The Bayes Factor vs. P-Value

Albert Assaf and Mike Tsionas

Abstract

The use of p-value for hypothesis testing has always been the norm in the tourism literature. The aim of this paper is to highlight some of the "misconceptions" of p-value and illustrate its performance using some simulated experiments. Importantly, the paper proposes the use of the Bayes factor as an attractive alternative for hypothesis testing in the tourism literature. As the Bayes factor is based on the Bayesian approach, which relies solely on the observed sample to provide direct and exact probability statements about the parameters of interest, it is more suited for the purpose of hypothesis testing. Importantly, in this paper we show that the Bayes factor is more objective and has nicer properties than the p-value, a fact that should be of interest irrespective of whether the user is Bayesian or not. We discuss in more details the concept of Bayes factor, and propose other related strategies aiming to improve the process of hypothesis testing in the tourism literature.

1. Introduction

> " A hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. This seems a remarkable procedure (Sir Harold Jeffreys, 1939, p. 316)"

While it is highly common in tourism research to use the concept of p-value to test hypotheses, we believe it is time to highlight the advantage of the Bayesian approach for hypothesis testing. The reliance of p-values should not be discouraged, but at the same time, p-values should not be blindly used, or poorly interpreted. As Goodman (2008, p.135) stated, the interpretation of p-value has been "made extraordinarily difficult because it is not part of any formal system of statistical inference". Recently, the American Statistical Association (ASA) has also released a statement on p-values, focusing on aspects and issues that are too often "misunderstood and misused" (Wasserstein and Lazar, 2016). We also support the view shared by Ionides et al. (2017) that any method that is very commonly used will often be misused and misinterpreted. This should not however be taken as a statement for researchers and practitioners to completely avoid p-value (Ionides et al. 2017). Rather the focus should be on highlighting some misconceptions about the method and suggest other alternatives to either complement it or replace the p-values with methods that have better statistical interpretations and properties. In this paper, we highlight the advantages of the Bayesian approach for such purposes, discuss how it can provide better statistical interpretations, and lead to more correct and consistent hypotheses statements even in the sampling – theory context.

Specifically, we focus on contrasting the p-value with its Bayesian counterpart, the Bayes factor, which is more suited to address the problem of comparing hypotheses. Across several fields of social science, the Bayesian approach is gaining increased popularity because it expands "the range of testable hypotheses and results can be interpreted in intuitive ways that do not rely on null hypothesis significance testing" (Zyphur and Oswald, 2013). From a Bayesian perspective, Bayes factors can be considered "as alternatives to p-values (or significance probabilities) for

testing hypotheses and for quantifying the degree to which observed data support or conflict with a hypothesis" (Lavine and Schervis, 1999, p.19).

So far, in tourism studies, there have been only very few applications of the Bayesian approach (e.g. Tsionas and Assaf, 2014; Assaf and Tsionas, 2015). With the traditional p-value approach, we assume uncertainty in the data (in the form of a sampling distribution), and we fix the parameter to a certain value (e.g. H_0: $\beta=0$). Hence, we conduct hypothesis testing "even though everyone knows the null hypothesis is false before conducting the analysis (the probability that an unknown parameter is any single value is always equal to zero)" (Zyphur and Oswald, 2015, p. 392).[1]This puts us in a difficult situation of testing against a null hypothesis that does not truly reflect our belief, and whether we reject the null or no, we cannot reflect how probable the null is. In fact, "p-values reflect only the probability of the estimated effect, assuming the null is true", and the calculation of p-value is based on an infinite number of replications that never really happened.

Suppose, for example, we have a certain parameter β in a regression model and we wish to test whether it is statistically different from zero. Suppose the p-value of the usual t-test is 0.001. Most researchers would argue that, at significance level $\alpha=0.05$, since $p<\alpha$, we must reject the null hypothesis (that $\beta=0$). Clearly, the p-value is not the probability that the null is true. The p-value is only the probability of observing results as extreme or more extreme than the observed data, given that the null holds true. The null $H_o:\beta=0$ involves an unknown parameter β, which is not a random variable but a constant. Hence, there can be no such thing as the probability of the null being true or false in sampling-theory based statistics. At best, the p-value provides indirect evidence about the null hypothesis, as the parameters are not allowed to be random variables. Moreover, it is known that the p-value might overstate the evidence against the null (Berry, 1996; 2005; Goodman, 1999, 2005, 2008; Kass and Raftery, 1995; Louis, 2005; Spiegelhalter et al., 2004). There are of course other potential misconceptions of p-value, which we aim to highlight in this paper.

The Bayesian approach enables us to make direct probability statements about the parameter of interest (e.g. β). We elaborate further on these issues in the paper. We start with a more detailed discussion of the various misconceptions about p-values. We then provide a background of the Bayesian approach and its plausible advantages over the traditional p-value approach for hypothesis testing. We introduce the concept of Bayes factor and provide some background details on how it can be calculated. Throughout the paper, we will provide various illustrations on the performance of both the p-value and the Bayes factor.

2. The p-value

We focus here on some on two important aspects: Interpretation and performance of p-value. Different misconceptions about p-values have been identified in the literature (Goodman, 2008,

[1] To be fair, most researchers do not really believe that the parameter is zero but rather "close enough" to zero for their purposes.

Rouder et al. 2009; Zyphur and Oswald, 2013). We elaborate on some of these misconceptions here. We also conduct an experiment to illustrate the performance of p-value before discussing the Bayesian approach and the concept of Bayes factor.

2.1. Misconceptions about the interpretation of p-value

1. Let us assume that we are using a 5% significance level to test the impact of tourism on economic growth, and find that the p-value is 0.04. In this case, we would reject the null. While this is statistically fine, it is of course one of the main mistakes we make in our understanding of p-value. A simple way to illustrate this, is that the p-value is derived with the assumption that the null is true, so how it can be also the *probability* that the null is false? The only way we can calculate the exact probability is by using the Bayesian approach. This illustrates immediately that the p-value is not the probability that the null is "wrong".

2. Let us, on the other hand, assume that we found a non-significant impact of tourism on economic growth ($p>0.05$). Hence, we would assume that the observed data support the null hypothesis. However as discussed by Goodman (2008, p.138), "this does not make the null effect the most likely. The effect best supported by the data from a given experiment is always the observed effect, regardless of its significance".

3. Going back to the case where we find significant impact of tourism on economic growth, we generally tend to say that the impact of tourism on economic growth is highly important. However, the p-value does not reflect the size of the impact.

4. If another study conducted on a different destination Y found a conflicting finding to our study conducted on destination X we tend to say that the results from this study contradict other findings from the literature. However, we all know that the results are only conflicting when the findings have not likely occurred by chance. For example, p-value is sensitive to sample size and the data used, as it is inherently a sampling-based concept. However, most researchers erroneously think that small p-values tend to persist in similar samples.

5. Even in the case when the two studies find exactly the same p-value, one cannot safely assume that the two tourism industries for destinations X and Y behave similarly. For example, even highly different effects can result in the same p-value.

6. The relationship between p-value and type I error needs also to be highlighted. For example, setting significance level $\alpha=0.05$ assumes that rejecting the null will result in 5% type I error. However, this goes back to limitation 1. For instance, a type I error simply means that there is a significant "difference" when no "difference" exists. However, "if

such a conclusion represents an error, then by definition there is no difference. So a 5% chance of a false rejection is equivalent to saying that there is a 5% that that the null hypothesis is true", which is in line with what we say in limitation 1. (Goodman, 2008, p.138).

7. A p-value of 0.05 is not the same as a p-value of <0.05. Recent studies have highlighted the importance of reporting the exact p-value and not write them as inequalities. If the interest is just on testing hypotheses, then reporting the p-value as an inequality is fine, but if we would like to assess the strength of the effect, then reporting the exact p-value is always recommended. Unfortunately, such practice is not yet common in the tourism or social science literature in general, but other related fields such as psychology have now moved in that direction.

2.2. Evaluating the performance of p-value

Not only the interpretation of p-value but also its performance has been a subject of criticism in the literature. We focus here on this issue using the following experiment: Let us assume that there is a single regressor $x_i, i = 1,...,n$ which we generate from a standard normal distribution and remains the same in repeated samples. The data generating process is

$$y_i = \beta x_i + u_i, i = 1,...,n,$$

where $u_i \sim iidN(0,1).$ and β will assume different values. We start the analysis setting β=0. The estimated model contains an intercept, viz.

$$y_i = \beta_1 + \beta_2 x_i + u_i, i = 1,..,n.$$

Of course, we expect that $\beta_1 = 0$ and $\beta_2 = \beta$. Let b_1 and $b_2 \equiv b$ denote the least squares (LS) estimates of β_1 and β_2, respectively. The t-statistic for testing $H_o : \beta_2 = 0$ is:

