Frequentist Comparison of the Bayesian Significance Test for Testing the Median of the Lognormal Distribution

1 Juliet Gratia D’Cunha

Research Scholar, Department of Statistics, Mangalore University, Mangalagangothri-574199

Karnataka, India

Email: gratiajuliet@gmail.com

2 K. Aruna Rao

Professor (Retd.), Department of Statistics, Mangalore University, Mangalagangothri-574199

Karnataka, India

Email: arunaraomu@gmail.com
Frequentist Comparison of the Bayesian Significance Test for Testing the Median of the Lognormal Distribution

Abstract: This paper focuses on the three Bayesian significance tests namely Bayes factor test, credible region test, full Bayesian significance test (FBST) and the reintroduced Bayesian/Non-Bayesian Likelihood Evidence test, with reference to the problem of testing specified value of the median of the two parameter lognormal distribution. These tests are compared in terms of type I error rate and power. The simulation results indicate that credible interval does not maintain level of significance for small samples. There is no difference in the power of the tests for the alternatives to the left and right of the null hypothesis. The other important conclusion that emerges is that these tests are robust against the specification of the prior distribution. Bayes factor test has the advantage of computational simplicity compared to the other three tests. The critical values for the Bayes factor and the evidence measure would be helpful for the scientists for taking a decision. In absence of critical values, guideline has been suggested for the use of FBST.

Key Words: Bayesian significance test, Bayes factor, credible region, full Bayesian significance test, likelihood evidence test, power of the test.

1. Introduction

Testing a sharp (precise) null hypothesis is a controversial and well debated topic in Bayesian inference. For a definition of sharp/precise null hypothesis see Pereira and Stern (1999). Acknowledging the fact that precise null hypothesis naturally occurs in scientific investigations, the interest boils down to the choice of an appropriate Bayesian test of significance. The commonly used Bayesian approaches are a) Bayes factor b) credible interval and the associated tests c) probability of null hypothesis under the posterior distribution $P(H_0|x)$ and d) Full Bayesian significance test (FBST). Standard text books on Bayesian inference like Berger (1985) and Ghosh et al. (2006) provide details regarding the former three procedures. FBST was introduced by Pereira and Stern (1999). The other salient papers on FBST are Pereira and Stern (2001a, b), Madruga et al. (2003) and Pereira et al. (2008).
In addition to these four approaches, following the idea of strong law of likelihood of Basu (1975), we also provide a new evidence measure namely likelihood evidence. This evidence measure is also related to the concept of fiducial inference introduced by Sir Ronald Fisher. Section 2 provides a critical comparison of these four procedures. Bayes factor and FBST provide evidence in support (or against) the null hypothesis given by the data. Unlike the well trained Bayesians, scientists find it difficult to take a decision based on these evidence measures. This has motivated us to provide critical values for these evidence measures.

Another difficulty faced by the researchers is the choice of appropriate Bayesian tests of significance. In this paper an attempt is made to compare the frequentist power of the Bayesian tests namely Bayes factor, credible interval and its associated tests and FBST for testing the median of the lognormal distribution. In Bayesian significance tests such a Frequentist approach is not adopted, which is part of Neyman Pearson school of inference. There are situations when pre-experimental information exists which cannot be used by a Frequentist but can be readily incorporated as a prior in the Bayesian approach. In these situations a Frequentist is justified in using the Bayesian tests which have got good power properties. As noted in Basu (1975), Neyman Pearson approach is valid in scientific investigations. Further he observes that Bayesians are not averse to the idea of precision of their measures and in tests of significance, from the frequentist view point, such a measure is the power of the test.

Frequentist comparison of Bayesian procedure is not new and has a long history. Bartholomew (1965) compared Bayesian and Frequentist approaches for the parameter $\theta$ of a continuous distribution. He discussed some examples for which it is not possible to find Bayesian methods possessing the properties required by Frequentist. Woodrofe (1976) examined the power properties of the approximate Bayes procedure with Sequential Probability Ratio test (SPRT) and fixed size test for testing the one sided hypothesis, that the mean of the normal distribution is less than or equal to a specified value. Hulting and Harville (1991) developed Bayesian procedures for estimating the small area mean. Using a Monte Carlo study they conclude for certain choices of the prior distribution, the Frequentist properties of the Bayesian prediction procedure compare favourably with those of their classical counterparts. Severini (1993) discussed the problem of constructing a credible interval for the real parameter of a distribution. He constructs procedures which have the same Frequentist coverage probability as the
one specified by the posterior distribution. Sweeting (2001) compares the coverage probability of Bayes procedure using objective priors with the confidence interval obtained using the likelihood principle. Genovese and Wasserman (2002) developed Bayesian approach for multiple testing. They showed that in contrast to the single hypothesis case where Bayes and Frequentist tests do not agree even asymptotically, in the multiple testing case, we do have asymptotic agreement. Stern and Zacks (2002) developed FBST for testing the hypothesis of independence in a Holgate Bivariate Poisson distribution. They compared the power of this test with several well known classical test (Non-Bayesian) and using Monte Carlo simulation showed that FBST has better power properties compared to the other tests. Agresti and Min (2005) studied the performance in a Frequentist sense, of Bayesian confidence intervals for the difference of proportions, relative risk, and odds ratio in $2 \times 2$ contingency tables. They recommended the use of tail credible intervals compared to highest posterior density (HPD) intervals. Moon and Schorfheide (2012) compared Bayesian and Frequentist inference in partially identified models. They recommended the use of Bayesian credible sets as it contains the identified set.

