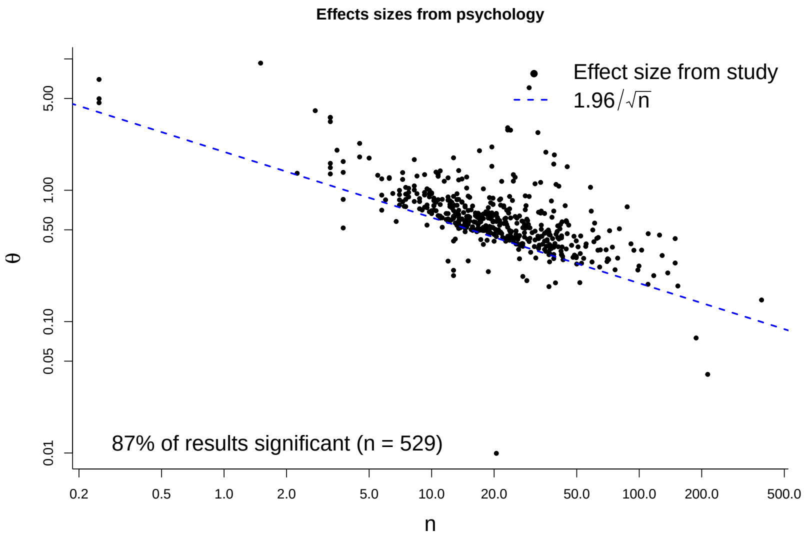
A Bayesian Meta-analysis Method that Corrects for Publication Bias

Jonas Moss

University of Oslo

June 26, 2018
Nordstat 2018

This is What p-hacking Looks Like!



Motyl et al. (2017): The state of social and personality science

Motivation and Setup

What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N(\theta_i, se_i)$

What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N(\theta_i, se_i)$
  - θ_i are normalized – they are *effect sizes*.

What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N(\theta_i, se_i)$
  - θ_i are normalized – they are *effect sizes*.
  - Most common is the *standardized mean difference*, $\frac{\mu_1 - \mu_2}{\sigma}$.

What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N(\theta_i, se_i)$
  - θ_i are normalized – they are *effect sizes*.
  - Most common is the *standardized mean difference*, $\frac{\mu_1 - \mu_2}{\sigma}$.
  - Fixed effects: $\theta_i = \theta$ for all i.

What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N(\theta_i, se_i)$
  - θ_i are normalized – they are *effect sizes*.
  - Most common is the *standardized mean difference*, $\frac{\mu_1 - \mu_2}{\sigma}$.
  - Fixed effects: $\theta_i = \theta$ for all i.
  - Random effects: $\theta_i \sim N(\theta_0, \sigma_0^2)$.

What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N(\theta_i, se_i)$
  - θ_i are normalized – they are *effect sizes*.
  - Most common is the *standardized mean difference*, $\frac{\mu_1 - \mu_2}{\sigma}$.
  - Fixed effects: $\theta_i = \theta$ for all i.
  - Random effects: $\theta_i \sim N(\theta_0, \sigma_0^2)$.
- The effect sizes are usually closely related:

What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N(\theta_i, se_i)$
  - θ_i are normalized – they are *effect sizes*.
  - Most common is the *standardized mean difference*, $\frac{\mu_1 - \mu_2}{\sigma}$.
  - Fixed effects: $\theta_i = \theta$ for all i.
  - Random effects: $\theta_i \sim N(\theta_0, \sigma_0^2)$.
- The effect sizes are usually closely related:
  - Effect of a class of anti-depressiva;

What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N(\theta_i, se_i)$
  - θ_i are normalized – they are *effect sizes*.
  - Most common is the *standardized mean difference*, $\frac{\mu_1 - \mu_2}{\sigma}$.
  - Fixed effects: $\theta_i = \theta$ for all i.
  - Random effects: $\theta_i \sim N(\theta_0, \sigma_0^2)$.
- The effect sizes are usually closely related:
  - Effect of a class of anti-depressiva;
  - effect of some psychological intervention.

