compusoft
An International Journal of Advanced Computer Technology

ISSN:2320-0790

A Study of Statistical Properties & Model Validation for Exponential Extension Model

Dr. Ashwini Kumar Srivastava

Department of Computer Application,
Shivharsh.Kisan P.G. College, Basti, U.P., India
ashwini.skpg@gmail.com

Abstract: In this paper, we study the statistical properties of Exponential Extension Model and then we also check the validity of proposed model for different real data sets through different techniques. We are using two main techniques which are easy to understand and implement, and are based on intuitive and graphical techniques such as Q-Q-plot test, Kolmogorov–Smirnov (K-S) test which plot the graph of empirical distribution function and fitted distribution function. These plots are used to investigate whether an assumed model adequately fits a set of data and we present power comparison between p-values of these data sets obtaining by K-S test for model validation to obtain feasible real data sets which are most suitable for parameter estimation of exponential extension model.

Keywords: Exponential Extension model, probability density function (pdf'), cumulative distribution function (cdf), model validation, quantile-quantile(Q-Q) test, goodness of fit test

I. INTRODUCTION

Exponential models play a central role in analyses of lifetime or survival data, in part because of their convenient statistical theory, their important 'lack of memory' property and their constant hazard rates. In circumstances where the one-parameter family of exponential distributions is not sufficiently broad, a number of wider families such as the gamma, Weibull and lognormal models are in common use[1]. Adding parameters to a well-established family of models is a time honored device for obtaining more flexible new families of models.

In recent times, Haghighi and Sadeghi[2], Nadarajah and Haghighi[3] introduced Exponential Extension model by adding a parameter to exponential model. The two parameter Exponential Extension model represent the shape and scale parameter. It is observed that this family always has a decreasing probability function like an exponential model but it allows for increasing, decreasing and constant hazard rates like a Weibull model or an Exponentiated Exponential model [4, 5 and 6]. The Exponential Extension model has an explicit expression of reliability function and failure rate hazard function.

II. THE STATISTICAL PROPERTIES OF EXPONENTIAL EXTENSION MODEL

The two-parameter Exponential Extension model has one shape and one scale parameter [7]. The random variable *x* follows exponential extension model with the shape and scale parameters as $\alpha > 0$ and $\lambda > 0$ respectively, if it has the following cumulative distribution function (cdf),

$$F(x;\alpha,\lambda) = \ 1-\exp\left\{1-\left(\left(1+\lambda x\right)^{\alpha}\right)\right\} \ ; \qquad (2.1)$$
$$\text{where, } x \ge 0, \alpha > 0, \lambda > 0.$$

The probability density function (pdf) can be written as

$$f(x;\alpha,\lambda) = \alpha\lambda\left(1+\lambda x\right)^{\alpha-1}\exp\left\{1-\left(\left(1+\lambda x\right)^{\alpha}\right)\right\} \ ; \qquad (2.2)$$
$$\text{where, } \ x \ge 0, \alpha > 0, \lambda > 0.$$

and it will be denoted by X~EE(α, λ). The R functions *dexpo.ext*( ) and *pexpo.ext*( ) given in [8] can be used for the computation of pdf and cdf, reapectively. Some of the typical EE density functions for different values of α and for λ = 1 are depicted in Figure 1.

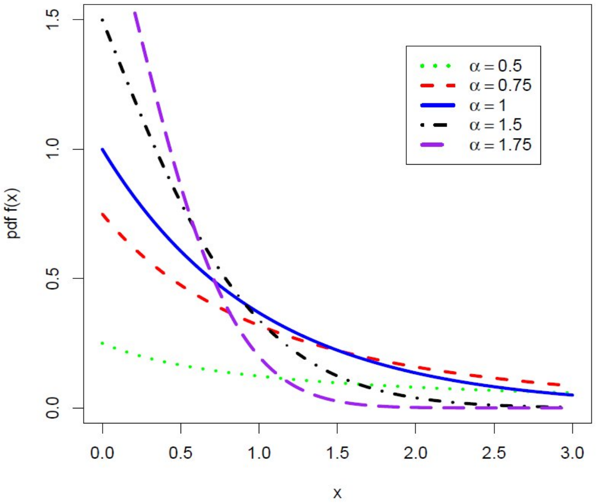


Fig 1. The PDF of EE model for λ =1 and different values of α.

