İSTATİSTİK: JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION
Vol. 7, No. 2, July 2014, pp. 33–42
ISSN 1300-4077 | 14 | 2 | 33 | 42

COMPARISON OF TWO-COMPONENT MIXTURE DISTRIBUTION MODELS FOR HETEROGENEOUS SURVIVAL DATASETS: A REVIEW STUDY

Ayça Hatice Türkan
Department of Statistics, Afyon Kocatepe University, Afyonkarahisar, Turkey

Nazif Çalış
Department of Management, Adıyaman University, Adıyaman, Turkey

Abstract: Heterogeneous survival data can have two different distributions before and after a certain time because many factors affect the life of the creatures or machines. For this purpose, we use a mixture of two identical (same kind of) distributions of Exponential, Gamma, Lognormal and Weibull and also all pairwise combinations of these distributions. In addition to the previous studies, we propose the mixture of Log-normal distribution with the Exponential, Gamma and Weibull distributions. Maximum likelihood estimations of parameters of the mixture distribution models are obtained by using the EM (Expectation Maximization) algorithm. Model performances are compared using goodness of fit tests and Akaike's information criterion (AIC). Results indicate that, mixtures of two non-identical (different kind of) distributions are as useful as mixtures of identical distributions.

Key words: Exponential, gamma, lognormal, mixture distribution models, survival analysis, weibull
History: Submitted: 4 April 2012; Revised: 1 September 2014; Accepted: 2 September 2014

1. Introduction

The statistical analysis of lifetime, survival time or failure time data is an important topic in many areas, including the biomedical, engineering, and social sciences. Various parametric families of models are used in the analysis of lifetime data. Among the univariate models, a few distributions occupy a central position because of their demonstrated usefulness in a wide range of situations. Foremost in this category are the Exponential, Lognormal, Gamma and Weibull distributions [13]. Also, mixture of two identical distributions (ID) and mixture of two non-identical distributions (NID) using different binaries of Gamma, Weibull and Exponential distributions have been recently used to model the heterogeneous survival data sets [7, 8, 9]. Mixture distribution models are useful because they are applied to represent heterogeneous data set when there is evidence of multimodality or simply unimodality [10]. Chen et al. [2] used a two-component mixture model for the analysis of cancer survival data generalizing an earlier idea in [1]. In [17], a similar model of a mixture of a Weibull component and a surviving fraction in the context of a lung cancer trial were considered. Marin et al. [15] illustrated how Bayesian methods can be used to fit a mixture of Weibull models with an unknown number of components to heterogeneous, possibly right-censored survival data using a birth death Markov chain Monte Carlo (MCMC) algorithm. Zhang [18] proposed and studied the usefulness of a parametric mixture model approach for the analysis of survival data.

As seen in the previous studies, mixture distribution models are more appropriate distribution models for the heterogeneous survival data sets. But, we need to compare the mixture distribution models which are mixture of two identical distributions and mixture of two different distributions.

Therefore, the purpose of this paper is to show that which kind of mixture distributions is more appropriate distribution for the heterogeneous survival times.

2. Survival Time Functions

Lifetime is the length of life measured from some particular starting point. In applications, other terms such as ”failure time” and ”survival time” are also frequently used. Survival time data measure the time to a certain event, such as failure, death, response, the development of a given disease. These times are subject to random variations, and like any random variables, form a distribution [14]. Let T denote the survival time. Survival function, denoted by $S(t)$, is defined as the probability that an individual survives longer than t:

$$S(t) = P(T \geq t) \tag{2.1}$$

where $S(t) \geq 0$, $S(0) = 1$ and $\lim\limits_{t \to \infty} S(t) = 0$.

If we define

$$F(t) = 1 - S(t) = P(T \rangle t) \tag{2.2}$$

where $F(0) = 0$ and then $F(t)$ is the probability that a fatality occurs before time t. We will refer to $F(t)$ as the cumulative distribution function (cdf). A third function, defined by

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{2.3}$$

is called the probability density function (pdf). The pdf, $f(t)$ has these two properties:

$$f(t) \geq 0 \; and \int_0^\infty f(t)\, dt = 1 \tag{2.4}$$

The mean lifetime is defined by

$$E(t) = \int_0^\infty t f(t)\, dt \tag{2.5}$$

which is the expected value of the probability distribution defined by $f(t)$ [5].

Probability density functions, distribution functions and means of the theoretical distributions used in this study are briefly summarized in Table1.

