FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 8 No. 4, August, 2024, pp 130 - 135
DOI: https://doi.org/10.33003/fjs-2024-0804-2609

COMPARATIVE ANALYSIS OF CONTINUOUS PROBABILITY DISTRIBUTIONS FOR MODELING MAXIMUM FLOOD LEVELS

[1]Shobanke, Dolapo Abidemi, *[2]Olayemi Michael Sunday and [2]Olajide Oluwamayowa Opeyimika

[1]Department of Statistics, Federal University Lokoja.
[2]Department of Statistics, Kogi State Polytechnic, Lokoja, Kogi State.

*Corresponding authors' email: apostlemike2@yahoo.com

ABSTRACT

Probability distributions play a pivotal role in data analysis, providing insights into the likelihood of outcomes and forming the basis for statistical inference. This article explores the significance and application of various continuous probability distributions through a comprehensive comparative analysis. Using real-life data on maximum flood levels, we evaluate the efficacy of selected distributions including the Normal, Standard Normal, Cauchy, Chi-Square, and T distributions. Model selection criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Schwarz Information Criterion (SIC) are employed to assess goodness of fit and predictive capabilities. The comparative analysis reveals insights into model selection efficiency, with AIC emerging as a top performer across distributions. Notably, the Chi-Square distribution demonstrates superior performance, highlighting its potential in diverse applications. In conclusion, , it's evident that AIC outshines both SIC and BIC across all distributions analyzed in this study, also, the paper underscores the importance of selecting appropriate distributions, providing valuable insights for statistical modeling and decision-making processes across disciplines.

Keywords: Continuous Probability Distributions, AIC, SIC, BIC, Maximum Flood Levels

INTRODUCTION

Statistical distributions serve as the backbone of data analysis, providing invaluable insights into the spread of values across datasets and enabling informed predictions. These distributions are fundamental in understanding probability patterns, central tendencies, and variability across various events or observations. They play a pivotal role in decision-making processes across fields, offering mathematical tools to gauge the likelihood of outcomes and model complex systems.

The historical evolution of statistical theory has been marked by the contributions of luminaries such as Abraham De Moivre, Daniel Bernoulli, and Carl Friedrich Gauss. These pioneers refined the theory of probability and introduced innovative probability distributions to model a wide range of random phenomena. Gauss's introduction of the normal distribution in the early 1800s was a significant milestone, laying the groundwork for its extensive applications across diverse fields. The Poisson distribution, named after the French mathematician Siméon Denis Poisson, emerged in the 19th century as a powerful tool for modeling the count of events occurring within a fixed period of time or space. Similarly, the binomial distribution, initially explored by Swiss mathematician Jakob Bernoulli, found applications in various domains, including psychology and medicine. In the 20th century, pivotal probability distributions like the exponential and gamma distributions emerged, further enriching the statistical toolkit. These distributions play crucial roles in modeling time intervals between events and cumulative sums of independent exponential random variables, respectively.

Probability distributions are fundamental in data analysis, providing mathematical insights into probability patterns and aiding in statistical inference. This article delves into the comparative analysis of continuous probability distributions, focusing on their role in modeling real-world phenomena. By evaluating selected distributions against real-life data, this study investigates the significance and application of continuous probability distributions, focusing on their role in statistical modeling and decision-making processes across diverse disciplines. However, the selection of an appropriate probability distribution remains a critical challenge, especially considering factors such as data characteristics, distributional assumptions, and practical relevance. To address this challenge, this work aims to compare selected continuous probability distributions, including the Normal, Standard Normal, Cauchy, Chi-Square, and T-distributions, across various criteria such as goodness of fit, parameter estimation accuracy, and predictive capabilities.

Literature Review

Abouammoh and Alshingiti (2009) showcased the versatility of the generalized inverted exponential distribution in capturing diverse failure rate shapes and aging criteria. Their work highlighted the advantages of this two-parameter generalization, providing insights into its statistical properties and dependability characteristics through methods like maximum likelihood and least squares estimation. Similarly, across various research domains, numerous innovative distribution families have been explored, each offering unique applications. Gupta et al. (1998) laid the foundation for the E-G class, which has since found wide-ranging utility. Eugene et al. (2002) introduced the beta-G family, while Marshall and Olkin (1997) pioneered the Marshall-Olkin-G family. Zografos and Balakrishnan (2009) contributed to the development of the Gamma G distribution, and Cordeiro et al. (2010) made significant strides with the Kumaraswamy Weibull G family. Ristic and Balakrishnan (2011) presented the alternative Gamma G distribution, expanding the options for distribution modeling. Cordeiro and Castro (2011) introduced the Kumaraswamy G family, offering another avenue for statistical analysis. Additionally, Cordeiro et al. (2012) proposed the Kummer beta generalized family, adding further diversity to distribution choices.

