The Bivariate Erlang and its Application in Modeling Recurrence Times of Kidney Dialysis Data

Norou Diawara[1,*], S.H. Sathish Indika[2], Melva Grant[1] and Edgard M. Maboudou-Tchao[3]

[1]Old Dominion University, 4700 Elkhorn Ave., Norfolk, VA 23529, USA

[2]Thomas Nelson Community College, 99 Thomas Nelson Drive, Hampton, VA 23666, USA

[3]University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816, USA

Abstract: Recent advances in computer modeling allows us to find closer fits to data. Our emphasis is on the interdependence between occurrence at kidney dialysis. The interdependence between kidney dialysis occurrences is modelled by a bivariate exponential that we propose in this article. The application is shown on the McGilchrist and Aisbett kidney data set with the use of the exponential distribution. The proposed bivariate exponential model has exponential marginal densities, correlated *via* a latent random variables and with finite probability of simultaneous occurrence. Extension of the model to a bivariate Erlang type distribution with same shape parameter is presented.

Keywords: Bivariate models, Erlang, exponential, Dirac delta.

1. INTRODUCTION

The exponential family of distributions is very important for modeling phenomena in life testing, reliability, and other types of medical applications. Exponential marginal densities have been known in the literature for some time. However, the majority of these models have been theoretically motivated rather than application oriented. A notable exception was the "fatal shock model" of Marshall and Olkin [1].

Fitting probability distributions to disease data is an essential part in its cure and in gaining control of such prevalent disease [2]. Patients with kidney and renal failures undergo dialysis processes. Knowing as much as possible on the aspects of dialysis help the patients prepare to control kidney disease, prevent possible failures of the dialysis and manage cost-effectiveness of initiating dialysis early [3]. Recent advances in computer modeling allows us to find closer fits. Emphasis is put on the interdependence between occurrence at kidney dialysis. Modeling the time to dialysis outcomes falls in the class of survival analysis.

The bivariate exponential model is proposed with an underlying linear relationship that enforces zero or positive correlation, including a finite probability of simultaneous occurrence. By including the nature of the underlying functional relationship of the data, structural mathematical models are capable of providing explanations for observed and projected changes, where correlation models only give limited explanatory capability. We propose an estimation approach and illustrate it with real data analysis.

The proposed bivariate exponential model extends ideas from [1, 4-5]. We investigate the bivariate Erlang type with a non-zero probability of simultaneous occurrence of the phenomena under investigation. More specifically, let X_1 and X_2 be Erlang random variables (rv's) with hazard rates λ_1 and λ_2 respectively. These two rv's are related linearly through a non-negative latent rv, Z. The authors in [6-7] showed that the distribution of Z can be completely and uniquely characterized as the product of two independent rv's, one being a Bernoulli rv with the parameter $a\lambda_2 / \lambda_1$, and the other being an exponential with parameter λ_2, when X_1 and X_2 are exponential with hazard rates λ_1 and λ_2, respectively.

The paper is organized as follows. Section 2 shows the main theoretical result. In Section 3, applications based on data from [8] are provided along with estimates and variance-covariance of the associated parameters.

2. BIVARIATE ERLANG DISTRIBUTIONS

Consider the univariate two parameter Erlang distribution, X. Its probability density function (pdf) is defined as:

$$f_X(x;\lambda,\alpha) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I_{[0,\infty)}(x), \quad (1)$$

where $\alpha \in \mathbf{N}$ is the shape parameter and $\lambda > 0$ is the scale parameter.

