MCMC for Estimating the Shortest-Length 95% Credible Intervals around Daily Ridership Estimates of Autonomous Vehicles: A Bayesian Approach

Kai-Fa Lu

***Abstract*—In Bayesian analysis of traffic data, credible intervals (CIs) are widely used for an inference on unknown parameters of interest, such as the average rate of arrival of passengers. Highest Posterior Density (HPD) sets are often used because they can guarantee the shortest length of CI estimates. However, in most of standard problems, there is no closed-form expression for exact HPD intervals, particularly for those posterior distributions with non-symmetric probability density function forms. To fill the gap, we used a Markov Chain Monte Carlo (MCMC) model such as the Random Walk Metropolis-Hastings algorithm to generate samples from the posterior distributions on an iteration basis and find the shortest length 95% CIs around mean estimates. Following a log-normal and chi distribution, MCMC simulations for daily average rates of arrival of passengers of autonomous vehicles were carried out to show their numerical behaviors. We compared the Bayesian HPD estimates with the frequentist HPD estimators and concluded that the proposed Bayesian HPD estimator brought the shortest length 95% CIs of the daily average rates of arrival of passengers.**

Keywords: Bayesian Inference; Highest Posterior Density; Markov Chain Monte Carlo; Random Walk Metropolis-Hastings; Log-Normal and Chi Distribution

I. Introduction

A credible interval (CI) is a measure of the uncertainty around the mean estimate [1, 2]. It is an interval made up of a lower and an upper limit, which indicates that the true (unknown) effect may be somewhere within the CI. The length of the interval represents the precision of the mean estimate and the shorter the length of the CI the more precise is the mean estimate [3]. The confidence (probability) level (i.e., 95%) of the CI represents the accuracy of the mean estimate. In general, the 99% CI is more accurate and wider than the 95% CI and at the same time, the 99% CI is less precise than the 95% CI. The trade-off between precision and accuracy is usually resolved by the shortest length 95% confidence level that is the most widely used [3]. In the field of transportation, precise 95% CI estimate of the daily average rate of arrival of passengers, particularly for autonomous vehicles (AVs), is vital to help decision-makers make scheduling plans to satisfy passengers' travel demand and reduce the AVs' state of zero passengers.

In the literature, two common approaches to approximate the 95% CI are the frequentist and the Bayesian [4], as shown in **Table I**. The frequentist is well suited to estimate the 95% CI for those sample distributions with no closed-form pdfs, for example, a naïve average estimator, an estimator based on density estimation [5], and normal approximation. The main idea of a naïve average estimator is to first sort out n samples from the smallest to the largest one and then find the 2.5% and 97.5% samples as the lower and upper limits of the 95% CI. In terms of density estimation, the idea is to first approximate the unobservable pdf based on the observed data and then identify the two data values corresponding to the 2.5% and 97.5% of the unobservable pdf [6, 7], respectively, as the 95% CI. The third method is to calculate the mean and standard deviation (SD) of the observed data and have $CI \approx mean \pm 1.96SD$, which is how the normal approximation works.

TABLE I
Method Summary of 95% Credible Interval Estimates

Type	Estimator	Description
Frequentist	A naïve average estimator	Sort out all samples from lowest to highest and find the 2.5% and 97.5% samples
	Density estimation	Approximate the unobservable pdfs based on the sampling distribution
	Normal approximation	$\theta \approx N(\hat{\theta}, SD)$
Bayesian	Posterior distribution	Prior knowledge × likelihood distribution
	Normal approximation	$\theta \vert y \approx N\left(\hat{\theta}, I_n(\hat{\theta})^{-1}\right)$

By contrast, Bayesian inference is a statistical approach that aims to estimate certain parameters of interest directly from the population distribution instead of estimating from the sampling distribution as the frequentist approach [3]. The Bayesian approach considers the parameters of interest as random variables with probability distributions. Thus, one of the main characteristics of the Bayesian approach is the compromise of prior evidence with the observed data, separately defined as prior knowledge and likelihood distribution in the Bayesian world [8]. The outcome of a Bayesian analysis is the posterior distribution characterized by a proportional product of prior and likelihood distributions, which is represented as $p(\theta|y)$ where θ denotes the parameters of interest and y is the observed data. As a result, the posterior distribution $p(\theta|y)$ is summarized as a measure of central tendency (e.g., mean or mode) and

Submitted for review on December 4, 2022.

