

Chapter 1: Introduction

1.1 Motivation

An important area of statistical applications is the estimation of mortality. The study of mortality in a population is essential for the knowledge of the evolution of its main indicators (life expectancy, infant mortality, etc.) as well as to make demographic projections. Mortality is an indicator of the situations involving health as well as the living conditions and the aging process of a population. The study of mortality is useful for analysing current demographic conditions and for determining the prospects of potential changes in mortality conditions of the future. In this chapter, some fundamental definition relating to our work and the mortality models for the graduation of mortality are introduced.

1.2 Life Table

The investigation of mortality which is the oldest subject in demography is first brought under systematic and rigorous analysis through the idea of a life table. A more entire picture of mortality is given by a life table which demonstrates the death rate independently for each age. A life table is a statistical device used by actuaries, demographers, public health workers and others to present the mortality experience of a population aggregate in a form that permits answering many related questions on

mortality [47]. It combines the mortality experience of a population at different ages in a single statistical model. The life table gives the life history of a hypothetical group of persons all born at the same point of time as this is gradually diminished by deaths. The life-table reveals to us what number of those persons survived to attain each birth day, the probability of surviving from one age to another, the person years lived at every age, the mortality rates and the expectation of life. Life expectancy of persons at various ages is derived from life tables and demonstrates the average number of years that persons (at birth or at a particular age) might expect to live if these mortality rates are constant over their lifetime. The derived life expectancies give an indication of the average longevity of the population, but do not necessarily reflect the longevity of an individual. The first life tables were created by John Graunt in 1662 [26] and Edmund Halley in 1693 [27].

Though a life table is designed essentially to measure mortality, it is employed by a variety of specialist in a variety of ways. The analyses of life table are important to various disciplines and fields, including health, demography, and actuary. It can be utilized for examining population growth and in addition for making projections of population size and structure. It is also used in studies of longevity, fertility, migration and population growth, as well as in studies of widowhood, orphan hood, length of married life, length of working life, and length of disability free life. In the simplest form of a life table, the entire table is generated from age-specific mortality rates and the resulting values are used to measure mortality, survivorship, and life expectation. In other applications the mortality rates in the life table are combined with other demographic data into a more complex model which measures the combined effect of mortality and changes in one or more socioeconomic characteristics, e.g., a table of working life, which combines mortality rates and labour force participation rates and measures their combined effects on working life [53].

Life tables can be classified in two ways as -

1. Complete life table.
2. Abridge life table.

A complete life table is usually one in which the values of the life table functions are given in single years of age. On the other hand, an abridged life table is one in which most functions are given only for certain pivotal ages, frequently spaced at five or ten year intervals after infancy. The principal columns in a life table are $l_x, {}_n d_x, {}_n q_x, {}_n L_x, T_x, e^0_x$ [35]. For abridged life table n is taken as 5 or 10 years. The variable 'x' means exact age x in this column. In case of complete life table the column is designed by only age x where $x = 0, 1, \dots, w$; where w is the highest attained age. The description of the life table functions of a life table are given below-

l_x : The number of persons living at exact age x at the beginning of the interval (x to $x + n$) out of total number of births given by the radix of the life table. This column starts with l_0 , the size of the birth cohort, i.e. the radix l_x is a decreasing function of age.

${}_n d_x$: The number of deaths out of l_x persons during the period of next n years.

${}_n q_x$: The probability of dying before reaching age $x + n$ for a person who is of exact age x . In this thesis, the symbol q_x has been used instead of ${}_1 q_x$ in case of a complete life table. Then q_x implies the probability that a person who has reached age x would die during the interval ($x, x + 1$). Sometimes one can calculate another life table function p_x as complement of q_x .