The reliability/survival function is

$$R(x;\alpha,\lambda) = \exp\left\{1-\left((1+\lambda x)^{\alpha}\right)\right\} ; \quad \text{where,} \quad x \ge 0, \alpha > 0, \lambda > 0. \tag{2.3}$$

The associated R function *sexpo.ext( )* given in [8], computes the reliability function.

The hazard function is

$$h(x;\alpha,\lambda) = \alpha\lambda(1+\lambda x)^{\alpha-1} ; \quad \text{where,} \quad x \ge 0, \alpha > 0, \lambda > 0. \tag{2.4}$$

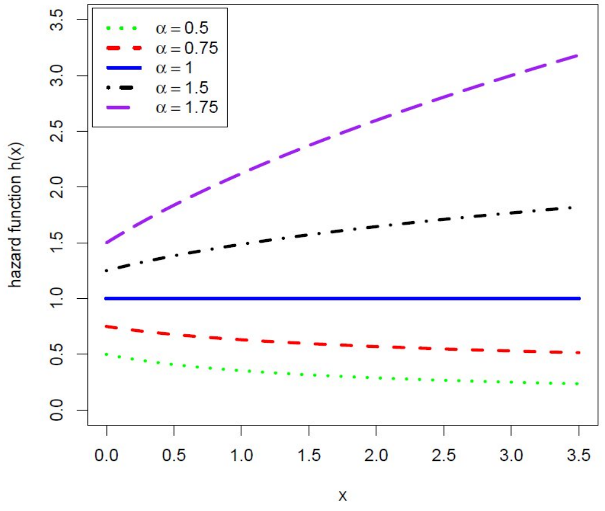


Fig 2. The Hazard function of EE model for λ =1 and different values of α.

The hazard rate function in equation (2.4) exhibits the following shapes:

1. if α < 1 then h(x) is monotonically decreasing with h(0)=α.λ and h(x) →0 as x→∞.
2. if α > 1 then h(x) is monotonically increasing with h(0)=α.λ and h(x) →0 as x→∞.
3. if α = 1 then h(x) =α.λ., ∀ x.

Some of the typical Exponential Extension model hazard functions for different values of α and for λ = 1 are depicted in Figure 2. The associated R function *hexpo.ext( )* given in [8].

The quantile function is

$$x_q = \frac{1}{\lambda}\left\{\left\{1-\log(1-q)\right\}^{1/\alpha}-1\right\} ; \quad 0<q<1. \tag{2.5}$$

The computation of quantiles, the R function *qexpo.ext*( ), given in [8] .

The median is

$$\text{Median}(x) = \frac{1}{\lambda}\left\{\left\{1-\log(0.5)\right\}^{1/\alpha}-1\right\} \tag{2.6}$$

Let U be the uniform (0,1) random variable and F(.) a cdf for which F^{-1}(.) exists. Then F^{-1}(u) is a draw from distribution F(.) .

Therefore, the random deviate can be generated from EE(α,λ) by

$$x = \frac{1}{\lambda}\left\{\left\{1-\log(1-u)\right\}^{1/\alpha}-1\right\} ; 0<u<1. \tag{2.7}$$

where u has the uniform distribution i.e. U(0, 1) distribution. The R function *rexpo.ext*( ), given in [8] , generates the random deviate from EE(α,λ).

III. Computation of Maximum Likelihood Estimation

To obtain maximum likelihood estimators of the parameters (α, λ), let $x_1, \ldots, x_n$ be the observation of a sample from a distribution with cumulative distribution function (2.1), and let $x_{(1)}, \ldots, x_{(n)}$ be the corresponding order statistics. The likelihood function of the parameter L(α, λ) based on the first k- order statistics is given by

$$L(\alpha, \lambda) = n \log \alpha + n \log \lambda + (\alpha - 1) \sum_{i=1}^{n} \log\left(1 + \lambda\, x_i\right) + n - \sum_{i=1}^{n} \left(1 + \lambda\, x_i\right)^{\alpha} \quad (3.1)$$