*Address correspondence to this author at the Old Dominion University, 4700 Elkhorn Ave., Norfolk, VA 23529, USA; Tel: (757) 683.3886;
Fax: (757) 683.3885; E-mail: ndiawara@odu.edu

E-ISSN: 1929-6029/14

The Laplace Stieltjes transform (LST) (the equivalent concept of moment generating function) provides a great deal of insight about the nature of a distribution. It is defined in [9] as:

$$L_X(s) = Ee^{-sX} = \int_0^\infty e^{-sx} dF_X(x) = (\frac{\lambda}{\lambda + s})^\alpha.$$

Let X_1 and X_2 be two rv's. We define:

$$X_2 = aX_1 + Z, \tag{2}$$

where a, a nonnegative fixed constant called the coefficient of linear relationship, and Z, an unknown random variable independent of X_1. The unknown rv Z is also called frailty in the survival model. It extends the ideas of the proportional Cox hazard model and allows positive correlation among survival times. The choice of Z as a gamma frailty is widely used. Hougaard in [10] discusses alternate choices of frailty distributions. We apply an alternate method based on the distributional forms of the covariates X_1 and X_2.

The coefficient a in Equation (2) is estimated by the ratio of the occurrences of the events. This model gives independent coordinates when $a = 0,$ and does not permit negative association as described in Iyer *et al*. (2002). The characterization described in which the marginal distributions are exponential was introduced in [1], and has been studied by authors such as in [11-12]. When X_1 and X_2 are exponential rv's with parameters λ_1 and λ_2, respectively, the rv Z is a product of a Bernoulli rv with parameter p and an exponential rv with parameter λ_2. Its pdf is given by:

$$f(z) = p\delta(z) + (1-p) f_{X_2}(z) I(z > 0), \tag{3}$$

where

- $p = P(X_2 = aX_1) = P(Z = 0) = a\lambda_2 / \lambda_1.$
- $\delta(t)$ refers to the Dirac delta function, i.e $\delta(t) = 0$,if $t \neq 0$, and $\int_{-\infty}^{+\infty} \delta(t) dt = 1$.
- and $f_{X_2}(t) = \lambda_2 e^{-\lambda_2 t}, t > 0.$

The subsequent output shows the result of the joint distribution when X_1 and X_2 are Erlang with the same shape parameter. First let's recall a related result.

Theorem 1. Let $Z_1, \ldots, Z_n$ be independent and identically distributed (iid) rv's with the pdf as in (3). Define $S_n = Z_1 + \ldots + Z_n$. Then the distribution of S_n can be written as a mixture of gamma and Dirac delta distributions with Binomial weights, i.e.

$$f_{S_n}(z) = \sum_{j=0}^{n} \binom{n}{n-j} p^{n-j} (1-p)^j f_{g_j}(z), \quad z \geq 0, \tag{4}$$

where $f_{g_0}(t) = \delta(t),$ and $f_{g_j}(t) = \frac{{\lambda_2}^j}{\Gamma(j)} t^{j-1} e^{-\lambda_2 t}, t > 0,$ for $1 \leq j \leq n.$

See [13].

We now present the joint distribution function of two Erlang distributions with same shape parameter.

Theorem 2. Let f_1 and f_2 represent the marginal Erlang densities of two random variables X_1 and X_2. More specifically, let $X_1 : Erlang(\alpha, \lambda_1)$ and $X_2 : Erlang(\alpha, \lambda_2)$. Then the joint probability density function of (X_1, X_2) is given by:

$$g(x_1, x_2) = \sum_{j=0}^{\alpha} \binom{\alpha}{\alpha - j} p^{\alpha - j} (1-p)^j f_1(x_1) f_{g_j}(x_2 - ax_1), \tag{5}$$

where

- the random variables X_1 and X_2 are related as in (2).
- $p = P(X_2 = aX_1) = a\lambda_2 / \lambda_1.$
- $f_{g_j}(t) = ({\lambda_2}^j / \Gamma(j)) t^{j-1} e^{-\lambda_2 t}, t > 0,$ for $1 \leq j \leq n.$
- $f_{g_0}(t) = \delta(t)$ refers to the Dirac delta i.e $\delta(t) = 0$,if $t \neq 0$, and $\int_{-\infty}^{+\infty} \delta(t) dt = 1.$

Proof. : The LST of Z is:

$$L_Z(s) = [\,(1-p)\frac{\lambda_2}{\lambda_2 + s} + p\,]^\alpha.$$

From the LST, Z is sum of α iid rv's that are the product of independent Bernoulli and exponential rv's.