K.-F. Lu is with International Center for Adaptation Planning and Design, College of Design, Construction and Planning, University of Florida, PO Box 115706, Gainesville, FL 32611-5706, USA (email: kaifa.lu@ufl.edu).

uncertainty (e.g., variance or standard deviation) [3]. Another widely used Bayesian method is a normal approximation of the posterior distribution $p(\theta|y)$ when the sample size is large enough, which does not take into account the prior [9]. This is reasonable because the shape of both likelihood and posterior distributions tends to become more and more Gaussian with an increase of sample size under standard regularity conditions [9].

In this study, our goal is to develop a Bayesian approach to estimate the shortest length 95% CI for a posterior distribution $p(\theta|y)$. In most cases, we do not have the exact expression of 95% CI for $p(\theta|y)$ in closed form and therefore, an alternative option is to use Markov Chain Monte Carlo (MCMC) [2, 10] to simulate M independent and identically distributed samples $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}$ from the posterior distribution and return the lower and upper limits of the 95% CI. Specifically, we apply the Bayesian approach into the case study of estimating the shortest length 95% CI of the daily average rate of arrival of passengers of autonomous vehicles to help make their specific scheduling plans.

The rest of this study is organized below: Section II presents theoretical framework and implementation of MCMC model to estimate the shortest-length 95% CI. Section III describes the experimental design and MCMC simulation results. Section IV summarizes the major findings.

II. Methodology And Implementation

In this section, we present theoretical concept and framework of a Bayesian approach to estimate the shortest length 95% CI and further demonstrate its MCMC implementation process.

A. Highest Posterior Density (HPD) Interval

To guarantee the shortest length 95% CI, we introduce the concept of Highest Posterior Density (HPD) set to find the most densely populated region of the posterior distribution $p(\theta|y)$ of the parameter of interest θ [3]. Then the Bayesian HPD interval is an interval with the probability of interest (95%) of the posterior distribution around the center of the distribution. This holds true the assumption that all values inside the HPD interval have a higher probability of representing the "truth" of the parameter of interest than all the values outside the HPD interval [3]. In other words, the 95% HPD interval has the shortest width and highest precision of the parameter of interest than all other 95% intervals, i.e., the 95% equal-tailed interval. **Fig. 1** further illustrates the difference between the 95% HPD interval and the 95% equal-tailed interval. In particular, there is no difference between them when the posterior distribution $p(\theta|y)$ follows a symmetric distribution.

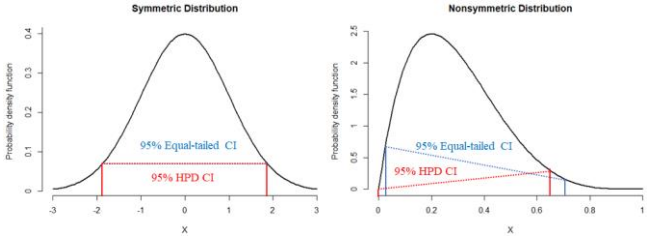

Fig. 1. Graphical representation of the 95% CI estimates for both symmetric and non-symmetric distributions.

B. Bayesian Approach for 95% CI Estimates

Assume that $y_1, y_2, \cdots, y_n$ are samples from $f_n(y|\theta)$ where $\theta \in \phi$ is an unknown scalar parameter and ϕ is the parameter space. The parameter of interest is θ that has different meanings in different fields. For instance, this study denotes θ as the daily rate of arrival of passengers of autonomous vehicles. Following the Bayesian approach, we assume that prior knowledge on θ is available based on the expert information or historical data, denoted as $\pi(\theta)$. Thus, based on a set of samples $y_1, y_2, \cdots, y_n$ from likelihood $f_n(y|\theta)$ and prior distributions $\pi(\theta)$, we have the expression of posterior distribution $f(\theta|y)$ as follows:

$$f(\theta|y) = \frac{f_n(y|\theta)\pi(\theta)}{\int f_n(y|\theta)\pi(\theta)d\theta} \quad (1)$$

