Mixture Model of the Exponential, Gamma and Weibull Distributions to Analyse Heterogeneous Survival Data

Yusuf Abbakar Mohammed[1,2*], Bidin Yatim[1] and Suzilah Ismail[1]

[1]*Department of Mathematics and Statistics School of Quantitative Sciences, University Utara Malaysia, Sintok, Malaysia.*
[2]*Department of Mathematics and Statistics, University of Maiduguri, Maiduguri, Nigeria.*

Authors' contributions

This work was carried out in collaboration between all authors. All authors read and approved the final manuscript.

ABSTRACT

Aims: In this study a survival mixture model of three components is considered to analyse survival data of heterogeneous nature. The survival mixture model is of the Exponential, Gamma and Weibull distributions.
Methodology: The proposed model was investigated and the Maximum Likelihood (ML) estimators of the parameters of the model were evaluated by the application of the Expectation Maximization Algorithm (EM). Graphs, log likelihood (LL) 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. The model was compared with the survival mixture models corresponding to each component.
Results: The graphs, LL and AIC values showed that the proposed model fits the real data better than the pure classical survival models corresponding to each component. Also the proposed model fits the real data better than the survival mixture models corresponding to each component.
Conclusion: The proposed model showed that survival mixture models are flexible and maintain the features of the pure classical survival model and are better option for modelling heterogeneous survival data.

_____

Corresponding author: Email: yusufabbakarm@yahoo.com;

1. INTRODUCTION

Survival analysis is concerned with the investigation of a particular event happening within a given duration of time. Survival analysis is widely applied in many fields such as Medical studies, biology, social sciences, economics and engineering to mention a few. The most commonly used methods in survival analysis are the nonparametric methods. Pure classical parametric survival models are commonly employed in survival analysis; they are better option when the chosen distribution seems to fit the data properly. The Exponential, Gamma, and Weibull distributions are commonly used in the literature for modeling survival data [1-4]. Survival mixture models are most appropriate for modeling survival data when the data are believed to be heterogeneous in nature. Recently, many research works employed the methods of survival mixture models to analyse survival data. A two component mixture model of Weibull distributions was proposed to anlaysed survival data where the parameters of the model were estimated by the weighted least squares method [5]. A two component survival mixture model of Weibull distributions was proposed; where the parameters of the models were estimated by graphical approach [6]. Also a new technique for evaluating the parameters of a two component survival mixture model of Weibull distribution was developed to analyse survival data [7].

The Expectation Maximization Algorithm (EM) was employed to evaluate the parameters of the Weibull-Weibull survival mixture model of two components and the EM stability was investigated [8]. Two components survival mixture models of Gamma-Gamma, Weibul-Weibull and Lognormal-lognormal were proposed to analyse survival data. Model selection method was used to select the model which better represents the real data [9]. A survival mixture of mixed distribution was employed for analyzing heterogeneous survival data. The mixed distribution model is a two components survival model of the Extended Exponential-Geometric (EEG) distribution where the EM was employed to estimate the model parameters [10]. Few researchers considered survival mixture models of different distributions. A two component parametric survival mixture model of different distributions of Exponentiated Pareto and Exponential distributions was used to model survival data [11]. Two components survival mixture models of different distributions consisting of an Exponential-Gamma, an Exponential-Weibull and a Gamma-Weibull models were proposed for analysing heterogeneous survival data by employing EM [12].

Three components parametric survival mixture models did not receive much attention. In a situation of an open heart surgery study; the risk of death after surgery was divided into three different time overlapping phases which are better analysed by a three component mixture model [13-15]. A parametric survival mixture model of the Exponential, Gamma and Weibull distributions was considered to fit heterogeneous survival data. Simulated data were used to investigate the stability and consistency of the EM [16]. In another study, model selection technique was employed to compare the parametric survival mixture model of the Exponential, Gamma and Weibull distributions with the parametric survival mixture model corresponding to each component [17]. A three component parametric survival mixture model of Weibull distributions was proposed to model survival data by applying Bayesian estimation method [18]. EM was usually employed on data believed to consist of some missing or unobserved observations [19]. The parameters of survival mixture models are commonly evaluated by implementing the EM Algorithm [20,21].