Because the columns of the life table are so closely related to one another, there are many arithmetic relations among the columns. For example:

$${}_n d_x = l_x - l_{x+n},$$

$${}_n q_x = \frac{{}_n d_x}{l_x} = 1 - \frac{l_{x+n}}{l_x},$$

$${}_n p_x = \frac{l_{x+n}}{l_x} = \frac{l_x - {}_n d_x}{l_x} = 1 - {}_n q_x .$$

${}_n L_x$: The number of person years lived by the l_x persons during the interval ($x, x + n$). If the interval is of one year, then it is denoted by L_x . If ${}_n a_x$ denotes the mean number of

person-years lived in the interval by those dying in the interval then ${}_nL_x$ can be written as:

$${}_nL_x = n \times l_{x+n} + {}_nd_x \times {}_na_x.$$

If the deaths at any age interval x are evenly distributed, the function l_x is linearly distributed over these ages and hence approximated as

$$L_x = \frac{1}{2}(l_x + l_{x+1}),$$

$${}_nL_x = \frac{n}{2}(l_x + l_{x+n}) \text{ for } x > 2.$$

Since linear relationship is not valid for ages 0 and 1, the value of L_0 becomes

$$L_0 = {}_1a_0 l_0 + (1 - {}_1a_0) l_1.$$

Often ${}_1a_0$ is taken as a value between 0.2 and 0.3 and, for instance, if ${}_1a_0 = 0.3$ then

$$L_0 = 0.3l_0 + 0.7l_1.$$

Similarly, L_1 can be written as $L_1 = {}_1a_1 l_1 + (1 - {}_1a_1) l_2$, often ${}_1a_1$ is taken as a value of 0.4 in which case,

$$L_1 = 0.4l_1 + 0.6l_2.$$

T_x : Total life time after age x . This is total number of person years lived by the survivors l_x in the future. This is given by the cumulative sum of ${}_nL_x$ values after age x i.e.,

$$T_x = {}_nL_x + T_{x+n},$$

and for single year age groups $T_x = L_x + T_{x+1}$.

For the last (open ended) age group, y , T_y is obtained as

$$T_y = \frac{l_y}{m_y}.$$

e^0_x : The expectation of life at exact age x . This is the average number of years to which the survivors l_x are expected to live. This is given by

$$e^0_x = \frac{T_x}{l_x}.$$

If $x = 0$, then $e^0_0 = \frac{T_0}{l_0}$, is the expectation of life at birth.

Mortality conditions of a period can be operationalized by converting the set of period age-specific death rates, ${}_n m_x$ into a set of age-specific probabilities of dying ${}_n q_x$. A basic formula to derive ${}_n q_x$ from age-specific death rates ${}_n m_x$ is [11]:

$${}_n q_x = \frac{n \times {}_n m_x}{\{1 + (n - {}_n a_x)\} \times {}_n m_x}.$$

This means for a cohort, the conversion from ${}_n m_x$ to ${}_n q_x$ depends on only one parameter: ${}_n a_x$, the average number of person-years lived in the interval by those dying in the interval. No other information is required to perform this conversion, and any other information is redundant [49]. If persons dying in the interval do so, on average, half-way through the interval, then Chiang's equation immediately becomes:

$${}_n q_x = \frac{(n \times {}_n m_x)}{\left\{1 + \left(\frac{n}{2}\right) \times {}_n m_x\right\}} = \frac{2n \times {}_n q_x}{2 + n \times {}_n m_x}.$$

1.3 Mortality Modelling by Model Life Tables

A typical life table, giving set of probabilities of dying at different ages (${}_n q_x$) corresponding to a given level of mortality (e^0_0) is called model life table. Model life tables are very useful in providing estimates of overall mortality conditions in countries for which vital registration is incomplete or of lesser quality. A set of model life tables ideally depicting the typical age-patterns of mortality for different levels may be of immense help to construct life tables for such countries if the level of mortality is indirectly estimated. There have been many attempts at developing model life tables. United Nations made the first attempt to develop model life tables [58]. Those in common use are Coale and Demeny [13], Ledermann [39] United Nations [58].

The first United Nations (U.N.) model life tables were prepared on the basis of 158 life tables collected from a wide selection of countries and representing different periods of time (United Nations, 1955). By fitting through the method of least squares, a second degree parabola of the type $y = a + bx + cx^2$, a relationship was developed between the mortality rates of the successive age groups. So each successor was estimated from the predecessor q_x value. Here q_0 becomes the core input and thereafter successive q_x values serve as input for further estimations. A system of model life tables were thus derived corresponding to the values of $q_0 = 20, 25, \dots, 100$ and thereafter $q_0 = 110, 120, \dots, 330$. This set of U.N. model life tables covered the range of e^0_0 from 20 to 73.9 years (for males and females combined). Up to age 55, the tables are spaced at intervals of 2.5 years in terms of e^0_0 . Intervals beyond $e^0_0 = 55$ have been used to reflect a hypothetical typical course of mortality decline over time.

