Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 34

December 2017

Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique

Oyebayo Ridwan Olaniran
Universiti Tun Hussein Onn Malaysia, Muar, Johor, Malaysia, rid4stat@yahoo.com

Waheed Babatunde Yahya
University of Ilorin, Ilorin, dr.yah2009@gmail.com

Recommended Citation
Olaniran, O. R., & Yahya, W. B. (2017). Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique. Journal of Modern Applied Statistical Methods, 16(2), 618-638. doi: 10.22237/jmasm/1509496440

Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique

Cover Page Footnote
My utmost thanks go to my parents, Ganiyu Oyebanji Olaniran and Fatimo Omowumi Olaniran. They bore me, raised me and supported me. To them I dedicate this paper.

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 618-638.
doi: 10.22237/jmasm/1509496440

ISSN 1538 − 9472

Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique

Oyebayo Ridwan Olaniran
Universiti Tun Hussein Onn Malaysia
Muar, Johor, Malaysia

Waheed Babatunde Yahya
University of Ilorin
Ilorin, Kawara State, Nigeria

The most important ingredient in Bayesian analysis is prior or prior distribution. A new prior determination method was developed under the framework of parametric empirical Bayes using bootstrap technique. By way of example, Bayesian estimations of the parameters of a normal distribution with unknown mean and unknown variance conditions were considered, as well as its application in comparing the means of two independent normal samples with several scenarios. A Monte Carlo study was conducted to illustrate the proposed procedure in estimation and hypothesis testing. Results from Monte Carlo studies showed that the bootstrap prior proposed is more efficient than the existing method for determining priors and also better than the frequentist methods reviewed.

Keywords: Prior, conjugacy, bootstrapping, hypothesis testing, Monte Carlo studies

Introduction

Bayesian statistics have several advantages over the traditional classical (frequentist) statistics ranging from proffering solution to problems related to estimation, testing hypotheses, or estimating confidence regions for unknown parameters. The reason is by use of Bayes' theorem probability density functions are obtained for the unknown parameters. These density functions allow for the estimation of unknown parameters, the testing of hypotheses, and the computation of confidence regions often referred to as the credible interval. Therefore, application of Bayesian statistics has been spreading (Koch, 2007). The process of inductive learning via Bayes' rule is referred to as Bayesian inference (Hoff, 2009).

The Bayesian inference utilizes the posterior distribution $p(\theta \mid y)$ which describes our belief that θ is the true value, having observed dataset y. The posterior

Oyebayo Ridwan Olaniran is a PhD student and Graduate Research Assistant in the Department of Mathematics and Statistics, Faculty of Applied Science and Technology. Email him at: rid4stat@yahoo.com. Waheed Babatunde Yahya is an Associate Professor in the Department of Statistics, Faculty of Physical Sciences. Email them at dr.yah2009@gmail.com.

distribution is obtained from the prior distribution and sampling model via Bayes' rule:

$$\mathrm{p}(\theta|\mathrm{y}) = \frac{\mathrm{p}(\mathrm{y}|\theta)\mathrm{p}(\theta)}{\int_{\theta}\mathrm{p}(\mathrm{y}|\theta)\mathrm{p}(\theta)d\theta} \tag{1}$$

The expression given by (1) above is the general Bayes theorem for inference and is the basis for making inferences from a Bayesian perspective in terms of estimation, hypothesis testing, and obtaining credible intervals, as well as making direct probability statements about the quantities in which we are interested (Spiegelhalter, Abrams, & Myles, 2004).

The Bayes theorem is commonly written in its proportional form as $\mathrm{p}(\theta \mid \mathrm{y}) \propto \mathrm{p}(\mathrm{y} \mid \theta)\mathrm{p}(\theta)$. Bayesian inference is based on the posterior distribution, which is the conditional distribution of the parameters or unobserved covariates given the observed data. The posterior distribution summarizes all the information about the parameters and covariates. For example, the mean, median, or mode of the posterior distribution could be used as point estimators. Bayesian inference for θ is then based on the posterior distribution $\mathrm{p}(\mathrm{y} \mid \theta)$. For example, a Bayesian estimator of θ is the posterior mean

$$\hat{\theta} = \mathrm{E}(\theta|\mathrm{y}) = \int_{\theta}\theta\mathrm{p}(\theta|\mathrm{y})d\theta$$

A Bayesian analogue to a confidence interval is the credible interval, which is a region with probability $1 - \alpha$ under the posterior distribution. Choices of prior distributions are important. In fact, much of the controversy regarding Bayesian methods revolves around the prior distributions (Wu, 2010).

Priors and Prior Distributions

Priors are carriers of information that is coherently incorporated via Bayes' theorem to the inference. At the same time, parameters are unobservable, and prior specification is subjective in nature. There are two different schools of thought to be considered when choosing priors in Bayesian analysis (Rouder, Speckman, Sun, & Morey, 2009). The first is the subjective Bayes school, which believe that priors should reflect the analyst's *a priori* beliefs about parameters. Usually, these beliefs are informed by the theoretical and experimental context. The second is the objective Bayes school, in which the priors are meant to reflect as few assumptions

as possible. Bayes himself proposes a class of uniform prior for the parameter *p* of a binomial distribution communicated to the Royal Statistical Society by Price in 1763 (Bayes & Price, 1763).

Laplace in the early 1770s extended this prior belief in his principle of insufficient reason and termed it as flat prior. Efron (2012) claimed that the Bayesian/frequentist controversy centers on the use of Bayes' rule in the absence of genuine prior experience. Due to the parameter invariant problem involved when using the uniform prior, Jeffreys (1949) proposed another class of prior that is invariant to parameter transformation (Lesaffre & Lawson, 2013). Both the uniform flat prior and Jeffrey prior are usually referred to as objective or non-informative prior.

