What's the evidence?

On *P*-values and Bayes factors for small samples

Leonhard Held

University of Zurich[UZH]

InSPiRe Conference on Methodology for Clinical Trials
in Small Populations and Rare Diseases
April 26-28, 2017
University of Warwick

Financial support by the Swiss National Science Foundation (SNF)
(project #159715) is gratefully acknowledged.

Outline

The "curse" of *P*-values

Goodman (2016), Science

STATISTICS

Aligning statistical and scientific reasoning

Misunderstanding and misuse of statistical significance impede science

By **Steven N. Goodman**

Imagine the American Physical Society convening a panel of experts to issue a missive to the scientific community on the difference between weight and mass. And imagine that the impetus for such a message was a recognition that engineers and builders had been confusing these concepts for decades, making bridges, buildings, and other components of our physical infrastructure much weaker than previously suspected.

POLICY

That, in a sense, is what happened with the recent release of a statement from the American Statistical Association (ASA), with the deceptively innocuous title, "ASA statement on statistical significance and p-values" (*1*). The scientific measure in need of clarification was the *P* value—perhaps the most ubiquitous statistical index used in scientific research to help decide what is true and what is not. The ASA saw misunderstanding and misuse of statistical significance as a factor in the rise in concern about the credibility of many scientific claims (sometimes called the "reproducibility crisis") and is hoping that its official statement on the matter will help set scientists on the right course.

Towards evidence-based medical statistics

Goodman (1999a,b), Annals of Internal Medicine

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

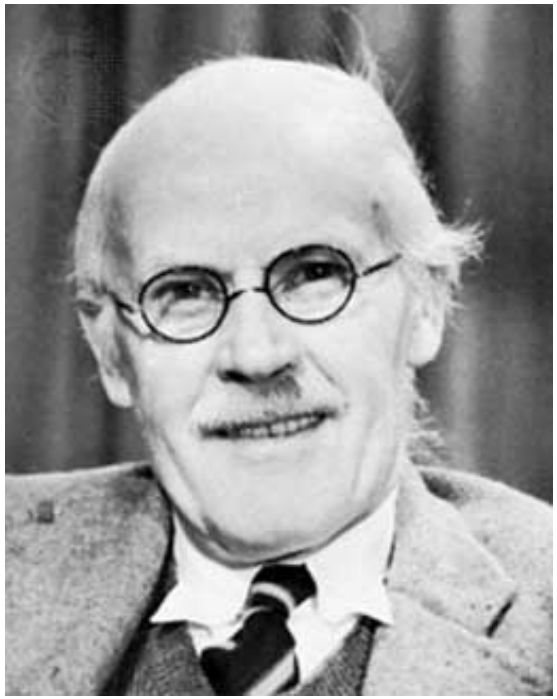
Steven N. Goodman, MD, PhD

Frequentist statistics can serve a useful purpose, but their limitations are many and serious. Some members of the biostatistical community have therefore worked long and hard to encourage the medical researchers and readers to use the Bayesian approach to statistical inference in the design and interpretation of their studies. Goodman's article is an elegant reflection of those efforts, providing both an explication of underlying theory and solid suggestions for practice. In our view, this article will contribute importantly to the task of standing statistical inference right side up. We recommend it to our readers' most serious attention.

Frank Davidoff, MD
Editor

Bayes factors

Harold Jeffreys
(1891-1989)



Consider two hypotheses H_0 and H_1 and some data y.

Bayes's theorem gives

$$\underbrace{\frac{\Pr(H_0 \mid y)}{\Pr(H_1 \mid y)}}_{\text{Posterior odds}} = \underbrace{\frac{f(y \mid H_0)}{f(y \mid H_1)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\text{Prior odds}}$$

Interpretation of Bayes factors

$$\underbrace{\frac{\Pr(H_0 \mid y)}{\Pr(H_1 \mid y)}}_{\text{Posterior odds}} = \underbrace{\frac{f(y \mid H_0)}{f(y \mid H_1)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{\Pr(H_0)}{\Pr(H_1)}}_{\text{Prior odds}}$$

- The Bayes factor $\text{BF}(y)$ provides a direct quantitative measure how the data y have increased or decreased the odds of H_0.
- The Bayes factor (or its logarithm) is therefore often referred to as the "strength of evidence" or "weight of evidence".
- We focus on the evidence against H_0 provided by small Bayes factors $\text{BF}(y) < 1$.

Standards of evidence in scientific investigation

Bayes factor BF	Strength of evidence against H_0
1 to 1/3	weak
1/3 to 1/10	moderate
1/10 to 1/30	substantial
1/30 to 1/100	strong
1/100 to 1/300	very strong
$< 1/300$	decisive

The evidence from a clinical trial

Example taken from Held and Ott (2017)

- Consider RCT to detect a clinically relevant difference with 80% power ($\beta = 0.2$) at the usual two-sided $\alpha = 0.05$ significance level.
- Two-sided P-value is $p = 0.01$.
- The PI asks the trial statistician to compute the corresponding Bayes factor.
- The statistician knows the distribution of the underlying test statistic $t = t(p) = \Phi^{-1}(1 - p/2) \sim \text{FN}(\mu, 1)$ where

$$\mu = \begin{cases} 0 & \text{under } H_0, \\ \Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta) = 2.8 & \text{under } H_1, \end{cases}$$

