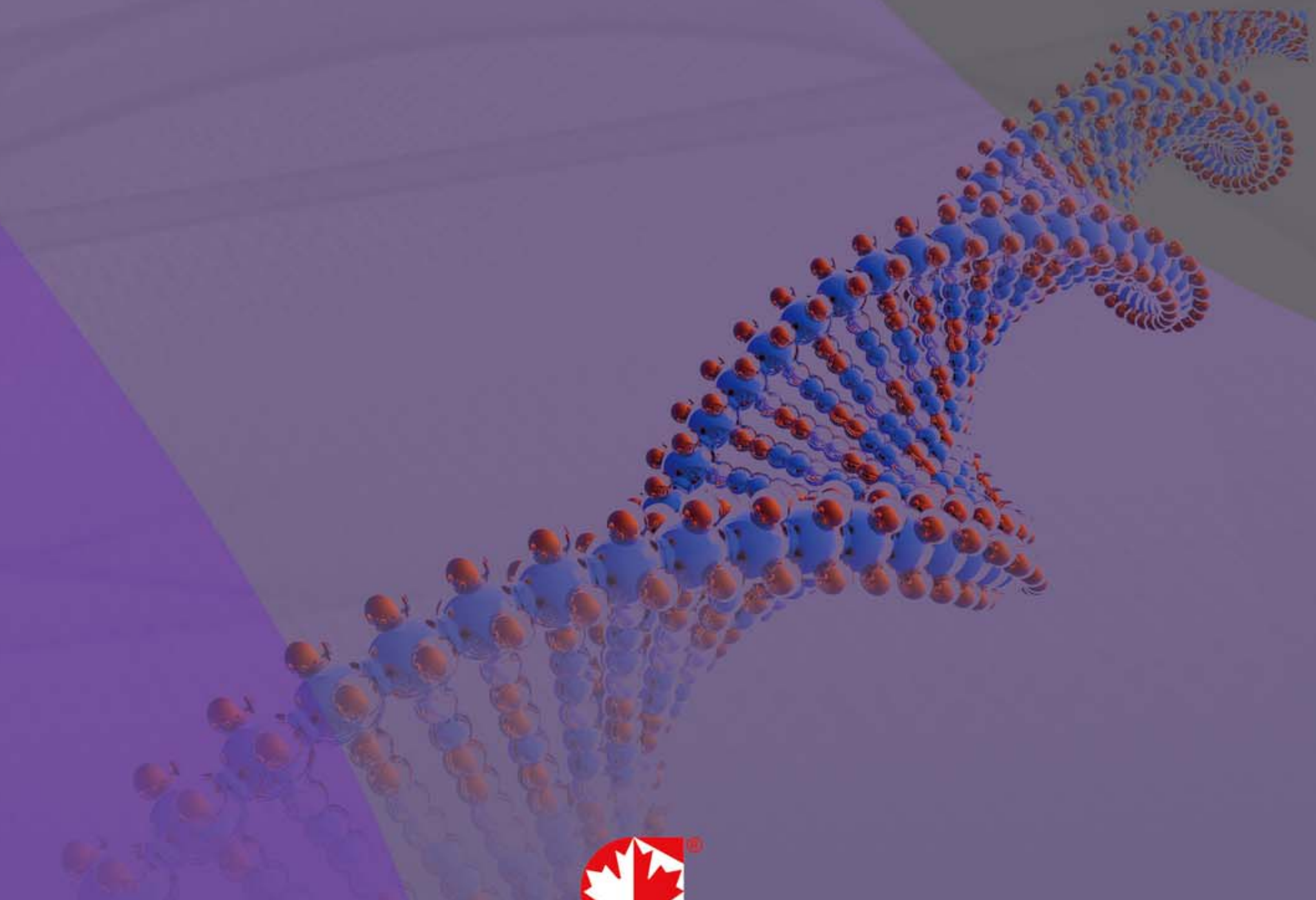
OA BOOKS

BIOSTATISTICAL ASPECTS OF POPULATION HEALTH

Edited By Ajay Pandey



The Canadian Center of Science and Education

BIOSTATISTICAL ASPECTS OF POPULATION HEALTH

Edited by Ajay Pandey

DOI:10.5539/9780978430160

ISBN: 978-0-9784301-6-0 (print)

ISBN: 978-0-9784301-7-7 (ebook)

Published by Canadian Center of Science and Education
1595 Sixteenth Ave, Suite 301
Richmond Hill, Ontario, L4B 3N9
Canada
Tel:1-416-642-2606

BIOSTATISTICAL ASPECTS OF POPULATION HEALTH

Edited by

Ajay Pandey Ph.D.

About the Editor

Dr. Ajay Pandey is an expert in the field of Public Health, Demography and Statistics. He has been contributing to the discipline for more than 20 years. His previous work association has been with India's highest policy making body i.e. the National Commission on Population (NCP), Government of India (GOI). NCP was constituted in the year 2000 under the Chairmanship on the Prime Minister of India and Deputy Chairman, Planning Commission (currently known as Niti Aayog) as its Vice Chairman. Dr. Pandey was involved in Mapping and Ranking of all the districts in India on the basis of key socio economic and demographic indicators that led to initiation of area specific plans for developmental planning. He also gained substantial experience working with community during his days with Futures Group International in India. It is at Futures Group that the public private model of Health Care service delivery was designed and implemented in the State of Uttar Pradesh. The PPP model of health care service delivery provided the policy options for expanding the reach and quality care services to the poor and marginalized sections of the society. Dr. Pandey is recipient of number of national and international fellowships and has chaired sessions at demographic conferences besides hosting them both in India and abroad. He has chaired a session during European Population Conference held at Stockholm University, Sweden in 2012. He has been part of the multi disciplinary team of international consultants documenting the efforts under USAID Innovation in Family Planning Services-1 project in Uttar Pradesh/India. He is currently working as Assistant Director at Population Research Centre at University of Lucknow in India and has contributed to three Books.

1. IDEAS, INSIGHTS, AND INNOVATIONS: Achievements and Lessons Learned from the Innovations in Family Planning Services (IFPS) Project, 1992–2004. 2006. ISBN: 1-59560-008-6
2. India's Billion Plus People: 2001 Census Highlights, Methodology and Media Coverage. By Ashish Bose. Assisted by Anita Haldhar, M S Bist & Ajay Pandey. 2001, ISBN: 8176462276, 9788176462273
3. "Demographic Scenario in India; Proceedings of the State Population Councils/commission", published by the National Commission on Population, Government of India. January, 2003. The book was laid on the table in the House of Parliament. 2003, National Commission on Population, GOI.

Contents

Preface

Bio-statistical aspect of Population Health book is important from the perspective of development of newer estimation procedures & analytical techniques that are able to address complex problems public health phenomenon simplistically. The book is an attempt in this direction as it not only suggests new estimation procedures but also demonstrates empirically, the estimates that are superior compared to earlier findings, using existing datasets. The multilevel estimation procedure used by Prof. K K Das, examined the effects of household and village environmental factors on the prevalence of diseases among individuals within and between households in North-Eastern parts of India. The indicators such as availability of doctors in the village, the distances of medical facility from the village, source of drinking water, separate kitchen facility, toilet facility, cooking fuel, type of house, urban-rural set-up have significant impact on disease prevalence and different types of diseases. Dr. Ajay Pandey through his analysis demonstrated the significant gain in neonatal survival among pregnant mothers who adhere to WHO recommended inter-birth birth-interval length of 33 months. The neonatal deaths were also found to be lower in communities that are connected to all weather roads compared to those who are not. This has policy implications for demographically poor performing Empowered Action Group States, as the infrastructure push is needed in these Sate. In recent times especially since 2014 there has been increased allocation by the Central Government for infrastructure development and construction of all weather roads in these states.

Dr. Vivek Verma using Rank Set Sampling (RSS) technique demonstrated the superiority of RSS in situations where the probability of occurrence of an event is not fixed but a random quantity. He showed that, in the estimation of the probability of infant deaths, Bayesian estimators based on ranked set sample not only proved more effective and efficient than any other estimator, but also consistent with the NFHS reported value. Dr. Aditi Baruah derived a statistical model for the distribution of closed birth interval by considering variation in post-partum amenorrhea (PPA) period. Using the model Dr. Aditi demonstrated low risk of conception among Adivasi (Tirbal) female tea gardeners of Assam regardless of their parity. It is found that the risk of conception is low in the population surveyed and is reasonable compared to other methods. Dr. H. Brojeshwor Singh using the data from rural Manipur estimated the average duration of PPA as 6.6 months. PPA is the time interval between the termination of women's pregnancy and the beginning of the first subsequent menstruation. This finding has immense value from the policy perspective for those designing family planning strategies, especially the PPIUCD/IUCD.

Professor M. Nazrul Islam of Bangladesh suggested new estimation procedure of estimating speed of aging in a population as a function of demographic components i.e. life expectancy and population fertility rates. He tested his estimates with the existing measures of aging velocity using census data of Bangladeshi population for the census years 1981 and 2001. The findings demonstrate that the method suggested are good alternative and consistent over the existing methods. The alternative approach suggests slower aging process than those obtained by the existing measures. Dr. Tandrima Chakraborty of NSSO-India using Weighted Epidemic Chain Binomial Model with one introductory case for four and five member households demonstrated the superiority of model fit to epidemic dataset in studying the pattern of spread of infectious diseases.

Dr. Ramesh K Vishwakarma using Liver Cirrhosis marker data demonstrated the feasibility of computing concordance correlation coefficient (CCC) through an application of prior information using Bayesian approach. The study demonstrated that the Bayesian counterpart of CCC estimates applied between serum bilirubin and albumin among liver cirrhosis patient's data and its 95% posterior interval for concordance correlation coefficient were found to be very narrow, indicating that the estimates obtained through the suggested method are very precise. Dr. Dharmendra Kumar Dubey determined the predictors of low birth weight among adolescent mothers in Assam-India. Predictors of Low Birth Weight were found to be low levels of education, being poor, fourth & above birth order and mothers being anemic. Prevalence of LBW varied across the districts, with highest reported from Kamrup and Dhubari and lowest reported from Sonitpur and Karbi Anglong districts of Assam in India. Dr. Padum Narayan studied the linkages of son preference over daughters in childbearing process among married couples in India. The findings from the study shows that the parity progression ratios were consistently higher among currently married women who had only daughters at all parities as compared to those who had only sons or both sons and daughters in NFHS-3 (2005-06) as well as NFHS-4 (2015-16) irrespective of the place of residence. The pace of progressions differs substantially between urban and rural areas. Greater parental preference for sons over daughters has been observed in the rural areas as compared to urban areas at all parties in 2015-16. The study suggest that the recent initiate of "BETI Bachao, BETI Padhao" which literally means educate girl child to save girl child is very timely to eliminate the son preference in the society.

Dr. Jagriti Das using Log Normal distribution as an actuarial risk model estimated important actuarial quantities like the probability of ultimate ruin, moments of the time to ruin, the surplus prior to ruin and the deficit at the time of ruin when the underlying claim severity distribution is Log Normal. Dr. Jaishree Prabha developed imputation methods to reduce the impact of non-response sampling error at both the occasions in two-occasion successive (rotation) sampling. Dr. Lipi B Mahanta improvised the Generalized System of Curves that is important in describing frequency distributions for wide a variety of observed distributions.

Scholarly work by authors not only demonstrate their innovative thinking but also is sincere gratitude towards an inspiring mentor, guide & motivator Professor Dilip C Nath who at the age of 70 keeps transforming our dreams into actions. He is currently Professor Emeritus at Royal Global University, Assam and was a Former Vice Chancellor, Assam University Silchar, Assam, India.

Contributors

Aditi Baruah,
Department of Statistics,
Bahona College, Jorhat, Assam, India

Atanu Bhattacharjee,
Statistician
Leicester Real World Evidence Unit
Diabetes Research Centre,
University of Leicester, UK

Tandrima Chakraborty,
Deputy Director
National Statistical Office
(Data Processing Divison),
Kolkata, India

Dharmendra Kumar Dubey,
Associate Professor
Sharda School of Allied Health Sciences
Sharda Hospital, Greater Noida, India

Jagriti Das,
Assistant Professor
Department of Statistics,
Gauhati University, Guwahati,
Assam, India.

Kishore K. Das,
Professor
Department of Statistics,
Gauhati University, Guwahati, India

M. Nazrul Islam,
Professor, Department of Statistics,
Shahjalal University of Science & Technology,
Sylhet, Bangladesh

Jaishree Prabha Karna,
Post Doctoral Fellow
Department of Statistics, Gauhati University,
Assam, India.

Lipi B Mahanta,
Associate Professor
Institute of Advanced Study in Science and Technology,
Guwahati, Assam, India

Dilip Chandra Nath,
Professor Emeritus,
School of Applied & Pure Sciences,
The Assam Royal Global University,
Former Vice Chancellor, Assam University,
Silchar, Assam India

Padum Narayan,
Data Manager/Analyst
ACCELERATE
Johns Hopkins University School of Medicine,
C-11, Green Park Extension,
New Delhi, India

Arvind Pandey,
Ex-Director, ICMR-National Institute of Medical Statistics,
New Delhi, India

Ajay Pandey,
Assistant Director
Population Research Centre,
University of Lucknow-226007, India.

H. Brojeshwor Singh,
Associate Professor
Department of Statistics,
Modern College, Imphal East. India

Richa Sharma
Deputy Director
National Institute of Labour Economics Research & Development
Narela, Delhi 110040 India

Ramesh K Vishwakarma,
Department of Biostatistics and Bioinformatics,
King Abdullah International Medical Research Center/
King Saud bin Abdulaziz University for Health Sciences-MNGHA,
Riyadh, Saudi Arabia

Vivek Verma,
Assistant Professor
Department of Statistics
Assam University. Silchar
Assam India

A Probability Model for Closed Birth Interval and Its Application to Adivasi Married Females

Abstract

Birth interval reflects the reproductive behaviour of a population. The study of the birth interval is useful in detecting and measuring the current changes in the natality pattern of women. The closed birth interval is a good index of fertility as it indicates at what spacing women have children. The fertility of a woman is inversely related to her mean closed birth interval. In this chapter, a model for the distribution of closed birth interval has been derived by considering variation in post-partum amenorrhoea (PPA). It is assumed that PPA follows a modified Pascal distribution. The risk of conception has been estimated by the scoring method. The model has been demonstrated for estimating the risk of conception irrespective of parity for Adivasi married female of tea garden areas of Assam. It is observed that model is suitable to describe a distribution of the closed birth interval. It is found that the risk of conception is low (0.61) for this survey population. The estimate of the parameter λ, the risk of conception obtained through the model is reasonable.

Keywords: Conception, Foetal wastage, Parity, Postpartum amenorrhea, Pregnancy

1. Introduction

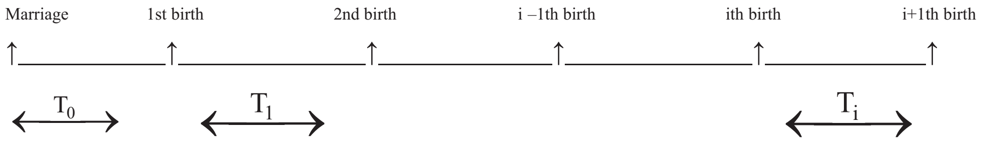
Birth interval reflects the reproductive behaviour of a population. The study of the birth interval is useful in detecting and measuring the current changes in the natality pattern of women. Birth interval influences the rate of population growth. There is an inverse relationship between the birth rate and birth intervals. A population with a higher birth interval may grow slowly than the other population. Thus birth intervals are related to population growth. It has also a direct impact on the health status of mothers as well as children. The birth interval can also be used to estimate certain bio social parameters of fertility such as fecundability, the incidence of foetal wastage, etc.

Mainly, there are two types of birth intervals-namely closed birth interval and open birth interval. These intervals are sensitive indices of fertility. Closed birth interval is the interval between the successive live births of a woman and open birth interval is the interval from the date of last live birth to the date of enquiry. Averaged closed birth interval indicates the extent of spacing between children. A study of closed birth interval is useful in the analysis of fertility change in the sense of change in spacing.

There are different types of closed birth intervals-- (i) All closed birth interval: In this case, all birth intervals obtained from the survey are taken together and analysis is carried out by birth order, age, and marital duration, (ii) Last closed birth interval: This is the interval between the last and last but one live birth of each woman. A woman who has given birth to at least one child will contribute one interval of this type. (iii) Straddling birth interval: An interval is considered to be straddling at a particular age or duration of marriage or at a particular point of time if one birth occurs before that age or point and the next birth occurs later, (iv) Interior birth interval: An interval that begins and ends in any segment of age group or marriage duration is called an interior birth interval.

A number of probability models have been developed to study birth interval which is more sensitive index for detecting current changes in the fertility pattern of women who are still in reproductive ages. Studies [1-3] had derived probability models for closed birth intervals. Some theoretical continuous time models [4-5] for closed birth intervals for any specific order with fixed marital duration have also developed. A model [6] was developed for the closed birth interval of women with specified marital duration. A parity dependent model [7] was derived for closed birth interval. Pathak (1983) proposed A continuous time probability model [8] was proposed for closed birth intervals of women of a specified marital duration. A probability distribution for the closed birth interval [9] was derived. A Probability Model [10] was derived to describe the variation in the length of any order of closed birth interval of females of large marital duration with the assumption that the fecundability for females of migrated and non-migrated couples is different and remains constant till the next birth.

A closed birth interval is the interval between two births i.e. the closed in the sense that the time is closed by two births. The closed birth interval can be represented as follows:

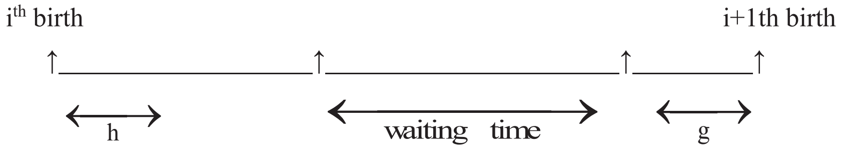


The symbol T_i denotes the length of closed birth intervals and T_i indicates the interval between the ith and (i+1) th births. Thus a woman of parity i can contribute i^{th} closed birth intervals The first closed birth interval T_0 is different from the remaining closed birth interval as it does not have the period of postpartum amenorrhoea which is an important component in other intervals.

Birth interval between two live births has four main components: (i) The period of post partum amenorrhea following the birth of the earlier child (except T_0 i.e. from marriage to first birth). (ii) Total waiting times between two live births. (iii) The period of pregnancy and post termination amenorrhea (if any) of abortion or stillbirths intervening the live births. (iv) Gestation period though variable but generally taken as nine months.

An analytical model has been developed for the probability distribution of the closed birth interval by considering the interval as the sum of these four components assuming known functional forms for each of the component distribution and their statistical independence [11].

Let X_i ($i \geq 1$) be the interval between ith and (i+1) th live birth, consisting of (i) the period of post-partum amenorrhea following the ith birth, (ii) waiting time for the conception (i+1) th and (iii) and the gestation period for the (i+1) th live birth.



Let h denotes the duration of non-susceptible period. The duration of non-susceptible period h after each parity is taken as constant, but in practice it is observed that the period of gestation of a conception resulting in a birth is almost constant. The duration of PPA following the birth varies from female to female, though for the same female the variation over parity may be assumed to be negligible. The non-susceptible period is a major determinant of birth interval in a population with low levels of contraception.

Thus in this paper an attempt has been made to develop a model of closed birth interval by considering variation in the non-susceptible period for estimating risk of conception for a low contracepting population. The application of this model has been illustrated through closed birth interval data of Adivasi (tea garden labourer) married females of Assam.

2. The Model

(i) Assumptions

A probability distribution for describing the variation in the length of the closed birth interval of a woman has been derived under the following assumptions.

(i) Closed birth intervals are considered only after at least one birth has taken place.

(ii) Only the non-contracepting women are considered

(iii) Foetal wastage, parity, and marital duration are ignored.

(iv) Assuming the one-to-one correspondence between a conception and a live birth

(v) Let the married women have a constant risk of conception λ ,

(vi) The time interval of the first conception after marriage follows an exponential distribution with density function

$$f_0(t) = \lambda\, e^{-\lambda t} \qquad , \qquad t > 0, \lambda > 0$$

(vii) The duration between ith and (i+1) th conception follows a displaced exponential distribution with probability density function

$$f_i(t) = \lambda\, e^{-\lambda\,(t-h)}, \qquad t > h, i = 1,2,3,\ldots.$$

where λ is a constant, called the conception rate, and is a measure of the expected risk of a conception with which the female proceeds for the next conception. And 'h' is the period of non-susceptibility including the gestation period and the period of post-partum amenorrhea (PPA).

We have considered only non-contracepting women for assessing the impact of variation of PPA on the closed birth interval.

(ii) Derivation of Model

The period of post-partum amenorrhea is the basic component of birth intervals. Usually, it is the longest in natural fertility. It is the most variable component of the birth interval. It is therefore important to measure precisely the length of amenorrhea in a fertility study. Data on amenorrhea have been obtained from the survey when women have been asked when they resumed menstruation after the birth of their last child.

Many models of amenorrhea [12-14] have been developed to adjust the data. A modified Pascal distribution [12], has been used most often in reproductive models and later [14] extended the work of Barrett [12].

A generalization of Barrett's distribution namely

A=C+ x_1 +x_2 + …+ x_k , where C and k are positive integers and the x_i are identically and independently distributed geometric variables with common parameter 'p'. The resulting modified Pascal distribution is

$$a(C+h) = \binom{h+k\text{-}1}{k\text{-}1}(1-p)^h\, p^k, h = 0,1,2,...$$

Parameter C, signifying the minimum length of amenorrhea, is typically set at 1 or 2, and h = the period of non-susceptibility.

Now suppose that 'h', instead of being regarded as a fixed constant, is also a random variable following the modified Pascal distribution. Hence the joint probability density function of closed birth interval X and 'h' the PPA is given by

$$f_1(x,h) = \lambda\, e^{-\lambda(x\text{-}h)} \binom{h+k\text{-}1}{k\text{-}1}(1-p)^h\, p^k, \quad h = 0,1,2,...; h < x < \infty.$$

Therefore the marginal function of x is given by:

$$f(x) = \sum_{h=0} f_1\,(x,h)$$

$$f(x) = \lambda p^k (1-e^{\lambda}q)^{-k} e^{-\lambda x}\ \text{, -k<x<}\infty$$

The probability density function of the truncated exponential distribution, truncated at both end '1' and 'T' is given as

$$g(x) = \frac{f(x)}{P(1 \le x \le T)}$$

$$= \lambda(\ e^{\text{-}\lambda} \text{-} e^{\text{-}T\lambda})^{-1}\, e^{\text{-}\lambda x},\ \ 1 \le x \le T.$$

$$F(T) = P(X \le T) = \ (\ e^{\text{-}\lambda} \text{-} e^{\text{-}T\lambda})^{-1}\,(1 \text{-} e^{\text{-}T\lambda}\ ).$$

3. Estimation Technique

It is seen that the proposed model is based on only one parameter, namely, λ ; an effort is made to estimate the parameter in the model by the method of maximum likelihood. The procedure to obtain the maximum likelihood estimate of λ is described. Let X_i (i=1,2,3,…N) be a random sample of size N from the distribution (1). Let N_r (r =

1,2,3…. n) be the observed frequencies, such that

$$\sum_{r=1}^{n} N_r = N \quad .$$

The likelihood function is given by

$$L = \frac{N}{\prod_{r=1}^{n} N_r} \prod_{r=1}^{n} P_r^{N_r},$$

$$\text{where } P_r = g(x)$$

which can be obtained from the corresponding distribution function $F_r(T)$ by successive subtractions. The maximum likelihood estimate of λ is obtained by solving the following normal equation,

$$\frac{\partial \log L}{\partial \lambda} = \sum_{r=1}^{n} \frac{N_r}{P_r} \frac{\partial P_r}{\partial \lambda} = 0 \qquad (2)$$

For this, the differentiation of P_r ($r \geq 1$) with respect to unknown parameters is required. Since the explicit solution of equation (2) is not possible, the scoring method can be used to obtain maximum likelihood estimators of the unknown parameter. So, the m. l. estimate of the parameter is calculated from the following matrix equation

$$I\partial\lambda = S \text{ , where Score (S)} = \sum_{r=1}^{n} \frac{N_r}{P_r} \frac{\partial P_r}{\partial \lambda}$$

$$\text{and Information (I)} = N\sum_{r=1}^{n} \frac{1}{P_r} \frac{\partial P_r}{\partial \lambda} \frac{\partial P_r}{\partial \lambda}$$

At first the differentiation of F_r (T)’ s for r=1, 2,…n is obtained and then the corresponding differentiation of the probabilities can be obtained by successive subtractions.

$$\frac{\partial F_r(T)}{\partial \lambda} = (-1)(e^{-\lambda} - e^{-T\lambda})^{-2}(Te^{-T\lambda} - e^{-\lambda})$$
$$+ Te^{-\lambda}(e^{-\lambda} - e^{-T\lambda})^{-1} + e^{-T\lambda}(e^{-\lambda} - e^{-T\lambda})^{-2}(Te^{-T\lambda} - e^{-\lambda})$$

$$\frac{\partial P_r(T)}{\partial \lambda} = \frac{\partial F_{r+1}(T)}{\partial \lambda} - \frac{\partial F_r(T)}{\partial \lambda} \qquad r = 1,2,3\ldots n$$

The pilot value of the unknown parameter, which is required for the scoring method, can be calculated by equating the relative frequency of one cell of the observed distribution having a significant number of observations to their respective theoretical expression. This equation can be solved by Newton Raphson’s iteration procedure. The pilot value obtained by this process is used in the scoring method.

For this analysis, the relative frequency of the first cell of the observed distribution is equated to the respective theoretical expression. The approximate value of the root is obtained and then the real root of the equation or the pilot value of the parameter λ is obtained by the Newton-Raphson method.

4. Application and Discussion

The data for this analysis has been taken from a survey conducted in 14 tea gardens of Jorhat district of Assam. The survey titled “A study on the fertility and reproductive health of tea garden female workers of Assam” was conducted in 2000 under the financial support of the University Grants Commission, New Delhi. The total number of women interviewed in this survey was 1015. Out of these 238 women of reproductive age group (15-49) of marital duration 6 years were taken for this analysis. Table 1 gives the frequency distribution of closed birth intervals in years for 238 Adivasi females with a marital duration of 6 years. The maximum likelihood estimate of the parameter is found to be 0.61.

Table 1. Observed and expected distribution of closed birth interval of Adivasi married females

Birth Intervals (in years)	Observed frequency	Expected frequency
1-2 2-3 3-4 4-5 5-6	113 64 33 18 10	114 62 33 19 10
Total	238	238
χ^2=0.124(calculated) $\chi^2_{0.05\,(3\text{ d.f})}$ =7.815	$\hat{\lambda} = 0.61$ $\hat{V}(\hat{\lambda})x10^3 = 13.78$	

The expected frequencies are computed with the help of the estimated value of parameter λ = 0.61. The expected frequencies corresponding to observed frequencies, the calculated value of chi square, the estimates of parameter λ and the variances of the estimated parameter are shown in the Table 1. To measure the closeness of the expected frequencies and the observed frequencies, the usual chi square test has been applied. The calculated value of chi-square is 0.124 for 3 d.f, which is insignificant. This shows that the proposed model provides an adequate fit to the given data.

Figure 1 displays both the observed and fitted distribution of closed birth intervals for Adivasi married females for this survey population. The estimate of the parameter λ, the risk of conception for an Adivasi married female obtained through this model is reasonable in Indian context, but it is lower in comparison to a female of developed countries. It is seen that this value which is 0.61, is similar to the value estimated [6] for a married woman in Uttar Pradesh, which was 0.58. Also, Singh et al [15] reported risk of conception from closed birth interval analysis is 0.78 for the same population. However, Singh [16] obtained the estimate of the risk of conception for women in Uttar Pradesh was very high (λ = 1.38).

However, in this model foetal wastage is not considered and hence the model can be extended by taking into incomplete conception. From the above analysis it is clear that proposed model is a suitable one to describe a distribution of the closed birth intervals for such survey-populations.

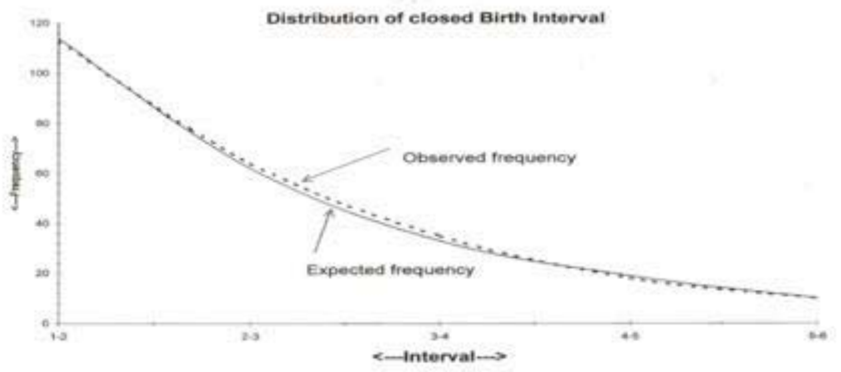


Figure 1

References

Barrett, J. C. (1969). A Monte Carlo simulation of human reproduction. *Genus, 25*, 1-22.

Chakroborty, K. C. (1976). *Some probability distributions for birth interval* (Unpublished doctoral dissertation). Banaras Hindu University, India.

D'Souza, S. (1974). *Closed birth intervals: A data analytic study*. Sterling Publishers Ltd., India.

Lesthaeghe, R. J., & Page, H. J. (1980). The post partum non susceptible period: Development and application of model schedules. *Population Studies, 34*(1), 143-170.

Mishra, R. N., Pandey, A., & Singh, K. K. (1983). A generalized probability distribution for closed birth distribution. *Health and Population Perspectives and Issues, 6*(1), 36-45.

Pathak, K. B. (1983). An extension of a probability model for close birth interval. *Health and Population: Perspectives and Issues, 6*(3), 133-142.

Potter, R. G., & Kobrin, F. E. (1981). Distribution of amenorrhoea and anovulation. *Population Studies, 35*, 85-99.

Sheps, M. C., & Menken, J. A. (1972). Distribution of birth intervals according to sampling frame. *Theoretical Population Biology, 3*(1), 1-26.

Sheps, M. C., & Perrin, E. B. (1964). The distribution of birth intervals under a class of stochastic models. *Population Studies, 18*, 321-331.

Sheps, M. C., Menken, J. A., & Radick, A. P. (1969). Probability models for family building: An analytical review. *Demography, 6*(2), 161-183.

Singh, A. S. (2016). Human fertility behavior through birth interval models: Overview. *American Journal of Theoretical and Applied Statistics, 5*(3), 132-137.

Singh, S. N. (2000). A probability model for closed birth interval to study the effect of migration on fertility. Paper presented at the International Conference on Statistics, Combinatorics and Related Areas, Indian Institute of Technology-Bombay.

Singh, S. N., Bhattacharya, B. N., Pandey, A., & Mishra, R. N. (1983). Parity dependent model for closed birth interval. *Journal of Indian Statistical Association, 21*, 67-72.

Singh, S. N., Pandey, A., & Mishra, R. N. (1981). A generalized probability distribution for open birth interval. *The Aligarh Journal of Statistics, 1*(2), 183.

Singh, S. N., Yadava, R. C., & Pandey, A. (1979). On a probability model for closed birth interval. *Health and Population Perspectives and Issues, 2*(3), 224-230.

Srinivasan, K. (1967). *Probability models for two types of birth intervals with application to Indian population* (Unpublished doctoral dissertation). Department of Demography, University of Kerala, India.

Inter-birth Interval Length and Neonatal Survival: A Study on Demographically Poor Performing EAG States

Abstract

World Health Organization (WHO) expert group on birth spacing recommended optimal birth spacing between live birth and the next pregnancy as 24 months in order to reduce maternal, peri-natal, and infant deaths. This means the Birth Interval between two consecutive births should at least be 33 months. Using the National Family Health Survey-II (1998-99) dataset the gain in neonatal survival is studied among those who adhered to WHO recommended minimum birth interval. The findings suggest substantial gain in neonatal survival among those who adhere to WHO-recommended birth spacing. As the NFHS-II dataset provides the opportunity to measure the developmental impact on neonatal survival the community connectivity with all-weather roads was studied using Hierarchical Linear Models. The odds of neonatal deaths were observed to be lower in communities that are connected by all-weather roads compared to those that are farther.

Keywords: Inter-birth Interval; All-weather Road; Neonatal Deaths.

1. Introduction

United Nations Sustainable Development Goals for 2030 aim at reducing neonatal mortality from its level to at least 12 per 1000 live births [1]. For the year 2015, India's medium-range estimates on neonatal mortality as estimated by the UN Inter-agency Group for Child Mortality Estimation (IGME) is 27.7 deaths per 1000 live births. This translates into 695,852 neonatal deaths in 2015 [2]. Globally during 2015 about 2,682,438 neonates die within 28 days of birth. India accounts for 26 percent of the global neonatal deaths. This reflects a disproportionately large burden of the global neonatal deaths, as approximately one in four neonatal deaths occur in India. Studies have shown that three-quarters of these neonatal deaths occur in the first seven days of life or the early neonatal period, which are largely preventable [3]. On analyzing the data from the successive rounds of the National Family Health Survey, a slow rate of decline in neonatal mortality in recent years is evident as the rate declined from 48 deaths per 1000 live births in 1995 to 27.7 in 2015 [2]. In order to increase the neonate's survival outcomes and speed up the rate of decline of neonatal mortality rates saw a marked policy shift and the launch of National Rural Health Mission (NRHM) programs in 2005. The State received direct grant from the central government to establish across district state of art Sick Neonatal Care Unit's (SNCU) within the District Hospitals. The SNCU units are established to provide medical emergencies to sick neonates (both in-born & out-born) within the district. India's progress in achieving the SDG goals hinges largely on the progress made by the eight demographically backward north Indian states, together termed as Empowered Action Group (EAG) States. Whether or not India achieves the SDG target of at least 12 neonatal deaths per 1000 live births, by the year 2030, depends largely on the progress made by these States in reducing neonatal deaths.

Poor birth spacing is among the leading causes of high levels of neonatal and maternal mortality. Births in India and especially in the EAG States are poorly spaced leading to high levels of maternal, infant, and child mortality. Studies including those done by USAID recommend 3 to 5 years of birth spacing compared to 2 to 3 years spacing recommended by the WHO [4]. In the year 2005, in order to have a uniform recommendation on birth spacing for improved survival outcomes of neonates, WHO held a technical consultation and scientific review of Birth Spacing at its headquarters in Geneva. The consultation meeting comprised 37 international experts. The expert group recommended the optimal birth spacing between live birth and the next pregnancy as 24 months at least, in order to reduce maternal, peri-natal, and infant deaths [4]. It is in this backdrop that the inter-birth interval (IBI) length in EAG States is studied keeping in view the WHO recommended minimum birth interval length. This chapter therefore examines the impact of successive inter-birth interval length on neonatal mortality keeping other socio-economic and demographic variables constant using the proportional hazards model.

2. Review

Several risk factors that influence the survival chances of neonates have been identified and reported in the literature. This includes women's parity, maternal age, caste, religion, birth weight, size of the baby at birth, frequency of antenatal visits, TT injections during pregnancy, stillbirth, and previous birth interval length [5]. There are research studies that advocate a birth interval length of 2 years between two consecutive births for better child health [6] while others advocate intervals of 3 to 5 years as safe for both the mother and the baby compared to ≤ 2 years of the birth interval [7]. Studies have also shown that too short birth intervals (<2 years) are associated with high levels of infant and child mortality [8-10]. Using conditional logistic regression Kozuki& Walker [11] analyzed 47 DHS country data and predicted that children that are born at less than 18 months & less than 24 months preceding birth interval have a higher risk of neonatal and under-five mortality compared to those who were born with a preceding birth interval greater than or equal to 24 to 60 months. Rutstein [12], study using DHS data from 17 developing countries showed the increased risk of mortality and under-nutrition for births with birth intervals less than 36 months. The study suggested optimal birth interval length to be in-between 36 and 59 months. Using the data from the rural northern Indian State of Uttar Pradesh Williams et.al. [13] examined the length of the preceding birth interval and neonatal outcome. The study reported higher odds of neonatal deaths for those with birth intervals less than <18 months and 18–35 months, compared to births that are spaced at 36–59 months. However, studies have also shown that too-long birth intervals (>5 years) are associated with an increased risk of pregnancy complications such as preeclampsia as the mother loses the protective effect from the previous pregnancy [14]. Exavery et al. [15] using data from rural Tanzania studied the factors responsible for adherence to WHO recommended inter-birth interval length. It reported young maternal age, low levels of education, multiple births to the index, home delivery of the index child, being an in-migrant, higher parity and married as factors that are associated with non-adherence to the WHO recommended minimum inter-birth interval of 33 months. The impact of adherence to WHO minimum inter-birth interval length on neonatal survival and the impact of all-weather road connectivity within the community is studied.

3. Objectives

•To estimate the neonatal mortality for each successive inter-birth interval less than 33 months vis-a-vis greater than equal to 33 months using life tables

•To study the hazards ratio using the Cox Proportional Hazards model for each successive inter-birth interval less than 33 months vis-a-vis greater than equal to 33 months

•To study the community connectivity by pucca (all-weather) road on neonatal deaths in respect of marriage to first inter-birth interval

4. Data & Methods

Data collected from the National Family Health Survey (NFHS-2) [16] is used to investigate the risk of neonatal survival for the successive inter-birth intervals among those who adhere to WHO recommendations for 24 months and those who did not, keeping all other socio-economic & demographic variables constant. The data is analyzed for the EAG states comprising Bihar, Madhya Pradesh, Odisha, Rajasthan, and Uttar Pradesh. Life- tables are used to estimate the levels of neonatal mortality for birth to pregnancy intervals greater than 24 months or less than equal to 24 months. For ease of analysis, we added 9 months of gestation to the recommended 24 months of birth to pregnancy interval making inter-birth interval length as 33 months. STATA version 15 was used to carry out the analysis.

5. Findings

Table 1 presents the life table estimates of neonatal mortality for each successive inter-birth interval that adhere to the WHO recommended 33 months compared to those that do not. The findings suggest strikingly high levels of neonatal mortality for inter-birth interval lengths that did not adhere to the recommended minimum inter-birth interval length of 33 months. The difference in neonatal mortality is more pronounced for higher-order inter-birth interval lengths. Among marriage to first birth interval length, there is a gain of seven neonates for those who adhere to the WHO recommended minimum compared to those who do not. Among first to second inter-birth interval, the gain is of 40 neonates among those who adhere to WHO recommended inter-birth interval length of greater than 33 months. Those who adhere to the WHO minimum of greater than 33 months of inter-birth interval reported 31 neonatal deaths compared to 71 neonatal deaths among those who did not adhere to the WHO-recommended IBI. Similarly, for the second to third inter-birth interval, those who adhere to the recommended minimum reported 28 neonatal deaths compared to 64 among those who did not, thereby gaining 36 neonatal deaths among those who adhere to the WHO recommended minimum. In the third to fourth inter-birth interval, there is a gain of 39 neonatal deaths. Those who adhere to greater than 33 months of inter-birth interval reported 30 neonatal deaths per 1000 live births compared to 69 deaths among those who did not adhere to the recommended greater than 33 months of inter-birth interval length. The gain in neonatal mortality for those who adhere to WHO recommended IBI among fourth to fifth inter-birth interval is 41 neonates. Neonatal mortality in this segment of the inter-birth interval was recorded as 74 for those who did not adhere and 33 neonatal deaths for those who adhered to the WHO recommended minimum. Among fifth to sixth inter-birth interval length the neonatal mortality for those who adhere to greater than 33 months of birth interval reported 32 neonatal deaths compared to 81 deaths for those who do not adhere to WHO recommended IBI. We thus see a gain of 49 neonates in this segment of inter-birth interval length for those adhere to WHO recommended birth spacing. In the six-plus births segment, we see maximum gain in neonatal survival for those adhering to WHO recommendations. Those whose inter-birth interval length was less than equal to 33 months, reported 97 neonatal deaths compared to 39 deaths for those who did not adhere to the WHO recommended minimum.

Table 1 also presents the findings from the multivariate analysis using the Cox Proportional Hazards model. The Hazard ratios are presented in the table after consolidating the findings from the seven different sets of multivariate hazard models run separately for each successive inter-birth interval length. The Hazards model was run after testing the proportionality assumption among the controlled covariates. For marriage to first birth interval length, the hazard or the risk of neonatal deaths is not different among those who adhere to and those who do not adhere to WHO recommendations. This difference is also not statistically significant.

Among the second inter-birth interval, it can be seen that, for those adhering to the inter-birth interval of greater than 33 months the risk or hazard of neonatal deaths is 45 percent less and is statistically highly significant. Similarly, for the third inter-birth interval (between the second and third birth) the risk/hazard of neonatal deaths is 40 percent less for those who adhere to WHO recommendations and is statistically significant. Similarly, for the fourth inter-birth interval (between the third and fourth birth) among those who space births more than 33 months the risk/hazard of neonatal deaths is 38 percent less compared to those whose inter-birth interval length is less than 33 months. This is again highly significant. For the fifth inter-birth interval (between fourth and fifth birth) those above WHO recommended have 30 percent less risk/hazard of neonatal deaths compared to those below WHO recommended, and is again statistically significant. Similarly, for the sixth inter-birth interval (between the fifth and sixth birth) those above WHO recommended have 39 percent less risk/hazard of neonatal deaths than whose birth interval is below WHO recommended and is statistically significant. For births Six and above 37 percent less risk/hazard of neonatal deaths is reported among those whose length of inter-birth interval is more than WHO recommended, and is again statistically highly significant compared to those whose inter-birth interval length was less than equal to 33 months. These findings are strongly suggestive of the implementation of WHO guidelines with regard to the length of inter-birth interval length as the probability of neonatal survival is very high among the successive inter-birth interval length.

Road connectivity in rural areas was taken as one of the development indicators, as it was a big challenge at the time of the NFHS-II survey in accessing health care services. NFHS-II had asked questions to rural communities on the village distance from the all-weather road in KM. The village data file was merged with the birth file to study the effect of community connectivity with all-weather roads and its impact on neonatal survival. Table 2 below provides the intra-class correlation (ICC) calculated using the GLIMMIX procedure based on an unconstrained model. The unconstrained model does not include any predictor variable and the estimate of random intercept provides the ICC value based on the model specification shown by equation 1 below.

Level 1

$$\eta_{\mathrm{ij}} = \beta_{0j}$$

Level 2

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \dots\dots\dots \text{where} \mu_{0j} N(0, \tau_{00})$$

Combined

$$\eta_{\mathrm{ij}} = \gamma_{00} + \mu_{0j} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

From the above model 1 which is intercept only model, gives the value of ICC as

$$\mathrm{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

$$= \frac{2.7247}{2.7247 + (\frac{\pi^2}{3})}$$

$$= 0.4532$$

Equation 2 below provides the value of ICC as 0.4531 or 45 percent of the total variance in neonatal deaths is due to the between-community differences. As a next step in the model building covariates at level 1 (individual level) were introduced in the model and the intercept was allowed to vary across the community. For the model to converge, the covariates at level-1 included in the model were the sex of the child (0=male 1=female), mother literacy (0=illiterate 1=literate), tetanus toxoid injection during pregnancy (0=no 1=yes) and marriage to first birth interval. The model specification is given by equation 3 below

Level^{1}

$$\eta_{\text{ij}} = \beta_{0j} + \beta_{1j}(\text{sexc}_{\text{ij}}) + \beta_{2j}(\text{literacy}_{\text{ij}}) + \beta_{3j}(\text{TTinjection}_{\text{ij}}) + \beta_{4j}(\text{BI}_{\text{ij}})$$

Level^{2}

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \ldots\ldots\ldots \text{where} \mu_{0j} N(0, \tau_{00}).$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

Combined

$$\eta_{\text{ij}} = \gamma_{00} + \gamma_{10}(\text{sexc}_{\text{ij}}) + \gamma_{20}(\text{literacy}_{\text{ij}}) + \gamma_{30}(\text{TTinjection}_{\text{ij}}) + \gamma_{40}(\text{BI}_{\text{ij}}) + \mu_{0j} \ldots\ldots\ldots\ldots (3)$$

In the model-3 the intercepts are allowed to vary across communities but the slope is fixed. The findings presented in Table-2 under model-1 show the relative importance of the predictor variable at the individual level (level 1) in predicting neonatal deaths. The only covariate at level-1 that predicts neonatal deaths is whether or not the mother got TT vaccination during the pregnancy. To study the impact of distance of village connectivity from all-weather roads, a level-2 predictor was introduced in the model. It is hypothesized that closer the community to all weather roads better the neonatal survival chances. The model specification is given by equation-4 below

Level^{1}

$$\eta_{\text{ij}} = \beta_{0j} + \beta_{1j}(\text{sexc}_{\text{ij}}) + \beta_{2j}(\text{literacy}_{\text{ij}}) + \beta_{3j}(\text{TTinjection}_{\text{ij}}) + \beta_{4j}(\text{BI}_{\text{ij}})$$

Level^{2}

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{distance}_{j}) + \mu_{0j} \ldots\ldots\ldots \text{where} \mu_{0j} N(0, \tau_{00}).$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

Combined

$$\eta_{\text{ij}} = \gamma_{00} + \gamma_{10}(\text{sexc}_{\text{ij}}) + \gamma_{20}(\text{literacy}_{\text{ij}}) + \gamma_{30}(\text{TTinjection}_{\text{ij}}) + \gamma_{40}(\text{BI}_{\text{ij}}) + \gamma_{01}(\text{distance}_{j}) + \mu_{0j} \ldots\ldots\ldots\ldots (4)$$

In equation 4 the BI considered is marriage to first birth interval. The variable distance is the village distance to all-weather roads in kilometers and is a continuous variable. The findings from the equation 4 are presented in the table-2 under sub-heading model 2. Here also the intercept is allowed to vary across the communities and the slope is fixed. Even after the introduction of the level 2 variable, at level-1 the tetanus injection during pregnancy is the most important variable predicting the outcome variable i.e. neonatal deaths. The village distance to all-weather road is significant at a 10 percent level of significance and in the communities that are closer to all-weather road, the odds of neonatal deaths are 10% less (exp(-0.0968)). This is an important finding and indicates the importance of rural connectivity by all-weather roads. Since 2014, the government of India has been investing heavily in various infrastructure projects. Building roads at a pace never seen before any time in the past, has led to improved rural connectivity. Rural areas are better connected with expressways today, which reduces travel time immensely and has led to access to healthcare services rapidly. Kayode et al. [17] used the multilevel model to identify factors responsible for neonatal mortality in Ghana. Findings suggest community-based interventions such as investment in basic education, poverty alleviation, women empowerment, and infrastructural development will improve neonatal survival. Rayment et al [18] study findings support the idea of community-level policy intervention to increase the presence and continuity of community healthcare workers for improved outcomes for women at increased risk of health inequalities. He also suggested further research to study the relationship between community-based models of care and neonatal outcomes.

6. Conclusions

If India has to achieve the Sustainable Development Goal target set for 2030 of achieving at least 12 neonatal deaths per thousand live births, it is imperative that strong policies of educating couples (or families) to space births as per WHO recommendation needs to be evolved. The gain in neonatal deaths among higher-order inter-birth intervals is strikingly high among those who adhere to WHO recommendations. Besides the rural connectivity should improve as the odds of neonatal deaths is strikingly lower among villages that are closer to all-weather roads compared to those that are further away.

References

United Nations. (2015). Sustainable Development Goals 2015-2030. United Nations.

Kerber, K. J., de Graft-Johnson, J. E., Bhutta, Z. A., Okong, P., Starrs, A., & Lawn, J. E. (2007). Continuum of care for maternal, newborn, and child health: From slogan to service delivery. *The Lancet, 370*(9595), 1358-1369. https://doi.org/10.1016/S0140-6736(07)61578-5

World Health Organization. (2007). Report of a WHO technical consultation on birth spacing: Geneva, Switzerland 13-15 June 2005. World Health Organization. https://apps.who.int/iris/handle/10665/69855

Hinderaker, S. G., Olsen, B. E., Bergsjø, P. B., Gasheka, P., Lie, R. T., Havnen, J., ... & Lindmark, G. (2003). Avoidable stillbirths and neonatal deaths in rural Tanzania. *BJOG: An International Journal of Obstetrics & Gynaecology, 110*(6), 616-623.

Saumya, R., John, T., & Ian, A. (2006). Correlates of inter-birth intervals: Implications of optimal birth spacing strategies in Mozambique. Population Council.

Population Council. (n.d.). Correlates of inter-birth intervals: Implications of optimal birth spacing strategies in Mozambique. Retrieved July 4, 2023, from http://www.popcouncil.org/pdfs/frontiers/FR_FinalReports/Mozam_OBSI.pdf

Yohannis, F., & Yemane, B., & Alemayehu, W. (2003). Differentials of fertility in rural Butajira. *Ethiopian Journal of Health Development, 17*. https://doi.org/10.4314/ejhd.v17i1.9778

Miller, J. E., Trussell, J., Pebley, A. R., & Vaughan, B. (1992). Birth spacing and child mortality in Bangladesh and the Philippines. *Demography, 29*(2), 305-318. https://doi.org/10.2307/2061543

Winikoff, B. (1983). The effects of birth spacing on child and maternal health. *Studies in Family Planning, 14*(10), 231-245.

