Bayesian Statistics and Reproducibility

NCS 2018, Paris

5th October 2018

Leonhard Held & Manuela Ott

Epidemiology, Biostatistics and Prevention Institute and

Center for Reproducible Science

University of Zurich[UZH]

Reproducibility in Drug Development

***Nature Reviews Drug Discovery* 10(9), 712-713 (2011)**

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

"With reasonable efforts (sometimes the equivalent of 3–4 full-time employees over 6–12 months), we have frequently been unable to reconfirm published data."

Reproducibility in Biomedical Research

***Nature* (2014)**

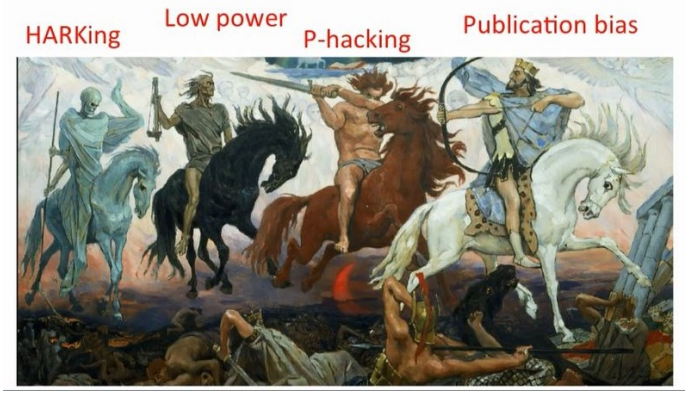
NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

"The recent evidence showing the irreproducibility of significant numbers of biomedical research publications demands immediate and substantive action."

The Four Horsemen of the Reproducibility Crisis

Dorothy Bishop, Oxford Reproducibility Lecture (2017)



Diagnostic Tests

Mr. Smith and Dr. Jones

Mr. Smith is worried that he has disorder D and consults Dr. Jones. Dr. Jones orders a test T that has only 2 outcomes, positive and negative, and Mr. Smith's test result comes back positive. Dr. Jones tells his patient that he has a 50% chance of having the disorder given this test result.

Mossman and Berger (2001)

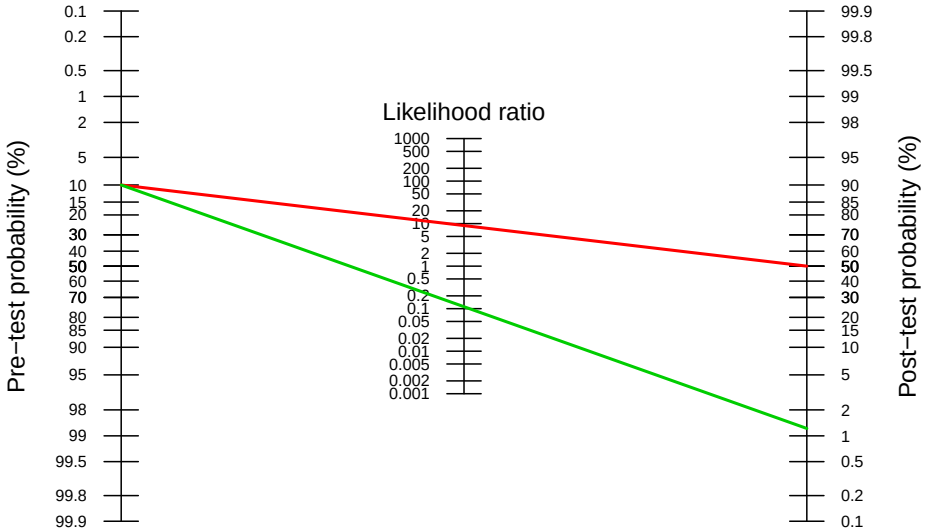
Diagnostic Tests

Mr. Smith and Dr. Jones cont.