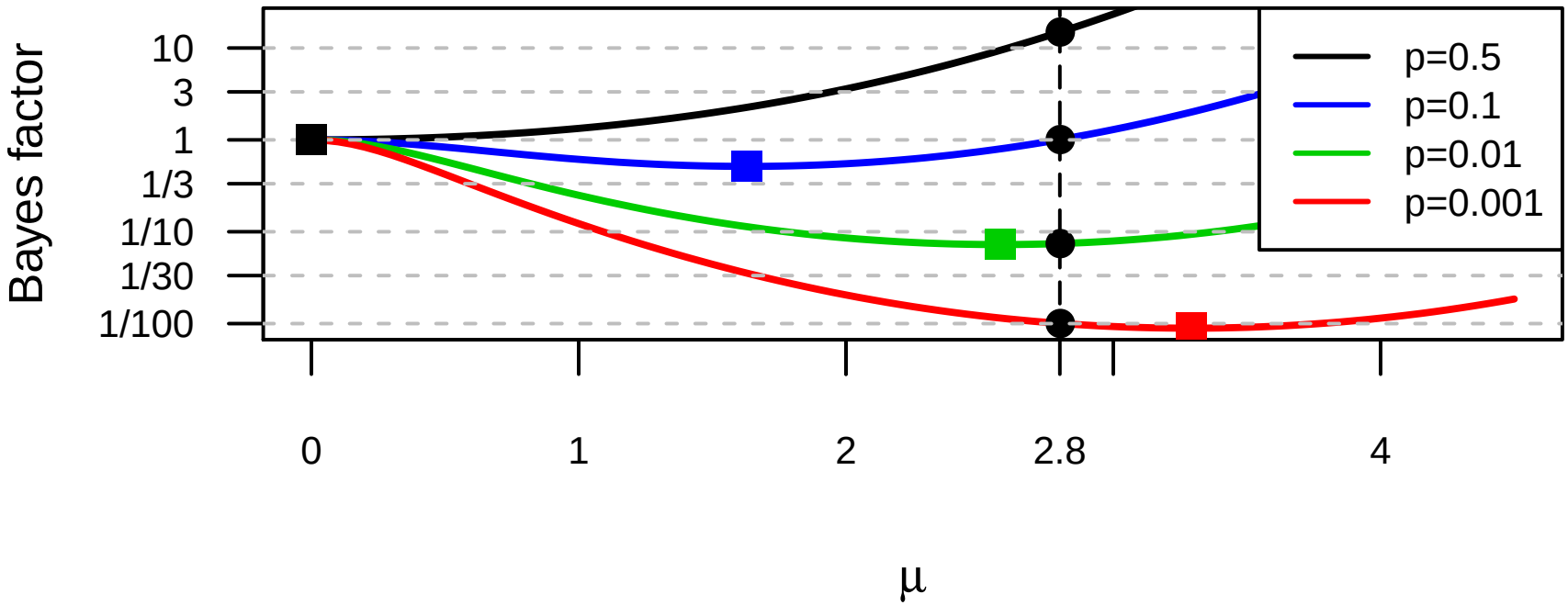
and obtains

$$\text{the Bayes factor } \text{BF}(p = 0.01) = \frac{f(t \mid H_0)}{f(t \mid H_1)} = 1/13.$$

The maximal evidence from a clinical trial

- The statistician is well aware that the power of the study may have been different from the assumed $1 - \beta$ value, if the true treatment effect is different from the pre-specified effect.
- She therefore maximizes $f(t \mid H_1)$ with respect to μ and obtains

  the minimum Bayes factor $\text{minBF}(p = 0.01) = 1/14$.



Calibration of P values

- Goal: Compute P-based Bayes factor

$$\text{BF}(p) = \frac{f(p \mid H_0)}{f(p \mid H_1)}$$

based on (two-sided) P value

$$p = \Pr(\text{data or more extreme data} \mid H_0).$$

for H_0: $\theta = \theta_0$.

- Advantage: Distribution of P-value under the null is known:

$$p \mid H_0 \sim \text{U}(0, 1)$$

Test-based Bayes factors

Johnson (2005, 2008)

- Problem: Distribution of P-value under the alternative is usually unknown.
- Idea: Replace P-based Bayes factor with test-based Bayes factors

$$\text{BF}(t) = \frac{f(t \mid H_0)}{f(t \mid H_1)}$$

where $t = t(p)$ is the underlying test statistic used to compute p.

- If $t(p)$ is one-to-one, then $\text{BF}(t) = \text{BF}(p)$.
- Maximizing

$$f(t \mid H_1) = \int f(t \mid \theta, H_1) f(\theta \mid H_1) d\theta$$

within a certain class of prior distributions $f(\theta \mid H_1)$ gives the minimum Bayes factor $\text{minBF}(t)$.

Minimum Bayes factor for simple alternatives

Berger and Sellke (1987)

- For a two-sided P-value p with underlying test statistic $t = \Phi^{-1}(1 - p/2)$, the Bayes factor is bounded by the minimum Bayes factor

$$\text{minBF}_{\text{simple}} \approx 2 \exp(-t^2/2)$$

within the class of all prior distributions $f(\theta \mid H_1)$.

- The minimum is attained if the prior density under the alternative hypothesis is concentrated at the place most favored by the data, *i. e.* for a simple alternative.
- Note: If the direction of the effect is also known, then the minimum Bayes factor is half as large (Goodman, 1999b).

Minimum Bayes factor for simple alternatives

Berger and Sellke (1987)

```
p <- c(0.05, 0.01, 0.001)
t <- qnorm(1-p/2)
minBF <- 2*exp(-t^2/2)
```

	P value p		
	0.05	0.01	0.001
$\text{minBF}_{\text{simple}}$	1/3.4	1/14	1/112
Maximal evidence against H_0	moderate	substantial	very strong

"A p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis."

Wasserstein and Lazar (2016)

"Even the utmost generosity to the alternative hypothesis cannot make the evidence in favor of it as strong as classical significance levels might suggest."

Edwards, Lindman, and Savage (1963)

Minimum Bayes factor for local alternatives

Edwards et al. (1963)

- Assume a local $\mathrm{N}(\theta_0, \tau^2)$ prior for $\theta \mid H_1$.
- Calculus shows that a lower limit on BF (for all τ^2) is

$$\text{minBF}_{\text{local}} = \begin{cases} t \exp(-t^2/2)\sqrt{e} & \text{for } t > 1 \\ 1 & \text{otherwise.} \end{cases}$$

```
minBF <- t*exp(-t^2/2)*exp(1/2)
```

	P value p		
	0.05	0.01	0.001
$\text{minBF}_{\text{local}}$	1/2.1	1/6.5	1/41
Maximal evidence against H_0	weak	moderate	strong

The "-e p log p" calibration

Sellke, Bayarri, and Berger (2001)