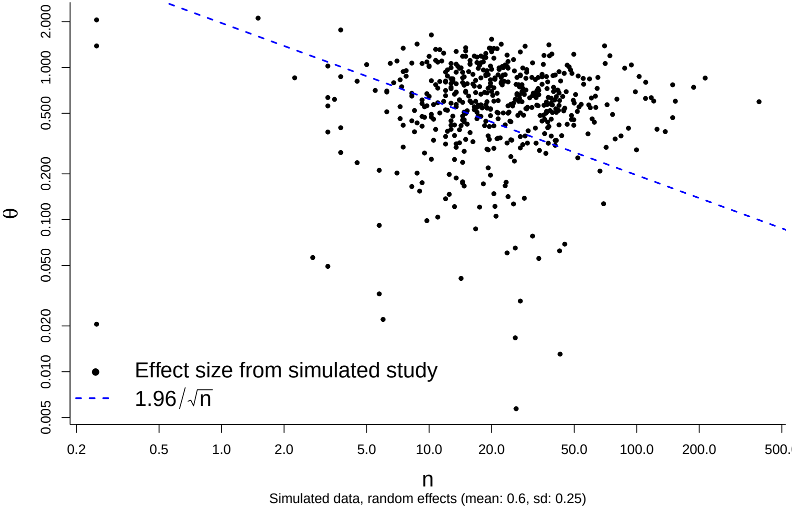
What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N\left(\theta_i, se_i\right)$
  - θ_i are normalized – they are *effect sizes*.
  - Most common is the *standardized mean difference*, $\frac{\mu_1 - \mu_2}{\sigma}$.
  - Fixed effects: $\theta_i = \theta$ for all i.
  - Random effects: $\theta_i \sim N(\theta_0, \sigma_0^2)$.
- The effect sizes are usually closely related:
  - Effect of a class of anti-depressiva;
  - effect of some psychological intervention.
  - ... but they don't have to be *that* closely related!

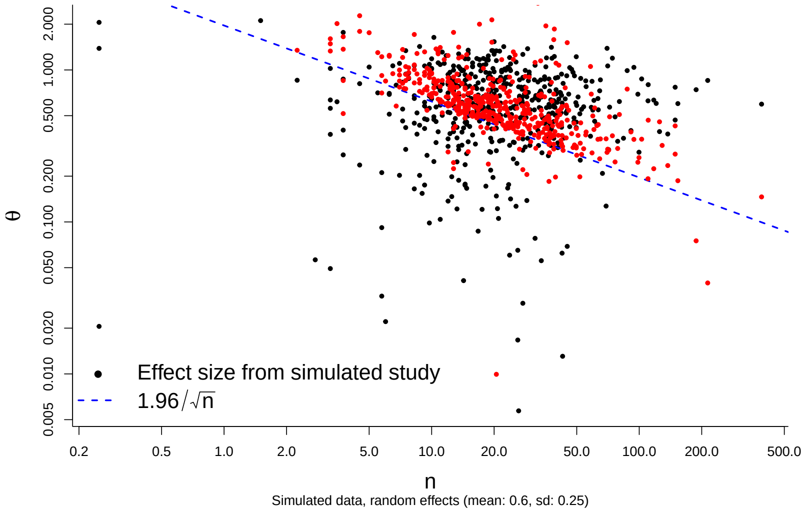
What is Classical Meta-analysis?

- Studies x_i are drawn according $x_i \sim N(\theta_i, se_i)$
  - θ_i are normalized – they are *effect sizes*.
  - Most common is the *standardized mean difference*, $\frac{\mu_1 - \mu_2}{\sigma}$.
  - Fixed effects: $\theta_i = \theta$ for all i.
  - Random effects: $\theta_i \sim N(\theta_0, \sigma_0^2)$.
- The effect sizes are usually closely related:
  - Effect of a class of anti-depressiva;
  - effect of some psychological intervention.
  - ... but they don't have to be *that* closely related!
- **Question:** Is the classical model realistic in presence of p-hacking?