In this study, real data were used to investigate the flexibility and appropriateness of a three component survival mixture of the Exponential, Gamma and Weibull distributions in modelling heterogeneous survival data. The arrangement of this article is as follows; in section two, the survival analysis and some important probability functions were highlighted. The development of survival mixture model of three components and the application of the EM in estimating the ML parameters of the model were discussed. Section three was devoted to data application to evaluate the parameters of the proposed model and the discussion of the result. Finally in section four the summary and conclusion were presented.

2. SURVIVAL ANALYSIS AND THREE COMPONENTS MIXTURE MODEL

Survival analysis deals with applying particular statistical methods to model and analyse survival data. The focus of interest is the occurrence of a particular event of interest within a given period of time. The response of primary interest is the random variable *T* which is non-negative and gives the survival time of an object or an individual. The survival time can be represented by three functions which are interchangeable The probability density function (pdf) is denoted by $f(t)$, which is expressed as

$$f(t) = \frac{dF(t)}{dt} \tag{1}$$

Where $F(t)$ is the distribution function of the random variable *T*. The graphical representation of the probability density function is frequently used in the literature, the graph of $f(t)$, is commonly referred to as the density curve. The area between the curve and the *t* axis of the non-negative density function $f(t)$ is equal to 1. The survival function $S(t)$ is commonly expressed as

$$S(t) = 1 - F(t) \tag{2}$$

Which estimated the probability of an individual surviving beyond a specified time t. The survival function $S(t)$ is a continuous monotonic decreasing function with $S(0) = \lim_{t \to 0} S(t) = 1$ and $S(\infty) = \lim_{t \to \infty} S(t) = 0$. The hazard function denoted by $h(t)$, and is given by

$$h(t) = \frac{f(t)}{S(t)} \tag{3}$$

Which gives the probability of an individual will fail within a small interval $(t, t + \Delta t)$, provided that the individual was alive until the beginning of that interval.

Pure classical parametric survival models are powerful methods in survival analysis. They are preferred when the chosen probability distribution appropriately represents the data. The Exponential, Gamma and Weibull distributions are among the most important and frequently used distributions in survival analysis [1,2,3,4].

The probability density function $f(t)$ and survival function $S(t)$ of these distributions are highlighted below.

Exponential Distribution

$$f_E(t) = \lambda e^{-\lambda t} \quad t, \lambda > 0 \tag{4}$$

Where λ is a scale parameter

$$S_E(t) = e^{-\lambda t} \tag{5}$$

Gamma distribution

$$f_G(t) = t^{\alpha - 1} \frac{e^{-t/\beta}}{\beta^\alpha \Gamma(\alpha)} \quad t, \alpha, \beta > 0 \tag{6}$$

Where α is the shape parameter and β is the scale parameter

$$S_G(t) = 1 - \frac{\Gamma_x(\alpha)}{\Gamma(\alpha)} \tag{7}$$

Where $\Gamma_x(\alpha) = \int_0^x t^{\alpha-1} e^{-1} dt$ is known as the

incomplete Gamma function.

Weibull Distribution

$$f_W(t) = \frac{\alpha}{\beta} \left(\frac{t}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{t}{\beta}\right)^\alpha\right) \quad t, \alpha, \beta > 0 \tag{8}$$

Where α is the shape parameter and β is the scale parameter

$$S_W(t) = \exp\left(-\left(\frac{t}{\beta}\right)^\alpha\right) \tag{9}$$

In survival analysis, mixture models are frequently used because they are flexible. They are the best option where pure classical parametric survival models do not fit the data of heterogeneous nature [20,22]. Survival mixture model of three components is used when it is believed that the data consist of three subpopulation or subgroups. Equation (10) represents a survival mixture model of three components.

