Chapter 7
A Classical Approach to Modeling of Coal Mine Data

Mehmet Yılmaz, Nihan Potas, and Buse Buyum

Abstract Data sets such as the occurrence time of random events or the lifetime of a certain product (or a system) are modelled by compound or mixture distributions especially in the last years. This situation is led to encounter proposal of more complex distribution models in the literature. One of the data set made a model proposal by in this way is coal mine data set. In this study, Two Component Mixed Exponential Distribution (2MED) model had more easier interpretation on this data set is used and compared with the other study results. Also, the extended coal mine data set with 191 observations is modelled by 2MED and the results are given.

7.1 Introduction

Mining is an important source of foreign exchange for many developing countries. But this important source of foreign exchange is also one of the sectors in which it occurs the most accidents all over the world. These coal mine accidents is still under investigation by many different disciplines in today as in the past. Coal mine accidents are often used as a real data set in studies in the field of statistics as in other fields. In these studies, some researchers have made trend analysis by taking the occurrence time of coal mine accidents, some of them have tried to model of accident occurrence time. Purpose of modeling studies is obtained the model that can best forecast (or model) these processes.

M. Yılmaz (✉)
Faculty of Science, Department of Statistics, Ankara University, 06100, Ankara, Turkey
e-mail: yilmazm@science.ankara.edu.tr

N. Potas
Faculty of Economics and Administrative Sciences, Department of Health Care Management, Gazi University, 06100, Ankara, Turkey
e-mail: nihanp@gazi.edu.tr

B. Buyum
Department of Statistics, Graduate School of Natural and Applied Science, Ankara University, 06100, Ankara, Turkey
e-mail: busebuyum@gmail.com

© Springer International Publishing Switzerland 2016

Ş.Ş. Erçetin (ed.), *Chaos, Complexity and Leadership 2014*,
Springer Proceedings in Complexity, DOI 10.1007/978-3-319-18693-1_7

Because of estimate of the occurrence time of these accidents is of vital importance to prevent them addition to the measures that can be taken.

In the current study, the data set, which is one of the most widely used in the literature, is obtained firstly by Maguire et al. [8] is firstly analyzed by Cox and Lewis [2], is handled. The data set obtains the time intervals (in days) between coal mine accidents concluded death of 10 or more men. In later, this data set is arranged by Jarrett [6] and is extended to 191 observations. Some researchers who try to model the data set with 109 observations are Adamidis and Loukas [1], Kus [7], Mirhossaini and Dolati [10], and Rodriguesa et al. [11]. Some of them use non-mixture distributions such as Exponential, Gamma and Weibull and the others use mixture distributions such as Exponential-Poisson (EP), Exponential-Gamma (EG) and Exponential Conway-Maxwell Poisson (ECOMP).

In this study, Two Component Mixed Exponential Distribution (2MED) is used to model this famous data set. Aim of this study, propose 2MED as a new distribution (model) in addition to distributions used in modeling study. First of all, properties of 2MED, then parameter estimations of Maximum Likelihood (MLE) and the Least Squares (LSE) will be introduced. In here, MLE is obtained by Expectation-Maximization (EM) algorithm which is one of the numeric way. The results that is obtained from other studies and from this study will be compared with Kolmogorov Smirnov Test Statistic (KS) and it is tried to indicate that how 2MED is successful about modeling this data set. Parameter estimations, KS values and p values (p) are obtained by MATLAB.

7.2 Parameter Estimations Methods for 2MED

In this section, some basic properties of 2MED are introduced and then the methods of maximum likelihood and the least squares will be given.

