A Bayesian Perspective On The Reproducibility Project: Psychology

Alexander Etz & Joachim Vandekerckhove

@alxetz ←My Twitter (no ‘e’ in alex)
alexanderetz.com ← My website/blog

Purpose

- Revisit Reproducibility Project: Psychology (RPP)

- Compute Bayes factors
  - Account for publication bias in original studies

- Evaluate and compare levels of statistical evidence

TLDR: Conclusions first

- 75% of studies find qualitatively similar levels of evidence in original and replication
  - 64% find weak evidence (BF < 10) in both attempts
  - 11% of studies find strong evidence (BF > 10) in both attempts

TLDR: Conclusions first

- 75% of studies find qualitatively similar levels of evidence in original and replication
  - 64% find weak evidence (BF < 10) in both attempts
  - 11% of studies find strong evidence (BF > 10) in both attempts

- 10% find strong evidence in replication but not original

- 15% find strong evidence in original but not replication

The RPP

- 270 scientists attempt to closely replicate 100 psychology studies
  - Use original materials (when possible)
  - Work with original authors

- Pre-registered to avoid bias
  - Analysis plan specified in advance
  - Guaranteed to be published regardless of outcome

The RPP

- 2 main criteria for grading replication

The RPP

- 2 main criteria for grading replication

- Is the replication result statistically significant ($p < .05$) in the same direction as original?
  - 39% success rate

The RPP

- 2 main criteria for grading replication

- Is the replication result statistically significant ($p < .05$) in the same direction as original?
  - 39% success rate

- Does the replication's confidence interval capture original reported effect?
  - 47% success rate

The RPP

- Neither of these metrics are any good
  - (at least not as used)

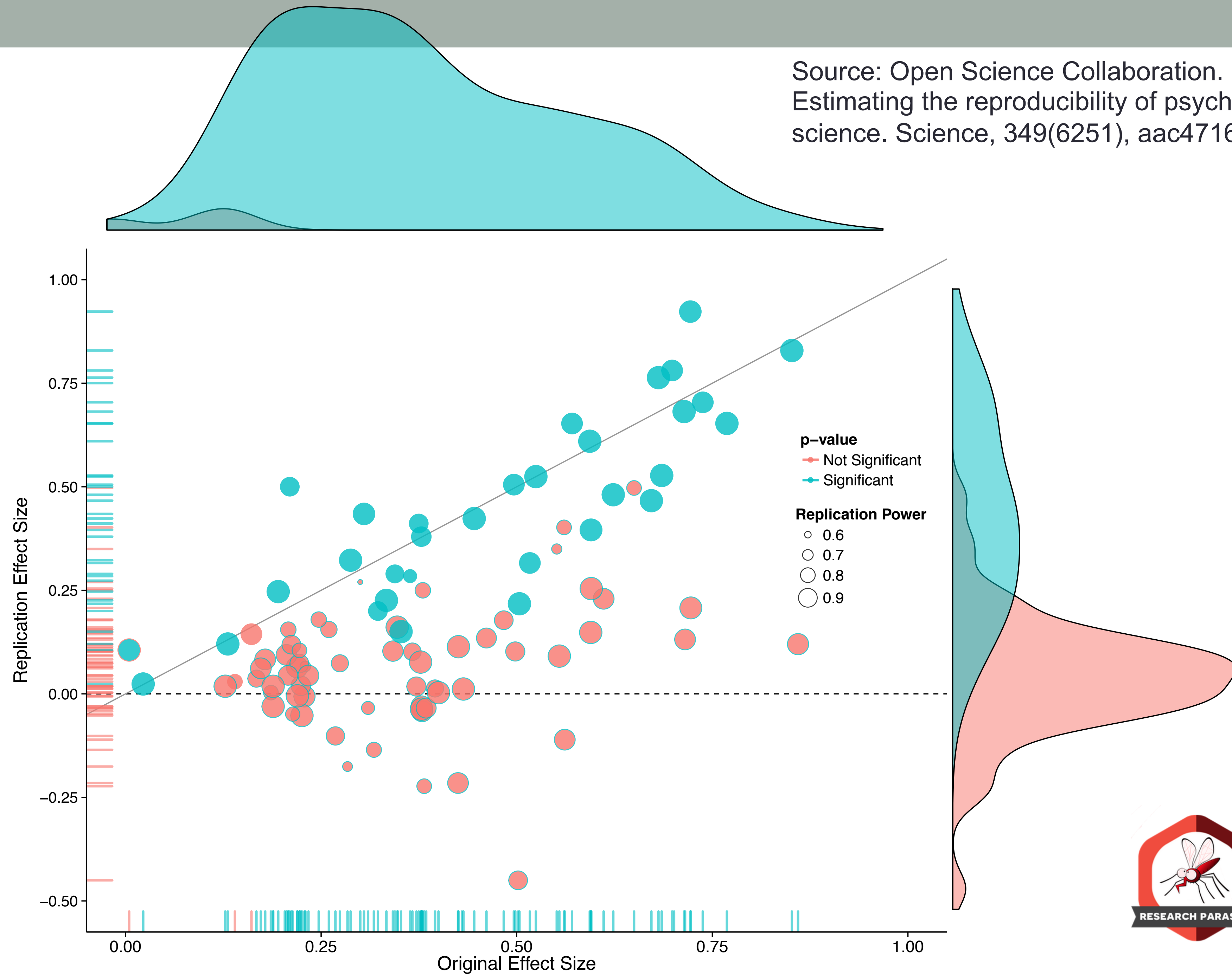
- Neither make predictions about out-of-sample data

- Comparing significance levels is bad
  - "The difference between significant and not significant is not necessarily itself significant"
    - -Gelman & Stern (2006)

The RPP

- Nevertheless, .51 correlation between original & replication effect sizes
- Indicates at least some level of robustness

Source: Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.



What can explain the discrepancies?

Moderators

- Two study attempts are in different contexts
  - Texas vs. California

- Different context = different results?
  - Conservative vs. Liberal sample

Low power in original studies

- Statistical power:
  - The frequency with which a study will yield a statistically significant effect in repeated sampling, assuming that the underlying effect is of a given size.

- Low powered designs undermine credibility of statistically significant results
  - Button et al. (2013)
  - Type M / Type S errors (Gelman & Carlin, 2014)

Low power in original studies

- Replications planned to have minimum 80% power
  - Report average power of 92%

Publication bias

- Most published results are "positive" findings
  - Statistically significant results
- Most studies designed to reject H_0
  - Most published studies succeed
- Selective preference = bias
  - "Statistical significance filter"

Statistical significance filter

- Incentive to have results that reach $p < .05$
  - "Statistically significant"
  - Evidential standard

- Studies with large effect size achieve significance
  - Get published

RESEARCH PARASITE

Statistical significance filter

- Studies with smaller effect size don’t reach significance
  - Get suppressed
- Average effect size inevitably inflates
- Replication power calculations meaningless

Can we account for this bias?

