RT&A, No 4(80)
Volume 19, December, 2024

A NEW ALGORITHM FOR MODELING ASYMMETRICAL DATA – AN EMPIRICAL STUDY

K.M. Sakthivel and Vidhya G

•

Department of Statistics, Bharathiar University, Coimbatore 641046, Tamil Nadu, India
sakthithebest@buc.edu.in,vidhyastatistic96@gmail.com

Abstract

In the current era, it is quite challenging to find symmetric data, as the form of most real-world data is asymmetric, meaning it tends to slant towards one side or another. These types of data emerge from various fields, including finance, economics, medicine, and reliability. Traditional statistical models often fail to handle such type of data as most of the statistical procedures are developed under normality assumptions. Therefore, the usual way of modeling these data results in incorrect predictions or leads to wrong decisions. There is no familiar methodology available in the research for modeling asymmetric data. Hence, there is a need to address this research gap as an emerging area of research in statistical modeling. In this paper, we propose a new systematic approach called the Model Selection Algorithm for modeling asymmetric data. In this algorithm, we incorporate various statistical tools and provide a guideline for a step-by-step procedure. Further, we have applied maximum likelihood estimation for parameter estimation, and model selection criteria such as Cramer Von Mises, Anderson Darling, and Kolmogorov Smirnov tests. We used real-time data to demonstrate the effectiveness of the algorithm.

Keywords: Lifetime distributions, Estimation, Information Criteria, Goodness of fit, Model selection.

1. Introduction

In this information era, data plays a significant role in policy and decision-making in various fields. With the advancement of technology, data generation, and its utilization have been increasing exponentially. However, it is important to note that the behavior of data is dynamic and depends on several factors. Therefore, it is crucial to understand the intricacies of data and its behavior to utilize it effectively. As part of our data analysis process, we use statistical models to gain a better understanding of the patterns and trends present in the data. These models help us to delve deeper into the underlying structure of the data and make informed decisions based on the findings from the statistical inference.

The foundation of the probability models is to capture the dynamics and variability of data. A vast amount of literature has been written about probability models, and new results are being produced daily. Choosing the right model for a particular asymmetrical dataset may be difficult for even statistics experts. To facilitate this process, we have developed a framework that considers choosing the best model for asymmetric data under study. Both statisticians and data analysts may benefit from this approach to make sensible Inference.

Data that has an uneven pattern due to an unequal distribution of data points' frequencies is referred to as asymmetrical data or skewed data. This kind of data is not symmetrical since the mean, median, and mode are not equal and also not at the same location. Consequently, the distribution takes on an extended form on one side and a longer or fatter tail on the other. We

will encounter many fields including Finance, Economics, Medicine, etc.

To provide a more sophisticated depiction of asymmetrical data, a combination of probability models may be employed. This technique proves especially valuable when handling data from a range of sources or with varying patterns. Blending different distributions into one mixture distribution enables the model to more effectively capture the distinctive features of the data. This proves particularly advantageous when the true underlying distribution remains unknown. By embracing a mixture model, analysts can more precisely assess the probability of different outcomes and make informed decisions based on the data at hand.

Karl Pearson [25] a prominent biometrician, proposed one of the earliest mixture models by fitting a proportionate mixture of two normal density functions. Subsequently, several writers used the finite mixture model to create bimodal distributions. This approach was limited to bimodal or multimodal datasets until Lindley [22] used the finite mixture model to generate a single parameter distribution for unimodal data. Furthermore, a mixed distribution was created by utilizing different proportions of gamma and exponential distributions, resulting in an improved outcome.

For the past few years, the authors of the study initially developed a model, which was subsequently applied to real-time data. However, the current situation presents a challenge, as there is no clarity on the most appropriate model to use for the available data. The data at hand are asymmetric, which makes it difficult to determine the most suitable distribution. To address this issue, a framework was established to identify the model that best aligns with the data. This involved a meticulous examination of the data properties to select a model that accurately captures the data features. Consequently, a model was developed that can provide reliable predictions through this approach. The study aims to provide a concise and comprehensible process for selecting an appropriate model for asymmetric real-time data. Considerable efforts were made to ensure that the method is straightforward and easy to understand, with the ultimate goal of facilitating the decision-making process for businesses and academics alike.

In our paper, we begin by presenting the framework that we followed for fitting skewed data in section 2. In Section 3, we also provide additional insight into the assumptions and limitations of our approach. We discuss a real-time application to demonstrate the effectiveness of our framework in finding the most suitable mixture model by our proposed methodology. Finally, in Section 4, we discuss the basic properties of the proposed model and also discuss the simulation work that was done for the proposed model.

2. Model Selection Algorithm

This algorithm outlines a comprehensive and systematic approach to analyzing asymmetrical data. It involves a model selection process that is designed to provide a detailed understanding of the data. The first step in this process is to gather the data and partition it using clustering techniques. This allows for the identification of distinct groups within the data, which can then be analyzed separately. After the data has been partitioned, the next step is to fit a probability distribution to each partition. This is done to determine the best-fit model for each group of data. The selection of the best-fit model is based on a variety of factors, including the goodness of fit, the complexity of the model, and the interpretability of the results. Finally, the best-fit models from each partition are combined to propose a comprehensive hybrid model. This model provides a detailed understanding of the data and allows for the identification of patterns and trends that may not be apparent when analyzing the data as a whole. Sakthivel and Vidhya ([26]-[27]) have discussed the algorithm and given different applications.

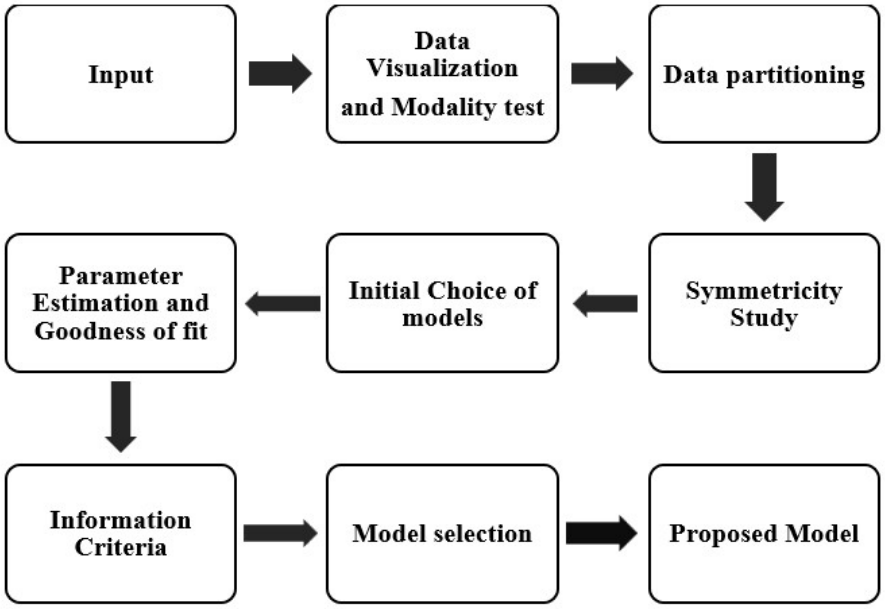

Figure 1: *Framework for selecting a better model*

The steps involved in this algorithm are as follows:

Step 1: Consider the asymmetrical data.