7.2.1 Mixed Exponential Distribution with Two-Component (2MED)

Probability density function (p.d.f) of 2MED is given below.

$$\begin{aligned} f(x;\alpha,\theta_1,\theta_2) &= \alpha f_1(x;\theta_1) + (1-\alpha) f_2(x;\theta_2) \\ &= \alpha \frac{1}{\theta_1} \exp(-x/\theta_1) + (1-\alpha)\frac{1}{\theta_2} \exp(-x/\theta_2) \end{aligned}$$

where $\alpha \in (0,1)$, $\theta_i > 0$ $(i = 1,2), x > 0$. Similarly the cumulative distribution function (c.d.f) is as follows.

$$F(x;\alpha,\theta_1,\theta_2) = \alpha F_1(x;\theta_1) + (1-\alpha)F_2(x;\theta_2)$$
$$= \alpha(1-\exp(-x/\theta_1)) + (1-\alpha)(1-\exp(-x/\theta_2))$$

Survival function,

$$S(x;\alpha,\theta_1,\theta_2) = \alpha S_1(x;\theta_1) + (1-\alpha)S_2(x;\theta_2)$$
$$= \alpha\exp(-x/\theta_1) + (1-\alpha)\exp(-x/\theta_2)$$

and the hazard function,

$$h(x;\alpha,\theta_1,\theta_2) = \frac{\alpha h_1(x)S_1(x;\theta_1) + (1-\alpha)h_2(x)S_2(x;\theta_2)}{\alpha S_1(x;\theta_1) + (1-\alpha)S_2(x;\theta_2)}$$
$$= h_1(x)\frac{\alpha S_1(x;\theta_1)}{\alpha S_1(x;\theta_1) + (1-\alpha)S_2(x;\theta_2)}$$
$$+h_2(x)\frac{(1-\alpha)S_2(x;\theta_2)}{\alpha S_1(x;\theta_1) + (1-\alpha)S_2(x;\theta_2)}$$
$$= h_1(x)w_1(x;\alpha,\theta_1,\theta_2) + h_2(x)w_2(x;\alpha,\theta_1,\theta_2)$$

where $h_i(x) = \frac{1}{\theta_i}$ and $w_1(.) + w_2(.) = 1$.

7.2.2 Maximum Likelihood Method for 2MED

Let $\underline{X} = \{X_1, X_{2,\ldots,}X_n\}$ be a random sampling with independent and identically distributed as 2MED having a p.d.f $f(\underline{x};\underline{\Phi})$ where $\underline{\Phi} = (\alpha,\theta_1,\theta_2)$ is a parameter vector. The likelihood function and the logarithmic form of the likelihood function of $\underline{\Phi}$ are respectively given as below:

$$L(\underline{\Phi};\underline{x}) = \prod_{j=1}^{n}\left[\sum_{i=1}^{2}\alpha_i\frac{1}{\theta_i}\exp(-x_j/\theta_i)\right]$$

$$\log L = \sum_{j=1}^{n}\log\left[\sum_{i=1}^{2}\alpha_i\frac{1}{\theta_i}\exp(-x_j/\theta_i)\right] - \lambda\left(\sum_{i=1}^{2}\alpha_i - 1\right)$$

where $\sum_{i=1}^{2}\alpha_i = 1$. If the derivative of this function respect to $\alpha_i, i = 1,2$ is equalized to zero,

$$\frac{\partial \log L}{\partial \alpha_i} = \sum_{j=1}^{n} \frac{\frac{1}{\theta_i} \exp(-x_j/\theta_i)}{\sum_{i=1}^{2} \alpha_i \frac{1}{\theta_i} \exp(-x_j/\theta_i)} - \lambda = 0$$

then

$$\sum_{j=1}^{n} \frac{\frac{1}{\theta_i} \exp(-x_j/\theta_i)}{\sum_{i=1}^{2} \alpha_i \frac{1}{\theta_i} \exp(-x_j/\theta_i)} = \lambda \tag{7.1}$$

