STATISTICS IN TRANSITION new series, March 2023
Vol. 24, No. 2, pp. 39–57, DOI 10.59170/stattrans-2023-019
Received – 08.06.2020; accepted – 07.04.2022

Bayesian estimation of fertility rates under imperfect age reporting

Vivek Verma[1], Dilip C. Nath[2], S. N. Dwivedi[3]

ABSTRACT

This article outlines the application of the Bayesian method of parameter estimation to situations where the probability of age misreporting is high, leading to transfers of an individual from one age group to another. An essential requirement for Bayesian estimation is prior distribution, derived for both perfect and imperfect age reporting. As an alternative to the Bayesian methodology, a classical estimator based on the maximum likelihood principle has also been discussed. Here, the age misreporting probability matrix has been constructed using a performance indicator, which incorporates the relative performance of estimators based on age when reported correctly instead of misreporting. The initial guess of performance indicators can either be empirically or theoretically derived. The method has been illustrated by using data on Empowered Action Group (EAG) states of India from National Family Health Survey-3 (2005–2006) to estimate the total marital fertility rates. The present study reveals through both a simulation and real-life set-up that the Bayesian estimation method has been more promising and reliable in estimating fertility rates, even in situations where age misreporting is higher than in case of classical maximum likelihood estimates.

Key words: Fisher information, square error loss function, age-specific marital fertility rate, Bayes estimator, maximum likelihood principle.

1. Introduction

The purpose of any demographic or health sample survey is to provide information on the demographic parameters of the concerned population. In demographic studies, the age of an individual plays an important role, and misreporting leads to transfers of an individual from one age to another. Misreporting causes subjective biases due to random and systematic errors in data that influence the estimate of the population parameters. Earlier studies by Hussey and Elo (1997), Narasimhan et al. (1997) and Denic et al. (2004), Yi (2008), and Neal et al. (2012) show that age misreporting is still highly prevalent in many countries including India. As a result of misreporting, various measures and vital indicators that are age-dependent get influenced (Coale and Li (1991), Szoitysek et al. (2017)). To overcome this problem many alternative methods have been discussed by Bhat (1990), Dechter and Preston (1991), Bhat (1995), and Nwogu and Okoro (2017), which are based on the

[1]Corresponding Author. Department of Statistics, Assam University, Silchar, Assam, India. E-mail: viv_verma456@yahoo.com. ORCID: https://orcid.org/0000-0003-2537-4431.
[2]School of Applied and Pure Sciences, Royal Global University, Guwahati, Assam, India. E-mail: dilipc.nath@gmail.com.
[3]All India Institute of Medical Sciences, New Delhi, India. E-mail: dwivedi7@gmail.com. ORCID: https://orcid.org/0000-0003-4262-6143.

requirement and availability of the other related information to detect and measure these errors.

The total marital fertility rate (TMFR) is considered as one of the important measures of the overall summary of marital fertility. The measure of TMFR is basically a linear function of the number of live births to the women in each group. Therefore, the distribution for the total marital fertility rate of a population is difficult to get in an explicit form. Hence, the total fertility rate is estimated by using this linear function. The procedure for estimation and prediction of TMFR using various alternative methods was already explored in studies by Garenne et al. (2001), Yadava and Kumar (2002), Martin et al. (2011), and Pathak and Verma (2013).

Under the assumption that TMFR is an unknown but fixed quantity and there is no age missreporting, many studies have been derived and investigated this. But, in practice, TMFR is a random quantity, and can quantify the randomness specifying suitable prior distribution for it. As such, the Bayesian approach could be successfully applied for making statistical inference on TMFR.

Fertility is regarded as one of the essential demographic measures and is influenced by age misreporting. Imperfect or wrong age reporting has been remained a methodological problem (Murray et al., 2018; Singh et al., 2020; Schoumaker, 2020), and for the sake of analysis, sophisticated methodological techniques are needed to address this situation during estimation. For situations like the estimation of age-specific mortality (Bhatta and Nandram, 2013), projecting populations (Daponte et al., 1995), school completion (Barakat et al., 2021), where age-misreported, the Bayesian methodology has been found very effective to estimate the population characteristics.

Under the assumption that age was correctly reported in recent years, various Bayesian methodology-based estimates of fertility rates have been also introduced by Oh (2018), Liu and Raftery (2018), Borges (2019), and Schmertmann and Hauer(2019). But the problem of age misreporting remains unexplored. The Bayesian inference on TMFR, based on the linear function of birth in married women in each age group, has not been considered much in the literature. The study is different from the existing one as it considered limited assumptions on the structure of data and choice of prior distribution in terms of hyper-parameters values. The present study attempts to progress in the same direction of utilizing the Bayesian paradigm to estimate TMFR considering that the age has been misreported. The present study aimed to derive a prior TMFR using the same linear function following Fishers' information. As an alternative to the Bayesian methodology, a classical estimator using the maximum likelihood principle has also been discussed. The performance of the derived posterior distributions is also generalized and investigated for both perfect and imperfect age-reporting situations. Here, we hypothesized that the Bayesian estimation method might provide a more promising and reliable estimation of fertility rates, even in cases where age misreporting is higher than classical maximum likelihood estimates.

This article is organized in the following way. Section 2 provides classical and Bayesian estimates of TMFR, based on the maximum likelihood principle and the Bayesian method, respectively, under perfect age reporting. In Section 3, the procedure is generalized for imperfect estimates of TMFR when age is misreported. Section 4 illustrates the performance of the derived prior and its associated posterior distribution through numerical simulation.

Section 5 illustrates the proposed estimate through real-life data of women belonging to the childbearing age-group, *i.e.* 15-44 years, from third rounds of the National Family Health Survey(NFHS-3) of 2005-2006 in India. Section 6 provides the results and discussion. Lastly, Section 6 gives a summary and conclusion.

