Open access • Journal Article • DOI:10.1080/00031305.2016.1209128

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

Leonhard Held, Manuela Ott

Institutions: University of Zurich

Published on: 21 Nov 2016 - The American Statistician (Taylor & Francis)

Topics: Sample size determination, Minimum chi-square estimation, Size, Insensitivity to sample size and Z-test

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

Held, Leonhard ; Ott, Manuela

DOI: https://doi.org/10.1080/00031305.2016.1209128

How the maximal evidence of P values against point null hypotheses depends on sample size

Leonhard Held & Manuela Ott

Department of Biostatistics

Epidemiology, Biostatistics and Prevention Institute

University of Zurich

Hirschengraben 84, 8001 Zurich, Switzerland

Email: {leonhard.held, manuela.ott}@uzh.ch

7th June 2016

Minimum Bayes factors are commonly used to transform two-sided P values to lower bounds on the posterior probability of the null hypothesis. Several proposals exist in the literature, but none of them depends on the sample size. However, the evidence of a P value against a point null hypothesis is known to depend on the sample size. In this paper we consider P values in the linear model and propose new minimum Bayes factors that depend on sample size and converge to existing bounds as the sample size goes to infinity. It turns out that the maximal evidence of an exact two-

sided P value increases with decreasing sample size. The effect of adjusting minimum Bayes factors for sample size is shown in two applications.

Key Words: Evidence, F-Test, Minimum Bayes factor, P value, t-Test, Sample size

1. Introduction

A common misconception of applied researchers is the widespread belief that the P value is the (posterior) probability of the null hypothesis, for some discussion see e.g. Freeman (1993); Goodman (1999a); Held (2013); Greenland and Poole (2013). Of course, this is not true and it is now well discussed in the literature that two-sided P values overstate the evidence against the null hypothesis. This can be shown using replication probabilities (Goodman, 1992) or using the concept of Bayes factors (Berger and Sellke, 1987; Sellke, Bayarri, and Berger, 2001). For some recent discussion on the use and misuse of P-values see the supplementary material of Wasserstein and Lazar (2016).

In this article we focus on Bayes factors and consider a point null hypothesis H_0: $\theta = \theta_0$ with prior probability $\pi = \mathsf{Pr}(H_0)$, so $\mathsf{Pr}(H_1) = 1 - \pi$ is the prior probability of the alternative hypothesis H_1. The alternative hypothesis may be simple, *i. e.* $H_1 : \theta = \theta_1 \neq \theta_0$ or composite, usually $H_1 : \theta \neq \theta_0$. The Bayes factor BF transforms the prior odds $\mathsf{Pr}(H_0)/\mathsf{Pr}(H_1)$ to the corresponding posterior odds $\mathsf{Pr}(H_0 \,|\, \text{data})/\mathsf{Pr}(H_1 \,|\, \text{data})$ in the light of the data:

$$\frac{\mathsf{Pr}(H_0 \,|\, \text{data})}{\mathsf{Pr}(H_1 \,|\, \text{data})} = \text{BF} \times \frac{\mathsf{Pr}(H_0)}{\mathsf{Pr}(H_1)}. \tag{1}$$

In (1), the Bayes factor

$$\text{BF} = \frac{f(\text{data} \,|\, H_0)}{f(\text{data} \,|\, H_1)} \tag{2}$$

is the ratio of the likelihood $f(\text{data} \,|\, H_0)$ under the null hypothesis H_0 and the likeli-

hood (or marginal likelihood for composite alternatives) $f(\text{data} \mid H_1)$ under the alternative hypothesis H_1 (Kass and Raftery, 1995). Thus, the Bayes factor provides a direct quantitative measure of whether the data 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" (Bernardo and Smith, 2000). In this paper we focus on the evidence *against* a point null hypothesis provided by small Bayes factors BF < 1. To categorize such Bayes factors, we use the (admittedly somewhat arbitrary) scale provided in Table 1, adapted from Kass and Raftery (1995) and Goodman (1999b).

[Table 1 about here.]

A P value is a quantitative measure of the degree of conflict of the data with the null hypothesis (Goodman, 1992). A transformation of a P value to a Bayes factor is possible, but reflects a fundamental change in interpretation from aleatory to epistemic uncertainty. Indeed, the Bayes factor shows how the probability of the null hypothesis *changes* after the data (with associated P value) have been observed.

It has long been recognized that for a given P value, Bayes factors also depend on the sample size (Spiegelhalter, Abrams, and Myles, 2004; Wagenmakers, 2007). For example, if the alternative hypothesis is simple, the evidence of a given P value against the null hypothesis has been shown to increase with decreasing sample size (Royall, 1986). For a composite alternative hypothesis the Bayes factor will depend on the prior on θ and there is in general no monotonic relationship between the Bayes factor and sample size for a given P value (Spiegelhalter et al., 2004, Section 4.4.3). However, it is possible to compute lower bounds on the Bayes factor, so-called *minimum Bayes factors* (Edwards, Lindman, and Savage, 1963; Berger and Sellke, 1987; Sellke et al., 2001). They quantify the *maximal evidence* of a P value against the null hypothesis within a certain class of prior distributions for θ, but all the minimum Bayes factors suggested in the literature (see the above references for details) do not depend on the sample size - in the sense that a fixed P value is transformed to the same minimum Bayes factor

no matter what the underlying sample size is.

