Building bridges: Bayesian approaches for increasing reproducibility in Null Hypothesis Significance Testing

María-Eglée Pérez
(Joint work with L.R.Pericchi and D. Vélez)

Department of Mathematics
Universidad de Puerto Rico, Río Piedras Campus

OBayes 2022
September 8, 2022

Contents

Fishers's scale of evidence, particularly the $\alpha = 0.05$ threshold, has been used in literally millions of serious scientific studies, and takes a good claim to being the 20th century's most influential piece of applied mathematics.

Bradley Efron, 2010

A “Passport for Publication”

Current strategy for claiming new scientific discoveries is based on the appearance of a single study with a “statistically significant” result.

p- value of less than 0.05 has become a “passport for publication” (Cox 2015).

"Negative” studies are extremely difficult to publish.

The crisis of $p < 0.05$

Bayesian literature has been criticizing for several decades the implementation of hypothesis testing with fixed significance levels, and the use of the scale p value < 0.05.

The natural alternative to Null Hypothesis Significance (NHST) methods would be using exact posterior probabilities for the hypothesis, which automatically incorporate adjustments by sample size; unfortunately, these probabilities are rarely available to scientists, while tools for calculating p-values are widely available.

This fact suggests finding bridges between p-values and posterior probabilities of hypothesis, thus improving the decision making process.

The Adaptive α level

One such bridge is the the *Adaptive* α *Level* (Pérez and Pericchi 2014, *Stat Probabil Lett*), a correction for p-values developed to obtain the same asymptotic behavior of posterior probabilities.

A classical problem in the theory of statistics has been: How can the p-values be corrected from their dependence on the Sample Size?

In our view it can be solved as: How can the p-values be calibrated so that to have, the same asymptotic behavior as posterior probabilities?

The adaptive α level proposed by Pérez and Pericchi is

$$\alpha_{n^*}(q) = \frac{[\chi^2_\alpha(q) + q\log(n^*)]^{\frac{q}{2}-1}}{2^{\frac{q}{2}-1} n^{*\frac{q}{2}} \Gamma(\frac{q}{2})} \times C_\alpha \tag{1}$$

which is a simple (approximate) calibration for the significance level.

Still the constant C_α has to be determined.

Special case: $q = 1$

$$\alpha(n) = \frac{C_\alpha}{\sqrt{n \times (\log(n^*) + \chi^2_\alpha(1))}},$$

the **square root** $n \times \log(n)$ **formula**.

This formula has antecedents in Cox and Hinkley (1974), Good (1992) (both with typos), and gives a clear guidance on how to decrease the scale of p-values with the sample size.

n^* is the *Effective Sample Size*, (Berger, Bayarri and Pericchi, 2012).

A strategy to find C_α

The calibration will assume that we believe the α level is adequate for a specific sample size.

$$\alpha(n) = \frac{\alpha * \sqrt{n_0 \times (\log(n_0) + \chi^2_\alpha(1))}}{\sqrt{n^* \times (\log(n^*) + \chi^2_\alpha(1))}}.$$

Here n_0 is the sample size of a **reference experiment**, which is the result of a experimental design where the experimenter has specified a Type I error α and a Type II error β for a specific point of statistical importance.

For n_0, $\alpha(n) = \alpha$, but decreases (increases) as the sample size grows (decreases).

Variation of the significance level with sample size

Assume $n_0 = 10$ and $\alpha = 0.05$

Sample Size	$\alpha(n^*)$
10	0.0500
50	0.0199
100	0.0135
500	0.0055
1000	0.0038
10000	0.0011

Adaptive Alpha for linear Models

(Vélez *et al*, 2022)
We want to compare the following two nested linear models

$$M_i : \mathbf{y} = \mathbf{X}_i \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim N(0, \sigma_i^2 \mathbf{I}_n)$$

and

$$M_j : \mathbf{y} = \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j, \qquad \boldsymbol{\epsilon}_j \sim N(0, \sigma_j^2 \mathbf{I}_n).$$

Our hypothesis test is, then

$$H_0 : \text{Model } M_i \quad \textit{versus} \quad H_1 : \text{Model } M_j,$$

The Bayes Factor is:

$$B_{01}(\mathbf{y}) = \frac{\int f(\mathbf{y}|\mathbf{X}_i\boldsymbol{\delta}_i, \sigma_i^2\mathbf{I}_n)\pi^N(\boldsymbol{\delta}_i, \sigma_i)d\boldsymbol{\delta}_i d\sigma_i}{\int f(\mathbf{y}|\mathbf{X}_j\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I}_n)\pi^N(\boldsymbol{\beta}_j, \sigma_j)d\boldsymbol{\beta}_j d\sigma_j}.$$

The construction of adaptive alpha is based on $B_{01}(\mathbf{y})$, but this does not require the assessment of prior distributions by the user. Instead we will use well established statistical practices to directly construct summaries of evidence.

