ISSN:0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 8, Issue, 10, pp. 20813-20824, October, 2017

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

TWO- COMPONENT OF NON- IDENTICAL MIXTURE DISTRIBUTION MODELS FOR HETEROGENEOUS SURVIVAL DATA

Uma maheswari R and Leo Alexander T*

Department of Statistics, Loyola College, Chennai–34, India

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0810.0965

ARTICLE INFO

Article History:

Received 15th July, 2017
Received in revised form 25th August, 2017
Accepted 23rd September, 2017
Published online 28th October, 2017

Key Words:

Mixture distribution models, Two component, EM algorithm, Maximum likelihood, Simulation, Survival analysis.

ABSTRACT

Heterogeneous Survival time data can have two different distributions before and after a certain time due to many factors which affects the life of the creatures or machines. For this purpose, we examine a mixture of two non-identical (different) distributions of Exponential, Gamma, Log-normal, Weibull and Gompertz distributions. In addition to the previous studies, we propose the mixture of Gompertz distribution with the Exponential, Gamma, Weibull and Lognormal distributions. Some properties of the proposed parametric mixture of Exponential, Gamma, Weibull, Lognormal and Gompertzare investigated. Both simulated and real data set were used to estimate the maximum likelihood estimators of the model by employing the Expectation Maximization (EM) algorithm method. The simulations are performed by generating data, sampled from a population of two component parametric mixture of two different distributions. The parameters estimated by the proposed EM Algorithm which are closer to the parameters of the postulated model. To investigate the consistency and stability of the EM algorithm, the simulations are repeated several times. The repetitions of the simulation give estimators closer to the values of postulated models, with relatively small standard errors. Graphs, goodness of fit tests and the Akaike Information Criterion (AIC) were used to compare the proposed model with the pure classical parametric survival models corresponding to each component using real survival data. Results revealed that the proposed model showed that a parametric mixture models are more flexible and maintainthe features of the pure classical survival model and are better option for modelling heterogeneous survival data.

1. INTRODUCTION

Survival analysis is a method to analyze the occurrence of a given event in which the individuals will be observed from the time they experience a particular event. It can be a development or diagnosis of a disease, treatment outcome, recurrence of a disease, death etc. Survival analysis has been a widely used in diverse fields of research and development such as medicine, economics, political science, etc. It has broader applications in Medical Studies, Social Sciences, Engineering, and Biology. Both Nonparametric and classic parametric survival models are commonly used to handle survival data over the past. Various parametric families of models are most frequently used in the analysis of lifetime data. Among the univariate models, distributions like Exponential, Gamma and Weibull distributions take primary chance for their demonstrated applications in wide range of scenarios.(Lee and wang, 2013; Lawless, 2003; Kleinbanm and Klein, 2005). Besides these pure classical statistical distribution models, other novel models for survival data have been developed recently. Especially in cases of data with heterogeneous structure, mixture distributions are more convenient to handle such data. Mixture models are usually applied to model failure-time data in a variety of situations. As a flexible way of modeling data, the mixture approach is directly applicable in situations where the adoption of a single parametric family for the distribution of failure time is inadequate.

Recently, a considerable number of authors applied mixture model technique to analyze survival time data.(Robert E. Colvert *et al*., 1976) discussed a unique and different type of hazard rate function along with maximum likelihood estimation of the parameters from the resulting life time distribution by considering an illustrative example data dealing with failure times for oral irrigators.(Chen *et al*.,1985)used a two-component mixture model for the analysis of cancer survival data by generalizing an earlier idea formulated by (Berkson *et al*., 1952).The parameter of a mixed Weibull distribution using graphical cdf curves was estimated by (Jiang *et al*., 1992a). Also, (Jiang *et al*., 1992b) developed a new algorithm for estimating the parameters of a mixture model of

Corresponding author:* **Leo Alexander T
Department of Statistics, Loyola College, Chennai–34, India

Weibull distributions for censored data.(Philips *et al.*, 2002) suggested an idea for estimating cancer prevalence using mixture models.(Ng.S.K *et al.*,2004**)** proposed a two-component survival mixture model to analyze a set of ischaemic stroke-specific mortality data.(Erisoglu.M *et al.*,2011) showed that the mixture of the same and different distributions of Weibull, Gamma, and Exponential is the appropriate distributions for the earthquake inter occurrence times.(Erisoglu.U *et al.*,2011) proposed a mixture of two different distributions such as Exponential-Gamma, Exponential-Weibull and Gamma-Weibull are the appropriate distributions to model heterogeneous survival data.(Yusuf Abbakar Mohammed *et al.*,2013) developed a parametric mixture model of three different distributions namely, Exponential, Gamma and Weibull to analyze heterogeneous survival time data and applied these mixture models for Kidney Cather data (2015).(AyçaHaticeTùrkan *et al.*,2014) showed a comparison study of two-component Mixture model distribution for heterogeneous survival time dataset by taking a mixture of two identical (same kind of) distributions of Exponential, Gamma, Lognormal and Weibull and also all pairwise combinations of these distribution and analyzed which kind of mixture model distributions is more appropriate for the heterogeneous survival times.(Sri AstutiThamrin *et al.*,2014) developed and applied the Bayesian Weibull mixture approach to model dengue fever patient's survival.(Tabasam Sultana *et al.*, 2016) suggested a 3-component mixture of the inverse Rayleigh distributions under Bayesian perspective.(NavidFeroze, 2016) developed the Bayes estimation of the inverse Weibull mixture distributions under doubly censoring data. Also, we discussed mixture of two identical distributions of Exponential, Gamma, Lognormal, Weibull and Gompertz model to analyze heterogeneous survival time data. (Uma maheswari and Leo Alexander, 2017)

In this paper, we investigate the consistency and stability of EM in estimating the parameters. Also it shows that the mixture of two different distributions is the appropriate distribution for heterogeneous survival data. Our paper is organized as follows: In Section 2, we define the functions of Survival analysis. Also, parametric survival models that have been used to describe survival time namely Exponential, Gamma, Weibull, lognormal, Gompertz distributions are discussed and their properties are also summarized. Section 3, devoted to discussing mixture model of two different distributions in survival analysis and the maximum likelihood estimators of the parameters are obtained by employment of EM algorithm. In Section 4, Simulations are performed by generating data, sampled from a population of two component parametric mixture of two different distributions. Then the simulations will be repeated 500, 1000 and 5000 times with the sample size of 100 observations for each mixture model to investigate the convergence of the EM, consistency and stability of EM algorithm. Also mixture of two different distributions is applied on illustrative examples based on heterogeneous survival real dataset successfully. The data got from National Institute for Research in Tuberculosis, Chetput, Chennai. Finally in Section 5, the summary and conclusion were presented. All computations are carried out using R language.