As discussed above, when the posterior distribution $f(\theta|y)$ is symmetric, the 95% HPD interval is the same as the 95% equal-tailed interval. Then we need to find the 2.5% and 97.5% samples as the lower and upper limits of the shortest length 95% CI. By contrast, when the posterior distribution $f(\theta|y)$ is non-symmetric, the 95% HPD interval is narrower than the 95% equal-tailed interval and the following algorithm is designed to find the shortest length 95% CI [11]:

Step 1: Given the posterior distribution $f(\theta|y)$, all highest posterior density sets are of the form $\{\theta|f(\theta|y) \geq h\}$. The total probability of any such HPD set should be:

$$p_f(h) = \int I(f(\theta|y) \geq h) f(\theta|y)\, d\theta \quad (2)$$

As a result, computing the 95% HPD intervals is a matter of solving the equation $p_f(h) = 0.95$ and returning the root h.

Step 2: Based on the derived h from *Step 1*, we further solve the equation $f(\theta|y) = h$ to return the lower and upper limits of θ as the shortest length 95% CI.

C. Random Walk Metropolis-Hastings (RWMH) Algorithm

Since it is highly challenging to obtain the roots of the above equations analytically, an alternative approach is to numerically acquire M samples $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}$ from $f(\theta|y)$ and return the lower and upper limits of the shortest length 95% CI. The most commonly used sampling approach is the Random Walk Metropolis-Hastings (RWMH) algorithm [11]. The objective of RWMH is to draw a set of samples from the target posterior distribution, which can be specifically implemented as follows:

Step 1: Let $p(\theta|y) \propto q(\theta|y)$ be the target distribution and $\theta^{(t)}$ be the current draw from $p(\theta|y)$, and assume the proposal pdf $g(\theta)$ be symmetric such that $g(\theta^*|\theta^{(t)}) = g(\theta^{(t)}|\theta^*)$.

Step 2: Propose $\theta^* \sim g(\theta|\theta^{(t)})$.

Step 3: Accept $\theta^{(t+1)} = \theta^*$ with a probability of $\min(1, r)$ where $r = \frac{q(\theta^*|y)g(\theta^{(t)}|\theta^*)}{q(\theta^{(t)}|y)g(\theta^*|\theta^{(t)})} = \frac{q(\theta^*|y)}{q(\theta^{(t)}|y)}$; Otherwise, set $\theta^{(t+1)} = \theta^{(t)}$.

Step 4: Run M iterations across *Step 2* and *Step 3* to return M samples $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}$ from $f(\theta|y)$.

D. MCMC Model for 95% CI Estimates

Since there is no exact expression of the shortest length 95% CI for most posterior distributions, we put forward a Bayesian inferential approach with numerical MCMC simulations of the parameters of interest to return the 95% HPD interval. To be specific, the MCMC model for estimating 95% HPD interval is implemented according to the following steps:

Step 1: Compute the likelihood function $f_n(y_1, y_1, \cdots, y_n|\theta)$, assume the prior distribution $\pi(\theta)$ based on expert knowledge or historical data.

Step 2: Calculate the posterior distribution $f(\theta|y)$ using **Eq. (1)**.

Step 3: Acquire M samples $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}$ from $f(\theta|y)$ based on the RWMH algorithm.

Step 4: Return the shortest length 95% CI of the parameter θ of interest based on the M samples where the lower and upper limits of the 95% CI almost have the same probability density as discussed in the section II-B.

III. Experiments And Results

A. Examples: The Log-normal and Chi Distributions of Daily Average Rate of Arrival of Passengers of Autonomous Vehicles

In this study, we select the daily ridership data of autonomous vehicles in Lake Nona area, City of Orlando, FL as an example to demonstrate how the daily ridership changes by day of week and how to approximate the shortest length 95% CI of the daily average rate of arrival of passengers using MCMC simulations. As illustrated in **Fig. 2**, the daily ridership is mainly distributed in the range of 10-90 passengers and Saturday displays the most passengers in each day of week, followed by Sunday, Friday, Thursday, Tuesday, Wednesday, and Monday, partly reflecting the travel patterns of AV users in the study area.