Table 1. The features of the theoretical distributions used in the study

Distribution	**Probability Density Function**	**Distribution Function**	**Mean**
Exponential distribution	$\frac{1}{\lambda}e^{-\frac{t}{\lambda}} \quad , t > 0 , \lambda > 0$	$1 - e^{-\frac{t}{\lambda}}$	λ
Gamma distribution	$t^{\alpha-1}\frac{e^{-t/\beta}}{\beta^\alpha \Gamma(\alpha)} \quad , t > 0 , \alpha, \beta > 0$	$\frac{\int_0^x t^{\alpha-1} e^{-t} dt}{\Gamma(\alpha)}$	$\alpha\beta$
Lognormal distribution	$\frac{e^{\left(-\frac{1}{2}\left\{\frac{\ln t - \mu}{\sigma}\right\}^2\right)}}{t\sigma\sqrt{2\pi}} \quad , t > 0 , \mu, \sigma > 0$	$\Phi\left\{(\ln t - \mu)/\sigma\right\}\Phi = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{\ln t - \mu}{\sigma}} e^{-\frac{u^2}{2}} du$	$e^{\mu + \frac{\sigma^2}{2}}$
Weibull distribution	$\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} e^{-\left(\frac{t}{\beta}\right)^\alpha} , t > 0 , \alpha, \beta > 0$	$1 - e^{-\left(\frac{t}{\beta}\right)^\alpha}$	$\beta\Gamma\left(1 + \frac{1}{\alpha}\right)$

3. Model Description

Choosing a theoretical distribution to approximate survival data is as much an art as a scientific task. In this paper, the mixture of several theoretical distributions that have been used widely to describe survival time are suggested. In Sec. 3.1, we define mixture of two identical distributions in survival analysis. In Sec. 3.2, we define mixture of two different distributions in survival analysis.

3.1. Mixture of Two Identical Distributions

To model the heterogeneous survival data set, we used the mixtures of the two same kind of distributions which are identical pairs of Exponential, Gamma, Lognormal and Weibull defined as

$$f_{\exp-\exp}(t) = \pi_1 f_{\exp}(t) + \pi_2 f_{\exp}(t) \tag{3.1}$$

$$f_{gam-gam}(t) = \pi_1 f_{gam}(t) + \pi_2 f_{gam}(t) \tag{3.2}$$

$$f_{\log n-\log n}(t) = \pi_1 f_{\log n}(t) + \pi_2 f_{\log n}(t) \tag{3.3}$$

$$f_{wbl-wbl}(t) = \pi_1 f_{wbl}(t) + \pi_2 f_{wbl}(t) \tag{3.4}$$

where π_1, π_2 are the mixing weights and $\pi_1 + \pi_2 = 1$, $0 < \pi_1, \pi_2 < 1$ for each of the mixture distribution models.

3.2. Mixture of Two Non-identical Distributions

To model the heterogeneous survival data set, we used the mixtures of two different distributions which are non-identical pairs of Exponential, Gamma, Lognormal and Weibull defined as

$$f_{\exp-gam}(t) = \pi_1 f_{\exp}(t) + \pi_2 f_{gam}(t) \tag{3.5}$$

$$f_{\exp-logn}(t) = \pi_1 f_{\exp}(t) + \pi_2 f_{logn}(t) \tag{3.6}$$

$$f_{\exp-wbl}(t) = \pi_1 f_{\exp}(t) + \pi_2 f_{wbl}(t) \tag{3.7}$$

$$f_{gam-\log n}(t) = \pi_1 f_{gam}(t) + \pi_2 f_{\log n}(t) \tag{3.8}$$

$$f_{gam-wbl}(t) = \pi_1 f_{gam}(t) + \pi_2 f_{wbl}(t) \tag{3.9}$$

$$f_{\log n-wbl}(t) = \pi_1 f_{\log n}(t) + \pi_2 f_{wbl}(t) \tag{3.10}$$

where π_1, π_2 are the mixing weights and $\pi_1 + \pi_2 = 1$, $0 < \pi_1, \pi_2 < 1$ for each of the mixture distribution models.