In the search of genuine or informative prior, Raifa and Schlaifer proposed the conjugate prior in 1961 (as reported in Bolstad, 2004). The conjugate prior ensures that the posterior distribution class is the same as the prior distribution. Conjugacy is mathematically convenient in that the posterior distribution follows a known parametric form (Gelman et al., 2013). If information is available that contradicts a conjugate parametric family, then it may be necessary to use a more realistic but often inconvenient prior distribution. A conjugate prior can be made informative or non-informative depending on the parameter value assumed. Yahya, Olaniran, and Ige (2014) claimed that the conjugate prior approach needs to be updated when the genuine prior parameter is not available. Using a conjugate prior does not necessarily guarantee an adequate posterior unless the parameter of the prior distribution is correctly specified. The adverse effect of incorrect prior specification is when the prior information did not agree with the data information which might lead to incorrect estimation or inference about the unknown parameter. Solving this problem led to the proposition of empirical Bayes in the early 1950s by Robbins as reported in Robbins (1956), Martiz (1970), Efron and Morris (1973, 1975, 1976), Morris (1983), Casella (1985), Bishop (2005), and recently in Efron (2012, 2013, 2014), Lee (2012), and Lesaffre and Lawson (2013).

Empirical Bayes methods are procedures for statistical inference in which the prior distribution is estimated from the data (Lee, 2012). This method is often perceived in two forms: the parametric (known functional form) and non-parametric (unknown functional form). Parametric empirical Bayes usually involve the use of conjugate prior with the prior parameters estimated from the data using the maximum likelihood estimation (MLE) method or method of moment (MM) (Lee, 2012). Efron (2014) reported the use of empirical Bayes methods is increasing, although still suffers from an uncertain theoretical basis, enjoying neither the safe haven of Bayes' theorem nor the steady support of frequentist optimality. Their

rationale is often reduced to inserting more or less obvious estimates into familiar Bayesian formulas. This conceals the essential empirical Bayes task: learning an appropriate prior distribution from ongoing statistical experience, rather than knowing it by assumption.

Efficient learning requires both Bayesian and frequentist modeling strategies. Bayesian statistics with well-known distributions are often smooth and easy with the use of conjugate priors with adequate prior parameter specification using subjective or empirical Bayes method. However, in most real life situations, it is often difficult to describe using existing or known functional form. Posterior distributions under this situation are often estimated using Monte Carlo integration or methods popularly referred to as Markov chain Monte Carlo (MCMC) methods. MCMC methods ranges from the Gibbs sampler (Casella & George, 1992), expectation maximization (EM) algorithm, to the Metropolis-Hasting (MH) algorithm (Lee, 2012).

Currently, the focus is on updating the parametric empirical Bayes procedure using bootstrapping resampling procedures. The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one's data (Efron & Tibshirani, 1993). It amounts to treating the data as if they were the population for the purpose of evaluating the distribution of interest. Under mild regularity conditions, the bootstrap yields an approximation to the distribution of an estimator or test statistic that is at least as accurate as the approximation obtained from first-order asymptotic theory. Thus, the bootstrap provides a way to substitute computation for mathematical analysis if calculating the asymptotic distribution of an estimator or statistic is difficult.

Therefore, the aim of this study is to develop efficient alternative methods for determining the priors within the Bayesian framework using bootstrapping techniques. The usefulness of the proposed method in classical hypothesis testing is demonstrated for comparing two population means from two independent samples. The efficiencies of the proposed methods shall be determined and compared with some of the existing frequentists and Bayesian test methods under different parameters combinations.

Methodology

Consider random sample $y_1, y_2, \ldots, y_n$ from $N(\mu, \sigma^2)$. The density function of y is given as

$$\mathrm{p}\left(y \mid \mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\left(y-\mu\right)^2\right] \tag{2}$$

The problem is how to effectively estimate the location and scale parameters μ and σ^2, respectively. The Bayes estimation procedures for μ and σ^2 require estimation of the posterior distribution of μ and σ^2 given y. The posterior density following Bayes' theorem is

$$\mathrm{p}\left(\mu,\sigma^2 \mid y\right)=\frac{\mathrm{p}\left(\mu,\sigma^2\right)\mathrm{p}\left(y \mid \mu,\sigma^2\right)}{\int_{\mu}\int_{\sigma^2}\mathrm{p}\left(\mu,\sigma^2\right)\mathrm{p}\left(y \mid \mu,\sigma^2\right)d\sigma^2 d\mu} \tag{3}$$

Bolstad (2004), Murphy, (2007), and Lesaffre and Lawson (2013), among others, used the Normal-Gamma NG(μ_0, n_0, α_0, β_0) natural conjugate prior for μ and $\lambda = \sigma^{-2}$, given as

$$\mathrm{p}\left(\mu,\lambda\right)=\frac{\left[\lambda n_0\right]^{\frac{1}{2}}}{\sqrt{2\pi}}\exp\left[-\frac{\lambda n_0}{2}\left(\mu-\mu_0\right)^2\right]\times\frac{\beta_0^{\alpha_0}\lambda^{\alpha_0-1}\exp\left(-\lambda\beta_0\right)}{\Gamma\alpha_0}$$

To characterize the information from the data D = $y_1, y_2, \ldots, y_n$, define the likelihood function L(D | μ, σ^2):

$$\begin{aligned}\mathrm{L}\left(\mathrm{D} \mid \mu,\lambda\right)&=\prod_{i=1}^{n}\frac{\lambda^{\frac{1}{2}}}{\sqrt{2\pi}}\exp\left[-\frac{\lambda}{2}\left(y_i-\mu\right)^2\right]\\&=\left(\frac{\lambda^{\frac{1}{2}}}{\sqrt{2\pi}}\right)^n\exp\left[-\frac{\lambda}{2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right]\end{aligned}$$