- Consider publication as part of data collection process

- This enters through likelihood function
  - Data generating process
  - Sampling distribution

Mitigation of publication bias

- Remember the *statistical significance filter*

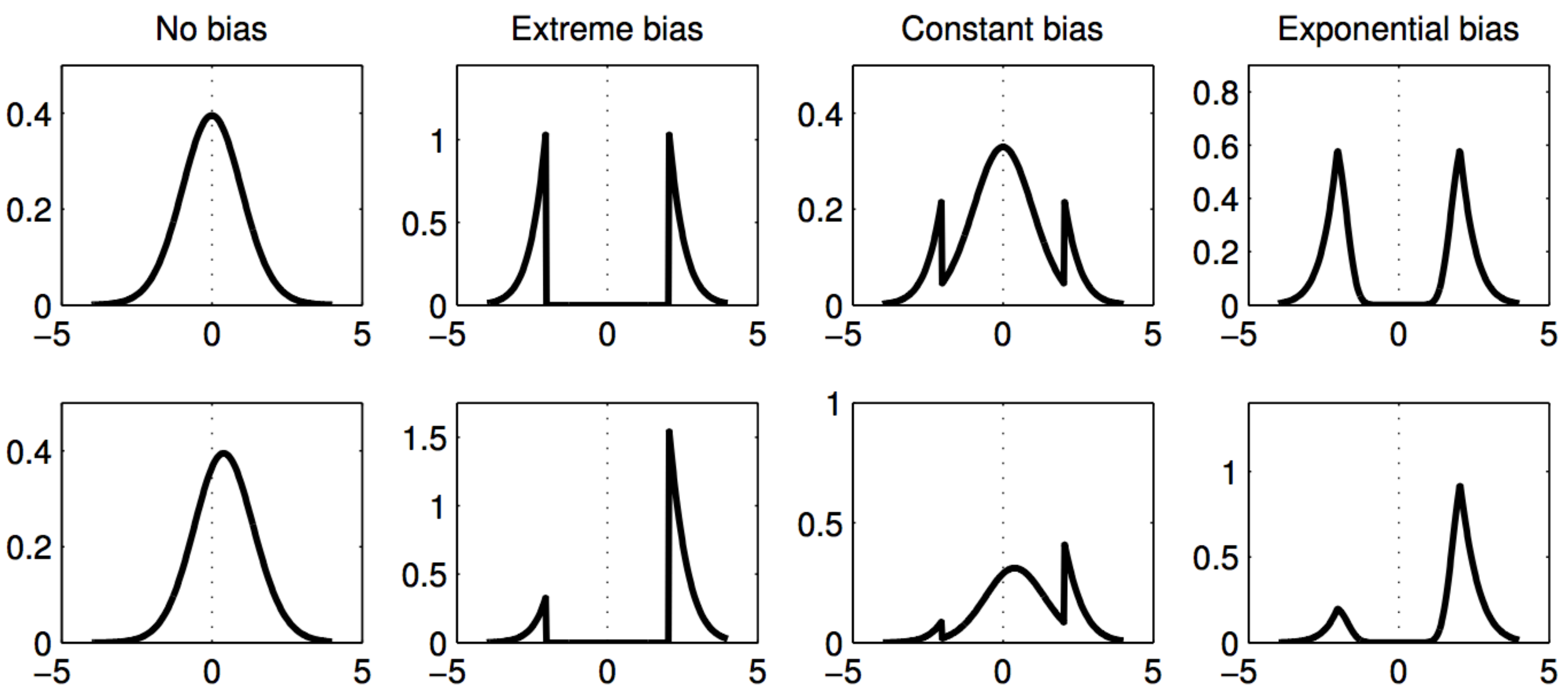
- We try to build a statistical model of it

Mitigation of publication bias

- We formally model 4 possible significance filters
  - 4 models comprise overall H_0
  - 4 models comprise overall H_1
- If result consistent with bias, then Bayes factor penalized
  - Raise the evidence bar

Mitigation of publication bias

- Expected distribution of test statistics that make it to the literature.



Mitigation of publication bias

- None of these are probably right
  - (Definitely all wrong)

- But it is a reasonable start

- Doesn't matter really
  - We're going to mix and mash them all together
  - "Bayesian Model Averaging"

RESEARCH PARASITE

The Bayes Factor

- How the data shift the balance of evidence
- Ratio of predictive success of the models

$$BF_{10} = \frac{p(data \mid H_1)}{p(data \mid H_0)}$$

RESEARCH PARASITE

The Bayes Factor

$$BF_{10} = \frac{p(data \mid H_1)}{p(data \mid H_0)}$$

- H_0: Null hypothesis
- H_1: Alternative hypothesis

The Bayes Factor

$$BF_{10} = \frac{p(data \mid H_1)}{p(data \mid H_0)}$$

- $BF_{10} > 1$ means evidence favors H_1
- $BF_{10} < 1$ means evidence favors H_0
- Need to be clear what H_0 and H_1 represent

The Bayes Factor

$$BF_{10} = \frac{p(data \mid H_1)}{p(data \mid H_0)}$$

- H_0: d = 0

The Bayes Factor

$$BF_{10} = \frac{p(data \mid H_1)}{p(data \mid H_0)}$$

- H_0: d = 0
- H_1: d ≠ 0

The Bayes Factor

$$BF_{10} = \frac{p(data \mid H_1)}{p(data \mid H_0)}$$

- H_0: d = 0

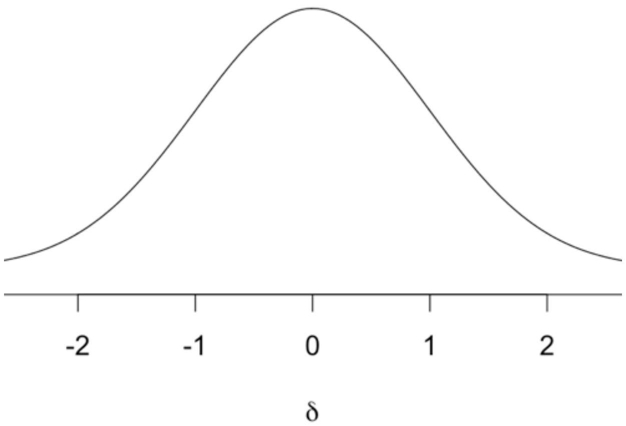
- H_1: d ≠ 0 (BAD)
  - Too vague
  - Doesn't make predictions

The Bayes Factor

$$BF_{10} = \frac{p(data \mid H_1)}{p(data \mid H_0)}$$

- H_0: d = 0

- H_1: d ~ Normal(0, 1)
  - The effect is probably small
  - Almost certainly -2 < d < 2"



Statistical evidence

- Do independent study attempts obtain similar amounts of evidence?

