Review of null hypothesis significance testing (NHST) & effect size

Qi-Wen Ding

Outline

1. Limitations of NHST
2. Definition and types of effect size (ES)
3. Reporting guidelines of ES

Null hypothesis significance test (NHST)

Now NHST is a makeshift mix of Fisher and Neyman-Pearson approaches:

Planning stage (Neyman-Pearson)

1. Set H_0, H_1 (, and expected ES)
2. Select a proper test statistic
3. Set α
4. Determine the sample size to maintain enough statistical power

Decision stage (Fisher)

5. Compute the *p*-value
6. Compare *p*-value with α, and make decision

However, in practice, the planning stage is often overlooked…

Criticism of (the misuse of) NHST

- The importance of power is often overly neglected in practice.
- As long as the sample size is large, NHST can easily yield “significant” result even when the effect is very small. (Amrhein, Greenland & McShane, 2019)
- Even if the null hypothesis is rejected, it still cannot provide more information about the parameters. (Cohen, 1994; Kirk, 1996; Meehl, 1967)
- P-value is not $P(H_0$ is true) or $P(H_0$ is true|Data)
- …

Definition of effect size (ES)

Therefore, many scholars recommend focusing more on the **magnitude of the effect** when interpreting results.

(e.g., APA, 2010, 2020; Citrome & Ketter, 2013; Grissom & Kim, 2012; Kirk, 1996; Maxwell, Camp, & Arvey, 1981)

Kelley & Preacher (2012)

- **Effect size** is defined as a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest.

- The question of interest might refer to central tendency, variability, association, difference, odds, rate, duration, discrepancy, proportionality, superiority, or degree of fit or misfit, ...

Types of ES

1 *d*-type

Difference of scores

- Cohen's *d*
- Raw mean difference
- …

2 *r*-type

Strength of association

- Pearson's correlation
- Regression coefficient
- …

3 Others

- Relative risk
- Odds ratio
- …

For the advantages, disadvantages, statistical properties, and confidence interval calculations of each ES, please refer to Supplementary Table 1 in https://osf.io/n69xs/.

Utilities of ES

- Conduct power analysis to determine the sample size.
  (Cohen, 1962, 1969)
- Quantify the strength of effect in the study
- Facilitates researchers in conducting meta-analyses.

Researchers' expectations for ES

- Its point estimate should be independent of sample size and statistical significance. (Kelley & Preacher, 2012; Rosenthal, 1994)
- Has good statistical properties such as unbiasedness and efficiency (Kelley & Preacher, 2012)
- Easy to understand (Pek & Flora, 2018); can express the usefulness of the study (Kirk, 1996)
- Can highlight the theoretical, practical, or clinical significance of the research findings. (APA, 2010; Thompson, 2002; Wilkinson & the Task Force on Statistical Inference, 1999)

Reporting guidelines of ES

1. Authors should report ESs in their research results to help readers understand the strength of the effects or the importance of the findings.
   (APA, 2010, p. 34; APA, 2020, p. 89)

2. If the measurement units have practical meaning, ESs expressed in the original units can be reported.
   (APA, 2010, p. 34; APA, 2020, p. 89; Wilkinson & the Task Force on Statistical Inference, 1999, p. 599)

3. Effect sizes, their confidence intervals or standard errors, statistical test results, and their significance should all be reported together.
   (AERA, 2006, p. 37; APA, 2010, p. 34; APA, 2020, p. 89)

4. Provide readers with sufficient information to evaluate the practical utility of the results or their theoretical or clinical significance.
   (APA, 2010, p. 34; APA, 2020, p. 89)

Reporting guidelines of ES

5. Compare the reported ESs with those from similar past studies to help readers assess the stability of the results across different samples, designs, and analyses.
   (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599)

6. When interpreting ESs, the contextual factors of the study should be considered, such as the characteristics of participants in the experimental and control groups and the events experienced by each group.
   (APA, 2010, p. 36; APA, 2020, p. 90)

7. When there is a discrepancy between statistical significance and the strength of ES, possible reasons should be provided.
   - E.g., NHST did not reach significance due to insufficient sample size, even though the ES is large.

   (Chen et al., 2020, p. 576)