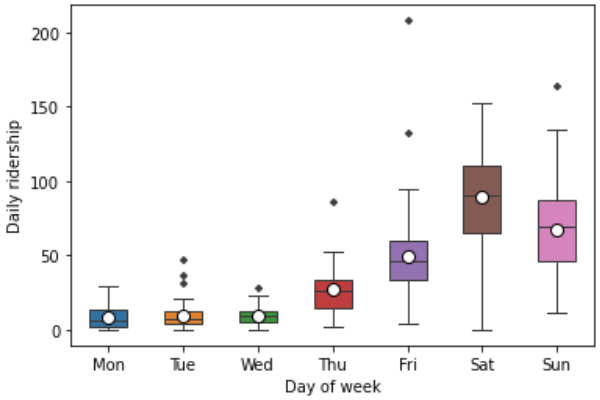

Fig. 2. Boxplot of daily passengers of autonomous vehicles in Lake Nona area, City of Orlando, FL.

Before performing MCMC simulations, it is vital to know the distributions of daily ridership in a day of week. Since the first four days (Mon-Thu) and the last three days (Fri-Sun) of a week separately display the similar distributions, we divide all data points into two groups: weekday and weekend. Inverse-gamma, log-norm, gamma, chi, and alpha distributions are chosen to fit the daily ridership data. The R-Squared (R^2) and Mean Squared Error (MSE) are used to compare the goodness-of-fit of each distribution and decide which one performs the best [12]. **Fig. 3** illustrates the rate of arrival of passengers of autonomous vehicles using the above five distribution fits. The results for the fitting curve, R^2, and MSE indicate that the log-normal distribution provides the best goodness of fit for the weekday group of the daily ridership data with $R^2 = 0.7970$ and $MSE = 0.0019$, while the chi distribution for the weekend group with $R^2 = 0.4845$ and $MSE = 0.0019$.

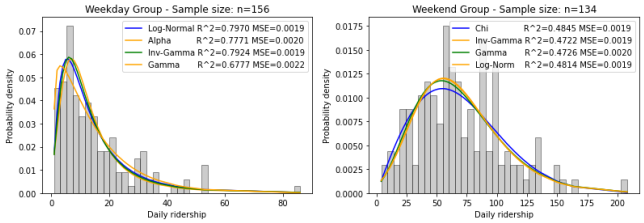

Fig. 3. Histogram and fitting plots of the daily rate of arrival of passengers for the weekday and weekend groups.

Therefore, the probability density functions of the daily rate of arrival of passengers for weekday and weekend groups are:

(1) For weekday group (log-normal distribution):

$$f(y|\mu,\sigma^2) = \frac{1}{y\sigma\sqrt{2\pi}} exp\left(-\frac{(lny-\mu)^2}{2\sigma^2}\right) \quad (3)$$

Where $f(y|\mu,\sigma^2)$ is the probability density function (pdf) of the rate of arrival of passengers for the weekday group, y is the daily ridership, and μ, σ^2 are parameters of the pdf. The mean of the rate of arrival is computed by $exp\left(\mu + \frac{\sigma^2}{2}\right)$. Here, the maximum likelihood estimates (MLE) $\hat{\mu}, \hat{\sigma}^2$ of $\mu, \sigma^2$is 2.3629 and 0.7904, respectively.

(2) For weekend group (chi distribution):

$$f(y|k,s) = \frac{y^{k-1}e^{-y^2/2s^2}}{2^{k/2-1}s^{k-1}\Gamma(k/2)}, y \geq 0 \quad (4)$$

Where $f(y|k,\sigma)$ is the pdf of rate of arrival of passengers, y is the daily ridership, and k, s is the parameters in the pdf. The mean of daily rate of arrival are computed by $\sqrt{2}\frac{s\Gamma((k+1)/2)}{\Gamma(k/2)}$. Here, the MLE $\hat{k}, \hat{s}$ of k, s is 1.9907 and 55.3774, respectively.

