CHAPTER IV

4.0 METHODOLOGY

This chapter includes the details about operational definition of sex ratio, female infanticide, source of data, statistical models used in the analysis and chapters of thesis.

4.1 DEFINITION

SEX RATIO

Sex Ratio is defined as the number of females per thousand males in a population.

\[
Sex \, Ratio = \frac{Number \, of \, females}{Number \, of \, males} \times k
\]

where k is the radix and it takes the values 1000 or multiples of 1000.

FEMALE INFANTICIDE

UNICEF defined, female infanticide as the abortion of foetus or the killing of an infant by a relative because it is female.

4.2 SOURCE OF DATA

This study uses the data from 2001 and 2011 census. The district and taluk the unit of analysis. All the districts in Tamil Nadu State are considered for the analysis. There are 30 districts in 2001 and 32 districts in 2011 after bifurcation of two districts (Coimbatore and Dharmapuri) in Tamil Nadu. The bifuricated districts in 2011 are combined with the respective parent districts in 2001 and the sex ratio is calculated from
the districts which were in 2001 so that the study of sex ratio at 2001 and 2011 is appropriate. There is no change in the number of taluks between 2001 and 2011 and the sex ratio is calculated from 215 taluks at both 2001 and 2011 census.

**DISTRICT**

District in India are local administrative units inherited from the British Raj. They generally form the tier of local government immediately below that of India’s sub-national states and territories. Districts may further be grouped into administrative divisions, which form an intermediate level between the districts and the sub-national state (or union territory).


**TALUKS**

A Tehsil or tahril/tahasil, known as Taluka (or taluq/taluk) or mandal, is an administrative division of India, it is an area of land with a city or town that serves as its administrative center, with possible additional towns, and usually a number of villages.


**4.3 VARIABLE**

Sex ratio is the dependent variable for the study. The changes in the sex ratio over space and time are due to various forces operating in the human population.
4.4 METHOD OF ANALYSIS

MOBILITY OF GENDER POPULATION

Mobility is the process of movement of population from one size to other size. In the case of gender population size, mobility is the movement for gender population from one size class to other size classes. The movement of population size is random in nature and motivate to develop the Markov-Chain model for describing the movement of gender population size.

The following distributions are considered in the analysis of sex ratio and briefly presented here:

- Exponential distribution
- Truncated Exponential distribution
- Log-Normal distribution
- Markov-Chain model

4.4.1 Exponential Distribution

Definition

A continuous random variable $x$ is said to have the exponential probability distribution if its probability density function (p.d.f.) is given by

$$f(x) = \sigma e^{-(\sigma x)}; \quad 0 < x < \infty, \sigma > 0 ...(1)$$

$$= 0; \text{ elsewhere}$$
Another form of the p.d.f. is stated as,

\[
f(x) = \frac{1}{\sigma} e^{-\frac{x}{\sigma}} \quad ; \quad 0 < x < \infty, \sigma > 0 \quad \text{...(2)}
\]

\[
= 0 \quad ; \text{elsewhere}
\]

The exponential distribution is completely determined by the single parameter (\(\sigma\)).

**Distribution Function**

The Distribution Function \(F\) of \(x\) is,

\[
F(x) = 0 \quad \text{for} \quad x < 0
\]

\[
= \int_{0}^{x} \sigma e^{-\sigma t} dt
\]

\[
= 1 - e^{-\sigma x} \quad \text{for} \quad 0 \leq x
\]

\[
F(x) = 0 \quad \text{for} \quad x < 0
\]

\[
= 1 - e^{-\sigma x} \quad \text{for} \quad x \geq 0
\]

Note that \(P_r [a \leq x \leq b] = F(b) - F(a)\)

\[
= e^{-\sigma a} - e^{-\sigma b}.
\]
Mean and Variance of Exponential distribution

An exponential random variable $x$ with p.d.f. is given by

$$f(x) = \sigma e^{-\sigma x}; x < 0$$

$$= 0 \; ; \text{ elsewhere}$$

The Mean of the random variable $x$

$$E(x) = \sigma \int_{0}^{\infty} xe^{-\sigma x} dx$$

$$= [-xe^{-\sigma x}]_{0}^{\infty} + \int_{0}^{\infty} e^{-\sigma x} dx,$$

$$= 0 + \left[ -\frac{1}{\sigma} e^{-\sigma x} \right]_{0}^{\infty} \left[ e^{-\sigma x} \rightarrow 0 \; as \; x \rightarrow \infty \right]$$