Step 2: Visualize the data to identify the distribution of the random variable. And test the modality of the data.

Step 3: Divide the data into two parts using the clustering technique.

Step 4: Calculate the skewness for both parts of the data to capture the asymmetry.

Step 5: Consider the basic distributions for modeling based on the data characteristics.

Step 6: For the first part of the data, estimate the value of parameters using MLE for suitable probability distributions and determines the model's adequacy by computing the goodness of fit and information metrics.

Step 7: Repeat the process for the second part of the data.

Step 8: Choose a better model from the considered distributions in steps 6 & 7, based on minimized -2LL, AIC, BIC, and AICc values.

Step 9: Propose a new model by combining the selected models from Step 8.

An advanced framework has been developed to effectively analyze asymmetrical data using a suite of analytical tools. In our framework, we incorporate the fundamental models to choose a better model. The distributions incorporated in this framework such as symmetric, heavy-tailed, light-tailed, positively skewed, and negatively skewed models. The framework employs a range of statistical techniques, including K-mean Clustering for initial data partitioning, maximum likelihood estimation for parameter estimation, and statistical tests such as Cramer Von Mises, Anderson Darling, and Kolmogorov Smirnov for rigorous model evaluation. The model selection process is further refined through the application of information criteria, which includes Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Corrected Akaike Information Criteria (AICc). These criteria help to identify the most suitable model for analyzing the data,

ensuring that the insights gained are both accurate and reliable.

This comprehensive approach ensures an efficient methodology for gaining valuable insights from asymmetrical datasets and can be used in a variety of applications, including Engineering, Finance, Marketing, and Healthcare. By utilizing a range of analytical tools and statistical techniques, the framework can effectively analyze complex data sets, providing valuable insights that can be used to inform decision-making processes.

Before using the framework, it is important to consider certain assumptions. Firstly, the data used for analysis should be asymmetrical with dual peaks and should not be perfectly bimodal. Additionally, the framework is specifically designed for situations where the distribution of the data significantly deviates from a normal distribution. To be suitable for analysis using this framework, the data must exhibit some degree of skewness, either positive or negative. Furthermore, the framework is best suited for data that exhibits heavy tails and allows for the existence of outliers in the dataset. It is important to note that this framework is designed for univariate data analysis only.

It is important to keep in mind that there are certain limitations to this procedure. Especially, this framework is designed to handle asymmetrical data and requires dividing the data into exactly two distinct groups. This method has been found to produce superior results for smaller datasets. We use the limited models to choose a better model. Manual processing of this framework with significant amounts of data can present numerous challenges and be time-consuming. In future research, we aim to automate this process through statistical software and produce results using simple codes. This approach will help streamline the process and make it more efficient for larger datasets.

3. Model Selection and Validation

In this section, we would like to apply our framework to the datasets on the glass strength of aircraft windows, as originally reported by Fuller et al. (1994). This dataset is widely recognized and has been utilized by various authors in the literature. The strength data involved can potentially originate from different underlying probability distributions, each representing varying conditions or modes.

It is important to note that the strength of aircraft windows is a crucial factor in ensuring the safety and reliability of air travel. Hence, it is essential to accurately represent the underlying distribution of the strength data to better understand and predict its behavior. By utilizing a mixture of probability distribution, we can more effectively capture the inherent variability in the data and simply fit standard probability distribution. The proposed approach will allow us to identify the most suitable model for the actual strength data to ensure the safety and reliability of aircraft windows.

Further, a mixture of probability distributions is a statistical model that combines two or more probability distributions into a single probability distribution. This concept of mixture enables to modeling of multimodal data, which present different types of defects or variations in the manufacturing process. By incorporating multiple distributions, we can capture the full range of variability in the data, which is crucial for modeling the strength of aircraft windows under different conditions.

The summary statistic of the data is given in Table 1.

Table 1: *Summary of data*

Minimum	First Quartile (Q1)	Median	Mean	Third Quartile (Q3)	Maximum	Skewness
18.83	25.51	29.90	30.81	35.83	45.38	0.4263

It is observed that the mean and median values are different. This indicates that the shape of the density plot is likely to be skewed and not symmetric. These findings can be ensured with the support of other statistical plots given below.

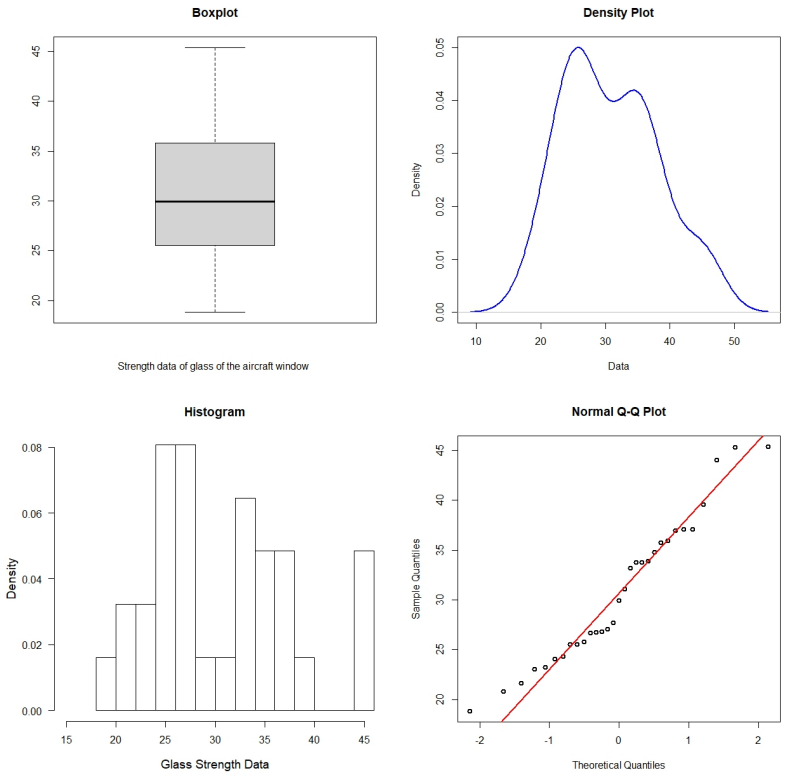

Figure 2: *Graphical representation of the data*

The analysis of the dataset through graphical representation has revealed the presence of two modes, which are depicted in Figure 2. Additionally, we check the modality of the data by using the Hartigan dip test (1985). Dip statistic=0.081364 ($p - value$ =0.09169). The Hartigan dip test confirms the data's unimodality. The Q-Q plot, which shows that the data is not symmetric, indicates that the data is asymmetric. In light of these findings, we have proceeded to step 3 of our analytical framework, which involves the division of the data into two parts using K-mean clustering. Figure 3 presents the clustered data along with the cut line that was used to divide the data.