A set of model life tables was developed by Coale and Demeny [13]. These models classify the life tables into four different sets, labeled West, East, North, and South, according to the patterns of mortality in the predominant regions of Europe represented in the original data. Methodologically, in each of these sets, the life expectancy at age 10 (e_{10}) was correlated with the probability of dying at different ages (${}_nq_x$), and these correlations provided the basis for estimating a series of “nested” life tables at different overall levels of e^0_x but with different age patterns of mortality.

Ledermann [39] published seven sets of model life tables, each with one and two parameters, by using 154 individual life tables as input. Ledermann's tables are more refined sets of model life tables than the United Nations and Coale and Demeny's model life tables, but they are not easy to use as the choice of the life table becomes difficult when the input parameters cannot be reliably estimated.

1.4 A Review of Parametric Models for Graduation of Mortality

The parametric graduating of mortality by different laws of mortality was addressed by various researchers [[16], [23], [46], [50], [3], [29],[57]]. Mostly the mathematical models are deterministic. Such modelling not only serves the purpose of smoothing the irregular fluctuations in data but also provides some parameters for comparison.

Mathematical models with fewer parameters (parsimony) are superior as long as they are capable of representing the observed pattern.

In 1725 Abraham De Moivre [16] suggested that the probability of survival from birth until age x could be expressed as a linear function of age. In terms of the force of mortality rate, the model can be written as

$$\mu_x = \left(\frac{1}{w-x}\right) \text{ for } 0 \leq x < w, \quad (1.1)$$

where w is the highest attainable age. De Moivre used this model to make various actuarial calculations. But the model did not give accurate representation of human survival across all ages.

The first mathematical formula describing the mortality rates was proposed by Benjamin Gompertz in 1825 [23]. In 1825, British actuary Benjamin Gompertz made a simple but important observation: "A law of geometrical progression pervades, in an approximate degree, large portions of different tables of mortality." This observation was based largely on observed death and population records for people in England, Sweden, and France between ages 20 and 60 in the nineteenth century. The simple formula describing the exponential rise in death rates between sexual maturity and extreme old age, $[\gamma(t) = \exp(\gamma t)]$ [48], is now commonly referred to as the Gompertz equation. Basically the "law" was derived by postulating a relationship between the rate of change of the force of mortality at any age and its value at that age. Gompertz modelled the aging or senescent component of mortality with two parameters: a positive parameter ' a ' that varies with level of mortality and a positive parameter ' b ' that measures the rate of increase in mortality with age. The mathematical model of Gompertz expressed in terms of the force of mortality is

$$\mu_x = ae^{bx}. \quad (1.2)$$

Gompertz's law constitutes one of the most influential proposals in the early times of survival modelling. Actually, many contributions in the field of mortality laws,

throughout the latter half of the 19th century, generalize or proceed from Gompertz's ideas. The Gompertz model of mortality focused on older ages (beyond infant and young adult periods where accidental deaths have a major contribution to mortality rates). This is seen as a weakness of the model and focusing on the problem of representing the mortality over the whole lifetime span. The paper by S. Jay Olshansky and Bruce A. Carnes [48] reviewed the literature on Gompertz's law of mortality and discussed the importance of his observations. For many purposes the Gompertz model provided a satisfactory fit to adult mortality rates. However, close inspection of the difference between model estimates and observed death rates reveals systematic underestimation of actual mortality at youngest adult ages (under 40) and overestimation at the oldest ages (over 80) [7]. Gompertz's law fits observed mortality rates very well at the adult ages, and it is a good tool for comparing mortality tables, as Wetterstrand [63] demonstrated. Brilinger argued that if the human body was considered as a series system of independent components, then the force of mortality may follow Gompertz law. The paper by Jack C. Yue [70], proposed a standard operating procedure for testing the Gompertz assumption using yearly age-specific mortality data. The paper by W.H. Wetterstrand [63] discussed the use of Gompertz's law to describe the mortality between the ages of 30 and 90. R.E Beard [3] in his paper was mainly concerned with-

- (a) Accidental deaths (assumed to be at a constant rate at all ages),
- (b) An upper limit to the rate of mortality, and
- (c) A progression in time.