$$E(x) = \frac{1}{\sigma}$$

The Variance of the random variable $x$

$$Var(x) = E(x^2) - [E(x)]^2$$

$$E(x^2) = \sigma \int_{0}^{\infty} x^2 e^{-\sigma x} dx$$

$$= [-x^2 e^{-\sigma x}]_{0}^{\infty} + 2 \int_{0}^{\infty} xe^{-\sigma x} dx,$$

$$= 0 + \frac{2}{\sigma^2} \cdot \int_{0}^{\infty} xe^{-\sigma x} dx = \frac{1}{\sigma^2}$$

$$E(x^2) = \frac{2}{\sigma^2}$$

$$Var(x) = \frac{2}{\sigma^2} - \frac{1}{\sigma^2} = \frac{1}{\sigma^2}.$$
The mean is $\frac{1}{\sigma}$ and the variance is $\frac{1}{\sigma^2}$

### 4.4.2 Truncated Exponential distribution

The exponential density function of the random variable $X$ is defined as

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & ; \lambda > 0, 0 < x < \infty \\ 0 & ; \text{ otherwise} \end{cases}$$

Since, the sex ratio distribution of gender discrimination is skew in nature, truncated exponential model was proposed to study the skew nature of the sexratio distribution of gender discrimination.

The truncated model for the random variable $X$ is defined as

$$g(x) = \frac{f(x)}{p(x \geq a)} ; x \geq a$$

In the case of exponential distribution it is obtained as

$$g(x) = \frac{f(x)}{p(x \geq a)} ; x \geq a$$

Where,

$$p(x \geq a) = \int_a^\infty \lambda e^{-\lambda x} \, dx$$

$$= \lambda \int_a^\infty e^{-\lambda x} \, dx$$

$$= \lambda \left[ \frac{e^{-\lambda x}}{-\lambda} \right]_a^\infty$$
The truncated model of the random variable $X$ is obtained as

$$g(x) = \frac{f(x)}{P(x \geq a)}; x \geq a$$

$$= \frac{\lambda e^{-\lambda x}}{e^{-\lambda a}}$$

$$= \lambda e^{-\lambda x + \lambda a}$$

The truncated model of the random variable $X$ is obtained as

$$g(x) = \lambda e^{-\lambda(x-a)}; x \geq a$$

The distribution of a random variable $X$ is said to be truncated on the right at the point $x = b$, if all the values of $x \geq b$ are discarded. Hence the probability density function of right truncated exponential distribution

$$f(y, \theta) = \theta e^{-\theta y} (1 - e^{-\theta b})^{-1} ; a \leq x \leq b$$

Left Truncated exponential model has been proposed to describe the sex ratio distribution of the gender discrimination in Tamil Nadu State.
\[ G(x) = p\{X \leq x\} \]
\[ = \lambda \int_a^x e^{-\lambda(t-a)} \, dt \]
\[ = \lambda e^{a\lambda} \int_a^x e^{-\lambda t} \, dt \]
\[ = \lambda e^{a\lambda} \left[ \frac{e^{-\lambda t}}{-\lambda} \right]_a^x \]
\[ = \lambda e^{a\lambda} \left( \frac{e^{-\lambda x} - e^{-a\lambda}}{-\lambda} \right) \]
\[ = -\left\{ e^{a\lambda} e^{-\lambda x} - e^{a\lambda} e^{-a\lambda} \right\} \]
\[ = -\{e^{-\lambda(x-a)} - e^0\} \]
\[ = -\{e^{-\lambda(x-a)} - 1\} \]

\[ G(x) = 1 - e^{-\lambda(x-a)}; x \geq a \]

where \( a \) is the threshold sex ratio of gender discrimination.

**Estimates of the parameters**

The parameters of the Truncated exponential distribution are estimated using method of maximum likelihood and presented as follows.