In this paper we propose adjusted minimum Bayes factors that depend on sample size. In Section 2 we first show that two commonly used minimum Bayes factors are special cases of so-called test-based Bayes factors (TBFs). Using results on minimum Bayes factors in the linear model, we derive in Section 3 adjusted minimum Bayes factors that do incorporate the sample size n. It turns out that for exact (two-sided) P values, these sample size adjusted minimum Bayes factors are monotonic functions of the sample size that converge to the traditional minimum Bayes factors for large sample size. Interestingly, the sample size adjusted minimum Bayes factors decrease with decreasing sample size, which means that the maximal evidence of an exact P value increases with decreasing sample size. The effect of sample size adjusted minimum Bayes factors is shown in Section 4 in two applications. We close with some discussion in Section 5.

2. Methodology

2.1. Minimum Bayes factors

Consider a point null hypothesis H_0: $\theta = \theta_0$ and the composite alternative hypothesis H_1: $\theta \neq \theta_0$, where a prior distribution $f(\theta \mid H_1)$ has to be specified to compute the marginal likelihood

$$f(\text{data} \mid H_1) = \int f(\text{data} \mid \theta, H_1)\, f(\theta \mid H_1)\, d\theta.$$

The Bayes factor (2) will thus depend on the prior $f(\theta \mid H_1)$. To eliminate this dependency on the prior, an upper bound on $f(\text{data} \mid H_1)$ is often derived within a certain class of priors, which can be transformed to a lower bound on the Bayes factor.

For example, suppose a two-sided P value p has been computed based on a normally

distributed observation x with known mean $\theta = \theta_0$ (the null hypothesis H_0) and known variance σ^2. For the class of normal priors on θ under H_1, centered around θ_0, the corresponding minimum Bayes factor is

$$\text{minBF}(p) = \begin{cases} \sqrt{z}\,\exp(-z/2)\sqrt{e} & \text{for } z = z(p) > 1 \\ 1 & \text{otherwise,} \end{cases} \tag{3}$$

see Edwards et al. (1963); Berger and Sellke (1987). Here $z = z(p) = t^2 = Q_{\chi^2(1)}(1 - p)$ is the squared normal test statistic $t = (x - \theta_0)/\sigma$, $Q_{\chi^2(d)}(.)$ denotes the quantile function of the χ^2-distribution with d degrees of freedom and e is Euler's number. The original derivation of (3) is described in Appendix A. Note that the setting considered here is not as restrictive as it seems, since many statistical test procedures are often based on Gaussian approximations (Goodman, 1999b). For example, x could represent a mean outcome, a difference between means, or a proportion, say. Of course, the class of all normal prior distributions with mean θ_0 is perhaps too restrictive, so Berger and Sellke (1987) have also derived minimum Bayes factors in more general classes of prior distributions. The most general case is the class of all possible prior distributions (Edwards et al., 1963), on which we comment in Section 5.

Another popular calibration, see for example Bayarri, Benjamin, Berger, and Sellke (2016), directly links a two-sided P value p to a lower bound on the Bayes factor, as first proposed in Vovk (1993, Section 9):

$$\text{minBF}(p) = \begin{cases} -e\,p\log p & \text{for } p < 1/e \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

Sellke et al. (2001) describe a simple derivation of (4). Briefly, under the null hypothesis H_0, p is known to be uniformly distributed. Under the alternative hypothesis small P values are expected, so a class of beta prior distributions with monotonically

decreasing density functions is considered for p. The minimum Bayes factor (4) can then be derived as outlined in Appendix B.

2.2. Test-based Bayes factors

The above minimum Bayes factors (3) and (4) are based on a test statistic (z) (or a P value p) and they are special cases of so-called (minimum) test-based Bayes factors (TBFs) (Johnson, 2005, 2008). Suppose a P value p has been obtained from a likelihood ratio test statistic (or deviance) $z = z_d(p) = Q_{\chi^2(d)}(1-p)$ with d degrees of freedom. Under certain assumptions, Johnson (2008) has shown that the minimum test-based Bayes factor (minTBF) for such likelihood ratio test statistics is

$$\text{minTBF}_d(p) = \min\left\{\left(\frac{z}{d}\right)^{d/2} \exp\left(-\frac{z-d}{2}\right), \quad 1\right\}. \tag{5}$$

It is easy to see that for $d = 1$, (5) reduces to (3), *i.e.* $\text{minTBF}_1(p) = \sqrt{z}\,\exp(-z/2)\sqrt{e}$ for $z = z_1(p) > 1$. Moreover, Held, Sabanés Bové, and Gravestock (2015, Appendix B) show that for $d = 2$, (5) is equivalent to (4), so $\text{minTBF}_2(p) = -e\,p\log p$ for $p < 1/e$.