Laplace's asymptotic method, under regularity conditions, gives the following approximation

$$B_{01}^L = \frac{f(\mathbf{y}|\mathbf{X}_i\widehat{\boldsymbol{\delta}}_i, S_i^2\mathbf{I}_n)|\hat{I}_i|^{-1/2}}{f(\mathbf{y}|\mathbf{X}_j\widehat{\boldsymbol{\beta}}_j, S_j^2\mathbf{I}_n)|\hat{I}_j|^{-1/2}} \cdot \frac{(2\pi)^{i/2}\pi^N(\widehat{\boldsymbol{\delta}}_i, S_i)}{(2\pi)^{j/2}\pi^N(\widehat{\boldsymbol{\beta}}_j, S_j)}, \tag{2}$$

where $\widehat{\boldsymbol{\delta}}_i, S_i^2, \widehat{\boldsymbol{\beta}}_j, S_j^2$, are MLE's and $\hat{I}_i, \hat{I}_j$ observed information matrices respectively for M_i and M_j.

Since the first factor typically goes to ∞ or to 0 as the sample size accumulates, but the second factor stays bounded, it is useful to rewrite (2) as:

$$-2\log(B_{01}) = -2\log\left(\frac{f(\mathbf{y}|\mathbf{X}_i\widehat{\delta}_i, S_i^2\mathbf{I}_n)}{f(\mathbf{y}|\mathbf{X}_j\widehat{\beta}_j, S_j^2\mathbf{I}_n)}\right) - 2\log\left(\frac{|\hat{I}_j|^{1/2}}{|\hat{I}_i|^{1/2}}\right) + C. \quad (3)$$

$$\frac{f(\mathbf{y}|\mathbf{X}_i\widehat{\delta}_i, S_i^2\mathbf{I}_n)}{f(\mathbf{y}|\mathbf{X}_j\widehat{\beta}_j, S_j^2\mathbf{I}_n)} = \left(\frac{S_j^2}{S_i^2}\right)^{\frac{n}{2}} = \left(\frac{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_j)\mathbf{y}}{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_i)\mathbf{y}}\right)^{\frac{n}{2}},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ and the Observed Fisher Information Matrix (OFIM) with i adjustable parameters is

$$\hat{I}_i(\delta_i) = \frac{1}{S_i^2}\cdot \mathbf{X}_i^t\mathbf{X}_i.$$

So (3) can be written as:

$$-2\log(B_{01}) = -(n-1)\log\left(\frac{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_j)\mathbf{y}}{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_i)\mathbf{y}}\right) - \log\left(\frac{|\mathbf{X}_j^t\mathbf{X}_j|}{|\mathbf{X}_i^t\mathbf{X}_i|}\right) + C. \quad (4)$$

If we denote by $g_{n,\alpha}(q)$ the quantile of the test statistic fixed by α, using

$$-2\log(B_{01}) = -(n-1)\log\left(\frac{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_j)\mathbf{y}}{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_i)\mathbf{y}}\right) - \log\left(\frac{|\mathbf{X}_j^t\mathbf{X}_j|}{|\mathbf{X}_i^t\mathbf{X}_i|}\right) + C,$$

instead of letting the quantile fixed (as under the significance principle) we re-define significance as a quantity according to the following rule:

$$g_{\alpha(\mathbf{x}_i,\mathbf{x}_j,n)}(q) = g_{n,\alpha}(q) + \log\left(\frac{|\mathbf{X}_j^t\mathbf{X}_j|}{|\mathbf{X}_i^t\mathbf{X}_i|}\right). \tag{5}$$

Then the Bayes factor will converge to a constant (and $g_{\alpha(\mathbf{x}_i,\mathbf{x}_j,n)}(q)$ replace the fixed quantile).

