y_i|\theta) = \prod_{j=1}^{n_i} p(y_{ij}|\theta) = \prod_{j=1}^{n_i} (1-\theta)^{1-y_{ij}}\, \theta^{y_{ij}} = (1-\theta)^{n_i - z_i}\, \theta^{z_i}, \tag{B.3}$$

where

$$z_i := \sum_{j=1}^{n_i} y_{ij} \tag{B.4}$$

is the number of widgets in box i.

Here is an example of a truncation. Let $n_i = 4$ and let

$$\Lambda_i = \{y_i \in \{0,1\}^4 : z_i = 0 \ \vee z_i = 4\}. \tag{B.5}$$

Note

$$\mathbb{P}_{\Lambda_i}(\theta) = (1-\theta)^4 + \theta^4. \tag{B.6}$$

Null hypothesis significance testing. A researcher at the factory examines box i as it comes off the production line, observes the sequence of occupied and empty slots, and records the dataset y_i. Given this dataset, the researcher conducts a null hypothesis significance test regarding the efficiency of the machine. The null hypothesis is $H_0 : \theta = 1/2$ and the alternative hypothesis is $H_1 : \theta \neq 1/2$.

To conduct the test, the researcher chooses z_i as the test statistic. The sample space for z_i is

$$\mathcal{Z}_i = \{0, 1, \ldots, n_i\}, \tag{B.7}$$

and the sampling distribution for this test statistic is

$$p(z_i|n_i,\theta) = \mathsf{Binomial}(z_i|n_i,\theta) = \binom{n_i}{z_i} (1-\theta)^{n_i - z_i}\, \theta^{z_i}. \tag{B.8}$$

Note that $\sum_{z_i \in \mathcal{Z}_i} p(z_i|n_i,\theta) = 1$. Consider a subset $\Omega_i \subset \mathcal{Z}_i$ and the probability

$$\Pr[z_i \in \Omega_i \,|\, \theta] = \sum_{z_i \in \Omega_i} p(z_i|n_i,\theta). \tag{B.9}$$

The researcher happened to choose a box with $n_i = 4$ slots. For this box, the probability distribution for the number of widgets in box i is

$$p(z_i|n_i = 4,\theta) = \begin{cases} (1-\theta)^4 & z_i = 0 \\ 4\,(1-\theta)^3\,\theta & z_i = 1 \\ 6\,(1-\theta)^2\,\theta^2 & z_i = 2 \\ 4\,(1-\theta)\,\theta^3 & z_i = 3 \\ \theta^4 & z_i = 4 \end{cases}. \tag{B.10}$$

Given $n_i = 4$ and $\mathcal{Z}_i = \{0,1,2,3,4\}$, the researcher wishes to associate extreme values of the test statistic with rejection of the null hypothesis. To this end, the researcher constructs the critical region (also known as the rejection region) as follows:

$$\Omega_i = \{z_i \in \mathcal{Z}_i : |z_i - 2| > 1\} = \{0, 4\}. \tag{B.11}$$

Note that Ω_i is a proper subset of $\mathcal{Z}_i$. If $z_i \in \Omega_i$ then the null hypothesis will be rejected. The power function is the probability that the null hypothesis is rejected (as a function of the probability of success):

$$\mathcal{P}_{\Omega_i}(\theta) = \sum_{z_i \in \Omega_i} \mathsf{Binomial}(z_i | n_i = 4, \theta) = (1-\theta)^4 + \theta^4. \tag{B.12}$$

Note we are sure to reject if either $\theta = 0$ or $\theta = 1$: $\mathcal{P}_{\Omega_i}(\theta = 0) = \mathcal{P}_{\Omega_i}(\theta = 1) = 1$.

The significance level of the hypothesis test, denoted α_i, equals the power function evaluated at the null hypothesis:

$$\alpha_i = \mathcal{P}_{\Omega_i}(\theta = 1/2) = 1/8. \tag{B.13}$$

So the probability of rejecting the null hypothesis when the null hypothesis is true is 1/8. If the null hypothesis is rejected then the experimental result will be declared significant with a significance level of 1/8.

The researcher submits the paper to the *Journal of Significant Widget Results*. This journal only publishes papers that report significant results (with a significance level of 1/8). As it turns out, $z_i = 4$ and the paper is published.

Inference. Inference regarding θ based on the distribution for the test statistic z_i is equivalent to inference based on the distribution for the dataset y_i because the ratio

$$\frac{p(z_i | n_i, \theta)}{p(y_i | \theta)} = \binom{n_i}{z_i} \tag{B.14}$$

does not depend on θ. Interpreted as likelihoods for θ, the two sampling distributions contain the same information. This may be summarized by noting that z_i (along with n_i) is a sufficient statistic for y_i. The upshot is that for any prior distribution $p(\theta)$ the two posterior distributions are the same:

$$p(\theta | n_i, z_i) = \frac{p(z_i | n_i, \theta)\, p(\theta)}{\int p(z_i | n_i, \theta)\, p(\theta)\, d\theta} = \frac{p(y_i | \theta)\, p(\theta)}{\int p(y_i | \theta)\, p(\theta)\, d\theta} = p(\theta | y_i). \tag{B.15}$$

In the published paper, the researcher reports the likelihood for θ given the outcome $z_i = 4$ based on the sampling distribution (B.10) as

$$p(z_i = 4 | n_i = 4, \theta) = \mathsf{Binomial}(z_i = 4 | n_i = 4, \theta) = \theta^4. \tag{B.16}$$

The researcher points out that if the prior for θ were $p(\theta) = 1$, then the posterior for θ would be

$$p(\theta | n_i = 4, z_i = 4) = \frac{p(z_i = 4 | n_i = 4, \theta)\, p(\theta)}{\int_0^1 p(z_i = 4 | n_i = 4, \theta)\, p(\theta)\, d\theta} = 5\,\theta^4. \tag{B.17}$$

What can a reader of the *JSWR* infer about the efficiency of the machine from this published study? As far as the reader is concerned, the data-generating mechanism includes the *selection process*: the journal does not publish studies with insignificant results. Therefore, datasets for which $z_i \in \mathcal{Z}_i \setminus \Omega_i = \{1, 2, 3\}$ cannot be observed. Consequently, the sample space for datasets is restricted to

$$\Lambda_i = \{y_i \in \mathcal{Y}^{n_i} : z_i \in \Omega_i\}. \tag{B.18}$$

Because z_i is a sufficient statistic, we may continue to conduct inference based on its relevant sample space, which now is restricted to Ω_i.

The sampling distribution needs to be normalized by the probability of the restricted sample space. This normalization amounts to a form of inverse probability weighting using the power function. Accordingly, the sampling distribution for the restricted sample space is given by

$$p(z_i|n_i = 4, \theta, \mathcal{S}_i) = \frac{\mathsf{Binomial}(z_i|n_i = 4, \theta)}{\mathcal{P}_{\Omega_i}(\theta)} = \begin{cases} \frac{(1-\theta)^4}{(1-\theta)^4+\theta^4} & z_i = 0 \\ \frac{\theta^4}{(1-\theta)^4+\theta^2} & z_i = 4 \end{cases}, \tag{B.19}$$

where $\mathcal{S}_i$ denotes the selection process. The likelihood based on the sampling distribution (B.19) given $z_i = 4$ is

$$p(z_i = 4|n_i = 4, \theta, \mathcal{S}_i) = \frac{\mathsf{Binomial}(z_i = 4|n_i = 4, \theta)}{\mathcal{P}_{\Omega_i}(\theta)} = \frac{\theta^4}{(1-\theta)^4 + \theta^4}. \tag{B.20}$$

If the prior for θ were $p(\theta) = 1$, then the posterior for θ would be

$$p(\theta|n_i = 4, z_i = 4, \mathcal{S}_i) = \frac{p(z_i = 4|n_i = 4, \theta, \mathcal{S}_i)\, p(\theta)}{\int_0^1 p(z_i = 4|n_i = 4, \theta, \mathcal{S}_i)\, p(\theta)\, d\theta} = \frac{2\,\theta^4}{(1-\theta)^4 + \theta^4}. \tag{B.21}$$

See Figure 16 for plots of (B.17) and (B.21).