The p.d.f of Truncated Exponential distribution is

\[ g(x) = \lambda e^{-\lambda(x-a)}; x \geq a \]
Likelihood function is defined as

\[ L(\lambda; X) = \prod_{i=1}^{n} f(X_i; \lambda) \]

When random variable \( X_i \)'s are independently distributed.

\[ L(\lambda; x) = \lambda^n e^{-\lambda \sum_{i=1}^{n} (X_i - a)} \]

on logarithmic scale, it is observed as

\[ \log L = n \log \lambda - \lambda \sum_{i=1}^{n} (X_i - a) \]

The likelihood equation \( \frac{\partial \log L}{\partial \lambda} = 0 \) \( \Rightarrow \) solution as \( \hat{\lambda} \)

\[ \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} (X_i - a)} \]

Where \( n = \sum f \)

The Maximum Likelihood Estimate of \( \lambda \) is

\[ \hat{\lambda} = \min_i (x_i) \quad \text{and} \quad \hat{\lambda} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} f_i (x_i - a)} \]

### 4.4.3 Log-Normal Distribution

Log-Normal distribution is one of the skew distributions. It has the characteristics, decreasing a long upper tail or lower tail. Skewness of the distribution is
described by the existence of non-vanishing odd-order moments. It has been described as follows:

Let \( X \) be a positive Random variable, let a new random variable \( Y \) be defined as \( Y = \log_e X \). If \( Y \) has a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), then \( X \), is said to have Log-Normal distribution. Then the density function is given by

\[
f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} ; \quad x \geq 0; \quad -\infty < \mu < \infty; \quad \sigma > 0
\]

It arises in problems of Economics, Biology, Geology and reliability theory. In particular, it arises in the study of dimensions of particles under pulverization. Statistical measures are obtained as

**Mean and Variance of Log-Normal distribution**

The \( r \) th order moment from mean origin is obtained.

\[
\mu_r = E[X^r] = \int_0^\infty x^r f(x; \mu, \sigma) \, dx
\]

\[
= \int_0^\infty x^r \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} \, dx
\]

When

\[
y = \log x \quad \Rightarrow \quad dy = \frac{dx}{x}
\]

\[\text{ie., } e^y = e^{\log x} = x\]

Limits

\[\text{when } x = 0; \quad x = \infty; \quad y = -\infty; \quad y = \infty\]
\[ \mu'_r = E[X^r] = \int_{-\infty}^{\infty} e^{\gamma^r} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \, dy \]

\[ \mu'_r = E[X^r] = \int_{-\infty}^{\infty} e^{\gamma^r} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \, dy \]

The variable \( y \) is transformed by the relation.

\[ \frac{y-\mu}{\sigma} = z \Rightarrow y = z\sigma + \mu \]

\[ \frac{dy}{\sigma} = dz \Rightarrow dy = \sigma \, dz \]

\[ E[X^r] = \int_{-\infty}^{\infty} e^{(\mu+\sigma z)^r} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2} \sigma^2} \sigma \, dz \]

\[ = e^{r \mu} \int_{-\infty}^{\infty} e^{z\sigma r} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2z^2}} \, dz \]

\[ = \frac{e^{r \mu}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[z^2-2xr]} \, dz \]

\[ = \frac{e^{r \mu}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-r\sigma)^2-r^2\sigma^2} \, dz \]

\[ = \frac{e^{r \mu \frac{r^2\sigma^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-r\sigma)^2} \, dz \]

\[ = e^{r \mu \frac{r^2\sigma^2}{2}} \sqrt{2\pi} \left[ \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-r\sigma)^2} \, dz = \sqrt{2\pi} \right] \]

\[ \mu'_r = E[X^r] = e^{r \mu \frac{r^2\sigma^2}{2}} \]

when \( r = 1, \) in \( E[X^r] \)

\[ \mu'_1 = \text{Mean} = E(X) = e^{\mu \frac{\sigma^2}{2}} \]
when $r = 2$, in $E[X^r]$

$$\mu'_2 = E(X^2) = e^{2\mu + 4\sigma^2}$$

$$= e^{2\mu + 2\sigma^2}$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$V(X) = E(X^2) - [E(X)]^2$$

$$= e^{2\mu + 2\sigma^2} - \left[ e^{\mu + \frac{\sigma^2}{2}} \right]^2 \quad [\because (e^m)^n = e^{mn}]$$

$$= e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2}$$

$$= e^{2\mu} [e^{2\sigma^2} - e^{\sigma^2}]$$

$$V(X) = e^{2\mu + \sigma^2}[e^{\sigma^2} - 1]$$

The mean and variance are obtained as

$$E(X) = e^{\mu + \frac{\sigma^2}{2}} \text{ and } V(X) = e^{2\mu + \sigma^2}[e^{\sigma^2} - 1]$$

While applying lognormal distribution practically, the best way is to transform it by taking its natural logarithm, $Y = \log X$. $Y$ would be normally distributed with mean $\mu$ and variance $\sigma^2$.