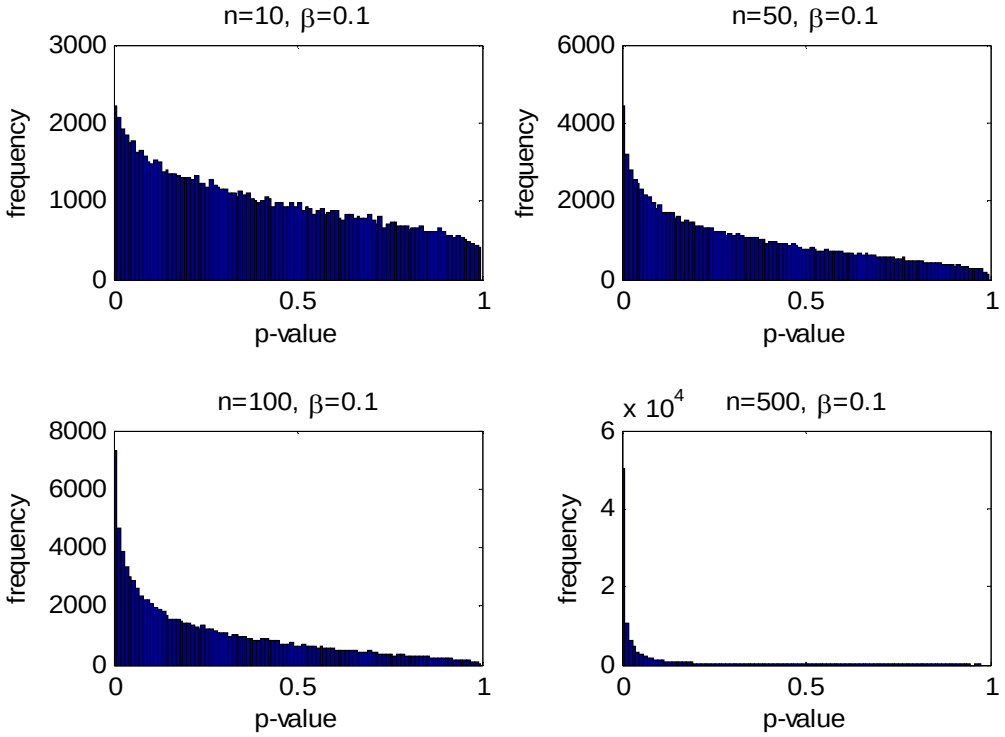
$$t = \frac{b}{\sqrt{\text{var}(b)}},$$

where $\text{var}(b) = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and $s^2 = (n-2)^{-1}\sum_{i=1}^{n}(y_i - b_1 - bx_i)^2$. The t-statistic follows the Student-*t* distribution with $n-2$ degrees of freedom.

The question that we wish to address concerns the sampling behavior of p-value in relation to H_o. We begin our illustration with the case $\beta = \beta_2 = 0.1$, that is the null hypothesis is incorrect but it is unlikely that we will be able to reject the null since the error variance is relatively large.

For this purpose, we conduct 100,000 simulations. The results are reported[2] in Figure 1. We can see that even in samples of size n=100 or n=500 there is a high probability that the null will be accepted.

FIGURE 1. Sampling behavior of p-values when β=0.1



Suppose now $\beta=0$ so that the null is correct. The sampling distribution of p-values is reported in Figure 2. It turns out that the sampling distributions are uniform! Therefore, decision—making based on p-values will be, more or less, arbitrary. This result is well known in the statistical literature (see for example Rouder et al. 2009) but not in applied research. We find it particularly disturbing because what it means is that, when the null is correct, one can obtain any p-value with equal probability in (repeated or observed) samples.

Suppose now we set $\beta=0.5$. The results are reported in Figure 3. Although the results are "better", there are still samples in which the null will be accepted even when n=100. In samples of size n=500 this will not happen. Suppose now that $\beta=1$ but the errors follow a Cauchy distribution, that is a Student-*t* with one degree of freedom (defined as the ratio of two independent standard normal random variables). The results are reported in Figure 4. Even in samples of size n=500 there is a sizable probability that we will obtain samples in which the null will be accepted even though it is wrong.

[2] All software used in the present study is available from the authors. All programs are written in WinGauss and Matlab.

FIGURE 2. Sampling behavior of p-values when β=0

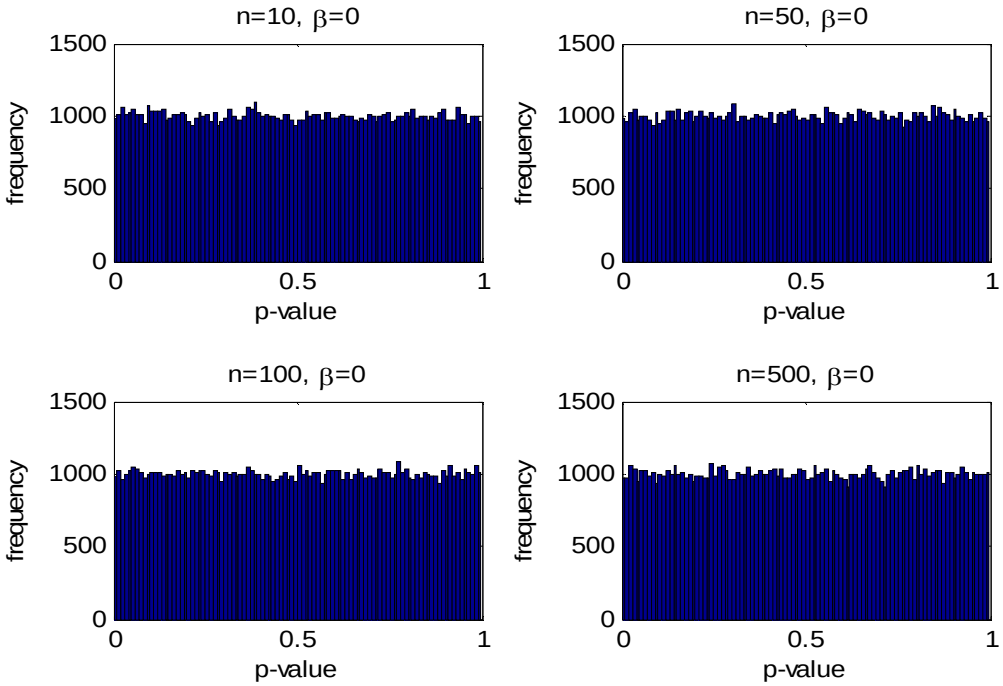


FIGURE 3. Sampling behavior of p-values when β=0.5

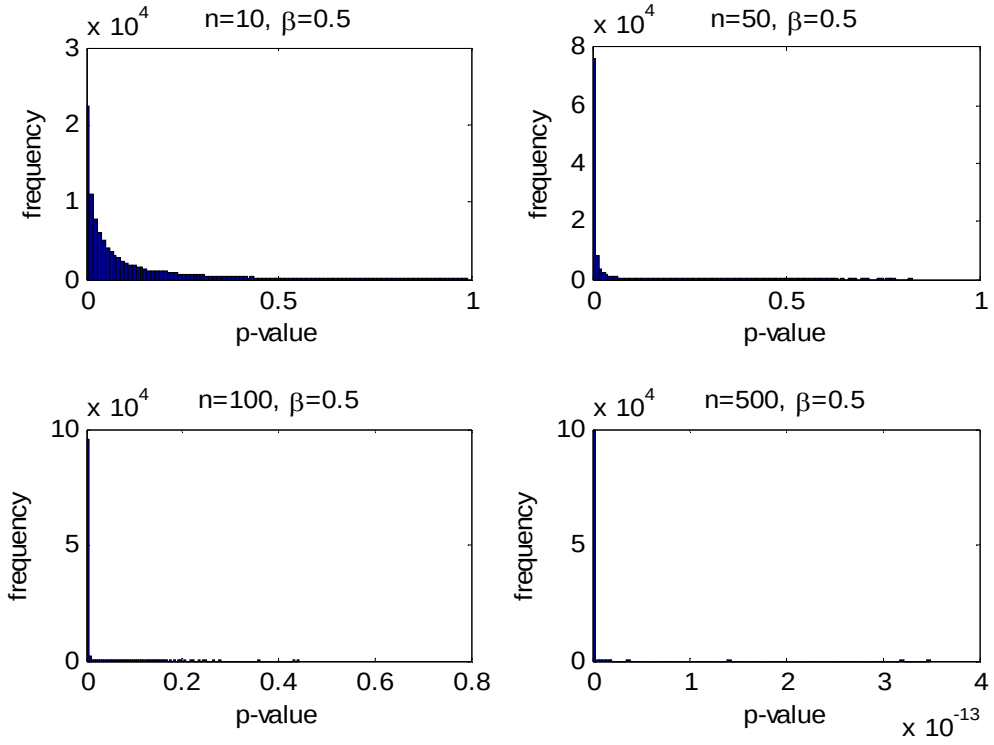
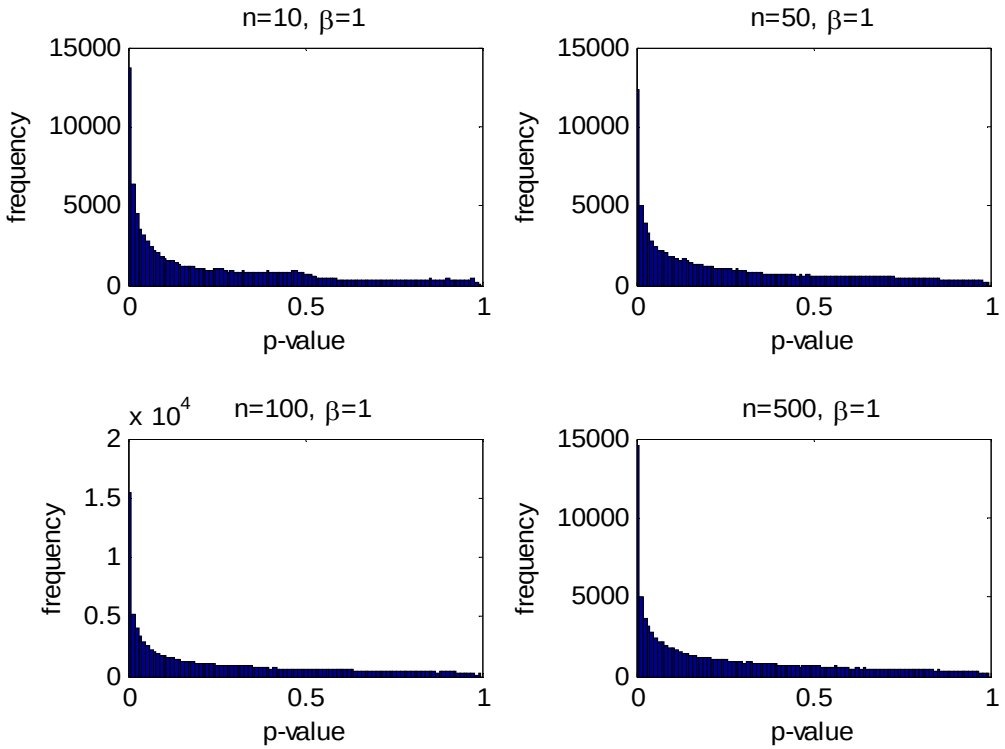


FIGURE 4. Sampling behavior of p-values when β=1 and errors are Cauchy



The fact that *p-values overstate the evidence against the null* is best illustrated in Figure 2 where the null is correct but the probability that we can obtain a value near zero or a value near one, are equal! Even when the null is incorrect, Figures 3 and 4 illustrate nicely the fact that there are samples in which we falsely accept. Clearly, this does not have to do with type-I or type-II errors; it is an "independent" property of p-value. In the words of Goodman (1992): "[T]he probability of repeating a statistically significant result, the 'replication probability', is substantially lower than expected". Therefore, using the conventional rule "$p < 0.05$" is suspect, and cannot be used in practice to provide evidence of "significance"[3].

