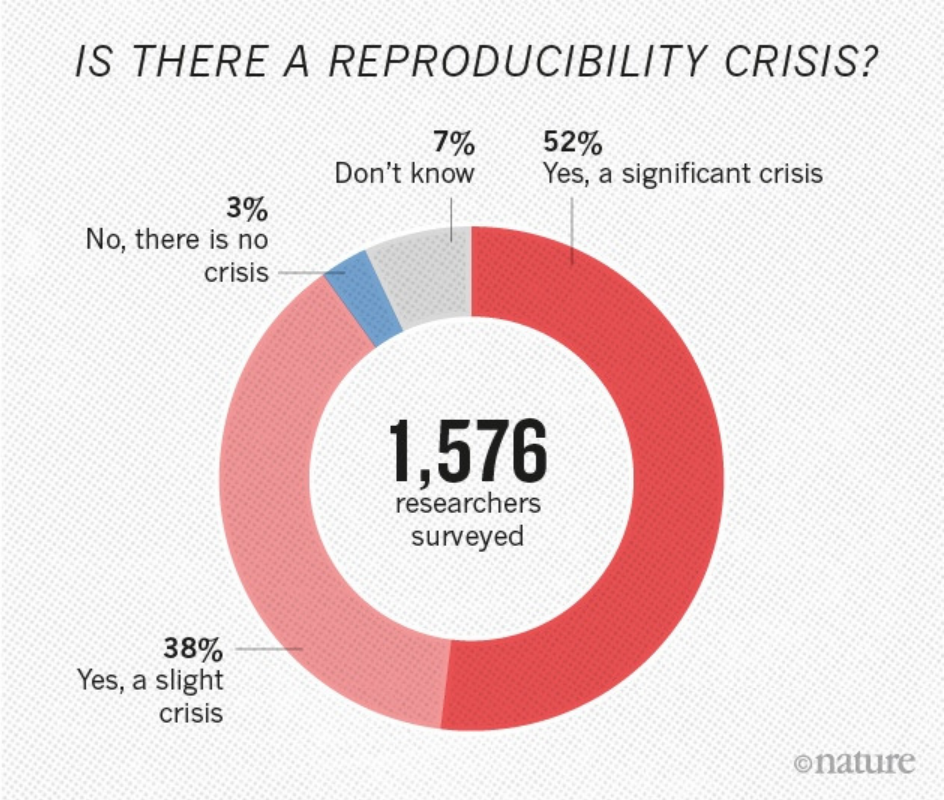
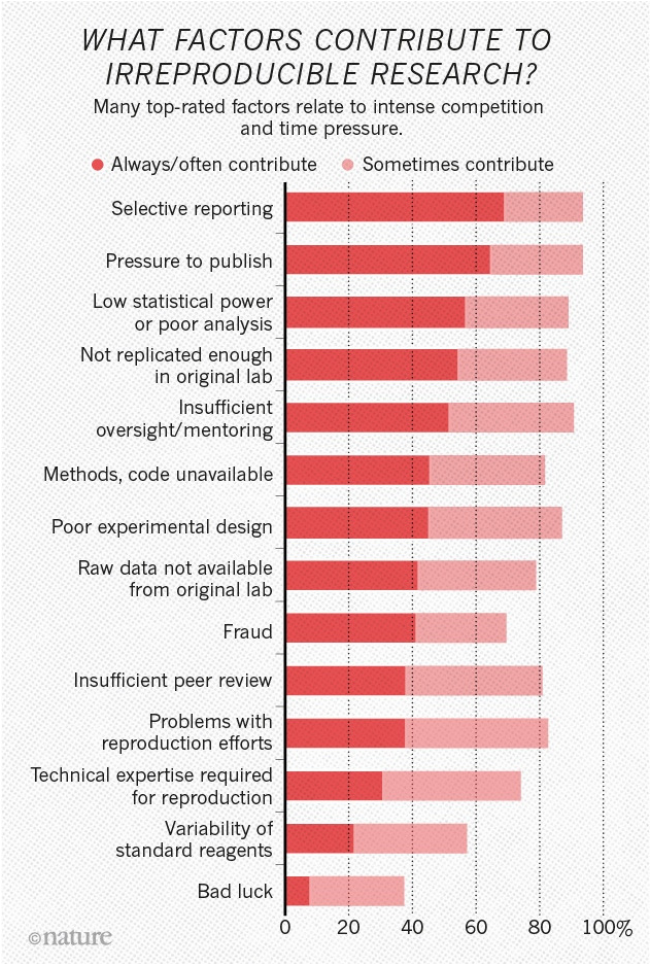
Reflecting about a magic formula converting P-values into Bayes Factors

Reproductibility crisis (Baker, Nature, 2016,533)