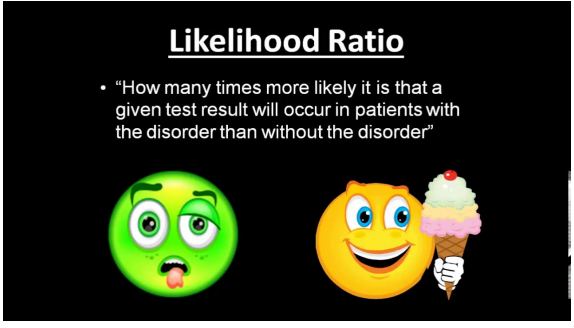
Mr. Smith asks his doctor how he came to this conclusion. Dr. Jones – an unusual physician who uses Bayes' theorem and discusses it with his patients – answers:

"Ten percent of men who are your age have D. Test T has 90% sensitivity and 90% specificity. Applying Bayes' theorem, your positive test result means that your chance of having D is 50%."

The Fagan Nomogram



Likelihood Ratio



https://www.youtube.com/watch?v=OssWmN9pm4Q

Diagnostic Tests

Bayes' theorem

$$\underbrace{\frac{\text{PPV}}{1-\text{PPV}}}_{\text{Posterior odds}} = \underbrace{\frac{\text{Sensitivity}}{1-\text{Specificity}}}_{\text{Likelihood ratio}} \times \underbrace{\frac{\text{Prevalence}}{1-\text{Prevalence}}}_{\text{Prior odds}}$$

PPV : positive predictive value

Statistical Tests

Two interpretations

- Test a point null hypothesis H_0 against an alternative H_1
- Use a p-value to quantify the evidence against the null hypothesis H_0

Two different interpretations of the evidence:

1. "p-less-than", *e. g.* $p < 0.05$
   - Dichotomization into "significant" and "non-significant"
   - Allows to study frequentist properties of "significant" tests
2. "p-equals", *e. g.* $p = 0.04$
   - Adequate interpretation for a single test, no loss of information

Colquhoun (2014, 2017)

Statistical Tests

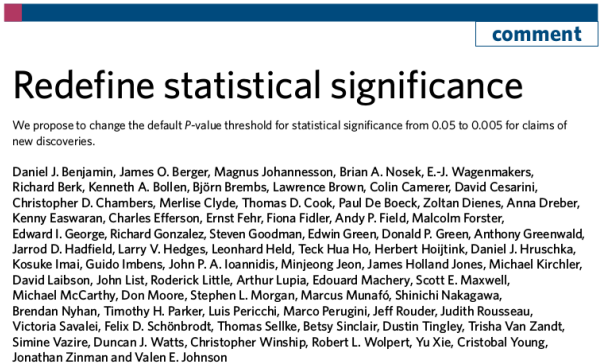
Bayes' theorem for "p-less-than" interpretation

$$\frac{\text{TPR}}{1-\text{TPR}} = \underbrace{\frac{\text{Power}}{\text{Type-I error rate}}}_{\text{Pre-experimental odds}} \times \frac{\Pr(H_1)}{\Pr(H_0)}$$

Bayarri *et al.* (2016)

$$\text{TPR} = \Pr(H_1 \mid \text{significance}) : \text{true positive rate}$$
$$\text{FPR} = 1 - \text{TPR} : \text{false positive rate}$$

Claims of New Discoveries



comment

Redefine statistical significance

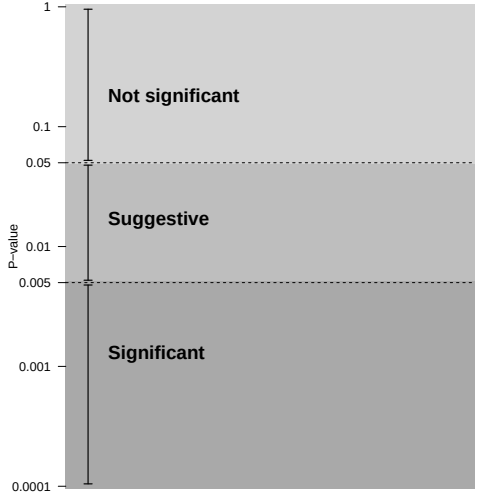
We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

"This simple step would immediately improve the reproducibility of scientific research in many fields."

