Bayesian approaches to designing replication studies
Supplementary materials

Samuel Pawel[⋆]**, Guido Consonni**[†]**, and Leonhard Held**[⋆]
⋆ Department of Biostatistics, University of Zurich
† Dipartimento di Scienze Statistiche, Universita Cattolica del Sacro Cuore
E-mail: samuel.pawel@uzh.ch

October 6, 2022

In this document we provide additional information on methods for analyzing replication data. For each method we derive the success region which are needed for sample size determination of the replication study.

1 The two-trials rule

Assume the data model $\hat{\theta}_i \,|\, \theta \sim \mathrm{N}(\theta, \sigma_i^2)$ where $\hat{\theta}_i$ is an estimate of the effect size θ from study i and σ_i is the corresponding standard error (assumed to be know). The p-value for testing $H_0 \colon \theta = 0$ versus $H_1 \colon \theta > 0$ is then given by $p_i = 1 - \Phi(\hat{\theta}_i/\sigma_i)$. Suppose the original effect estimate was positive ($\hat{\theta}_o > 0$) and statistically significant at some level α, i. e., $p_o \leq \alpha$. Replication success at level α is then established if also the replication effect estimate $\hat{\theta}_r$ is positive and statistically significant at level α, i. e., $p_r \leq \alpha$. This implies that replication success is achieved if the replication effect estimate $\hat{\theta}_r$ is contained in the success region

$$S_{\text{2TR}} = [z_\alpha \, \sigma_r, \infty) \, .$$

Conversely, if the original effect estimate was negative ($\hat{\theta}_o < 0$), the one-sided p-values are computed for the lower tail of the null distribution and the success region is therefore given by

$$S_{\text{2TR}} = [-\infty, -z_\alpha \, \sigma_r) \, .$$

1.1 The multisite two-trials rule

If multiple replication studies of the same original study are conducted, the two-trials rule is usually modified in a way that the replication effect estimates are first synthesized using either fixed or random effects meta-analysis (see e. g., the "Many labs" projects from Klein et al., 2014, 2018). That is, a weighted average $\hat{\theta}_{r*} = \{\sum_{i=1}^m \hat{\theta}_{ri}/(\sigma_{ri}^2 + \tau_r^2)\} \, \sigma_{r*}^2$ with standard error $\sigma_{r*} = 1/\sqrt{\{\sum_i^m 1/(\sigma_{ri}^2 + \tau_r^2)\}}$ is computed from the m replication effect estimates $\hat{\theta}_{ri}$ and standard errors σ_{ri}. The between replication heterogeneity variance τ_r^2 can be either be set to zero (fixed effects) or estimated from the data. Assuming again that the Replication success at level α is then established if the replication p-value is smaller than α, i. e., $p_{r*} = 1 - \Phi(\hat{\theta}_{r*}/\sigma_{r*}) \leq \alpha$. This implies a success region for the weighted average replication effect estimate $\hat{\theta}_{r*}$ given by

$$S_{\text{2TR}} = [z_\alpha \, \sigma_{r*}, \infty) \, .$$

for positive original effect estimates ($\hat{\theta}_o > 0$) and

$$S_{\text{2TR}} = [-\infty, -z_\alpha \, \sigma_{r*})$$

for negative original effect estimates ($\hat{\theta}_o < 0$).

2 Fixed effects meta-analysis

Assume again the data model $\hat{\theta}_i \,|\, \theta \sim \text{N}(\theta, \sigma_i^2)$ where $\hat{\theta}_i$ is an estimate of the effect size θ from study $i \in \{o, r\}$ and σ_i is the corresponding standard error (assumed to be know). In the fixed effects meta-analysis approach replicability is assessed in terms of the pooled effect estimate $\hat{\theta}_m$ and standard error σ_m which are

$$\hat{\theta}_m = \left(\hat{\theta}_o/\sigma_o^2 + \hat{\theta}_r/\sigma_r^2\right) \sigma_m^2 \qquad \text{and} \qquad \sigma_m = \left(1/\sigma_o^2 + 1/\sigma_r^2\right)^{-1/2},$$

