P-value and Replicability: A talk by Dr. Yoav Benjamini

Angel Garcia de la Garza

November 15th, 2017

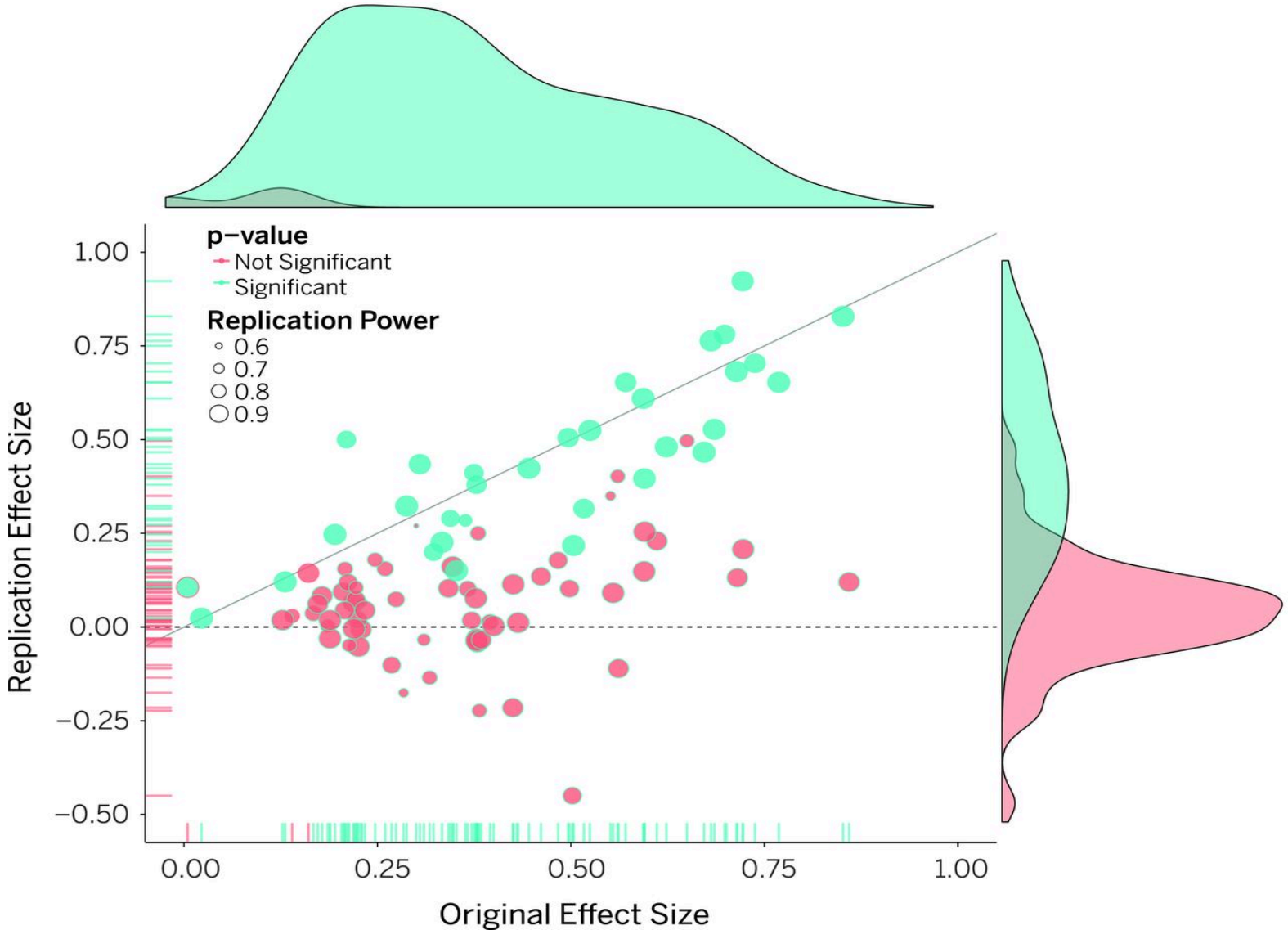
The original presentation can be found online at: http://www.replicability.tau.ac.il/index.php/links.html

Replicability vs. Reproducibility

- Reproduce the study: from the original data, through analysis, to get same figures and conclusions
- Replicability of results: replicate the entire study, from enlisting subjects through collecting data, and analyzing the results, in a similar but not necessarily identical way, yet get essentially the same results.

(Biostatistics, Editorial 2010, Nature Editorial 2013, NSF 2015)

Replicability crisis



Estimating the reproducibility of psychological science (Science, 2015)

Controversy surrounding p-values

- ***Psychological Science:*** "… seeks to aid researchers in shifting from reliance on NHST to estimation and other preferred techniques"
- **Basic and Applied Social Psychology**: *"From now on, BASP is banning the NHSTP…prior to publication, authors will have to remove all vestiges of the NHSTP (p -values, t -values, F –values, statements about ''significant'' differences or lack thereof, and so on)."*

Use other alternatives to p-value:

- Likelihood ratios
- Bayesian methods
- Prediction intervals
- Confidence intervals
- Effect size

How to address replicability?

1. Well and transparently designed experiment
2. Reproducible data analysis and computation
3. Statistical methodology that enhances replicability

    But what is it?

    What problems should it address?

How to address statistical obstacles of replicability?