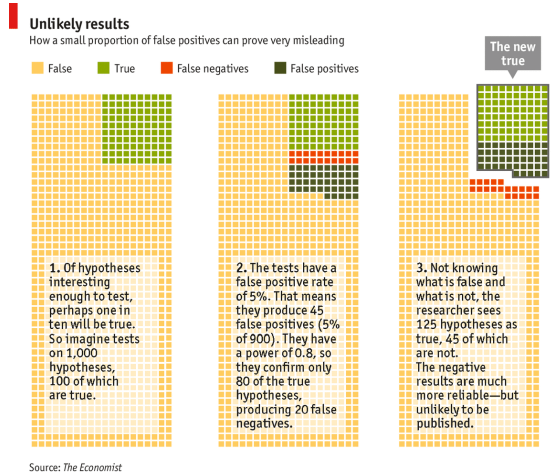
Lower the *p*-Value Threshold for Significance!

Suggestive evidence emphasizes the need for replication



Why Would You Do This?

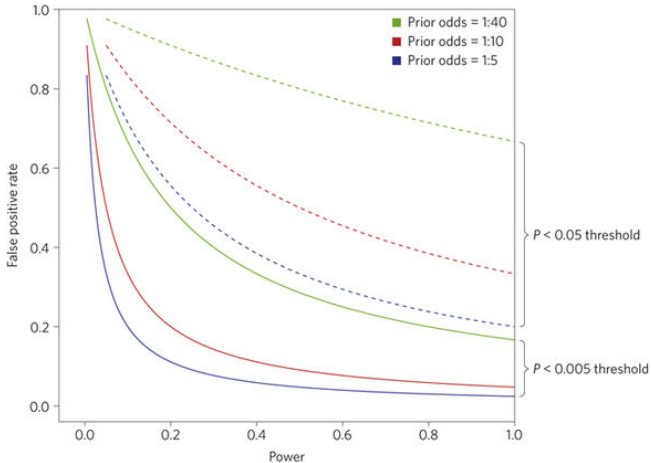
Why Most Published Resarch is Probably False



https://www.economist.com/blogs/graphicdetail/2013/10/daily-chart-2

The False Positive Rate

70% sample size increase for $p < 0.005$ threshold



Benjamin *et al.* (2018)

This Argument is Not New

Staquet et al (1979, Cancer Treat Rep, 1917–21)

The Delta and Epsilon Errors in the Assessment of Cancer Clinical Trials [1,2,3]

M. J. Staquet,[4,*] M. Rozencweig,[5,6] D. D. Von Hoff,[5,7] and F. M. Muggia [5,8,9]

SUMMARY

The error probabilities α and β are widely used to compute sample sizes and to analyze results of clinical trials. These errors are, however, not the only probabilities to consider when assessing results of clinical studies. The rate of false positive (δ) and false negative (ε) results allows one to determine if an experimental finding is likely to reflect the true situation in the population of interest. The δ error is generally high in randomized phase III and in early phase II clinical trials in cancer patients, whereas the ε error is relatively low in these settings. This is essentially due to the small probability of detecting a more effective treatment or a new chemotherapeutic agent active in cancer. The δ error could be considerably reduced by increasing the sample sizes and by restricting the allowance made for the α error, which should be set at a 1% level as a minimum requirement.

[Cancer Treat Rep 63:1917–1921, 1979]