Kozuki, N., & Walker, N. (2013). Exploring the association between short/long preceding birth intervals and child mortality: Using reference birth interval children of the same mother as comparison. *BMC Public Health, 13*(Suppl 3), S6. https://doi.org/10.1186/1471-2458-13-S3-S6

Rutstein, S. O. (2005). Effects of preceding birth intervals on neonatal, infant and under-five years mortality and nutritional status in developing countries: Evidence from the demographic and health surveys. *International Journal of Gynecology & Obstetrics, 89*(Suppl 1), S7-S24. https://doi.org/10.1016/j.ijgo.2004.11.012

Williams, E. K., Hossain, M. B., Sharma, R. K., Kumar, V., Pandey, C. M., & Baqui, A. H. (2008). Birth interval and risk of stillbirth or neonatal death: Findings from rural North India. *Journal of Tropical Pediatrics, 54*(5), 321-327. https://doi.org/10.1093/tropej/fmn027

Orji, E. O., Shittu, A. S., Makinde, O. N., & Sule, S. S. (2004). Effect of prolonged birth spacing on maternal and perinatal outcome. *East African Medical Journal, 81*(8), 388-391. https://doi.org/10.4314/eamj.v81i8.9198

Exavery, A., Mrema, S., Shamte, A., Bietsch, K., Mosha, D., Mbaruku, G., & Masanja, H. (2012). Levels and correlates of non-adherence to WHO recommended inter-birth intervals in Rufiji, Tanzania. *BMC Pregnancy and Childbirth, 12*(1), 152. https://doi.org/10.1186/1471-2393-12-152

International Institute for Population Sciences, & ORC Macro. (2000). National Family Health Survey (NFHS-2), 1998-99: India. Mumbai: IIPS.

Kayode, G. A., Ansah, E., Agyepong, I. A., Amoakoh-Coleman, M., Grobbee, D. E., & Klipstein-Grobusch, K. (2014). Individual and community determinants of neonatal mortality in Ghana: A multilevel analysis. *BMC Pregnancy and Childbirth, 14*(1), 165. https://doi.org/10.1186/1471-2393-14-165

Rayment-Jones, H., Dalrymple, K., Harris, J., Harden, A., Parslow, E., Georgi, T., ... & Lewis, J. (2021). Project20: Does continuity of care and community-based antenatal care improve maternal and neonatal birth outcomes for women with social risk factors? A prospective, observational study. *PLOS ONE, 16*(5), e0250947. https://doi.org/10.1371/journal.pone.0250947

UNICEF. (n.d.). Neonatal mortality. Retrieved July 23, 2023, from https://data.unicef.org/topic/child-survival/neonatal-mortality

Table 1. Neonatal Deaths & Hazards Ratio: EAG States

Inter-birth Interval	**Neonatal Deaths**	**Hazard Ratio**	**SE**	**Sig.**
First (Marriage to First Birth)				
<=33 months (ref)	81			
>33 months	74	0.978	0.0445	0.626
Second (First to Second Birth)				
<=33 months (ref)	71			
>33 months	31	0.5562	0.038	0.000
Third (Second to Third Birth)				
<=33 months (ref)	64			
>33 months	28	0.5935	0.0484	0.000
Fourth (Third to Fourth Birth)				
<=33 months (ref)	69			
>33 months	30	0.6272	0.0592	0.000
Fifth (Fourth to Fifth Birth)				
<=33 months (ref)	74			
>33 months	33	0.6987	0.0779	0.001
Sixth (Fifth to Sixth Birth)				
<=33 months (ref)	81			
>33 months	32	0.6134	0.0868	0.001
Six Plus (Six and above)				
<=33 months (ref)	97			
>33 months	39	0.6357	0.0769	0.000

Table 2 Neonatal Deaths as an outcome of Community distance to All-Weather Roads

GLIMMIX Estimates	**Estimate**	**SE**	**Pr>|t|**	**CI**	
				Lower	**Upper**
Model 1 (Unconstrained)					
Intercept	-2.7247	0.13	<.0001	-2.9906	-2.4588
Model 2					
(Random Intercept Fixed Slope Co-variate at Level 1)					
Level 1 predictor					
Intercept	-2.1186	0.18	<.0001	-2.4782	-1.7589
TT inj	-0.7310	0.16	<.0001	-1.0541	-0.4080
Literacy	-0.2151	0.18	0.2356	-0.5705	0.1404
Sex of Child	-0.2236	0.16	0.1561	-0.5326	0.08547
Marriage to First IBI	0.1085	0.17	0.5134	-0.2169	0.4339
Model 3					
(Random Intercept Fixed Slope Outcome dependent on Level 2 Covariate (connectivity)					
Level 1 predictor					
Intercept	-1.8416	0.22	<.0001	-2.2883	-1.3949
TT inj	-0.7181	0.16	<.0001	-1.0414	-0.3948
Literacy	-0.2156	0.18	0.2339	-0.5706	0.1395
Sex of Child	-0.2230	0.16	0.1574	-0.5322	0.0861
Marriage to First IBI	0.1060	0.17	0.5230	-0.2195	0.4316
Level 2 predictor					
Distance to all weather road (km)	-0.0968	0.06	0.0914	-0.21	0.01643

Environmental Factors Affecting Health in Northeastern Region of India: A Multilevel Analysis

Abstract

This study examines household and village level environmental effects on the prevalence of diseases among households in Northeastern India. It uses data from the National Family Health Survey-2 (1998-99). Results are obtained from the estimated multilevel logistic regression model. There are 12564 households of eight Northeastern states viz., Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Tripura and Sikkim under this study and persons belonging to 10.4 percent households have been reported suffering from some diseases during two weeks preceding the survey and out of which members belonging to 3.77 percent household are suffering from waterborne diseases. In these reported diseases have been classified into the following five categories viz., (i) respiratory disorder, (ii) diarrhea & gastroenteritis, (iii) fever, jaundice, typhoid, (iv) delivery-related injuries and diseases for newborn babies, and (v) other diseases. It was found that the availability of doctors, the distance to medical facilities from the village, sources of drinking water, separate kitchens, toilet facilities, type of house, and urban-rural set-up have a significant impact on disease prevalence. This study has the potential for a better understanding of environmental factors associated with disease prevalence in the survey population and implementing of the National Population and Health Policies.

Keywords: Prevalence of diseases, Environmental determinants, logistic regression.

1. Introduction

In the constitution of the World Health Organization, health is defined as "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity" [21, 22]. In the context of health, Last [13] defined environment as "all that which is external to the individual human host. It can be divided into physical, biological, social, cultural, any or all of which can influence health status in a population." This definition is based on the notion that a person's health is determined by genetics and the environment. There are close links between the environment and people's health [1, 3, 5, 11]. A high-quality environment enables people to live longer in good health. Environmental factors have a huge impact on people's health. Globally, every year hundreds of millions of people suffer from respiratory and other diseases associated with indoor and outdoor pollution. Four million infants and children die every year from diarrheal diseases, largely as a result of contaminated water or food. Two million people die from malaria every year while 267 million are ill with it at any given time [4]. Three million people die each year from tuberculosis and 20 million are actively ill with it [4]. Half a million die as a result of road accidents. Hundreds of millions suffer from poor nutrition. This picture is far gloomier for any developing country like India. Almost all these health problems could be prevented [22]. Only a few studies on environmental factors affecting morbidity have been reported in India [12, 18]. In the international context, Saathoff et al [17], Wakou and Bell [20], Tipayamongkholgulet al [19], Dale et al [2], Dyer [4], and Sack et al [16] identified different environmental factors associated with prevalence[*1] of various diseases.

In this paper, we use both traditional and multilevel logistic regression models to identify a set of environmental determinants of morbidity in households using the Indian National Family Health Survey (NFHS) [14] data collected during 1998-99 on diseases in every household member. This study also analyzed the impact of different environment covariates on the prevalence of waterborne diseases in households for the same set of data. Our analysis is restricted to all Northeastern states including Sikkim.

2. Materials and Methods

2.1 Data Sources

To achieve the objective, data has been extracted from the second round of NFHS. It was conducted in 1998-99 under the auspices of the Ministry of Health and Family Welfare, India, and funded by the United States Agency for International Development (USAID). Data collection from 26 states was carried out in two phases, in the first phase data collection began in November 1998 in 10 states, and the second phase began in March 1999 in the remaining states.

2.2 Methodology

[1*] The prevalence rate (P) for disease is calculated as

$$P = \frac{\text{Number of people with the disease at a specified time}}{\text{Number of people in the population at risk at the specified times}} \times 100$$

Multilevel models [6,7,8,10,15,23] existence of hierarchical data by allowing for residual components at each level in the hierarchy. In the present study, a two-level model is considered, which allows for the grouping of household outcomes within village quantifying variations at each level. Thus, the residual variance is partitioned into a between-village component (the variance of the village-level residuals) and a within-village component (the variance of the household-level residuals). The village residuals, often called 'village effects,' represent unobserved village characteristics that affect household outcomes. It is these unobserved variables that lead to a correlation between outcomes for the households from the same village.

Many explanatory variables affecting the prevalence of disease are reported in this survey. Though the hierarchical structure of NFHS data involved many levels viz., households, village, district, state. In the present study, a two-level viz., household and village are considered for analysis with households nested within villages. There may be variation at the household level and between villages. The multilevel analysis may be used to quantify variations at each at each level, i.e., household and village. This may help to assess the contribution of covariates at a level regarding variability present at that level as well as at other levels. Considering 2-level data structures where we have a sample of households are nested within the village (level 2 units), multilevel logistic regression [6,7,8,10,15] is given as follows:

For the ith household in the $j$thvillage, observe a binary response

$$Y_{ij} = \begin{cases} 1, healthy\ household\ (disease\ free\ household\ during\ last\ five\ years), \\ 0, non-healthy\ (at\ least\ one\ of\ the\ members\ was\ reported\ suffering \\ \qquad from\ some\ kind\ of\ disease\ during\ last\ five\ years). \end{cases}$$

$$Y_{ij}|p_{ij} \sim Bernoulli(p_{ij}), \text{ where } p_{ij} = Pr(Y_{ij} = 1), \text{ and}$$

$$\text{logit}\,(p_{\text{ij}}) = \text{logit}\left(\frac{p_{\text{ij}}}{1-p_{\text{ij}}}\right) = \beta_{00} + \beta X_{\text{ij}} + \delta W_j + u_j + e_{\text{ij}}$$

where $u_j \sim N(0, \sigma_u^2)$ and $e_{\text{ij}} \sim N(0, \sigma_{\text{ij}}^2)$, $p_{\text{ij}}$is the probability that $i$thhousehold in the $j$thvillage having diseases during the last five years; $X_{\text{ij}}$and w_i are vectors of households and village level characteristics; and βandδ are vectors of estimated parameter coefficients.

The level-2 random variation is described by the termu_j, i.e., unobserved variation at the village level and $e_{\text{ij}}$is an error term at the household level.

The variability unexplained by the considered village-level variables is thus estimated in a multilevel approach through the estimation of σin the following form

$$\text{logit}\left(\frac{p_{\text{ij}}}{1-p_{\text{ij}}}\right) = \beta_{00} + \beta X_{\text{ij}} + \delta W_j + \sigma V_j + e_{\text{ij}},$$

where$V_j$follows standard normal distribution $N(0,1)$and hence the variability estimated by the multilevel approach known as "Multilevel Effect" given by the term σV_j. Accordingly, if $V_j = 1$, then there will be an increase in $\text{logit}\left(\frac{p_{\text{ij}}}{1-p_{\text{ij}}}\right)$ byσ. On the other hand, if $V_j = -1$, then there will be a decrease in $\text{logit}\left(\frac{p_{\text{ij}}}{1-p_{\text{ij}}}\right)$ byσ. This is how the consideration of multilevel analysis, in case data involves hierarchical structure, helps in adjusting and getting accurate results.

3. Results and Discussion

3.1 Descriptive Statistics

Table 1 presents the frequency distribution of households wherein one or more members were suffering from any kind of disease (10.4 percent) in eight different states of the northeastern region at any time during the last five years from the reference date of the survey. The percentage of households having any kind of disease is minimum (7.5 percent) in Nagaland and maximum in Arunachal Pradesh which is double in that of Nagaland. From Table 1, it is also seen that 51 percent of the households of Arunachal Pradesh suffer from waterborne diseases out of the total number of households having any kind of disease. On the other hand, Sikkim has the lowest, i.e., 19.3 percent of household suffering from waterborne diseases out of the total number of households affected by any kind of disease.

This study considers all Northeastern States, viz., Assam, Meghalaya, Manipur, Mizoram, Nagaland, Sikkim, Arunachal Pradesh, and Tripura.

Table 1. Household Health Status of different states of North Eastern Region

States	Health status		Total	Waterborne diseases*
	Good health	Disease present		
	Count (Per cent)	Count (Per cent)	Count (Per cent)	Count (Per cent)
Assam	2821 (90.4)	300 (9.6)	3121 (100)	98 (32.7)
Manipur	1509 (89.3)	180 (10.7)	1689 (100)	47 (26.1)
Meghalaya	1079 (87.0)	161 (13.0)	1240 (100)	78 (48.4)
Mizoram	1263 (92.0)	110 (8.0)	1373 (100)	46 (41.8)
Nagaland	1048 (92.5)	85 (7.5)	1133 (100)	35 (41.2)
Sikkim	1164 (89.6)	135 (10.4)	1299 (100)	26 (19.3)
Arunachal Pradesh	1213 (85.5)	206 (14.5)	1419 (100)	105 (51.0)
Tripura	1156 (89.6)	134 (10.4)	1290 (100)	39 (29.1)
Total	11253 (89.6)	1311 (10.4)	12564 (100)	(36.2)

**Percentage distribution of waterborne diseases is done out of the all kinds of diseases of the respective states.*

The analysis is restricted to the morbidity pattern of the persons in a household reported for the last five years from the reference date of the survey. The reported diseases are primarily respiratory disorder (12.9 percent), Diarrhoea and gastroenteritis (7.8 percent), Malaria (8.9 percent), fever (10.6 percent), Heart disease (8.4 percent), cancer (6.2 percent), Jaundice and Cirrhosis of the lever (4.7 percent), Senility (15.9 percent), delivery related disease of the newborn (2.4 percent), and others diseases (Influenza, Typhoid, pneumonia, measles, tetanus, poliomyelitis, diabetes, malnutrition, and others) having 22.2 percent (Table 2).

Households are dichotomized as healthy (disease-free household) and non-healthy (at least one person suffering from any kind of disease anytime during five years before the reference date) concerning selected background characteristics of households. Henceforth we may refer to the healthy household as a good-health household and the non-healthy household as sick-household. The categorization of all the environmental characteristics such as place of residence, type of house, toilet facility, source of drinking water, village-level health facility, the main source of light, and the main source of cooking fuel-saving separate kitchen facility was found to be significant.

Table 2. Distribution of diseases

Name of diseases	Frequency	Per cent
Respiratory disorder	169	12.9
Diarrhoea & Gastroenteritis	102	7.8
Malaria	117	8.9
Fever not classifiable	139	10.6
Heart disease	110	8.4
Jaundice & Cirrhosis of Liver	61	4.7
Cancer (Malignant neoplasm)	81	6.2
Delivery-related disease of newborn	32	2.4
Senility	209	15.9
Others	291	22.2
Total	1311	100.0

Here we want to know whether the type of residence has any relation with the health status of households in terms of selected background characteristics of households. Here our null hypothesis is that the health status of households is independent of the type of place of residence. It is seen that the value of the Pearson Chi-Square test is 35.881 for 1 degree of freedom and the two-sided asymptotic *p*-value is 0.000, which is significant, and we reject our null hypothesis. Also, it is seen that rural households (11.4 percent) are more susceptible to sickness than urban households (7.7 percent). Next, we want to see whether the type of house has any relation to the health status of households. In this case, our null hypothesis is that the type of house and health status of households are independent. It is seen that the value of the

Pearson Chi-Square test statistic is 16.951 for 2 degrees of freedom and the two-sided asymptotic *p*-value is 0.000, which is significant, and we reject our null hypothesis. Now, we want to verify whether the type of toilet has any relation to the health status of households. Here our null hypothesis is that the type of toilet used by the household residents is independent of the health status of households. It is seen that the value of the Pearson Chi-Square test is 30.274 for 2 degrees of freedom and the two-sided asymptotic *p*-value is 0.0001, which is significant, and we reject our null hypothesis. Again, we check whether the source of drinking water has any relation to the health status of households. Here our null hypothesis is that the source of drinking water used by the persons in the households is independent of the health status of households. It is seen that the value of the Pearson Chi-Square test is 12.533 for 3 degrees of freedom and the two-sided asymptotic *p*-value is 0.006, which is significant, and we reject our null hypothesis. Then we want to find whether the health status of households has any relation with where members go for treatment. Here our null hypothesis is that the medical service facility for treatment is independent with healthy and non-healthy households. It is seen that the value of the Pearson Chi-Square test is 14.209 for 2 degrees of freedom and the two-sided asymptotic *p*-value is 0.001, which is significant, and we reject our null hypothesis. Next, we want to study whether the main source of lighting has any relation to the health status of households. Here our null hypothesis is that the main source of lighting is independent with the health status of households. It is seen that the value of the Pearson Chi-Square test is 13.642 for 1 degree of freedom and the two-sided asymptotic *p*-value is 0.000, which is significant, and we reject the null hypothesis. Also, it is seen that households using other than electricity and gas as the main source of lighting (i.e., Kerosene, Oil, and others) (11.7 percent) are more susceptible to diseases than households using electricity and gas (9.6 percent).

Then, we want to verify whether the main cooking fuel has any relation to the health status of households. Here our null hypothesis is that the main cooking fuel is independent of the health status of households. It is seen that the value of the Pearson Chi-Square test is 21.506 for 1 degree of freedom and the two-sided asymptotic *p*-value is 0.000, which is significant, and we reject the null hypothesis. Also, it is seen that the households using other than electricity, LPG, and bio-gas as the main cooking fuel (i.e., wood, crop residues, dung cakes, coal, coke, lignite, charcoal, and others) (11.1 percent) are more susceptible to diseases than the households using electricity, LPG, and bio-gas (7.9 percent). Finally, we want to check whether having a separate kitchen has any relation to the health status of households. Here our null hypothesis is that having a separate kitchen is independent of the health status of households. It is seen that the value of the Pearson Chi-Square test is 0.024 for 1 degree of freedom and the two-sided asymptotic *p*-value is 0.877, which is large, and we have no evidence against the null hypothesis.

Table 3. Tukey's HSD test for pairwise comparisons of between groups

Comparisons	*p*-values (all diseases)	*p*-values (Waterborne diseases)
Type of house		
Pucca vs. semi-pucca	0.198	0.004
Pucca vs. kutcha	0.000	0.000
Semi-pucca vs. kutcha	0.044	0.001
Toilet facility		
Pit toilet vs. flush toilet	0.001	0.000
Pit toilet vs. other	0.011	0.000
Flush toilet vs. others	0.000	0.000
Source of drinking water		
Hand pump vs. well water	0.633	
Hand pump vs. piped water	0.495	
Hand pump vs. others	0.196	
Well water vs. piped water	0.025	
Well water vs. others	0.008	
Piped water vs. others	0.797	
Medical service facilities		
Government vs. private	0.007	0.007
Government vs. others	0.157	0.996
Private vs. others	0.002	0.231

Dependent variable: Health status

From the chi-square test, it is not possible to find out which group of toilet facilities, source of drinking water, and medical service facilities are significantly different for all diseases and as well as for waterborne diseases, but in this case, waterborne diseases were found to be insignificant. It needs to perform multiple comparison analyses [9, 24]. Here we apply Tukey's honestly significant difference (HSD) test for pairwise comparisons of different groups for covariates under this analysis. The significant difference between pairs of groups (more than two) of a covariate as indicated by *p*-values is shown in Table 3.

Similarly, the households can be dichotomised as healthy and suffering at least one person suffering from waterborne disease for the last five years before the reference date, concerning selected background characteristics of the household. According to the above, it is found from the Pearson Chi-Square test that all the explanatory variables, viz., Type of residence, Type of house, Toilet facility, Medical service facilities, Main source of lighting, Main cooking oil saving Source of drinking water and Separate room used as a kitchen are found to be significant at 5 percent level of significance.

Table 4. Logistic analysis of healthy and non-healthy households

Category	Odd ratio	S.E.	Sig.	95.0% C.I. for odd ratio	
				Lower	Upper
Type of residence					
Urban	1				
Rural	1.357***	.087	.000	1.144	1.611
Type of house					
Pucca	1				
Semi-pucca	1.036	.104	.731	.846	1.270
Kachha	1.109	.105	.321	.904	1.362
Source of drinking water					
Piped water	1				
Hand pump	.761***	.087	.002	.641	.902
Well water	.649***	.091	.000	.543	.776
Others	.876*	.081	.100	.748	1.027
Type of toilet					
Flush toilet	1				
Pit toilet	1.146	.092	.138	.957	1.372
Others	1.286**	.109	.021	1.038	1.593
Medical Service Facilities					
Govt. health facility	1				
Private health facility	.924	.082	.338	.787	1.086
Others	1.199*	.108	.094	.970	1.482
Main source of lighting					
Electricity & Gas	1				
Others	1.107	.072	.157	.962	1.274
Main cooking fuel					
Electricity, LPG & Bio-Gas	1				
Others	1.085	.102	.425	.888	1.324
Separate kitchen					
No	1				
Yes	1.067	.064	.313	.941	1.210
Constant	.075***	.112	.000		

*Here the symbols *,**, and *** represent the level of significance at 10 percent, 5 percent and 1 percent respectively.*

3.2 Logistic Regression Analysis:

In this study, households are dichotomized as healthy (disease-free during the last five years) and non-healthy (at least one of the members was reported as suffering from some kind of disease during the last five years). It is reported that 10.5 percent of the total households in the northeastern region are non-healthy households. In this study, the healthy and non-healthy status of the household is considered as the response variable. Logistic regression is suitable for understanding the relationship between this response variable and some explanatory variables considered in the following Table 4.

Table 5. Logistic analysis of healthy and households suffering from waterborne disease

Category	Odd ratio	S.E.	Sig.	95.0% C.I. for odd ratio	
				Lower	Upper
Type of residence					
Urban	1				
Rural	.741**	.152	.049	.550	.998
Type of house					
Pucca	1				
Semi-pucca	.672	.206	.050	.449	1.006
Kachha	.543***	.204	.003	.364	.809
Source of drinking water					
Piped water	1				
Hand pump	1.761***	.148	.000	1.317	2.355
Well water	1.461***	.137	.006	1.116	1.912
Others	1.373**	.130	.015	1.064	1.770
Type of toilet					
Flush toilet	1				
Pit toilet	.672**	.171	.020	.481	.939
Others	.506***	.191	.000	.348	.734
Medical Service Facilities					
Govt. health facility	1				
Private health facility	1.123	.138	.400	.857	1.472
Others	1.090	.187	.645	.756	1.572
Main source of lighting					
Electricity & Gas	1				
Others	.875	.112	.233	.702	1.090
Main cooking fuel					
Electricity, LPG & Bio-Gas					
Others	.726*	.194	.100	.496	1.063
Separate kitchen					
No	1				
Yes	.911	.102	.360	.746	1.113
Constant	84.023***	.229	.000		

*Here the symbols *,**, and *** represent the level of significance at 10 percent, 5 percent, and 1 percent respectively.*

The fitted logistic regression equation for the general health status of households is written as follows:

$$logit\left(\frac{\Pr\{healthy\ household\}}{1-\Pr\{healthy\ household\}}\right)=-2.587+0.306\ place\ of\ residence_{rural}$$
$$+\ 0.036\ type\ of\ house_{semi-pucca}+0.104\ type\ of\ house_{kuchha}-0.274\ Source$$
$$of\ drinking\ water_{hand\ pump}-0.432\ Source\ of\ drinking\ water_{well\ water}$$
$$-0.132\ Source\ of\ drinking\ water_{others}+0.136\ Type\ of\ toilet_{pit\ toilet}$$
$$+\ 0.252\ Type\ of\ toilet_{others}-0.09\ Where\ do\ go\ for\ treatment_{private}$$
$$+\ 0.181\ Where\ do\ go\ for\ treatment_{others}+0.101\ Main\ source\ of\ lighting_{others}$$
$$+0.081\ Main\ cooking\ fuel_{others}+0.065\ Separate\ kitchen_{yes}$$

Explanatory variables such as place of residence, source of drinking water, type of toilet, and place of treatment were found to be significantly associated with morbidity. The rural population in the northeastern region was 36% more likely to suffer from any kind of disease in comparison to that of the urban population. Surprisingly, it is observed that people using other than piped drinking water have less likely to suffer from any sort of disease. We cannot provide any plausible explanation for such a type of result. In a recent study [21] it is reported that the quality of water (hard or soft) has a definite impact on disease susceptibility. In contrast when we consider only the waterborne disease then the population consuming piped drinking water is less likely to be sick.

The estimated logistic regression equation for waterborne disease households is written as follows:

$$logit\left(\frac{\Pr\{healthy\ household\}}{1-\Pr\{healthy\ household\}}\right)=4.431-0.300\ place\ of\ residence_{rural}$$
$$-\ 0.398\ type\ of\ house_{semi-pucca}-0.611\ type\ of\ house_{kuchha}+0.566\ Source$$
$$of\ drinking\ water_{hand\ pump}-0.379\ Source\ of\ drinking\ water_{well\ water}$$
$$+\ 0.317\ Source\ of\ drinking\ water_{others}-0.397\ Type\ of\ toilet_{pit\ toilet}$$
$$-\ 0.683\ Type\ of\ toilet_{others}+\ 0.116\ Where\ do\ go\ for\ treatment_{private}$$
$$+\ 0.086\ Where\ do\ go\ for\ treatment_{others}-0.320\ Main\ source\ of\ lighting_{others}$$
$$-\ 0.320\ Main\ cooking\ fuel_{others}-0.093\ Separate\ kitchen_{yes}$$

3.3 Multilevel Logistic Analysis

Household with pit toilet/kachha toilet were 15% /27% more likely to have a sick person as compared to those with flush toilets.

Explanatory variables such as type of residence, source of drinking water, type of toilet, main cooking fuel, and type of house were found to be significantly associated with morbidity due to waterborne diseases. The rural population was at lower risk of suffering from waterborne diseases in comparison to their counterparts living in urban areas. The rural population was at a 48% higher risk of suffering from any sort of disease in comparison to those of urban areas.

Table 6. Multilevel logistic analysis of healthy and non-healthy households

Category	Odd ratio	S.E.	Sig.	95.0% C.I. for odd ratio	
				Lower	Upper
Type of residence					
Urban	1				
Rural	1.483***	0.113	.000	1.188	1.851
Type of house					
Pucca	1				
Semi-pucca	0.99	0.107	.720	0.803	1.221
Kachha	1.041	0.11	.311	0.839	1.291
Source of drinking water					
Piped water	1				
Hand pump	0.79***	0.095	.001	0.656	0.951
Well water	0.685***	0.097	.000	0.567	0.829
Others	0.912**	0.087	.050	0.769	1.082
Type of toilet					
Flush toilet	1				
Pit toilet	1.145	0.095	.125	0.95	1.379
Others	1.273**	0.113	.016	1.02	1.588
Medical Service Facilities					
Govt. health facility	1				
Private health facility	0.931	0.087	.238	0.785	1.105
Others	1.192*	0.113	.085	0.956	1.488
Main source of lighting					
Electricity & Gas	1				
Others	1.102	0.076	.150	0.949	1.279
Main cooking fuel					
Electricity, LPG & Bio-Gas	1				
Others	1.067	0.107	.400	0.865	1.316
Separate kitchen					
No	1				
Yes	1.089	0.067	.215	0.955	1.241
Constant	0.071	0.134	.000		

*Here the symbols *,**,*** represent the level of significance at 10 percent, 5 percent and 1 percent respectively.*

Binary multilevel logistic regression equation is given as follows:

$$genhealth_{household,tehsil} = Binimial(denom_{household,tehsil}, \pi_{household,tehsil})$$

$$\begin{gathered} logit(\pi_{household,tehsil)} = -2.648 + \beta_{1,tehsil}\, cons + 0.394\, Rural_{household,tehsil} \\ -0.01\, Semi - pucca_{household,tehsil} + 0.04\, kachha_{household,tehsil} - 0.236\, hand \\ pump_{household,tehsil} - 0.378\, well\ water_{household,tehsil} - 0.092\, Sdrink: \\ others_{household,tehsil} - 0.135\, Pit\ toilet_{household,tehsil} + 0.241\, Toilet: \\ others_{household,tehsil} - 0.071\, Private\ health\ facility_{household,tehsil} + 0.176 \\ Health\ facility_{household,tehsil} + 0.097\, Slight: others_{household,tehsil} + 0.065 \\ Cookfuel: others_{household,tehsil} + 0.085\, Separate\ Kitchen: Yes_{household,tehsil}, \end{gathered}$$

where $\beta_{1,tehsil} = 0.000 + u_{1,tehsil}$

$$\begin{bmatrix} u_{0,tehsil} \\ u_{1,tehsil} \end{bmatrix} \sim N(0, \Omega_u): \Omega_u = \begin{bmatrix} 0.109 & \\ 0.000 & 0.000 \end{bmatrix}$$

Among village-level variables, households having access to the private health facility in the village is 8 percent less likely and those having access to other health facilities (quack and other traditional medical facilities) are 20 percent more prone to sickness in comparison to the village having government health facilities.

Table 7. Multilevel logistic analysis of healthy and households suffering from waterborne diseases

Category	Odd ratio	S.E.	Sig.	95.0% C.I. for odd ratio	
				Lower	Upper
Type of residence					
Urban	1				
Rural	0.728**	0.16	.035	0.532	0.997
Type of house					
Pucca	1				
Semi-pucca	0.685**	0.206	.045	0.458	1.026
Kachha	0.562***	0.208	.002	0.374	0.845
Source of drinking water					
Piped water	1				
Hand pump	1.732***	0.16	.000	1.265	2.369
Well water	1.449***	0.148	.002	1.084	1.937
Others	1.265***	0.138	.005	0.965	1.658
Type of toilet					
Flush toilet	1				
Pit toilet	0.703**	0.17	.015	0.504	0.981
Others	0.567***	0.193	.000	0.388	0.827
Where do go for treatment					
Govt. health facility	1				
Private health facility	1.127	0.143	.350	0.852	1.492
Others	1.089	0.192	.540	0.747	1.586
Main source of lighting					
Electricity & Gas	1				
Others	0.857	0.118	.230	0.68	1.08
Main cooking fuel					
Electricity, LPG & Bio-Gas	1				
Others	0.739*	0.198	.090	0.501	1.089
Separate kitchen					
No	1				
Yes	0.856	0.106	.350	0.696	1.054
Constant	82.85***	0.255	.000		

*Here the symbols *,**,*** represent the level of significance at 10 percent, 5 percent and 1 percent respectively.*

Binary multilevel logistic regression equation for the response variable waterborne diseases is given as follows

$$waterborne\ disease_{household,tehsil} = Binimial(denom_{household,tehsil}, \pi_{household,tehsil})$$

$$\begin{gathered}logit(\pi_{household,tehsil)} = \ 4.417 + \ \beta_{1,tehsil}\ cons - 0.317\ Rural_{household,tehsil} \\ -0.378\ Semi - pucca_{household,tehsil} - 0.576\ kachha_{household,tehsil} + 0.549\ hand \\ pump_{household,tehsil} + 0.371\ well\ water_{household,tehsil} + 0.235\ Sdrink: \\ others_{household,tehsil} - 0.352\ Pit\ toilet_{household,tehsil} - 0.568\ Toilet: \\ others_{household,tehsil} + 0.120\ Private\ health\ facility_{household,tehsil} + 0.085 \\ Health\ facility_{household,tehsil} - 0.154\ Slight: others_{household,tehsil} - 0.303 \\ Cookfuel: others_{household,tehsil} - 0.155\ Separate\ Kitchen: Yes_{household,tehsil},\end{gathered}$$

where $\beta_{1,tehsil} = 0.000 + \ u_{1,tehsil}$

$$\begin{bmatrix} u_{0,tehsil} \\ u_{1,tehsil} \end{bmatrix} \sim N(0, \Omega_u): \Omega_u = \begin{bmatrix} 0.290 & \\ 0.000 & 0.000 \end{bmatrix}$$

The lone village level variable was found to be significantly associated with disease prevalence under the multilevel model. This sentence does not make any sense. Which variable? Moreover, only 11 percent of the variability in disease prevalence (all diseases) could not be explained by the considered set of covariates. Table 8 presents results under multilevel analysis using the same set of covariates as described above. Comparing the results under both analytical methods, the results were virtually equivalent in terms of the same direction under traditional logistic regression analysis with those obtained under multilevel analysis. As under traditional regression analysis, the variables residence, source of drinking water, and type of toilet were found to be significantly associated with disease prevalence also under multilevel analysis. However, in traditional analysis, the village-level variable is only significant with other health facilities. The main source of lighting and the main source of cooking fuel was found to have an insignificant effect on morbidity status. The village-level variable namely the private health facility in the village had an insignificant impact on disease prevalence.

In multilevel analysis also the source of drinking water had a strong impact on the waterborne disease prevalence. Households having hand pump/well water as a source of drinking water have 73/45 percent higher risk of suffering from waterborne diseases in comparison to that of piped water.

Whether people are healthy or not is determined by many different factors including their circumstances, behaviour, and environment. This paper identifies a few environmental factors affecting disease prevalence. It is suggested that the community at the Panchayat level should be empowered to control the few key environmental determinants viz., the supply of drinking water, sanitation system, and health service facility for health promotion in this part of India.

In this study, the variables for the diseases reported in NFHS 2 were seen as respiratory disorder, diarrhea and gastroenteritis, fever, jaundice, and diseases for newborn babies. To study the environmental factors affecting health, these factors played an important role. But, in the later part of the NFHS surveys, many of the variables were missing. So, in my opinion, the importance of the study with NFHS 2 survey data lies here. A few recent works have been noticed in [1], [3], [5] and [11].

Acknowledgement

The authors are extremely thankful to Professor A. Gore for constructive suggestions which led to the clarity of the presentation and the anonymous reviewer for making this paper into presentable form.

References

Connolly, M. A., Gayer, M., Ryan, M. J., Salama, P., Spiegel, P., & Heymann, D. L. (2004). Communicable diseases in complex emergencies: Impact and challenges. *The Lancet, 364*(9448), 1974-1983. https://doi.org/10.1016/S0140-6736(04)17481-3

Dale, P., Sipe, N., Anto, S., Hutajulu, B., & Ndoen, E. (2005). Malaria in Indonesia: A summary of recent research into its environmental relationships. *Southeast Asian Journal of Tropical Medicine and Public Health, 36*(1), 1-13.

Doherty, J. (2000). Establishing priorities for national communicable disease surveillance. *Canadian Journal of Infectious Diseases, 11*(1).

Dyer, O. (2003). Environmental hazards kill five million children a year. *British Medical Journal (BMJ), 326*(7393), 782. https://doi.org/10.1136/bmj.326.7393.782

Edemekong, P. F., & Huang, B. (2023). Epidemiology of prevention of communicable diseases. *StatPearls*. StatPearls Publishing. https://doi.org/10.1016/j.statpearls.2023.10.024

Goldstein, H. (1995). *Multilevel statistical models*. Halsted Press.

Goldstein, H. (2003). *Multilevel statistical models*. Oxford University Press Inc.

Goldstein, H., & McDonald, R. (1998). A general model for the analysis of multilevel data. *Psychometrika, 53*(3), 455-467. https://doi.org/10.1007/BF02294437

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. John Wiley & Sons.

Hox, J. J. (1995). *Applied multilevel analysis*. TT-Publikaties.

Jafari, N., Shahsanai, A., Memarzadeh, M., & Loghmani, A. (2011). Prevention of communicable diseases after disaster: A review. *Journal of Research in Medical Sciences, 16*(7), 956-962.

Keraka, M. N., & Wamicha, W. N. (2003). Child morbidity and mortality in slum environments along Nairobi River. *Eastern Africa Social Science Research Review, 19*(1), 41-57.

Last, J. (1995). *A dictionary of epidemiology* (3rd ed.). Oxford University Press.

International Institute for Population Sciences (IIPS). (1998-1999). *National Family Health Survey*. Bombay: IIPS.

Pain, Pardoe. (2003). Model assessment plots for multilevel logistic regression. *Computational Statistics & Data Analysis*.

Sack, R. B., Siddique, A. K., Longini, I. M., Jr., Nizam, A., & Yunus, M. (2003). A 4-year study of the epidemiology of Vibrio cholera in four rural areas of Bangladesh. *Journal of Infectious Diseases, 187*(1), 96-101. https://doi.org/10.1086/367669

Saathoff, E., Olsen, A., Kvalsvig, J. D., Appleton, C. C., & Sharp, B. (2005). Ecological covariates of *Ascaris lumbricoides* infection in schoolchildren from rural KwaZulu-Natal, South Africa. *Tropical Medicine and International Health, 10*(5), 412-422. https://doi.org/10.1111/j.1365-3156.2005.01414.x

Srivastava, H. C., & Mishra, N. R. (2004). A comparative study of morbidity pattern and life style among the elderly in Kerala and Andhra Pradesh. Paper presented at the International Seminar on Demographic Changes and Implications, Department of Demography, University of Kerala, Trivandrum, India.

Tipayamongkholgul, M., Podhipak, A., Chearskul, S., & Sunakorn, P. (2005). Factors associated with the development of tuberculosis in BCG immunized children. *Southeast Asian Journal of Tropical Medicine and Public Health, 36*(1), 145-150.

Wakou, B. B. (2005). An examination of the combined effects of maternal characteristics, environment and treatment programs on the prevalence of diarrhea amongst infants and children in Uganda. *Population Review, 44*(2), 51.

Watts, C., Fawell, J., Sartory, D., Leaman, J., & Tuffin, A. (2006). Evaluation of the Drinking Water Quality and Health Research Programme (1996-2004) for Defra. Report. Watts and Crane Associates.

World Health Organization (WHO). (1992). *Panel report on food and agriculture*. Geneva: World Health Organization.

Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80*(391), 513-524. https://doi.org/10.1080/01621459.1985.10478158

Zar, J. H. (1999). *Biostatistical analysis* (4th ed.). Prentice Hall.

Estimating the Duration of Postpartum Amenorrhea through Bayesian Approach

Abstract

Postpartum amenorrhea (PPA) is the interval between the termination of a woman's pregnancy and the resumption of menstruation, during which conception does not occur. PPA plays a crucial role in birth spacing and fertility. Understanding the factors influencing PPA duration is essential for family planning and reproductive health programs. Previous studies have reported variations in PPA duration across different populations, but there is a lack of research on PPA in the specific context of Manipur, India. The objective of this study was to investigate the relationship between mother's age and PPA duration in women residing in Manipur, India, using a Bayesian approach. Additionally, the study aimed to examine the variability of PPA duration based on religion and family status by applying zero-inflated models. The utilization of Bayesian techniques and longitudinal data analysis would provide valuable insights into the determinants of PPA duration in this population. The analysis of longitudinal data from 1296 eligible women in rural areas of Manipur revealed that the average duration of PPA was 6.6 months. There was no significant difference in PPA duration among different districts of Manipur. The study identified a positive relationship between mother's age and PPA duration, indicating that older mothers tend to have longer PPA. Furthermore, religion and family status were found to influence PPA duration, with variations observed between different subgroups.

The study contributes to the understanding of PPA duration in women residing in Manipur, India. The findings highlight the importance of considering the mother's age, religion, and family status in predicting PPA duration. These insights can inform family planning interventions and reproductive health programs in the region. The application of Bayesian modeling techniques allowed for a comprehensive analysis of PPA duration and provided robust results. Further research is warranted to explore additional factors influencing PPA and to validate these findings in other populations.

Keywords: PPA, cross-sectional data, ZIP model, Manipur

1. Introduction

In the field of statistical modeling, the Bayesian approach has gained widespread acceptance, particularly with the advent of WinBUGS software. Ghosh and colleagues [1] have introduced a Bayesian Zero-Inflated Poisson (ZIP) model specifically designed for cross-sectional data. Additionally, a semi-parametric ZIP model has been developed to enhance the Poisson component [2]. Various studies [3] have explored the application of zero-inflated models, including the Poisson hurdle model and the Zero-alter model, highlighting their versatility and effectiveness. The Bayesian methodology has also been employed in mixed distribution models for analyzing cross-sectional data [4]. Furthermore, it has been utilized in zero-inflated models to examine NFHS-3 data, focusing on child mortality distribution [5]. These advancements underscore the robustness and applicability of Bayesian techniques in addressing complex statistical challenges. Moreover, most population-based studies addressing women's reproductive health issues rely on cross-sectional analyses rather than longitudinal data analysis. Postpartum amenorrhea (PPA), a physiological process following each conception, refers to the interval between the end of a woman’s pregnancy and the onset of her next menstrual cycle. During this period, conception does not occur. As it tends to increase the birth interval and hence to reduce women's fertility over her life span, especially in societies where the use of contraceptive methods is not widespread. It depends on a number of factors which vary from woman to woman in a population [6-13]. The most important conclusion of these studies is that breast feeding increases the duration of PPA. This chapter delves into the application of the Bayesian approach to address the complexities of longitudinal zero-inflated data concerning women living in Manipur, a state in eastern India that shares its border with Myanmar. By focusing on this specific population, the chapter aims to provide a nuanced understanding of how Bayesian techniques can effectively manage and interpret the unique statistical challenges presented by this demographic. The study utilizes a longitudinal dataset that captures information on successive pregnancies. The primary objective of this research is two-fold: firstly, to investigate the relationship between a mother's age and the duration of postpartum amenorrhea using a Bayesian approach with cluster models, and secondly, to examine the variability in the duration of postpartum amenorrhea among mothers based on their religion and family status by employing zero-inflated models. By employing these methods, the study aims to gain insights into the factors influencing postpartum amenorrhea duration in this population.

2. Materials and Methods

The study is based on the retrospective reporting data of 1296 eligible women surveyed during nine months (April-December, 2009) in rural areas of Manipur valley. Summary measures, including the mean and median, have been calculated to analyze the data. The duration of postpartum amenorrhea (PPA) has been studied in relation to the mother's age, incorporating prior assumptions to enhance the analysis. This approach allows for a more detailed examination of how maternal age influences the length of PPA, providing valuable insights into the reproductive health patterns of the population under study. A survival analysis technique has been applied through Bayesian approach to identify the variability in PPA. The zero inflated models corresponding to the number of children have been dealt with Bayesian approach under longitudinal data analysis. The longitudinal observation of the duration of PPA is Y_{ij} for the p-dimensional covariates (X_{ij}) for the i^{th} indexed and j^{th} units within clusters. The within-cluster and between-cluster specific regression coefficients can be expressed as the function of E(Y). Larmbert's [14] zero mixed un-truncated Poisson distribution is further applied in this analysis. The frequencies of PPA durations are denoted with Y. The individual and time-specific data are also separated with Y_{ij}.

2.1 Cluster Models

The longitudinal response says the duration of PPA is Y_{ij} for the p-dimensional covariates (X_{ij}) for the i^{th} indexed and j^{th} units within clusters. The within-cluster and between-cluster specific regression coefficients can be expressed as the function of E(Y). The widely used cluster-specific approaches are GLMMs in the class of generalized linear models by including random effects in the linear predictor by given a vector b_i of parameters to the i^{th} cluster, for the j^{th} unit, the conditional density of Y_{ij} is of the form

$$f(Y_{ij}|b_i,X_{ij},Z_{ij},\beta)=\exp[\{Y_{ij}\theta_{ij}-c(\theta_{ij})\}\phi+d(y_{ij},\phi)] \quad \dots (1)$$

where θ_{ij} is the canonical parameter, c and d are functions of known form and ϕ is a positive scale factor. McCullagh and Nelder (1989) have formulated the θ_{ij} as the function of the parameter β. Neuhaus et al. (2006) have assume that

$$E(Y_{ij}/b_i,x_{ij},z_{ij})=g^{-1}(z_{ij}b_i+x_{ij}\beta) \quad \dots (2)$$

Where x_{ij} is the design matrix corresponding to β, z_{ij} is the design matrix corresponding to $b_i$and g is a monotonic differentiable function. Given b_i, the model assumes that the responses $Y_{i1}, \dots, Y_{ini}$ for i^{th} individuals n_i^{th} PPA duration are independent. In this work, the random effects b and $G(\acute{b}D)$ have been assumed, where D is the independent parameter. To compute the model we have assumed the distribution for the random effects b, G(b/D), which is independent to the parameters value D. Let the response observation $Y_i= (Y_{i1},\dots.Y_{ini})$, the duration of PPA are extended to n_i-dimensional vectors. The likelihood has been assumed there with m independent cluster with n_i units by

$$L(\beta,D)=\prod_{i=1}^{m}\int_b f\left(y_i\middle|b,x_i,z_{i,}\beta\right)dG(b|D) \quad \dots (3)$$

$$=\prod_{i=1}^{m}\int\prod_{j=1}^{n_i}f(yi|b,xij,xij,\beta)dG(b\vee D) \quad \dots (4)$$

2.2 Zero-Inflated Models

In 1992, Lambert developed the zero mixed un-truncated Poisson distribution, a significant advancement in statistical modeling. The formulation of this model is based on

$$P(Y_i=0)=(1-p)+pe^{-\mu},\ 0\underline{<}p\underline{<}1 \quad \dots (5)$$

$$P(Y_i=0)=p\frac{\mu^k e^{-\mu}}{k!},\ k=1,\dots\alpha,\ 0<\mu<\alpha \quad \dots (6)$$

We extended the model in equation (7.6) by i^{th} individuals and j^{th} observation with

$$(1-p_{ij})+p_{ij}e^{-\mu ij}=1-p_{ij}(1-e^{-\mu ij})=1-\theta_{ij} \quad \dots (7)$$

where, i, j is used for i^{th} individual j^{th} observation. Heilbron (1994) has separated the "zero-altered model" in two parts by $P(Y_i=0;\mu_{1i})$ and $P(Y_i=y_i;\mu_{1i})$ for $y_i>0$. The model can be modified into

$$P(Y_i=0;\mu_{1i})=e^{-\mu 1i},\ \log(\mu 1i)=x_{i'}\beta_1 \quad \dots (8)$$

$$\text{and} \quad P(Y_i=y_i;\mu_{1i})=e^{-\mu 1i},\ \log(\mu 1i)=x_{i'}\beta_1 \quad \dots (9)$$

$$P(Y_i=y_i/y_i>0;\mu_{2i})=\frac{y_i{}^{-\mu_{2i}e^{-\mu_{2i}}}}{y_i!(1-e^{-\mu_{2i}})},log(\mu_{2i})=x_i'\beta_2 \quad \dots (10)$$

In 1994, Heilborn developed a model specifically designed to test the hypothesis regarding the presence of zero inflation in data sets. This model operates under the assumption that

$$\log(\mu_{1i}) = \beta+\log(\mu_{2i}) \quad \dots (11)$$

If $\beta=0$, the model conforms to the standard Poisson distribution. When $\beta<1$, the data exhibit zero inflation, indicating an excess of zero counts. Conversely, if $\beta>$n 1, the data are zero deflated, reflecting a scarcity of zero counts. Brain and colleagues (2010) have thoroughly discussed the implications and applications of zero-inflated count data by

$$P(Y_i=0)=1-p_i,\ 0\leq p\leq 1 \qquad \ldots (12)$$

$$P(Y_i=0)=p\frac{\mu^k e^{-\mu}}{k!},\ k=1,\ldots,\alpha,\ 0<\mu<\alpha \qquad \ldots (13)$$

If (1-p)>$e^{-\mu}$, the data is inflated by zero and when (1-p)<$e^{-\mu}$ the data is deflated with zero. When p=1, the data contain no zero counts, aligning with the characteristics of a standard Poisson distribution. The above-mentioned models are applied in the analysis on the duration of PPA. The frequencies of PPA durations are denoted with Y. The individual and time-specific data are also separated with Y_{ij}.

3. Analysis

The duration of PPA and mother age as a subset of data from a more extensive longitudinal study involves the duration of PPA to n = 108 mothers, all of whom have at least T = 5 live births. The analysis here is focused on the impact of duration of PPA that is Y_{it}, ith mothers duration of PPA at tth baby born on the mother's age at birth that is w_{it}, ith mothers with tth baby born and the extent to which there is heterogeneity in the overall smooth by a function of S(w_{it}).

Thus for each five birth history for ith mother, the stipulated equations are like

$$y_{it}=\beta+b_{1i}+S(w_{it})+b_{2i}S(w_{it})+u_{it} \qquad \ldots (14)$$

where (b_{1i}, b_{2i}) is assumed to follow the Normal distribution with N(0,D). The hyper parameter D-1 is further assumed to follow the Wishart prior with an identity scale matrix and two degrees of freedom. The random walk smooth is estimated over all (i, t) pairs using a normal prior with a single variance parameter, rather than the basis of successive ages within each fertility sequence, which would permit distinct variance parameters for each subject. A two chain run of 5000 iterations shows early convergence, with significant heterogeneity in the b_{2i}, namely a posterior mean for var(b_2) of 1.35, and 95% credible interval (0.95, 1.74). The estimated posterior mean of the β_{1i} is 0.009 with 95%CIs (0.010, 0.009) and β_{2i}, the estimated posterior mean is 0.33 with 95%CIs (0.472, 0.188). It shows that β_{2i} has some useful effect on Y_{it} and β_{2i} has no effect on Y_{it}. The analysis has been carried out through the present data. Here, we assume the distributions of $\beta_{i1,}$ $\beta_{i2,}$ and β_0 to be conventional non-informative prior. In the present data, Wishart prior for the deviation index is consistent with continuous and binary outcome measures (Nehulaus et al., 2006). In order to derive the posterior distribution of $\beta_{0,}$ we use the posterior mean value of the deviation index and generate the 1000 random sample values. The MCMC algorithm has been used in Gibbs sampling through the WinBUGS software. The algorithm has been run for 50,000 iterations, which is composed of a burn-in of 25,000 iterations and posterior inference based on the next 25,000 iterations which were thinned at 25 in order to have 1000 Pseudo-independent posterior sample values.