Theorem

(Casella et al, *2009)*

Under H_0, the sampling distribution of $\dfrac{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_j)\mathbf{y}}{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_i)\mathbf{y}}$ *is a beta distribution,*

$$\frac{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_j)\mathbf{y}}{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_i)\mathbf{y}} \sim Beta\left(\frac{n-j}{2}, \frac{j-i}{2}\right).$$

Theorem

Under H_0, when $n \to \infty$, $-(n-1)\log\left(\dfrac{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_j)\mathbf{y}}{\mathbf{y}^t(\mathbf{I}-\mathbf{H}_i)\mathbf{y}}\right)$ *converges in distribution to a Gamma*$\left(\frac{q}{2}, \frac{\frac{n-j}{n-1}}{2}\right)$ *with $q = j - i$*

Now the α to the approximate upper tail in a Gamma $Ga\left(\frac{q}{2}, \frac{\frac{n-j}{n-1}}{2}\right)$:

$$\alpha \approx \frac{g_{n,\alpha}(q)^{\frac{q}{2}-1} \exp\{-\frac{n-j}{2(n-1)} \cdot g_{n,\alpha}(q)\}}{\left(\frac{2(n-1)}{n-j}\right)^{q/2-1} \Gamma\left(\frac{q}{2}\right)}.$$

If we replace the fixed quantile $g_{n,\alpha}(q)$ by $g_{\alpha_{(\mathbf{x}_i,\mathbf{x}_j,n)}}(q)$, the following result is obtained:

$$\alpha_{(\mathbf{x}_i,\mathbf{x}_j,n)}(q) = \frac{[g_{n,\alpha}(q) + \log(b)]^{\frac{q}{2}-1}}{b^{\frac{n-j}{2(n-1)}} \cdot \left(\frac{2(n-1)}{n-j}\right)^{q/2-1} \Gamma\left(\frac{q}{2}\right)} \times C_\alpha, \ \textit{with} \quad b = \frac{|\mathbf{X}_j^t \mathbf{X}_j|}{|\mathbf{X}_i^t \mathbf{X}_i|}. \quad (6)$$

Strategies for Selecting the Calibration Constant

- **The strategy of a minimal balanced experimen**

  Consider the one-way layout. Here the minimal balanced experiment has $n_i = 2$ for each group and $n = 2m = 2(q+1)$.
  Then,

$$C_\alpha = \alpha \cdot \frac{(2^q/(q+1))^{\frac{q+1}{2(2q+1)}}\,\Gamma\left(\frac{q}{2}\right)}{[g_\alpha(q) + \log(2^q/(q+1))]^{\frac{q}{2}-1}}$$

  where α is the desired level for the minimal sample. The case $m = 2$ is of particular interest since $q = 1$, then the calibration constant C_α is:

$$C_\alpha = \alpha \cdot \sqrt{\pi \cdot g_\alpha(1)}$$

- **The strategy of a simple approximation**.
  The simplest approximation in (2), which is implicit in the BIC approximation, comes from assuming priors $\pi^N(\boldsymbol{\beta}_j, S_j)$, $\pi^N(\boldsymbol{\delta}_i, S_i)$ to be $N((\beta_j, \sigma_j)|(\boldsymbol{\beta}_j, S_j), \hat{I}_j^{*(-1)})$, $N((\delta_i, \sigma_i)|(\boldsymbol{\delta}_i, S_i), \hat{I}_i^{*(-1)})$ respectively, where $\hat{I}_k^* = \hat{I}_k/n^*$, with $\hat{I}_k$ being the observed Fisher Information matrix, but noting that n^* is the Effective Sample Size, mentioned before. This leads to a $C = 1$ in (4) and then a

$$C_\alpha = \exp\left\{-\frac{n-j}{2(n-1)} \cdot g_{n,\alpha}(q)\right\}$$

- **The strategy based on the Prior Based Information Criterion PBIC**
  (Bayarri *et al.* 2019).

  PBIC improves BIC type approximations by

  i) replacing the "sample size" n by a more precise "The Effective Sample Size" TESS n^e (Berger, Bayarri and Pericchi, 2014) and
  ii) retaining the effect of the prior in the final expression using a flat-tailed no-normal prior.