The posterior distribution is of the form

$$\mathrm{p}\left(\mu,\lambda \mid y\right)=\frac{\mathrm{p}\left(\mu,\lambda\right)\mathrm{L}\left(\mathrm{D} \mid \mu,\lambda\right)}{\int_{\mu}\int_{\lambda}\mathrm{p}\left(\mu,\lambda\right)\mathrm{L}\left(\mathrm{D} \mid \mu,\lambda\right)d\lambda d\mu}$$

Murphy (2007) gave the solution of the posterior density which is also Normal-Gamma, i.e. NG(μ_n, n_n, α_n, β_n) where

$$\mu_n = \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \tag{4}$$

$$n_n = n_0 + n \tag{5}$$

$$\alpha_n = \alpha_0 + \frac{n}{2} \tag{6}$$

$$\beta_n = \beta_0 + \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 + \frac{n_0 n\left(\bar{y} - \mu_0\right)^2}{2\left(n_0 + n\right)} \tag{7}$$

and

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Thus, the Bayes estimate of μ is

$$\hat{\mu} = \mu_n$$

and, from (4),

$$\begin{aligned} \hat{\mu} &= \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \\ &= \frac{n\bar{y}}{n_0 + n} + \frac{n_0\mu_0}{n_0 + n} \end{aligned} \tag{8}$$

Let

$$w = \frac{n}{n_0 + n}$$

Then

$$\begin{aligned}\hat{\mu} &= w\bar{y} + (1-w)\mu_0 \\ &= w(\bar{y} - \mu_0) + \mu_0\end{aligned}$$

Similarly, the Bayes estimate of σ^2 is determined by

$$\hat{\sigma}^2 = \frac{\beta_n}{\alpha_n}$$

and from (6) and (7)

$$\hat{\sigma}^2 = \frac{\beta_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{n_0 n(\bar{y} - \mu_0)^2}{2(n_0 + n)}}{\alpha_0 + \frac{n}{2}} \tag{9}$$

The empirical Bayes version of the above estimate involves estimating the prior parameters $\boldsymbol{\pi} = (\mu_0, n_0, \alpha_0, \beta_0)$ from the data. Thus the empirical Bayes estimate of μ and σ^2 are

$$\hat{\mu}_{\text{EB}} = \frac{\hat{n}_0\hat{\mu}_0 + n\bar{y}}{\hat{n}_0 + n} \tag{10}$$

and

$$\hat{\sigma}^2_{\text{EB}} = \frac{\hat{\beta}_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{\hat{n}_0 n(\bar{y} - \hat{\mu}_0)^2}{2(\hat{n}_0 + n)}}{\hat{\alpha}_0 + \frac{n}{2}} \tag{11}$$

The proposed bootstrap Bayesian version of the estimate of μ and σ^2 involves the following steps:

1. Generation of bootstrap samples from the original data a desired number of times *B*,

2. Estimating the hyperparameters (prior parameters) each time the samples are generated using the maximum likelihood (ML) method,
3. Updating the posterior estimates using the hyperparameters in step 2 above using (8) and (9), and
4. Then obtaining the proposed bootstrap empirical Bayesian estimates $\hat{\mu}_{\mathrm{BT}}$ and $\hat{\sigma}^2_{\mathrm{BT}}$ using

$$\hat{\mu}_{\mathrm{BT}} = \frac{1}{B}\sum_{j=1}^{B}\hat{\mu}_{\mathrm{EB}j}$$
$$\hat{\sigma}^2_{\mathrm{BT}} = \frac{1}{B}\sum_{j=1}^{B}\hat{\sigma}^2_{\mathrm{EB}j}$$

The $\hat{\mu}_{\mathrm{BT}}$ proposed here has good statistical properties in terms of bias and mean square error (MSE).

To evaluate bias,

$$\begin{aligned}\hat{\mu}_{\mathrm{BT}} &= \frac{1}{B}\sum_{j=1}^{B}\hat{\mu}_{\mathrm{EB}j} \\ &= \frac{1}{B}\sum_{j=1}^{B}\frac{\hat{n}_{0j}\hat{\mu}_{0j} + n\bar{y}}{\hat{n}_{0j} + n} \\ &= \frac{1}{B}\sum_{j=1}^{B}\left[\frac{n\bar{y}}{\hat{n}_{0j} + n} + \frac{\hat{n}_{0j}\hat{\mu}_{0j}}{\hat{n}_{0j} + n}\right]\end{aligned}$$

Fixing $\hat{n}_{0j} = B$ and $\hat{\mu}_{0j} = \bar{y}_{bj}$, where $\bar{y}_{bj}$ is the j^{th} ML estimate based on the j^{th} y_b bootstrap sample drawn, i.e.

$$\bar{y}_{bj} = \frac{\sum_{i=1}^{n} y_{ib}}{n}$$

then

$$\begin{aligned}\hat{\mu}_{\mathrm{BT}} &= \frac{1}{B}\sum_{j=1}^{B}\left[\frac{n\bar{y}}{B+n}+\frac{B\bar{y}_{bj}}{B+n}\right] \\ &= \frac{1}{B(B+n)}\sum_{j=1}^{B}\left[n\bar{y}+B\bar{y}_{bj}\right] \\ &= \frac{n\bar{y}}{(B+n)}+\frac{\sum_{j=1}^{B}\bar{y}_{bj}}{(B+n)}\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{E}\left[\hat{\mu}_{\mathrm{BT}}\right] &= \mathrm{E}\left[\frac{n\bar{y}}{(B+n)}+\frac{\sum_{j=1}^{B}\bar{y}_{bj}}{(B+n)}\right] \\ &= \frac{n\,\mathrm{E}\left[\bar{y}\right]}{(B+n)}+\frac{\sum_{j=1}^{B}\mathrm{E}\left[\bar{y}_{bj}\right]}{(B+n)}\end{aligned}$$

Because $\bar{y}$ and $\bar{y}_{bj}$ are known unbiased estimates of μ,

$$\begin{aligned}\mathrm{E}\left[\hat{\mu}_{\mathrm{BT}}\right] &= \frac{n\mu}{(B+n)}+\frac{\sum_{j=1}^{B}\mu}{(B+n)} \\ &= \frac{1}{(B+n)}\left[n\mu+B\mu\right] \\ &= \mu\end{aligned}$$

Therefore, $\hat{\mu}_{\mathrm{BT}}$ is unbiased for estimating μ.