Meta-studies. Over time, the *JSWR* published 100 studies of the widget machine, each with $n_i = 4$ and $\Omega_i = \{0, 4\}$. (It is straightforward to combine studies with different sample sizes and different selection criteria into a single meta study.) Of the 100 studies, 96 reported $z_i = 4$ and four reported $z_i = 0$.

A first meta-study aggregates the data without considering the restrictions on the sample spaces imposed by the selection criteria. This study reports a likelihood of

$$p(z|n, \theta) = p(z_i = 0|n_i = 4, \theta)^4\, p(z_i = 4|n_i = 4, \theta)^{96} = (1-\theta)^{16}\, \theta^{384}, \tag{B.22}$$

where $z = (z_1, \ldots, z_{100})$ and $n = (n_1, \ldots, n_{100})$. The maximum of $p(z|n, \theta)$ occurs at $\theta = 0.96$. (In passing, it may be noted that the studies that reported $z_i = 0$ were considered by many researchers to have been conducted incompetently because the probability that a correctly conducted study would have produced $z_i = 0$ given the well-established value of $\theta \geq 0.96$ is less than 3×10^{-6}.)

A second meta-study takes the restricted sample space Ω_i into account and reports a likelihood of

$$p(z|n, \theta, \mathcal{S}) = \frac{p(z|n, \theta)}{\prod_{i=1}^{100} \mathcal{P}_{\Omega_i}(\theta)} = \frac{(1-\theta)^{16}\, \theta^{384}}{((1-\theta)^4 + \theta^4)^{100}}, \tag{B.23}$$

where $\mathcal{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_{100})$. The factor $\left(\prod_{i=1}^{100} \mathcal{P}_{\Omega_i}(\theta)\right)^{-1}$ in the likelihood acts like a strong prior that shrinks the likely values of θ toward $1/2$. The maximum of $p(z|n, \theta, \mathcal{S})$ occurs as $\theta = 0.69$.

See Figure 17 for posterior distributions computed using $p(\theta) = 1$, so that $p(\theta|n, z) \propto p(z|n, \theta)$ and $p(\theta|n, z, \Omega) \propto p(z|n, \theta, \Omega)$.

Increased sample size. Holding the significance level fixed, an increase in the sample size will increase the power. If the power is increased sufficiently, the power function becomes relatively flat over the rejection region with the result that the normalization factor has very little effect.

Example. Suppose

$$\Omega_i = \{z_i \in \mathcal{Z}_i : |z_i - n_i/2| > \zeta_i\}. \tag{B.24}$$

For $n_i = 4$ we set $\zeta_i = 1$ which produced $\Omega_i = \{0, 4\}$ and delivered $\alpha_i = \mathcal{P}_{\Omega_i}(\theta = 1/2) = 0.125$. For $n_i = 408$ and $\zeta_i = 15$, we get $\alpha_i \approx 0.125$. Suppose a single study finds $z_i = 279$. In this case, the adjustment to the likelihood, $1/\mathcal{P}_{\Omega_i}(\theta)$, produces no noticeable effect on the posterior distribution.

Uncertainty about the significance level. Suppose that there is uncertainty about the significance level used in the selection process. In particular, suppose there are two possible values for the significance level:

$$\alpha_i \in A = \{1/8, 1\}. \tag{B.25}$$

It is convenient to express the power function in terms of the significance level α_i instead of the restricted set Ω_i. Consider the case where $n_i = 4$. Then

$$\alpha_i = \mathcal{P}_{\Omega_i}(\theta = 1/2) = \begin{cases} 1/8 & \Omega_i = \{0, 4\} \\ 1 & \Omega_i = \{0, 1, 2, 3, 4\} \end{cases}. \tag{B.26}$$

With this correspondence, we can express the power function as

$$\mathcal{P}(\theta, n_i = 4, \alpha_i) = \begin{cases} (1-\theta)^4 + \theta^4 & \alpha_i = 1/8 \\ 1 & \alpha_i = 1 \end{cases}. \tag{B.27}$$

This allows us to express the likelihood associated with the selection process in terms of the significance level:

$$p(z_i|n_i, \theta, \alpha_i) = \frac{p(z_i|n_i, \theta)}{\mathcal{P}(\theta, n_i, \alpha_i)}. \tag{B.28}$$

Assuming prior independence between θ and α_i [i.e., $p(\theta, \alpha_i) = p(\theta)\, p(\alpha_i)$], the joint posterior distribution for θ and α_i is given by

$$p(\theta, \alpha_i|z_i, n_i, \mathcal{S}_i) = \frac{p(z_i|n_i, \theta, \alpha_i)\, p(\theta)\, p(\alpha_i)}{p(z_i|n_i, \mathcal{S}_i)}, \tag{B.29}$$

where the marginal likelihood is given by

$$p(z_i|n_i, \mathcal{S}_i) = \sum_{\alpha_i \in A} p(\alpha_i) \int_0^1 p(z_i|n_i, \theta, \alpha_i)\, p(\theta)\, d\theta. \tag{B.30}$$

Let the prior be given by $p(\theta) = 1$ and

$$p(\alpha_i) = \begin{cases} 1/2 & \alpha_i = 1/8 \\ 1/2 & \alpha_i = 1 \end{cases}. \tag{B.31}$$

Then the marginal likelihood is

$$p(z_i = 4|n_i = 4, \mathcal{S}_i) = \frac{1}{2}\int_0^1 \frac{\theta^4}{(1-\theta)^4 + \theta^4} + \theta^4 \, d\theta = \frac{7}{20}, \tag{B.32}$$

and consequently the joint posterior distribution is

$$p(\theta, \alpha_i|z_i = 4, n_i = 4, \mathcal{S}_i) = \begin{cases} \left(\frac{5}{7}\right) \frac{2\,\theta^4}{(1-\theta)^4+\theta^4} & \alpha_i = 1/8 \\ \left(\frac{2}{7}\right) 5\,\theta^4 & \alpha_i = 1 \end{cases}. \tag{B.33}$$

The marginal posterior for θ is a mixture of the two posterior distributions displayed in Figure 16:

$$p(\theta|z_i = 4, n_i = 4, \mathcal{S}_i) = \sum_{\alpha_i \in A} p(\theta, \alpha_i|z_i = 4, n_i = 4, \mathcal{S}_i) = \left(\frac{5}{7}\right) \frac{2\,\theta^4}{(1-\theta)^4 + \theta^4} + \left(\frac{2}{7}\right) 5\,\theta^4. \tag{B.34}$$

The marginal posterior for α_i is

$$p(\alpha_i|z_i = 4, n_i = 4, \mathcal{S}_i) = \int_0^1 p(\theta, \alpha_i|z_i = 4, n_i = 4, \mathcal{S}_i) \, d\theta = \begin{cases} 5/7 & \alpha_i = 1/8 \\ 2/7 & \alpha_i = 1 \end{cases}. \tag{B.35}$$

Thus the probability of $\alpha_i = 1/8$ increases from $1/2$ (according the prior) to $5/7$ (according to the posterior).

Additional expression. In passing, it is interesting to note that we can express the joint posterior distribution for θ and α_i in terms of the posterior for θ absent the selection process and an additional factor:

$$\begin{aligned} p(\theta, \alpha_i|z_i, n_i, \mathcal{S}_i) &\propto p(z_i|n_i, \theta, \alpha_i)\, p(\theta)\, p(\alpha_i) \\ &= p(z_i|n_i, \theta)\, p(\theta)\, \frac{p(\alpha_i)}{\mathcal{P}(\theta, n_i, \alpha_i)} \\ &\propto p(\theta|z_i, n_i)\, \frac{p(\alpha_i)}{\mathcal{P}(\theta, n_i, \alpha_i)}. \end{aligned} \tag{B.36}$$

Therefore, the marginal posterior for θ can be expressed as

$$\begin{aligned} p(\theta|z_i = 4, n_i = 4, \mathcal{S}_i) &\propto p(\theta|z_i = 4, n_i = 4) \sum_{\alpha_i \in A} \frac{p(\alpha_i)}{\mathcal{P}(\theta, n_i = 4, \alpha_i)} \\ &= 5\,\theta^4 \times \frac{1}{2}\left(\frac{1}{(1-\theta)^4 + \theta^4} + 1\right). \end{aligned} \tag{B.37}$$

An even simpler example to illustrate an alternative approach. Let θ denote the probability of success for a Bernoulli trial (e.g., flipping a coin and getting heads). Consider a single trial (coin flip), $n_i = 1$. Then $\mathcal{Y}^{n_i} = \{0, 1\}$ and $\mathcal{Z}_i = \{0, 1\}$. Let the critical region be $\Omega_i = \{1\}$. The power function is

$$\mathcal{P}_{\Omega_i}(\theta) = \theta. \tag{B.38}$$

Let the null hypothesis be $H_0 : \theta = 0.05$ and let the alternative be $H_1 : \theta > .05$. If $k_i = 1$ then the result is (just) significant at the 5% level and the result is published. What can

be inferred about the probability of success from the published result? Given the restricted sampling space, the sampling distribution is

$$p(k_i = 1|\theta, \Omega_i) = \frac{\mathsf{Bernoulli}(k_i = 1|n_i = 1, \theta)}{\mathcal{P}_{\Omega_i}(\theta)} = 1. \tag{B.39}$$

Reinterpreting the sampling distribution as the likelihood for θ, we see that the likelihood contains no information about θ and nothing can be learned (beyond learning that heads is indeed possible; i.e., $\theta > 0$).