$$f_{X,Y,Q}(t; \Theta) = \pi_1 f_X(t; \theta_X) + \pi_2 f_Y(t; \theta_Y) + \pi_3 f_Q(t; \theta_Q) \tag{10}$$

Where the vector $\Theta = (\pi_1, \pi_2, \pi_3, \theta_X, \theta_Y, \theta_Q)$, represents the vector the parameters of the mixture model. The functions $f_X(t; \theta_X), f_Y(t; \theta_Y)$ and $f_Q(t; \theta_Q)$ are the probability density function corresponding to each component with some parameters θ_X, θ_Y and θ_Q respectively.

In this paper, a survival mixture model of three components of different distributions is proposed to analyse survival data which is believed to be heterogeneous. The paper proposed a survival mixture model of the Exponential, Gamma and Weibull distributions. The proposed model is defined as

$$f_{E\_G\_W}(t; \Theta) = \pi_1 f_E(t; \lambda) + \pi_2 f_G(t; \alpha_1, \beta_1) + \pi_3 f_W (t; \alpha_2, \beta_2)$$

(11)

Where $\pi_i's$ represent the mixing probabilities of the three subpopulations with $\sum_{i=1}^{3} \pi_i = 1$. The functions f_E, f_G and f_W, as defined in (4), (6) and (8), represent the probability functions of the Exponential, Gamma and Weibull distributions respectively.

One of the most efficient and effective methods commonly used to estimate the ML estimators of finite mixture models is the EM Algorithm [21].

Let $t_1, t_2, ..., t_n$ be a set of observations of n incomplete data and z_1, z_2, z_3 be a set of missing observations, where $z_{ki} = z_k(t_i) = 1$, if the observation belongs to the kth component and 0 otherwise for $k = 1, 2, 3$ and $i = 1, 2, ..., n$. On the implementation of the EM to the mixture model, the variables $z`s$ are considered as missing values. The EM consists of two different steps, the first one is the Expectation step or the E-step and the second one is the Maximization step or the M-step.

The $z`s$ variables are treated as missing observations in the E-step, the hidden variable vector $z_i = [z_{1i}, z_{2i}, z_{3i}]$ are estimated by the evaluation of the conditional expectation $E(z_{ki} | t_i)$.

Thus

$$\hat{z}_{1i} = E(z_{1i} | t_i) = \frac{\pi_1 f_X(t_i; \theta_X)}{\pi_1 f_X(t_i; \theta_X) + \pi_2 f_Y(t_i; \theta_Y) + \pi_3 f_Q(t_i; \theta_Q)}$$ (13)

$$\hat{z}_{2i} = E(z_{2i} | t_i) = \frac{\pi_2 f_Y(t_i; \theta_Y)}{\pi_1 f_X(t_i; \theta_X) + \pi_2 f_Y(t_i; \theta_Y) + \pi_3 f_Q(t_i; \theta_Q)}$$ (14)

$$\hat{z}_{3i} = E(z_{3i} | t_i) = \frac{\pi_3 f_Q(t_i; \theta_Q)}{\pi_1 f_X(t_i; \theta_X) + \pi_2 f_Y(t_i; \theta_Y) + \pi_3 f_Q(t_i; \theta_Q)}$$ (15)

The functions $E(z_{1i} | t_i), E(z_{2i} | t_i)$ and $E(z_{3i} | t_i)$ calculated in the E-step will be maximized in the M-step of the EM under the condition $\sum_{i=1}^{3} \pi_i = 1$. The evaluation of the mixing probabilities π_i and vector of parameter $\theta = [\theta_X, \theta_Y, \theta_Q]$, is by the implementation of the Lagrange method. The mixing probabilities will be obtained by;

$$\hat{\pi}_1 = \frac{\sum_{i=1}^{n} \hat{z}_{1i}}{n}$$ (16)

$$\hat{\pi}_2 = \frac{\sum_{i=1}^{n} \hat{z}_{2i}}{n}$$ (17)

$$\hat{\pi}_3 = \frac{\sum_{i=1}^{n} \hat{z}_{3i}}{n}$$ (18)