However, the minimum test-based Bayes factor (5) does not depend on the sample size n. Indeed, derivation of (5) is based on the asymptotic distribution of the likelihood ratio test statistic, so n is assumed to be large. In Section 3 we propose to replace (5) with the corresponding minimum Bayes factors in the linear model to obtain minimum Bayes factors adjusted for sample size. This step is based on the fact that the assumptions in Johnson (2008) are equivalent to the generalized g-prior (Sabanés Bové and Held, 2011) in generalized linear models (Held et al., 2015). The generalized g-prior reduces to the ordinary g-prior in the linear model (Copas, 1983; Zellner, 1986), where a minimum Bayes factor is analytically available and depends on the sample size n (Liang, Paulo, Molina, Clyde, and Berger, 2008). To make a fair transformation of P values to minimum Bayes factors it is necessary to assume that the P value has

been computed with the exact global F-test, which also takes sample size into account, rather than with the approximate χ^2-distribution of the deviance, which holds only for large sample sizes. Note that the F-test is for $d = 1$ equivalent to the commonly used two-sided t-test.

3. Adjusting minimum Bayes factors for sample size

Consider the standard linear model

$$\boldsymbol{y} = \alpha\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{6}$$

where the response vector $\boldsymbol{y}$ is of length n, the regression coefficient vector $\boldsymbol{\beta}$ is of dimension d and the errors $\boldsymbol{\epsilon}$ are assumed to be independent and normally distributed with unknown residual variance σ^2. Under the g-prior for $\boldsymbol{\beta} \,|\, \sigma^2 \sim \mathrm{N}(\mathbf{0}, g\,\sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ (Zellner, 1986) combined with a reference prior $f(\alpha, \sigma^2) \propto \sigma^{-2}$ for the intercept α and the residual variance σ^2, the Bayes factor of the null model (with intercept only) against the linear model (6) has the form

$$\mathrm{BF}_d = (g+1)^{-(n-d-1)/2}\left\{1 + g(1-R^2)\right\}^{(n-1)/2}, \tag{7}$$

here R^2 is the usual coefficient of determination (Liang et al., 2008). BF_d is minimized for $\hat{g} = \max\{F - 1, 0\}$ where

$$F = \frac{R^2/d}{(1-R^2)/(n-d-1)} \tag{8}$$

is the usual F-statistic for testing H_0: $\boldsymbol{\beta} = \mathbf{0}$, see, e.g. Liang et al. (2008, equation (9)). By plugging-in $\hat{g}$ into (7) it then follows that the minimum Bayes factor (for $n \geq d + 2$)

is

$$\text{minBF}_d(n) = \begin{cases} (n-1)^{(n-1)/2} \left(\frac{R^2}{d}\right)^{d/2} \left(\frac{1-R^2}{n-d-1}\right)^{(n-d-1)/2} & \text{for } R^2 > \frac{d}{n-1} \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$

We now explain how the minimum Bayes factor (9) can be used to transform a P value p to a minimum Bayes factor that depends on sample size n. Since (9) depends on sample size n, a fair comparison requires to apply an exact frequentist procedure, taking into account sample size as well. For H_0: $\boldsymbol{\beta} = \mathbf{0}$ in the linear model (6) this is the F-test, so we transform a P value p to the corresponding F value $F = F_d(p, n) = Q_{\text{F}(d,n-d-1)}(1-p)$ of the F-test, here $Q_{\text{F}(d_1,d_2)}(.)$ denotes the quantile function of the $\text{F}(d_1, d_2)$-distribution with degrees of freedom d_1 and d_2. The F value can then be transformed to R^2 via inversion of (8),

$$R^2 = \left(1 + \frac{n-d-1}{d} F^{-1}\right)^{-1}, \tag{10}$$

and R^2 is finally inserted into (9). The procedure is illustrated in Figure 1.

[Figure 1 about here.]

Consider first a P value obtained from a standard two-sample t-test ($d = 1$). Since there is no analytic formula for the quantile function of the $\text{F}(1, n-2)$-distribution, there is no explicit formula for the minimum Bayes factor $\text{minBF}_1(p, n)$ as a function of p and n. However, we can still calculate $F = Q_{\text{F}(1,n-2)}(1-p)$ numerically and transform F to the minimum Bayes factor using (10) and (9).

For $d = 2$, an analytic formula for the quantile function of the $\text{F}(2, n-3)$-distribution is available:

$$F = Q_{\text{F}(2,n-3)}(1-p) = \frac{n-3}{2}\left(p^{-2/(n-3)} - 1\right). \tag{11}$$

A derivation of (11) is given in Appendix C. Transforming (11) to R^2 via (10) yields

$R^2 = 1 - p^{2/(n-3)}$ and inserting this into (9) gives an analytic formula for the minimum Bayes factor as a function of the P value p and the sample size n:

$$\text{minBF}_2(p, n) = \begin{cases} \frac{1}{2} \left(\frac{(n-1)^{(n-1)}}{(n-3)^{(n-3)}} \right)^{1/2} \left(1 - p^{2/(n-3)} \right) p & \text{for } p < (\frac{n-1}{n-3})^{-(n-3)/2} \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