Suppose 100 such studies (based on the same coin) are published. What can we infer about the probability of heads? Because each study has no information about θ, all studies together have no information.

Alternative approach. An alternative approach, in the spirit of Rosenthal (1979), is to consider the studies that remain in the file drawers. Suppose there are N_0 studies in the file drawer where $k_i = 0$ and N_1 published studies where $k_i = 1$. Conditional on N_1 and θ, the sampling distribution for N_0 is

$$p(N_0|N_1, \theta) = \mathsf{Neg\text{-}Binomial}(N_0|N_1, \theta) = \binom{N_0 + N_1 - 1}{N_1 - 1} \theta^{N_1} (1 - \theta)^{N_0}. \tag{B.40}$$

The mean of this distribution is $E[N_0|N_1, \theta] = N_1 (1 - \theta)/\theta$. Under the null hypothesis $\theta = 1/20$ we would expect $19\, N_1$ studies in file drawers. For example if $N_1 = 100$ then we expect 1900 studies in file draws. If we assume there are actually far fewer such studies in file drawers then we can reject the null hypothesis. This appears to involve prior information regarding N_0.

Given such prior information, it is more direct use it to compute the posterior distribution for θ. Note that

$$p(N_0, N_1, \theta) = p(N_1|N_0, \theta)\, p(N_0, \theta), \tag{B.41}$$

where

$$p(N_1|N_0, \theta) = \mathsf{Neg\text{-}Binomial}(N_1|N_0, 1 - \theta) = \binom{N_0 + N_1 - 1}{N_0 - 1} \theta^{N_1} (1 - \theta)^{N_0}. \tag{B.42}$$

Suppose $p(N_0, \theta) = p(N_0)\, p(\theta)$. Then

$$p(\theta|N_1) = \frac{p(N_1|\theta)\, p(\theta)}{\int_0^1 p(N_1|\theta)\, p(\theta)\, d\theta}, \tag{B.43}$$

where

$$p(N_1|\theta) = \sum_{N_0=0}^{\infty} p(N_1|N_0, \theta)\, p(N_0) = \theta^{N_1} \sum_{N_0=0}^{\infty} \binom{N_0 + N_1 - 1}{N_0 - 1} (1 - \theta)^{N_0}\, p(N_0). \tag{B.44}$$

For example, suppose

$$p(N_0) = \mathsf{Neg\text{-}Binomial}(N_0|500/199, 1/200), \tag{B.45}$$

so that N_0 has a prior mean of 500 and variance of 10^5. Also suppose $p(\theta) = 1$. The posterior distribution for θ is shown in Figure 18.

Appendix C. A story

The section was the original introductory illustration. Currently it provides a story for the example.

Consider an *effect* that has been established and confirmed by a twenty published studies. Nineteen of the published studies have statistically significant positive results; the remaining study has a statistically significant negative result. The results of the latter study are viewed as having been produced by bad luck. The experiments range in size from seven to fifty subjects/measurements. See Figure 8 for the distribution of the published results.

A graduate student who is familiar with the published results conducts his own experiment but fails to get a significant result. He supposes that chance and/or lack of competence has played a role. He decides to investigate further and finds a meta-analysis combining all twenty published results. See Figure 19 for the posterior distribution of the effect computed from the meta-analysis. The probability that the effect is less than 0.2 is quite small, about 2×10^{-5}.

After some additional study, the student concludes the meta-analysis was incorrectly done because it did not account for the selection process by which only studies with statistically significant results are published. One tell-tale feature of the selection process (which is visible in Figure 8) is that smaller experiments tend to show larger measured effects. He decides to do his own meta-analysis.

The back story. The published studies are consistent with a true effect equal to 0.1 and a standard deviation for each observation equal to 1. Two hundred experiments were conducted of which twenty produced statistically significant results, where significance is judged according to a two-tailed test with a significance of 5%. The rate of rejection of the null hypothesis is 10%, which is in line with the true rejection rate of about 8% (since the true effect does not equal zero). See Figure 9. The 180 studies with insignificant results were not published and the results of those experiments constitute missing datasets.

See Figure 13 for a comparison of the posterior distribution that properly takes the selection process into account with the distribution that does not. The evidence in favor of the adjusted distribution (relative to the unadjusted distribution) is overwhelming. The publication selection process has a large effect on results in this example because the power of the hypothesis tests involved is low. (Power is the probability of rejecting the null hypothesis when it is false.) The sample sizes are too small given the size of the true effect relative to the noise in the experiments. The adjustment required to properly account for the selection process amounts to dividing by the power for each experiment. Because the power reaches its minimum when the effect is zero, dividing by the power shrinks the distribution for the effect toward zero.

Appendix D. Bayesian inference absent the selection process

In this section I analyze the data assuming there is no selection process.

The likelihood for (μ, τ) is given by

$$p(y_i|\mu,\tau) = \prod_{j=1}^{n_i} \mathsf{N}(y_{ij}|\mu,\tau^2), \tag{D.1}$$

where y_i is fixed. The likelihood is proportional to the sampling distribution shown in (4.1):

$$p(y_i|\mu,\tau) \propto p(\widehat{\mu}_i, \widehat{\sigma}_i|\mu,\tau,n_i). \tag{D.2}$$

This demonstrates that $(n_i, \widehat{\mu}_i, \widehat{\sigma}_i)$ is a sufficient statistic for the data. The posterior distribution for (μ, τ) can be expressed as

$$p(\mu,\tau|y_i,\mathcal{I}) \propto p(\widehat{\mu}_i, \widehat{\sigma}_i|\mu,\tau,n_i)\, p(\mu,\tau), \tag{D.3}$$

where $\mathcal{I}$ denotes the prior information.

Jeffreys prior. If we adopt the Jeffreys prior,[11]

$$p(\mu,\tau) \propto 1/\tau, \tag{D.4}$$

then the marginal posterior for μ is[12]

$$p(\mu|y_i,\mathcal{J}) = \int p(\mu,\tau|y_i,\mathcal{J})\, d\tau = \mathsf{Student\text{-}t}(\mu|\widehat{\mu}_i, \widehat{\sigma}_i^2, n_i - 1), \tag{D.5}$$

where $\mathcal{J}$ denotes the Jeffreys prior. The location and scale parameters of the posterior distribution for μ are $\widehat{\mu}_i$ and $\widehat{\sigma}_i$. In addition, if $n_i \geq 3$, then the mean is $\widehat{\mu}_i$ and if $n_i \geq 4$ then the standard deviation is

$$\sqrt{\frac{n_i - 1}{n_i - 3}}\, \widehat{\sigma}_i. \tag{D.6}$$

We can express the posterior in (D.5) as

$$\frac{\mu - \widehat{\mu}_i}{\widehat{\sigma}_i} \,\Big|\, y_i, \mathcal{J} \sim \mathsf{Student\text{-}t}(0, 1, n_i - 1). \tag{D.7}$$

Note the similarity in form between (D.7) and the sampling distribution

$$\frac{\widehat{\mu}_i - \mu}{\widehat{\sigma}_i} \,\Big|\, \mu, n_i \sim \mathsf{Student\text{-}t}(0, 1, n_i - 1), \tag{D.8}$$

which can be derived from (4.1). Also note the conceptual distinctions: In (D.8) we are conditioning on the unknown true effect (and the sample size), while in (D.7) we are conditioning on the observed data.