Multiplying the both side of (7.1) by α_i and taking the sum over index i:

$$\sum_{j=1}^{n} \sum_{i=1}^{2} \frac{\alpha_i \frac{1}{\theta_i} \exp(-x_j/\theta_i)}{\sum_{i=1}^{2} \alpha_i \frac{1}{\theta_i} \exp(-x_j/\theta_i)} = \lambda \alpha_i$$

then $n = \lambda$. Based on Bayes' rule, the probability that x_j belongs to ith component when $X_j = x_j$ is observed is as follows:

$$P(i \mid x_j) = \frac{\alpha_i \frac{1}{\theta_i} \exp(-x_j/\theta_i)}{\sum_{i=1}^{2} \alpha_i \frac{1}{\theta_i} \exp(-x_j/\theta_i)}$$

Thus,

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{n} P(i \mid x_j)}{n}$$

where $i = 1, 2$. If the derivative of $\log L$ with respect to θ_i is equalized to zero,

$$\frac{\partial \log L}{\partial \theta_i} = \sum_{j=1}^{n} \frac{\frac{\alpha_i}{\theta_i} \left(\frac{x_j}{\theta_i^2}\right) \exp\left(\frac{-x_j}{\theta_i}\right) - \frac{\alpha_i}{\theta_i^2} \exp\left(\frac{-x_j}{\theta_i}\right)}{\sum_{i=1}^{2} \alpha_i \frac{1}{\theta_i} \exp\left(\frac{-x_j}{\theta_i}\right)} = 0$$

$$\hat{\theta}_i = \frac{\sum_{j=1}^{n} x_j P(i \mid x_j)}{\sum_{j=1}^{n} P(i \mid x_j)}$$

where $i = 1, 2$. $\hat{\theta}_i$ is obtained and reminded that $P(2 \mid x_j) = 1 - P(1 \mid x_j)$, then the solutions will be

$$\hat{\theta}_1 = \frac{1}{n\hat{\alpha}_i} \sum_{j=1}^{n} x_j P(i \mid x_j)$$

$$\hat{\theta}_2 = \frac{1}{n(1-\hat{\alpha}_i)} \sum_{j=1}^{n} x_j (1 - P(i \mid x_j)).$$

As seen in the above, the parameter estimations can not been obtained directly from derivative equations. Therefore numeric ways are preferred for solving of these equations. In this study, EM algorithm which is one of the numeric way is taken into account [3, 4, 9]. These are step solutions obtained by EM which steps are given:

1. Input the initial values $(\alpha_i^{(0)}, \theta_i^{(0)})$, $i = 1, 2$.
2. Calculate the $P(i \mid x_j)$.
3. Calculate $\hat{\alpha}_i^{(k)}$, $\hat{\theta}_i^{(k)}$
4. After calculations of $\hat{\alpha}_i$ and $\hat{\theta}_i$, the values replace in $logL$ and get the value of function. For $\epsilon > 0$ selected small enough $\log L^{(k)} - \log L^{(k-1)} \leq \epsilon$ is provided then the values on the kth step will be used for parameter estimations. Steps 2–4 are repeated until converge is accomplished.

7.2.3 The Least Squares Method for 2MED

This method is based on the idea that there is a regression relationship between empirical $\hat{F}$ and parametric F distributions. Considering ordered observations $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ versus empirical distribution $\hat{F}(x_{(i)}) \equiv i/(n+1)$, the vector $\underline{\Phi}$ which minimizes the following expression is tried to determine. Detailed study was given in Gupta and Kundu [5] for non-mixture Generalized Exponential Distribution. System of equations that is occurred for the solutions for this optimization problem is as follows.

$$Q(\underline{\Phi}) = \sum_{i=1}^{n} \left(\hat{F}(x_{(i)}) - F(x_{(i)}; \underline{\Phi}) \right)^2$$

For solving of this optimization problem, since the expressions after derivative are related to parameters, it is difficult to obtain the solutions. Therefore it is necessary to use numerical ways. The values minimized $Q(\underline{\Phi})$ function are calculated numerically by current command in MATLAB. The stopping rule can be based on absolute value of the difference between the function values in the previous iteration and next iteration. So, when the measured absolute difference becomes less than $10^{(-21)}$ the search can be stopped.