In order to detect whether a specific distribution is preferred, we used two different goodness-of-fit tests: the mean square error (MSE) test and the Kolmogorov-Smirnov (KS) test. Firstly, we use the MSE test. The MSE value is defined as

$$MSE = \frac{\sum_{i=1}^{n} [F_e(t_i) - F(t_i)]^2}{n-k} \tag{3.11}$$

where $F_e(t)$ is the emprical distribution and $F(t)$ is the cumulative distribution function that is proposed to model the heterogeneous survival data set. k is the number of free parameters in the distribution. As is known, the most appropriate distribution guides the smallest MSE value. Secondly, the KS test is performed. The Kolmogorov_Smirnov statistic KS is defined by

$$KS = \max |F_e(t) - F(t)| \tag{3.12}$$

It is well recognized that the preferred distribution has the smallest value of KS. Furthermore, we used AIC as goodness of fit test because it is one of the most commonly used model selection criteria. AIC value is given as follows

$$AIC = -2LogL + 2d \tag{3.13}$$

where d represents estimated parameters [16]. The smallest AIC value represents the best model.

4. Parameter Estimation

In this study we tend to select mixture distributions with two components. Therefore this study only includes mixtures with two components. Suppose the density of a random variable T has a 2-component mixture form which is given as

$$f\left(t\left|\psi\right.\right)=\sum_{k=1}^{2}\pi_k f_k\left(t\left|\theta_k\right.\right) \tag{4.1}$$

where $\psi=(\pi_1,\pi_2,\theta_1,\theta_2)$ is the vector containing all the unknown parameters in the mixture model. The function $f_k\left(t\left|\theta_k\right.\right)$ is called component density function for $k=1,2.$ with parameter θ_k and π_k is called mixing proportion of kth class satisfying conditions $\pi_k\in(0,1)$ and $\sum_{k=1}^{2}\pi_k=1$ [16].

In finite mixture distribution models, the EM algorithm is a broadly applicable algorithm that provides an iterative procedure for computing maximum likelihood estimators of unknown parameters [3].

Suppose $t_1,t_2,...,t_n$ is an incomplete data and z_1,z_2 are the label values of observed data or latent class variables where the kth element of z_i, z_{ki} is defined to be one or zero, according to whether t_i did or did not arise from the kth component of the mixture. Thus z_i is distributed according to a multinomial distribution with probabilities π_1,π_2 and its density function is given by

$$f\left(z_i\right)=\prod_{k=1}^{2}\pi_k^{z_{ki}} \tag{4.2}$$

The probability density function of t_i, given z_i can be written as

$$f\left(t_i\left|z_i\right.\right)=\prod_{k=1}^{2}\left(f_k\left(t_i;\theta_k\right)\right)^{z_{ki}} \tag{4.3}$$

Therefore the complete data likelihood function is given by

$$f\left(t_1,t_2,...,t_n,z_1,z_2,...z_n\right)=\prod_{i=1}^{n}\prod_{k=1}^{2}\left(f_k\left(t_i;\theta_k\right)\right)^{z_{ki}}\pi_k^{z_{ki}} \tag{4.4}$$

From equation (4.4) the maximum log-likelihood function for complete data can be written as

$$\ln L\left(\psi\left|t_1,t_2,...,t_n,z_1,z_2,...,z_n\right.\right)=\sum_{i=1}^{n}\sum_{k=1}^{2}z_{ki}\ln\left(\pi_k f_k\left(t_i;\theta_k\right)\right) \tag{4.5}$$

The EM algorithm is applied to this problem by treating the z_i as unobservable or missing data. EM algorithm consists of two steps, E and M steps. The E step simply requires to calculate the current conditional expectation of z_{ki} given the observed data $t_1,t_2,...,t_n$,

$$\hat{z}_{ki}=E\left(z_{ki}\left|t_i\right.\right)=\frac{\pi_k f_k\left(t_i\left|\theta_k\right.\right)}{\sum_{k=1}^{2}\pi_k f_k\left(t_i\left|\theta_k\right.\right)} \tag{4.6}$$

$f\left(z_{ki}\left|t_i\right.\right)$ is given by

$$f\left(z_{ki}\left|t_i\right.\right)=\frac{\left(f_k\left(t_i;\theta_k\right)\right)^{z_{ki}}\pi_k^{z_{ki}}}{\sum_{k=1}^{2}\left(f_k\left(t_i;\theta_k\right)\right)^{z_{ki}}\pi_k^{z_{ki}}} \tag{4.7}$$

Therefore the expectation of the complete data log-likelihood function is given by

$$E\left(\ln L\right)=\sum_{i=1}^{n}\sum_{k=1}^{2}\hat{z}_{ki}\ln\left(\pi_k f_k\left(t_i;\theta_k\right)\right) \tag{4.8}$$

In M step, $E(z_{ki}|t_i)$ function which is calculated in E step is maximized. To maximize the $E(\ln L)$, we introduce the Lagrange multiplier λ with the constraint that $\sum_{k=1}^{2}\pi_k = 1$ and we take the derivative of equation (4.9) with respect to all parameters [9].