B. Simulations: The Shortest Length 95% CI Estimates of the Daily Average Rate of Arrival of Passengers

Based on the fitted log-normal and chi distributions of the rate of arrival for the weekday and weekend groups, we assume the priors and calculate the posterior distributions. Then we perform MCMC simulations for sampling from the posterior distributions and return the shortest length 95% CI estimates.

(1) For weekday group:

Assuming the joint prior distribution $p(\mu,\sigma^2) \propto 1/\sigma^2$, the posterior distribution is $p(\mu,\sigma^2|y) \propto p(y|\mu,\sigma^2)p(\mu,\sigma^2) \propto \frac{1}{y\sigma\sqrt{2\pi}} exp\left(-\frac{(lny-\mu)^2}{2\sigma^2}\right) 1/\sigma^2$. Then we use a normal proposal $\begin{pmatrix}\mu^* \\ \sigma^2\end{pmatrix} \sim N\left(\begin{bmatrix}\mu^{(t)} \\ \sigma_{(t)}\end{bmatrix}, S\right)$ to perform MCMC simulations where S is a tunning parameter and is estimated by $Cov(\mu,\sigma^2|y)$. For the first four days of week (Mon-Thu), we separately run the

RWMH algorithm for $n = 1000$ iterations to sample from the posterior distribution, and the results are shown in **Fig. 4**.

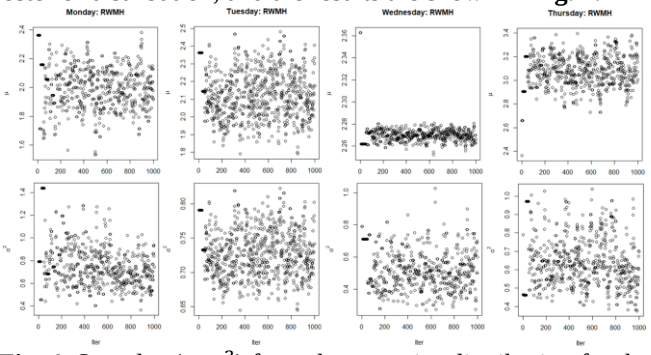

Fig. 4. Samples (μ, σ^2) from the posterior distribution for the first four days of week (Mon-Thu).

(2) For weekend group:

Assuming the joint prior distribution $p(k,s) \propto Ca^{+}(s;0,1)$, the posterior distribution is $p(k,s|y) \propto p(y|k,s)p(k,s) \propto \frac{y^{k-1}e^{-y^2/2s^2}}{2^{k/2-1}s^{k-1}\Gamma(k/2)}\frac{1}{1+s^2}$. We use a normal proposal $k^{*}\sim N\left(k^{(t)}, S\right)$ to conduct MCMC simulation where S is the tunning parameter and is estimated by $Var(k|y)$. Regarding the last three days of week (Fri-Sun), we separately run the RWMH algorithm for $n = 1000$ iterations to sample from the posterior distribution, and the results are illustrated in **Fig. 5**.

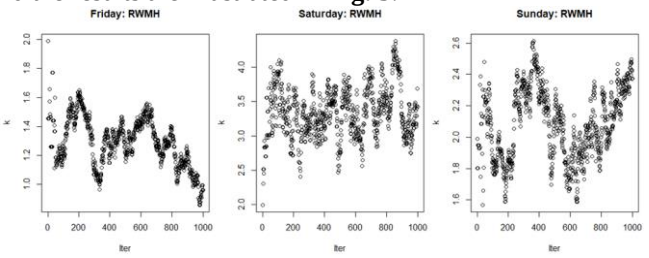

Fig. 5. Samples (k) from the posterior distribution for the last three days of week (Fri-Sun).

Based on the samples from the posterior distributions, we can compute the mean and the shortest length 95% CI estimates of the rate of arrival of passengers of autonomous vehicles using the method in the sections II-D and III-A, and the results of each day of week from Monday to Sunday are all presented in **Table II**.

