Quantitative Methods for The Replication Crisis: "The New Statistics" …and ”Some Even Newer Statistics"

Brenden Tervo-Clemmens, Ph.D.

Massachusetts General Hospital, Harvard Medical School

btervo-clemmens@mgh.harvard.edu

code for simulations and animations:
github.com/tervoclemmensb/newstatsdemo

Quantitative Expertise...

Quantitative Expertise...

"Choose the appropriate statistical model"

"You won't get grants or pubs if you don't use [fancy method]"

"The best analysis is a good research design"

"All models are wrong, some are useful!"

Quantitative Expertise...
Matching Methods to Questions and Data

"Choose the appropriate statistical model"

"You won't get grants or pubs if you don't use [fancy method]"

"The best analysis is a good research design"

"All models are wrong, some are useful!"

Quantitative Expertise...
Matching Methods to Questions and Data

Longitudinal data → growth curve/mixed effects

High dimensional data/correlated measures → latent variable analysis

Optimizing prediction → regularization/machine learning

Concerns of reproducibility ?

Quantitative Expertise...
Matching Methods to Questions and Data

- *Longitudinal data → growth curve/mixed effects*
- *High dimensional data/correlated measures → latent variable analysis*
- *Optimizing prediction → regularization/machine learning*

Concerns of reproducibility

Quantitative Methods for The Replication Crisis: "The New Statistics" …and "Some Even Newer Statistics"

Outline

1. Background and quantitative foundations of the reproducibility crisis.

2. “The New Statistics” to address these challenges.

3. The “Even Newer Statistics” and bringing across quantitative areas.

Outline

1. Background and quantitative foundations of the reproducibility crisis.
2. "The New Statistics" to address these challenges.
3. The "Even Newer Statistics" and bringing across quantitative areas.

Psychological Science in a Crisis of Crises

- *Replication Crisis*[1]


RESEARCH ARTICLE

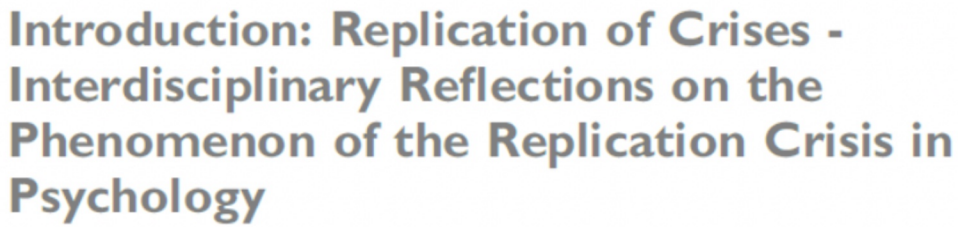
Estimating the reproducibility of psychological science

Open Science Collaboration*,†

- *Generalizability Crisis*[2]


The generalizability crisis

- *Replication Crises/A Crisis of Crises*[3]

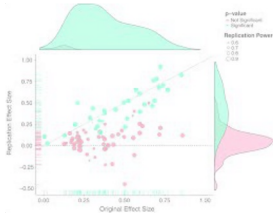

Introduction: Replication of Crises - Interdisciplinary Reflections on the Phenomenon of the Replication Crisis in Psychology

[1]*Open Science Collab. 2015, Science;* [2]*Yarkoni, 2020, Behavioral and Brain Sciences;*
[3]*Malich & Munafo 2022, Review of General Psychology*

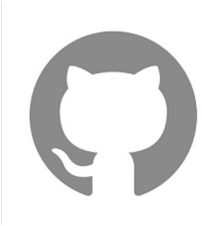
“The New Tools” for The Replication Crisis

- *Pre-registration*

- *Direct replications*



- *Open data & code*



- *Reporting checklists*

"The New Tools" for Experimenter Bias and Experimenter Degrees of Freedom

- *Pre-registration*

- *Direct replications*

- *Open data & code*

- *Reporting checklists*

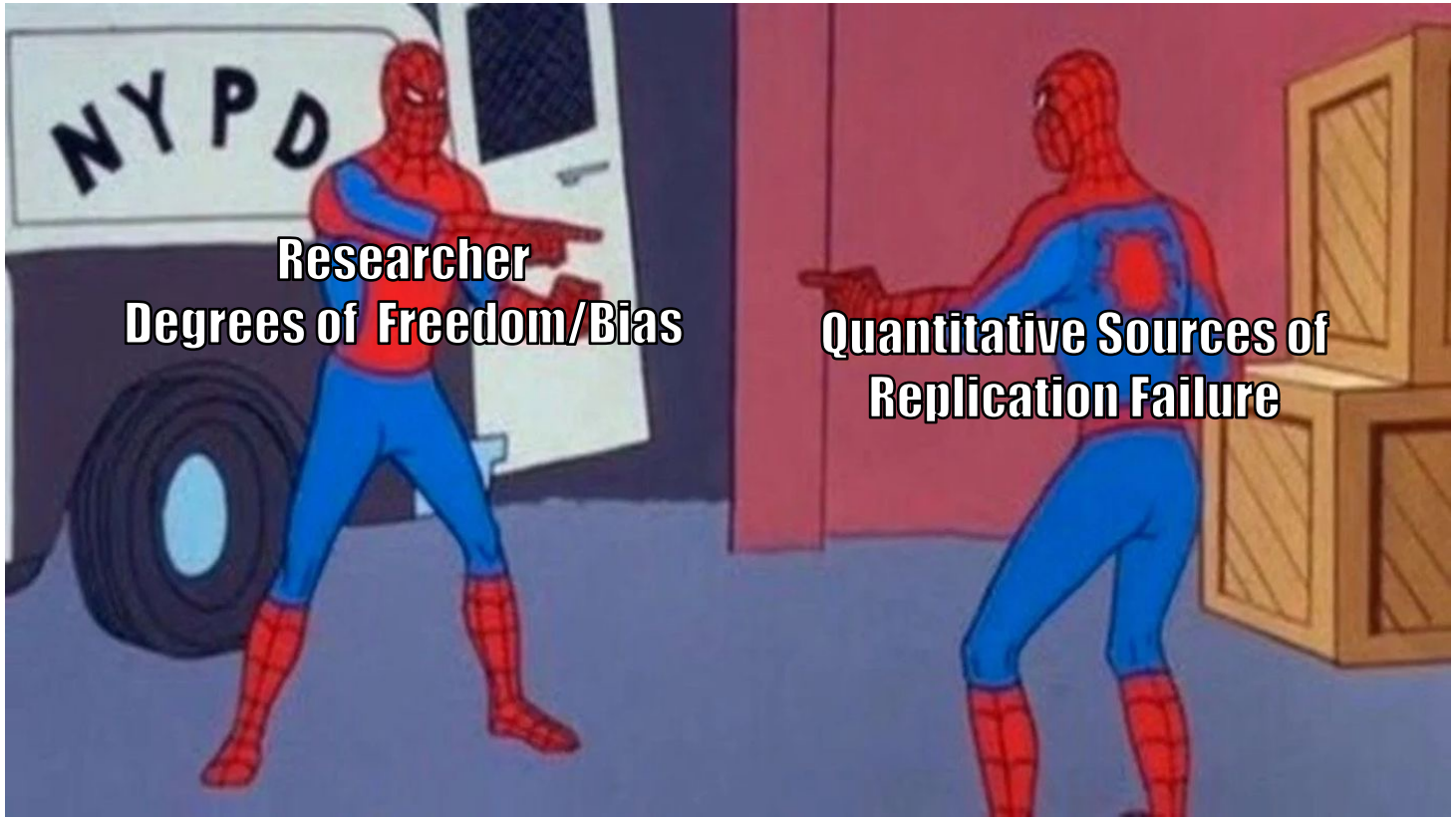
Artist: Benita Epstein

Quantitative Insights into Reproducibility Challenges

- *Psychometrics, power, and research design.*

Quantitative Insights into Reproducibility Challenges

- *Psychometrics, power, and research design.*



Quantitative Insights into Reproducibility Challenges

Quantitative Insights into Reproducibility Challenges

- *[Re]creating a reproducibility crisis….*

Quantitative Insights into Reproducibility Challenges

- *[Re]creating a reproducibility crisis….*
    - Sampling variability, selective reporting, and statistical power

Quantitative Insights into Reproducibility Challenges

- *[Re]creating a reproducibility crisis….key references:*

The New Statistics: Why and How

Geoff Cumming (Psychological science, 2014)

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò (Nature Reviews Neuroscience, 2017)

At what sample size do correlations stabilize?

Felix D. Schönbrodt [a,*], Marco Perugini [b] (Journal of Research in Personality, 2013)

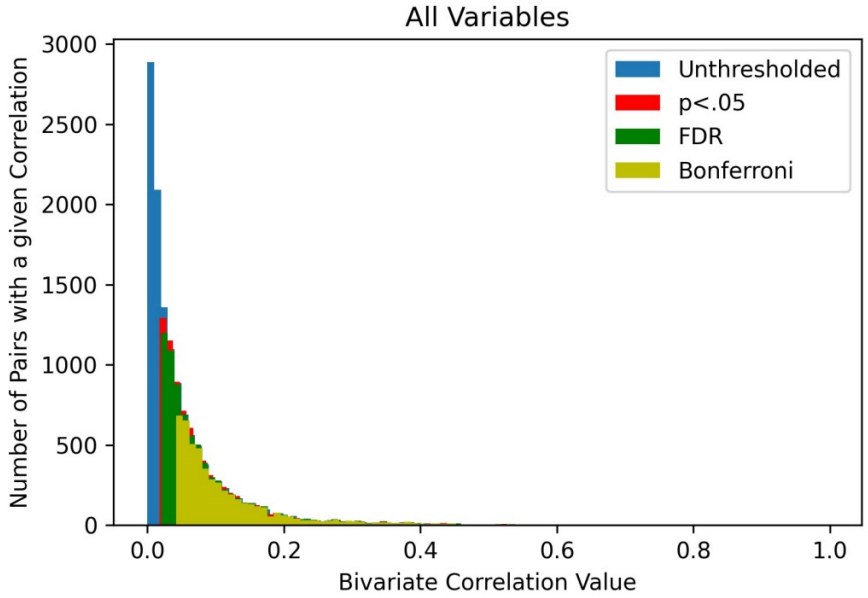
BTC code for simulations and animations:
github.com/tervoclemmensb/newstatsdemo

Quantitative Insights into Reproducibility Challenges

- *[Re]creating a reproducibility crisis….*
    - How big are the effects in psychology anyway?

Quantitative Insights into Reproducibility Challenges

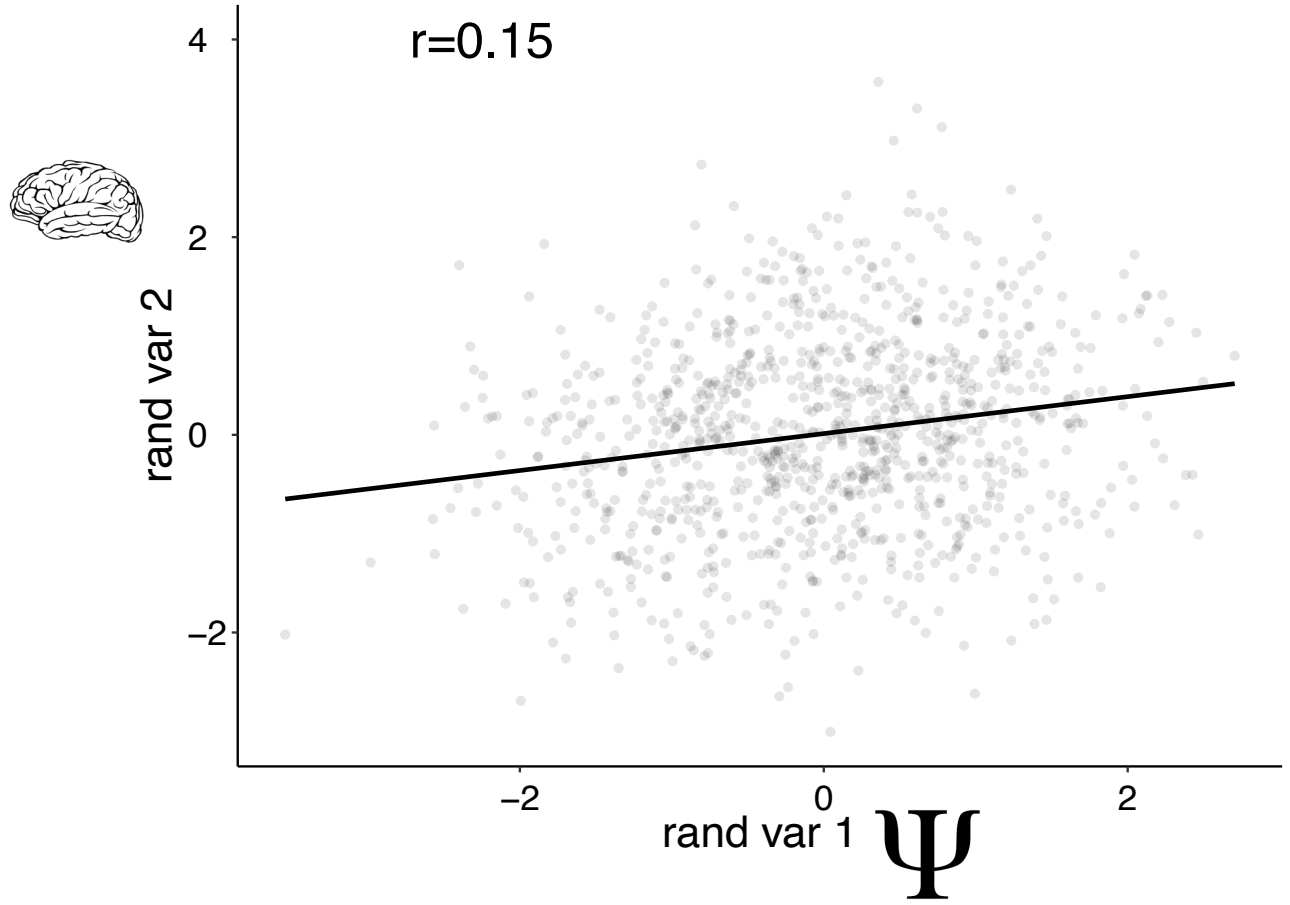
- *[Re]creating a reproducibility crisis….*
  - How big are the effects in psychology anyway?
  - Smaller than we once thought…
  - r ~.15
  - Consult your literature!