2. *Basics of Survival Analysis*

Survival time data measure the time taken for a certain event to occur such as response, failure, death, relapse and the development of a given disease. These times are subject to random variations and form a distribution like any random variables. The distribution of survival times is usually described or characterized by three factors namely:

i. The survival function $S(t)$,

ii. The probability density function $f(t)$ and

iii. The hazard function $h(t)$.

It is to be noted that these three functions are mathematically equivalent. If one of them is given, then other two can be derived.
Let T denote the survival time. It is a non-negative and absolutely continuous random variable that represents the life time of individuals. If $F(t)$ is the cumulative distribution of T , then survival function $S(t)$ defined as,

$$S(t) = P(\text{An individual survives longer than } t) = P(T > t), \quad 0 < t < \infty.$$

Here $S(t)$ is a non-increasing function of time $t$with the probability of surviving at least at the time zero is 1 and that of surviving an infinite time is 0. Cumulative distribution function $F(t)$, is defined as the probability that an individual fails before time t,that is

$$F(t) = P(T \le t), 0 < t < \infty.$$

The hazard function $h(t)$ of survival time T gives the conditional failure rate. This is defined as the probability of failure during a very small time interval, given that the individual has survived to the beginning of the interval. It can be expressed as

$$h(t) = \lim_{\Delta t \to 0} \frac{P\{\text{An individual of age t fails in the interval } (t, t+\Delta t)\}}{\Delta t}.$$

The cumulative hazard function is defined as, $H(t) = -\log(S(t)) = \int_0^1 h(u)du.$ Given any one of them, the other two can be derived (Lee and Wang, 2003) $S(t) = 1 - F(t) = \exp(-H(t)).$

2.1 Pure Parameteric Survival Models

Pure parametric Survival models play an important role in Survival analysis and these models are preferred when the chosen probability distribution aptly represents the data. A parametric survival model is a model in which survival time, thus the outcome, is assumed to follow a known distribution. By reviewing the literature about modeling the survival data, it can be seen that the Exponential, Gamma, Weibull, Lognormal and Gompertz probability distribution functions are commonly used in survival analysis. The probability density function $f(t)$ and the survival function $S(t)$ of Exponential, Gamma, Weibull, Lognormal and Gompertz distributions are mentioned in the following Table 1 (Lee and wang, 2013; Lawless, 2003).

Table 1

Distribution	Probability Density Function	Survival function
Exponential	$f_{\exp}(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}}, \; t > 0, \lambda > 0$	$s_{\exp}(t) = 1 - e^{-\frac{t}{\lambda}}$
Weibull distribution	$f_{wbl}(t) = \frac{\gamma}{\eta}\left(\frac{t}{\eta}\right)^{\gamma-1} e^{-\left(\frac{t}{\eta}\right)^{\gamma}}, t, \eta, \gamma > 0$	$s_{wbl}(t) = e^{-\left(\frac{t}{\gamma}\right)^{\eta}}.$
Gamma distribution	$f_{gam}(t) = \frac{t^{\alpha-1} e^{-\frac{t}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}, t, \alpha, \beta > 0$	$S_{gam}(t) = 1 - \frac{\Gamma_x(\alpha)}{\Gamma(\alpha)}$
Lognormal distribution	$f_{\mathrm{logn}}(t) = \frac{1}{\sigma t\sqrt{2\pi}} e^{-\frac{(\log t-\mu)^2}{2\sigma^2}}, \; t > 0, \mu, \sigma > 0$	$S_{\mathrm{logn}}(t) = 1 - \phi\left(\frac{\log t - \mu}{\sigma}\right)$
Gompertz distribution	$f_{gomp}(t) = be^{at} e^{-\frac{b}{a}\left(e^{at}-1\right)}, t > 0, a, b > 0$	$s_{gomp}(t) = e^{-\frac{b}{a}\left(e^{at}-1\right)}$

where $\Gamma_x(\alpha)$ is called an incomplete Gamma function and ϕ is cumulative distribution function of normal probability distribution function .

3. *Parametric Mixture of Two Different Distributions*

Mixture models are frequently used to analyze Survival time data in a variety of situations, because of their high flexibility and they are good choice in situations where a single parametric distribution may not be sufficient. A mixture model of two different distributions is considered where it is assumed that it is sampled from a population consists of $g(\geq 2)$ distinct subgroups or subclasses. The mixture model can be written as

$$f_{1,2}(t;\psi) = \pi \; f_1(t;\theta_1) + (1-\pi) f_2(t;\theta_2) \; ,$$

where the vector $\psi' = (\pi, \theta)$ contains all the unknown parameters for π and $\theta = (\theta_1, \theta_2)'$ in the mixture model(Mclachlan and Pell, 2000; Hogg Mckean Craig,2005). The function $f_1(t;\theta_1)$ is called mixture component density function for the first population with parameter θ_1 and $f_2(t;\theta_2)$ is called mixture component density function for the Second population with parameter θ_2.

