Estimating replicability of science by taking statistical significance into account

Robbie C.M. van Aert & Marcel A.L.M. van Assen

R.C.M.vanAert@tilburguniversity.edu

Tilburg University
Department of Methodology and Statistics

Tilburg University

NWO Nederlandse Organisatie voor Wetenschappelijk Onderzoek Maatschappij- en Gedragswetenschappen

The Problem

Example (Maxwell et al., 2015)

Independent sample t-test

Original: $d = 0.5$, $t(78) = 2.24$, $p = 0.028$

Replication (power = .8 at $d = 0.5$): $d = 0.23$, $t(170) = 1.50$, $p = 0.135$

Conclusion?!?

Questions considered relevant

1) Does effect exist? (0 or not)
2) What is magnitude of effect? (best guess)

Tilburg University 2

Omnipresent and Relevant

- Reproducibility Project Psychology (RPP):
  - Significant original study and non-significant replication in 63.9%
- Experimental Economics Replication Project (EE-RP):
  - Significant original study and non-significant replication in 31.2%
- Replication is often a starting point of a multi-study paper

Tilburg University 3

Problem and Solution

Problem

How to evaluate results of original study and replication?

Solution

Accurate evaluation of effect size …

… taking statistical significance of the original study into account

Tilburg University 4

Overview

1. Publication bias
2. Snapshot Bayesian Hybrid Meta-Analysis Method
3. Statistical properties of snapshot method
4. Application: RPP and EE-RP
5. Conclusion and discussion

Tilburg University 5

1. Publication bias

- Publication bias is 'the selective publication of studies with a statistically significant outcome'
- Overwhelming evidence of publication bias:
  - 95% of published articles contain significant results in psychology
- Consequences of publication bias:
  - False impression that effect exists
  - Overestimation of effect sizes
  - Questionable research practices

Tilburg University 6

2. Snapshot method: Basic idea

- **Snapshot** Bayesian Hybrid Meta-Analysis Method
  - Assume four effect sizes (zero, small, medium, large [Cohen]) → *snapshots*
- Snapshot **Bayesian** Hybrid Meta-Analysis Method
  - Compute posterior probability of these four effects → *Bayesian*
- Snapshot Bayesian **Hybrid** Meta-Analysis Method
  - Take statistical significance of original study into account → *hybrid*
- Snapshot Bayesian Hybrid **Meta-Analysis** Method
  - Combine original study with replication → *meta-analysis*

2. Snapshot method: Basic idea

- Density of the replication is "normal" pdf because no selection:

$$f_R = f(y = y_R; \theta)$$

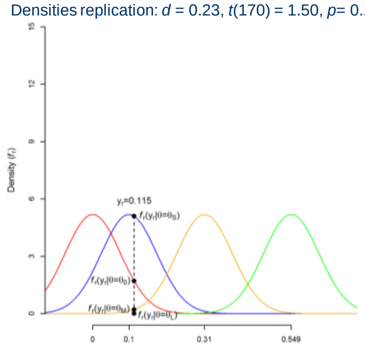
- Density of the original study is pdf conditional on effect size being statistically significant:

$$f_O = \frac{f(y = y_O; \theta)}{P(y \geq y_{CV}; \theta)}$$

- Assumptions:
  - Original study is statistically significant
  - Both studies estimate the same effect (fixed-effect)
  - No questionable research practices

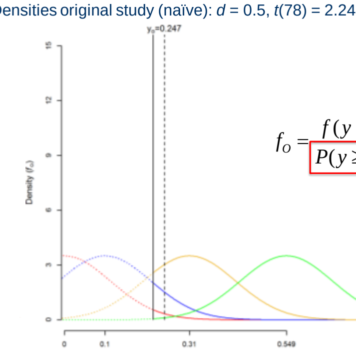
2. Snapshot method: Basic idea

Densities replication: *d* = 0.23, *t*(170) = 1.50, *p*= 0.135



2. Snapshot method: Basic idea

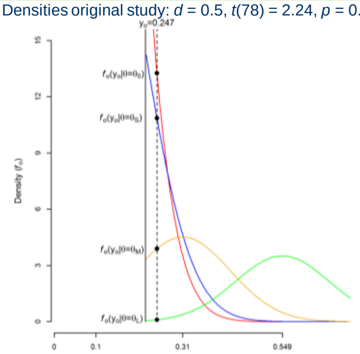
Densities original study (naïve): *d* = 0.5, *t*(78) = 2.24, *p* = 0.028



$$f_O = \frac{f(y = y_O; \theta)}{P(y \geq y_{CV}; \theta)}$$

2. Snapshot method: Basic idea

Densities original study: *d* = 0.5, *t*(78) = 2.24, *p* = 0.028



2. Snapshot method: Basic idea

- Combined likelihood:

$$L(\theta) = f_O(\theta) \times f_R(\theta)$$

- Posterior probabilities assuming a uniform prior for each snapshot are computed with:

$$\pi_x = \frac{L(\theta = x)}{L(\theta = \theta_0) + L(\theta = \theta_S) + L(\theta = \theta_M) + L(\theta = \theta_L)}$$