After partitioning the data, the next process will evaluate the skewness of both partitions of the datasets. We can select the model accordingly if the data is positively skewed or negatively skewed.

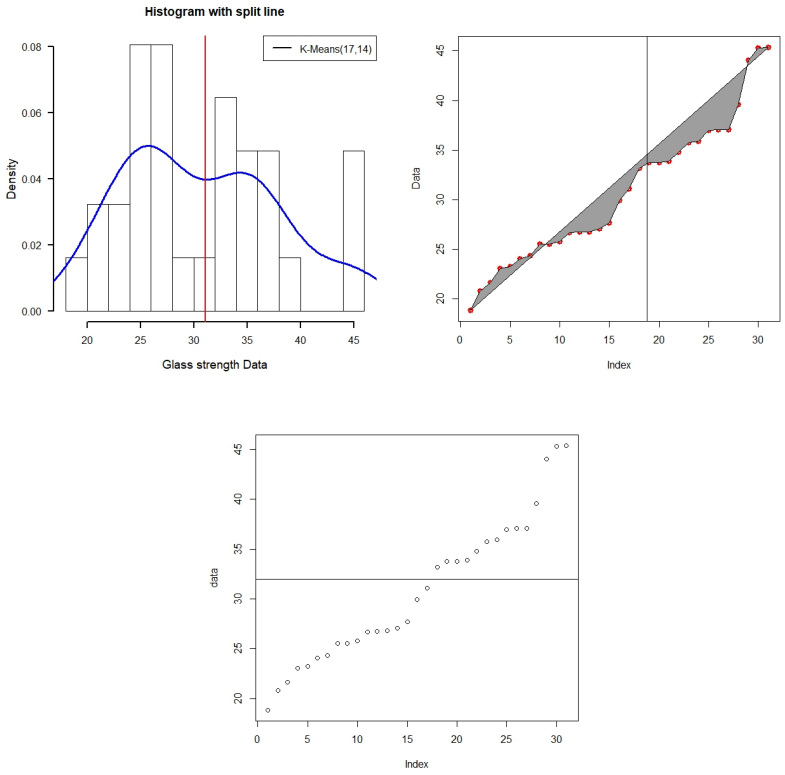

Figure 3: *Data partitioning for data*

Table 2: *Summary of the first part of the data*

Minimum	Q1	Median	Mean	Q3	Maximum	Skewness
18.83	23.23	25.52	25.22	26.78	31.11	-0.1538

Since the initial portion of the data is unimodal and negatively skewed, we can consider a probability distribution that also portrays this particular characteristic. The low value of skewness indicates that the distribution may be nearly symmetric. Therefore, we considered the normal distribution as one of the choices. The present section deals with the computation of the tools mentioned in Section 2 for diverse distributions. These computations are crucial to accomplishing step 6 of the analysis. All statistical findings and calculations have been obtained through the R programming language, and the resulting outcomes have been documented in Table 3.

Lower values for KS, CVM, and AD suggest a better fit between the model and data. Higher values imply a poor fit. We can compare p-values to determine the goodness of fit. A p-value closer to zero denotes a weaker fit, while a p-value closer to unity denotes a better fit.

From Table 3, we are unable to identify which model best matches the data. On the other hand, we might go on to the next process if we filter the model using Table 3. After calculating the information criteria, the data's best fit is selected.

Table 3: *Estimated parameters value and Goodness of fit for the first part of the dataset.*

Model	Estimated Parameter	-2logL	CVM (p values)	AD (p-values)	KS (p-values)
Gamma	$\hat{\alpha}$=66.3065 $\hat{\beta}$=2.6293	86.4328	0.0419 (0.9287)	0.2466 (0.972)	0.1406 (0.8451)
Lognormal	$\hat{\mu}$=3.2200 $\hat{\sigma}$=0.1239	86.7240	0.0468 (0.9014)	0.2746 (0.9553)	0.1480 (0.7992)
Weibull	$\hat{k}$=9.2248 $\hat{\lambda}$=26.5551	86.4999	0.0362 (0.9561)	0.2449 (0.9729)	0.1291 (0.9058)
Cauchy	$\hat{a}$=25.6887 $\hat{b}$=1.7157	91.4211	0.0546 (0.8542)	0.3931 (0.8537)	0.1259 (0.9195)
Logistic	$\hat{a}$=25.2943 $\hat{b}$ =1.73282	86.4194	0.0316 (0.9742)	0.1950 (0.9919)	0.1179 (0.9500)
Normal	$\hat{\mu}$=25.2181 $\hat{\sigma}$=3.0433	86.0835	0.0344 (0.9636)	0.2100 (0.9876)	0.1251 (0.9231)
Gompertz	$\hat{k}$=0.1834 $\hat{\lambda}$=0.0016	96.5179	0.3806 (0.0803)	1.9689 (0.0962)	0.2722 (0.1328)
Gumbel	$\hat{k}$=23.6725 $\hat{\lambda}$=3.0430	89.1066	0.0793 (0.7019)	0.4724 (0.7727)	0.1660 (0.6767)
Laplace	$\hat{\lambda}$=25.5200 $\hat{\beta}$=2.3737	86.9575	0.0408 (0.9343)	0.2566 (0.9666)	0.1100 (0.9717)

Table 4: *Model selection criteria for the first part of the dataset*

Model	AIC	BIC	AICc
Gamma	90.4328	92.0993	91.2899
Lognormal	90.7240	92.3905	91.5812
Weibull	90.4999	92.1664	91.3571
Cauchy	95.4211	97.0875	96.2782
Logistic	90.4194	92.0858	91.2766
Normal	90.0835	91.7499	90.9407
Gompertz	100.5179	102.1843	101.3750
Gumbel	93.1066	94.7730	93.9637
Laplace	90.9575	92.6239	91.8147

From Table 4, selecting a model is a simpler process. It was determined that the distribution with the lowest values of AIC, BIC, and AICc provided the best fit. Hence, the normal distribution is the most suitable fit for the first segment of the data. This process is then repeated for the second part of the data.