Owens et al., 2021 (Figure); Marek* Tervo-Clemmens* et al., 2022; Funder & Orzer, 2019

[Re]creating a reproducibility crisis….

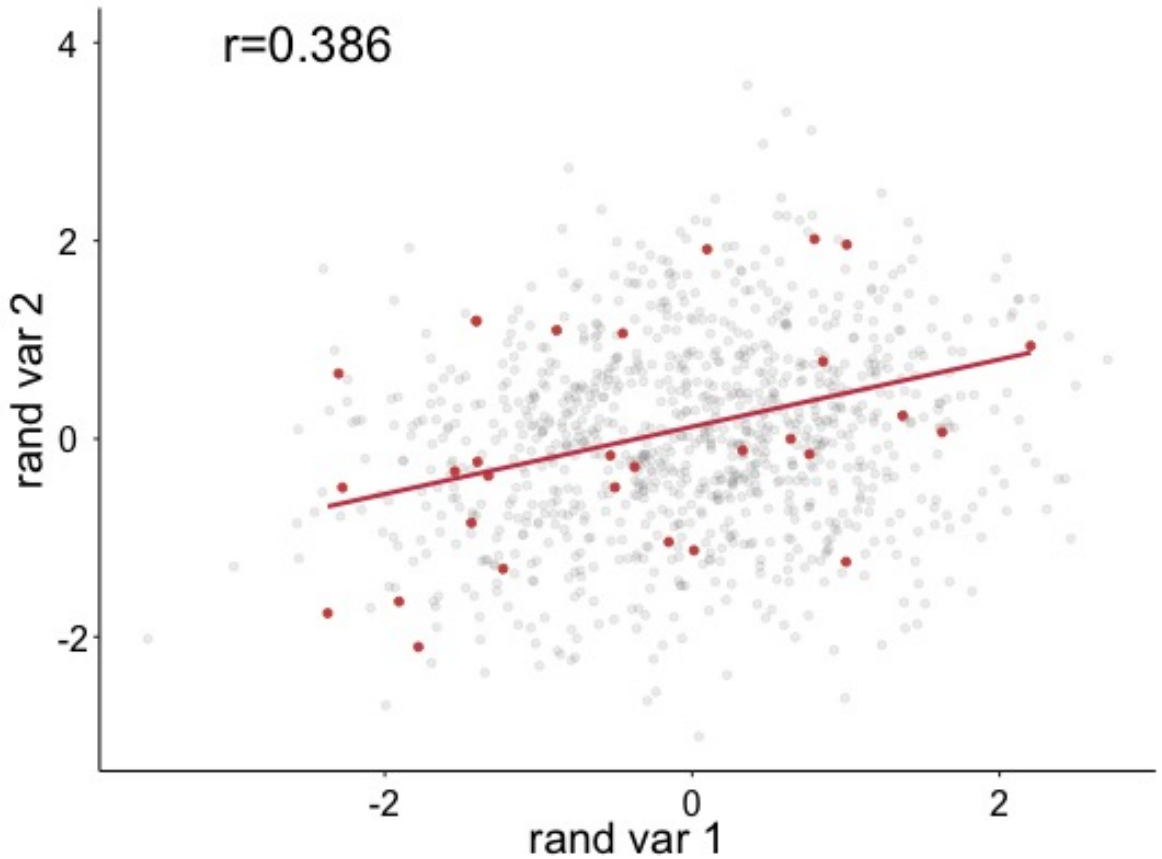
- *Simulated Population Effect*



Population of 1,000
Population correlation = .15

[Re]creating a reproducibility crisis….

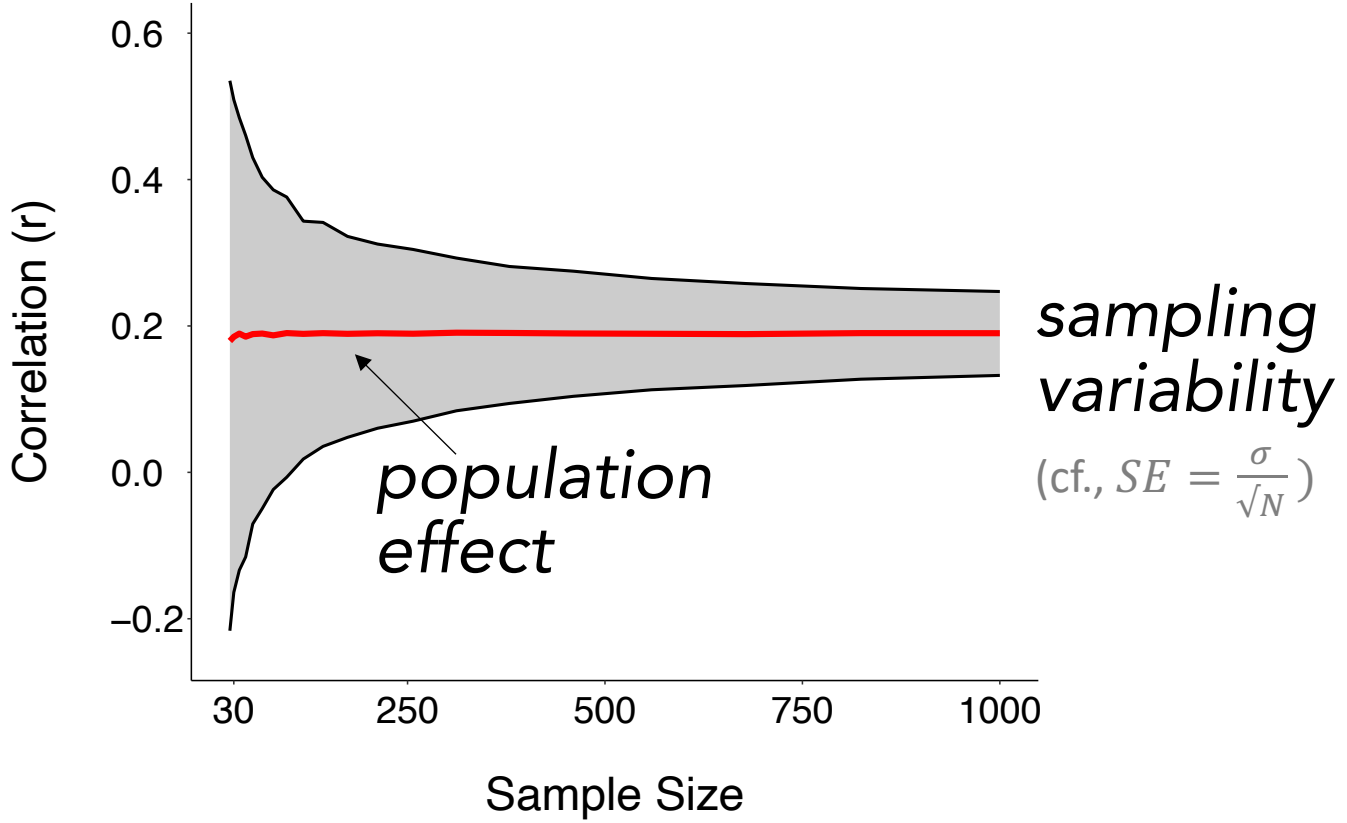
- *Simulated Samples*



Samples of n=30
Population correlation = .15

[Re]creating a reproducibility crisis….

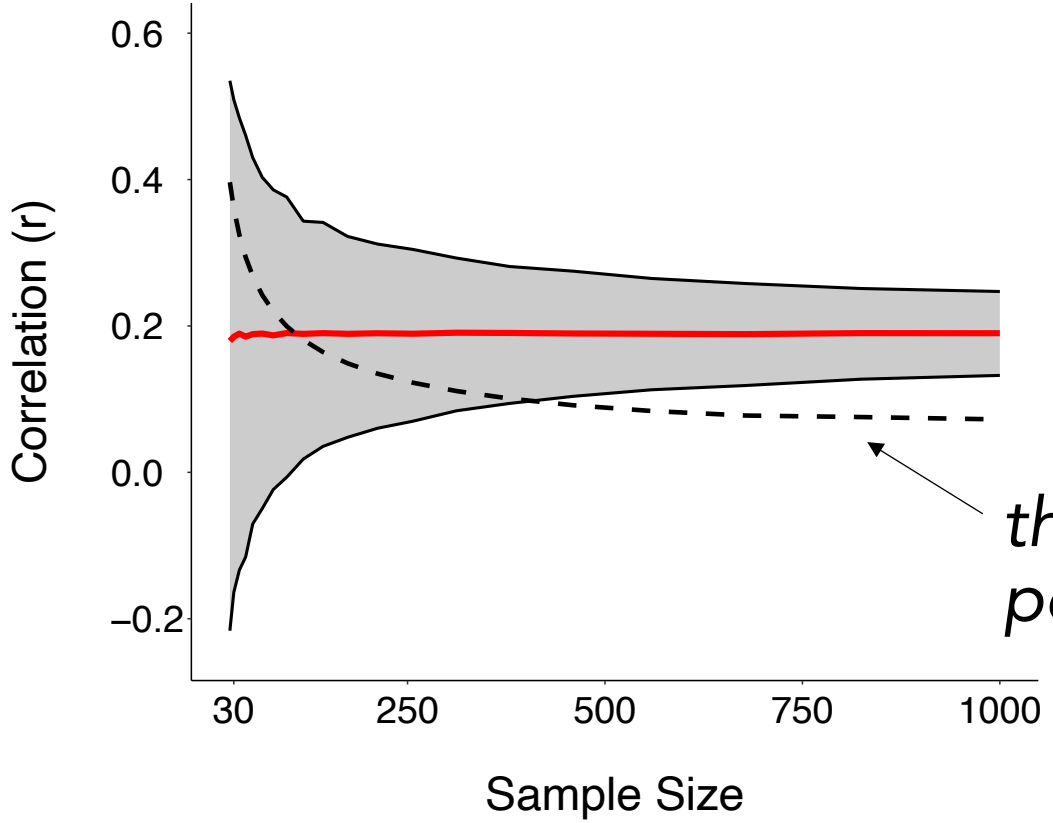
- *Correlation by Sample Size*



cf., Schönbrodt & Perugini , 2013

[Re]creating a reproducibility crisis….

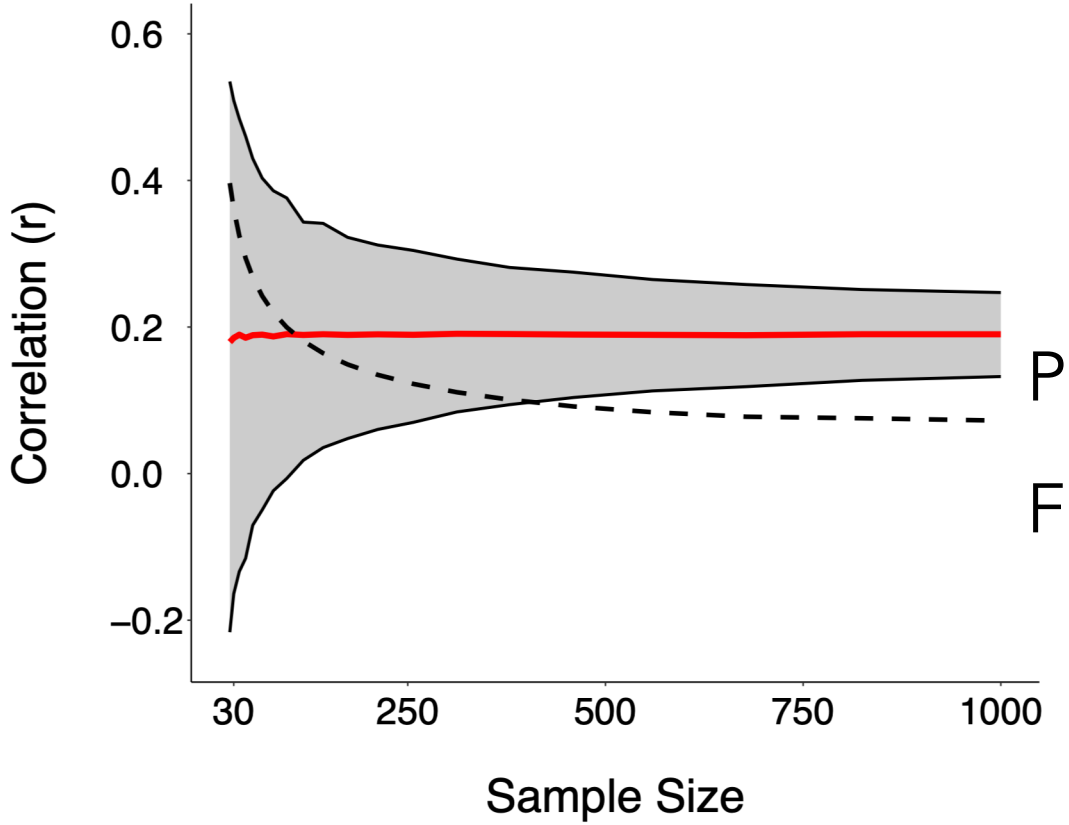
- *Correlation by Sample Size*



threshold for significant positive correlation

[Re]creating a reproducibility crisis….