Statistical Tests

Bayes' theorem for "p-equals" interpretation

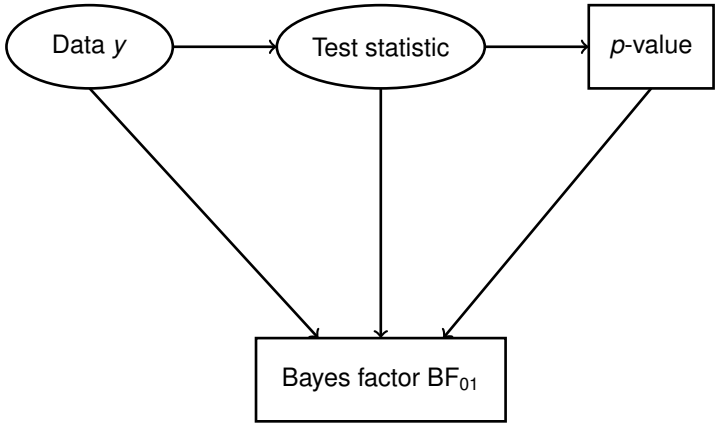
$$\frac{\text{FPR}}{1-\text{FPR}} = \underbrace{\frac{f(p \mid H_0)}{f(p \mid H_1)}}_{\text{Bayes factor BF}_{01}} \times \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\text{Prior odds}}$$

$$\text{FPR} = \Pr(H_0 \mid p) : \text{false positive risk}$$

Colquhoun (2017, 2018)

Bayes Factors

Data-based, test-based or *p*-based



Held & Ott (2018)

Interpretation of Bayes Factors

Table 2 Categorization of Bayes factors BF ≤ 1 into levels of evidence against H_0

Bayes factor	Strength of evidence against H_0		
	Jeffreys (1961)	Goodman (1999b)	Held & Ott (2016)
1 to 1/3	Bare mention		Weak
1/3 to 1/10	Substantial	Weak to moderate	Moderate
1/10 to 1/30	Strong	Moderate to strong	Substantial
1/30 to 1/100	Very strong	Strong	Strong
1/100 to 1/300	Decisive	Very strong	Very strong
<1/300			Decisive

Bayes Factors

Choice of the alternative

- Point null hypothesis H_0: $\theta = \theta_0$
- Two types of alternatives H_1:
  - Local, *e. g.* H_1: $\theta \sim \mathrm{N}(\theta_0, \sigma^2) \Rightarrow$ for exploratory studies
  - Simple, *i. e.* H_1: $\theta = \theta_1 \neq \theta_0 \Rightarrow$ for confirmatory studies
- The simple alternative is often chosen such that a given power (80%) is reached under a pre-specified type I error rate (5%).

Colquhoun (2017, 2018)

Minimum Bayes Factors

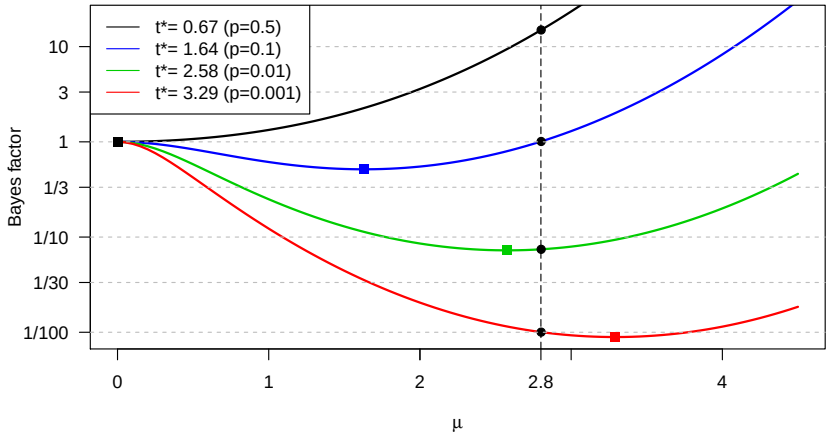
- Minimum Bayes factor (minBF): the smallest possible Bayes factor within a specific class of alternatives

$\rightarrow$ The smaller the minBF, the stronger the maximal evidence against H_0 within that class

- The minBF also depends on the
  - type of test *e. g.* z-test vs. t-test
  - sample size n
  - dimension of the parameter of interest.