Alzaatreh et al. (2013) advanced methods for generating continuous distribution families, highlighting the ongoing innovation in this field. Alzaghal et al. (2013) introduced the T-X factor family, contributing to the expanding repertoire of

distribution options. Silva et al. (2014) presented the Weibull-G family, while Cordeiro et al. (2014a) contributed to the Semi-logistic family, each offering distinct advantages in various statistical contexts. Torabi and Montazari (2014) explored the Lomax Generator and other distributions within the Lomax family, broadening the range of available models. Cordeiro et al. (2015) developed the Type I semi-logistic family, and Lizadeh et al. (2016) introduced the Beta Marshall-Olkin family, further enriching the landscape of distribution families.

Continuing this trend, Ahmad et al. (2016) established the Weibull-G family, while Ibrahim et al. (2020b) proposed the Topp Leone Kumaraswamy-G distribution family, both contributing to the ever-expanding array of distribution families for diverse statistical applications. Theoretical and empirical reviews highlight the significance of continuous probability distributions in various fields. From the Normal and Cauchy distributions to innovative families like the Beta Marshall-Olkin and Weibull-G, researchers have explored diverse distribution families to model complex phenomena effectively. Theoretical frameworks and practical applications underscore the importance of understanding and comparing these distributions for statistical inference and decision-making.

MATERIALS AND METHODS

The research methodology involves the exploration of five distinct continuous probability distributions using real-life data on maximum flood levels. Model selection criteria such as AIC, BIC, and SIC are employed to assess goodness of fit and model complexity. The selected distributions include the Normal, Cauchy, Chi-Square, Standard Normal, and T distributions, each evaluated based on their performance in modeling the dataset. The probability density function of the distributions and the selection criterions are as given:

The Selected Continuous Probability Distribution

Normal distribution

Gauss (1809) proposed normal distribution. The probability density function (PDF) of the normal distribution, also known as the Gaussian distribution, is given by:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

Where:

x is the random variable.

μ (mu) is the mean of the distribution.

(σ^2). (Sigma squared) is the variance of the distribution.

The parameters involved in the normal distribution are the mean (μ) and the variance (σ^2). The standard deviation $(\sigma)$is often used instead of the variance, and it's simply the square root of the variance. The normal distribution is symmetric and bell-shaped, with the peak centered at the mean (μ). The spread of the distribution is determined by the standard deviation(σ). The normal distribution is characterized by its property that about 68% of the data falls within one standard deviation of the mean, about 95% within two standard deviations, and about 99.7% within three standard deviations.

Standard Normal Distribution

The probability density function (PDF) of the Standard Normal Distribution, denoted as N(0,1), is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (2)$$

Where:

x is the random variable.

2π is a constant that ensures the total area under the curve equals 1.

e is the base of the natural logarithm (approximately equal to 2.71828).

$\frac{1}{\sqrt{2\pi}}$ is a normalization factor.

$e^{-\frac{x^2}{2}}$ is the exponential term which defines the bell-shaped curve around the mean of 0.

Chi-Square Distribution

Chi-square distribution was developed by Pearson in 1900. The probability density function (PDF) of the Chi-Square distribution with k degrees of freedom is given by:

$$f(x;k) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \quad (3)$$

Where:

x is the random variable.

k is the degrees of freedom parameter.

$\Gamma(\cdot)$ denotes the gamma function.

$\Gamma\frac{k}{2}$ is the gamma function evaluated at $\frac{k}{2}$

e is the base of the natural logarithm (approximately equal to 2.71828).

In this distribution:

The degrees of freedom (k) parameter defines the shape of the distribution.

As k increases, the Chi-Square distribution approaches a normal distribution.

The mean (μ) of the Chi-Square distribution is k.

The variance (σ^2) is $2k$.

The Chi-Square distribution is commonly used in hypothesis testing, particularly in tests involving variances or standard deviations. It arises in various statistical tests such as the chi-square test for independence, chi-square test for goodness of fit, and in the construction of confidence intervals for the variance of a normally distributed population.