Therefore, to obtain the MLE's of α and λ [9], we can maximize (3.1) directly with respect to α and λ or we can solve the following two non-linear equations using Newton-Raphson method. We have,

$$\frac{\partial \log L}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} \log\left(1 + \lambda\, x_i\right) + \left\{1 - \left(1 + \lambda\, x_i\right)^{\alpha}\right\} \quad (3.2)$$

and,

$$\frac{\partial \log L}{\partial \lambda} = \frac{n}{\lambda} + \left(\alpha - 1\right) \sum_{i=1}^{n} \frac{x_i}{\left(1 + \lambda\, x_i\right)} - \alpha \sum_{i=1}^{n} x_i \left(1 + \lambda\, x_i\right)^{\alpha - 1} \quad (3.3)$$

IV. Data Analysis

In this section we present five real data sets for illustration of the proposed methodology. These are

Data Set 1: The following data set includes the time intervals (in days) of the successive earthquakes in the last century in Iran and this data are taken by International Institute of Earthquake Engineering and Seismology. [2].

284, 246, 139, 2280, 95, 308, 355, 607, 11, 563, 553

Data Set 2: The following data represent the number of million revolution before failing for each of the 23 ball bearings in the life test [9].

17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40

Data Set 3: Aarset MV. *How to identify bathtub hazard rate*. IEEE Trans Reliability 1987;R-36(1):106 -108. (Failure time of 50 items)[10].

0.1, 0.2, 1.0, 1.0, 1.0, 1.0, 1.0, 2.0, 3.0, 6.0, 7.0, 11.0, 12.0, 18.0, 18.0, 18.0, 18.0, 18.0, 21.0, 32.0, 36.0, 40.0, 45.0, 45.0, 47.0, 50.0, 55.0, 60.0, 63.0, 63.0, 67.0, 67.0, 67.0, 67.0, 72.0, 75.0, 79.0, 82.0, 82.0, 83.0, 84.0, 84.0, 84.0, 85.0, 85.0, 85.0, 85.0, 85.0, 86.0, 86.0

Data Set 4: The data represent 46 repair times (in hours) for an airborne communication transceiver Chhikara and Folks [11]. The data are as follows:

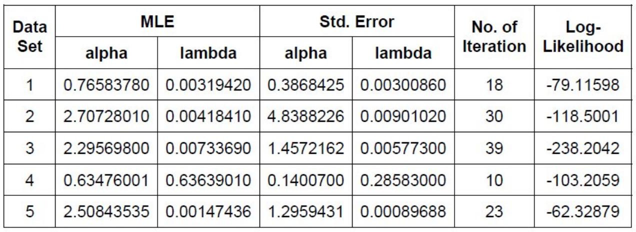
0.2, 0.3, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0, 1.1, 1.3, 1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.2, 2.5, 2.7, 3.0, 3.0, 3.3, 3.3, 4.0, 4.0, 4.5, 4.7, 5.0, 5.4, 5.4, 7.0, 7.5, 8.8, 9.0, 10.3, 22.0, 24.5

Data Set 5: This data set is from McCool (1974) giving the fatigue life in hours of ten bearing of a certain type[12]. These data are as follows:

152.7, 172.0, 172.5, 173.3, 193.0, 204.7, 216.5, 234.9, 262.6, 422.6

A. *Maximun Likelihood (ML) Estimation*

For obtaining the MLE (maximum likelihood estimation) and standard error, we have started the iterative procedure by maximizing the log-likelihood function given in (3.1) directly with an initial guess for α=1.0 and λ=0.5, far away from the solution[13]. We have used *optim*( ) function in R with option Newton-Raphson method[14, 15]. The iterative process stopped only after various no. of iterations depend on used data set[16]. The Table 1 shows the ML estimates, standard error(SE) with number of Iterations and Log-Likelihood value of the parameters alpha and lambda.