Table 5: *Summary of the second part of the dataset*

Minimum	Q1	Median	Mean	Q3	Maximum	Skewness
33.20	34.11	36.45	37.60	38.96	45.38	0.90099

Since the data's second component is skewed towards positive values, we should choose a probability distribution that also displays positive skewness. This will help us accurately represent the shape of the data and ensure more representative statistical analysis and modeling.

Table 6: *Estimated parameters value and Goodness of fit for the second part of the dataset.*

Model	Estimated Parameter	-2logL	CVM (p-values)	AD (p-values)	KS (p-values)
Rayleigh	$\hat{\sigma}$=26.753	110.6552	0.8793 (0.0039)	4.1262 (0.0078)	0.5370 (0.0003)
Lindley	$\hat{\theta}$=0.0519	119.5955	0.9050 (0.0034)	4.3004 (0.0065)	0.5289 (0.0004)
Exponential	$\hat{\theta}$=0.0266	129.5586	1.1305 (0.0009)	5.2124 (0.0024)	0.5866 (0.0000)
Gamma	$\hat{\alpha}$=85.9100 $\hat{\beta}$=2.2846	78.7922	0.1428 (0.4163)	0.9128 (0.4045)	0.2506 (0.2914)
Weibull	$\hat{k}$=8.8277 $\hat{\lambda}$=39.5728	82.47142	0.1983 (0.2725)	1.1507 (0.2861)	0.2829 (0.1745)
Pareto	$\hat{\alpha}$=8.4226	69.7299	0.0427 (0.9254)	0.2184 (0.9324)	0.1266 (0.9573)
Lomax	$\hat{\alpha}$=185.6654 $\hat{\lambda}$=5.3694	132.0715	1.1395 (0.0008)	5.2563 (0.0023)	0.5866 (0.0000)
Lognormal	$\hat{\mu}$=3.6213 $\hat{\sigma}$=0.1064	78.3804	0.1344 (0.4453)	0.8709 (0.4305)	0.2440 (0.3208)
Cauchy	$\hat{a}$=35.8128 $\hat{b}$=1.9214	81.5577	0.1156 (0.5192)	0.8219 (0.4631)	0.2018 (0.5517)
Logistic	$\hat{a}$=36.9773 $\hat{b}$=2.3644	80.0735	0.1038 (0.5729)	0.8596 (0.4378)	0.2024 (0.5483)
Normal	$\hat{\mu}$=37.6033 $\hat{\sigma}$=4.1682	79.7004	0.1608 (0.3614)	1.0064 (0.3524)	0.2633 (0.2402)
Gompertz	$\hat{k}$=0.1200 $\hat{\lambda}$=0.0014	91.5035	0.5587 (0.0269)	2.6816 (0.0407)	0.4545 (0.0037)
Gumbel	$\hat{k}$=35.7494 $\hat{\lambda}$=2.8798	75.6418	0.0814 (0.6912)	0.6522 (0.5971)	0.1805 (0.6872)
Laplace	$\hat{\mu}$=36.6535 $\hat{\beta}$=3.1747	79.7542	0.0815 (0.6908)	0.7508 (0.5152)	0.1684 (0.7628)

Table 7: *Model selection criteria for the second part of dataset*

Model	AIC	BIC	AICc
Gamma	82.7922	84.0703	83.8831
Weibull	86.4714	87.7495	87.5623
Pareto	71.7299	72.3689	72.0633
Lognormal	82.3804	82.3804	83.6585
Cauchy	85.5577	86.8358	86.6486
Logistic	84.0735	85.3516	85.1644
Normal	83.7004	84.9785	84.7913
Gumbel	79.6418	80.9199	80.7328
Laplace	83.7542	85.0323	84.8451

Applying the same methodology to the second half of the data, we can conclude from Tables 6 and 7 that the Pareto distribution is the best fit. We may go on to the next stage once we have successfully obtained the two components required for the two sections of our data. We may infer from step 7 that a Pareto distribution would be appropriate for the second half of the data, and a normal distribution for the first part. Step 9 involves creating a new distribution called the Normal-Pareto distribution (NPD) by combining these two distributions using the finite mixture

model. Section 4 describes the function of NPD.

We must test our suggested model to ensure that it functions accurately and efficiently with data. We will be able to demonstrate that our framework performs better than the current model and yields better outcomes through this approach. Thus, we carry out step 6 again using the suggested model, and Table 8 presents the outcomes.

Table 8: *Estimated parameters value, Goodness of fit, and Model selection criteria for the dataset*

Model	Estimated Parameter	-2LL	CVM	AD	KS	AIC	BIC	AICc
Rayleigh	$\hat{\sigma}$=22.36	236.4	0.8388 (0.0055)	4.3918 (0.0057)	0.3189 (0.0027)	238.4	239.8	238.5
Lindley	$\hat{\theta}$=0.063	253.9	1.1389 (0.0010)	5.8718 (0.0011)	0.3655 (0.0003)	255.9	257.4	256.1
Exponential	$\hat{\theta}$=0.032	274.5	1.7891 (0.000)	8.5303 (0.000)	0.4587 (0.0000)	276.5	277.9	276.6
Gamma	$\hat{\alpha}$=18.93 $\hat{\beta}$=0.614	208.2	0.0816 (0.6863)	0.4387 (0.8085)	0.1349 (0.5785)	212.2	215.1	212.6
Lognormal	$\hat{\mu}$=3.401 $\hat{\sigma}$=0.231	208.0	0.0791 (0.7007)	0.4136 (0.8341)	0.1246 (0.6759)	212.0	214.8	212.4
Weibull	$\hat{k}$=4.635 $\hat{\lambda}$=33.674	210.9	0.0908 (0.6353)	0.5973 (0.6492)	0.1526 (0.4238)	214.9	217.8	215.4
Pareto	$\hat{\alpha}$=2.146	225.5	0.6215 (0.0192)	3.2534 (0.000)	0.25419 (0.0298)	227.5	228.9	227.6
Lomax	$\hat{\alpha}$=113.53 $\hat{\lambda}$=4.392	281.5	2.0613 (0.0000)	9.6293 (0.000)	0.4903 (0.000)	285.5	288.3	285.9
Cauchy	$\hat{a}$=29.258 $\hat{b}$=5.093	225.6	0.1696 (0.3363)	1.2009 (0.2669)	0.1612 (0.3573)	229.6	232.4	230.0
Logistic	$\hat{a}$=30.44 $\hat{b}$=4.224	211.6	0.0966 (0.6051)	0.5729 (0.6727)	0.1425 (0.5094)	215.6	218.5	216.0
Normal	$\hat{\mu}$=30.81 $\hat{\sigma}$=7.135	209.8	0.0936 (0.6203)	0.5559 (0.6892)	0.154 (0.4125)	213.8	216.6	214.2
Gompertz	$\hat{k}$=0.117 $\hat{\lambda}$=0.002	216.7	0.1332 (0.4473)	1.0257 (0.3434)	0.1549 (0.4053)	220.7	223.6	221.1
Gumbel	$\hat{k}$=27.399 $\hat{\lambda}$=5.986	208.2	0.0757 (0.7204)	0.3980 (0.8497)	0.1358 (0.5704)	212.2	215.1	212.7
Laplace	$\hat{\mu}$=29.900 $\hat{\beta}$=6.124	217.3	0.1477 (0.3984)	0.8691 (0.4328)	0.1599 (0.367)	221.3	224.2	221.7
Normal-Pareto (NPD)	$\hat{\theta}$=0.973 $\hat{\mu}$=29.690 $\hat{\sigma}$=8.117 $\hat{\alpha}$=152.23	202.3	0.062 (0.8019)	0.551 (0.6590)	0.122 (0.6984)	210.3	216.0	211.8