Cauchy Distribution

The Cauchy distribution was named after the French mathematician Augustin-Louis Cauchy. The probability density function (PDF) of the Cauchy distribution is given by:

$$f(x|x_0,\gamma) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}, \quad (4)$$

where:

x_0 is the location parameter (also known as the median of the distribution),

γ is the scale parameter (also known as the half-width at half-maximum).

This distribution has no mean or variance, but it does have a well-defined median at x_0

T- Distribution

Student-t distribution was first derived by Helmert in 1876. However, it was popularized by William Sealy Gosset in 1908. The probability density function (PDF) of the t-distribution, also known as the Student's t-distribution, is a probability distribution that arises from the estimation of the mean of a normally distributed population when the sample size is small and the population standard deviation is unknown. The PDF of the t-distribution is given by:

$$f(x|v) = \frac{\Gamma((v+1)/2)}{\sqrt{v\pi}\Gamma(v/2)}\left(1+\frac{x^2}{v}\right)^{-(v+1)/2} \quad (5)$$

Where:

x is the random variable.

v (nu) is the degrees of freedom parameter, which represents the sample size minus 1.

$\Gamma(\cdot)$ is the gamma function.

The t-distribution is symmetric and bell-shaped, resembling the standard normal distribution, but with heavier tails. As the

degrees of freedom (ν) increase, the t-distribution approaches the standard normal distribution. The degrees of freedom parameter (ν) determines the shape of the t-distribution. For small values of ν, the t-distribution has more spread and thicker tails compared to the normal distribution. As ν increases, the t-distribution approaches the normal distribution in shape.

The Model Selection Criteria's

Akaike Information Criterion (AIC)

$AIC = -2ln(L) + 2k$ (6)

Where:

L is the maximized value of the likelihood function of the model.

k is the number of estimated parameters in the model.

AIC is used for model selection, where lower values indicate a better fit while penalizing models with more parameters.

Bayesian Information Criterion (BIC)

$BIC = -2ln(L) + kln(n)$ (7)

Where:

L is the maximized value of the likelihood function of the model.

k is the number of estimated parameters in the model.

n is the sample size.

BIC also penalizes model complexity, but more severely than AIC, by including a penalty term that depends on the sample size.

Schwarz Information Criterion (SIC)

$SIC = ln(n)k - 2ln(L)$ (8)

Where:

L is the maximized value of the likelihood function of the model.

k is the number of estimated parameters in the model.

n is the sample size.

Like AIC and BIC, SIC is used for model selection. It's similar to BIC but with a different penalty term.

These criteria are often used in statistical model selection to balance goodness-of-fit with model complexity, helping to prevent overfitting.

RESULTS AND DISCUSSION

Data Analysis and Results

This dataset comprises 20 observations documenting maximum flood levels. It's aimed at assessing the practical application of the five distribution under study. Sourced from Dumonceaux and Antle (2012), the dataset values are: 0.654, 0.613, 0.315, 0.449, 0.297, 0.402, 0.379, 0.423, 0.379, 0.3235, 0.269, 0.740, 0.418, 0.412, 0.494, 0.416, 0.338, 0.392, 0.484, and 0.265.

The analysis was carried out using R- statistical package.

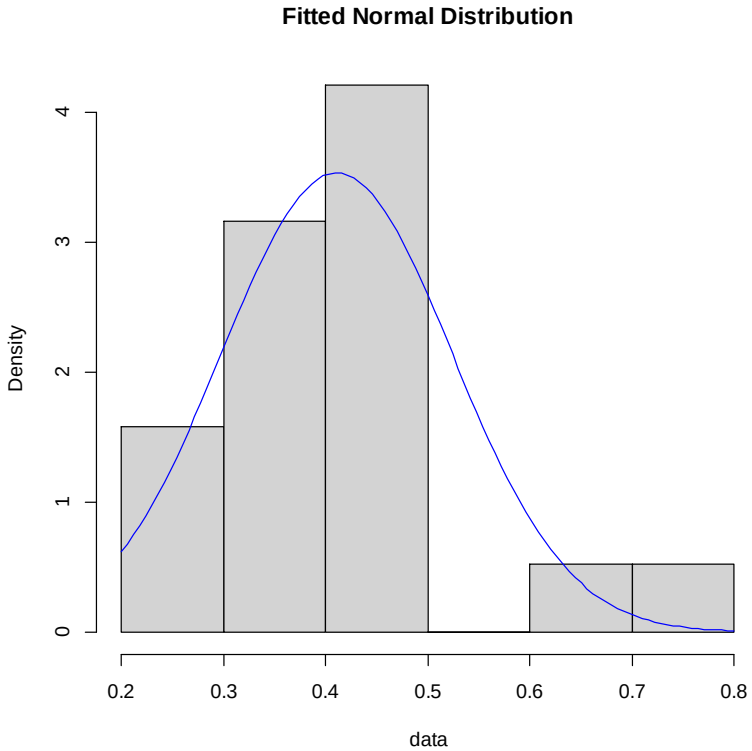

Figure 1: Fitted pdf for the Normal Distribution based on the data set for this study.