For the power comparison of the three Bayesian tests of significance, we have restricted our attention to the lognormal distribution. The choice of lognormal distribution stems from the fact that it is widely used in applied research, see (D'Cunha and Rao (2014a,b); Harvey et al. (2011); Harvey and Merwe (2010, 2012); Krishnamoorthy and Mathew (2003); Limpert et al. (2001); Zhou and Tu (2000); Zellner (1971); Aitchison and Brown (1957)). This paper also examines how far the Bayesian tests are robust for the specification of the prior density. Extensive simulation is used to seek answers to these queries. The results show that Bayes tests are robust for the specification of the prior density. The power comparison indicates that the test based on Bayes Factor is more powerful compared to FBST and credible region approach and the classical t-test.

The paper is divided in seven sections. A critical comparison of the Bayesian tests of significance is given in section 2. This section also introduces the likelihood evidence. Section 3 provides the details of Bayesian tests of significance for the median of the lognormal distribution. Section 4 gives details of the simulation experiment. The power comparison of the tests is presented in section 5. Analysis of stock prices for the script TATA STEEL listed in the National Stock Exchange (NSE) limited, India for the
period May and June 2015 is analysed in section 6. Concluding remarks and suggestions for the choice of the test are given in section 7.

2. Bayesian significance tests

2.1 Bayes factor

Let \( \theta \) be the real parameter of interest and \( \lambda \) be the nuisance parameter. \( \lambda \) may be real or vector valued. The hypothesis of interest is \( H_0: \theta = \theta_0 \) against the alternative \( H_1: \theta \neq \theta_0 \). In this approach a positive probability \( p_0 \) is assigned to the null hypothesis. Let \( \pi_0(\lambda) \) be the prior density of the nuisance parameter \( \lambda \), under the null hypothesis and let \( \pi_1(\lambda) \) and \( \pi_1(\theta|\lambda) \) denote the prior densities for \( \lambda \) and \( \theta \) respectively, under the alternative hypothesis.

The Bayes factor \( BF_{01} \) is the ratio of the posterior odds to the prior odds in favour of \( H_0 \). When \( p_0 = \frac{1}{2} \), it is given by,

\[
BF_{01} = \frac{\int_{\lambda} f(x|\theta_0, \lambda)\pi_0(\lambda)d\lambda}{\int_{\theta} \int_{\lambda} f(x|\theta, \lambda)\pi_1(\theta|\lambda)\pi_1(\lambda)d\theta d\lambda}
\]

(1)

Throughout the paper \( \pi(\cdot) \) is used to denote the prior density and \( \pi(\cdot|x) \) is used to denote posterior density given the data and \( \pi \) is a generic notation. Bayes factor has the following features:

a) Positive prior probability is specified for the null hypothesis.

b) The prior density need not be the same for the null and alternative hypothesis.

Bayes factor is related to another concept namely posterior probability of \( H_0 \) given the data. Thus \( P(H_0|data) = \frac{BF_{01}}{1+BF_{01}} \).

Decision can be taken regarding the null hypothesis using the Bayes factor \( BF_{10} \) which is the reciprocal of \( BF_{01} \). A large value of \( BF_{10} \) leads to the rejection of the null hypothesis. Bounds for the Bayes factor can be obtained with reference to a specific problem (See Ghosh et al. (2006) page164 to 165 in chapter 6). The Bayes factor \( BF_{10} \) has a resemblance to the Neyman Pearson approach. Without having a critical value it is difficult to take a decision. Nevertheless, guideline has been suggested by Jeffrey (1961) and Kass and Raftery (1995) for taking decision.

2.2 Test based on credible interval
In Frequentist approach there is a one to one relation between the acceptance region of a level \( \alpha \) test and the \((1 - \alpha)\) level confidence interval (see Ghosh et al. (2006); Guddattu and Rao (2010)). This reasoning is used to develop a Bayesian significance test. The null hypothesis is rejected if a specified value \( \theta_0 \) falls outside the \((1 - \alpha)\) credible interval. The credible interval can be either an equitailed or highest posterior density (HPD) credible interval. Credible intervals are based on the critical values of posterior distribution \( \pi(\theta|x) \) and in most cases have to be obtained numerically when closed form expression for the posterior density does not exist. The features of this approach are the following:

a) No positive prior probability is assigned to the null hypothesis.
b) Single prior \( \pi(\theta) \) covering both null and alternative hypothesis is considered.
c) As in the Frequentist approach if the level of the credible interval is specified, decision can be taken easily.

2.3 Full Bayesian significance test

This procedure was introduced by Pereira and Stern (1999). In the present notation let

\[
\gamma = P((\theta, \lambda) : \pi(\theta, \lambda|x) > \sup_{\lambda} \pi(\theta_0, \lambda|x))
\]

\[
= \int_{\pi(\theta, \lambda|x) > \sup_{\lambda} \pi(\theta_0, \lambda|x)} \pi(\theta, \lambda|x) \, d\theta \, d\lambda
\]

(2)

The evidence measure \( Ev \) is given by \( Ev = 1 - \gamma \). This is the evidence against the null hypothesis given the data. As in the case of \( p \)-value we reject the null hypothesis for small values of \( Ev \). The following observations can be made regarding this test.

a) This test was developed so as to avoid ambiguities in the \( p \)-value. The tangential set accounts for the null as well as alternative parameter space and thus is based not only on the null hypothesis but also on the alternative hypothesis.