What the Previous Plot *Should Have* Looked Like!



What the previous plot *should have* looked like!



What is Selection for Significance?

- Happens when only statistically significant results ($p < 0.05$) are published.

What is Selection for Significance?

- Happens when only statistically significant results ($p < 0.05$) are published.
  - In this case

$$p\left(x_i \mid \theta_i, se_i\right) = \phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right)$$

What is Selection for Significance?

- Happens when only statistically significant results ($p < 0.05$) are published.
  - In this case

$$p\left(x_i \mid \theta_i, se_i\right) = \phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right)$$

  - $\phi_{(a,b)}$ is a truncated normal.

What is Selection for Significance?

- Happens when only statistically significant results ($p < 0.05$) are published.
  - In this case

$$p\left(x_i \mid \theta_i, se_i\right) = \phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right)$$

  - $\phi_{(a,b)}$ is a truncated normal.
  - Can be caused by both p-hacking and publication bias.

What is Selection for Significance?

- Happens when only statistically significant results ($p < 0.05$) are published.
  - In this case

$$p\left(x_i \mid \theta_i, se_i\right) = \phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right)$$

  - $\phi_{(a,b)}$ is a truncated normal.
  - Can be caused by both p-hacking and publication bias.
- How to account for selection for significance?

What is Selection for Significance?

- Happens when only statistically significant results ($p < 0.05$) are published.
  - In this case

$$p\left(x_i \mid \theta_i, se_i\right) = \phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right)$$

  - $\phi_{(a,b)}$ is a truncated normal.
  - Can be caused by both p-hacking and publication bias.
- How to account for selection for significance?
  - Use the random / fixed effects model with a truncated normal!

What is Selection for Significance?

- Happens when only statistically significant results ($p < 0.05$) are published.
  - In this case

$$p\left(x_i \mid \theta_i, se_i\right) = \phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right)$$

  - $\phi_{(a,b)}$ is a truncated normal.
  - Can be caused by both p-hacking and publication bias.
- How to account for selection for significance?
  - Use the random / fixed effects model with a truncated normal!
  - Goes back to Hedges (1984); idea back to Lane & Dunlap (1978).

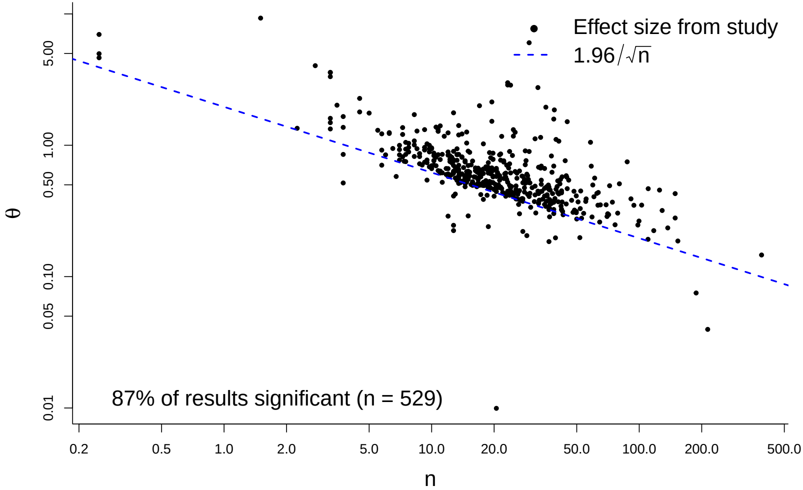
What is Selection for Significance?

- Happens when only statistically significant results ($p < 0.05$) are published.
  - In this case

$$p\left(x_i \mid \theta_i, se_i\right) = \phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right)$$

  - $\phi_{(a,b)}$ is a truncated normal.
  - Can be caused by both *p*-hacking and publication bias.
- How to account for selection for significance?
  - Use the random / fixed effects model with a truncated normal!
  - Goes back to Hedges (1984); idea back to Lane & Dunlap (1978).
- **Problem**: The first plot also contains studies that weren't affected by selection for significance!