---

[11] This prior produces the same posterior for (μ, τ) as does the prior $p(\mu, \tau^2) \propto 1/\tau^2$. In the first case $p(\mu, \tau|y,\mathcal{J}) \propto p(y_i|\mu,\tau)/\tau$ while in the second case $p(\mu, \tau^2|y_i,\mathcal{J}) \propto p(y_i|\mu,\tau^2)/\tau^2$. By the the change of variables formula, $p(\mu,\tau|y_i,\mathcal{J}) = 2\,\tau\, p(\mu,\tau^2|y_i,\mathcal{J})$. I will adopt whichever is more convenient to the task at hand.

[12] The marginal posterior for τ is characterized by $\tau^2 \sim \mathsf{Inv\text{-}Gamma}\big((n_i-1)/2, \widehat{\sigma}_i^2\, n_i\, (n_i-1)/2\big)$.

Meta-analysis. It is straightforward to combine the results from difference studies/experiments into the results from one large study/experiment. The meta-analysis merges the results of the individual experiments into a single large experiment.

Let $y = (y_1, \ldots, y_N)$. A sufficient statistic for each study is $(n_i, \widehat{\mu}_i, \widehat{\sigma}_i)$. Let $\mathcal{M}$ denote the meta-analysis. The combined likelihood for (μ, τ) is

$$p(y|\mu, \tau, \mathcal{M}) = \prod_{i=1}^{N} p(y_i|\mu, \tau) \propto \mathsf{N}\Big(\overline{\mu}\,\Big|\,\mu, \frac{\tau^2}{n}\Big)\,\mathsf{Nakagami}\Big(\overline{\sigma}\,\Big|\,\frac{n-1}{2}, \frac{\tau^2}{n}\Big), \tag{D.9}$$

where $(n, \overline{\mu}, \overline{\sigma})$ is a sufficient statistic for y, and where $n = \sum_{i=1}^{N} n_i$ and

$$\overline{\mu} = \sum_{i=1}^{N} \Big(\frac{n_i}{n}\Big)\,\widehat{\mu}_i \tag{D.10}$$

$$\overline{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{N} \Big(\frac{n_i}{n}\Big)\,\Big((n_i - 1)\,\widehat{\sigma}_i^2 + (\widehat{\mu}_i - \overline{\mu})^2\Big). \tag{D.11}$$

The posterior distribution can be expressed as

$$p(\mu, \tau|y, \mathcal{M}, \mathcal{I}) \propto p(y|\mu, \tau, \mathcal{M})\,p(\mu, \tau). \tag{D.12}$$

Given the Jeffreys prior, the marginal posterior for μ is

$$p(\mu|y, \mathcal{M}, \mathcal{J}) = \mathsf{Student}(\mu|\overline{\mu}, \overline{\sigma}^2, n-1). \tag{D.13}$$

The associated p-value is $\overline{\pi} = f_{\overline{n}-1}(|\overline{t}|)$, where $\overline{t} = \overline{\mu}/\overline{\sigma}$.

This is the meta-analysis that is (incorrectly) applied to the published data as discussed in Section C [see Figure 19]. The application of this meta-analysis to the published data is inappropriate because it ignores the effects of the selection process on the likelihood.

Appendix E. Expected power [Incomplete]

When designing an experiment, one may wish to consider the power one can expect.

Let $p(\lambda|\mathcal{I})$ denote the prior for λ where $\mathcal{I}$ denotes the prior information. The prior expected power is given by

$$\mathcal{P}(n_i, \alpha|\mathcal{I}) = \int \mathcal{P}(\lambda, n_i, \alpha)\,p(\lambda|\mathcal{I})\,d\lambda. \tag{E.1}$$

Given a prior for (μ, τ^2) of the following form,[13]

$$p(\mu, \tau^2|\mathcal{I}) = \mathsf{N}(\mu|m, \tau^2/\kappa)\,p(\tau^2|\mathcal{I}), \tag{E.2}$$

the prior distribution for (λ, τ^2) is

$$p(\lambda, \tau^2|\mathcal{I}) = \mathsf{N}(\lambda|m/\tau, 1/\kappa)\,p(\tau^2|\mathcal{I}). \tag{E.3}$$

[13]The informed editor's prior for (μ, τ^2) has the form of (E.2) [see (**??**)]. The general expression for $p(\lambda|\mathcal{E})$ is complicated [see (A.20)].

If $m = 0$ then $p(\lambda|\mathcal{I}) = \mathsf{N}(\lambda|0, 1/\kappa)$ (where κ is the precision) and λ and τ are independent. In this case, the prior expected power has these features:

$$\lim_{\kappa \to 0} \; \mathcal{P}(n_i, \alpha|\mathcal{I}) = 1 \tag{E.4}$$

$$\lim_{\kappa \to \infty} \; \mathcal{P}(n_i, \alpha|\mathcal{I}) = \alpha. \tag{E.5}$$

Figure 20 illustrates how $\mathcal{P}(n_i, \alpha|\mathcal{I})$ increases with sample size, given $m = 0$, $\alpha = 5\%$, and various values of κ.

Given the results of an experiment, the posterior expected power is

$$\mathcal{P}(n_i, \alpha|y_i, \mathcal{I}) = \int \mathcal{P}(\lambda, n_i, \alpha)\, p(\lambda|y_i, \mathcal{I})\, d\lambda. \tag{E.6}$$

The informed editor's posterior distribution for (μ, τ^2) has the same form as the prior, and consequently the posterior distribution for λ can be computed from (A.20).

Appendix F. Linear regression [Incomplete]

We now examine the case of linear regression with normally-distributed errors. Suppose

$$y_i = X_i\, \beta_i + \varepsilon_i, \tag{F.1}$$

where $\varepsilon_i \sim \mathsf{N}(0, \tau_i^2\, I_{n_i})$. Then

$$p(y_i|\beta_i, \tau_i) = \mathsf{N}(y_i|X_i\, \beta_i, \tau_i^2\, I_{n_i}) = p(\widehat{\beta}_i, s_i|\beta_i, \tau_i, X_i^\top X_i, n_i)\, h(y_i, X_i), \tag{F.2}$$

where

$$\widehat{\beta}_i = (X_i^\top X_i)^{-1} X_i^\top y_i \tag{F.3}$$

and

$$s_i^2 = (y_i - X_i\, \widehat{\beta}_i)^\top (y_i - X_i\, \widehat{\beta}_i). \tag{F.4}$$

References

Bayarri, M. J. and M. H. DeGroot (1987). Bayesian analysis of selection models. *The Statistician 36*, 137–146.

Bayarri, M. J. and M. H. DeGroot (1991). The analysis of published significant results. Technical Report 91-21, Department of Statistics, Perdue University.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience 14*, 356–376.

Gelman, A. and J. Carlin (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science 9*(6), 641–651.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian data analysis* (Third ed.). CRC Press.

Gelman, A. and E. Loken (2014). The statistical crisis in science. *American Scientist 106*, 460–465.

Gelman, A. and F. Tuerlinckx (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics 15*, 373–390.

Iyengar, S. and J. B. Greenhouse (1988). Selection models and the file drawer problem. *Statistical Science 3*(1), 109–117.

Rosenthal, R. (1979). The "File Drawer Problem" and tolerance for null results. *Psychological Bulletin 86*(3), 638–641.

