

# Reproducibility — Replicability: P-values and the Larger Questions

Andreas Buja

Department of Statistics, The Wharton School  
University of Pennsylvania  
Philadelphia, USA

With the PoSI group:  
Richard Berk, Larry Brown, Linda Zhao, Kai Zhang, Ed George

NRC 2015/02/26-27

Boos & Stefanski's contribution:

- Raising awareness of sampling variability in p-values.
- Showing that it can be quantified.
- Fundamental question: Seeing a p-value, do we believe that under replication something close to it would appear again and again? (see Steve Goodman citing Fisher)
- Use more stringent cut-offs than 0.05 to achieve replicable 0.05.

Basic pedagogical problems:

- P-values **are** random variables!  
They smell like probabilities but are transformed/inverted test statistics.
- The sense of “random variable”: “sampling variability”  
⇒ a tragically belittling term for a deep concept!
- “Sampling variability” = **dataset-to-dataset variability**  
= **possible-worlds variability**

# P-Values: From Variability to Bias

Source of Bias: Standard Error **SE** (tie to Benjamini and  $G \times L$ )

- Generic **SE**<sup>2</sup>:  $X_i$  standardized  $\Rightarrow \mathbf{V}[\bar{X}] = 1/n$
- Assumption:  $X_i \sim$  uncorrelated
- Consider exchangeable dependence: **Corr** $[X_i, X_j] = \rho > 0$   
 $\Rightarrow \mathbf{V}[\bar{X}] = (1-\rho)/n + \rho \geq \rho > 0$
- Example:  $\rho = 0.01 \Rightarrow \mathbf{SE} \geq \sqrt{\rho} = 0.1$ , never mind  $n$ .
- Random effects model for research studies:  $\mathbf{X}_{study,i} = \alpha_{study} + \epsilon_{study,i}$   
 $\Rightarrow \mathbf{Corr}[X_{study,i}, X_{study,j}] = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\epsilon^2) = \rho$   
 $\Rightarrow$  Exchangeable intra-study correlation
- Message: Don't ask for larger studies; ask for multiple studies.

# Statistical and Economic Thinking for Replicability

- Statistical Thinking: Statistics = “quantitative epistemology”  
Statistics = the science that creates **protocols** for the acquisition of qualified knowledge.
  - ▶ Absence of protocols is damaging.
  - ▶ Important distinctions made today: replicability vs reproducibility; empirical, computational, statistical
- Economic Thinking: Research = “economic system”  
To solve the replicability problem, we must set incentives right.
- Points of attack:
  - ▶ Economic incentives: Journals and their policies
  - ▶ Statistical protocols: Researchers and their protocols

# Two Types of Reform: (1) Economics → Journals

**Journals:** Stop the chase of “breakthrough science”.

- Publish, solicit, and treat favorably:
  - ▶ replicated results,
  - ▶ negative outcomes.
- Insidious:
  - ▶ Researchers will self-censor if journals treat replicated results and negative outcomes even slightly less favorably.
  - ▶ Researchers lose interest as soon as negative outcomes are apparent.
- Ideal protocol: Journals should accept/reject **NOT** knowing outcomes (Young & Karr, Significance Mag. 2011). Accept/reject based on:
  - ▶ merit and interest of the research problem,
  - ▶ study design,
  - ▶ quality of researchers.
- Goal: **No outcome-based deselection and a share of replication.**

## Two Types of Reform: (2) Statistics → Researchers

**Researchers:** Account for *all* data-analytic activity.

- Reveal all exploratory data analysis, in particular visualizations.
- Reveal all model searching (lasso, forward/backward/all-subsets, Bayesian, ...; CV, AIC, BIC, RIC...)
- Reveal all model diagnostics and actions resulting from them.
- Attempt inference that accounts for all of the above.
- Principle: Any data-analytic action that could result in a different outcome in another dataset needs to be accounted for.
- Goal: **“Whole-Data-Analysis inference”**

# Some Attempts

- Post-selection inference:

- ▶ Did you ever write a contract with yourself to try just one selection method?
- ▶ **PoSI**: Inference that is inferentially insured against all attempts at model selection, including significance hunting (a form of p-hacking).

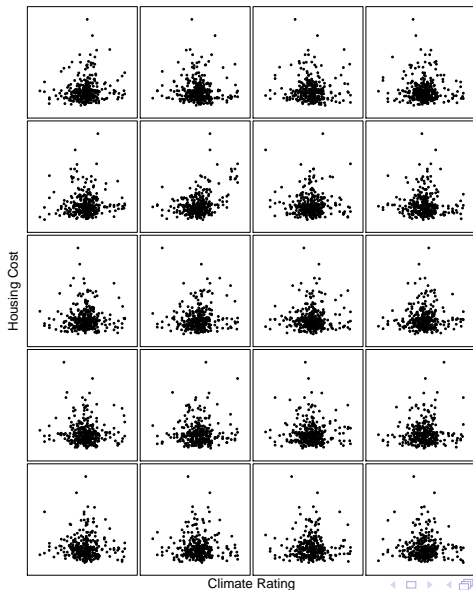
Berk et al., “Valid Post-Selection Inference,” AoS, 2013

- Inference for data visualization: a beginning

- ▶ Principle: Plot synthetic data and compare with the actual data.
- ▶ Sources of synthetic data: Permutations for independence tests, parametric bootstrap for model diagnostics, sampling conditional on sufficient statistics, ...
- ▶ **Line-up** protocol: insert the actual plot among 19 synthetic plots  
⇒ 5% significance

Buja et al., “Statistical Inference for Exploratory Data Analysis and Model Diagnostics,” Philosophical Transactions of the Royal Society A., 2009

# Line-Up: 5% significance if you find the actual data





# Summary

- P-values: variability and bias
- Institutional Reforms (1): Outcome-blind policies for journals
- Institutional Reforms (2): Whole-data-analysis protocols for researchers
- To achieve replicability, replicate.

THANKS!