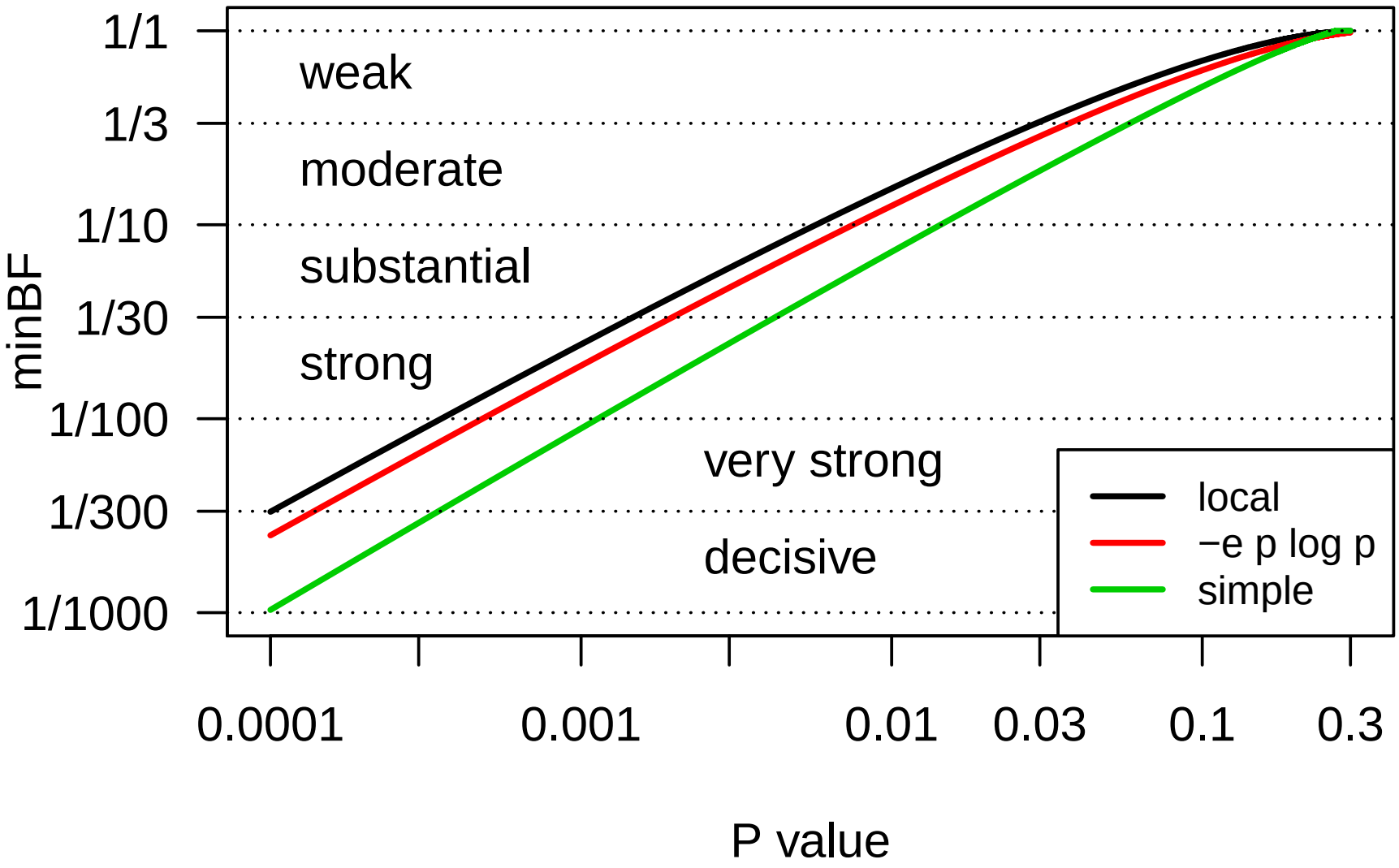
- Based on $p \,|\, H_0 \sim \mathrm{U}(0,1)$ vs. $p \,|\, H_1 \sim \mathrm{Be}(\xi \leq 1, 1)$.
- Minimization with respect to ξ gives

$$\mathrm{minBF}_{\mathrm{SBB}} = \begin{cases} -e\, p\, \log(p) & \text{for } p < e^{-1} \approx 0.37 \\ 1 & \text{else} \end{cases}$$

```
minBF <- -exp(1)*p*log(p)
```

	P value		
	0.05	0.01	0.001
$\mathrm{minBF}_{\mathrm{SBB}}$	1/2.5	1/8	1/53
Maximal evidence against H_0	weak	moderate	strong

Comparison of minimum Bayes factors



Test-based Bayes factors

Johnson (2008); Held et al. (2015)

- Consider the deviance statistic z in a GLM with associated P-value

$$p = \Pr(\chi^2(d) \geq z) \text{ at } d \text{ degrees of freedom.}$$

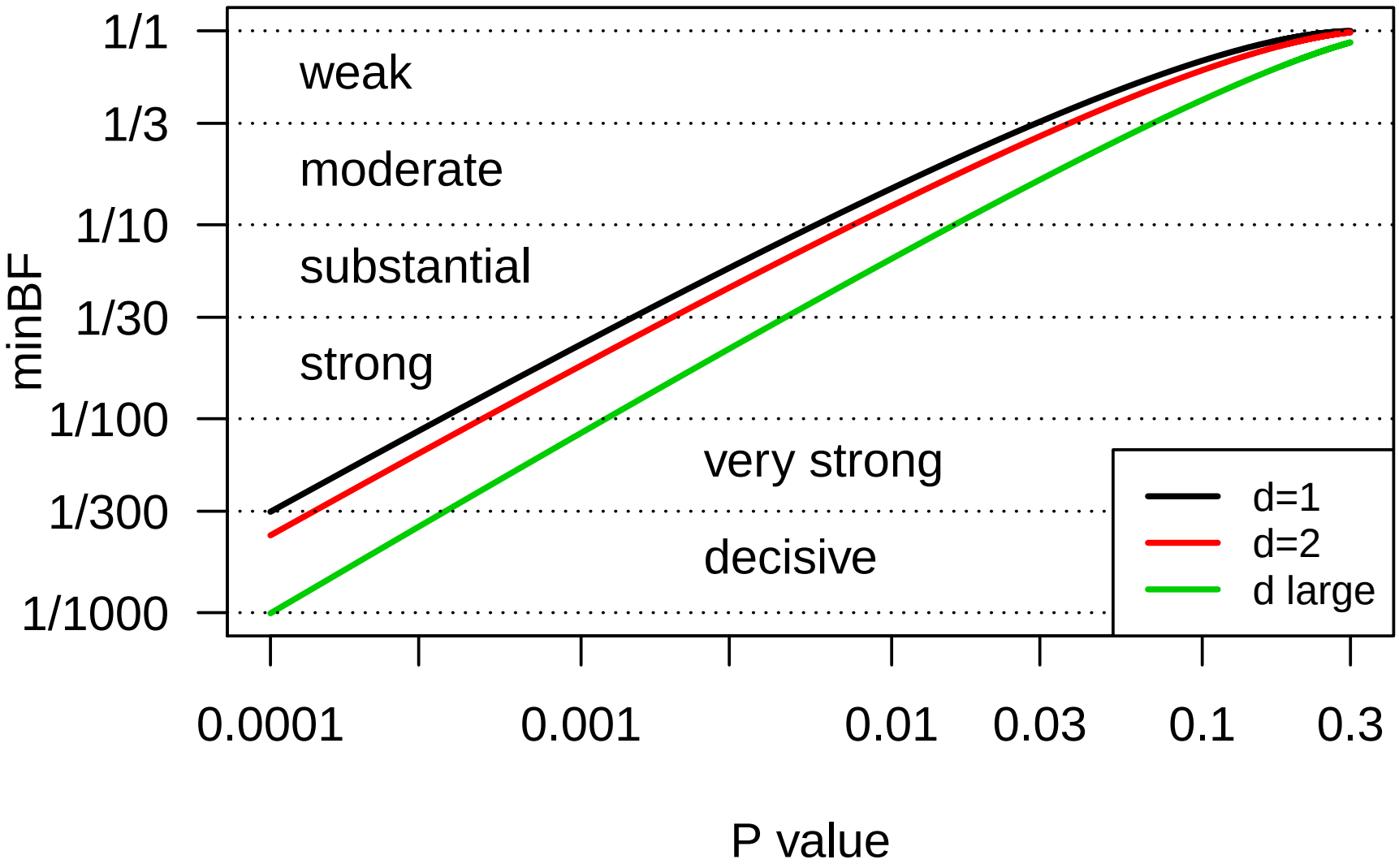
- Under local g-priors, the minimum Bayes factor based on z is

$$\text{minTBF}(z) = \begin{cases} \left(\frac{z}{d}\right)^{d/2} \exp\left(-\frac{z-d}{2}\right) & \text{for } z > d \\ 1 & \text{otherwise.} \end{cases}$$