Federal Reserve Bank of Atlanta, Research Department, 1000 Peachtree Street N.E., Atlanta, GA 30309–4470

E-mail address: mark.fisher@atl.frb.org

URL: http://www.markfisher.net

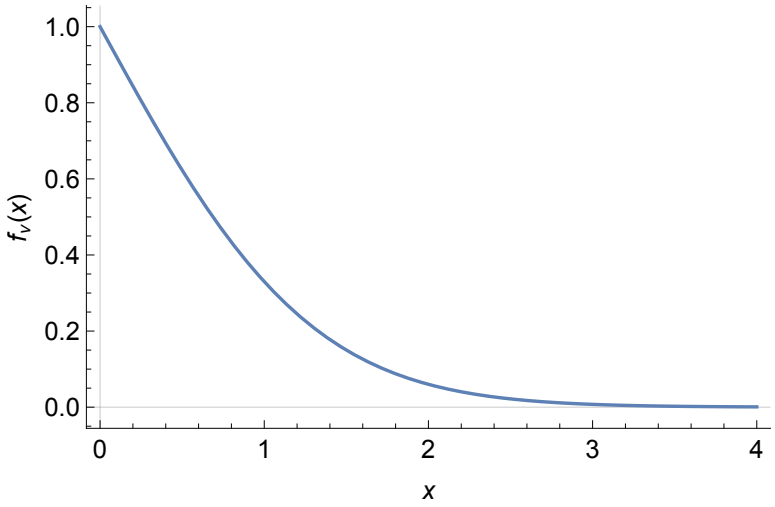

FIGURE 1. Plot of $f_\nu(x)$ given $\nu = 19$.

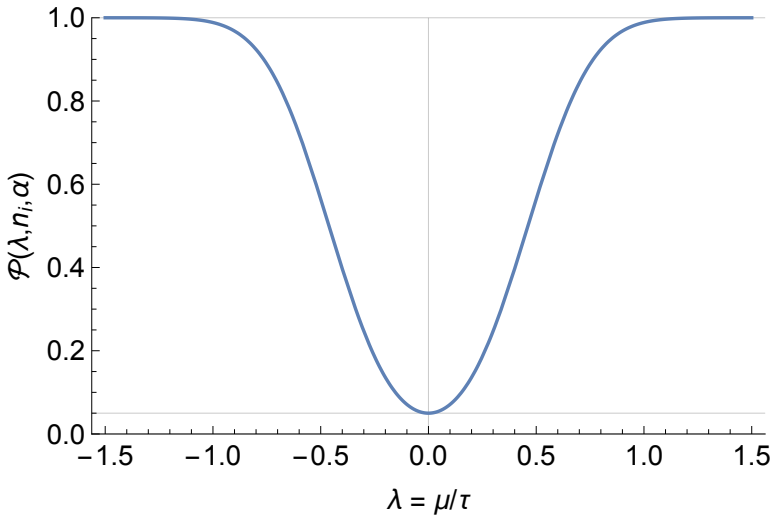

FIGURE 2. Plot of the power $\mathcal{P}(\lambda, n, \alpha)$ given $n = 20$ and $\alpha = 0.05$.

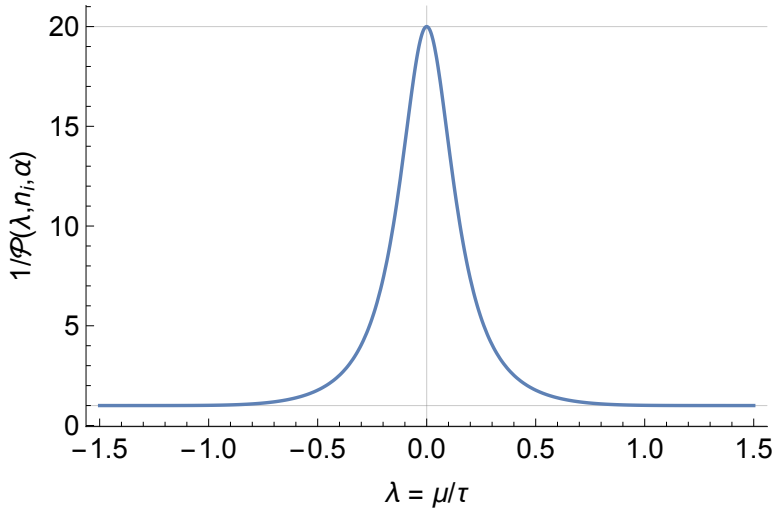

FIGURE 3. Plot of $1/\mathcal{P}(\lambda, n, \alpha)$ given $n = 20$ and $\alpha = 0.05$.

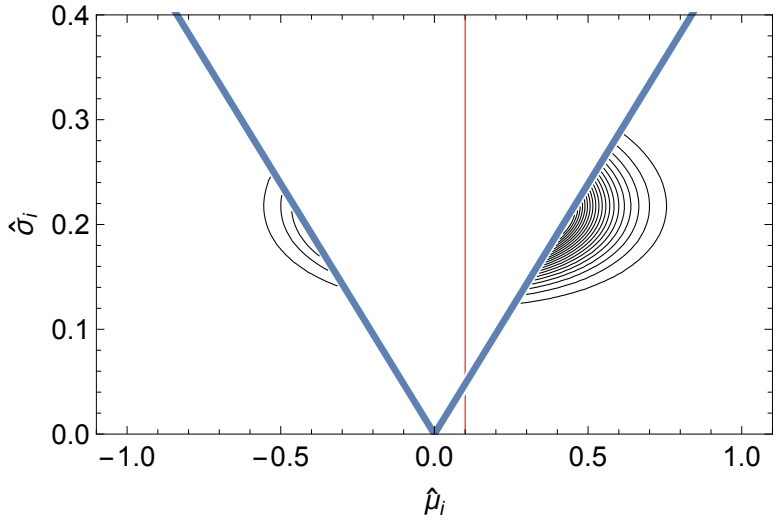

FIGURE 4. Contours of $p(\widehat{\mu}, \widehat{\sigma}|\mu, \tau^2, n, \alpha, \mathbb{S})$, the sampling distribution for $(\widehat{\sigma}, \widehat{\mu})$ subject to the selection process. The values for the parameters in this illustration are $\mu = 0.1$, $\tau = 1$, $n = 20$, and $\alpha = 5\%$. The line $\widehat{\sigma} = |\widehat{\mu}|/c_\alpha$ is shown. The probability below the line is given by the power $\mathcal{P}(\mu/\tau, n, \alpha) = 0.07$.

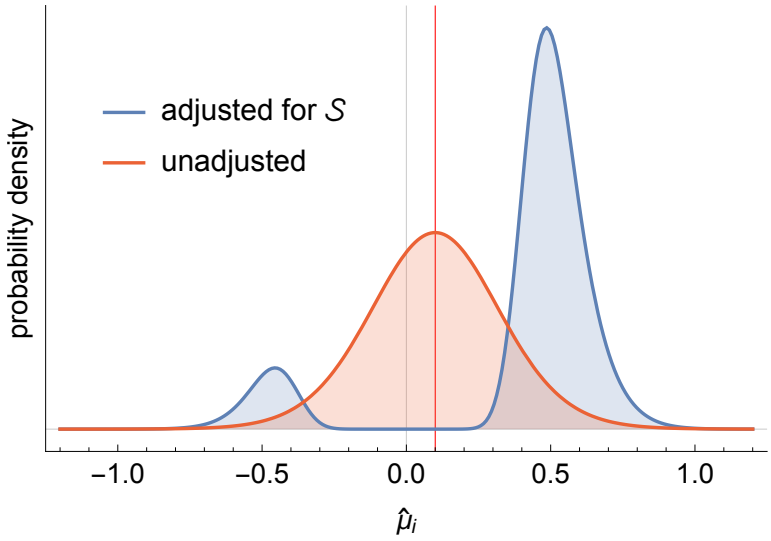

FIGURE 5. Plot of $p(\widehat{\mu}|\mu, \tau^2, n, \alpha, \mathcal{S})$, the marginal sampling distribution for $\widehat{\mu}$ given $\mathcal{S}_\alpha$ (in blue) computed from the joint distribution shown in Figure 4, where $\mu = 0.1$, $\tau = 1$, $n = 20$, and $\alpha = 5\%$. The average value of $|\widehat{\mu}|$ is 0.51, and the probability that $\widehat{\mu} < 0$ is about 12%. The unadjusted distribution is shown for comparison.