- *Publication Bias*

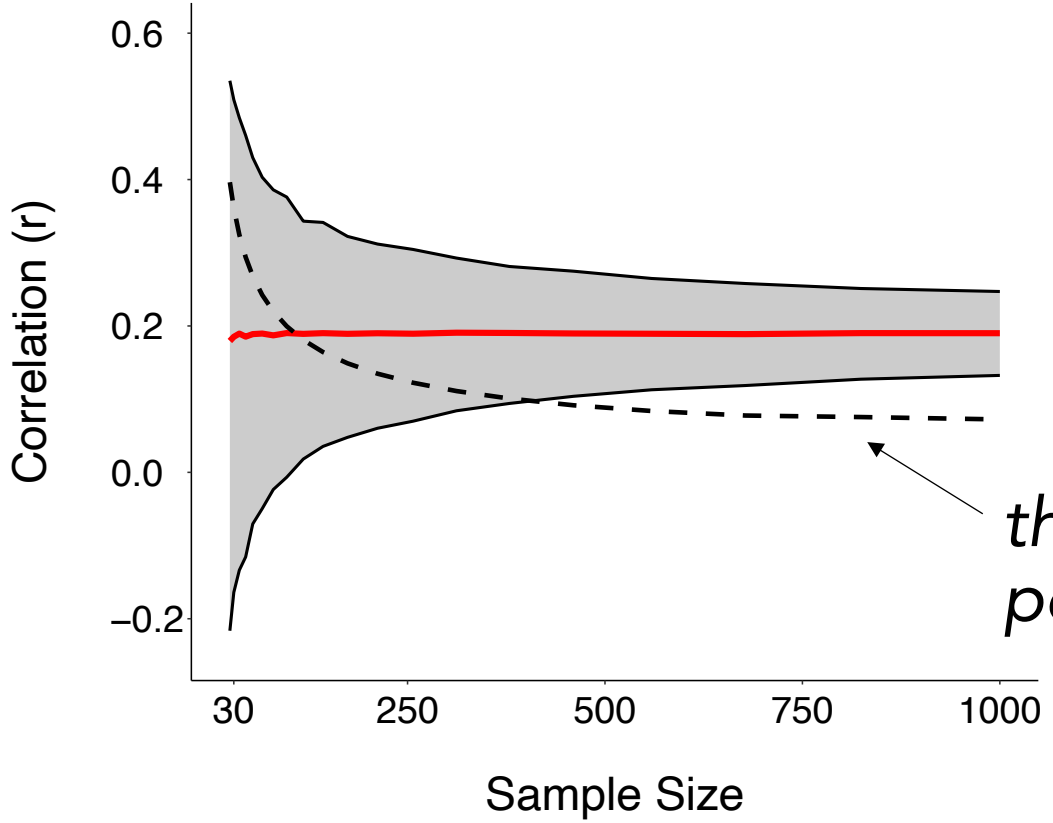


Publish significant result ☺

File drawer non sig. result ☹

[Re]creating a reproducibility crisis….

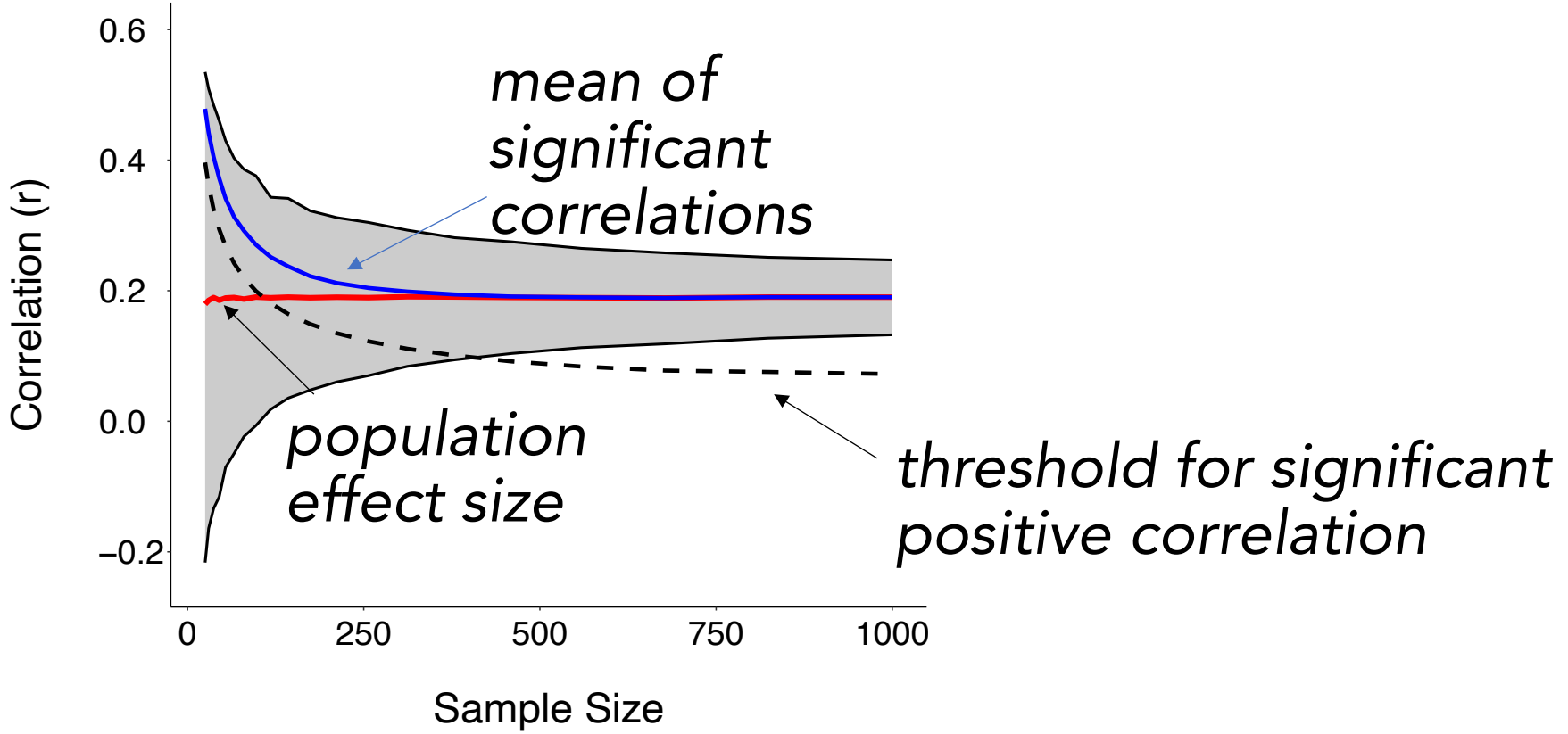
- *Correlation by Sample Size*



threshold for significant positive correlation

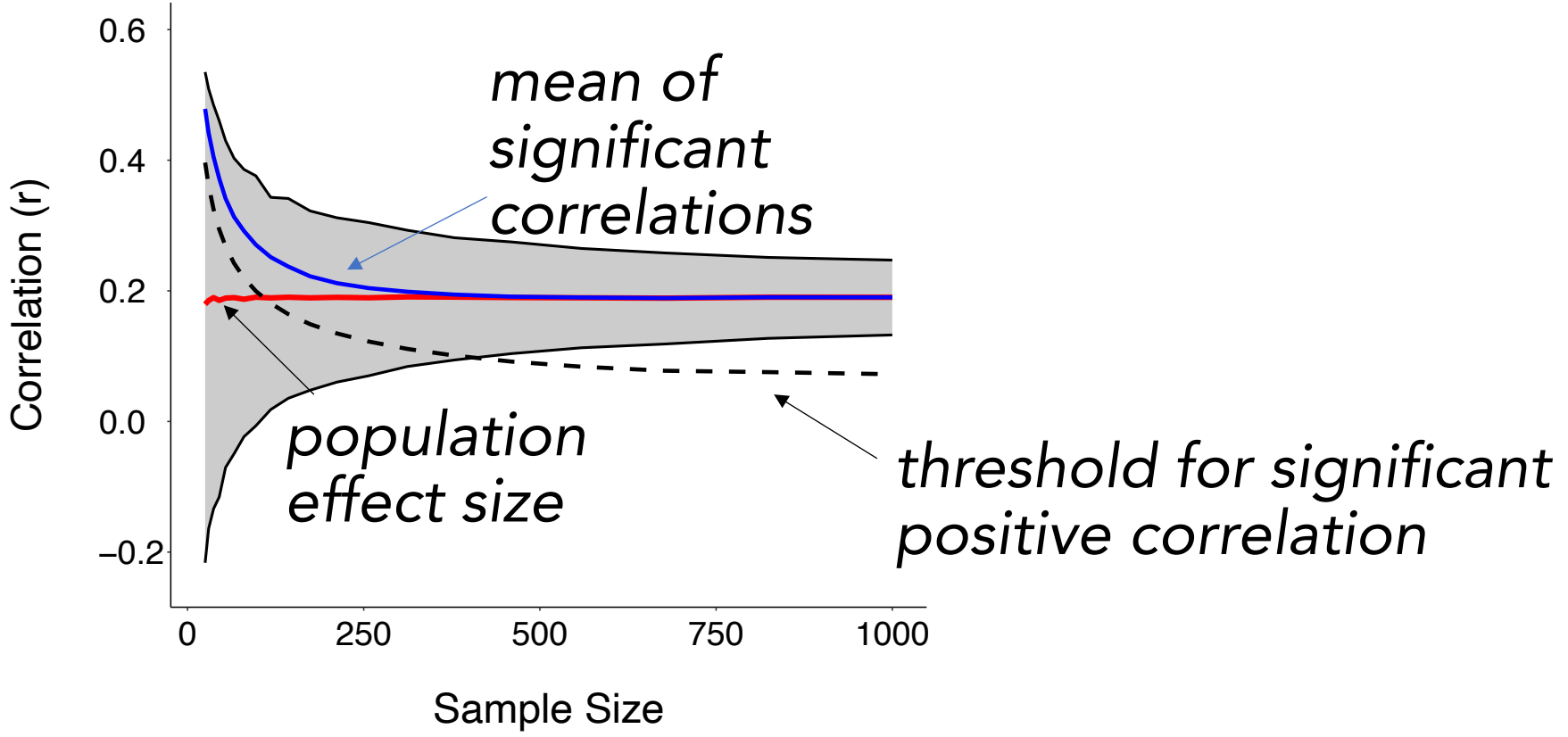
[Re]creating a reproducibility crisis….

- *Correlation by Sample Size*



[Re]creating a reproducibility crisis….

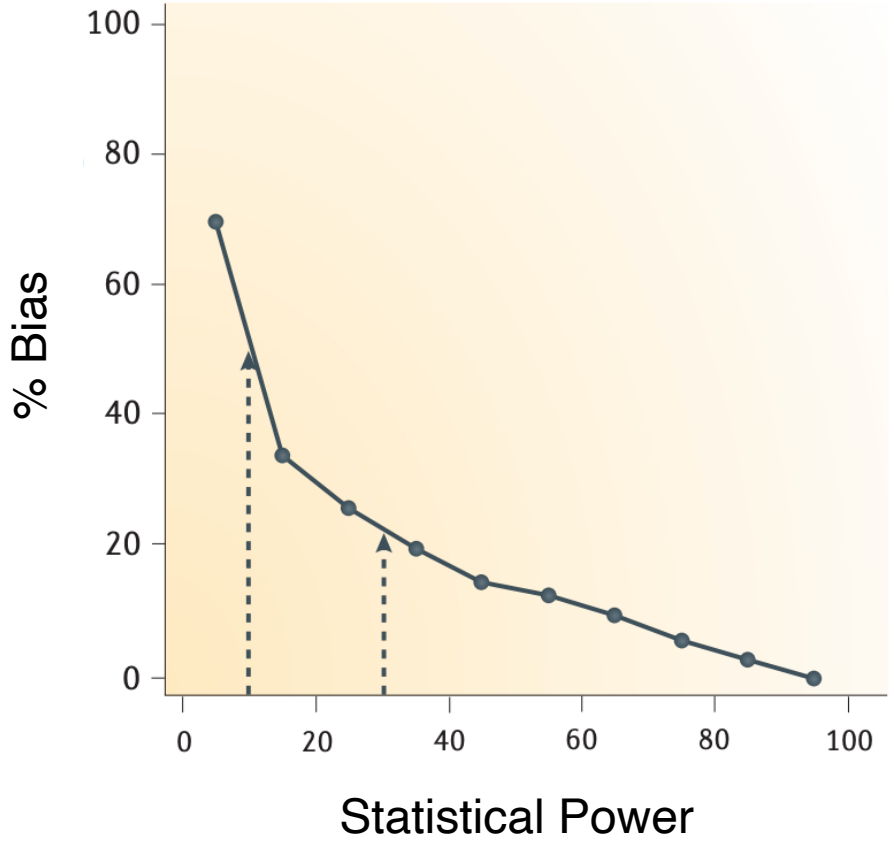
- *Effect size inflation; Winner's curse*



cf., Button et al., 2013

[Re]creating a reproducibility crisis….

- *Effect size inflation; Winner's curse*



Inflation reflects underpowered studies.

Button et al., 2013

Quantitative Insights into the Reproducibility Crisis

*Expected variation across samples, *sampling variability*, and selective reporting of significant effects, *publication bias*, (particularly in designs with low statistical power]) provide a quantitative basis of challenges to reproducibility.*

Outline

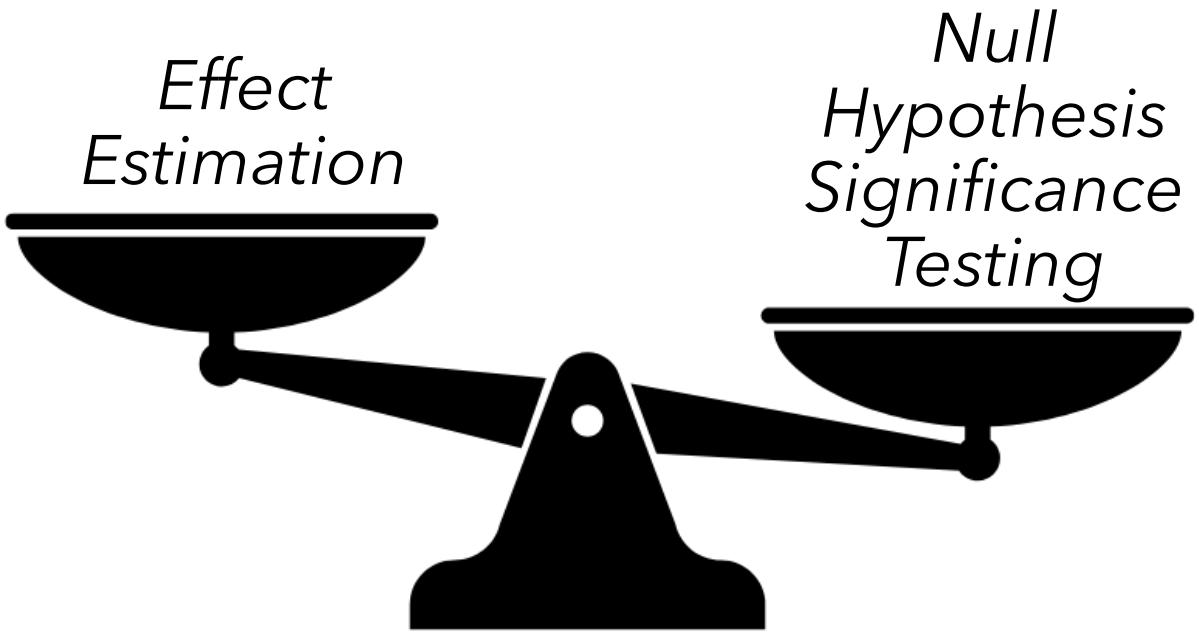
1. Background and quantitative foundations of the reproducibility crisis.
2. "The New Statistics"[1] to address these challenges.
3. The "Even Newer Statistics" and bringing across quantitative areas.