Revisiting the First Plot



Motyl et al. (2017): The state of social and personality science

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!
- **Prototypical likelihood:**

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!
- **Prototypical likelihood:**

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- The parameter p is the *propensity to p-hack*!

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!
- **Prototypical likelihood:**

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- The parameter p is the *propensity to p-hack*!
  - Beta is a reasonable prior for $p \mid p_0, \psi \sim \mathrm{Beta}\left(p_0, \psi\right)$

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!
- **Prototypical likelihood:**

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- The parameter p is the *propensity to p-hack*!
  - Beta is a reasonable prior for $p \mid p_0, \psi \sim \mathrm{Beta}\left(p_0, \psi\right)$
  - But can and will depend on covariates!

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!
- **Prototypical likelihood:**

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- The parameter p is the *propensity to p-hack*!
  - Beta is a reasonable prior for $p \mid p_0, \psi \sim \mathrm{Beta}\left(p_0, \psi\right)$
  - But can and will depend on covariates!
- Model for θ_i could be normal: $\theta_i \mid \theta_0, \sigma_0 \sim N\left(\theta_0, \sigma_0^2\right)$

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!
- **Prototypical likelihood:**

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- The parameter p is the *propensity to p-hack*!
  - Beta is a reasonable prior for $p \mid p_0, \psi \sim \mathrm{Beta}\left(p_0, \psi\right)$
  - But can and will depend on covariates!
- Model for θ_i could be normal: $\theta_i \mid \theta_0, \sigma_0 \sim N\left(\theta_0, \sigma_0^2\right)$
  - Other models possible, for instance skew-normal.

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!
- **Prototypical likelihood:**

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- The parameter p is the *propensity to p-hack*!
  - Beta is a reasonable prior for $p \mid p_0, \psi \sim \mathrm{Beta}\left(p_0, \psi\right)$
  - But can and will depend on covariates!
- Model for θ_i could be normal: $\theta_i \mid \theta_0, \sigma_0 \sim N\left(\theta_0, \sigma_0^2\right)$
  - Other models possible, for instance skew-normal.
  - Can also depend on covariates.

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!
- **Prototypical likelihood:**

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96\cdot se_i,\infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- The parameter p is the *propensity to p-hack*!
  - Beta is a reasonable prior for $p \mid p_0, \psi \sim \mathrm{Beta}\left(p_0, \psi\right)$
  - But can and will depend on covariates!
- Model for θ_i could be normal: $\theta_i \mid \theta_0, \sigma_0 \sim N\left(\theta_0, \sigma_0^2\right)$
  - Other models possible, for instance skew-normal.
  - Can also depend on covariates.
- Computationally feasible due to STAN.

The Mixture Model Skeleton

- *Key idea:* Use *partial* selection for significance!
- **Prototypical likelihood:**

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- The parameter p is the *propensity to p-hack*!
  - Beta is a reasonable prior for $p \mid p_0, \psi \sim \mathrm{Beta}\left(p_0, \psi\right)$
  - But can and will depend on covariates!
- Model for θ_i could be normal: $\theta_i \mid \theta_0, \sigma_0 \sim N\left(\theta_0, \sigma_0^2\right)$
  - Other models possible, for instance skew-normal.
  - Can also depend on covariates.
- Computationally feasible due to STAN.
- On to examples!

Example I: Meta-analysis of a Field

The Effect Size Distribution in Psychology

- Data from (Motyl 2017); same as in the plots.

The Effect Size Distribution in Psychology

- Data from (Motyl 2017); same as in the plots.
- **Likelihood:**

$$p\left(x_i \mid \theta_i, p\right) = p\phi^{f}_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1 - p)\,\phi^{f}\left(\theta_i, se_i\right)$$

The Effect Size Distribution in Psychology

- Data from (Motyl 2017); same as in the plots.
- **Likelihood:**

$$p\left(x_i \mid \theta_i, p\right) = p\phi^f_{\left(1.96 \cdot se_i, \infty\right)}\left(\theta_i, se_i\right) + \left(1 - p\right)\phi^f\left(\theta_i, se_i\right)$$

  - Here ϕ^f is the *folded normal*, distribution of $|Z|$ where Z is normal.