By comparing the newly created probability distribution against conventional distributions using every criterion that was used to choose the model, its goodness of fit was assessed. Table 8 and Figure 4 make it clear that our suggested mixed probability model yields the best results, and our methodology helps select a more appropriate model for the skewed data.

However, comparing the traditional distribution alone is not enough to prove that our proposed model is a better fit for the data so we extended the study and collected various distributions using various techniques, and compared it with our proposed model. The results are given in Table 9.

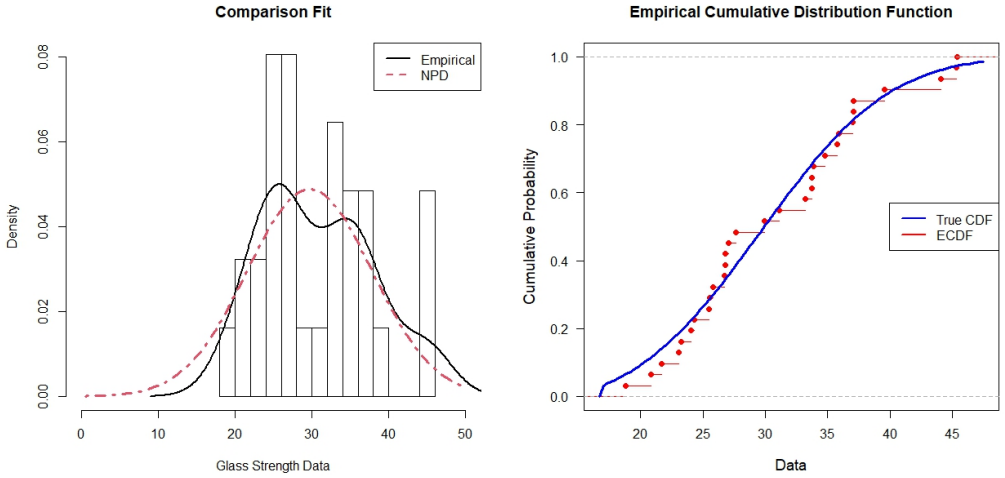

Figure 4: *Comparison fit for the dataset*

Other than the model listed in Table 9, many distributions are taken into account for comparison. The distributions are Shanker, Akash, Rama, Suja, Sujatha, Amarendra, Devya, Shambhu, Aradhana, Akshya, Inverse Rayleigh, Inverse Exponential, Kpenadidum, Iwok-Nwi, Two parameter Pranav, Two parameter Sujatha, Weibull Extended Pranav, Extended Pranav, Weibull-Lindley, Weibull- Pranav, Exponentiated Exponential, Exponentiated Hypoexponential, Generalized Inverted Exponential, Inverted Exponential, Beta Generalized Inverted Exponential, Exponentiated Exponential, Exponentiated Lindley, Exponentiated Akash, Gold, Power Size Biased Two Parameter Akash distributions.

In summary, the Normal-Pareto distribution provides the best fit to the data when compared to other distributions.

4. Key Properties of Normal-Pareto Distribution

From the above statistical analysis, we have examined our proposed model key features in depth. The model satisfy the essential condition for the distribution function, this shows that they can be used for more research and application in pertinent areas.

4.1. Normal-Pareto Distribution (NPD)

The model can be obtained using a finite mixture model.

$$f(x) = w_1 g_1(x) + w_2 g_2(x) \tag{1}$$

Equation (1) is used for developing the Normal-Pareto model. Where $g_1(x) \sim Normal(\mu, \sigma)$, $g_2(x) \sim Pareto(x_m, \alpha)$ and $w_1 = \theta\,; w_2 = 1 - w_1 = 1 - \theta$ and x_m is the minimum of x. To obtain a perfect density function, we utilized a normalizing constant.

Let $X \sim NPD(\theta, \mu, \sigma, \alpha)$ then the probability density function (pdf) and cumulative distribution function (cdf) for the NPD are

Table 9: *Estimated value of parameters, and model selection criteria for data for various distributions*

Model	Estimated Parameter	-LL	AIC	BIC	AICc
A Distribution (A)	$\hat{\alpha}$=125.662	107.950	217.901	219.335	218.039
Inverse Gompertz (IG)	$\hat{\alpha}$=1.249, $\hat{\beta}$=119.762	107.884	219.768	222.636	220.196
Kumaraswamy Inverse Gompertz (KuIG)	$\hat{\alpha}$=79.042, $\hat{\beta}$=18.694 $\hat{\gamma}$=26.554	103.988	213.976	218.278	214.865
Exponentiated Aradhana	$\hat{\alpha}$=19.1870, $\hat{\theta}$=0.2200	104.083	212.165	215.033	212.594
Inverse Weibull (IW)	$\hat{\alpha}$=446.1827, $\hat{\beta}$=4.655	105.323	214.647	217.515	215.075
Lomax Gumbel Type-Two (LGTT)	$\hat{\alpha}$=31.7086, $\hat{\beta}$=0.4549 $\hat{\theta}$=89.5227, $\hat{k}$=0.8379	104.818	217.636	223.372	219.174
Lomax-Gompertz (LomGo)	$\hat{\alpha}$=0.2952, $\hat{\beta}$=3.7704 $\hat{\theta}$=0.0005, $\hat{k}$=0.2523	105.729	219.457	225.193	220.996
Weighted Quasi Akash Distribution	$\hat{\theta}$=0.6152, $\hat{\alpha}$=3.7439 $\hat{\beta}$=16.9691	104.117	214.234	217.528	215.123
Three-parameter Weighted Lindley distribution (TWLD)	$\hat{\theta}$=0.6198, $\hat{\alpha}$=18.300 $\hat{\beta}$=16.9691	104.119	214.238	217.558	215.127
Weighted New quasi-Lindley	$\hat{\alpha}$=4.7687, $\hat{\theta}$=0.6146 $\hat{c}$=-16.9412	104.116	214.232	218.534	215.120
Harris Extended Generalized Exponential Distribution (HEGED)	$\hat{c}$=-0.1121, $\hat{\varphi}$=0.0761 $\hat{\lambda}$=0.1566, $\hat{\theta}$=5.0976	104.093	216.186	221.922	217.724
Marshall Olkin Extended Generalized Exponential Distribution (MOEGE)	$\hat{\varphi}$=0.0761, $\hat{\lambda}$=0.1566 $\hat{\theta}$=5.0976	105.776	217.552	221.854	218.441
Inverse Flexible Weibull (IFW)	$\hat{\alpha}$=61.167, $\hat{\beta}$=0.0859	104.963	213.927	216.795	214.355
Exponentiated Inverse Flexible Weibull (EIFW)	$\hat{\alpha}$=2.376,$\hat{\beta}$=0.164 $\hat{\gamma}$=81.51	104.141	214.282	218.584	215.171
Normal-Pareto (NPD)	$\hat{\theta}$=0.973, $\hat{\mu}$=29.6904 $\hat{\sigma}$=8.1176, $\hat{\alpha}$=152.232	101.167	210.334	216.069	211.872