Actually, many contributions in the field of mortality laws, throughout the latter half of the 19th century, generalize or proceed from Gompertz's ideas. The law of Makeham [46], which improves the Gompertz law by adding an extra parameter to this model to take into account the force of accidental death, assumed to be a constant independent of age. Makeham noticed that Gompertz's model was not adequate for higher ages and amended it in an effort to correct this deficiency [24], [46]. The force of mortality in the Makeham model is-

$$\mu_x = c + ae^{bx}. \quad (1.3)$$

The constant c can be explained as the risk of death from all causes which do not depend on age. The Makeham model represented a clear improvement over the Gompertz model at younger ages, but it still overestimates at the oldest ages [7].

The next significant modification to the Makeham law was the system of curves devised by Perks [49] and of which the important formula was the Logistic. Many human life-tables have been graduated by this basic curve. The Logistic model is known under a variety names. It was first discovered by Perks [49], who found empirically that the values of μ_x in a life table which he was examining could be fitted by a certain curve, which was in fact a logistic function (though he did not describe it such a time). The deviation of Makemam model can be addresses in a number of ways, most simply by the following logistic model [56], [57]:

$$\mu_x = \frac{ae^{bx}}{1 + de^{bx}} + c. \quad (1.4)$$

At lower adult ages the force of mortality estimated with models (1.3) and (1.4) are very similar, because the denominator of the first term in (1.4) is close to 1.0. At the oldest ages, however, the two models diverge as (1.4) levels off at $1 + c$ while ((1.3) has no limit. More complex Logistic models with additional parameters were also proposed [[56], [57]]. On the basis of detailed comparison of different models, Thatcher [57] and Thatcher et al. [56] recommend (1.4) because it provided an excellent fit to mortality rates over the entire adult age range with relatively few parameters.

By assuming that the parameter $c = 0$ in the logistic model, Beard [4] obtained the 3-parameter model as

$$\mu_x = \frac{ae^{bx}}{1 + de^{bx}}. \quad (1.5)$$

In 1992, Kannisto [34] noticed that the modern data for μ_x at high ages are very close to one of the simplest forms of the Logistic model, in which *logit* (μ_x) is a similar function of x . This fact was also used independently by Himes, Preston and Condran in 1994 [31]. The force of mortality in the Kannisto model is-

$$\mu_x = \frac{ae^{bx}}{1 + ae^{bx}} + c. \quad (1.6)$$

For Kannisto's model, Doray [18] proposed a weighted least-squares estimator which can easily be calculated with any regression software; the estimator is shown to be consistent, asymptotically unbiased and normally distributed.

It is observed that the Gompertz ($c = 0$; $d = 0$), Makeham ($d = 0$), Beard ($c = 0$) and Kannisto ($c = 0$; $d = a$) models are all special cases of the Logistic model. The paper by Doray [19] discussed that Logistic type models for the force of mortality provides a better fit to mortality data of people aged over 85 than Makeham's models where the force of mortality increases exponentially with age.

A descriptive model used by Coale & Kisker [14] in a limited range of ages. They fitted $\ln(\mu_x)$ by a quadratic function of over x for the purpose of interpolating in the range of ages from 85 to 110. The force of mortality in the Coale & Kisker model is-

$$\mu_x = ae^{bx}k^{x^2}, \quad (1.7)$$

$$\Rightarrow \ln \mu_x = A + bx + Cx^2, C < 0, \quad (1.8)$$

where $A = \ln a$ and $C = \ln k$.

This model is also known as quadratic model. Below age 85, this model would conflict with findings of Horiuchi & Coale [32]. Wilmoth [64] used the model for estimating μ_x at age 110 from data which extended above age 85.

Thatcher et al. [56] fitted the Gompertz, logistic, Kannisto, and Weibull models also two descriptive models Heligman and Pollard [29] and the quadratic model to mortality data of aged people in 13 industrialized countries for the periods 1960-70, 1970-80, 1980-90 and for the cohort born in 1871-80 by using maximum likelihood method. The data utilized were deaths at ages 85 and over for the quadratic model and ages 80 and over for all the other models. The best fit was reliably given by the Kannisto and logistic models for all countries in every period and for the cohort data. All the models mentioned above create close estimations of μ_x at ages 80 to 95. After age 95, the Gompertz and Makeham forces of mortality continue to increase exponentially with age, while for the Kannisto, Beard and logistic models, μ_x tends asymptotically to a constant as x increases.