The Effect Size Distribution in Psychology

- Data from (Motyl 2017); same as in the plots.
- **Likelihood:**

$$p\left(x_i \mid \theta_i, p\right) = p\phi^f_{(1.96\cdot se_i,\infty)}\left(\theta_i, se_i\right) + \left(1-p\right)\phi^f\left(\theta_i, se_i\right)$$

  - Here ϕ^f is the *folded normal*, distribution of $|Z|$ where Z is normal.
  - Used because the effect sizes have no inherent sign.

The Effect Size Distribution in Psychology

- Data from (Motyl 2017); same as in the plots.
- **Likelihood:**

$$p\left(x_i \mid \theta_i, p\right) = p\phi^f_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + \left(1 - p\right)\phi^f\left(\theta_i, se_i\right)$$

  - Here ϕ^f is the *folded normal*, distribution of $|Z|$ where Z is normal.
  - Used because the effect sizes have no inherent sign.
  - *Note:* θ_i can still be negative.

The Effect Size Distribution in Psychology

- Data from (Motyl 2017); same as in the plots.
- **Likelihood:**

$$p\left(x_i \mid \theta_i, p\right) = p\phi^f_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + \left(1-p\right)\phi^f\left(\theta_i, se_i\right)$$

  - Here ϕ^f is the *folded normal*, distribution of $|Z|$ where Z is normal.
  - Used because the effect sizes have no inherent sign.
  - *Note:* θ_i can still be negative.
- **Effect size distribution:** $\theta_i \sim \mathrm{Gumbel}\left(\mu, \sigma\right)$

The Effect Size Distribution in Psychology

- Data from (Motyl 2017); same as in the plots.
- **Likelihood:**

$$p\left(x_i \mid \theta_i, p\right) = p\phi^f_{(1.96\cdot se_i,\infty)}\left(\theta_i, se_i\right) + \left(1-p\right)\phi^f\left(\theta_i, se_i\right)$$

  - Here ϕ^f is the *folded normal*, distribution of $|Z|$ where Z is normal.
  - Used because the effect sizes have no inherent sign.
  - *Note:* θ_i can still be negative.
- **Effect size distribution:** $\theta_i \sim \mathrm{Gumbel}\left(\mu, \sigma\right)$
  - Has positive skewness, which we probably want.

The Effect Size Distribution in Psychology

- Data from (Motyl 2017); same as in the plots.
- **Likelihood:**

$$p\left(x_i \mid \theta_i, p\right) = p\phi^f_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + \left(1 - p\right)\phi^f\left(\theta_i, se_i\right)$$

  - Here ϕ^f is the *folded normal*, distribution of $|Z|$ where Z is normal.
  - Used because the effect sizes have no inherent sign.
  - *Note:* θ_i can still be negative.
- **Effect size distribution:** $\theta_i \sim \mathrm{Gumbel}\left(\mu, \sigma\right)$
  - Has positive skewness, which we probably want.
  - Has much better fit than the normal distribution.

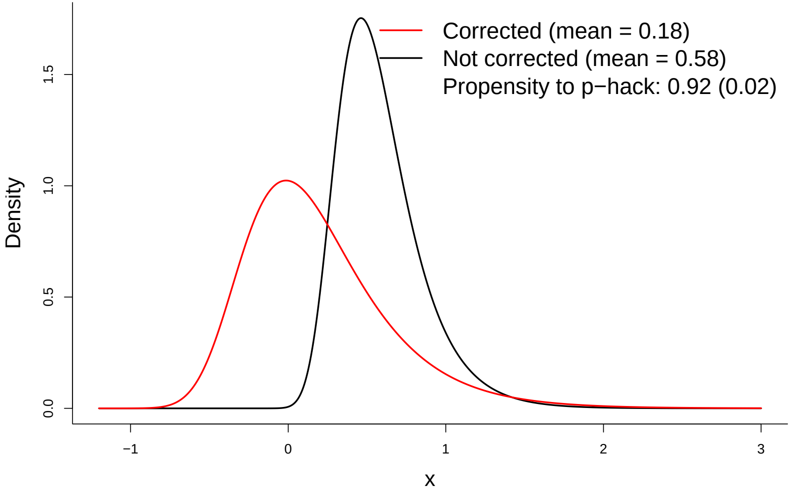
The Effect Size Distribution in Psychology