which are also equivalent to the mean and standard deviation of a posterior distribution for the effect size θ based on the data from original and replication study and an initial flat prior for θ. Replication success at level α is established if the one-sided meta-analytic p-value (in the direction of the original effect estimate $\hat{\theta}$) is significant at level α, i. e., $p_m = 1 - \Phi(\hat{\theta}_m/\sigma_m) \leq \alpha$ (assuming $\hat{\theta}_o > 0$). This criterion implies a success region S_{MA} for the replication effect estimate $\hat{\theta}_r$ given by

$$S_{\text{MA}} = \left[\sigma_r z_\alpha \sqrt{1 + \sigma_r^2/\sigma_o^2} - (\hat{\theta}_o \sigma_r^2)/\sigma_o^2, \infty\right),$$

respectively

$$S_{\text{MA}} = \left(-\infty, -\sigma_r z_\alpha \sqrt{1 + \sigma_r^2/\sigma_o^2} - (\hat{\theta}_o \sigma_r^2)/\sigma_o^2\right]$$

for negative original effect estimates ($\hat{\theta}_o < 0$).

3 Effect size equivalence

Assume the data model $\hat{\theta}_i \,|\, \theta_i \sim \text{N}(\theta_i, \sigma_i^2)$ for study $i \in \{o, r\}$. A $(1 - 2\alpha)$ confidence interval for the effect size difference $\delta = \theta_r - \theta_o$ is then given by

$$\hat{\theta}_r - \hat{\theta}_o \pm z_\alpha \sqrt{\sigma_r^2 + \sigma_r^2}$$

4 The *Q*-test

5 The replication Bayes factor

5.1 The multisite replication Bayes factor

The data model for the replication effect estimate vector $\hat{\boldsymbol{\theta}}_r = (\hat{\theta}_{r1}, \ldots, \hat{\theta}_{rm})^\top$ with standard error vector $\boldsymbol{\sigma}_r = (\sigma_{r1}, \ldots, \sigma_{rm})^\top$ is then $\hat{\boldsymbol{\theta}}_r \,|\, \theta \sim \text{N}_m\{\theta \, \mathbf{1}_m, \text{diag}(\boldsymbol{\sigma}^2 + \tau_r^2 \, \mathbf{1}_m\}$ where τ_r^2 is the heterogeneity variance for the replication effect sizes and $\mathbf{1}_m$ is a vector of m ones.

To determine the multisite version of the replication Bayes factor we need to know the marginal density of the replication effect estimates $\hat{\theta}_r \,|\, \theta \sim \mathrm{N}_n\{\theta \mathbf{1}_m, \mathrm{diag}(\sigma_r^2 + \tau_r^2 \mathbf{1}_m)\}$ under a normal prior $H_k \colon \theta \sim \mathrm{N}(m, v)$. Let $\mathrm{N}(x; m, v)$ denote the normal density function mean m and variance v evaluated at x. Define also $\hat{\theta}_{r*} = \left\{\sum_{i=1}^n \hat{\theta}_{ri}/(\sigma_{ri}^2 + \tau_r^2)\right\} \sigma_{r*}^2$ and $\sigma_{r*}^2 = 1/\left\{\sum_{i=1}^n 1/(\sigma_{ri}^2 + \tau_r^2)\right\}$, i. e., the weighted average of the replication effect estimates and its variance. The marginal density is then given by

$$
\begin{aligned}
f(\hat{\theta}_r \,|\, H_k) &= \int f(\hat{\theta}_r \,|\, \theta) f(\theta \,|\, H_k) \,\mathrm{d}\theta \\
&= \int \frac{\exp\left[-\frac{1}{2}\left\{\sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \theta)^2}{\sigma_{ri}^2 + \tau_r^2} + \frac{(\theta - m)^2}{v}\right\}\right]}{\left\{2\pi v \prod_{i=1}^n 2\pi\left(\sigma_{ri}^2 + \tau_r^2\right)\right\}^{1/2}} \,\mathrm{d}\theta \\
&= \int \frac{\exp\left[-\frac{1}{2}\left\{\sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \hat{\theta}_{r*})^2}{\sigma_{ri}^2 + \tau_r^2} + \frac{(\hat{\theta}_{r*} - \theta)^2}{\sigma_{r*}^2} + \frac{(\theta - m)^2}{v}\right\}\right]}{\left\{2\pi v \prod_{i=1}^n 2\pi\left(\sigma_{ri}^2 + \tau_r^2\right)\right\}^{1/2}} \,\mathrm{d}\theta \\
&= \frac{\exp\left[-\frac{1}{2}\left\{\sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \hat{\theta}_{r*})^2}{\sigma_{ri}^2 + \tau_r^2}\right\}\right]}{\left\{2\pi v \prod_{i=1}^n 2\pi\left(\sigma_{ri}^2 + \tau_r^2\right)\right\}^{1/2}} \underbrace{\int \exp\left[-\frac{1}{2}\left\{\frac{(\hat{\theta}_{r*} - \theta)^2}{\sigma_{r*}^2} + \frac{(\theta - m)^2}{v}\right\}\right] \mathrm{d}\theta}_{=\mathrm{N}(\hat{\theta}_{r*}; m, v + \sigma_{r*}^2) 2\pi\sqrt{v}\sigma_{r*}} \\
&= \left\{(1 + v/\sigma_{r*}^2) \prod_{i=1}^n 2\pi\left(\sigma_{ri}^2 + \tau_r^2\right)\right\}^{-1/2} \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^n \frac{(\hat{\theta}_{ri} - \hat{\theta}_{r*})^2}{\sigma_{ri}^2 + \tau_r^2} + \frac{(\hat{\theta}_{r*} - m)^2}{\sigma_{r*}^2 + v}\right\}\right].
\end{aligned}
$$