Using PBIC, constant C in (3) is replaced by

$$C = 2\sum_{m_i=1}^{q_i} \log \frac{(1 - e^{-v_{m_i}})}{\sqrt{2} v_{m_i}} - 2\sum_{m_j=1}^{q_j} \log \frac{(1 - e^{-v_{m_j}})}{\sqrt{2} v_{m_j}},$$

where $v_{m_l} = \frac{\hat{\xi}_{m_l}}{[d_{m_l}(1+n^e_{m_l})]}$ with $l = i, j$ corresponding to the Model M_i and M_j respectively. The ξ_{m_l}'s come from orthogonal transformations of the models parameters (see Bayarri et al, 2019)

Hence

$$\alpha_{(b,n)}(q) = \frac{\left[g_{n,\alpha}(q) + \log(b) + C\right]^{\frac{q}{2}-1}}{b^{\frac{n-j}{2(n-1)}} \cdot \left(\frac{2(n-1)}{n-j}\right)^{q/2-1} \Gamma\left(\frac{q}{2}\right)} \times C_\alpha, \tag{7}$$

and

$$C_\alpha = \exp\left\{-\frac{n-j}{2(n-1)}\left(g_{n,\alpha}(q) + C\right)\right\}.$$

Example: Balanced One Way Anova

Suppose we have k groups with n observations each, for a total sample size of kn and let $H_0 : \mu_1 = \ldots = \mu_k = \mu$ *vs* H_1 : At least one μ_i different.
Then the design matrices for both models are

$$\mathbf{X_1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \mathbf{X_2} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}, b = \frac{|\mathbf{X_2^t X_2}|}{|\mathbf{X_1^t X_1}|} = k^{-1} n^{k-1}$$

The effective sample size is r, the number of replications.

In order to compare with the approach of Pérez and Pericchi (2014), we will use the strategy of a fixed sample size for a designed experiment. Here we used an effect size of $f = 0.25$, which according to Cohen (1988) represents a medium effect size. We fixed $\alpha = 0.05$ and the power at 0.8 . The sample sizes obtained were $n_0 = 64, 40$ and 26 for $k = 2, 5$ and 10, respectively.

	k					
	Adaptive α for linear model			Adaptive α (PP 2014)		
r	2	5	10	2	5	10
50	0.057	0.0327	3.6×10^{-3}	0.058	0.0333	3.8×10^{-3}
100	0.038	0.0087	2.2×10^{-4}	0.038	0.0093	2.4×10^{-4}
500	0.016	0.0004	3.1×10^{-7}	0.015	0.0005	3.4×10^{-7}
1000	0.011	0.0001	1.8×10^{-8}	0.010	0.0001	2.0×10^{-8}

Different calibration strategies for $k = 2$. All strategies yield comparable results, with the strategy based on PBIC being somewhat more drastic in its penalization for higher samples (for this particular case).

	$k = 2$		
r	Minimal sample	Simple Calibration	PBIC Calibration
4	0.0523	0.0412	0.0283
10	0.0342	0.0235	0.0159
50	0.0130	0.0090	0.0061
100	0.0087	0.0060	0.0041
500	0.0035	0.0024	0.0017
1000	0.0024	0.0017	0.0011

Simulation

Inspired by an example in Sellke et al (2001)

- Simulate r data points from each of two normal distributions $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$. We replicate this $2K$ times. For K of the simulations, $\mu_1 - \mu_2 = 0$, while for the other K $\mu_1 - \mu 2 = \Delta > 0$.
- For all $2K$ replications, test the hypotheses $H_0 : \mu_1 = \mu 2$ vs $H_1 : \mu_1 \neq \mu_2$, and then count how many of the p-values lie between $0.05 - \varepsilon$ and 0.05. Note that all these p-values would be deemed enough for rejecting H_0 if $\alpha = 0.05$ is selected.
- Determine the proportion of "significant" p-values obtained from samples where H_0 is true (false discoveries) .
- Repeat the whole experiment R times for stability.

Median percentage of false discoveries for $R = 100$ replicates of the simulation scheme with $K = 4000$, $\Delta = 0.25$, $\sigma = 1$ and $\varepsilon = 0.04$, for $r = 10, 50, 100, 500$ and 1000.

	% of samples with $0.01 < p < 0.05$	
	without adjustment	with PBIC calibration
r	2-groups	2-group
10	39.06%	34.18%
50	21.43%	8.57%
100	15.73%	3.07%
500	39.04%	0.22%
1000	97.15%	0.11%

The proportion of false positives is not monotonic with r, but always far higher than 5%. For $r = 1000$, almost 100% of these significant values near 0.05 are generated from H_0.

When the α level is corrected according to the method suggested using a calibration strategy based on PBIC, the proportion of false positives decreases steadily, providing a more reliable Type I error control.

