Statistical Methods for Replicability Assessment

Yi Zhao and William Wen

March 31, 2022
ROLES OF REPRODUCIBILITY IN SCIENTIFIC RESEARCH

- A necessary characteristic of correctness and rigor for scientific discoveries

- Irreproducible findings can erode public trust in science and cause damages to society

- Growing awareness in public since 2010s ("Replication crisis" Wikipedia)
CHALLENGES

- Confusions in different modes of reproducibility
- Lack of consensus on specifying “replication successes”
- Lack of rigorous statistical method for replicability assessment
Different modes of reproducibility

By emphasizing the interplays between data and methods:

- **Methods Reproducibility**: consistency between results generated from same data, same method

- **Inferential Reproducibility**: consistency under same data, different methods

- **Results Reproducibility** or **Replicability**: consistency under different data, same method
Inference Principles

- Concern variation of underlying true effects in different experiments:
  - Define an acceptable extent of variability

- Role of experimental random noise:
  - Non-informative principle: extremely noisy observations contain little information for assessment
Defining replication success based on repeated statistically significant findings

Utilize the compound quantity: signal-to-noise ratio (i.e., p-values)

Violates non-informative principle
Define Replication Success

DC Criterion

With a high probability, the underlying effects of replicable signals are expected to have the same (positive or negative) sign.

- Emphasize on true underlying effect
- Range of acceptable heterogeneity
- Establishing a baseline for assessing extent of heterogeneity
APPLICATION SCENARIOS

- Two-group scenario:
  - Original study and replication follows a chronological order
    - E.g., Reproducibility Project Psychology (RPP), a systematic replication attempt for findings in psychology

- Exchangeable group scenario:
  - A group of multiple experiments is gathered
    - E.g., systematic review
**Model Criticism Strategy**

- Define a family of reference models for characteristics of replicable results (i.e. high DC probability)

- Fit the replicable model with the observed data

- Evaluate the goodness-of-fit via Bayesian predictive checking procedures
  - Prior/Posterior-predictive replication p-value
  - \((1 - \alpha)\%\) predictive interval

- Poor-fitting indicates rejection of “the observed data are likely replicable”
APPLICATION: RE-ANALYSIS OF RPP

- Reproducibility Project: Psychology

- Goal: attempt to replicate 100 psychology studies published in three top psychology journals during 2008

- Findings reported: more than half of the scientific results are not reproducible because $p_{val_{orig}} < 0.05$ and $p_{val_{rep}} > 0.05$. 
RE-ANALYSIS OF RPP

estimated effect
study ID

- $p_{prior} < 0.05$
- $p_{prior} > 0.05$
Re-analysis of RPP

- Test statistic: $T_{pb} = \hat{\beta}_{rep}/\hat{\beta}_{orig}$

R CRAN package is available: https://CRAN.R-project.org/package=PRP