2. TMFR Estimation under Perfect Age Reporting

Let us consider a population of married women who are in the childbearing age group (*i.e.*, 15-44 years) at a particular period. Let X_{ai} denote a binary form of the event of ever occurrence of birth to the i^{th} women during the study period within the a^{th} childbearing age-interval, where $i = 1,2,\cdots,n_a$ and $a = 1,2,\cdots,c$. Here, c denotes the number of non-overlapping age-groups and n_a is the number of women in the a^{th} age group. The cases of twin births in a particular interval are not considered a serious issue in reality as these events are rare and found to be one out of 240 births in the database. The probability mass function (p.m.f.) of age-specific birth occurrence to a woman is given by

$$f(x_{ai}|p_a) = p_a^{x_{ai}}(1-p_a)^{1-x_{ai}}, \;\; x_{ai} = 0,1, \;\; 0 < p_a < 1, \tag{2.1}$$

where p_a denotes the probability that a child was born to a married woman belonging to a^{th} childbearing age-group, referred to as the age-specific married fertility rate (ASMFR) of mothers belonging to a^{th} age-group, for all $a = 1,2,\cdots,c$. For any age-group, say a, let $Y_a \left[= \sum_{i=1}^{n_a} X_{ai}\right]$ denote the total number of children born to n_a women belonging to that age-group, then Y_a is assumed to follow the Binomial (n_a, p_a) distribution. The estimate of probability that a child was born to a married woman in a^{th} age-group, p_a, is obtained using the observed sample, say Y_a.

2.1. Estimator of TMFR based on Maximum Likelihood Principle

Let $f(y_a|p_a)$ denote the p.m.f. of Y_a and by applying the standard maximum likelihood (ML) principle, the ML estimate of p_a, for all $a = 1,2,\cdots,c$, is obtained as

$$\hat{p}_a = \underset{p_a}{arg\;max}\; f(y_a|p_a) = \frac{y_a}{n_a} \tag{2.2}$$

and if the condition

$$\sum_{y_a=0}^{n_a} \frac{\delta}{\delta p_a} f(y_a|p_a) = 0 \tag{2.3}$$

is satisfied, then the variability explained by the estimator of p_a is given by

$$V(\hat{p}_a) \geq \{\mathscr{I}(\hat{p}_a)\}^{-1} = \frac{\hat{p}_a(1-\hat{p}_a)}{n_a}. \tag{2.4}$$

Classically, the estimate of TMFR has been obtained using the estimates of the probabilities, $p_1, p_2, \cdots p_c$, using the linear function:

$$\psi(p) = \sum_{a=1}^{c} \alpha_a \, p_a, \; \alpha_a \geq 0. \tag{2.5}$$

Fisher's information of the probability that a child born to a married woman belonging to a^{th} childbearing age-group, p_a, using standard notation, has been obtained as

$$\mathscr{I}(p_a) = \sum_{y=0}^{n_a} \left(\frac{\delta}{\delta p_a} log f(y_a|p_a) \right)^2 f(y_a|p_a) \tag{2.6}$$

and the inverse of Fisher's information matrix of age classified probabilities vector, say $p = (p_1 \,, \cdots, p_c)$, has been given by

$$\mathscr{I}^{-1}(p) = \mathscr{I}^{-1}(p_1, p_2 \cdots, p_c) = \begin{bmatrix} \frac{p_1(1-p_1)}{n_1} & 0 & \dots & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{p_c(1-p_c)}{n_c}. \end{bmatrix}$$

For the given linear function, $\psi(p)$, of TMFR in equation (2.5), the gradient of p has been obtained as

$$\begin{aligned} D_{\psi}^{T}(p) &= \begin{bmatrix} \frac{\partial \psi(p)}{\partial p_1} & \frac{\partial \ \psi(p)}{\partial p_2} & \dots & \frac{\partial \ \psi(p)}{\partial p_a} & \dots & \frac{\partial \ \psi(p)}{\partial p_c} \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_a & \dots & \alpha_c \end{bmatrix}. \end{aligned}$$

Let $v_a = \left(\mathscr{I}^{-1}(p)\right)_{aa}$=($a^{th}$ diagonal element of $\mathscr{I}^{-1}(p))) = \frac{p_a(1-p_a)}{n_a}$, then

$$\begin{aligned} D_{\psi}^{T}(p)\mathscr{I}^{-1}(p) &= \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_c \end{bmatrix} \begin{bmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & v_c \end{bmatrix} \\ &= \begin{bmatrix} \alpha_1 v_1 & \alpha_2 v_2 & \dots & \alpha_c v_c \end{bmatrix} \qquad (2.7) \\ D_{\psi}^{T}(p)\mathscr{I}^{-1}(p) D_{\psi}(p) &= \sum_{a=1}^{c} \alpha_c^2 v_a. \qquad (2.8) \end{aligned}$$

The mean and variance of $\psi(p)$ based on the ML estimates are of the form

$$\hat{\psi}(p)^M = \sum_{a=1}^{c} \alpha_a \hat{p}_a \quad (2.9)$$

$$V(\hat{\psi}(p)^M) = D_\psi^T(p)\mathscr{I}^{-1}(p)D_\psi(p) = \sum_{a=1}^{c} \alpha_a^2 \frac{\hat{p}_a(1-\hat{p}_a)}{n_a}. \quad (2.10)$$

As $Y_a \sim$ Binomial(n_a, p_a), for all $a = 1(1)c$, and $\psi(p)$ is estimated as $\hat{\psi}(p) = \sum_{a=1}^{c} \alpha_a \hat{p}_a$, from the central limit theorem, we have

$$\frac{\psi(p) - \hat{\psi}(p)}{\sqrt{D_\psi^T(p)\mathscr{I}^{-1}(p)D_\psi(p)}} \sim N(0,1). \quad (2.11)$$

The confidence interval for TMFR, $\psi(p)$, has been obtained using the above equation as

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\psi(p) - \hat{\psi}(p)}{\sqrt{D_\psi^T(p)\mathscr{I}^{-1}(p)D_\psi(p)}} \leq z_{\frac{\alpha}{2}}\right) \approx 1 - \alpha, \quad (2.12)$$

where $z_{\frac{\alpha}{2}}$ is the $(\frac{\alpha}{2})^{th}$ quantile from the top of the standard normal distribution.