In this study, to model heterogeneous Survival times, we consider mixture of two different distributions namely Exponential-Gamma, Exponential-Weibull, Exponential-Lognormal,Exponential-Gompertz,Gamma-Weibull,Gamma-Lognormal,Gamma-Gompertz,Weibull-Lognormal,Weibull-Gompertz and Lognormal-Gompertz which are represented as follows,

$$f_{\exp-gam}(t) = \pi \; f_{\exp}(t;\lambda) + (1-\pi) f_{gam}(t;\alpha,\beta) \quad (1)$$

$$f_{\exp-wbl}(t) = \pi \; f_{\exp}(t;\lambda) + (1-\pi) f_{wbl}(t;\eta,\gamma) \quad (2)$$

$$f_{\exp-\mathrm{logn}}(t) = \pi \; f_{\exp}(t;\lambda) + (1-\pi) f_{\mathrm{logn}}(t;\mu,\sigma) \quad (3)$$

$$f_{\exp-gomp}(t) = \pi \; f_{\exp}(t;\lambda) + (1-\pi) f_{gomp}(t;a,b) \quad (4)$$

$$f_{gam-wbl}(t) = \pi \; f_{gam}(t;\alpha,\beta) + (1-\pi) f_{wbl}(t;\eta,\gamma) \quad (5)$$

$$f_{gam-\mathrm{logn}}(t) = \pi \; f_{gam}(t;\alpha,\beta) + (1-\pi) f_{\mathrm{logn}}(t;\mu,\sigma) \quad (6)$$

$$f_{gam-gomp}(t) = \pi \; f_{gam}(t;\alpha,\beta) + (1-\pi) f_{gomp}(t;a,b) \quad (7)$$

$$f_{wbl-\mathrm{logn}}(t) = \pi \; f_{wbl}(t:\eta,\gamma) + (1-\pi) f_{\mathrm{logn}}(t;\mu,\sigma) \quad (8)$$

$$f_{wbl-gomp}(t) = \pi \; f_{wbl}(t;\eta,\gamma) + (1-\pi) f_{gomp}(t;a,b) \quad (9)$$

and $f_{\mathrm{logn}-gomp}(t) = \pi \; f_{\mathrm{logn}}(t;\mu,\sigma) + (1-\pi) f_{gomp}(t;a,b)$, (10) where π is the mixture weight of the distributions and $\pi \in (0,1)$. The maximum likelihood estimators of parameters of these mixture distributions are estimated using Expectation-Maximization (EM) algorithm.

3.1 Expectation Maximization Algorithm (Em) and Parameter Estimation

EM (Expectation-Maximization) algorithm is one of the most effective method that is used to estimate the Maximum likelihood estimators in finite mixture models likelihood (Mclachlan and Pell, 2000; Hogg Mckean Craig, 2005; McLachlan and Krishnan, 1997). Let $t_1, t_2, \ldots, t_n$ be a set of observations of n incomplete data and z_1, z_2 be a set of missing observations where $z_{1i} = z_1(t_i) = 1$ for $i = 1, \ldots, n$ if the observation t_i belongs to 1st class and 0 otherwise.

The EM algorithm is applied to the mixture distributions by treating z_i as unobserved or missing data. It consists of two steps, E (for Expectation) and M (for Maximization).

In E- step, to estimate the hidden variable vector $z_i = (z_{1i}, z_{2i})$, conditional expectation function $E(Z_{1i} | t_i)$ and $E(Z_{2i} | t_i)$ are used.

So, $$\hat{z}_{1i} = E_{\psi_0}(z_{1i} | t_i) = \frac{\pi\, f_{1,0}(t_i\,;\theta_1)}{\pi\, f_{1,0}(t_i;\theta_1) + (1-\pi)\, f_{2,0}(t_i;\theta_2)}$$

and $$\hat{z}_{2i} = E_{\psi 0}(z_{2i} | t_i) = \frac{(1-\pi)\, f_{2,0}(t_i\,;\theta_2)}{\pi\, f_{1,0}(t_i;\theta_1) + (1-\pi)\, f_{2,0}(t_i;\theta_2)}\,.$$

In M-step, $E(Z_{1i} | t_i)$and $E(Z_{2i} | t_i)$ function which are calculated in E-step is maximized. The M-step and E- step should be iterated alternatively till the convergence criterion is met. The estimator of π_k $(k = 1,2)$ is obtained as $\hat{\pi}_k = \dfrac{\sum_{i=1}^{n} \hat{z}_{ki}}{n}$.

By using Eqs (1), (2),(3) and (4), the maximum likelihood estimator of λ parameter can be obtained in Equation (11) for Exponential-Gamma, Exponential-Weibull, Exponential-Lognormal, Exponential-Gompertz mixture distributions. The MLE of λ parameter is given by,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \hat{z}_{1i} t_i}{\sum_{i=1}^{n} \hat{z}_{1i}}\,. \qquad (11)$$

The maximum likelihood estimators of α and β parameters can be obtained in Eqs (12) and (13) for Exponential-Gamma mixture distribution. The MLE of α and β parameter are as follows,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \hat{z}_{2i} t_i}{\hat{\alpha} \sum_{i=1}^{n} \hat{z}_{2i}} \qquad (12\,)$$

and $$\hat{\alpha}^{r+1} = \hat{\alpha}^{r} - \frac{\log(\hat{\alpha}^{r}) - \psi^{'}(\hat{\alpha}^{r}) - \log\left(\dfrac{\sum_{i=1}^{n} \hat{z}_{2i} t_i}{\sum_{i=1}^{n} \hat{z}_{2i}}\right) + \dfrac{\sum_{i=1}^{n} \hat{z}_{2i} \log t_i}{\sum_{i=1}^{n} \hat{z}_{2i}}}{\dfrac{1}{\hat{\alpha}^{r}} - \psi^{'}(\hat{\alpha}^{r})}. \qquad (13\,)$$

It is to be noted that r is the number of Newton-Raphson iterations within EM algorithm. Also $\psi(.)$ and $\psi^{'}(.)$ are a digamma and trigamma functions respectively.

The maximum likelihood estimators of η and γ parameters can be obtained in Eqs (14) and (15) for Exponential-Weibull and Gamma-Weibull mixture distributions. The MLE of η and γ parameter are given by,

$$\hat{\eta} = \left(\left(\sum_{i=1}^{n} \hat{z}_{2i} \right)^{1} \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\gamma}} \right)^{1/\hat{\gamma}} \tag{14}$$

and $$\hat{\gamma}^{r+1} = \hat{\gamma}^{r} + \frac{A_r + \left(1/\hat{\gamma}^{r}\right) - \left(C_r / B_r\right)}{(1/(\hat{\gamma}^{r})^2) + (B_r D_r - C_r^2)/B_r^2}, \tag{15}$$

where $A_r = \left(\sum_{i=1}^{n} \hat{z}_{2i} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{2i} \log t_i$, $B_r = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\gamma}^r}$, $C_r = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\gamma}^r} \log t_i$ and $D_r = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\gamma}^r} (\log t_i)^2$.