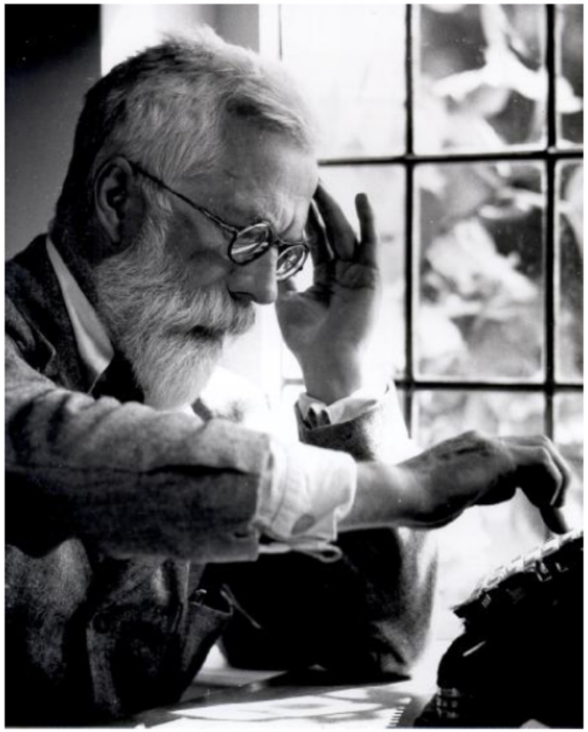


Reproductibility crisis: factors involved



Significance testing & P-value

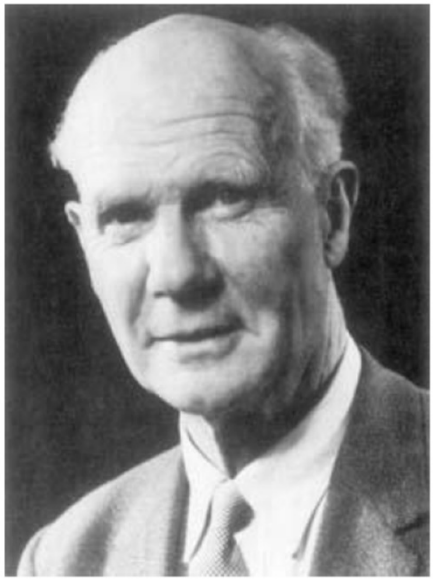
Sir Ronald Fisher (1890-1962)



Hypothesis testing/NP

Jerzy Neyman (1894-1981)

Egon Pearson (1895-1980)



NHST as a synthesis

- Most statisticians do not make the difference between Fisher P-values/Significant testing and Neyman-Pearson hypothesis testing « Null Hypothesis Significance Testing » (Lecoutre & Poitevineau, 2014. The Significant Test Controversy Revisited)
- « It is an incoherent mismah of some of Fisher's ideas on one hand and some of the ideas of Neyman and ES Pearson on the other » Gigerenzer, 1993
- « I don't care about the people, Neyman, Fisher, and Pearson. I care about what researchers do. They do something called NHST, and it's a disaster", Gelman, 2019
- Ex: ICH E9 Guidelines for Clinical Trials
- Synthesis responsible for many bad uses of p-values (Greenland, Senn et al, 2016)

P-value vs Tests d'hypothèses (Biau et al, 2010)

Fisher's p value	Hypothesis testing
Ronald Fisher	Jerzy Neyman and Egon Pearson
Significance test	Hypothesis test
p Value	α
The p value is a measure of the evidence against the null hypothesis	α and β levels provide rules to limit the proportion of errors
Computed a posteriori from the data observed	Determined a priori at some specified level
Applies to any single experiment	Applies in the long run through the repetition of experiments
Subjective decision	Objective behavior
Evidential, ie, based on the evidence observed	Nonevidential, ie, based on a rule of behavior

P-value/The 2016 ASA statement

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016): The ASA's statement on p-values: context, process, and purpose, The American Statistician, DOI: 10.1080/00031305.2016.1154108

1. *P-values can indicate how incompatible the data are with a specified statistical model.*

2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*

3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*

4. *Proper inference requires full reporting and transparency*

5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*

6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

What to do?

- Cleaver use of p-values (ASA statement, 2016)
- Threshold of significance
  - Move from 5 p 100 to 5p1000 (Benjamin, Berger et al, 2018)
  - Total abandon (Amrheim, Greenland, Mc Shane, 2019, Nature)
- Ban of p-values (Ex New England Journal of Medicine)
- Ban of hypothesis testing (Mc Shane, Gelman, Robert, 2019)
- Other criteria
  - Effect size & confidence interval
  - Bayes Factor
  - …

What about it in XKCD?

Propositions

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Moving to a World Beyond "$p < 0.05$"

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p<0.05$", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

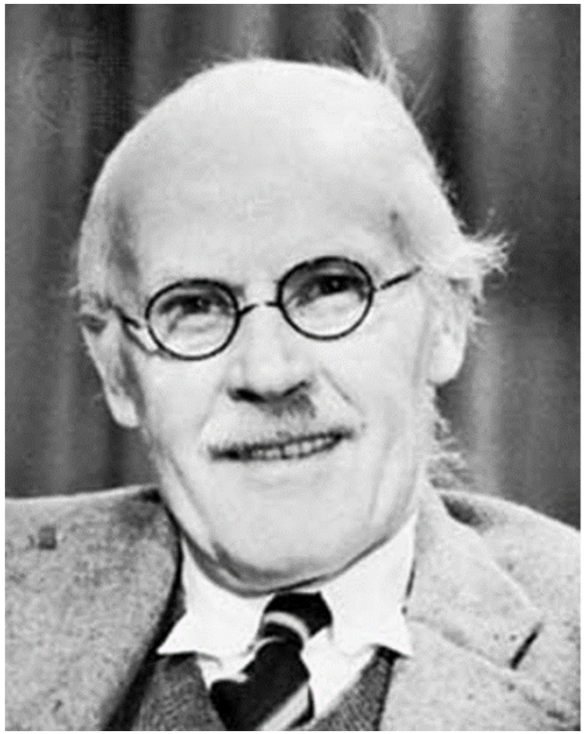
To link to this article: https://doi.org/10.1080/00031305.2019.1583913

Three recommendations for improving the use of p-values

- Benjamin & Berger (2019) *The American Statistician*, 73, 186-191
- R 0.1-Refer to discoveries,with a p-value between 0.05 and 0.005 as « suggestive » rather than « significant »
- R 0.2-When reporting a p-value, p, in a test of the null hypothesis H0 vs an alternative H1, also report that the data-based odds of H1 to H0 being true are at most 1/[-eplogp]
- R 0.3 –Determine and report your prior odds of H1 to H0 and derive and report the final (posterior) odds of H1 to H0