2.2. Bayes Estimators of TMFR

In the previous sub-section, it has been assumed that the numbers of live births that occurred to women belong to a^{th} age-interval, say Y_a, follow the distribution denoted as $f(y_a|p_a)$. TMFR, $\psi(p)$, is defined as a linear function of unknown but fixed probabilities of having a live birth to a married woman in a^{th} age-group. But in practical situations, TMFR might be a random quantity and can model that randomness through the Bayesian approach by specifying suitable prior distribution for $\psi(p)$. To suggest a prior distribution for $\psi(p)$, a linear function of p_a's is difficult to be obtained directly. Here, an attempt has been made to derive a prior distribution for $\psi(p)$, based on the linear functions of p_a's as

Theorem-1: *Suppose $\tau(\cdot)$ defines the prior distribution for $\psi(p) = \sum_{a=1}^{c} \alpha_a p_a$, a linear function of probabilities that a child born to a married woman in a^{th} age-group, say $p_1, \cdots, p_a, \cdots p_c$, is given by*

$$\tau(p) = \tau(p_1, \cdots, p_c) \propto \left\{\sum_{a=1}^{c} \alpha_a^2 \, p_a(1-p_a)\right\}^{1/2} \prod_{a=1}^{c} p_a^{-1}(1-p_a)^{-1} \quad (2.13)$$

Proof: The proof of Theorem 1 is given in the Appendix.

The posterior distribution based on the matching prior of TMFR, $\psi(p)$, for the given sample has been obtained as

$$\begin{aligned} q(p|data) &\propto L(p|data)\ \tau(p) \\ q(p|y) &\propto \left\{\sum_{a=1}^{c}\alpha_a^2\ p_a(1-p_a)\right\}^{1/2}\prod_{a=1}^{c}p_a^{y_a-1}(1-p_a)^{n_a-y_a-1}, \end{aligned} \tag{2.14}$$

where $y_a=\sum_{i=1}^{n_a}x_{ai}$. Here, the posterior distribution, $q(p|y)$, does not have any explicit form. For this reason one has to get samples from $q(p|x)$ to get the posterior distribution of $\psi(p)$. This is done by simulating $N(=100000)$ (with burning period 10000) values from the posterior distribution as $\{p_1^{(l)},p_2^{(l)}\cdots,p_c^{(l)};\ l=1,2,\cdots,N\}$ for fixed values of α_a's, then these samples have been used for the computation of TMFR as $\psi^{(1)}(p),\psi^{(2)}(p),\cdots,\psi^{(N)}(p)$, where $\psi^{(l)}(p)=\sum_{a=1}^{c}\alpha_a p_a^{(l)}$.

The procedure of the Monte Carlo simulation technique was adopted to estimate an empirical HPD interval of $\psi(p|x)$ using the posterior samples by the following procedure:

1. $\psi^{(1)}(p),\psi^{(2)}(p),\cdots,\psi^{(N)}(p)$ are sorted $\psi_{(1)}(p)\leq\psi_{(2)}(p)\leq,\cdots,\leq\psi_{(N)}(p)$
2. Computation of the credibility interval of $100(1-\alpha)\%$ is done as

$$\Delta_l=(\psi_{(l)}(p),\psi_{(l+[(1-\alpha)N])}(p));\forall l=1,2,\cdots,N-[1-\alpha]N$$

3. The $100(1-\alpha)\%$ credible interval is denoted as Δ_l^*, and is the one which has the smallest interval width among all credible intervals.

Note : The posterior mean and variance of $\psi(p)$ can be approximated as

$$\hat{\psi}(p)^B=E(\psi(p)|x)\simeq\frac{1}{N}\sum_{l=1}^{N}\psi^{(l)}(p) \tag{2.15}$$

and

$$V(\hat{\psi}(p)^B)=V(\psi(p)|x)\simeq\frac{1}{N}\sum_{l=1}^{N}[\psi^{(l)}(p)]^2-\left[\frac{1}{N}\sum_{l=1}^{N}\psi^{(l)}(p)\right]^2. \tag{2.16}$$

3. Effect of Age Misreporting

The obtained estimates and their related discussions are enough to infer TMFR if each woman correctly reported ages. But, the works of Narasimhan et al. (1997) and Denic et al. (2004), Yi (2008), and Neal et al. (2012) have suggested that age misreporting is still highly prevalent in many countries, including India and hence the error in age reporting is inevitable. As a result, the fertility measures, including TMFR, might get highly underestimated or overestimated, which is likely to inappropriately influence related policy planning leading to poor health care and/or undue economic burden. So, it is necessary to study how robust the proposed estimates are when the prevalence of misreporting of age is high.

Let 'a' symbolize the age-interval reported by a woman and a^* denote true age-interval, where $a, a^* = 1, 2, \cdots, c$. The probability that a child born to a married woman in a^{th} reported age-group may be formulated as

$$p^*_{a^*} = \sum_a \pi_{a,a^*} \; p_a, \; a, a^* = 1, 2, \cdots, c \tag{3.1}$$

where $p^*_{a^*}$ denotes the probability that a child born to a married woman in her true age-interval a^*, and π_{a,a^*} is the probability of shifting from the true age class a^* to a due to misreporting. Equation (3.1) can be represented in a matrix form as follows:

$$p^* = \pi p, \tag{3.2}$$

where $p = (p_1, p_2, \cdots, p_c)'$ is a column vector representing probabilities of birth to a woman based on their reported ages, $p^* = (p^*_1, p^*_2, \cdots, p^*_c)'$ is a column vector of probabilities of birth to a women as per their true ages, and π is assumed to be a stochastic transition probabilities matrix was (π_{a,a^*}) of order $c \times c$. The π_{a,a^*}'s are such that

$$0 \leq \pi_{a,a^*} \leq 1, \sum_{a^*} \pi_{a,a^*} = \sum_a \pi_{a,a^*} = 1, \; \forall a, a^*$$

Based on the above probabilistic model for misreporting of age, we have the following observations:

Theorem 2: *If $\alpha_a = \alpha^*_a$, then the estimate of TMFR based on the classical procedure of estimation does not take into account the age misreporting mechanism, i.e.i.e.,*

$$\sum_{a^*=1}^{c} p^*_{a^*} = \sum_{a=1}^{c} p_a \Rightarrow \psi(p^*) = \psi(p) \tag{3.3}$$

Proof: Let the coefficients of the linear function of TMFR for both prefect age reporting and misreporting are the same *i.e.* $\alpha^*_a = \alpha_a$, for all $a, a^* = 1, 2, \cdots, c$, then

$$\begin{aligned} \sum_{a^*} p^*_{a^*} &= \sum_{a^*}\sum_a \pi_{a,a^*} \; p_a = \sum_a \left(\sum_{a^*} \pi_{a,a^*} \right) p_a = \sum_{a=1}^{c} p_a \\ \Rightarrow \psi(p^*) &= \sum_{a^*} \alpha^*_a \; p^*_{a^*} = \sum_a \alpha_a \; p_a = \psi(p). \end{aligned} \tag{3.4}$$