The maximum likelihood estimators of μ and σ parameters can be obtained in Equation (16) for Exponential-Lognormal, Gamma-Lognormal, Weibull-Lognormal mixture distributions. The MLE of μ and σ parameter are as follows,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \hat{z}_{2i} \ln t_i}{\sum_{i=1}^{n} \hat{z}_{2i}} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{z}_{2i} (\ln t_i - \hat{\mu})^2}{\sum_{i=1}^{n} \hat{z}_{2i}}. \tag{16}$$

The maximum likelihood estimators of a and b parameters can be obtained in Equations (17) and (18) for Exponential-Gompertz,Gamma-Gompertz,Weibull-Gompertz mixture distributions. The MLE of a and b parameter are given as,

$$\hat{b} = \frac{\hat{a} \sum_{i=1}^{n} \hat{z}_{2i}}{\sum^{n} \hat{z}_{2i} e^{\hat{a} t_i} - \sum^{n} \hat{z}_{2i}} \quad \text{and} \tag{17}$$

$$\hat{a}^{r+1} = \hat{a}^{r} + \frac{E_r + \left\{ (F_r G_r - \hat{a}^r F_r H_r - (F_r)^2) / \hat{a}^r (G_r - F_r) \right\}}{\left\{ \frac{(F_r (G_r)^2 - 2(F_r)^2 G_r + (\hat{a}^r)^2 G_r F_r I_r - (\hat{a}^r)^2 (F_r)^2 I_r - (\hat{a}^r)^2 F_r (H_r)^2 + (F_r)^3)}{(\hat{a}^r (G_r - F_r))^2} \right\}}, \tag{18}$$

where $E_r = \sum_{i=1}^{n} \hat{z}_{2i} t_i$, $F_r = \sum_{i=1}^{n} \hat{z}_{2i}$, $G_r = \sum_{i=1}^{n} \hat{z}_{2i} e^{\hat{a}^r t_i}$, $H_r = \sum_{i=1}^{n} \hat{z}_{2i} t_i e^{\hat{a}^r t_i}$ and $I_r = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\,2} e^{\hat{a}^r t_i}$.

By using Equations (5), (6) and (7), the maximum likelihood estimators of α and β parameters of Gamma-Weibull, Gamma-Lognormal, Gamma-Gompertz mixture distributions are estimated using $\hat{z}_{1i}$ instead of $\hat{z}_{2i}$ in equations (12) and (13).

By using Equations (8) and (9), the maximum likelihood estimators of η and γ parameters of Weibull-Lognormal, Weibull-Gompertz mixture distributions are estimated using $\hat{z}_{1i}$ instead of $\hat{z}_{2i}$ in equations (14) and (15).

Also by equation (10),the MLE of μ and σ parameters of Lognormal-Gompertz is estimated using $\hat{z}_{1i}$ instead of $\hat{z}_{2i}$ in equation (16).

3.2 Criteria for Model Selection

To find the appropriate distribution, we use two different goodness of fit tests: the mean square error (MSE) test and the Kolmogorov-Smirnov (KS) test. Let us first use the MSE test. The MSE value is defined as

$$MSE = \frac{\sum_{i=1}^{n} \left[F_e(t_i) - F(t_i) \right]^2}{n-k},$$

where $F_e(t)$ is the empirical distribution and $F(t)$ is the cumulative distribution function that is proposed to model the heterogeneous survival data set. Here k is the number of free parameters in the distribution. As it is known, the smallest MSE value reveals the most appropriate distribution. Then The Kolmogorov-Smirnov statistic KS is defined by

$$KS = \max\left|F_e(t) - F(t)\right| .$$

It is known that the preferred distribution has the smallest value of KS. Also, we use AIC as goodness of fit test for model selection criteria. AIC value is as follows

$AIC = -2\log L + 2d$, where d represents estimated parameters (Mclachlan and Pell, 2000). The smallest AIC value represents the best model.

4. *Data analysis*

4.1 *Simulated data*

In this section, samples of size 100 observations were generated from two component parametric mixture of two different distributions. The mixture modelsinclude Exponential-Gamma, Exponential-Weibull, Exponential-Lognormal, Exponential-Gompertz, Gamma-Weibull, Gamma-Lognormal, Gamma-Gompertz, Weibull-Lognormal, Weibull-Gompertz and Lognormal-Gompertz. The simulations will be repeated 500, 1000 and 5000 times with the sample size of 100 observations for each mixture model to investigate the convergence of the EM, consistency and stability of EM algorithm.

There is no restriction imposed on the maximum number of iterations and convergence was achieved when the differences between successive estimates were less than 10^{-4}. The results from the simulated data sets are listed in the following Table 2 – 11which gives the averages of the maximum likelihood estimators $av(\hat{\pi},\hat{\theta})$ and standard errors $se(\hat{\pi},\hat{\theta})$. Also, the graphs of mixture of two different distributions for simulation parameters are shown in the following Figures 1 -10. Figures 1 - 10, exhibits the comparison between the probability density function of the parametric mixture model and the probability density functions of each single distribution. Also it can be seen in the graph, the mixture model fits the simulated data far better than the single distributions.

4.1.1 Mixture Model of Exponential- Gamma

Table 2

	Exponential – Gamma				
Parameters		π	λ	α	β
Postulated model		$\pi=0.6$	$\lambda=3$	$\alpha=10$	$\beta=15$
5000 times	$av(\hat{\pi},\hat{\theta})$	0.600	3.003	10.211	15.319
	$se(\hat{\pi},\hat{\theta})$	0.049	0.041	0.017	0.003

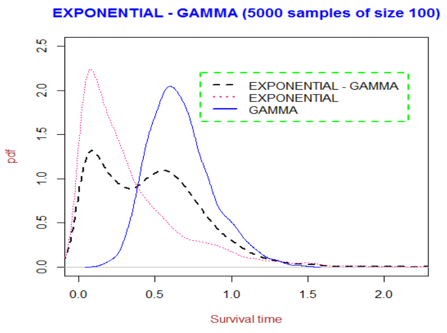


Figure 1

4.1.2 Mixture Model of Exponential- Weibull

Table 3

	Exponential –Weibull				
Parameters		π	λ	η	γ
Postulated model		$\pi=0.6$	$\lambda=0.75$	$\eta=1$	$\gamma=5$
5000 times	$av(\hat{\pi},\hat{\theta})$	0.600	0.750	1.002	5.084
	$se(\hat{\pi},\hat{\theta})$	0.050	0.167	0.030	0.030