[1]*Cumming 2014, Psychological Science*

An overarching goal:

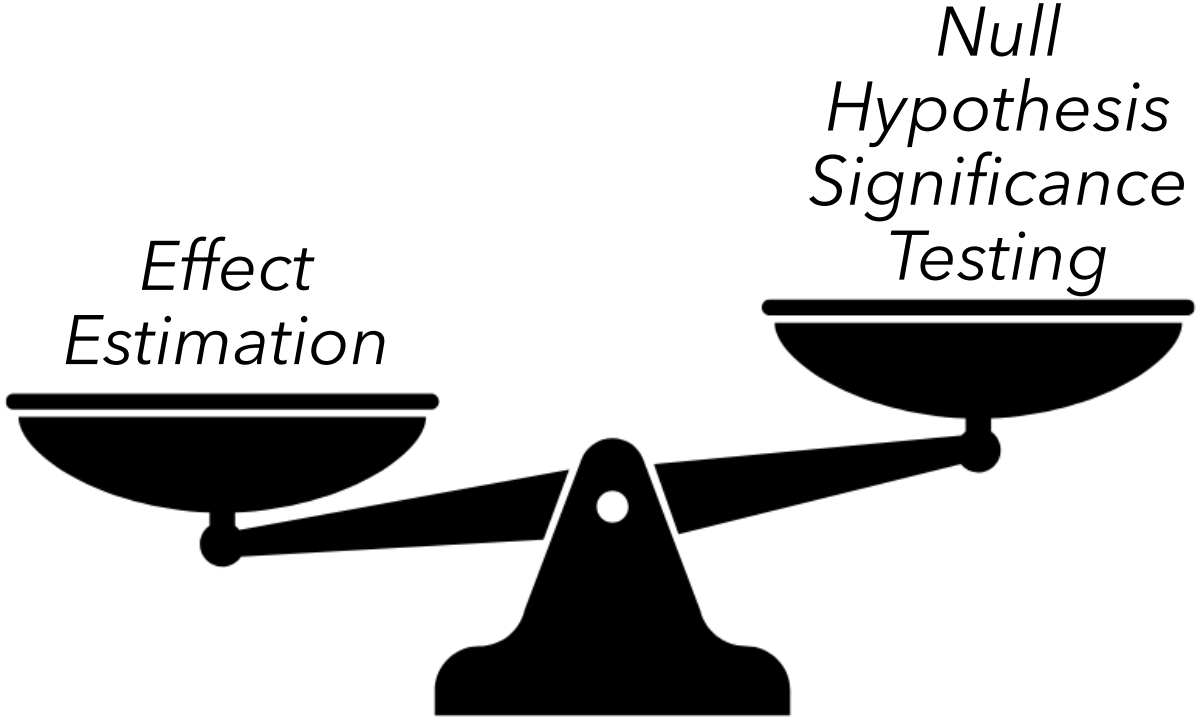
Address sampling variability, publication bias, and misleading effects via “The Winner’s Curse”

Quantitative Insights into Reproducibility Challenges



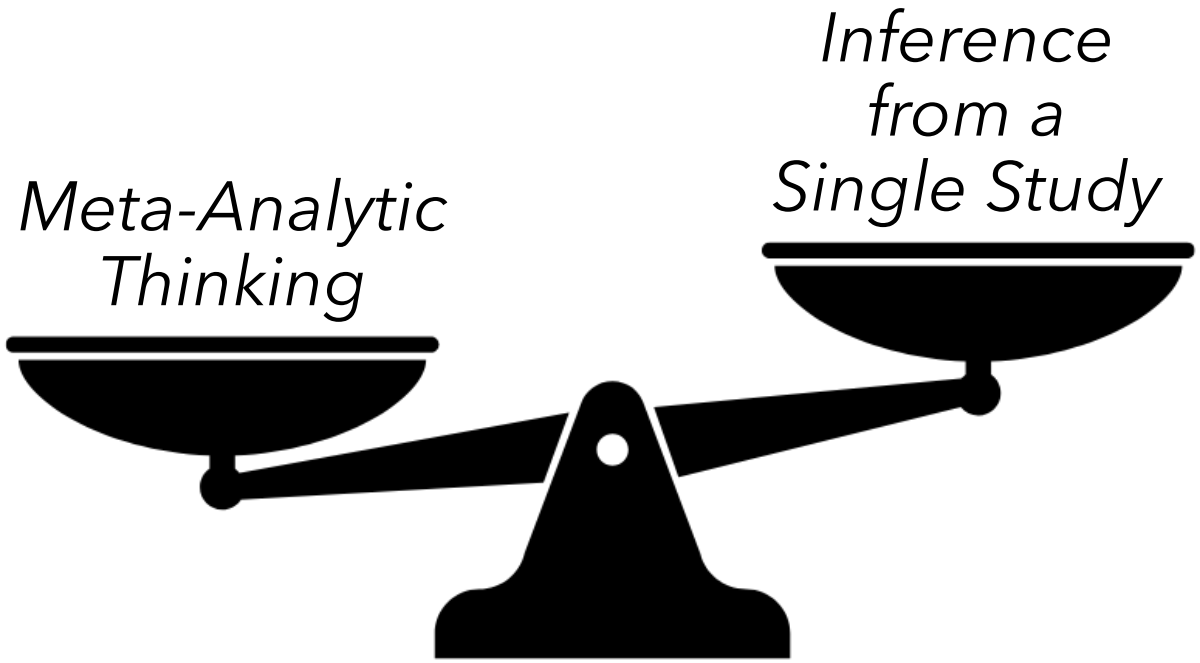
Cumming 2014, Psychological Science; Ioannidis 2005, PLoS Medicine

The New Statistics



Cumming 2014, Psychological Science; Ioannidis 2005, PLoS Medicine

The New Statistics

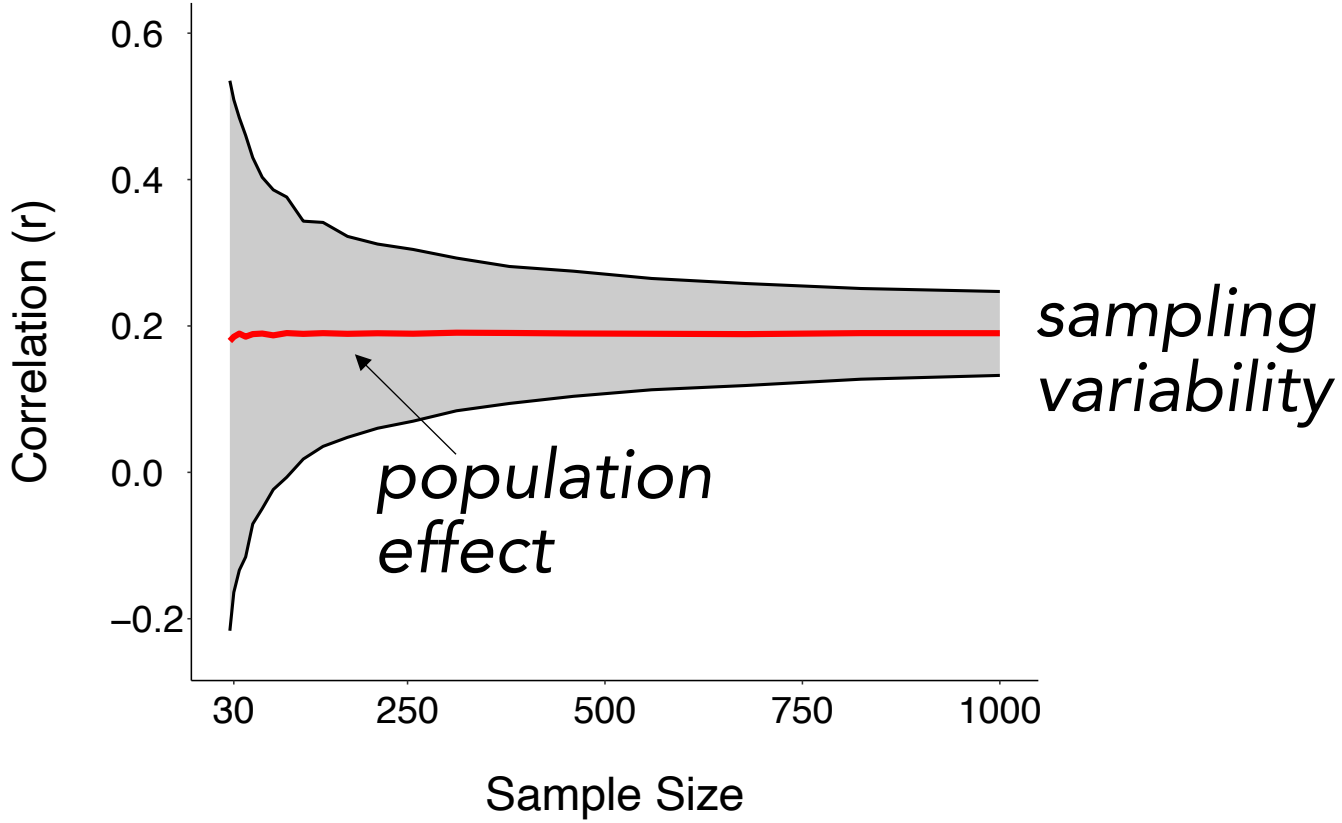


Cumming 2014, Psychological Science

The New Statistics: Meta-analytic Thinking

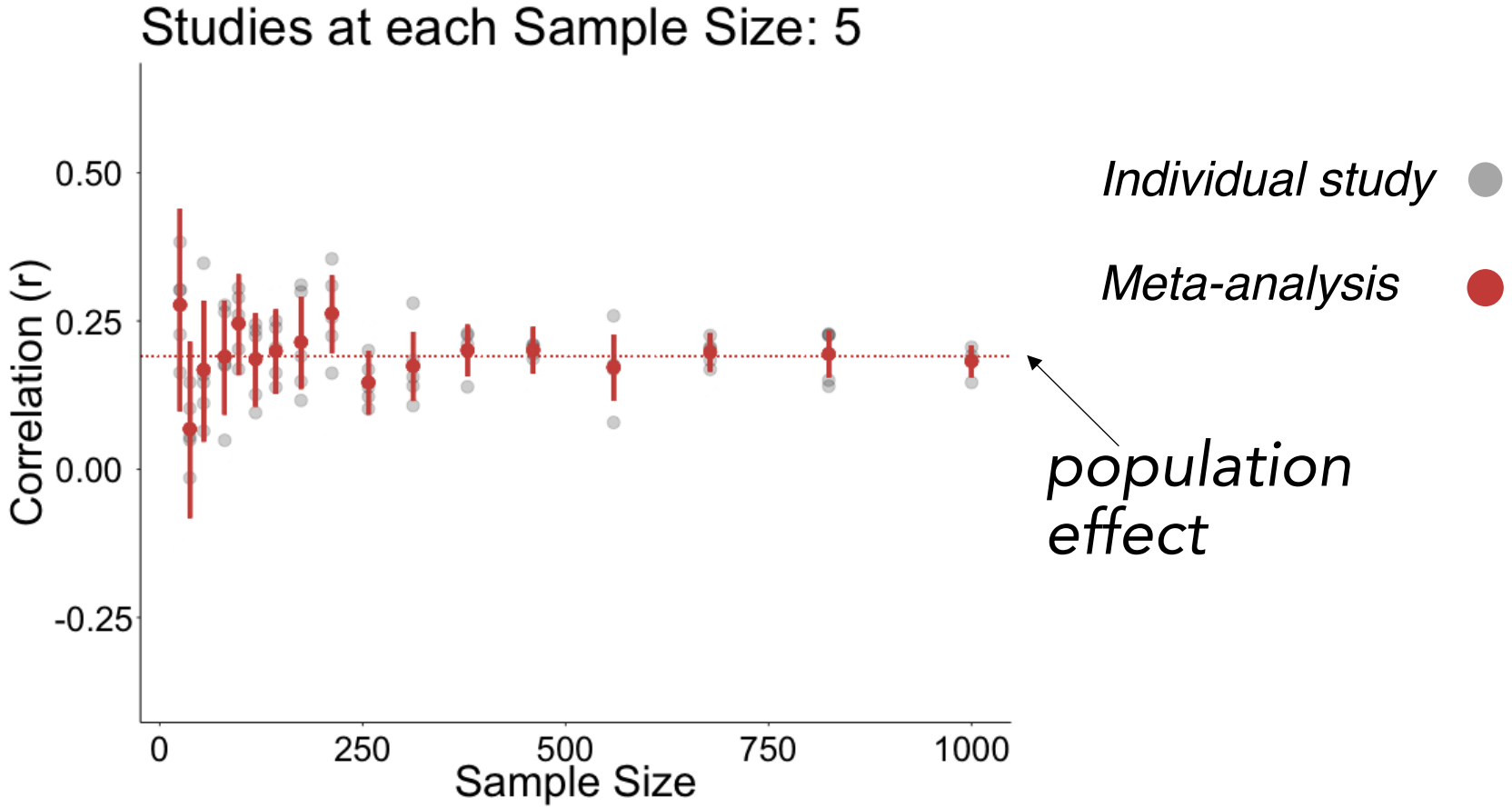
The New Statistics: Meta-analytic Thinking