This formula can be further simplified using the Stirling approximation $n^n \approx e^n \, n! / \sqrt{2\pi \, n}$ applied to

$$\left(\frac{(n-1)^{n-1}}{(n-3)^{n-3}} \right)^{1/2} \approx \left(e^2 \sqrt{(n-1)(n-3)} (n-2) \right)^{1/2} \approx e \, (n-2)$$

to obtain

$$\text{minBF}_2(p, n) \approx \begin{cases} \frac{e}{2}(n-2) \left(1 - p^{2/(n-3)} \right) p & \text{for } p < (\frac{n-1}{n-3})^{-(n-3)/2} \\ 1 & \text{otherwise.} \end{cases} \tag{13}$$

Note that $(n-2)(1 - p^{2/(n-3)}) \uparrow -2\log(p)$ as $n \to \infty$ so that - for fixed P value $p < 1/e$ - the approximate minimum Bayes factor (13) converges monotonically *from below* to the asymptotic minimum Bayes factor (4). It is easy to check that the exact minimum Bayes factor (12) is even slightly smaller than (13). We therefore conclude that the "$-e \, p \, \log(p)$" Bayes factor (4) is not necessarily "a best-case scenario for the strength of the evidence in favor of H_1 that can arise from a given p-value" (Bayarri et al., 2016). In fact, once the sample size n is incorporated, the adjusted Bayes factor bound (12) is *always smaller* than (4).

[Figure 2 about here.]

Figure 2 compares the sample size adjusted minimum Bayes factor $\text{minBF}_d(p, n)$ for $d = 1$ (left) and $d = 2$ (right) and $n = 5, 10, 20$ to the asymptotic bounds (3) and (4), respectively. The x-axis in Figure 2 gives a (two-sided) P value p from the t- (left) or

F-test (right) for the null hypothesis H_0: $\boldsymbol{\beta} = \mathbf{0}$. Of note, not only for $d = 2$, but also for $d = 1$ the minimum Bayes factors increase with increasing sample size. For example, for $p = 0.05$ and $d = 1$ the minimum Bayes factor (3) is 0.47, while the finite sample sizes minimum Bayes factors (9) are 0.44, 0.40 and 0.30 for $n = 20, 10, 5$. So according to Table 1, only for $n = 5$, $p = 0.05$ from a standard t-test provides moderate evidence against the null hypothesis, whereas the evidence is weak for $n = 10$ or larger. For $p = 0.05$ and $d = 2$ the minimum Bayes factor (4) is 0.41, while both the exact (equation (12)) and approximate (equation (13)) finite sample sizes minimum Bayes factors are 0.36, 0.31 and 0.19 for $n = 20, 10, 5$. Again taking the categorization from Table 1, $p = 0.05$ now provides moderate evidence against H_0 for $n = 10$ and $n = 5$, but only weak evidence for $n = 20$ and larger.

We have also investigated the dependence of $\text{minBF}_d(p, n)$ on n for larger values of d and have always observed the same pattern: the minimum Bayes factors for a given P value p converge (for $n \to \infty$) monotonically from below to the asymptotic minimum Bayes factor (5).

4. Applications

4.1. Bayesian interpretation of P values

We revisit Table 1 in Goodman (2001), who has used the "$-e\, p\, \log(p)$" calibration (4) to transform the P values $p = 0.1$, 0.05, 0.03, 0.01 and 0.001 to minimum Bayes factors and eventually to lower bounds on the posterior probability of the null hypothesis. From the results presented in Section 3 it is clear that these bounds are valid for large n, but will be too large for small sample size. To illustrate the effect of a small sample size, we therefore extend Table 1 in Goodman (2001) using the sample size adjusted Bayes factors (12) with sample size n equal to 20 or 10.

[Table 2 about here.]

Note that we have added a P value of 0.005 to the original Table in Goodman (2001). This is the significance threshold recently proposed by Johnson (2013) instead of the conventional 0.05 threshold. Applying the categories from Table 1 to Table 2, $p = 0.005$ represents substantial evidence against H_0 with minimum Bayes factor $1/17$ (0.06) for $n = 20$ and $1/25$ (0.04) for $n = 10$. This correspond to a decrease in the probability of the null hypothesis from 75% *a priori* to no less than 15% (for $n = 20$) and no less than 11% (for $n = 10$) *a posteriori*, respectively. If n is large, then the minimum Bayes factor is $1/14$ (0.07) and the decrease is from 75% to no less than 18%.

4.2. Reverse-Bayes analysis

In Bayesian inference, posterior information is obtained by combining prior information and observed data. Given a certain posterior and the data, one can back-calculate the prior that would yield this posterior. Such a reasoning is called a "reverse-Bayes analysis". Several authors have recently suggested to use a reverse-Bayes analysis to check the plausibility of scientific findings (Matthews, 2001; Greenland, 2006, 2011). Held (2013, Section 3) applies a reverse-Bayes analysis to derive an upper bound on the prior probability of the null hypothesis assuming the posterior probability of the null hypothesis equals the P value. He shows that the common misinterpretation of the P value as (posterior) probability of the null hypothesis implies strong and often unrealistic assumptions on the prior probability of H_0.