Bayes Factor

Sir Harold Jeffreys (1891-1989)



Posterior odds and BF

Bayes rule $\underbrace{\frac{\Pr(H_1 \mid y)}{\Pr(H_0 \mid y)}}_{\text{"Posterior odds"}} = \underbrace{\frac{\Pr(Y = y \mid H_1)}{\Pr(Y = y \mid H_0)}}_{\text{"Bayes Factor" } B_{10}} \times \underbrace{\frac{\Pr(H_1)}{\Pr(H_0)}}_{\text{"Prior odds"}}$

("K^{-1}" de Jeffreys,1939), proposed independently by Turing at Blechtley Park (Good,1979)

"Weight of evidence" defined as $10\log_{10} B_{10}$ (deciban)

Posterior probability & BF

$$\rho_{10} = \Pr(\mathrm{H}_1) / \Pr(\mathrm{H}_0)$$

$$\Pr(\mathrm{H}_1 \mid y) = \frac{B_{10}\rho_{10}}{1 + B_{10}\rho_{10}} \quad \Pr(\mathrm{H}_0 \mid y) = \frac{1}{1 + B_{10}\rho_{10}}$$

$$\Pr(\mathrm{H}_1 \mid y, \rho_{10} < 1) < \Pr(\mathrm{H}_1 \mid y, \rho_{10} = 1)$$

BF Calibration

BF Calibration according to Jeffreys

$K^{-1} = B_{10}$	$\Pr(H_1 \mid \mathbf{y})^*$	Deciban (dB)	Deviance (ΔD)	Strength of evidence against the null
<1	$>1/2$	<0		0-Null hypothesis supported
1.0 à 3.16	0.50 à 0.76	0 à 5	0 à 2.3	1-Not worth than a bare mention
3.16 à 10	0.76 à 0.91	5 à 10	2.3 à 4.6	2-Substantial
10.0 à 31.62	0.91 à 0.97	10 à 15	4.6 à 6.9	3-Strong
31.62 à 100	0.97 à 0.99	15 à 20	6.9 à 9.2	4-Very strong
>100	>0.99	>20	>9.2	5-Decisive

$K = B_{01} = f(y \mid H_0) / f(y \mid H_1)$: « grade of decisiveness of evidence », Jeffreys (1961) Appendix B

$\Pr(H_1 \mid \mathbf{y})^*$ setting $\Pr(H_0) = \Pr(H_1) = 1/2$

Deciban pertaining to K^{-1}: $dB = 10 \log_{10} B_{10}$ (Turing A, Good IJ, 1940)

Deviance defined as : $\Delta D_{01} = D_0 - D_1 = -2 \log \mathrm{L}_0 / \mathrm{L}_1$ (reduced vs complete models)

Upper bound of BF10

Sellke, Bayarri & Berger (2001) proposed

$$BF_{10}(p) \leq -\frac{1}{ep \log p}$$

for $p < 1/e$, else $B_{10}(p) = 1$

see Vovk (1993) Held & Ott (2018)

Distribution of P-values

Example: unilateral test under normality

$H_0 : \mu = \mu_0$ vs $H_1 : \mu \geq \mu_0$

$X_i \sim_{id} \mathcal{N}(\mu, \sigma^2)$ $i = 1,..,n$ σ known

$p_{obs} = \Pr(T_{H_0} \geq t_{obs})$ $t_{obs} = \sqrt{n}(\bar{x} - \mu_0)/\sigma$

$T_{H_0} \sim \mathcal{N}(0,1)$: Distribution of statistical test under H_0

If now $\bar{x}$ sampled from $\bar{X} \Rightarrow$ P random variable

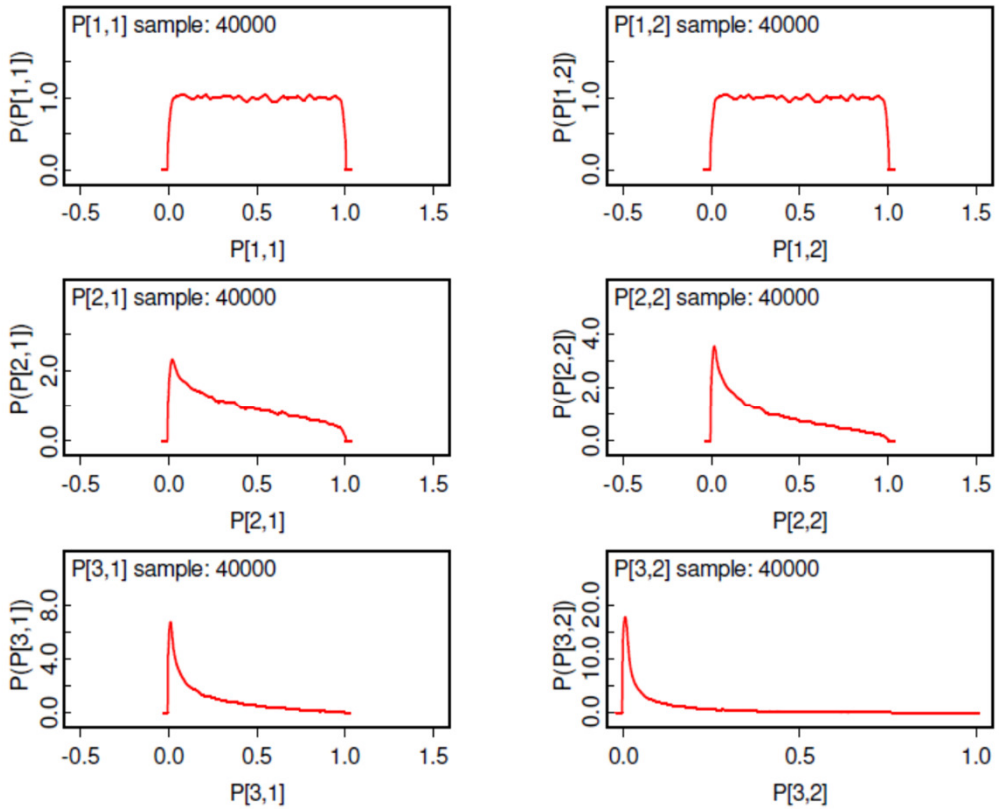
$\bar{X} = \mu_0 + \Delta\mu + Z\sigma/\sqrt{n}$ where $Z \sim \mathcal{N}(0,1)$