Reinforcing this issue, Goodman (1992) considered a similar experiment and asked the following question which is of central interest in applied tourism research: "*If we repeat this experiment under identical conditions (similar groups, same sample size, same intervention), what is the probability of observing another statistically significant result in the same direction as the first*?" On his results Goodman (1992) wrote: "*if an experiment produces a correct result with $p = 0.01$, there is still greater than a one in four chance that a repetition of this experiment will not reach statistical significance at $a = 0.05$. We do not achieve a 95 per cent probability of replication until $p = 0.00032$, and 99 per cent probability at $p = 2 \times 10^{-5}$. That this is a best-case scenario is confirmed by the Bayesian calculations, which more accurately reflect the fact that there is still uncertainty about the true difference after the first experiment. They show that when $\boldsymbol{p} < 0.05$, the replication*

[3] This is, of course, a great problem in applied research because we cannot hope to increase the sample size in order to gain more information about the null.

probabilities are even lower." (pp. 876-877). Therefore, the conventional rule "$p < 0.05$" or even "$p < 0.01$" might not be totally unjustified, and we may need much lower bounds for p-values such as 10^{-5}. For more details, the interested reader can consult Edwards et al. (1963); Goodman (1999); Jeffreys (1961); Sellke, Bayarri, & Berger (2001) and Wagenmakers & Grȯnwald (2006). A related problem is that in large samples even phenomenally small differences become statistically significant. This suggests that bounds for p-values *cannot*, in fact, be independent of the sample size. In other words, critical values for *t*-tests *cannot*, in fact, be independent of the sample size. The problem does not seem to be well known in applied tourism research.

To avoid all the above limitations and dilemmas, one can resort to the Bayesian approach. We will illustrate below the main differences between the Bayes factor and p-value for hypothesis testing.

4. The Bayesian Approach and the Concept of Bayes Factor

Fundamentally, the Bayesian approach and the frequentist (i.e. p-value) approach have each their unique characteristics. The Bayesian approach however is more realistic as it provides inferences and compares hypotheses for the given data we analyze. With the frequentist approach, the data always carrying uncertainty. However, the "Bayesian thinking places the focus on whether or not parameters and models are sensible for a set of data (e.g., credibility intervals instead of confidence intervals), rather than whether a specific "correct" model has been specified or a null model is rejected" (Zyphur and Oswald, 2013, p.3). Hence, with the Bayesian approach we do not need to worry about setting up the null and we can make direct probability statements about a certain parameter β. The Bayesian approach does this in a somehow reverse order, by referring "directly to the parameter β itself (instead of treating parameters as fixed null hypotheses), and observed data in *y* are treated as fixed" (Zyphur and Oswald, 2013, p.3). In other words, the Bayesian approach makes direct probability statements about the parameters using the observed sample, whereas the p-value is calculated based on the assumption of drawing a hypothetical infinite number of samples (i.e. sampling distribution) that we never really observe. Finally, the Bayesian approach is also known to deal better with small samples, where on the contrary, it is highly unlikely to support hypothesis with p-value due to statistical power problems. The Bayesian approach has also the advantage of incorporating prior information (about previous findings and theory) into the estimation, which sometimes can prove to be highly useful.

All the above advantages can translate to the concept of Bayes factor, which as mentioned, is the Bayesian counterpart of the p-value for hypothesis testing. The Bayes theorem can be written as:

$$p(\theta \mid y) \propto p(\mathrm{y} \mid \theta)\, p(\theta) \tag{1}$$

where $\propto$ is the proportionality symbol. Here, $p(\theta \mid y)$ is the posterior distribution, and is proportional to the product of the prior $p(\theta)$ and the likelihood function $p(\mathrm{y} \mid \theta)$, which summarizes the information from the data. Hence, from here, we can draw the distinction between the Bayesian and the frequentist approach. For example, with the Bayesian approach,

the posterior is conditional on the data, while with the frequentist approach "we consider the data [as] random, and we investigate the behavior of test statistics over imaginary samples from $p(\mathrm{y}|\theta)$" (Rossi and Allenby, 2003, 305).

To place the Bayes factor into the equation, the theorem in (1) can be written in words as :

Odds of the null hypothesis after obtaining the data

=Odds of the null hypothesis before obtaining the data× Bayes factor

or in more technical details as:

Posterior odds (H_0, given the data) (3)

$$= \text{Posterior odds } (H_0, \text{given the data}) \times \frac{P(Data,\ under\ H_0)}{P(Data,\ under\ H_1)}$$

Hence, the Bayes factor is nothing but the ratio of the posterior probabilities of the two hypotheses, viz.:

$$BF_{1:2} = \frac{P(H_0|Y)}{P(H_1|Y)}. \quad (4)$$

This is naturally defined in Bayesian analysis but it has no meaning in sampling—theory statistics[4]. How do interpret the Bayes factor?

- Bayes factor of 1 indicates no evidence (i.e. equal support for both hypotheses).
- Bayes factor between 1 and 3 indicates anecdotal evidence for H0.
- Bayes factor between 3 and 10 indicates substantial evidence for H0.
- Bayes factor between 1/3 and 1 indicates anecdotal evidence for H1.
- Bayes factor between 1/10 and 1/3 indicates substantial evidence for H1.

We should notice here the difference between Bayes factor and Posterior Odds Ratio. The POR is the BF multiplied by the ratio of prior probabilities that we attach to the two hypotheses. Since we rarely have information to come up with a reasonable prior odds ratio we set it to unity (so that the two hypotheses receive the same prior probability) and, therefore the BF is simply the POR.

We provide in Appendix A some more technical details about the Bayes factor for regression models under alternative priors. Hence, from this discussion, it turns out that in Bayesian

[4] It is, perhaps, worthwhile to compare the marginal likelihood to the maximum value of the likelihood function, $L(\hat{\theta};Y)$ where $\hat{\theta}$ is the maximum likelihood (ML) estimate. Even if the likelihood is multimodal, the ML estimate is the global maximum. In contrast, the marginal likelihood weights all values of θ using the prior as the weighting function and thus takes into account multimodality in the proper manner

analysis we *compare* hypotheses rather than test a null against an alternative. In other words, we are interested in the *plausibility* of one hypothesis relative to another, given the data.

5. Why is the Bayes factor an attractive alternative to the p-value?

1- If we look at (3) and (4), we can see that the Bayes factor is not probability itself but a ratio of probabilities, ranging from zero to infinity. It necessitates two different hypotheses: "for evidence to be against the null hypothesis, it must be for other alternative" (Goodman, 1999, p.1006). In other words, it has a more objective interpretation. For example, a Bayes factor of 0.5 simply indicates that the results we observe have half the probability under the null as they are under the alternative. To explain further: P- values "are based on calculating the probability of observing test statistics that are as extreme or more extreme than the test statistic actually observed, whereas Bayes factors represent the relative probability assigned to the observed data under each of the competing hypotheses. The latter comparison is perhaps more natural because it relates directly to the posterior probability that each hypothesis is true" (Johnson, 2013, p.19313).

2- The Bayes factor relies *only* on the observed data at hand, and *not* on some hypothetical repeated samples, which we do not observe and they are the essence of the calculation of the p-value.

3- The Bayes factor accounts for the likelihood under both H_0 and H_1 and provide evidence for and not only against H_0. This is contrary to the (admittedly peculiar) habit of frequentists who do not reject a null but cannot accept it.

4- The Bayes factor is more convenient at it shows the size of an effect. This is because the Bayes factor is a ratio of probabilities, while the p- value is a probability to obtain a more significant result in repeated sampling which never occurs.

5- It is a widespread belief that the Bayes Factor approach significantly depends on correctness of parametric assumptions and on prior selections (Vexler et al. 2016). This is true but the p-value approach also depends on correctness of parametric assumptions. If the parametric model is misspecified there is no way for a Bayesian or a non-Bayesian researcher to reach the correct conclusion. We also agree that the Bayes Factor approach significantly depends on prior assumptions. We show through our simulations below that for given 'hostile' priors (the term 'hostile' will be defined below) and a nearly-flat prior the sampling behavior of the Bayes factor is much better to the sampling behavior of p-values. For example the Bayes factors do not exhibit the 'dancing phenomenon' that is well known for p-values and which we also document in this paper. The performance of Bayes factors is excellent and certainly better compared to the sampling-theory performance of p-values.