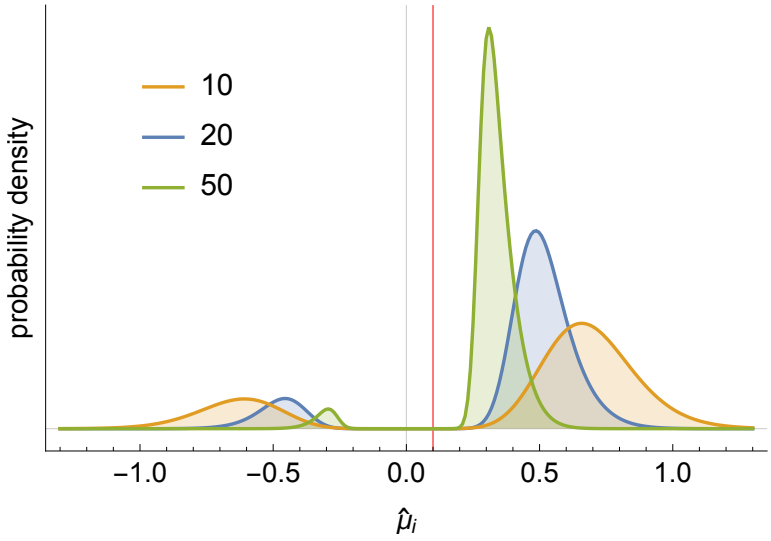

FIGURE 6. Plots of $p(\widehat{\mu}|\mu, \tau^2, n, \alpha, \mathcal{S})$, the sampling distribution for $\widehat{\mu}$ given the selection process (assuming $\mu = 0.1$, $\tau = 1$, and $\alpha = 5\%$) for $n \in \{10, 20, 50\}$. The blue curve is the same as shown in Figure 5. Smaller studies tend to have larger measured effects in absolute value (reflecting Type M errors) and larger probability of incorrect sign (reflecting Type S errors).

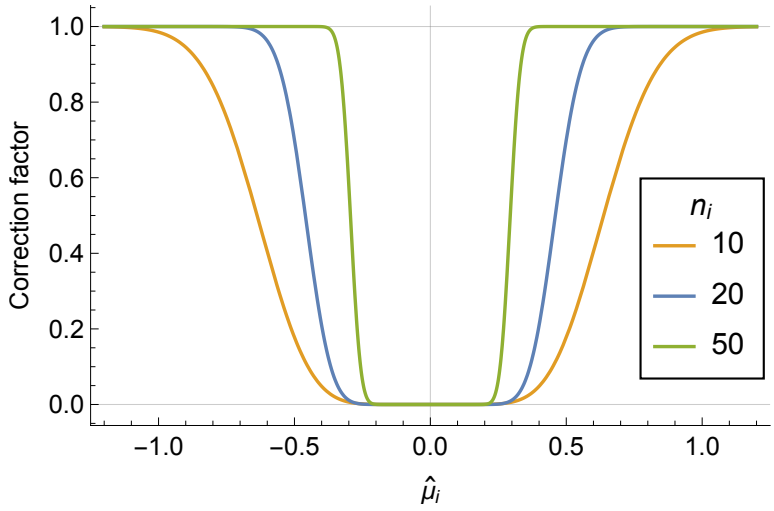

FIGURE 7. Plot of the correction factor $\mathcal{C}(\widehat{\mu}, \tau^2, n, \alpha) = \int_0^{|\widehat{\mu}|/c_\alpha} p(\widehat{\sigma}|\tau^2, n)\, d\widehat{\sigma}$ for $n \in \{10, 20, 50\}$, assuming $\mu = 0.1$, $\tau = 1$, and $\alpha = 5\%$.

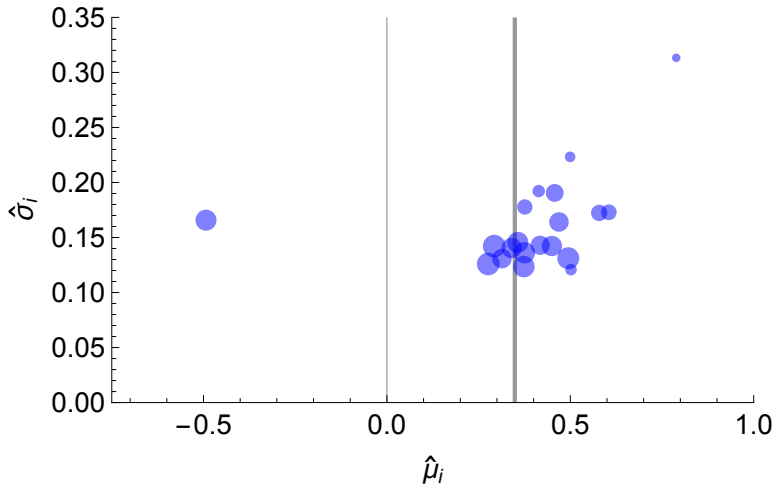

FIGURE 8. Plot of $(\widehat{\mu}_i, \widehat{\sigma}_i)$ for 20 published studies, where $\widehat{\mu}_i$ is the measured effect and $\widehat{\sigma}_i$ is a measure of the uncertainty regarding $\widehat{\mu}_i$. The sample size is proportional to the area of the dot. The weighted mean from all the studies equals 0.349.

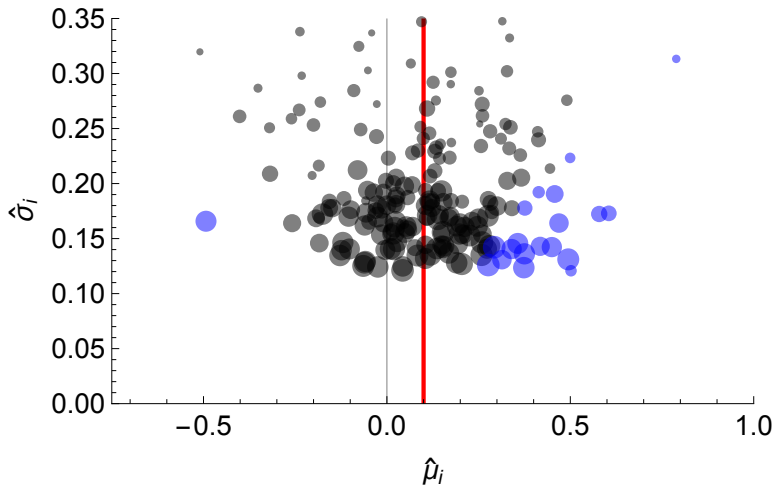

FIGURE 9. Plot of $(\widehat{\mu}_i, \widehat{\sigma}_i)$ for all 200 experiments, where $\widehat{\mu}_i$ is the measured effect and $\widehat{\sigma}_i$ is a measure of the uncertainty regarding $\widehat{\mu}_i$. The sample size is indicated by the area of the dot. The weighted mean from all the studies equals 0.109, which is in line with the true effect.

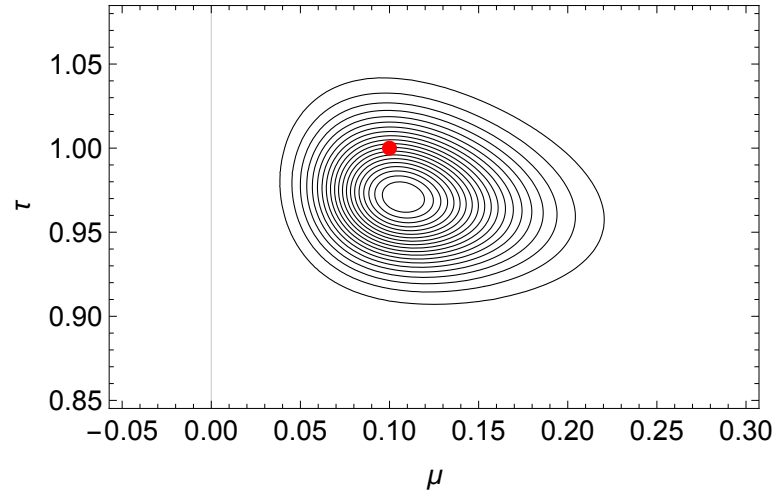

FIGURE 10. Contours of the joint posterior distribution $p(\mu, \tau | y, \mathcal{S}_\alpha, \mathcal{M}, \mathcal{J})$ from the meta-analysis $\mathcal{M}$ taking into account the selection process $\mathcal{S}_\alpha$ and using the Jeffreys prior $\mathcal{J}$, given $\alpha = 5\%$.

