UNIVERSITY OF AMSTERDAM

University of Amsterdam / Centrum Wiskunde & Informatica / JASP

Discussion of "Building Bridges: Bayesian Approaches for Increasing Reproducibility in Null Hypothesis Significance Testing" by Maria-Eglée Pérez

Alexander Ly

Santa Cruz, 8 September 2020+2


VALENCIA 6
6th Valencia International Meeting on

Outline

Recap rough ideas in an oversimplified manner

Highlight the following:

- OBayes: Bayes factors depend crucially on priors
- Session: Luis Pericchi's importance on my (and many others') view of Bayes factors/model selection

Mention "alternative" approach to improve replicability

- Safe testing

Rough recap

- Goal: Increase replicability with a familiar tool: *p*-value
  - Replace: $p < \alpha$ for all n
  - By: $p < \alpha(n) \downarrow 0$ as $n \to \infty$
    - Subgoal: Hide the Bayesian stuff
    - Avoid discussion on priors
- Method: Derive $\alpha(n)$ using
  1. (Asymptotic) sampling distribution of Bayes factors
  2. Approximate tail prob of the (asym) sampling distribution

Laplace approximation

$$-2\log \mathrm{BF}_{01}(Y,n) \approx \overbrace{-2\log\left(\frac{f(Y\mid X_0,\hat{\delta}_0,S_0^2 I_n)}{f(Y\mid X_1,\hat{\beta}_1,S_1^2 I_n)}\right)}^{\text{GLR}} \quad (1)$$

$$-2\log\left(\frac{|\hat{I}_1|^{1/2}}{|\hat{I}_0|^{1/2}}\right) - C \quad (2)$$

as $n \to$, where

$$C = m\log(2\pi) - \log\left(\frac{\pi_0(\hat{\delta}_0,S_0)}{\pi_1(\hat{\beta}_1,S_1)}\right), \quad m := m_1 - m_0 \quad (3)$$

Linear model

$$-2\log \mathrm{BF}_{01}(Y,n) \approx -(n-1)\log\left(\frac{Y^T(I-H_1)Y}{Y^T(I-H_0)Y}\right) \tag{4}$$

$$-\log\left(\frac{X_1^T X_1}{X_0^T X_0}\right) - C \tag{5}$$

Vélez, Pérez, Pericchi (2022) show under $\mathcal{M}_0$

$$-(n-1)\log\left(\frac{Y^T(I-H_1)Y}{Y^T(I-H_0)Y}\right) \xrightarrow{d} \mathrm{Gam}\left(\frac{m}{2}, \frac{\frac{n-m_1}{n-1}}{2}\right) \tag{6}$$

as $n \to \infty$

Gamma tail probability

Richter and Schumacher (2000)

$$\alpha \approx \frac{g_{n,\alpha}(m)^{\frac{m}{2}-1} \exp\left(-\frac{n-m_1}{2(n-1)} g_{n,\alpha}(m)\right)}{\left(\frac{2(n-1)}{n-m_1}\right)^{\frac{m}{2}-1} \Gamma\left(\frac{m}{2}\right)} \tag{7}$$

Replace $g_{X_0,X_1,n}(m) := g_{n,\alpha}(m) + \log(b) + C$, $b := \frac{X_1^T X_1}{X_0^T X_0}$

$$\alpha(n) \approx \frac{\left(g_{n,\alpha}(m) + \log(b) + C\right)^{\frac{m}{2}-1}}{b^{\frac{n-m_1}{2(n-1)}} \left(\frac{2(n-1)}{n-m_1}\right)^{\frac{m}{2}-1} \Gamma\left(\frac{m}{2}\right)} C_\alpha \tag{8}$$

Choosing $C_\alpha \approx$ choosing ratio of priors

1. Simple approximation/“BIC”: Normal (unit info) priors
   - Set $C = 0$
   - $C_\alpha = \exp(-\frac{n-m_1}{2(n-1)} g_{n,\alpha}(m))$
2. Minimal balanced experiment: Normal (unit info) priors
   - Set $C = 0$
   - Plugin $n_{\min}$, tolerable α, and solve for C_α
3. PBIC (Bayarri, Berger, Jang, Ray, Pericchi, Visser, 2019)/Tail of the robust priors (e.g. Bayarri et al, 2012)
   - Set $C = 2\sum_{i=1}^{m_0} \log \frac{1-e^{-v_i}}{\sqrt{2}v_i} - 2\sum_{j=1}^{m_1} \log \frac{1-e^{-v_j}}{\sqrt{2}v_j}$
   - $C_\alpha = \exp(-\frac{n-m_1}{2(n-1)} g_{n,\alpha}(m) + C)$

Ex: Balanced one-way ANOVA $K = 2$, i.e. two-sample t-test

$$\mathcal{H}_0 := \mu_1 = \mu_2 \text{ vs } \mathcal{H}_1 := \mu_1 \neq \mu_2 \tag{9}$$

$n_1 = n_2$	PBIC $\alpha(n_1, n_2)$ [%]	False rejections(?) [%]
10	2.83	34.18
50	1.59	8.57
100	0.61	3.07
500	0.41	0.22
1000	0.17	0.11

Some questions

- Q1: For $n_1 = n_2 \leq 50$ PBIC $\alpha(n)$ problematic?
- Q2: How does PBIC work for unbalanced designs? TESS (Berger, Bayarri, Pericchi, 2014) in this setting?
  - Ly (2018), Victor Peña (2018): Two-sample Bayes factor should converge to a one-sample Bayes factor
  - Dablander, van den Bergh, Wagenmakers, Ly (2022): $\text{plim}_{n_2 \to \infty} \text{BF}_{10}^{(2)}(s_1, n_1, S_2, n_2) = \text{BF}_{10\,;\,\sigma_2}^{(1)}(s_1, n_1)$
  - ~~Conjugate priors~~, but right Haar priors on nuisance parameters suffices.
- Q3: PBIC $\alpha(n)$ under optional stopping/continuation?