$$\ln \tilde{L} = \sum_{i=1}^{n}\sum_{k=1}^{2}\hat{z}_{ki}\ln\left(\pi_k f_k\left(t_i;\theta_k\right)\right) - \lambda\left(\sum_{k=1}^{2}\pi_k - 1\right) \tag{4.9}$$

The estimate of π_k (k=1,2.) is defined by

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} z_{ki} \tag{4.10}$$

Suppose $f_k\left(t\,|\theta_k\right)$, in equation (4.1), as the Exponential pdf of kth group for observed data. We can evaluate the maximum likelihood estimator of λ parameter of Exponential distribution with EM for kth group in rth iteration as

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n}\hat{z}_{ki}t_i}{\sum_{i=1}^{n}\hat{z}_{ki}} \tag{4.11}$$

Suppose $f_k\left(t\,|\theta_k\right)$, in equation (4.1), as the Gamma pdf of kth group for observed data. We can evaluate the maximum likelihood estimators of parameters α and β of Gamma distribution with EM for kth group in rth iteration as

$$\hat{\alpha}_{k,(r+1)} = \hat{\alpha}_{k,r} - \frac{\ln(\hat{\alpha}_{k,r}) - \psi(\hat{\alpha}_{k,r}) - \ln\left(\frac{\sum_{i=1}^{n}\hat{z}_{ki}t_i}{\sum_{i=1}^{n}\hat{z}_{ki}}\right) + \frac{\sum_{i=1}^{n}\hat{z}_{ki}\ln(t_i)}{\sum_{i=1}^{n}\hat{z}_{ki}}}{\frac{1}{\hat{\alpha}_{k,r}} - \psi'\left(\hat{\alpha}_{k,r}\right)} \tag{4.12}$$

$$\hat{\beta}_k = \frac{\sum_{i=1}^{n}\hat{z}_{ki}t_i}{\hat{\alpha}_k\sum_{i=1}^{n}\hat{z}_{ki}} \tag{4.13}$$

where $\psi(.)$ and $\psi'(.)$ are a digamma and trigamma functions respectively.

Suppose $f_k\left(t\,|\theta_k\right)$, in equation (4.1), as the Log-normal pdf of kth group for observed data. We can evaluate the maximum likelihood estimators of parameters μ and σ^2 of Log-normal distribution with EM for kth group in rth iteration as

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n}\hat{z}_{ki}\ln t_i}{\sum_{i=1}^{n}\hat{z}_{ki}} \tag{4.14}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n}\hat{z}_{ki}\left(\ln t_i - \hat{\mu}_k\right)^2}{\sum_{i=1}^{n}\hat{z}_{ki}} \tag{4.15}$$

Suppose $f_k\left(t\,|\theta_k\right)$, in equation (4.1), as the Weibull pdf of kth group for observed data. We can evaluate the maximum likelihood estimators of parameters α and β of Weibull distribution with EM for kth group in rth iteration as

$$\hat{\alpha}_{k,(r+1)} = \hat{\alpha}_{k,r} + \frac{A_{k,r} + (1/\hat{\alpha}_{k,r}) - (C_{k,r}/B_{k,r})}{(1/\hat{\alpha}_{k,r}^2) + \left(B_{k,r}D_{k,r} - C_{k,r}^2\right)/B_{k,r}^2} \tag{4.16}$$

$$\hat{\beta}_k = \left(\frac{\sum_{i=1}^{n}\hat{z}_{ki}t_i^{\hat{\alpha}_k}}{\sum_{i=1}^{n}\hat{z}_{ki}}\right)^{1/\hat{\alpha}_k} \tag{4.17}$$

where $A_{k,r} = \frac{\sum_{i=1}^{n}\hat{z}_{ki}\ln t_i}{\sum_{i=1}^{n}\hat{z}_{ki}}$, $B_{k,r} = \sum_{i=1}^{n}\hat{z}_{ki}t_i^{\hat{\alpha}_{k,r}}$, $C_{k,r} = \sum_{i=1}^{n}\hat{z}_{ki}t_i^{\hat{\alpha}_{k,r}}\ln t_i$ and $D_{k,r} = \sum_{i=1}^{n}\hat{z}_{ki}t_i^{\hat{\alpha}_{k,r}}\left(\ln t_i\right)^2$.