The posterior mean and 95% credible interval (CIs) for β_0, $\beta_{1,}$ and β_2 are presented in Table 2 along with the resulting regression coefficients. For this Model the DinV[1, 1] of the posterior mean is 97.21 with large CIs (132,49.93). Similarly, the between-cluster DinV[1, 2] had a relatively narrow CIs(5.04,5.26). The D [1, 1] had a posterior mean of 0.01 and a 95 percent CI of (0.02, 0.02). For the difference in model 2, the conclusions are similar, with a similar posterior mean for the within and between regression coefficients. It can be concluded that the posterior mean for the between-cluster coefficient is less in comparison to the within-subject regression coefficient. Table 1 depicts the estimates of the duration of postpartum amenorrhea as calculated by the current status method for women at their different ages. The findings have been compared with those of earlier studies to indicate any change in the pattern of postpartum amenorrhea in the population for different subgroups though the levels may differ due to differences in data sources and methods employed. In this work, the density function f(x) has been assumed to be a mixture distribution. The distribution of f(x) is again expressed as the equation by

$$f(x)=wf_1(x)+(1-w)f_2(x),\ \text{where } 0<w<1 \qquad \ldots (15)$$

Here, $f_1(x)$ and $f_2(x)$ are assumed to follow the beta distribution and w is the additive parameter to separate f(x) into a mixture distribution of functions $f_1(x)$ and $f_2(x)$ respectively. The ***vague*** prior is used to reduce the effect of prior influences on the posterior mean. The gamma distribution has been obtained for $f_1(x)$ and $f_2(x)$. The value of w has been computed through a simulation procedure. The posterior mean of PPA duration has been obtained by w=0.42. A total of 10000 burns have been carried out to obtain the mean value of PPA. The computed age-wise PPA durations have been shown in Table 5.

4. Findings

The average duration of postpartum amenorrhea (PPA) has been observed to be approximately 6.6 months. There is no significant difference in postpartum amenorrheic periods in different districts of Manipur. Among the four districts, the duration of amenorrhea ranged from a high of 6.6 months among women from Imphal West districts to a visibly low of 6.4 months in Imphal East. The mean length of amenorrhea was also observed to increase from 6.4 months for women for their first child to 6.9 months for the fifth baby's birth in Imphal West district. The postpartum amenorrhoeic duration of the observed women has also increased from 5.4 months for 1st child to 6.6 months for 5th child in the Thoubal District. The women of Imphal East and Bishnupur have reduced the mean duration of postpartum amenorrhea from 5.3 to 3.9 and from 6.7 to 6.3 months in 1st live birth to 5th live birth shown in Table-4. The duration of PPA plays an important role in controlling the birth spacing between two children in an overpopulated country like India.

Table - 8 presents insightful data regarding the percentage of zero counts and mean zero count across different durations of postpartum amenorrhea (PPA) and types of family structures. Notably, at the first delivery, joint families exhibit a higher percentage of zero counts (6.32% compared to 4.55% in nuclear families), although nuclear families tend to have longer observed durations of PPA. Table 2 illustrates a non-linear trend over time in PPA durations. Among joint families, zero observations notably decrease during the 4th and 5th births. Demographic characteristics of women, including family status, adoption of sterilization methods, and religious affiliation, are detailed in Table 7. The majority of women (57.5%) are from nuclear families, with a small proportion (2.2%) having adopted sterilization methods. Hindu women comprise 85% of the sample, followed by 12.1% identifying as Meitei.

Comparing estimates from Table - 6, it is evident that the estimated values of the Binomial and Poisson distributions are identical. Specifically, the average β11 is 0.27 (1.1) for the Binomial distribution and 1.45 (0.09) for the Poisson distribution, indicating strong agreement between the two methodologies. Incorporating prior evidence into the Zero-Inflated Poisson (ZIP) model estimates for β12 and β22 are -2.03 (0.98) in the Binomial model and 0.09 (0.12) in the Poisson model. The Deviance Information Criterion (DIC) values further illustrate the superior fit of the Zero-Altered Poisson (ZAP) model (234.67) compared to the ZIP model (286.32) and Hurdle model (293.65), as obtained through WINBUGS software. Table 6 underscores that including the adoption of sterilization in the Binomial distribution yields results comparable to those of the Poisson distribution within the ZIP model. The posterior distributions in mean (SD) across different models reveal consistent outputs. For instance, parameter estimates and standard deviations (SD) are nearly identical for the Binomial distribution in relation to family status within the ZIP model.

Examining the adoption of sterility, the posterior distributions of β28 are estimated at -0.14 (0.14) and 1.13 (0.57) for the Binomial and Poisson distributions, respectively. This indicates a negative relationship in the Binomial model and a positive relationship in the Poisson model with respect to PPA. These contrasting outcomes highlight the divergent conclusions drawn by each model on this aspect. Furthermore, the posterior distributions for caste and time of delivery yield estimates of 0.68 (0.94) and 2.42 (31.9) for the Binomial and Poisson distributions, respectively. Meanwhile, the posterior mean (SD) for variance components ρ, σ1, and σ2 are found to be -0.04 (0.09), 1.29 (0.3), and 1.65 (0.16), respectively. These findings provide a comprehensive view of the statistical analyses conducted, shedding light on the nuanced relationships and variances explored within the study framework.

5. Discussion

The average duration of postpartum amenorrhea (PPA) was found to be 6.6 months in this study. However, previous research has reported considerable variation in PPA duration across different countries. For example, studies from Bangladesh have documented prolonged lactational amenorrhea lasting from 12 to 17 months [15-17]. A retrospective study conducted in the Philippines reported an average PPA duration of 8.5 months [18-20]. In contrast, some developed countries have reported median durations as low as 3 months [21]. Among Nepalese mothers residing in rural areas, the tri-mean of the amenorrheic period was found to be 9.6 months, with a median of 8.4 months and a mean of 10.4 months [22]. On the other hand, Bangladesh exhibited a longer mean amenorrheic period of 12.6 months [23].

Factors such as parity, age, survival status of the child, breastfeeding practices, and socioeconomic status of the mothers have been identified as influential factors in the timing of amenorrhea among Nepalese mothers [24]. The present findings support the notion that a mother's age significantly affects the duration of PPA. This study extends existing approaches to analyzing PPA duration in women under observation [25-26], using cluster-specific modeling on response observations. The methods employed here can be further extended to include different covariates related to PPA among women aged between 15 and 42 years.

One advantage of the proposed methods is their ability to incorporate prior information, which allows for posterior inference in cluster-specific modeling. Alternatively, multivariate analysis using different distribution assumptions, such as frequentist approaches, could prove useful. The results of the relationship between the duration of PPA and mothers' age demonstrate that the choice of link function and the assumed distribution of random effects can influence the functional form of the relationship.

Table 1. Observed distribution of PPA with respect to mothers Age

PPA	Age of mother (in year)						
(in month)	Upto 20	20 to 25	25 to 31	31 to 35	36 to 40	41 to 45	Total
1	3	12	5	1	1	0	22
2	3	3	1	0	0	0	7
3	59	111	95	28	3	1	297
4	5	5	2	2	1	0	15
5	0	4	2	0	0	0	6
6	1	4	4	3	0	0	12
7	3	2	4	1	1	0	11
8	1	8	2	1	0	0	12
9	1	0	0	0	0	0	1
10	3	7	0	0	0	0	10
12	17	39	32	27	7	0	122
18	0	3	2	1	0	0	6
24	4	5	7	2	0	0	18
36	1	0	0	0	0	0	1
Total	101	203	156	66	13	1	540

Table-2. Average statistics of PPA for the observed women

Parameter	**Mean(SD)**	**HPD (97.5%, 2.5%)**
B_0(Intercept)	0.00(0.013)	(0.03, -0.01)
B_{1i}(RW_1)	0.009(0.14)	(0.010, 0.009)
B_{1i}(RW_1)	0.33(0.19)	(0.472, 0.188)
D[1,1]	0.01(0.00)	(0.02, 0.00)
D[1,2]	5.9E-4(0.01)	(0.02, -0.02)
D[2,2]	0.34(0.15)	(0.713, 0.122)
Dinv[1,1]	97.21(19.65)	(132.0, 49.93)
Dinv[1,2]	-0.07(2.64)	(5.04, -5.26)
Dinv[2,2]	3.64(1.83)	(8.22, 1.46)

Table 3. Distribution of observed mothers with duration of PPAand order of birth

Order of birth	PPA (in month)														Total
	1	2	3	4	5	6	7	8	9	10	12	18	24	36	
1st	5	2	62	3	1	2	1	1	1	2	22	1	4	1	108
2nd	4	2	59	3	3	1	3	1	0	2	24	2	4	0	108
3rd	4	1	59	3	1	2	2	5	0	2	24	1	4	0	108
4th	5	1	58	3	1	3	3	2	0	2	26	1	3	0	108
5th	4	1	59	3	0	4	2	3	0	2	26	1	3	0	108
Total	22	7	297	15	6	12	11	12	1	10	122	6	18	1	540

Table 4. Observed duration of PPA according to order of birth

District	Mean & Median	Orders of Birth				
		1st	2nd	3rd	4th	5th
Imphal West	Mean	6.39	6.80	7.82	8.44	6.88
	Median	3.00	3.00	5.00	7.00	3.00
Imphal East	Mean	5.34	5.86	6.10	6.94	3.92
	Median	3.00	3.00	3.00	3.00	3.00
Bishnupur	Mean	6.71	7.11	7.18	7.28	6.29
	Median	3.00	3.00	3.00	3.00	4.50
Thoubal	Mean	5.83	6.24	6.81	6.77	6.52
	Median	3.00	3.00	3.00	3.00	3.00

Table 5. Estimated duration of PPA with mother's age

Duration of PPA	Age of mother (in year)						
	Up to 20	20 to 25	25 to 31	31 to 35	36 to 40	41 to 45	Total
Posterior mean	4.64	4.50	4.61	4.73	4.83	4.89	4.35
SD	0.04	0.03	0.05	0.04	0.02	0.00	0.03

Table 6. Posterior mean of PPA with parameters in Bayesian Approach

Model	**Parameter**	**Mean (SD)**	(97.5%, 2.5%)
Binomial	(Intercept)β_{11}	0.27(1.10)	(2.36, -0.95)
	(Order of birth$_5$)β_{12}	-2.03(0.98)	(-0.55, -3.57)
	(Order of birth$_{15}$)β_{13}	2.42(31.90)	(66.52, -57.41)
	(caste)β_{14}	0.68(0.94)	(2.09, -0.52)
	(caste * Order of birth$_5$)β_{15}	1.84(0.97)	(3.30, 0.34)
	(caste * Order of birth$_{15}$)β_{16}	-1.48(30.87)	(58.60, -62.55)
	(Adoption of sterilization)β_{17}	0.49(0.12)	(0.73, 0.25)
	(Sex(male/female) of child first ever born)β_{18}	-1.13(0.57)	(-0.02, -2.35)
Poisson	(Intercept)β_{21}	1.45(0.09)	(1.63, 1.31)
	(Order of birth$_5$)β_{22}	0.09(0.12)	(0.35, -0.12)
	(Order of birth$_{15}$)β_{23}	1.07(30.09)	(59.78, -59.63)
	(caste)β_{24}	0.12(0.15)	(0.34, -0.19)
	(caste * Order of birth$_{15}$)β_{25}	-0.08(0.12)	(0.17, -0.33)
	(caste * Order of birth$_{15}$)β_{26}	-0.34(32.04)	(61.7, -64.28)
	(Adoption of sterilization)β_{27}	-0.03(0.08)	(0.16, -0.17)
	(Sex(male/female) of child first ever born)β_{28}	-0.14(0.14)	(0.11, -0.44)
Variance component	P	-0.04(0.09)	(0.12,-0.23)
	σ_1	1.29(0.30)	(1.92, 0.79)
	σ_2	1.65(0.16)	(1.99, 1.34)

Table 7. Descriptive statistics of family

Characteristics	Frequency (%)
Type of Family	
Joint	551(42.5%)
Nuclear	745(57.5%)
Adoption of Sterilization	
Yes	29(2.2%)
No	1267(97.8%)
Religion of wife	
Hindu	1101(85%)
Meitei	157(12.1%)
Muslim	18(1.4%)
Christian and others	20(1.5%)
Sex of child first ever born	
Male	562(43.4%)
Female	594(45.8%)

Table 8. Summary statistics for the fertility data

Time of delivery	Type of family	N	Percent of zeros	Mean non-zero months of PPA(SD)
1st	Nuclear	59	4.55	6.11(5.35)
	Joint	82	6.32	5.71(4.79)
2nd	Nuclear	166	12.80	6.33(5.31)
	Joint	246	18.98	6.56(5.12)
3rd	Nuclear	364	28.08	6.81(5.60)
	Joint	373	28.78	7.15(5.49)
4th	Nuclear	560	43.20	7.12(6.04)
	Joint	446	34.41	7.64(5.77)
5th	Nuclear	679	52.39	6.30(5.23)
	Joint	509	39.27	6.26(4.85)

References

Aguirre, G. P. (1996). The determinants of postpartum amenorrhea: A multi-state hazard approach. Centre for Demography and Ecology Working Paper, University of Wisconsin-Madison, 96-03.

Aguirre, G. P., & Jones, R. E. (2005). Breastfeeding and post-partum amenorrhea in rural Guatemala. *Población y Salud en Mesoamérica* [Internet]. Retrieved October 2009, from http://ccp.ucr.ac.cr/revista/.

Ahmed, L. C., Geshe, S. M., & Mosley, W. H. (1974). A prospective study of birth interval dynamics in Chen Rural Bangladesh. *Population Studies, 28*(2), 277-296.

Aryal, T. R. (2005). Differentials of female age at marriage in rural Nepal. *Nepalese Journal of Development and Rural Studies, 2*(1), 90-95.

Aryal, T. R. (2006). Retrospective reporting of the duration of postpartum amenorrhea: A survival analysis. *Kathmandu University Medical Journal, 14*(1), 211-217.

Aryal, T. R. (2007). Post-partum amenorrhea among Nepalese mothers. *Journal of Population and Social Studies, 16*(1), 35-63.

Bongaarts, J., & Potter, R. G. (1983). *Fertility, biology and behaviour: A theoretical analysis* (1st ed.). New York: Academic Press.

Del, M. F., & Adiao, A. C. (1970). Lactation and child spacing as observed among 2,102 rural Filipino mothers. *Philippine Journal of Pediatrics, 19*(2), 128-132.

Fahrmeir, L., & Osuna, E. L. (2006). Structured additive regression for over-dispersed and zero-inflated count data. *Applied Stochastic Models in Business and Industry, 22*(4), 351-369.

Ghosh, S. K., Mukhopadhyay, P., & Lu, J. C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference, 36*(13), 1360-1375.

Huffman, S. L., Chowdhary, A. K., Chakraborty, J., & Simpson, N. K. (1980). Breastfeeding patterns in rural Bangladesh. *American Journal of Clinical Nutrition, 19*(1), 128-132.

Janet, A., Gary, K., & Richard, H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research, 11*(4), 341-355.

Jones, R. E. (1989). Breastfeeding and postpartum amenorrhea in Indonesia. *Journal of Biosocial Science, 21*(1), 83-100.

Karim, O., & Hajian, T. (2002). Factors affecting the pattern of postpartum amenorrhea. *Annals of Saudi Medicine, 22*(1), 5-6.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*(1), 1-14.

Mannan, H. R. (1998). Differential patterns and correlates of postpartum amenorrhea in Bangladesh: A multivariate analysis. *Journal of Family Welfare, 44*(3), 28-35.

Min, Y., & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling, 5*(1), 1-19.

Nath, D. C., & Atanu, B. (2011). A Bayesian approach for autoregressive models in longitudinal data analysis: An application to type 2 diabetes drug comparisons. *Asian Journal of Applied Sciences, 4*(5), 640-648.

Nath, D. C., Land, K. C., & Singh, K. K. (1994). The role of breastfeeding beyond postpartum amenorrhea on the return of fertility in India: A life table and hazards model analysis. *Journal of Biosocial Science, 26*(2), 191-206.

Pinto, G., Palloni, A., & Jones, R. E. (1998). Effects of lactation on post-partum amenorrhea: Re-estimation using data from a longitudinal study in Guatemala. *Population Studies, 52*(2), 231-248.

Prema, K., Naidu, A. N., & Neela, K. S. (1979). Lactation and fertility. *American Journal of Clinical Nutrition, 32*(5), 1298-1303.

Rahman, M. M. (1992). Measurement of post-partum amenorrhea in Bangladesh. *Journal of Biosocial Science, 24*(1), 17-24.

Salway, S., Roy, N. C., Koenig, M. A., & Cleland, J. (1993). Levels and trends in post-partum amenorrhea, breastfeeding and birth intervals in Matlab, Bangladesh: 1978-1989. *Asia-Pacific Population Journal, 8*(2), 3-22.

Singh, S. N., Singh, N. S., & Narendra, R. K. (2012). Postpartum amenorrhea among Manipuri women: A survival analysis. *Journal of Health, Population, and Nutrition, 30*(1), 93-98.

Srinivasan, K., Pathak, K. B., & Pandey, A. (1989). Determinants of breastfeeding and postpartum amenorrhea in Orissa. *Journal of Biosocial Science, 21*(3), 365-371.

Weis, P. (1993). The contraceptive potential of breastfeeding in Bangladesh. *Studies in Family Planning, 24*(2), 100-108.

Changing Impact of Son-Preference on the Family Building Process in India: A Parity Progression Ratio Analysis

Abstract

The manifestations of son preference over daughter have been common in India. This is expected to lose its rheostat with time and developmental activities. The pace may be greater in urban than rural areas. The present chapter is an attempt in this direction; it aims to assess the impact of son preference over daughter over time and between urban and rural India. For this, it adopts the concept of parity progression ratios based on Kaplan-Meier life table approach as the measure of fertility. It uses the birth history data of currently married women aged 40-49 years covered in two consecutive rounds of India's National Family Health Survey, i.e., NFHS-3 (2005-06) and NFHS-4 (2015-16). Each parity transition (up-to-fifth-order births) was analyzed, separately, with respect to the number of sons at each specific parity. The PPRs were consistently higher among currently married women who had only daughters at all parities as compared to those who had only sons or both sons and daughters in NFHS-3 (2005-06) as well as NFHS-4 (2015-16) irrespective of the place of residence. The pace of progressions differs substantially between urban and rural areas. Greater parental preference for sons over daughters has been observed in rural areas as compared to urban areas at all parties in 2015-16. A comprehensive family welfare behaviour change communication packages are required to counsel couples that both daughters and sons, are equally important and valuable, and emphasize the slogan "BETI Bachao, BETI Padhao" against the prevailing perception in India.

Keywords: Son preferences, PPR, Birth History, Life table, and NFHS.

1. Introduction

Preferences for sons over daughters are deep routed in Indian societies [1-3]. Many studies have shown that married couples keep childbearing on until they achieve the desired number of sons which results in adding up additional births [4-7]. Studies have also shown that biologically, only 26% of married couples fulfill their desire for two sons after two births and around 10% of married couples are unable to achieve their desire for two sons ever after having six children [3, 5, 8]. Such preferences for sons over daughters have many implications on the social and demographic structure of the population; it has a direct impact on population growth and sex ratio at birth [3, 5, 9-11] and is positively associated with the desire for more children and thereby the total fertility rate. Also, it is negatively associated with contraceptive uses which resulted in higher fertility rates delaying the demographic transitions in India [3, 12-14]. However, evidence suggests that gender preferences have declined and a moderate preference for a daughter has been noted in India [3, 4]. This declining trend in sex preference might be due to the influence of education, urbanization, exposure to mass media, rapid economic growth, and a rise in women's status [3, 9]. It again varies in levels with respect to the geographical, social, and economic status of couples [15-16]. There are a few studies that explored the relationship between the son's desire, the sex ratio of living children in the family, contraceptive uses, and intended fertility [9-10, 17-18]. There are attempts to analyze the relationship between the sex composition of living children and realized fertility using parity progression ratios in India [3, 11, 19]. In the present study, we examined the impact of preferences for sons over daughters on fertility in India by examining parity progression ratios at different parities by the sex composition of children and women's urban-rural residence over a decade between 2005-2006 and 2015-16. The choice of examining parity progression ratios (PPRs) over the conventional measures of fertility is to study step-by-step incremental aspects of family building process. It encapsulates the contingent nature of fertility behavior including the decision when to have the next child and family size in terms of the number of children. Relative contributions of gender preferences at different birth orders/parity to overall family size are well explained within the parity progression ratio framework.

2. Objectives

The specific objective of the study is to estimate and compare the impact of son preferences over daughters on the family-building process in India using the parity progression ratios.

3. Materials and Methods

3.1 Data

The complete birth history data of currently married women aged 40-49 years are used to examine the effects of son preferences on the parity progressions ratios. The Birth history data was extracted from the third & fourth rounds of Indian National Family Health Survey (NFHS) [15-16] conducted during the periods of 2005-2006 and 2015-16 across the country. NFHS is a nationally representative household survey and provides information on a wide range of indicators on health, diseases, nutrition, and demographics along with the birth history data. In these surveys, all women aged 15-49 years present at the household at the time of the survey were interviewed irrespective of their marital status (i.e., ever married and never married). Uniform sampling design was used across the country, to ensure comparability and maintain the highest data quality. In NFHS-3, a two-stage sampling method in rural areas and a three-stage sampling method in urban areas were adopted to select the sampling units, however, in NFHS-4, a two-stage sampling method is employed at both urban-rural areas. It is worth mentioning that NFHS-3 had state representative sample while NFHS-4 was designed to provide estimates of certain RCH-related service indicators at the district level and outcome indicators at the state level [15-16]. In analysis, national women's weights derived and provided in the NFHS data are used in the analysis to maintain the representativeness of the samples.

3.2 Inclusion and Exclusion Criteria

In the analysis, we have included those currently married women who have given birth to at least one child and completed their childbearing process and excluded from the analysis to those currently married women who were married more than once and had multiple births (twins or more) at any parity to control the displacement of date of births and the variations arose due to different marriages, marital status and sex-composition of children. Women who married before the age of 10 years and had child births before the age of 13 years, and had premarital births are also excluded from the analysis. To progress from any specific parity or birth to the next parity/birth, women must have their next birth within 10 years of the preceding birth. In the sample, around 18,309 currently married women aged 40-49 years from NFHS-3 and around 1,19,065 currently married women (CMW) aged between 40 -49 years from NFHS-4 have satisfied the inclusion and exclusion criteria and included in the analysis. These CMWs from NFHS-3 and NFHS-4 respectively had 65,996 and 3,69,522 live births.

3.3 Study Variables

3.3.1 Dependent Variable

Parity progression ratios (PPRs) at each parity are the main study variable and have been used as a method to estimate the propensity of subsequent childbearing after having specific-combination of the number of children assessing the impact of son-preferences over daughters. The parity progression ratio describes the family-building process by considering the incremental aspect of childbearing. It describes that out of women with the specific parity or birth order, how many will move to the next child.

In analysis, PPRs are estimated at parity 1 (p_1), parity 2 (p_2), parity 3 (p_3), and parity 4 (p_4) representing the progression from first to second birth, second to third birth, third to fourth birth and fourth to fifth births. The choice of these four parity transitions is governed by the intention to look at the effect of preferences for sons on parity transitions during which a considerable proportion of married women have already achieved their desired sex composition and desired family size, in order that the estimated effects of the son-preferences on parity progressions would be relatively large and straightforward to spot and, it might be of interest to policymakers and program managers.

3.3.2 Independent Variables

Fourteen predictor variables as the combination of number of sons and parity are created as: parity 1 (0 or 1 son), parity 2 (0, 1 or 2 sons), parity 3 (0, 1, 2, or 3 sons), and parity 4 (0, 1, 2, 3 or 4 sons). The number of sons refers to all possible combinations of sons and daughters at a given parity. The other independent variable included in the analysis by which the effects of son-preferences are examined is the urban-rural place of residence.

3.4 Methods

3.4.1 Estimation of Parity Progression Ratios (PPRs)

Kaplan-Meier life table approach is employed to estimate the parity progression ratios at each parity by the number of sons using the birth history of currently married women in the sample. Parity-specific life tables are constructed by pooling the closed and open birth intervals of specific birth orders for all parities up-to 5th order births or beyond [4, 20-21]. Each life table is truncated at 10 years with the assumption that after 10 years of each birth, the probability of the next birth is negligible [4, 20-21].

It is assumed that the minimum parity progression ratios would prevail at a specific parity if the currently married women would not have any son preferences over daughters at that parity. Let us assume that CMW_i is the actual number of currently married women with parity i^{th} and $min(p_i)$ is the minimum parity progression ratio at i^{th} parity, where i=1, 2, 3, 4. Then, under the assumption of minimum parity progression in the absence of gender preferences, the number of currently married women at the start of $(i+1)^{th}$ parity is calculated as follows:

$$\boldsymbol{Est(CMWi+1)=(CMWi)*[min(pi)]}$$

Hence, the number of additional child births attributed by (i+1)th parity due to son-preferences over daughter, is equal to $\{CMW_{i+1} - Est(CMW_{i+1})\}$, where, CMW_{i+1} is the actual number of the currently married women at the starting of (i+1)th parity; and $Est(CMW_{i+1})$ is the estimated number of the currently married women at the starting of (i+1)th parity. The approach is earlier employed by Chaudhuri S., 2012 [3].

4. Results

4.1 Sample Distribution

There were 18, 309 CMW in NFHS-3 (2005-06) and 1,19,065 CMW in NFHS-4 (2015-16) aged 40-49 years who satisfied the inclusion and exclusion criteria and were included in the analysis. In NFHS-3, around 34% lived in urban areas and 66% CMW lived in rural areas, however, in NFHS-4, around 36% lived in urban areas and 64% lived in rural areas.

4.2 Currently Married Women by the number of sons in 2005-2006 and 2015-16

Table 1 shows the percentage distribution of CMW aged 40-49 years by the number of sons by place of residence in NFHS-3 (2005-06) and NFHS-4 (2015-16).In NFHS-3, among women with parity 1, around 47% did not have any son and 53% had one son. Among women with parity 2, around 23% had no son and 25% had one son. By place of residence, in urban areas, 24% have no son, 26% have one son and in rural areas around 22% CMW have no son and 25% have only one son. At parity 3, only 13% CMW in urban areas and 12% CMW in rural areas had no sons, and 25% in urban areas and 28% in rural areas had 3 sons. Among CMW who had 4 children, only 8% have no son in urban areas and 6% have no son in rural areas. The majority of CMW (47% in urban areas and 49% in rural areas) have 2 sons and around one-fourth have 3 sons at both places. In the 2015-16 cohort of CMW, at parity 1, around 46% CMW do not have any sons and 54% of CMW have one son. The percentages of CMW with no son in urban and rural areas are 45% and 46% respectively. At parity 2, around 22% have no son, and 25% have only one son while in urban areas, 21% have no son, 26% have one son and in rural areas around 22% CMW have no son and 25% have only one son. At parity 3, around 48% CMW have 2 sons. Only 24% CMW in urban areas and 26% CMW in rural areas have 3 sons and around 11% CMW have no sons in urban areas as well as in rural areas. Among CMW who have 4 children, only 8% have no son in urban areas and 7% CMW have no son in rural areas. The majority of CMW (48%) have 2 sons, around 24% have 3 sons at both places, around 10%CMW in urban areas and 12% CMW in rural areas have 4 sons at parity 4.

4.3 Differentials in Parity Progression Ratios (PPRs) by the Number of Sons

Table 2 presents the differentials in PPRs by the number of sons at different parities in NFHS-3 (2005-06) & NFHS-4 (2015-16) cohorts. In 2005-06, at first parity transition, the percentage of going on to second birth (p_1) is 95% for CMW who have first birth but have no son and it is 94% for CMW who have first birth and have one son. At parity 2, the value of p_2 is higher for CMW who had no son (88%) as compared to those who have 1 son (78%) and 2 sons (77%). Among CMW who have third-order births, the PPRs are around 86% for those who have no son, 71% for those who have only one son, 70% for those who have 2 sons, and 65% for those who have 3 sons. At parity 4, the values of p4 are substantially lower for CMW who have sons than those who have no son at all; the percentage of going on to fifth order birth is 85% for CMW who have no sons, 66% for those who have one son, 64% for those who have 2 sons, 60% those who have 3 sons and 64% for those who have 4 sons. In 2015-16, among CMW who had their first birth, the percentage of going on to second parity is 93% among CMW who have no son as compared to 89% CMW with one son. Whereas, at parity 2, around 81% CMW who have no son would progress to parity 3 as compared to 58% CMW who had 1 son, and 55% CMW who have 2 sons, respectively. At parity 3, the conditional probability for progression from 3rd to 4th order births is 81% among CMW who have no son, 52% among those who have 1 son, 53% among those who have 2 sons, and 46% among those who have 3 sons. At parity 4, the likelihood of progressions from 4th order to 5th order births is around 81% among those CMW who have no son, 54% among those who have 1 and 2 sons, 44% among those who have 3 sons, and 52% among those who have 4 sons.

4.4 Impact of Son-Preferences on the Family Building Process

Table 2 also shows the impact of son-preferences on parity progressions at different parity by 2005-06 and 2015-16 cohorts. In the 2005-06 cohort, under the assumption of the minimum parity progression in the absence of son-preferences, around 94% CMW would like to progress to parity 2 from parity 1; 77% would like to progress from parity 2 to parity 3; 65% would like to progress to parity 4 from parity 3 and 60% would like to progress to parity 5 from parity 4. By this approach, around 1% less birth at parity 2; 4% less birth at parity 3; 8% less birth at parity 4, and 7% less birth at parity 5 would occur. Overall, around 63,959 births would occur by the parity 5 which is 2,040 births fewer than the observed total 65, 996 births by parity 5 which indicates that around 3% additional births are attributed to parity 5 in 2005-2006 cohorts due to son-preferences. However, among the 2015-16 cohort, under the assumption of minimum parity progression at a parity and absence of son desire at that parity, around 89% CMW would like to progress from parity 1 to parity 2; 55% would like to progress to parity 3 from parity 2; 46% CMW would like to move from parity 3 to parity 4 and 44% CMW would like to progress to parity 5 from parity 4 which induced 3% fewer births at parity 2; 18% fewer births at parity 3; 21% fewer births at parity 4 and parity 5. By this approach, around 3, 38,132 childbirths would have occurred by parity 5 which is 31,390 lesser than the observed total live births of 3,69,522 by parity 5 in the study sample which indicates around 8% additional births are attributed due to son preferences over daughter by parity 5 in 2015-16. The above analysis clearly shows that the trends in progressions by the number of sons are same in the both cohorts. The CMW who have no sons have substantially higher progressions at all parities than those who have one or more sons or sons and daughters in both cohorts. Though the parity progressions are consistently higher among CMW from 2005-2006 than CMW from 2015-16 at all parities, a higher proportion of additional births are taking place among CMW from 2015-16 than CMW from 2005-2006 due to son-preferences.

4.5 Differentials in PPRs by Number of Sons and the Place of Residence in 2005-06

Table 3 shows the differentials in estimated PPRs by the number of sons and the place of residence for the 2005-2006 cohort. Urban-rural variations in the PPRs by the number of sons have been seen at all parities. In urban areas, PPR is slightly lower for CMW who have one son than those who have no son at parity 1. The percentage of going on to second birth is 90% for CMW who have one son and 92% for those who have no son at parity 1. Among CMW who had second births, the PPR is substantially lower among CMW who have 2 sons or one son than those who have no sons; the values of PPRs are 82%, 67%, and 66% respectively for those who had no sons, one son, and 2 sons. At parity 3, the percentage of going on to fourth birth is 79% for those who have no son, 58% for those who have one son, 61% for those who have 2 sons, and 55% for those who have 3 sons. At parity 4, around 86% CMW who had no son are progressed to fifth-order birth whereas the parity progressions to the next parity among CMW who had one son, 2 sons, 3 sons, and 4 sons are around 52%, 52%, 48%, and 60% respectively. However, in rural areas, the PPRs at parity 1 are around 97% for CMW who have no son and 95% for CMW who have one son. At parity 2, among CMW who had 2 births, the likelihood of progressions to third birth is 91% for CMW who have no son, 84% for those who have one son, and 82% for those who have 2 sons. The PPRs at parity 3 are 89% for those who have no son, 78% for CMW who have one son, 73% for those who have 2 sons, and 68% for those who have 3 sons. Among CMW who had 4 births, around 85% CMW who had no son would like to progress to fifth birth, whereas, the percentage of going on to fifth birth is 71% for CMW who have one son, 68% for those who have 2 sons, 64% those who have 3 sons, and 65% those who have 4 sons. The pattern and trend in the parity progressions by the number of sons at different parity are same in both the place of residence.

4.6 Impact of Son-Preferences on Family Building Process by the Place of Residence in 2005-06

Table 3 also shows the estimated impact of son preferences on the family building process by urban-rural residence in 2005-2006.In urban areas, under the assumption of minimum progression in the absence of son-preferences, around 90% CMW would like to move from parity 1 to parity 2; 66% from parity 2 to parity 3; 55% from parity 3 to parity 4, and 48% from parity 4 to parity 5 which come about a 2% less childbirth at parity 2; 7% less childbirth at parity 3; 11% childbirth at parity 4 and 13% less childbirth at parity 5. Overall, around 119,201 childbirths would occur by parity 5 as against the actual 20,035 childbirths by parity 5. It indicates that around 4% of additional births are attributed to parity 5 due to preferences for sons over daughters in urban areas in 2005-2006 in India. However, in rural areas among CMW from 2005-2006, if there are no gender preferences and minimum parity progression prevails at each parity then around 95% CMW would like to progress to parity 2 from parity 1; 82% would like to move from parity 2 to parity 3; 68% would like to move from parity 3 to parity 4 and 64% would move to parity 5 from parity 4. By this approach, around 2% fewer child births at parity 2; 4% fewer child births at parity 3; 9% fewer child births at parity 4, and 6% fewer child births at parity 5 will occur in rural areas. Overall, around 3% of additional births are attributed to son preferences in rural areas in 2005-2006. The analysis clearly shows that the preferences for sons are stronger in urban areas than in rural areas in 2005-2006 in India.

4.7 Differentials in Parity Progression Ratios (PPRs) by the Number of Sons and the Place of Residence in 2015-16

Table 4 shows the differentials in estimated PPRs by the number of sons and the place of residence for the 2015-16 cohort of CMW. Urban-rural differentials in the percentages of going on to the next parity at different parities by the number of sons exist and increase with increasing parity in 2015-16. In urban areas, around 89% CMW who have no son have progressed to second-order birth from first birth whereas around 84% CMW who have one son have moved to second-order birth at parity 1. Among CMW who have two births, the PPRs are 70%, 45%, and 44%, respectively for CMW who have no son, one son, and two sons. At parity 3, around 72% CMW who have no son have progressed to the next higher order birth, as compared to 40% CMW who have one son, 44% CMW who have 2 sons, and 39% CMW who have 3 sons. At parity 4, around 74% CMW who have no son progress to fifth order birth whereas the parity progressions to next parity among CMW who have one son, two sons, three sons, and four sons are around 45%, 46%, 39%, and 48% respectively. However, in rural areas, PPRs at parity 1 are 95% for CMW who have no son and 92% for CMW who have one son. At parity 2, among CMW who have two births, the percentages of going on to third birth are 86% for CMW who have no son, 64% for CMW who have one son, and 60% for those who have 2 sons. The PPRs at parity 3 are 85% for CMW who have no son, 57% for CMW who have one son, 56% for CMW who have two sons, and 49% for CMW who have three sons. At parity 4, around 84% CMW who have no son have progressed to fifth-order birth, whereas, 56% CMW who have one son, 56% CMW who have two sons, 45% CMW who have three sons, and 53% CMW who have four sons have progressed to next order parity (table 4). The above findings clearly indicate that PPR among CMW who have no sons are consistently higher at all parities than those who have one or more sons or sons and daughters both, in both places of residence. The levels of progressions are quite high in rural areas than in urban areas at all parities.

4.8 Impact of Son-Preferences on Family Building Process by the place of Residence in 2015-16

Table 4 also shows the estimated impact of son preferences at different parities by urban-rural residence in 2015-16 in India. In urban areas, under the assumption of no gender preferences, around 84% CMW would like to progress from parity 1 to parity 2; 44% would like to progress from parity 2 to parity 3; 39% would like to progress from parity 3 to parity 4 and 39% from parity 4 to parity 5. By this approach, around 4% less childbirth at parity 2, 20% less childbirth at parity 3, 19% fewer childbirths at parity 4, and 20% less childbirth at parity 5 will occur in urban areas in 2015-16. Overall, around 1,08,196 births would occur by parity 5 as against the observed 1,16,703 births by parity 5 in urban areas. In other words, around 7% additional births are attributed to son preferences in urban areas among CMW aged 40-49 years in India in 2015-16 (table 4). However, in rural areas, under the assumption of no gender preferences, the minimum parity progressions at parity 1, 2, 3, and 4 would be 92%, 60%, 49%, and 45% respectively, which come about to a 3% less childbirth at parity 2, 19% less childbirth at parity 3, 22% less childbirth at parity 4 and 22% less childbirth at parity 5 will occur in rural areas. Overall, there would be around 2,29,438 births as against the observed total 2,52,819 births by parity 5 in rural areas. It means around 9% additional births are attributed by parity 5 due to the preferences for sons in rural areas among CMW aged 40-49 years in 2015-16 in India (table 4). A substantial proportion of additional childbirths are taking place in both urban and rural areas, but it is stronger in rural areas as compared to urban areas in 2015-16 in India.

5. Discussion and Conclusions

The foregoing analysis echoed the preferences for sons over daughters have been influencing the fertility behavior and family building process in India at all parties. The PPRs are consistently higher among those who have only daughters at all parties in comparison to those who have only sons or both sons and daughters (male and female children) in 2005-06 as well as in 2015-16. In other words, those who have no sons are more prone to continue their childbearing process till they achieve their desired number of sons which is consistent with findings of prior research in India [3, 4, 13].The results conform to the established norm that preferences for sons over daughters are a significant motivational factor for the continuous childbearing process among married couples in India [1, 3, 4] and abroad as well. Many couples continue the childbearing process to have sons even after having the desired family size which results in adding up more additional births and a slow pace of fertility declines [3, 4, 11]. However, at higher parity, the progression is quite high among CMW who have only 4 sons than CMW who have 1 daughter or 2 daughters in the study samples (2005-06 & 2015-16), which confirms earlier findings [4, 13] which found that though in India the parental preferences for sons over daughters are higher, but parents do want one or more daughters correlating it with the Hindu customs of Kanydaan (rituals of giving away a daughter in marriage). The results based on PPR model clearly show that sons-preferences over daughters have added a good proportion of additional children at all parity. The trends in progressions by the number of sons are same in the both cohorts. The CMW who have no sons have substantially higher progressions at all parity than those who have one or more sons or sons and daughters in both cohorts. Though the parity progressions are consistently higher among CMW from 2005-2006 than CMW from 2015-16 at all parities, a higher proportion of additional births due to gender preferences have taken place in 2015-16 than 2005-2006 at all parities. It may be due to differences in the sample sizes in 2005-2006 and 2015-16. It may also be that most of the progressions among CMW from 2005-2006 are not due to son preferences. The gender preferences have added around 1% additional childbirths at parity 2, 4% additional childbirths at parity 3, 8% additional childbirths at parity 4, and 7% additional childbirths at parity 5 in 2005-2006, whereas, around 3% additional childbirths at parity 2; 18% additional childbirths at parity 3; 21% additional childbirths at parity 4 and 21% additional childbirths at parity 5 have added-up in 2015-16 in India. Overall, the gender preferences have added around 8% additional births by parity 5 in 2015-16 as compared to 3% in 2005-06 in India. It indicates that the gender preferences have substantially increased at all parity between 2005-06 to 2015-16 in India. Though urban-rural differentials exist in gender preferences in India, however, the pace of transitions is different in urban areas as compared to rural areas. A high parental preference for sons over daughters has been observed in the rural areas at all parties in 2015-16, however, the preferences for sons are stronger in urban areas than rural areas in 2005-06 in India. The PPRs are lower among CMW who live in the urban areas and had atleast one daughter and among CMW who has one daughter in the rural areas in 2005-06 and 2015-16. The results based on PPR model clearly show that sons-preferences over daughters have added a good proportion of additional children at both the place of residence in both cohorts. Around 4% additional childbirths in urban areas and 3% additional childbirths in rural areas have been attributed by parity 5 due to parental preferences for sons over daughters in 2005-06. However, in 2015-16, around 7% additional childbirths in urban areas and 9% additional childbirths in rural areas were attributed due to son preferences among CMW aged 40-49 years in India. This shows that the gender preferences for sons are stronger in rural areas than urban areas in 2015-16 whereas in 2005-2006, the preferences for sons over daughters are stronger in urban areas than rural areas. As preferences for sons over daughters appear to be a significant determinant and motivational factor for the continuous childbearing process, comprehensive family welfare behaviour change communication packages are required to counsel couples that both daughters and sons, are equally important and valuable. It further needs to emphasize the slogan of "BETI Bachao, BETI Padhao" to reduce the prevailing perception.

References

Arnold, F., Choe, M. K., & Roy, T. K. (1998). Son preference, the family-building process and child mortality in India. *Population Studies, 52*(3), 301-315. http://www.jstor.org/stable/2584732

Arokiasamy, P. (2002). Gender preference, contraceptive use and fertility in India: Regional and development influences. *International Journal of Population Geography, 8*(1), 49-67.

Bairagi, R. (2001). Effects of sex preference on contraceptive use, abortion and fertility in Matlab, Bangladesh. *International Family Planning Perspectives, 27*(3), 137-143. https://doi.org/10.2307/2673835

Basu, D., & De Jong, R. (2010). Son targeting fertility behavior: Some consequences and determinants. *Demography, 47*(2), 521-536.

Bhatia, J. C. (1978). Ideal number and sex preference of children in India. *Journal of Family Welfare, 24*(4), 3-16.

Bhattacharya, P. C. (2006). Economic development, gender inequality, and demographic outcomes: Evidence from India. *Population and Development Review, 32*(2), 263-291. http://www.jstor.org/stable/20058874

Chaudhuri, S. (2012). The desire for sons and excess fertility: A household-level analysis of parity progression in India. *International Perspectives on Sexual and Reproductive Health, 38*(4), 178-186. https://doi.org/10.1363/3817812

Clark, S. (2000). Son preference and sex composition of children: Evidence from India. *Demography, 37*(1), 95-108. https://doi.org/10.2307/2648099

Das Gupta, M., Zhenghua, J., Bohua, L., Zhenming, X., Chung, W., & Hwa-Ok, B. (2003). Why is son preference so persistent in East and South Asia? A cross-country study of China, India and the Republic of Korea. *The Journal of Development Studies, 40*(2), 153-187.

Gray, E., & Evans, A. (2005). Parity progression in Australia: What role does sex of existing children play? *Australian Journal of Social Issues, 40*(4), 505-520.

International Institute for Population Sciences. (2007). *National Family Health Survey (NFHS-3), 2005-06: India*. Mumbai: International Institute for Population Sciences.

International Institute for Population Sciences. (2017). *National Family Health Survey (NFHS-4), 2015-16: India*. Mumbai: International Institute for Population Sciences (IIPS) and ICF.

Jayaraman, A., Mishra, V., & Arnold, F. (2009). The relationship of family size and composition to fertility desires, contraceptive adoption and method choice in South Asia. *International Perspectives on Sexual and Reproductive Health, 35*(1), 29-38. http://www.jstor.org/stable/25472413

Leone, T., Matthews, Z., & Zuanna, G. D. (2003). Impact and determinants of sex preference in Nepal. *International Family Planning Perspectives, 29*(2), 69-75. https://doi.org/10.2307/3181060

Mari Bhat, P. N., & Zavier, A. J. F. (2003). Fertility decline and gender bias in northern India. *Demography, 40*(4), 637-657. https://doi.org/10.2307/1515201

Mutharayappa, R., Choe, M. K., Arnold, F., & Roy, T. K. (1997). Son preference and its effect on fertility in India. *National Family Health Survey Subject Reports Number 3*. Mumbai: International Institute for Population Sciences and Honolulu: East-West Center.

Narayan, P., Nath, D. C., Das, K. K., & Pandey, A. (2020). Effect of son preferences on the family building process in India: A parity progression ratio analysis. *IJRAR - International Journal of Research and Analytical Reviews (IJRAR), 7*(2), 449-453. Retrieved from http://www.ijrar.org/IJRAR19W1184.pdf

Narayan, P., Pandey, A., & Nath, D. C. (2017). Parity progression analysis to study the urban-rural differentials in fertility in Uttar Pradesh. *Janasamkhya, 35*(1), 1-16.

Narayan, P., Pandey, A., & Nath, D. C. (2020). Effect of contraceptive uses on fertility in India: A parity progression ratios life table analysis. *International Journal of Research and Analytical Reviews, 7*(2), 468-473.

Seidl, C. (1995). The desire for a son is the father of many daughters: A sex ratio paradox. *Journal of Population Economics, 8*(2), 185-203. http://www.jstor.org/stable/20007464

Sheps, M. C. (1963). Effects on family size and sex ratio of preferences regarding the sex of children. *Population Studies, 17*(1), 66-72.

Table 1. Number and Percentage Distribution of Currently Married Women (CMW) aged 40-49 years at each specific Parity by Number of Sons they had and Urban-Rural Place of Residence, India, NFHS-3 (2005-2006) and NFHS-4 (2014-15), India

Parity/ No. of Sons/ No. of CMW	NFHS-3			NFHS-4		
	Urban	Rural	All	Urban	Rural	All
Parity 1	6285	12024	18309	43,236	75,829	1,19,065
0 Son	49%	47%	47%	45%	46%	46%
1 Son	51%	53%	53%	55%	54%	54%
Parity 2	5,771	11,650	17,421	37,770	71,599	1,09,369
0 Son	24%	22%	23%	21%	22%	22%
1 Son	26%	25%	25%	26%	25%	25%
2 Son	51%	53%	52%	53%	53%	53%
Parity 3	4074	9903	13977	20,805	52,997	73,802
0 Son	13%	12%	12%	11%	12%	12%
1 Son	15%	13%	13%	16%	14%	15%
2 Sons	48%	48%	48%	49%	48%	48%
3 Sons	25%	28%	27%	24%	26%	25%
Parity 4	2,520	7,377	9,896	10,023	33,152	43,175
0 Son	8%	6%	7%	8%	7%	8%
1 Son	9%	8%	8%	9%	9%	9%
2 Sons	47%	49%	48%	48%	48%	48%
3 Sons	25%	25%	25%	24%	24%	24%
4 Sons	11%	12%	12%	10%	12%	11%

Table 2 Number of Currently Married Women aged 40-49 years, Parity Progression Ratios, and Number of currently married women who would have continued childbearing if the minimum parity progression ratios prevailed at all parity by NFHS-3 & NFHS-4, India

Parity/Number of Sons	NFHS-3			NFHS-4		
	No. of CMW at the start of Parity	PPR (%)	No. of CMW if Minimum PPR prevailed	No. of CMW at the start of Parity	PPR(%)	No. of CMW if Minimum PPR prevailed
Parity 1	18,309		18,309	1,19,065		1,19,065
0 Son		95%			93%	
1 Son		94%			89%	
Parity 2	17,421		17,210	1,09,369		1,05,968
0 Son		88%			81%	
1 Son		78%			58%	
2 Sons		77%			55%	
Parity 3	13,977		13,414	73,802		60,153
0 Son		86%			81%	
1 Son		71%			52%	
2 Sons		70%			53%	
3 Sons		65%			46%	
Parity 4	9,896		9,085	43,175		33,949
0 Son		85%			81%	
1 Son		66%			54%	
2 Sons		64%			54%	
3 Sons		60%			44%	
4 Sons		64%			52%	
Parity 5	6,393		5,938	24,112		18,997
Total	65,996		63,956	3,69,522		3,38,132

Table 3. Number of Currently Married Women aged 40-49 years, Parity Progression Ratios and Number of currently married women who would have continued childbearing if the minimum parity progression ratios prevailed at all parity by Place of Residence, India, NFHS-3, 2005-06

	Urban			Rural		
Parity/Number of Sons	**No. of CMW at starting of Parity**	**PPR (%)**	**No. of CMW if Minimum PPR prevailed**	**No. of CMW at starting of Parity**	**PPR (%)**	**No. of CMW if Minimum PPR prevailed**
Parity 1	6,285		6,285	12,024		12,024
0 Son		92%			97%	
1 Son		90%			95%	
Parity 2	5,771		5,657	11,650		11,423
0 Son		82%			91%	
1 Son		67%			84%	
2 Sons		66%			82%	
Parity 3	4,074		3,809	9,903		9,553
0 Son		79%			89%	
1 Son		58%			78%	
2 Sons		61%			73%	
3 Sons		55%			68%	
Parity 4	2,520		2,241	7,377		6,734
0 Son		86%			85%	
1 Son		52%			71%	
2 Sons		52%			68%	
3 Sons		48%			64%	
4 Sons		60%			65%	
Parity 5	1,385		1210	5,009		4,721
Total	20,035		19,201	45,962		44,455

Table 4. Number of Currently Married Women aged 40-49 years, Parity Progression Ratios, and Number of currently married women who would have continued childbearing if the minimum parity progression ratios prevailed at all parity by Woman's Place of Residence, India, NFHS-4, 2015-16

	Urban			**Rural**		
Parity/Number of Sons	**No. of CMW at starting of Parity**	**PPR (%)**	**No. of CMW if Minimum PPR prevailed**	**No. of CMW at starting of Parity**	**PPR (%)**	**No. of CMW if Minimum PPR prevailed**
Parity 1	43,236		43,236	75,829		75,829
0 Son		89%			95%	
1 Son		84%			92%	
Parity 2	37,770		36,318	71,599		69,763
0 Son		70%			86%	
1 Son		45%			64%	
2 Sons		44%			60%	
Parity 3	20,805		15,980	52,997		42,959
0 Son		72%			85%	
1 Son		40%			57%	
2 Sons		44%			56%	
3 Sons		39%			49%	
Parity 4	10,023		8,114	33,152		25,969
0 Son		74%			84%	
1 Son		45%			56%	
2 Sons		46%			56%	
3 Sons		39%			45%	
4 Sons		48%			53%	
Parity 5	4,871		3,909	19,241		14,918
Total	1,16,703		1,08,196	2,52,819		2,29,438

Probability of Ultimate Ruin for the Log Normal Distribution and the Computation of Some of its Related Actuarial Quantities with Real Data Applications

Abstract

The chapter illustrates the application of the Log Normal distribution as an actuarial risk model. Here, more emphasis has been on the computational features of some of the key actuarial numbers, such as the minutes of time to ruin and the likelihood of final ruin when the underlying claim severity distribution is log normal. The classical risk model does not accommodate the case when there is an interest rate and tax payment acting upon the surplus process. In this chapter, the probability of ultimate ruin in the presence of interest rates and tax payments is also being evaluated for the case when the claim severity is distributed as Log Normal. The implementation of the algorithms aimed at evaluating these quantities presents computational challenges with associated sources of error. The quantities thus obtained are found to be exhibiting trends which are logical and consistent in terms of actual functioning of an insurance company. The parameters of the Log Normal distribution are estimated from a set of real- life insurance claim data.