6. Evaluating the Behavior of the Bayes factor and comparing its performance with the p-value

To illustrate the behavior of the Bayes factor we consider in this section several experiments. In the next section, we also provide the results from a real data set.

We start with the following model:

$$y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + u_i, i = 1,...,n = 50,$$

where we generate $x_{i1} \sim iidN(0,1)$, $x_{i2} = x_{i1} + e_i, e_i \sim iidN(0,0.1^2)$ and $u_i \sim iidN(0,1)$. We set $\beta_1 = \beta_2 = \beta_3 = 1$, and $\beta \mid \sigma \sim N(0, C*\sigma^2)$, where $\sigma^2 = 1$ and C is a constant that changes the conjugate prior. Therefore, we have the regressors correlated. We also assume that they are not fixed in repeated samples. We use one million simulated data sets. We are interested in the Bayes factor of the full model (M1) against a model where x_{i1} is omitted (model M2) and a model where x_{i2} is omitted (M3). With probability 95% the parameters β_j will be in the interval [$-2\sigma\sqrt{C}$, $2\sigma\sqrt{C}$] since their prior mean is zero. For ease of presentation, we present the log Bayes factors in Figure 5. We use various prior choices for C including C= 0.1, C=10, C=10^{-4}, C=1, and C=100, which we believe are common prior choices given that $\sigma^2 = 1$. As the actual values of parameters β_j are equal to one, then it follows that conditional on σ=1, priors with C>1 imply 95% Bayes probability intervals that include the actual values of the parameters. Setting C=0.1 or C=10^{-4} implies more or less dogmatic priors (extremely so in the second case) while 100 is a practically flat prior (relative to the likelihood). Setting C=10 is also flat relative to the prior. So the only 'reasonable' prior is the one with C=1 which implies a 95% Bayes probability interval [-2, 2]. By 'reasonable' we mean a prior corresponding to some educated guess about the parameter[5]. The other choices of C lead to either dogmatic or flat priors. For example with C=10 the 95% Bayes probability interval becomes [$-2\sqrt{10}$, $2\sqrt{10}$]. With C=100 it becomes [-20, 20]. The likelihood is, clearly, far more concentrated around the actual parameter values (unity) so these priors are flat and practically uninformative in this example. The choice C=0.2 implies a 95% Bayes probability interval of the form [$-2\sqrt{0.2}$, $2\sqrt{0.2}$] which is far from the actual parameter value and the prior with C=10^{-4} is practically dogmatic, almost excluding the actual parameter value (which is unity) with very high prior probability.

So our priors are not helpful in 'driving' the 'correct result' (viz. that parameters are close to unity). Moreover, in this instance the situation for testing is hostile. Not only we have collinearity among the regressors but the error variance (σ) is very large whereas the sample size (n) is small. A 'helpful' prior would have been like: $\beta \sim N(1, C\sigma^2)$ with a value of C like, for example, 1, 5 or 10.

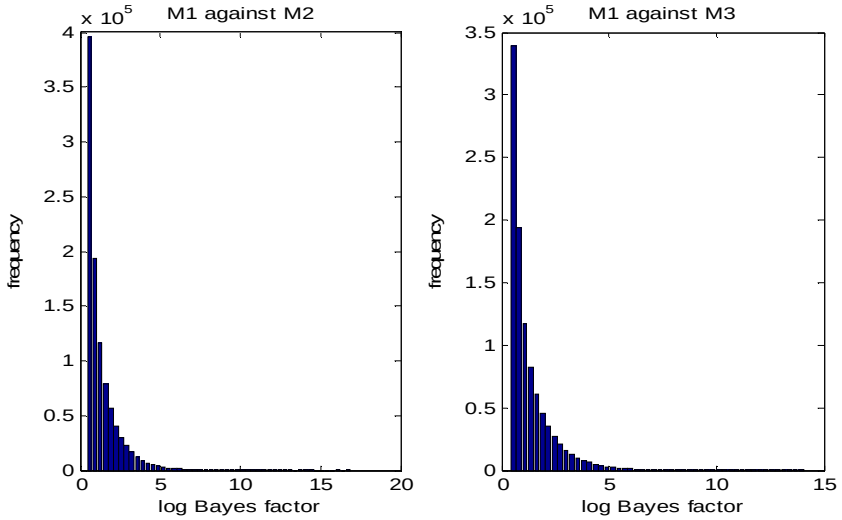
[5] An educated guess may come from previous studies of the same problem or associated economic theory information, e.g. in the case of elasticities, the marginal propensity to consume and several other similar situations.

Before proceeding we should mention that we present the sampling distribution of log Bayes factors and not Bayes factors themselves as the magnitudes of the latter prevent us from visual clarity. So in order for Bayes factor to be greater than 1 we need the log Bayes factor to be positive. For Bayes factor to be greater than 10 we need the log Bayes factor to be greater than 2.33 etc. If the sampling distribution of the log Bayes factor has a long tail to the right it means that extremely large values occur that apparently are in favor of the correct model.

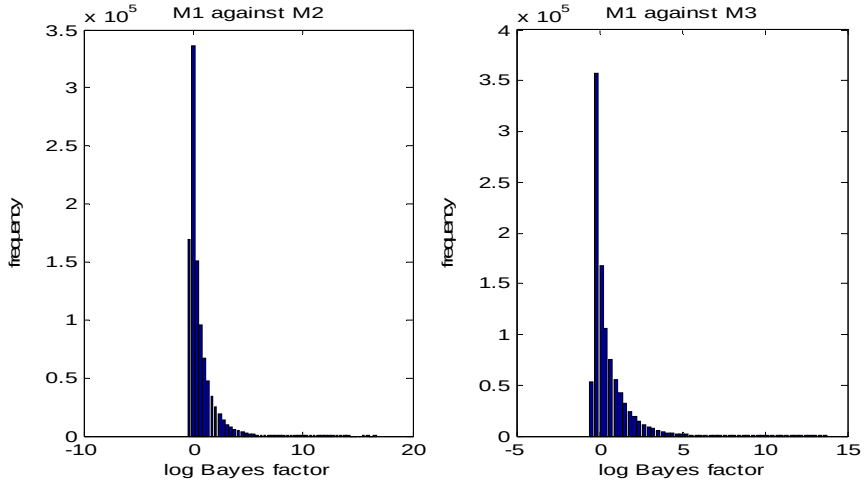
Note that with the error variance being equal to one, the "signal" is buried" into noise. Despite this fact, we can see that the Bayes factor performs consistently across various prior choices. For instance, if we look at the case with C=0.1, we can see that the Bayes factors are always favoring the full model. The minimum log Bayes factor, for example, is larger than 0 (i.e. Bayes factor are greater than 1), and this seems to be consistent across all priors. Importantly, we can see that the median log Bayes factors is approximately 3 10^5., Hence, the message from Figure 5 is clearly that Bayes factors favor the correct model as they should be, and seem to perform well across various priors despite the "hostile conditions" for testing and model comparison in this context.

FIGURE 5. Sampling distribution of log Bayes factors.