Statistical evidence

- Do independent study attempts obtain similar amounts of evidence?
  - Same prior distribution for both attempts
  - Measuring general evidential content
  - We want to evaluate evidence from outsider perspective

Interpreting evidence

- "How convincing would these data be to a neutral observer?"
  - 1:1 prior odds for H_1 vs. H_0
  - 50% prior probability for each

Interpreting evidence

- BF > 10 is sufficiently evidential
  - 10:1 posterior odds for H_1 vs. H_0 (or vice versa)
  - 91% posterior probability for H_1 (or vice versa)

Interpreting evidence

- BF > 10 is sufficiently evidential
  - 10:1 posterior odds for H_1 vs. H_0 (or vice versa)
  - 91% posterior probability for H_1 (or vice versa)

- BF of 3 is too weak
  - 3:1 posterior odds for H_1 vs. H_0 (or vice versa)
  - Only 75% posterior probability for H_1 (or vice versa)

Interpreting evidence

- It depends on context (of course)

- You can have higher or lower standards of evidence

Interpreting evidence

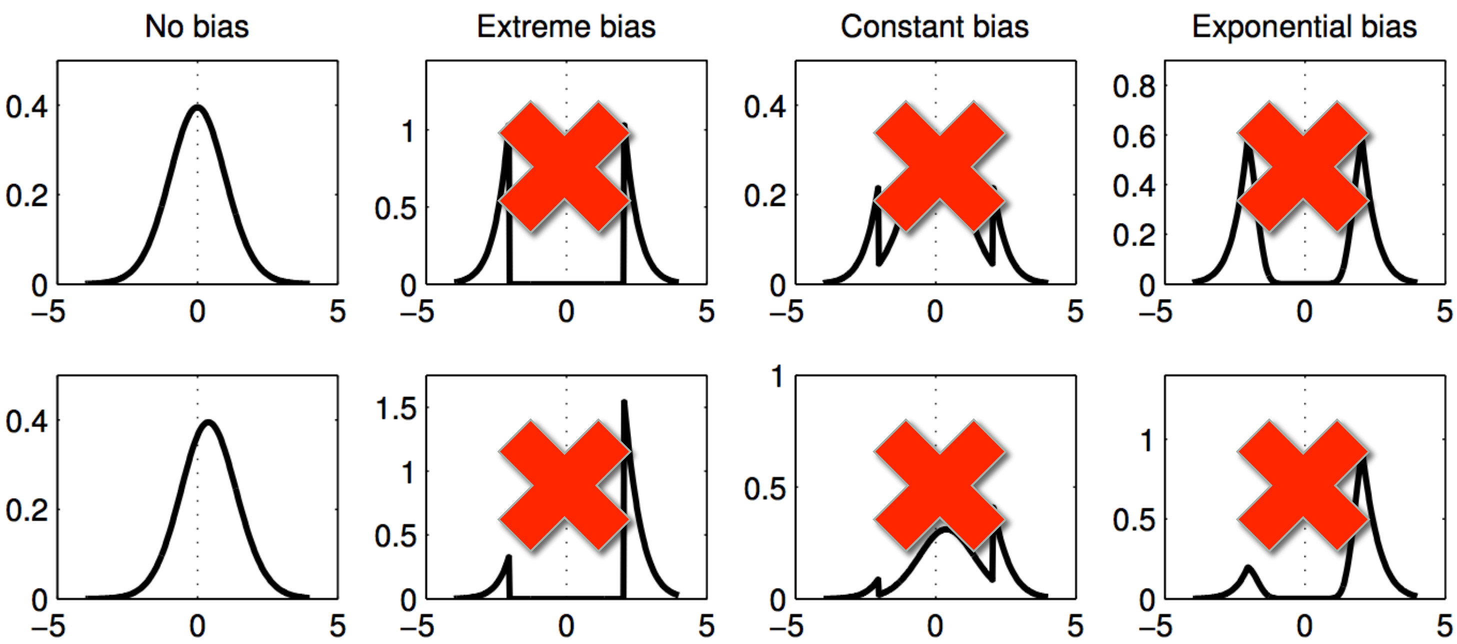
- How do p values stack up?
- American Statistical Association:
  - “Researchers should recognize that a p-value … near 0.05 taken by itself offers only weak evidence against the null hypothesis. “

Interpreting evidence

- How do p values stack up?
  - $p < .05$ is weak standard
  - $p = .05$ corresponds to $BF \le 2.5$ (at BEST)
  - $p = .01$ corresponds to $BF \le 8$ (at BEST)