Bayes and Minimum Bayes Factors

Simple alternative, 80% power for $\mu = 2.8$



Minimum Bayes Factors

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 4, 335–341
http://dx.doi.org/10.1080/00031305.2016.1209128

Taylor & Francis
Taylor & Francis Group

How the Maximal Evidence of *P*-Values Against Point Null Hypotheses Depends on Sample Size

Leonhard Held and Manuela Ott

Department of Biostatistics Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

Annual Review of Statistics and Its Application

On *p*-Values and Bayes Factors

Leonhard Held and Manuela Ott

Epidemiology, Biostatistics and Prevention Institute, University of Zurich, CH-8001 Zurich, Switzerland; email: leonhard.held@uzh.ch, manuela.ott@uzh.ch

The R package `pCalibrate`

- Implements various calibrations of *p*-values as minimum Bayes factors
- Available on CRAN (https://cran.r-project.org/package=pCalibrate)
- Covers different classes of alternatives and one- and two-sided tests

Examples of some functions:

- *p*-based minBFs: `pCalibrate(p, ...)`
- test-based minBFs
  - *z*-test: `zCalibrate(p, ...)`
  - *t*-test: `tCalibrate(p, ...)`
  - likelihood ratio test: `LRCalibrate(p, ...)`

p-Based Minimum Bayes Factors

For a two-sided p-value p

The "$-e\,p\log(p)$" calibration

$$\text{minBF}(p) = \begin{cases} -e\,p\,\log p & \text{for } p < 1/e \approx 0.37 \\ 1 & \text{otherwise.} \end{cases}$$

Sellke, Bayarri & Berger (2001)

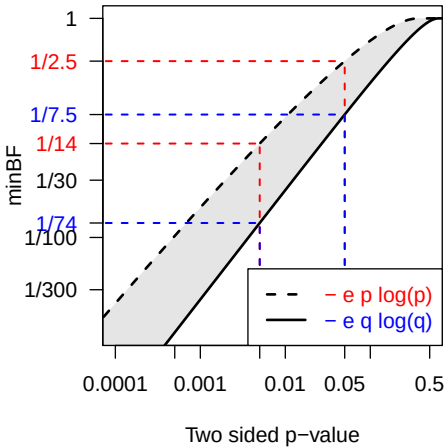
```
pCalibrate(p)
```

The "$-e\,q\log(q)$" calibration, where $q = 1 - p$.

Held & Ott (2018)

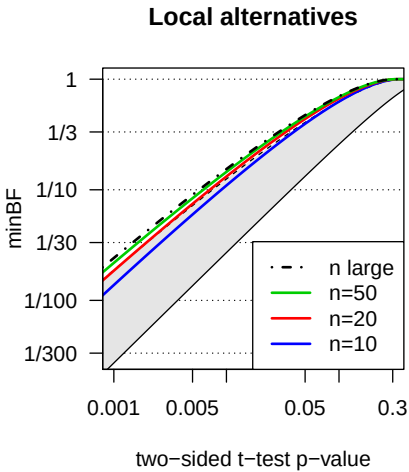
```
pCalibrate(p, alternative="informative")
```

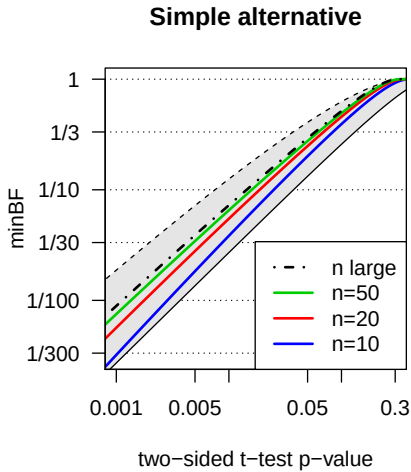
p-Based Minimum Bayes Factors



Minimum Bayes Factors

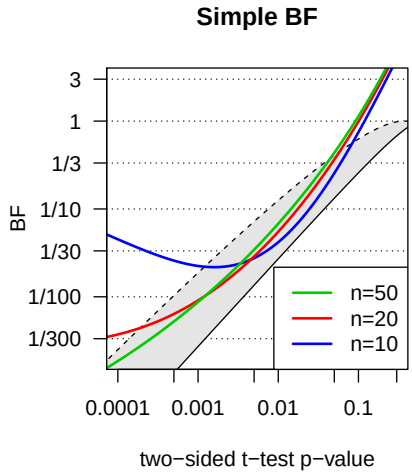
Two-sample t-test setting: `tCalibrate(...)`