Held (2013) uses different calibration schemes for converting P values to minimum Bayes factors, including (3) and (4). Having obtained sample size adjusted versions for these two calibrations in the linear model in Section 3, we can now derive upper bounds for the prior probability $\pi = \mathsf{Pr}(H_0)$ of the null hypothesis H_0 which depend on the sample size n. To this end, note that the assumption $\mathsf{Pr}(H_0 \,|\, \text{data}) = p$ leads to the inequality

$$\frac{\pi}{1-\pi} \leq \frac{1}{\text{minBF}_d(p,n)} \frac{p}{1-p},$$

which gives the desired upper bound on the prior probability π of H_0. For $d = 2$, we obtain a simple analytic upper bound on π using the approximate formula (13) for $\text{minBF}_2(p, n)$:

$$\pi \leq 1/\left\{1 + \frac{e}{2}(n-2)\left(1 - p^{2/(n-3)}\right)(1-p)\right\}.$$

For large n this upper bound on π converges to $1/\{1 - e(1-p)\log(p)\}$, which is the asymptotic upper bound derived in Held (2013). For example, if $p = 0.01$ and $n = 10$, the prior probability π of the null hypothesis must have been no more than 11.3% to obtain a posterior probability of the null hypothesis equal to 1%, *i. e.* equal to the P value. It will depend on the scientific context if such a small prior probability can be considered as reasonable. For $n = 20$ the upper bound for π reduces to 9% and for large n it is only 7.5%.

[Figure 3 about here.]

Figure 3 shows the upper bound on the corresponding difference $\pi - p$ (in percentage points) as a function of the P value p and $n = 5, 10, 20$ for the two calibration schemes considered in this paper. Now the exact formula (12) is used for $d = 2$. The upper bound increases with decreasing sample size to values around 12 percentage points for $d = 1$ and around 20 percentage points for $d = 2$ (both for $n = 5$). This implies that the common misinterpretation of the P value as posterior probability of the null hypothesis requires less stringent assumptions on the corresponding prior probability π for smaller sample size n.

5. Discussion

In this paper we have derived minimum Bayes factors for point null and composite alternative hypotheses, that depend on the sample size n. The work thus extends methods originally proposed by Edwards et al. (1963), Berger and Sellke (1987) and

Sellke et al. (2001) to the finite sample setting. We have shown that the maximal evidence of an exact P value against a point null hypothesis increases with decreasing sample size. The results will be useful to adjust P value calibration methods for sample size, for example the P value nomogram proposed by Held (2010).

We note that the same relationship between evidence and sample size has been observed for point alternative hypotheses (Royall, 1986). However, point alternative hypothesis tests are rarely used in applications. For composite alternatives, Spiegelhalter et al. (2004, Section 4.4.3) observe a non-monotonic relationship of Bayes factors to sample size for fixed P value assuming a normal likelihood and a normal prior on the mean. Also the Bayes factor (7) in the linear model under the unit-information prior (Kass and Wasserman, 1995) with $g = n$ shows such a non-monotonic relationship. However, the corresponding *minimum* Bayes factor does have a simple monotonic relationship to sample size, as we have shown in this paper.

The proposed methodology was motivated by the correspondence of test-based Bayes factors in regression models (Johnson, 2008; Hu and Johnson, 2009) to methods for the calibration of P values. Specifically, we have considered two calibration schemes, which correspond to minTBFs with $d = 1$ and $d = 2$ degrees of freedom, respectively. Held et al. (2015) have also considered the case $d \to \infty$, which gives the bound

$$\text{minBF} = \exp(-t^2/2) \tag{14}$$

where the standard normal test statistic $t = t(p) = \Phi^{-1}(1-p)$ now corresponds to a *one-sided* test, here $\Phi(.)$ denotes the cumulative standard normal distribution function. The minimum Bayes factor (14) is always smaller than (5) and can be derived in the class of all possible prior distributions for the population mean θ of a normal observation (Edwards et al., 1963). Therefore, (14) is the natural choice for a "universal" bound on the Bayes factor. The minimum is obtained if the alternative hypothesis has all its prior density at the Maximum Likelihood estimate (MLE) of θ. Because the

MLE is always on one side of the null hypothesis, a one-sided rather than a two-sided P value is usually used (Edwards et al., 1963). However, if (14) is viewed as the limit of the minimum Bayes factor (5) for $d \to \infty$, the sample size $n \geq d + 2$ is implicitly also assumed to be infinite, so a finite sample size adjustment of this universal bound is not possible with our approach.

We close with some cautionary comments. First, the Bayes and minimum Bayes factors discussed in this paper assume - explicitly or implicitly - a normal prior on the regression coefficients $\boldsymbol{\beta}$ centered around the null value. Other priors on $\boldsymbol{\beta}$ will lead to other (minimum) Bayes factors. Secondly, we have repeatedly emphasized that it is important to make a fair comparison and to transform exact P values obtained from the classical F test to (exact) sample size adjusted minimum Bayes factors, as shown in Figure 1. We have shown that minimum Bayes factors then decrease with decreasing sample size. If instead P values are transformed to the deviance $z = Q_{\chi^2(d)}(1-p)$ and inserted into (9) using $R^2 = 1 - \exp(-z/n)$, the minimum Bayes factors *increase* with decreasing sample size. This is a consequence of equation (17) in Held et al. (2015), which - transformed to minimum (rather than maximum) Bayes factors - states that (9) is always larger than (5) if we evaluate both at the same P value. However, sample size adjustments need to be made both for the P value and the minimum Bayes factor to find the exact relationship between P values and minimum Bayes factors.