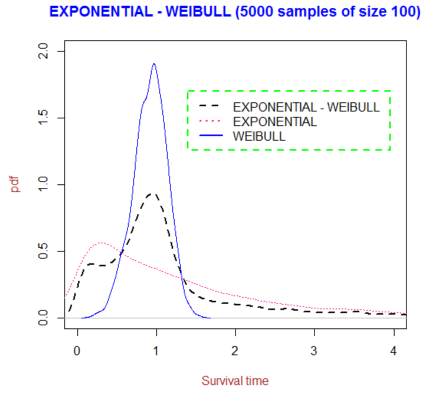


Figure 2

4.1.3 Mixture Model of Exponential- Lognormal

Table 4

Exponential – Lognormal					
Parameters		π	λ	μ	σ
Postulated model		$\pi=0.7$	$\lambda=0.20$	$\mu=0.5$	$\sigma=1.5$
5000 times	$av\left(\hat{\pi},\hat{\theta}\right)$	0.700	0.200	0.499	1.492
	$se\left(\hat{\pi},\hat{\theta}\right)$	0.046	0.576	0.263	0.193

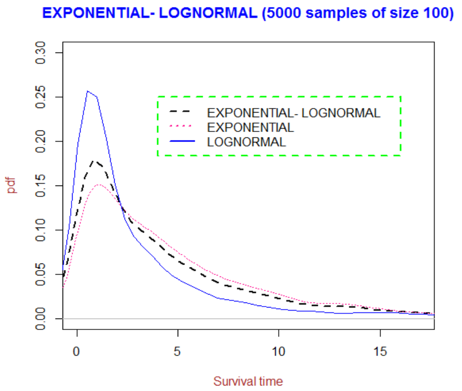


Figure 3

4.1.4 Mixture Model of Exponential- Gompertz

Table 5

Exponential-Gompertz					
Parameters		π	λ	a	b
Postulated model		$\pi=0.6$	$\lambda=2$	$a=0.9$	$b=0.03$
5000 times	$av\left(\hat{\pi},\hat{\theta}\right)$	0.600	1.996	0.862	0.029
	$se\left(\hat{\pi},\hat{\theta}\right)$	0.049	0.062	0.001	0.005

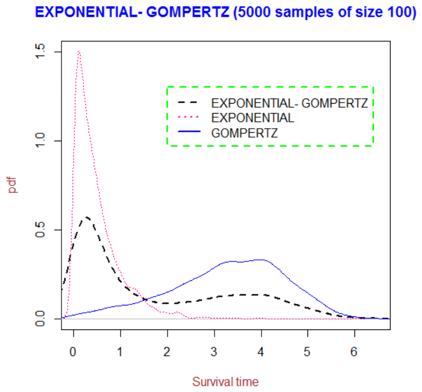


Figure 4

4.1.5 Mixture Model of Gamma-Weibull

Table 6

Gamma-Weibull						
Parameters		π	α	β	η	γ
Postulated model		$\pi=0.3$	$\alpha=12$	$\beta=2$	$\eta=15$	$\gamma=4$
5000 time	$av\left(\hat{\pi},\hat{\theta}\right)$	0.300	12.357	1.978	14.984	4.019
	$se\left(\hat{\pi},\hat{\theta}\right)$	0.047	0.277	0.104	0.412	0.009

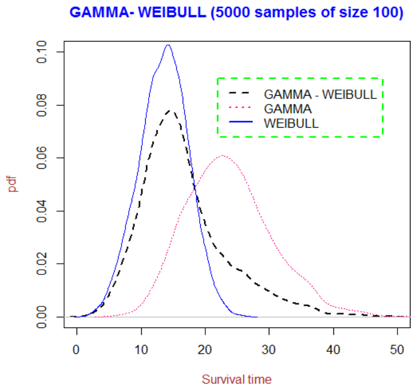


Figure 5

4.1.6 Mixture Model of Gamma-Lognormal

Table 7

Gamma-Lognormal						
Parameters		π	α	β	μ	σ
Postulated model		$\pi=0.4$	$\alpha=15$	$\beta=5$	$\mu=1$	$\sigma=0.25$
5000 times	$av\left(\hat{\pi},\hat{\theta}\right)$	0.400	15.318	5.102	1.000	0.249
	$se\left(\hat{\pi},\hat{\theta}\right)$	0.050	0.247	0.008	0.030	0.023

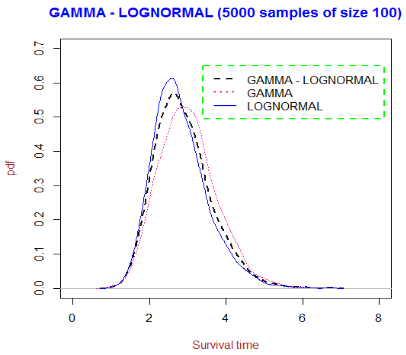


Figure 6

4.1.7 Mixture Model of Gamma-Gompertz

Table 8

Gamma-Gompertz					
Parameters	π	α	β	a	b
Postulated model	$\pi = 0.3$	$\alpha = 15$	$\beta = 2$	$a = 0.35$	$b = 0.05$
5000 times $av\left(\hat{\pi},\hat{\theta}\right)$	0.300	15.425	2.058	0.356	0.035
$se\left(\hat{\pi},\hat{\theta}\right)$	0.047	0.367	0.023	0.001	0.004

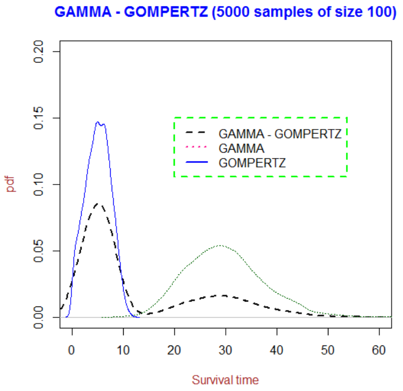


Figure 7

4.1.8 Mixture Model of Weibull-Lognormal

Table 9

Weibull-Lognormal						
Parameters		π	η	γ	μ	σ
Postulated model		$\pi = 0.8$	$\eta = 8$	$\gamma = 1.5$	$\mu = 1.5$	$\sigma = 2$
5000 times	$av\left(\hat{\pi},\hat{\theta}\right)$	0.800	7.891	1.511	1.502	1.993
	$se\left(\hat{\pi},\hat{\theta}\right)$	0.040	0.543	0.003	0.444	0.314