A recent attempt to represent mortality over the course of the entire life span, using a single analytical expression, was made by Heligman-Pollard model [29]. The idea behind the H-P model is that the cause of death can be divided into three classes, namely those affecting childhood, early and middle adult life, and old age. The mathematical function H&P suggest is given by the formula

$$\frac{q_x}{p_x} = A^{(x+B)^c} + D \exp\left(-E \left(\ln \frac{x}{F}\right)^2\right) + GH^x. \quad (1.9)$$

Here q_x is the probability of dying within one year for a person aged x exactly, and $p_x = 1 - q_x$. A, B, C, D, E, F, G, H are positive parameters. The curve is continuous and applicable over the entire age range $0 \leq x < \infty$ and allows q_x lies between 0 and 1 only. The HP model contains three terms, each representing a distinct component of mortality. The first, a rapidly declining exponential, reflects the fall in mortality during the early childhood years as the child adapts to its new environment and gains immunity from the diseases of the outside world. This component of mortality has three parameters A, B, C . The third term in the formula, the well-known Gompertz exponential, reflects the near geometric rise in mortality at the adult ages, and is generally considered to represent the ageing or deterioration of the body, i.e., senescent mortality. The remaining term, a function similar to the lognormal, reflects accident mortality for males and accident plus maternal mortality for the female population; i.e., additional mortality superimposed on

the ‘natural curve of mortality as described by the other two terms. The ‘accident hump’ is found in all populations, and appears either as a distinct hump in the mortality curve or at least ages 10 and 40. The accident term has three parameters D, E, F . In order to fit the H&P formula to the empirical q_x - values in a complete life table, the parameters of the model need to be estimated. Letting $Q(x, c)$ to denote the right hand side of (6), it is reasonable to expect that the parameters can be estimated by minimizing the sum of squares:

$$\sum_x \left(\frac{q_x}{p_x} - Q(x, c) \right)^2, \quad (1.10)$$

with respect to $c = (A, B, C, D, E, F, G, H)$. Hartmann [28] explained that this procedure cannot be used because it will give a negative estimate of B , which is not permitted. In order to estimate the parameters Heligman-Pollard minimized the function was

$$S^2 = \sum_x \left(\frac{q_x}{\hat{q}_x} - 1.0 \right)^2, \quad (1.11)$$

where q_x is fitted value at age x and \hat{q}_x is the observed mortality rate. That is, the sum of the squares of the proportional difference between the fitted and observed rates was minimized. The paper by [29] estimated the parameters by least squares using Gauss Newton iteration procedure. The paper by Kostaki [37] used a nonlinear least squares algorithm which was based upon a modification of the Gauss-Newton iteration method in order to estimate the parameters. The paper by Ibrahim [33] used Levenberg-Marquardt iteration procedure in order to estimate the parameters of the H&P model. The Heligman-Pollard model as discussed above is quite complex, contain large number of parameters and often do not yield stable solutions. Though the Heligman-Pollard model fits the entire age range, the parameters pertaining to the middle term often do not converge. Congdon (1993) argued that over parameterization is a concern in the Heligman-Pollard model. The model is also valid for that data set only, which provide single year probability of dying (or age-specific death rate for single year).

1.5 Forecasting Modelling: Lee Carter Model

In the early 1990's researchers began to look at modeling mortality using time series to extrapolate the time trend based on historic mortality experience. These sorts of models make the implicit assumption that past trends identified in the data will continue into the future. The first and most recognized of these types of models is the Lee Carter mortality model [40] which models the time trend using a one factor stochastic model. The Lee and Carter model, published in 1992, was the first attempt to model longevity data in a stochastic fashion by fitting the past mortality data and modeling the time trend as a stochastic process. The benefit of this for an actuary or an end user of mortality forecasts is that the uncertainty associated with mortality forecast can also be visualized as well as the mean or expected value. The Lee Carter model takes no account of cause of death or any explanatory modeling of mortality. Instead it models the data as a stochastic time series. The Lee-Carter model became one of the most well-known models and it is applied in different countries around the world to forecast age specific death rates. The model was developed as a simple one factor model which ensured the plausibility of projected age patterns. The Lee-Carter model has a relatively simple formulation as a one-parameter family of life tables, one for each age x . The model postulated by Lee and Carter is given by:

$$\ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t},$$

where

$m_{x,t}$: observed central death rate at age x in year t

a_x : average age-specific pattern of mortality

b_x : pattern of deviations from the age of profile as the k_t varies

k_t : a time-trend index of general mortality level

$\varepsilon_{x,t}$: the residual term at age x and time t .