Acknowledgments

This work was supported by the Swiss National Science Foundation [project #159715]. We thank an Associate Editor and several referees for numerous comments that improved the presentation of the results in this article.

References

M. J. Bayarri, D. J. Benjamin, J. O. Berger, and T. M. Sellke. Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 2016. in press.

J. O. Berger and T. Sellke. Testing a point null hypothesis: The irreconcilability of P values and evidence (with discussion). *Journal of the American Statistical Association*, 82:112–139, 1987.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 2000.

J. B. Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):311–354, 1983.

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

P. R. Freeman. The role of p-values in analysing trial results. *Statistics in Medicine*, 12: 1443–1452, 1993.

S. N. Goodman. A comment on replication, P-values and evidence. *Statistics in Medicine*, 11(7):875–879, 1992.

S. N. Goodman. Towards evidence-based medical statistics. 1.: The P value fallacy. *Annals of Internal Medicine*, 130:995–1004, 1999a.

S. N. Goodman. Towards evidence-based medical statistics. 2.: The Bayes factor. *Annals of Internal Medicine*, 130:1005–1013, 1999b.

S. N. Goodman. Of P-values and Bayes: a modest proposal. *Epidemiology*, 12(3):295–297, 2001.

S. Greenland. Bayesian perspectives for epidemiological research: I. foundations and basic methods. *International Journal of Epidemiology*, 35:765–775, 2006.

S. Greenland. Null misinterpretation in statistical testing and its impact on health risk assessment. *Preventive Medicine*, 53:225–228, 2011.

S. Greenland and C. Poole. Living with p values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, 24:62–68, 2013.

L. Held. A nomogram for P values. *BMC Medical Research Methodology*, 10(1):21, 2010.

L. Held. Reverse-Bayes analysis of two common misinterpretations of significance tests. *Clinical Trials*, 10:236–242, 2013.

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

J. Hu and V. E. Johnson. Bayesian model selection using test statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(1):143–158, 2009.

V. E. Johnson. Bayes factors based on test statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67(5):689–701, 2005.

V. E. Johnson. Properties of Bayes factors based on test statistics. *Scandinavian Journal of Statistics*, 35(2):354–368, 2008.

V. E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313—19317, 2013.

R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (430):773–795, June 1995.

R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90 (431):928–934, 1995.

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

R. Matthews. Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, 35:1469–1478, 2001.

R. M. Royall. The Effect of Sample Size on the Meaning of Significance Tests. *The American Statistician*, 40(4):313–315, Nov. 1986.

D. Sabanés Bové and L. Held. Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410, 2011.

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

D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, New York, 2004.

V. G. Vovk. A logic of probability, with application to the foundations of statistics (with discussion and a reply by the author). *J. Roy. Statist. Soc. Ser. B*, 55(2):317–351, 1993.

E.-J. Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5):779–804, 2007.

R. L. Wasserstein and N. A. Lazar. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 2016. in press.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, volume 6 of *Studies in Bayesian*

Econometrics and Statistics, chapter 5, pages 233–243. North-Holland, Amsterdam, 1986.

Appendix

A. Derivation of (3)

Consider a normal likelihood for the observation x with mean θ and variance σ^2 and the null hypothesis $H_0 : \theta = \theta_0$. Under the alternative H_1 we choose a normal prior with mean θ_0 and variance $\tau^2 > 0$ for θ. Then, the marginal likelihood under the alternative is normal with mean θ_0 and variance $\sigma^2 + \tau^2$. The Bayes factor is therefore

$$\text{BF} = \frac{\frac{1}{\sigma}\varphi\left(\frac{x-\theta_0}{\sigma}\right)}{\frac{1}{\sqrt{\sigma^2+\tau^2}}\varphi\left(\frac{x-\theta_0}{\sqrt{\sigma^2+\tau^2}}\right)} = \frac{\varphi(t)}{\alpha\varphi(\alpha t)},$$

where φ denotes the density of the standard normal distribution, $t = (x - \theta_0)/\sigma$ is the normal test statistic and

$$\alpha = \frac{\sigma}{\sqrt{\sigma^2 + \tau^2}} \in (0, 1).$$

The explicit form of the Bayes factor as a function of α and t is thus

$$\text{BF} = \frac{1}{\alpha}\exp\left[-\frac{1}{2}(1-\alpha^2)t^2\right] \tag{15}$$

and minimizing this function with respect to α for fixed t gives

$$\alpha = \begin{cases} |t|^{-1} & \text{for } |t| > 1 \\ 1 & \text{otherwise.} \end{cases}$$

Plugging the above value of α into (15) yields the minimum Bayes factor (3) with $z = t^2$.