Under imperfect age reporting scenario, the variance of maximum likelihood estimate has been obtained after replacing p_a by p^*_a and $n_a = n_{a^*}$ as

$$V(\hat{\psi}^*(p^{*M})) = \sum_{a^*=1}^{c} \alpha^2_{a^*} \; \frac{\hat{p}^*_{a^*}(1 - \hat{p^*}_{a^*})}{n_{a^*}}. \tag{3.5}$$

and the posterior distribution of equation (2.14) will be of the form

$$h(p^*|x) \quad \propto \quad \left\{\sum_{a=1}^{c} \alpha_a^2 p_a^*(1-p_a^*)\right\}^{1/2} \prod_{a=1}^{c} (p_a^*)^{y_a-1}(1-p_a^*)^{n_a-y_a-1}. \tag{3.6}$$

In the context of the Bayesian framework, the distribution of p_i^*, for each i, is a mixture of the distributions of c independent variables p_a, $a=1,2,\cdots,c$ with mixing proportions $\pi_{i1},\pi_{i2},\cdots,\pi_{ic}$ respectively. Here again, under age misreporting scenario, the posterior, $h(p^*|x)$, does not have any explicit form and hence it is evaluated by the following Monte Carlo simulation technique.

4. Numerical Study

In this section the proposed procedures have been illustrated numerically through a simulation study. For demonstration purpose we first draw a random observation from Uniform(0,1) of size c, say p_a, for all $a=,1,2,\cdots,c$, where 'c' denotes the numbers of groups. By using the using the same p_a a random number has been generated from $Binomial(n,p_a)$, where assumed $n_1=n_2=\cdots=n_c=n$, *i.e.*, number of individuals corresponding to each group is the same and $\alpha_c=1$ for all $a=,1,2,\cdots,c$. The suggested prior and posterior distribution of $\psi(p)=\sum_{a=1}^{c}\alpha_a\ p_a,\ \alpha_a\geq 0$, defined in equations (2.13) and (2.14) not have any explicit forms, therefore, the simulation procedure discussed in Section (2.2) will be followed to characterize of $\psi(p)$.

Here we have been computed both ML and Bayesian estimators of $\hat{\psi}(p)^M$ and $\hat{\psi}(p)^B$, respectively, for both perfect and imperfect classification frameworks. Under the assumption of perfect classification of individuals into groups, the comparison among the ML and Bayes' estimators of $\psi(p)$ can be made based using their MSEs under the square risk function as

$$R_{\hat{\psi}(p)^B}(\psi(p)) \quad = \quad E(\hat{\psi}(p)^B-\psi(p))^2, \tag{4.1}$$

$$R_{\hat{\psi}(p)^M}(\psi(p)) \quad = \quad E(\hat{\psi}(p)^M-\psi(p))^2. \tag{4.2}$$

As the posterior mean is obtained by minimizing the Bayes risk under the squared error loss function, the procured Bayes estimator of an unknown parameter has often been found superior to the corresponding ML estimator concerning MSE. It is to be emphasized that the estimator based on ML principal neither depends on any prior distribution for the parameter nor it requires any particular loss function. Thus, in such a situation, the comparison among the ML and Bayes estimator ought to be made so that the criteria do not depend on the nature of prior information regarding unknown parameters. As the MSE of an estimator is also considered risk under squared error loss, it has been treated as a risk function for comparison purposes. The comparison is done by calculating the estimated relative risk of

Bayes estimators concerning $\hat{\psi}(p)^M$ and is defined by

$$\hat{\theta}_{\hat{\psi}(p)^B} = \frac{\hat{R}(\hat{\psi}(p)^M)}{\hat{R}(\hat{\psi}(p)^B)} \tag{4.3}$$

For generalization of the suggested methodology in the imperfect classification of group situation, and comparison of ML and Bayesian technique, the misclassification matrix,π, is known. For the demonstration purpose we have considered the particular form of π, misclassification transition probabilities matrix *i.e.* (π_{a,a^*}) of order $c \times c$ as,

$$\pi = \begin{pmatrix} \rho & \delta & \delta & \dots & \delta \\ & \rho & \delta & \dots & \delta \\ & & \rho & \dots & \delta \\ & & & \ddots & \vdots \\ & & & & \rho \end{pmatrix}, \quad 0 < \rho < 1, \quad \delta = \frac{1-\rho}{s-1}, \tag{4.4}$$

where ρ denotes the probability of an accurately classified group and δ denotes the inaccuracy, which has been assumed as equally distributed across the remaining groups. Here '$\rho = 1$' corresponds to the case of perfect classification. To illustrate the performance of both ML and Bayesian estimators under perfect and imperfect classification frameworks, for different choices of group size $c \in \{3,5,7\}$, the number of observation in each group $n = \{50,100\}$ and $\rho = 0.8, 0.9, 1.0$, estimates $(\hat{\psi}(p)^M, \hat{\psi}(p)^B)$, 95 % confidence and credible intervals and relative risk of Bayes estimators $\hat{\theta}_{\hat{\psi}(p)^B}$ have been obtained. Based on the simulation of 100000 times the obtained results have been depicted in Table 2.

5. Application to Real life data

In this section, an illustration of the proposed procedure using real-life data on Indian married women has been discussed. For this study, we took the data set for the third round of the National Family Health Survey-3 (NFHS-3) for the years 2005-06 from the Measure DHS Demographic and Health Surveys (DHS). DHS provides a nationally representative state survey that helps estimate various key indicators of fertility, infant mortality, family planning practice, maternal and child care, and access to mother and child services (NFHS-III(2005-2006)). NFHS-3 is conducted by the Ministry of Health and Family Welfare (MoHFW), Government of India, and managed by the International Institute of Population Sciences (IIPS), Mumbai, covering 29 states and 7 Union Territories of India (NFHS-III(2005-2006)). Here, the samples of NFHS-3 are treated as our population of interest and the study population comprised of the women residing at the Empowered Action Groups (EAG) states of India *viz.* (a) Bihar ($n = 3818$) (b) Uttaranchal ($n = 2953$) (c) Chhatisgarh ($n = 3810$) (d) Jharkhand ($n = 2983$) (d) Orissa ($n = 4540$) (e) Rajasthan ($n = 3892$) (f) Madhya Pradesh ($n = 6427$) and (g) Uttar Pradesh ($n = 12183$), which are considered as socio-economically backward and have high fertility rates compare to other states.