$$P = \Pr(T_{H_0} \geq \sqrt{n}\Delta\mu/\sigma + Z) = 1 - \Phi(\delta + Z) = \Phi(-\delta - Z)$$

$$\boxed{\text{P-val} \mid H_1 = \Phi[\mathcal{N}(-\delta, 1)]}$$

$$\delta = 0 \Rightarrow \boxed{\text{P-val} \mid H_0 = \Phi[\mathcal{N}(0,1)] \sim \mathcal{U}(0,1)}$$

Distribution of p-values/suite



P[i,j]: P-value for i=1 (d=0) , 2 (d=0.10), 3 (d=0.5) and j=1 (N=20) and 2 (N=50)

Deriving BF10 from p-values

$$H_0 : p \sim U(0,1)$$

Choice of a generic function

$H_1 : f(p) \sim$ "Standard power function distribution"

$$H_1 : p \sim Beta(\xi, 1),\ 0 < \xi \leq 1$$

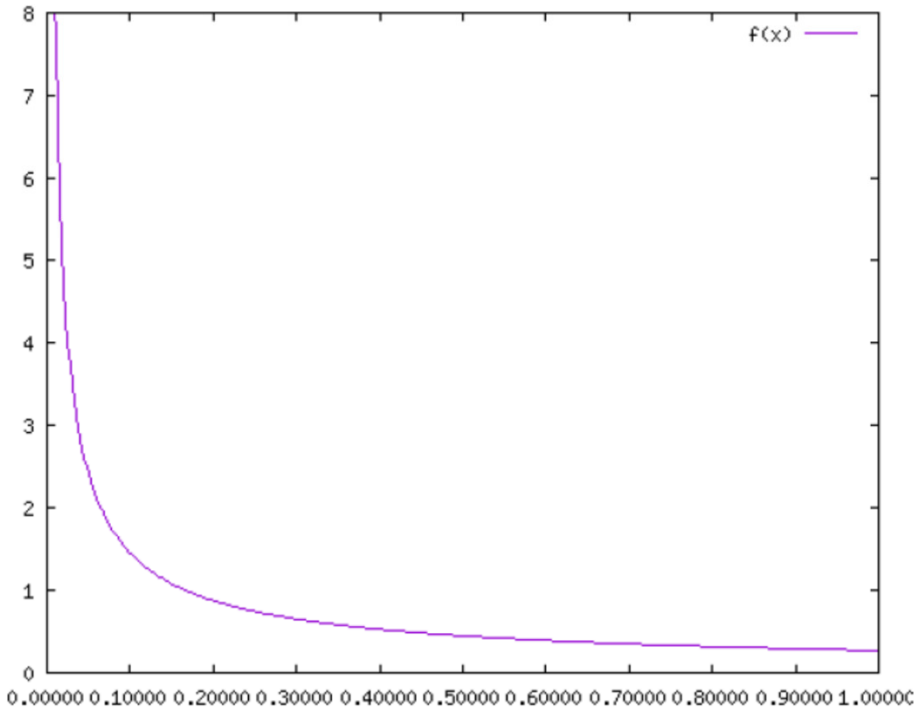
$f(p \mid \xi) = \xi p^{\xi - 1}$ (decreasing function)

Distribution of p-values under H1

You have entered : $f(x) = 0.27(x^{-0.73})$.

f(0.005) = 12.91563195995249156 . (click on a value to check its meaning in another window)

$$\int_{0.0}^{1.00} f(x)\mathrm{dx} = 0.99999999999999976313$$



Upper bound of BF10/proof

$$BF_{10}(p) = \frac{\int_0^1 f(p \mid \xi)\pi(\xi)d\xi}{f(p \mid \xi = 1)}$$

$$0 \le \int_0^1 f(p \mid \xi)\pi(\xi)d\xi \le \left[\text{Max}_\xi \, f(p \mid \xi)\right] \underbrace{\int_0^1 \pi(\xi)d\xi}_{1}$$

$$Ici\ f(p \mid \xi) = \xi p^{\xi-1}, 0 < \xi \le 1 \Rightarrow \xi_{ML} = \min(-1/\log p, 1)$$