The marginal density under the null hypothesis $H_0 \colon \theta = 0$ is then obtained by setting $m = 0$ and $v = 0$, whereas the marginal density under the alternative hypothesis is obtained by setting $m = \hat{\theta}_o$ and $v = \sigma_o^2$. This then leads to the replication Bayes factor

$$
\mathrm{BF}_{01}(\hat{\theta}_r) = \frac{f(\hat{\theta}_r \,|\, H_0)}{f(\hat{\theta}_r \,|\, H_1)} = \sqrt{1 + \sigma_o^2/\sigma_{r*}^2} \exp\left[-\frac{1}{2}\left\{\frac{\hat{\theta}_{r*}^2}{\sigma_{r*}^2} - \frac{(\hat{\theta}_{r*} - \hat{\theta}_o)^2}{\sigma_{r*}^2 + \sigma_o^2}\right\}\right],
$$

which is equivalent to the unisite replication Bayes factor from (**??**) using the weighted average $\hat{\theta}_{r*}$ and its standard error σ_{r*} as the replication effect estimate $\hat{\theta}_r$ and standard error σ_r.

6 The sceptical *p*-value

(Held, 2020)

7 The sceptical Bayes factor

Pawel and Held (2022) showed that the success region based on $\mathrm{BF}_{\mathrm{R}} \leq \gamma$ is given by

$$
S_{\mathrm{BF_S}} = \begin{cases} (-\infty, -\sqrt{B} - M] \bigcup [\sqrt{B} - M, \infty) & \text{for } s_\gamma < 1 \\ [\hat{\theta}_o - \{(\sigma_o^2 + \sigma_r^2) \log \gamma\}/\hat{\theta}_o, \infty) & \text{for } s_\gamma = 1 \\ [-\sqrt{B} - M, \sqrt{B} - M] & \text{for } s_\gamma > 1 \end{cases} \tag{1}
$$

with

$$B = \left\{ \frac{\hat{\theta}_o^2}{\sigma^2(1-s_\gamma)} + 2\log\left(\frac{\sigma_r^2+\sigma_o^2}{\sigma_r^2+s\sigma_o^2}\right) - 2\log\gamma \right\} \frac{(\sigma_r^2+s\sigma_o^2)(\sigma_r^2+\sigma_o^2)}{\sigma_o^2(1-s_\gamma)}$$

$$M = \frac{\hat{\theta}_o(\sigma_r^2+s_\gamma\sigma_o^2)}{\sigma_o^2(1-s_\gamma)}$$

$$s_\gamma = \begin{cases} -\dfrac{z_o^2}{q} - 1 & \text{if } -\dfrac{z_o^2}{q} \geq 1 \\ \text{undefined} & \text{else} \end{cases}$$

$$\text{where } q = W_{-1}\left(-\frac{z_o^2}{\gamma^2}\cdot\exp\left\{-z_o^2\right\}\right)$$

References

Held, L. (2020). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):431–448. doi:10.1111/rssa.12493.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, v., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45:142–152. doi:10.1027/1864-9335/a000178.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490. doi:10.1177/2515245918810225.

Pawel, S. and Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):879–911. doi:10.1111/rssb.12491.