- Data from (Motyl 2017); same as in the plots.
- **Likelihood:**

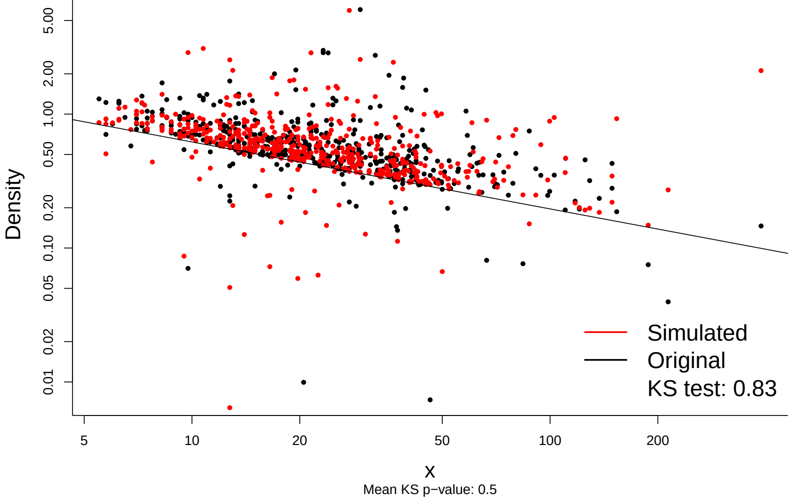
$$p\left(x_i \mid \theta_i, p\right) = p\phi^f_{(1.96\cdot se_i,\infty)}\left(\theta_i, se_i\right) + \left(1-p\right)\phi^f\left(\theta_i, se_i\right)$$

  - Here ϕ^f is the *folded normal*, distribution of $|Z|$ where Z is normal.
  - Used because the effect sizes have no inherent sign.
  - *Note:* θ_i can still be negative.
- **Effect size distribution:** $\theta_i \sim \mathrm{Gumbel}\left(\mu, \sigma\right)$
  - Has positive skewness, which we probably want.
  - Has much better fit than the normal distribution.
- **Priors:** $\theta_0 \sim N\left(0,1\right)$, $\sigma \sim \mathrm{Exp}\left(1\right)$, $p \sim \mathrm{Uniform}$

Posterior Predictive Distributions

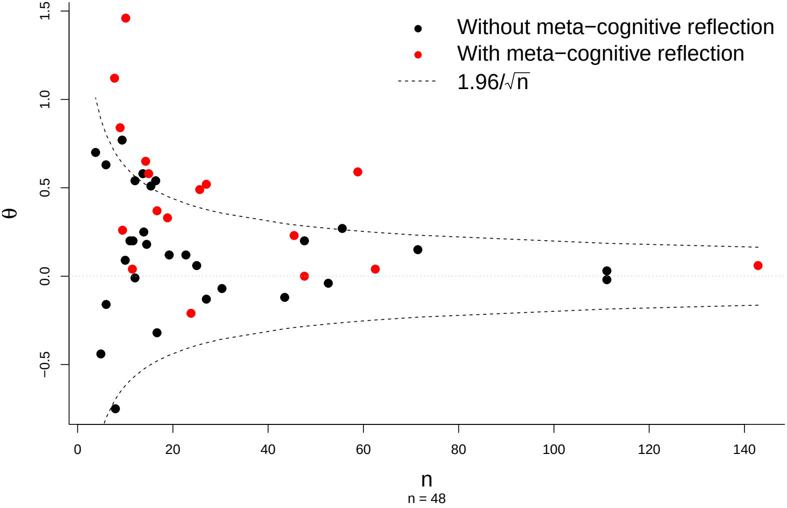


A Simulation from the Posterior



Example II: Ordinary Meta-analysis

Bangert-Drowns, Hurley & Wilkinson (2004)



Is there a difference between red and black?