The pdf of Z is a mixture of two types of functions: the Dirac Delta and gamma type pdf. The probability density function of Z is:

$$f_Z(z)=\sum_{j=0}^{\alpha}\binom{\alpha}{\alpha-j}p^{\alpha-j}(1-p)^j f_{g_j}(z),\ z\geq 0,\ by\ Theorem\ 2.1.$$

Using the independence of X_1 and Z, we have:

$$f_{X_1,Z}(x_1,z)=f_1(x_1)f_Z(z)$$

$$=p^{\alpha}f_1(x_1)\delta(z)+\sum_{j=1}^{\alpha}\binom{\alpha}{\alpha-j}p^{\alpha-j}(1-p)^j f_{g_j}(z)f_1(x_1).$$

$$Then\quad g(x_1,x_2)=\int_{-\infty}^{+\infty}f_{X_1,Z}(x_1,z)\delta(ax_1+z-x_2)dz$$

$$=\int_{-\infty}^{+\infty}p^{\alpha}f_1(x_1)\delta(z)\delta(ax_1+z-x_2)dz+$$

$$\int_{-\infty}^{+\infty}\sum_{j=1}^{\alpha}\binom{\alpha}{\alpha-j}p^{\alpha-j}(1-p)^j f_1(x_1)f_{g_k}(z)\delta(ax_1+z-x_2)dz$$

$$=Part1+Part2,$$

with

$$Part1=\int_{-\infty}^{+\infty}p^{\alpha}f_1(x_1)\delta(z)\delta(ax_1+z-x_2)dz$$

$$=p^{\alpha}f_1(x_1)\int_{-\infty}^{+\infty}\delta(z)\delta(ax_1+z-x_2)dz$$

$$=p^{\alpha}f_1(x_1)\delta(x_2-ax_1),\ and$$

$$Part2=\int_{-\infty}^{+\infty}\sum_{j=1}^{\alpha}\binom{\alpha}{\alpha-j}p^{\alpha-j}(1-p)^j f_1(x_1)f_{g_k}(z)\delta(ax_1+z-x_2)dz$$

$$=\sum_{j=1}^{\alpha}\binom{\alpha}{\alpha-j}p^{\alpha-j}(1-p)^j f_1(x_1)\int_{-\infty}^{+\infty}f_{g_k}(z)\delta(ax_1+z-x_2)dz$$

$$=\sum_{j=1}^{\alpha}\binom{\alpha}{\alpha-j}p^{\alpha-j}(1-p)^j f_1(x_1)f_{g_k}(x_2-ax_1).$$

Putting together Part1 and Part 2, we obtain,

$$g(x_1,x_2)=\sum_{j=0}^{\alpha}\binom{\alpha}{\alpha-j}p^{\alpha-j}(1-p)^j f_1(x_1)f_{g_j}(x_2-ax_1).$$

The following result can then be deduced.

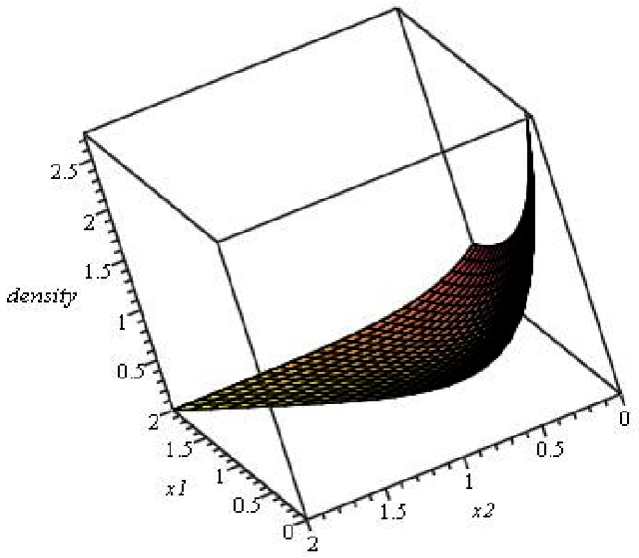

Figure 1: Graph of the joint pdf when $\lambda_1=4$, $a=1$, and $\lambda_2=1$.