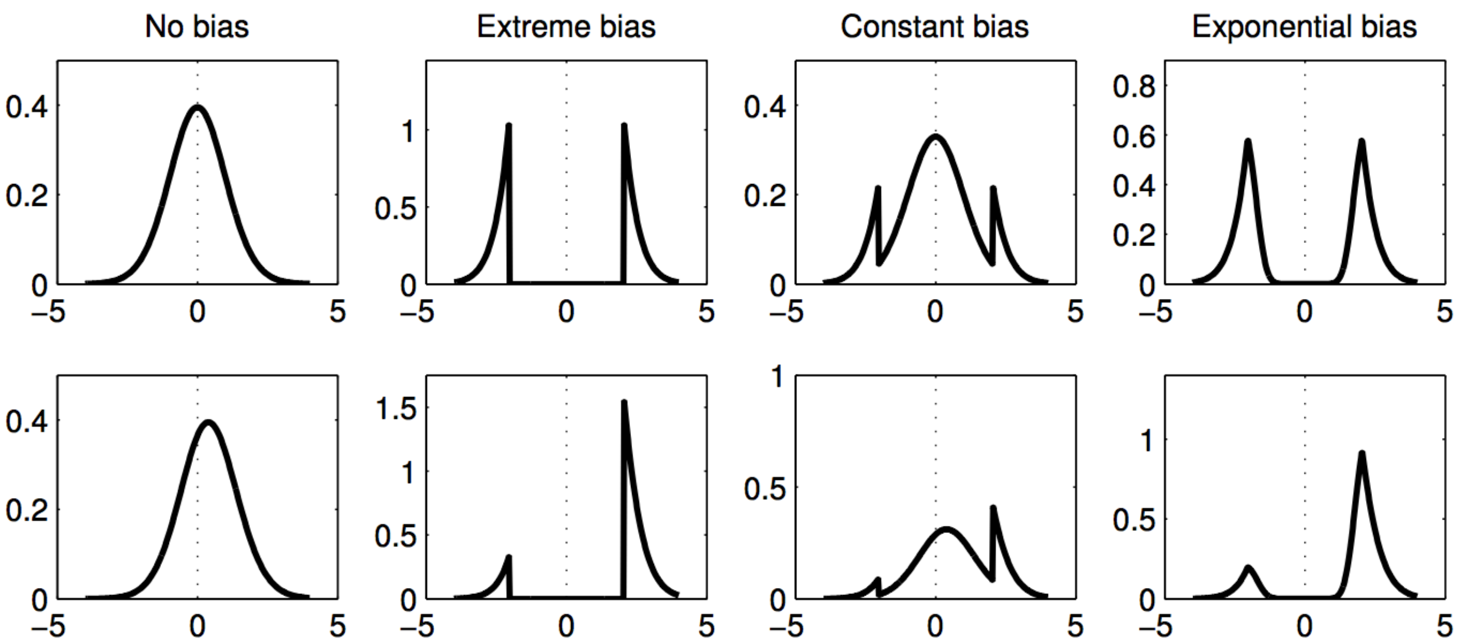
Face-value BFs

- Standard Bayes factor
- Bias free
- Results taken at face-value



Bias-mitigated BFs

- Bayes factor accounting for possible bias

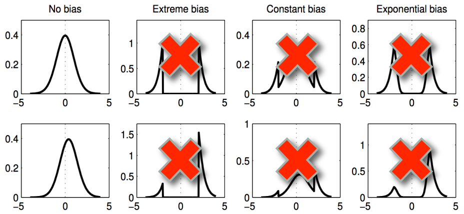


Illustrative Bayes factors

- Study 27
  - $t(31) = 2.27$, $p=.03$
  - Maximum $BF_{10} = 3.4$

Illustrative Bayes factors

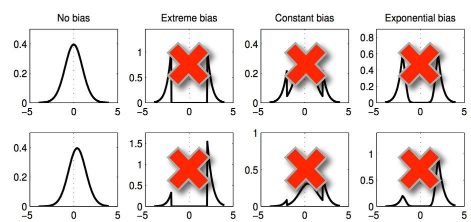
- Study 27
  - t(31) = 2.27, p=.03
  - Maximum BF_{10} = 3.4
  - Face-value BF_{10} = 2.9



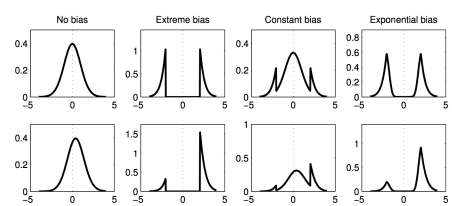
Illustrative Bayes factors

- Study 27
  - t(31) = 2.27, p=.03
  - Maximum BF_{10} = 3.4

  - Face-value BF_{10} = 2.9



  - Bias-mitigated BF_{10} = .81

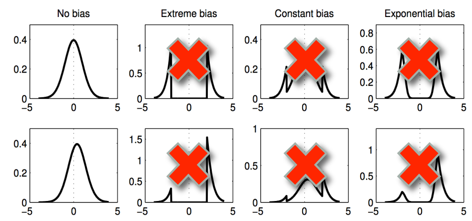


Illustrative Bayes factors

- Study 71
  - $t(373) = 4.4$, $p < .001$
  - Maximum BF_{10} = ~2300

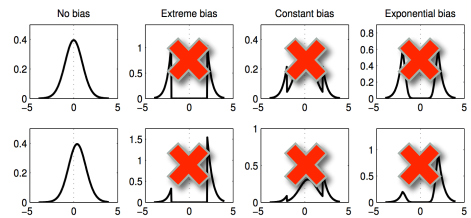
Illustrative Bayes factors

- Study 71
  - $t(373) = 4.4$, $p < .001$
  - Maximum BF_{10} = ~2300
  - Face-value $BF_{10} = 947$

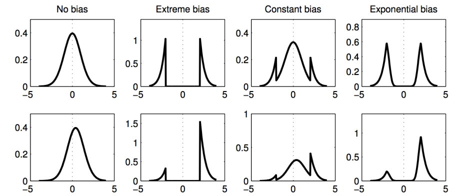


Illustrative Bayes factors

- Study 71
  - $t(373) = 4.4, p < .001$
  - Maximum BF_{10} = ~2300
  - Face-value BF_{10} = 947



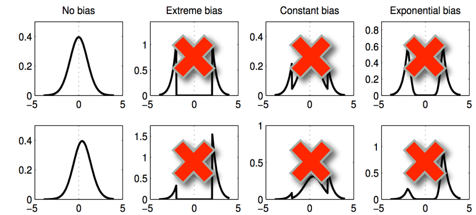
  - Bias-mitigated BF_{10} = 142



RPP Sample

- N=72
  - All univariate tests (t test, anova w/ 1 model df, etc.)

Results



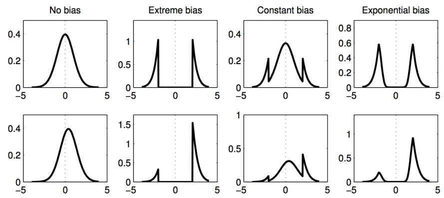
- Original studies, face-value
  - Ignoring pub bias