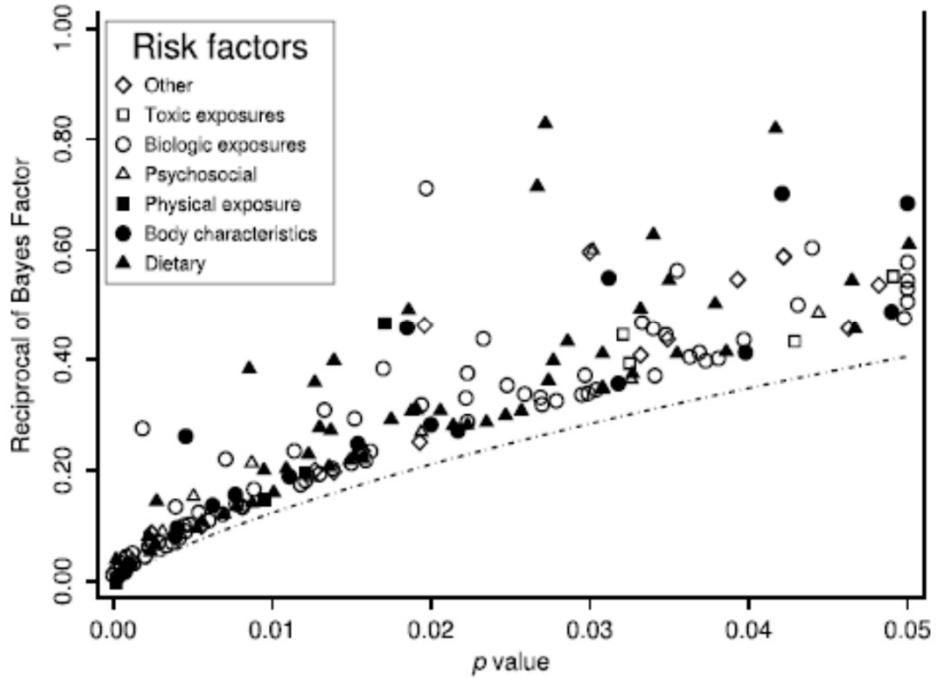
$$\boxed{BF_{10}(p) \le BFB(p) = -\frac{1}{ep\log p}} \text{ for } p < 1/e \text{ , 1 otherwise}$$

$$p = 0.05 \Rightarrow BFB_{10} = 2.46 \quad (\Pr(H1 \mid p) \le 0.71$$

BFB viewed as GLR (W. Edwards et al,1963)

Checking BFB10(p) on real data

Bayarri et al, 2016, J Math Psycho 272 studies (Ioannidis, 2008)



BFB10(p)/Bias towards H1

- BFB10 favors H1
- Quantify amount of bias towards H1
- Test priors on ξ (Foulley,2019)
  - Uniform
  - PC prior (Simpson & al, 2015)

Upper bound of BF10: uniform prior

Jeffreys' prior $\pi_J(\xi) \propto \xi^{-1}$ improper

$\xi \in]0,1[$ uniform

Analytical expression of $\int_0^1 f(p \mid \xi)\, d\xi$

$$\boxed{BFU_{10} = \frac{1}{\log p}\left(\frac{1-p}{p \log p} + 1\right)}$$

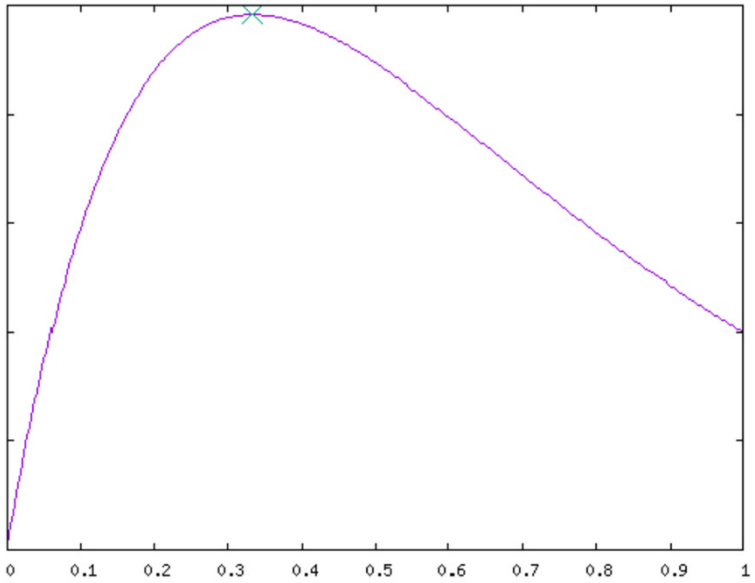
$$p = 0.05 \Rightarrow BFU_{10} = 1.78 \quad (\Pr(\mathrm{H1} \mid \mathrm{p}) \leq 0.64$$

BFB vs BFU (p fixed)

You have entered : $f(x) = x \times 0.05^{x-1}$.