As a result, when f_1 and f_2 represent exponential densities of two random variables X_1 and X_2 with parameters λ_1 and λ_2, respectively, then the joint probability density function of (X_1,X_2) is given by:

$$g(x_1,x_2)=p\lambda_1 e^{-\lambda_1 x_1}\delta(x_2-ax_1)+ (1-p)\lambda_1\lambda_2 e^{-\lambda_2 x_2}e^{-(\lambda_1-a\lambda_2)x_1}I(x_2>ax_1), \quad (6)$$

where the random variables X_1 and X_2 are related as in (2).

Figure **1** describes the joint distribution of (X_1,X_2) for $\lambda_1=4$, $a=1$, and $\lambda_2=1$.

The reliability of our approach is tested in the next section using McGilchrist and Aisbett (1991) kidney data [8].

3. APPLICATION AND ESTIMATION

The analysis of the approach was carried out for the kidney data [8]. The maximum likelihood for the correlated data and the variance for the parameters of the bivariate model were calculated. The results are presented next. Let $(x_{1i},x_{2i}),i=1,\ldots,n,$ represent a random sample from (6). The joint maximum likelihood estimator of (λ_1,λ_2) is given by:

$$\hat{\lambda}_1 = \frac{a}{\bar{x}_2} + \frac{n-k}{n\bar{x}_1} \quad and \quad \hat{\lambda}_2 = \frac{1}{\bar{x}_2}, \tag{7}$$

where

- $k = \sum_{i=1}^{n} I(x_{2i} - ax_{1i} = 0)$, represents the number of times of proportional occurrence between X_1 and X_2.
- $\bar{x}_1 = \sum_{i=1}^{n} x_{1i} / n$, and $\bar{x}_2 = \sum_{i=1}^{n} x_{2i} / n$.

The variance covariance matrix of $\hat{\lambda}_1$, and $\hat{\lambda}_2$, is:

$$\Sigma = \frac{1}{n}\begin{pmatrix} \lambda_1(\lambda_1 - a\lambda_2) + a^2\lambda_2^2 & a\lambda_2^2 \\ a\lambda_2^2 & \lambda_2^2 \end{pmatrix}. \tag{8}$$

When the two rv's are independent, then it is reasonable to assume that $a = k = 0$, and then the estimators will represent those proposed in [14].

Simulated example: The above mentioned estimators were developped under the Exponential-Exponential case. A simulation study of the same setting is assessed. The data were generated from a bivariate exponential with $\lambda_1 = 4, \lambda_2 = 1$ *and* $a = 1$ from sample of size $n = 25$. Using the simulated data the parameters are estimated. The estimated parameters and the estimated variance-covariance matrix are given as:

$$\hat{\lambda} = \begin{pmatrix} 4.162 \\ 1.041 \end{pmatrix} \quad and \quad \hat{\Sigma} = \begin{pmatrix} 0.599 & 0.047 \\ 0.047 & 0.0472 \end{pmatrix}.$$

As we can see, there is a close correspondence with its original parameters. Now a real data set is considered.

Real example: Here we consider the complete data from [8]. The data set describes the recurrence times to infection at point of insertion of the catheter for kidney patients who are using portable dialysis equipment. The sample data correlation between the recurrence times is found to be 0.794 with a significant p-value smaller than 0.001. To fit the model, we assume that the underlying functional relation between the recurrence times is linear. The interdependence is captured by the joint distribution of the suggested model and fitted using R. The specifications for the coefficient of linear relationship is chosen to be 1 and data censored is taken into account. The joint MLE's, and the estimated asymptotic variance/covarince matrix of the MLE's(from (7) and (8)) are as follows:

$$\hat{\lambda} = \begin{pmatrix} 0.0138651 \\ 0.0049134 \end{pmatrix} and \quad \hat{\Sigma} = \begin{pmatrix} 0.000003902 & 0.000000635 \\ 0.000000635 & 0.000000635 \end{pmatrix}.$$

The log-likelihood of the data is given in Figure **2**.