The ML estimator of the parameter λ of the Exponential distribution for the proposed model can be obtained by the equation (19) [12,16,17].

$$\hat{\lambda} = \frac{\left(\sum_{i=1}^{n} \hat{z}_{1i} \right)}{\left(\sum_{i=1}^{n} \hat{z}_{1i} t_i \right)}$$ (19)

The maximum likelihood estimators of the parameters α_1 and β_1 of the Gamma distribution for proposed model are evaluated using equations (20) and (21) respectively [9,12,16,17].

$$\hat{\alpha}_{(r+1)} = \hat{\alpha}_r - \frac{\ln(\hat{\alpha}_r) - \Psi(\hat{\alpha}_r) - \ln\left(\frac{\sum_{i=1}^{n} \hat{z}_{2i} t_i}{\sum_{i=1}^{n} \hat{z}_{2i}}\right) + \frac{\sum_{i=1}^{n} \hat{z}_{2i} \ln(t_i)}{\sum_{i=1}^{n} \hat{z}_{2i}}}{\frac{1}{\hat{\alpha}_r} - \Psi'(\hat{\alpha}_r)} \quad (20)$$

And

$$\hat{\beta} = \left(\frac{\sum_{i=1}^{n} \hat{z}_{2i} t_i}{\hat{\alpha} \sum_{i=1}^{n} \hat{z}_{2i}}\right) \quad (21)$$

Where r is the number of Newton-Raphson iteration within EM Algorithm and $\Psi(.)$ and $\Psi'(.)$ are a digamma and trigamma functions respectively.

The shape and scale parameters α_2 and β_2 of the Weibull distribution in the proposed model are obtained by solving the equations (22) and (23) respectively [9,12,16,17].

$$\hat{\alpha}_{(r+1)} = \hat{\alpha}_r + \frac{A_r + \left(\frac{1}{\hat{\alpha}_r}\right) - \frac{C_r}{B_r}}{\left(\frac{1}{\hat{\alpha}_r^2}\right)\frac{(B_r D_r - C_r^2)}{B_r^2}} \quad (22)$$

Where

$$A_r = (\sum_{i=1}^{n} \hat{z}_{3i})^{-1} \sum_{i=1}^{n} \hat{z}_{3i} \ln t_i \ , \ B_r = \sum_{i=1}^{n} \hat{z}_{3i} t_i^{\hat{\alpha}_r}$$

$$C_r = \sum_{i=1}^{n} \hat{z}_{3i} t_i^{\hat{\alpha}_r} \ln t_i$$

$$D_r = \sum_{i=1}^{n} \hat{z}_{3i} t_i^{\hat{\alpha}_r} (\ln t_i)^2 \text{ and } r \text{ is the number of}$$

Newton-Raphson iteration within EM.

$$\hat{\beta} = \left(\frac{\sum_{i=1}^{n} \hat{z}_{3i} t_i^{\hat{\beta}}}{\sum_{i=1}^{n} \hat{z}_{3i}}\right)^{\frac{1}{\hat{\alpha}}} \quad (23)$$

3. REAL DATA APPLICATION AND DISCUSSION

The real data analysed in this section are the Kidney Catheter data which is included as one of the datasets in famous *survival* package [23] of the R statistical software [24]. The data were studied originally in [25]. The data give the recurrence times to infection, at the time of inserting catheters of kidney patients using portable dialysis equipment. It consists of 76 observations and 7 variables. The proposed model was used to analyse the data and then it was compared with the pure classical parametric survival models corresponding to each component using Log-likelihood (LL) and Akiake Information Criterion (AIC) value. Table 1. shows that the LL value of the proposed model is higher than that of the pure classical survival models and also the AIC value of the proposed model is lower than that of the pure classical survival models which makes the proposed model suitable for the real data used.

