Too good to be false: Nonsignificant results revisited

CHRIS HJ HARTGERINK, JELTE M WICHERTS, MARCEL ALM VAN ASSEN

Highlights

- Relatively more nonsignificant results are reported in psychology papers now than in 1985
- Fisher test is a powerful method to detect presence of false negatives among nonsignificant *p*-values (3 *p*s when medium population effect)
- Evidence for false negatives in eight psychology journals, providing empirical evidence that nonsignificant results should not be discarded by psychology researchers.
- 2 out of 3 psychology papers has sufficient evidence for at least one false negative nonsignificant result

Testing for false negatives

Adjusted Fisher method

$$\chi^2_{2k} = -2\sum_{i=1}^{k} \ln\left(\frac{p_i - \alpha}{1 - \alpha}\right)$$

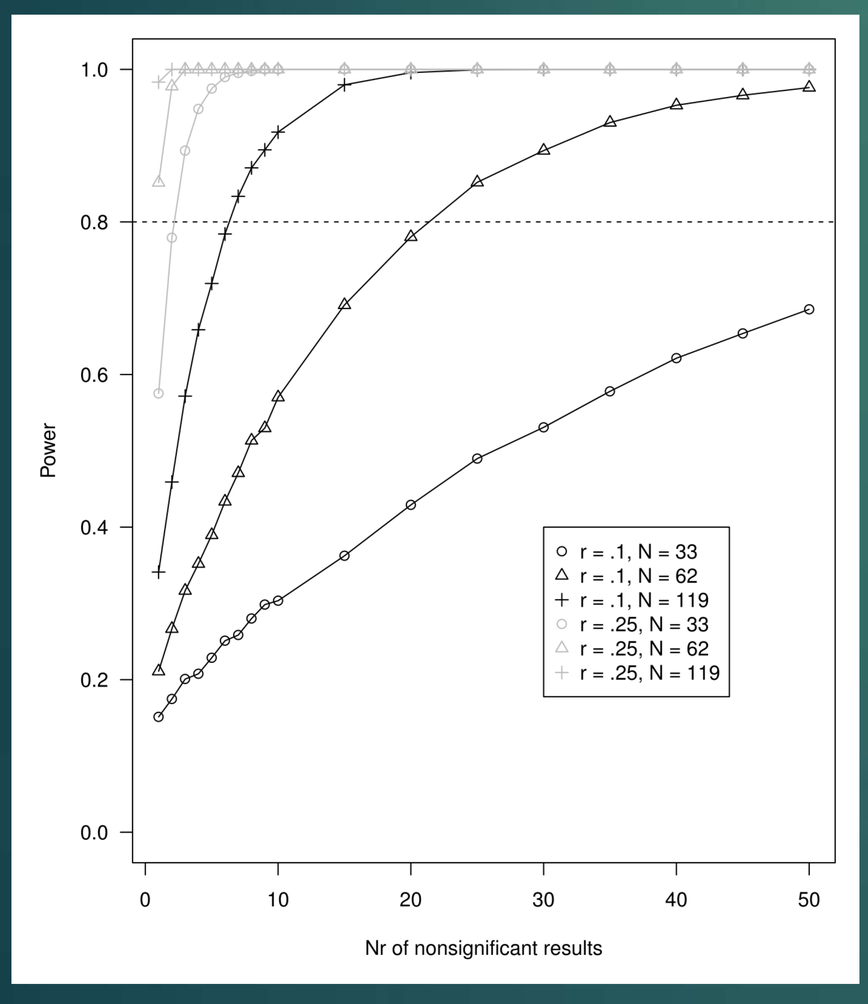
H_0: *p*-values are uniformly distributed (i.e., no effect)
H_1: *p*-values are right-skew distributed (i.e., an effect)

Simulation study

What is the power to detect a false negative among a set of nonsignificant p*-values?*

Simulation design:

1. Sample size single study (N = 33, 62, 119)
2. Number of nonsignificant studies (k = 1, 2, .. 10, 15, 20, ..., 50)
3. Underlying effect size (r = .00, .01, .02, ..., .99)



Application

- Use 54,595 nonsignificant results over 6,591 papers
- Data collected with `statcheck` for ~30,000 papers in eight psychology papers from 1985-2013 (data also used in Nuijten et al. 2015)
- Extracted 54,595 nonsiginificant results across 14,759 papers
- 6,951 of these 14,759 papers reporting nonsignificant results show evidence for at least one false negative (66.7%)
- Percentage of papers with evidence for false negatives varies across journals, from 49.4% (*Journal of Applied Psychology*) to 81.3% (*Journal of Personality and Social Psychology*)
- Correlation between median number of nonsignificant results strongly correlates with evidence of at least one false negative (*r* = .617)
- Sample sizes in psychology highly stable throughout thirty years