$$f(x) = \frac{2\left(\frac{\theta}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} + (1-\theta)\frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}}\right)}{\left(erf\left(\frac{\sqrt{2}\mu-\sqrt{2}x_m}{2\sigma}\right)-1\right)\theta+2} \tag{2}$$

$$F(x) = \frac{x^{\alpha}\left(\theta\left(erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)+erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-2\right)+2\right)+2x_m^{\alpha}(\theta-1)}{\left(\left(erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-1\right)\theta+2\right)x^{\alpha}} \tag{3}$$

For $x_m \leq x, \alpha > 0, \mu \in \mathbb{R}, \sigma > 0, 0 \leq \theta \leq 1$. Where, x_m is the minimum of x.

The survival function of X is

$$S(x) = \frac{1}{\left(\left(erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-1\right)\theta+2\right)x^{\alpha}} \times x^{\alpha}\left(\left(\left(erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-1\right)\theta+2\right)-\left(\theta\left(erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)+erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-2\right)+2\right)\right) - 2x_m^{\alpha}(\theta-1) \tag{4}$$

The hazard function of X is

$$h(x) = \frac{2x^{\alpha}\left(\frac{\theta}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}+(1-\theta)\frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}}\right)}{x^{\alpha}\left(\left(\left(erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-1\right)\theta+2\right)-\left(\theta\left(erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)+erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-2\right)+2\right)\right)-2x_m^{\alpha}(\theta-1)} \tag{5}$$

Figure 5 displays the possible shapes of the pdf, cdf, sf, and HF of the Normal-Pareto distribution for the various parameter values.

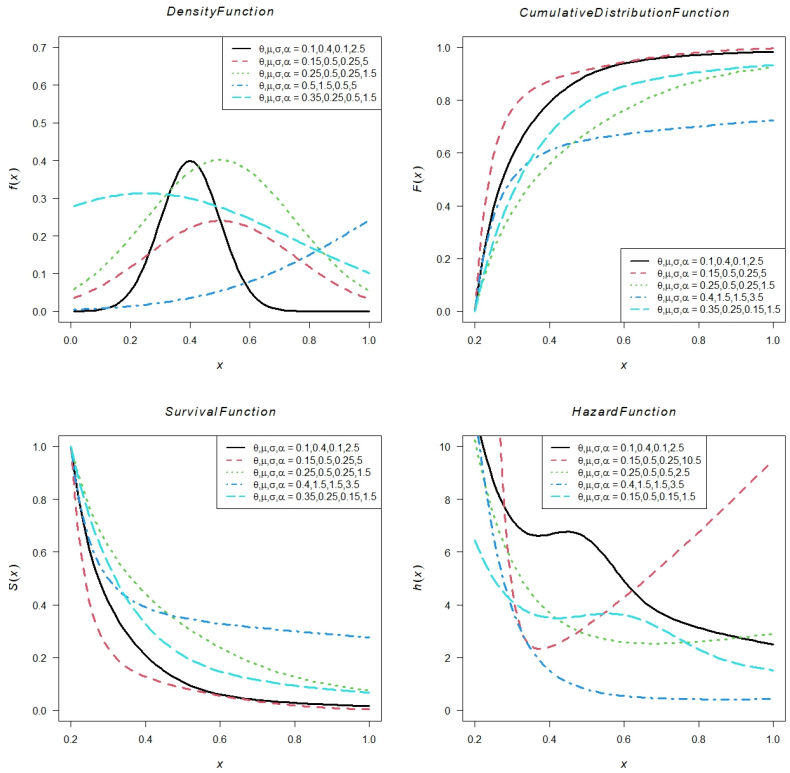

Figure 5: *The shapes of the pdf, cdf, sf, and HF of NPD for different values of the parameters*

The mean value can be calculated using the following equation.

$$Mean = \frac{2^{\frac{3}{2}}\sqrt{\pi}(\alpha\mu\theta-\mu\theta-\alpha x_m\theta+\alpha x_m)+(\alpha-1)\left(2\Gamma\left(1,\frac{(\mu-x_m)^2}{2\sigma^2}\right)\sigma-\sqrt{2}\Gamma\left(\frac{1}{2},\frac{(\mu-x_m)^2}{2\sigma^2}\right)\mu\right)\theta}{\sqrt{2}\sqrt{\pi}(\alpha-1)\left(\left(erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-1\right)\theta+2\right)} \tag{6}$$

The n^{th}-order statistic is given as

$$f_{X_{(n)}}(x) = n\left(\frac{2\left(\frac{\theta}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}+(1-\theta)\,\frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}}\right)}{\left(erf\left(\frac{\sqrt{2}\mu-\sqrt{2}x_m}{2\sigma}\right)-1\right)\theta+2}\right)\left[\frac{x^{\alpha}\left(\theta\left(erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)+erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-2\right)+2\right)+2x_m^{\alpha}(\theta-1)}{\left(\left(erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-1\right)\theta+2\right)x^{\alpha}}\right]^{(n-1)} \tag{7}$$

The first-order statistic is obtained as

$$f_{X_{(1)}}(x) = n\frac{2\left(\frac{\theta}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}+(1-\theta)\,\frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}}\right)}{\left(erf\left(\frac{\sqrt{2}\mu-\sqrt{2}x_m}{2\sigma}\right)-1\right)\theta+2}\times \left[\frac{x^{\alpha}\left(\left(\left(erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-1\right)\theta+2\right)-\left(\theta\left(erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)+erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-2\right)+2\right)\right)+2x_m^{\alpha}(\theta-1)}{\left(\left(erf\left(\frac{\mu-x_m}{\sqrt{2}\sigma}\right)-1\right)\theta+2\right)x^{\alpha}}\right]^{(n-1)} \tag{8}$$