- Special cases:

  $d = 1$: $\text{minTBF}(z) \equiv \text{minBF}_{\text{local}}$
  $d = 2$: $\text{minTBF}(z) \equiv \text{minBF}_{\text{SBB}}$ ("- e p log p")
  $d \to \infty$: $\text{minTBF}(z) \equiv \exp(-t^2/2)$ with one-sided $t = \Phi^{-1}(1-p)$

Comparison of test-based minimum Bayes factors



Small samples

Royall (1986); Held and Ott (2016)

The Effect of Sample Size on the Meaning of Significance Tests

RICHARD M. ROYALL*

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

Taylor & Francis
Taylor & Francis Group

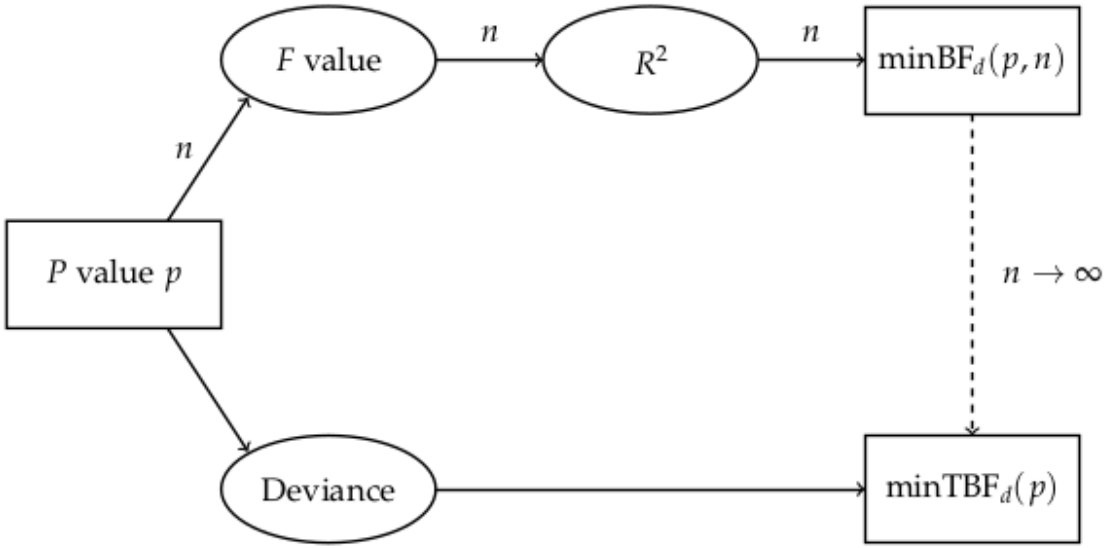
How the Maximal Evidence of *P*-Values Against Point Null Hypotheses Depends on Sample Size

Leonhard Held and Manuela Ott

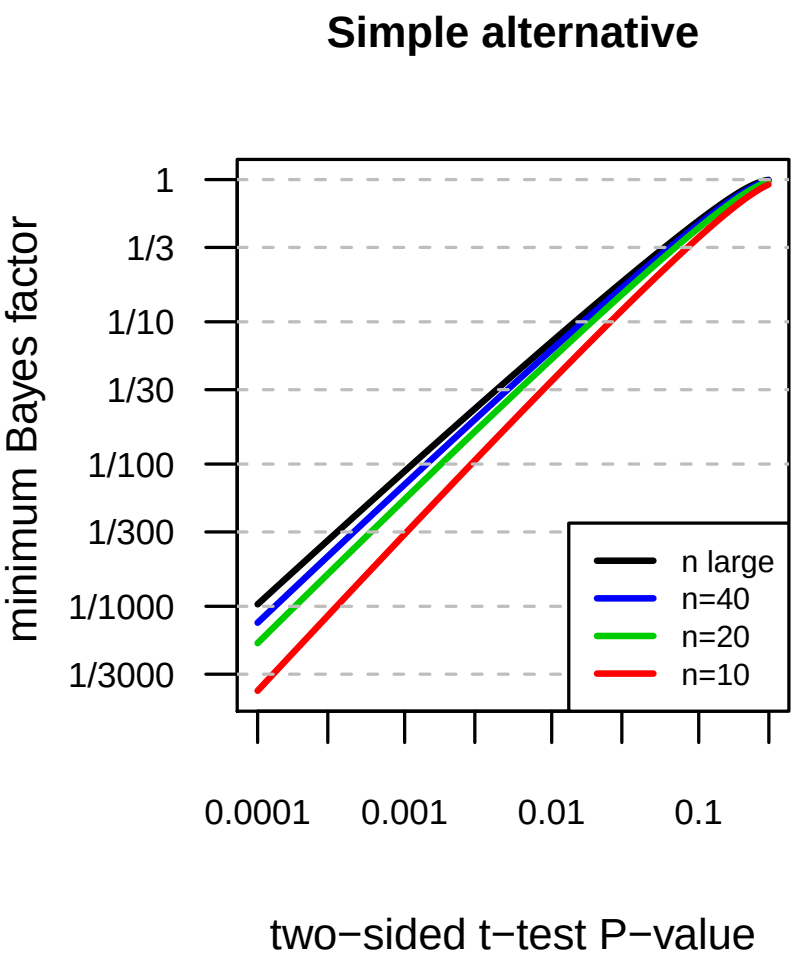
Department of Biostatistics Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

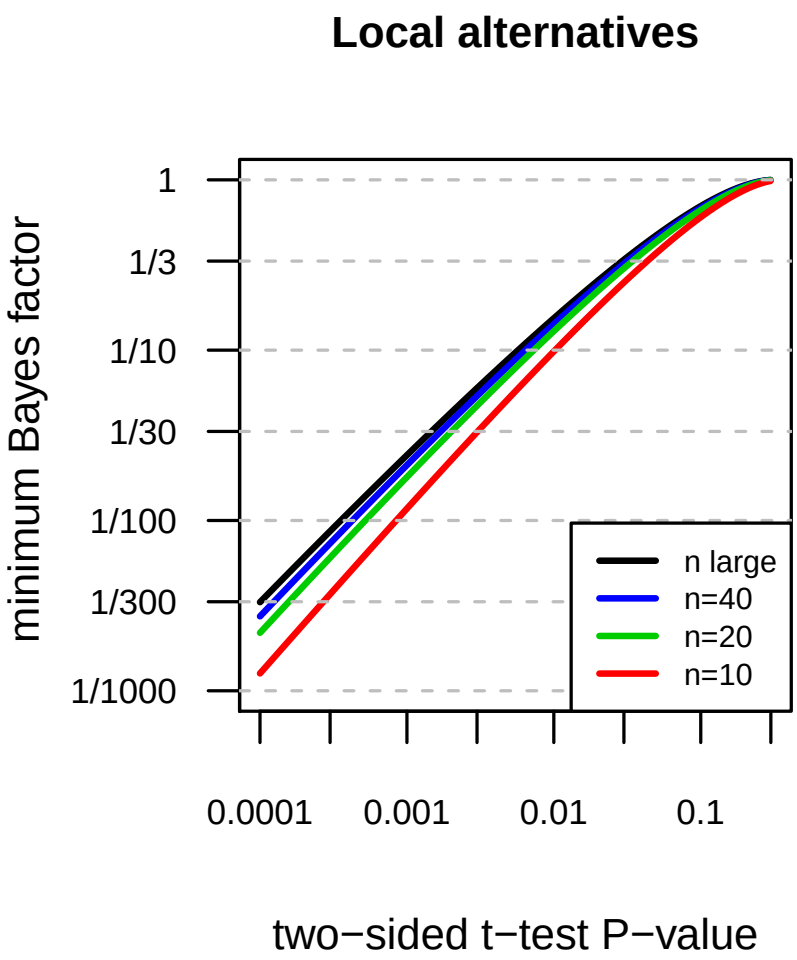
Dependence of minimum Bayes factors on sample size