Table 7.1 The time intervals (in days) between coal mine accidents

378	96	59	108	54	275	498	228	217	19	156
36	124	61	188	217	78	49	271	120	329	47
15	50	1	233	113	17	131	208	275	330	129
31	120	13	28	32	1,205	182	517	20	312	1,630
215	203	189	22	23	644	255	1,613	66	171	29
11	176	345	61	151	467	195	54	291	145	217
137	55	20	78	361	871	224	326	4	75	7
4	93	81	99	312	48	566	1,312	369	364	18
15	59	286	326	354	123	390	348	338	37	1,357
72	315	114	275	58	457	72	745	336	19	

Table 7.2 Parameter estimations, KS and p-values for 2MED

LSE			MLE		
$\hat{\alpha}$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\alpha}$	$\hat{\theta}_1$	$\hat{\theta}_2$
0.9162	238.5660	27.8652	0.1757	592.0210	166.1348
KS Stat.		p-value	KS Stat.		p-value
0.0594		0.8138	0.0578		0.8386

7.3 Suggested and Current Models for Coal Mine Data

The data set, which is obtained firstly by Maguire et al. [8] and obtains the time intervals (in days) between coal mine accidents concluded death of 10 or more men is given in Table 7.1. In this section, the results obtained from the studies in the literature and from the current study will be compared.

First of all, in terms of providing comparison and ease of comment the parameter estimations, KS and p-values obtained by modeling with 2MED is given in Table 7.2.

In Adamidis and Loukas [1], firstly they are suggested Weibull and Gamma which are used frequently as non-mixture distributions and then they are tried to model the data set with EG. After modeling, KS value of EG is found as **0.076** and they said that the EG distribution fits the data set at least as good as the two popular alternatives. When this value and the KS value for 2MED according to two methods, it can be said that 2MED is more successful than EG distribution about modeling the data set.

In Kus [7], EP distribution is used in addition to the distributions used in Adamidis and Loukas [1]. The KS and p-value of EP is given in.

When the KS values for 2MED and EP given in Table 7.3 are compared, it can be seen that 2MED values are smaller than EP values. Therefore 2MED is the best amongst four distributions handled so far according to KS criteria.

Table 7.3 KS and p-values for Kus [7]

Distribution	KS value	p-value
EP	0.0625	0.7876
EG	0.0761	0.5524
WEIBULL	0.0773	0.5325
GAMMA	0.0852	0.4076

Table 7.4 KS values for Mirhossaini and Dolati [10]

Distribution	KS value
EXPONENTIAL	0.0776
ME	0.0667
GAMMA	0.0796
WEIBULL	0.2965

Table 7.5 A* values for Rodriguesa et al. [11]

Distribution	A*
ECOMP	0.432
EG	0.439
EP	0.480
EXPONENTIAL	0.658

Exponential distribution model is discussed in addition to the above non-mixture distributions in Mirhossaini and Dolati [10]. Besides non-mixture distributions, ME is used in modelling study and the results is given in the Table 7.4.

Considering the proposed model in the above, it is thinkable that ME is the closest model to 2MED. Even in this thought, it is clear that the KS values for 2MED is lower than the KS values for ME. The results is same for the other three distributions. The comment made on the results of other studies is also applied here.

The ECOMP distribution is used as well as commonly used for modeling this data set in Rodriguesa et al. [11]. In their study, the modified Cramer-von Mises (W*) and Anderson-Darling (A*) test statistics are taken into account but it is decided that which distribution is more successful according to the value. Therefore the A* value is calculated for 2MED while comparing with distributions in Rodriguesa et al. [11]. Computational code is taken from the first author Josemar Rodriguesa. The A* value is given in the Table 7.5.

The A* values calculated according to MLE and LSE methods for 2MED are found **0,426** and **0,722** respectively. A* value found according to MLE seems to be smaller than the value calculated for distribution in the above table. Accordingly, 2MED is more suitable for this data set.

Coal mine data set is arranged and is extended to 191 observations by Jarrett [6]. However any modelling study for 191 observations has not reached in the literature search. Here we try to show how this extended data set (given in Table 7.6) is modelled by 2MED in addition to the studies above.