b) Madruga et al. (2003) developed an extended version of the FBST which possess invariance properties under coordinate transformations. They also prove that \( \gamma \) is a Bayes rule under certain loss function. They also derive an upper bound for \( \gamma \) under this loss function. Scientists may find it difficult to specify the loss function. Thus this critical value is difficult to use in practice. Guideline is required regarding a suitable critical value for taking decision.
d) Pereira et al. (2008) prove that FBST is a significant test. For a discussion on significant test refer Cox (1977). They argue that since it is a significant test, there is no need to specify a positive probability for the null hypothesis.

e) FBST shares some common features with likelihood ratio test. The hard line Bayesians persist that positive prior probability had to be assigned for the null hypothesis. Their argument is that if this probability is zero, the posterior probability of $H_0$ will also be equal to zero (See the discussion on Bayesian significance test in Ghosh et al. (2006), page 41-49). Despite of this, the popularity of FBST is increasing as is evident in the papers of Stern and Zacks (2002), Lauretto et al. (2003) and Andrade et al. (2014). Motivated from these papers and the classical paper of Basu (1975) we propose the following measure of evidence, which we call likelihood evidence.

2.4 Likelihood evidence ($L_{ev}$)

Let $L(\theta, \lambda; x)$ denote the likelihood given the data $x = x_1, ..., x_n$. Basu (1975) vehemently argued that likelihood $L(\theta, \lambda; x)$ is a set function, and is a measure and can be treated as probability. Further he proves that, when $\theta$ is real and the parameter space $\Theta$ is an interval, the likelihood obeys the additive law of probability. He disagrees with the argument of Fisher and others that likelihood is not a probability but only a relative frequency. He also proposes the strong law of likelihood. The philosophical issues involved in the strong law of likelihood are beyond the scope of this paper. This has motivated us to propose the likelihood evidence in the sequel.

Let $\gamma$ be a generic notation. We define $\gamma = P((\theta, \lambda): L(\theta, \lambda; x) > \sup_\lambda L(\theta_0, \lambda; x)). 1 - \gamma$ is the likelihood evidence against the null hypothesis. The smaller value of $L_{ev}$ indicates that the data does not support the null hypothesis but on the other hand supports the alternative hypothesis. Of course we are aware that this is a particular case of FBST when uniform prior is used. Thus Bayesians are likely to dismiss this new measure. But it should be clearly noted that the philosophical reasoning is different. Given the data the parameter is treated as random, and the normalized likelihood is viewed as a density function of $\theta$. Of course, it is important that the likelihood to be well behaved. The same criticism applies to the posterior density of $\pi(\theta, \lambda|x)$ in FBST. The posterior density may not be a proper density function, if improper prior distribution is used in the investigation. The advantage of $L_{ev}$ is that it does not require the specification of the prior density and thus
is appealing to non-Bayesians (See the discussion of Barnard in Basu’s (1975) paper). Although the choice of a suitable prior is well discussed in Bayesian literature (See Berger (1985); Ghosh et al. (2006) and the references cited therein), the applied researchers and data analyst, face the problem of choosing a suitable prior density. Further, in the application of FBST in the papers cited previously, uniform prior is used in the analysis of data which is then equal to the Likelihood evidence, $L_{ev}$ (see Stern and Zacks (2002); Agresti and Min (2005)). $L_{ev}$ is more closely related to the fiducial argument of Sir Ronald Fisher (see the remark of Edwards (2000) while reviewing the book “Statistical Evidence: A Likelihood Paradigm” of Royall (1997) and Edwards (1992)). Unless a critical value is specified, as in FBST, it may be difficult to take a suitable decision.

3. Bayes test for the median of the lognormal distribution

Bayesian inference of the parameters of lognormal distribution is widely discussed in the literature starting from Zellner (1971), Harvey and Merwe (2010, 2012), Harvey et al. (2011) and D’Cunha and Rao (2014 a, b, 2015). For comparing the power of the Bayes tests, the problem chosen is testing for the specified value of the median of the two parameter lognormal distribution with parameters $\mu$ and $\sigma^2$. The median of the lognormal distribution is $e^\mu$. If we make log transformation of the random variable under consideration, the transformed distribution is normal with location parameter $\mu$ which is the median of the normal distribution and is equal to $\ln e^\mu$. Thus median is invariant under log transformation of the random variable. Thus the null hypothesis $H_0: \theta = \theta_0$ of the lognormal distribution, where $\theta = e^\mu$ is equivalent to testing $H_0: \mu = \mu_0$ of the normal distribution, where $\mu_0 = \ln \theta_0$. The problem is chosen, as Uniformly Most Powerful Unbiased (UMPU) test exists for the latter hypothesis and it would be meaningful to compare the power of the Bayes tests with this test. Given a random sample $X = X_1, …, X_n$ from lognormal distribution, denoting log $X_i$ as $Z_i$, the minimal sufficient statistic for $\mu$ and $\sigma^2$ are $\bar{Z} = \frac{\sum_{i=1}^n Z_i}{n}$ and $S_Z^2 = \frac{\sum_{i=1}^n (Z_i - \bar{Z})^2}{n-1}$. Hence the likelihood for $\mu$ and $\sigma^2$ is given by
\[
L(\mu, \sigma^2; z) = \frac{1}{\sqrt{2\pi} \sigma^2} e^{-\frac{1}{2} \left( \frac{(z-\mu)^2}{\sigma^2} - \frac{(n-1)S^2}{2} \right)} \cdot \frac{1}{\Gamma\left(\frac{n+1}{2}\right)} e^{-\frac{1}{2} \left( \frac{n+1}{2} \right) \sigma^2} 
\]

\[-\infty < \mu < \infty, \sigma^2 > 0 \quad (3)\]