5. Analysis of Survival Data

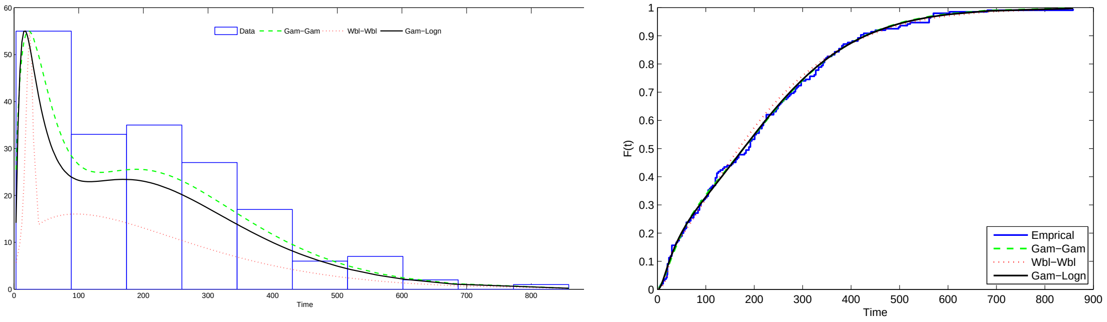
5.1. Analysis of Lung Cancer Survival Data

Lung cancer survival data set consists of survival times of 183 lung cancer patients [4]. A variety of mixture models have been proposed for the data set. Models proposed, parameter estimations of proposed models and the comparison values to select the most adequate model are given in Table 2. The histogram and three probability densities having better fits than others for the survival times of 183 lung cancer patients are given in Figure 1(a). The empirical distribution function and three distribution functions having better fits than others are shown in Figure 1(b).

TABLE 2. The estimated parameters, AIC values, KS test statistics and MSE values for survival times of 183 lung cancer patients

Distributions	**Estimations of Parameters**					**AIC**	**KS***	**MSE** ($\times 10^{-4}$)
Exp-Exp	$\lambda_1 = 206.81$	$\lambda_2 = 205.09$			$\pi_1 = 0.414$	2321.6	0.099	12.000
Gam-Gam	$\alpha_{g1} = 3.795$	$\beta_{g1} = 74.531$	$\alpha_{g2} = 1.684$	$\beta_{g2} = 31.583$	$\pi_1 = 0.665$	2306.0	0.041	**0.6955**
Logn-Logn	$\mu_1 = 4.307$	$\sigma_1 = 1.101$	$\mu_2 = 5.690$	$\sigma_2 = 0.378$	$\pi_1 = 0.583$	2310.0	0.037	1.0500
Wbl-Wbl	$\alpha_{w1} = 5.323$	$\beta_{w1} = 25.168$	$\alpha_{w2} = 1.377$	$\beta_{w2} = 244.65$	$\pi_1 = 0.092$	**2301.4**	0.045	2.0662
Exp-Gam	$\lambda = 126.40$	$\alpha_g = 5.571$	$\beta_g = 57.826$		$\pi_1 = 0.594$	2307.8	0.037	0.8057
Exp-Logn	$\lambda = 146.13$	$\mu = 5.741$	$\sigma = 0.365$		$\pi_1 = 0.680$	2307.8	0.037	0.7316
Exp-Wbl	$\lambda = 120.25$	$\alpha_w = 2.064$	$\beta_w = 332.69$		$\pi_1 = 0.489$	2308.8	0.045	1.5509
Gam-Logn	$\alpha_g = 3.685$	$\beta_g = 76.612$	$\mu = 3.850$	$\sigma = 1.029$	$\pi_1 = 0.626$	2306.6	**0.034**	0.7567
Gam-Wbl	$\alpha_g = 2.041$	$\beta_g = 17.267$	$\alpha_w = 1.668$	$\beta_w = 278.38$	$\pi_1 = 0.201$	2305.6	0.036	0.7488
Logn-Wbl	$\mu = 3.850$	$\sigma = 1.071$	$\alpha_w = 1.874$	$\beta_w = 301.98$	$\pi_1 = 0.335$	2306.2	0.035	0.6957

*p values are 0.72, 1, 1, 0.99, 1, 1, 0.99, **0.99**, 1, 1 at the 5% significance level