In order to estimate TMFR, defined in equation (2.5), corresponding to each of the selected Indian states under both ML and Bayesian methods, here married women belonging to childbearing age interval (15-44 years) have been grouped into six ($c = 6$) non-overlapping equal subgroups of 5 year interval *viz.* 15-19,$\cdots$, 40-44. Age interval 45-49 could not be considered due to the lack of a sufficient number of women. Further, the information on the birth status in the last year of the survey has been considered a study period. Corresponding to each selected woman, information regarding their age and whether any birth occurred or not during the study period has been collected.

The problem of estimation of TMFR for the situation where age has been misreported as in equation (3.1), through ML and Bayesian technique, is possible only when the π matrix is known. The present study suggested two different methods to obtain π matrix.

Firstly, we considered the particular form of π, misclassification transition probabilities matrix presuming that the correct age reporting was done at five different levels *viz.* $\rho = 100\%, 90\%,$ and 80% in equation (4.4), where '$\rho = 100\%$' corresponds to the case of perfect or correct age reporting. The impact of perfect or imperfect age reporting has been presented as Table 2 and change in pattern of p_a has been depicted in Figure 1.

Alternatively: The π matrix can be simulated empirically by using independent observation from the same underlying population corresponding to each c age class. The π_{a,a^*} has been estimated as the proportion of women out of total women whose reported age belonging to the age-interval a belongs to the true age-interval a^* in the set of c class. Here, the true age of the mother is determined using the other additional reported information *viz.* age at first marriage(A_M), duration of Gauna (return marriage) if performed(A_G), marriage to first birth duration(A_{FB}) and age of the first child(A_C). The difference among the reported age(A_R) and age calculated using above information *i.e.* $A_R - (A_M + A_G + A_{FB} + A_C)$, has been considered as error in reporting. The empirical estimates of π_{a,a^*}, for all a^*, a, have been obtained by repeatedly observing the set of values for sufficiently large number of times, and, finally computed the proportion of cases where the age a^* has been reported as a. The approximate π^E matrix following this procedure based on the available information can be estimated empirically using the whole population. Obtained estimates of TMFR $(\psi(p))$ under different model assumptions are presented as Table 3 and 4. All computations are carried out using Statistical Analysis System (SAS) package, University edition and R package (version-3.4.0).

6. Findings and Discussion

Table 2 depicts the results under both perfect and imperfect age misreporting situations, where Bayesian estimates $\psi(p)^B$ are not only found to be more reliable but also always provide compact and efficient credible interval as compared to $\psi(p)^M$. It also shows that the Bayesian methodology is capable enough to capture the change in estimates due to misclassification in terms of estimation with better accuracy.

The performance of the estimation procedures, as far as TMFR is concerned, has been presented in Tables 3 and 4. Overall, the results indicate that the Bayes estimates of TMFR

Table 1: Empirical estimate of the age misreporting error probability matrix for India

Reported Age-interval (a)	True Age-interval (a^*) 15-19	20-24	25-29	30-34	35-39	40-44
15-19	0.992	0.008	0.000	0.000	0.000	0.000
20-24	0.067	0.923	0.010	0.000	0.000	0.000
25-29	0.002	0.113	0.868	0.017	0.000	0.000
30-34	0.000	0.003	0.145	0.832	0.021	0.000
35-39	0.000	0.000	0.005	0.182	0.808	0.004
40-44	0.000	0.000	0.001	0.006	0.173	0.819

for all EAG states have shown a decreasing trend until correct age reporting decreases to 90% and starts increasing after that. As theoretically shown, the ML estimates have shown no impact due to misreporting. Table 3 shows that the Bayes estimates of TMFR under both perfect and imperfect age reporting have been found more precise *i.e.* with lesser risk than those of the ML estimates. The 95% credible intervals based on the Bayes estimators have been found narrower than those obtained using the ML estimates. It implies that the proposed Bayes estimators based on the suggested prior provide estimates more precisely and accurately address the issues of misreporting while estimating TMFR. Among the ML and Bayes estimates generated by using empirically estimated transition probabilities matrix, π^E, in Table 4, results also reveal that Bayes' estimates of TMFR of selected Indian states are comparatively more precise (with narrower credible intervals). It is also to be emphasized that Bayes' estimates of TMFR (Table 3) under the presumption that the age has been perfectly reported ($\rho = 1$), corresponding to each Indian state, have been found close to the values of TMFR obtained during 2005-06 *viz.* Uttaranchal (4.0), Uttar Pradesh (5.7), Bihar (5.2), Jharkhand (4.9), Orissa (4.4), Chhattisgarh (4.9), Madhya Pradesh (4.9) and Rajasthan (4.6).

Since the probabilities that a child born to a married woman belonging to a^{th} childbearing age-group, p_a, are sensitive towards age reporting, they are affected immensely due to misreporting. Figure 1 depicts the estimates of p_a based on the Bayesian principle, which shows a significant variation in the pattern in Bayes' estimates of p_a with a change in levels of the inaccuracy of age reporting corresponding to each Indian state. No systematic pattern has been observed in the obtained estimates, as all states are demographically distinct. Still, variation in Bayes' estimates of TMFR is expected with a change in levels of misreporting. The degree of distortions in the Bayes estimates of p_a at age a has been noticed comparatively higher than those obtained using the principle of maximum likelihood. In particular, as the proportion of misreporting increases from 5% to 20%, the Bayes estimates of p_a are getting more distorted.