- 43% obtain $BF_{10} > 10$

- 57% obtain $1/10 < BF_{10} < 10$

- 0 obtain $BF_{10} < 1/10$

Results



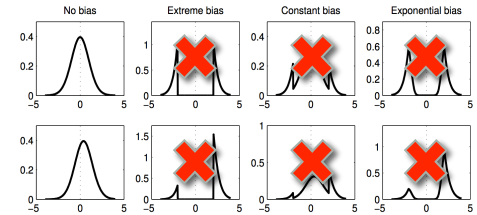
- Original studies, bias-corrected

- 26% obtain $BF_{10} > 10$

- 74% obtain $1/10 < BF_{10} < 10$

- 0 obtain $BF_{10} < 1/10$

Results

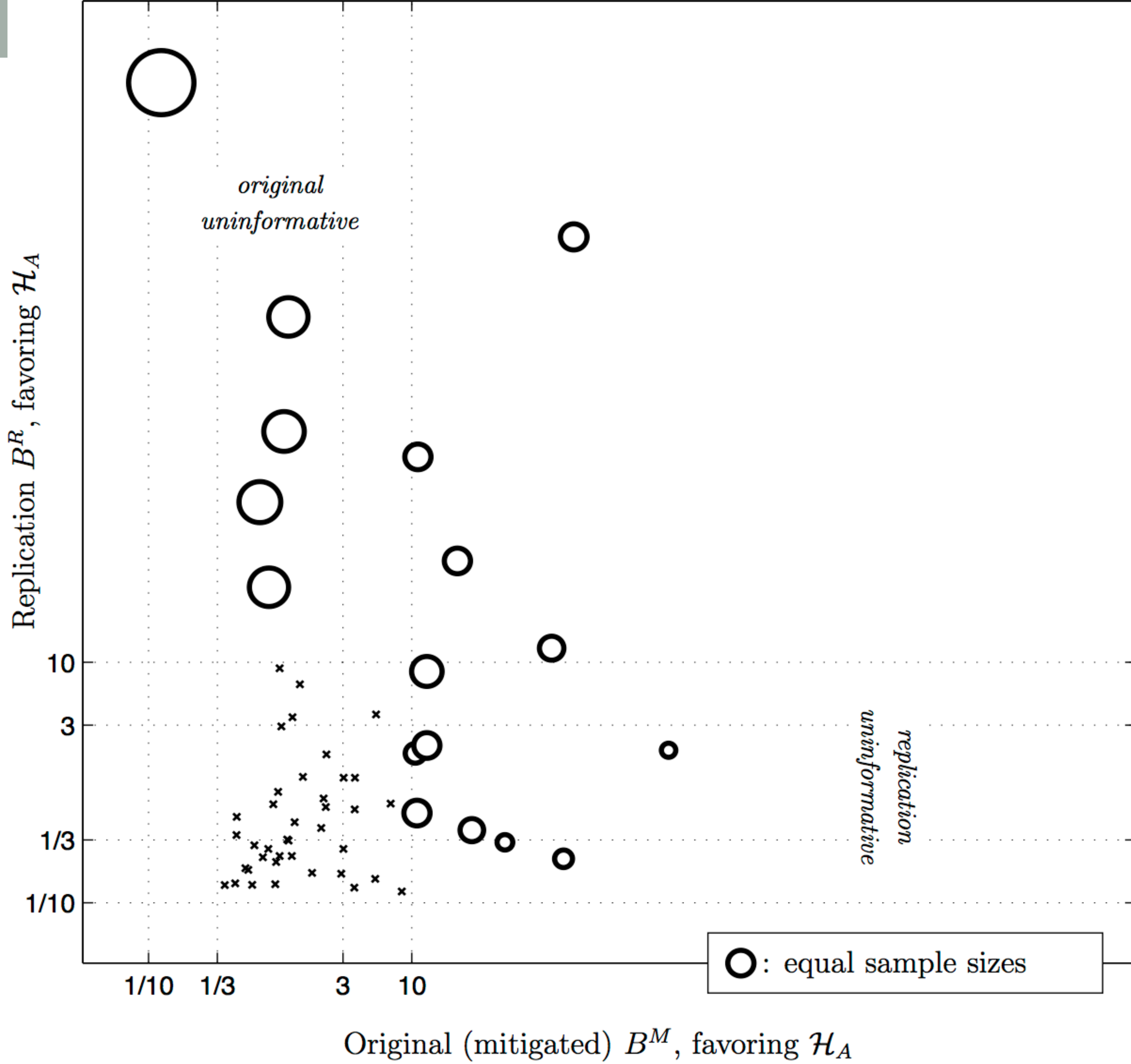


- Replication studies, face value
  - No chance for bias, no need for correction

- 21% obtain $BF_{10} > 10$

- 79% obtain $1/10 < BF_{10} < 10$

- 0 obtain $BF_{10} < 1/10$



original
uninformative
replication
uninformative
Replication B^R, favoring $\mathcal{H}_A$
10
3
1/3
1/10
1/10
1/3
3
10
Original (mitigated) B^M, favoring $\mathcal{H}_A$
: equal sample sizes

Consistency of results

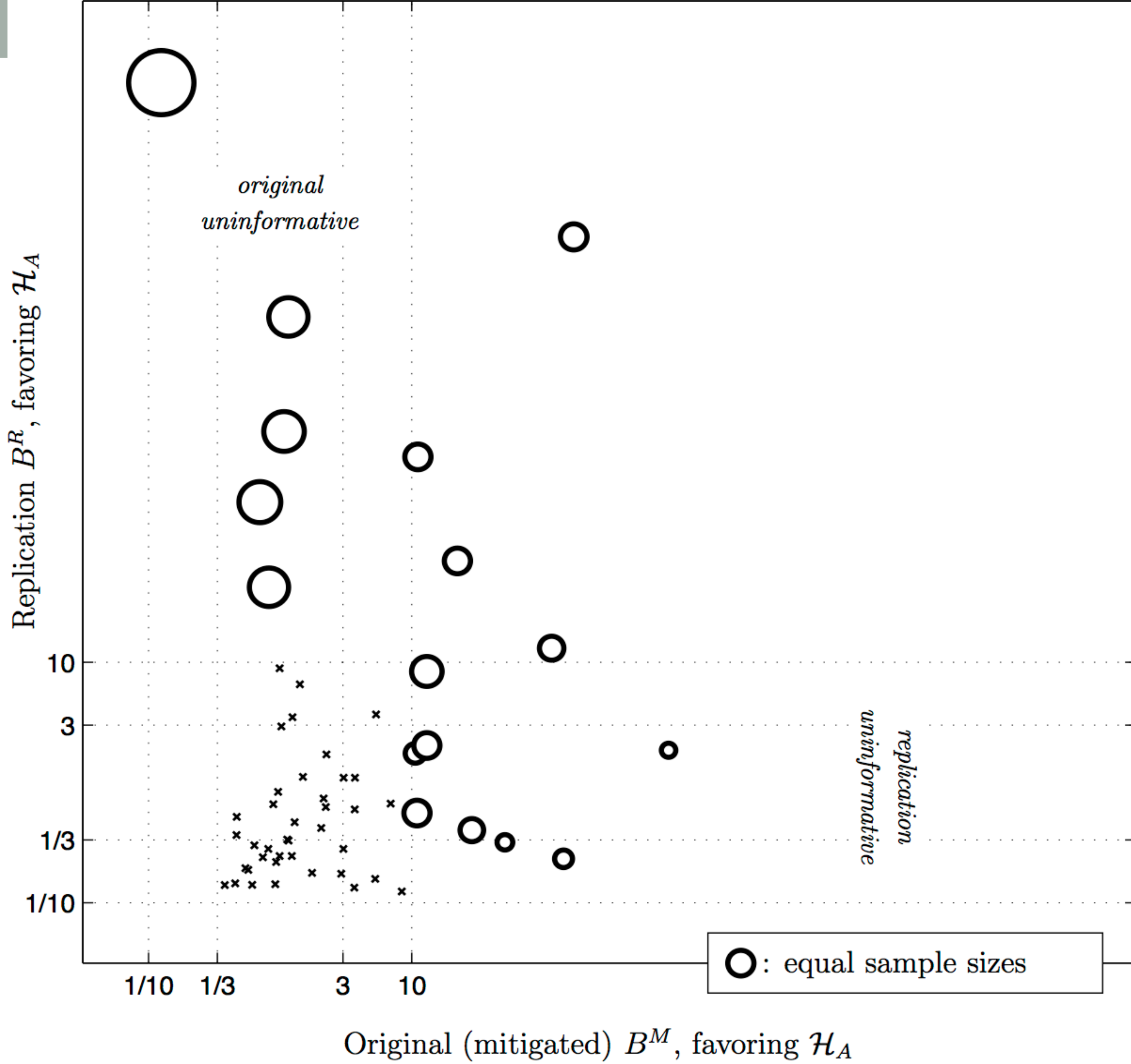
- No alarming inconsistencies
- 46 cases where both original and replication show only weak evidence
- Only 8 cases where both show $BF_{10} > 10$

Consistency of results

- 11 cases where original $BF_{10} > 10$, but not replication

- 7 cases where replication $BF_{10} > 10$, but not original

- In every case, the study obtaining strong evidence had the larger sample size


original
uninformative
replication
uninformative
10
3
1/3
1/10
1/10
1/3
3
10
: equal sample sizes
Replication B^R, favoring $\mathcal{H}_A$
Original (mitigated) B^M, favoring $\mathcal{H}_A$

Moderators?