When the results given in Table 7.7 are examined, modelling of this extended data set with 2MED for two methods is also successful according to KS criteria.

Table 7.6 The extended coal mine data set

157	123	2	124	12	4	10	216	25	19	33
66	232	826	40	12	29	190	53	17	186	23
92	197	431	16	154	95	250	80	78	202	36
110	276	16	88	225	24	91	538	187	34	101
41	139	42	1	112	43	3	324	56	31	96
70	41	93	2	0	143	16	27	144	45	6
208	29	15	72	193	134	420	95	125	34	127
218	59	315	378	36	15	31	215	11	137	4
286	114	96	124	50	120	203	176	55	93	326
275	59	61	1	13	189	345	20	81	354	307
108	188	233	28	22	61	78	99	123	456	54
217	113	32	388	151	361	312	462	228	275	78
17	1,205	644	467	871	48	217	120	498	49	131
182	255	194	224	566	19	329	806	517	1,643	54
326	1,312	348	745	156	47	275	20	66	292	4
368	307	336	952	632	330	312	536	145	75	364
129	1,630	29	217	7	18	1,358	2,366	65	17	19
37	19	12								

Table 7.7 Parameter estimations, KS and p-values for 2MED of extended data set

LSE			MLE		
$\hat{\alpha}$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\alpha}$	$\hat{\theta}_1$	$\hat{\theta}_2$
0.1940	39.9254	224.5533	0.8210	134.7353	574.3638
KS Stat.		p-value	KS Stat.		p-value
0.0304		0.9926	0.0443		0.8334

7.4 Conclusion

In this study is handled seven different distributions used modelling of the time intervals (in days) between coal mine accidents. It can be seen that the data set is modelled mostly by Weibull, Gamma and Exponential as non-mixture distributions and by EG, EP as mixture distributions. Except for these distributions, ME and ECOMP are used in modelling study.

The results of KS statistic and the A* test statistic are used as a measure to compare three of these modelling studies. When the results obtained from both LSE and MLE methods for 2MED and the results of these distributions are compared, 2MED seems to be the best model according to KS and A* criteria (except for LSE). 2MED is the best model between mixture and non-mixture distributions used in Adamidis and Loukas [1], Kus [7], Mirhossaini and Dolati [10], and Rodriguesa et al. [11]. In addition to comparison, the extended data set is also analyzed. But

studies for this data set is generally based on analyzing as a stochastic process. Therefore this extended data set is modelled only by 2MED and the results is given Table 7.7.

Finally, we can say that as an uncomplicated model 2MED can be recommended for the coal mine data set which is studied by many researches.

References

1. Adamidis K, Loukas S (1998) A lifetime distribution with decreasing failure rate. Stat Probab Lett 39:35–42
2. Cox DR, Lewis PAW (1966) The statistical analysis of series of events. Chapman and Hall, London/Methuen
3. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B 39:1–38
4. Everitt ES, Hand DJ (1981) Finite mixture distributions. Chapman and Hall, London
5. Gupta RD, Kundu D (2000) Generalized exponential distribution: different method of estimations. J Stat Comput Simul 00:1–22
6. Jarrett RG (1979) A note on the intervals between coal mining disasters. Biometrika 66:191–193
7. Kus C (2007) A new lifetime distribution. Comput Stat Data Anal 51:4497–4509
8. Maguire BA, Pearson ES, Wynn AHA (1952) The time intervals between industrial accidents. Biometrika 39:168–180
9. McLachlan GJ, Krishnan T (2008) The EM algorithm and extensions. Wiley, Hoboken
10. Mirhossaini SM, Dolati A (2008) On a new generalization of the exponential distribution. J Math Ext 3(1):27–42
11. Rodriguesa J, Balakrishnan N, Cordeiro GM (2011) A unified view on lifetime distributions. Comput Stat Data Anal 55(12):3311–3319