Also, the MSE is the combination of square of bias and variance of the estimate, then following from the above derivation the MSE is just the variance of the estimate. Thus

$$\mathrm{var}\left[\hat{\mu}_{\mathrm{BT}}\right] = \mathrm{var}\left[\frac{n\bar{y}}{(B+n)}+\frac{\sum_{j=1}^{B}\bar{y}_{bj}}{(B+n)}\right]$$

$$= \frac{n^2 \operatorname{var}\left[\bar{y}\right]}{\left(B+n\right)^2} + \frac{\sum_{j=1}^{B} \operatorname{var}\left[\bar{y}_{bj}\right]}{\left(B+n\right)^2}$$
$$= \frac{n^2 \dfrac{\sigma^2}{n}}{\left(B+n\right)^2} + \frac{\displaystyle\sum_{j=1}^{B} \frac{\sigma^2}{n}}{\left(B+n\right)^2}$$
$$= \left[\frac{n^2+B}{\left(B+n\right)^2}\right]\frac{\sigma^2}{n}$$

Hence, it can be show that the limiting form of

$$\left[\frac{n^2+B}{\left(B+n\right)^2}\right]$$

is 0 by applying L'Hôpital's rule (Weisstein, 2003):

$$\lim_{B\to\infty}\left[\frac{n^2+B}{\left(B+n\right)^2}\right] = \lim_{B\to\infty}\left[\frac{\dfrac{d\left[n^2+B\right]}{dB}}{\dfrac{d\left(B+n\right)^2}{dB}}\right]$$
$$= \lim_{B\to\infty}\left[\frac{1}{2\left(B+n\right)}\right]$$
$$\to \frac{1}{\infty}$$

Hence

$$\lim_{B\to\infty}\left[\frac{n^2+B}{\left(B+n\right)^2}\right] = 0$$

The above derivation implies that at a fixed sample size *n*, the

$$\lim_{B\to\infty} \text{var}\left[\hat{\mu}_{\text{BT}}\right] = 0$$

This affirms that the experimenter can control the stability of the estimator by increasing the number of bootstrap sample *B*. In addition,

$$\text{var}\left[\hat{\mu}_{\text{BT}}\right] = \left[\frac{n^2 + B}{\left(B+n\right)^2}\right]\frac{\sigma^2}{n} < \text{var}\left[\hat{\mu}_{\text{ML}}\right] = \frac{\sigma^2}{n}$$

which implies that the proposed estimator is more efficient than the frequentist ML estimator. This comparison is reasonable because they are both unbiased. Also within the Bayesian realm, it could be observe that the proposed estimator is also more efficient since it minimizes the MSE in terms of bias and variance reduction. The traditional Bayesian estimator minimizes the MSE by reducing the variance alone.

Application to Two-Sample Hypothesis Testing

Consider the situation in which we have independent samples from two normal distributions

$$x_1, x_2, \ldots, x_{n_1} \sim \text{N}\left(\mu_1, \sigma_1^2\right)$$
$$y_1, y_2, \ldots, y_{n_2} \sim \text{N}\left(\mu_2, \sigma_2^2\right)$$

which are independent of each other, and the quantity of interest really is the posterior distribution of

$$\delta = \mu_1 - \mu_2$$

The hypothesis of interest under this scenario might be of the form

$$\text{H}_0 : \mu_1 = \mu_2 \text{ against } \text{H}_1 : \mu_1 \neq \mu_2 \tag{12}$$

or similarly in terms of δ

$$\text{H}_0 : \delta = 0 \text{ against } \text{H}_1 : \delta \neq 0 \tag{13}$$

Testing the above hypotheses in (12) and (13) using the Bayesian method requires computing p(δ | D) defined as the posterior distribution of δ given data D (Lee, 2012). The posterior probability of the null hypothesis H_0 can then be estimated using

$$\mathrm{p}\left(\mathrm{H}_0 : \delta \le 0 \,|\, \mathrm{D}\right) = \int_{-\infty}^{0} \mathrm{p}\left(\delta \,|\, \mathrm{D}\right) d\delta \tag{14}$$

If this probability is less than the chosen α, reject the null hypothesis and conclude that H_1 holds. But (14) will fail if the null hypothesis is simple as in the case of (12) and (13) above (because the probability of a specific point on a continuous interval is 0). Bolstad (2004) and Lee (2012) suggested the use of credible interval under this condition. The credible interval for a specified significance level α for parameter δ is

$$\mathrm{p}\left(a \le \delta \le b \,|\, \mathrm{D}\right) = \int_{a}^{b} \mathrm{p}\left(\delta \,|\, \mathrm{D}\right) d\delta = 1 - \alpha$$