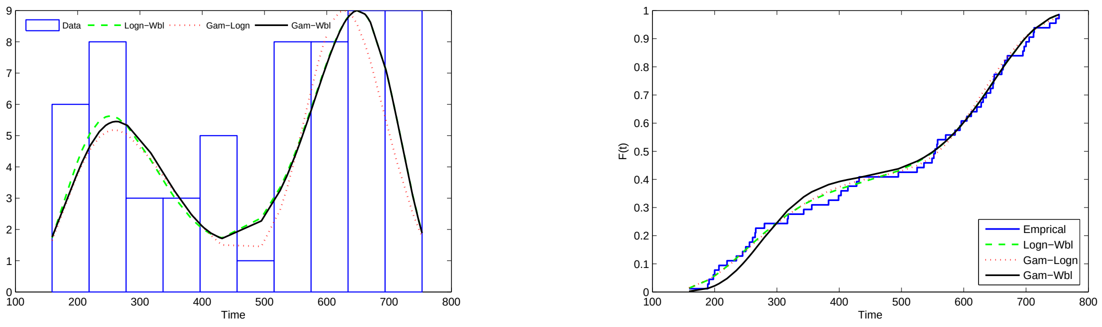
(a) (b)

FIGURE 1. (a) The probability densities of the fitted distributions and a histogram (b) The empirical distribution function and the fitted distribution functions for survival times of 183 lung cancer patients

Table 3. The estimated parameters, AIC values, KS test statistics and MSE values for actual time-of-death for irradiated mice

Distributions	**Estimations of Parameters**					**AIC**	**KS***	**MSE** $(\times 10^{-4})$
Exp-Exp	$\lambda_1 = 488.89$	$\lambda_2 = 488.69$			$\pi_1 = 0.614$	869.038	0.293	337.00
Gam-Gam	$\alpha_{g1} = 10.421$	$\beta_{g1} = 28.609$	$\alpha_{g2} = 92.975$	$\beta_{g2} = 6.824$	$\pi_1 = 0.433$	773.542	0.054	4.8282
Logn-Logn	$\mu_1 = 5.650$	$\sigma_1 = 0.313$	$\mu_2 = 6.447$	$\sigma_2 = 0.104$	$\pi_1 = 0.434$	773.07	0.047	4.6731
Wbl-Wbl	$\alpha_{w1} = 3.761$	$\beta_{w1} = 313.53$	$\alpha_{w2} = 9.731$	$\beta_{w2} = 657.20$	$\pi_1 = 0.397$	774.922	0.061	4.3954
Exp-Gam	$\lambda = 380.18$	$\alpha_g = 5.690$	$\beta_g = 85.905$		$\pi_1 = 0.0001$	809.65	0.188	86.000
Exp-Logn	$\lambda = 405.09$	$\mu = 6.102$	$\sigma = 0.452$		$\pi_1 = 0.0001$	815.276	0.201	98.000
Exp-Wbl	$\lambda = 328.74$	$\alpha_w = 3.051$	$\beta_w = 549.21$		$\pi_1 = 0.0002$	802.078	0.157	73.000
Gam-Logn	$\alpha_g = 94.19$	$\beta_g = 6.743$	$\mu = 5.661$	$\sigma = 0.323$	$\pi_1 = 0.559$	772.762	0.048	4.3470
Gam-Wbl	$\alpha_g = 24.481$	$\beta_g = 11.693$	$\alpha_w = 9.732$	$\beta_w = 657.55$	$\pi_1 = 0.401$	772.446	0.072	11.000
Logn-Wbl	$\mu = 5.619$	$\sigma = 0.299$	$\alpha_w = 9.695$	$\beta_w = 657.50$	$\pi_1 = 0.404$	**771.708**	**0.044**	**3.0028**

*p values are 0.00, 0.99, 1, 0.99, 0.16, 0.10, 0.34, 1, 0.98, **1** at the 5% significance level



(a) (b)

Figure 2. (a) The probability densities of the fitted distributions and a histogram (b) The empirical distribution function and the fitted distribution functions for actual time-of-death for irradiated mice

5.2. Analysis of Actual Time-of-Death for Irradiated Mice Data

The data are actual time-of-death for irradiated mice given in Elandt-Johnson and Johnson (1980) [6]. Jiang and Murthy (1995) [11] confine their attention to deaths due to thymic lymphoma (22 data points) and reticulum cell sarcoma (38 points). Proposed models, parameter estimations of proposed models and the comparison values to select the most adequate model are given in Table 3. The histogram and three probability densities having better fits than others are given in Figure 2(a). The empirical distribution function and three distribution functions having better fits than others are shown in Figure 2(b).

