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
- How many papers reporting nonsignificant results contain evidence for ≥1 false negative?
- Applied the Fisher method, which is a powerful method to detect presence of false negatives among nonsignificant *p*-values (3 *p*s when medium population effect)
- Using distributions of *p*-values is great! See also *p*-uniform (Van Assen et al., 2015)
- 2 out of 3 psychology papers has sufficient evidence for at least one false negative nonsignificant result.
- False negatives problematic, just like false positives; can solve both by increasing the sample size.
- Sample sizes have hardly changed in psychology since 1985.
- Nonsignificant replications in Reproducibility Project: Psychology (RPP) do not allow strong conclusions about true underlying effect.

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)

K = the number of nonsignificant results

Simulation study

How many nonsignificant results needed to get > .8 power to detect a false negative among a set of nonsignificant p-values?

	N=33	N=62	N=119
r = .1 (small)	**>50**	**25**	**7**
r = .25 (medium)	**3**	**1**	**1**
r = .5 (strong)	**1**	**1**	**1**

Application 1 – false negatives in Ψ

- 54,595 nonsignificant *p*-values extracted from 6,591 papers in 8 well-known psychology journals (data from Nuijten et al., 2015)
- 66.7% of the 6,951 papers show evidence for at least one false negative, varying across journals (49.4%-81.3%), and dependent on the number of nonsignificant results reported (*r* = .617)
- More nonsignificant results reported throughout the years, but less evidence for false negatives. Data indicate more smaller effects are reported over time.

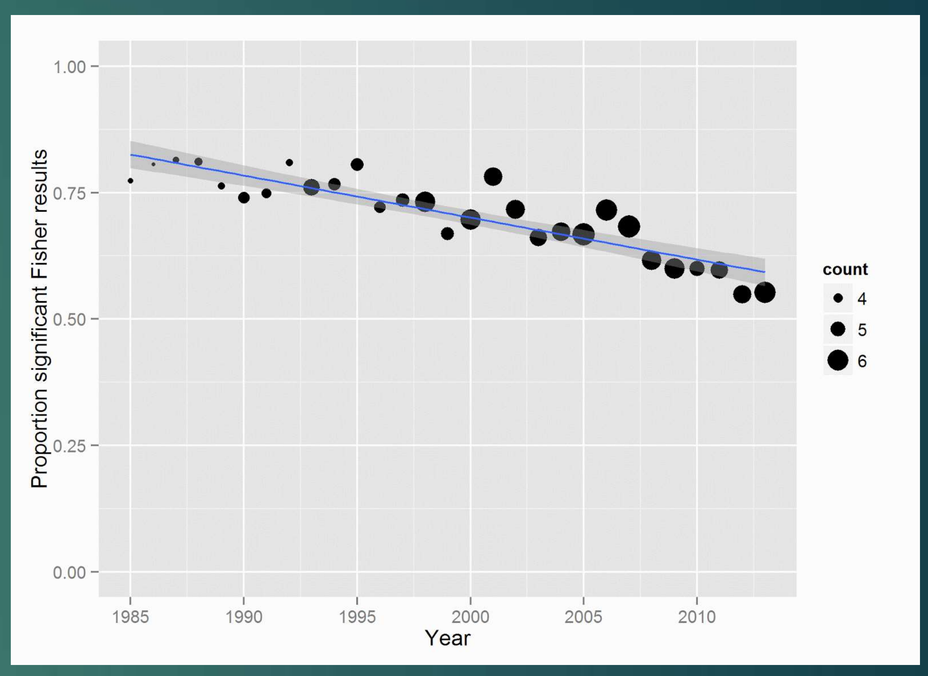

Figure 1. Proportion of papers reporting nonsignificant results in a given year, showing evidence for false negative results. Larger point size indicates a higher mean number of nonsignificant results reported in that year.

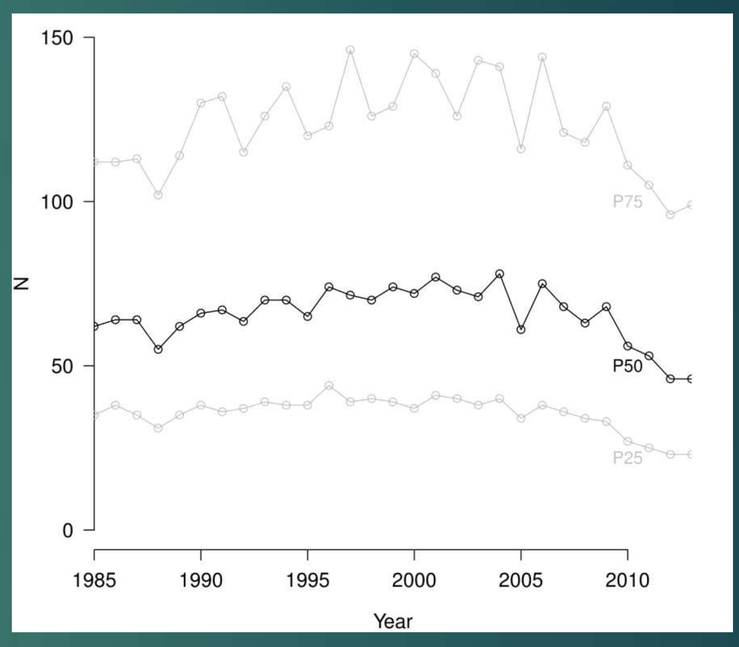

Figure 2. Sample size development throughout 1985-2013, based on degrees of freedom across 258,050 test results. P25 = 25th percentile. P50 = 50th percentile (i.e., median). P75 = 75th percentile.

Application 2 – nonsignificant replications in RPP

- Of the 100 replications, 64 statistically nonsignificant
- Of the 63 with test statistics, what would be the 95% confidence interval of false negatives when imposing a small, medium, or strong population effect?
- Small (r = .1) → 0-63 false negatives
- Medium (r = .3) → 0-21 false negatives
- Strong (r = .5) → 0-13 false negatives

TILBURG UNIVERSITY
Understanding Society