- We have studied the relationship between a P-value from the t/F-test in the linear model and the minimum Bayes factor.
- Minimum Bayes factors under a simple alternative can be calculated numerically.
- Minimum Bayes factors under local g-priors are also available (Johnson, 2005; Liang et al., 2008).

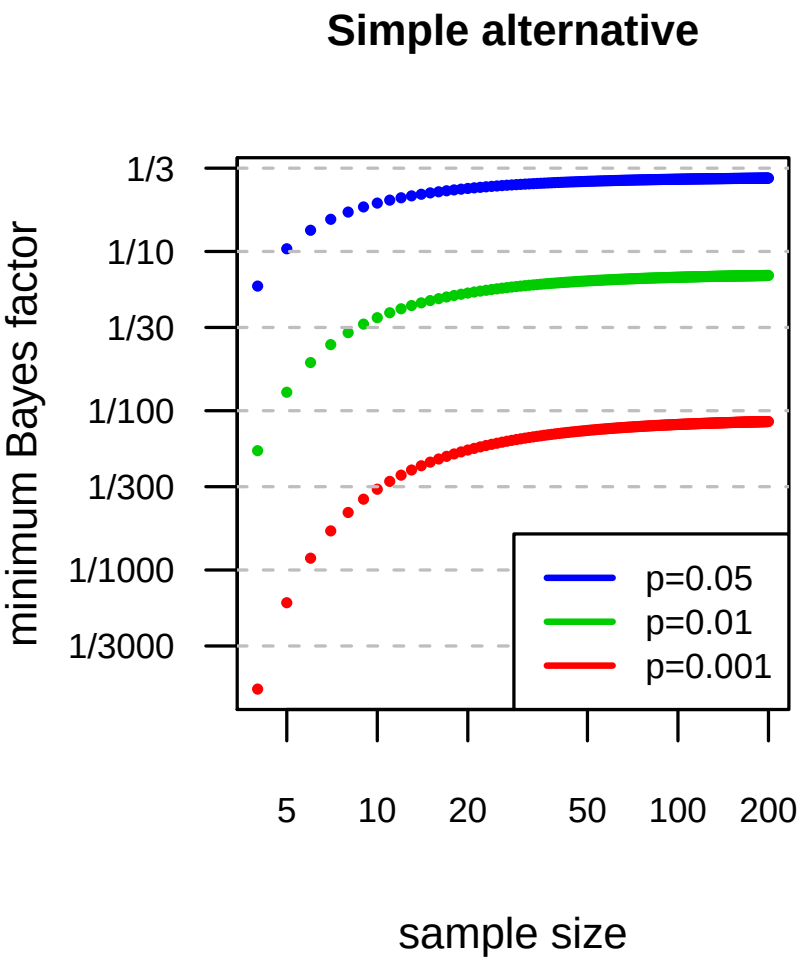


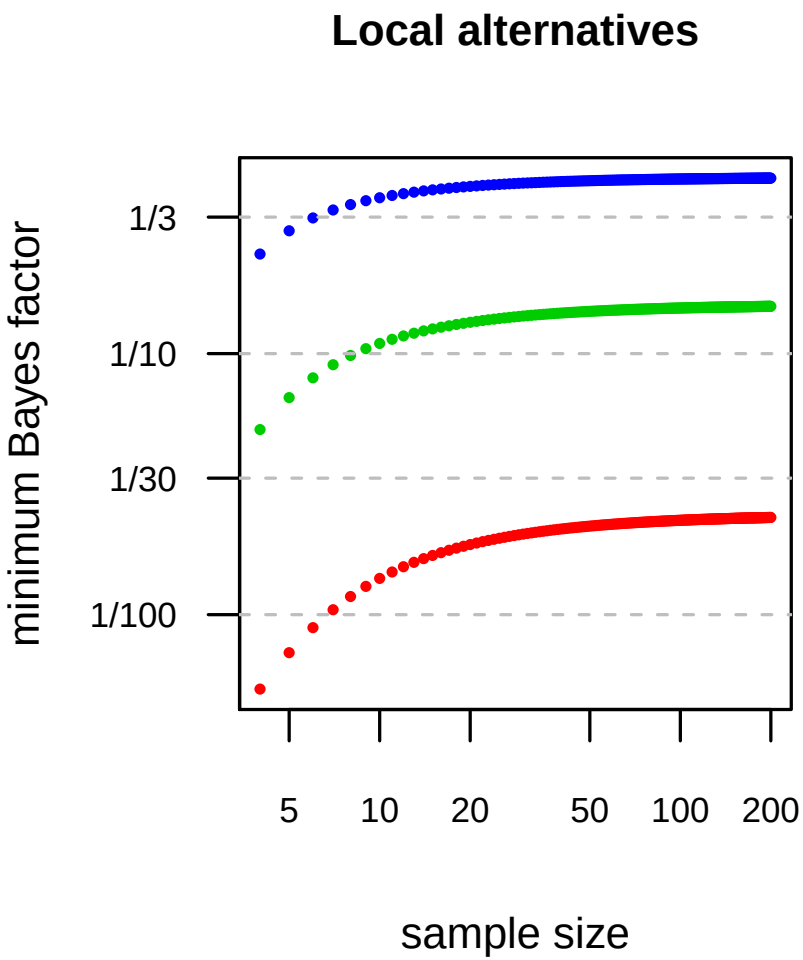
Dependence of minimum Bayes factors on sample size