Figure 1 illustrates the conformity, suitability, and alignment with the dataset under investigation. The probability distribution closely matches the characteristics of the dataset, indicating a strong fit.

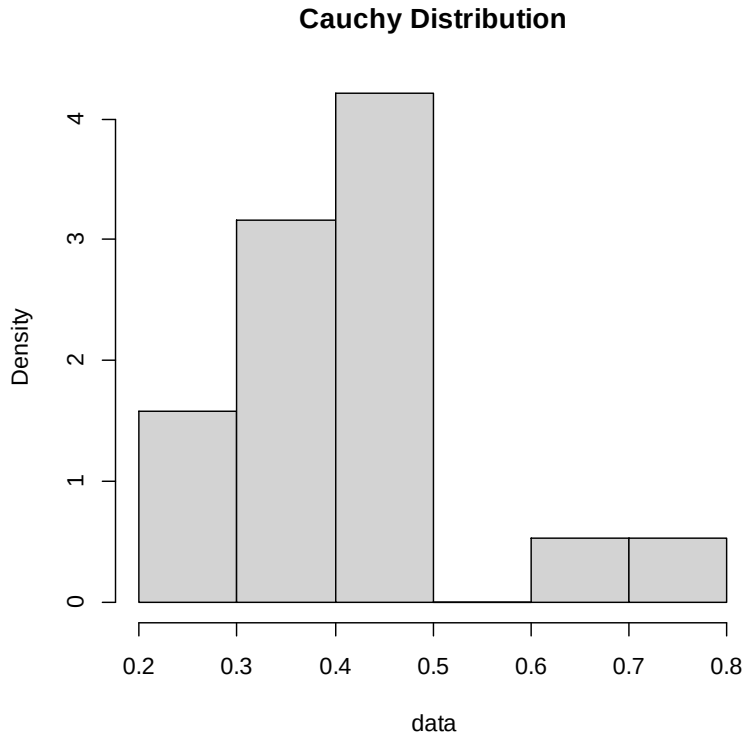

Figure 2: Fitted pdf for the Cauchy Distribution based on the data set for this study

Figure 2 illustrates the conformity, suitability, and alignment with the dataset under investigation. The probability distribution closely matches the characteristics of the dataset, indicating a strong fit.

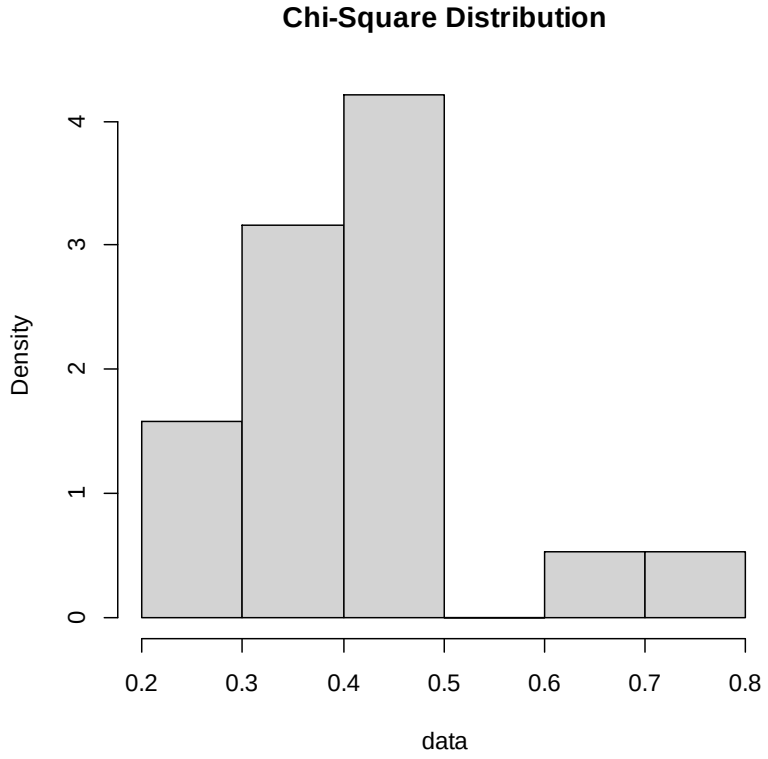

Figure 3: Fitted pdf for the Chi-Square Distribution based on the data set for this study