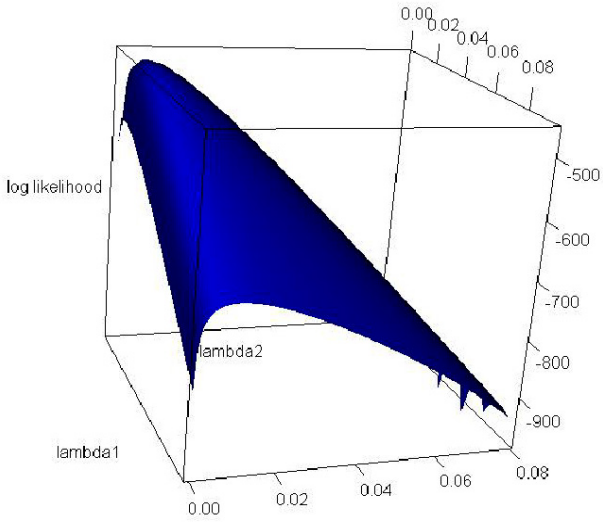

Figure 2: The log-likelihood.

Typically, an analysis of disease types can lead to debate of the differences between associated mortalities, estimates of disease risks and variations. Clustering may lead to significantly less stability. The inherent correlation between the dependent variables can be further studied with regard to this model. We believe that a correlation type test can be developed to test whether interdependency is statistically significant for this data set. By repeating the analysis to the different types of kidney diseases, it turns out that a little more light can be shed about this disease. The data was further aggregated into disease types. To address such issues related to the kidney data in [8], our model that allows estimates of risk parameters associated with the four types of diseases has been fitted. The four disease types are $0 =$ GN, $1 =$ AN, $2 =$ PKD, and $3 =$ other. Using our construction, each disease type model was fitted separately. The results are presented below. We observed conjugate property that shows differences between those diseases.

For disease type 0, the estimates and the variance-covariance are respectively:

$$\hat{\lambda} = \begin{pmatrix} 0.0142799 \\ 0.0055590 \end{pmatrix} and \quad \hat{\Sigma} = \begin{pmatrix} 0.000017271 & 0.000003434 \\ 0.000003434 & 0.000003434 \end{pmatrix}.$$

For disease type 1, the estimates and the variance-covariance are respectively:

$$\hat{\lambda} = \begin{pmatrix} 0.0283395 \\ 0.0071006 \end{pmatrix} and \quad \hat{\Sigma} = \begin{pmatrix} 0.000054360 & 0.000004202 \\ 0.000004202 & 0.000004202 \end{pmatrix}.$$

For disease type 2, the estimates and the variance-covariance are respectively:

$$\hat{\lambda} = \begin{pmatrix} 0.0145298 \\ 0.0038059 \end{pmatrix} and \quad \hat{\Sigma} = \begin{pmatrix} 0.000042575 & 0.000003621 \\ 0.000003621 & 0.000003621 \end{pmatrix}.$$

For disease type 3, the estimates and the variance-covariance are respectively:

$$\hat{\lambda} = \begin{pmatrix} 0.0095673 \\ 0.0038530 \end{pmatrix} and \quad \hat{\Sigma} = \begin{pmatrix} 0.000005347 & 0.000001142 \\ 0.000001142 & 0.000001142 \end{pmatrix}.$$

The log-likelihoods for the different disease types are given in Figure **3**. Such figures along with the parameter estimates show that going from disease type 0 to 3, the likelihood becomes flattened and the estimates match that difference except for disease type 0. There appears to be substantial differences across the disease types. The estimated variance from our suggested model is smaller than the one proposed in [8]. Without use of prior distribution, the relationship between recurrence time to infection at point of insertion shows that there are substantial differences found and since maximum likelihood estimation was used variance is stable. Having such results can lead to the development of benchmarks where the different kinds of kidney diseases are compared, and one could discriminate between them.