- *Correlation by Sample Size*



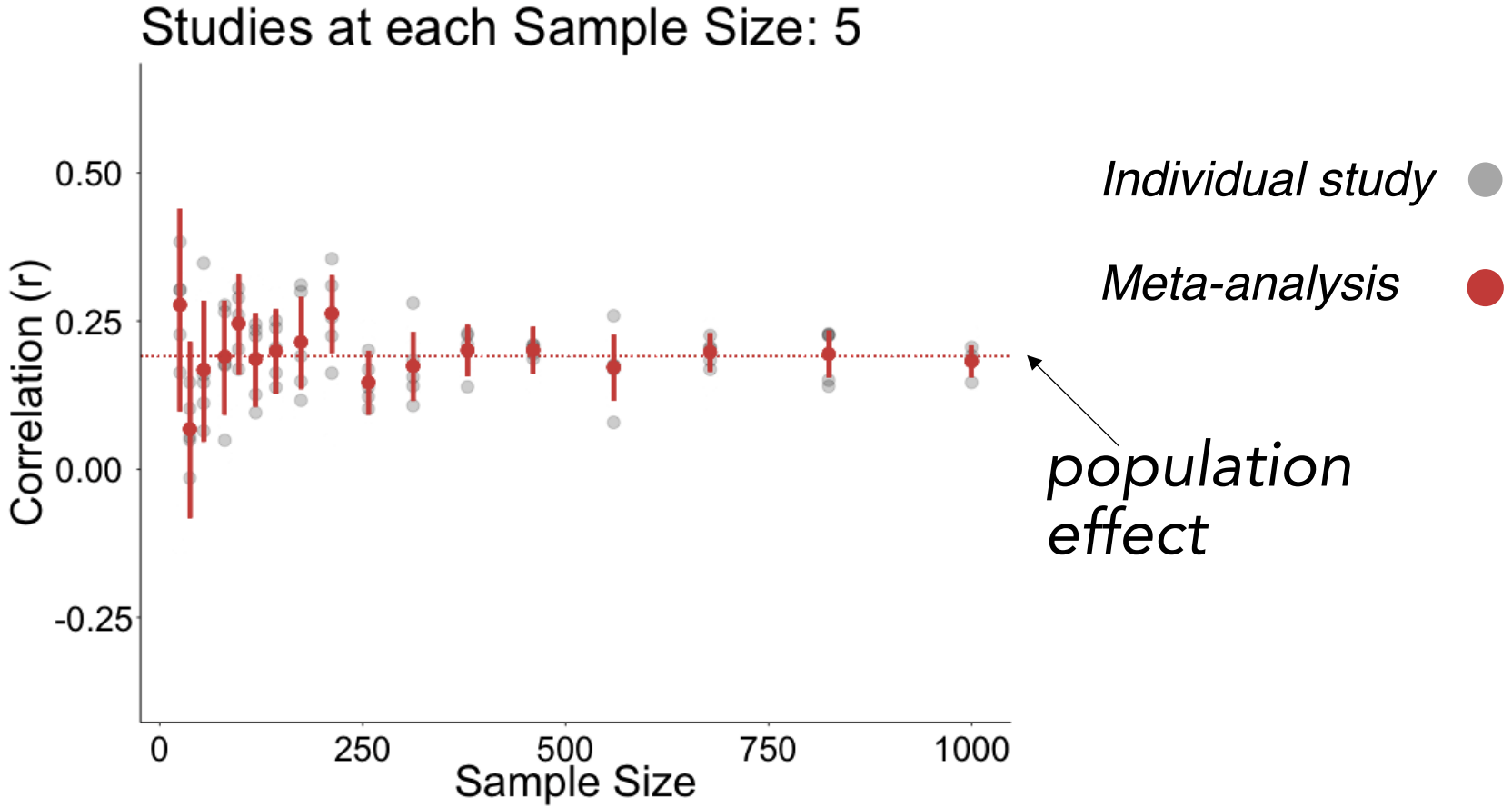
The New Statistics: Meta-analytic Thinking

- *Correlation by Sample Size with Meta-Analysis*



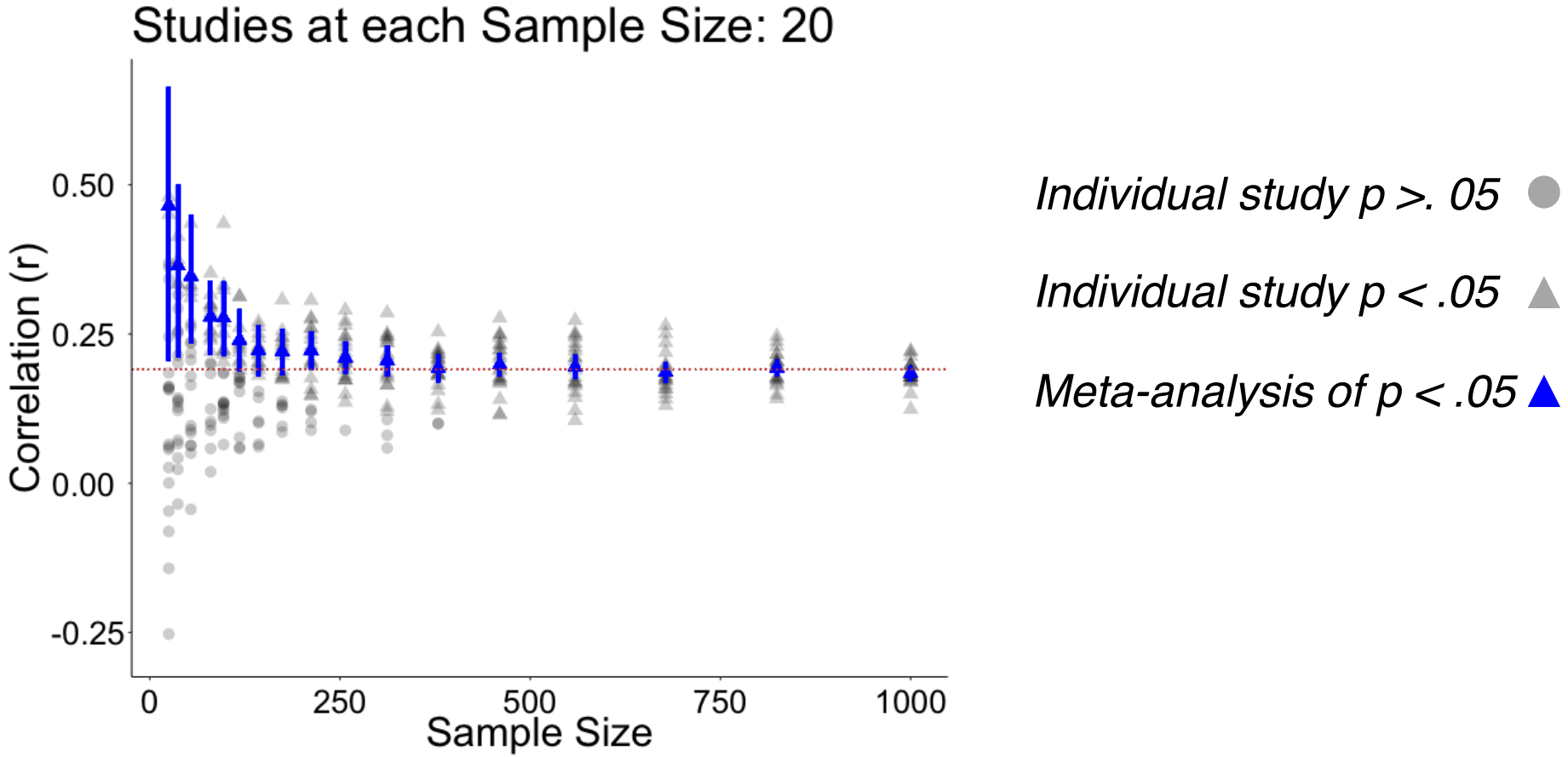
The New Statistics: Meta-analytic Thinking

- *Correlation by Sample Size with Meta-Analysis*



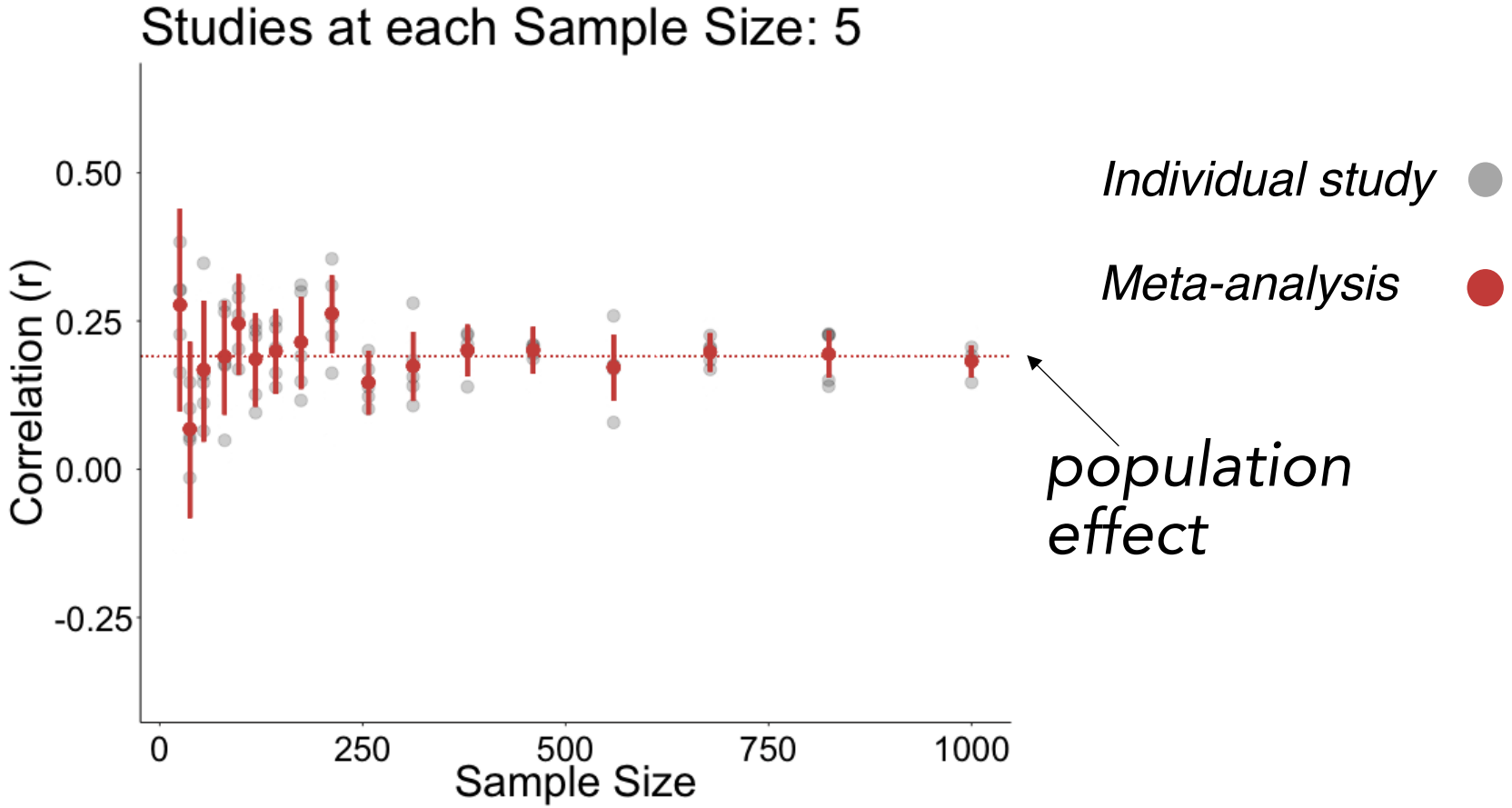
The New Statistics: Meta-analytic Thinking

- *Meta-analysis reflect the "Winner's curse" with publication bias.*



The New Statistics: Meta-analytic Thinking

- *Correlation by Sample Size with Meta-Analysis*



The New Statistics: Meta-analytic Thinking

Unbiased meta-analysis (effect sizes reported without publication bias) with enough studies recover the true, population effect, even with relatively small sample sizes.

[1]*Cumming 2014, Psychological Science*

The New Statistics: Meta-analytic Thinking

Unbiased meta-analysis (effect sizes reported without publication bias) with enough studies recover the true, population effect, even with relatively small sample sizes.

Meta-analytic thinking: *"Any one study is most likely contributing rather than determining; it needs to be considered alongside any comparable past studies and with the assumption that future studies will build on its contribution. "*[1]

[1]*Cumming 2014, Psychological Science*

The New Statistics: Meta-analytic Thinking

Unbiased meta-analysis (effect sizes reported without publication bias) with enough studies recover the true, population effect, even with relatively small sample sizes.

Meta-analytic thinking: *"Any one study is most likely contributing rather than determining; it needs to be considered alongside any comparable past studies and with the assumption that future studies will build on its contribution. "*[1]

Highlights the importance of large initiatives and our contribution as individual analysis to publish null results and report effect sizes!

[1]*Cumming 2014, Psychological Science*

The New Statistics: Meta-analytic Thinking

Unbiased meta-analysis (effect sizes reported without publication bias) with enough studies recover the true, population effect, even with relatively small sample sizes.

Meta-analytic thinking: *"Any one study is most likely contributing rather than determining; it needs to be considered alongside any comparable past studies and with the assumption that future studies will build on its contribution. "*[1]

Highlights the importance of large initiatives and our contribution as individual analysis to publish null results and report effect sizes!

How we read the literature and make broad judgements about theory.