Four prior distributions are considered for the comparison. They are: i) Right invariant prior \( \pi_1(\mu, \sigma) = \frac{1}{\sigma} \) ii) Left invariant Jeffrey’s prior \( \pi_2(\mu, \sigma) = \frac{1}{\sigma^2} \) iii) Jeffrey’s rule prior \( \pi_3(\mu, \sigma) = \frac{1}{\sigma^3} \) and iv) Uniform prior \( \pi_4(\mu, \sigma) = 1 \). Jeffrey’s rule prior \( \pi_3(\mu, \sigma) \propto |I(\mu, \sigma^2)|^{1/2} \), where \( I(\mu, \sigma^2) \) is the Fisher information matrix corresponding to \((\mu, \sigma^2)\) for a single observation and is given by \( I(\mu, \sigma^2) = diag\left[ \frac{1}{\sigma^2} \frac{1}{\sigma^4} \right] \). For details of this prior refer to Harvey and Merwe (2012). Although Basu (1975) criticizes the use of improper priors, they have been widely used in objective Bayesian analysis. These priors were also used in the papers cited in the beginning of this section. The posterior distribution is the product of gamma for \( \eta = \frac{1}{\sigma^2} \) and the conditional normal distribution for \( \mu \). To save space we give below the posterior distribution of \( \mu \) and \( \sigma^2 \) for the right invariant prior.

\[
\pi(\mu, \sigma | z) = \frac{1}{\sqrt{2\pi} \sigma^2} e^{-\frac{1}{2} \left( \frac{(z-\mu)^2}{\sigma^2} - \frac{(n-1)S^2}{2} \right)} \cdot \frac{1}{\Gamma\left(\frac{n+2}{2}\right)} e^{-\frac{1}{2} \left( \frac{n+2}{2} \right) \sigma^2} \quad (using \ right \ invariant \ prior) \]

\[-\infty < \mu < \infty, \sigma^2 > 0 \quad (4)\]

The difference with respect to various priors is only in the shape parameter of the gamma distribution for \( \eta = \frac{1}{\sigma^2} \) and is symbolically given below,

\[
\pi(\mu, \sigma | z) = N(\mu; Z, \sigma^2 / n) \times G_\eta\left(\frac{n+3}{2}, \frac{n-1}{2} \right) S^2 Z \quad (using \ left \ invariant \ prior) \]

\[-\infty < \mu < \infty, \sigma^2 > 0 \quad (5)\]

\[
\pi(\mu, \sigma | z) = N(\mu; Z, \sigma^2 / n) \times G_\eta\left(\frac{n+4}{2}, \frac{n-1}{2} \right) S^2 Z \quad (using \ Jeffrey's \ rule \ prior) \]

\[-\infty < \mu < \infty, \sigma^2 > 0 \quad (6)\]

\[
\pi(\mu, \sigma | z) = N(\mu; Z, \sigma^2 / n) \times G_\eta\left(\frac{n+1}{2}, \frac{n-1}{2} \right) S^2 Z \quad (using \ Uniform \ prior) \]

\[-\infty < \mu < \infty, \sigma^2 > 0 \quad (7)\]
where \( z = z_1, z_2, ..., z_n \). The Bayes factor \( BF_{10} \) for the four posterior distributions after simplification is given by

\[
BF_{10} = \frac{\sqrt{\pi} \Gamma\left(\frac{n+3}{2}\right)}{\sqrt{2\pi} \Gamma\left(\frac{n+2}{2}\right)} \left(\frac{1}{2(n-1)\sigma^2}\right)^{\frac{n+2}{2}} (using \ right \ invariant \ prior)
\]

(8)

\[
BF_{10} = \frac{\sqrt{\pi} \Gamma\left(\frac{n+4}{2}\right)}{\sqrt{2\pi} \Gamma\left(\frac{n+2}{2}\right)} \left(\frac{1}{2(n-1)\sigma^2}\right)^{\frac{n+3}{2}} (using \ left \ invariant \ prior)
\]

(9)

\[
BF_{10} = \frac{\sqrt{\pi} \Gamma\left(\frac{n+5}{2}\right)}{\sqrt{2\pi} \Gamma\left(\frac{n+4}{2}\right)} \left(\frac{1}{2(n-1)\sigma^2}\right)^{\frac{n+4}{2}} (using \ Jeffreys' \ rule \ prior)
\]

(10)

\[
BF_{10} = \frac{\sqrt{\pi} \Gamma\left(\frac{n+2}{2}\right)}{\sqrt{2\pi} \Gamma\left(\frac{n+1}{2}\right)} \left(\frac{1}{2(n-1)\sigma^2}\right)^{\frac{n+1}{2}} (using \ Uniform \ prior)
\]

(11)

Jeffrey (1961) suggested that the same prior \( \pi(\sigma) \) be used in the numerator and denominator of \( BF_{01} \) in (1), while an informative prior be used for \( \pi(\mu|\sigma) \) in the denominator. Use of informative prior does not permit a fair comparison of the various tests and in the present paper we have used independent prior for \( \mu \) and \( \sigma^2 \). Non informative priors are defined upto a multiplicative constant and if we use \( \pi(\mu) = c \), the Bayes factor and the frequentist critical values changes by a factor \( \frac{1}{c} \), while the power of the test remains unaltered. Thus Lindley’s paradox does not affect the power comparison of the Bayes factor test.