Data Set	MLE		Std. Error		No. of Iteration	Log-Likelihood
	alpha	lambda	alpha	lambda		
1	0.76583780	0.00319420	0.3868425	0.00300860	18	-79.11598
2	2.70728010	0.00418410	4.8388226	0.00901020	30	-118.5001
3	2.29569800	0.00733690	1.4572162	0.00577300	39	-238.2042
4	0.63476001	0.63639010	0.1400700	0.28583000	10	-103.2059
5	2.50843535	0.00147436	1.2959431	0.00089688	23	-62.32879

Table1. ML estimates, Standard Error with no. of Iterations and Log-Likelihood value

V. Model Validation

Most statistical methods assume an underlying model in the derivation of their results. However, when we presume that the data follow a specific model, we are making an assumption. If such a model does not hold, then the conclusions from such analysis may be invalid. Although hazard plotting and the other graphical methods can guide the choice of the parametric distribution, one cannot of course be sure that the proper model has been selected. Hence model validation is still necessary to check whether we have achieved the goal of choosing the right model[17]. In this paper we outline some of the methods used to check model appropriateness.

A. *Kolmogorov–Smirnov Test*

The Kolmogorov–Smirnov test (K–S test) is a nonparametric test for the equality of continuous and that can be used to compare a sample with a reference probability model. The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution[18].

- **The Empirical Distribution Function(EDF)**

An estimate of F(x) = P[X ≤ x] is the proportion of sample points that fall in the interval [-, x]. This estimate is called the empirical distribution function(EDF). The EDF of an observed sample $x_1, x_2, \ldots, x_n$ is defined by

$$F_n(x) = \begin{cases} 0 & \text{for} \quad x < X_{1:n} \\ \frac{i}{n} & \text{for} \quad X_{i:n} \le x < X_{i+1:n}\,;\, i = 1, \ldots, n-1 \\ 1 & \text{for} \quad x \ge X_{n:n} \end{cases}$$

where $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$ is the ordered sample.

The Kolmogorov–Smirnov (K-S) test is a nonparametric *goodness-of-fit* test and is used to determine whether an underlying probability distribution ($F_n(x)$) differs from a hypothesized distribution ($F_0(x)$).

- **Kolmogorov-Smirnov (K-S) distance**

The K-S distance between two distribution functions is defined as

$$D_n^+ = \max_{1 \le i \le n} \left| F_n(x) - F_0(x_i) \right|, \text{ and}$$

$$D_n^- = \max_{1 \le i \le n} \left| F_0(x_i) - F_n(x) \right|,$$

where $F_0(x_i)$ is the cumulative distribution function evaluated at x_i and $F_n(x)$ is the EDF. To perform the two-sided goodness of fit test $H_0 : F(x) = F_0(x)$ for all x, where F is a completely specified continuous distribution function against the alternative $H_1 : F(x) = F_0(x)$, for some x, the K-S statistic is

$$D_n = \max_{1 \le i \le n} \left\{ D_n^+, D_n^- \right\}$$

The distribution of the K-S statistic does not depend on F_0 as long as F_0 is continuous.

To study the goodness-of-fit of the Exponential Extension model, we compute the Kolmogorov-Smirnov statistic between the empirical distribution function and the fitted distribution function when the parameters are obtained by method of maximum likelihood. We shall use the *ks.expo.ext*( ) function in R given in [8] to perform the test. Now, we plot the empirical distribution function and the fitted distribution function using proposed data sets in Figure 3-7 and the result of K-S test is shown in table 2.

Table2. D and its Corresponding p-value using KS-test

Data Set	D - value	P - value
1	0.1862	0.77620
2	0.2885	0.04348
3	0.1915	0.05108
4	0.1309	0.40930
5	0.4853	0.01084

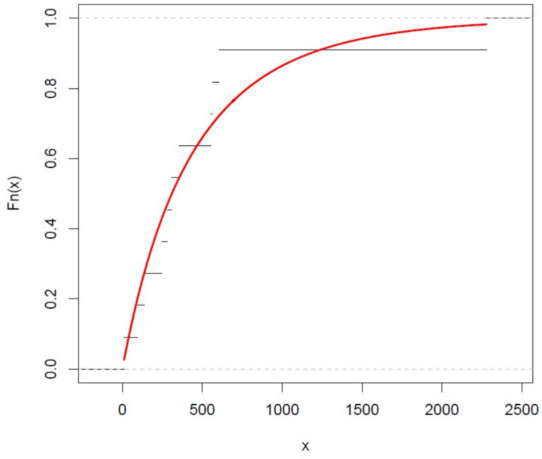


Fig 3. The graph for empirical distribution function and fitted distribution function for data set-1.