[1]*Cumming 2014, Psychological Science*

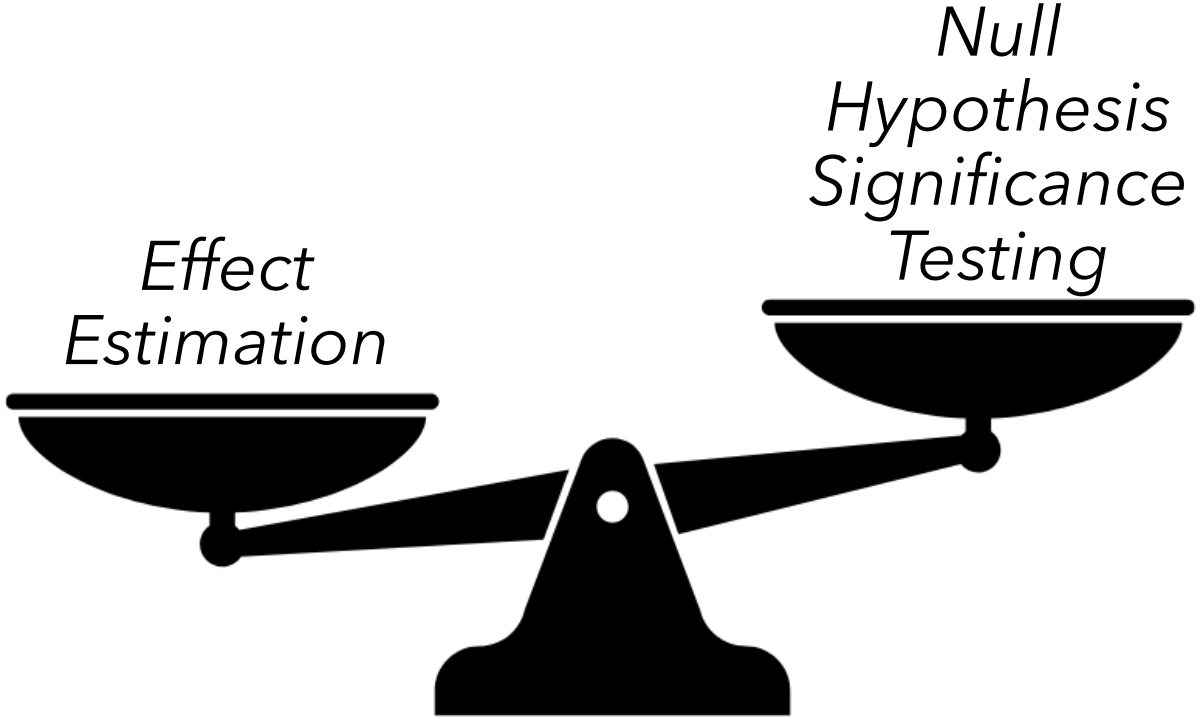
The New Statistics: Meta-analytic Thinking

Additional references for meta-analysis…..meta-analytic "thinking" is important but no substitute for meta–analytic details and "doing"

Balduzzi, Rücker, & Schwarzer, 2019, How to perform a meta-analysis with R: a practical tutorial, *BMJ Ment Health*
Hedges & Olkin 2014, Statistical Methods for Meta-Analysis (Academic Press)

“Meta-analytic thinking is great,
but what can I do better in an individual study?”

The New Statistics

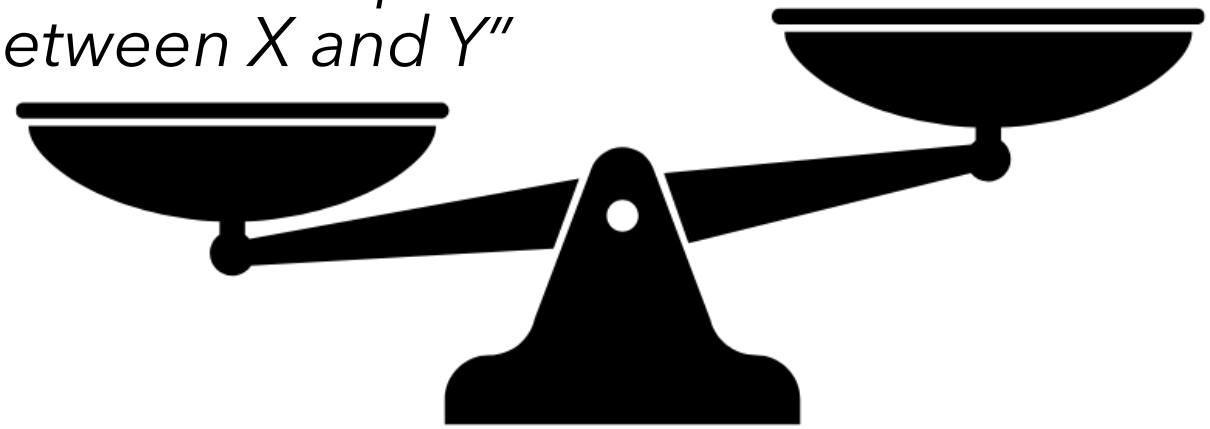


Cumming 2014, Psychological Science; Ioannidis 2005, PLoS Medicine

The New Statistics

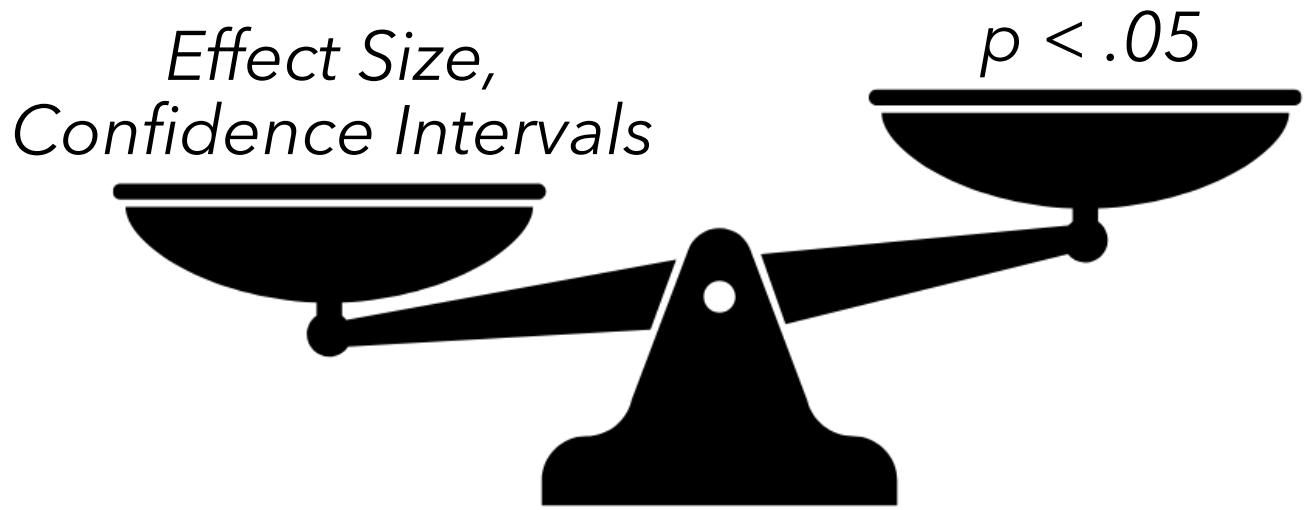
"The nature of the relationship between X and Y"

"Whether there is/isn't a relationship between X and Y"



Cumming 2014, Psychological Science

The New Statistics



Cumming 2014, Psychological Science

The New Statistics: Effect Sizes and Confidence Intervals

Effect size: value measuring the strength of the relationship/difference between variables (e.g., correlation).

Dimensional interpretation of a studied effect vs. binary $p < .05$

The New Statistics: Effect Sizes and Confidence Intervals

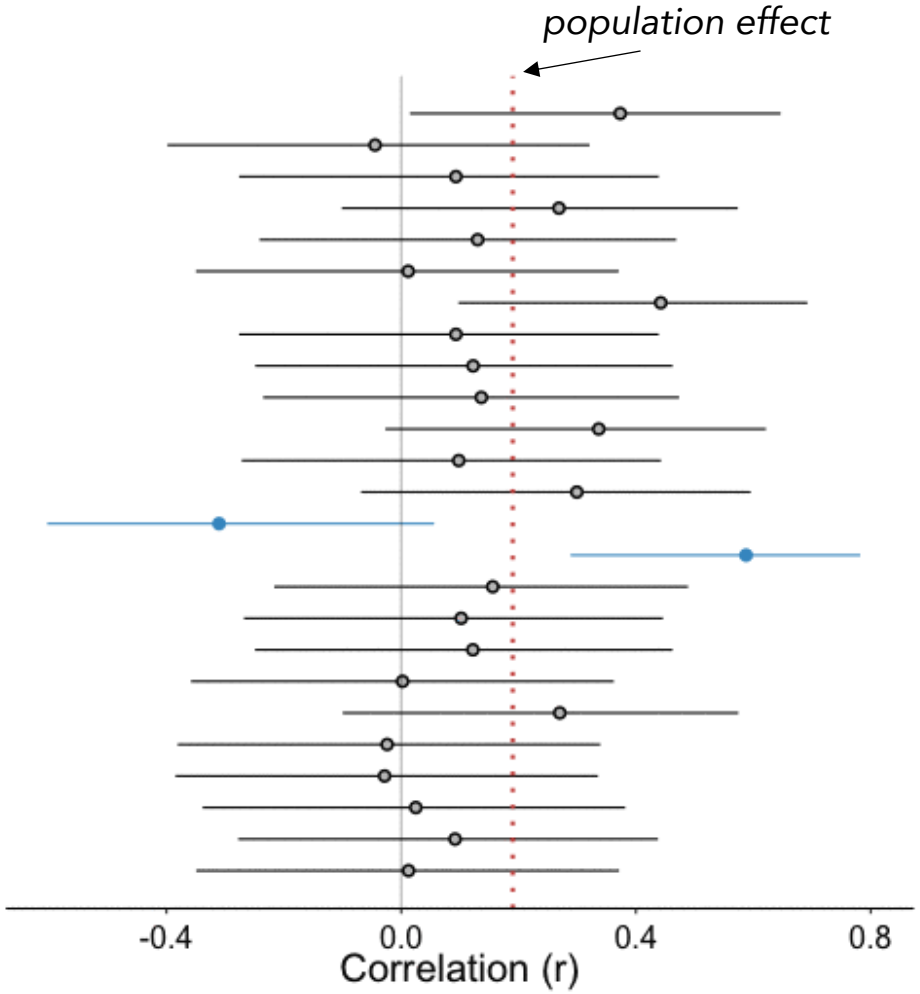
Effect size: value measuring the strength of the relationship between two variables (e.g., correlation).

Dimensional interpretation of a studied effect vs. binary $p < .05$

Confidence interval: above and below an effect size/point estimate and estimates other possible values of the effect size.

Uncertainty estimates for a given study that support reproducibility

The New Statistics: Effect Sizes and Confidence Intervals



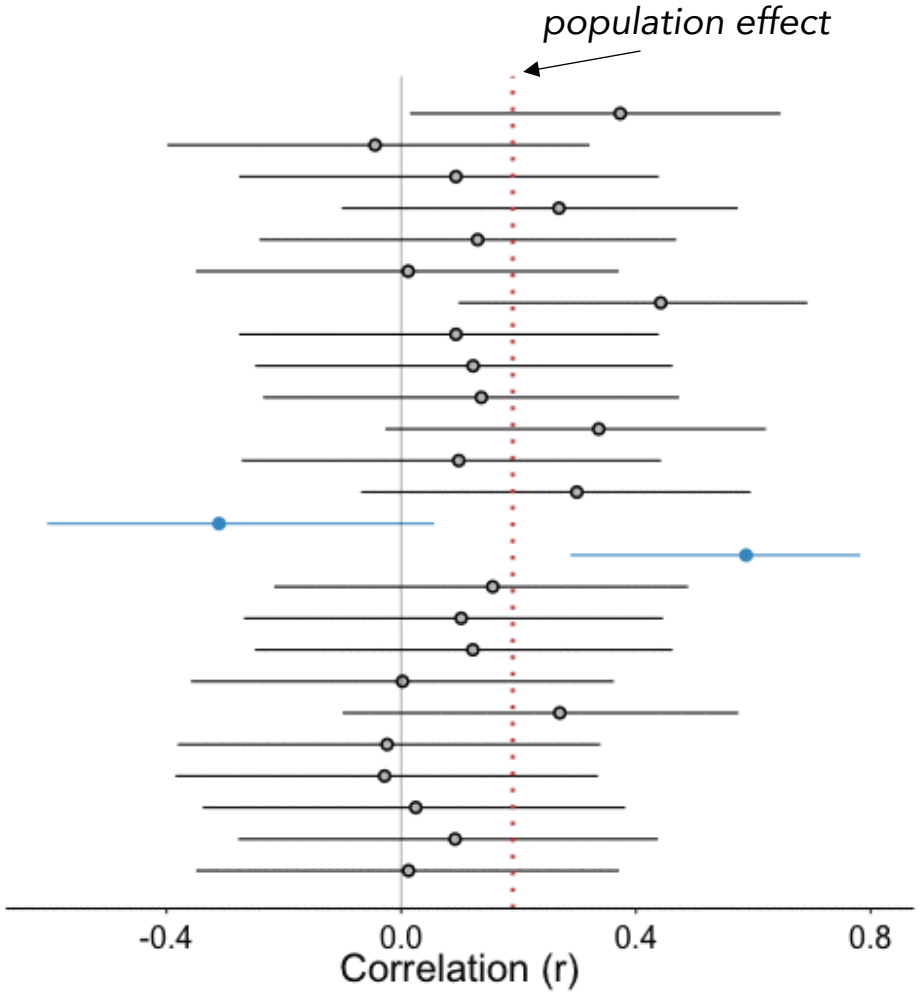
Correlations with 95% Confidence Intervals from 25 studies (n=30).