Dependence of minimum Bayes factors on sample size

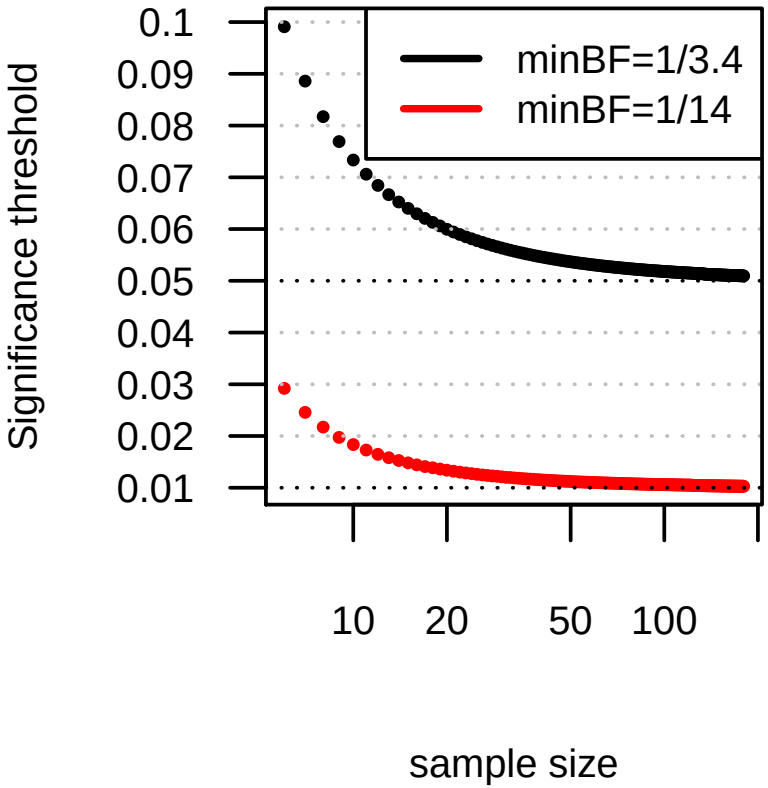




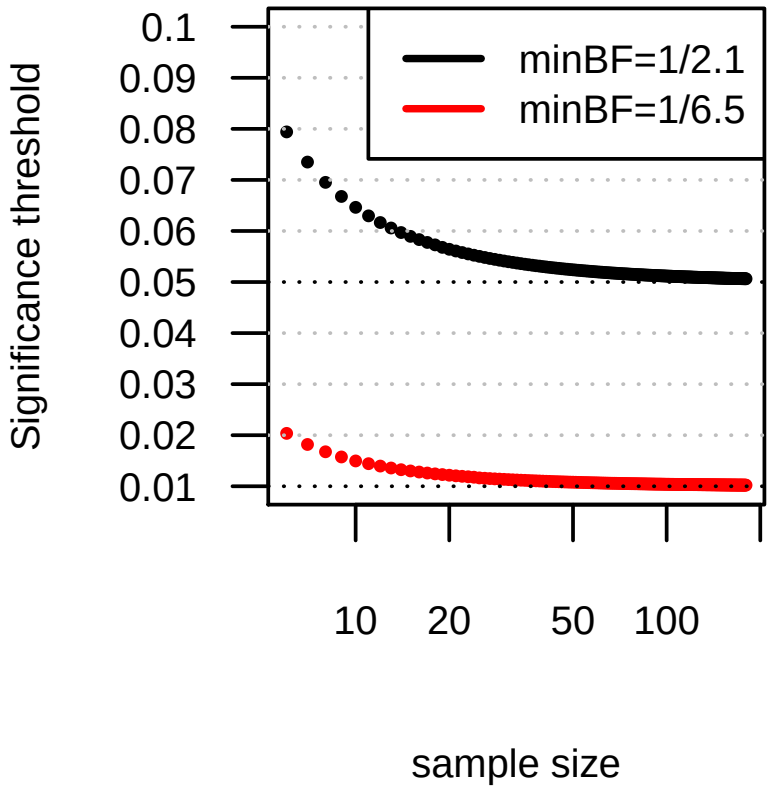
Adjusted significance thresholds

Achieving the same minimum Bayes factor

Simple alternative



Local alternatives



Sample size adjustments of "$-e\,p\,\log(p)$"

For $d = 2$ (and $n > 4$) there is an analytic formula with Stirling approximation

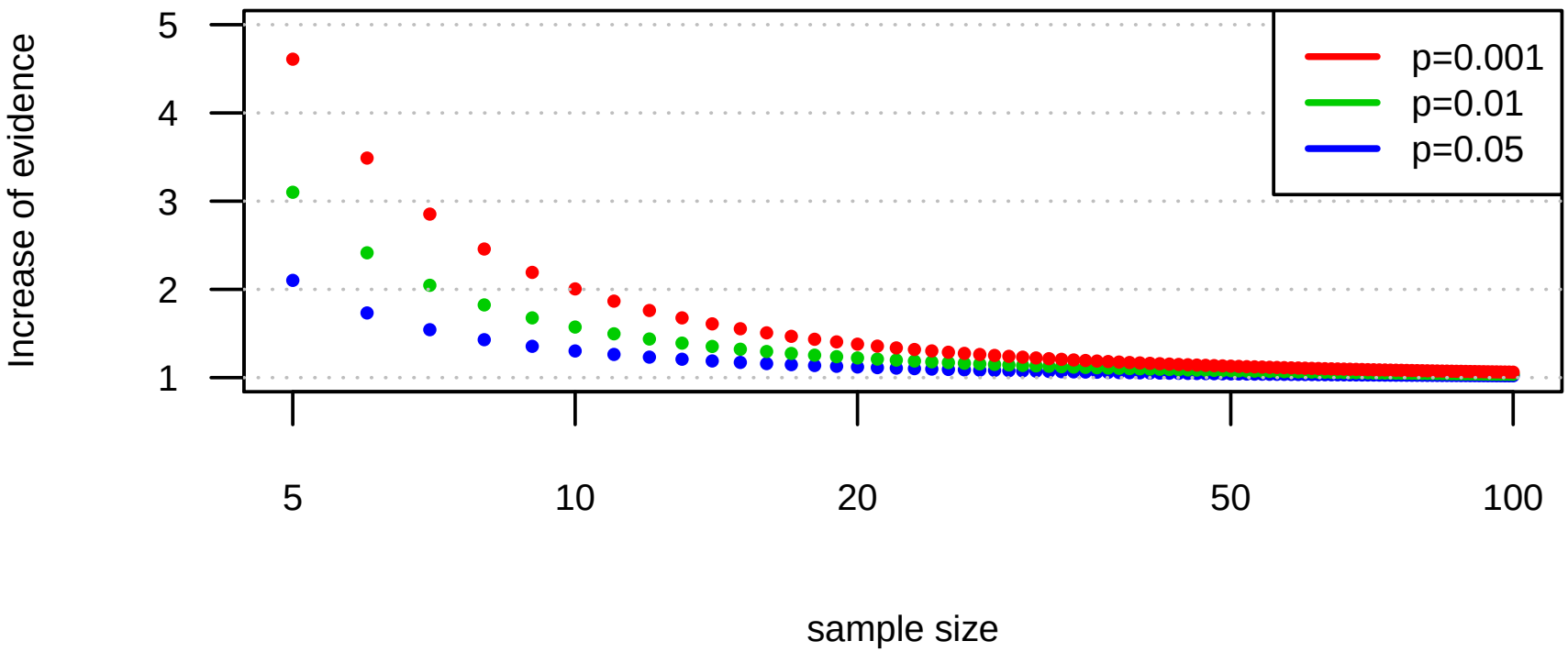
$$\text{minBF}_{\text{SBB}}(n) \approx \begin{cases} -\frac{e}{2}\, p\, \overbrace{(n-2)\left(1 - p^{2/(n-3)}\right)}^{\uparrow 2\log(p) \text{ as } n \to \infty} & \text{for } p < \left(\frac{n-1}{n-3}\right)^{-(n-3)/2} \\ 1 & \text{otherwise.} \end{cases}$$