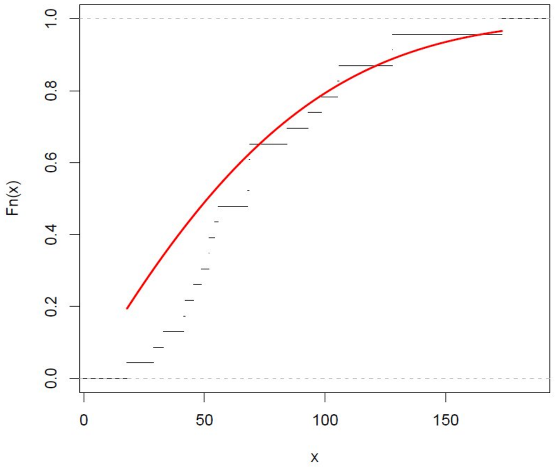


Fig 4. The graph for empirical distribution function and fitted distribution function for data set-2.

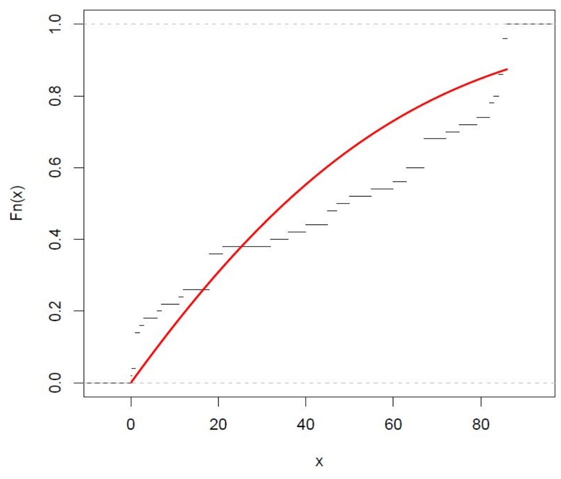


Fig 5. The graph for empirical distribution function and fitted distribution function for data set-3.

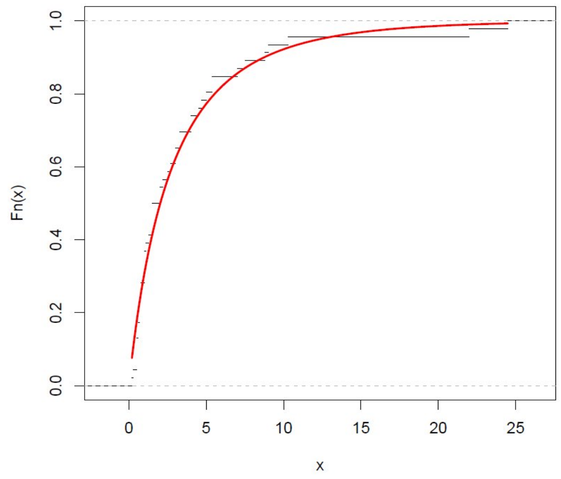


Fig 6. The graph for empirical distribution function and fitted distribution function for data set-4.

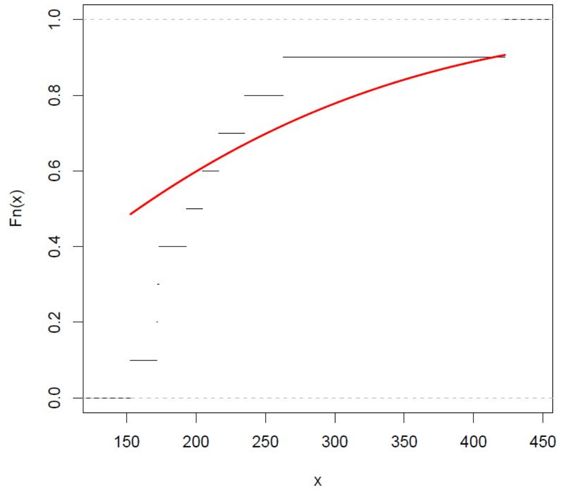


Fig 7. The graph for empirical distribution function and fitted distribution function for data set-5.