B. Derivation of (4)

Under the null hypothesis H_0, an exact two-sided P value p is known to be uniformly distributed on the unit interval, so the likelihood $f(p \mid H_0)$ is one. Under the alternative hypothesis, small P values are expected, so the density of p should be monotonically decreasing. A class of decreasing densities on the unit interval is the class of $\text{Be}(\xi, 1)$ densities with parameter $\xi \in (0, 1)$. The density of a $\text{Be}(\xi, 1)$-distribution has the form $f(p \mid \xi) = \xi\, p^{\xi-1}$ with MLE $\hat{\xi}_{\text{ML}} = -1/\log(p)$ if $p < 1/e$ and 1 elsewhere. It is then easy to see that the marginal likelihood

$$f(p \mid H_1) = \int_0^1 f(p \mid \xi) f(\xi) d\xi \tag{16}$$

has the upper bound

$$f(p \mid \hat{\xi}_{\text{ML}}) = 1/(-e\, p \log p) \tag{17}$$

if $p < 1/e$ and 1 elsewhere, for any prior distribution $f(\xi)$. The upper bound (17) is obtained from (16) if $f(\xi)$ is a point mass prior at $\hat{\xi}_{\text{ML}}$. With $f(p \mid H_0) = 1$ it directly follows that the minimum Bayes factor is

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

C. Derivation of the quantile function (11)

If the first parameter of the F-distribution is equal to 2, then the corresponding density function has a simpler form than in the general case. We take advantage of this fact to

derive an analytic formula for the quantile function of the $\mathrm{F}(2, m)$ distribution in this special case. Indeed, the density of the $\mathrm{F}(2, m)$ distribution can be written as

$$f(x) = \left(1 + \frac{2x}{m}\right)^{-(1+m/2)}$$

and the cumulative distribution function turns out to be

$$F(x) = 1 - \left(1 + \frac{2x}{m}\right)^{-m/2}.$$

The inverse function of F is then the quantile function

$$Q_{\mathrm{F}(2,m)}(p) = \frac{m}{2}\left\{(1-p)^{-2/m} - 1\right\}.$$

Setting $m = n - 3$ and replacing p by $1 - p$ then yields (11).

List of Figures

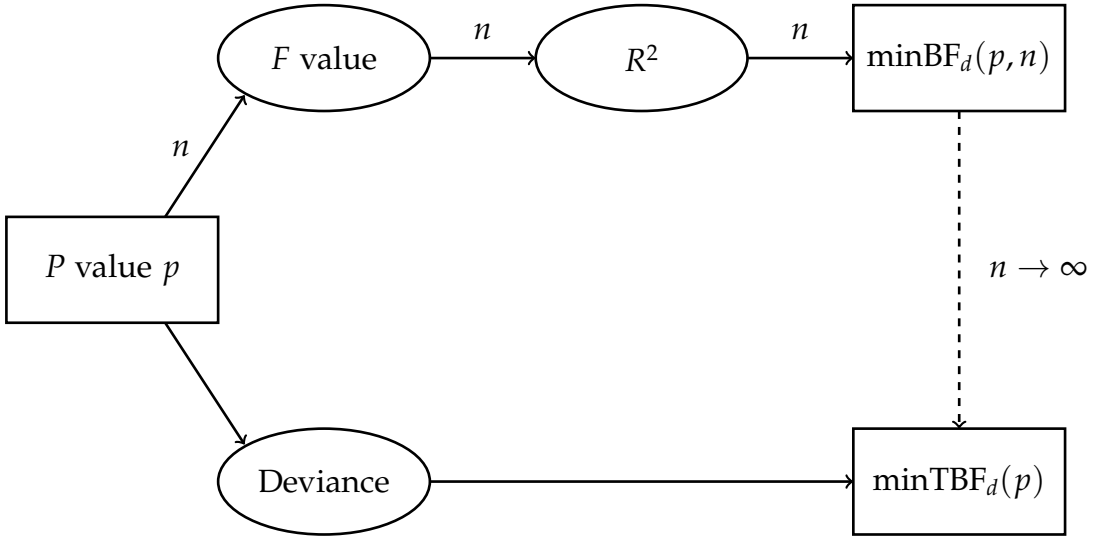

Figure 1: Schematic illustration how to compute minimum Bayes factors for a given P value p. The lower path to $\text{minTBF}_d(p)$ is the traditional way and does not take the sample size n into account. The upper path to $\text{minBF}_d(p, n)$ is the proposed way to incorporate the sample size n. Asymptotically we have $\lim_{n \to \infty} \text{minBF}_d(p, n) = \text{minTBF}_d(p)$.