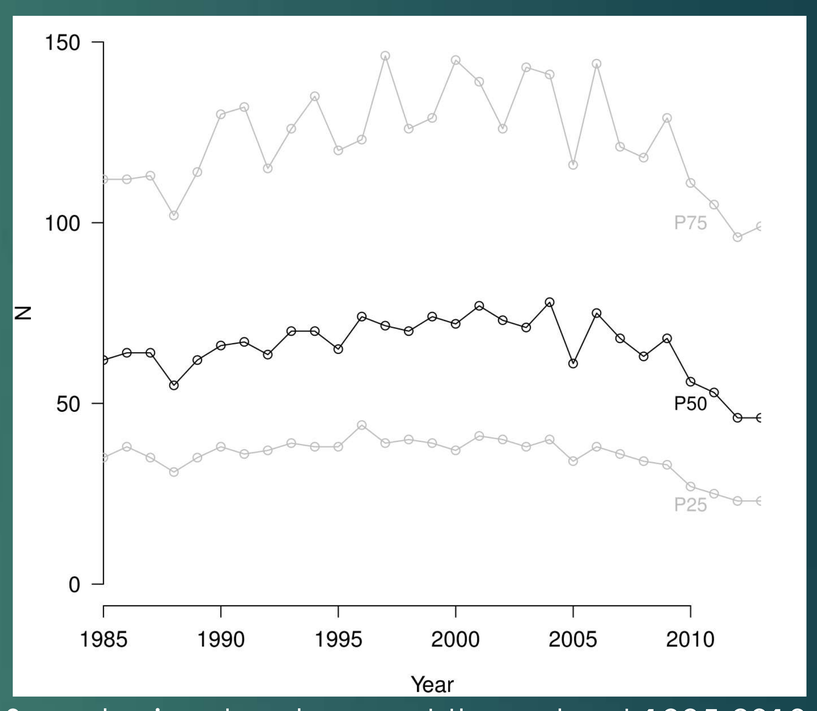


Figure 2. Sample size development throughout 1985-2013, based on degrees of freedom across 258,050 test results. P25 = 25th percentile. P50 = 50th percentile (i.e., median). P75 = 75th percentile.

- More nonsignificant results reported throughout the years, but less evidence for false negatives. Data indicate more smaller effects are reported over time.

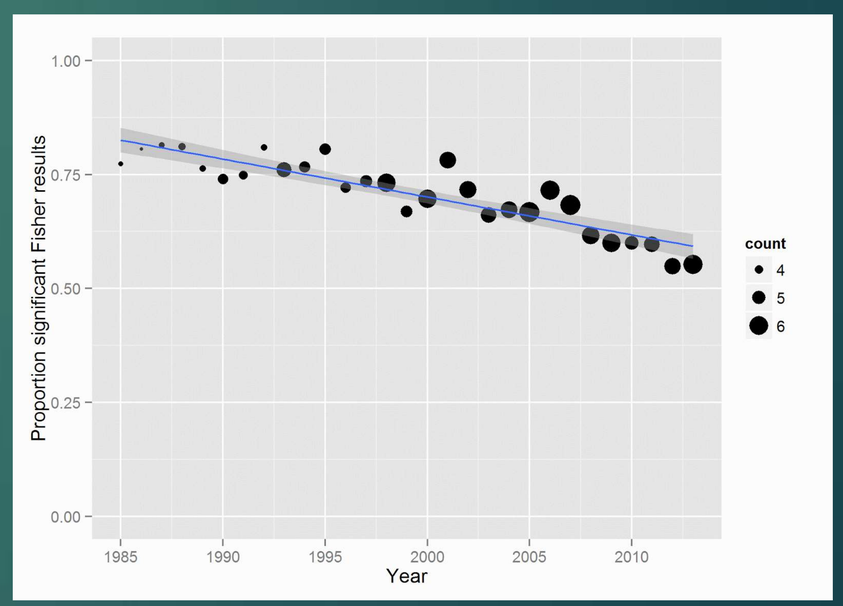


Figure 3. Proportion of papers reporting nonsignificant results in a given year, showing evidence for false negative results. Larger point size indicates a higher mean number of nonsignificant results reported in that year.

TILBURG UNIVERSITY
Understanding Society