Since, the high p-value clearly indicates that this data set can be used to analyze EE model, and in this analysis data set-1 and data set-4 having high p-value. Therefore from above result and Figure 3-7, it is clear that the estimated EE model provides excellent good fit to the given data set-1 and data set-4.

B. The Q-Q Plots Test

The Q-Q plot test is used to investigate whether an assumed model adequately fits a set of data. It helps the analyst to assess how well a given theoretical distribution fits the data.

Let $x_l, x_2, . . . , x_n$ be a sample from a given population with cdf F(x). Let $x_{l:n}, x_{2:n}, . . . , x_{n:n}$, be the corresponding order statistics and $p_{l:n}, p_{2:n}, . . . , p_{n:n}$ be the plotting positions. Define the plotting positions by [19, 20],

$$p_{1:n} = \frac{i-0.5}{n} \ ; \ i=1, 2, . . ., n.$$

Finally, let $\hat{F}(x)$ be an estimate of F(x) based on $\underline{x} = (x_l, x_2, . . . , x_n)$. Thus, $\hat{F}^{-1}(p_{1:n})$ is the estimated quantile corresponding to the ith order statistic, $x_{i:n}$ Similarly, $\hat{F}(x_{i:n})$ is the estimated probability corresponding to $x_{i:n}$.

again,

Let $\hat{F}(x)$ be an estimate of F(x) based on $x_l, x_2, . . . , x_n$. The scatter plot of the points

$$\hat{F}^{-1}(p_{1:n}) \ \text{ versus } \ x_{i:n} \ , \ \ i = 1, 2, . . . , n,$$

is called a Q-Q plot. Thus, the Q-Q plots show the estimated versus the observed quantiles. If the model fits the data well, the pattern of points on the Q-Q plot will exhibit a 45-degree straight line. Note that all the points of a Q-Q plot are inside the square

$$\left[\hat{F}^{-1}(p_{1:n}),\ \hat{F}^{-1}(p_{n:n})\right] \times \left[x_{1:n}, x_{n:n}\right].$$

Now, we shall use the R function *qq.expo.ext( )* given in [8] to perform the proposed test. We draw Quantile-Quantile(Q-Q) plot using MLEs as estimate for used different data set in given Figure 8-12.

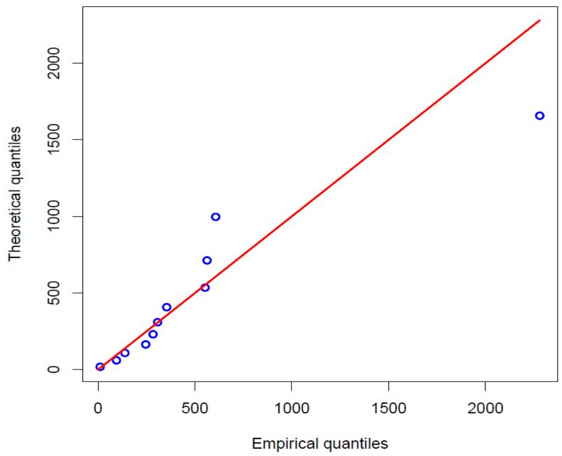


Fig 8. Quantile-Quantile(Q-Q) plot using MLEs as estimate for data set-1.

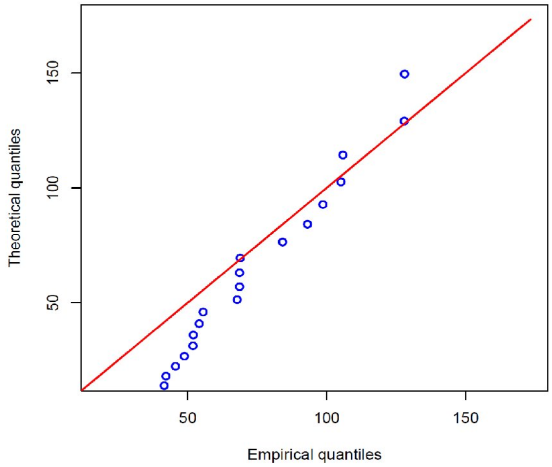


Fig 9. Quantile-Quantile(Q-Q) plot using MLEs as estimate for data set-2.