The time component k_t captures the main time trend on the logarithmic scale in mortality rates at all ages. The model includes no assumption about the nature of the trend in k_t . The age component b_x modifies the main time trend according to whether change at a particular age is faster or slower than the main trend. In principle, not all the b_x need have the same sign, in which case movement in opposite directions could occur. In practice, all the do have the same sign, at least when the model is fit over fairly long periods. The model assumes that b_x is invariant over time.

The various mortality models have been studied and tried to fit Assam mortality data. But some models fail to fit our data and those models are not considered here. The selection of models is based on an extensive study of the proper literature and examinations with other workers in this area.

1.6 Objectives

In this research work, the main aim is to investigate the suitable mortality model for Assam population. To achieve the main goal of the study to be presented in the thesis the following objectives have been undertaken.

1. To construct the complete life tables of 23 districts of Assam for both male and female.
2. To select the best-fit mortality model for extrapolating survivors in a life table up to the last age for total, rural and urban area Assam population for both the genders.
3. To compare different parametric models and to find a suitable mortality model at some selected age groups 20-60 and 60-100 for people of mortality of districts of Assam for both male and female.
4. To find a suitable mortality model for projection of oldest-old force of mortality rates based on our constructed complete life table of Assam for the period 2009-13.
5. To investigate the feasibility of using the Lee-Carter model for projecting mortality for the Assam population for total, male and female.

1.7 Outline and Organization of the Thesis

With relevant to the objectives mentioned above, the thesis comprised of six chapters.

The first introductory chapter gives the fundamental definition and introduces the mortality models which are used to find a suitable model for Assam population.

In chapter 2, a method has been introduced for constructing complete life tables for 23 districts of Assam for both male and female. The name of 23 districts of Assam considered here are Kokrajhar, Dhubri, Goalpara, Bongaigaon, Barpetra, Kamrup, Nalbari, Darang, Marigaon, Nagaon, Sonitpur, Lakhimpur, Dhemaji, Tinsukia, Dibrugarh, Sibsagar, Jorhat, Golaghat, Karbi angling, North Cachar, Cachar, Karimganj and Hailakandi. These life tables could be used for the study of mortality, longevity, fertility, and population growth, as well as in making projections of population.

In chapter 3, a comparison has been made between Gompertz and Makeham model for extrapolating survivors in a life table past beyond the last age for Assam for total, rural and urban population for both the genders.

In chapter 4, the most commonly used parametric mortality models namely Gompertz, Makeham, Logistic and Beard have been used for describing the mortality pattern of districts of Assam. The mortality pattern is divided by two age groups 20-60 and 60-100 for the purpose of separating working years and retirement so that separate laws of mortality can apply to these groups. In this chapter, a suitable mortality model is found for people of mortality of districts of Assam. Ten districts have a sufficiently reliable data to be useful for the specialized purpose of the present analysis: five high population growth rate districts: Darang, Dhubri, Goalpara, Morigaon and Nagaon and five low population growth rate districts: Golghat, Jorhat, Sibsagar, Dibrugarh, and Tinsukia.

In chapter 5, six mortality models: Gompertz, Makeham, Logistic, Kannisto, Beard and Coale-Kisker models have been examined for the oldest-old force of mortality rates for Assam.

In chapter 6, the feasibility of the Lee Carter model has been examined for projecting the mortality for the Assam population for total, male and female. The model has been fitted to the matrix of Assam death rates based on 15 years data separately for Assam male and female populations in the form of life tables for the period 1995-99 to 2009-13. The

Singular Value Decomposition (SVD) methodology has been applied to estimate the parameters of the model. A time-varying index of mortality is forecasted up to 2025 year using random walk drift (RWD) model and is used to generate projected life tables.