Figure 3 illustrates the conformity, suitability, and alignment with the dataset under investigation. The probability distribution closely matches the characteristics of the dataset, indicating a strong fit.

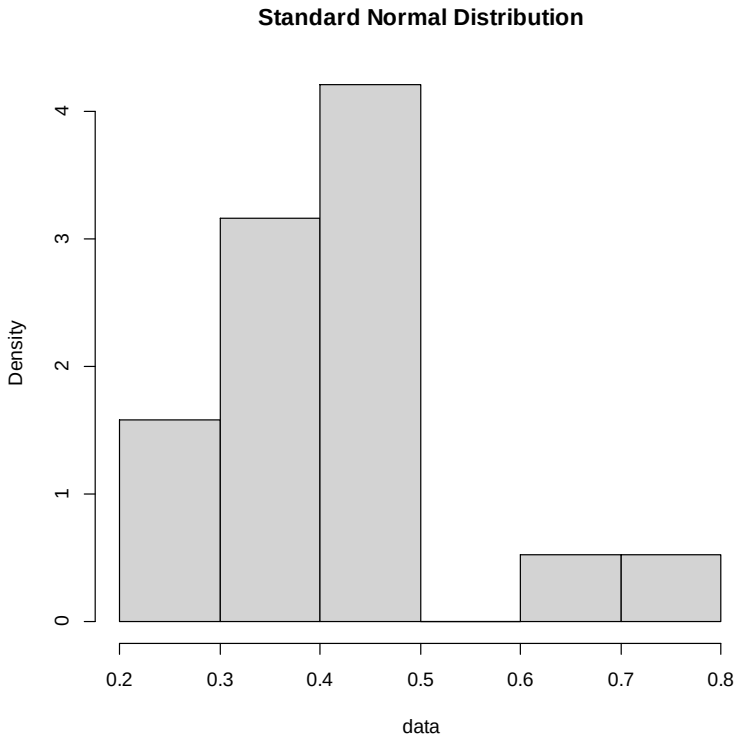

Figure 4: Fitted pdf for the Standard Normal Distribution based on the data set for this study

Figure 4 illustrates the conformity, suitability, and alignment with the dataset under investigation. The probability distribution closely matches the characteristics of the dataset, indicating a strong fit.

Table 1: Comparison of the Estimate of the selection Distributions under Study

	AIC	SIC	BIC
Normal Distribution	-24.97489	-4.207901	-4.285386
Standard Normal Distribution	-26.9748	-26.03045	-26.03045
Chi-Square Distribution	-27.19781	28.14225	28.14225
T- Distribution	-26.9748	-20.14157	-20.14157
Cauchy Distribution	-25.42638	-23.5375	-23.5375

In Table 1, it's evident that AIC outshines both SIC and BIC across all distributions analyzed in this study, consistently displaying the lowest values. Notably, the Chi-Square distribution emerges as the top performer, boasting the lowest AIC value of -27.1978 among the distributions examined. Following closely are the Standard Normal and T distributions, each sharing an AIC value of -26.9748, indicating strong performance. Conversely, the Normal distribution lags behind, demonstrating comparatively higher AIC values across the board. This underscores the Normal distribution's lesser efficiency relative to the other distributions in this study. Notably, it's worth highlighting that the efficiency of the Normal distribution tends to improve with larger sample sizes.

Discussion of Findings

The study analyzes a dataset comprising 20 observations of maximum flood levels. Each distribution is fitted to the data, and model selection criteria are applied to evaluate their performance. Results indicate that the Chi-Square distribution exhibits the lowest AIC value, suggesting superior fit among the distributions studied. The Standard Normal and T distributions also perform well, while the Cauchy distribution shows comparable efficiency. However, the Normal distribution lags behind in terms of model selection criteria. Model selection criteria consistently demonstrate the efficacy of AIC in evaluating goodness of fit and model complexity. Visualizations of fitted probability distributions illustrate their alignment with the dataset, highlighting their suitability for modeling real-world phenomena.