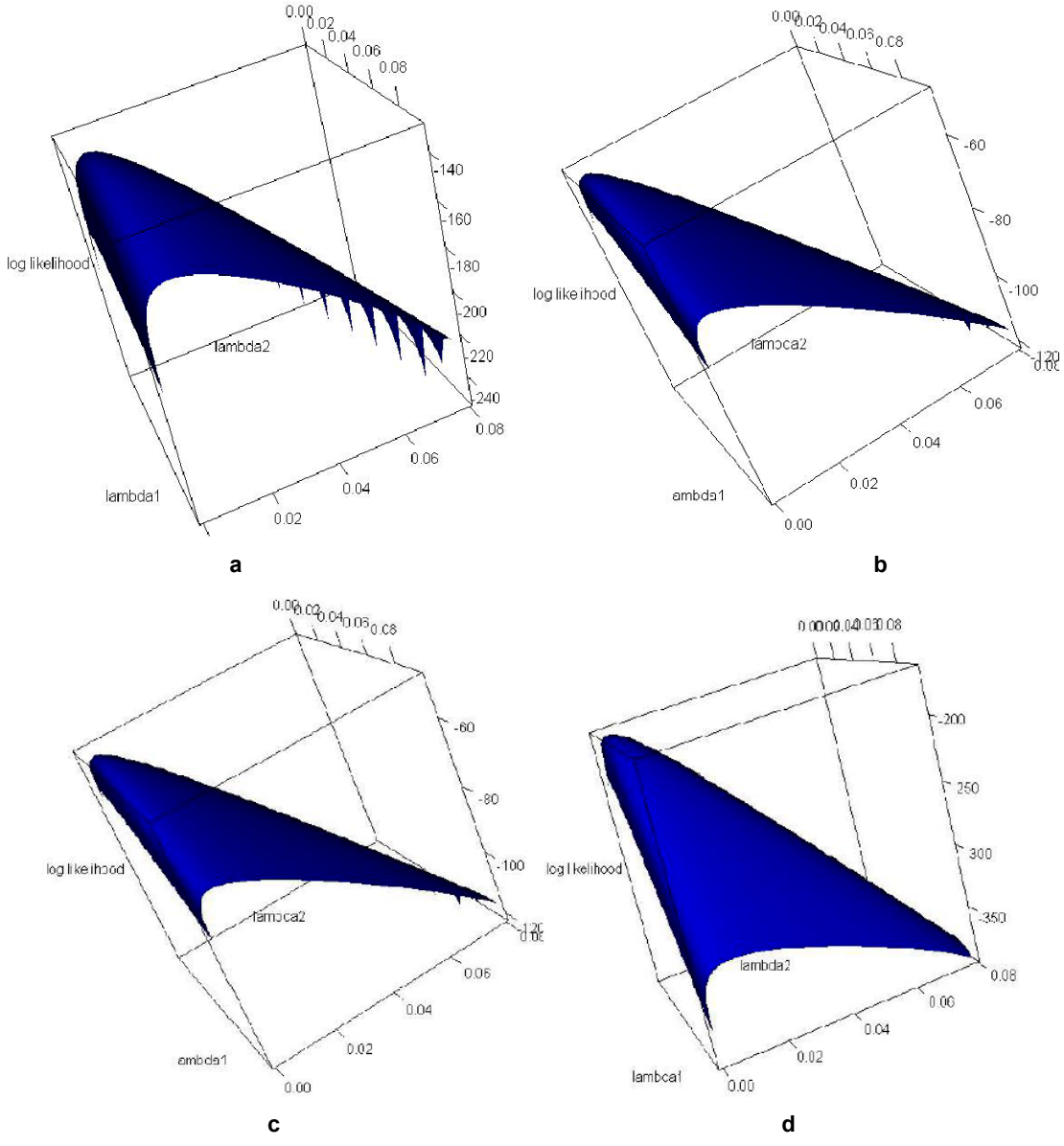

Figure 3: Graphs of the log-likelihood functions.

a) The log-likelihood for the disease type 0. **b**) The log-likelihood for the disease type 1. **c**) The log-likelihood for the disease type 2. **d**) The log-likelihood for the disease type 3.