Linear regression

Consider the models

$$\begin{aligned} M_i : y_v &= \beta_1 + \beta_2 x_{v2} + \cdots + \beta_i x_{vi} + \varepsilon_v \\ M_j : y_v &= \beta_1 + \beta_2 x_{v2} + \cdots + \beta_i x_{vi} + \beta_{i+1} x_{v(i+1)} \beta_j + \cdots + x_{vj} + \epsilon_v \end{aligned}$$

with $1 \leq v \leq n$ and $2 \leq j \leq k$, then

$$|\mathbf{X}_j^t \mathbf{X}_j| = n(n-1)^{j-1} \prod_{l=2}^{j} s_l^2 |R_j|$$

where s_l^2 and R_j are the variance and the correlation matrix of the predictors in model M_j respectively.

Then

$$b = (n-1)^{j-i} \left(\prod_{l=i+1}^{j} s_l^2 \right) |R_{j-i} - R_{ij}^t R_i^{-1} R_{ij}|,$$

Here R_{ij} is the correlation matrix between predictors of the models M_j that are not in M_i with predictors of the model M_i, and R_{j-i} is the correlation matrix of the predictors of the models M_j that are not in M_i,

Example: Car Mileage Data

(Acuña 2013)

We want to predict the average mileage per gallon (denoted by `mpg`) of a set of $n = 82$ vehicles using four possible predictor variables: cabin capacity in cubic feet (`vol`), engine power (`hp`), maximum speed in miles per hour (`sp`) and vehicle weight in hundreds of pounds (`wt`).

1. $H_0 : M_2 :(\text{mpg}{=}\beta_1{+}\beta_2\text{wt}_i{+}\epsilon_i)$ vs $H_1 : M_3 :(\text{mpg}{=}\beta_1{+}\beta_2\text{wt}_i{+}\beta_3\text{sp}_i{+}\epsilon_i)$
2. $H_0 : M_2 :(\text{mpg}{=}\beta_1{+}\beta_2\text{wt}_i{+}\epsilon_i)$ vs $H_1 : M_3 :(\text{mpg}{=}\beta_1{+}\beta_2\text{wt}_i{+}\beta_3\text{hp}_i{+}\epsilon_i)$
3. $H_0 : M_2 :(\text{mpg}{=}\beta_1{+}\beta_2\text{wt}_i{+}\epsilon_i)$ vs $H_1 : M_3 :(\text{mpg}{=}\beta_1{+}\beta_2\text{wt}_i{+}\beta_3\text{vol}_i{+}\epsilon_i)$

For these tests

$$b = (n-1)s_3^2(1-\rho_{23}^2),$$

where s_3^2 is the variance of the entering predictor in model M_3 and ρ_{23} is the correlation between wt y the new predictor in M_3.

Test	Predictor	Var(·)	Cor(wt,·)	b	p-value	α_{Simple}	α_{PBIC}
1	`sp`	197.1	0.68	8612.9	0.0325	0.0004	0.0134
2	`hp`	3230.9	0.83	80449.5	0.1661	0.0001	0.0046
3	`vol`	491.3	0.38	33901.1	0.6482	0.0002	0.0087

In all cases, the significance level is substantially reduced, specially using the simple calibration. Note that the strongest correction corresponds to the engine power (`hp`). This variable has both the largest variance and the highest correlation with the weight.

Adaptive α revisited: PBIC strategy

The PBIC strategy for the calibration of the constant could also have been used in (1), leading to

$$\alpha_n(q) = \frac{[\chi^2_\alpha(q) + q\log(n) + C]^{\frac{q}{2}-1}}{n^{\frac{q}{2}} 2^{\frac{q}{2}-1} \Gamma\left(\frac{q}{2}\right)} \times \exp\left\{-\frac{1}{2}\left(\chi^2_\alpha(q) + C\right)\right\}. \tag{8}$$

Here C is calculated similarly to the expression given for linear models.

Note that this adaptive α is still of BIC structure, since, the expression $\chi^2_\alpha(q) + q\log(n)$ remains.

From p-Values to Posterior Probabilities of Null Hypothesis

It is by now well known by practitioners, that p-values are not posterior probabilities of a null hypothesis, which is what science needs to declare a scientific finding.

The so-called Robust Lower Bound $BF \geq -e \cdot p \cdot \log(p)$ (Vovk 1993,Sellke et al 2001) links the p-value with real model probabilities, but does not take into account the dependence of the evidence on the sample size

Our goal is to adapt the Robust Lower Bound, make it dependent on the sample size, and approximate actual Bayes Factors, for any sample size.

A further complication arises when the null hypotheses depend on unknown nuisance parameters. Here, what are usually called p-values do not follow an Uniform distribution.