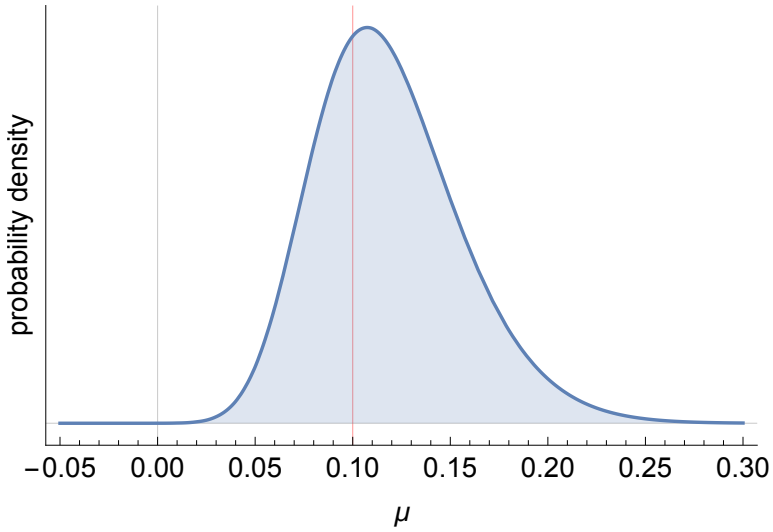

FIGURE 11. The marginal posterior distribution $p(\mu|y, \mathcal{S}_\alpha, \mathcal{M}, \mathcal{J})$ from the meta-analysis $\mathcal{M}$ taking into account the selection process $\mathcal{S}_\alpha$ and using the Jeffreys prior $\mathcal{J}$, given $\alpha = 5\%$.

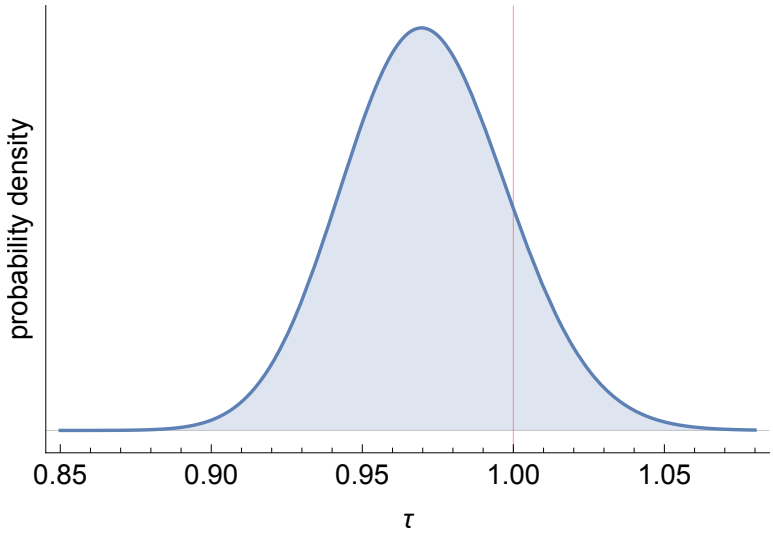

FIGURE 12. The marginal posterior distribution $p(\tau|y, \mathcal{S}_\alpha, \mathcal{M}, \mathcal{J})$ from the meta-analysis $\mathcal{M}$ taking into account the selection process $\mathcal{S}_\alpha$ and using the Jeffreys prior $\mathcal{J}$, given $\alpha = 5\%$.

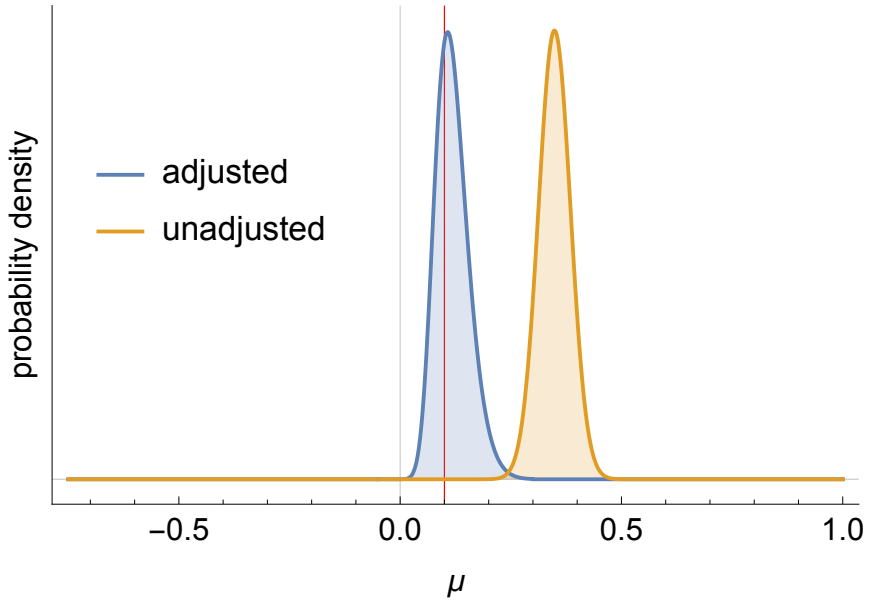

FIGURE 13. Posterior distributions $p(\mu|y, \alpha, \mathcal{S}, \mathcal{M}, \mathcal{J})$ and $p(\mu|y, \mathcal{M}, \mathcal{J})$.

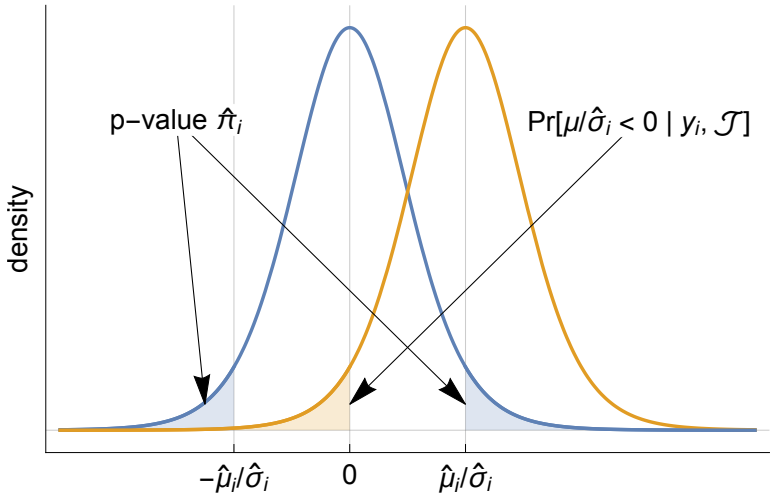

FIGURE 14. Given $\widehat{\mu} > 0$, the p-value $\widehat{\pi}$ equals twice the posterior probability $\Pr[\mu/\widehat{\sigma} < 0|y_{\mathcal{J}}]$.

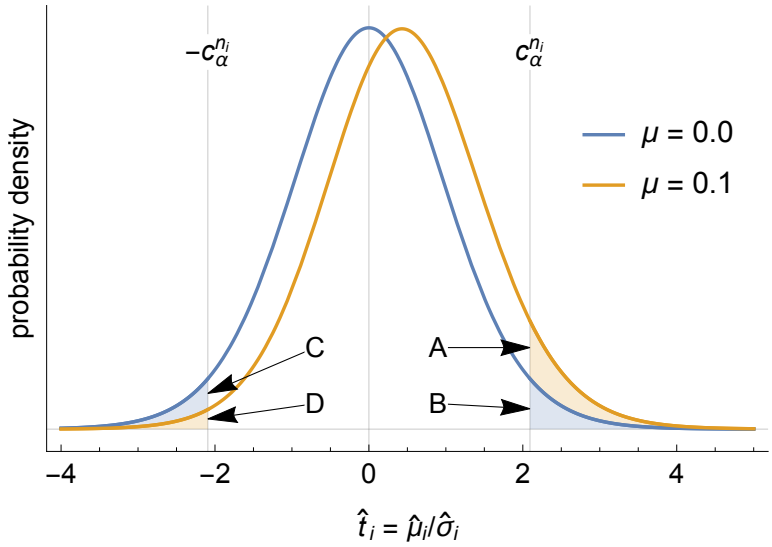

FIGURE 15. Plots of $p(\widehat{t_i}|\mu, \tau, n_i)$ for two values of μ, given $\tau = 1$ and $n_i = 20$. The critical values are shown given $\alpha = 0.05$. Note that $\mathcal{P}(0, n_i, \alpha) = B + C + D = 0.05$ and $\mathcal{P}(0.1, n_i, \alpha) = A + B + D = 0.07$.