The primary reason for accepting the suggested Bayesian estimates of fertility rates is that the derived prior distribution is subjective and empirical. Here, we have also discussed its formalization and update for imperfect age reporting situations, which demographers or policy-makers routinely experience. Further, we compared the proposed Bayesian estimates

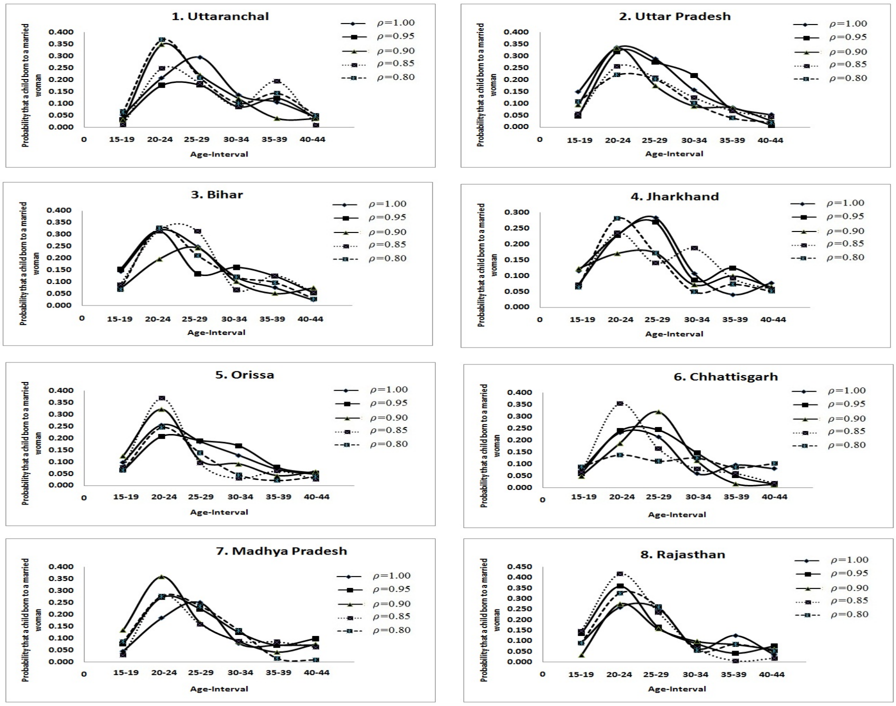

Figure 1: Bayes estimates of the probability that a child born to a married woman belonging to a^{th} childbearing age-group (p_a) in EAG States of India, when there is perfect age reporting($\rho = 1$) and misreporting lies in 5%- 20%.

with the classical through relative risk, and an attempt has been made to generalize this comparison for the imperfect age-reporting situations. As the likelihood function contains the observation and combining the Bayesian approach with the classical model, the Bayesian approach can incorporate more realistic conditions and data into the estimation.

Table 2: Simulation results of relative risks and their 95% confidence and credible intervals under perfect classification($\rho = 1$) and under misclassification of 0%, 10% and 20%

Class Size	n	ρ	ML			Bayes'			RR
			$\hat{\psi}(p^*)^M$	95% Interval		$\hat{\psi}(p^*)^B$	95% Interval		$(\hat{\theta}_{\hat{\psi}(p^*)^B})$
3	50	1.00	1.960	1.750	2.170	1.954	1.885	2.023	11
		0.90	1.960	1.745	2.175	1.957	1.880	2.035	6
		0.80	1.960	1.740	2.180	1.940	1.856	2.024	6.5
3	150	1.00	1.947	1.833	2.060	1.933	1.893	1.974	3
		0.90	1.947	1.828	2.066	1.949	1.906	1.993	4
		0.80	1.947	1.823	2.070	1.938	1.894	1.983	4
5	50	1.00	3.340	3.105	3.575	3.324	3.275	3.374	14
		0.90	3.340	3.090	3.590	3.261	3.208	3.313	16
		0.80	3.340	3.079	3.601	3.351	3.289	3.413	18
5	150	1.00	3.447	3.303	3.591	3.449	3.415	3.483	5
		0.90	3.447	3.297	3.596	3.440	3.412	3.469	6
		0.80	3.447	3.293	3.600	3.428	3.394	3.462	6
7	50	1.00	4.540	4.289	4.791	4.217	4.179	4.255	16
		0.90	4.540	4.264	4.816	4.565	4.549	4.581	20
		0.80	4.540	4.244	4.836	4.392	4.357	4.427	23
7	150	1.00	4.480	4.336	4.624	4.230	4.193	4.267	5
		0.95	4.480	4.328	4.632	4.346	4.316	4.375	6
		0.90	4.480	4.321	4.639	4.286	4.250	4.322	7
		0.80	4.480	4.309	4.651	4.484	4.467	4.501	8

$\hat{\theta}^*{}_{\hat{\psi}(p)^B} = \frac{\hat{R}(\hat{\psi}(p^*)^M)}{\hat{R}(\hat{\psi}(p^*)^B)}$

7. Conclusion

In the present article, we have derived a prior for total marital fertility rate using Fishers' information and its related posterior distributions under perfect age reporting and generalized for misreporting scenarios. Since the posterior distributions of TMFR (in the Bayesian paradigm) are complicated, a direct comparison with the maximum likelihood principle (in connection with classical framework) is not straightforward. Thus, through simulation, a comparison among classical and Bayes' estimates of TMFR is presented. Both the simulated and real-life based results show that the suggested Bayesian estimators of $\psi(p)$ and TMFR lead to population parameters more closely than classical ML estimators and are much more precise than maximum likelihood estimates, even in imperfect scenarios. As evident from the obtained results, even with inaccuracy in age reporting, the Bayesian technique has been found most promising for estimating TMFR, and obtained Bayes' estimates are more precise and reliable than those obtained using the maximum likelihood procedure.

To conclude, apart from the estimation of transition probabilities, the Bayesian technique has been found to be more useful in estimating the pattern of fertility rates even in situations where there is inaccuracy in age reporting.

Table 3: Estimates of TMFR ($\psi(p)$) in EAG States of India, when there is perfect age reporting($\rho = 1$) and misreporting is 10%, and 15%