**Estimates of the parameters in the mean and variance**

Parameters of Log-Normal distribution are estimated using maximum likelihood method as

The p.d.f of Log- Normal distribution is

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} ; x > 0 ; -\infty < \mu < +\infty ; \sigma > 0$$
Likelihood function is defined as

$$L(\mu, \sigma^2; x) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2)$$

When random variable $x_i$’s independently distributed

$$L(\mu, \sigma^2; x) = \left(\frac{1}{x\sqrt{2\pi} \sigma} \right)^n e^{-\frac{(\log \frac{x}{\sigma} - \mu)^2}{2\sigma^2}}$$

On logarithmic scale, it is observed as

$$\log L = n \log \frac{1}{x\sqrt{2\pi} \sigma} - \frac{(\log \frac{x}{\sigma} - \mu)^2}{2\sigma^2}$$

The likelihood equation $\frac{\partial \log L}{\partial \mu} = 0 \Rightarrow \text{solution as } \hat{\mu}$

i.e., $\Rightarrow \frac{\Sigma \log e x_i}{\sigma^2} - \frac{n \mu}{\sigma^2} = 0$

$\Rightarrow \Sigma \log e x_i - n \hat{\mu} = 0$

$\Rightarrow \Sigma \log e x_i = n \hat{\mu}$

$\hat{\mu} = \frac{\Sigma \log e x_i}{n}$ where n= $\Sigma f$

$\hat{\mu} = \frac{\Sigma \log e x_i f_i}{\Sigma f_i}$

The likelihood equation $\frac{\partial \log L}{\partial \sigma^2} = 0 \Rightarrow \text{solution as } \hat{\sigma}^2$

i.e., $\Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \Sigma (\log e x_i - \mu)^2 = 0$

$\Rightarrow \frac{1}{2\sigma^4} \Sigma (\log e x_i - \mu)^2 = \frac{n}{2\sigma^2}$

$\Rightarrow \frac{1}{\sigma^2} \Sigma (\log e x_i - \mu)^2 = n$

$\Rightarrow \frac{\Sigma (\log e x_i - \mu)^2}{n} = \sigma^2$

$\hat{\sigma}^2 = \frac{\Sigma (\log e x_i)^2}{n} - \hat{\mu}^2$

$\hat{\sigma}^2 = \frac{\Sigma (\log e x_i)^2 f_i}{\Sigma f_i} - \hat{\mu}^2$ where n= $\Sigma f$
The maximum likelihood estimate of $\mu$ and $\sigma^2$ is

$$\hat{\mu} = \frac{\sum \log e x_i f_i}{\sum f_i} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum (\log e x_i)^2 f_i}{\sum f_i} - \hat{\mu}^2$$

### 4.4.4 Markov-Chain

A family of random variable $\{X_n\}$ is a stochastic process where the value of the random variable $\{i \in I\}$ and $\{n \in N\}$ are called states space and parametric space respectively.

The stochastic process $\{X_n\}$ is said to be a Markov-Chain, if for all $i_0, ... i_{n+1} \in I$, and $\forall n$

$$P[X_{n+1} = i_{n+1} | X_0 = i_0, ... X_n = i_n] = P[X_{n+1} = i_{n+1} | X_n = i_n] \quad \rightarrow (1)$$

The above equation is also called as Markov property.

The space $I$ is called the state space and $i_j$ are called as the states of the Markov-Chain. If $I$ is finite, then the Markov-Chain is finite. Markov property implies that the future value depends only on the present value and not on the past values.

**Transition probability matrix**

The transition probabilities $P_{jk}$ satisfy $P_{jk} \geq 0, \sum_k P_{jk} = 1$ for all $j$,

The transition probabilities may be written in the matrix form as
This is called the transition probability matrix (t.p.m) of the Markov -Chain.  P is a stochastic matrix. i.e., a square matrix with non-negative elements with unit row sums.

Chapman- Kolmogorov equation. The m-step transition probability is described as

\[ P_{jk}^{(m)} = \Pr\{X_{n+m}=k/X_n=j\} \]

It means probability of transition from the state j to the state k in exactly m steps.

When m=2 \[ P_{jk}^{(2)} = \Pr\{X_{n+2}=k/X_n=j\} \]

This is called 2-step transition probability.