TABLE II
MEAN AND 95% CI ESTIMATES OF THE RATE OF ARRIVAL OF PASSENGERS FOR EACH DAY OF WEEK

Day of week	The rate of arrival of passengers		
	Mean estimate	95% CI estimate	Length
Monday	10.9590	[7.4228,15.7699]	8.3471
Tuesday	12.1793	[9.2325,16.0327]	6.8002
Wednesday	12.5507	[11.4696,13.9418]	2.4722
Thursday	30.1085	[22.9268,40.0984]	17.1716
Friday	52.3384	[43.0149,60.1134]	17.0985
Saturday	93.8304	[81.3748,105.2867]	23.9119
Sunday	71.0079	[62.0814,79.0076]	16.9262

Finally, we compare the results of the 95% CI estimates and lengths in Table II derived from the Bayesian approach with those from the frequentist approaches (presented in **Table I** and **Fig. 2**) and conclude that there is a 95% probability that the population mean ("truth") of the rate of arrival of passengers would lie between the lower and upper limits of the 95% CI estimates. In addition, the length of the 95% CIs derived from this study is significantly shorter than that from the frequentist estimators in **Table I**. Therefore, the 95% CI estimates in **Table II** are considered as the shortest length 95% CI of the daily rate of arrival of passengers of autonomous vehicles.

IV. CONCLUSIONS

In this study, we developed a Bayesian approach to estimate the shortest length 95% credible intervals (CIs) of the daily rate of arrival of passengers of autonomous vehicles. Specifically, given the observed data, we first derived that the daily ridership of Monday-Thursday and Friday-Sunday separately followed a log-normal and chi distribution as likelihood function. Then we assumed their priors and computed the posterior distributions to perform Markov Chain Monte Carlo simulations for sampling from the posterior distributions, particularly from the Highest Posterior Density regions. Based on the simulation results, we estimated the mean and 95% CIs of rate of arrival of passengers for each day of week and the 95% CIs from the Bayesian estimator are confirmed as the shortest length and more precise than those derived from the frequentist approaches.

REFERENCES

1 Manski, C.F.: 'Credible interval estimates for official statistics with survey nonresponse', Journal of Econometrics, 2016, 191, (2), pp. 293-301
2 Lu, D., Ye, M., and Hill, M.C.: 'Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification', Water Resources Research, 2012, 48, (9)
3 Hespanhol, L., Vallio, C.S., Costa, L.M., and Saragiotto, B.T.: 'Understanding and interpreting confidence and credible intervals around effect estimates', Braz J Phys Ther, 2019, 23, (4), pp. 290-301
4 Albers, C.J., Kiers, H.A.L., van Ravenzwaaij, D., Savalei, V., and Savalei, V.: 'Credible confidence: A pragmatic view on the frequentist vs Bayesian debate', Collabra: Psychology, 2018, 4, (1)
5 Eberly, L.E., and Casella, G.: 'Estimating Bayesian credible intervals', Journal of Statistical Planning and Inference, 2003, 112, (1-2), pp. 115-132
6 Deng, H., and Wickham, H.: 'Density estimation in R', in Editor (Ed.)^(Eds.): 'Book Density estimation in R' (2011, edn.), pp.
7 Degnan, J.: 'Kernel density estimation in R', in Editor (Ed.)^(Eds.): 'Book Kernel density estimation in R' (2016, edn.), pp.
8 Bickel, D.R.: 'Empirical Bayes interval estimates that are conditionally equal to unadjusted confidence intervals or to default prior credibility intervals', Stat Appl Genet Mol Biol, 2012, 11, (3), pp. Article 7
9 De Santis, F., and Gubbiotti, S.: 'Sample size requirements for calibrated approximate credible intervals for proportions in clinical trials', Int J Environ Res Public Health, 2021, 18, (2)
10 Ji, W., and AbouRizk, S.M.: 'Credible interval estimation for fraction nonconforming: Analytical and numerical solutions', Automation in Construction, 2017, 83, pp. 56-67
11 Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D.: 'Bayesian Data Analysis (3rd Edition)' (Taylor & Francis Group, 2014. 2014)
12 Gong, H., Chen, X., Yu, L., and Wu, L.: 'An application-oriented model of passenger waiting time based on bus departure time intervals', Transportation Planning and Technology, 2016, 39, (4), pp. 424-437