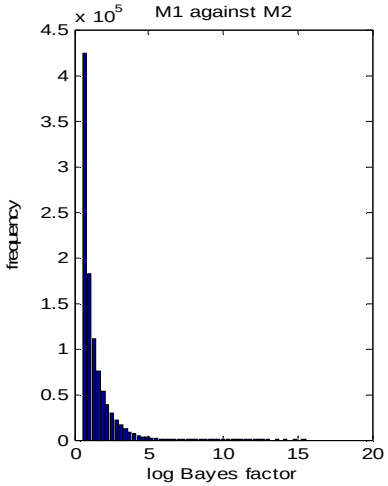
C=0.1

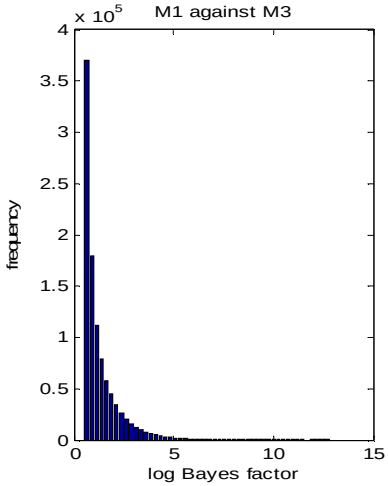


C=10

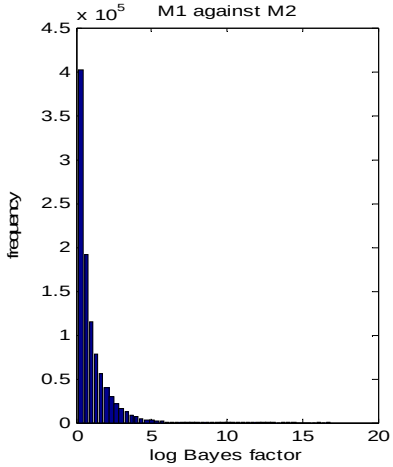


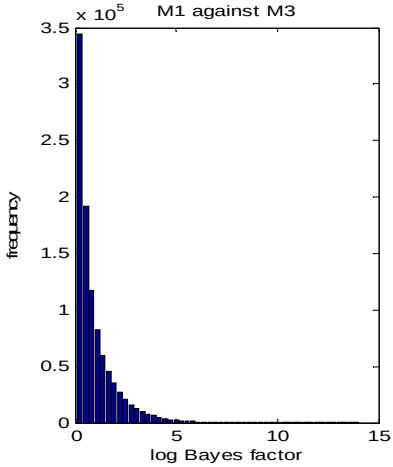
C= 10^{-4}


x 10^5
M1 against M2
frequency
log Bayes factor

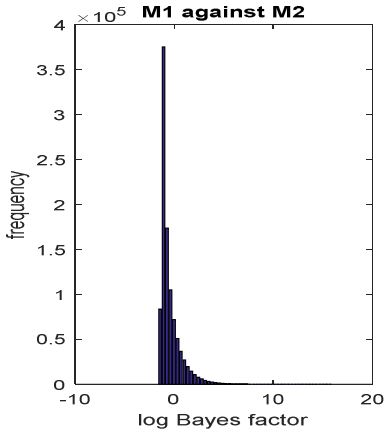

x 10^5
M1 against M3
frequency
log Bayes factor

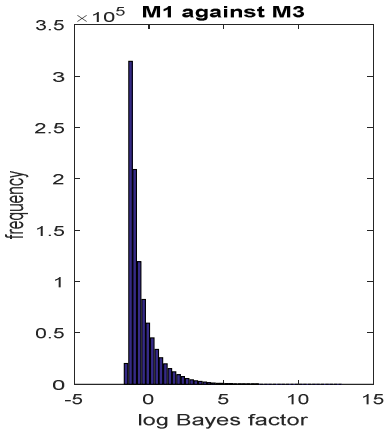
C=1


x 10^5
M1 against M2
frequency
log Bayes factor


x 10^5
M1 against M3
frequency
log Bayes factor

C=100


×10^5
M1 against M2
frequency
log Bayes factor


×10^5
M1 against M3
frequency
log Bayes factor

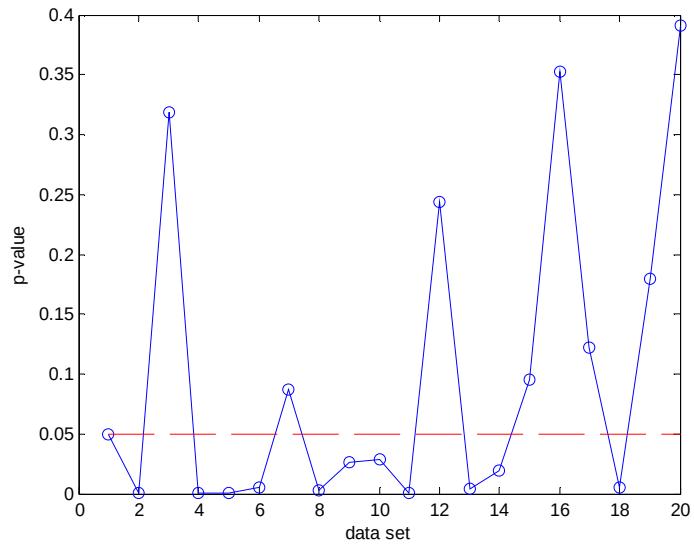
To further illustrate, we consider another experiment to compare directly between the p-value and Bayes factor. We start with the following data generating process, where β=1 so that there is an effect in the population:

$$y = \beta x_i + u_i, u_i \sim iidN(0, \sigma^2),\ i = 1,...,n = 100,\ \beta = 1,\ \sigma = 5.$$

We chose the value of standard deviation (σ) so that the power of the least square (LS) estimator for β is close to 0.50[6] (i.e. we assume only 50 out of 100 tests to be significant, where in fact there is real effect. Of course, in reality, we do not know such information). The standard deviation of the LS parameter of β is 0.5 (since σ=5 and the sum of squares of x_is follows a chi-square distribution with n=100 degrees of freedom, and therefore its expectation is 100). So the average t-statistic for testing the null that β=0 should be -1/0.5=-2. So this is not too bad for the sampling – theory estimator (OLS and associated p-value of the usual t-statistic).

We are interested in $H_o : \beta = 0$ against the two-sided alternative. In Figure 6 we can see the fluctuation or what is known as "the dance of p-values" (Cummings, 2011, Dienes, 2014). The dotted (red) line corresponds to the conventional significance level of 0.05. The actual p-values are reported in Table 1. We can see that in some situations, the actual p-value is showing significant effect, while in others do not. In other words, the p-value is making the mistake of accepting the null.

Figure 6. Dance of p-values



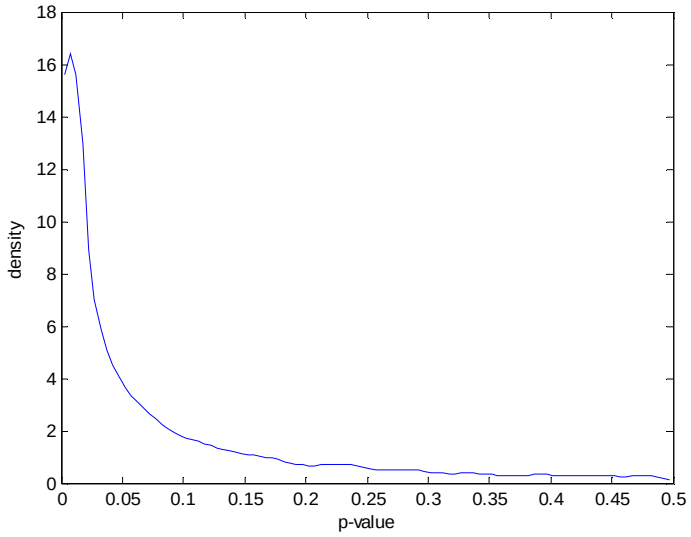
[6] We keep the x's fixed in repeated samples.

Table 1. Actual p-values from Figure 6

0.049370	0.000112	0.318299	0.000083	0.000831
0.005598	0.087391	0.003340	0.025822	0.028236
0.000929	0.756712	0.003966	0.018895	0.094902
0.353184	0.121725	0.005472	0.179371	0.391317

In Figure 7 we report the sampling distribution of p-values from a simulation using 10,000 replications.

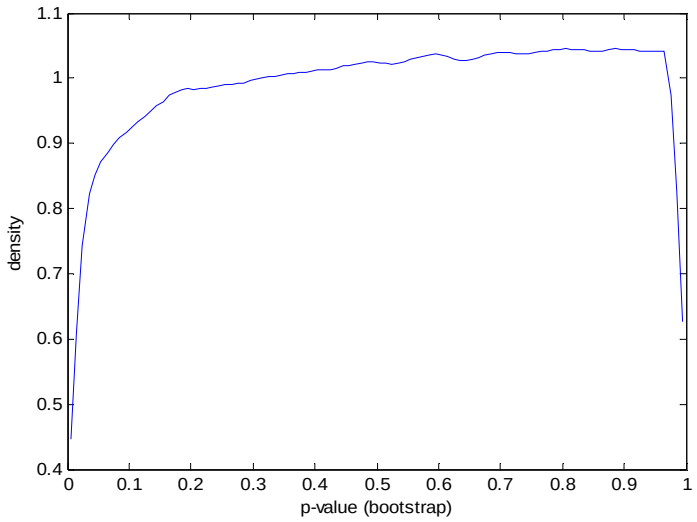
Figure 7. Sampling distribution of p-values



Suppose now we keep the 20th dataset for which the p-value was large, indicating that β=0. We perform a bootstrap[7] using a million replications and the resulting sampling density of p-values is reported in Figure 8. Contrary, perhaps, to expectations, this distribution is nearly uniform. Therefore, we observe, again, the "dancing p-values" phenomenon. The fact that the bootstrap does *not* work in this instance (in the sense that p-values do not appear to be low to indicate statistical significance) is because *the bootstrap has only an asymptotic justification*, and, in fact, it cannot perform well independently of $\sigma / \sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$ or $\sigma / \sqrt{n}$ in the case a simple mean is tested (i.e. $x_i = 1, i = 1,...,n$). The asymptotic justification of the bootstrap is not well known, and most applied researchers are inclined to believe that it provides a panacea.

[7] This works as follows. We resample n values from the y_i's randomly. The x_i's are fixed by assumption. For each data set generated we apply LS and we get an estimate of β, its t-statistic and the p-value. The p-values are saved and to these a kernel density smooth is applied.

Figure 8. Sampling density of p-values (bootstrap)



Next, we examine what happens when n=10 and σ=1, in which case the power is approximately 75%. Clearly, a lot of them indicate non-significance despite the fact that the null is false, that is β is not zero.

From Figure 10, the sampling and bootstrapped distributions are much improved but, due to low power, there is still a non-trivial probability to accept the null. To conclude, *there is little hope to rely on the bootstrap for computing corrected p-values* when the power of test statistic is low (even a power of 75% is low as we have seen). In fact the bootstrap-based distribution is nearly uniform, again.Therefore, the case for small-sample econometrics, that is a case that does not focus on asymptotic theory but rather small-sample based exact inference, does not appear to be promising. This is despite the facts that some Bayesians have advised in favor of it to cure, *at least*, the "asymptotic theory" disease that plagues modern applied econometrics practice. In fact, the bootstrap is not a panacea; the bootstrapped distributions of p-values can still be nearly uniform when the data at hand yield a large p-value in the original LS estimation.

More fundamentally, a p-value that indicates non-significance (that is, β=0) is quite common when, in fact, the hypothesis is correct (we have had β=1 in our case). Therefore, *large p-values cannot be taken as evidence supporting the null.* First, the dancing p-values phenomenon arises quite commonly. Second, the sampling and bootstrapped distributions of p-values can still allow for large p-values giving rise to the dancing p-values. Both phenomena are due to low power, and 75% is low in what we described in this section.

Figure 9. Dance of p-values when n=10 and σ=1.