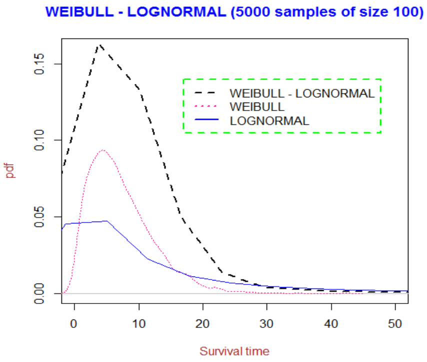


Figure 8

4.1.9 Mixture Model of Weibull-Gompertz

Table 10

Weibull-Gompertz						
Parameters		π	η	γ	a	b
Postulated model		$\pi = 0.6$	$\eta = 1$	$\gamma = 3$	$a = 10$	$b = 5$
5000 times	$av\left(\hat{\pi},\hat{\theta}\right)$	0.600	1.001	3.034	9.864	5.049
	$se\left(\hat{\pi},\hat{\theta}\right)$	0.050	0.041	0.002	0.002	0.796

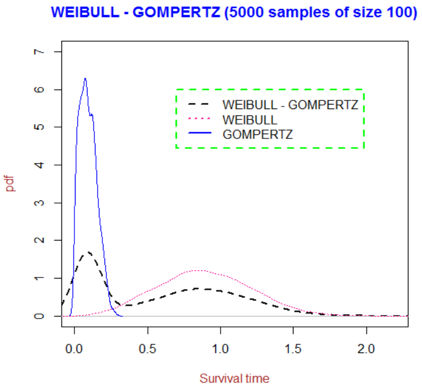


Figure 9

4.1.10 Mixture Model of Lognormal-Gompertz

Table 11

Lognormal-Gompertz					
Parameters	π	μ	σ	a	b
Postulated model	$\pi = 0.5$	$\mu = 0.8$	$\sigma = 0.15$	$a = 7$	$b = 0.10$
5000 times $av\left(\hat{\pi},\hat{\theta}\right)$	0.500	0.800	0.150	6.967	0.198
$se\left(\hat{\pi},\hat{\theta}\right)$	0.051	0.020	0.015	0.041	0.024

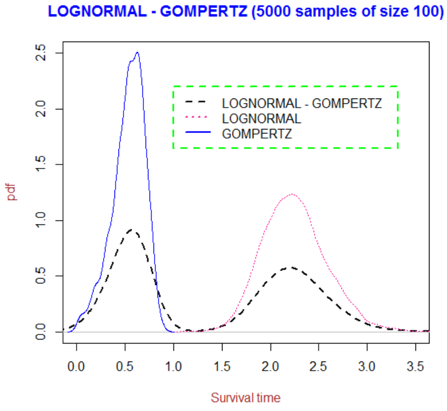


Figure 10

The results of the parameter estimation listed from Table 2- 11, show the averages of the estimated parameters of Exponential-Gamma,Exponential-Weibull,Exponential-Lognormal,Exponential-Gompertz,Gamma-Weibull,Gamma-Lognormal,Gamma-Gompertz,Weibull-Lognormal,Weibull-Gompertz and Lognormal- Gompertz mixture model and its corresponding standard error respectively. It can be observed that the estimators get closer to the true values (postulated model) of the mixture model as the number of repetitions increases i.e., the averages of the estimators are very close to the true values of the parameters and their standard errors are relatively small which suggests that the EM algorithm estimators performed consistently. Convergence was achieved in all the cases, even though when the starting values are poor and this emphasizes the numerical stability of the EM algorithm.

Figure 1 - 10exhibits the comparison between the probability density function of the parametric mixture model Exponential, Gamma, Weibull, Lognormal and Gompertz distributions and the probability density functions of each single distribution. As it can be seen in the graph, the mixture model fits the simulated data far better than the single distributions. Simulation results revealed that EM algorithm approach works well with Non- identical mixture proportions.

4.2 AnApplication of Non-Identical Mixture Models

Bone marrow survival data set consists of survival times of 137 patients. The data has been collected from National Institute for Research in Tuberculosis (NIRT), Chennai. Mixtures of Non-identical distributions have been fitted for the data set. The estimated parameter, Log-likelihood (LL), K-S test statistic, mean square error (MSE) values , Akaike information Criterion (AIC) for mixture of non-identical distribtuions such as Exponential-Gamma, Exponential-Weibull, Exponential-Lognormal, Exponential-Gompertz, Gamma-Weibull, Gamma-Lognormal, Gamma-Gompertz,Weibull-Lognormal, Weibull-Gompertz and Lognormal-Gompertz are mentioned in Table 12

From Table 12, it can be noted that based on KS statistic, MSE and AIC values, Exponential-Gamma mixture has the smallest KS test statistic and smallest MSE value and it is the best model for bone marrow survival data set. According to MSE and AIC comparison values Exponential-Weibull mixture has least value which is also considered another best model for the same data set. Therefore, it can be observed from Table 12, Exponential-Gamma and Exponential-Weibull mixture models are best models for bone marrow survival data set.