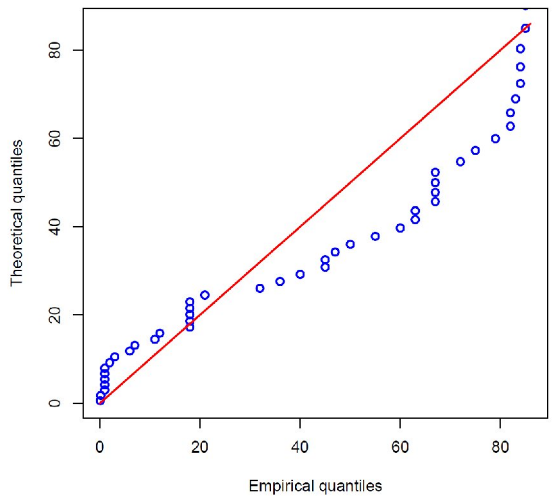


Fig 10. Quantile-Quantile(Q-Q) plot using MLEs as estimate for data set-3.

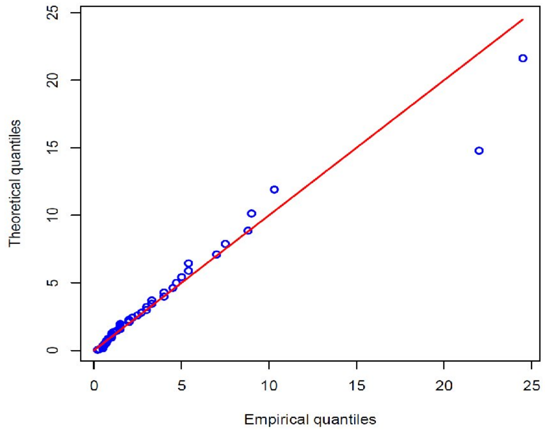


Fig 11. Quantile-Quantile(Q-Q) plot using MLEs as estimate for data set-4.

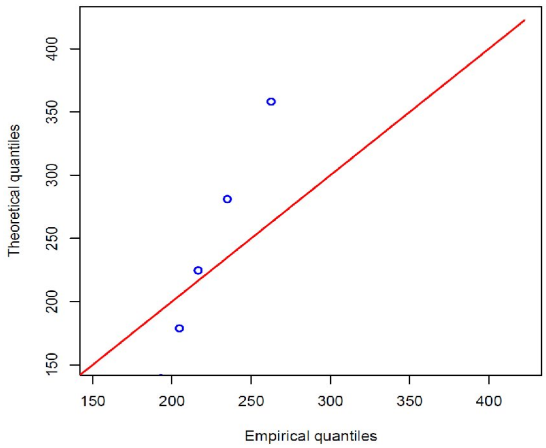


Fig 12. Quantile-Quantile(Q-Q) plot using MLEs as estimate for data set-5.

Thus, as can be seen from the straight line pattern in Figure 8-12, the EE model fits the data very well for data set-1 and data set-4.

VI. Conclusion

An attempt has been made to incorporate Exponential Extension model for software reliability data. We have presented the statistical tools for empirical modeling of the data in general. These tools are developed in R language and environment for model analysis, model validation and estimation of parameters using method of maximum likelihood. To check the validity of the model, we have plotted a graph of empirical distribution function and fitted distribution function using KS-test for different data set and also we have to present power comparison between p-values of these data sets obtaining by K-S test for receiving feasible real data sets which are excellent good fit for analysis of Exponential Extension model. We have also discussed the Quantile-Quantile (Q-Q) plots for model validation. Thus, from both used techniques of model validation for EE model on different data set, the Exponential Extension model fits the data very well only for data set-1 and data set-4.