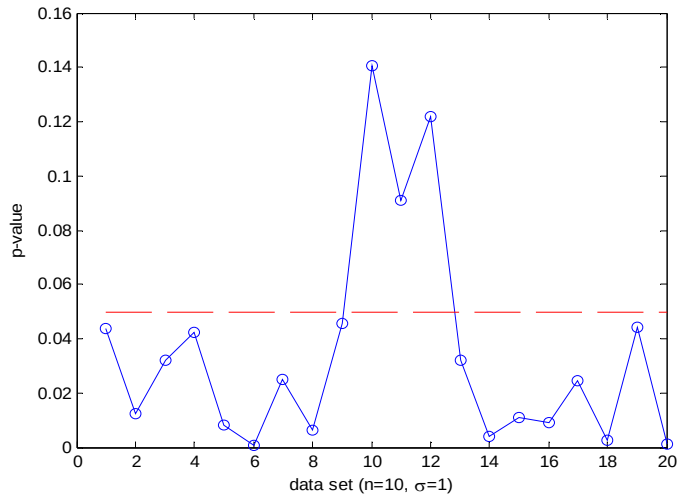


Figure 10. Sampling and bootstrap density of p-values when n=10 and σ=1.

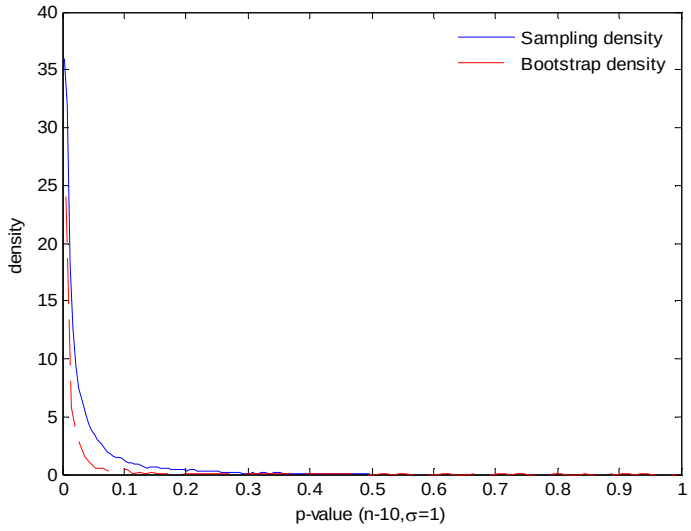


Table 7. The actual p-values for Figure 9

0.043521	0.012516	0.032072	0.042491	0.008104
0.000835	0.025236	0.006302	0.045610	0.140445
0.090976	0.121967	0.032085	0.003787	0.011040
0.009092	0.024772	0.002497	0.044316	0.001395

Finally, for the same experiment in Table 7, we report the Bayes factors in favor of the null hypothesis (Table 8). Our interest is to see to see whether the Bayes factor will reveal any non-significant outcome when the null is false. From the results, we can see that *the Bayes factor is less sensitive and does not fluctuate as much as much as the p-value*. For instance, we can see that the Bayes factor is never supporting the null, while the p-value does so, in many instances. Such results are in line with Dienes (2014), and of course, we are not claiming that the Bayes factor never produce inconsistent estimates, but at least it is less sensitive to this error than the p-value.

Table 8. p-values vs. Bayes Factor

p-value	Bayes Factor
0.043521	0.271805
0.012516	0.304833
0.032072	0.183034
0.042491	0.262004
0.008104	0.177460
0.000835	0.141651
0.025236	0.254825
0.006302	0.186759
0.045610	0.267727
0.140445	0.267324
0.090976	0.271406
0.121967	0.294731
0.032085	0.345209
0.003787	0.129610
0.011040	0.165317
0.009092	0.246224
0.024772	0.241461
0.002497	0.150977
0.044316	0.300740
0.001395	0.130336

7. An Illustration with a real dataset

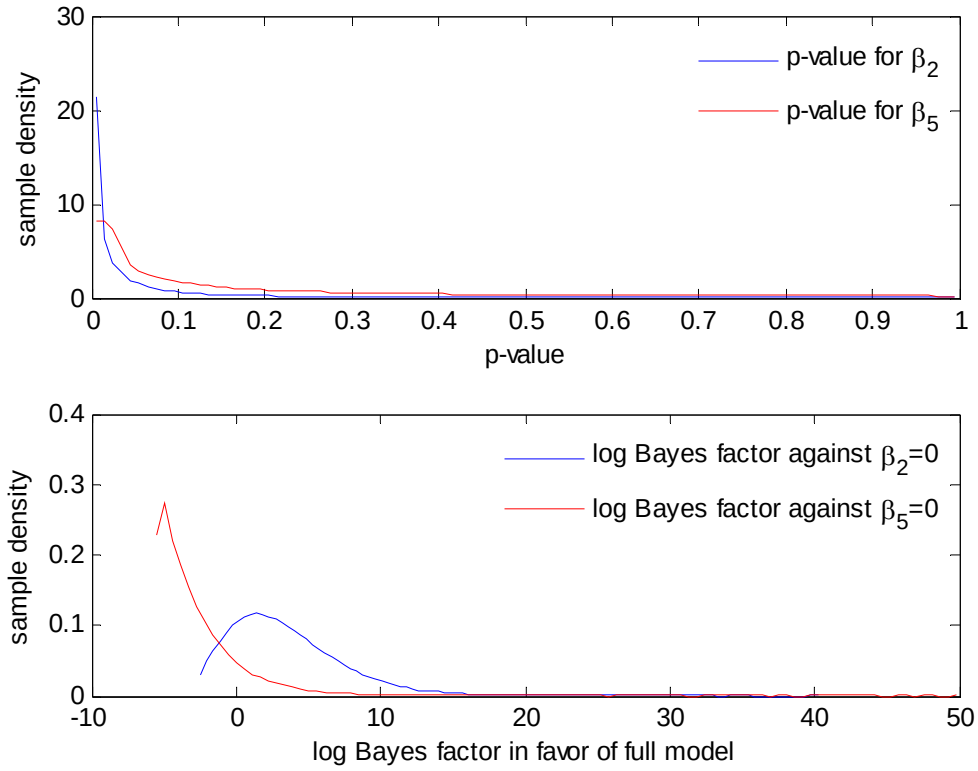
Finally, we compare between the Bayes factor and p-value using a real application on the impact advertising spending on firm value. We use data on US hotels and restaurants covering an unbalanced sample of 31 publicly traded companies[8] from 2001 to 2012 (314 observations). In our estimation we also control for advertising spending, firm size and financial leverage (Assaf et al. 2017). We used the COMPUSTAT database to collect advertising spending and performance data. Following previous research (Luo and De Jong, 2012), we measured advertising spending as the reported firm advertising expenditure in the COMPUSTAT database. For CSR data, we used

[8] We focus here on publicly traded firms as we have firm value as one of our performance measures.

the KLD Research and Analytics' KLD STAT, one of the most frequently used databases in the strategy and management literature (Kang et al. 2010). In line with Assaf et al. (2017), we measured firm value using the Market Value Added (MVA), calculated as: MVA= market value-capital , where market value reflects the equity market valuation of the firm and capital reflects the debt and equity invested in the firm.

Hence, in our model we have four variables: advertising spending, CSR, firm size and financial leverage. We report in Figure 11 the sample densities of the Bayes factor and p-value for the coefficients associated with advertising spending (β_2) and financial leverage (β_5), because this is where we observe interesting differences[9]. The sampling distribution of p-value for $\beta_2=0$ has a peak at zero indicating that this parameter can be significant but on average the p-value is exceeds 0.1. The Bayes factor (lower panel) is approximately exp(7)=1096 on the average suggesting that the parameter is definitely not zero. For β_5 the sampling distribution of p-values suggests a similar thing and has a long right tail implying that in most samples the parameter will be non-significant. The Bayes factor (lower panel) in favor of $\beta_5=0$ is less than unity reaching the same conclusion but its sampling distribution has considerable probability to the right of zero suggesting that in many samples the conclusion will be reversed.

Figure 11. Sample density of the Bayes factor and p-value for selected coefficients



[9] Results for other variables can be obtained from the authors upon request.

8. Concluding Remarks

The aim of this paper was to highlight some of the misconceptions about p-value and illustrate its performance using some simulated experiments. Importantly, the paper also discussed the concept of Bayes factor and illustrated how it can serve as an attractive alternative to p-value in the tourism literature.

The paper is not necessarily advising against abandoning the p-value approach from the tourism literature. However, we suggest the following strategies moving forward, echoing the recommendations of other researchers in related fields (Stern, 2016):

1. As tourism researchers, we need to be more careful in how we interpret the p-value and highlight its limitations. Along with the p-value we should also provide more reflection on the effect size, and the distribution of data (Khalilzadeh and Tasci, 2017). We need also to start reporting the exact p-values and not write them as inequalities. However, the p-value has important drawbacks as we have discussed.

2. We need to start transitioning more quickly toward the use of the Bayesian approach for hypothesis testing. This is happening at a faster pace in other fields such as management and psychology (Zyphur and Oswald, 2013). We discussed some of the strengths of the Bayesian approach. It is simply more suitable and provides a more objective interpretation for hypothesis testing. More importantly, it is not based on the idea of repeating sampling like the p-value approach[10].

3. The Bayes factor discussed in this study is one counterpart to p-value offered by the Bayesian approach. In fact the Bayesian approach provides the exact probability that a given hypothesis H is true given the data Y, viz. $P(H|Y)$. We discussed above some of the main advantages of the Bayes factor. Deriving these Bayes factors can be somehow challenging. However, some well-known software such as SPSS now provides extension commands for the derivation of Bayes factors[11]. To simplify the process, we also provide in Appendix B, WinGauss code for the derivation of Bayes factors in a regression model, where the parameters have a normal prior and σ has an inverted gamma prior. As the prior of β is non-conjugate (that is, it does not depend on σ) results are not available analytically and we use a Gibbs sampler to perform MCMC. The Bayes factor, in turn, is computed using the Verdinelli and Wasserman (1995) approach known as Savage – Dickey ratio.