To overcome the difficulty of computing the evidence measure Pereira et al. (2008) have obtained a first order approximation for \( \gamma \). Cabras et al. (2015) have improved the approximation by obtaining the third order expression for \( \gamma \). In this paper we have derived third order approximation for \( \gamma \) in the context of lognormal distribution for testing the median. This expression is directly obtained without using Laplace transformation and requires mild regularity conditions.
Theorem 1: The third order approximation for \( \gamma \) for testing \( H_0: \mu = \mu_0 \) of the lognormal distribution is given by
\[ \gamma = \int_{\eta} \left\{ \Phi \left( \frac{g_1(\sigma^2)}{\sigma} \right) + 1 - \Phi \left( \frac{g_2(\sigma^2)}{\sigma} \right) \right\} G_\eta(.) d\eta + O(n^{-3/2}), \]
where \( \Phi(.) \) denotes the cumulative distribution function (CDF) of the standard normal distribution and \( G_\eta(.) \) denotes the gamma density for \( \eta = \frac{1}{\sigma^2} \) with scale parameter \( \left( \frac{n-1}{2} \right) S_Z^2 \), and shape parameter \( v = n + 1 \) (Uniform), \( v = n + 2 \) (Right invariant), \( v = n + 3 \) (Left invariant) and \( v = n + 4 \) (Jeffreys’ rule) priors.

Proof: Notation and proof of the theorem is sketched in the appendix.

4. Simulation experiment

Extensive simulation is carried out to estimate the type 1 error rate and power of three of the Bayesian tests namely i) Bayes factor ii) Credible interval (equitailed) and iii) FBST. A sample of size \( n \) was initially generated from Normal distribution with mean \( \mu \) and variance \( \sigma^2 \). Since the posterior density is the product of independent gamma distribution for \( \eta = \frac{1}{\sigma^2} \) and the conditional normal distribution for \( \mu \) given \( \sigma^2 \), a pair of observation \((\mu, \sigma)\) is generated by first generating the observation \( \sigma^2 \) from the gamma distribution and then generating an observation from the normal distribution using the value of \( \sigma^2 \) for each selected sample. Equitailed credible intervals are constructed using importance sampling approach. The complement of the coverage probability under the null hypothesis yielded an estimate of the type 1 error rate and power of the test under the alternative hypothesis which is estimated using 1000 simulations. The 5\(\%\) critical value for the Bayes factor is estimated using 1000 simulations. The critical value corresponds to the upper \( \alpha^{th} \) percentile value of the simulated distribution of \( B_{F10} \). For the estimation of the critical value of \( Ev = 1 - \gamma \), the simulation had to be carried out in two stages. For each sample, \( 1 - \gamma \) is estimated using 10,000 simulations (Monte Carlo integration) and the critical value is estimated using 1000 simulations. The critical value refers to lower \( \alpha^{th} \) percentile value of the simulated distribution of \( 1 - \gamma \). For this purpose program is written using Matlab software version 7.0. Power of the Bayes Factor test and FBST is estimated using these critical values. The coefficient of variation (CV) of lognormal distribution is given by
\[ CV = \left( e^{\sigma^2} - 1 \right)^{1/2} \]
and the values of CV used in the simulations are 0.1, 0.3, 0.5, 0.7, 1, 1.5, 2 and 2.5. Sample sizes considered are \( n = 10, 20, 40, 60, 80, 100, 150 \) and 200.
The nominal level is fixed at 0.05. Thus the specified value of \( e^\mu \) in \( H_0 \) is chosen as 1000. This value of \( e^\mu \) corresponds to the analysis of stock prices reported in D’Cunha and Rao (2015).

5. Results and discussion

5.1 Null behaviour of the tests

The estimated value of the upper 5\(^{th}\) percentile for the Bayes factor \( BF_{10} \) for selected sample sizes is presented in table 5.1, although it is computed for all sample sizes considered in the simulation.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Prior</th>
<th>Critical Values for Bayes Factor when CV equal to</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Right Invariant</td>
<td>2.0219</td>
<td>2.0491</td>
<td>2.0980</td>
<td>2.1865</td>
<td>2.3221</td>
<td>2.5179</td>
<td>2.6354</td>
<td>2.736</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left Invariant</td>
<td>1.9766</td>
<td>2.0040</td>
<td>2.0538</td>
<td>2.1432</td>
<td>2.2808</td>
<td>2.4823</td>
<td>2.5993</td>
<td>2.7597</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeffrey’s Rule</td>
<td>1.9342</td>
<td>1.9619</td>
<td>2.0125</td>
<td>2.1031</td>
<td>2.2423</td>
<td>2.4489</td>
<td>2.5692</td>
<td>2.7288</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>2.0704</td>
<td>2.0975</td>
<td>2.1455</td>
<td>2.2331</td>
<td>2.3668</td>
<td>2.5568</td>
<td>2.6756</td>
<td>2.8311</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Right Invariant</td>
<td>0.3016</td>
<td>0.6397</td>
<td>1.3244</td>
<td>1.7065</td>
<td>2.1008</td>
<td>3.3078</td>
<td>2.8540</td>
<td>3.4899</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left Invariant</td>
<td>0.3102</td>
<td>0.6536</td>
<td>1.3596</td>
<td>1.7515</td>
<td>2.1533</td>
<td>3.4046</td>
<td>2.9212</td>
<td>3.5747</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeffrey’s Rule</td>
<td>0.2933</td>
<td>0.6678</td>
<td>1.3959</td>
<td>1.7979</td>
<td>2.2074</td>
<td>3.5048</td>
<td>2.9903</td>
<td>3.6620</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.3191</td>
<td>0.6257</td>
<td>1.2903</td>
<td>1.6629</td>
<td>2.0493</td>
<td>3.2141</td>
<td>2.7888</td>
<td>3.4075</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>Right Invariant</td>
<td>0.1604</td>
<td>0.5344</td>
<td>0.8011</td>
<td>1.1138</td>
<td>1.4670</td>
<td>1.6144</td>
<td>2.0769</td>
<td>2.3937</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left Invariant</td>
<td>0.1626</td>
<td>0.5428</td>
<td>0.8131</td>
<td>1.1295</td>
<td>1.4859</td>
<td>1.6335</td>
<td>2.1041</td>
<td>2.4284</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jeffrey’s Rule</td>
<td>0.1648</td>
<td>0.5513</td>
<td>0.8254</td>
<td>1.1455</td>
<td>1.5051</td>
<td>1.6529</td>
<td>2.1317</td>
<td>2.3600</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.1583</td>
<td>0.5262</td>
<td>0.7893</td>
<td>1.0984</td>
<td>1.4484</td>
<td>1.5957</td>
<td>2.0501</td>
<td>2.4637</td>
<td></td>
</tr>
</tbody>
</table>