Keywords: Loss Modelling, Stable Recursive Algorithm, Time to ruin, Surplus process under Interest rate and tax payments

1. Introduction

In a general insurance portfolio, two quantities of interest are the number of claims arriving in a particular period of time and the amount of each claim. We model the uncertainty in these quantities by random variables; specifically, a counting distribution is used to model the claim arrival pattern whereas a continuous distribution is used to model the claim severity. Loss modeling is a vital component of Mathematical modeling in general insurance since as specified in [1], in the most general sense, all of Actuarial science is about loss distribution modelling. Loss modelling is considered to be one of the most important aspects of risk modelling for it constitutes the basis, on which depends the accuracy of various other actuarial quantities related to the long-term solvency of the insurance company.

Log Normal distribution is a right skewed heavy tailed distribution which often arises as a potential model for modeling the claim severity. However, the drawback of the Log Normal distribution is that its Laplace transformation and hence its moment generating function does not have a closed form expression. Moreover, even its cumulative distribution function does not have a closed form expression. Log Normal distribution is a heavy tailed distribution i.e. it has relatively high probabilities in the right hand tail. The applications of the Log Normal distribution are commonly found in a variety of fields-Physics, Reliability theory, Biology, Economics, to name a few. There is enough evidence on the use of this distribution in property/casualty insurance to model claim sizes ([2].[3])

A good introduction to the subject of fitting distribution to losses is given in [2]. Other references on this subject include [1] and [3]. A typical characteristic of the claim data arising in the general insurance sector is that it is skewed to the right and hence heavy tailed distributions like Lognormal, Weibull, and others are considered to be potentially good candidates for modelling such data. However, as stated in [4] there is still no adequate framework confirming which class of distributions is appropriate to which category of insurance.

Based on our broad objective of illustrating the application of Log Normal distribution as an Actuarial risk model, we present the following objectives for our current work:

a) To fit the Log Normal distribution to a set of Insurance claim data.

b) To compute the probability of ultimate ruin for the Log Normal distribution using the stable recursive algorithm.

c) To compute the first two moments of the time to ruin for the fitted Log Normal distribution.

d) To compute the probability of ultimate ruin for the Log Normal distribution under the presence of interest earnings and tax payments.

The second section of the chapter is devoted to methodology which is again sub-divided into a few subsections, first sub section deals with fitting the Log Normal distribution whereas the second sub section deals with the classical risk model. The stable recursive algorithm for computing the probability of ultimate ruin is presented in the third sub section. The moments of the time to ruin for the Log Normal distribution is dealt with in the fourth sub section. The fifth subsection's content focuses on calculating the likelihood of ultimate ruin when interest and taxes are paid. The chapter's third section covers the findings and discussions, while the fourth and last section focus on the conclusion.

2. Methodology

2.1 Fitting of the Log Normal Distribution:

The probability density function of the Log Normal distribution is given by

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma y} exp\left\{\frac{-1}{2}\frac{(log(y) - \mu)^2}{\sigma^2}\right\}, y > 0, \sigma > 0, -\infty < \mu < \infty \tag{1}$$

The cumulative distribution function (cdf) of the Log normal distribution is given by

$$F(y) = \Phi\left(\frac{log(y) - \mu}{\sigma}\right), y > 0 \tag{2}$$

where Φ is the cdf of the standard Normal distribution.

The k^{th} raw moment of the Log normal distribution is given by

$$E(x^k) = p_k = exp\left(\mu k + \frac{\sigma^2 k^2}{2}\right), k = 1,2,3, \dots \tag{3}$$

The maximum likelihood estimators of the parameters are given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} log\ (x_i)\,, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\{log(x_i) - \hat{\mu}\}^2 \tag{4}$$

2.2 The Classical Risk Model

Let $\{U(t)\}_{t\geq 0}$ denote the surplus process of an insurer as

$$U(t) = u + ct - S(t) \tag{5}$$

where $u \geq 0$ is the initial surplus, c is the rate of premium income per unit time and $\{S(t)\}_{t\geq 0}$ is the aggregate claim process and we have $S(t) = \sum_{i=1}^{M(t)} X_i$, where $\{M(t)\}_{t\geq 0}$ is a homogeneous Poisson process with parameter λ, X_i denotes the amount of the ith claim and $\{X_i\}_{i=1}^{\infty}$ is a sequence of iid random variables with a distribution function F such that $F(0) = 0$ and probability density function f. We denote $E(X_1^k)$ by $p_k(k = 1,2,3, \dots)$. Also, we have $c = (1 + \theta)\lambda p_1$, where θ is the security loading factor. [5], [6], [7]

Let T_u denote the time to ruin from initial surplus u so that $T_u = inf\ \{t: U(t) < 0\}$ and define $\psi(u) = Pr\{T_u < \infty | U(0) = u\} = 1 - \chi(u)$ and $\psi(u, t) =$ Pr $(T_u \leq t | U(0) = u)$. $\psi(u)$ is known as the ultimate ruin probability whereas $\psi(u, t)$ is the finite time ruin probability. A detailed discussion on the Classical Risk model and the probability of ruin can be found in ([1],[8],[9],[10])The classical Risk model is the base for many mathematical models in insurance mathematics, but it is formulated with many simplifying criteria which make it deviate to some extent, from the actual picture observed in the insurance scenario. Some other computational techniques and related concepts on probability of ultimate ruin can be found in ([11], [12]).

2.3 A Stable Recursive Algorithm for the Evaluation of the Ultimate Ruin Probabilities:

The probability of ruin can be obtained as the solution of an integro differential equation [13]. A stable recursive algorithm implements a recursive algorithm to solve this integro differential equation for the Probability of ultimate ruin. It basically targets at yielding solution for the convolution part of the integro differential equation and in the process, produces an estimate of the probability of the ultimate ruin by finding the average of the bounds (lower and upper bounds) of this probability. Since the algorithm does not lead to the propagation of error, therefore it is stable and also, it produces bounds within a prescribed tolerance level.[14]

The following integral equation must be solved in order to compute the infinite time ruin probability numerically, as derived in [13].

$$\psi(x) = \frac{\lambda}{c}\int_0^x \psi(x - y)\{1 - F(y)\}dy + \frac{\lambda}{c}\int_x^{\infty}\{1 - F(y)\}dy \tag{6}$$

where,

$$\psi(0) = \frac{\lambda p_1}{c}, c = \lambda p(1 + \theta)\,, p_1 = \int_0^{\infty}\{1 - F(y)\}\,dy \tag{7}$$

Let

$$h(x)=\int_x^\infty\{1-F(y)\}dy\,,x\geq 0,h(0)=p_1(8)$$

An interested reader is recommended to refer to ([14], [6], and [7]) for a detailed description on the algorithm. Following is an outline of the steps to be carried out for implementing the stable recursive algorithm [14].

(i) First, divide the interval $[0,x]$ into some smaller sub intervals as shown below.

$$\left[0,\frac{x}{n}\right],\left[\frac{x}{n},\frac{2x}{n}\right],\dots\dots.\left[\frac{(n-1)x}{n},x\right]$$

where "n" the number of intervals and is chosen to be sufficiently large.

(ii) For every $y\epsilon\left[\frac{ix}{n},\frac{(i+1)x}{n}\right]$, let

$$h_u(y)=h\left(i\frac{x}{n}\right),h_l(y)=h\left((i+1)\frac{x}{n}\right)(9)$$

then $h(y)\leq h_u(y)$ for every $y\geq 0$ and $h(y)\geq h_l(y)$ for every $y\geq 0$, since $h(y)$ is a decreasing function of y. As given in [[14], [5], [6]], the upper bound to the probability of ultimate ruin is given by

$$\psi_u\left(j\frac{x}{n}\right)=\frac{\lambda}{c}h\left(j\frac{x}{n}\right)+\frac{\lambda}{c}\sum_{i=1}^{j}\{h((i-1)\frac{x}{n})-h\left(i\frac{x}{n}\right)\}\,\psi_u\left((j-i)\frac{x}{n}\right)\tag{10}$$

And the lower bound to the probability of ultimate ruin is given by

$$\psi_l\left(j\frac{x}{n}\right)=\frac{\lambda}{c}\{1-\frac{\lambda}{c}(p_1-h\left(\frac{x}{n}\right)\}^{-1}[\mathrm{h}\,(j\frac{x}{n})+\sum_{i=1}^{j-1}[h(i\frac{x}{n})-h((i+1)\frac{x}{n}]\psi_l((j-i)\frac{x}{n})]\tag{11}$$

with

$$\psi_u(0)=\psi_l(0)=\frac{\lambda p_1}{c},\quad \psi_l^{(n)}(x)\leq\psi(x)\leq\psi_u^{(n)}(x)(12)$$

where $\psi_l^{(n)}(x)$ and $\psi_u^{(n)}(x)$are respectively the lower bound and upper bound to the approximation for $\psi(x)$ at the n^{th} iteration.

$\psi(x)$ can be approximated by

$$\psi(x)\approx\frac{1}{2}\psi_l^{(n)}(x)+\frac{1}{2}\psi_u^{(n)}(x)(13)$$

An upper bound to the error of estimation is given by

$$\psi_u^{(n)}(x)-\psi_l^{(n)}(x)(14)$$

For a discussion on the stability of this algorithm, one can refer to [14]. For the implementation of this algorithm, we need to compute the function $h(x)$ for the Log Normal distribution which is as follows

Computing the function $h(x)$ for the Log Normal Distribution

Here

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_0^x\frac{1}{y}e^{\frac{-1}{2\sigma^2}(log(y)-\mu)^2}\,dy,\sigma>0,-\infty<\mu<\infty,x>0(15)$$

Hence

$$h(x)=\int_x^\infty\{1-F(y)\}dy=p_1-\int_0^x\{1-F(y)\}dy$$

where

$$\int_0^x\{1-F(y)\}dy=\int_0^x\{1-plnorm(y,\mu,\sigma)\}dy\tag{16}$$

and

$$plnorm(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\int_0^x \frac{1}{y} e^{\frac{-1}{2\sigma^2}(log(y)-\mu)^2}\, dy \qquad (17)$$

This integral has to be computed numerically.

2.3 Moments of the Time to Ruin

The moments of the time to ruin are another set of important quantities which are invariably related to the probability of ruin. An insight into the average time to ruin and the intensity of ruin can be deduced based on these quantities. Closed form expressions for the distribution of the time to ruin don't exist for most of the claim severity distributions except for Exponential, Mixture of Exponentials, and the Erlang group of distributions (15). Hence, numerical computation of the moments of the time to ruin is the only viable method for all other distributions.

Two non–negative random variables in connection to the time to ruin are the surplus just prior to the time of ruin and the deficit at the time of ruin. While the former is denoted by $U(T^-)$, where T^- is the left-hand limit of T, the deficit at the time of ruin T is denoted by $|U(T)|$. ([5]. [7],[16]).

A theme of significant focus in Actuarial Science is the expected discounted penalty function denoted by $\emptyset(u)$ and defined as

$$\emptyset(u) = E\{e^{-\delta T}W(U(T^-), |U(T)|)I(T < \infty)|U(0) = u\} \qquad (18)$$

where $I(T < \infty) = 1, if\ T < \infty$ and $I(T < \infty) = 0,$ otherwise.

Here, $W(x_1, x_2), 0 \le x_1, x_2 < \infty$ is a non negative function and $\delta \ge 0$ is interpreted as the force of interest. The quantity $W(U(T^-), |U(T)|)$ can be interpreted as the penalty at the time of ruin. For the special case when $\delta = 0$ and $W(x_1, x_2) = 1,\ , \emptyset(u)$ reduces to $\psi(u)$, the probability of ultimate ruin when initial surplus is "u" since $\psi(u) = E\{I(T < \infty)|\ U(0) = u\}$. For a further understanding on the moments of the time to ruin and the expected discounted penalty function, one can refer to [17], [18].

The reference [19] shows that the function $\emptyset(u)$ satisfies the following defective renewal equation

$$\boldsymbol{\emptyset(u) = \frac{\lambda}{c}\int_0^u \emptyset(u - x)\int_x^\infty e^{-\rho(y-x)}\, dF(y)dx + \frac{\lambda}{c}e^{\rho u}\int_u^\infty e^{-\rho u}\int_x^\infty W(x, y - x)dF(y)dx} \qquad \mathbf{(19)}$$

where $\rho = \rho(\delta)$ is the unique non-negative solution of the equation

$$c\rho - \delta = \lambda - \lambda \tilde{F}(\rho)$$

and $\tilde{F}(.)$ denotes the Laplace transform of the function $F(.)$.

The results of [20] lack mathematical tractability for their practical implementation and this problem is solved in [21] which simplified the results of [20] to make them mathematically tractable for numerical computation and have used them to calculate the approximate values for the moments of the time to ruin when explicit solutions for the probability of ultimate ruin do not exist. In their numerical computations, values of $\psi(u)$ have been calculated from the stable algorithms described in [[22], [5]].

The reference [20] shows that the k^{th} moment of the distribution of the time to ruin is given by

$$E(T^k) = \frac{\psi_k(u)}{\psi(u)}, k = 1,2,3, \dots$$

where

$$\psi_k(u) = \frac{k}{\lambda p_1 \theta}\left[\int_0^u \psi(u - x)\psi_{k-1}(x)dx + \int_u^\infty \psi_{k-1}(x)dx - \psi(u)\int_0^\infty \psi_{k-1}(x)dx\right] \qquad (20)$$

Let L be the maximum of the aggregate loss process so that $\psi(u) = pr(L > u)$ (see[23], formula (13.6.2)).

In [13], it has been shown that

$$E(L) = \int_0^\infty \psi(x)dx = \frac{p_2}{2\theta p_1}, E(L^2) = 2\int_0^\infty x\psi(x)dx = \frac{p_3}{2\theta p_1} + \left(\frac{1}{2}\right)\left(\frac{p_2}{2\theta p_1}\right)^2 \qquad (21)$$

$\psi_1(u)$ appearing in [20] has been simplified in [21] as

$$\psi_1(u) = \frac{1}{\lambda p_1 \theta}\{E(L)\delta(u) - \int_0^u \psi(x)\delta(u-x)dx\} \tag{22}$$

where, $\psi_1(u)$ can be evaluated numerically in the absence of any explicit form of expression for $\psi(u)$

Similarly, $\psi_2(u)$ appearing in [20], has been simplified in [21], and the simplified form is given by

$$\psi_2(u) = \frac{2}{\lambda p_1 \theta}\{\frac{E(L^2)\delta(u)}{2\lambda p_1 \theta} - \int_0^u \psi_1(x)\delta(u-x)dx\} \tag{23}$$

As has been indicated earlier, explicit expressions for the moments of the time to ruin can't be obtained owing to the lack of analytical ways for solving the integrals appearing in Eq. (22) and Eq. (23), the exception being in the cases of Mixture of Exponential distribution and Erlang group of distributions. For any distribution barring these, the integrals can be evaluated only numerically. Due to the non-explicit form of [21], the method of numerical integration has been used, and the values of $\psi(u)$ is obtained through a stable recursive algorithm mentioned in section (2.3).

2.5 Probability of Ruin in the Presence of Interest Earnings and Tax Payments

The classical risk model excludes the influence of interest earnings and tax payments on the surplus process but under the influence of these factors that is interest earnings and tax payments, the surplus process exhibits some interesting properties and consequently, a modified approach to compute the probability of ruin under these influences is required. Albrecher- Hipp tax identity (see [24], [25]) is an important result in this context where the modified surplus at a time t is denoted by $U_\gamma(t)$ and this modified surplus carries within it a component of tax, paid at a fixed rate γ whenever the insurer is in a profitable position.

In the context of this modification let $\psi_{\delta,\gamma}(u)$ and $\varphi_{\delta,\gamma}(u)$ respectively denote the ruin and non ruin probabilities. The following result from [26] has been used to compute the probability of ultimate ruin in the presence of interest earnings and tax payments. We assume that the claim size distribution F and its equilibrium distribution $F_1$are both sub exponential and $1 < J_*(F) \leq \infty$, where $J_*(F)$ is defined in eq (4.1) of [26]. It can be noted that all these assumptions are true for log normal distribution. Under this setup, the probability of ultimate ruin is given by

$$\psi_{\delta,\gamma}(u)\int_u^\infty \frac{\lambda \acute{F}(x)}{(c+\delta x)\big(1-\gamma(x)\big)}dx \tag{24}$$

In practical situations, the choice of investment made by the insurance company determines the interest rate δ and fiscal policies of the country concerned governs the tax rate γ. However, with an objective to simplify, we have taken the tax structure as given in [26] for our computation and δ is fixed at a level of $\delta = 0.05$

The tax structure used is

$$\gamma(x) = \begin{Bmatrix} 0.10, & 0 < x \leq 10^4 \\ 0.18, & 10^4 < x \leq 10^5 \\ 0.30, & 10^5 < x \leq 10^6 \\ 0.50, & x > 10^6 \end{Bmatrix} \tag{25}$$

The following is the computation of the probability of ruin [5,6] for a lognormal distribution under interest earnings and tax payments.

Here, we have

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_0^x \frac{1}{y} e^{\frac{-1}{2\sigma^2}\left(\frac{logy-\mu}{\sigma}\right)^2} dy$$

Therefore,

$$\psi_{\delta,\gamma}(u) = \int_u^\infty \frac{\lambda\left\{1 - \frac{1}{\sqrt{2\pi}\sigma}\int_0^x \frac{1}{y} e^{\frac{-1}{2\sigma^2}\left(\frac{logy-\mu}{\sigma}\right)^2} dy\right\}}{(c+\delta x)\big(1-\gamma(x)\big)}dx \tag{26}$$

Changing the scale, we have

$$\psi_{\delta,\gamma}(u) = \frac{\lambda}{0.90}\int_{0}^{10^4-u}\frac{1-plnorm(y+u,\mu,\sigma)}{\{c+\delta(y+u)\}}dy + \frac{\lambda}{0.82}\int_{10^4-u}^{10^5-u}\frac{1-plnorm(y+u,\mu,\sigma)}{\{c+\delta(y+u)\}}dy$$
$$+\frac{\lambda}{0.70}\int_{10^5-u}^{10^6-u}\frac{1-plnorm(y+u,\mu,\sigma)}{\{c+\delta(y+u)\}}dy + \frac{\lambda}{0.50}\int_{10^6-u}^{\infty}\frac{1-plnorm(y+u,\mu,\sigma)}{\{c+\delta(y+u)\}}dy \qquad (27)$$

where,

$$plnorm(y,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\int_{0}^{x}\frac{1}{y}e^{\frac{-1}{2\sigma^2}\left(\frac{log y-\mu}{\sigma}\right)^2}dy$$

The above integral has been computed numerically

All the computations are done through R programming [27].

3. Results and Discussions

Data: Our data consists of a collection of 160000 claim amounts that were collected from an Indian general insurance company's motor insurance portfolio, which included all of its Indian branches, during the course of six months, from April 2013 to September 2013. No adjustment is made for inflation for the time horizon is narrow. It needs to be mentioned that the data is utilized more for the illustration of the various methodologies rather than for the extraction of any concrete meaningful conclusion from the data itself. Since the inter arrival time of the claim is difficult to track, an illustrative value of the intensity parameter is taken as $\lambda = 32.427$.

Summary statistics shown in Table 1 as well as the graphical display shown in Figure 1, indicate the existence of a high degree of positive skewness in the data and this in turn is indicative of the fact that Log Normal could be a potential model for the data.

Table 1. Summary Statistics for the Insurance claim data [5]

Sample Size	Mean	Standard deviation	Min	25% Quantile	Median	75% Quantile	Max	Skewness	Kurtosis
160000	1.78834e+04	22805.81	523	6043.00	10583.00	19374.25	188209	3.576628	18.94972

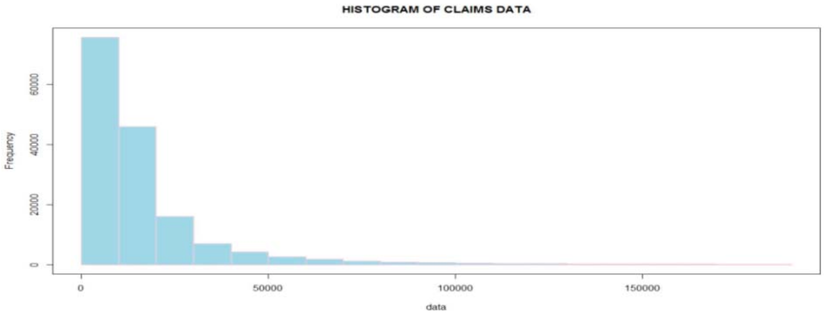


Figure 1. Histogram of the observed claim data on motor insurance

The parameter estimates of the Log Normal distribution have been found as $\hat{\mu} = 9.327069$ and $\hat{\sigma} = 0.9254849.$ However, the Log Normal distribution could not qualify the goodness of fit tests by means of two EDF(Empirical Distribution Function) statistics which we are being used, namely the Cramer Von Mises (the value is found to be 70.57441 with p-value<0.05) and Anderson Darling tests (The value of this statistic is found to be 484.3411 with p-value<0.05). However, in conjunction with the real objectives of this work, we have retained these estimated values for the parameters of the Log Normal distribution. These estimated values are then, used as the values of the parameters of the Log Normal for computing the various actuarial quantities under consideration.

Table 2 shows the probability of ultimate ruin for the Log Normal distribution obtained through the stable recursive algorithm with an illustrative value of security loading factor as $\theta = 0.3.$ These values are found to be decreasing with an increase in the initial surplus and this is consistent with its expected behaviour since a larger initial surplus should diminish the chance of ruin, if any. The greatest advantage of the stable recursive algorithm is that it is fast, accurate, stable and produces the bounds to the error of estimation for the probability of ultimate ruin. The greatest challenge of this method lies in computing the function $h(x)$ which implies integrating the equilibrium distribution of the Log Normal which has no closed form and the problem is compounded by the fact that even the distribution function of the Log Normal distribution does not have a closed form expression. Hence, the emergence of error from these two sources has contributed to the overall error in the evaluation of the probability of ruin by the stable recursive algorithm.

Table 2. The probability of ultimate ruin for the Log Normal distribution fitted to our set of insurance data ($\hat{\mu} = 9.327069$ and $\hat{\sigma} = 0.9254849$) computed through the stable recursive algorithm

Initial Surplus ($u\ in\ Rs$)	**Lower bound to the probability of ultimate ruin**	**Upper bound to the probability of ultimate ruin**	**Probability of ultimate Ruin ($\psi(u)$)**
10	0.7691278	0.7691278	0.7691278
20	0.7690248	0.7690248	0.7690248
30	0.7689218	0.7689218	0.7689218
40	0.7688187	0.7688187	0.7688187
50	0.7687155	0.7687155	0.7687155
60	0.7686123	0.7686124	0.7686124
70	0.7685091	0.7685091	0.7685091
80	0.7684058	0.7684059	0.7684058
90	0.7683025	0.7683025	0.7683025
100	0.7681991	0.7681992	0.7681992
200	0.7671628	0.7671630	0.7671629
500	0.7640260	0.7640275	0.7640268
1000	0.7587123	0.7587182	0.7587152

Table 3 shows the first moment of the time to ruin for our fitted Log Normal distribution and it is found to be increasing with an increase in the initial surplus. This trend is obvious, for with the increase in the initial surplus, the time to ruin should get delayed. An interpretation of a typical value in Table 3 is that starting with an initial surplus of Rs 100, it would on an average take.0.1216053 years for the surplus process to be less than or equal to zero for the first time, thereby leading to ruin in the sense of its definition.

Table 3. The first moment of the time to Ruin for the Log Normal distribution fitted to our set of insurance data with an illustrative value for the intensity parameter as λ=32.427

Initial Surplus ($u\ in\ Rs$)	**First moment (Mean in years)**
10	0.1210953
20	0.1211518
30	0.1212084
40	0.1212649
50	0.1213215
60	0.1213782
70	0.1214349
80	0.1214917
90	0.1215484
100	0.1216053

Table 4 shows the second moment of the time to ruin. It is also found to be increasing with an increase in the initial surplus. We have not gone to the extent of identifying the cause behind this increase in heterogeneity in the time to ruin, with an increase in the initial surplus. However, it is also a fact there is no intuitive assumption regarding the behaviour of the second moment of the time to ruin.

Table 4. Second moment of the time to Ruin for the Log Normal distribution fitted to our set of insurance data with intensity parameter λ=32.427

Initial Surplus ($u\ is\ in\ Rs$)	Second Moment
10	0.1437515
20	0.1438315
30	0.1439116
40	0.1439913
50	0.1440715

It needs to be noted that the first moment of the time to ruin is computed using Eq.(22) and Eq. (23) and the quantities$\psi(x)$ and $\delta(x)$ appearing therein, are computed through the stable recursive algorithm as described in section (2.3). The execution procedure for the computation of the second moment is rather more complex for it amounts to simultaneous handling of three numerical computations in a nested order aimed at the evaluation of $\psi(x)$, $\psi_1(x)$ and then finally $\psi_2(x)$. This gives rise to several issues pertaining to the occurrence of error and the overall execution time. Suitable adjustments are made in the interval of discretization in the numerical integration which in our case is Simpson's $\frac{1}{3}^{rd}$ rule for numerical integration. The steep increase in the execution time with an increase in the value of initial surplus (u)constraints us to limit the values of u to relatively small amounts.

Table 5 shows the probability of ruin in the presence of interest rates and taxes for the Log Normal distribution and it is found to be decreasing with an increase in the initial surplus and this is consistent with its expected behaviour. An occurrence of some amount of error is inevitable owing to the numerical integration being carried out to evaluate the integral in Eq. (27). We have used the *integrate* function in R to evaluate these integrals and the error accumulated is not more than 1e-15.

Table 5. The probability of ultimate ruin for Log Normal Distribution fitted to our set of insurance data under the tax structure given by Eq. (25) and rate of interest δ=0.05

Value of the initial surplus u (in Rs)	$\psi_{\delta,\gamma}(u)$
10	0.8548930
20	0.8543974
30	0.8539018
40	0.8534062
50	0.8529106
60	0.8524150
70	0.8519194
80	0.8514239
90	0.8509283
100	0.8504327
200	0.8494415
500	0.8479547
1000	0.8058814

It is found that the net impact of interest earnings and tax payments is positive; that is, the probability of ultimate ruin under the presence of interest earnings and tax payments is found to be increasing as compared to that obtained in the absence of these two factors. It further needs to be mentioned that the interest rate used is also purely illustrative and the tax structure is taken from [26], although the tax structure prevalent in India during the time, the data was collected would have been more realistic.

4. Concluding Remarks

We have computed the probability of ultimate ruin when the underlying claim severity is distributed as Log Normal distribution. Apart from it, we have dealt with the computation of another important actuarial quantity namely the moments of the time to ruin for this claim severity distribution. An important aspect of this work has been to reassess the Probability of ultimate ruin in the presence of interest rates and tax payments. Each of the computed quantities has its importance in assessing the operational dimension of an insurance company. The computations are constrained by technical difficulties including the limited capacity of a personal computer. The work focuses on the difficulty of handling multidimensional integrals where the integrand itself does not have a closed form expression and is to be computed numerically through algorithms which are themselves loaded with several challenges. The algorithms that have been used can serve to provide guidance in constructing their improved versions with some consolidated effort to quantify the error associated with the computation and to minimize it.

References

Albrecher, H., & Hipp, C. (2007). Lundberg's risk process with tax. *Blätter der DGVFM, 28*(1), 13-28. https://doi.org/10.1007/s11857-007-0004-4

Asmussen, S., & Albrecher, H. (2010). *Ruin Probabilities* (2nd ed.). Advanced Series on Statistical Science and Applied Probability, Vol. 14.

Assmusen, S. (2000). *Ruin Probabilities*. Singapore: World Scientific. https://doi.org/10.1142/2779

Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A., & Nesbitt, C. J. (1998). *Actuarial Mathematics*. Itasca, Illinois: Society of Actuaries.

Čížek, P., Weron, R., Härdle, W., Burnecki, K., Misiorek, A., & Weron, R. (2005). Loss distributions. In *Statistical Tools for Finance and Insurance* (pp. 289-317). https://doi.org/10.1007/3-540-27395-6_13

Das, J. (2016). Actuarial Risk Process and the Probability of Ruin for non Life insurance [Doctoral dissertation]. Retrieved from http://hdl.handle.net/10603/192119

Das, J. (2022). Computation of the probability of ultimate ruin and some other actuarial quantities under the classical risk model via fast Fourier transform. *Model Assisted Statistics and Applications, 17*(1), 15-25. https://doi.org/10.3233/mas-220004

Das, J., & Nath, D. C. (2016). Burr distribution as an actuarial risk model and computation of some of its actuarial quantities related to the probability of ruin. *Journal of Mathematical Finance, 6*(2), 213-231. http://dx.doi.org/10.4236/jmf.2016.61019

Das, J., & Nath, D. C. (2019). Weibull distribution as an actuarial risk model: Computation of its probability of ultimate ruin and moments of the time to ruin, deficit at ruin and surplus prior to ruin. *Journal of Data Science, 17*(1), 161-149. https://doi.org/10.6339/jds.201901_17(1).0008

Dickson, D. C., & Waters, H. R. (2002). The distribution of the time to ruin in the classical risk model. *ASTIN Bulletin: The Journal of the IAA, 32*(2), 299-313. https://doi.org/10.2143/AST.32.2.1031

Dickson, D. C., dos Reis, A. D., & Waters, H. R. (1995). Some stable algorithms in ruin theory and their applications. *ASTIN Bulletin: The Journal of the IAA, 25*(2), 153-75. https://doi.org/10.2143/AST.25.2.563245

Gerber, H. U. (1979). *An Introduction to Mathematical Risk Theory*. Homewood: Richard D. Irwin Inc.

Gerber, H. U., & Shiu, E. S. (1998). On the time value of ruin. *North American Actuarial Journal, 2*(1), 48-72. https://doi.org/10.1080/10920277.1998.10595671

Goovaerts, M., & de Vylder, F. (1984). A stable recursive algorithm for evaluation of ultimate ruin probabilities. *ASTIN Bulletin: The Journal of the IAA, 14*(1), 53-9. https://doi.org/10.1017/S0515036100004803

Grandell, J. (1991). *Aspects of risk theory*. New York: Springer. https://doi.org/10.1007/978-1-4613-9058-9

Hogg, R. V., Klugman, S. A., & Loss, D. (1984). *Loss Distributions*. New York: Wiley. https://doi.org/10.1002/9780470316634

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (1998). *Loss Models: From Data to Decisions*. New York: Wiley.

Lin, X. S., & Willmot, G. E. (1999). Analysis of a defective renewal equation arising in ruin theory. *Insurance: Mathematics and Economics, 25*(1), 63-84. https://doi.org/10.1016/S0167-6687(99)00026-8

Lin, X. S., & Willmot, G. E. (2000). The moments of the time of ruin, the surplus before ruin, and the deficit at ruin. *Insurance: Mathematics and Economics, 27*(1), 19-44. https://doi.org/10.1016/S0167-6687(00)00038-X

Nath, D. C., & Das, J. (2016). Modelling of insurance data through two heavy tailed distributions: Computation of some of their actuarial quantities through simulation from their equilibrium distribution and the use of their convolutions. *Journal of Mathematical Finance, 6*(3), 378-400. https://doi.org/10.4236/jmf.2016.63031

Nath, D. C., & Das, J. (2017). Modelling of claim severity through the mixture of exponential distribution and computation of its probability of ultimate ruin. *Thailand Statistician, 15*(2), 128-148. Retrieved from https://ph02.tci-thaijo.org/index.php/thaistat/issue/view/8422

Panjer, H., & Willmot, G. (1992). *Insurance risk models*. Schaumburg: Society of Actuaries.

Promislow, D. S. (2006). *Fundamental of Actuarial Statistics*. Chichester: John Wiley and Sons.

R Development Core Team. (2009). *A language and environment for statistical computing*. Retrieved from http://www.R-project.org

Thampi, K. K., & Jacob, M. J. (2008). Moments of the time of ruin in a renewal risk model with discounted penalty. *International Journal of Management Science and Engineering Management, 9*(2), 173-187. https://doi.org/10.1108/15265940810853922

Wei, L. (2009). Ruin probability in the presence of interest earnings and tax payments. *Insurance: Mathematics and Economics, 45*(1), 133-138. https://doi.org/10.1016/j.insmatheco.2009.05.004

Willmot, G. E., & Woo, J. K. (2017). *Surplus Analysis of Sparre Andersen Insurance Risk Processes*. Springer Actuarial Series.

A Note on the Discrete Analogue of the Pearsonian System of Curves

Abstract

Frequency distributions are the arrangement of statistics to show the number of times or frequency with which, an event occurs in a particular way. A generalized System of Curves is important as it describes frequency distributions for as wide a variety of observed distributions as possible. One of these systems that became the most successful was the Pearsonian System of Curves in 1850. The discrete analogue of the Pearsonian System of Curves was first discussed in 1967 by J.K.Ord, where the difference equation employed to define the class of distributions was based on four parameters. This distribution was the basis of many important studies. However, there exists some discrepancies in the values of the parameters as observed in this study. Considering the importance of the distribution, the present paper gives a detail of the errors and corrected values of the parameters. Some discussion on the resultant value of kuppa (the parameter determining the 'type' of the distribution) and the complete distribution in the system has been added too. Further studies based on this distribution may follow the corrected values of the parameters.

Keywords: Pearsonian System of Curves, discrete case, J.K.Ord, correction

1. Introduction

The Pearson system of continuous distribution is defined by the differential equation

$$\frac{df}{dx} = \frac{(a-r)f_r}{b_0+b_1x+b_2x^2} \tag{1}$$

where f(x) is the density function when the random variable *X*=*x* and *a*,*b*$_0$,*b*$_1$,*b*$_2$ are parameters. This system provides a framework for discussing several important continuous distributions. Its utility has been exponentiated time to time [1-2].

Analogously, it was derived by [3] that we may employ the difference equation

$$\frac{df}{dx} = \frac{(a-r)f_r}{b_0+b_1r+b_2r(r-1)} \tag{2}$$

to define a class of discrete distributions, based on a lattice of unit width (without loss of generality). This system was studied earlier to Ord by [4-5] and a special case by [6-7].

The moments relation was found for the distribution (2) by multiplying throughout by

$$r^{(j)} = r(r-1)(r-2)\,\ldots\ldots\ldots.(r-j+1) \tag{3}$$

and summing over [u,v], the range of r, the factorial moments about the origin, $\mu'_{(j)}$ satisfy the relation

$$\{(j+2)b_2-1\}\mu'_{(j+1)} + \{(j+1)(b_1+2jb_2)+a-2j-1\}\mu'_{(j)} + \{b_0+b_1j+b_2j(j-1)+a-j\}j\mu'_{(j-1)} + E_{j+1} = 0 \tag{4}$$

for $j = 0,1,2,3,\ldots..$

where

$$E_{j+1} = \{b_0+b_1u+b_2u(u-1)\}u^{(j)}f_u - \{a+b_0+(v+1)(b_1-1+vb_2)\}(v+1)^{(j)}f_v \tag{5}$$

and

$$\mu'_{(-1)} = 0, \mu'_0 = 1$$

The calculations of the parameters of the difference equation done by Ord were followed and few discrepancies were noted. The corrected values are detailed as follows:

The complete distributions may be taken to have the range [0, *N*] where *N* may be infinite or $(-\infty,\infty)$. If the range is

i) Doubly infinite; or

ii) Has $b_0 = 0$, lower terminal zero and has when *N* is finite

$a + (b_1 + b_1N - 1)(N + 1) = 0$

Then $E_{j+1}$is zero for all *j*.

Calculation of parameters

Case I: Discrete with range [0, *N*]

Considering the specifications, we proceed as follows:

Putting $j = 0$ in (4)

$$(2b_2 - 1)\mu'_{(1)} + (b_1 + a - 1)\mu'_{(0)} = 0$$

$$\xrightarrow{yields} a = (1 - 2b_2)\mu + 1 - b_1 \tag{6}$$

Next putting $j = 1$ in (4), we get

$$(3b_2 - 1)\mu'_{(2)} + (2b_1 + 4b_2 + a - 3)\mu'_{(1)} + (b_0 + b_1 a - 1) = 0$$

$$\xrightarrow{yields} b_1 = \frac{\mu_2 - b_2(3\mu_2 - \mu + \mu^2)}{\mu} \tag{7}$$

Again putting $j = 2$ in (4), we get

$$(4b_2 - 1)\mu'_{(3)} + (3b_1 + 12b_2 + a - 5)\mu'_{(2)} + (b_0 + 2b_1 + 2b_2 + a - 2)2\mu'_{(1)} = 0$$

$$\xrightarrow{yields} b_2 = \frac{\mu\mu_3 + \mu\mu_2 - 2\mu_2^2}{4\mu\mu_3 + 2\mu\mu_2 + 2\mu_2\mu^2 - 6\mu_2^2} \tag{8}$$

Table (**Table 1**) of compares values of the parameters (in terms of the first three (or four) moments is given below:

Table 1. Comparison of values of parameter of discrete PSC for range [0, N]

Parameter	**By Ord**	**Corrected**
a	$(1 - 2b_2)(1 + \mu) - b_1$	$(1 - 2b_2)\mu + 1 - b_1$
b_0	0	0
b_1	$\dfrac{\mu_2 - b_2(3\mu_2^2 - \mu + \mu^2 - 2)}{\mu}$	$\dfrac{\mu_2 - b_2(3\mu_2 - \mu + \mu^2)}{\mu}$
b_2	$\dfrac{\mu_3 + \mu\mu_2 - 2\mu_2^2}{D_G}$	$\dfrac{\mu_3 + \mu\mu_2 - 2\mu_2^2}{D_G^M}$

where

$D_G = 4\mu\mu_3 + 2\mu_2(2 + \mu_2 - 3\mu_2)$

and

$D_G^M = 4\mu\mu_3 + 2\mu_2(\mu + \mu^2 - 3\mu_2)$

Case II: Discrete with range $(-\infty, \infty)$

Here, we consider $\mu = 0$, for the result of making $\mu = 0$is to change the origin of the system to the mean of the distribution.

Putting $j = 0$ in equation (4), we get

$$(2b_2 - 1)\mu'_{(1)} + (b_1 + a - 1)\mu'_{(0)} + (b_0 + a).0.\mu'_{(-1)} = 0$$

$$\xrightarrow{yields} b_1 + a - 1 = 0 \tag{9}$$

Putting $j = 1$ in equation (4), we get

$$(3b_2 - 1)\mu'_{(2)} + (2b_1 + 4b_2 + a - 3)\mu'_{(1)} + (b_0 + b_1 a - 1).1.\mu'_{(0)} = 0$$

$$\xrightarrow{yields} 3b_2\mu_2 - \mu_2 + b_0 + b_1 + a - 1 = 0 \tag{10}$$

Putting $j = 2$ in equation (4), we get

$$(4b_2 - 1)\mu'_{(3)} + (3b_1 + 12b_2 + a - 5)\mu'_{(2)} + (b_0 + 2b_1 + 2b_2 + a - 2).2.\mu'_{(1)} = 0$$

$$\xrightarrow{yields} 4b_2\mu_3 + 3b_1\mu_2 + a\mu_2 - \mu_3 - 2\mu_2 = 0 \tag{11}$$

Putting $j = 3$ in equation (4), we get

$$(5b_2 - 1)\mu'_{(4)} +$$

$$\xrightarrow{yields} 5b_2\mu_4 + 6b_2\mu_3 + b_2\mu_2 + 4b_1\mu_3 - 3b_1\mu_2 + a\mu_3 + 3b_0\mu_2 - \mu_4 - \mu_3 + \mu_2 = 0 \tag{12}$$

Now from equation (9) we get

$$b_1 = 1 - a \tag{13}$$

From equation (10) we get

$$b_0 = -3b_1\mu_2 + \mu_2 \tag{14}$$

From equation (11) we get

$$4b_2\mu_3 + 3b_1\mu_2 + a\mu_2 - \mu_3 - 2\mu_2 = 0$$

$$\xrightarrow{yields} a = \frac{-\mu_3 + \mu_2 + 4b_2\mu_3}{2\mu_2} \tag{15}$$

From equation (12) we get

$$5b_2\mu_4 + 6b_2\mu_3 + b_2\mu_2 + 4b_1\mu_3 - 3b_1\mu_2 + a\mu_3 + 3b_0\mu_2 - \mu_4 - \mu_3 + \mu_2 = 0$$

$$\xrightarrow{yields} b_2 = \frac{2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3 + \mu_2^2}{2(5\mu_2\mu_4 - 9\mu_2^3 - 6\mu_3^2 + \mu_2^2)}$$

Given, $\sqrt{\mu_2}$, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ and $\beta_2 = \frac{\mu_4}{\mu_2^2}$

The above equation can be written as

$$b_2 = \frac{\left(2\beta_2 - 3\beta_1 - 6 + \frac{1}{\mu_2}\right)}{2\left(5\beta_2 - 6\beta_1 - 9 + \frac{1}{\mu_2}\right)} \tag{16}$$

Putting value of b_2 in equation (15), we get

$$a = \frac{-\mu_3 + \mu_2 + 4\mu_3\left[\frac{2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3 + \mu_2^2}{2(5\mu_2\mu_4 - 9\mu_2^3 - 6\mu_3^2 + \mu_2^2)}\right]}{2\mu_2}$$

$$= \frac{\frac{-\mu_3}{\mu_2}\left(\beta_2 + 3 - \frac{1}{\mu_2}\right)}{2\left(5\beta_2 - 6\beta_1 - 9 + \frac{1}{\mu_2}\right)} + \frac{1}{2} \tag{17}$$

Putting value of b_2 in equation (14), we get

$$b_0 = -3\mu_2\left[\frac{2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3 + \mu_2^2}{2(5\mu_2\mu_4 - 9\mu_2^3 - 6\mu_3^2 + \mu_2^2)}\right] + \mu_2$$

$$= \frac{\mu_2\left(4\beta_2 - 3\beta_1 - \frac{1}{\mu_2}\right)}{2\left(5\beta_2 - 6\beta_1 - 9 + \frac{1}{\mu_2}\right)} \tag{18}$$

The results obtained in this case are found to tally with the results obtained by Ord, except a possible typing mistake of “/”. The comparison of the results is shown in Table 2.

Table 2. Comparison of values of parameter of discrete PSC for range $(-\infty, \infty)$

Parameter	By Ord	Corrected
a	$\frac{1}{2} - \frac{\mu_3}{\mu_2}\left(\beta_2 + 3 - \frac{1}{\mu_2}\right) D_F$	$\frac{1}{2} - \frac{\frac{\mu_3}{\mu_2}\left(\beta_2 + 3 - \frac{1}{\mu_2}\right)}{D_F}$
b_0	$\mu_2\left(4\beta_2 - 3\beta_1 - \frac{1}{\mu_2}\right) D_F$	$\frac{\mu_2\left(4\beta_2 - 3\beta_1 - \frac{1}{\mu_2}\right)}{D_F}$
b_1	$1 - a$	$1 - a$
b_2	$\left(2\beta_2 - 3\beta_1 - 6 + \frac{1}{\mu_2}\right) D_F$	$\frac{\left(2\beta_2 - 3\beta_1 - 6 + \frac{1}{\mu_2}\right)}{D_F}$

where

$$D_F = 2\left(5\beta_2 - 6\beta_1 - 9 + \frac{1}{\mu_2}\right)$$

Distributions contained in the system

The alternative form of equation (2) can be written as

$$\Delta f_{r-1} = \frac{(a-r)f}{(a+b_0)+(b_1-1)r+b_2 r(r-1)} \tag{19}$$

The form of the density function will depend on the behavior of the roots of the denominator in equation (2). We may use this behavior to distinguish different forms for the continuous distributions.

Given, $(\kappa) = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)}$, we get

$$\kappa = \frac{(b_1-b_2-1)^2}{4b_2(b_0+a)} \quad \text{, based on equation} \tag{19}$$

Ord defined 1 as the index of dispersion by $I = \frac{\mu_2}{\mu_1'}$ and also uses $S = \frac{\mu_3}{\mu_2}$.

Using these two measures, the corrected values of kappa for the two ranges (based on the corrected values of the four parameters) are as follows:

i) Range [0,N]

$$\kappa = \frac{\left[\beta_2(I+12) + 15\beta_1 + \frac{1}{\mu_2}(I+3) - 3(I+8)\right]^2}{4\left(2\beta_2 - 3\beta_1 - 6 + \frac{1}{\mu_2}\right)\left[\beta_2(I - 4\mu_2 - 10) + \beta_1(3\mu_2 + 16) - \frac{1}{\mu_2}(I+2) + 3(I+6) - \mu_2\right]}$$

ii) Range $(-\infty, \infty)$

$$\kappa = \frac{[\mu_3(I - \mu - 4) + \mu_2(4\mu_2 + 5I - 3\mu - 2)]^2}{4\left(\frac{\mu_3}{\mu} + I - 4\right)\left[\mu_3\left(3 + \frac{3}{\mu} - \frac{I}{\mu}\right) + \mu_2\left(2\mu - 4I + \frac{1}{\mu} + 3 - 3\frac{I}{\mu}\right)\right]}$$

The complete distributions in the system, obtained by [8], are summarized in Table 3, along with their density functions and κ, I values, where relevant. The type numbers associated with each distribution correspond, as far as possible, to those of the analogous Pearson curve; the letter c(d) being written after the type number to distinguish the continuous (discrete) form.

Table 3. Distributions contained in the system with range, density function and criteria values

Type	Name	Density	Criteria	Range	Comments
I (a)	Hypergeometric	$\frac{\binom{Np}{r}\binom{Nq}{n-r}}{\binom{N}{n}}$	$I < 1.\kappa > 1$	$[0, m]$ $m = \min(n, Np)$	J- or bell- shaped
I (b)	Negative hypergeometric or beta-binomial	$\frac{\binom{k+r-1}{r}\binom{N-k-r}{Np-r}}{\binom{N}{Np}}$	$\kappa < 1$	$[0, Np]$	J- or bell- shaped
I (e)	-	$\frac{\binom{A}{r}\binom{C}{B-r}}{\binom{A+C}{B}}$	$\kappa > 1$	$[0, \infty)$	A, C non-integer, but have the same integral part
I (u)	-	$\alpha\left\{\binom{A}{C+r}\binom{B}{D-r}\right\}^{-1}$	$\kappa > 1$	$[0, n)$	U- shaped
VI	Beta-Pascal	$\frac{A}{(k+a)}\frac{\binom{k+r-1}{r}\binom{A+B-1}{A}}{\binom{k+A+B+r-1}{k+a}}$	$I > 1.\kappa > \infty$	$[0, \infty)$	J- or bell- shaped
IV	-	$\frac{\alpha Q(r, a, d)}{Q(r, k = a, b)}, r > 0$	$0 < \kappa < 1$	$(-\infty, \infty)$	K a positive integer, bell-shaped
II (a) II (b) II (u)	As for I(.)	Similar expression for $r < 0$	$\left\{\begin{matrix} I < 1, \kappa = 1 \\ \kappa = 0 \\ \kappa = 1 \end{matrix}\right\}$	As for type I(.)	Symmetric forms of type I(.)
V	-	As type IV, but $b = 0$	$\kappa = 0$	$[0, \infty)$ or $(-\infty, \infty)$	Limiting form of IV
III (b)	Binomial	$\binom{n}{r} p^r (1-p)^{n-r}$	$I < 1.\kappa \rightarrow \infty$	$[0, n)$	Limiting form of I(a), I(b)
III (n)	Negative binomial or Pascal	$\binom{k+r-1}{r} p^k (1-p)^r$	$I > 1.\kappa \rightarrow \infty$	$[0, \infty)$	Limiting form of I(b), VI
III (p)	Poisson	$e^{-m}m^r/_{r!}$	$I = 1.\kappa \rightarrow \infty$	$[0, \infty)$	Limiting form of III(b), III(n)
VII	Discrete Student's t	$\alpha\left[\prod_{j=1}^{k}\{(j+r+a)^2 + b^2\}\right]^{-1}$	$0 < \kappa < 1$	$(-\infty, \infty)$	'Nearly symmetric form of IV

Acknowledgements

This work is a part of the thesis study of the first author under the guidance of the second author. The first author expresses her gratitude to the second author for all the support.