Pseudo-P-Values and Robust Lower Bound

We will use the general definition of p-value given by Casella and Berger (2001, p.397)

Definition

A p-**value** $p(\mathbf{X})$ is a statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every sample point $\mathbf{x}$. Small values of $p(\mathbf{X})$ give evidence that H_1 is true. A p-value is **valid** if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$P_\theta(p(\mathbf{X}) \leq \alpha) \leq \alpha.$$

We will consider any p-value complying the definition with strict inequality for a non-zero measure set of α values a **pseudo-**p**-value**.

The "Robust Lower Bound" (RLB) is:

$$B_L(p) = \begin{cases} -e \cdot p \cdot \log(p) & p < e^{-1} \\ 1 & \text{otherwise} \end{cases}$$

when $p|H_0 \sim U(0,1)$ and $p|H_1 \sim Beta(\xi, 1)$ for $0 < \xi < 1$ (see Sellke *et al* 2001)

Consider now the Hypothesis test

$$H_0 : p \sim Beta(\xi_0, 1) \quad vs \quad H_1 : p \sim f(p|\xi)$$

with ξ_0 fixed but arbitrary and $f(p|\xi) \sim Beta(\xi, 1)$ for $0 < \xi < 1$

Using an argument very similar to the one used in Sellke *et al* (2001) It can be shown that a Robust Lower Bound in this setting is

$$B_L(p, \xi_0) = \begin{cases} -e \cdot \xi_0 \cdot p^{\xi_0} \log(p) & p < e^{-1} \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

where ξ_0 has to be estimated or calculated theoretically, but we know that $\xi_0 = 1$ when the p-value is not pseudo-p-value.

For any Bayes Factor B_{01},

$$B_{01} \geq B_L(p) > B_L(p, \xi_0) \text{ with } \xi_0 > 1$$

Theorem

The RLB_ξ is a valid p-value, for $\xi \geq 1$, that is,

$$P(B_L(p,\xi) \leq \alpha | p \sim f(p|\xi)) \leq \alpha, \quad \text{for each } 0 \leq \alpha \leq 1.$$

Adjusting RLB_ξ with Adaptive α

When we subtitute the adaptive α (8) in the RLB_ξ given by equation (9), we obtain the following Bayes Factor

$$B(\alpha, q, n, \xi_0) = -\alpha^{\xi_0}\log(\alpha)\Gamma(q/2)^{\xi_0} n^{\frac{\xi_0 q}{2}}\left[\frac{2}{\chi^2_\alpha(q) + q\cdot\log(n) + C}\right]^{\frac{\xi_0 q}{2}-(\xi_0-1)} \quad (10)$$

For an uniform p-value , $\xi_0 = 1$ and the Bayes factor simplifies to

$$B(\alpha, q, n) = -\alpha\log(\alpha)\Gamma(q/2)n^{\frac{q}{2}}\left[\frac{2}{\chi^2_\alpha(q) + q\cdot\log(n) + C}\right]^{\frac{q}{2}}. \quad (11)$$

The refined version to linear models, for this calibration is obtained when evaluated in (7)

$$B(\alpha, q, n, b) = -\alpha \log(\alpha)\Gamma(q/2) b^{\frac{n-j}{2(n-1)}} \left[\frac{2(n-1)}{(g_{n,\alpha}(q) + \log(b) + C)(n-j)}\right]^{\frac{q}{2}} \quad (12)$$

in this case, we only consider $\xi_0 = 1$.
In all cases, n is the *effective sample size* (Berger *et al* 2014)

Obtaining bounds for $P(H_0|Data)$

The lower bounds for Bayes Factors RLB_ξ in (10) and (12) can be used to produce bounds for the posterior probability of the null hypothesis H_0. . Since for any Bayes factor B_{01}

$$B_{01} \geq B_L(p, \xi_0) \quad \text{con} \quad \xi_0 \geq 1, \text{ fixed but arbitrary,}$$

a lower bound for the posterior probability of the null hypothesis can be obtained as:

$$\min P(H_0|Data) = \left[1 + \frac{1}{B_L(p, \xi_0)}\right]^{-1}. \tag{13}$$

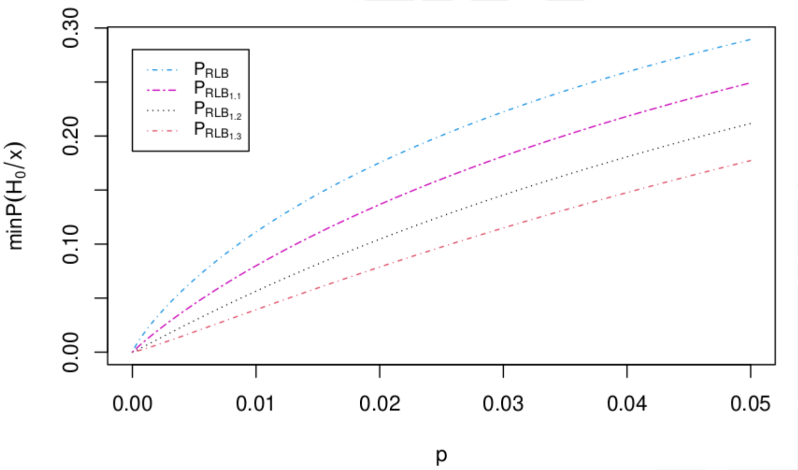


Figure: Lower bound for posterior probabilities for the null hypothesis H_0 (in 13) for $\xi_0 = 1, \xi_0 = 1.1, \xi_0 = 1.2, \xi_0 = 1.3$.

Example:Testing Equality of Two Normal Means

$H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$, with known non-equal variances, σ_1^2 and σ_2^2

Define $\alpha = (\mu_1 + \mu_2)/2$ and $\beta = (\mu_1 - \mu_2)/2$. Then we can write this problem using linear models.

$$\mathbf{Y} = \mathbf{B}\binom{\alpha}{\beta} + \epsilon \text{ with } \mathbf{B} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{pmatrix}$$

We want to compare $M_0 : \beta = 0$ versus $M_1 : \beta \neq 0$.

Then for the adjustments of the Bayes factor based in the PBIC strategy,

$$C = -2 \log \frac{(1 - e^{-v})}{\sqrt{2} v}$$

$v = \frac{\hat{\beta}^2}{d(1+n^e)}, d = \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right), n^e = \max \left\{ \frac{n_1^2}{\sigma_1^2}, \frac{n_2^2}{\sigma_2^2} \right\} \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right).$

A special case is the standard test of equality of means when $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Then

$$n^e = \min \left\{ n_1 \left(1 + \frac{n_1}{n_2} \right), n_2 \left(1 + \frac{n_2}{n_1} \right) \right\}.$$

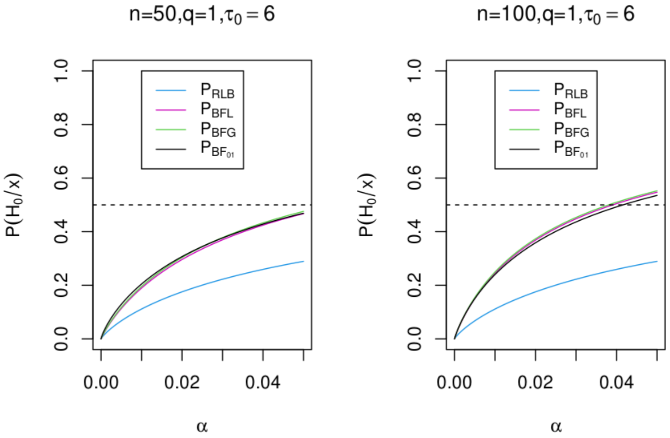
On the other hand, define $\delta = \mu_1 - \mu_2$ with $\sigma_1^2 = \sigma_2^2 = \sigma^2$
$H_0 : \mu_1 = \mu_2 \leftrightarrow \delta = 0$ vs $H_0 : \mu_1 \neq \mu_2 \leftrightarrow \delta \neq 0$
Assigning priors

- $\delta|\sigma^2, H_1 \sim Normal(0, \sigma^2/\tau_0), \tau_0 \in (0, \infty)$
- $\pi(\sigma^2) \propto 1/\sigma^2$ for both H_0 and H_1.

The Bayes factor is:

$$BF_{01} = \left(\frac{n+\tau_0}{\tau_0}\right)^{1/2} \left(\frac{t^2 \frac{\tau_0}{n+\tau_0} + l}{t^2 + l}\right)^{\frac{l+1}{2}}$$

where $t = \frac{|\bar{\mathbf{Y}}|}{s/\sqrt{n}}$ is a t-statistic with $l = n - 1$ degrees of freedom and $n = n_1 + n_2$ (see Roger 2018).



Posterior probability for the null hypothesis H_0 for $n = 50$ and $n = 100$ using the Bayes factor RLB_{ξ_0} with $\xi_0 = 1$, the Bayes factor BF_{01}, the Bayes factor BFL and BFG

Example: Fisher's Exact Test

(Example of pseudo-p-value, see the example 8.3.30 in Casella and Berger 2001).