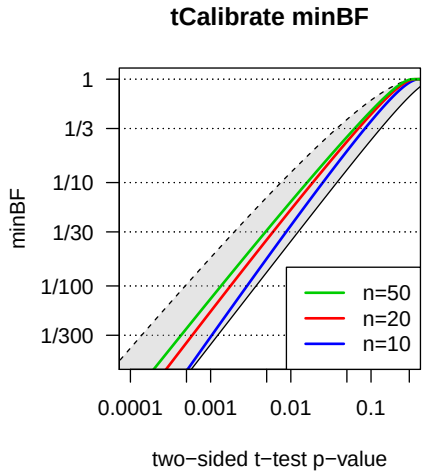




Bayes Factors vs. Minimum Bayes Factors

Simple alternative (at 80% power) in two-sample *t*-test





Back to Probabilities

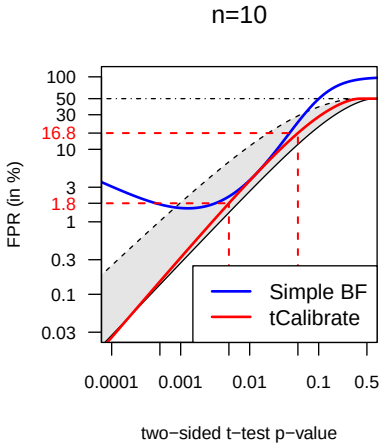
Forward – Reverse – Inverse Bayes

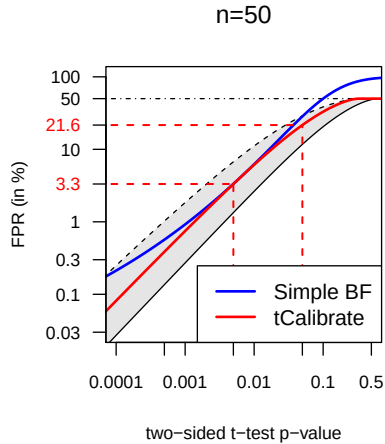
$$\frac{\text{FPR}}{1-\text{FPR}} = \text{BF}_{01}(p) \times \frac{\Pr(H_0)}{\Pr(H_1)}$$

Forward Bayes:	$\Pr(H_0) = 50\%$	$\xrightarrow{p}$	FPR
Reverse Bayes:	$\text{FPR} = 5\%$	$\xrightarrow{p}$	$\Pr(H_0)$
Inverse Bayes:	$\Pr(H_0) = 50\%$	$\xrightarrow{\text{FPR}}$	p

Forward-Bayes

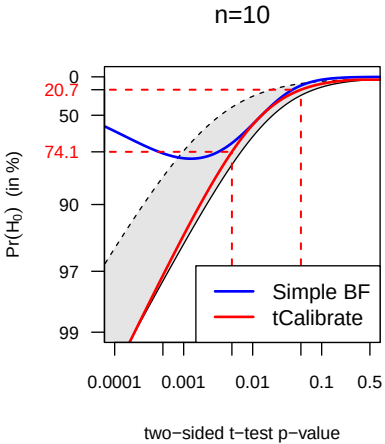
Under the assumption of equipoise: $\Pr(H_0) = \Pr(H_1) = 50\%$