4. CONCLUSION

We have proposed a method for the linearly related type events that have simultaneous or proportional occurrence. Such a procedure was originally proposed by Marshall and Olkin (1964) for tracking systems that can evolve simultaneously in a linear relation. To demonstrate the usefulness of the method for occurrence of events using a non-zero probability of simultaneous occurrence, we have applied it to McGilchrist and Aisbett (1991) epidemiology case and found results that were not apparent in previously proposed models.

Our results support the use of linear relationship in describing related events, due to its relative simplicity and comparative ease of implementation.

REFERENCES

[1] Marshall AW, Olkin I. A Multivariate Exponential Distribution. J Am Stat Assoc 1967; 63: 30-44. http://dx.doi.org/10.1080/01621459.1967.10482885

[2] Coresh J, Selvin E, Stevens LA, *et al*. Prevalence of Chronic Kidney Disease in the United States. J Am Med Assoc 2007; 298(17): 2038-47. http://dx.doi.org/10.1001/jama.298.17.2038

[3] Harris A, Cooper BA, Jing Li J, *et al*. Cost-Effectiveness of Initiating Dialysis Early: A Randomized Controlled Trial. Am J Kidney Diseases 2011; 57(5): 707-15. http://dx.doi.org/10.1053/j.ajkd.2010.12.018

[4] Csorgo S, Welsh AH. Testing for Exponential and Marshall-Olkin Distributions. J Stat Plan Infer 1989; 23; 287-300. http://dx.doi.org/10.1016/0378-3758(89)90073-6

[5] Zhao P, Balakrishnan N. Ordering Properties of Convolutions of heterogeneous Erlang and Pascal Random Variables. Stat Probability Lett 2010; 80: 969-74. http://dx.doi.org/10.1016/j.spl.2010.02.010

[6] Iyer SK, Manjunath D. Correlated Bivariate Sequence for queueing and reliability applications. Commun Stat 2004; 33: 331-50. http://dx.doi.org/10.1081/STA-120028377

[7] Iyer SK, Manjunath D, Manivasakan R. Bivariate Exponential Distributions Using Linear Structures. Sankhya 2002; 64(A): 156-66.

[8] McGilchrist CA, Aisbett CW. Regression with Frality in Survival Analysis. Biometrics 1991; 47; 461-66. http://dx.doi.org/10.2307/2532138

[9] Abramowitz M, Stegun IA. Handbook of Mathematical Functions, Selected Government Publications, Chapter 6, New York, Dover 1972.

[10] Hougaard P. A class of multivariate failure time distributions. Biometrika 1986; 73(3): 671-78. http://www.dx.doi.org/10.1093/biomet/73.3.671

[11] Pickands J. Multivariate Extreme Value Distributions. Bull Int Stat Instit 1981; 49: 859-78.

[12] Johnson RA, Evans JW, Green DW. Some Bivariate Distributions for Modeling the strength properties of Lumber 1999; US Department of Agriculture, Forest Service, Forest Products laboratory, Res. Pap. pp. 11.

[13] Diawara N, Indika SHS, Maboudou-Tchao EM. The Distribution of the Sum of Independent Product of Bernoulli and Exponential. Am J Math Manag Sci 2013; 32(1): 75-89.

[14] Carpenter M, Diawara N, Han Y. A New Class Of Bivariate Survival and Reliability Models. Am J Math Manag Sci 2006; 26: 163-84.

Received on 24-02-2014 Accepted on 25-04-2014 Published on 30-04-2014

http://dx.doi.org/10.6000/1929-6029.2014.03.02.2