References

Solomon, H., & Stephens, M. A. (1978). Approximations to density functions using Pearson curves. *Journal of the American Statistical Association, 73*(361), 153-160. https://doi.org/10.2307/2286537

DasGupta, A. (2010). Standard discrete distributions. In *Fundamentals of Probability: A First Course*. Springer Texts in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-5780-1_6

Ord, J. K. (1976). On a system of discrete distributions. *Biometrika, 54*(3/4), 649-656. https://doi.org/10.2307/2335056

Carver, H. C. (1919). On the graduation of frequency distribution. *Proceedings of the Casualty Actuarial Society, 6*, 52-72. Retrieved from https://www.casact.org/pubs/proceed/proceed19/19052.pdf

Katz, L. (1948). Frequency functions defined by the Pearson difference equation. *Annals of Mathematical Statistics Abstracts, 19*, 120.

Tweedie, M. C. K. (1947). Functions of a statistical variate with given means, with special reference to Laplace distribution. *Proceedings of the Cambridge Philosophical Society, 43*, 41-49. https://doi.org/10.1017/S0305004100023185

Noack, A. (1950). A class of random variables with discrete distributions. *Annals of Mathematical Statistics, 21*, 127-132. https://doi.org/10.1214/aoms/1177729894

Measuring Speed of Aging Process: An Illustration with the Population of Bangladesh

Abstract

The conventional aging measures are crude as these are silent on the demographic components of population change. This chapter is an exercise to improve the speed of aging measures as a function of the demographic components that are responsible for population growth. The measures of aging acceleration, used here, are functional relationship between life table mortality and fertility rate under the assumption of stable population. These are tested with the existing measures of aging velocity using census data of Bangladeshi population for the census years 1981 and 2001. The results show that these are good alternative and consistent measures over the existing methods. These alternative approaches indicate a slower aging process than those obtained by the existing measures. The gender and urban-rural gap in the rising aging process is noticeable for Bangladeshi population. Therefore, a consideration of both the existing and proposed modified measures of pace of demographic aging will be helpful for assessing the process precisely.

Keywords: Population aging, pace of aging, measures of aging process, Population of Bangladesh

1. Introduction

Population aging from a demographic point of view is a natural process, generated by demographic transitions. It is not a demographic crisis, it mirrors a general trend of human development aimed at achieving longevity and wellbeing. On the contrary, population aging poses long-term challenges to society. Therefore, demographic aging should be rated as one of the most important processes in developed countries and even worldwide.

Over the 20th century, the population of Bangladesh was young and at the dawn of the 21st century, its population has entered into the intermediate level of aging. The age structure of Bangladeshi population is changing markedly. A small proportion (over 6 percent) of the total population constitutes the elderly population, but the absolute number of them is quite significant (7.9 million) and by the year 2050 elderly share will be 16 percent of the total population (1). This change in the population characteristics will have serious consequences on society as well as on the overall socio-economic development of the country.

In recent years, in response to an increasing concern in the developed world with the rapid increase in the aging of populations, there has occurred a great expansion in the literature trying to measure the population aging, forecast trends and analysis of socio-economic implications. Yet, despite this great proliferation the existing measures of aging continue to be crude (2). The population size, growth, structure and aging all are influenced by the demographic components of fertility, mortality, and migration. The conventional aging measures are crude as these are silent on these components. Measuring aging as a function of demographic components of population change is an attempt to improve these crude measures. This work is an exercise to improve such type of measures. The work of Preston, Himes and Eggers (1989) was a milestone on such type of measures (3). Liao (1996) suggested such measures of aging as a function of crude birth rate (CBR), crude death rate (CDR) and migration rate (4). Some improved measures have been proposed by several authors (5–7). Here an attempt has been made to modify the measures of aging velocity as a function of demographic components. The basic idea of these measures is to use the stable population fertility (adjusted with NRR and population growth) and life table mortality. In this study, life table death rate has been used as an alternative to the crude death rate and the net reproduction rate (NRR) has been used as an alternative to the birth rate respectively. All the measures have been applied to the Bangladesh census population of 1981 and 2001 to measure the speed of aging process and a comparison is made among these measures.

2. Methods of the Study

2.1 Existing Measures

2.1.1 Aging measure given by Preston, Himes and Eggers (1989)

The rate of change of mean age in terms of birth and death rates per unit of time as

$$\frac{\mathrm{dA}p(t)}{\mathrm{dt}} = 1 - d(t)\left[A_D(t) - A_p(t)\right] - b(t) \times A_p(t) \tag{1}$$

where A_p is the mean age of the population, A_D is the mean age at death, b (t) is the birth rate of the population at time t (number of births per person-year lived), d (t) is the death rate of the population at time t (number of births per person-year lived).

2.1.2 Aging Measures Given by Liao (1996)

On the basis of the rate of change of mean age (3), Liao (1996) proposed the rate of change of the remaining four common measures of aging (4). For closed population, these are:

i) Rate of change in the proportion of persons age 60 and over at time t (P_{60}):

$$\frac{dP_{60}(t)}{dt} = \frac{N_{60}(t)}{N(t)}\{[d(t) - d_{60}(t)] + [a_{60}(t) - b(t)]\} \tag{2}$$

where, N(t): size of the total population at time t.

$d_{60}(t)$: corresponding death rate of the sub-population aged 60 and over.

$a_{60}(t)$: rate of new members added to sub-population aged 60 and over.

ii) Rate of change in the proportion of persons aged below 15 at time t (P_{15}):

$$\frac{dP_{15}(t)}{dt} = \frac{N_{15}(t)}{N(t)}\{[d(t) - a_{15}(t) - d_{15}(t)] + [b_{15}(t) - b(t)]\} \tag{3}$$

where, $a_{15}(t)$: members of the age group aging out of the group at time t.

$d_{15}(t)$: biological death rate in that age group (below 15) at time t.

$b_{15}(t)$: new birth added to the age group at time t.

iii) Rate of change in the aged-child ratio at time t:

$$\frac{dR(t)}{dt} = \frac{N_{60}(t)}{N_{15}(t)}\{[a_{15}(t) + d_{15}(t) - d_{60}(t)] + [a_{60}(t) - b_{15}(t)]\} \tag{4}$$

iv) Rate of change in the Median age at time t:

Liao (1996) conceptualized as the rate of change in the proportion aged median and below at time t (4), dPMd(t)/dt and the expression is

$$\frac{dP_{Md}(t)}{dt} = \frac{N_{Md}(t)}{N(t)}\{[d(t) - a_{Md}(t) - d_{Md}(t)] + [b_{Md}(t) - b(t)]\} \tag{5}$$

under usual notations.

2.2 Modified Measures

Demographic measure varies among different sub-populations (age, sex, region, education level, economic status, marital status, etc.). Nath and Deka (2006) suggested a few improved aging indices considering birth and death rates for male, female, and male-female simultaneously (5). Here an attempt has been made to expand these measures with respect to the region (urban and rural) and sex specific alternative birth rates, the birth rate under the assumption of stable population, [$b(t)^*$] and life expectancies [$e_x(t)$] i.e., death rate under the assumption of the stationary population (reciprocal to the life expectancy). The indices presented here are for a closed population.

2.2.1 Measures Related to Proportion of Aged 60 and Above (P_{60})

$$\text{i)}\ \frac{dP_{60}(t)}{dt} = \frac{N_{60}(t)}{N(t)}\left\{\left[\frac{1}{e_0(t)} - \frac{1}{e_{60}(t)}\right] + [a_{60}(t) - b(t)^*]\right\} \tag{6}$$

$$\text{ii)}\ \frac{dP_{60}(t)}{dt} = \frac{N_{60}^m(t)}{N(t)}\left[a_{60}^m(t) - \frac{1}{e_{60}^m(t)}\right] + \frac{N_{60}^f(t)}{N(t)}\left[a_{60}^f(t) - \frac{1}{e_{60}^f(t)}\right]$$
$$+ \frac{N_{60}(t)}{N(t)}\left\{\left[\frac{1}{e_0^m(t)} \cdot \frac{N^m(t)}{N(t)} + \frac{1}{e_0^f(t)} \cdot \frac{N^f(t)}{N(t)} - b^m(t)^* - b^f(t)^*\right]\right\} \tag{7}$$

This measure considers simultaneous effects of overall male-female birth and life expectancies.

$$\text{iii)}\ \frac{dP_{60}(t)}{dt} = \frac{N_{60}^u(t)}{N(t)}\left[a_{60}^u(t) - \frac{1}{e_{60}^u(t)}\right] + \frac{N_{60}^R(t)}{N(t)}\left[a_{60}^R(t) - \frac{1}{e_{60}^R(t)}\right]$$
$$+ \frac{N_{60}(t)}{N(t)}\left[\frac{1}{e_0^u(t)} \cdot \frac{N^u(t)}{N(t)} + \frac{1}{e_0^R(t)} \cdot \frac{N^R(t)}{N(t)} - b^u(t)^* . \frac{N^u(t)}{N(t)} - b^R(t)^* . \frac{N^R(t)}{N(t)}\right] \tag{8}$$

This measure considers simultaneous effects of urban-rural birth and life expectancies.

The superscripts u, R, m and f stand for urban, rural, male and female respectively.

2.2.2 Measures related to proportion of age below 15 (P_{15})

$$\text{i)}\ \frac{dP_{15}(t)}{dt} = \frac{N_{15}(t)}{N(t)}\left\{\left[\frac{1}{e_0(t)} - a_{15}(t) - d_{15}(t)\right] + \left[b_{15}(t)^* - b(t)^*\right]\right\} \tag{9}$$

$$\text{ii)}\ \frac{dP_{15}(t)}{dt} = \frac{N^m_{15}(t)}{N(t)}\left[b^m_{15}(t)^* - a^m_{15}(t) - d^m_{15}(t)\right] + \frac{N^f_{15}(t)}{N(t)}\left[b^f_{15}(t)^* - a^f_{15}(t) - d^f_{15}(t)\right]$$

$$+\frac{N_{15}(t)}{N(t)}\left\{\left[\frac{1}{e^m_0(t)}\frac{N^m(t)}{N(t)} + \frac{1}{e^f_0(t)}\frac{N^f(t)}{N(t)} - b^m(t)^* - b^f(t)^*\right]\right\} \tag{10}$$

This measure considers the simultaneous effects of overall male-female birth and life expectancies.

$$\text{iii)}\ \frac{dP_{15}(t)}{dt} = \frac{N^u_{15}(t)}{N(t)}\left[b^u_{15}(t)^* - a^u_{15}(t) - d^u_{15}(t)\right] + \frac{N^R_{15}(t)}{N(t)}\left[b^R_{15}(t)^* - a^R_{15}(t) - d^R_{15}(t)\right]$$

$$+\frac{N_{15}(t)}{N(t)}\left[\frac{1}{e^u_0(t)}\cdot\frac{N^u(t)}{N(t)} + \frac{1}{e^R_0(t)}\cdot\frac{N^R(t)}{N(t)} - b^u(t)^*.\frac{N^U(t)}{N(t)} - b^R(t)^*.\frac{N^R(t)}{N(t)}\right] \tag{11}$$

This measure considers simultaneous effects of urban-rural birth and life expectancies.

2.2.3 Measures Related to Aged-Child Ratio(R)

$$\text{i)}\ \frac{dR(t)}{dt} = \frac{N_{60}(t)}{N_{15}(t)}\left\{\left[a_{15}(t) + d_{15}(t) - \frac{1}{e_{60}(t)}\right] + \left[a_{60}(t) - b_{15}(t)^*\right]\right\} \tag{12}$$

$$\text{ii)}\ \frac{dR(t)}{dt} = \frac{N_{60}(t)}{N_{15}(t)}\cdot\frac{N^m_{15}(t)}{N_{15}(t)}\left[a^m_{15}(t) + d^m_{15}(t) - b^m_{15}(t)^*\right] + \frac{N_{60}(t)}{N_{15}(t)}\cdot\frac{N^f_{15}(t)}{N_{15}(t)}\left[a^f_{15}(t) + d^f_{15}(t) - b^f_{15}(t)^*\right]$$

$$+\frac{N^m_{60}(t)}{N_{15}(t)}\left[a^m_{60}(t) - \frac{1}{e^m_{60}(t)}\right] + \frac{N^f_{60}(t)}{N_{15}(t)}\left[a^f_{60}(t) - \frac{1}{e^f_{60}(t)}\right] \tag{13}$$

This measure considers the simultaneous effects of overall male-female birth and life expectancies.

$$\text{iii)}\ \frac{dR(t)}{dt} = \frac{N_{60}(t)}{N_{15}(t)}\left\{\left[a^u_{15}(t) + d^u_{15}(t) - b^u_{15}(t)^*\right]\frac{N^u_{15}(t)}{N_{15}(t)} + \left[a^R_{15}(t) + d^R_{15}(t) - b^R_{15}(t)^*\right]\frac{N^R_{15}(t)}{N_{15}(t)}\right\}$$

$$+\frac{N^u_{60}(t)}{N_{15}(t)}\left[a^u_{60}(t) - \frac{1}{e^u_{60}(t)}\right] + \frac{N^R_{60}(t)}{N_{15}(t)}\left[a^R_{60}(t) - \frac{1}{e^R_{60}(t)}\right] \tag{14}$$

This measure considers simultaneous effects of urban-rural birth and life expectancies.

2.2.4 Measures related to Median age (Md)

$$\text{i)}\ \frac{dP_{Md}(t)}{dt} = \frac{N_{Md}(t)}{N(t)}\left\{\left[\frac{1}{e_0(t)} - a_{Md}(t) - d_{Md}(t)\right] + \left[b_{Md}(t)^* - b(t)^*\right]\right\} \tag{15}$$

and

$$\frac{dMd(t)}{dt} = \frac{-(wdP_{Md}(t)/dt)}{P_{Md-w}} \tag{16}$$

$$\text{ii)}\ \frac{dP_{Md}(t)}{dt} = \frac{N^m_{Md}(t)}{N(t)}\left[b^m_{Md}(t)^* - a^m_{Md}(t) - d^m_{Md}(t)\right] + \frac{N^f_{Md}(t)}{N(t)}\left[b^f_{Md}(t)^* - a^f_{Md}(t) - d^f_{Md}(t)\right] + \frac{N_{Md}(t)}{N(t)}\left\{\left[\frac{1}{e^m_0(t)}\cdot\frac{N^m(t)}{N(t)} +\right.\right.$$

$$\left.\left.\frac{1}{e^f_0(t)}\cdot\frac{N^f(t)}{N(t)} - b^m(t)^* - b^f(t)^*\right]\right\} \tag{17}$$

This measure considers the simultaneous effects of overall male-female birth and life expectancies.

$$\text{iii)}\ \frac{dP_{Md}(t)}{dt} = \frac{N^u_{Md}(t)}{N(t)}\left[b^u_{Md}(t)^* - a^u_{Md}(t) - d^u_{Md}(t)\right] + \frac{N^R_{Md}(t)}{N(t)}\left[b^R_{Md}(t)^* - a^R_{Md}(t) - d^R_{Md}(t)\right]$$

$$+\frac{N_{Md}(t)}{N(t)}\left[\frac{1}{e^u_0(t)}\cdot\frac{N^u(t)}{N(t)} + \frac{1}{e^R_0(t)}\cdot\frac{N^R(t)}{N(t)} - b^u(t)^*.\frac{N^U(t)}{N(t)} - b^R(t)^*.\frac{N^R(t)}{N(t)}\right] \tag{18}$$

This measure considers the simultaneous effects of urban-rural birth and life expectancies.

2.2.5 Measures Related to Mean age (A_p)

$$\text{i) } \frac{\mathrm{dA}p(t)}{\mathrm{dt}} = 1 - \frac{1}{e_0(t)}\left[A_D(t) - A_p(t)\right] - b(t)^* \times A_p(t) \tag{19}$$

$$\text{ii)} \frac{\mathrm{dA}p(t)}{\mathrm{dt}} = 1 - A_p(t)\left[b^m(t)^* + b^f(t)^*\right] + \frac{1}{e_0^m(t)}\left[A_p(t) - A_D^m(t)\frac{N^m(t)}{N(t)}\right] + \frac{1}{e_0^f(t)}\left[A_p(t) - A_D^f(t)\frac{N^f(t)}{N(t)}\right] \tag{20}$$

This measure considers the simultaneous effects of overall male-female birth and life expectancies.

iii)

$$\frac{\mathrm{dA}p(t)}{\mathrm{dt}} = 1 - A(t)[b^u(t)^* + b^R(t)^*] + \frac{1}{e_0^u(t)}\left[A_p(t) - A_D^u(t).\frac{N^u(t)}{N(t)}\right] + \frac{1}{e_0^R(t)}\left[A_p(t) - A_D^R(t).\frac{N^R(t)}{N(t)}\right] \tag{21}$$

This measure considers the simultaneous effects of urban-rural birth and life expectancies.

2.3 Calculation of alternative Birth Rates

Alternative birth rates b(t)* have been calculated from its functional relationship with NRR, r, and T. The relation is stated below:

A fundamental equation of population dynamics is

$$r(t) = b(t) - d(t) \tag{22}$$

In which r(t) is the "instantaneous" per capita growth rate, b(t) is the per capita birth rate, and d (t) is the per capita death rate at time t.

Bertran and Murray (1997) proposed the following relationship (8):

$$e^{r(t)} = 1 + b(t) - d(t) \tag{23}$$

Here we use,

$$d(t) = \frac{1}{e_0(t)} \tag{24}$$

which is known as the life table death rate.

Now we need to find out the value of r(t), intrinsic rate at time t. For this, we use the following approximation (9):

$$\mathrm{NRR} = e^{\mathrm{rT}} \tag{25}$$

Where NRR is the net reproduction rate, r is the intrinsic growth rate and T is the mean length of generation.

Again, T is very close to the average age of childbearing of the female population for the stationary population (10). Here T is assumed as average age of childbearing of female population and this is calculated from age-specific fertility rate of Bangladesh female population. Therefore,

$$e^{r(t)} = \frac{\mathrm{NRR}}{T} \tag{26}$$

Therefore, the alternative birth rate b(t)* has been calculated from equation (23) after putting the value of $e^{r(t)}$ from equation (26) and d(t) from equation (24).

After calculating b(t)* for the nation, urban, and rural areas we have computed the sex specific birth rate adjusting with sex ratio at birth. Similarly age specific birth rate, for example, birth rate for persons under fifteen and persons aged at or below the median, which is not actually a demographic birth rate but used for calculating the rate of change of aging measures.

3. Materials of the Study

The measures discussed above have been applied to the Bangladeshi population for the census years 1981 and 2001 (11). Various demographic rates have also been collected from other Bangladesh Bureau of Statistics (2003) publication (12). Sex specific birth rates have been calculated using the sex ratio at birth. Age specific death rates and the average age at death have been computed from the distribution of age specific death rates by age, sex, and locality respectively for the years 1981 and 2001. Normally various rates are expressed in terms of per 100 or per 1000 populations. However, for ease of calculation, these rates are presented per unit population in our study.

4. Results and Discussions

This study intends to measure the population aging as a function of alternative fertility and mortality measures given by Preston et al. (1989) and Liao (1997), of a person (3,4). The rate of change of conventional aging measures alternative to those of existing measures has been exercised to observe the greying process of Bangladeshi population for two census years 1981 and 2001. A comparison has been made among different formulae used here. We try to find out the best method of measuring the rate of change in the aging process with respect to demographic components. Here, all the measures have been calculated for the closed population of Bangladesh. Various alternative birth and other rates and life expectancies for 1981 and 2001 are presented in Table 1.

Table 1. Alternative birth, death, and other rates of Bangladesh, 1981- 2001

Year	Locality	Sex	Life expectancy, alternative birth, and other rates				
			$b(t)$	e_0	e_{60}	$b_{Md}(t)$	$b_{15}(t)$
1981	National	Male	0.0210	55.30	16.20	0.0816	0.0879
		Female	0.0206	54.50	16.00	0.0852	0.0910
		Person	0.0417	54.80	16.10	0.0833	0.0894
	Urban	Person	0.0268	60.30	17.80	0.0535	0.0641
	Rural	Person	0.0432	54.30	16.00	0.0865	0.0911
2001	National	Male	0.0121	64.00	16.14	0.0470	0.0487
		Female	0.0110	64.50	16.72	0.0456	0.0591
		Person	0.0232	64.20	16.44	0.0463	0.0589
	Urban	Person	0.0065	66.40	18.07	0.0130	0.0190
	Rural	Person	0.0281	63.20	16.97	0.0562	0.0687

4.1 Aging Through Alternative Demographic Components

In this study, the rate of change of various conventional aging measures has been computed with stable population birth and life table death rate of a person. These measures indicate per year change i.e., 1980-1981 and 2000-2001. The basic considerations of these formulae are:

(a) First: Stable population birth and life table death rates of population.

(b) Second: Stable population birth and life table death rates for both male and female populations.

(c) Third: Stable population birth and life table death rates for both urban and rural populations.

4.2 Aging Process of Bangladeshi Population

Various alternative measures of the rate of change in the aging process of the Bangladeshi population for 1981 and 2001 have been presented in Table 2.

Table 2. National level alternative rate of change aging measures, 1981 - 2001

Measures	Year	Person-1*	Person-2**	Person-3***
$\frac{dp_{60}(t)}{dt}$	1981	- 0.00118	- 0.00105	- 0.00246
	2001	- 0.00042	- 0.00039	- 0.00046
$\frac{dp_{15}(t)}{dt}$	1981	- 0.02021	- 0.00988	- 0.01737
	2001	- 0.01604	- 0.01977	- 0.02916
$\frac{dR(t)}{dt}$	1981	0.00270	0.00265	0.00292
	2001	0.00529	0.00520	0.00584
$\frac{dp_{Md}(t)}{dt}$	1981	0.00132	- 0.00705	- 0.01077
	2001	- 0.00835	- 0.01059	- 0.00992
$\frac{dMd(t)}{dt}$	1981	- 0.05000	0.26000	0.40000
	2001	0.43000	0.55000	0.51000
$\frac{dA_p(t)}{dt}$	1981	- 0.50000	- 0.09000	- 0.74000
	2001	- 0.15000	0.23000	- 0.05000

*Measures considering overall alternative birth rates and life expectancies.

**Measures considering male-female alternative birth rates and life expectancies.

***Measures considering urban-rural alternative birth rates and life expectancies.

The yearly change of proportion of persons aged 60 or more (P_{60}) according to the first, second, and third formulae were -0.00118, -0.00105, and -0.00111 respectively in 1981. The corresponding figures were -0.00042, -0.00039, and -0.00046 respectively for the year 2001. All the measures show a declining peak aging process having a very slow pace in 1981 and in 2001 it is almost zero. Therefore, it can be said that the greying process of Bangladeshi population is more or less stable in 2001 according to the peak aging measures with alternative approaches.

The rates of change of base aging in 1981 were -0.02021, -0.00988, and -0.0173 according to the four formulae under consideration. In 2001, the corresponding figures were -0.01604, -0.01977, and -0.00292 respectively. All the measures indicate a decreasing young population for both 1981 and 2001 with different magnitudes.

The values of aged-child ratio (R) were 0.0027, 0.00265, and 0.00292 corresponding to three formulae in 1981. In 2001, the corresponding values were 0.00529, 0.00520, and 0.00584 respectively. All the formulae under consideration indicate more or less similar aging processes at both 1981 and 2001 where the aging process was faster in 2001 than that of 1981.

Median age, a combined measure of aging, is popular for its non-parametric nature. Most of the measures show an increasing aging process in both 1981 and 2001 though the difference in magnitude among the various formulae of this process is mentionable. The increasing rate was significantly higher in 2001 than in 1981. According to the rate of change of the median age of the population (both existing and alternative), it is evident that the Bangladeshi population is becoming aged over time.

The rate of change of mean age (A_p) with alternative measures gives somewhat confusing results because of the skewed distribution of population age. According to this measure, the aging process was faster in 2001 than that in 1981.

With the above discussions on the alternative rate of change of conventional aging measures, we observe that all of them indicate an upward aging process of Bangladeshi population except for those on the peak aging as the population of Bangladesh is not mature enough yet.

4.3 Gender Differences

The rate of change of conventional aging measures at the national level with respect to sex for 1981 and 2001 has been presented in Table 3.

Table 3. National level alternative measures with respect to sex, 1981 - 2001

Measures	1981		2001	
	Male	Female	Male	Female
$\frac{dp_{60}(t)}{dt}$	- 0.00115	- 0.00121	- 0.00054	- 0.00032
$\frac{dp_{15}(t)}{dt}$	- 0.02179	- 0.01810	- 0.01682	- 0.01488
$\frac{dR(t)}{dt}$	0.00371	0.00164	0.00556	0.00483
$\frac{dp_{Md}(t)}{dt}$	- 0.00867	- 0.00768	- 0.00588	- 0.01097
$\frac{dMd(t)}{dt}$	0.36000	0.30000	0.30000	0.58000
$\frac{dA_p(t)}{dt}$	- 0.04000	- 0.03000	0.13000	0.14000

It is observed that the aging process of Bangladesh with respect to sex is more or less stable in terms of the speed of peak aging measures with the alternative approaches. A clear base aging process with respect to sex in Bangladeshi population has been observed because of the successful family planning program of this country where male aging is faster than female aging.

An alternative measure of velocity in the aged-child ratio with respect to sex also indicates an increasing aging process. This process is faster in 2001 than in 1981.

The speed of changing median age with the alternative measure indicates an increasing aging process in Bangladesh with respect to sex. Female aging was faster (almost double) than that of males in 2001 though the picture was the opposite in 1981.

In 2001, we observed a positive aging process for the male and female population of Bangladesh with the modified measures of rate of change of mean age. The gender gap is not visible with these measures. From the rate of change of average age (median and mean), it can be said that both the male and female populations of Bangladesh are greying over time.

4.4 Urban-Rural Differences

Alternative rate of change of conventional aging measures for urban and rural populations have been presented in Table 4.

Table 4. Urban and rural level alternative measures, 1981 - 2001

Measures	Location	1981		2001	
		Person-1*	Person-2**	Person-1*	Person-2**
$\frac{dp_{60}(t)}{dt}$	Urban	- 0.00016	0.00013	0.00103	0.00161
	Rural	- 0.00130	- 0.00117	- 0.00073	- 0.00068
$\frac{dp_{15}(t)}{dt}$	Urban	- 0.02069	- 0.02610	- 0.02226	- 0.04536
	Rural	- 0.02056	- 0.00913	- 0.01493	- 0.01352
$\frac{dR(t)}{dt}$	Urban	0.00543	0.00544	0.01212	0.01261
	Rural	0.00251	0.00252	0.00406	0.00403
$\frac{dp_{Md}(t)}{dt}$	Urban	- 0.01353	- 0.01515	- 0.01882	- 0.02478
	Rural	- 0.00766	- 0.00533	- 0.00776	- 0.00885
$\frac{dMd(t)}{dt}$	Urban	0.66000	0.73000	0.84000	1.11000
	Rural	0.29000	0.20000	0.42000	0.48000
$\frac{dA^{p}(t)}{dt}$	Urban	- 0.05000	0.33000	0.27000	0.63000
	Rural	- 0.54000	0.13000	- 0.26000	0.12000

*Measures considering overall alternative birth rates and life expectancies.

**Measures considering rural and urban male-female alternative birth rates and life expectancies.

A notable urban-rural gap has been observed in both 1981 and 2001. The urban old-age aging process was almost stable in 1981 whereas that of rural was somewhat decreasing. The trend of increasing peak aging process was observed in the period 1981-2001 for both urban and rural areas where the rural aging was almost stable and the urban aging was slightly up-warding.

A mentionable urban-rural gap was found in the base aging process at 2001 according to the alternative measures where the urban aging was faster than that of rural though the gap between these two was negligible at 1981. The decreasing rate in the proportion of the urban young female population was faster than that of rural in 1981 but an opposite direction was found in the urban-rural male population. Both the urban young male and female populations decreasing rate was faster than that of rural in 2001. There is a huge gap between the results of the base aging in urban and rural areas with the alternative approaches in this analysis. This finding is an outcome of successful family planning programs in urban areas.

An alternative measure of change in the speed of the aged-child ratio indicates a wide urban-rural gap in the aging process for both in 1981 and in 2001 where the urban aging was remarkably faster than that of rural.

The per year increase in the median and mean age with the alternative measures show a wide urban-rural gap in the aging process in both 1981 and 2001 where a faster urban aging process than that of rural is mentionable.

4.5 Comparison Between Alternative and Existing Measures

A close look at the measures of Liao and Nath-Deka with their corresponding measures indicates the same direction of changing the aging process having different magnitudes (Table 5). Though the speed of the peak aging process is almost stable for both the existing and alternative Liao and Nath-Deka measures, alternative measures show a slightly faster aging process. Again, both the alternative approaches of Liao and Nath-Deka indicate a slow base aging process though the aging at base is increasing over the year for all the measures.

Table 5. Existing and alternative measures of speed of aging process

Measures	Year	Liao	Alternative to Liao	Nath and Deka	Alternative to Nath and Deka
$\frac{dp_{60}(t)}{dt}$	1981	- 0.00302	- 0.00118	- 0.00458	- 0.00105
	2001	- 0.00077	- 0.00042	- 0.00058	- 0.00039
$\frac{dp_{15}(t)}{dt}$	1981	- 0.02713	- 0.02021	- 0.02695	- 0.00988
	2001	- 0.02271	- 0.01604	- 0.03239	- 0.01977
$\frac{dR(t)}{dt}$	1981	0.00056	0.00270	0.00064	0.00265
	2001	0.00704	0.00529	0.00717	0.00520
$\frac{dp_{Md}(t)}{dt}$	1981	- 0.00558	0.00132	- 0.01224	- 0.00705
	2001	- 0.01587	- 0.00835	- 0.01348	- 0.01059
$\frac{dMd(t)}{dt}$	1981	0.21000	- 0.05000	0.46000	0.26000
	2001	0.82000	0.43000	0.70000	0.55000
$\frac{dA^p(t)}{dt}$	1981	- 0.13000	- 0.50000	0.13000	- 0.09000
	2001	0.36000	- 0.15000	0.08000	0.23000

More speed of increasing aging process of Bangladeshi population has been found with alternative to Liao and Nath-Deka in 1981 in terms of the measure of the aged-child ratio. But a measure in 2001 reflects the opposite picture. Again, the alternative measures represent a slow pace of increasing median age in both 1981 and 2001. The measures regarding the rate of change of mean age of the population are confusing with all the measures (existing and alternative ones) because of skewed age distribution. Only the alternative measure of mean age to Nath-Deka gives a somewhat accepting result compared to the rate of change of median age for the year 2001.

An empirical study finds that the yearly linear rate of change of aging at base, aged-child ratio and median age over the years 1981-2001 are -0.0036, 0.0017, and 0.22 respectively. The alternative measures have a narrow gap with these yearly rates of change compared to those measured with Liao and Nath-Deka's approaches. So, we can claim that these are good alternative measures. Therefore, the use of both the measures, existing and alternative, will be helpful for a better understanding of the aging process.

4.6 Comparison Among Alternative Indices

We have proposed several alternative approaches to conventional aging measures in this chapter. These are the alternative index to Liao, the alternative index to Preston and Himes et al., the alternative index to Nath and Deka, alternative modified-1 index (this modification is based on the consideration discussed in 'c'). The results are presented in Table 6.

Table 6. Various alternative measures of Bangladesh population, 1981 - 2001

Measures	Year	Liao	Nath and Deka	Modified-1*
$\frac{dp_{60}(t)}{dt}$	1981	- 0.00118	- 0.00105	- 0.00246
	2001	- 0.00042	- 0.00039	- 0.00046
$\frac{dp_{15}(t)}{dt}$	1981	- 0.02021	- 0.00988	- 0.01737
	2001	- 0.01604	- 0.01977	- 0.02916
$\frac{dR(t)}{dt}$	1981	0.00270	0.00265	0.00292
	2001	0.00529	0.00520	0.00584
$\frac{dp_{Md}(t)}{dt}$	1981	0.00132	- 0.00705	- 0.01077
	2001	- 0.00835	- 0.01059	- 0.00992
$\frac{dMd(t)}{dt}$	1981	- 0.05000	0.26000	0.40000
	2001	0.43000	0.55000	0.51000
$\frac{dA^p(t)}{dt}$	1981	- 0.50000	- 0.09000	- 0.74000
	2001	- 0.15000	0.23000	- 0.05000

*Measures considering urban-rural alternative birth rates and life expectancies.

The entire alternative measures exhibit a slightly decreasing peak aging process in both 1981 and 2001 where the decreasing rate is almost negligible in the later year. The magnitudes of this process are more or less the same in alternative measures except for the alternative modified-1 index in 1981.

Alternative measures of base aging show a different aging process among the formulae in both 1981 and 2001 though the direction (decreasing proportion of young) is the same for all the approaches. In 1981, the alternative to Nath and Deka's formula showed a very slow decreasing rate while the alternative to Liao's measure is close to each other and these slightly differ from the alternative modified-1 measure. The variations are poor among different alternative measures except for modified-1 in 2001.

The speed of aging is faster with alternative modified-1 measures than those of Liao and Nath-Deka's measure in both 1981 and 2001 based on the rate of change of the aged-child ratio. All alternative measures based on the median age indicate an upward aging process in Bangladesh in both 1981 and 2001 except for the alternative to Liao's method in 1981. The alternative modified-1 measure shows a faster aging process than those of alternative to Liao and alternative to Nath-Deka's measures in 1981.

All other alternative measures of the rate of change of mean age (Preston et al., Nath-Deka, and modified-1) show a confusing aging process while alternative to Nath-Deka's approach shows an acceptable aging process in 2001. From the results, it is clear that all the alternative measures show the same direction of Bangladeshi population aging except for the measures regarding the mean age. The modified-1 measure captures more clear pictures than other measures.

5. Conclusions

Rigorous and robust measures are indispensable for population aging research. Conventional aging measures are not consistent and they suffer from their limitations. This study is an effort to extend some measures of the speed of aging process and to assess the Bangladeshi aging process using them. Bangladesh is not an exception to the global phenomenon of demographic aging. It is a relatively new issue in the country as its demographic transition has entered into the third stage recently. Although it has not reached at an alarming stage yet, there should be no room for complacency. The country is now experiencing declining fertility and gradually improving mortality rates, especially in infant and maternal levels. The demographic transition represents a dynamic character of the country's population age structure. From the analysis, it is found that these extended measures are good alternatives to existing measures of aging velocity and alternative approaches show the somewhat different pace of changing aging process. Our proposed modified-1 measure can be considered as the best one as it gives consistent results on the speed of aging with the majority of the conventional aging measures. Since the level of changing aging process is different among alternative measures hence it is recommended to consult all the measures carefully and to take a decision on the basis of the combined idea of these measures. So, the alternative measures will be helpful for a better understanding of the demographic aging process. This work also reveals that the Bangladeshi population is greying over time with a noticeable gender and urban-rural gap. The urban population is aging at faster pace than those of rural. The male demographic aging is faster than that of females in both 1981 and 2001. Therefore, this type of study on the population aging process of a nation will add some weight on taking sustainable aging policy.

Conflict of Interest

The authors declare no conflict of interest.

References

Bangladesh Bureau of Statistics. (2003). *National Report (provisional) of Population Census 2001*. Ministry of Planning, Dhaka, Bangladesh.

Bangladesh Bureau of Statistics. (2004). *Report on Bangladesh Sample Vital Statistics 2002*. Ministry of Planning, Dhaka, Bangladesh.

Basu, A., & Basu, K. (1987). The greying of populations: Concepts and measurement. *Demography India, 16*(1), 79-89.

Bertram, G., & Murray Jr, (1997). On calculating birth and death rates. *Oikos, 78*(2), 384-387.

Coale, A. J. (1955). The calculation of approximate intrinsic rates. *Population Index, 21*(2), 94-97.

Islam, M. N., & Nath, D. C. (2010). Application of demographic components for measuring the aging velocity: An explanation with Bangladesh context. *Demography India, 39*(2), 297-313.

Keyfitz, N. (1971). On the momentum of population growth. *Demography, 8*(1), 71-80.

Liao, T. F. (1996). Measuring population aging as a function of fertility, mortality, and migration. *Journal of Cross-Cultural Gerontology, 11*(1), 61-79.

Nath, D. C., & Deka, A. K. (2006). A few improved aging indices: An application to elderly population in Assam. *Assam Statistical Review, 20*(1), 11-40.

Nath, D. C., & Islam, M. N. (2009). New indices: An application of measuring the aging process of some Asian countries with special reference to Bangladesh. *Journal of Population Ageing, 2*(1-2), 23-39).

Nations U. Population Division. (2001). *World Population Ageing: 1950-2050*. New York, NY: Population (English Edition). (For books or reports, include the full title and name of the publisher.)

Preston, S. H., Himes, C., & Eggers, M. (1989). Demographic conditions responsible for population aging. *Demography, 26*, 691-704.

Handling Non-Response in Presence of p(p≥2) Auxiliary Variables in Two Occasion Rotation Patterns

Abstract

The successive sampling is a well-known technique that can be used in longitudinal surveys to estimate population parameters. The present work is an attempt to develop imputation methods to reduce the impact of non-response at both the occasions in two-occasion successive (rotation) sampling. Utilizing the information on p (p ≥ 2) auxiliary variates, which are available at both occasions, estimators have been proposed for estimating the population mean at the current occasion. Behaviour of the proposed estimators is studied and the optimum replacement strategy is discussed in detail. To study the effectiveness of the suggested imputation methods are examined by comparing the performances of the proposed estimators in two different situations: with and without non-response. Empirical studies validate the results thus obtained.

Keywords: Non-response, imputation, successive sampling, auxiliary character, chain-type, optimum replacement policy.

1. Introduction

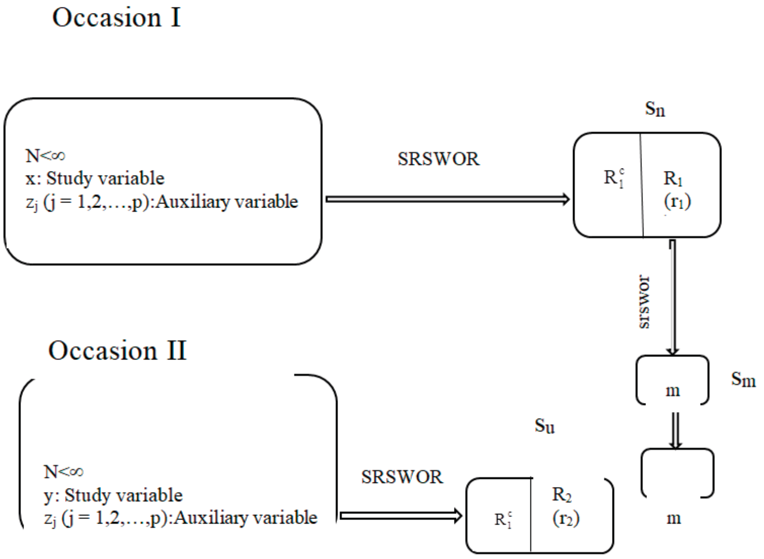
By repeatedly sampling the same population over different periods and measuring the same study variables, surveys can track development over time. This approach, known as successive (or rotation) sampling, involves sampling across successive occasions—such as different years, seasons, or months—following a specific pattern and allowing for partial replacement of units. Successive (rotation) sampling is a robust method for obtaining reliable estimates at various time points. Key contributions to the theory of estimating population mean through successive (rotation) sampling have been made by Jessen [1], Sen [2-4], and Singh and Singh [5]. Furthermore, in many cases, additional information from an auxiliary variate is available at both the first and second occasions. Several estimators that utilize this auxiliary information to estimate the population mean at the current occasion have been proposed by researchers [6-12].

Non-response is a prevalent issue faced by survey researchers, and it tends to be more problematic in repeated surveys compared to single-occasion surveys. The patterns and causes of non-response or missing data can vary significantly. One effective method for addressing missing data is imputation, which involves filling in the missing values with reasonable estimates. This technique can simplify the analysis of incomplete datasets. Kalton et al. [13] have recommended various imputation methods designed to render incomplete data sets structurally complete and easier to analyse. Additionally, imputation can be enhanced using auxiliary variates [14, 15]. Ahmed et al. [16] and Singh [17] ave introduced several novel imputation techniques that leverage auxiliary variates for this purpose.

The aim of the current study is to explore the impact of non-response across two occasions in successive (rotation) sampling. Singh and Karna [18] developed estimators for calculating the population mean at the current occasion while considering non-response at both occasions using imputation methods. This study proposes new imputation techniques that utilize information from p (where p≥2) auxiliary variates available at both occasions to address the non-response issue in two-occasion successive (rotation) sampling. The performance of these new estimators is evaluated under two scenarios: one with non-response and one without, and appropriate recommendations are provided

2. Sample Designs on Two-Occasion

Consider a finite population U = (U_1, U_2, - - -, U_N) that has been sampled on two separate occasions. The variable of interest is denoted as x on the first occasion and y on the second occasion. It is assumed that there is a significant time gap between the two occasions, and at both times, information on p (a non-negative integer constant ≥ 2) auxiliary variables z_j (j = 1, 2, - - -, p) is available with known population means. The presence of non-response at both occasions is also assumed. The pictorial representation of sampling design is given below:



λ (=m/n) and μ (=u/n), given that $\lambda + \mu = 1$, are the fractions of the matched and fresh sample, respectively, at the current occasion. For every unit $i \in R_k (k = 1, 2)$ the value x_i (y_i) is observed, but for the units $i \in R_k^c$ (k = 1, 2) the values x_i (y_i) are missing and instead, imputed values are derived. The mean imputation method is employed for the matched sample, while the imputation method for the unmatched sample utilizes p auxiliary variables z_j available only during the current occasion. The following notations are defined for use in the subsequent sections:

$\bar{X}, \bar{Y}, \bar{Z}_j$: Population means of the variates x, y and z_j (j = 1, 2, - - p) respectively.

$\bar{x}_n, \bar{y}_m, \bar{x}_m, \bar{z}_{mj}, \bar{y}_u$: Sample means of the respective variates

$\bar{x}_{r_1}, \bar{y}_{r_2}, \bar{z}_{r_2 j}$: Mean of responses of the respective variates

$\rho_{yx}, \rho_{xz_j}, \rho_{yz_j}$: Correlation coefficients between the variates.

$b_{yx}, b_{yz_j}(m), b_{yz_j}(r_2)$: Sample regression coefficient based on the sample sizes given in braces.

$S_x^2 = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \bar{X})^2$: The population mean square of the variate x.

$S_y^2, S_{z_j}^2$: Population mean squares of the variates y and z_j respectively.

$f\left(= \frac{n}{N}\right)$: The sampling fraction.

$f_1\left(= \frac{r_1}{n}\right), \; f_2\left(= \frac{r_2}{u}\right)$: Portions of respondents in the sample of sizes n and u respectively.

$t_1 (= 1 - f_1)$, $t_2 (= 1 - f_2)$: Portions of non-respondents in the sample of sizes n and u respectively.

3. Design of Estimator

Consider two estimation designs for the population mean at the current occasion, employing inputs from p auxiliary variables z_j (j = 1, 2, - - -, p). These estimators are developed to address non-response issues encountered at two sampling occasions. The initial estimator relies on a fresh sample, s_u, collected during the second occasion. For this estimator, unavailable data points are imputed using a specific method proposed for handling the missing data at the current occasion. The imputation technique recommended for addressing missing data on the second occasion is described as follows:

$$y_{眏} = \begin{cases} y_i & \text{if } i \in R_2 \\ \bar{y}_{r_2} + \sum\limits_{j=1}^{p} b_{yz_j}\left(r_2\right)\left[\dfrac{u}{\left(u - r_2\right)}\left(\bar{Z}_j - \bar{z}_{r_2 j}\right) + z_{ij} - z_{r_2 j}\right] & \text{if } i \in R_2^c \end{cases} \tag{1}$$

Following the above imputation procedures, the estimator for $\bar{Y}$ based on fresh sample s_u at the current (second) occasion is given by

$$T_1 = \frac{1}{u}\sum_{i \in s_u} y_{眏} = \frac{1}{u}\left[\sum_{i \in R_2} y_{眏} + \sum_{i \in R_2^c} y_{眏}\right] = \bar{y}_{r_2} + \sum_{j=1}^{p} b_{yz_j}\left(r_2\right)\left(\bar{Z}_j - \bar{z}_{r_2 j}\right) \tag{2}$$

where $\bar{y}_{r_2} = \dfrac{1}{r_2}\sum\limits_{i \in R_2} y_i$.

The second estimator is based on a sample s_m that is observed at each occasion and makes use of insights from the first occasion. Given that there are incomplete responses at the first occasion, the unrecorded data are imputed using the mean imputation method.

As a result, the data, after imputation, is structured as follows:

$$x_{眏} = \begin{cases} x_i & \text{if } i \in R_1 \\ \bar{x}_{r_1} + \sum\limits_{j=1}^{p} b_{xz_j}\left(r_1\right)\left[\dfrac{n}{\left(n - r_1\right)}\left(\bar{Z}_j - \bar{z}_{r_1 j}\right) + z_{ij} - z_{r_1 j}\right] & \text{if } i \in R_1^c \end{cases} \tag{3}$$

By employing the imputation procedures presented, the estimator for a sample s_n obtained during the first occasion is expressed as

$$\bar{x}_n^* = \frac{1}{n}\sum_{i \in s_n} x_{眏} = \frac{1}{n}\left[\sum_{i \in R_1} x_{眏} + \sum_{i \in R_1^c} x_{眏}\right] = \bar{x}_{r_1} + \sum_{j=1}^{p} b_{xz_j}\left(r_1\right)\left(\bar{Z}_j - \bar{z}_{r_1 j}\right) \tag{4}$$

where $\bar{x}_{r_1} = \dfrac{1}{r_1}\sum\limits_{i \in R_1} x_i$.

Therefore, the estimator of $\bar{Y}$ based on a sample s_m ($m = n\lambda$), which is common to each occasion and utilizes the above discussed imputation methods for imputing the unrecorded data at the first occasion are defined as:

$$T_2 = \frac{\bar{y}_m^*}{\bar{x}_m^*}\ \bar{x}_n^* \tag{5}$$

where $\bar{y}_m^* = \bar{y}_m + \sum\limits_{j=1}^{p} b_{yz_j}(m)\left(\bar{Z} - \bar{z}_{jm}\right)$, $\bar{x}_m^* = \bar{x}_m + \sum\limits_{j=1}^{p} b_{yz_j}(m)\left(\bar{Z} - \bar{z}_{jm}\right)$

Considering the convex linear combination of estimators T_1 and T_2; the final estimator T for estimating the population mean $\bar{Y}$ at the current (second) occasion is defined as:

$$T = \varphi T_1 + \left(1 - \varphi\right) T_2 \tag{6}$$

where φ is an unknown real constant to be determined by the minimization of the mean square error of the estimator T.

4. Behaviour of the Proposed Estimator

The linear regression nature of both T1 and T2 implies that they are biased for $\bar{Y}$, and hence, the estimator T, specified in equation (6) is also subject to the bias for $\bar{Y}$. The bias B (.) and mean square error MSE (.) up-to the first order of approximations and for large population (ignoring fpc) are derived under large sample approximations using the following transformations:

$$\bar{y}_{r_2} = \bar{Y}\left(1 + e_1\right) \quad , \quad \bar{y}_m = \bar{Y}\left(1 + e_2\right) \quad , \quad \bar{x}_{r_1} = \bar{X}\left(1 + e_3\right) \quad , \quad \bar{x}_m = \bar{X}\left(1 + e_4\right) \quad , \quad \bar{z}_{r_1 j} = \bar{Z}_j\left(1 + e_{5j}\right) \quad ,$$

$$\bar{z}_{r_2 j} = \bar{Z}_j\left(1 + e_{6j}\right) \quad , \quad \bar{z}_{mj} = \bar{Z}_j\left(1 + e_{7j}\right) \quad , \quad s_{yz_j}\left(r_2\right) = S_{yz_j}\left(1 + e_{8j}\right) \quad , \quad s_{yz_j}\left(m\right) = S_{yz_j}\left(1 + e_{9j}\right) \quad ,$$

$$s_{xz_j}\left(r_1\right) = S_{xz_j}\left(1 + e_{10j}\right) \quad , \quad s_{xz_j}\left(m\right) = S_{xz_j}\left(1 + e_{11j}\right) \quad , \quad s^2_{z_j}\left(r_2\right) = S^2_{z_j}\left(1 + e_{12j}\right) \quad , \quad s^2_{z_j}\left(r_1\right) = S^2_{z_j}\left(1 + e_{13j}\right) \quad ,$$

$$s^2_{z_j}\left(m\right) = S^2_{z_j}\left(1 + e_{14j}\right).$$

With the transformations applied, Tu and Tm assume the following forms:

$$T_1 = \left[\bar{Y}\left(1 + e_1\right) - \sum_{j=1}^{p} \beta_{yzj}\, \bar{Z}_j\, e_{6j}\left(1 + e_{8j}\right)\left(1 + e_{12j}\right)^{-1}\right] \tag{7}$$

$$T_2 = \bar{Y}\left(1 + e_2\right) + \beta_{yx}\bar{X}\left(1 + e_5\right)\left(1 + e_6\right)^{-1}\left(e_3 - e_4\right) - \sum_{j=1}^{p} \beta_{yzj}\, \bar{Z}_j\, e_{10j}\left(1 + e^*_{7j}\right)\left(1 + e^*_{8j}\right)^{-1} \tag{8}$$

We can thus formulate the following theorems from the discussion above:

Theorem 4.1: The Bias(T) approximated up to the order O(n^{-1}) in the estimation of $\bar{Y}$, can be expressed as

$$B\left(T\right) = \varphi\, B\left(T_1\right) + \left(1 - \varphi\right) B\left(T_2\right) \tag{9}$$

$$\text{where} \quad B\left(T_1\right) = -\left(\frac{1}{r_2}\right)\sum_{j=1}^{p}\beta_{yzj}\left[\frac{C_{012}}{S_{yzj}} - \frac{C_{003}}{S^2_{zj}}\right] \tag{10}$$

$$\text{and} \quad B\left(T^*_2\right) = -\frac{1}{m}\sum_{j=1}^{p}\beta_{yzj}\left[\frac{C_{012}}{S_{yzj}} - \frac{C_{003}}{S^2_{zj}}\right] + \left(\frac{1}{m} - \frac{1}{r_1}\right)\left[\beta_{yx}\left\{\frac{C_{300}}{S^2_x} - \frac{C_{210}}{S_{yx}}\right\}\right.$$

$$\left. + \beta_{yx}\sum_{j=1}^{p}\beta_{xzj}\left\{\frac{C_{111}}{S_{yx}} - \frac{C_{201}}{S^2_x} - \frac{C_{003}}{S^2_{zj}} + \frac{C_{102}}{S_{xzj}}\right\}\right] \tag{11}$$

where $C_{rst} = E\left[\left(x_i - \bar{X}\right)^r\left(y_i - \bar{Y}\right)^s\left(z_{ij} - \bar{Z}_j\right)^t\right]$; $r \geq 0$, $s \geq 0$, $t \geq 0$, $j = 1, 2, ---, p$.