- Likelihood:

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{\left(1.96 \cdot se_i, \infty\right)}\left(\theta_i, se_i\right) + \left(1 - p\right)\phi\left(\theta_i, se_i\right)$$

Is there a difference between red and black?

- Likelihood:

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- Effect size distribution:

$$\theta_i \sim N(\theta_0 + \theta_M \cdot \mathrm{Meta?}, \sigma)$$

Is there a difference between red and black?

- Likelihood:

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- Effect size distribution:

$$\theta_i \sim N(\theta_0 + \theta_M \cdot \mathrm{Meta?}, \sigma)$$

- Priors: $\theta_0, \theta_M \sim N(0,1)$, $\sigma \sim \mathrm{Exp}(1)$, $p \sim \mathrm{Uniform}$

Is there a difference between red and black?

- Likelihood:

$$p\left(x_i \mid \theta_i, se_i, p\right) = p\phi_{(1.96 \cdot se_i, \infty)}\left(\theta_i, se_i\right) + (1-p)\,\phi\left(\theta_i, se_i\right)$$

- Effect size distribution:

$$\theta_i \sim N(\theta_0 + \theta_M \cdot \mathrm{Meta?}, \sigma)$$

- Priors: $\theta_0, \theta_M \sim N\left(0, 1\right)$, $\sigma \sim \mathrm{Exp}\left(1\right)$, $p \sim \mathrm{Uniform}$

Model	θ_0	θ_M	σ	p
Corrected	0.02(0.06)	0.14(0.16)	0.1(0.1)	0.3(0.07)
Not corrected	0.14(0.05)	0.21(0.10)	0.2(0.05)	*NA*

-

Without Meta-cognition

- Logistic regression for the propensity to p-hack!

Without Meta-cognition

- Logistic regression for the propensity to p-hack!
- $p \mid n = \operatorname{logit}\left(p_0 + p_1 \log n\right)$

Without Meta-cognition

- Logistic regression for the propensity to p-hack!
- $p \mid n = \mathrm{logit}\left(p_0 + p_1 \log n\right)$
- $p_0, p_1 \sim N\left(0, 1\right)$

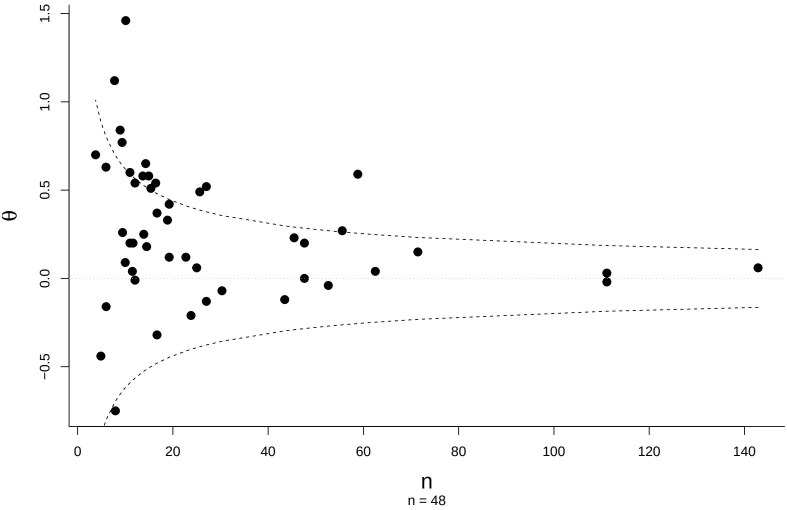
Without Meta-cognition

- Logistic regression for the propensity to *p*-hack!
- $p \mid n = \mathrm{logit}\left(p_0 + p_1 \log n\right)$
- $p_0, p_1 \sim N\left(0, 1\right)$
- *p*-hacking becomes *harder* with increasing sample size.