From the table it follows that for fixed sample size and priors the critical value (percentile) increases with CV. Further for each CV, for sample sizes \( n=10, 20 \) and 40, the critical value is marginally higher for uniform prior, while for sample sizes \( n=60, 80, 100 \) and 200, Jeffreys’ rule prior has marginally higher critical value. Table 5.2
presents the coverage probability of equitailed credible interval for median for selected sample sizes and priors.

A close glance at the table reveals that irrespective of the priors the credible interval maintains coverage probability when the sample size is ≥ 100. As in Guddattu and Rao (2010) we say that a 95% credible interval maintains coverage probability if the estimated coverage probability ranges between 0.945 to 0.955, i.e., 10% difference in type 1 error rate when the nominal level is 0.05. For each prior and sample size the coverage probability changes with CV, but no systematic pattern can be identified regarding the fluctuations across CV. For each combination of sample size and CV, the difference in coverage probability is marginal for the 4 priors under consideration.
Table 5.3 Estimated critical values for the evidence measure for various combinations of CV of the distribution and sample size $n=10, 20, 60$ and $100$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Prior</th>
<th>Critical Values for Evidence measure when CV equal to</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>10</td>
<td>Right Invariant</td>
<td>0.0730</td>
</tr>
<tr>
<td></td>
<td>Left Invariant</td>
<td>0.0559</td>
</tr>
<tr>
<td></td>
<td>Jeffrey's Rule</td>
<td>0.0410</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.1003</td>
</tr>
<tr>
<td>20</td>
<td>Right Invariant</td>
<td>0.1079</td>
</tr>
<tr>
<td></td>
<td>Left Invariant</td>
<td>0.0979</td>
</tr>
<tr>
<td></td>
<td>Jeffrey's Rule</td>
<td>0.0852</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.1209</td>
</tr>
<tr>
<td>60</td>
<td>Right Invariant</td>
<td>0.1348</td>
</tr>
<tr>
<td></td>
<td>Left Invariant</td>
<td>0.1275</td>
</tr>
<tr>
<td></td>
<td>Jeffrey's Rule</td>
<td>0.1261</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.1411</td>
</tr>
<tr>
<td>100</td>
<td>Right Invariant</td>
<td>0.1320</td>
</tr>
<tr>
<td></td>
<td>Left Invariant</td>
<td>0.1278</td>
</tr>
<tr>
<td></td>
<td>Jeffrey's Rule</td>
<td>0.1179</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0.1350</td>
</tr>
</tbody>
</table>

From the table 5.3 which presents the estimated lower $5^{th}$ percentile (critical value) of the evidence measure $Ev = 1 - \gamma$, we observe that critical values for sample size $n=10$ ranges from 0.03 to 0.12 while for sample size $n=20$ it ranges from 0.05 to 0.13 for various values of CV. As sample size increases the critical values fluctuate around the value of 0.10 for all the priors irrespective of the values of CV.

Jeffrey (1961) and Kass Raftery (1995) have provided guideline for taking decision using Bayes factor. In this paper an attempt is made to provide guideline for the FBST. Table 5.4 displays the quantiles of the distribution of $Ev = 1 - \gamma$ when CV=1.5 and $n=100$. 


Table 5.4 Estimated Quantiles of the distribution of $E_v = 1 - \gamma$ when CV=1.5 and $n=100$

<table>
<thead>
<tr>
<th>Prior</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right Invariant</td>
<td>0.0268</td>
<td>0.1239</td>
<td>0.2402</td>
<td>0.4310</td>
<td>0.5727</td>
<td>0.6944</td>
<td>0.7708</td>
<td>0.8600</td>
<td>0.9336</td>
<td>0.9671</td>
<td>0.9929</td>
</tr>
<tr>
<td>Left Invariant</td>
<td>0.0219</td>
<td>0.1218</td>
<td>0.2392</td>
<td>0.4245</td>
<td>0.5689</td>
<td>0.6937</td>
<td>0.7675</td>
<td>0.8591</td>
<td>0.9334</td>
<td>0.9659</td>
<td>0.9929</td>
</tr>
<tr>
<td>Jeffrey’s Rule</td>
<td>0.0234</td>
<td>0.1181</td>
<td>0.2308</td>
<td>0.4256</td>
<td>0.5672</td>
<td>0.6898</td>
<td>0.7683</td>
<td>0.8579</td>
<td>0.9319</td>
<td>0.9657</td>
<td>0.9925</td>
</tr>
<tr>
<td>Uniform</td>
<td>0.0260</td>
<td>0.1273</td>
<td>0.2443</td>
<td>0.4312</td>
<td>0.5761</td>
<td>0.6968</td>
<td>0.7698</td>
<td>0.8606</td>
<td>0.9351</td>
<td>0.9669</td>
<td>0.9930</td>
</tr>
</tbody>
</table>

Based upon the table the following guideline is suggested. When $E_v \leq 0.03$ there is concluding evidence against the null hypothesis, when $E_v$ ranges between 0.03 to 0.13 there is a strong evidence against the null hypothesis, when $E_v$ ranges between 0.13 to 0.24, although there is evidence against the null hypothesis it is not strong enough and conclusion has to be taken with care, when $E_v$ ranges between 0.24 to 0.43 there is substantial evidence in support of the null hypothesis, when $E_v \geq 0.43$, there is very strong evidence in support of the null hypothesis.