Proof: $B\left(T\right) = E\left[T - \bar{Y}\right] \; = \; \varphi\, B\left(T_1\right) + \left(1 - \varphi\right) B\left(T_2\right)$

where $B\left(T_1\right) = E\left[T_1 - \bar{Y}\right]$

$$= E\left[\bar{Y}\left(1 + e_1\right) - \sum_{j=1}^{p}\beta_{yzj}\, \bar{Z}_j\, e_{9j}\left(1 + e_{7j}\right)\left(1 + e_{8j}\right)^{-1} - \bar{Y}\right]$$

Assuming $\left|e_{8j}\right| < 1$, expanding the right-hand side of the above expression binomially, taking expectations, and collecting the terms up-to the order o(n^{-1}), we have

$$B\left(T_1\right) = -\left(\frac{1}{r_2} - \frac{1}{N}\right)\sum_{j=1}^{p}\beta_{yzj}\left[\frac{C_{012}}{S_{yzj}} - \frac{C_{003}}{S_{zj}^2}\right] \tag{12}$$

similarly

$$B\left(T_2\right) = E\left[T_2 - \bar{Y}\right]$$

$$= E\left[\bar{Y}\left(1+e_2\right) + \beta_{yx}\bar{X}\left(1+e_5\right)\left(1+e_6\right)^{-1}\left(e_3 - e_4\right) - \sum_{j=1}^{p}\beta_{yzj}\,\bar{Z}_j\; e_{10j}\left(1+e_{7j}^{*}\right)\left(1+e_{8j}^{*}\right)^{-1} - \bar{Y}\right]$$

Assuming $\left|e_6\right| < 1$ and $\left|e_{8j}\right| < 1$, expanding the right hand side of the above expression binomially, taking expectations and retaining the terms up-to the first order of approximations, we have

$$B\left(T_2\right) = -\left(\frac{1}{m} - \frac{1}{N}\right)\sum_{j=1}^{p}\beta_{yzj}\left[\frac{C_{012}}{S_{yzj}} - \frac{C_{003}}{S_{zj}^2}\right] + \left(\frac{1}{m} - \frac{1}{r_1}\right)\left[\beta_{yx}\left\{\frac{C_{300}}{S_x^2} - \frac{C_{210}}{S_{yx}}\right\} + \beta_{yx}\sum_{j=1}^{p}\beta_{xzj}\left\{\frac{C_{111}}{S_{yx}} - \frac{C_{201}}{S_x^2} - \frac{C_{003}}{S_{zj}^2} + \frac{C_{102}}{S_{xzj}}\right\}\right] \tag{13}$$

For sufficiently large populations, if the finite population correction is disregarded in equations (12) and (13), the bias of the estimators T, T_1 and T_2 approximated up-to the order $o(n^{-1})$ can be found in equations (9), (10), and (11).

Theorem 4.2: The MSE (T) approximated up to the first-order in the estimation of $\bar{Y}$, can be obtained as

$$MSE\left(T\right) = \varphi^2 MSE\left(T_1\right) + \left(1-\varphi\right)^2 MSE\left(T_2\right) + 2\varphi\left(1-\varphi\right)Cov\left(T_1, T_2\right) \tag{14}$$

where

$$MSE\left(T_1\right) = \left(\frac{1}{r_2} - \frac{1}{N}\right)A_1\, S_y^2 \tag{15}$$

$$MSE\left(T_2\right) = \left[\left(\frac{1}{m} - \frac{1}{N}\right)A_1 + \left(\frac{1}{m} - \frac{1}{r_1}\right)A_2\right]S_y^2 \tag{16}$$

and

$$Cov\left(T_1, T_2\right) = E\left[\left(T_1 - \bar{Y}\right)\left(T_2 - \bar{Y}\right)\right] = -\frac{A_1 S_y^2}{N} \tag{17}$$

where

$$A_1 = 1 - \sum_{j=1}^{p}\rho_{yz_j}^2 + \sum_{j\neq k=1}^{p}\rho_{yz_j}\rho_{yz_k}\rho_{z_jz_k}$$

and

$$A_2 = 1 - 2\rho_{yx} + \sum_{j=1}^{p}\rho_{yz_j}^2 - \sum_{\text{j筴=1}}^{p}\rho_{yz_j}\rho_{yz_k}\rho_{z_jz_k} \quad .$$

Proof: By the definition of mean square error, we have

$$MSE\left(T\right) = E\left[T - \bar{Y}\right]^2 = E\left[\varphi\left(T_1 - \bar{Y}\right) + \left(1-\varphi\right)\left(T_2 - \bar{Y}\right)\right]^2$$

$$=\varphi^2 \mathrm{MSE}\left(\mathrm{T}_1\right)+\left(1-\varphi\right)^2 \mathrm{MSE}\left(\mathrm{T}_2\right)+2\varphi\left(1-\varphi\right)\mathrm{E}\left[\left(\mathrm{T}_1-\bar{\mathrm{Y}}\right)\left(\mathrm{T}_2-\bar{\mathrm{Y}}\right)\right] \tag{18}$$

where

$$\mathrm{MSE}\left(\mathrm{T}_1\right)=\mathrm{E}\left[\mathrm{T}_1-\bar{\mathrm{Y}}\right]^2 \text{ and } \mathrm{MSE}\left(\mathrm{T}_2\right)=\mathrm{E}\left[\mathrm{T}_2-\bar{\mathrm{Y}}\right]^2$$

Now, using the expressions given in equations (7) and (8), expanding binomially, taking expectations, taking expectations up to o(n^{-1}), we have the expression of mean square error of the estimator Δas given in equation (14).

Since mean square error of the estimator T defined in equation (14) is a function of the unknown constant φ , therefore, it is minimized with respect to φ and subsequently the optimum value of φ is obtained as

$$\varphi_{\text{opt.}}=\frac{\mathrm{MSE}\left(\mathrm{T}_2\right)-\mathrm{Cov}\left(\mathrm{T}_1,\mathrm{T}_2\right)}{\mathrm{MSE}\left(\mathrm{T}_1\right)+\mathrm{MSE}\left(\mathrm{T}_2\right)-2\mathrm{Cov}\left(\mathrm{T}_1,\mathrm{T}_2\right)} \tag{19}$$

Substituting the optimum value $\varphi_{\text{opt.}}$ in equation (14) we obtain the optimum mean square error of Tas

$$\mathrm{MSE}\left(\mathrm{T}\right)_{\text{opt.}}=\frac{\mathrm{MSE}\left(\mathrm{T}_1\right).\mathrm{MSE}\left(\mathrm{T}_2\right)-\left\{\mathrm{Cov}\left(\mathrm{T}_1,\mathrm{T}_2\right)\right\}^2}{\mathrm{MSE}(\mathrm{T}_1)+\mathrm{MSE}(\mathrm{T}_2)\text{-}2\mathrm{Cov}\left(\mathrm{T}_1,\mathrm{T}_2\right)} \tag{20}$$

Further, substituting the values from equations (15), (16) and (17) in equation (20) yields the simplified value of $\mathrm{MSE(T)}_{\text{opt.}}$ as shown below in Theorem 4.3.

Theorem 4.3: The $\mathrm{MSE(T)}_{\text{opt.}}$ is derived as

$$\mathrm{MSE}\left(\mathrm{T}\right)_{\text{opt.}}=\frac{\left[\mathrm{A}_4+\mu\mathrm{A}_5+\mu^2\mathrm{A}_6\right]}{\mathrm{n}[\mu^2\mathrm{f}_2\mathrm{A}_2+\mu\mathrm{A}_3+\mathrm{f}_1\mathrm{A}_1]}\mathrm{S}_y^2 \tag{21}$$

where

$$\mathrm{A}_3=\mathrm{f}_1\mathrm{f}_2\left(\mathrm{A}_1+\mathrm{A}_2\right)-\mathrm{f}_1\mathrm{A}_1-\mathrm{f}_2\mathrm{A}_2,\ \mathrm{A}_4=\mathrm{f}_1\left(1-\mathrm{f}\right)\mathrm{A}_1^2+\left(\mathrm{f}_1-1\right)\mathrm{A}_1\mathrm{A}_2,$$

$$\mathrm{A}_5=\mathrm{ff}_1\left(1-\mathrm{f}_2\right)\mathrm{A}_1^2+\left\{\mathrm{ff}_2\left(1-\mathrm{f}_1\right)+1\right\}\mathrm{A}_1\mathrm{A}_2 \text{ and } \mathrm{A}_6=-\mathrm{ff}_2\mathrm{A}_1\mathrm{A}_2$$

In the context of equation (21), $\mathrm{MSE(T)}_{\text{opt.}}$ is a function of μ. Setting μ to 1 (representing no matching) is effective for estimating the population mean at single occasion, and setting μ to 0 (indicating full matching) is suitable for measuring changes over time. To design a strategy that is effective for both purposes, it is important to determine the optimal value for μ.

5. Optimum Replacement Strategy

In equation (21), $\mathrm{MSE(T)}_{\mathrm{opt}}$ is dependent on the fraction of new samples at the current survey occasion (μ), which plays a significant role in reducing survey costs. Therefore, it is important to minimize $\mathrm{MSE(T)}_{\mathrm{opt}}$ with respect to μ. The optimal value of μ, denoted μ_0, is calculated as

$$\mu_o=\frac{-\mathrm{Q}_2\ ?\ \sqrt{\mathrm{Q}_2^2\text{-}\mathrm{Q}_1\mathrm{Q}_2}}{\mathrm{Q}_1} \tag{22}$$

where $\mathrm{Q}_1=\mathrm{A}_3\mathrm{A}_6-\mathrm{f}_2\mathrm{A}_2\mathrm{A}_5$, $\mathrm{Q}_2=\mathrm{f}_1\mathrm{A}_1\mathrm{A}_6-\mathrm{f}_2\mathrm{A}_2\mathrm{A}_4$ and $\mathrm{Q}_3=\mathrm{f}_1\mathrm{A}_1\mathrm{A}_5-\mathrm{A}_3\mathrm{A}_5$.

For μ_o to be a valid solution, it must satisfy $\mathrm{Q}_2^2-\mathrm{Q}_1\mathrm{Q}_2\ \geq\ 0$. Admissibility of $\mu_{\text{min.}}$ depends on the condition, $0\leq\mu_{\text{min.}}\leq 1$, being met. By substituting the admissible value μ_o from equation (22) into equation (21), we get

$$\mathrm{MSE}\left(\mathrm{T}\right)_{\text{opt.}^*}=\frac{\left[\mathrm{A}_4+\mu_o\mathrm{A}_5+\mu_o^2\mathrm{A}_6\right]}{\mathrm{n}[\mu_o^2\mathrm{f}_2\mathrm{A}_2+\mu_o\mathrm{A}_3+\mathrm{f}_1\mathrm{A}_1]}\mathrm{S}_y^2 \tag{23}$$

7. Efficiency Comparison

In order to judge the effect of non-response on the precision of estimates under the two-occasion successive (rotation) sampling, the percent relative loss in efficiency of the estimator T has been obtained with respect to the estimator Δ. The estimator Δ is defined under the same circumstances as of T but in the case of complete information (with no missing data).

Consider the estimator Δ of $\overline{Y}$:

$$\Delta = \psi\Delta_1 + \left(1 - \psi\right)\Delta_2 \tag{24}$$

where $\Delta_1 = \overline{y}_u + \sum_{j=1}^{p} b_{yzj}(u)\left(\overline{Z}_j - \overline{z}_{ju}\right)$, $\Delta_2 = \dfrac{\overline{y}^*_m}{\overline{x}^*_m}\,\overline{x}^*_n$, $\overline{y}^*_m = \overline{y}_m + \sum_{j=1}^{p} b_{yzj}(m)\left(\overline{Z}_j - \overline{z}_{jm}\right)$,

$\overline{x}^*_n = \overline{x}_n + \sum_{j=1}^{p} b_{xzj}(n)\left(\overline{Z}_j - \overline{z}_{jn}\right)$ and $\overline{x}^*_m = \overline{x}_m + \sum_{j=1}^{p} b_{xzj}(m)\left(\overline{Z}_j - \overline{z}_{jm}\right)$.

ψ is an unknown real constant to be determined by the minimization of the mean square error of the estimator Δ.

The optimum mean square error of Δ following Sukhatme et al. [19] is given by

$$M(\Delta)_{opt^*} = A_1\left[\frac{A_1+\mu A_2}{A_1+\mu^2 A_2} - f\right]\frac{S_y^2}{n} \tag{25}$$

where μ_1 is the admissible value of $\mu_{min.}$ which is obtained as

$$\mu_1 = \frac{-A_1 \ ? \ \sqrt{A_1^2 + A_1 A_2}}{A_2} \tag{26}$$

Remark 1: The permissible value of μ_1 in equation (26) is obtained in the similar manner as μ_0.

The percent relative loss in precision of the estimators T with respect to the estimator Δ under their respective optimality conditions are given by

$$L = \frac{MSE\left(T\right)_{opt.^*} - MSE\left(\Delta\right)_{opt.^*}}{MSE\left(T\right)_{opt.^*}} \ ? \ 00$$

7.1 Empirical Study

The expressions of the minimum μ and the percent relative losses L are in terms of population correlation coefficients. Therefore, the values of minimum μ and L have been computed for different choices of positive correlations. For empirical studies few cases arise:

Case 7.1.1: Assuming p = 1, the proposed estimator aligns with the specific estimator developed by Singh and Karna [18].

Case 7.1.2: when p = 2, the values A1 and A2 take the form as $A_1 = 1 - \rho_1^2 - \rho_2^2 + 2\rho_1\rho_2\rho_{z1z2}$ and $A_2 = 2\rho_{yx}\left(\rho_1^2 + \rho_2^2\right) - \rho_{yx}^2$ where $\rho_{yz_1} = \rho_{xz_1} = \rho_1$, and $\rho_{yz_2} = \rho_{xz_2} = \rho_2$. Using these we have the values of μ_0, μ_1 and L for a range of correlation values displayed in Tables 1 and 2.

Case 7.1.3: choosing p = 2, the values of A1 and A2 change as $A_1 = 1 - \rho_1^2 - \rho_2^2 - \rho_3^2 + 2\rho_1\rho_2\rho_{z_1z_2} + 2\rho_2\rho_3\rho_{z_2z_3} + 2\rho_1\rho_3\rho_{z_1z_3}$ and $A_2 = 2\rho_{yx}\left(\rho_1^2 + \rho_2^2 + \rho_3^2\right) - \rho_{yx}^2$ where $\rho_{yz_1} = \rho_{xz_1} = \rho_1$, $\rho_{yz_2} = \rho_{xz_2} = \rho_2$, $\rho_{yz_3} = \rho_{xz_3} = \rho_3$. Again the values of μ_0, μ_1 and L for a variety of correlation cases are outlined in Tables 3.

8. Existence of M Under Various Values of μo and r1

For a random value of r1 (the count of non-participants in the initial n sized sample) and optimum value of μo, it is obvious to verify the condition m ≤ r1. Table 4 illustrates the computed m values for different scenarios involving n, r1, μo and various correlations for the estimator T.

Remark 2: "*" in the Tables 1 – 4 denotes that no admissible values for μ exist.

Table 1. Percentage of relative loss L in precision of T concerning Δ at permissible values of μ, when p = 2 and ρ_{z1z2} = 0.7

for t_1 = 0.05 and t_2 = 0.05										
ρ_{yx}		0.5			0.7			0.9		
ρ_1	ρ_2	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.5	0.5	0.65	0.47	3.10	0.49	0.54	3.18	0.67	0.67	4.38
	0.7	0.57	0.46	2.73	0.44	0.52	2.91	0.65	0.66	4.28
	0.9	0.50	0.43	2.16	*	-	-	0.62	0.63	4.04
0.7	0.5	0.57	0.46	2.73	0.44	0.52	2.91	0.66	0.66	4.28
	0.7	0.55	0.45	2.59	0.39	0.52	2.72	0.65	0.65	4.23
	0.9	0.50	0.43	2.20	*	-	-	0.62	0.63	4.05
0.9	0.5	0.50	0.43	2.16	*	-	-	0.62	0.63	4.03
	0.7	0.50	0.43	2.20	*	-	-	0.62	0.63	4.05
	0.9	0.48	0.41	1.95	0.66	0.48	3.13	0.61	0.61	3.93
For t_1 = 0.05 and t_2 = 0.15										
ρ_{yx}		0.5			0.7			0.9		
ρ_1	ρ_2	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.5	0.5	*	-	-	0.33	0.54	7.44	0.63	0.67	9.91
	0.7	0.76	0.46	10.5	0.19	0.52	6.27	0.61	0.66	9.77
	0.9	0.59	0.43	9.08	*	-	-	0.57	0.63	9.40
0.7	0.5	0.76	0.46	10.5	0.19	0.52	6.27	0.61	0.66	9.77
	0.7	0.71	0.45	10.1	0.05	0.52	5.13	0.61	0.65	9.69
	0.9	0.60	0.43	9.16	*	-	-	0.57	0.63	9.43
0.9	0.5	0.59	0.43	9.08	*	-	-	0.57	0.63	9.40
	0.7	0.60	0.43	9.16	*	-	-	0.57	0.63	9.43
	0.9	0.56	0.41	8.72	*	-	-	0.55	0.61	9.23
For t_1 = 0.15 and t_2 = 0.15										
ρ_{yx}		0.5			0.7			0.9		
ρ_1	ρ_2	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.5	0.5	*	-	-	0.40	0.54	9.20	0.67	0.67	13.2
	0.7	0.79	0.46	10.4	0.27	0.52	7.87	0.65	0.65	12.9
	0.9	0.64	0.43	8.27	*	-	-	0.61	0.62	12.2
0.7	0.5	0.79	0.46	10.4	0.27	0.52	7.87	0.65	0.65	12.9
	0.7	0.74	0.45	9.85	0.15	0.52	6.72	0.65	0.65	12.7
	0.9	0.64	0.43	8.40	*	-	-	0.62	0.63	12.2
0.9	0.5	0.64	0.43	8.27	*	-	-	0.61	0.62	12.2
	0.7	0.64	0.43	8.40	*	-	-	0.62	0.63	12.2
	0.9	0.61	0.41	7.61	*	-	-	0.60	0.61	11.9

Table 2. Percentage of relative loss L in precision of T concerning Δ at permissible values of μ, when p = 2 and ρ_{z1z2} = 0.9

For t_1 = 0.05 and t_2 = 0.05										
ρ_{yx}		0.5			0.7			0.9		
ρ_1	ρ_2	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.5	0.5	*	-	-	0.52	0.55	3.35	0.68	0.68	4.47
	0.7	0.72	0.48	3.33	0.50	0.54	3.25	0.67	0.67	4.42
	0.9	0.57	0.46	2.73	0.44	0.52	2.91	0.65	0.65	4.28
0.7	0.5	0.72	0.48	3.33	0.50	0.54	3.25	0.67	0.67	4.42
	0.7	0.75	0.48	3.42	0.51	0.55	3.27	0.68	0.67	4.43
	0.9	0.63	0.47	3.03	0.48	0.54	3.14	0.67	0.67	4.37
0.9	0.5	0.57	0.46	2.73	0.44	0.52	2.91	0.65	0.65	4.28
	0.7	0.63	0.47	3.03	0.48	0.54	3.14	0.67	0.67	4.37
	0.9	0.64	0.47	3.04	0.49	0.54	3.15	0.67	0.67	4.37

For t_1 = 0.05 and t_2 = 0.15										
ρ_{yx}		0.5			0.7			0.9		
ρ_1	ρ_2	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.5	0.5	*	-	-	0.40	0.55	8.02	0.65	0.68	10.0
	0.7	*	-	-	0.36	0.54	7.71	0.64	0.67	9.96
	0.9	0.76	0.46	10.5	0.19	0.52	6.27	0.61	0.65	9.77
0.7	0.5	*	-	-	0.36	0.54	7.71	0.64	0.67	9.96
	0.7	*	-	-	0.37	0.55	7.78	0.64	0.67	9.98
	0.9	0.94	0.47	12.0	0.32	0.54	7.32	0.63	0.67	9.89
0.9	0.5	0.76	0.46	10.5	0.19	0.52	6.27	0.61	0.65	9.77
	0.7	0.94	0.47	12.0	0.32	0.54	7.32	0.63	0.67	9.89
	0.9	0.95	0.47	12.1	0.32	0.54	7.35	0.63	0.67	9.90

For t_1 = 0.15 and t_2 = 0.15										
ρ_{yx}		0.5			0.7			0.9		
ρ_1	ρ_2	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.5	0.5	*	-	-	0.46	0.55	9.94	0.69	0.68	13.4
	0.7	*	-	-	0.43	0.54	9.53	0.68	0.67	13.3
	0.9	0.79	0.46	10.4	0.27	0.52	7.87	0.65	0.65	12.9
0.7	0.5	*	-	-	0.43	0.54	9.53	0.68	0.67	13.3
	0.7	*	-	-	0.44	0.55	9.62	0.68	0.67	13.3
	0.9	0.95	0.47	12.0	0.39	0.54	9.04	0.67	0.67	13.1
0.9	0.5	0.79	0.46	10.4	0.27	0.52	7.87	0.65	0.65	12.9
	0.7	0.95	0.47	12.0	0.39	0.54	9.04	0.67	0.67	13.1
	0.9	0.96	0.47	12.1	0.39	0.54	9.08	0.67	0.67	13.2

Table 3. Percentage of relative loss L in precision of T concerning Δ at permissible values of μ, when p = 3

For $t_1 = 0.05$, $t_2 = 0.05$, $\rho_{yx} = 0.9$, $\rho_{z1z2} = 0.9$, $\rho_{z1z3} = 0.9$, $\rho_{z2z3} = 0.9$										
ρ_1		0.5			0.7			0.9		
ρ_2	ρ_3	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.5	0.5	0.74	0.73	4.82	0.74	0.74	4.87	0.75	0.74	4.88
	0.7	0.74	0.74	4.87	0.75	0.75	4.93	0.76	0.76	4.96
	0.9	0.75	0.74	4.88	0.76	0.76	4.96	0.77	0.76	5.00
0.7	0.5	0.74	0.74	4.87	0.75	0.75	4.93	0.76	0.76	4.96
	0.7	0.75	0.75	4.93	0.77	0.76	5.00	0.77	0.77	5.04
	0.9	0.76	0.76	4.96	0.77	0.77	5.04	0.78	0.78	5.09
0.9	0.5	0.75	0.74	4.88	0.76	0.76	4.96	0.77	0.76	5.00
	0.7	0.76	0.76	4.96	0.77	0.77	5.04	0.78	0.78	5.09
	0.9	0.77	0.76	5.00	0.78	0.78	5.09	0.79	0.79	5.15
For $t_1 = 0.1$, $t_2 = 0.1$, $\rho_{yx} = 0.9$, $\rho_{z1z2} = 0.9$, $\rho_{z1z3} = 0.9$, $\rho_{z2z3} = 0.9$										
ρ_1		0.5			0.7			0.9		
ρ_2	ρ_3	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.5	0.5	0.74	0.73	9.65	0.75	0.74	9.73	0.75	0.74	9.75
	0.7	0.75	0.74	9.73	0.76	0.75	9.85	0.76	0.76	9.91
	0.9	0.75	0.74	9.75	0.76	0.76	9.91	0.77	0.76	9.99
0.7	0.5	0.75	0.74	9.73	0.76	0.75	9.85	0.76	0.76	9.91
	0.7	0.76	0.75	9.85	0.77	0.76	9.98	0.78	0.77	10.0
	0.9	0.76	0.76	9.91	0.78	0.77	10.0	0.79	0.78	10.1
0.9	0.5	0.75	0.74	9.75	0.76	0.76	9.91	0.77	0.76	9.99
	0.7	0.76	0.76	9.91	0.78	0.77	10.0	0.79	0.78	10.1
	0.9	0.77	0.76	9.99	0.79	0.78	10.1	0.80	0.79	10.2
For $t_1 = 0.15$, $t_2 = 0.15$, $\rho_{yx} = 0.9$, $\rho_{z1z2} = 0.9$, $\rho_{z1z3} = 0.9$, $\rho_{z2z3} = 0.9$										
ρ_1		0.5			0.7			0.9		
ρ_2	ρ_3	μ_0	μ_1	L	μ_0	μ_1	L	μ_0	μ_1	L
0.5	0.5	0.74	0.73	14.4	0.75	0.74	14.5	0.75	0.74	14.6
	0.7	0.75	0.74	14.5	0.76	0.75	14.7	0.77	0.76	14.8
	0.9	0.75	0.74	14.6	0.77	0.76	14.8	0.77	0.76	14.9
0.7	0.5	0.75	0.74	14.5	0.76	0.75	14.7	0.77	0.76	14.8
	0.7	0.76	0.75	14.7	0.77	0.76	14.9	0.78	0.77	15.0
	0.9	0.77	0.76	14.8	0.78	0.77	15.0	0.79	0.78	15.2
0.9	0.5	0.75	0.74	14.6	0.77	0.76	14.8	0.77	0.76	14.9
	0.7	0.77	0.76	14.8	0.78	0.77	15.0	0.79	0.78	15.2
	0.9	0.77	0.76	14.9	0.79	0.78	15.2	0.80	0.79	15.3

Table 4. The values of m for different choices of r_1 and μ_0, when non response occurs at both the occasions, when p = 2 and ρ_{z1z2} = 0.9

n = 50, t_1 = 0.05, t_2 = 0.05							
ρ_1		0.7			0.9		
ρ_2	ρ_{yx}	μ_0	m	r_1	μ_0	m	r_1
0.7	0.5	0.7519	12	48	0.6392	18	48
	0.7	0.5131	24	48	0.4894	25	48
	0.9	0.6809	16	48	0.6718	16	48
0.9	0.5	0.6392	18	48	0.6435	18	48
	0.7	0.5056	25	48	0.4911	25	48
	0.9	0.6778	16	48	0.6724	16	48
n = 50, t_1 = 0.05, t_2 = 0.15							
ρ_1		0.7			0.9		
ρ_2	ρ_{yx}	μ_0	m	r_1	μ_0	m	r_1
0.7	0.5	*	-	48	0.9459	3	48
	0.7	0.3773	31	48	0.3209	34	48
	0.9	0.6472	18	48	0.6359	18	48
0.9	0.5	0.9459	3	48	0.9585	2	48
	0.7	0.3209	34	48	0.3250	34	48
	0.9	0.6359	18	48	0.6366	18	48
n = 50, t_1 = 0.15, t_2 = 0.05							
ρ_1		0.7			0.9		
ρ_2	ρ_{yx}	μ_0	m	r_1	μ_0	m	r_1
0.7	0.5	0.7780	11	43	0.6772	16	43
	0.7	0.5644	22	43	0.5432	23	43
	0.9	0.7145	14	43	0.7064	15	43
0.9	0.5	0.6772	16	43	0.6810	16	43
	0.7	0.5432	23	43	0.5446	23	43
	0.9	0.7064	15	43	0.7069	15	43
n = 50, t_1 = 0.15, t_2 = 0.15							
ρ_1		0.7			0.9		
ρ_2	ρ_{yx}	μ_0	m	r_1	μ_0	m	r_1
0.7	0.5	*	-	43	0.9516	2	43
	0.7	0.4429	28	43	0.3924	30	43
	0.9	0.6843	16	43	0.6742	16	43
0.9	0.5	0.9516	2	43	0.9628	2	43
	0.7	0.3924	30	43	0.3960	30	43
	0.9	0.6742	16	43	0.6749	16	43

9. Conclusions

The following insights can be drawn from Tables 1-4:

- Table 1 and 2 show that for fixed values of $t_{2,}$ ρ_1 and ρ_{yx}, as t_1 increases, μ_0 rises while L decreases. This implies that a higher non-response rate at the initial occasion necessitates a larger fresh sample in the current occasion to improve the precision of the estimates. Conversely, for fixed t_1, ρ_1 and ρ_{yx} , both μ_0 and L increase with t_2. When t_1, t_2 and ρ_{yx} are constant, an increase in ρ_1 leads to a decrease in both μ_0 and L. This suggests that a stronger correlation between the study and auxiliary variate reduces the amount of fresh sample needed and the loss in precision.
- Table 3 reveals that for p=3, the behaviour of the estimator mirrors that observed for p=2, as indicated in previous point.
- Table 4 confirms that for permissible values of μ_o, the condition $m \leq r_1$ is consistently met.

A review of all scenarios indicates that the loss in precision is generally minimal. In some instances, a negative loss is observed, demonstrating the effectiveness of the proposed imputation methods. Hence, the imputation techniques introduced in this study are effective for addressing non-response issues across the occasions in two-occasion successive (rotation) sampling.

References

Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. In *Iowa Agricultural Experiment Station Road* (Vol. 304, pp. 1-104).

Sen, A. R. (1971). Successive sampling with two auxiliary variables. *Sankhya, B, 33*, 371-378.

Sen, A. R. (1972). Successive sampling with p (p ≥ 1) auxiliary variables. *Annals of Mathematical Statistics, 43*, 2031-2034.

Sen, A. R. (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics, 29*, 381-385.

Singh, G. N., & Singh, V. K. (2001). On the use of auxiliary information in successive sampling. *Journal of the Indian Society of Agricultural Statistics, 54*(1), 1-12.

Feng, S., & Zou, G. (1997). Sample rotation method with auxiliary variable. *Communications in Statistics - Theory and Methods, 26*(6), 1497-1509.

Singh, G. N. (2003). Estimation of population mean using auxiliary information on recent occasion in h occasions successive sampling. *Statistics in Transition, 6*(4), 523-532.

Singh, G. N. (2005). On the use of chain-type ratio estimator in successive sampling. *Statistics in Transition, 7*(1), 21-26.

Singh, G. N., & Karna, J. P. (2009a). Estimation of population mean on current occasion in two-occasion successive sampling. *Metron, 67*(1), 69-85.

Singh, G. N., & Karna, J. P. (2009b). Search of effective rotation patterns in presence of auxiliary information in successive sampling over two occasions. *Statistics in Transition - New Series, 10*(1), 59-73.

Karna, J. P., & Nath, D. C. (2016). Rotation sampling scheme using transformed auxiliary variable. *Journal of Statistics and Management Systems, 19*(6), 739-754.

Karna, J. P., & Nath, D. C. (2018). Improved rotation patterns using two auxiliary variables. *Statistics in Transition - New Series, 19*(1), 25-44.

Kalton, G., Kasprzyk, D., & Santos, R. (1981). Issues of non-response and imputation of income and program participation. In *Current Topics in Survey Sampling* (pp. 455-480). Academic Press.

Lee, H., Rancourt, E., & Särndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics, 10*(3), 231-243.

Lee, H., Rancourt, E., & Särndal, C. E. (1995). Variance estimation in the presence of imputed data for the generalized estimation system. In *Proceedings of the American Statistical Association* (Social Survey Research Methods Section, pp. 384-389).

Ahmed, M. S., Al-Titi, O., Al-Rawi, Z., & Abu-Dayyeh, W. (2006). Estimation of a population mean using different imputation methods. *Statistics in Transition, 7*(6), 1247-1264.

Singh, S. (2009). A new method of imputation in survey sampling. *Statistics: A Journal of Theoretical and Applied Statistics, 43*(5), 499-511.

Singh, G. N., & Karna, J. P. (2010). Some imputation methods to minimize the effect of non-response in two-occasion rotation patterns. *Communications in Statistics - Theory and Methods, 39*(18), 3264-3281.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., & Asok, C. (1984). *Sampling theory of surveys with applications*. Iowa State University Press, Ames, IA, USA, and Indian Society of Agricultural Statistics.

A Weighted Epidemic Chain Binomial Model (WECBM) with One-Introductory Case & Its Application

Abstract

This paper emphasizes the theoretical development of Weighted Epidemic Chain Binomial Model (WECBM), its properties and its application. Here, a more detailed comparison of the fits provided by Heasman & Reid (1961) and Becker (1980) for Reed-Frost chain binomial model and Becker's ECM (general and with β=1) with the fitting of WECBM for four and five member households with one introductory case are discussed. For drawing an exhaustive comparison, both Heasman-Reid data (1961) and current epidemic data (2016) have been used.

For this study, the chi-square test for goodness of fit and R-software version 3.3.2 were used and applications of the above mentioned models were shown.

From the results, it has been observed that in the case of fitting of three epidemic chain models to current epidemic data for five member households, WECBM gives the best fit amongst the three. For four member households it gives a better fit more or less similar to Becker's ECM (general). Again, in the case of fitting of the four epidemic chain models to Heasman & Reid data, Becker's ECM (general) gives the best fit and WECBM gives the third best fit. However, Becker's ECM with β =1 gives not so not-so-good fit in all three cases.

Keywords: infectious Diseases (IDs), WECBM

1. Introduction

Bailey(1975)[1] in his book viewed that, in the simplest continuous-time models, the latent period is assumed to be zero, so that the infected individual becomes infectious to others immediately after the receipt of the infection. And on the other hand, in the simplest discrete-time models like chain-binomial model, the latent period is considered to be constant, and an infectious period is assumed to be short.

According to Bailey[2-5], the chain binomial models can be used quite successfully in the statistical fitting of certain epidemic theories to real –life data relating to smaller group such as families in statistical theory. But, so far the analysis of epidemic processes in large groups is concerned, the discrete-time models are rather difficult, and it is easy to rely on the insights provided by continuous-time models for understanding the behavior of epidemics in reasonably large groups. As the interest is to capture information from infected households having three, four or five members, therefore the discrete-time models are considered for the study. It is perhaps essential to take a quick look at the way in which the discrete-time models were constructed, as future developments may enable them to be used as a basis for the investigation of the corresponding stochastic processes.

Modeling the spread of these diseases among individuals in a population is a complex task and it becomes necessary to make several mathematical and biological assumptions about the factors which control the disease process.

Cairoli (1988)[6] discussed that, the mathematical formulation of discrete time epidemic models flows from attempts by several investigators to present models which realistically describe the progress of a disease through a population. The usual starting point in model building is the set of assumptions about those factors which control the spread of a disease. These assumptions create a model which describes actual disease patterns. The epidemic model is then useful as a predictive tool for epidemiologists.

This paper presents an extension of the Becker's epidemic chain model [7-9], where Becker assumes the probability of being infected to follow a beta distribution of first kind.

In case of application of chain binomial model to real-life epidemic data, so far no attempt has been made to assign weights to chains of infections. In analyzing the epidemic data it is evident that some chains of infections are more prevalent in occurrence than others. It may be the case that these frequently occurring chains consists of some unexplored infective factors and may prove of importance in epidemiological studies. It will therefore be of interest to study the pattern of infection through a weighted distribution where we assign some weight to each of the chain of infection. In this work, first a general weight expression has been used to develop the weighted epidemic chain binomial model from Becker's ECM[7] and to study the properties of the WECBM. Thereafter, a specific weight, proportional to the occurrence of the corresponding chain, has been used to fit the WECBM to the epidemic data. However, it may be extended to any weight assigned to the chains of infection.

In order to make a more exhaustive comparison with the other existing models, an attempt has been made in this paper to develop a new discrete time model named as weighted epidemic chain binomial model (WECBM) by assuming a weighted beta distribution of first kind for the probability of being infected by contact with a given infective from the same household. The chain probabilities of the WECBM for three, four and five member household with one introductory case are also developed.

The chain binomial models (Bailey, 1975)[1] have met with logical success, whenever fitted to infectious diseases data for households, for example diseases like common cold or influenza. Also, Heasman and Reid (1961)[10] demonstrated that, the Reed-Frost chain binomial model can provide an adequate fit to data on outbreaks of the common cold in households of size five. And, by comparing the observed frequencies with the expected frequencies for the total number of cases, they also demonstrated that, the stochastic version of the Kermack-McKendrick epidemic model [1] may provide an even better fit. In the later stage, a detailed comparison of the fits provided by these two models was attempted by Becker (1980) [7] by formulating an ECM. The Becker's ECM includes, as a particular case (with $\beta=1$), the ECM corresponding to the stochastic version of the Kermack-McKendrick epidemic model [1] and, as a limiting case, the Reed-Frost chain binomial model. Becker (1980)[7] studied the advantages of the more general model and illustrated the same with an application to Heasman-Reid common cold data. In fact, the assumptions made were found similar in many ways to those used by Ludwig (1975)[11] in his derivations of the final size distributions for epidemics with arbitrary time-dependent infectiousness.

This paper emphasizes on the theoretical development of the model, its properties and its application. A more detailed comparison of the fits provided by Heasman & Reid (1961)[10] and Becker (1980)[7] for the Reed-Frost chain binomial model and Becker's ECM (general and with $\beta=1$) with the fitting of the WECBM for four and five member households with one introductory case will be discussed. The same is shown in the paper using both the Heasman-Reid data (1961)[10] and current epidemic data(2016)[12] for drawing the exhaustive comparison.

2. Objectives

The main objectives of the paper are

(i) to develop a discrete time model named as weighted epidemic chain binomial model(WECBM) from Becker's ECM(1980)[7].

(ii) to derive the conditional probability and epidemic chain probabilities of WECBM for three, four and five member household with one introductory case.

(iii) to fit the model WECBM to both the traditional Heasman-Reid epidemic data (1961)[10] and current epidemic data (2016)[12] for four and five member households with one introductory case.

(iv) to compare the results of WECBM with other epidemic chain model viz. Reed-Frost, Becker's ECM (general) and Becker's ECM with $\beta=1$.

3. Material and Methods

3.1 Data Used

The Heasman-Reid data [10] is a classic set of data in the area mathematical epidemiological studies. Thus in the present study, this data is considered as the secondary data, so that further comparison can be drawn with the current epidemic data.

This household data provide the ideal population for testing the adequacy of chain binomial models. The application of this data was shown by Heasman and Reid in their work [10]. Later on the same set of data was used by Becker (1980)[7] for application to ECM. It is only for such small groups as households that the different possible chains can be readily classified.

Along with the above mentioned epidemic data, a new set of epidemic data so collected during 2015-2016 from the 4 selected wards of Guwahati Municipality Corporation, Guwahati, Assam, India are being used for application to the WECBM, Becker's ECM (1980)[7] and stochastic version of the Kermack-McKendrick epidemic model (Bailey, 1975)[1] i.e., Becker's ECM with $\beta=1$.

3.2 Chi-square Test for Goodness of Fit

The usual chi-square test for goodness of fit is used in this paper for the purpose of model fitting.

3.3 Statistical Software Used: R Version 3.3.2

R-software has been used to find the factorial powers, to fit the models through parameter estimation of different distributions under study, to find the expected frequencies as well as to calculate the chi-square for testing goodness of fit.

4. Some Related Theories

In this section, some theories related to this paper are given.

4.1 Chain binomial models (Bailey, 1975)[1]

This is a model that satisfies both the criteria of being mathematically accurate as well as relatively simple in describing essential features of the epidemic. In describing viral diseases such as measles, chicken pox, influenza, and the common cold, these models were found to be very useful.

The theory behind the generation of this model is given below. This model is set up on some basic assumptions like, the population under consideration is assumed to be closed and homogeneously mixed.

As, the basic idea of the chain binomial models is that, an infectious period is contracted to a single point and the latent period is fixed, therefore this may be used as a unit of time. The population consists of two classes of individuals, susceptible and infectives. The models assume that all individuals have equal susceptibility, i.e., the capability to transmit the disease and have the ability to be removed from observation when the transmitting period of the disease is over.

The theory given by Bailey[1] assumed that,$S_t$is the number of susceptible in the group just before time t, $I_t$is the number of infected individuals just before time t who actually become infectious at that instant.

It was further defined as, a chance of adequate contact p(=1-q), which is the probability of a contact at any time between any two specified members of the group sufficient to produce a new infection if one of them is susceptible and one infectious.

It was followed from the above that, the chance that any given susceptible will escape infection at time t is q^{i_t}, i.e., will have adequate contact with none of the I_t infectives.

Thus $1 - q^{i_t}$ is the chance of adequate contact with at least one infective, and this is what is required for infection to occur.

The conditional probability of I_{t+1} new infections taking place (who will become infectious at time t+1) is therefore given by the binomial distribution

$$P\{I_{t+1} = i_{t+1} | I_t = i_t, S_t = s_t\} = \binom{s_t}{i_{t+1}} \left(1 - q^{i_t}\right)^{i_{t+1}} q^{i_t s_{t+1}} \tag{1}$$

where, $S_t = I_{t+1} + S_{t+1}$

The process develops in a series of binomial distributions like Eq. (1). Hence the name of the chain binomial model is adopted from the chain binomial process. The different chain binomial models are discussed below in detail.

In this paper, basically, the three types of distribution models are used to test the goodness of fit and to draw a comparison with other existing results. They are given below.

4.1.1 Reed-Frost Model (Bailey, 1975)[1]

It is used here only for comparison of the results. The final results given by Heasman & Reid (1961)[10] using the traditional epidemic data are used here for final comparison.

4.1.2 Becker's ECM (Becker, 1980)[7]

The model theory will be discussed here, as this model is considered to be the most essential and base model of the present study and the theoretical developments to be shown in this paper.

(a) Chains of Infection -Becker (1980)[7] described that, as it is not always possible to determine which infective is actually responsible for a certain infection. But, it can be made simpler by making use of the gaps between the cases, to actually partition the cases of a household into generations: the susceptible who gets infected by direct contact with the introductory cases are said to form the first generation of cases; the susceptibles which gets infected by direct contact with the first generation cases are said to form the second generation and so on. By an epidemic chain, it truly means the enumeration of the number of cases in each of the generation.

(b) Number of possible chains - This theory was reviewed as a part of the published work of Nath et. al. (2017)[13]. In general, for a binomial distribution, the formula for the total number of epidemic chains actually possible for the households of size *m* containing *j* introductory cases is given as

$$\sum_{i=0}^{m-j} C_i = 2^{m-j} \tag{2}$$

where *m* and *j* take the values as *m=3,4 or 5* and *j=1,2 or 3.*

(c) Conditional and unconditional probability -Corresponding to a given infective *A*, the conditional probability that *r* out of *k* susceptible of the household escape infection by *A* is given by $C_r^k \varepsilon^r (1-\varepsilon)^{k-r}$.Here ε is allowed to vary according to some known distribution. Becker (1980)[7] considered beta distribution of the first kind having the density

$$f(\varepsilon)=\frac{1}{B(\alpha,\beta)}\varepsilon^{\alpha-1}(1-\varepsilon)^{\beta-1}, o<\varepsilon<1, \alpha,\beta>0 \tag{3}$$

and$E\{\varepsilon^r(1-\varepsilon)^{k-r}\}$ is simplified as

$$E\{\varepsilon^r(1-\varepsilon)^{k-r}\}=\frac{\alpha^{(r)}\beta^{(k-r)}}{(\alpha+\beta)^{(k)}} \tag{4}$$

Then unconditional probability is given by

$$C_r E\{\varepsilon^r(1-\varepsilon)^{k-r}\}=C_r\frac{\alpha^{(r)}\beta^{(k-r)}}{(\alpha+\beta)^{(k)}} \tag{5}$$

(d) Chain probabilities - Becker [7] explained that, out of the sixteen(16) possible epidemic chains, let *1-2-1-0* be one such chain out of the sixteen(16) combinations, where this chain *1-2-1-0* actually denotes the chain consisting of one introductory case, two first-generation cases, one second-generation case and no cases in later generation. In the chain *1-2-1-0,* 1: Introductory case; 2: First Generation case; 1: Second Generation case; 0: Third Generation case.

To explain the calculation of the probabilities associated with the various possible epidemic chains, Becker [7] considered the chain *1-1-2-0* in a five-member household including one introductory case. The actual probability of this chain, conditional on the probabilities $\varepsilon_1, \varepsilon_2, \varepsilon_3 \wedge \varepsilon_4$that a given susceptible escape infection by each of the four infected individuals, respectively was found to be

$$C_3\varepsilon_1^3(1-\varepsilon_1)C_1\varepsilon_2^1(1-\varepsilon_2)^2C_1(\varepsilon_3\varepsilon_4)^1(1-\varepsilon_3\varepsilon_4)^0 \tag{6}$$

And the unconditional probability of the chain *1-1-2-0* in a five-member household including one introductory case is expressed by Becker [7], as

$$E\{C_3\varepsilon_1^3(1-\varepsilon_1)C_1\varepsilon_2^1(1-\varepsilon_2)^2C_1(\varepsilon_3\varepsilon_4)^1(1-\varepsilon_3\varepsilon_4)^0\}=12\frac{\alpha^3\alpha^{(3)}\beta^{(2)}\beta}{(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}(\alpha+\beta)^2} \tag{7}$$

It is used in this study for the model fitting of both general and a particular case *(β=1)* of Becker's model.

4.1.3 WECBM- The Formulation of WECBM is Given in Next Section

5. Weighted Epidemic Chain Binomial Model (WECBM)

As discussed earlier, why it is important to assign weights to the observed data, keeping in view the practical situation of data collection. An attempt has been made in the present study to introduce the concept of assigning weights to the existing distribution i.e., to Becker's ECM [7]. In general, let us consider $w(\varepsilon)=\varepsilon^\gamma$ as the weights to be assigned to the Becker's model (1980)[7]. Therefore the probability density function for the new model i.e., the weighted epidemic chain binomial model (WECBM) is written as,

$$Q(\varepsilon)=\frac{1}{B(\alpha+\gamma,\beta)}\varepsilon^{\gamma+\alpha-1}(1-\varepsilon)^{\beta-1}, o<\varepsilon<1, \alpha,\beta>0 \tag{8}$$

Then $E\{\varepsilon^r(1-\varepsilon)^{k-r}\}$ is expressed as

$$E\{\varepsilon^r(1-\varepsilon)^{k-r}\}=\frac{(\alpha+\gamma)^{(r)}\beta^{(k-r)}}{(\alpha+\beta+\gamma)^{(k)}} \tag{9}$$

Special case: Let us consider$\gamma=1$ which implies $w(\varepsilon)=\varepsilon$,

For practical purposes putting$\gamma=1$ in Eq. (9), $E\{\varepsilon^r(1-\varepsilon)^{k-r}\}$ is written as

$$E\{\varepsilon^r(1-\varepsilon)^{k-r}\}=\frac{(\alpha+1)(\alpha+2)_{\dots}(\alpha+r)\beta^{(k-r)}}{(\alpha+\beta+1)(\alpha+\beta+2)_{\dots}(\alpha+\beta+k)} \tag{10}$$

$$E\{\varepsilon^r(1-\varepsilon)^{k-r}\}=\frac{(\alpha+\beta)}{\alpha}\frac{\alpha^{(r+1)}\beta^{(k-r)}}{(\alpha+\beta)^{(k+1)}} \tag{11}$$

where

$$\alpha^{(r)} = \alpha(\alpha+1)(\alpha+2)_{\dots}(\alpha+r-1)$$

Then unconditional probability is given by

$$C_r E\{\varepsilon^r(1-\varepsilon)^{k-r}\} = C_r \frac{(\alpha+\beta)}{\alpha} \frac{\alpha^{(r+1)}\beta^{(k-r)}}{(\alpha+\beta)^{(k+1)}} \tag{12}$$

5.1 Chain Probabilities of WECBM

Let us consider an illustration of a five-member household with one introductory case, as per the formula Eq. (1), wherein the total no. of possible epidemic chains are $2^{5-1} = 16$. Similar to that of the chain probabilities calculated by Becker [7], an attempt has been made to find the chain probabilities for the weighted epidemic chain binomial model. Furthermore, let, 1-2-1-0 be one such chain out of the total 16(sixteen) possible combinations, where this chain 1-2-1-0 actually denotes the chain consisting of one introductory case, two first-generation cases, one second-generation case, and no cases in the last generation.