The proposed model was graphically compared with pure classical parametric survival models corresponding to each component of the mixture model. The probability functions of proposed model and the pure classical parametric survival models along with the histogram of the Kidney Cather data were presented in Fig. 1. Fig. 1. shows that the proposed model analysed the real data better than the individual pure parametric survival models.

Table 1. The parameters, LL and AIC values for the Kidney Catheter data

Model	Estimates	LL	AIC
Exponential	$\hat{\lambda}$ = 132.95	-341.7	685.4
Weibull	$\hat{\alpha}$ = 0.89, $\hat{\beta}$ = 128.00	-340.9	685.8
Gamma	$\hat{\alpha}$ = 0.86, $\hat{\beta}$ = 156.96	-341.20	686.40
Mixture model	$\hat{\lambda}$ = 32.79	-331.50	677.01
	$\hat{\alpha}_1$ = 20.38, $\hat{\beta}_1$ = 7.73		
	$\hat{\alpha}_2$ = 3.91, $\hat{\beta}_2$ = 441.56		
	$\hat{\pi}_1$ = 0.52, $\hat{\pi}_2$ = 0.29		

The Kidney Catheter data were used to compare the proposed model with the survival mixture models of the Exponential, Gamma and Weibull distributions, respectively to select the model that fits the data appropriately. Table 2. displays the estimated parameters of each model together with the LL and AIC values. It is observed that the proposed model represents the real data better than the other models. Also proposed model was compared graphically with the survival mixture models of the Exponential, Gamma and Weibull distributions, respectively. Fig. 2. shows the comparison of the density function of the proposed model with the other models. It is also observed that the proposed model represents the real data better than the other models.
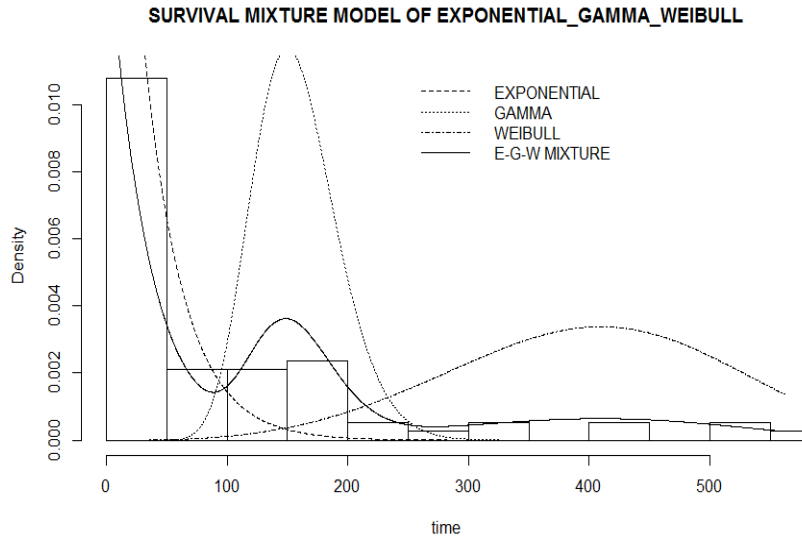