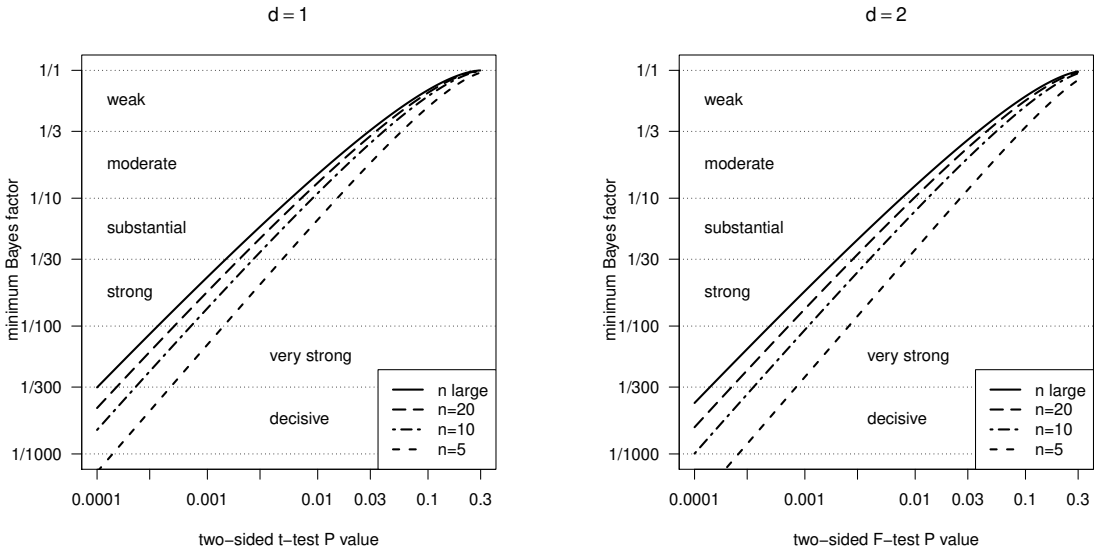

Figure 2: Minimum Bayes factors as a function of the P value from a t- or F-test. Shown are the asymptotic bounds (3) and (4) and the proposed bounds for $d = 1$ (left, no analytic formula) and $d = 2$ (right, formula (12)) for sample size $n = 5, 10, 20$. The areas delineated are the levels of evidence proposed in Table 1

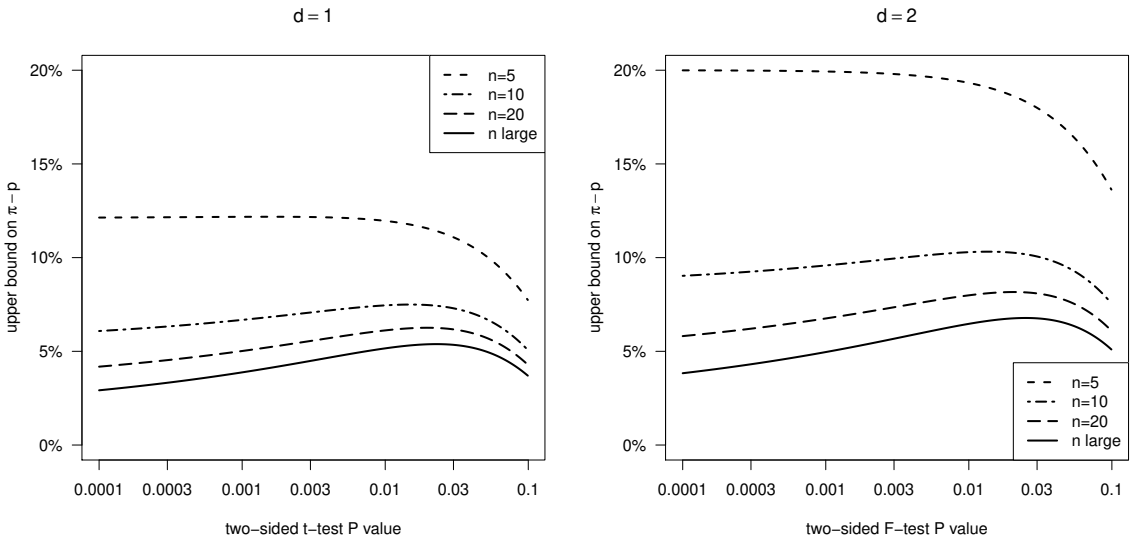

Figure 3: Upper bound on the difference $\pi - p$ (in percentage points) for exact P values p, assuming that p equals the posterior probability of the null hypothesis. The upper bound is shown as a function of the P value p for different values of the sample size n and the two different calibration schemes: The bound (3) ($d = 1$) and its sample size adjusted version shown in Figure 2 on the left and the bound (4) ($d = 2$) and its sample size adjusted version (12) on the right.

List of Tables

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

Table 1: Categorization of Bayes factors BF < 1 into levels of evidence against H_0

	Minimum Bayes factor			Decrease in probability of the null hypothesis from 50% to no less than			Decrease in probability of the null hypothesis from 75% to no less than		
P value	n large	n=20	n=10	n large	n=20	n=10	n large	n=20	n=10
0.1	0.63	0.58	0.52	38%	37%	34%	65%	64%	61%
0.05	0.41	0.36	0.31	29%	27%	24%	55%	52%	48%
0.03	0.29	0.25	0.21	22%	20%	17%	46%	43%	38%
0.01	0.13	0.1	0.08	11%	9%	7%	27%	23%	19%
0.005	0.07	0.06	0.04	7%	5%	4%	18%	15%	11%
0.001	0.02	0.01	0.01	2%	1%	1%	5%	4%	3%

Table 2: Bayesian interpretation of P values. Table adapted from Goodman (2001). For large n, calibration (4) was used and for $n = 20, 10$ calibration (12)