We obtained the maximum likelihood estimator of parameters $(\theta,\mu,\sigma,\alpha)$ of the NPD. Consider the following log-likelihood function l of a random sample $X_1, X_2, \ldots, X_n$ from the density of NPD $(\theta,\mu,\sigma,\alpha)$ given in Equation (9).

$$l = n\ log2 - nlog\left(\theta\left(erf\left(\frac{\sqrt{2}\mu-\sqrt{2}x_m}{2\sigma}\right)-1\right)+2\right) + log\sum\left(\frac{\theta}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}+(1-\theta)\,\frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}}\right) \tag{9}$$

On differentiating Equation (9) with respect to the parameters θ, μ, σ, and α and equating to zero, we obtain the following likelihood equations.

$$\frac{\partial l}{\partial\theta} = \sum\left(\frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}-\sqrt{2\pi}\sigma\alpha x_m^{\alpha}x^{-\alpha-1}}{\sqrt{2\pi}\sigma\frac{\theta}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}+(1-\theta)\,\frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}}}\right) - \frac{n\left(erf\left(\frac{\sqrt{2}\mu-\sqrt{2}x_m}{2\sigma}\right)-1\right)}{\left(\theta\left(erf\left(\frac{\sqrt{2}\mu-\sqrt{2}x_m}{2\sigma}\right)-1\right)+2\right)} = 0 \tag{10}$$

$$\frac{\partial l}{\partial\mu} = \sum\left(\frac{\theta(x-\mu)e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2}\sqrt{\pi}\sigma^3\frac{\theta}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}+(1-\theta)\,\frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}}}\right) - \frac{\sqrt{2}n\theta e^{-\frac{(\mu-x_m)^2}{2\sigma^2}}}{\left(\theta\left(erf\left(\frac{\sqrt{2}\mu-\sqrt{2}x_m}{2\sigma}\right)-1\right)+2\right)\sqrt{\pi}\sigma} = 0 \tag{11}$$

$$\frac{\partial l}{\partial\sigma} = \sum\left(\frac{\theta(x-\mu)^2e^{\frac{-(x-\mu)^2}{2\sigma^2}}-\theta\sigma^2e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2}\sqrt{\pi}\sigma^4\frac{\theta}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}+(1-\theta)\,\frac{\alpha {x_m}^{\alpha}}{x^{\alpha+1}}}\right) + \frac{\left(\sqrt{2}\mu-\sqrt{2}x_m\right)n\theta e^{-\frac{(\mu-x_m)^2}{4\sigma^2}}}{\left(\theta\left(erf\left(\frac{\sqrt{2}\mu-\sqrt{2}x_m}{2\sigma}\right)-1\right)+2\right)\sqrt{\pi}\sigma^2} = 0 \tag{12}$$

and

$$\frac{\partial l}{\partial \alpha} = \sum \left(\frac{x_m^{\alpha} \ln(x_m)(1-\theta) x^{-\alpha-1}\alpha - x_m^{\alpha} \ln(x)(1-\theta) x^{-\alpha-1}\alpha + x_m^{\alpha}(1-\theta) x^{-\alpha-1}}{\frac{\theta}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} + (1-\theta)\frac{\alpha x_m{}^{\alpha}}{x^{\alpha+1}}} \right) = 0 \qquad (13)$$

Now the MLEs $\hat{\theta}, \hat{\mu}, \hat{\sigma}$,and $\hat{\alpha}$ of the parameters θ, μ, σ,and α of NPD can be obtained by solving the above four likelihood equations with the help of statistical software R.

5. Simulation Study

In this section, we evaluate the performance of ML estimates using a simulation study. For this purpose, we carry out a replication of 1000 times with various sample sizes ranging from 25 to 250 for the Normal-Pareto Distribution (NPD) parameters. We created a random sample of NPD using the Monte Carlo simulation method to generate the samples with the help of R programming. For each sample, we compute the mean value, average bias, and root-mean-square error (RMSE) to assess the performance of the MLEs, and these values are presented in Table 10.

From Table 10, it is observed that the sample size of n increases, and the bias and RMSE tend to decrease. Therefore, a larger sample size indicates more accurate results.

Table 10: *Simulation analysis: Mean, Bias, and RMSE values of NPD for various sample sizes*

		Case (i): θ=0.1, μ =0.5, σ =0.7, α=0.5			**Case (ii): θ=0.5, μ =1.5, σ =0.5, α=0.1**		
n	**Parameters**	**Mean**	**Average Bias**	**RMSE**	**Mean**	**Average Bias**	**RMSE**
25	θ	0.1355	0.0645	0.0907	0.4968	0.0732	0.1079
	μ	26.5063	0.1122	2.6756	6.3569	1.0943	2.3405
	σ	0.1943	0.1941	3.1157	0.0308	0.0303	1.8590
	α	1.0902	0.2871	0.5930	1.1431	0.1124	1.6121
50	θ	0.1062	0.0619	0.0396	0.4959	0.0610	0.0931
	μ	26.6213	0.0046	0.0021	6.2519	0.0575	0.0018
	σ	0.0002	$6.25e^{-06}$	$8.77e^{-05}$	0.0004	0.0001	0.0003
	α	1.0292	0.1007	0.1891	1.0482	0.1078	0.2333
75	θ	0.0997	0.0197	0.0378	0.4908	0.0608	0.0625
	μ	26.6190	0.0015	0.0018	6.2507	0.0011	0.0015
	σ	0.0002	$3.13e^{-06}$	$7.3e^{-05}$	0.0003	0.0003	0.0002
	α	1.0172	0.0374	0.1299	1.0273	0.1011	0.1993
100	θ	0.0985	0.0185	0.0341	0.4900	0.0151	0.0515
	μ	26.6192	0.0012	0.0011	6.2506	0.0009	0.0009
	σ	0.0002	$3.12e^{-06}$	$6.85e^{-05}$	0.0003	$9.5e^{-05}$	0.0001
	α	1.0166	0.0368	0.1246	1.0229	0.0663	0.1604
250	θ	0.0900	0.0004	0.0191	0.4902	0.0098	0.0325
	μ	26.6108	0.0006	0.0007	6.2506	0.0003	0.0007
	σ	0.0002	$6.15e^{-07}$	$5.84e^{-05}$	0.0003	4.92 e^{-05}	7.8 e^{-05}
	α	1.0084	0.0285	0.0748	1.0031	0.0248	0.0932

6. Conclusion

We developed an algorithm that is a comprehensive and specifically designed framework to select the most appropriate model for asymmetric data. This framework is based on a unique combination of probability distributions, which allows us to determine the best possible mixture of probability models. To ensure that our mixture model is better than existing models, we

have used various goodness of fit tests and information criteria. Our approach utilizes a finite mixture model, which combines multiple probability models. We used the maximum likelihood estimation method to estimate the parameters of NPD. To demonstrate the effectiveness of the algorithm, we conducted an experiment where we utilized real-time data and ran it through the algorithm. This enabled us to analyze the data and come up with a new appropriate model based on the finite mixture. To further test the efficiency of the proposed model, we compared it with other models used for the same data sets available in the literature. The algorithm proposed NPD model is found most suitable in this comparison study. Finally, we have obtained the statistical characteristics of our NPD model.