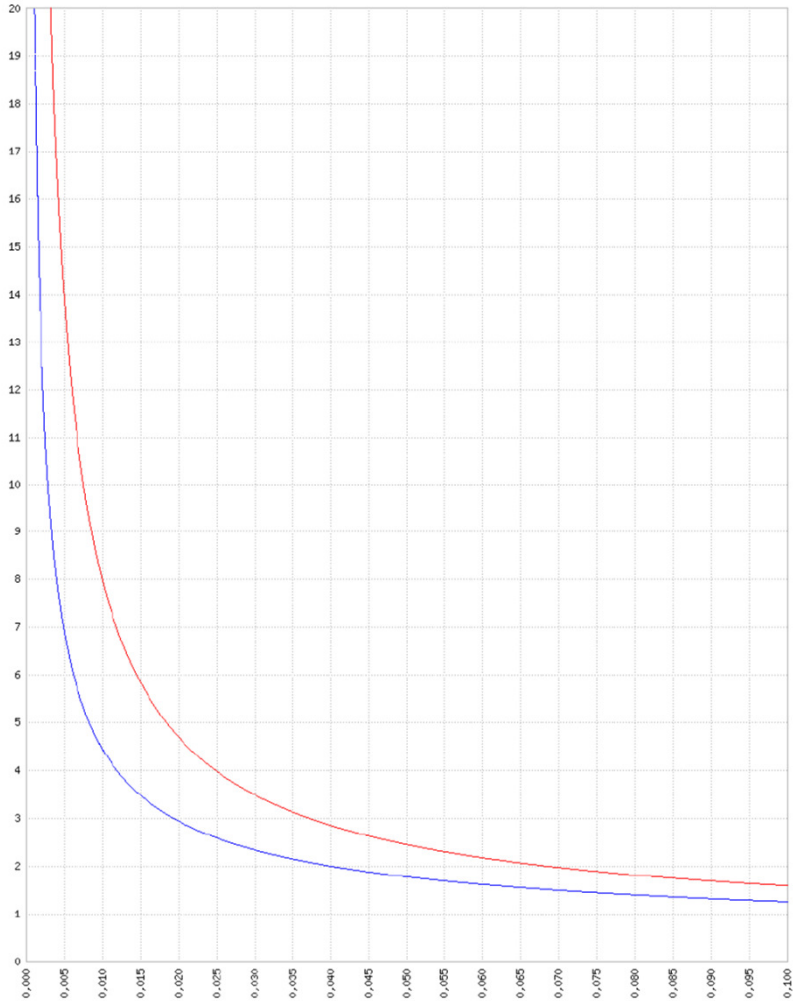
$\int_{0.001}^{0.999} f(x)\mathrm{dx} = 1.78231120307$

Click on a root or an extremum of the curve, to compute its value up to 12 digits.

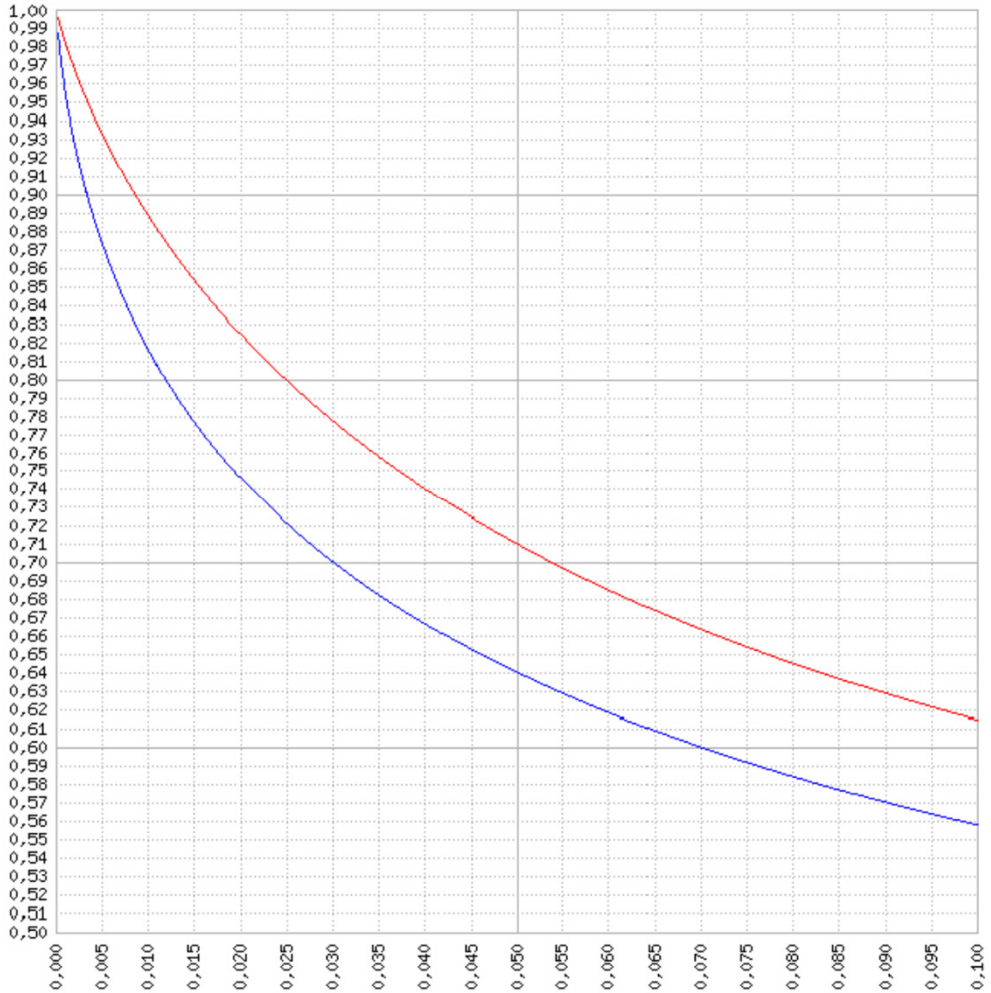


Maximum local : f(0.333808200695) = 2.4560234866. (click on a value to check its meaning)

BFB (red) vs BFU (blue)



Pr(H1/p) bsup (red) vs Pr(H1/p) unif (blue)



PC priors

Looking for priors decreasing functions of $\xi = 1$ (H_0) to $\xi \to 0$ (H_1)

A priori PC penalizing H_1 vs H_0 as a function of the Kullback-Leibler distance between $f_1(p)$ and $f_0(p)$ (Simpson et al, 2015)

$$KL(f_1, f_2) = \int f_1(t) \ln \frac{f_1(t)}{f_0(t)} dt \quad \text{D}=\sqrt{2KL}$$

Constant rate of complexity penalty $\pi_d(D+\delta) / \pi_d(D) = r^{\delta}$

$(0 < r < 1)$ implies an exponential form of the prior on D

$$\pi_d(D) = \lambda \exp(-\lambda D) \; ; \; \pi(\xi) = \lambda \exp\left[-\lambda D(\xi)\right] \left| D'(\xi) \right|$$