CI:

includes Pop. r

does not include Pop. r

The New Statistics: Effect Sizes and Confidence Intervals

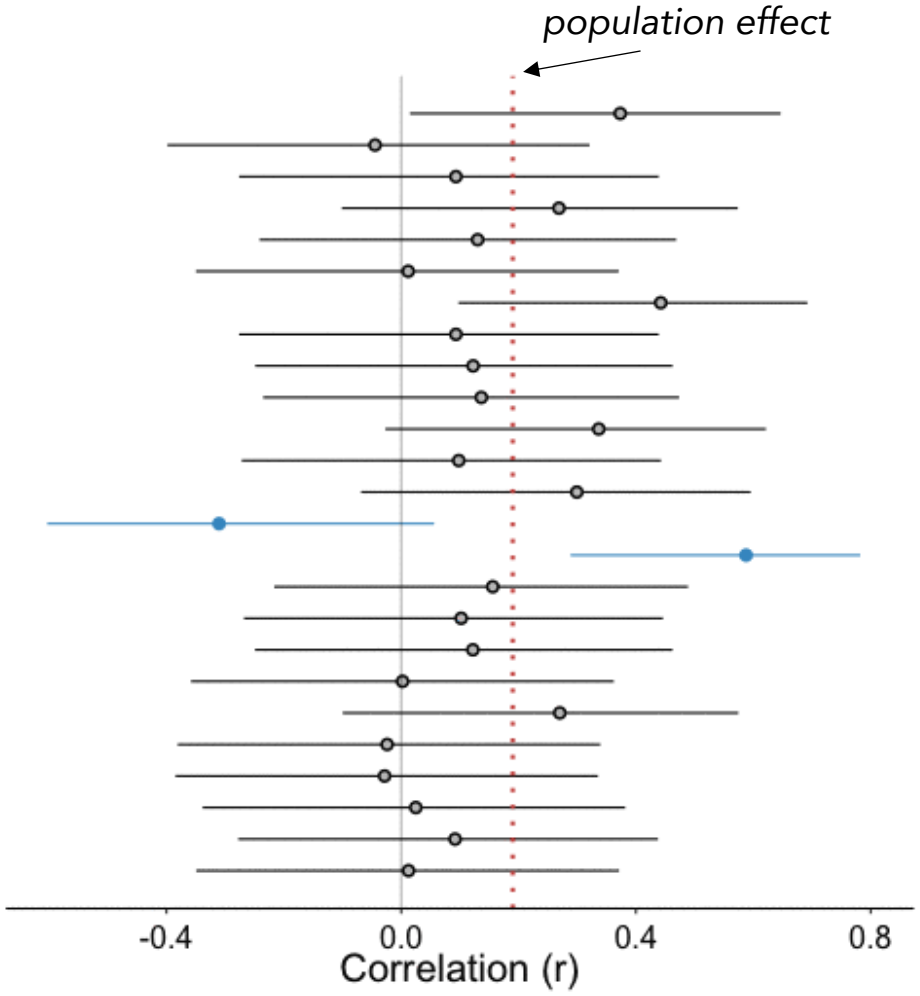


Correlations with 95% Confidence Intervals from 25 studies (n=30).

CIs that do not include zero (grey line) are statistically significant at $p < .05$.

CI:
includes Pop. r
does not include Pop. r

The New Statistics: Effect Sizes and Confidence Intervals



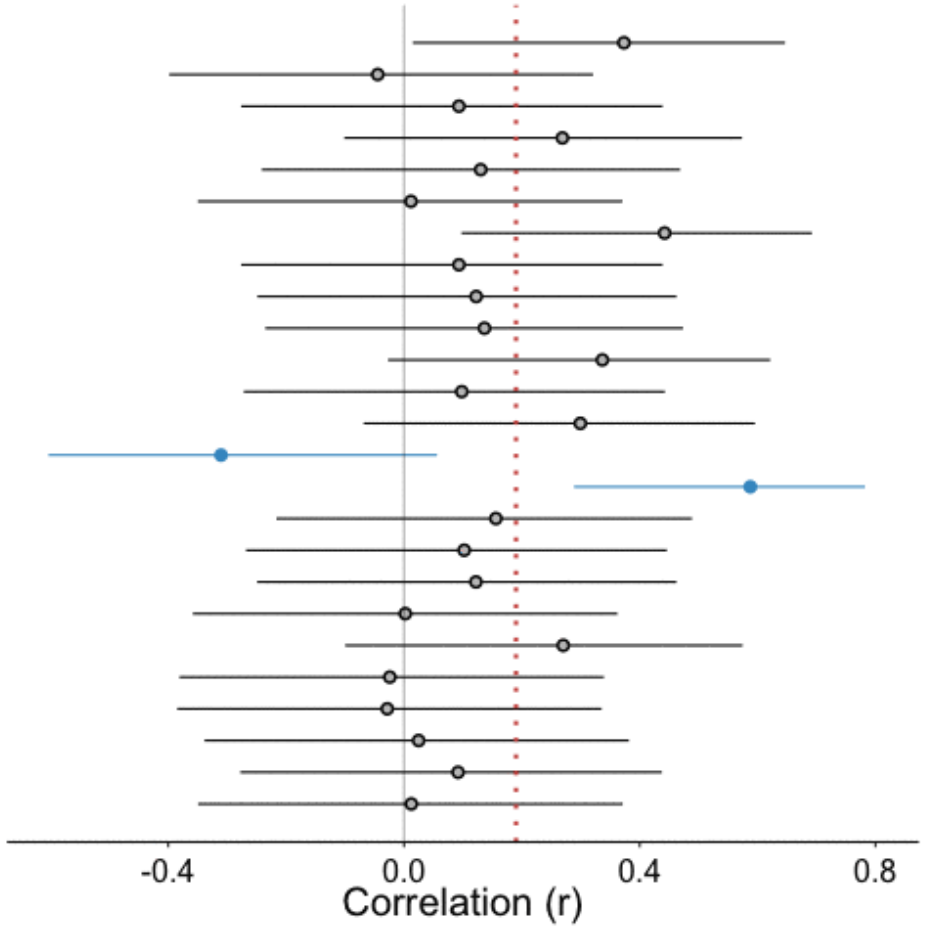
Correlations with 95% Confidence Intervals from 25 studies (n=30).

CIs that do not include zero (grey line) are statistically significant at $p < .05$.

In infinite replications, 95% of CIs will include the population effect (r=.15; red dotted line).

CI:
includes Pop. r
does not include Pop. r

The New Statistics: Effect Sizes and Confidence Intervals



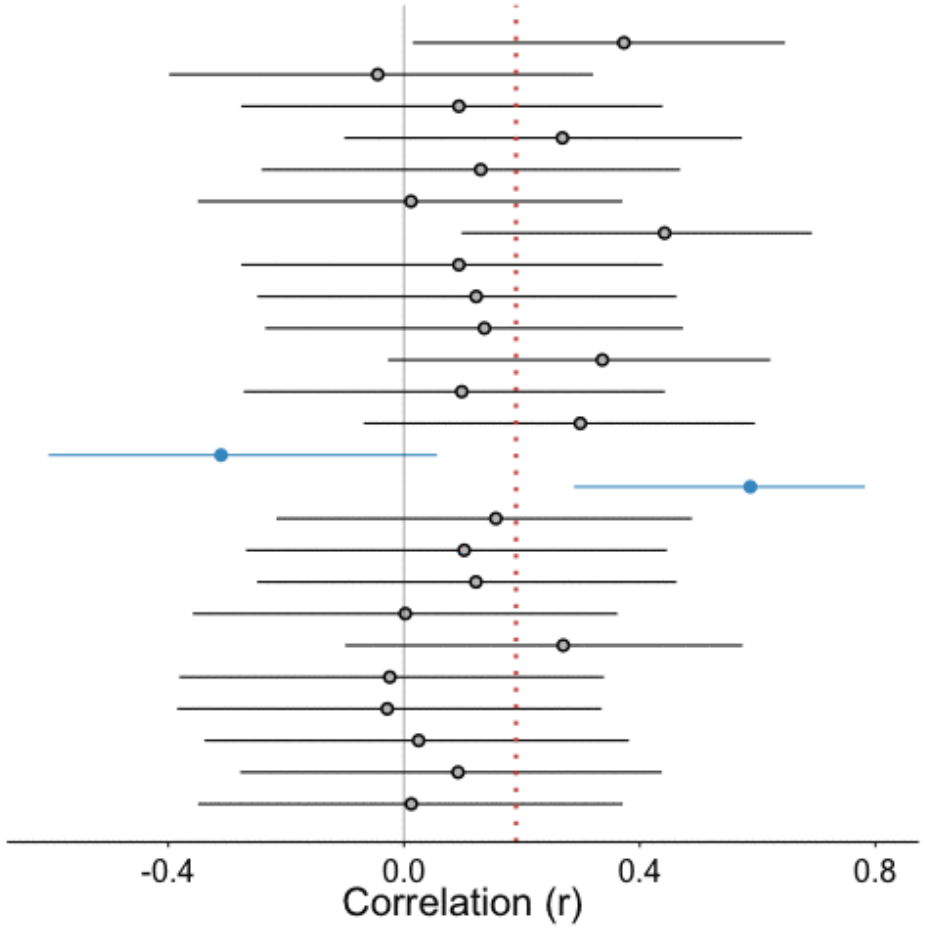
Groups of 25 studies.

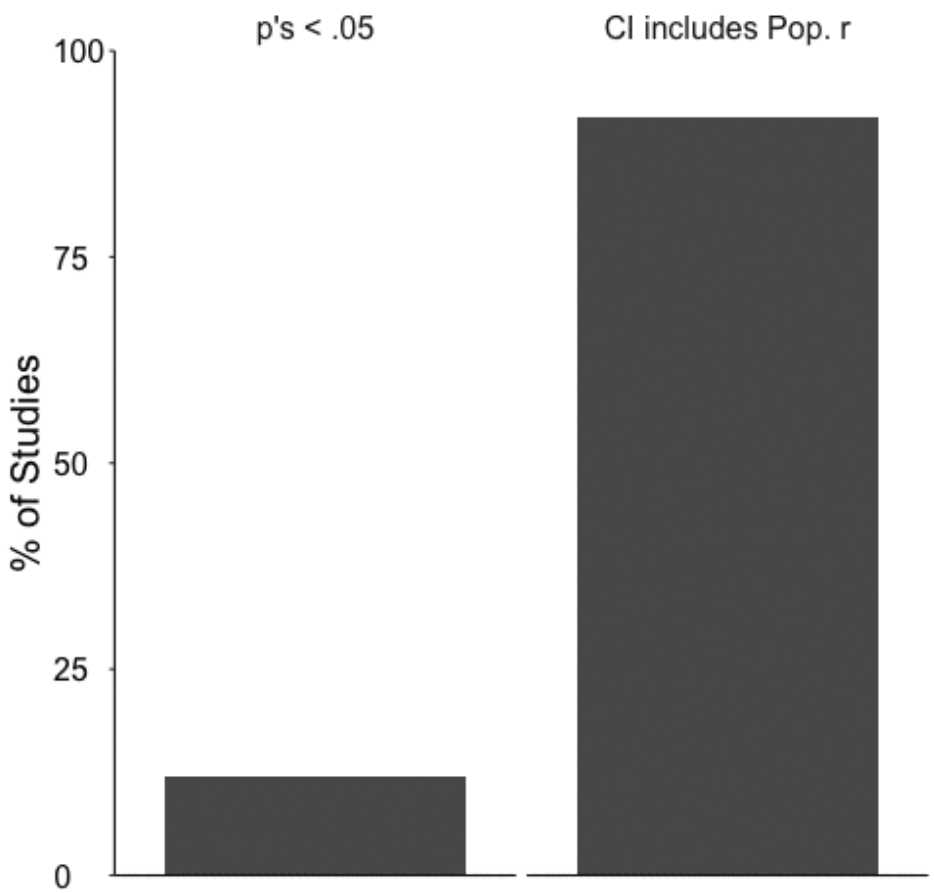
~95% of CIs include population r (red line).

Most are not statistically significant (include zero: grey line; power ~20%).

CI:
includes Pop. r
does not include Pop. r

The New Statistics: Effect Sizes and Confidence Intervals





The New Statistics: Effect Sizes and Confidence Intervals

In an individual study, effect sizes and confidence intervals (*estimation*) versus binary null hypothesis significance testing ($p < .05$), provide a plausible interval of values that will better capture the true population effect.

Cumming 2014, Psychological Science

The New Statistics: Effect Sizes and Confidence Intervals

In an individual study, effect sizes and confidence intervals (*estimation*) versus binary null hypothesis significance testing ($p < .05$), provide a plausible interval of values that will better capture the true population effect.

Embracing confidence intervals incorporates study-level error estimates that better protect against potentially biased inferences from effect sizes or significance alone.