On construction of the credible interval [a, b] using a specified significance level α, δ is said to be credible if it lies within the interval [a, b], and thus H_0 holds; otherwise, H_1 holds. The bootstrap Bayesian estimates can be used here to determine the posterior density or posterior samples of δ by using the formulae above. The bootstrap Bayesian estimate of parameter δ is $\hat{\delta}_{\mathrm{BT}} = \hat{\mu}_{1\mathrm{BT}} - \hat{\mu}_{2\mathrm{BT}}$. The posterior density of δ using the bootstrap Bayesian approach will likely approach the Gaussian distribution if one is to follow the central limit theorem (since the bootstrap prior distribution used is the sampling distribution of means which is Gaussian). Thus the posterior density of δ under this assumption is $\mathrm{N}\left(\hat{\mu}_{1\mathrm{BT}} - \hat{\mu}_{2\mathrm{BT}}, \mathrm{var}\left[\hat{\mu}_{1\mathrm{BT}}\right] + \mathrm{var}\left[\hat{\mu}_{2\mathrm{BT}}\right]\right)$. This then implies we can construct a frequentist-like test-statistic for the unknown parameter δ as

$$Z_{\mathrm{BT}} = \frac{\delta - \hat{\delta}_{\mathrm{BT}}}{\sqrt{\mathrm{var}\left[\hat{\mu}_{1\mathrm{BT}}\right] + \mathrm{var}\left[\hat{\mu}_{2\mathrm{BT}}\right]}} \sim \mathrm{N}(0,1) \,|\, \mathrm{H}_0 \tag{15}$$

In another parlance, Lee (2012) claimed that to correct for small sample bias and unequal variance bias, the Student's t-distribution with v degree of freedom would provide a better approximation than the Gaussian distribution. Thus a modification to (15) above is

$$t_{\mathrm{BT}} = \frac{\delta - \hat{\delta}_{\mathrm{BT}}}{\sqrt{\mathrm{var}\left[\hat{\mu}_{1\mathrm{BT}}\right] + \mathrm{var}\left[\hat{\mu}_{2\mathrm{BT}}\right]}} \sim t_v \mid \mathrm{H}_0 \tag{16}$$

The parameter v of the Student's t-distribution used here indicates the effective sample size to be used for the hypothesis testing. Hence, using $v = \min(n_1, n_2)$ is proposed here.

It is pertinent to note that the above equations (15) and (16) are approximate distributions of $\hat{\delta}_{\mathrm{BT}}$. The hypothesis can be tested directly by computing $\mathrm{p}(\delta_{\mathrm{BT}} \mid \mathrm{D})$ using the difference of the posteriors (generated using bootstrap priors) for the parameters μ_1 and μ_2. In this regard, the posterior probability of the null hypothesis H_0 can then be estimated using

$$\mathrm{p}\left(\mathrm{H}_0 : \delta = 0 \mid \mathrm{D}\right) \approx 2\left[\min\left(\int_{-\infty}^{0} \mathrm{p}\left(\delta_{\mathrm{BT}} \mid \mathrm{D}\right) d\delta, \int_{0}^{\infty} \mathrm{p}\left(\delta_{\mathrm{BT}} \mid \mathrm{D}\right) d\delta\right)\right] \tag{17}$$

In the same parlance, posterior probability of the null hypothesis H_0 for (15) and (16) are, respectively,

$$\mathrm{p}\left(\mathrm{H}_0 : \delta = 0 \mid \mathrm{D}\right) \approx 2\left[\min\left(\Phi\left(z\right), 1 - \Phi\left(z\right)\right)\right] \tag{18}$$

and

$$\mathrm{p}\left(\mathrm{H}_0 : \delta = 0 \mid \mathrm{D}\right) \approx 2\left[\min\left(\psi_v\left(z\right), 1 - \psi_v\left(z\right)\right)\right] \tag{19}$$

where $\Phi(z)$ is the cumulative distribution of the standard normal variate z and ψ_v is the cumulative distribution of the Student's t-distribution with mean 0, variance 1, and degrees of freedom v.

The above procedures in (15), (16), and (17) will be evaluated to ascertain which to recommend under specific situation. Consideration of the Bayesian MCMC approach to estimation and testing of equality in means for two groups proposed by Kruschke (2011, 2013) and Kruschke, Aguinis, and Joo (2012) was also achieved. This approach is already implemented in the R statistical package via the package BEST (Kruschke & Meredith, 2014). As a standard check, two frequentist procedures were also considered. The frequentist procedures considered are the pooled variance t-test and unequal variance Welch test (Montgomery & Runger, 2003).

Simulation

To illustrate the proposed bootstrap empirical Bayesian procedures in estimation and hypothesis testing, two Monte Carlo samples were generated from univariate normal distributions with the following mean structures: $\mu_1 = 10$ and $\mu_2 = \mu_1 + \delta$, where $\delta = 0, 1, 2$. The cases of equal and unequal variances were considered with equal variance case define as $\sigma_1^2 = \sigma_2^2 = 4$ and unequal variance case define as $\sigma_1^2 = 4$ and $\sigma_2^2 = 16$. Under equal sample condition, five sample size, $n_1 = n_2 = 5$, 10, 20, 30, and 50, were used representing sample ranges of extreme low to large sample. Similarly for unequal sample condition, three sample structures were considered, namely $n_1 = 5$, $n_2 = 10$; $n_1 = 10$, $n_2 = 30$; and $n_1 = 20$, $n_2 = 80$. The bootstrap size (B) and number of iterations used were fixed at 1000.

Results

The empirical type-I error rate (false positive rate) and power (true positive rate) were computed using the frequentist and Bayesian procedures discussed. The role of sample size cannot be overemphasized in estimation and hypothesis testing, therefore more emphasis will be laid on the sample size regarding the results obtained from various procedures used here.