- Addressing selective Inference:
  - P-values and related analyses should not be reported selectively
  - In study and out-of-study selection.
- Addressing the relevant variability:
  - You fail to account for true variability in study design.

Selective Inference

- “When inferring on a selected subset of the parameters, that turned out to be of interest after viewing the data the original properties no longer hold”

Out-of-study selection - may not be evident in published work: publication bias, Cherry-picking, Data Snooping, p-Hacking.

In-study selection – by highlighting specific results in the abstract, a table, a figure or the way you model your data.

The issue of selective inference and multiple comparisons.

In 100 papers from the NEJM 2002-2010. (Cohen and YB ‘16)

- # of endpoints in a paper 4-167 ; mean=27
- In 80% multiplicity entirely ignored: $p \leq 0.05$ (in none fully addressed.)
- All studies designated primary endpoints, conclusions were based on secondary endpoints when the primary failed

- From YB analysis of 100 papers:
- # of inferences per study (4-700, average 72)
- Only 11 (partially) addressed selection

False Discovery Rate

- Also known as the Benjamini-Hochberg Procedure
- FDR is designed to control the expected number of discoveries that are actually false.
- You order your p-values in ascending order and find the largest k such that: $P_k \leq \frac{k}{m}\alpha$
- Similarly, there's the False Coverage Rate (FCR)
  - You select k "interesting" features from m total
  - Construct confidence interval at a marginal $1 - \alpha * \frac{k}{m}$

In favor of the p-value

- “ First defense line against being fooled by randomness” (Benjamini)
- Significance testing gives sign determination
- This might be one of the only ways to compare across conditions (GWAS, Brain imaging)
- Thresholds are not ideal but needed, you need to emphasize that those results close to the threshold are less convincing than those away
- You should accompany p-values by the CI of the effect size
- Avoid selective inference

In spite of large lab differences

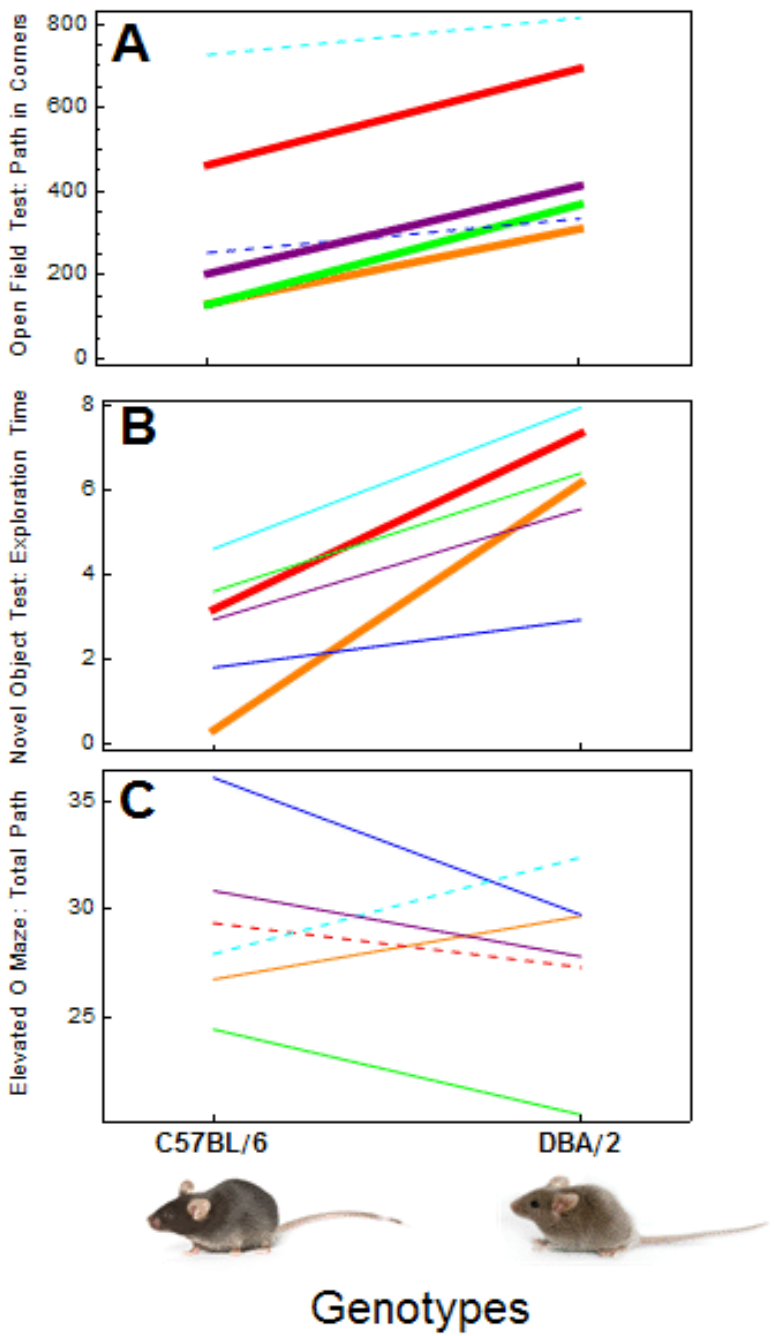
- Significant difference in 4/6
- Same direction same size
- Replicable

Significant difference in 6/6
Same direction different size

- Replicable (True)

Significant difference in 4/6
Different directions

- Non- Replicable (False)



Laboratories


Giessen


Muenster


Zürich


Mannheim


Munich


Utrecht