$\lambda > 0$ parameter monitoring the change of $\pi(\xi)$ $\lambda = -\log r$

PC priors/suite

$\pi(\xi) = \lambda \exp\left[-\lambda D(\xi)\right] \left| D'(\xi) \right|$ où $\mathrm{D}=\sqrt{2KL}$

$KL(\xi) = \ln \xi + \xi^{-1} - 1 \quad D'(\xi) = \xi^{-2}(\xi - 1)(2KL)^{-1/2}$

$\pi(\xi), f$ monotically incresing function for $\lambda \geq 4/3$

$\lambda = 4/3$ least defavorable to H_1 among those penalising them

Numerical computation de $\int_0^1 f(p \mid \xi)\pi(\xi) d\xi$ via

$https://wims.auto.u-psud.fr/wims/$

PC priors/suite

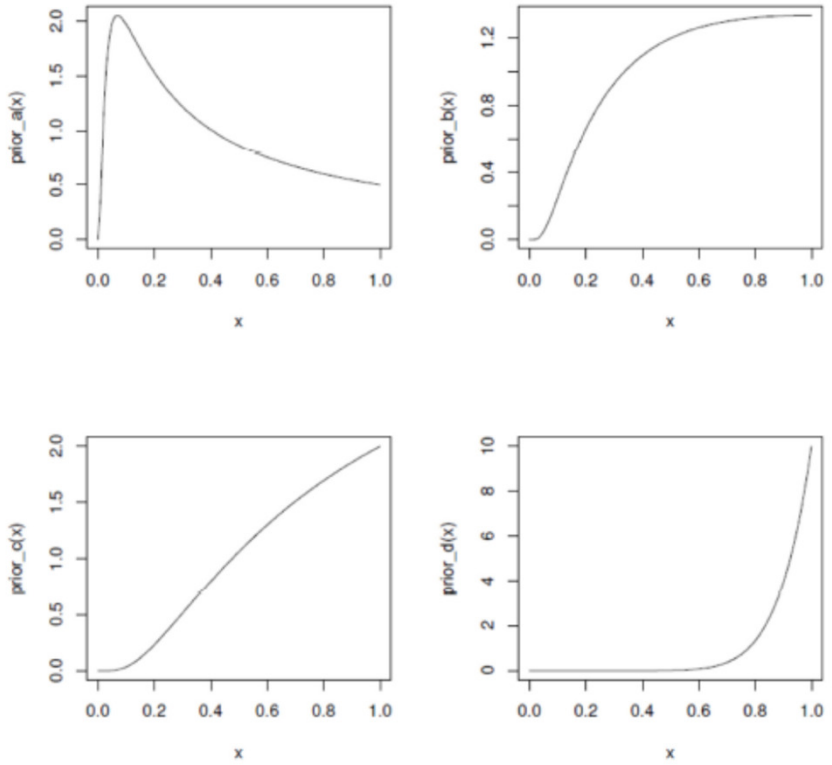


Figure 1. Examples of PC prior pdf's for the following λ values: a=0.5, b=4/3; c=2 and d=10

BF10 under PC priors

p	0.10	0.05	0.01	0.005	0.001
BFB	**1.60**	**2.46**	**7.99**	**13.9**	**53.3**
Pr(H1|p)	**0.62**	**0.71**	**0.89**	**0.93**	**0.981**
BFU	1.26	1.78	4.45	6.90	20.8
Pr(H1|p)	0.56	0.64	0.82	0.87	0.954
BFP	1.36	1.84	4.04	5.89	15.3
Pr(H1|p)	0.57	0.65	0.80	0.85	0.939

Table 1. Bayes Factors and corresponding probabilities of the alternative hypothesis Pr(H1|p) under different prior distributions

BFB: (upper) bound based on generalized likelihood ratio by Benjamin & Berger (2019)

BFU: uniform prior

BFP: penalizing complexity prior with $\lambda=4/3$

Checking the rule in standard testing situations

- Rule BFB(p) works well for
  - Two-sided z-tests
  - One-sided z-tests
  - Two-sided t-tests with df≥10
- Rule plausible for Chi-squared tests
- “Even if the BFB(0.05) = 2.46 and BFB(0.01) = 7.99 values are dubious in some cases, they are still likely to give a better assessment of the evidence against H_0 than the typical “transpose the conditional" interpretation of .05 and .01”

  (Sellke, 2012)

Discussion

- Issues with BFB10 as BF10 upper bounds
  - OK if bounds small then do not reject H0; uncertainty otherwise
  - Bias towards H1 for BFB10, but BFB10 remains less optimistic wrt H1 than supposed from the p-value: good warning against H1 (adjustable according to choice of other priors)
- A fixed P-value does mean the samething at different sample sizes (decreasing evidence with n incresing): not taken into account here
- BFB can be implemented without a prior elicitation & is a good entry to a complete Bayesian analysis (Goodman, 1999)
- Choice of prior under H1 very influential on BF
- Computing marginal likelihood not so easy (ban harmonic mean)

Discussion : other criteria

- Killeen replication and Lecoutre predictive probabilities (Lecoutre et al,2010)
- Posterior predictive p value (Gelman et al, 2004)
- Mixture models (Kamary et al, 2018)
- Analysis of credibility (AnCred, Matthews,2018)
- Severe testing (Mayo & Spanos, 2006)

Intuition about probability of replicating results (Tversky & Kahneman, 1971)

"Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory ($z = 2.23$, $p < .05$, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?"