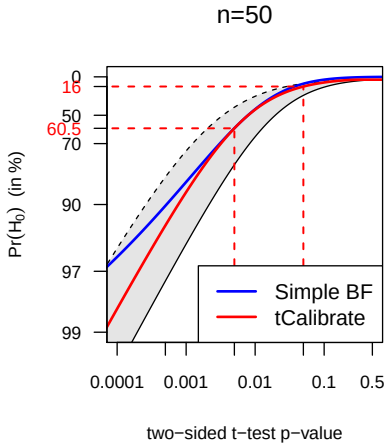




Reverse-Bayes

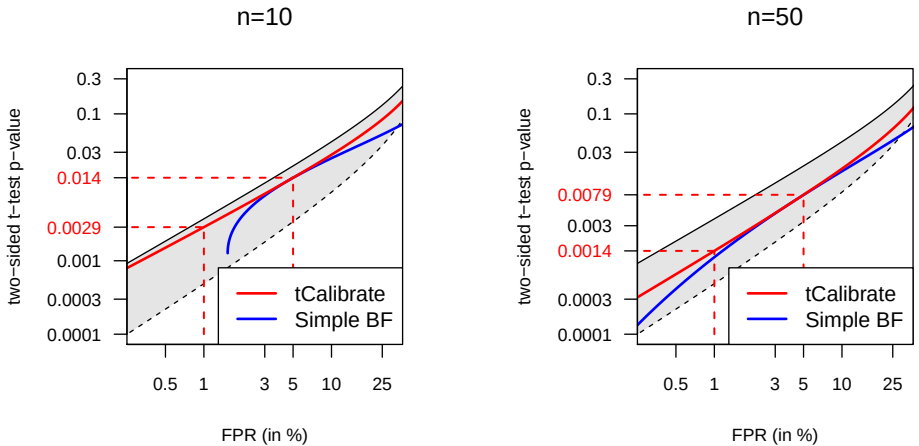
For a given p-value, determine (max) $\Pr(H_0)$ such that FPR $= 5\%$.





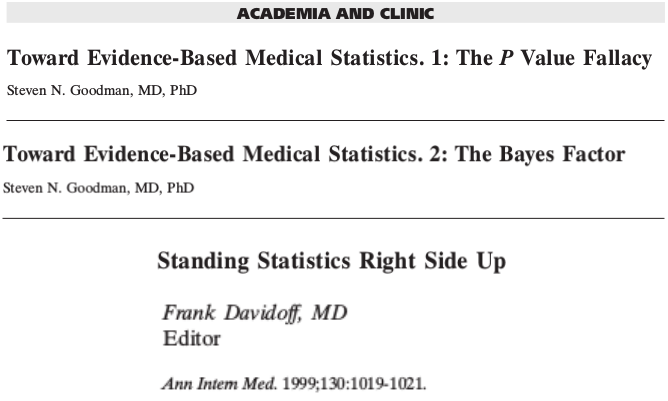
Inverse-Bayes

Under the assumption of equipoise: $\Pr(H_0) = \Pr(H_1) = 50\%$



This Argument is Not New Either

Goodman (1999, Ann Intern Med, 995-1013)



ACADEMIA AND CLINIC

Toward Evidence-Based Medical Statistics. 1: The *P* Value Fallacy

Steven N. Goodman, MD, PhD

Toward Evidence-Based Medical Statistics. 2: The Bayes Factor

Steven N. Goodman, MD, PhD

Standing Statistics Right Side Up

Frank Davidoff, MD
Editor

Ann Intern Med. 1999;130:1019-1021.

Misunderstanding of Significance Tests

Steve Goodman in *Science* (2016)

INSIGHTS | *PERSPECTIVES*

STATISTICS

Aligning statistical and scientific reasoning

Misunderstanding and misuse of statistical significance impede science

"Let us hope the next century will see as much progress in the inferential methods of science as in its substance."

Conclusion

A Bayesian perspectives on statistical significance tests offers important insights:

- A significance threshold at $\alpha = 0.05$ combined with low power produces high false positive rates.
- A substantial reduction in false positive rate can be obtained with $\alpha = 0.005$ and a sample size increase of 70%.
- *P*-values equal or slightly below 0.05 carry only weak evidence against the null hypothesis.

References

- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J. et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2:6–10.
- Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, 1: 140216.
- Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, 4:171085.
- Colquhoun, D. (2018). The false positive risk: a proposal concerning what to do about p-values. arXiv:1802.04888v4 [stat.AP].
- Held, L. & Ott, M. (2018). On *p*-values and Bayes factors. *Annual Review of Statistics and Its Application*, 5:393–419.
- Sellke, T., Bayarri, M. J. & Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71.