Without Meta-cognition

- Logistic regression for the propensity to p-hack!
- $p \mid n = \mathrm{logit}\left(p_0 + p_1 \log n\right)$
- $p_0, p_1 \sim N\left(0, 1\right)$
- p-hacking becomes *harder* with increasing sample size.
- Publication is easier with larger n.

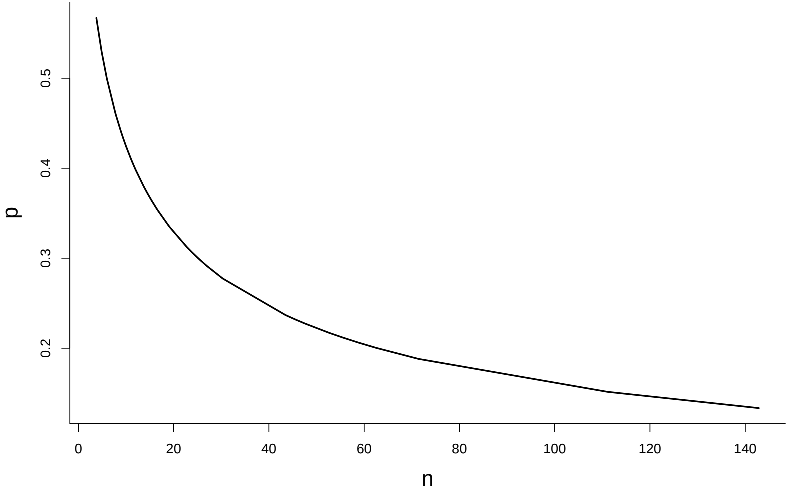
Visually Evidence for Logistic Regression on *p*



Results without Meta-cognition

Model	θ_0	σ	p_0	p_1
Random effects	$0.07(0.07)$	$0.1(0.05)$	$-0.7(0.3)$	$-0.5(0.3)$
Fixed effects	$0.04(0.07)$	NA	$-0.8(0.3)$	$-0.6(0.4)$

The Shape of the p-hacking Propensity



Conclusion

The End

- p-hacking is everywhere and must be accounted for!

The End

- p-hacking is everywhere and must be accounted for!
- When correcting for publication bias we should

The End

- p-hacking is everywhere and must be accounted for!
- When correcting for publication bias we should
  - Care about selection for significance!

The End

- *p*-hacking is everywhere and must be accounted for!
- When correcting for publication bias we should
  - Care about selection for significance!
  - Try out the mixture model to do the correction!

The End

- p-hacking is everywhere and must be accounted for!
- When correcting for publication bias we should
  - Care about selection for significance!
  - Try out the mixture model to do the correction!
  - Stay Bayesian!

The End

- p-hacking is everywhere and must be accounted for!
- When correcting for publication bias we should
  - Care about selection for significance!
  - Try out the mixture model to do the correction!
  - Stay Bayesian!
- An R-package at GitHub:

`straussR`

Statistical Reanalysis under Selection for Significance

`https://github.com/JonasMoss/straussR`

References

Bangert-Drowns, Robert L, Marlene M Hurley, and Barbara Wilkinson. 2004. "The Effects of School-Based Writing-to-Learn Interventions on Academic Achievement: A Meta-Analysis." *Review of Educational Research*.

Hedges, Larry V. 1984. "Estimation of Effect Size Under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences." *Journal of Educational Statistics*.

Lane, David M, and William P Dunlap. 1978. "Estimating Effect Size: Bias Resulting from the Significance Criterion in Editorial Decisions." *British Journal of Mathematical and Statistical Psychology*.

Motyl, Matt et al. 2017. "The State of Social and Personality Science." *Journal of Personality and Social Psychology*.