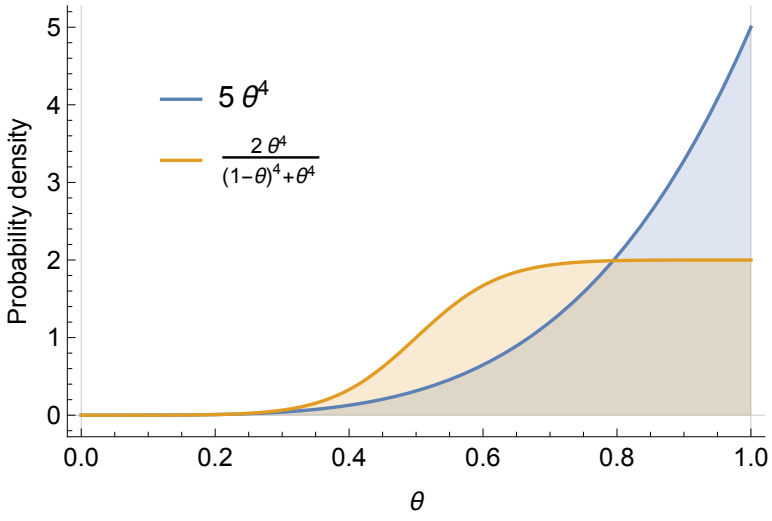

FIGURE 16. Posterior distributions $p(\theta|n_i = 4, z_i = 4)$ and $p(\theta|n_i = 4, z_i = 4, \mathcal{S}_i)$.

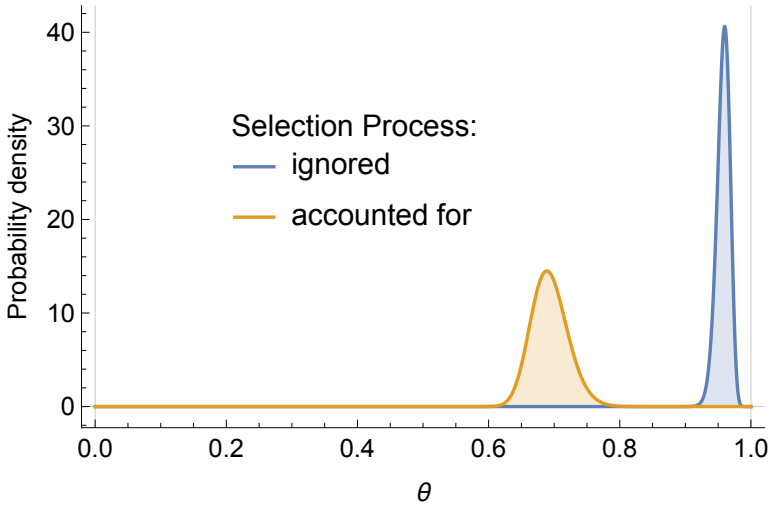

FIGURE 17. Posterior distributions from two meta-studies of the same 100 studies.

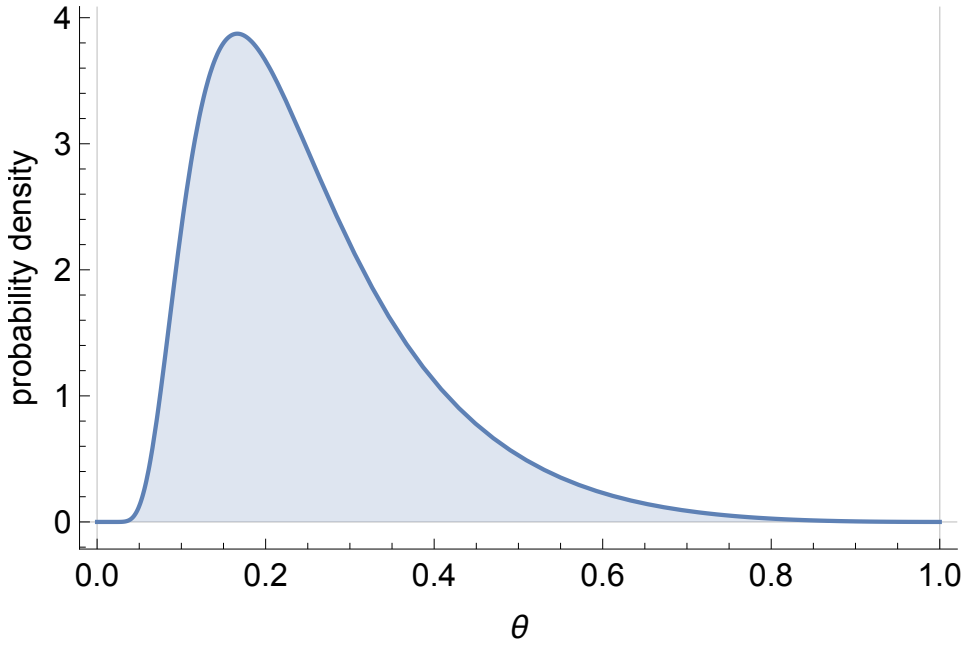

FIGURE 18. Posterior distribution.

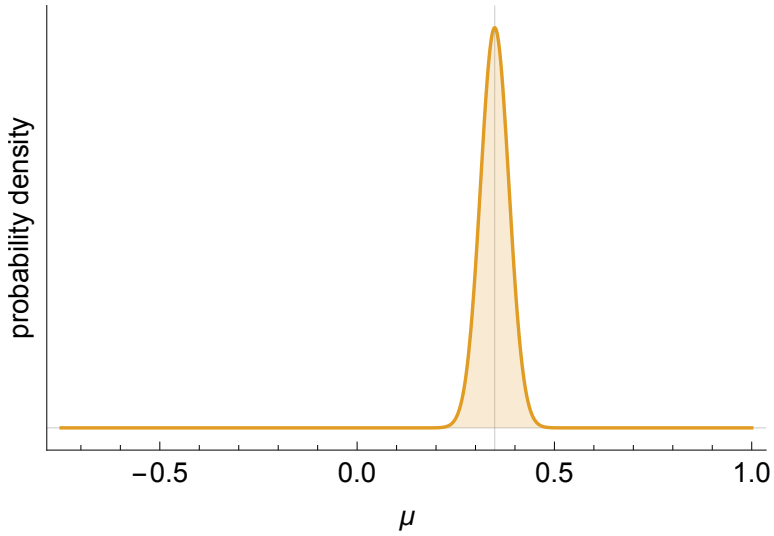

FIGURE 19. Meta-analysis based on all 20 published studies. The posterior distribution for μ is Student t with a mean of 0.349, a standard deviation of 0.036, and 661 degrees of freedom.

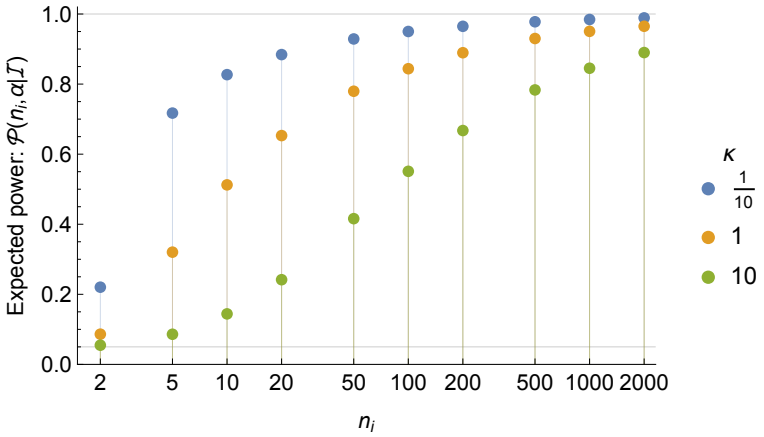

FIGURE 20. Expected power $\mathcal{P}(n_i, \alpha|\mathcal{I})$ as a function of n_i, given $\alpha = 5\%$ and $p(\lambda|\mathcal{I}) = \mathsf{N}(\lambda|0, 1/\kappa)$, where $\kappa = \frac{1}{10}, 1, 10$.