Therefore, similar to Becker's illustration, the computation of the probabilities associated with the different possible epidemic chains, the chain 1-1-2-0 is considered in a five-member household including one introductory case. The probability of this chain, conditional on the probabilities $\varepsilon_1, \varepsilon_2, \varepsilon_3 \wedge \varepsilon_4$that a given susceptible escape infection by each of the four infected individuals, respectively is found to be $C_3\varepsilon_1^3(1-\varepsilon_1)C_1\varepsilon_2^1(1-\varepsilon_2)^2C_1(\varepsilon_3\varepsilon_4)^1(1-\varepsilon_3\varepsilon_4)^0$ (13)

The unconditional probability of the chain 1-1-2-0 in a household of size five including one introductory case is expressed to be as

$$E\{C_3\varepsilon_1^3(1-\varepsilon_1)C_1\varepsilon_2^1(1-\varepsilon_2)^2C_1(\varepsilon_3\varepsilon_4)^1(1-\varepsilon_3\varepsilon_4)^0\} = 12\frac{(\alpha+\beta)^4\alpha^{(4)}\left[\alpha^{(2)}\right]^3\beta^{(2)}\beta}{\alpha^4(\alpha+\beta)^{(5)}(\alpha+\beta)^{(4)}(\alpha+\beta)^{(2)}} \tag{14}$$

The probabilities for all the possible epidemic chains for WECBM for three, four, and five member households with one introductory case are presented in the following three tables 1, 2, and 3 respectively. In all three tables, the chain probabilities of WECBM and Becker's ECM (1980)[7] are presented for theoretical comparison.

Table 1. Epidemic chain probabilities for 3-member households with one introductory case

Type of Chain	Chain probabilities	
	ECM(Becker,1980)	WECBM
1-0	$\frac{\alpha^{(2)}}{(\alpha+\beta)^{(2)}}$	$\frac{(\alpha+\beta)\alpha^{(3)}}{\alpha(\alpha+\beta)^{(3)}}$
1-1-0	$\frac{2\alpha^2\beta}{(\alpha+\beta)(\alpha+\beta)^{(2)}}$	$\frac{2(\alpha+\beta)^2\{\alpha^{(2)}\}^2\beta}{\alpha^2(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$
1-2	$\frac{\beta^{(2)}}{(\alpha+\beta)^{(2)}}$	$\frac{(\alpha+\beta)\beta^{(2)}}{(\alpha+\beta)^{(3)}}$
1-1-1	$\frac{2_{\alpha\beta}{}^2}{(\alpha+\beta)(\alpha+\beta)^{(2)}}$	$\frac{2(\alpha+\beta)^2\alpha^{(2)}\beta^2}{\alpha(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$

Table 2. Epidemic chain probabilities for 4-member households with one introductory case

Type of Chain	Chain probabilities	
	ECM(Becker,1980)	WECBM
1-0	$\frac{\alpha^{(3)}}{(\alpha+\beta)^{(3)}}$	$\frac{(\alpha+\beta)\alpha^{(4)}}{\alpha(\alpha+\beta)^{(4)}}$
1-1-0	$\frac{3\{\alpha^{(2)}\}^2\beta}{(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$	$\frac{3(\alpha+\beta)^2\{\alpha^{(3)}\}^2\beta}{\alpha^2(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}}$
1-2-0	$\frac{3\alpha^2\beta^{(2)}}{(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$	$\frac{3(\alpha+\beta)^2\{\alpha^{(2)}\}^2\beta^{(2)}}{\alpha^2(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}}$
1-1-1-0	$\frac{6\alpha^2\alpha^{(2)}\beta^2}{(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}(\alpha+\beta)}$	$\frac{6(\alpha+\beta)^3\alpha^{(3)}\{\alpha^{(2)}\}^2\beta^2}{\alpha^3(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$
1-3	$\frac{\beta^{(3)}}{(\alpha+\beta)^{(3)}}$	$\frac{(\alpha+\beta)\beta^{(3)}}{(\alpha+\beta)^{(4)}}$
1-2-1	$\frac{3_{\alpha\beta\beta}{}^{(2)}}{(\alpha+\beta)(\alpha+\beta)^{(3)}}$	$\frac{3(\alpha+\beta)^2\alpha^{(2)}{}_{\beta\beta}{}^{(2)}}{\alpha(\alpha+\beta)^{(4)}(\alpha+\beta)^{(2)}}$
1-1-2	$\frac{3\alpha^{(2)}{}_{\beta\beta}{}^{(2)}}{(\alpha+\beta)^{(2)}(\alpha+\beta)^{(3)}}$	$\frac{3(\alpha+\beta)^2\alpha^{(3)}{}_{\beta\beta}{}^{(2)}}{\alpha(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}}$
1-1-1-1	$\frac{6_{\alpha\alpha}{}^{(2)}\beta^3}{(\alpha+\beta)(\alpha+\beta)^{(2)}(\alpha+\beta)^{(3)}}$	$\frac{6(\alpha+\beta)^3\alpha^{(3)}\alpha^{(2)}\beta^3}{\alpha^2(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$

Table 3. Epidemic chain probabilities for 5- member households with one introductory case

Type of Chain	Chain probabilities	
	ECM(Becker,1980)	WECBM
1-0	$\frac{\alpha^{(4)}}{(\alpha+\beta)^{(4)}}$	$\frac{(\alpha+\beta)\alpha^{(5)}}{\alpha(\alpha+\beta)^{(5)}}$
1-1-0	$\frac{4\beta\{\alpha^{(3)}\}^2}{(\alpha+\beta)^{(3)}(\alpha+\beta)^{(4)}}$	$\frac{4(\alpha+\beta)^2\{\alpha^{(4)}\}^2\beta}{\alpha^2(\alpha+\beta)^{(5)}(\alpha+\beta)^{(4)}}$
1-2-0	$\frac{6\beta^{(2)}\{\alpha^{(2)}\}^3}{\{(\alpha+\beta)^{(2)}\}^2(\alpha+\beta)^{(4)}}$	$\frac{6(\alpha+\beta)^2\{\alpha^{(3)}\}^2\beta^{(2)}}{\alpha^2(\alpha+\beta)^{(5)}(\alpha+\beta)^{(3)}}$
1-1-1-0	$\frac{12^{\beta^2\alpha^{(3)}\{\alpha^{(2)}\}^2}}{(\alpha+\beta)^{(2)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(4)}}$	$\frac{12^{(\alpha+\beta)^3\alpha^{(4)}\beta^2\{\alpha^{(3)}\}^2}}{\alpha^3(\alpha+\beta)^{(5)}(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}}$
1-3-0	$\frac{4\alpha^4\beta^{(3)}}{(\alpha+\beta)(\alpha+\beta)^{(4)}}$	$\frac{4(\alpha+\beta)^2\{\alpha^{(2)}\}^2\beta^{(3)}}{\alpha^2(\alpha+\beta)^{(5)}(\alpha+\beta)^{(2)}}$
1-1-2-0	$\frac{12^{\alpha^3}{}_{\beta\alpha}{}^{(3)}\beta^{(2)}}{(\alpha+\beta)^2(\alpha+\beta)^{(3)}(\alpha+\beta)^{(4)}}$	$\frac{12^{(\alpha+\beta)^4\alpha^{(4)}\{\alpha^{(2)}\}^3\beta^{(2)}\beta}}{\alpha^4(\alpha+\beta)^{(5)}(\alpha+\beta)^{(4)}\{(\alpha+\beta)^{(2)}\}^2}$
1-2-1-0	$\frac{12^{\alpha^3}{}_{\beta\alpha}{}^{(2)}\beta^{(2)}(2\alpha+\beta+2)}{(\alpha+\beta)^3(\alpha+\beta+1)^2(\alpha+\beta)^{(4)}}$	$\frac{12^{(\alpha+\beta)^3\alpha^{(3)}\{\alpha^{(2)}\}^2\beta^{(2)}\beta}}{\alpha^3(\alpha+\beta)^{(5)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$
1-1-1-1-0	$\frac{24^{\alpha^2\beta^3\alpha^{(2)}\alpha^{(3)}}}{(\alpha+\beta)(\alpha+\beta)^{(2)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(4)}}$	$\frac{24^{(\alpha+\beta)^4\alpha^{(4)}\alpha^{(3)}\{\alpha^{(2)}\}^2\beta^3}}{\alpha^4(\alpha+\beta)^{(5)}(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$
1-4	$\frac{\beta^{(4)}}{(\alpha+\beta)^{(4)}}$	$\frac{(\alpha+\beta)\beta^{(4)}}{(\alpha+\beta)^{(5)}}$

Type of Chain	Chain probabilities	
	ECM(Becker,1980)	WECBM
1-3-1	$\frac{4\alpha\beta^{(3)}}{(\alpha+\beta)^{(4)}}\{1-\alpha^3/(\alpha+\beta)^3\}$	$\frac{4(\alpha+\beta)^2\alpha^{(2)}\beta^{(3)}\beta}{\alpha(\alpha+\beta)^{(5)}(\alpha+\beta)^{(2)}}$
1-1-3	$\frac{4\beta\alpha^{(3)}\beta^{(3)}}{(\alpha+\beta)^{(3)}(\alpha+\beta)^{(4)}}$	$\frac{4(\alpha+\beta)^2\alpha^{(4)}\beta^{(3)}\beta}{\alpha(\alpha+\beta)^{(5)}(\alpha+\beta)^{(4)}}$
1-2-2	$\frac{6\alpha^{(2)}\beta^{(2)}}{(\alpha+\beta)^{(4)}}\left[1-\frac{2\alpha^2}{(\alpha+\beta)^2}+\left\{\frac{\alpha^{(2)}}{(\alpha+\beta)^{(2)}}\right\}^2\right]$	$\frac{6(\alpha+\beta)^2\alpha^{(3)}\{\beta^{(2)}\}^2}{\alpha(\alpha+\beta)^{(5)}(\alpha+\beta)^{(3)}}$
1-2-1-1	$\frac{12\alpha^2\beta^2\alpha^{(2)}\beta^{(2)}(2\alpha+\beta+2)}{(\alpha+\beta)^3(\alpha+\beta+1)^2(\alpha+\beta)^{(4)}}$	$\frac{12(\alpha+\beta)^3\alpha^{(3)}\beta^{(2)}\beta^2}{\alpha^2(\alpha+\beta)^{(5)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$
1-1-2-1	$\frac{12\alpha\beta^2\alpha^{(3)}\beta^{(2)}(2\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta)^{(3)}(\alpha+\beta)^{(4)}}$	$\frac{12(\alpha+\beta)^3\alpha^{(4)}\alpha^{(2)}\beta^{(2)}\beta^2}{\alpha^2(\alpha+\beta)^{(5)}(\alpha+\beta)^{(4)}(\alpha+\beta)^{(2)}}$
1-1-1-2	$\frac{12\beta^2\alpha^{(2)}\alpha^{(3)}\beta^{(2)}}{(\alpha+\beta)^{(2)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(4)}}$	$\frac{12(\alpha+\beta)^3\alpha^{(4)}\alpha^{(3)}\beta^{(2)}\beta^2}{\alpha^2(\alpha+\beta)^{(5)}(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}}$
1-1-1-1-1	$\frac{24\alpha\beta^4\alpha^{(2)}\alpha^{(3)}}{(\alpha+\beta)(\alpha+\beta)^{(2)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(4)}}$	$\frac{24(\alpha+\beta)^4\alpha^{(4)}\alpha^{(3)}\alpha^{(2)}\beta^4}{\alpha^3(\alpha+\beta)^{(5)}(\alpha+\beta)^{(4)}(\alpha+\beta)^{(3)}(\alpha+\beta)^{(2)}}$

The number of possible cases, i.e., the combination of the epidemic chains for the above three tables were calculated by using the formula Eq. (1) and further the probabilities were calculated by using the Eq. (12).

It can be viewed from the above tables that, the expressions for the chain probabilities of the WECBM are a bit complicated as compared to the Becker's ECM. It gives us an idea about the complicacy of WECBM than that of Becker's ECM.

5.2 Method of Estimation

The parameter estimate of both Becker's ECM (general and with $\beta = 1$) and WECBM, $\hat{\alpha}$ and $\hat{\beta}$ has been calculated by using the method of maximum likelihood estimation (MLE) for multinomial distribution approach as described below:

Let there be k chains of infections for a given size of household. Since the chains are mutually exclusive they can be assumed to follow a multinomial distribution. Suppose $C_1, C_2, \ldots, C_k$ be k mutually exclusive and exhaustive chains of infection with respective probabilities $p_1, p_2, \ldots, p_k$. Here, each p_i , $i=1, 2, \ldots, k$ is in turn a function of α and β as given in Table 1, Table 2 and Table 3.

The probability that C_1 occurs x_1 times, C_2 occurs x_2 times,…, C_k occurs x_k times in n independent observations, is given by

$$p(x_1, x_2, \ldots, x_k) = Bp_1{}^{x_1}p_2{}^{x_2}\ldots p_k{}^{x_k}, \tag{15}$$

where $\sum x_i = n$ and B is the number of permutation of the chains of infection $C_1, C_2, \ldots, C_k$.

To determine B, it is required to find the number of permutations of n objects of which x_1 are of one kind, x_2 of another kind,…, x_k of the k^{th} kind, which is given by,

$$B = \frac{n!}{x_1!x_2!\ldots x_k!} \tag{16}$$

$$\text{Hence,} p(x_1, x_2, \ldots, x_k) = \frac{n!}{\prod_{i=1}^{k} x_i!}\prod_{i=1}^{k} p_i{}^{x_i}, \sum_{i=1}^{k} x_i = n \tag{17}$$

which is the required probability function of the multinomial distribution. It is so called since Eq. (17) is the general term in the multinomial expansion,

$$(p_1+p_2+\cdots+p_k)^n, \sum_{i=1}^{k} p_i = 1$$

Since, the total probability is 1, so

$$\textstyle\sum_x p(x) = \sum_x \left[\frac{n!}{x_1!x_2!\ldots x_k!} p_1{}^{x_1} p_2{}^{x_2} \ldots p_k{}^{x_k}\right] = (p_1+p_2+\cdots+p_k)^n = 1 \qquad (18)$$

For MLE, the likelihood function is given by,

$$logL(x,p) = logB + \textstyle\sum_{i=1}^{k} x_i logp_i \qquad (19)$$

where$B = \frac{n!}{x_1!x_2!\ldots x_k!}$. Since, p_i , $i=1, 2,\ldots,k$ are functions of α and β , the ML estimates of α and β are obtained by the maximum likelihood equations

$$\frac{\partial}{\partial\alpha}\log^{L} = 0 \qquad (20)$$

and

$$\frac{\partial}{\partial\beta}\log^{L} = 0 \qquad (21)$$

6. Applications to Epidemic Data

In this section, the applications of WECBM and Becker's ECM with one introductory case to both sets of epidemic data are shown in detail. R-programming is used to calculate the estimates of the two parameters of both models (by the MLE method described in Section 5.2).

In order to test the adequacy of these models, the using the usual chi-square goodness of fit has been used. Pooling was done for all the chains having expected frequencies of less than five.

Let, H_0: there is no significant difference between the observed and expected values of the distribution.

As an application, the WECBM is fitted to epidemic data quoted in Section 3.1 and reproduced in Table 4. The value of the estimated parameters for the WECBM is calculated to be $\hat{\alpha} = 13.59598$ and $\hat{\beta} = 2.018984$ respectively.

Table 4. Fitting of distribution to Heasman-Reid data(1961) for 5- member households with one introductory case

Type of chain	Observed Frequency	Expected Frequency			
		Reed –Frost model (Heasman& Reid 1961)	Epidemic chain model (Becker 1980)		WECBM
			General	β =1	
1-0	423	409.9	410.5	435.7	413.2
1-1-0	131	146.2	141.5	117.7	131.3
1-2-0	24	24.8	27.3	32.0	40.2
1-1-1-0	36	44.2	42.0	29.0	37.3
1-3-0	3	2.3	3.5	8.2	7.8
1-1-2-0	8	5.6	6.2	6.6	8.2
1-2-1-0	11	6.8	13.3	13.9	9.5
1-1-1-1-0	14	10.1	9.6	5.9	8.7
1-4	0	0.1	0.2	1.5	0.9
1-3-1	0	1.1	1.6	3.6	1.1
1-1-3	2	0.4	0.5	1.1	1
1-2-2	1	1.8	2.3	3.3	1.1
1-2-1-1	3	6.7	1.7	1.8	0.2
1-1-2-1	2	1.6	1.8	1.9	1.2
1-1-1-2	2	0.9	0.8	0.9	1
1-1-1-1-1	4	1.5	1.2	0.9	1.3
Total	664	664	664	664	664
χ^2 (Chi-square)		12.6(7df)	6.1(5 df)	20.9(7df)	18.159 (5df)

χ^2 tabulated for 5df at 95% level of significance=1.145

χ^2 tabulated for 7df at 95% level of significance=2.167

Table 4 gives the results of fitting of different models viz., Reed-Frost model, Becker's ECM (general and with $\beta = 1$), and WECBM to Heasman- Reid data. The observed and expected frequencies for various types of chains, the value of χ^2 measure for goodness of fit along with the df are shown in the table. In a five member household with one introductory case there are 16(sixteen) possible chains observed over 664 households. Although all the fitted values are found to be non-significant, it can be found that Becker's ECM (general) gives the best fit to the observed data amongst the four models given in the table. The Reed-Frost model is found to give the second best fit to the data. Restricting the parameter of Becker's ECM to $\beta = 1$ reduces its efficiency to fit the data to a huge extent. The WECBM also appears to perform not very satisfactorily in estimating the distribution of the number of chains of infections in a five member household with a single introductory case although it outperforms Becker's ECM with $\beta = 1$.

Again, for comparison of the three models, by applying the current epidemic data for four and five member households with one introductory case, the expected values of the WECBM, Becker's ECM (general and with $\beta = 1$) as given in Table 5 and Table 6 has been computed.

Table 5. Fitting of distribution to current epidemic data (2016) for a 4-member household with one introductory case

Type of chain	Observed Frequency	Expected Frequency		
		Epidemic chain model		Weighted epidemic chain binomial model
		General	β =1	
1-0	130	137.12	143.79	136.99
1-1-0	53	45.26	38.45	45.47
1-2-0	16	9.44	11.42	9.33
1-1-1-0	10	11.81	8.77	11.89
1-3	0	0.79	1.99	0.76
1-2-1	0	1.27	1.53	1.26
1-1-2	0	1.47	1.73	1.45
1-1-1-1	0	1.84	1.32	1.85
Total	209	209	209	209
χ^2 (Chi-square)		11.899(2 df)	15.408(3 df)	11.992(2 df)
Parameter estimates $\hat{\alpha}$		31.860242	6.615281	34.919342
$\hat{\beta}$		4.953511	-	5.578844

χ^2 tabulated for 2df at 95% level of significance=0.103

χ^2 tabulated for 3df at 95% level of significance=0.352

Table 6. Fitting of distribution to current epidemic data (2016) for 5- member household with one introductory case

Type of chain	Observed Frequency	Expected Frequency		
		Epidemic chain model		Weighted epidemic chain binomial model
		General	β =1	
1-0	64	71.97	78.06	70.76
1-1-0	30	25.84	21.09	24.95
1-2-0	10	5.00	5.75	6.74
1-1-1-0	13	7.99	5.21	8.24
1-3-0	0	0.62	1.47	0.87
1-1-2-0	1	1.19	1.19	1.45
1-2-1-0	0	2.53	2.49	1.65
1-1-1-1-0	1	1.90	1.07	2.01
1-4	0	0.04	0.28	0.25
1-3-1	0	0.29	0.66	0.12
1-1-3	0	0.08	0.20	0.09
1-2-2	0	0.44	0.59	0.38
1-2-1-1	0	0.35	0.33	0.32
1-1-2-1	0	0.35	0.33	0.20
1-1-1-2	0	0.16	0.16	0.18
1-1-1-1-1	0	0.25	0.12	0.79
Total	119	119	119	119
χ^2 (Chi-square)		14.382(2df)	26.426(3df)	10.786(2df)
Parameter estimates $\hat{\alpha}$		70.095296	7.627815	65.458672
$\hat{\beta}$		9.587776	-	9.426828

χ^2 tabulated for 2df at 95% level of significance=0.103

χ^2 tabulated for 3df at 95% level of significance=0.352

Table 5 and Table 6 give the results of fitting the current epidemic data to three different models viz., Becker's ECM (general and with $\beta =1$) and WECBM. The observed and expected frequencies for various types of chains, the value of χ^2 measure for goodness of fit along with the df are shown in the tables. In a four and five member household with one introductory case, there are 8(eight) and 16(sixteen) possible chains observed over 209 and 119 households, respectively.

From Table 5, it has been observed that WECBM and Becker's ECM (general) give more or less similar levels of best fit to the data. And, as compared to the other two models, Becker's ECM with $\beta =1$ is found to give a good fit to the data.

From Table 6, it has been observed that WECBM gives the best fit to the observed data amongst the three models given in the table. Becker's ECM (general) is found to give the second best fit to the data. Similar to that of the application using Heasman – Reid data, Becker's ECM with $\beta =1$ reduces its efficiency to fit the data to a huge extent as compared to the other two models.

7. Conclusion

The present study is an attempt to develop a solution to the difficulty of epidemic processes and also to study the pattern of the spread of IDs using some real-life ID data or epidemic data. For this purpose, the WECBM was developed to simply provide an alternative approach to Becker's ECM (1980)[7]. But as compared to Becker's model it appears to be a complicated model. All the expressions for chain probabilities worked out for households with sizes three, four, and five having one introductory case in a closed population were found to be complicated ones.

This study facilitates us to draw a more in-depth inference of the theory so developed for WECBM. In such applications, a better result can always be expected with a large and standard set of sample data. Due to the limited availability of data and resources for the study, like manpower, money, and time, a larger set of primary data could not be collected and the survey and analysis were restricted to 600 sample data only.

Acknowledgments

In the first place, I am expressing my deep sense of gratitude with reverence to Prof. Dilip Chandra Nath, Vice Chancellor, Administration, Assam University, Silchar, and Supervisor of my research for his initiative, benevolence, endurance, constructive criticism, encouragement, and constant monitoring during the period of my investigation and also in the preparation of this paper as a part of my thesis. I am also thankful to Prof. Kishore K. Das, Department of Statistics, Gauhati University, Guwahati for his advice, help, and contribution which indeed strengthened the work.

Also, I am extending my sincere thanks to University Grants Commission, New Delhi, under UGC-BSR for their financial support provided for the data collection for the study through the one time grant(No. F 19-145/2015(BSR)) provided to my supervisor.

References

Bailey, N. T. J. (1953). The use of chain-binomials with a variable chance of infection for the analysis of intra-household epidemics. *Biometrika, 40*, 177-185.

Bailey, N. T. J. (1957). *The mathematical theory of epidemics* (Chapter 6, pp. 75-108). Charles Griffin & Co., Ltd., London.

Bailey, N. T. J. (1958). A perturbation approximation to the simple stochastic epidemic in a large population. *Biometrika, 55*, 199-210.

Bailey, N. T. J. (1964). *The elements of stochastic processes with applications to the natural sciences*. Wiley, New York.

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its application*. Griffin, London.

Becker, N. (1977). Estimation for discrete time branching processes with applications to epidemics. *Biometrics, 33*, 515-522.

Becker, N. (1980). An epidemic chain model. *Biometrics, 36*(2), 249-254.

Becker, N. (1981). A general chain binomial model for infectious diseases. *Biometrics, 37*(2), 251-258.

Cairoli, L. H. (1988). Chain binomial epidemic models. An abstract of Master's report. Department of Statistics, Kansas State University, Manhattan, KS.

Heasman, H. A., & Reid, D. D. (1961). Theory and observation in family epidemics of the common cold. *British Journal of Preventive and Social Medicine, 15*, 12-16.

Ludwig, D. (1975). Final size distributions for epidemics. *Mathematical Biosciences, 23*, 33-46.

Nath, D. C., Das, K. K., & Chakraborty, T. (2016). A modified epidemic chain binomial model. *Open Journal of Statistics, 6*(1).

Nath, D. C., Das, K. K., & Chakraborty, T. (2017). A modified epidemic chain binomial model (MECBM) and its 2, 3-introductory probabilities. *Open Journal of Statistics, 7*, 225-239.

Ranked Set Sampling: An Estimation of Infant Mortality Using Bayesian Method

Abstract

The technique of Ranked set sampling (RSS) in the estimation of the probability of occurrence of a dichotomous event in a population is found to be effective and reliable. In this work, the superiority of RSS has been discussed for situations where the probability of occurrence, say p, of an event is not fixed but a random quantity, by using the method of Bayesian estimation. The closeness of the estimates thereof obtained through maximum likelihood procedures (when p is assumed as a fixed quantity) and Bayesian estimation in the sample design of the SRS and RSS are evaluated using Pitman closeness criteria. The performance of the proposed procedure has been illustrated through numerical simulation as well as in the estimation of the probability of infant death in India using the real-life demographical data from National Family Health Survey-III(2005-06).

Keywords: Bayes estimator, Pitman closeness criteria, square error loss, risk function, and infant deaths

1. Introduction

McIntyre's[1] proposed method of Ranked Set sampling (RSS) is referred to as a suitable alternative of the sampling procedure for situations where obtaining information is difficult in terms of the cost incurred and time required during collection. In studies [2-6] where the population of concern is dichotomous and the probability of occurrence of an event have discussed that the estimators obtained through the design of RSS, which is found to be comparatively more efficient and reliable inference than better known simple random sampling (SRS). In a few recent works [7-11], the performance of estimators of p based on the ranked set sample for situations where p is assumed to be an unknown and random quantity, have discussed.

In existing literature [7-11], more emphasis has been given to the estimation of p under RSS, which is based on the assumption that the parameter p is an unknown but a fixed quantity. In practical situations, some prior information about p and p is treated as an unknown and random quantity. The present work is an attempt to highlight the performance of a Bayesian estimate of p based on the ranked set sample, and comparing the closeness of the estimates obtained through the procedures of maximum likelihood(ML) under both SRS and RSS through Pitman nearness criteria.

We organize the chapter in the following way. In section 2, both the classical version estimator based on the ML principle and the Bayes estimator of the population proportion have been discussed. Section 3 discussed the Pitman Closeness Criterion for comparison of the estimators in terms of risks obtained through both SRS and RSS procedures. In section 4 the proposed procedure is used to estimate the probability of infant death in India using real-life demographical data from National Family Health Survey-III. Lastly, section 5 gives a brief concluding remark.

2. Estimation of Parameters

2.1 Maximum Likelihood Estimation

Consider a population where the variable of interest (X) is dichotomous then there will be two possible outcomes, success or event occurred (denoted as 1) and failure or non-occurrence of event (denoted as 0). Thus, X follows Bernoulli (p), where p denotes the probability of occurrence of an event or proportion of occurrence of an event. Let $X_1, X_2, \cdots X_n$ constitutes a random sample of size n and each $X_i; i = 1(1)n$ follows independent and identically (iid) Bernoulli (p), where p is an unknown parameter that lies between $(0,1)$.

For $r = 1(1)s$, a sample of m sets of size, s, observed from the dichotomous population and are ranked within each set. Classification (ranking) within each set can be the result of judgment ranking or through concomitant variables. The r^{th} smallest ranked item in each of the m sets are quantified to be either 1 ("occurrence of an event") or 0 ("non-occurrence") for $r = 1,2, \cdots, s$. Let $X_{[r]}$ represents the class of r^{th} judgment order statistics and $X_{j[r]}$ denote the j^{th} observation from that class, then this sampling scheme yields the ranked set sample denoted as, $X_{j[r]}$, for $r = 1,2, \cdots, s$ and $j = 1,2, \cdots, m$, under the assumption that the judgemental identification of ranks is perfect and is done with negligible cost. The $X_{j[r]}$'s constitutes a typical balanced ranked set sample of size $n = ms$ and can be represented as:

$$X_{1[1]}, X_{1[2]}, \cdots, X_{1[r]}, \cdots, X_{1[s]}$$
$$X_{2[1]}, X_{2[2]}, \cdots, X_{2[r]}, \cdots, X_{2[s]}$$
$$\cdots$$

$$X_{k[1]}, X_{k[2]}, \cdots, X_{k[r]}, \cdots, X_{k[s]}$$
$$\cdots$$
$$X_{m[1]}, X_{m[2]}, \cdots, X_{m[r]}, \cdots, X_{m[s]}.$$

Corresponding to each r^{th} class, due to the nature of RSS, $X_{1[r]}, X_{2[r]}, \cdots, X_{m[r]}$, constitutes independent and identical Bernoulli random variables with probability of occurrence of the event, say $\pi_{[r]}$, for all $r = 1,2,\cdots,s$, where $\pi_{[r]}$ is given by [12]

$$\pi_{[r]} = P(X_{j[r]} = 1) = \sum_{l=s-r+1}^{s} \frac{s!}{(s-l)!\,(l!)} p^l (1-p)^{(s-l)}; r = 1,2,\cdots,s(1)$$

The ML estimator of the population parameter p using the data $\boldsymbol{X} = (X_{j[r]})_{jr}; j = 1(1)m, r = 1(1)s$, is one of the possible estimation procedures and is given by Terpstra (2004). The natural unbiased estimator and the variance for p based on ranked set sample[10] can be given by

$$\hat{p}_{RSS} = \sum_{r=1}^{s} \frac{\pi_{[r]}}{s} = \frac{1}{ms} \sum_{r=1}^{s} \sum_{j=1}^{m} x_{j[r]}$$

$$V(\hat{p}_{RSS}) = V\left(\sum_{r=1}^{s} \frac{\pi_{[r]}}{s}\right) = \frac{1}{s^2 m} \sum_{r=1}^{s} \pi_{[r]} (1 - \pi_{[r]}) \tag{2}$$

If the same samples are assumed to be obtained from SRS procedure and let it be symbolized as $Z_i; i = 1(1)n$, where $n = ms$, be the independent and identically Bernoulli distributed random variables having occurrence probability p, then MLE of the p and its variance will be given by

$$\hat{p}_{SRS} = \sum_{i=1}^{n} \frac{z_i}{n}$$

$$V(\hat{p}_{SRS}) = \frac{\hat{p}_{SRS} * (1 - \hat{p}_{SRS})}{n} \tag{3}$$

Here, we have opted for the alternative method based on the sampling structure for Bayesian estimation of the population proportion (p). In next section, a comparison among estimators obtained by MLE and Bayes using SRS and RSS procedures is discussed.

2.2 Bayesian Estimation

2.2.1 Under RSS

Under the assumption that the proportion parameter p is a random variable and can be explained through a suitable prior density, say $\tau(p)$ of p defined over the interval [0,1]. An attempt has been made to derive a Bayesian estimator of p by utilizing the available prior information $\tau(p)$ and sample information gathered through RSS methodology. Let the observations obtained following the RSS procedure is denoted as, $\boldsymbol{X}$=$(\boldsymbol{X}_{[1]}, \cdots, \boldsymbol{X}_{[s]})'$, where $\boldsymbol{X}_{[r]} = (X_{1[r]}, X_{2[r]}, \cdots, X_{m[r]})$. Corresponding to each r^{th} class, $X_{j[r]}$s are independent and identical Bernoulli ($\pi_{[r]}$) variate, for all $j = 1(1)m$. Let us define the variables

$$Y_r = \sum_{j=1}^{m} X_{[r]j}, \forall r = 1(1)s.$$

By the virtue of ranked set sampling, the variables $Y_1, Y_2, \cdots, Y_s$ are independently distributed as $Y_r \sim$ Binomial $(m, \pi_{[r]})$ and is such that

$$\frac{1}{s} \sum_{r=1}^{s} \pi_{[r]} = p. \tag{4}$$

Estimation of the p under Bayesian paradigm can segregated based on the definition of $\pi_{[r]}$ as:

Case I: $\boldsymbol{\pi}_{[r]}$ is a function of $\boldsymbol{p}$

For each r, $1 \leq r \leq s$, $\pi_{[r]}$ is a function of the basic parameter p, so we denote it as $\pi_{[r]}(p)$. The likelihood function of $L^R(p|\boldsymbol{x})$ will be given by

$$L^R(p|\boldsymbol{x}) = \prod_{r=1}^{s}\prod_{j=1}^{m} P(x_{j[r]}|\pi_{[r]}(p)) = \prod_{r=1}^{s} \pi_{[r]}(p)^{\sum_{j=1}^{m} x_{j[r]}}(1-\pi_{[r]}(p))^{m-\sum_{j=1}^{m} x_{j[r]}}. \tag{5}$$

where $0 \leq \pi_{[r]} \leq 1$ and $y_r = \sum_{j=1}^{m} x_{j[r]}$ for all $r = 1(1)s$. Suppose prior density for p is Beta distribution and is of the form

$$\tau(p) = \frac{1}{\mathcal{B}(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \;; 0 < p < 1, ; \alpha > 0;\ \beta > 0. \tag{6}$$

The posterior density of p, given $\boldsymbol{Y} = \boldsymbol{y}$, with respect to the prior $\tau(p)$ for p is given by

$$f(p|\boldsymbol{y}) = \frac{L^R(p|\boldsymbol{y})\tau(p)}{\int_0^1 L^R(p|\boldsymbol{y})\tau(p)dp}$$

$$\Leftrightarrow f(p|\boldsymbol{y}) \propto L^R(p|\boldsymbol{y})\tau(p)$$

$$\Leftrightarrow f(p|\boldsymbol{y}) \propto \left[\prod_{r=1}^{s} [p_{[r]}(p)]^{y_r}[1-p_{[r]}(p)]^{(m-y_r)}\right] p^{\alpha-1}(1-p)^{\beta-1}. \tag{7}$$

The derived form of the posterior distribution is not explicitly, therefore Monte Carlo simulation technique has been adopted for characterization. For a sufficiently large number of replications, say N observations have been randomly drawn from the posterior distribution $f(p|\boldsymbol{y})$ and let it be denoted as $p^{(1)}, p^{(2)}, \cdots, p^{(N)}$. Then the posterior mean and variance of p can be approximated as

$$E(p|\boldsymbol{y}) = \int_0^1 pf(p|\boldsymbol{y})dp \simeq \frac{1}{N}\sum_{j=1}^{N} p^{(j)} \tag{8}$$

and

$$V(p|\boldsymbol{y}) = \int_0^1 \{p - E(p|\boldsymbol{y})\}^2 f(p|\boldsymbol{y})dp \simeq \frac{1}{N}\sum_{j=1}^{N} [p^{(j)}]^2 - \left[\frac{1}{N}\sum_{j=1}^{N} p^{(j)}\right]^2 \tag{9}$$

Thus, under square error loss function the Bayes estimate of p with respect to the prior $\tau(p)$ has obtained as

$$\hat{p}_{RSS}^{B} \simeq \frac{1}{N}\sum_{j=1}^{N} p^{(j)}. \tag{10}$$

2.2.2 Case II- $\boldsymbol{\pi}_{[r]}$ Is Independent and Identically Distributed

According to the nature of the RSS, $Y_1, Y_2, \cdots, Y_s$ are independently distributed as $Y_r|\pi_{[r]} \sim f(y_r|\pi_{[r]})$ for each $r = 1,2,\cdots,s$. So, in order to find estimate of the parameter of interest p and satisfying the relation (2.2), it is assumed that $\pi_{[r]}$'s are independent and identically distributed with a common prior density, say $\tau(.)$, which is defined over the compact set [0,1]. The posterior density of $\pi_{[r]}$, given $Y_r = y_r$, with respect to the prior density $\tau(p) = \tau(\pi_{[r]})$, for all $r = 1,2,\cdots s$ is given by

$$h(\pi_{[r]}|y_r) \propto L^R(\pi_{[r]}|\boldsymbol{x_r})\tau(\pi_{[r]})$$

$$\Leftrightarrow h(\pi_{[r]}|y_r) \propto \left[\prod_{j=1}^{m} P(x_{j[r]}|\pi_{[r]})\right] \pi_{[r]}^{\alpha-1}(1-\pi_{[r]})^{\beta-1},$$

$$\Leftrightarrow h(\pi_{[r]}|y_r) \propto \pi_{[r]}^{y_r+\alpha-1}(1-\pi_{[r]})^{m-y_r+\beta-1}. \tag{11}$$

Bayes estimator of $\pi_{[r]}$ has been obtained using squared error loss function as

$$\hat{\pi}_{[r]}^{*B} = \frac{Y_r + \alpha}{m + \alpha + \beta}, \qquad forr = 1, ..s. \qquad (12)$$

After having the estimators $\hat{\pi}_{[r]}^{*B}; r = 1,2,\cdots,s$, by virtue of the relation (**2.2**), a Bayesian estimator of p as

$$\hat{p}_{RSS}^{*B} = \frac{1}{s}\sum_{r=1}^{s} \hat{\pi}_{[r]}^{*B}(13)$$

2.2.3 Under SRS

The prior distribution of the parameter of p of the data $\boldsymbol{Z}$ obtained using SRS, is assumed to follow the same Beta distribution with parameters α, β, then under this setup the posterior of $(p|\boldsymbol{z})$ will be given by

$$g(p|\boldsymbol{y}) \sim Beta\left(\sum_{i=1}^{n} z_i + \alpha, n - \sum_{i=1}^{n} z_i + \beta\right). \quad (14)$$

Using the equation-(14) the posterior mean and variance of $(p|\boldsymbol{z})$ will be given by

$$\hat{p}_{SRS}^{B} = E(p|\boldsymbol{z}) = \frac{\sum_{i=1}^{n} z_i + \alpha}{n + \alpha + \beta} \qquad (15)$$

$$V(p|\boldsymbol{z}) = \frac{(\sum_{i=1}^{n} z_i + \alpha)(n - \sum_{i=1}^{n} z_i + \beta)}{(n + \alpha + \beta)^2(n + \alpha + \beta + 1)} \qquad (16)$$

3. Pitman Closeness Criterion for Comparison

The goal of this section is to compare the estimators of p derived in the previous section. To estimate an ML estimator no of prior specification is required and it does not involve any particular loss function, as converse of Bayesian method. Therefore, Thus, mean square error (MSE) corresponding to both ML and Bayes' estimators has been calculated, as the MSE of an estimator can be regarded as a risk function under squared error loss and can be used for comparison purposes [7]. Under square error loss the risk of $\hat{\theta}_B$ for parameter θ is given by

$$R_{\hat{\theta}_B}(\theta) = E(\hat{\theta}_B - \theta)^2. \qquad (17)$$

On the other hand, the risk of $\hat{\theta}_M$ has a theoretical expression obtained as

$$R_{\hat{\theta}_M}(\theta) = E(\hat{\theta}_M - \theta)^2. \qquad (18)$$

To compare the closeness of the estimates of the parameters obtained for this MLE and Bayesian estimation methods based on samples of SRS and RSS, the concept of Pitman measure of closeness or Pitman closeness[13] is adopted. Pitman's measure of closeness is a probability that measures the frequency with which one estimator is closer to the value of a parameter than another competing estimator within the same class of estimators.

$\boldsymbol{Definition} -$ Let $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators of the parameter θ an estimator $\hat{\theta}_1$ will be said to be Pitman closer (to θ) than another estimator $\hat{\theta}_2$ for all $\theta \in \Theta$ if

$$P(|\hat{\theta}_1 - \theta| > |\hat{\theta}_2 - \theta|) > 0.5.$$

The necessary condition[14] to show that $\hat{\theta}_1$ is more peak (Pitman closer)to θ than $\hat{\theta}_2$ is that

$$V(\hat{\theta}_1) \leq V(\hat{\theta}_2) \qquad (19)$$

$$\Leftrightarrow R_{\hat{\theta}_1}(\theta) \leq R_{\hat{\theta}_2}(\theta) \qquad (20)$$

Following comparison among the estimators in general is possible to evaluate the method of estimation (MLE and Bayes) based on sampling techniques (SRS and RSS):

$\boldsymbol{Case - 1}$: Let $R_{\hat{p}_{SRS}}(p)$ and $R_{\hat{p}_{RSS}}(p)$ denotes the risks under SRS and RSS, respectively then

$$R_{\hat{p}_{RSS}}(p) \leq R_{\hat{p}_{SRS}}(p).$$

$\boldsymbol{Proof}$: The risks under SRS and RSS are $R_{\hat{p}_{SRS}}(p)$ and $R_{\hat{p}_{RSS}}(p)$, respectively, and can be written as

$$R_{\hat{p}_{RSS}}(p) = \frac{1}{ms^2}\sum_{r=1}^{s} \hat{\pi}_{[r]}\left(1 - \hat{\pi}_{[r]}\right) = \frac{1}{ns}\sum_{r=1}^{s} \hat{\pi}_{[r]}\left(1 - \hat{\pi}_{[r]}\right) \qquad (21)$$

Since, $\pi_{[r]}$'s, for all $r = 1,2,\cdots,s$ is a non-decreasing sequence and subsequently, among two sequences $\pi_{[r]}$ and

$(1-\pi_{[r]})$, one is non-decreasing and the other is non-increasing. So, from Chebyshev's inequality for $\pi_{[r]}$'s and $(1-\pi_{[r]})$'s we have

$$\frac{1}{s}\sum_{r=1}^{s}\pi_{[r]}(1-\pi_{[r]}) \le \left\{\frac{1}{s}\sum_{r=1}^{s}\pi_{[r]}\right\}\left\{\frac{1}{s}\sum_{r=1}^{s}(1-\pi_{[r]})\right\}.$$

As we know that the variance of $\hat{p}$ in SRS is $\frac{p(1-p)}{n}$, the required justification follows from the fact that

$$\sum_{r=1}^{s}\pi_{[r]} = sp, \qquad \sum_{i=1}^{s}(1-\pi_{[r]}) = s(1-p).$$

$\boldsymbol{Case-2}$: Relationship among risks of ML and Bayes estimators, $R_{\hat{p}_{SRS}}(p)$ and $R_{\hat{p}^B_{SRS}}(p)$ under SRS

$$\begin{aligned}
R_{\hat{p}^B_{SRS}}(p) &= E(\hat{p}^B(SRS)-p)^2 = E\left(\frac{n\bar{Z}+\alpha}{n+\alpha+\beta}-p\right)^2\\
&= (n+\alpha+\beta)^{-2}[n^2E(\bar{Z}-p)^2+(\alpha-p(\alpha+\beta))^2]\\
&= (n+\alpha+\beta)^{-2}\left[n^2\frac{p(1-p)}{n}+\{\alpha-p(\alpha+\beta)\}^2\right]\\
&= (n+\alpha+\beta)^{-2}\left[n^2R_{\hat{p}_{SRS}}(p)+\{\alpha-p(\alpha+\beta)\}^2\right]\\
&= \left[1+\frac{\alpha+\beta}{n}\right]^{-2}\left[R_{\hat{p}_{SRS}}(p)+\left\{\frac{\alpha-p(\alpha+\beta)}{n}\right\}^2\right] \quad (22)
\end{aligned}$$

$\boldsymbol{Case-3}$: Relationship among risks of ML and Bayes estimators, $R_{\hat{p}_{RSS}}(p)$ and $R_{\hat{p}^{*B}_{RSS}}(p)$under RSS

$$\begin{aligned}
R_{\hat{p}^{*B}_{RSS}}(p) &= E(\hat{p}^{*B}_{RSS}-p)^2 = E\left(\frac{m\bar{X}+\alpha}{m+\alpha+\beta}-p\right)^2\\
&= (m+\alpha+\beta)^{-2}E\{(m\bar{X}-mp)+\alpha-p(\alpha+\beta)\}^2\\
&= (m+\alpha+\beta)^{-2}\left[m^2R_{\hat{p}_{SRS}}(p)+\{\alpha-p(\alpha+\beta)\}^2\right]\\
&= \left[1+\frac{\alpha+\beta}{m}\right]^{-2}\left[R_{\hat{p}_{RSS}}(p)+\left\{\frac{\alpha-p(\alpha+\beta)}{m}\right\}^2\right]. \qquad (23)
\end{aligned}$$

Note: Under square error loss function the risk of $\hat{p}^B_{RSS}$ is given as

$$R_{\hat{p}^B_{RSS}}(p) = E(\hat{p}^B_{RSS}-p)^2, \qquad (24)$$

that cannot be further simplified analytically.

To compare the risks among the ML and Bayes estimators obtained following SRS *viz.*,$\hat{p}_{SRS}, \hat{p}^B_{SRS}$ and RSS procedures $\hat{p}_{RSS}, \hat{p}^B_{RSS}, \hat{p}^{*B}_{RSS}$, illustration has been proposed numerically through a simulation study. For the numerical computation we have considered, in particular, $(s,m) = (4,25), (6,50)$ and Beta distribution parameters $(\alpha,\beta) = \left(\frac{1}{2},\frac{1}{2}\right)$, (2,2), (3,5), (5,3). The computation of the risk values, $R_{\hat{p}^{*B}_{RSS}}(p)$, for the Bayes estimator $\hat{p}^{*B}_{RSS}$ under RSS, has obtained by following the simulation technique through Metropolis-Hasting's algorithm and then plotted along with other risks for ML and Bayes, under SRS and RSS, over the whole range of $p \in [0,1]$. The plotted risk functions have been presented in Figure 1 of the Appendix section. Obtained figures have shown that the risk curves corresponding to the ML estimator based on RSS, $\hat{p}_{RSS}$, is lying below the risk curve of both ML and Bayes estimators, $\hat{p}_{SRS}$ and $\hat{p}^B_{SRS}$, which are based on simple random samples. It has also depicted that the risk curves corresponding to Bayes estimators $\hat{p}^B_{RSS}$ and $\hat{p}^{*B}_{RSS}$ are completely lie below the risk curve of $\hat{p}_{RSS}$ based on the ranked set sample, implying that the Bayes estimators are uniformly better than the other proposed estimators. We also observe that the risk curves based on Bayes estimators $\hat{p}^B_{RSS}$ and $\hat{p}^{*B}_{RSS}$ are not significantly different and can conclude that both of the Bayes estimators based on ranked set samples are more or less equally good. For the given parametric combinations of α and β are concerned, it is found that both Bayes' estimators $\hat{p}^B_{RSS}$ and $\hat{p}^{*B}_{RSS}$ are more pitman close to p than any other estimator.

4. Illustration with Real-life Data

In public health related studies, infant survival is regarded as an important indicator that defines the health status of a country or State. Here, our objective is to illustrate the performance of proposed procedures (ML and Bayes), and their properties while estimating the probability of infant death, under both SRS and RSS, in India. For the present study, the database of Demographic and Health Surveys (DHS) has been utilized. DHS provides national and state estimates of various demographic measures that help in different family planning. Here, is the data set of National Family Health Survey-III (NFHS-III) of India for the year 2005-06 of a few selected states from different areas of India $viz.$ Bihar (East), Assam (northeast), Rajasthan (north), Madhya Pradesh (MP) (central), and Orissa (East) have been considered. And, according to the NFHS-III report[15] these selected states have experienced comparatively high infant death rates than other states in the region. Those children who were born in a specific period (2001-2005) of five years prior to the period of the NFHS-III survey have been considered as our population of interest.

Earlier works[16-17] suggested that the survival of a child is positively associated with the age of the mother, and lowering the mother's age while childbearing would lower the survival chance of the child. Therefore, the age of the mother (in months) has been used for ranking purposes in ranked set sampling. The following steps have been followed to obtain the samples through the ranked set sampling principle [2]:

- A simple random sample of s^2 units are drawn from the study population and partitioned randomly into s sets each having s units.
- In each of s non-overlapping sets according to the mother's age the units were ranked. In the case of ties, the observations are ordered systematically in the sequence.[3]
- From the first set, the unit corresponding to the mother with the lowest age is selected. From the second set, the unit corresponding to the mother with the second lowest age is selected, and so on. Finally, from the s^{th} set, the unit corresponding to the mother with the highest age is selected. The remaining $s(s-1)$ sampled units are discarded from the data set.
- Steps 1 - 3, called a cycle, are repeated m times to obtain a ranked set sample of size $n = ms$.

Corresponding to each selected mother, information regarding the status of her infant survival status has been collected. If the infant is not alive then X will take the value '1' and '0' otherwise. With this notation, we have the sampled observations from each of the selected states for different of set sizes, s, and number of cycles, m. Here, $\pi_{[r]}$ denotes the probability that an infant who is not alive in the r^{th} class, and p is the probability that an infant dies before reaching one year in the entire population. The implementation of the proposed ML and Bayes approach of estimation, based on both a simple random sample and a ranked set sample, has been demonstrated. Here we use Beta (α, β) priors with (α, β)= $(0.5,0.5), (2,2), (3,5)$ and $(5,3)$. For these parametric combinations of (α, β), estimates compute the estimates $\hat{p}_{SRS}, \hat{p}^{B}_{SRS}$, $\hat{p}_{RSS}, \hat{p}^{B}_{RSS}$ and $\hat{p}^{*B}_{RSS}$ and all computed results are summarized in Tables 1 and 2 in the Appendix section. It has been observed that the Bayes, estimate of the probability that an infant dies before the completion of the year is very close to estimates, obtained through the ML approach. Also, both estimates are quite near to the value which is the estimated value of p reported by NFHS-III(2005-06)[15]. It is also clear that the proposed Bayes procedure, especially the estimators $\hat{p}^{B}_{RSS}$ and $\hat{p}^{*B}_{RSS}$ based on the ranked set samples are showing comparatively greater precision than the ML estimates. The sampling of ranked set sample units is done by using the Statistical Analysis System (SAS) package, University edition, and all other computation works are carried out by using the R package (version 3.4.4).