State	n	ML			Bayes'			$\hat{\theta}_{\hat{\psi}(p^*)^B}$
		$\hat{\psi}(p^*)^M$	95% Interval		$\hat{\psi}(p^*)^B$	95% Interval		
				$\rho = 1.00$				
Uttaranchal	2953	3.08	2.76	3.39	4.17	4.00	4.34	3.7
Uttar Pr.	12183	3.54	3.37	3.71	5.30	5.24	5.37	7.0
Bihar	3818	3.77	3.45	4.08	4.66	4.57	4.75	6.5
Jharkhand	2983	3.26	2.93	3.58	4.24	4.14	4.34	6.8
Orissa	4540	2.38	2.15	2.60	3.90	3.82	3.99	6.5
Chhattisgarh	3810	2.69	2.43	2.96	3.79	3.71	3.86	3.6
Madhya Pr.	6427	2.92	2.71	3.14	3.50	3.36	3.63	2.4
Rajasthan	3892	3.40	3.11	3.69	4.12	3.98	4.26	4.4
				$\rho = 0.90$				
Uttaranchal	2953	3.08	2.76	3.40	3.95	3.88	4.03	6.8
Uttar Pr.	12183	3.54	3.37	3.72	3.97	3.89	4.06	4.0
Bihar	3818	3.77	3.45	4.09	4.20	4.13	4.26	6.8
Jharkhand	2983	3.26	2.93	3.58	3.49	3.41	3.58	7.0
Orissa	4540	2.38	2.15	2.61	3.72	3.60	3.84	3.5
Chhattisgarh	3810	2.69	2.42	2.96	3.48	3.44	3.51	6.3
Madhya Pr.	6427	2.92	2.71	3.14	4.29	4.29	4.52	4.0
Rajasthan	3892	3.40	3.10	3.70	3.51	3.41	3.62	7.7
				$\rho = 0.85$				
Uttaranchal	2953	3.08	2.75	3.40	3.68	3.59	3.76	9.3
Uttar Pr.	12183	3.54	3.37	3.72	3.78	3.63	3.93	1.3
Bihar	3818	3.77	3.44	4.09	4.86	4.80	4.92	5.4
Jharkhand	2983	3.26	2.92	3.59	3.91	3.80	4.03	9.7
Orissa	4540	2.38	2.14	2.61	3.28	3.17	3.38	4.7
Chhattisgarh	3810	2.69	2.42	2.96	3.72	3.64	3.79	9.5
Madhya Pr.	6427	2.92	2.71	3.14	3.49	3.39	3.59	4.0
Rajasthan	3892	3.40	3.10	3.70	4.44	4.44	4.75	4.6

Table 4: Estimates of TMFR ($\psi(p)$) in EAG States of India, under imperfect age-reporting using empirical π^E.

State	ML			Bayes'			$\hat{\theta}_{\hat{\psi}(p)^B}$
	$\hat{\psi}(p)^M$	95% Interval		$\hat{\psi}(p)^B$	95% Interval		
Uttaranchal	3.187	2.861	3.514	4.108	3.995	4.221	9.3
Uttar Pardesh	3.656	3.482	3.83	4.513	4.37	4.656	1.6
Bihar	3.875	3.552	4.198	3.758	3.649	3.868	9.0
Jharkhand	3.369	3.036	3.701	4.243	4.145	4.341	9.7
Orissa	2.464	2.232	2.697	3.697	3.587	3.808	4.7
Chhattisgarh	2.793	2.52	3.066	3.988	3.932	4.043	6.3
Madhya Pradesh	3.027	2.808	3.246	3.742	3.606	3.878	2.4
Rajasthan	3.513	3.214	3.812	4.022	3.858	4.187	3.3

Acknowledgements

The authors would like to thank all anonymous reviewers and the editorial team for their valuable comments and suggestions that helped us improve the article's quality.

References

Barakat, B. F., Dharamshi, A., Alkema, L. and Antoninis, M., (2021). Adjusted Bayesian Completion Rates (ABC) Estimation (No. at368). Center for Open Science.

Bhat, P. M., (1990). Estimating transition probabilities of age misstatement. *Demography*, 27(1), pp. 149–163.

Bhat, P. N., (1995). Age misreporting and its impact on adult mortality estimates in South Asia. *Demography India*, 24(1), pp. 59–80.

Bhatta, D., Nandram, B., (2013). A Bayesian adjustment of the HP law via a switching nonlinear regression model. *Journal of Data Science*, 11(1), pp. 85–108.

Borges, G. M., (2019). A Bayesian framework for estimating fertility from multiple data sources. *Anais*, pp. 1–8.

Coale, A. J., Li, S., (1991). The effect of age misreporting in China on the calculation of mortality rates at very high ages. *Demography*, 28(2), pp. 293–301.

Coal Ansley J., Trussell T. James, (1974). Model Fertility Schedules: Variation in the Age Structure of Childbearing in Human Populations. *Population Index*. 40(2), pp. 191–228.

Daponte, B. O., Kadane, J. B. and Wolfson, L. J., (1997). Bayesian demography: projecting the Iraqi Kurdish population, 1977-1990. *Journal of the American Statistical Association*, 92(440), pp. 1256–1267.

Datta, G. S., Ghosh, J. K., (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika*, 82(1), pp. 37–45.

Dechter, A. R., Preston, S. H., (1991). Age misreporting and its effects on adult mortality estimates in Latin America. *Population Bulletin of the United Nations*, (31-32), pp. 1–16.

Denic, S., Saadi, H. and Khatib, F., (2004). Quality of age data in patients from developing countries. *Journal of Public Health*, 26(2), 168–171.

Garenne, M., Tollman, S., Kahn, K., Collins, T. and Ngwenya, S., (2001). Understanding marital and premarital fertility in rural South Africa. *Journal of Southern African Studies*, 27(2), pp. 277–290.

Hussey, J. M., Elo, I. T., (1997). Cause specific mortality among older African Americans: Correlates and consequences of age misreporting. *Social Biology*, 44(3-4), pp. 227–246.

Liu, P., Raftery, A. E., (2018). Accounting for Uncertainty About Past Values In Probabilistic Projections of the Total Fertility Rate for All Countries. *arXiv preprint arXiv*: 1806.01513.

Martin, J. A., Hamilton, B. E., Ventura, S. J., Osterman, M. J., Kirmeyer, S., Mathews, T. J. and Wilson, E. C., (2011). Births: final data for 2009. National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 60(1), pp. 1–70.

Murray, C. J., Callender, C. S., Kulikoff, X. R., Srinivasan, V., Abate, D., Abate, K. H., ... and Bililign, N., (2018). Population and fertility by age and sex for 195 countries and territories, 1950?2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159), pp. 1995–2051.

NFHS-III. (2005-2006). National Family Health Survey, International Institute for Population Sciences, Bombay.

Narasimhan, R. L., Retherford, R. D., Mishra, V. K., Arnold, F. and Roy, T. K., (1997). Comparison of Fertility Estimates from India's Sample Registration System and National Family Health Survey.

Neal, S., Matthews, Z., Frost, M., Fogstad, H., Camacho, A. V. and Laski, L., (2012). Childbearing in adolescents aged 12–15 years in low resource countries: a neglected issue. New estimates from demographic and household surveys in 42 countries. *Acta obstetricia et gynecologica Scandinavica*, 91(9), pp. 1114–1118.