$\rightarrow$ $\text{minBF}_{\text{SBB}}(n)$ converges monotonically from below to the asymptotic minimum Bayes factor $\text{minBF}_{\text{SBB}} = -e\,p\,\log(p)$.

Increase of evidence

- It is convenient to study the increase of evidence for sample size n for the "-e p log p" calibration:

$$\text{Increase of evidence}(n) = \frac{\text{minBF}_{\text{SBB}}}{\text{minBF}_{\text{SBB}}(n)}$$



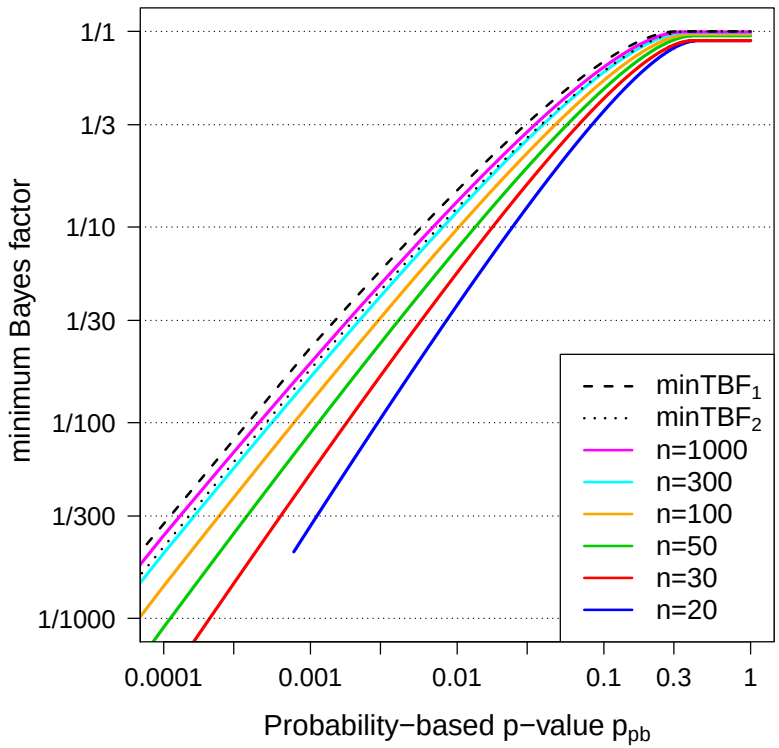
Discrete data: Fisher's exact test

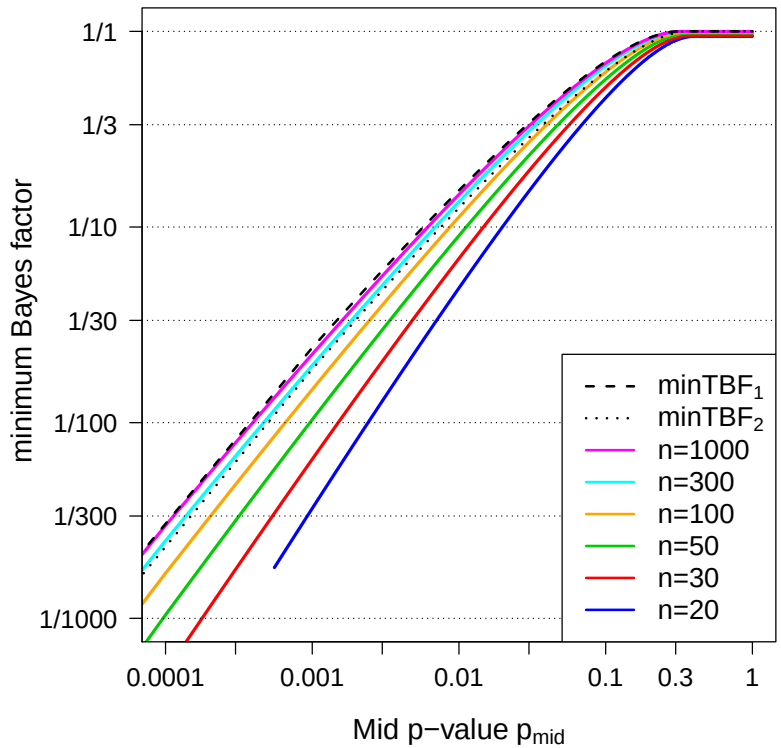
Ott and Held (2017)

- We have also studied the relationship between a P-value from Fisher's exact test and the corresponding minimum Bayes factor (Li and Clyde, 2016) for local alternatives.
- The very same relationship as for the t-test has been observed: the (mean) minimum Bayes factors decrease with decreasing sample size, but convergence to the asymptotic limit is slower.

Discrete data: 2×2 table

Fisher's exact test (probability based and mid p-value)





Liebermeister's test (for direction)

- Consider H_0: $\pi_1 \leq \pi_0$ vs. H_1: $\pi_1 > \pi_0$ for two binomial samples

$$\begin{aligned} \text{Treatment: } x_1 \,|\, \pi_1 &\sim \text{Bin}(x_1 + y_1, \pi_1) \\ \text{Placebo: } x_0 \,|\, \pi_0 &\sim \text{Bin}(x_0 + y_0, \pi_0) \end{aligned}$$

- Carl von Liebermeister (1833-1901) has computed in 1877:

$$\Pr(H_0 \,|\, \text{data}) = \Pr(\pi_1 \leq \pi_0 \,|\, \text{data})$$

for uniform priors $\pi_1, \pi_0 \overset{ind}{\sim} \text{U}(0, 1)$

- $\Pr(H_0 \,|\, \text{data})$ equals the one-sided P-value p^- from Fisher's exact test for

	Survived	Died
Treatment	x_1+1	y_1
Placebo	x_0	y_0+1

Liebermeister's test (for existence)