The validity of test procedures can be assessed using the empirical type-I error rate which is the probability that a test function wrongly rejects the null hypothesis when it is true. A test procedure yielding a false positive that is close to the nominal level is often regarded as having been valid. In light of this, the first situation, or Case I in Table 1, considered the common assumptions (equal sample and homoscedasticity) involved while comparing two normal samples. The frequentist traditional pooled *t*-test produced on average (over all sample sizes) false positive rates that are relatively close to the nominal (0.05) level.

However, this result was not the best if the comparison is made over all the test procedures employed in this paper. For instance, within the Bayesian test procedures, the proposed t_{BT} test procedures produced false positive rates that are closer to the nominal (0.05) level at all the sample sizes considered. Therefore, it can be re-affirmed that the traditional pooled *t*-test is valid but the proposed t_{BT} is more valid as it yielded an overall average of the empirical type-I error (0.051) that is relatively closer to the nominal 5% level set for the test than the overall average of 0.046 provided by the pooled *t*-test. The performances of all the tests considered

as described above are clearly depicted in the various graphical plots provided in Figure 1 for better understanding.

Table 1. Proportion of empirical type-I error (false positive) rate based on 1000 simulations at varying sample sizes (n_1, n_2) and $\delta = 0$ for the various methods under equal and unequal variance conditions

			Bayesian			
	Frequentist		Existing	Proposed		
Sample sizes	**Welch**	**Pooled**	**MCMC**	**BT**	Z_{BT}	t_{BT}
Case I: Equal sample sizes and equal variance						
$n_1 = n_2 = 5$	0.040	0.051	0.007	0.145	0.119	0.051
$n_1 = n_2 = 10$	0.050	0.052	0.032	0.106	0.080	0.050
$n_1 = n_2 = 20$	0.040	0.042	0.036	0.077	0.062	0.048
$n_1 = n_2 = 30$	0.045	0.046	0.045	0.080	0.066	0.051
$n_1 = n_2 = 50$	0.041	0.041	0.039	0.069	0.067	0.057
Average	0.043	0.046	0.032	0.095	0.079	0.051
Case II: Equal sample sizes and unequal variance						
$n_1 = n_2 = 5$	0.046	0.054	0.013	0.156	0.137	0.056
$n_1 = n_2 = 10$	0.057	0.059	0.045	0.105	0.098	0.058
$n_1 = n_2 = 20$	0.032	0.035	0.028	0.065	0.057	0.039
$n_1 = n_2 = 30$	0.050	0.052	0.046	0.067	0.063	0.055
$n_1 = n_2 = 50$	0.040	0.040	0.041	0.066	0.058	0.051
Average	0.045	0.048	0.035	0.092	0.083	0.052
Case III: Unequal sample sizes and equal variance						
$n_1 = 5, n_2 = 10$	0.041	0.044	0.011	0.109	0.106	0.034
$n_1 = 10, n_2 = 30$	0.062	0.053	0.041	0.093	0.088	0.058
$n_1 = 20, n_2 = 80$	0.051	0.056	0.051	0.069	0.069	0.055
Average	0.051	0.051	0.034	0.090	0.088	0.049
Case IV: Unequal sample sizes and unequal variance with large variance in large sample direction						
$n_1 = 5, n_2 = 10$	0.040	0.019	0.018	0.095	0.084	0.029
$n_1 = 10, n_2 = 30$	0.049	0.006	0.034	0.075	0.069	0.040
$n_1 = 20, n_2 = 80$	0.048	0.005	0.045	0.064	0.062	0.053
Average	0.046	0.010	0.032	0.078	0.072	0.041
Case V: Unequal sample sizes and unequal variance with large variance in small sample direction						
$n_1 = 5, n_2 = 10$	0.066	0.120	0.024	0.136	0.128	0.073
$n_1 = 10, n_2 = 30$	0.042	0.163	0.037	0.097	0.094	0.050
$n_1 = 20, n_2 = 80$	0.044	0.159	0.040	0.062	0.063	0.052
Average	0.051	0.147	0.034	0.098	0.095	0.058

Table 2. Proportion of power (true positive rate) based on 1000 simulations at varying sample sizes (n_1, n_2) and $\delta = 1$ for the various methods under equal and unequal variance conditions

			Bayesian			
	Frequentist		Existing	Proposed		
Sample sizes	**Welch**	**Pooled**	**MCMC**	**BT**	**Z_{BT}**	**t_{BT}**
Case I: Equal sample sizes and equal variance						
$n_1 = n_2 = 5$	0.087	0.094	0.020	0.245	0.202	0.093
$n_1 = n_2 = 10$	0.171	0.176	0.126	0.262	0.257	0.178
$n_1 = n_2 = 20$	0.327	0.328	0.296	0.389	0.374	0.339
$n_1 = n_2 = 30$	0.485	0.485	0.458	0.531	0.532	0.504
$n_1 = n_2 = 50$	0.690	0.691	0.691	0.740	0.741	0.730
Average	0.352	0.355	0.318	0.433	0.421	0.369
Case II: Equal sample sizes and unequal variance						
$n_1 = n_2 = 5$	0.075	0.084	0.015	0.196	0.167	0.085
$n_1 = n_2 = 10$	0.100	0.105	0.076	0.157	0.144	0.106
$n_1 = n_2 = 20$	0.162	0.168	0.145	0.200	0.202	0.175
$n_1 = n_2 = 30$	0.225	0.232	0.215	0.265	0.255	0.238
$n_1 = n_2 = 50$	0.353	0.358	0.348	0.416	0.414	0.396
Average	0.183	0.189	0.160	0.247	0.236	0.200
Case III: Unequal sample sizes and equal variance						
$n_1 = 5, n_2 = 10$	0.114	0.133	0.043	0.227	0.221	0.102
$n_1 = 10, n_2 = 30$	0.253	0.262	0.205	0.333	0.338	0.256
$n_1 = 20, n_2 = 80$	0.505	0.496	0.480	0.570	0.569	0.521
Average	0.291	0.297	0.243	0.377	0.376	0.293
Case IV: Unequal sample sizes and unequal variance with large variance in large sample direction						
$n_1 = 5, n_2 = 10$	0.074	0.034	0.028	0.147	0.137	0.050
$n_1 = 10, n_2 = 30$	0.166	0.040	0.132	0.211	0.219	0.145
$n_1 = 20, n_2 = 80$	0.355	0.093	0.331	0.409	0.411	0.366
Average	0.198	0.056	0.164	0.256	0.256	0.187
Case V: Unequal sample sizes and unequal variance with large variance in small sample direction						
$n_1 = 5, n_2 = 10$	0.065	0.159	0.036	0.174	0.165	0.082
$n_1 = 10, n_2 = 30$	0.117	0.260	0.098	0.176	0.172	0.133
$n_1 = 20, n_2 = 80$	0.177	0.399	0.170	0.218	0.218	0.192
Average	0.120	0.273	0.101	0.189	0.185	0.136