Nwogu, E. C., Okoro, C., (2017). Adjustment of Nigeria population censuses using mathematical methods. *Canadian Studies in Population*, 44(3-4), pp. 149–64.

Oh, J., (2018). A comparison and prediction of total fertility rate using parametric, nonparametric, and Bayesian model. *The Korean Journal of Applied Statistics*, 31(6), pp. 677–692.

Pullum, T. W., (2006). An assessment of age and date reporting in the DHS Surveys 1985-2003.

Pathak, P., Verma, V., (2013). Projection of Indian Population by Using Leslie Matrix with Changing Age Specific Mortality Rate, Age Specific Fertility Rate and Age Specific Marital Fertility Rate. In Advances in Growth Curve Models. *Springer, New York, NY*, pp. 227–240.

Schoumaker, B., (2020). Fertility estimates from full birth histories and HDSS. In United Nations Expert Group Meeting on Methods for the World Population Prospects 2021 and Beyond.

Schmertmann, C. P., Hauer, M. E., (2019). Bayesian estimation of total fertility from a population's age sex structure. *Statistical Modelling*, 19(3), pp. 225–247.

Singh, B. P., Singh, N. and Singh, S., (2020). Estimation of total fertility rate: an indirect approach using auxiliary information. *Journal of the Social Sciences*, 48(3), pp. 789–798.

Szoitysek, M., Poniat, R. and Gruber, S., (2017). Age heaping patterns in Mosaic data. Historical Methods: A *Journal of Quantitative and Interdisciplinary History*, pp. 1–26.

Yadava R. C., Kumar A., (2002). On an Indirect Estimation of Total Fertility Rate from Open Birth Interval Data, *Demography India*, 31(2), pp. 211–222.

Yi, Z., (2008). Reliability of age reporting among the Chinese oldest-old in the CLHLS datasets. In Healthy Longevity in China. *Springer, Dordrecht*, pp. 61–78.

Appendix

Proof of Theorem 1: Let X_{ai} be a binary variable, denoting the birth status of i^{th} woman belonging to a^{th} age-group and (p_a) be the probability that a child birth occurred to a married woman in the same age-group, for all $i = 1,2,\cdots,n_a$ and $a = 1,2,\cdots,c$. Let $\psi(p) = \sum_{a=1}^{c} \alpha_a \ p_a$ be the linear function of probabilities, $p_1,\cdots p_c$, $\mathscr{I}^{-1}(p)$ be the inverse of Fisher's information matrix and $D_\psi^T(p)$ denote the gradient of $\psi(p)$, where $p = \{p_1,\cdots,p_a,\cdots p_c\}$. Let us consider

$$\begin{aligned} \gamma^T(p) &= \frac{D_\psi^T(p)\mathscr{I}^{-1}(p)}{\sqrt{D_\psi^T(p)\mathscr{I}^{-1}(p)D_\psi(p)}} = \left[\frac{\frac{\alpha_1 p_1(1-p_1)}{n_1}}{\sqrt{\sum_{a=1}^{c} \frac{\alpha_a^2 p_a(1-p_a)}{n_a}}} \quad \cdots \quad \frac{\frac{\alpha_c p_c(1-p_c)}{n_c}}{\sqrt{\sum_{a=1}^{c} \frac{\alpha_a^2 p_a(1-p_a)}{n_a}}} \right] \\ &= \left[\gamma_1(p) \quad \gamma_2(p) \quad \ldots \quad \gamma_c(p) \right], \end{aligned} \tag{7.1}$$

where $\gamma_a(p) = \frac{\alpha_a p_a(1-p_a)}{n_a} \left(\sqrt{\sum_{a=1}^{c} \frac{\alpha_a^2 p_a(1-p_a)}{n_a}} \right)^{-1}$. In the context of deriving a prior distribution of a parameter, Dutta and Ghose (1995) has suggested the criteria that must be satisfied to establish the posterior distribution for a parametric function under which $\sum_{a=1}^{c} \frac{\partial}{\partial p_a} \gamma_a(p)\tau(p) = 0$.

Let

$$\tau(p) = \left(\sum_{a=1}^{c} \frac{\alpha_a^2 p_a(1-p_a)}{n_a} \right)^{1/2} \prod_{a=1}^{c} p_a^{-1}(1-p_a)^{-1}$$

then

$$\begin{aligned} \gamma_1(p)\tau(p) &= \frac{\frac{\alpha_1 p_1(1-p_1)}{n_1}}{\sqrt{\sum_{a=1}^{c} \frac{\alpha_a^2 p_a(1-p_a)}{n_a}}} \left(\sum_{a=1}^{c} \frac{\alpha_a^2 p_a(1-p_a)}{n_a} \right)^{1/2} \prod_{a=1}^{c} p_a^{-1}(1-p_a)^{-1} \\ &= \frac{\alpha_1}{n_1} \prod_{a\neq 1}^{s} p_a^{-1}(1-p_a)^{-1} \Rightarrow \frac{\partial}{\partial p_1}\gamma_1(p)\tau(p) = 0 \end{aligned} \tag{7.2}$$

and

$$\begin{aligned} \gamma_j(p)\tau(p) &= \frac{\frac{\alpha_j p_j(1-p_j)}{n_j}}{\sqrt{\sum_{a=1}^{c} \frac{\alpha_a^2 p_a(1-p_a)}{n_a}}} \left(\sum_{a=1}^{c} \frac{\alpha_a^2 p_a(1-p_a)}{n_a} \right)^{1/2} \prod_{a=1}^{c} p_a^{-1}(1-p_a)^{-1} \\ &= \frac{\alpha_j}{n_j} \prod_{a\neq j=1}^{c} p_a^{-1}(1-p_a)^{-1} \Rightarrow \frac{\partial}{\partial p_j}\gamma_j(p)q(p) = 0 \end{aligned} \tag{7.3}$$

From the above equations (7.2) and (7.3) we have

$$\sum_{a=1}^{c} \frac{\partial}{\partial p_a} \, \gamma_a(p)\tau(p) = 0,$$

which satisfied the condition required to be a prior distribution, $\tau(p)$, of a parameter. Therefore,

$$\tau(p) \propto \left\{ \sum_{a=1}^{c} \alpha_a^2 p_a(1-p_a) \right\}^{1/2} \prod_{a=1}^{c} p_a^{-1}(1-p_a)^{-1} \; ; \; 0 < p_a < 1$$

and hence we get the required proof.