5.2 Power comparison

The power curves for the three tests and four priors along with the t-test, when the sample size is $n=100$ and CV=0.1 is presented in figures 5.1 to 5.4.
Fig 5.2 Power Curve using Left Invariant prior when n=100, CV=0.1

Fig 5.3 Power Curve using Jeffreys Rule prior when n=100, CV=0.1
The sample size $n=100$ is chosen for the power comparison so that the test based on credible region maintains level of significance. The power curves are also estimated when CV=1 and 1.5 and to save space it is not shown here. For these sample sizes and values of CV, the credible interval maintains confidence level. For comparison of Bayes tests, the power of t-test is taken as the gold standard as it is the Uniformly Most Powerful Unbiased (UMPU) Test. The conclusion that follows depends on all the power curves and not necessarily to those which are presented here. When the alternatives are close to the null hypothesis, the power of the test based on Bayes factor is marginally higher irrespective of the fact that the alternative is either left or right of the null hypothesis. The power of all the tests converges to the value 1 at the same rate. The shape of the power curve is almost symmetrical around the null hypothesis. For the power curves when CV=1 and 1.5 (not presented here) the pattern does not change but the power curve attains the value 1 at a slower rate compared to small values of CV=0.1. The overall conclusion that emerges is that for the alternatives close to the null hypothesis Bayes factor test is slightly more powerful compared to the other tests including t-test, while there is no difference in the power of all the tests for far away alternatives. Power of each of the test does not depend on the specification of the prior distribution.
6. Example
(Source: http://www.nseindia.com/products/content/equities/equities/eq_security.htm)

To illustrate the difference in the evidence provided by the Bayes tests and classical t-test, we have analysed the stock market data. The script chosen is TATA STEEL which is listed in the National Stock Exchange (NSE) limited, India for the period May and June, 2015. The price of the stock is the average price per day and the data is available from the above link. Although analysis of stock prices is a complex problem, we have limited our investigation to check whether the median stock price for the month of June 2015 is the same as that of May 2015.

The Q-Q plot for the log transformed data for the month of June 2015 is presented in figure 6.1 which shows that the transformed data is normal which in turn implies that the distribution of stock prices is lognormal.

![Q-Q plot for log transformed stock price data for the script TATA STEEL](image)

For the month of May 2015 the median value turns out to be Rs.361.035. Thus the null hypothesis of interest is that

\[ H_0: \text{Median stock price for the month of June 2015,} \]
\[ i.e. e^\mu = 361.035 \text{ or equivalently } \mu = 5.8890 \]

Table 6.1 presents the calculated p-value of the t-test and the evidence measures for the three Bayesian tests. All the tests conclusively suggests that the null hypothesis has to be rejected. The p-value for the t-test is equal to zero. For this example the degree of evidence against the null hypothesis provided by the t-test is equally strong as the Bayes factor test and the FBST. Therefore the conclusion that can be drawn is that for
standard examples use computationally simpler procedure such as the t-test or the Bayes factor test rather than computationally intensive FBST.

<table>
<thead>
<tr>
<th>Test</th>
<th>Prior</th>
<th>Statistic/Credible Interval</th>
<th>Critical value</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes factor ($B_{10}$)</td>
<td>Right invariant</td>
<td>$2.8148 \times 10^{17}$</td>
<td>2.7936</td>
<td>Reject</td>
</tr>
<tr>
<td></td>
<td>Left invariant</td>
<td>$1.6284 \times 10^{18}$</td>
<td>2.7597</td>
<td>Reject</td>
</tr>
<tr>
<td></td>
<td>Jeffrey’s rule</td>
<td>$9.4286 \times 10^{18}$</td>
<td>2.7288</td>
<td>Reject</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>$4.8698 \times 10^{16}$</td>
<td>2.8311</td>
<td>Reject</td>
</tr>
<tr>
<td>Credible interval</td>
<td>Right invariant</td>
<td>(5.7124, 5.7366)</td>
<td>-</td>
<td>Reject</td>
</tr>
<tr>
<td></td>
<td>Left invariant</td>
<td>(5.7125, 5.7359)</td>
<td>-</td>
<td>Reject</td>
</tr>
<tr>
<td></td>
<td>Jeffrey’s rule</td>
<td>(5.7128, 5.7357)</td>
<td>-</td>
<td>Reject</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>(5.7122, 5.7365)</td>
<td>-</td>
<td>Reject</td>
</tr>
<tr>
<td>FBST</td>
<td>Right invariant</td>
<td>0</td>
<td>0.1017</td>
<td>Reject</td>
</tr>
<tr>
<td></td>
<td>Left invariant</td>
<td>0</td>
<td>0.0909</td>
<td>Reject</td>
</tr>
<tr>
<td></td>
<td>Jeffrey’s rule</td>
<td>0</td>
<td>0.0820</td>
<td>Reject</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>0</td>
<td>0.1209</td>
<td>Reject</td>
</tr>
<tr>
<td>T-test</td>
<td>-</td>
<td>-26.6785</td>
<td>$t_{0.025}(21) = 2.0796$</td>
<td>Reject</td>
</tr>
</tbody>
</table>