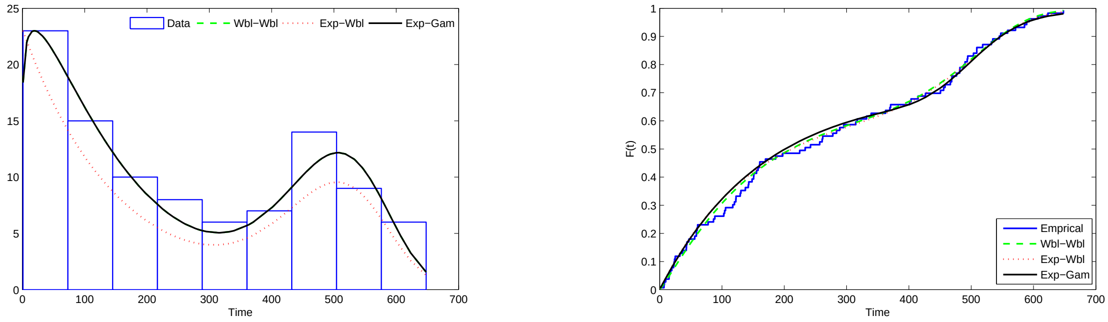
5.3. Analysis of Failure Times for Oral Irrigators

Another real data set previously studied by [12] consists of failure times for oral irrigators. This data set is dealing with failure times for oral irrigators. Models proposed, parameter estimations of proposed models and the comparison values to select the most adequate model are given in Table 4. The histogram and three probability densities having better fits than others are given in Figure 3(a). The empirical distribution function and three distribution functions having better fits than others are shown in Figure 3(b).

TABLE 4. The estimated parameters, AIC values, KS test statistics and MSE values for oral irrigator data set

Distributions	**Estimations of Parameters**					**AIC**	**KS***	**MSE** $(\times 10^{-4})$
Exp-Exp	$\lambda_1 = 263.32$	$\lambda_2 = 265.76$			π_1=0.493	1295.302	0.113	34.000
Gam-Gam	$\alpha_{g1} = 1.127$	$\beta_{g1} = 145.27$	$\alpha_{g2} = 65.224$	$\beta_{g2} = 7.790$	$\pi_1 = 0.707$	1261.932	0.049	3.975
Logn-Logn	$\mu_1 = 4.548$	$\sigma_1 = 1.203$	$\mu_2 = 6.199$	$\sigma_2 = 0.146$	$\pi_1 = 0.684$	1276.998	0.090	15.000
Wbl-Wbl	$\alpha_{w1} = 522.19$	$\beta_{w1} = 6.871$	$\alpha_{w2} = 149.28$	$\beta_{w2} = 1.115$	$\pi_1 = 0.351$	1260.654	**0.045**	**3.225**
Exp-Gam	$\lambda = 165.74$	$\alpha_g = 65.505$	$\beta_g = 7.748$		$\pi_1 = 0.711$	1260.506	0.059	4.791
Exp-Logn	$\lambda = 167.73$	$\mu = 6.226$	$\sigma = 0.119$		$\pi_1 = 0.717$	1260.646	0.059	4.866
Exp-Wbl	$\lambda = 148.17$	$\alpha_w = 520.98$	$\beta_w = 6.839$		$\pi_1 = 0.656$	**1259.608**	0.058	4.058
Gam-Logn	$\alpha_g = 53.299$	$\beta_g = 9.386$	$\mu = 4.566$	$\sigma = 1.206$	$\pi_1 = 0.308$	1276.284	0.091	16.000
Gam-Wbl	$\alpha_g = 1.143$	$\beta_g = 126.21$	$\alpha_w = 521.01$	$\beta_w = 6.792$	$\pi_1 = 0.648$	1260.996	0.050	3.441
Logn-Wbl	$\mu = 4.412$	$\sigma = 1.199$	$\alpha_w = 505.64$	$\beta_w = 5.686$	$\pi_1 = 0.386$	1272.738	0.081	11.000

*p values are 0.43, 0.99, 0.78, **0.99**, 0.99, 0.99, 0.99, 0.78, 0.99, 0.78 at the 5% significance level



(a) (b)

FIGURE 3. (a) The probability densities of the fitted distributions and a histogram (b) The empirical distribution function and the fitted distribution functions for oral irrigators data set

6. Discussion and Conclusions

The survival analysis has become an increasingly active and very important area of research. Various theoretical distributions are used to model survival time. In this paper, mixtures of these distributions are considered for modelling heterogeneous survival time data. Heterogeneous survival time data can have two different distributions before and after a certain time because many factors affect 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. Each of the different parts 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. In this study, we also apply the mixture of Log-normal distribution with the Exponential, Gamma and Weibull distributions and we compare the mixture of two identical distributions and two different distributions. The AIC values, KS test statistics and MSE are calculated to determine the most appropriate distribution for the present data sets.