References

- Assaf AG, Tsionas M (2018) Bayes Factors vs P-values. *Tourism Management*, 67, 17-31
- **Benjamin DJ, Berger JO, (2019) Three recommandations for improving the use of p-values. The American Statistician, 73, 186-191**
- Benjamin DJ, Berger JO, Johannesson M et al (2017) Redefine Statistical Significance. *Nature Human Behaviour*
- Berger JO (1985) Statistical theory and Bayesian analysis. Springer
- Berger JO (2003) Could Fisher, Jeffreys and Neyman have agreed on testing. (with discussion), *Statistical Science*, 18, 1-32
- Berger JO, Sellke T (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. *JASA*, 82, 112-122
- Berger JO, Delampady M (1987) Testing precise hypotheses. *Statistical Science*, 2, 317-352
- Berger JO, Perrichi LR (1996) The intrinsic Bayes factor for model selection and prediction. *JASA*, 91, 109-122
- Berger JO, Boukai B, Wang Y (1997) Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statistical Science*, 3, 133-160
- Bernardo JM (1999) Nested hypothesis testing: the Bayesian reference criterion. *Bayesian Statistics*, 6, 101-130
- Biau DJ, Jolles BM Porcher R (2010) P value and the Theory of Hypothesis Testing. *Clin Orthop Relat Res ,* 468,885-892
- Cumming G (2005) Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*. 16, 1002–1004
- Cumming G, Fidler F (2009) Confidence intervals. *Journal of Psychology*, 217, 15-26.
- Cumming G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. Routledge, New York

References/Cont.

- Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242
- Foulley JL (2013) Le paradoxe de Jeffreys-Lindley: pierre dans le jardin des fréquentistes ou épine dans le pied des bayésiens. Applibugs, 19/12/2013, https://www.researchgate.net/publication/259575305_applibugs13_12_19JLF
- Foulley JL (2019) Comment on « Three Recommandations for Improving the Use of p-values ». *The American Statistician*, https://doi.org/10.1080/00031305.2019.1668850

  https://www.tandfonline.com/eprint/KAJWATZ25ZPZRFT2P3TM/full?target=10.1080/00031305.2019.1668850
- Gelman A, Shalizi CR (2013) Philosophy and the practice of Bayesian statistics. *British J of Mathematical and Statistical Psychology*, 66, 8-38
- Goodman SN (1999) Toward evidence based statistics. 1: the P value fallacy. *Annals of Internal Medicine*, 130, 996-1004
- Goodman SN (1999) Toward evidence based statistics. 2: the Bayes Factor. *Annals of Internal Medicine*, 130, 1005-1013
- Greenland S, Senn SJ, Rothman K, Carlin JB, Poole C, Goodman SN, Altman DG (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337-350
- Held, L., and Ott, M. (2018), On p-values and Bayes Factors. *Annual Review of Statistics and Its Application*, 5, 393-519.
- Hubbard R, Bayrarri MJ (2003) Confusion over measures of evidence (p's) versus errors (α's) in classical testing. *The American Statistician*, 57, 171-178
- Hung HM, O'Neill RT, Bauer P, Kohne K (1997) The behavior of the P-value when the alternative hypothesis is true. *Biometrics*, 57, 11-22
- Ioannidis JPA (2005) Why most published research findings are false. *PloS Medicine*, 2, (8), 696-701
- Jeffreys H (1961) Theory of probability (3rd edition) Oxford-Clarendon Press
- Johnson VE (2013) Revised standards for statistical evidence. PNAS, 110 (48);19313-1931
- Kamary K, Mengersen K, Robert CP, Rousseau J (2014) Testing hypotheses via a mixture model; arXiv
- Kass RE, Raftery AE (1995) Bayes factors. *JASA*, 90, 773-795
- Killeen PR (2005) An alternative to null hypothesis tests. *Psychological Science*, 16, 345-353

References/Cont.

- **Lecoutre B, Poitevineau J (2014) The significance test controversy revisited: the fiducial Bayesian alternative. Springer Briefs in Statistics**
- Lecoutre B, Lecoutre M-P, Poitevineau J (2010) Killeen's probability of replication and predictive probabilities: How to compute, use and interpret them. *Psychological Methods*, 15, 158-171.
- Lindley DV (1957) A statistical paradox. *Biometrika*, 44, 187-192
- Mayo DG, Spanos A (2006) Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British J of Philosophy of Science*, 57, 323-357
- Rougier, J. (2019), p-values, Bayes Factors, and Sufficiency, *The American Statistician*, 73: sup1, 148-151.
- Royall RM (1986) The effect os sample size on the meaning of significance tests. *The American Statistician*, 40, 313-315.
- Sellke, TM. (2012) On the interpretation of p-values. Technical report #17-01, Department of Statistics, Purdue University.
- **Sellke, T., Bayarri, M.J., and Berger, J.O. (2001), Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55, 62–71.**
- Senn S (2001) Two cheers for P-values. *J of Epidemiology & Statistics*, 6, 193-204
- Spanos A (2013) Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science*, 80, 73-93
- Simpson, D.P., Rue, H., Martins, T. G., Riebler, A., Martins. T. G., and Sørbye, S.H. (2015), Penalising model component complexity: A principled, practical approach to constructing priors. arXiv:1403.4630v4
- Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian approaches to clinical trials and health-care evaluation. J Wiley &Sons.
- Tversky A., Kahneman, D (1971) Belief in the law of small numbers. *Psychological Bulletin*, 6, 105-110.
- Vovk V . (1993) Mellin transforms and asymptotics: harmonic sums. *Journal of the Royal Statistical Society*, B, 55, 317-351
- Wagenmakers EJ (2007) A practical solution to the pervasive problem of p values. *Psychonomic Bulletin & Review*, 14, 779-804