VII. ACKNOWLEDGEMENT

The author is thankful to Dr. Vijay Kumar, Associate Professor in Department of Mathematics and Statistics in DDU Gorakhpur University, Gorakhpur, the editor and the referees for their valuable suggestions, which improved the paper to a great extent.

VIII. REFERENCES

[1] Murthy, D.N.P., Xie, M. and Jiang, R. (2003). *Weibull Models*, Wiley, New York.

[2] Haghighi, F. and Sadeghi, S., (2009) ' An Exponential Extension', URL: http://hal.inria.fr/docs/00/38/67/55/PDF/p186.pdf

[3] NADARAJAH, S., and Haghighi, F. (2009): An extension of the exponential distribution, Statistics, doi: 10.1080/02331881003678678.

[4] Gupta, R.D., Kundu, D., 2001. Exponentiated exponential family: an alternative to gamma and Weibull distributions. Biometrical Journal 43, 117-130.

[5] Srivastava, A.K. and Kumar V.(2011). *Analysis of Software Reliability Data using Exponential Power Model*. International Journal of Advanced Computer Science and Applications, Vol. 2, No. 2, February 2011, 38-45.

[6] Kumar, R., Srivastava, A.K. and Kumar, V. (2012). *Analysis of Gumbel Model for Software Reliability Using Bayesian Paradigm*, International Journal of Advanced Research in Artificial Intelligence, Vol. 1 (9), 39-45.

[7] Kumar, V. (2010). *Bayesian analysis of exponential extension model*, J. Nat. Acad.Math., Vol. 24, 109-128.

[8] Kumar, V. and Ligges, U. (2011). *reliaR : A package for some probability distributions*. http://cran.r-project.org/web/packages/reliaR/index.html.

[9] Lawless, J. F., (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed., John Wiley and Sons, New York.

[10] Aarset, M. V. (1987). How to identify bathtub hazard rate. *IEEE Transactions Reliability*, 36,106-108.

[11] Chhikara, R. S. & Folks, J. L. (1977). The inverse Gaussian distribution as a lifetime model. Technometrics, 19, 461–468.

[12] J. I. McCool, "Inferential Techniques for Weibull Populations", *Aeropace Research Laboratories Report ARL TR74-0180, Wright-Patterson Air Force Base, Dayton, Ohio*, 1974.

[13] Ihaka, R. and Gentleman, R.R. (1996). R: *A language for data analysis and graphics*, Journal of Computational and Graphical Statistics, 5, 299–314.

[14] Venables, W. N., Smith, D. M. and R Development Core Team (2010). *An Introduction to R*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-12-7, http://www.r-project.org.

[15] Srivastava, A.K. and Kumar V. (2011). *Markov Chain Monte Carlo methods for Bayesian inference of the Chen model*, International Journal of Computer Information Systems, Vol. 2 (2), 07-14.

[16] Eastman, J. and Bain, L.J., (1973).*A property of maximum likelihood estimators in the presence of location-scale nuisance parameters*, Communications in Statistics, 2, 23-28.

[17] Srivastava, A.K. and Kumar V. (2011). Software reliability data analysis with Marshall-Olkin Extended Weibull model using MCMC method for non-informativeset of priors, International Journal of Computer Applications, Vol. 18(4), 31-39.

[18] Hazewinkel, Michiel, ed. (2001), "Kolmogorov-Smirnov test", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4

[19] Evans, M., Hastings, N., and Peacock, B.(2000). *Statistical Distributions*, 3rd ed. New York, Wiley.

[20] Thode, H. C. (2002). Testing For Normality. CRC Press, ISBN: 0-8247-9613-6.

AUTHOR'S PROFILE

ASHWINI KUMAR SRIVASTAVA received his M.Sc in Mathematics from D.D.U.Gorakhpur University, MCA(Hons.) from U.P.Technical University, M. Phil in Computer Science from Allagappa University and Ph.D. in Computer Science from D.D.U.Gorakhpur University, Gorakhpur. Currently working as Assistant Professor in Department of Computer Application in Shivharsh Kisan P.G. College, Basti, U.P. He has got 8 years of teaching experience as well as 4 years research experience. His main research interests are Software Reliability, Artificial Neural Networks, Bayesian methodology and Data Warehousing.