Fig. 1. The probability density functions of proposed model and the pure classical distribution of the Kidney Catheter data
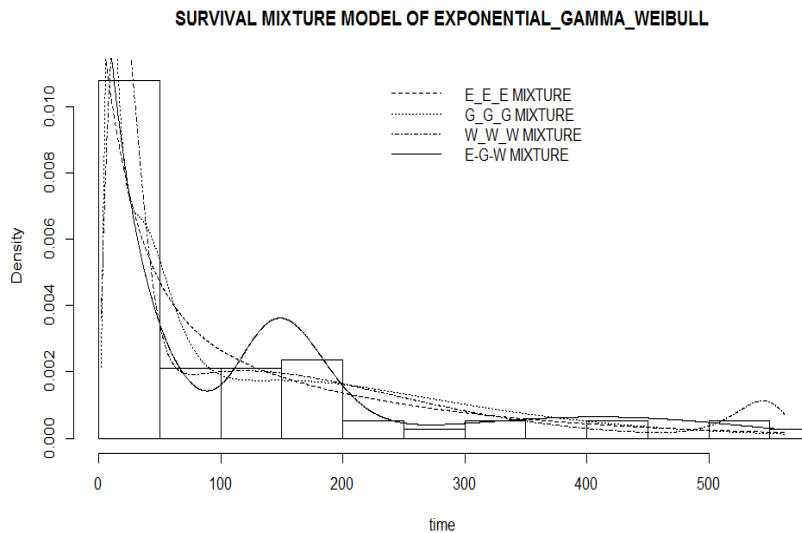

Fig. 2. The probability density functions of the proposed model and the mixture models of each component of the Kidney Catheter data

Table 2. LL and AIC of the proposed model and parametric survival mixture models of the same distribution

	E_G_W		E1_E2_E3		G1_G2_G3		W1_W2_W3
$\hat{\pi}_1$	0.52	$\hat{\pi}_1$	0.26	$\hat{\pi}_1$	0.19	$\hat{\pi}_1$	0.41
$\hat{\pi}_2$	0.29	$\hat{\pi}_2$	0.22	$\hat{\pi}_2$	0.30	$\hat{\pi}_2$	0.53
$\hat{\lambda}$	32.79	$\hat{\lambda}_1$	0.04	$\hat{\alpha}_1$	3.59	$\hat{\alpha}_1$	1.82
$\hat{\alpha}_1$	20.38	$\hat{\lambda}_2$	0.01	$\hat{\alpha}_2$	4.11	$\hat{\alpha}_2$	1.69
$\hat{\alpha}_2$	3.91	$\hat{\lambda}_3$	0.01	$\hat{\alpha}_3$	2.97	$\hat{\alpha}_3$	26.56
$\hat{\beta}_1$	7.73			$\hat{\beta}_1$	3.44	$\hat{\beta}_1$	26.19
$\hat{\beta}_2$	441.56			$\hat{\beta}_2$	11.09	$\hat{\beta}_2$	202.39
				$\hat{\beta}_3$	79.42	$\hat{\beta}_3$	545.34
LL	*-331.50*	LL	-339.46	LL	-337.30	LL	-331.91
AIC	*677.01*	AIC	682.91	AIC	690.61	AIC	679.83

4. CONCLUSION

This article proposed a survival mixture model of three components of the Exponential, Gamma and Weibull distributions to analyse survival data which is believed to be heterogeneous. Real data were used to estimate the parameters of the model. EM algorithm was employed in estimating the ML parameters of the proposed model. The comparison of the proposed model with the pure classical survival models and the survival mixture models corresponding to each distribution showed that the proposed model represents the data better than the other models. The proposed model showed that the survival mixture models are flexible and maintain the feature of pure classical survival models and they are better option to model heterogeneous survival data.

*Note: The R language version 3.0.2 (2013-09-25) http://CRAN.R-project.org was used for all the calculations and graphs

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Ibrahim JG, Chen MH, Sinha D. Bayesian survival analysis. New York: Springer-verlag; 2001.
2. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. New Jersey: John Wiley & Sons, Inc. Hoboken; 2002.
3. Lawless JF. Statistical models and methods of lifetime data. 2nd ed. New Jersey: John Wiley and Sons, Inc. Hoboken; 2003.
4. Lee ET, Wang JW. Statistical methods for survival time data analysis. 3rd ed. New Jersey: John Wiley & son; 2003.
5. Cheng SW, Fu JC. Estimation of mixed weibull parameters in life testing. Reliability, IEEE Transactions. 1982;31(4):377-81. Available:http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=522138210.1109/TR.1982.5221382
6. Jiang S, Kececioglu D. Graphical representation of two mixed-Weibull distributions. IEEE Transaction on Reliability. 1992a; 41:241-47. Available:http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=25778910.1109/24.257789
7. Jiang S, Kececioglu D. Maximum likelihood estimates, from censored data, for mixed-Weibull distributions. IEEE Transaction on Reliability. 1992b;41,248-55. DOI:10.1109/24.257791.
8. Zhang Y. Parametric mixture models in survival analysis with application, (Doctoral Dissertation) UMI Number: 3300387, Graduate School, Temple University; 2008.
9. Erisoglu U. Erisoglu M. Erol H. Mixture model approach to the analysis of heterogeneous survival time data. Pakistan Journal of Statistics. 2012;28(1):115-30.