A variety of mixture models have been proposed for each of the three data sets and majority of these mixture models fit these data sets successfully. The best model among the two component mixture distribution models is the mixture of Gamma and Log-normal for survival times of 183 lung cancer patients according to KS test statistics and alternative models are determined as the mixture of two Gamma distributions and two Weibull distributions according to MSE and AIC values for this data set respectively.

For the mice data, the mixture of Log-normal and Weibull distributions is found the best model according to all comparison criteria. Mixture of Gamma and Log-normal distributions and mixture of Gamma and Weibull distributions can be used alternative models for the mice data according to AIC values. Finally for the oral irrigator data set, the best fit model is the mixture of two Weibull distributions according to all comparison criteria. Mixture of Exponential and Gamma distributions, mixture of Exponential and Log-normal distributions and mixture of Exponential and Weibull distributions are alternative models also for the oral irrigator data set.

In conclusion, it's known that mixture distribution model approaches are provided better fit for the heterogeneous data set. Mixture of identical distributions can be used with different number of components for the heterogeneous data sets having two or more modes. Also as seen in the previous studies, the mixture models of two different distributions approach is a new method for heterogeneous data sets. However, it can be difficult to model the heterogeneous data sets with the mixture models consisted of more than two different distributions. It will be subject of a new study to see how the mixture models can be generalized for the heterogeneous data sets which have more than two modes and how the parameters of these models are estimated. As a result of this study, the mixture models of two different distributions are powerful alternative models compared to mixture of two identical distributions in case of heterogeneous survival data sets which have two components.

References

[1] Berkson, J. and Gage, RP. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259), 501-515.

[2] Chen, WC., Hill, B., Greenhouse, J. and Fayos, J. (1985). Bayesian analysis of survival curves for cancer patients following treatment. *Bayesian statistics*, 2, 299-328.

[3] Dempster, AP., Laird, NM. and Rubin, DB. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, Series B (Methodological), 1-38.

[4] Dirican, A. (2004). The patients with lung cancer were diagnosed in our clinic as a prospective evaluation and determination of the factors that affect survival. Master's thesis, Ondokuz Mayis University.

[5] Ebeling, CE. (1997). *An introduction to reliability and maintainability engineering.* The Mcgraw-Hill Companies, Inc., New York.

[6] Elandt-Johnson, RC. and Johnson, NL. (1980). *Survival models and data analysis.* John Wiley & Sons.

[7] Erişoğlu, M., Çalış, N., Servi, T., Erişoğlu, Ü. and Topaksu, M. (2011a). The mixture distribution models for interoccurence times of earthquakes. *Russian Geology and Geophysics*, 52, 685-692.

[8] Erişoğlu, Ü., Erişoğlu, M. and Erol, H. (2011b.) A mixture model of two different distributions approach to the analysis of heterogeneous survival data. *International Journal of Computational and Mathematical Sciences*, 5(2) ,75-79.

[9] Erişoğlu, Ü., Erişoğlu, M. and Erol, H. (2012). Mixture model approach to the analysis of heterogeneous survival data. *Pakistan Journal of Statistics*, 28(1), 115-130.

[10] Everitt, BS. and Hand, DJ. (1981). *Finite mixture distributions.* London: Chapman and Hall.

[11] Jiang, R. and Murthy, D. (1995). Modeling failure-data by mixture of 2 weibull distributions: a graphical approach. *Reliability, IEEE Transactions on*, 44(3), 477-488.

[12] Jiang, R.and Murthy, D. (1997). Two sectional models involving three weibull distributions. *Quality and Reliability Engineering International*, 13(2), 83-96.

[13] Lawless, JF. (2011). *Statistical models and methods for lifetime data.* John Wiley & Sons.

[14] Lee, ET. and Wang, J. (2003). *Statistical methods for survival data analysis.* John Wiley & Sons.

[15] Marin, J., Rodriguez-Bernal, M. and Wiper, M. (2005). Using weibull mixture distributions to model heterogeneous survival data. *Communications in Statistics-Simulation and Computation*, 34(3), 673-684.

[16] McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

[17] Qian, J. (1994). A bayesian weibull survival model. PhD thesis, Institute of Statistical and Decision Sciences Duke University: North Corolina.

[18] Zhang, Y. (2008). Parametric mixture models in survival analysis with applications. PhD thesis, Temple University.