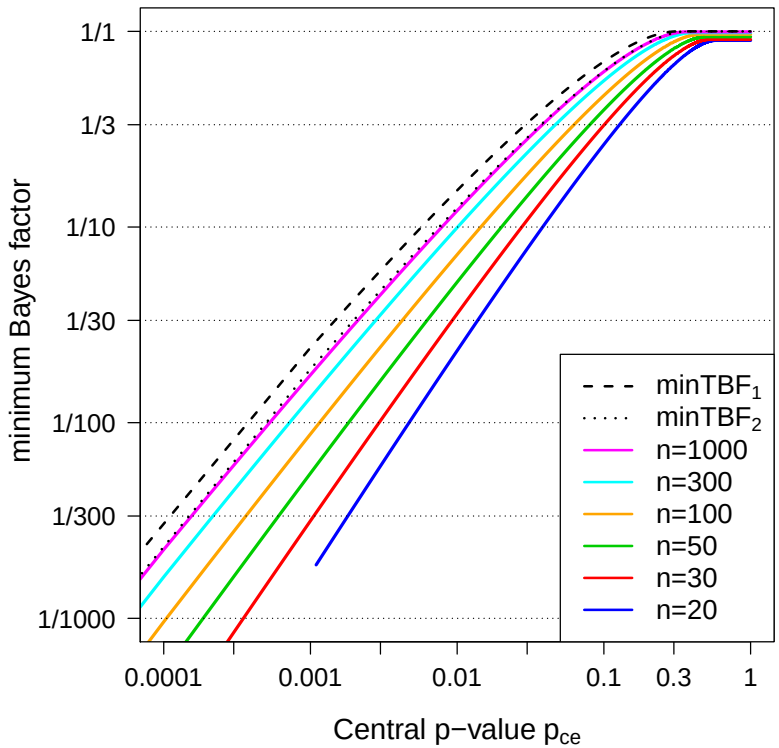
- Point null hypothesis H_0: $\pi_1 = \pi_0$ vs. H_1: $\pi_1 \neq \pi_0$
- Two-sided P-value can be defined as

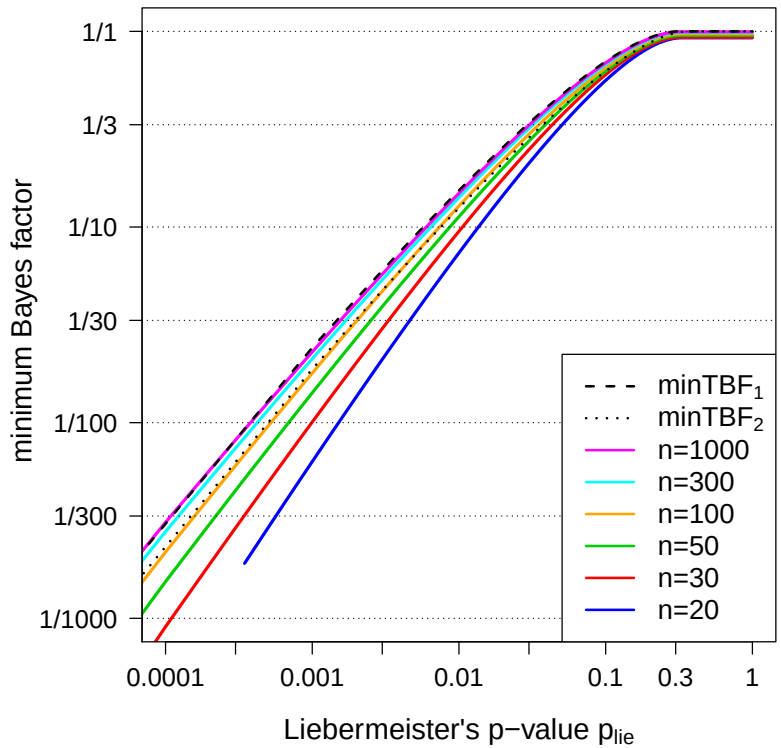
$$p = \min\left\{2\min\left\{p^-, p^+\right\}, 1\right\},$$

as for the central P-value in Fisher's exact test.

Discrete data: 2×2 table

Fisher's exact test (central) and Liebermeister's test





Summary points

1. Minimum Bayes factors are useful to quantify the maximal evidence of a two-sided P-value against a point null hypothesis.
2. The maximal evidence of a two-sided P-value depends on how the P-value has been calculated. It generally increases with decreasing sample size.
3. The maximal evidence of a two-sided P-value also depends on the underlying study design: It matters whether the P-value comes from
   - a confirmatory study with a well-defined simple alternative,
   - or from an exploratory analysis used to generate hypotheses $\rightarrow$ local alternatives

More details in Held and Ott (2017) review paper

Literature I

J. O. Berger and T. Sellke. Testing a point null hypothesis: The irreconcilability of P values and evidence (with discussion). *J. Am. Stat. Assoc.*, 82:112–139, 1987.

W. Edwards, H. Lindman, and L. J. Savage. Bayesian statistical inference in psychological research. *Psychol. Rev.*, 70:193–242, 1963.

S. N. Goodman. Towards evidence-based medical statistics. 1.: The P value fallacy. *Ann. Intern. Med.*, 130:995–1004, 1999a.

S. N. Goodman. Towards evidence-based medical statistics. 2.: The Bayes factor. *Ann. Intern. Med.*, 130(12):1005–1013, 1999b.

S. N. Goodman. Aligning statistical and scientific reasoning. *Science*, 352:1180–1181, 2016.

L. Held and M. Ott. How the maximal evidence of P values against point null hypotheses depends on sample size. *Am. Stat.*, 70(4):335–341, 2016.

L. Held and M. Ott. On P values and Bayes factors. Technical report, University of Zurich, 2017. invited Article for the Annual Review of Statistics and its Applications.

L. Held, D. Sabanés Bové, and I. Gravestock. Approximate Bayesian model selection with the deviance statistic. *Stat. Sci.*, 30(2):242–257, 2015.

V. E. Johnson. Bayes factors based on test statistics. *J. R. Stat. Soc. Ser. B*, 67(5):689–701, 2005.

Literature II

V. E. Johnson. Properties of Bayes factors based on test statistics. *Scand. J. Stat.*, 35(2): 354–368, 2008.

Y. Li and M. A. Clyde. Mixtures of g-priors in generalized linear models. Technical report, Clemson/Duke University, October 2016.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for Bayesian variable selection. *J. Am. Stat. Assoc.*, 103(481):410–423, 2008.

C. Liebermeister. Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. *Sammlung klinischer Vorträge* (Innere Medicin No. 31-64), 110:935–962, 1877.

M. Ott and L. Held. Sample size adjusted minimum Bayes factors for 2×2 contingency tables. Technical report, University of Zurich, 2017. submitted.

R. M. Royall. The effect of sample size on the meaning of significance tests. *Am. Stat.*, 40(4): 313–315, Nov. 1986.

T. Sellke, M. J. Bayarri, and J. O. Berger. Calibration of p values for testing precise null hypotheses. *Am. Stat*, 55:62–71, 2001.

R. L. Wasserstein and N. A. Lazar. The ASA's statement on p-values: context, process, and purpose. *Am. Stat.*, 70(2):129–133, 2016.