Cumming 2014, Psychological Science

The New Statistics: Effect Sizes and Confidence Intervals

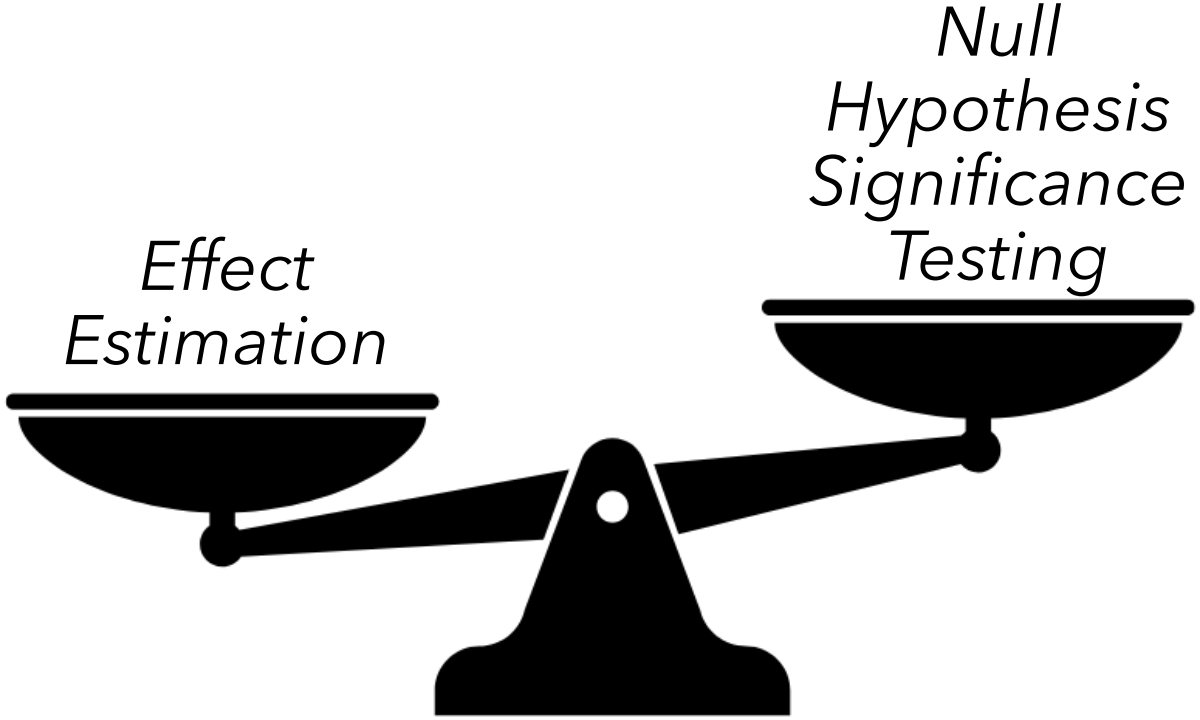
Additional references for effect sizes and confidence intervals

r-family (association), versus d-family (difference) effect sizes

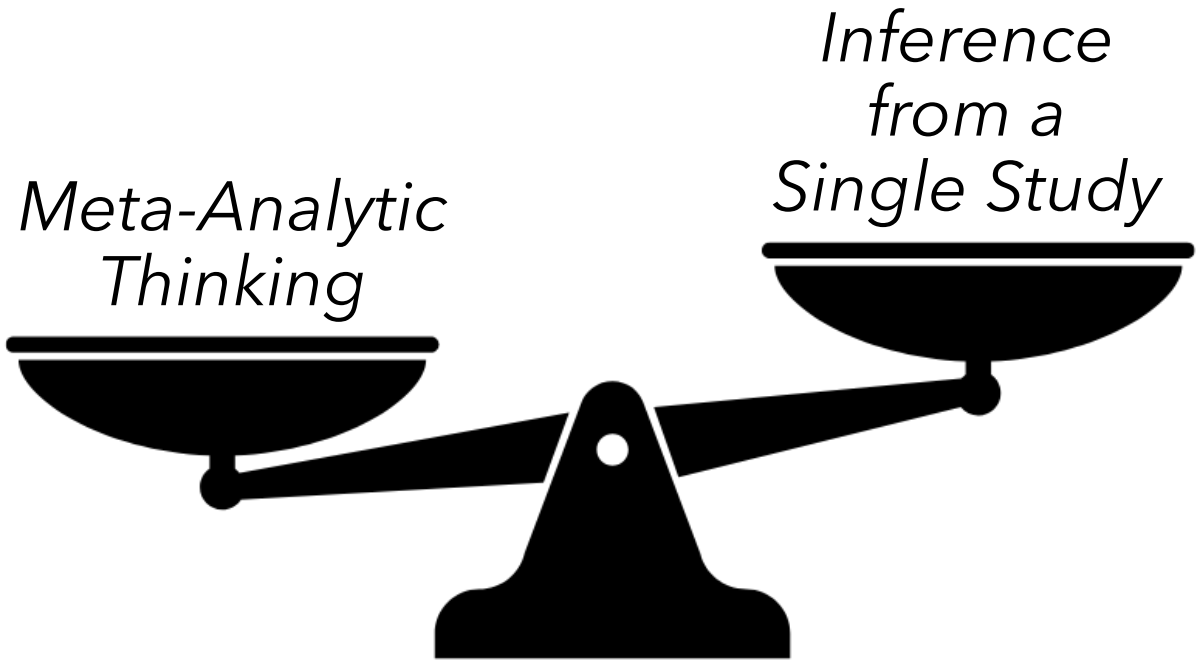
Confidence intervals can have different levels (e.g., 95%, 99%), calculations/assumptions (e.g., r to z transformation for correlation), estimated through resampling methods: *bootstrapping,* and can be semantically tricky.

Cumming 2014, Psychological Science;
Cumming 2013, Understanding the New Statistics (Routledge);
Goulet-Pelletier & Cousineau 2018, *The Quantitative Methods for Psychology;*
Thomas 2007, Psychology in Schools

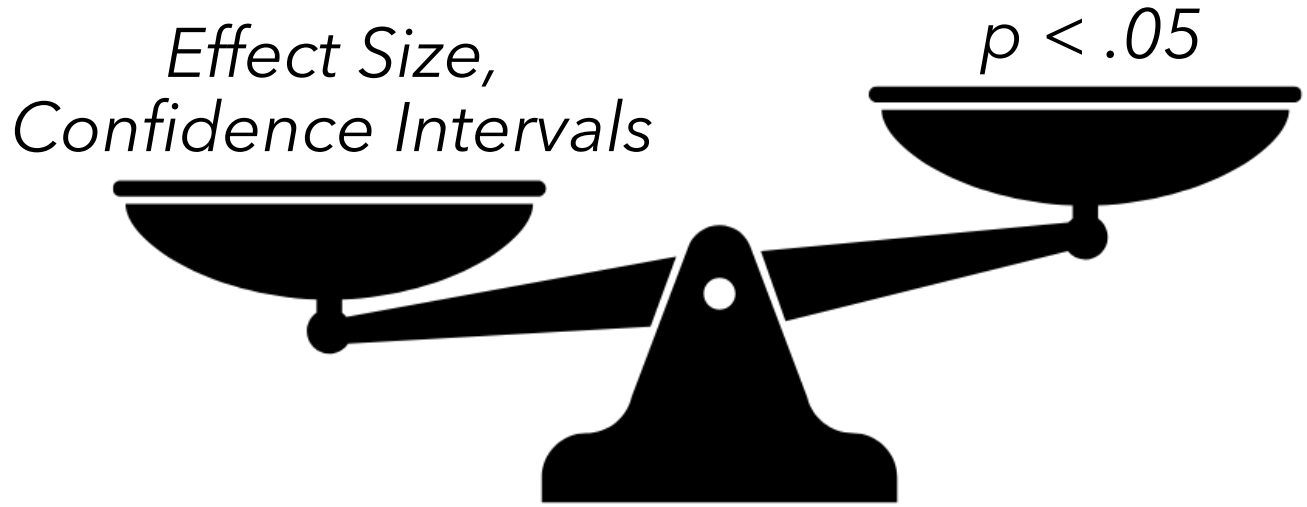
The New Statistics: Recommendations



The New Statistics: Recommendations



The New Statistics: Recommendations



Outline

1. Background and quantitative foundations of the reproducibility crisis.
2. “The New Statistics” to address these challenges.
3. The “Even Newer Statistics” and bringing across quantitative areas.

Outline

1. Background and quantitative foundations of the reproducibility crisis.
2. “The New Statistics” to address these challenges.
3. The “Even Newer Statistics” and bringing across quantitative areas.

The Even Newer Statistics (in Psychology)

Replication as a part of the analysis plan/study design.

Planned replication studies

Large, generalizable samples/consortia

Meta/mega-analysis

Bayesian versus Frequentist approaches

Cross-validation & train-test splits

The Even Newer Statistics (in Psychology)

Replication as a part of the analysis plan/study design.

Planned replication studies

Large, generalizable samples/consortia

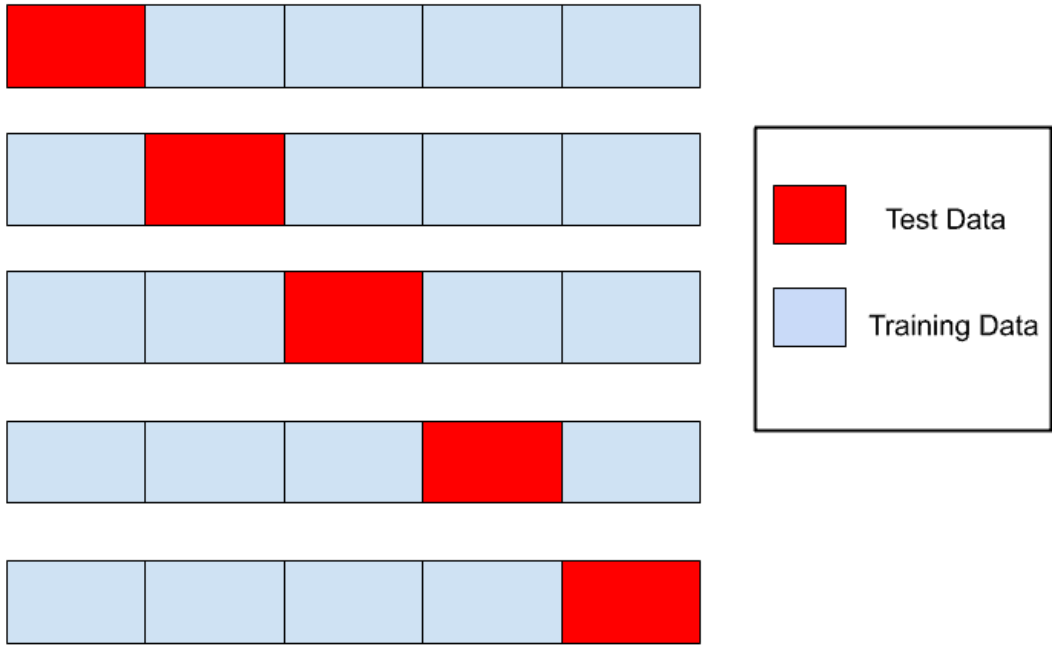
Meta/mega-analysis

Bayesian versus frequentist approaches

Cross-validation & train-test splits

The Even Newer Statistics (in Psychology)

Cross-validation & train-test splits (cf., machine learning)



cf., Browne 2000, *J. Mathematical Psychology*

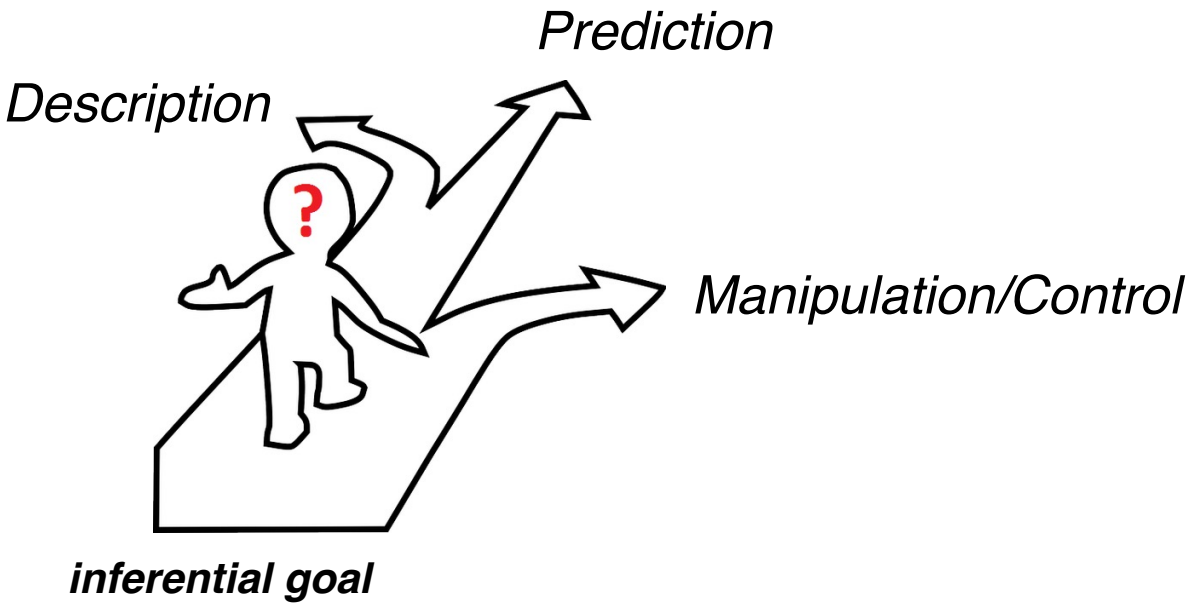
The Even Newer Statistics (in Psychology)

Contextualizing the relative importance of hypothesis generation versus hypothesis testing, reproducibility, and generalizability in the "maturity of a given research field".

Yarkoni & Westfall 2017, Perspect Behav Sci; Killeen 2019, Perspect Behav Sci
Tervo-Clemmens, Marek, & Barch, in press

The Even Newer Statistics (in Psychology)

Considering the specific demands for reproducibility and generalizability research based on the inferential goals of a given research design.



Yarkoni & Westfall 2017, Perspect Behav Sci; Killeen 2019, Perspect Behav Sci

Outline

1. Background and quantitative foundations of the reproducibility crisis.

2. “The New Statistics” to address these challenges.

3. The “Even Newer Statistics” and bringing across quantitative areas.

Summary

Psychological science faces ongoing challenges to reproducibility and robustness.

New tools/procedures (e.g., pre-registration, open data and code) to limit experimenter bias are increasingly important.

Quantitative insights into reproducibility challenges and "The New Statistics" (effect sizes, confidence intervals, meta-analysis) and contemporary practices to incorporate reproducibility as a central part of an analysis plan (e.g., cross-validation) are essential for methodologists.

Quantitative Expertise...
Matching Methods to Questions and Data

"The New Statistics"

- *Longitudinal data → growth curve/mixed effects*
- *High dimensional data/correlated measures → latent variable analysis*
- *Optimizing prediction → regularization/machine learning*

Quantitative Methods for The Replication Crisis: "The New Statistics" …and ”Some Even Newer Statistics"

Brenden Tervo-Clemmens, Ph.D.

Massachusetts General Hospital, Harvard Medical School

btervo-clemmens@mgh.harvard.edu

code for simulations and animations:
github.com/tervoclemmensb/newstatsdemo

The New Statistics: Effect Sizes and Confidence Intervals

A note of caution on the interpretation of confidence intervals

The New Statistics: Effect Sizes and Confidence Intervals

A note of caution on the interpretation of confidence intervals

*The true population effect has a single value that is almost always unknown (real-world data versus simulation). *

<u>Our effect size and confidence intervals are estimates</u>

[1]*Cumming 2014, Psychological Science*

The New Statistics: Effect Sizes and Confidence Intervals

Interpretations of Confidence Intervals

- "If the interval is calculated for an infinite number of replications, 95% of these replications will include the population value."

- "The CI is a set of values that are plausible for the population value."

- "We can be 95% confident that the interval contains the population value. We can can think of the lower and upper CI limits as likely lower and upper bounds for the population parameter."

Cumming 2014, Psychological Science