10. Erisoglu U. Erol H. Modelling heterogeneous survival data using mixture of extended exponential-geometric distributions. Communications in Statistics -Simulation and Computation. 2010;39(10):1939-52. Available:http://www.dx.doi.org/10.1080/03610918.2010.524335

11. Abu -Zinadah HH. A study on mixture of exponentiated pareto and exponential distributions. Journal of Applied Sciences Research. 2010;6(4):358-76.

12. Erisoglu U, Erisoglu M, Erol H. A mixture model of two different distributions approach to the analysis of heterogeneous survival data. International Journal of Computational and Mathematical Sciences.2011;5(2). Available:http://www.scopus.com/inward/record.url?eid=2-s2.0-78449258657&partnerID=40&md5=901faa5759d0767b0b2676000e17839c

13. Blackstone EH, Naftel DC, Turner ME. The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. Journal of the American Statistical Association. 1986;81(395):615-624. Available:http://www.jstor.org/stable/2288989

14. Ng ASK, McLachlan GJ, Yau KKW, Lee AH. Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. Statistics in Medicine. 2004;23(17):2729-44. Available:http://www.scopus.com/inward/record.url?eid=2-s2.0-4444257472&partnerID=40&md5=139f3273239b2d5525b68c728faf99e3

15. Phillips N, Coldman A, McBride ML. Estimating cancer prevalence using mixture models for cancer survival. Statistics in Medicine. 2002;21(9):1257-70. Available:http://dx.doi.org/10.1002/sim.1101 DO - 10.1002/sim.1101

16. Mohammed YA, Yatim B, Ismail S. A simulation study of parametric mixture model of three different distributions to analyse heterogeneous survival data. Modern Applied Science. 2013;7(7):1-9. Available:http://dx.doi.org/10.5539/mas.v7n7p1

17. Mohammed YA, Yatim B, Ismail S. A parametric mixture model of three different distributions: An approach to analyse heterogeneous survival data. Proceedings of the 21st National Symposium on Mathematical Sciences (SKSM21) AIP Conf. Proc. 2014;1605:1040-45. DOI:10.1063/1.4887734.

18. Marín JM, Rodríguez-Bernal MT, Wiper MP. Using weibull mixture distributions to model heterogeneous survival data. Communications in Statistics: Simulation and Computation. 2005;34(3):673-84.

19. Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). Journal of Royal Statistical Society. Series B. 1977;39:1-38. Available:http://www.jstor.org/stable/2984875

20. McLachlan GJ, Peel D. Finite mixture models. New Jersey: John Wiley & Sons, Inc.; 2000.

21. McLachlan GJ, Krishnan T. The EM algorithm and extensions 2nd ed. New Jersey: John Wiley & Sons, Inc.; 2008.

22. Fruhwirth-Schnatter S. Finite mixture and markovs switching models. Springer; 2006.

23. Therneau T. A package for survival analysis in S. R package version. 2.37-4; 2013 Retrieved from Available:http://CRAN.R-project.org/package=survival

24. Team RC. R: A language and environment for statistical computing: ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria; 2013. Available: http://www. R-project. org

25. Mc Gilchrist CA, Aisbett CW. Regression with frailty in survival analysis. Biometrics 1991;47:461-66.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=748&id=22&aid=7196