5. Conclusion

The focus of the present study lies on the problem of estimating unknown population proportion or probability of occurrence of an event(p), where p is a random quantity, based on a ranked set sample drawn from a dichotomous population. An application of the Bayesian method of estimation and existing maximum likelihood procedure (incorporating classical framework), for estimating p, have been discussed. Under the assumption about the parameters, $\pi_{[r]}$, that it is a function of p, no explicit form of posterior has been found, whereas, if one follows the structural independence of RSS procedure, if $\pi_{[r]}$'s considered as an independent variable, then the Beta conjugate posterior will be obtained. To compare the closeness of the estimators towards the estimation of parameters, both simulation and real-life-based results confirmed that for the given parametric combinations of α and β, both the Bayes' estimators ($viz.$,$\hat{p}^{B}_{RSS}$ and $\hat{p}^{*B}_{RSS}$), based on ranked set samples are more pitman close to p than any other estimator. In the estimation of the probability of infant deaths, Bayesian estimators based on a ranked set sample not only proved more effective and efficient than any other estimator but also consistent with the NFHS reported value. This study showed that the Bayesian estimation of the probability of occurrence of an event in the population using a ranked set sample is not only suitable and applicable for the estimation of demographic parameters but also provides greater efficiency and accuracy than the corresponding other procedure.

Table 1. ML and Bayes estimators of the probability of infant death under SRS and RSS of selected states in India and NFHS-III reported infant death, for different $s; m; \alpha$ and β

State	*s*	*m*	ML Risk		α	β	Bayesian Estimates			Population
			$\hat{p}_{SRS}$	$\hat{p}_{RSS}$			$\hat{p}^{B}_{SRS}$	$\hat{p}^{*B}_{RSS}$	$\hat{p}^{B}_{RSS}$	
Assam	3	170	0.06	0.061	0.5	0.5	0.062	0.063	0.059	0.066
					2	2	0.064	0.071	0.059	
					3	5	0.066	0.075	0.059	
					5	3	0.069	0.086	0.059	
Bihar	6	230	0.06	0.06	0.5	0.5	0.061	0.062	0.059	0.062
					2	2	0.063	0.067	0.059	
					3	5	0.063	0.07	0.059	
					5	3	0.066	0.077	0.059	
M.P.	5	120	0.07	0.067	0.5	0.5	0.067	0.07	0.066	0.07
					2	2	0.07	0.081	0.066	
					3	5	0.071	0.083	0.066	
					5	3	0.074	0.083	0.066	
Orissa	3	200	0.06	0.06	0.5	0.5	0.061	0.062	0.058	0.065
					2	2	0.063	0.069	0.058	
					3	5	0.064	0.072	0.058	
					5	3	0.067	0.082	0.058	
Rajasthan	5	80	0.06	0.055	0.5	0.5	0.056	0.06	0.054	0.065
					2	2	0.059	0.076	0.054	
					3	5	0.061	0.084	0.054	
					5	3	0.066	0.071	0.054	

Table 2. Risks (in 10^4) for the ML and Bayes estimators of probability of infant death under SRS and RSS of selected states in India and NFHS-III reported infant death, for different $s; m; \alpha$ and β

State	s	m	ML Risk		α	β	Bayesian Estimates		
			$\hat{p}_{SRS}$	$\hat{p}_{RSS}$			$\hat{p}^{B}_{SRS}$	$\hat{p}^{*B}_{RSS}$	$\hat{p}^{B}_{RSS}$
Assam	3	170	1.119	1.108	0.5	0.5	1.113	0.383	0.03
					2	2	1.117	0.418	0.028
					3	5	1.118	0.43	0.03
					5	3	1.115	0.488	0.026
Bihar	6	230	0.726	0.624	0.5	0.5	0.7	0.246	0.014
					2	2	0.715	0.261	0.013
					3	5	0.715	0.267	0.012
					5	3	0.718	0.293	0.012
M.P.	5	120	1.037	1.016	0.5	0.5	1.024	0.212	0.01
					2	2	1.027	0.235	0.008
					3	5	1.029	0.242	0.009
					5	3	1.015	0.281	0.007
Orissa	3	200	0.94	0.84	0.5	0.5	0.848	0.321	0.024
					2	2	0.874	0.346	0.024
					3	5	0.886	0.356	0.024
					5	3	0.826	0.399	0.021
Rajasthan	5	80	1.299	1.192	0.5	0.5	1.274	0.276	0.013
					2	2	1.28	0.33	0.012
					3	5	1.241	0.345	0.011
					5	3	1.251	0.428	0.011

Appendix of Figures

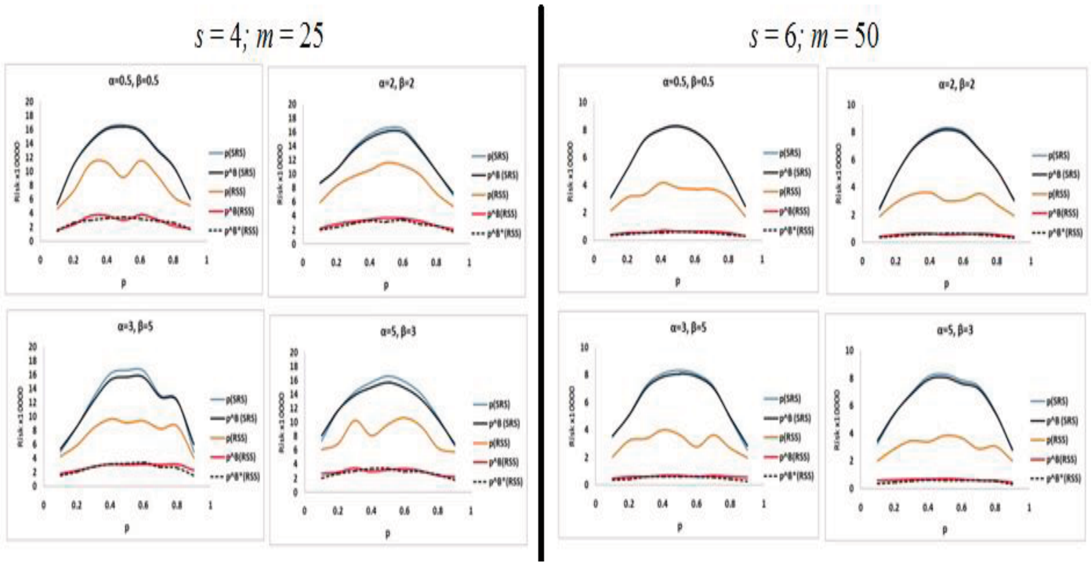


Figure 1. Risk curves (in 10^4) for the estimators $\hat{p}_{SRS}$, $\hat{p}^B_{SRS}$, $\hat{p}_{RSS}$, $\hat{p}^B_{RSS}$ and $\hat{p}^{*B}_{RSS}$when s = 4*;* m = 25 and s = 6*;* m = 50 at different choices of α and β

References

Chen, H. (2008). Alternative ranked set sample estimators for the variance of a sample proportion. In *Applied Statistics Research Progress* (pp. 35-38). Nova Publishers.

Chen, H., Stasny, E. A., & Wolfe, D. A. (2005). Ranked set sampling for efficient estimation of a population proportion. *Statistics in Medicine, 24*, 3319-3329.

Chen, H., Stasny, E. A., & Wolfe, D. A. (2007). Improved procedures for estimation of disease prevalence using ranked set sampling. *Biometrical Journal, 49*(4), 530-538.

Chen, Z., Z.D., & Sinha, B. K. (2004). *Ranked set sampling: Theory and applications*. Lecture Notes in Statistics, 176. Springer-Verlag, New York.

Das, R., Verma, V., & Nath, D. C. (2017). Bayesian estimation of measles vaccination coverage under ranked set sampling. *Statistics in Transition new series, 18*(4), 589-608.

Feeney, G. (1980). Estimating infant mortality trends from child survivorship data. *Population Studies: A Journal of Demography, 34*(1), 109-128.

Finlay, J. E., Özaltin, E., & Canning, D. (2011). The association of maternal age with infant mortality, child anthropometric failure, diarrhoea and anaemia for first births: evidence from 55 low-and middle-income countries. *BMJ Open, 1*(2), e000226.

International Institute for Population Sciences (IIPS), & Macro International. (2007). *National Family Health Survey (NFHS-3), 2005-06: India*. Mumbai: IIPS.

Jozani, M. J., Balakrishnan, N., & Davies, F. K. (2013). Some Pitman closeness properties pertinent to symmetric populations. *Statistics: A Journal of Theoretical and Applied Statistics*. https://doi.org/10.1080/02331888.2013.809721

McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research, 3*, 385–390. https://doi.org/10.1080/03610918.2019.1662043

Pitman, E. J. G. (1937). The closest estimates of statistical parameters. *Proceedings of the Cambridge Philosophical Society, 33*, 212-222.

Terpstra, J. T., & Nelson, E. J. (2005). Optimal rank set sampling estimates for a population proportion. *Journal of Statistical Planning and Inference, 127*, 309-321.

Verma, V., & Nath, D. C. (2019). Characterization of the sum of binomial random variables under ranked set sampling. *Statistics in Transition new series, 20*(3), 1-30.

Verma, V., & Nath, D. C. (2019). Estimation of measles immunization coverage in Guwahati by ranked set sampling. In *Viruses* (pp. 84382). IntechOpen. https://doi.org/10.5772/intechopen.84382

Verma, V., Nath, D. C., & Das, R. (2018). Bayesian Cramer-Rao lower bound of variances under ranked set sampling. *Materials Today: Proceedings, 5*(1), 1747-1758.

Verma, V., Nath, D. C., & Das, R. (2019). Bayesian bounds for population proportion under ranked set sampling. *Communications in Statistics - Simulation and Computation, 48*(2), 478-493. https://doi.org/10.1080/03610918.2017.1387659

Vivek, V., Das, R., & Nath, D. C. (2019). Representativeness of ranked set sampling based on Bayesian score. *Communications in Statistics - Simulation and Computation*, 1-16.

Bayesian Computation for the Concordance Correlation Coefficient: An Illustration Through Liver Cirrhosis Patients

Abstract

Clinical trials are alarmed very often to assess whether different raters/instruments produce similar results to measure a quantitative variable. The assessment of agreement between two raters for continuous responses plays a crucial role in setting decisions about medical diagnostic tools. In the current practice, more emphasis has been given to using the concordance correlation coefficient as a measure of reproducibility. These assessments are commonly carried by the concordance correlation coefficient. This paper carries the Bayesian counterpart to compute the concordance correlation coefficient estimator and establishes the performance of the proposed estimator. The methodology is illustrated on Liver Cirrhosis marker data. It is found feasible to compute the concordance correlation coefficient through an application of prior information. The Bayesian counterpart of CCC estimates applied between serum bilirubin and albumin among liver cirrhosis patients data and its 95% posterior interval for concordance correlation coefficient found to be very narrow, which indicates that estimates are very precise.

Keywords: Agreement analysis, ICC, Bayesian, repeated measurements

1. Introduction

The experiment of reproducible research is important to support different scientific research. The idea behind the development of a new experimental tool is to get at least the same level of output as the available gold standard. The measurement of the performance of the newly developed tool can be accessed through reproducible research with an available gold standard. Indeed, the requirement to quantify agreement between two tools is a matter of interest. If performances between two tools are captured through categorical observation, then the agreement analysis plays the role of demonstrating a level of equal performance. Particularly, in medical research, the application of biomarkers is important to detect the stage of any disease. The exploration of the relation between two markers is useful to know the actual medical scenario condition of any disease. When two continuously measured markers are positively correlated, then it is natural that the high level of one marker is automatically influencing other presence. Pearson's correlation coefficient is widely useful to detect the agreement between two markers. A study by [1] identified limitations in the system's ability to detect poor agreement in certain scenarios. For instance, in a laboratory setting where blood cell counters are used for hematology analysis, duplicate measurements of the same blood sample are routinely performed. The system struggled to flag inconsistencies in these duplicates.

on multiple occasions. Plotting the first measurement against the second measurement of the red blood cell counts for all available blood samples, they anticipated that the measurements would lie on a 45° line through the origin, within a tolerable error. While measuring a linear relationship, the Pearson correlation coefficient is unable to identify any potential deviations from the 45o line. There are various methods for assessing raters' agreement. For example, kappa statistic [2] and the weighted kappa statistic [3] are the most popular indices for measuring agreement between two rates. However, the Pearson correlation coefficient is widely used to capture the linear relationship between variables but fails to explore the departure from a 45^o line from the origin. The t-test measures the effectiveness of response between two categories but fails to measure case-by-case agreement. Instead of treating duplicate readings as two separate readings, these approaches treat them as replicates (random). If there were a first reading (earlier) and a second reading (later), there would be two different readings. The discussion about the performance of Intra-class correlation [4, 5] and within-subject [1] had been used conventionally as indices to evaluate reproducibility. The linear relation between two random variables can be accessed through the concordance correlation coefficient (CCC) when the intercept is zero and the slope is one. Review and comparison of different methods (i.e. agreement measurement are also discussed by [1]. The concordance correlation coefficient has been extended to address more general types of outcomes such as categorical data and complex study designs involving multiple observers and repeated measures [6]. It is found that the generalized CCC is the same to the weighted kappa coefficient for ordinal data [Cohen, 1968] and the kappa coefficient for binary data [2].

Agreement between continuous data measured from different observers or measurement methods is a question that has received a great deal of consideration from the scientific community. The methodology to compute agreement assessment between two rater's observers through continuous data is established by CCC [1]. Continuous outcomes as well as those that can be so treated such as count response [1]. However, literature for categorical data is established comparatively earlier [2, 3]. However, the Bayesian counterpart to compute the concordance correlation coefficient (CCC) is yet to be explored. The computational approaches through frequency on Intra-class Coefficient (ICC) are well organized by [7, 8]. The Bayesian approach for ICC is also developed by [9]. Recently, the Bayesian Estimator of the Intra cluster Correlation Coefficient from correlated binary responses is explored by [10]. The application of weighted concordance correlation for longitudinal data is discussed in [11]. The ICC quantifies the overall data variance due to between-subject variability. But CCC measured the distance in the plane of each pair of data to 45° line from the origin (the concordance line) [12]. Particularly, CCC is useful for more than two observers by adjusting the confounders. Moreover, CCC for more than two observers and adjusted by confounding covariates would be desirable for many real problems [1, 2, 13]. There are different techniques to evaluate agreement between two raters. For example, kappa statistic [2] and the weighted kappa statistic [3] are the most popular indices for measuring agreement between two rates. However, the Pearson correlation coefficient is widely used to capture the linear relationship between variables but fails to explore the departure from a 45° line from the origin. The t-test measures the effectiveness of response between two categories but fails to measure case-by-case agreement. In a case of very scattered data least square fails to find the departure from intercept equal to 0 and equal slope to 1. It is true, that concordance correlation can precisely execute the reproducibility between two readings through exploring departure from intercept equal to 0 and slope equal to 1.

The CCC is useful to calculate sample size for study validation [14]. It has an advantage for detecting goodness of fit in mixed effect modeling [11]. A class of CCC estimators is useful to handle the outlier data [6]. The CCC is also useful to specify the level of agreement for more than two raters [6].

The extension of concordance correlation coefficients for repeated measurement is studied through weighted extension [15]. The assessment of generalized nonlinear mixed effects is also addressed by the concordance correlation coefficient [11]. The modified version of the concordance correlation coefficient for the generalized estimating equation is proposed by [13]. When a significant clinical range is well-known and the study is conducted over that range, the CCC offers a meaningful interpretation and is unit-free. In addition, the accuracy and precision components of the CCC offer more insight. Therefore, the CCC, accuracy, and precision remain very useful tools. Therefore, we applied Bayesian counterpart of the concordance correlation coefficient. In this paper, our objective is to apply Bayesian concordance correlation coefficient on Liver cirrhosis data and explore its performance. The prior information about the relation between biochemical markers for Liver Cirrhosis patients has been used to illustrate the method. We propose an estimator for the concordance correlation coefficient and establish the performance of the proposed estimator. The standard error of the estimation is also derived and calculated. The application of the proposed methodology is illustrated with liver cirrhosis patient data.

2. Data Methodology

2.1 Concordance Correlation Coefficient

Suppose the observation of two different biochemical parameters are observed $X$and Y. Now the set of two measurements $X = (x_1,\cdots,x_n)$and $Y = (y_1,\cdots,y_n)$.are independent and identically distributed bivariate populations with means $\mu_x$and $\mu_y$and covariance matrix

$$\begin{bmatrix} {\sigma^2}_x & \sigma_{xy} \\ \sigma_{xy} & {\sigma^2}_y \end{bmatrix}$$

Now the observations are called as perfectly agreed if $x_i = y_i$for $i = 1,\cdots,n.$ The presence of an angle between x and y can be denoted as$\theta = 0$, i.e. $cos\theta = 1$.

$$cos\theta = X,Y > \frac{}{||X||\vee Y \vee} \tag{1}$$

Now, $X,Y \geq \sum_{i=1}^{n} X_i Y_i$ and $\vee X \vee = \sum_{i=1}^{n} {X_i}^2$ are n-dimensional vectors in the Euclidean norm R^n.

Further, the above equation can be stated as

$$\rho(X,Y) = \tag{2}$$

whereρ is the Pearson correlation coefficient. If only if $\rho(X,Y) = 1$, when x and y are perfectly agreed. But it failed to work for the scale change. The degree of agreement between x and y can be observed through the expected value of the squared difference$E]$ i.e.

$$E]=(\mu_x - \mu_y \quad {}^2 + \left({\sigma^2}_x + {\sigma^2}_y - 2\sigma_{xy}\right)$$

$$=(\left(\mu_x - \mu_y\right)^2 + \left(\sigma_x - \sigma_y\right)^2 + 2(1-\rho)\sigma_{xy}(3)$$

If each pair, X, and Y, in the population are in perfect agreement, $E[(X,-Y)^2]$ would be 0. To scale the index value between -1 and 1, the following transformation is suggested: [16]:

$$\boldsymbol{\rho^c = 1 - \frac{E[(X-Y)^2]}{\sigma^2{}_x+\sigma^2{}_y+}} \tag{4}$$

2.2 Relation between Pearson Correlation and Concordance Correlation.

The relationship between Pearson (ρ) and Concordance correlation (ρ^C) can be described as below:

(I)$\rho^c \leq 1.$ Where, ρ^c and ρ having the same sign.

(II)$\rho^c = \rho$, if and only if $\left|\left|E(X) - E(Y)\right|\right| = 0$and $\left|\left|X - E(X)\right|\right| = \left|\left|Y - E(Y)\right|\right|$i.e. $\sigma_x = \sigma_y,.\mu_x = \mu_y$

(III)$\rho^c = 0$if and only if $\rho = 0.$

(iv)$\rho^c = \pm 1$ if and only if $\left(\mu_x - \mu_y\right)^2 + \left(\sigma_x - \sigma_y\right)^2 + 2(1-\rho)\sigma_{xy} = 0$

whereρ and $\rho^c$are the Pearson correlation coefficient and concordance correlation coefficient (CCC) respectively. The above equation (iv) reduced to Pearson correlation coefficient $\rho \pm 1$, if and only if $\sigma_x = \sigma_y \wedge \mu_x = \mu_y$

The CCC is substituted as sample moments of the independent bivariate sample into the above equation by ρ^c. The normal approximation is provided through the Fisher's Z-transformation as

$$Z = tanh^{-1}(\acute{\rho}^c) = \frac{1}{2}\left(\frac{1+\acute{\rho}^c}{1-\acute{\rho}^c}\right) \tag{5}$$

[16] proposed concordance correlation coefficient is based on the squared function of distance $g(z) = z^3$ when the underlying bivariate distribution is heavy-tailed.

2.3 Statistical Models.

Bayes's theorem is useful for robust statistical inference. It is useful to update current information to draw statistical inferences. The posterior probability value is computed through Baye's theorem for the concordance correlation coefficient. The posterior probability of concordance correlation is calculated by

$$P(concordancecorrelation|data) = \frac{P(data/concordancecorrelation)*P(concordancecorrelation)}{P(data)} \tag{6}$$

The term$P(data/concordancecorrelation)$is the likelihood function of concordance correlation. The value of $P(concordancecorrelation)$ is obtained from prior literature. The value of P (data) gives the cumulative measurements of all possible values of concordance observations. The posterior probability is obtained through

$PosteriorProbability\infty LikelihoodxPriorProbability(7)$

The posterior probability is proportional to the likelihood and prior probability.

Let the variable of interest are $X$and Y. The sample mean and standard deviation of$X$ and$Y$are given as

$$\acute{x} = \frac{\sum x_i}{n}, \quad \acute{y} = \frac{\sum y_i}{n} \quad ,S_{xx} = \frac{1}{n}\sum(x_i - \acute{x})^2 \ ,S_{yy} = \frac{1}{n}\sum(y_i - \acute{y})^2$$

and

$$S_{xy} = \frac{1}{n}\sum(x_i - \acute{x})(y_i - \acute{y})$$

The sample Pearson correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{8}$$

The sample concordance correlations

$$r^c = \frac{2S_{xy}}{S_x+S_y+\left(\mu_x-\mu_y\right)^2} \tag{9}$$

The posterior density value of the Concordance correlation is defined as

$$P(\rho^c \vee x, y) \propto P(\rho^c)\frac{\left(1-\rho^2\right)^{(n-1)/2}}{(1-\rho\times r)^{n-\left(\frac{3}{2}\right)}} \tag{10}$$

Here, r is the sample correlation coefficient. Now, the, CCC and sample coefficient are replaced as$\rho^c = tanh\xi$,$r =$

$tanhz$. The mean and variance of X and Y are μ_1, μ_2, σ_1^2 and σ_2^2 respectively. The mean (z) is derived from $e^z = \frac{\mu_1\sigma_2}{\mu_2\sigma_1}$. Now, ξ is assumed to follow Normal distribution with mean z and variance $(1/n)$ [17,18].

The hyperbolic tangent transformation is used to allow the conjugate prior of normal to build a posterior distribution of the correlation coefficient.

There are different ways to define the prior but the handy one is

$$P(\rho^c) \propto (1-\rho^2)^m (11)$$

The prior value of m should be supported based on prior information. However, the detailed about selection of m can be referred to from [19] and [20].

The sample values of r^c is defined as $\rho_{prior}^c$and$\rho_{likelihood}^c$. The value of ρ_{prior}^c is obtained from the support of prior information and $\rho_{likelihood}^c$ from the study data itself. Further, the likelihood function to the posterior estimates of CCC (ρ_{prior}^c) is defined a

$$\mathrm{X}\sigma_{posterior}^2\left(\eta_{prior} \times tanh^{-1}\rho_{prior}^c + \eta_{Likelihood} \times tanh^{-1}\rho_{likelihood}^c\right) \tag{12}$$

where the term is used to define the sample size.

Further, the sample value of $\sigma_{posterior}^2$ in the above equation is obtained as

$$\sigma_{posterior}^2 = \frac{1}{\eta_{Prior}+\eta_{likelihood}} \tag{13}$$

$$\text{The sample CCC is called as } \rho_{prior}^c = \frac{2S_{xy}}{S_x+S_y+\left(\mu_x-\mu_y\right)^2} \tag{14}$$

Further, the likelihood function to the posterior estimates of CCC ρ^c is defined as

$$\mu_{Posterior} = \sigma_{posterior}^2\left(\eta_{prior} \times tanh^{-1}\rho_{prior}^c + \eta_{Likelihood} \times tanh^{-1}\rho_{likelihood}^c\right) \tag{15}$$

The posterior interval is obtained by

$$(\mu_{Posterior} \pm 1.96 * \sqrt{\sigma_{posterior}^2} \tag{16}$$

The 95% posterior interval of CCC can be defined as

as $(\mu_{Posterior} - 1.96 * \sqrt{\sigma_{posterior}^2}, \mu_{Posterior} + 1.96 * \sqrt{\sigma_{posterior}^2})$

3. Application

3.1 Data Description

Data has been taken from the clinical trial of therapeutic drug development for the Liver Cirrhosis patients (from path http://www4.stat.ncsu.edu/~boos/var.select/pbc.html).

Following cell transplantation, patients were expected to attend follow-up appointments twice a week for the first week, once a week for the next eight weeks, once every four weeks for the next 24 weeks, and once every three months for the next two years. Following the Week 24 Visit, assessments of trough plasma tacrolimus levels and routine clinical laboratory testing (haematology, blood chemistry, and urinalysis) were also carried out once a month in between visits (planned every three months). More details about data can be cited with [21].

3.2 Computation of Bayesian Concordance Correlation Coefficient

The agreement between serum bilirubin and albumin has been measured to explore posterior estimates of CCC in a total of 172 patients. The preliminary exploratory data was tested and normality assumptions were found to be followed for those selected subjects. The generated information between serum bilirubin and albumin is used as prior information of a sample size of 418 individuals in the Primary Biliary Cirrhosis (PBC) study [22].

The computed value of r_{prior}^c is observed with 0.70 and a calculated value of$r_{likelihood}^c$ is 0.63.Further, the computed value of $r_{posterior}^c$is obtained as 0.001.

$$\mu_{posterior} =$$

$$\sigma^2_{posterior}\left(\eta_{prior} \times tanh^{-1}\rho^c_{prior} + \eta_{Likelihood} \times tanh^{-1}\rho^c_{likelihood}\right) \quad (17)$$

$$\mu_{posterior} = 0.001\left(418 \times tanh^{-1}(0.70) + 180 \times tanh^{-1}(0.63)\right) \quad (18)$$

$$\mu_{posterior} = 0.001\left(418 \times (0.87) + 180 \times (0.75)\right) = 0.49866 \quad (19)$$

The 95% posterior interval is $0.49866 \pm 1.96 * \sqrt{0.001}$ is [0.43, 0.56].

4. Result &Discussion

The baseline and demographic details of patients with liver cirrhosis in the treatment and control groups are shown in Table 1. Patients with liver cirrhosis in the therapy group had an average age of 48.6 years with a standard deviation (SD) of 9.38, whereas those in the control group had an average age of 49.85 years with an SD of 11.06. There are 10 (or 12.051%) males and 73 (or 87.958.0%) females in the treatment group of people. Comparably, in the control group, the distribution of males and females is 70 (76.09%) and 22 (23.91%), respectively. The male patients are more in the control group as compared to the treated group. Patients with liver cirrhosis in the therapy group had an average height of 164.89 with SD 4.46, while those in the control group had an average height of 165.38 with SD 5.84. Patients with liver cirrhosis in the therapy group had an average weight of 65.71 with SD 5.24, while those in the control group had an average weight of 69.51 with SD 8.87. Patients with liver cirrhosis in the therapy group had an average Respiratory Rate (RR) of 25.99 with an SD of 16.59, while patients in the control group had an average RR of 21.43 with an SD of 1.62. Patients with liver cirrhosis in the therapy group had a mean heart rate (HR) of 72.89 with a standard deviation of 15.61, while the mean HR in the control group was 77.22 with a standard deviation of 2.08.

Table 2 describes the posterior estimate of CCC along with a 95% credible interval.

The assessment of the relation between two variables plays a crucial role in setting decisions about medical diagnostics tools. The widely explored tools to explore the relationship between two continuous variables are the Pearson and Spearman correlation coefficients. However, the Pearson correlation coefficient is not useful for multivariate data analysis. Distance correlation which is a measure of statistical dependence between two random variables and not necessarily of equal dimension, is another choice for this gap. Bayesian techniques to compute the Distance correlation are also elaborated in [18]. Measurements of one variable can be waived off by consideration of another variable if the concordance correlation between them is highly positive.

This work explores the potential of the Bayesian approach for calculating CCC. Compared to the frequentist approach, Bayesian methods offer greater flexibility and can incorporate prior knowledge, leading to more realistic results. The authors demonstrate this by estimating the CCC between serum bilirubin and albumin in liver cirrhosis patients using a Bayesian framework.

The development of the robust extension of the concordance correlation coefficient is discussed by [6]. The generalized form of the concordance correlation coefficient is a useful and unified approach for agreement measurement. It is more appropriate for categorical and continuous data analysis. However, it is not as appropriate for ordinal measurements. Also, [6] has proposed a general index, the generalized concordance correlation coefficient, for evaluating agreement for continuous and categorical data. [6] also introduced a stratified concordance correlation coefficient that adjusts for categorical covariates in the marginal mean and an extended concordance correlation coefficient that measures agreement among more than two responses.

5. Conclusion

The application of Bayesian computation of the Concordance Correlation Coefficient is described by the agreement of measurement between two continuous random variables. This application is an effort on Biochemical markers to get prominent evidence about test statistics on a relation between variables. The concordance correlation coefficient may seem to be convenient to have a single measure of agreement, but it is very inconvenient when a low is obtained in a study. One is unsure whether the evident disagreement is due to the homogeneity of the sample, the systematic bias between methods, or great random error between methods. The solutions of the latter two causes of disagreement are very different in nature; after obtaining a low, a researcher would not know whether the measurement tool needs to correct systematic bias or is inherently associated with large random errors. Regardless of the above drawback, CCC has attractive characteristics. It is simple to use. For bivariate normal data, their asymptotic normality and consistency are guaranteed by utilizing the sample counterparts. However, the inverse hyperbolic tangent transformation (Z-transformation) could be used to enhance its statistical features (consistency and asymptotic normality) [6]. Even with tiny sample sizes, it also holds up well against samples from the Poisson and uniform distributions. The further studies should be done to explore the CCC approach in more than two raters.

Acknowledgment

The first author acknowledges the support of UGC-BSR one-time grant [F-19-145/2015 (BSR)] of the University Grant

Commission, New Delhi. The authors thank the two anonymous referees for their cautious reading and constructive suggestions which have led to improvement on earlier versions of the manuscript.

Competing Interest

None Declared

Appendix of tables

Table 1. Descriptive statistics about baseline observations of the patients

Parameters	**Treatment group**	**Control Group**
	Mean (SD)	**Mean (SD)**
Age	48.60(9.38)	49.85(11.06)
Gender Male	10 (12.1%)	70 (76.1%)
Female	73 (88.0%)	22 (23.9%)
Height	164.89(4.46)	165.38(5.84)
Weight	65.71(5.24)	69.51(8.78)
Respiratory rate (RR)	25.99(16.59)	21.43(1.62)
Heart Rate	72.89(15.61)	77.22(2.08)

Table 2. Summary Parameters estimate for CCC

Parameters	**Estimates**
$r_{prior}^{c}(observed)$	0.70
$r_{likelihood}^{c}$	0.63
$\sigma_{posterior}^{2}$	0.001
$\mu_{posterior}$	0.49866
95% posterior interval	[0.43, 0.56]

References

Ahmed, M., & Shoukri, M. (2010). A Bayesian estimator of the intracluster correlation coefficient from correlated binary responses. *Journal of Data Science, 8*, 127-137.

Barnhart, H. X., & Williamson, J. M. (2001). Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics, 57*, 931-940.

Bashir, S. A., & Duffy, S. W. (1997). The correction of risk estimated for measurement. *Annals of Epidemiology, 7*, 154-164.

Bhattacharjee, A. (2014). Distance correlation coefficient: An application with Bayesian approach in clinical data analysis. *Journal of Modern Applied Statistical Methods, 13*(1), Article 23.

Box, G. E. P., & Tao, D. R. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Carrasco, J. L., & Jover, J. L. (2003). The concordance correlation coefficient estimated through variance components. In *Proceedings of the IX Conferencia Española de Biometría*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review, 54*, 67-82.

Fisher, R. A. (1915). Frequency distributions of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*, 507-521.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: Wiley.

Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley, New York.

King, T. S., & Chinchilli, V. M. (2001). Robust estimators of the concordance correlation coefficient. *Journal of Biopharmaceutical Statistics, 31*, 83-105.

Lee, J., Koh, D., & Ong, C. N. (1989). Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Computers in Biology and Medicine, 19*, 61-70.

Lin, L. (1992). Assay validation using the concordance correlation coefficient. *Biometrics, 48*, 599-604.

Lin, L. I. (1989). Concordance correlation coefficients to evaluate reproducibility. *Biometrics, 45*, 255-268.

Lin, L., & Chinchilli, V. (1997). Rejoinder to the letter to the editor from Atkinson and Nevill. *Biometrics, 53*, 777-778.

Nath, D. C., Bhattacharjee, A., & Vishwakarma, R. K. (2015). Cure rate modeling: An application with Bayesian approach in liver cirrhosis patients. *Journal of Data Science, 13*, 421-42.

Quan, H., & Shih, W. J. (1996). Assessing reproducibility by the within-subject coefficient of variation with random effects models. *Biometrics, 52*, 1195-1203.

Shoukri, M. M., & Ward, R. H. (1985). Estimation of intraclass correlation. *Communications in Statistics - Theory and Methods, 13*, 1239-1255.

Spiegelhalter, D. J. (2001). Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine, 20*, 435-52.

Vonesh, E. F., Chinchilli, V. M., & Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics, 52*, 572-587.

Prevalence and Risk Factors of Low Birth Weight Among Adolescent Mothers in Assam

Abstract

Low birth weight (LBW), is a newborn baby weighing less than 2.5 kilograms and it is defined by the World Health Organization (WHO). Birth weight is an important indicator when assessing a child's health for early childhood morbidity and mortality exposure. This study aimed to identify the prevalence of low birth weight & its risk factors among the adolescent age group of mothers in Assam. The fourth round of the National Family Health Survey data (NHFS-4; 2015-2016) was used. Univariate and Multivariable logistic regression models were applied to identify the risk factors. For those mothers with a primary level of education, the prevalence of LBW was reported at 24.5 %, for poorest mothers at 21.4%, fourth & above birth order at 23.5%, and among anemic mothers at 17.2%. The prevalence of LBW was wide-ranging across the districts, it was highest in Kamrup and Dhubari and lowest in Sonitpur and Karbi Anglong districts. Education, wealth index, higher level of birth order, and method of reporting were risk factors for low birth weight in Assam.

Keywords: LBW, Adolescent, District, Prevalence, Assam.

1. Introduction

Low birth weight (LBW), of a baby, is a weight less than 2.5 kilograms according to the World Health Organization (WHO). Low birth weight continues to be a significant public health problem globally and is associated with a range of both short and long-term consequences. 'Low birth weight (LBW)' is an important health indicator for infants. LBW babies indicated a high incidence in a rural setup and an equal proportion of prematurity as a cause of LBW. In the Rural area of Assam, LBW babies, are identified among illiterate teenage mothers, anemic mothers, short inter-pregnancy intervals, and nutrition during pregnancy and it is an essential fact for a healthy mother and a healthy baby. Short inter-pregnancy intervals were the important risk factors for LBW **[1]**. Lack of true measurement of birth weight is a major problem of underestimating the prevalence of LBW **[2].**

LBW infants suffer more incidences of common childhood diseases and the curse of illness is more prolonged and serious. The incidence of morbidities was higher among LBW babies compared to normal birth weight (NBW) babies **[3]**. Low birth weight occurred due to poor socioeconomic development, maternal nutrition, and non-utilization of health services during pregnancy. LBW occurred due to poor awareness about the existing maternal services at the basic level of the community **[4]**. Digit preference is one of the measure causes of underreporting of LBW and due to this reason actual prevalence of LBW was influenced **[5]**.

The prevalence of low birth weight is high among those women who were underweight, anemic, and never visited for any ANC checkup and maternal nutritional factors are more persistent across India. In India, higher parity and birth order is a common issue of the lower reporting of birth weight particularly in the northern, central, and eastern regions of India **[6,7].** Keeping in the view literature review, there were no specific LBW research studies reported on the adolescent age group in Assam. Therefore, the authors aimed to identify the prevalence of low birth weight & its risk factors among the adolescent age group in Assam state of India.

2. Materials and Methods

The fourth round of the National Family Health Survey (NFHS-4) data was considered for the analysis and it was collected during 2015-2016. It is freely available at public forums for research. For the fulfillment of the objective, data was analyzed for the adolescent age group (15-24) for Assam states of India. The primary sample units (PSUs) were selected from the sampling frame which was created from the 2011 census. PSUs were 'villages' in rural areas and 'Census Enumeration Blocks (CEBs)' in urban areas. PSUs with 'fewer than 40 households' were linked to the nearest PSU. Within each rural stratum, villages were selected from the sampling frame with probability proportionate to size (PPS). The final sample PSUs and CEBs were selected with PPS sampling. In every selected rural and urban PSU, a complete household mapping and listing operation was conducted before the main survey **[7].**

2.1 Dependent Variables

Birth weight was considered as an outcome variable. It was categorized dichotomously [Low birth weight (LBW) and normal birth weight (NBW)] for analysis purposes.

2.2 Independent Variables

In this study number of predictor variables were considered; place of residence (rural/urban); Educational level (No education, Primary, Secondary, Higher); Wealth index (Poorest, Poorer, Middle, Richer, and Richest); Birth order (First order, Second order, Third order and Fourth & above order); Anemic mother (Anemic and Not-anemic); Method of reporting (From written card and mother's recall).

2.3 Statistical Analysis

Descriptive statistics (frequency and proportion), Bivariate logistic regression (unadjusted odds ratio, p-value, and confidence interval), and multivariable logistic regression model (adjusted odds ratio, p-value, and confidence interval) were used to find the prevalence and the risk factors for the occurrence of LBW.

3. Results and Discussion

In India, the prevalence of low birth weight among the adolescent age group of mothers was reported at 19.7% and among the reproductive age group (15-49) was reported at 18.2%. The prevalence of LBW among the adolescent age group was higher than the reproductive age group (15-49). Similarly, in Assam, the prevalence of LBW among the adolescent age group (17.2%) was higher than in the reproductive age group (15.8%) [Table 1]. In urban places, the prevalence of LBW was 18.1%, and in rural was 17.1%. Among the primary level of education, LBW was reported at 24.5 %, which was higher than other categories of education. The poorest level of the mother wealth index was reported at 21.4%, which was higher than the remaining category of the wealth index. Fourth & above birth order was reported 23.5% than first, second, and third birth order. LBW among the anemic mothers was 17.2%. Birth weight reported through mothers' memory recall among them LBW was 18.6%, which was higher than written health card (16.7%). Similar studies were reported about maternal nutritional factors and reporting of birth weight and important risk factors of LBW in India [8,9].

LBW is associated with high infant mortality, particularly within the first month of life. Birth weight is an important measure for determining the neonatal and infant's survival. Over the decades, several intervention programs including Reproductive and Child Health have been launched all over India to improve the health status of mothers and children [10,11].

In Assam, the state government is involved in achieving the target of the national health mission (NHM) and sustainable development goal (SDG-3) for improving maternal and child health goals. There were various health programs like; Janani shishu suraksha karyakram (JSSK) and Janani shishu yojana (JSY). In each district of Assam, low birth weight was reported in Table 2. The highest prevalence of low birth weight was reported (32.5%) in Kamrup district, and Dhubri district (24.8%), and the lowest was reported (7.8%) in Sonitpur, and KarbiAnglong districts (8%). Similarly, hospital-based selected studies reported [5,12].

The univariate and multivariable logistic regression model was developed to identify the risk factors of LBW in Assam. Unadjusted logistic regression model; When mothers were educated at primary level among them 1.54 times (OR: 1.54, 95% CI: 0.99-2.37, $p < 0.05$) more likely to have low birth weight than uneducated. The poorest group of mothers was more likely to have low birth weight than another category of the wealth index. Fourth and above level of birth orders were 1.56 times more likely to have low birth weight than first birth order. Birth weight reported through mothers' memory recall (birth size) was 1.14 times more likely to have low birth weight than reported through health cards (written cards). Adjusted logistics regression model; controlling the other factors, those who were residing in rural places were 0.22 times less likely to have low birth weight than those residing in urban places. When mothers were educated at the primary level, they were 1.59 times (OR: 1.59, 95% CI: 1.02-2.48) more likely to have low birth weight than uneducated [Table 3].

This study was carried out in Assam and its district to understand the prevalence and risk factors of LBW among the adolescent age group of the mothers. The authors utilized the Assam state representative cross-sectional secondary data to determine the objective. There were similar studies reported to find the clinical and nonclinical risk factors of low birth weight and its risk factors in Assam state. Almost all research has been completed in the state which is focused on hospital-based data but this is the cross-sectional survey-based secondary data [5,13,14]. Here, similar risk factors of low birth weight were reported [15,16,17]

4. Limitation of Study

Proper antenatal care plays an important role in the healthy outcome of the pregnancy. Since this is the cross-sectional survey-based data for India, states, and districts, retrospective birth history information is used. This data was analyzed with selected indicators for only the adolescent age group of the mothers. The large number of information available in this data set is open for further studies. This study focuses only on the Assam state in India.

Table 1 shows the prevalence of Low birth weight among the adolescent age group (15-24) through different background characteristics, State of Assam.

Predictors	**LBW**	**NBW**	**Total**
Place of Residence			
Urban	29 (18.1)	131 (81.9)	160
Rural	273 (17.1)	1320 (82.9)	1593
Total	302 (17.2)	1451 (82.8)	1753
Educational level			
No education	43 (17.6)	202 (82.4)	245
Primary	65 (24.5)	200 (75.5)	265
Secondary	189(15.7)	1012 (84.3)	1201
Higher	5 (11.9)	37 (88.1)	42
Total	302 (17.2)	1451 (82.8)	1753
Wealth index			
Poorest	89 (21.4)	327 (78.6)	416
Poorer	139 (17.0)	677 (83.0)	816
Middle	54 (15.5)	294 (84.5)	348
Richer	16 (11.0)	129 (89.0)	145
Richest	4 (14.3)	24 (85.7)	28
Total	302 (17.2)	1451 (82.8)	1753
Birth order			
First order	227 (17.9)	1043 (82.1)	1270
Second order	63 (15.9)	333 (84.1)	396
Third order	7 (10.1)	62 (89.9)	69
Fourth and above order	4 (23.5)	13 (76.5)	17
Anemic mother			
Anemic	149 (17.2)	718 (82.8)	867
Non-anemic	147 (17.3)	704 (82.7)	851
Method of reporting			
From written card	204 (16.7)	1021 (83.3)	1225
From mother's recall	98 (18.6)	430 (81.4)	528
India age group (15-24)	**14184 (19.7)**	**57869 (80.3)**	**72053**
India	**35475 (18.2)**	**159343 (81.8)**	**194818**

Table 2 shows the district-wise prevalence of low birth weight among adolescents age group (15-24), state of Assam.

District	**Low Birth Weight N (%)**	**Normal Birth Weight N (%)**	**Total**
Kokrajhar	7 (14.6)	41 (85.4)	48
Dhubri	30 (24.8)	91 (75.2)	121
Goalpara	10 (13.9)	62 (86.1)	72
Barpeta	22 (20.0)	88 (80.0)	110
Morigaon	12 (16.0)	63 (84.0)	75
Nagaon	29 (17.4)	138 (82.6)	167
Sonitpur	6 (7.8)	71 (92.2)	77
Lakhimpur	10 (18.5)	44 (81.5)	54
Dhemaji	8 (15.4)	44 (84.6)	52
Tinsukia	18 (21.40	66 (78.6)	84
Dibrugarh	16 (20.5)	62 (79.5)	78
Sivasagar	10 (17.5)	47 (82.5)	57
Jorhat	8 (10.1)	71 (89.9)	79
Golaghat	8 (12.1)	58 (87.9)	66
KarbiAnglong	2 (8.0)	23 (92.0)	25
Dima Hasao	1 (12.5)	7 (87.5)	8
Cachar	14 (16.5)	71 (83.5)	85
Karimganj	10 (14.3)	60 (85.7)	70
Hailakandi	5 (13.5)	32 (86.5)	37
Bongaigaon	7 (15.2)	39 (84.8)	46
Chirang	3 (15.0)	17 (85.0)	20
Kamrup	27 (32.5)	56 (67.5)	83
Kamrup Metropolitan	6 (15.4)	33 (84.6)	39
Nalbari	7 (20.6)	27 (79.4)	34
Baksa	13 (20.0)	52 (80.0)	65
Darrang	7 (11.1)	56 (88.9)	63
Udalguri	5 (13.5)	32 (86.5)	37
Assam	**301 (17.2)**	**1451 (82.8)**	**1752**

Table 3 shows the adjusted and unadjusted logistic regression model through different background characteristics for the adolescent age group (15-24), Assam.

	Adjusted			Unadjusted		
Predictors		**95% C. I.**			**95% C. I.**	
Place	OR	Lower	Upper	OR	Lower	Upper
Urban	1.00	-	-	1.00	-	-
Rural	0.78	0.49	1.25	1.06	0.69	1.62
Education						
No education	1.00	-	-	1.00	-	-
Primary	1.59	1.02	2.48	1.54*	0.99	2.37
Secondary	0.97	0.66	1.44	0.88	0.61	1.27
Higher	0.74	0.26	2.15	0.59	0.21	1.63
Wealth index						
Poorest	1.00	-	-	1.00	-	-
Poorer	0.78	0.57	1.07	0.75**	0.56	1.01
Middle	0.68	0.45	1.02	0.67*	0.46	0.97
Richer	0.45	0.24	0.86	0.45*	0.26	0.79
Richest	0.64	0.21	1.94	0.64	0.22	1.85
Birth order						
First order	1.00	-	-	1.00	-	-
Second order	0.75	0.55	1.03	0.87	0.64	1.18
Third order	0.43	0.19	0.94	0.56	0.26	1.20
Fourth and above order	1.19	0.39	3.61	1.56	0.53	4.64
Anemic mother						
Anemic	1.00	-	-	1.00	-	-
Non-anemic	1.01	0.78	1.30	1.00	0.78	1.29
Method of reporting						
From written card	1.00	-	-	1.00	-	-
From mother's recall	1.18	0.89	1.54	1.14	.87	1.48

5. Conclusion

The proportion of low birth weight among the adolescent age group of mothers was 17.2% in Assam. Which is lower than the national level. Primary education, poorest wealth index, fourth and above birth order, and mother memory recall (birth size) were identified as higher risk factors for low birth weight. Currently, a total of 35 districts in Assam but in this study 27 district prevalence of low birth weight reported. The maximum proportion of LBW was 32.5% in Kamrup and Dhubari (24.8%) districts. The minimum proportion of LBW was 7.8% in Sonitpur and Karbi Anglong (8.0%) districts.

Acknowledgments

We are thankful to the entire team of the National Family Health Survey of India.

Conflict of Interest

There is no conflict of interest.

References

Borah, M., & Agarwalla, R. (2016). Maternal and socio-demographic determinants of low birth weight (LBW): A community-based study in a rural block of Assam. *Journal of Postgraduate Medicine, 62*(3), 178–81.

Borah, M., & Baruah, R. (2015). Morbidity status of low birth weight babies in rural areas of Assam: A prospective longitudinal study. *Journal of Family Medicine and Primary Care, 4*(3), 380–3.

Chutia, D., & Sharma, H. (2017). A study of low birth weight in a primary health centre of Cachar district, Assam—a record-based study. *Journal of Evolutionary Medical and Dental Sciences, 6*(82), 5772–4.

Dey, P., Zahir, F., Jha, R. K., & Ranjan, R. (2019). Maternal antenatal profile in VLBW and ELBW babies. *Journal of Evolutionary Medical and Dental Sciences, 8*(15), 1237–9.

Dubey, D. K., & Nath, D. C. (2016). An epidemiological model investigating the association between mothers' nutritional status and low birth weight in India. *Health, 8*(3).

Dubey, D. K., & Nath, D. C. (2016). Prevalence of low birth weight & its change due to heaping problem in collected data on birth weight. *Asian Journal of Social Sciences & Humanities, 5*(1).

Dubey, D. K., & Nath, D. C. (2017). Measurement issues of low birth weight in India. *JB Economics, 3*(2), 31-40).

Dubey, D. K., & Nath, D. C. (2019). Regional models assessing region-specific determinants of low birth weight in India. *Current Science, 116*(10), 1674.

Gathimba, N. W., Wanjoya, A., Kiplagat, G. K., Mbugua, L., & Kibiwott, K. (2017). Modeling maternal risk factors affecting low birth weight among infants in Kenya. *American Journal of Theoretical and Applied Statistics, 6*(1), 22-31.

Gogoi, N. (2018). Maternal and neonatal risk factors of low birth weight in Guwahati metro, Assam, northeast India. *Academic Journal of Pediatrics & Neonatology, 6*(5).

Haq, M. I., Uddin, M. S. G., & Islam, M. (2022). Low birth weight baby and its associated factors among rural women in Bangladesh: A decision curve analysis. *Bangladesh Medical Research Council Bulletin, 47*, 42-49.

International Institute for Population Sciences (IIPS), & ICF. (2017). *National Family Health Survey (NFHS-4)*. Mumbai: IIPS.

Kumar, S. N., Raisuddin, S., Singh, K. J., Bastia, B., Borgohain, D., & Teron, L. (2020). Association of maternal determinants with low-birth-weight babies in tea garden workers of Assam. *Journal of Obstetrics and Gynaecology Research, 46*(5), 715–26.

Patale, J. P., Masare, S., & Bansode-Gokhe, S. (2018). A study of epidemiological co-relates of low birth weight babies born in tertiary care hospital. *International Journal of Research in Medical Sciences, 6*(3).

Paul, T., Chakraborty, K., Sarkar, N., Chatterjee, M., & Roy, S. (2020). Prevalence and correlates of low-birth-weight babies born in a tertiary care teaching hospital in Eastern India. *International Journal of Community Medicine and Public Health, 7*, 3052.

Phukan, R. K., & Mahanta, J. (1998). A study of neonatal deaths in the tea gardens of Dibrugarh district of upper Assam. *Journal of the Indian Medical Association, 96*(11), 333–4, 337.

Rahman, K., Bhuyan, A. R., & Ullah, M. A. (2015). Incidence and clinical profile of low birth weight (LBW) babies: A rural tertiary care hospital-based study. *The New Indian Journal of OBGYN, 2*(1), 43–5.

ISBN 978-0-9784301-6-0 (Print)
ISBN 978-0-9784301-7-7 (Ebook)

Published by OABooks (The Canadian Center of Science and Education)

1595 Sixteenth Ave, Suite 301

Richmond Hill, Ontario

L4B 3N9

Canada

9 780978 430160 >