- As Laplace would say, we have no need for that hypothesis

- Results adequately explained by:
  - Publication bias in original studies
  - Generally weak standards of evidence

Take home message

- Recalibrate our intuitions about statistical evidence
- Reevaluate expectations for replications
  - Given weak evidence in original studies

Thank you

Thank you

@alxetz ←My Twitter (no ‘e’ in alex)
alexanderetz.com ← My website/blog

@VandekerckhoveJ ← Joachim’s Twitter
joachim.cidlab.com ← Joachim’s website

Want to learn more Bayes?

- My blog:
  - alexanderetz.com/understanding-bayes
- "How to become a Bayesian in eight easy steps"
  - http://tinyurl.com/eightstepsdraft
- JASP summer workshop
  - https://jasp-stats.org/
  - Amsterdam, August 22-23, 2016

RESEARCH PARASITE

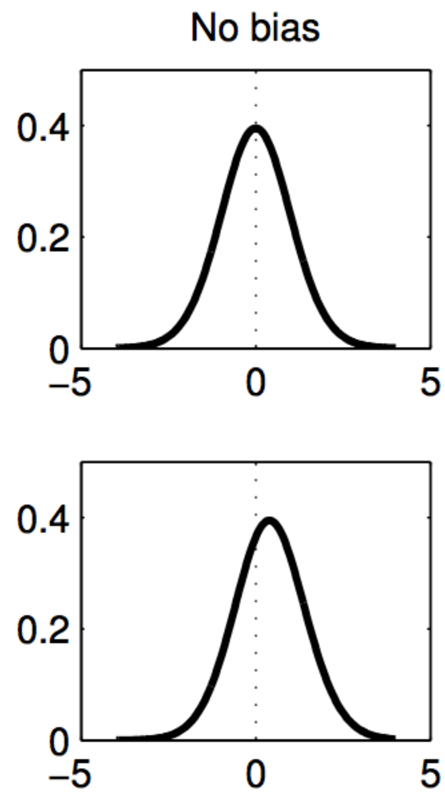
Technical Appendix

Mitigation of publication bias

- Model 1: No bias
  - Every study has same chance of publication

- Regular *t* distributions
  - H_0 true: Central *t*
  - H_0 false: Noncentral *t*

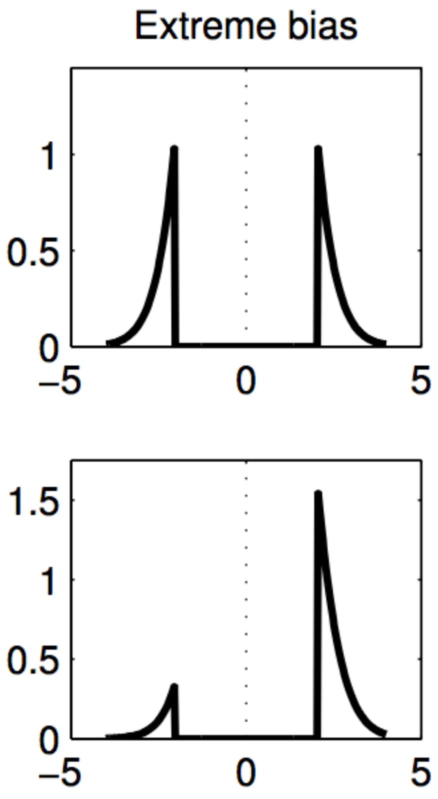
- These are used in standard BFs



Mitigation of publication bias

- Model 2: Extreme bias
  - Only statistically significant results published

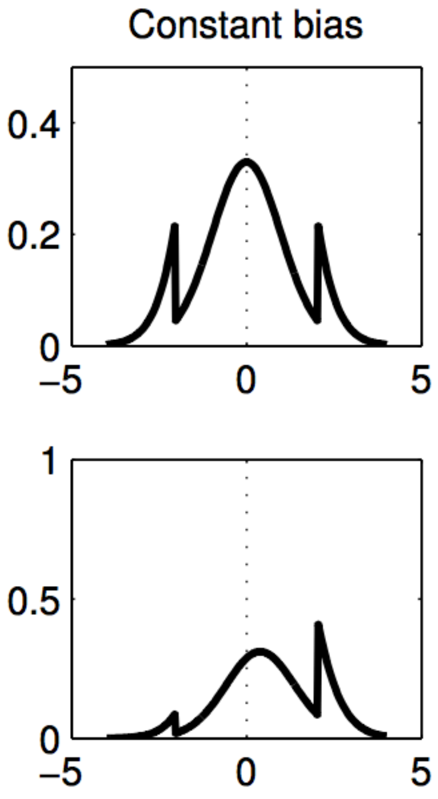
- *t* distributions but…
  - Zero density in the middle
  - Spikes in significant regions



Mitigation of publication bias

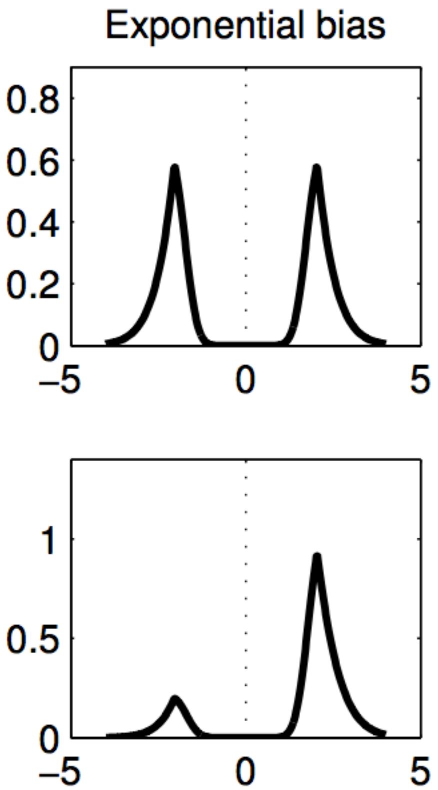
- Model 3: Constant-bias
  - Nonsignificant results published x% as often as significant results

- *t* distributions but…
  - Central regions downweighted
  - Large spikes over significance regions

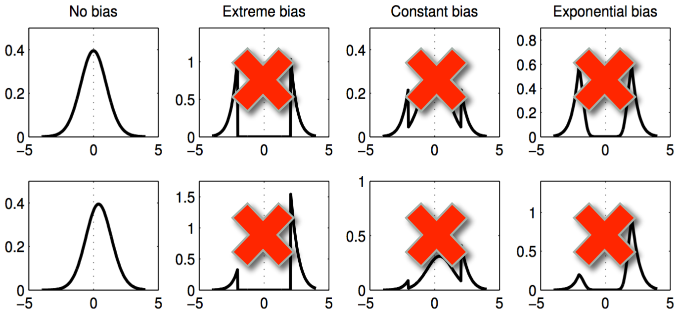


Mitigation of publication bias

- Model 4: Exponential bias
  - "Marginally significant" results have a chance to be published
  - Harder as p gets larger
- *t* likelihoods but…
  - Spikes over significance regions
  - Quick decay to zero density as
    (p – α) increases