4. The Bayesian approach also provides other nice tools for hypothesis that can be used in tourism literature, and have better properties than the p-value. For example, one can use

[10] It is encouraging that most statistical software packages now provide some Bayesian estimation for regression models.

[11] See for example: (https://developer.ibm.com/predictiveanalytics/2015/10/07/new-bayesian-extension-commands-for-spss-statistics/)

the 95% higher posterior density (HPD) to reflect on the significance of a certain effect[12]. Like the confidence interval, the HPD can also inform us about the magnitude of uncertainty about a parameter. However, the HPD has nicer properties than the confidence interval. We elaborate mote on this issue in Appendix C, and for more details refer to Congdon (2001).

5. Finally, we note that "scientific studies involve much more than the statistical analysis stage. There are numerous other points in the research process as well" (Stern, 2016, p.8). The above should not distract us from designing carefully each stage of the research process. Both the p-value and the Bayes factor will not produce reliable results if errors exist in the research process.

References

Assaf, A. G., & Tsionas, E. G. (2015). Incorporating destination quality into the measurement of tourism performance: A Bayesian approach. *Tourism Management*, 49, 58-71.

Berry DA. (1996) *Statistics: A Bayesian Perspective*. Pacific Grove, CA: Duxbury Press.

Berry DA. (2005). Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. *Clinical Trials*; *2*(4):295–300.

Congdon, P. (2001). Bayesian Statistical Modelling. Wiley, Chichester, UK.

Cumming G. (2011). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Abingdon: Routledge.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, *5,* 781.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. P*sychological Review*, *70* (3), 193-242.

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

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of internal medicine*, *130*(12), 995-1004.

[12] If the HPD includes zero, the effect is considered non-significant.

Goodman SN. (2005). Introduction to Bayesian methods I: Measuring the strength of evidence. *Clinical Trials*, *2* (4), 282-290

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology*, 45 (3), 135-140. WB Saunders.

Ionides, E. L., Giessing, A., Ritov, Y., & Page, S. E. (2017). Response to the ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *71*(1), 88-89.

Jeffreys, H. (1939). Theory of probability. Oxford: Clarendon.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press, Clarendon Press.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Khalilzadeh, J., & Tasci, A. D. (2017). Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. *Tourism Management*, *62*, 89-96.

Lavine, M., & Schervish, M. J. (1999). Bayes factors: what they are and what they are not. *The American Statistician*, *53*(2), 119-122.

Louis TA. (2005). Introduction to Bayesian methods II: fundamental concepts. *Clinical Trials,* 2(4), 291–294.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225-237.

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.

Rossi, P. E., & Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, *22*(3), 304-328.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, *90*(430), 614-618.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation* (Vol. 13). John Wiley & Sons.

Tsionas, E. G., & Assaf, A. G. (2014). Short-run and long-run performance of international tourism: Evidence from Bayesian dynamic models. *Tourism Management*, *42*, 22-36.

Vexler, A., Zou, L., & Hutson, A. D. (2016). Data-Driven Confidence Interval Estimation Incorporating Prior Information with an Adjustment for Skewed Data. *The American Statistician*, *70*(3), 243-249.

Wagenmakers, E.-J., & Grónwald, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science*, *17* (7), 641-642.

Wasserstein, R. L., and Lazar, N. (2016), The ASA's statement on *p*-values: Context, process, and purpose, *The American Statistician* 70, 129–133.

Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions, in Goel, P. K. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*. Amsterdam: North-Holland.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, *41*(2), 390-420.

Zyphur, M. J., & Oswald, F. L. (2013). Bayesian probability and statistics in management research: A new horizon. *Journal of Management*, *39*(1), 5-13.

Appendix A

More Technical Details about the Bayes Factor

Bayes factor in the linear regression model

Undoubtedly, the linear regression model is the workhorse of applied tourism research. The model we consider is

$$y_i = \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i,\ u_i \sim iidN(0, \sigma^2), i = 1, ..., n.$$

In familiar notation, we have:

$$y = X\beta + u,\ u \sim N_n(0, \sigma^2 I).$$

The vector notation facilitates the calculations. Here, y is the $n \times 1$ vector of observations on the dependent variable, X is the $n \times k$ matrix of data on the explanatory variables and β is the $k \times 1$ vector of parameters. It is convenient to set $h = \sigma^{-2}$, a quantity known as "precision". From frequentist analysis we know that the LS estimator is $b = (X'X)^{-1}X'y$. We also define $\nu s^2 = (y - Xb)'(y - Xb)$ and $\nu = n - k$.

Case A. The Natural Conjugate Prior

Suppose we adopt the following prior for the parameters, known as natural conjugate:

$$\beta \mid h \sim N\left(\underline{\beta}, h^{-1}\underline{V}\right),\ h \sim Gamma\left(\underline{s}^2, \underline{\nu}\right).$$

Here, $\underline{\beta}$ is the prior mean of β and $h^{-1}\underline{V}$ is its prior covariance matrix. Since we wish to compare models, it convenient to write $y_{(m)} = X_{(m)}\beta_{(m)} + u_{(m)},\ u_{(m)} \sim N_n(0, \sigma^2_{(m)} I)$, where $m = M_1, M_2$ denotes the models. The models can be nested or non-nested.

For this model we define:

$$\overline{V} = \left(\underline{V}^{-1} + X'X\right)^{-1},$$

$$\overline{\beta} = \overline{V}\left(\underline{V}^{-1}\underline{\beta} + X'Xb\right),$$

$$\overline{\nu} = \underline{\nu} + n,$$

$$\overline{\nu s}^2 = \underline{\nu}\,\underline{s}^2 + \nu s^2 + (b - \underline{\beta})'\left[\underline{V} + \left(X'X\right)^{-1}\right]^{-1}(b - \underline{\beta}).$$

The marginal likelihood is available in closed form:

$$p(y \mid M_m) = c_m \left(\frac{|\bar{V}_m|}{|\underline{V}_m|} \right)^{1/2} \left(\bar{\nu}_m \bar{s}_m^2 \right)^{-\bar{\nu}_m/2}.$$

$$c_m = \frac{\Gamma\left(\frac{\bar{\nu}_m}{2} \right) \left(\underline{\nu}_m \underline{s}_m^2 \right)^{\nu_m/2}}{\Gamma\left(\frac{\underline{\nu}_m}{2} \right) \pi^{n/2}}.$$

It is reasonable to set $\underline{\nu}_1 = \underline{\nu}_2 = 0$ so that we are non-informative on σ^2. In this case we have $c_1 = c_2$. Additionally, we may set

$$\underline{V}_m^{-1} = C_m I_{k_m}, m = M_1, M_2,$$

in which case we have

$$\left| \underline{V}_m \right| = C^{-k_m}.$$

The Bayes factor is:

$$BF_{1:2} = \frac{p(y \mid M_1)}{p(y \mid M_2)}.$$

Although these expressions are somewhat complicated they can be easily programmed in standard software. Regarding prior selection, one would, perhaps, like to set C_m to a small value (we use 0.1). By assumption, we have: $\underline{\nu}_m = 0.$ Moreover, we also set $\underline{\beta}_m = 0$.

From simple inspection of the marginal likelihood, it is clear that it accounts for model fit, parsimony (in terms of number of parameters) and "agreement" between the prior and the posterior.

B. Other Bayes factors

Besides using the Natural Conjugate priors, there are other alternatives and, therefore, different Bayes factors that we can consider. In the Natural Conjugate case, one has to specify different parameters. A prominent alternative is Zellner's (1996) g-prior which requires specifying a single parameter, g.

We continue to assume $\underline{\beta} = 0$. The prior covariance matrix is now specified as:

$$\underline{V} = h^{-1} \left(gX'X \right)^{-1}.$$

In this expression, $\left(X'X \right)^{-1}$ is used because it appears in the usual LS analysis of the linear model. The marginal likelihood of the model is:

$$p(Y \mid g) = \left(\frac{g}{g+1} \right)^{k/2} \left\{ \frac{1}{g+1} y'Py + \frac{g}{g+1} (y - \bar{y}\iota_n)'y - \bar{y}\iota_n \right\}^{-(n-1)/2},$$

where $P = I_n - X\left(X'X\right)^{-1}X'$, $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ and $\iota_n = [1,...,1]'$ is a vector of ones.

The value of g can be chosen so that is relatively low, for example between 0 and 1. It is also known that setting $g = \frac{1}{\left(\ln n\right)^3}$ then the log Bayes factor mimics the Hannan-Quinn criterion. To reduce arbitrariness, one can specify a prior for g, say $p(g)$, compute $p(Y \mid g)p(g)$ and select the value that maximizes this quantity, given the data.

C. Intrinsic Bayes factors

The general idea behind Intrinsic Bayes Factors (IBS) is due to Berger and Pericchi (1996). A portion of the data set (as training set) is set aside and posterior analysis is performed using a non—informative prior. In turn, the posterior is used as a prior in the remaining data set. As they mention the IBF: "*is fully automatic in the sense of requiring only standard noninformative priors for its computation and yet seems to correspond to very reasonable actual Bayes factors. The criterion can be used for nested or nonnested models and for multiple model comparison and prediction. From another perspective, the development suggests a general definition of a "reference prior" for model comparison*" (Berger and Pericchi, 1996, p.109).