A graphical comparison of the fitted (pure) pdf of Exponential, Gamma and Weibull distribution and fitted pdf of Non-identical mixture models of Exponential- Gamma and Exponential - Weibull for survival times of bone marrow data set is mentioned in Figure 4.11(a) and Figure 4.11(b)

The Estimated Parameters, LL values, K-S test statistics, MSE values and AIC for bone marrow dataset

Table 12

S.No.	Models	Estimates		π_1	π_2	LL	KS	MSE	AIC
1	Exp-Gam	$\lambda = 182.9933$	$\alpha = 15.5613$ $\beta = 43.8543$	0.6956	0.3044	-921.8529	**0.0273**	**0.0002**	1851.71
2	Exp-Wbl	$\lambda = 179.265$	$\eta = 740.4712$ $\gamma = 4.3949$	0.6848	0.3152	-920.3082	0.0279	**0.0002**	**1848.62**
3	Exp-Logn	$\lambda = 347.1548$	$\mu = 2.2240$ $\sigma = 1.4303$	0.9622	0.0378	-932.6449	0.0811	0.0012	1873.29
4	Exp-Gomp	$\lambda = 335.0073$	$a = 0.0067$ $b = 0.0030$	1.0000	0.0000	-933.5389	0.0820	0.0015	1875.08
5	Gam-Wbl	$\alpha = 1.4639$ $\beta = 264.5574$	$\eta = 232.2236$ $\gamma = 0.6769$	0.7705	0.2295	-939.4677	0.0799	0.0025	1888.94
6	Gam-Logn	$\alpha = 1.0350$ $\beta = 330.3848$	$\mu = 0.7482$ $\sigma = 0.4794$	0.9796	0.0204	-932.0110	0.0798	0.0014	1874.02
7	Gam-Gomp	$\alpha = 4.6968$ $\beta = 126.1662$	$a = 0.0062$ $b = 0.0024$	0.4510	0.5490	-961.9240	0.0474	0.0005	1933.85
8	Wbl-Gomp	$\eta = 316.8434$ $\gamma = 3.6349$	$a = 0.0057$ $b = 0.0031$	0.4134	0.5866	-921.7945	0.3366	0.0246	1853.59
9	Wbl-Logn	$\eta = 187.1038$ $\gamma = 0.9902$	$\mu = 6.5000$ $\sigma = 0.2527$	0.7046	0.2954	-922.4826	0.3130	0.0222	1854.97
10	Logn-Gomp	$\mu = 6.4192$ $\sigma = 0.2947$	$a = 0.0055$ $b = 0.0034$	0.4080	0.5920	-923.1831	0.0361	0.0003	1856.37

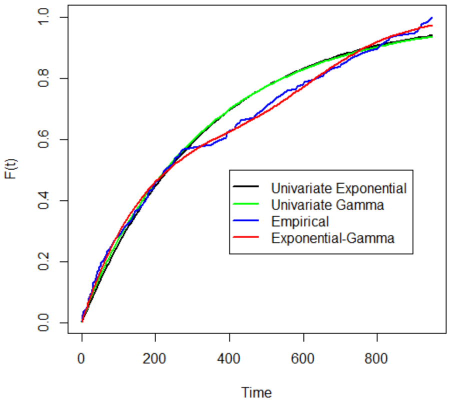


Figure 11 a

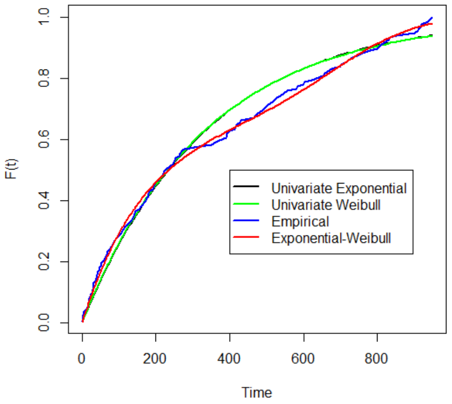


Figure 11 b

From Figure 11(a) and 11(b),Non-identical mixture models of Exponential-Gamma and Exponential-Weibull fit much better than (pure) Gamma and Weibull distributions for survival times of bone marrow data set.

5. Over All Conclusion

In this paper we proposed the mixture models of two Non-identical distributions such as Exponential-Gamma, Exponential-Weibull, Exponential-Lognormal, Exponential-Gompertz, Gamma-Weibull, Gamma-Lognormal, Gamma-Gompertz, Weibull-Lognormal, Weibull-Gompertz and Lognormal- Gompertzto represent the heterogeneous survival data sets. Heterogeneous survival time data can have two different distributions before and after a certain time due to many factors which affects the life of the creatures. For instance, a slowly growing tumor can grow faster after a particular process and this can be affect the life time. Each of the different phases of life will generate a peak in the mixture distribution.

Therefore, we try to model the heterogeneous survival time data with the most appropriate distributions among the mixture models. Mixtures of Exponential-Gamma, Exponential-Weibull, Exponential-Lognormal, Exponential-Gompertz, Gamma-Weibull, Gamma-Lognormal, Gamma-Gompertz, Weibull-Lognormal, Weibull-Gompertz and Lognormal- Gompertz were tested for the best fit to the simulated datasets as well as real survival datasets.

The maximum likelihood estimations of parameters of the mixture models obtained with EM algorithm. The repetitions of the Simulation give estimators closer and closer to the postulated models, as the number of repetitions increases with relatively small standard errors. The Table show that the EM algorithm converged to the true values (postulated model) of the mixture model parameters in 5000 repetitions and that emphasizes the stability of the algorithm in estimating the parameters with different proportion of mixing probabilities. The averages are close to the true values of the parameters and the standard errors are relatively small which suggest that the EM algorithm estimator performed consistently.

Also, the graphs for all the two component mixture model fits the simulated data far better than the single distributions. According to the simulation results, the EM algorithm successfully estimated the parameters of the two component mixture model of identical distributions.

Also, we employ mixture of Non-identical distributions for modeling survival times for bone marrow dataset. The AIC values, KS test statistics and MSE are calculated to determine the most appropriate distribution for the present data set. It can be noted from **Table 12** that the best model among the two component mixture models of Non-identical distribution is the mixture of Exponential-Gamma for Survival times of bone marrow patients according to KS test statistics and MSE value. And alternative models are determined as the mixture of Exponential-Weibull distributions according to MSE and AIC values for this data set respectively.