Advantages of method

- Easy and insightful
- Easy (re)computation posterior for other (than uniform) prior:

$$\pi_x^* = \frac{p_x \pi_x}{p_0 \pi_0 + p_S \pi_S + p_M \pi_M + p_L \pi_L}$$

3. Statistical Properties Snapshot Method

- Analytically approximated properties using numerical integration
- Effect size measure: Correlation coefficient
- 5,000 equally spaced cumulative probabilities given significance for original study (α=.025)
- 5,000 equally spaced cumulative probabilities for replication
- Converting probabilities to effect sizes: 5,000 x 5,000 = 25.000.000

3. Statistical Properties Snapshot Method

- Conditions:
  - ρ = 0; 0.1; 0.3; 0.5
  - Sample size (n_i): 31; 55; 96; 300; 1,000
  - Snapshots (ρ_S) = 0; 0.1; 0.3; 0.5
  - Snapshot and naïve method
- Outcome variables:
  - Expected value of posterior probability
  - Probability of strong evidence (π_x > .75 or Bayes Factor > 3)

4. Statistical Properties Snapshot Method

- Expected values of posterior probabilities:

		Snapshot method			
	n_i	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5
ρ=0	31	**0.466**	0.36	0.151	0.023
	55	**0.535**	0.375	0.089	0.002
	96	**0.601**	0.368	0.03	0
	300	**0.757**	0.243	0	0
	1,000	**0.948**	0.052	0	0

- Huge sample sizes (n_i~1,000) are required to distinguish 0 from small effect

3. Statistical Properties Snapshot Method

- Expected values of posterior probabilities (WRONG METHOD):

		Snapshot Naïve method			
	n_i	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5
ρ=0	31	**0.177**	0.336	0.411	0.076
	55	**0.212**	0.479	0.304	0.005
	96	**0.241**	0.648	0.112	0
	300	**0.338**	0.662	0	0
	1,000	**0.758**	0.242	0	0

- No correction for statistical significance → overestimation

3. Statistical Properties Snapshot Method

- Expected values of posterior probabilities:

	Snapshot method			
n_i	ρ=0	ρ=0.1	ρ=0.3	ρ=0.5
31	0.466	0.351	0.367	0.669
55	0.535	0.403	0.523	0.808
96	0.601	0.481	0.738	0.918
300	0.757	0.745	0.985	0.997
1,000	0.948	0.948	1	1

- Easier to distinguish medium and large effect

3. Statistical Properties Snapshot Method

- Probability of strong evidence (π_x > .75):

	Snapshot method			
n_i	ρ=0	ρ=0.1	ρ=0.3	ρ=0.5
31	0.04	0	0	0.498
55	0.142	0	0.115	0.732
96	0.291	0	0.645	0.895
300	0.641	0.625	0.982	0.997
1,000	0.935	0.933	1	1

- Large sample size needed for zero and small effect

3. Statistical Properties Snapshot Method

Conclusions:

- Not correcting for statistical significance (naïve method) is inappropriate
- Huge sample sizes are required to distinguish 0 from small effect
- Large sample sizes are required for medium and large effect

4. Application: RPP and EE-RP

- Initiatives to study the replicability of psychological and economic research
- RPP: Studies from JPSP, Psychological Science, and Journal of Experimental Psychology: 67 out of 100 studies were included
- EE-RP: Experimental research from the American Economic Review and Quarterly Journal of Economics: 16 out of 18 studies were included
- "High-powered" replication of a key effect

4. Application: RPP and EE-RP

- Probability of strong evidence (π_x > .75) using snapshot method:

	ρ_S				
	0	**0.1**	**0.3**	**0.5**	**Unknown**
EE-RP	0	0.062	0.312	0.438	0.188
RPP	0.134	0.030	0.045	0.164	0.627

- Conclusions:
  - Studied effects larger in EE-RP than in RPP
  - Only few studies have strong evidence for zero effect in RPP (13.4%)
  - Often not enough information for determining magnitude of effect size in RPP (62.7%)

5. Conclusion and discussion

(1) Methods *should* take statistical significance of original study into account

(2) We developed such a method within a Bayesian framework

(3) Need huge sample sizes (n_i~1,000) to distinguish 0 from small effect

→*With current sample sizes in psychology, one or two studies is not sufficient to accurately evaluate effect size*

(4) Application of method to RPP and EE-RP:

→ *Often not sufficient information for determining magnitude of effect size*

→ *Studied effects larger in EE-RP than RPP*

5. Conclusion and discussion

- R code for snapshot method in "puniform" package and web application: https://rvanaert.shinyapps.io/snapshot/
- Determining sample size of replication with snapshot method akin to computing required sample size with power analysis
- Intervals of effect sizes instead of discrete values as snapshots
- Future research:
  - Extend method such that it can deal with multiple original studies and replications

Thank you for your attention