The second scenario, Case II in Table 1, is the case where the frequentist Welch *t*-test has been established to be better. Here the equal sample sizes, but with heteroscedastic situation, was considered. As expected, the Welch *t*-test yielded

false positive rates that are relatively closer to the nominal (0.05) level. The result of the proposed Bayesian t_{BT} test is not worst-off here as it equally produced false positive rates that are quite close to the 5% nominal level and competes favorably with the results of the Welch test.

Moving to unequal sample sizes and heteroscedastic situations (Cases IV and V), similar results as observed with equal sample and unequal variance situations were observed. To assess the usability of the test procedures, the true positive (power) as assessment criteria was employed. The most powerful test procedures under the varying scenarios is the BT method which is the Bayesian method using the direct bootstrap distribution as can be observed from the various results in Tables 2 and 3 under various parameters combinations. The powers of this test method appreciated better as the values of the effect size, δ increases.

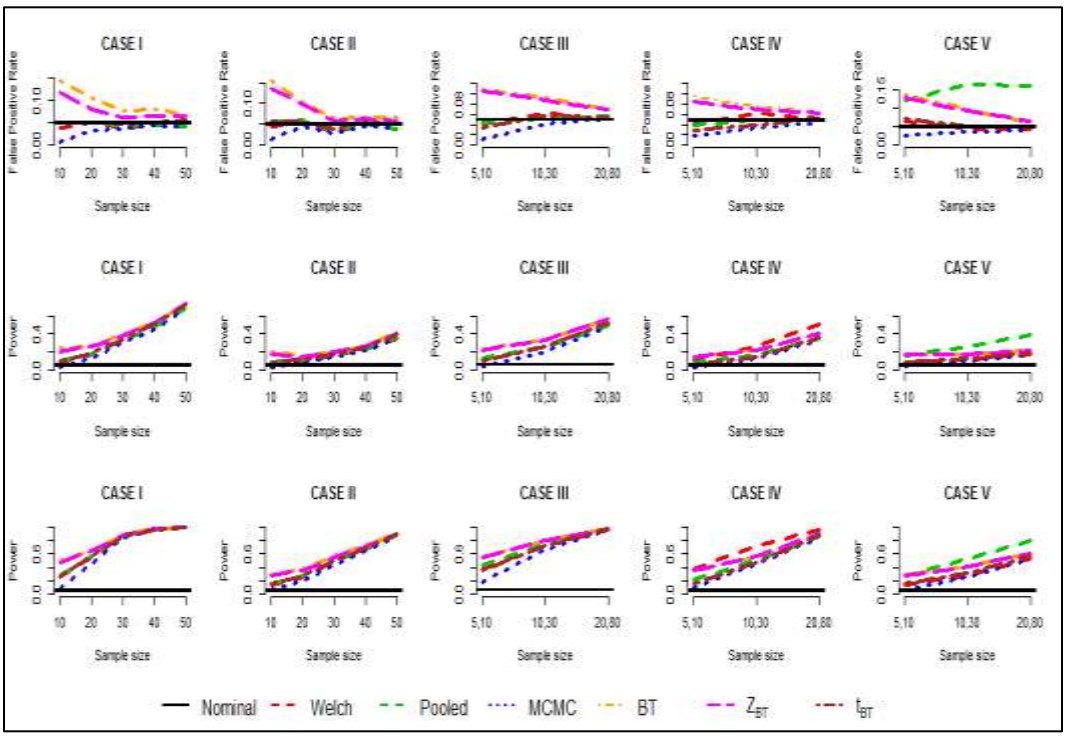

Figure 1. Plots of false positive rate and power (true positive rate) for various scenarios and sample sizes

Table 3. Proportion of power (true positive rate) based on 1000 simulations at varying sample sizes (n_1, n_2) and $\delta = 2$ for the various methods under equal and unequal variance conditions