The histogram and the two probability density functionsof Exponential–Gamma and Exponential-Weibull fits better than others for the survival times of 137 Bone marrow patientsthat is given in Figure 12(a). The empirical distribution function and two distribution functions of Exponential-Gamma and Exponential-Weibullfits better is shown in Figure 12(b)

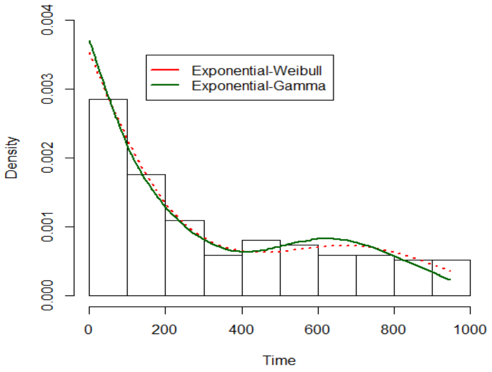
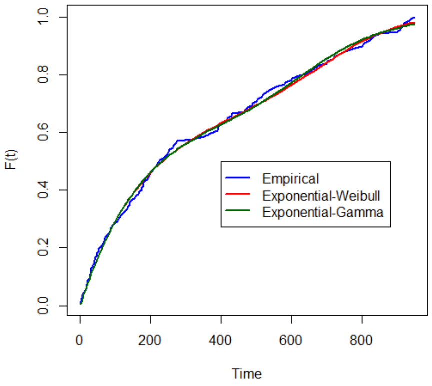


Figure 12 (a)

The probability densities of the fitted distributions and a histogram.

Figure 12 (b)



The empirical distribution function and the fitted distribution function for bone marrow dataset

Acknowledgement

We would like to thank Dr.Ponnuraja.Scientist-D, National Institute for Research in Tuberculosis, Chetput, Chennai who helped in providing data

References

1. AyçaHaticeTùrkan and Nazifçaliş (2014) Comparison of Two-Component Mixture Distribution Models for Heterogeneous Survival Datasets: A Review Study. ˙ISTAT˙IST˙IK: *Journal of the Turkish Statistical Association* Vol.7, No. 2, July 2014,pp. 33-42 ISSN 1300-4077 | 14 | 2 | 33 | 42
2. Berkson, J., Gage, R.P, Survival cure for cancer patients following treatment. *Journal of the American Statistical Association* 47, 501-515, 1952.
3. Chen W.C., Hill B.M., Greenhouse J.B. and Fayos J.V. (1985). Bayesian Analysis of Survival Curves for Cancer Patients Following Treatment. *Bayesian Statistics* 2, 299-328.
4. Eri¸so˘glu M, C¸ alı¸s N, Servi T, Eri¸saho˘glu ¨ U, Topaksu M (2011a) The mixture distribution models for interoccurence times of earthquakes. *Russian Geology and Geophysics* 52(2011):685-692.
5. Erişoğlu, Ü.,Erişoğlu, M., &Erol, H. (2011). A mixture model of two different distributions approach to the analysis of heterogeneous survival data. *International Journal of Computer, Electrical, Automation, Control and Information Engineering* Vol:5, No:6,
6. Hogg MckeanCraig. (2005) Introduction to Mathematical Statistics. Sixth Edition, Published by Dorling Kindersley (India) Pvt.Ltd., licensees of Pearson Education in south Asia.
7. Jiang, S., & Kececioglu, D. (1992a). Graphical representation of two mixed-Weibull distributions. IEEE Transaction on Reliability, 41, 241-247. http://dx.doi.org/10.1109/24.257789

8. Jiang, S., &Kececioglu, D. (1992b). Maximum likelihood estimates, from censored data, for mixed-Weibull distributions. *IEEE Transaction on Reliability, 41*, 248-255. http://dx.doi.org/10.1109/24.257791
9. Kleinbanm D.G. and Klein M., Survival Analysis: A Self-Learning Text, Second Edition, Springer, 2005
10. Lee E.T. and Wang J.W. (2013). Statistical Methods for Survival Data Analysis. Fourth Edition, John Wiley &Sons, Inc. All rights reserved
11. Lawless J.F., Statistics Models and Methods for Lifetime Data, Second Edition, John Wiley & Sons, New Jersey, 2003
12. Mclachlan G.J. and Peel D. (2000). *Finite Mixture Model*. Wiley, New York.
13. McLachlan G.J. and Krishnan T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.
14. NavidFeroze (2016) Bayesian Inference of a Finite Mixture of Inverse Weibull Distributions with an Application to Doubly Censoring Data .Pakistan Journal of Statistics and Operation Research pak.j.stat.oper.res.Vol.XII No.1 2016 pp 53-72. http://dx.doi.org/10.18187/pjsor.v12i1.877
15. Ng, A. S. K., McLachlan, G. J., Yau, K. K. W., & Lee, A. H. (2004). Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. *Statistics in Medicine, 23*(17), 2729-2744. http://dx.doi.org/10.1002/sim.1840
16. Phillips, N., Coldman, A., & McBride, M. L. (2002). Estimating cancer prevalence using mixture models for cancer survival. *Statistics in Medicine*, 21(9), 1257-1270. http://dx.doi.org/ 10.1002/sim.1101
17. Robert E.Colvert and Boardman, T.J., Estimation in the piece-wise constant hazard rate model. Communication in Statistics-Theory. *Methods.*11:1013-1029, 1976.
18. Sri AstutiThamrin, AndiKresna Jaya, La PodjeTalangko (2014) Bayesian Weibull Mixture Models for Dengue Fever. Conference Paper. Proc. ICCS-13, Bogor, Indonesia , December 18-21, 2014, Vol. 27, pp. 73-86
19. Tabasam Sultana, Muhammad Aslam (2016) A 3-Component Mixture of inverse Rayleigh Distributions: Properties and Estimation in Bayesian Framework. *International Journal of Basic and Applied Sciences*, 5 (2) (2016) 120-139. doi: 10.14419/ijbas.v5i2.5935
20. Uma maheswari. R and Leo Alexander.T (2017). Mixture of Identical Distributions of Exponential, Gamma, Lognormal, Weibull, Gompertz approach To Heterogeneous Survival Data. *International journal of Current Research*, Vol.9, Issue,09, pp.57521-575332.
21. Yusuf Abbakar Mohammed, BidinYatim and Suzilah Ismail (2013) A Parametric Mixture Model of Three Different Distributions: An Approach to Analyse Heterogeneous Survival Data. DOI: 10.1063/1.4887734

How to cite this article:

Uma maheswari R and Leo Alexander T.2017, Two- Component of Non- Identical Mixture Distribution Models For Heterogeneous Survival Data. *Int J Recent Sci Res.* 8(10), pp. 20813-20824.

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0810.0965

****