Replicability vs Questionable Research Practices

1. Optional continuation: 72% researchers decide whether to collect more data after looking to see whether the results were significant
2. Optional stopping: 36% researchers stop collecting data earlier than planned because one found the result that one had been looking for

Estimated prevalence of QRP (John et al. 2012)

Safe testing

- Peter Grünwald et al (CWI, Amsterdam)
- Aaditya Ramdas et al (CMU, Pittsburg)
- Glenn Shafer et al (Rutgers, New Jersey)

Ville/Robbin's inequality

- If $\mathrm{BF}_{10}(Y, n)$ is a super martingale wrt *all* $\mathbb{P}_0 \in \mathcal{M}_0$
- $\mathbb{E}_{Y \sim \mathbb{P}_0}[\mathrm{BF}_{10}(Y, n)] \leq 1$ for all n, then

$$\sup_{\mathbb{P}_0 \in \mathcal{M}_0} \mathbb{P}_0(\exists \tau, \mathrm{BF}_{10}(Y, \tau) \geq 1/\alpha) \leq \alpha \tag{10}$$

- Hence, tolerable 5% type I error, threshold $\mathrm{BF}_{10}(Y, n) > 20$
- If BF_{10} is an $\mathcal{M}_0$-NSM, then it is safe under optional stopping

Safe BF_{10} **for invariant hypotheses (Pérez-Ortiz, Lardy, de Heide, Grünwald, 2022)**

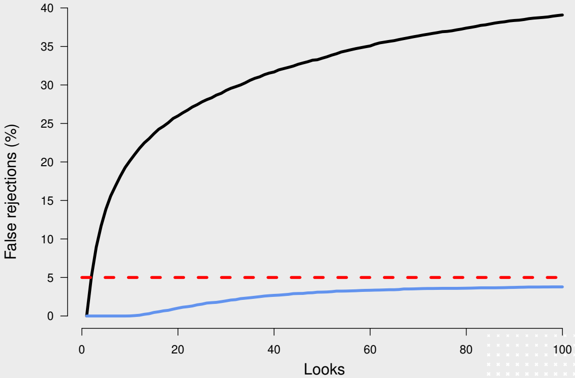
- Group invariance, e.g. location-shift invariance for two-sample *t*-test
- Put $\mu_1 = \mu_g + \delta\sigma/2$, $\mu_2 = \mu_g - \delta\sigma/2$
- Right Haar prior on nuisance parameters $\mu_g \propto 1, \sigma \propto \sigma^{-1}$
  - Conjugate priors σ don't work
  - One-dimensionalises the problem
  - Condition $\mathcal{M}_0$-NSM $\Rightarrow \mathbb{P}_0$-NSM
- Proper prior on δ, e.g. Zellner-Siow/Cauchy (Jeffreys 1948)
- Relevance of group structure for BF already highlighted by Berger, Pericchi, Varshavsky (1998)

Ex: Balanced one-way ANOVA $K = 2$, i.e. two-sample t-test

$$\mathcal{H}_0 := \mu_1 = \mu_2 \text{ vs } \mathcal{H}_1 := \mu_1 \neq \mu_2 \tag{11}$$

$n_1 = n_2$	PBIC $\alpha(n_1, n_2)$ [%]	Safe BF_{10} [%]
10	2.83	0.251
50	1.59	0.088
100	0.61	0.061
500	0.41	0.026
1000	0.17	0.019

Performance safe BF_{10}

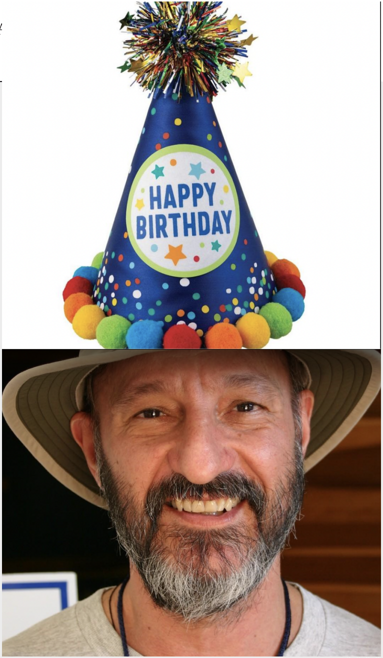


More questions

- Q1: For $n_1 = n_2 \leq 50$ PBIC $\alpha(n)$ problematic?
- Q2: How does PBIC work for unbalanced designs?
  - TESS (Berger, Bayarri, Pericchi, 2014) in this setting?
- Q3: PBIC $\alpha(n)$ under optional stopping/continuation?
- Q4: How to advice practitioners as a community?

References

- Bayarri, Berger, Jang, Ray, Pericchi, Visser (2019). Prior based Bayesian information criterion
- Berger, Bayarri, Pericchi (2014). The effective sample size
- Dablander, van den Bergh, Wagenmakers, Ly (2022). Default Bayes Factors for Testing the (In)equality of Several Population Variances.
- Turner, Ly, Perez-Ortiz, ter Schure, Grünwald (2022). `safestats` R package
- Perez, Pericchi (2014). Changing statistical significance with the amount of information The adaptive alpha significance level
- Perez-Ortiz, Lardy, de Heide, Grünwald (2022). E-statistics, group invariance and anytime valid testing
- Velez, Perez, Pericchi (2022). Increasing the Replicability for Linear Models via Adaptive Significance Levels


HAPPY
BIRTHDAY