Let S_1 and S_2 be independent observations with $S_1 \sim binomial(n_1, p_1)$ and $S_2 \sim binomial(n_2, p_2)$.

$$H_0 : p_1 = p_2 \text{ vs } H_1 : p_1 \neq p_2.$$

Under H_0, let p be the common value for $p_1 = p_2$. The the joint pmf of (S_1, S_2) is

$$f(s_1, s_2|p) = \binom{n_1}{s_1}\binom{n_2}{s_2} p^{s_1+s_2}(1-p)^{n_1+n_2-(s_1+s_2)}$$

The conditional pseudo-p-value is

$$p(s_1, s_2) = \sum_{j=s_1}^{\min\{n_1, s\}} f(j|s), \tag{14}$$

the sum of hypergeometric probabilities, with $s = s_1 + s_2$.

It does not seem to be simple to estimate the appropriate ξ_0 that best fits the pseudo-p-value in (14). Some arbitrary values will be used.

We will compare the performance of the approximate probabilities for H_0 with those based on a Bayes Factor calculate using the following prior .

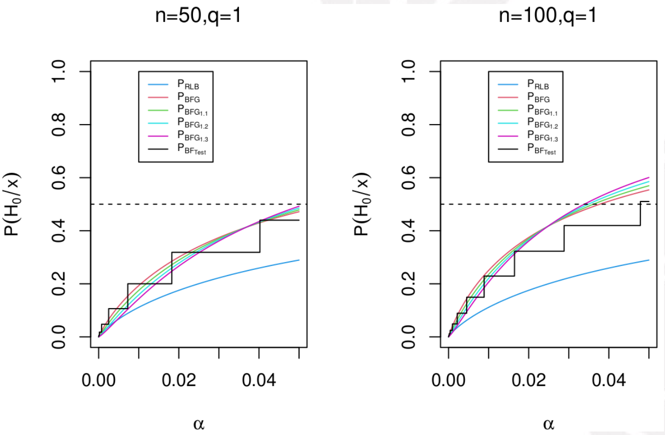
$$\pi(p) = \begin{cases} \pi_0 & p = (p_1 = p_2) \\ \pi_1 g_1(p) & p \neq (p_1 = p_2) \end{cases}$$

It can be shown that the Bayes Factor is is

$$B_{01} = \frac{f(s|(p_1 = p_2))}{m_1(s)}.$$

Now, if we take $g_1(p) = Beta(a, b)$ such that $E(p) = \dfrac{a}{a+b} = (p_1 = p_2)$, then

$$BF_{Test} = \frac{B(a, b)}{B(s + a, n_1 + n_2 - s + b)} p^s (1 - p)^{n_1 + n_2 - s}.$$



Posterior probability for the null hypothesis H_0 of equality of proportions in Fisher Exact Test for $n = 50$ and $n = 100$, using the Bayes factor RLB_{ξ_0} with $\xi_0 = 1$, the Bayes factor BF_{Test}, the Bayes factor BFG_{ξ_0}, and the Bayes factor BFG.

Discussion and Final Comments

- The adaptive α provides guidance for adjusting significance to the sample size. The Linear Model version incorporates not only the sample size and the difference of dimensions, but also the information provided by the predictors or the design, and particularly their correlations, correcting for co-linearity.
- The adaptive α is simple to use, and gives equivalent results than a sensible Bayes Factor, like Bayes Factors with Intrinsic Priors
- These results make use of state of the art large sample approximations of Bayes Factors like the PBIC and can be coupled with recent sensible base thresholds like $\alpha = 0.005$, Benjamin et al (2017)

Discussion and Final Comments

- Lower bounds have been an important development to give practitioners alternatives to classical testing with fixed α levels, but their usefulness is limited for large sample sizes. The calibrations proposed here can alleviate this problem.
- The (approximate) Bayes factors (10) and (12) are simple to use and provide results equivalent to the sensitive p-value-Bayes factors of hypothesis tests. We extend the validity of the approximation for "pseudo-p-values" which are ubiquitous in statistical practice.
- It's our hope that this methods will help to improve the replicabilitiy of scientific findings, making Bayesian Hypothesis Testing more practical and in line with familiar concepts of the traditional NHST.

The Department for Mathematics of the University of Puerto Rico, Rio Piedras Campus is announcing a tenure-track position in Statistics or Applied Probability.

- https://www.uprrp.edu/jobs
- maria.perez34@upr.edu
- **Deadline for documents: October 2, 2022**