In general,

\[ P_{jk}^{(m+n)} = \sum P_{rk}^{(n)} P_{jr}^{(m)} = \sum P_{jr}^{(n)} P_{rk}^{(m)} \]

This equation is called as Chapman- Kolmogorov equation. It may be also represented in the matrix form as follows

\[ p^{(2)} = p \cdot p = p^2 \]

\[ p^{(m+1)} = p^m \cdot p = p \cdot p^m \]

\[ \vdots \]

\[ p^{(m+n)} = p^m \cdot p^n = p^n \cdot p^m \]
4.4.5 MAXIMUM LIKELIHOOD METHOD

The random variables $X_1, X_2, \ldots, X_n$ are distributed independently with density function $f(X, \theta)$. Then the likelihood function of the sample values $X_1, X_2, \ldots, X_n$ usually denoted by $L = L(\theta; X)$ is there joint density function given by

$$L = f(X_1, \theta)f(X_2, \theta) \ldots \ldots f(X_n, \theta)$$

$$= \prod_{i=1}^{n} f(X_i, \theta)$$

$L$ gives the relative likelihood that the random variables assume a particular set of values $X_1, X_2, \ldots, X_n$. For a given sample $X_1, X_2, \ldots, X_n$, $L$ becomes a function of the variables $\theta$ the parameter.

The principle of maximum likelihood consists in finding an estimation for the unknown parameter $\theta = \theta_1, \theta_2, \ldots, \theta_k$ say, which maximizes the likelihood functions $L(\theta)$. For variations in parameter (i.e.) we wish to find

$$\hat{\theta} = \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$$

so that

$$L(\hat{\theta}) > L(\theta) \text{ for all } \theta \in \Omega$$

$$L(\hat{\theta}) > \text{Sup } L(\theta) \text{ for all } \theta \in \Omega$$

Thus, if there exists a function $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ of the sample values which maximizes $L$ for variations in $\theta$ then $\hat{\theta}$ is to be taken as an estimator of $\theta$. $\hat{\theta}$ is usually called maximum likelihood estimator (MLE) and it is the solution of

$$\frac{\partial L}{\partial \theta} = 0; \forall \theta \in \Omega$$
Since $L > 0$ and $\log L$ is a non-decreasing function of $L$; $L$ and $\log L$ attain their extreme values of same values of $\hat{\theta}$. 

\[
\left(\frac{1}{L}\right) \frac{\partial L}{\partial \theta} = 0 \Rightarrow \frac{\partial \log L}{\partial \theta} = 0
\]

which is a Likelihood equation

If $\theta$ is vector valued parameters, then

\[
\hat{\theta} = \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k \text{ is given by the solution of simultaneous likelihood equations.}
\]

\[
\frac{\partial \log L}{\partial \theta_i} = 0; i = 1, 2, \ldots, k
\]

4.4.6 $\chi^2$ – Test

The square of a standard normal variate is known as a chi-square variate with 1 degrees of freedom.

Thus, if $X \sim N(\mu, \sigma^2)$ then, $Z = \left(\frac{X - \mu}{\sigma}\right) \sim N(0, 1)$ and $Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2$ is a chi-square variate with 1 degrees of freedom.

In general, if $X_i$, $(i = 1, 2, \ldots, n)$ are ‘n’ independent normal variates with mean $\mu_i$ and variance $\sigma_i^2$, $(i = 1, 2, \ldots, n)$ the $\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ is a chi-square variate with $n$ degrees of freedom. It is used to test the goodness of fit.
Prof. Karl Pearson has developed a method to test the difference between the hypothesis and the observed value. The test is done by comparing the computed value of $\chi^2$ with the table value of $\chi^2$ for the desired degrees of freedom. A greek centre is used to describe the magnitude of difference between the fact and theory.

The $\chi^2$ may be defined as,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(n-1)} d.f$$

$O_i =$ Observed frequencies \\
$E_i =$ Expected frequencies \\

For the validity of chi-square test of “goodness of fit” between theory and experiment, the following conditions must be satisfied.

- The sample observations should be independent.
- All the events must be mutually exclusive.
- There must be large observations.
- For comparison purpose, the data must be in original units.
4.4.7 Diagrams and graphs are used in the presentation of data.

Diagrams

Diagrams are based on scale but nor confined to points or lines. There are various geometrical shapes such as bars, circles, square etc. Diagrams are visual presentations of categorical and geographical data. It furnishes only approximate information, diagrams are more appealing to the eyes and even laymen can understand the concept easily under study.

Graphs

Graphs are more appropriate to represent the time series data and the frequency distribution. Graphs are more precise and accurate than the diagrams. Graphs can be effectively used to study the slopes, rate of change and forecasting. Frequency curve, Ogive curve, Trend line etc., are graphs representing the data relates to the study.