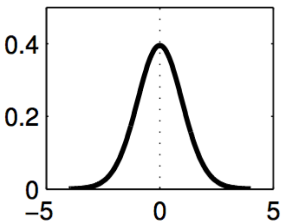


Calculate face-value BF



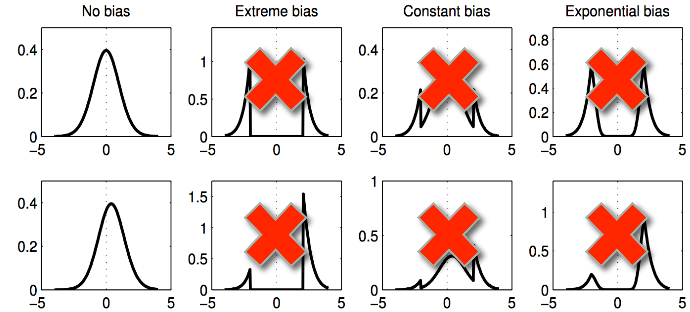
- Take the likelihood:

$$t_n(x \mid \delta)$$



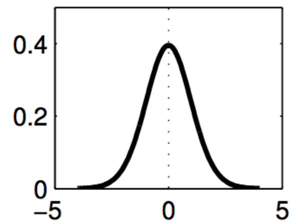
- Integrate with respect to prior distribution, p(δ)

Calculate face-value BF



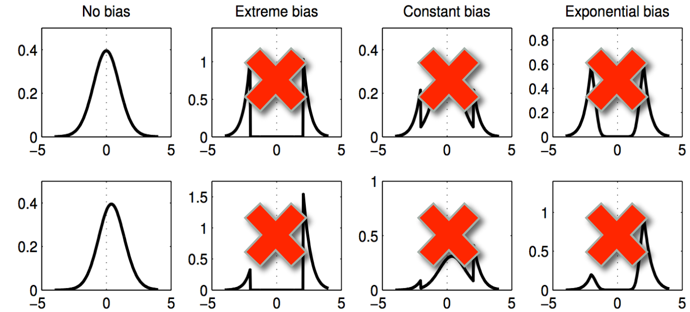
- Take the likelihood:

$$t_n(x \mid \delta)$$



- Integrate with respect to prior distribution, p(δ)
  - “What is the average likelihood of the data *given this model*?”
  - Result is marginal likelihood, M

Calculate face-value BF



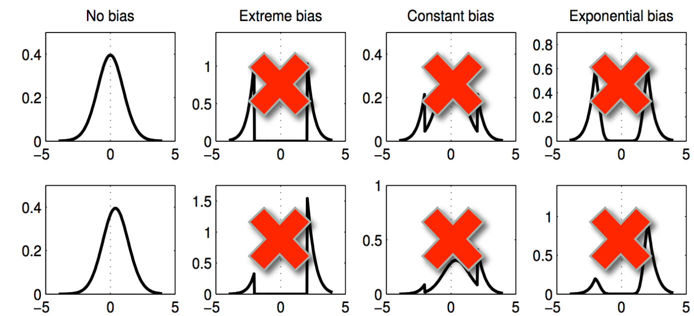
- For H_1: δ ~ Normal(0 , 1)

$$M_{+} = \int_{\Delta} t_n(x \mid \delta) p(\delta) d\delta$$

- For H_0: δ = 0

$$M_{-} = t_n(x \mid \delta = 0)$$

Calculate face-value BF



- For H_1: δ ~ Normal(0 , 1)

$$M_+ = \int_\Delta t_n(x \mid \delta)\, p(\delta)\, d\delta$$

$$\mathbf{BF_{10}} = \frac{p(data \mid H_1) = M_+}{p(data \mid H_0) = M_-}$$

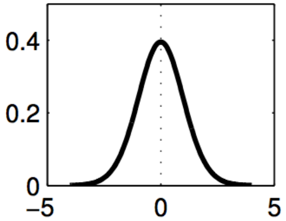
- For H_0: δ = 0

$$M_- = t_n(x \mid \delta = 0)$$

Calculate mitigated BF

- Start with regular *t* likelihood function

$$t_n(x \mid \delta)$$



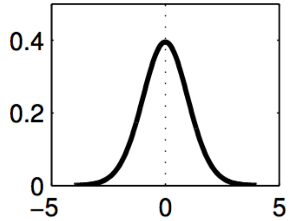
- Multiply it by bias function: w = {1, 2, 3, 4}
  - Where w=1 is no bias, w=2 is extreme bias, etc.

RESEARCH PARASITE

Calculate mitigated BF

- Start with regular *t* likelihood function

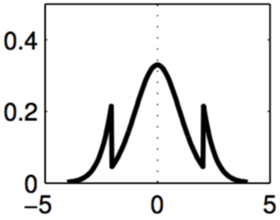
$$t_n(x \mid \delta)$$



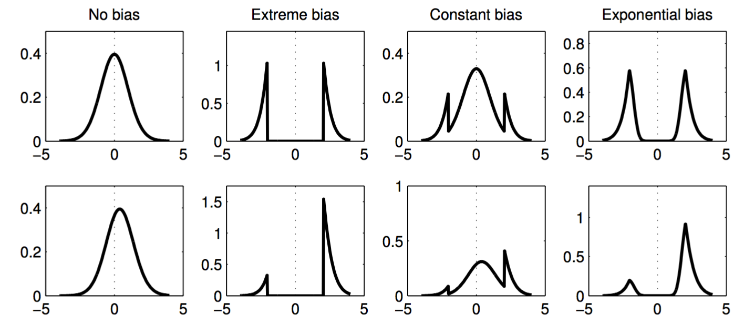
- Multiply it by bias function: w = {1, 2, 3, 4}
  - Where w=1 is no bias, w=2 is extreme bias, etc.