The size of the training sample is an important choice. One solution is to use a minimal training sample whose size is $k+1$. Unfortunately, Berger and Pericchi (1996) do not consider a linear regression model. This is taken up in Berger and Pericchi (1994). We consider the normal linear regression model:

$$M_m : y = X_m\beta_m + u_m, u_m \sim iidN(0, \sigma_m^2).$$

The prior is of the form:

$$p\left(\beta_m, \sigma_m\right) \propto \sigma_m^{-(q_m+1)}, q_m > -1.$$

Common choices are $q_m = 0$ or $q_m = k_m$ (the Jerffreys prior). Suppose that the training sample has length T and the corresponding data is $\left\{y_m^{(T)}, X_m^{(T)}\right\}$. Suppose we wish to compare models i and j. Berger and Pericchi (1994) show that the IBF is given by:

$$IBF_{i:j} = C_{ij}\left(\frac{\left| X_j^{(T)\prime} X_j^{(T)} \right|}{\left| X_i^{(T)\prime} X_i^{(T)} \right|} \right)^{1/2} \frac{RSS_j^{(T-k_j+q_j)/2}}{RSS_i^{(T-k_i+q_i)/2}},$$

where

$$C_{ij} = \frac{\pi^{(k_i-k_j)/2}}{2^{(q_j-q_i)/2}} \frac{\Gamma\left(\frac{T-k_i+q_i}{2} \right)}{\Gamma\left(\frac{T-k_j+q_j}{2} \right)}.$$

Here $RSS_i = \left(y_i^{(T)} - X_i^{(T)\prime} b\right)' \left(y_i^{(T)} - X_i^{(T)\prime} b\right)$ is the residual sum of squares.

With time series data it is reasonable to consider the first T observations. With cross-sectional data there is some ambiguity so Berger and Pericchi (1994) propose to use the arithmetic mean. Suppose we consider various training samples $\left\{y^{(m)}, X^{(m)}, m = 1,...,L\right\}$ consisting of T observations. Then, using a Jeffreys prior, we have:

$$IBF_{i:j}^{A} = \left(\frac{\left|X_j' X_j\right|}{\left|X_i' X_i\right|}\right)^{1/2} \left(\frac{RSS_j}{RSS_i}\right)^{n/2} L^{-1} \sum_{m=1}^{L} \left(\frac{\left|X_i^{(m)\prime} X_i^{(m)}\right|}{X_j^{(m)\prime} X_j^{(m)}}\right)^{1/2} \left(\frac{RSS_i^{(m)}}{RSS_j^{(m)}}\right)^{T/2}.$$

Instead of the arithmetic mean, one can take the geometric mean as well.

D. Nonlinear models

For nonlinear models with a general likelihood $L(\theta;Y)$ and a prior $p(\theta)$ the marginal likelihood is:

$$M(Y) = \int_{\Theta} L(\theta;Y)\, p(\theta) d\theta,$$

and the posterior is: $p(\theta \mid Y) \propto L(\theta;Y)\, p(\theta)$. We assume that the dimensionality of the parameter vector is k. Computation of the multivariate integral is, in general, impossible using analytical techniques. Therefore we have to resort to Markov Chain Monte Carlo (MCMC) methods. These methods produce a sample $\left\{\theta^{(s)}, s = 1,...,S\right\}$ which converges to the distribution whose density is $p(\theta \mid Y)$. Since

$$p(\theta \mid Y) = \frac{L(\theta;Y)\, p(\theta)}{\int_{\Theta} L(\theta;Y)\, p(\theta) d\theta} = \frac{L(\theta;Y)\, p(\theta)}{M(Y)}.$$

Therefore, we have the following remarkable identity:

$$M(Y) = \frac{L(\theta;Y)\, p(\theta)}{p(\theta \mid Y)}, \forall\, \theta \in \Theta.$$

Since this is an identity in θ we have:

$$M(Y) = \frac{L(\bar{\theta};Y)\, p(\bar{\theta})}{p(\bar{\theta} \mid Y)},$$

where $\bar{\theta}$ is, for example, the posterior mean. The numerator can be computed easily. Following DiCiccio et al. (1997). The denominator is unknown but we can use a normal approximation to obtain:

$$p(\bar{\theta} \mid Y) \approx (2\pi)^{-k/2} |V|^{-1/2},$$

where[13] $V = S^{-1}\sum_{s=1}^{S}\left(\theta^{(s)} - \bar{\theta}\right)\left(\theta^{(s)} - \bar{\theta}\right)'$ is an estimate of the posterior covariance matrix of θ. In turn, the log marginal likelihood can be approximated as follows:

$$\log M\left(Y\right) \approx \ln L(\bar{\theta};Y) + \ln p\left(\bar{\theta}\right) + \tfrac{k}{2}\ln(2\pi) + \tfrac{1}{2}\ln\left|V\right|.$$

The approximation is very easy to use in practice.

Appendix B

Gauss Code

```
/ this program computes log Bayes factors for testing that each
// regression coefficient is zero in a linear regression/
// We use a Verdinelli-Wassernan (1994) approach/
// The prior is:
// beta~N(beta_underbar, h_underbar*I),
// q_underbar/sigmasq ~chisquare(nu_underbar).
// There are default values for beta_underbar, h_underbar and
// nu_underbar
//

new; cls; library pgraph;
rndseed 11;

// define or load the data
n = 20;                    // number of observations
k = 5;                     // number of regressors
X = ones(n,1)~rndN(n,k-1);
y = X*ones(k,1)+rndN(n,1);

h_underbar    = 1e4;
beta_underbar = zeros(k,1);    // prior info and Gibbs specifications
q_underbar    = 1e-4;
nu_underbar   = 0;
npass         = 6000;          // Gibbs sampling passes
nburn         = 1000;          // length of Gibbs sampling burn-in
```

[13] The remaining term $\exp\left\{-\tfrac{1}{2}(\theta-\bar{\theta})V^{-1}(\theta-\bar{\theta})\right\}$ vanishes at $\theta = \bar{\theta}$.

```
// ************* USER INPUT ENDS HERE ***************
Draws           = zeros(npass,k+1);
Log_BayesFactors = zeros(k,1);
V_underbar      = h_underbar*eye(k);
V_underbar_inv  = (1/h_underbar)*eye(k);


XtX     = X'X;
Xty     = X'y;
beta    = invpd(XtX)*Xty;
u       = y-X*beta;
sigmasq  = u'u/(n-k);


// Gibbs sampler
ISIM=1; do while ISIM<=npass;
sigma = sqrt(sigmasq);
A=invpd(XtX+sigmasq*V_underbar_inv);
b = A*(Xty+sigmasq*V_underbar_inv*beta_underbar);
beta = b+sigma*chol(A)'rndN(k,1);
sigmasq = (u'u+q_underbar)/rndchisq(1, n-k+nu_underbar);
Draws[ISIM,.] = beta'~sigma;
ISIM=ISIM+1; endo;

Draws = trimr(Draws,nburn,0);
beta  = Draws[.,1:k];
sigma = Draws[.,k+1];


_output = 0;
// Bayes factors
i=1; do while i<=k;
  screen OFF; {x1,f1} = dens(beta[.,i]); screen ON;
   // compute log posterior at zero by interpolation
   lf0 = polyint(x1,ln(1e-12+f1),0);
   // compute log prior at zero
   lpri =
   -0.5*ln(2*pi*h_underbar)-(0.5/h_underbar)*beta_underbar[i]^2;
   // compute log Bayes factor
   lbf = lf0-lpri;
   Log_BayesFactors[i] = lbf;
i=i+1; endo;


call OLS("", y, X);

"";
"posterior statistics"; "";
"regression parameters";
```

```
"";"post. mean post. s.d. 95% Bayes prob. interval log Bayes factor Bayes factor";
meanc(beta)~stdc(beta)~Bayes_Interval(beta)~Log_BayesFactors~exp(Log_BayesFactors);
""; "sigma ";
meanc(sigma)~stdc(sigma)~Bayes_Interval(sigma);


proc rndchisq(n, nu);
/*
draws n random numbers from a chi-square with nu degrees of freedom
*/
retp(2*rndgam(n, 1, nu/2));
endp;


proc Bayes_Interval(x);
local n,k,BI,i,x1,a,b;
n=rows(x); k=cols(x);
BI = zeros(k,2);
i=1; do while i<=k;
  x1 = sortc(x[.,i],1);
  a = x1[0.025*n];
  b  = x1[0.975*n];
  BI[i,.] = a~b;
i=i+1; endo;
retp(BI); endp;
```

Appendix C

The Bayesian Higher Posterior Density (HPD)

FIGURE 12. HPD vs. Confidence Intervals

Suppose a parameter θ has the sampling density shown in the Figure above. Suppose also we identify correctly the global maximum through ML estimation but we resort to an asymptotic normal approximation[14] around the global maximum. Is the p-value OK in this context?

Of course not, because we miss the second mode. Therefore, what is needed, really, in the sampling-theory context is a *small-sample investigation of the problem at hand*. However, most of the time, if not always, the researcher uses asymptotic normal approximations. If the density above is also the marginal posterior of θ—or something fairly close to it, plotting the posterior cannot miss the important information conveyed from the second mode.

Also in this instance, the 95% sampling-theory confidence interval and the 95% highest posterior density interval, are not the same. *The HPD consists of two intervals which excludes the zero probability area between the two modes*. The 95% sampling-theory confidence interval includes this area which is, apparently, wrong.

For the Bayesian, the testing problem is really simple. Suppose we need to evaluate $H:\theta=\theta_o$. All the Bayesian has to do is examine the relation between θ_o and the density. Does it belong to the extreme tails or is it relatively likely? In essence, *Bayesian do not test hypotheses; they evaluate hypotheses*. This is precisely the purpose of a Bayes factor, viz. to compare the probability of a given model relative to another.

[14] We are justified to do this, from the sampling-theory perspective, as all inferences rely on the global optimum, which is defined as the ML estimate.