* Critical value refers to sample size $n=20$ and $CV=2.5$

7. Conclusion

The primary focus of this paper is the frequentist comparison of the three Bayesian tests namely the Bayes factor test, test based on credible intervals and Full Bayesian Significance Test (FBST) introduced by Pereira and Stern (1999), per se and not the comparison of Bayesian and frequentist tests. The reason for this is to recommend a Bayes test having frequentist optimal properties to the users who are neutral as to the choice of Frequentist or Bayesian tests. For this purpose we have selected lognormal distribution, and the null hypothesis of interest is to test for the specified value of median. This is a standard example for which uniformly most powerful unbiased t-test exists. Although frequentist comparison of the FBST and several classical tests for testing the independence of the Holgate distribution was
undertaken by Stern and Zacks (2002), they have not included Bayes factor test and credible region test in their comparison. Thus the investigation of the present paper is much more comprehensive. The overall conclusion is that, use the frequentist procedure for standard problems as it is computationally simpler compared to several Bayes tests. This conclusion coincides with the conclusion of Severini (1999) from a different context. For non standard problems the commonly used frequentist tests are the Likelihood ratio, Wald and Score tests. Finite sample comparison of these tests did not arrive at unified conclusion regarding the choice of the test (see Nairy and Rao 2003; Bhat and Rao 2007; Guddattu and Rao 2009, 2010; Sumathi and Rao 2009, 2010, 2011, 2013, 2014 and the references cited therein). A bootstrap version of these tests is computationally tedious therefore our suggestion in such situations is to use Bayes factor test with an objective prior. The reason for this is that the computational burden is minimal for the Bayes factor test and the frequentist may not have any objection for using Bayes procedures having good power. The comparison of Bayesian significance tests in non standard situations is beyond the scope of the present paper and is a topic for future research.

Jeffrey (1961) and Kass and Raftery (1995) have suggested guideline for taking decision using Bayes factor. In this paper guideline is also suggested for the use of FBST. Although the guideline is suggested with reference to lognormal distribution, we feel that the rule should be satisfactory in other cases also. The rationale of the FBST is appealing even to the frequentist. For each sample size and various values of CV, the average computational time required for simulation is less than 1 minute for Bayes factor and is approximately 30 minutes for the FBST. The computation of evidence measure in FBST involves 2 steps namely maximization and integration. Therefore it is computationally more tedious compared to the other procedures. The applied researchers can use FBST if a general purpose package is made available. In the absence of such a package, the evidence measure can be computed in the public domain software R involving 3 stages: a) Use maxLik package and carry out the maximization b) Use MCMCpack package and obtain 10,000 observations or the number of simulations specified by the user for the parameters from the posterior distribution c) Carry out Monte Carlo integration using the generated samples in b).
Acknowledgements

The first author would like to take the opportunity to thank Government of India, Ministry of Science and Technology, Department of Science and Technology, New Delhi, for sponsoring her with the award of INSPIRE fellowship, which enables her to carry out the research programme which she has undertaken. She is much honoured to be the recipient of this award.

References


Guddattu, V., Rao, K.A.: On the use of observed fisher information in Wald and Score test, INTERSTAT # 003, (2009)


Harvey, J., Merwe, A.J.V.: Bayesian confidence intervals for the ratio of means of lognormal data with zeros, A Microsoft word-Article (2010)


Limpert, E., Stahel, W.A., Abbt, M. Log-normal Distributions across the Sciences: Keys and Clues On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question. BioScience, 51(5), 341-352 (2001)


Appendix

Proof of Theorem 1:
For the computation of $\gamma$ we need to identify the region of the parameter space where

$$\pi(\mu, \sigma^2) > \pi(\mu_0, \tilde{\sigma}^2),$$

where $\tilde{\sigma}^2$ is the value where $\pi(\mu_0, \sigma^2)$ attains its maximum value. Taking Taylor series expansion of the L.H.S around the mode of the posterior distribution we have,

$$ax^2 + 2bxy + cy^2 + O_p\left(n^{-3/2}\right) > k$$

where $x = \mu - \bar{\mu}$ and $y = \sigma^2 - \bar{\sigma}^2$, where $(\bar{\mu}, \bar{\sigma}^2)$ denotes the modal vector of the posterior density and $a, b$ and $c$ refer to the second derivative of the posterior density.

Suppressing the order term and rearranging the inequality we have

$$(x + dy)^2 > k + h(\sigma^2),$$

for some constant $d$. The inequality is equivalent to $x < g_1(\sigma^2)$ or $x > g_2(\sigma^2)$, where $g_1(\sigma^2) = -k - h(\sigma^2) - dy$ and $g_2(\sigma^2) = k + h(\sigma^2) - dy$
∴ \gamma = \int_{\eta} \left\{ \Phi \left( \frac{g_1(\sigma^2)}{\sigma} \right) + 1 - \Phi \left( \frac{g_2(\sigma^2)}{\sigma} \right) \right\} G_\eta(\cdot) \, d\eta + O(n^{-3/2}),

where \Phi(\cdot) denotes the CDF of the standard normal distribution and \( G_\eta(\cdot) \) denotes the posterior marginal distribution for \( \eta = \frac{1}{\sigma^2} \). Note that for the evaluation of the integral to the desired degree of accuracy one has to use the expansion for the CDF of the standard normal distribution and subsequent Taylor series expansion for \( g_1(\sigma^2)/\sigma \) and \( g_2(\sigma^2)/\sigma \) around \( \sigma^2 = 0 \). The integral can also be evaluated by generating observations from the gamma distribution.