- E.g., when w=3 (constant bias):

$$t_n(x \mid \delta) \times w(x \mid \theta)$$

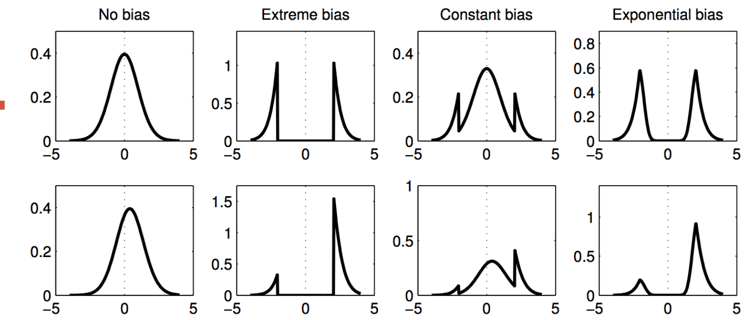


Calculate mitigated BF



- Too messy
- Rewrite as a new function

Calculate mitigated BF



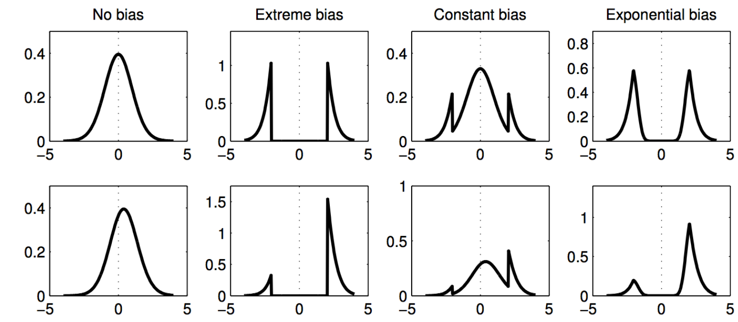
- Too messy
- Rewrite as a new function

- H_1: δ ~ Normal(0 , 1):

$$p_{w+}(x \mid n, \delta, \theta) \propto t_n(x \mid \delta) \times w(x \mid \theta)$$

RESEARCH PARASITE

Calculate mitigated BF



- Too messy
- Rewrite as a new function

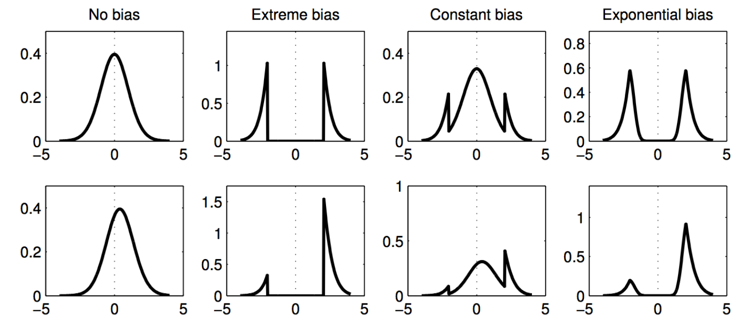
- H_1: δ ~ Normal(0 , 1):

$$p_{w+}(x \mid n, \delta, \theta) \propto t_n(x \mid \delta) \times w(x \mid \theta)$$

- H_0: δ = 0:

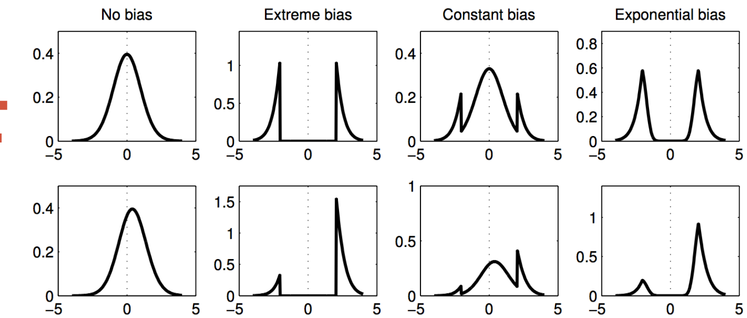
$$p_{w-}(x \mid n, \theta) = p_{w+}(x \mid n, \delta = 0, \theta)$$

Calculate mitigated BF



- Integrate w.r.t. p(θ) and p(δ)
  - “What is the average likelihood of the data *given each bias model*?”
  - p(data | bias model w):

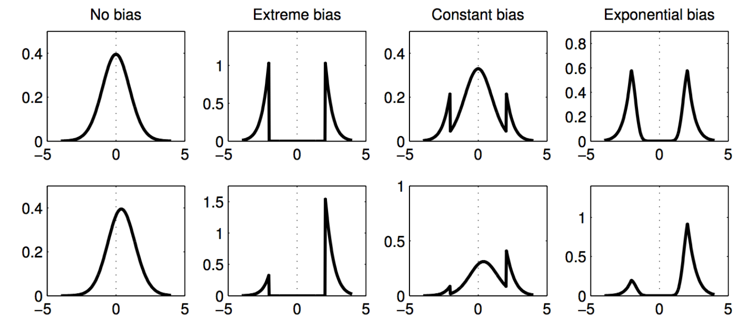
Calculate mitigated BF



- Integrate w.r.t. p(θ) and p(δ)
  - "What is the average likelihood of the data *given each bias model*?"
  - p(data | bias model w):

$$M_{w+} = \int_{\Theta}\int_{\Delta} p_{w+}(x \mid n, \delta, \theta)\, p(\delta)\, p(\theta)\, d\delta\, d\theta$$

Calculate mitigated BF



- Integrate w.r.t. p(θ) and p(δ)
  - “What is the average likelihood of the data *given each bias model*?”
  - p(data | bias model w):

$$M_{w+} = \int_{\Theta}\int_{\Delta} p_{w+}(x \mid n, \delta, \theta)\, p(\delta)\, p(\theta)\, d\delta\, d\theta$$

$$M_{w-} = \int_{\Theta} p_{w-}(x \mid n, \theta)\, p(\theta)\, d\theta$$

Calculate mitigated BF

- Take M_{w+} and M_{w-} and multiply by the weights of the corresponding bias model, then sum within each hypothesis

Calculate mitigated BF

- Take M_{w+} and M_{w-} and multiply by the weights of the corresponding bias model, then sum within each hypothesis

- For H_1:

$$p(w=1)M_{1+} + p(w=2)M_{2+} + p(w=3)M_{3+} + p(w=4)M_{4+}$$

RESEARCH PARASITE

Calculate mitigated BF

- Take M_{w+} and M_{w-} and multiply by the weights of the corresponding bias model, then sum within each hypothesis

- For H_1:

$$p(w=1)M_{1+} + p(w=2)M_{2+} + p(w=3)M_{3+} + p(w=4)M_{4+}$$

- For H_0:

$$p(w=1)M_{1-} + p(w=2)M_{2-} + p(w=3)M_{3-} + p(w=4)M_{4-}$$

Calculate mitigated BF

- Messy, so we restate as sums:

Calculate mitigated BF

- Messy, so we restate as sums:
- For H_1:

$$\sum_{w} p(w) M_{w+}$$

Calculate mitigated BF

- Messy, so we restate as sums:
- For H_1:

$$\sum_{w} p(w)M_{w+}$$

- For H_0:

$$\sum_{w} p(w)M_{w-}$$

Calculate mitigated BF

- Messy, so we restate as sums:
- For H_1:

$$\sum_w p(w)M_{w+}$$

$$\mathbf{BF_{10}} = \frac{p(data \mid H_1) = \sum_w p(w)M_{w+}}{p(data \mid H_0) = \sum_w p(w)M_{w-}}$$

- For H_0:

$$\sum_w p(w)M_{w-}$$

RESEARCH PARASITE