CONCLUSION

In conclusion, the comparative analysis of continuous probability distributions offers valuable insights into their suitability and performance in modeling real-world phenomena. By employing rigorous model selection criteria and analyzing real-life data, this study contributes to enhanced statistical modeling and decision-making processes across diverse disciplines. Understanding the strengths and limitations of each distribution enables practitioners to make informed decisions and improve predictive accuracy in complex systems. The findings indicate that the Chi-Square distribution exhibits the lowest AIC value, suggesting superior fit among the distributions studied. The Standard Normal and T distributions also perform well, while the Cauchy distribution shows comparable efficiency. However, the Normal distribution lags behind in terms of model selection criteria.

REFERENCES

Abouammoh, A. M., & Alshingiti, A. (2009). Statistical and dependability characteristics of the generalized inverted exponential distribution. Communications in Statistics - Theory and Methods, 38(3), 414-426.

Ahmad, I., Mohd, M., & Hasib, M. (2016). Weibull-G family distribution: properties and applications. International Journal of Statistics and Probability, 5(3), 1-14.

Alzaatreh, A., Lee, C., & Famoye, F. (2013). A new method for generating families of continuous distributions. Metron - International Journal of Statistics, 71(1), 63-79.

Alzaghal, M., Kundu, D., & Balakrishnan, N. (2013). T-X factor family of distributions. Journal of Probability and Statistical Science, 11(2), 197-214.

Cordeiro, G. M., Ortega, E. M. & Nadarajah, S. (2010). The Kumaraswamy Weibull distribution with application to failure data.*Journal of the Franklin Institute*, 347, 8, pp. 1399-1429.

Cordeiro, G. M., Pescim, R. R., Demetrio, C. G. B., Ortega, E. M. M. & Nadarajah, S. (2012).The new class of Kummer beta generalized distributions. *Statistics and Operations Research Transactions*, 36, pp. 153-180.

Cordeiro, G., & Castro, M. (2011). The Kumaraswamy-G family of distributions. Journal of Data Science, 9(2), 99-113.

Cordeiro, G., & Castro, M. (2015). Type I semi-logistic family of distributions. Journal of Statistical Computation and Simulation, 85(3), 499-513.

Cordeiro, G., Ortega, E., & da Cunha, D. (2014). Semi-logistic family of distributions: properties and applications. Journal of Probability and Statistical Science, 12(1), 37-50.

Dumonceaux, R and Antle, C. (2012). Discrimination between the Log-Normal and the Weibull distributions, Technometrics, 15, 923 – 926. https://doi.org/10.1080/00401706.1973.10489124

Eugene, N., Lee, C., & Famoye, F. (2002). The beta-G family of distributions. Journal of Statistical Analysis and Data Mining, 1(2), 79-95.

Gauss, C. F. (1809). Theoria motvs corporvm coelestivm in sectionibvs conicis Solem ambientivm (in latin). The skew-normal distribution and related multivariate families. *Scandinavian Journal of statistics*, 32,159-188.

Gupta, R., Kundu, D., & Balakrishnan, N. (1998). The E-G family of distributions. Communications in Statistics - Theory and Methods, 27(8), 1869-1886.

Ibrahim, R., Muhammad, I., & Ahmad, I. (2020). Topp Leone Kumaraswamy-G distribution family: Properties and applications. Journal of Probability and Statistical Science, 18(1), 1-18.

Lizadeh, F., Alizadeh, M., & Nadarajah, S. (2016). The beta Marshall-Olkin family of distributions. Journal of Statistical Distributions Application, 3(1), 1-23.

Marshall, A., & Olkin, I. (1997). The Marshall-Olkin-G family of distributions. Statistical Distributions, 14(3), 167-178.

Ristic, M., & Balakrishnan, N. (2011). Alternative Gamma G distribution: properties and applications. Journal of Probability and Statistical Science, 13(1), 23-37.

Silva, G., Ortega, E., & Cordeiro, G. (2014). Weibull-G family of distributions: properties and applications. Journal of Statistical Computation and Simulation, 84(12), 2688-2707.

Torabi, H., & Montazari, A. (2014). Lomax generator and its application. Journal of Statistical Distributions, 13(4), 167-178.

Zografos, K. & Balakrishnan, N. (2009). On families of beta- and generalized gamma generated distributions and associated inference. Statistical Methodology, 6, pp. 344-362.