			Bayesian			
	Frequentist		Existing	Proposed		
Sample sizes	**Welch**	**Pooled**	**MCMC**	**BT**	Z_{BT}	t_{BT}
Case I: Equal sample sizes and equal variance						
$n_1 = n_2 = 5$	0.254	0.279	0.078	0.485	0.468	0.282
$n_1 = n_2 = 10$	0.557	0.564	0.447	0.651	0.657	0.569
$n_1 = n_2 = 20$	0.869	0.871	0.846	0.889	0.891	0.874
$n_1 = n_2 = 30$	0.962	0.962	0.957	0.975	0.980	0.969
$n_1 = n_2 = 50$	0.999	0.999	0.997	0.999	0.999	0.999
Average	0.728	0.735	0.665	0.800	0.799	0.739
Case II: Equal sample sizes and unequal variance						
$n_1 = n_2 = 5$	0.135	0.154	0.049	0.297	0.274	0.152
$n_1 = n_2 = 10$	0.255	0.267	0.193	0.365	0.359	0.269
$n_1 = n_2 = 20$	0.480	0.488	0.433	0.560	0.540	0.494
$n_1 = n_2 = 30$	0.662	0.670	0.648	0.718	0.714	0.689
$n_1 = n_2 = 50$	0.880	0.882	0.872	0.907	0.908	0.900
Average	0.482	0.492	0.439	0.569	0.559	0.501
Case III: Unequal sample sizes and equal variance						
$n_1 = 5, n_2 = 10$	0.368	0.423	0.172	0.557	0.547	0.355
$n_1 = 10, n_2 = 30$	0.718	0.769	0.663	0.813	0.810	0.727
$n_1 = 20, n_2 = 80$	0.969	0.976	0.963	0.983	0.984	0.975
Average	0.685	0.723	0.599	0.784	0.780	0.686
Case IV: Unequal sample sizes and unequal variance with large variance in large sample direction						
$n_1 = 5, n_2 = 10$	0.205	0.101	0.094	0.332	0.352	0.161
$n_1 = 10, n_2 = 30$	0.506	0.250	0.444	0.576	0.572	0.475
$n_1 = 20, n_2 = 80$	0.869	0.598	0.849	0.903	0.897	0.879
Average	0.527	0.316	0.462	0.604	0.607	0.505
Case V: Unequal sample sizes and unequal variance with large variance in small sample direction						
$n_1 = 5, n_2 = 10$	0.131	0.276	0.061	0.290	0.279	0.157
$n_1 = 10, n_2 = 30$	0.280	0.530	0.248	0.410	0.402	0.316
$n_1 = 20, n_2 = 80$	0.536	0.812	0.521	0.611	0.613	0.573
Average	0.316	0.539	0.277	0.437	0.431	0.349

Conclusion

Efficient Bayesian methods were developed for testing equality of two population means from two independent samples. Among all the tests methods considered, it

can be concluded that the most suitable test method is the proposed Bayesian t_{BT} method giving its high level of validity as demonstrated by various results obtained.

References

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions, 53*, 370-418. doi: 10.1098/rstl.1763.0053

Bishop, C. M. (2005). *Neural networks for pattern recognition*. New York, NY: Oxford University Press

Bolstad, W. M. (2004). *Introduction to Bayesian statistics*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/047172212x

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician, 39*(2), 83-87. doi: 10.1080/00031305.1985.10479400

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*(3), 167-174. doi: 10.1080/00031305.1992.10475878

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics, 6*(4), 1971-1997. doi: 10.1214/12-aoas571

Efron, B. (2013). *Empirical Bayes modeling, computation, and accuracy* [Unpublished manuscript]. Stanford, CA: Department of Statistics, Stanford University.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science, 29*(2), 285-301. doi: 10.1214/13-sts455

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors – An empirical Bayes approach. *Journal of the American Statistical Association, 68*(341), 117-130. doi: 10.1080/01621459.1973.10481350

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association, 70*(350), 311-319. doi: 10.1080/01621459.1975.10479864

Efron, B., & Morris. C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *The Annals of Statistics, 4*(1), 22-32. doi: 10.1214/aos/1176343345

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York, NY: Springer Science+Business Media, LLC. doi: 10.1007/978-0-387-92407-6

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 186*(1007), 453 461. doi: 10.1098/rspa.1946.0056

Koch, K.-R. (2007). Introduction to Bayesian statistics. New York, NY: Springer Science+Business Media, LLC. doi: 10.1007/978-3-540-72726-2

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Amsterdam, the Netherlands: Elsevier.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*(2), 573-603. doi: 10.1037/a0029146

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*(4), 722-752. doi: 10.1177/1094428112457829

Kruschke, J. K., & Meredith, M. (2014). *BEST: Bayesian estimation supersedes the t-test* [R package, version 0.2.2]. Retrieved from http://CRAN.R-project.org/package=BEST

Lee, P. M. (2012). *Bayesian statistics: An introduction* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Lesaffre, E., & Lawson, A. B. (2013). *Bayesian biostatistics*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781119942412

Maritz, J. (1970). *Empirical Bayes methods*. London, UK: Methuen.

Montgomery, D. C., & Runger, G. C. (2007). *Applied statistics and probability for engineers* (4th ed.). Hoboken, NJ: Wiley.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association, 78*(381), 47-65. doi: 10.1080/01621459.1983.10477920

Murphy, K. P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution* [Unpublished manuscript] Vancouver, BC: University of British Columbia Department of Computer Science. Retrieved from https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf

Robbins, H. (1956). An empirical Bayes approach to statistics. In J. Neyman (Ed.), *Proceedings of the third Berkeley Symposium on Mathematical Statistics*

and Probability. Volume 1. Contributions to the theory of statistics (pp. 157-163). Berkeley, CA: University of California Press.

Rouder, J. N., Speckman, P. L., Sun, D., & Morey, R. D. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237. doi: 10.3758/pbr.16.2.225

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, UK: Wiley. doi: 10.1002/0470092602

Weisstein, E. W. (2003). L'Hospital's rule. Retrieved from http://mathworld.wolfram.com/LHospitalsRule.html

Wu, L. (2010). *Mixed effects models for complex data*. Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/9781420074086

Yahya, W. B., Olaniran, O. R., & Ige, S. O. (2014). On Bayesian conjugate normal linear regression and ordinary least square regression methods: A Monte Carlo study. *Ilorin Journal of Science, 1*(1). 216-227. Retrieved from http://www.unilorin.edu.ng/ejournals/index.php/ILJS/article/view/1611