References

[1] Adeyemi, A. O., Adeleke, I. A., and Akarawak, E. E. (2022). Lomax Gumble types two distributions with applications to lifetime data. *International Journal of Statistics and applied mathematics*, 7(1): 36-45.

[2] Adewara, J. A., Adeyeye, J. S., Khaleel, M. A., and Aako, O. L. (2021). Exponentiated Gompertz Exponential (EGoE) distribution: Derivation, properties, and applications. *ISTATISTIK: Journal of the Turkish Statistical Association* , 13(1), 12-28.

[3] Adeyemi, A. O., Adeleke, I. A., and Akarawak, E. E. (2022). Lomax Gumbel types two distributions with applications to lifetime data. *International Journal of Statistics and Applied Mathematics*, 7(1), 36-45.

[4] Ahmad, Z., Hamedani, G. G., and Butt, N. S. (2019). Recent developments in distribution theory: A brief survey and some new generalized classes of distributions. *Pakistan Journal of Statistics and Operation Research*, 15(1), 87-110.

[5] Akaike, H. (1973). Information theory and an extension of the Maximum Likelihood Principle. *On Information Theory (Akademia Kiado, Budapest),* 267 -281.

[6] Alhyasat, K., Kamarulzaman, I., Al-Omari, A. I., and Abu Bakar, M. A. (2020). Power size biased distribution. *Statistics in Transition new series,* 21(3), 73-91.

[7] Alshenawy, R. (2020). A new one-parameter distribution: Properties and estimation with applications to complete and type II censored data. *Journal of Taibah University of Science,* 14(1), 11-18.

[8] Al-Talib, M., Al-Nasser, A., and Ciavolino, E. (2023). Gold distribution is another look at the generalization of Lindley distribution. *Pakistan Journal of Statistics and Operation Research,* 19(2), 241-256.

[9] Alzaatreh, A., Lee, C., and Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron,* 7(11), 63-79.

[10] Badmus, N. I., Amusa, S. O., and Ajiboye, Y. O. (2021). Weibull - Extended Pranav distribution: application to lifetime data sets. *Unilag Journal of Mathematics and Applications,* 1(1), 104-120.

[11] Bakoban. R.A. and Hanaa H. Abu-Zinadah. (2017).. The beta generalized inverted exponential distribution with real data applications. *REVSTAT- Statistical Journal,* 15(1), 65-88.

[12] Chesneau, C. (2017). A new family of distributions based on the hypo-exponential distribution with fitting reliability data. *HAL Archives.* HAL-01519350v5

[13] El-Gohary, A., El-Bassiouny, A.H. and El-Morshedy M. (2015). Inverse Flexible Weibull Extension Distribution. *International Journal of Computer Applications,* 115(2), 46-51.

[14] El-Morshedy, M., El-Bassiouny, A. H., and El-Gohary, A. (2017). Exponentiated Inverse Flexible Weibull Extension Distribution. *Journal of Statistics Applications & Probability- An International Journal,* 6(1), 169-183.

[15] El-Morshedy, M., El-Faheem, A. A., and El-Dawoody, M. (2020). Kumaraswamy inverse Gompertz distribution: Properties and engineering applications to complete, type II right censored and upper record data. *PLoS ONE,* 15(12).

[16] El-Morshedy, M., El-Faheem, A. A., Al-Bossly, A., and El-Dawoody, M. (2021). Exponentiated generalized inverted Gompertz distribution: Properties and estimation methods with applications to symmetric and asymmetric data. *Symmetry,* 13(10), 1868.

[17] Eyob, T., Shanker, R., Shukla, K. K., and Leonida, T. A. (2019). Weighted quasi-Akash distribution: Properties and applications. *American Journal of Mathematics and Statistics,* 9(1), 30-43.

[18] Fuller, E. R., Jr., Freiman, S. W., Quinn, J. B., Quinn, G. D., and Carter, W. C. (1994). *Fracture mechanics approach to the design of glass aircraft windows: A case study.* Proceedings of SPIE 2286, Window and Dome Technologies and Materials IV.

[19] Ganaie, R. A., and Rajagopalan, V. (2022). Exponentiated Aradhana distribution with properties and applications in engineering sciences. *Journal of Scientific Research,* 6(1), 316-325.

[20] Ganaie, R. A., and Rajagopalan, V. (2020). Weighted new quasi-Lindley distribution with properties and applications. *Journal of Xi'an University of Architecture & Technology,* 12(2), 561-575.

[21] Iwok, I. A., and Nwikpe, B. J. (2021). The Iwok-Nwikpe distribution: Statistical properties and application. *European Journal of Statistics and Probability,* 10(1), 33-48.

[22] Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society. Series B,* 20, 102-107.

[23] Okereke, E. W., and Uwaeme, O. R. (2018). Exponentiated Akash distribution and its applications. *Journal of the Nigerian Statistical Association,* 30.

[24] Omale, A., Asiribo, O. E., and Yahaya, A. (2019). On properties and applications of Lomax-Gompertz distribution. *Asian Journal of Probability and Statistics,* 3(2), 1-17.

[25] Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society Series A,* 185, 71-110.

[26] Sakthivel, K.M. and Vidhya, G. (2024). A Systematic Procedure for Modeling Asymmetrical Data. *Advances and Applications in Statistics,* 91(10), 1241-1260.

[27] Sakthivel, K.M. and Vidhya, G. (2024). Statistical Framework for Modeling Asymmetrical data with Dual Peaks. *Indian Journal of Science and Technology,* 17(27), 2829-2840.

[28] Shanker, R., Shukla, K.K., and Mishra, A. (2017). A Three-Parameter weighted Lindley distribution and its applications to model survival time data. *Statistics in Transition- New Series,* 18(2), 291-300.

[29] Sharma, V., Shanker, R., and Shanker, R. (2019). On some one parameter lifetime distributions and their applications. *Annals of Biostatistics & Biometric Applications,* 3(2).