A STUDY OF PROPERTIES ON GENERALIZED BETA AND MIXTURE OF TWO MODIFIED LOG-NORMAL DISTRIBUTIONS

By

DENNIS NG WEN WEI

A dissertation submitted to the
Department of Mathematical and Actuarial Sciences,
Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman,
in partial fulfilment of the requirements for the degree of
Master of Science
March 2019

ABSTRACT

A STUDY OF PROPERTIES ON GENERALIZED BETA AND MIXTURE OF TWO MODIFIED LOG-NORMAL DISTRIBUTIONS

Dennis Ng Wen Wei

The objective of this research is to study the statistical properties of two newly proposed distributions which are the generalized Beta distribution from the Beta family and the mixture of 2 modified Log-Normal distributions from the Skew Normal family (Chuah, 2016). Properties such as the moment generating function are derived for the two mentioned distributions. The advantages of the proposed distributions are their versatility and flexibility where they could provide a good description to various data with properties such as unimodal/uniantimodal increasing, decreasing, bath-tub shape distributions and etc. Other distributions such as Beta, Gauss Hypergeometric, Exponential and Gamma distributions are selected to compare their fitting ability with the proposed distributions. An empirical study is performed using simulated data and real rainfall volume data collected from Sungai Lui (river) with Maximum Likelihood Estimation (MLE) as the parameter estimation method. Model selection criteria such as Kolmogorov-Smirnov K-S test, Akaike's Information Criteria (AIC) and Root Mean Square Error (RMSE) are used to identify the better fitted model in this study. The empirical results show that the proposed mixture is the better fit in its Skew Normal family while the proposed generalized Beta is the worst performed in its Beta family.

ACKNOWLEDGEMENTS

To begin with, I, Dennis Ng Wen Wei, would like to express my deepest gratitude to my supervisor, Dr. Koh Siew Khew and my co-supervisor, Dr. Sim Shin Zhu for all of the guidance, help and advice given to me for the entire duration of my research. I would like to thank them for all of the comments and suggestions given to me when I intended to publish a conference proceeding as well as writing up my dissertation where they have helped in improving my academic writing skills. I have learned a lot in understanding that the flow and the main points of the research are essential for a good publication.

Besides that, I would also like to express my upmost gratitude to my former supervisor, Dr. Lee Min Cherng for the opportunity given to me in experiencing a Statistics related research. I am truly thankful for all of the advice, guidance, and comments that was shared to me where I am able to acquire a much better understanding in Statistics and also the statistical software. Through his guidance, lots of knowledge and ideas were gain throughout the whole academic journey. Also, I would like to thank Universiti Tunku Abdul Rahman (UTAR) for providing me a chance to pursue my Masters studies as well as offering me with a staff scholarship for my programme.

Not to forget Mr. Chuah Hock Lung, for giving me the opportunity to continue his research and sharing with me the challenges faced during his journey. I am grateful for all of the references and guidance that was

shared to me as they played a very important role in solving major issues in my research. To my parents and family members, I would like to thank for giving me the opportunity to further my studies which is essential in achieving my goals in life. Last but not least, I would also want to express my gratitude towards all of my colleagues and friends for their kindness in sharing their opinions and knowledge in different aspects of life that guided me socially, academically as well as professionally.

APPROVAL SHEET

This dissertation entitled **"A STUDY OF PROPERTIES ON GENERALIZED BETA AND MIXTURE OF TWO MODIFIED LOG-NORMAL DISTRIBUTIONS"** was prepared by DENNIS NG WEN WEI and submitted as partial fulfilment of the requirements for the degree of Master of Science at Universiti Tunku Abdul Rahman.

Approved by:

(Dr. KOH SIEW KHEW) Date: .....
Assistant Professor/Supervisor
Department of Mathematical and Actuarial Sciences
Lee Kong Chian Faculty of Engineering and Sciences
Universiti Tunku Abdul Rahman

(Dr. SIM SHIN ZHU) Date: .....
Assistant Professor/Co-supervisor
Department of Mathematical and Actuarial Sciences
Lee Kong Chian Faculty of Engineering and Sciences
Universiti Tunku Abdul Rahman

LEE KONG CHIAN FACULTY OF ENGINEERING AND SCIENCE

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 01 MARCH 2018

SUBMISSION OF DISSERTATION

It is hereby certified that **Dennis Ng Wen Wei** (ID No: **16UEM07640**) has completed this dissertation entitled "A STUDY OF PROPERTIES ON GENERALIZED BETA AND MIXTURE OF TWO MODIFIED LOG-NORMAL DISTRIBUTIONS" under the supervision of Dr. Koh Siew Khew (Supervisor) from the Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, and Dr. Sim Shin Zhu (Co-Supervisor) from the Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science.

I understand that the University will upload a softcopy of my dissertation in PDF format into UTAR Institutional Repository, which may be accessible to UTAR community and public.

Yours truly,

_____

(Dennis Ng Wen Wei)

DECLARATION

I, Dennis Ng Wen Wei hereby declare that the dissertation is based on my original work except for quotations and citation which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

_____

(DENNIS NG WEN WEI)

Date: _____

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike's Information Criteria
BIC	Bayesian Information Criteria
CDF	Cumulative Distribution Function
IID	independent and identically distributed
K-S	Kolmogorov-Smirnov Test
MAE	Mean Absolute Error
MGF	Moment Generating Function
MLE	Maximum Likelihood Estimation
PDF	Probability Density Function
RMSE	Root Mean Square Error
VaR	Value at Risk

CHAPTER 1

INTRODUCTION

Continuous models enable researches to visualise, analyse and make predictions on the sample easily; see Johnson et al. (1994, 1995). Hundreds of continuous distributions are discovered and more new models with various applications are still being studied. Families of the continuous univariate and multivariate distributions have been examined by many researchers where they could be applied to many different field of studies. For example, application of new Beta-type models towards various fields could be seen from McDonald and Xu (1995), Chotikapanich et al. (2007) and Lima et al. (2016). For the Skew Normal family distributions, see Woolhiser and Roldan (1982), Cho et al. (2004) and Suhaila et al. (2011).

Three new continuous distributions were proposed by Chuah (2016) to study the frequency of rainfall volume data. The proposed distributions include two Beta-type distributions and a mixture distribution where they are generalized versions of various statistical distributions known. The first Beta-type distribution introduced is the generalized Beta distribution consisting of 6 parameters. It could be reduced to the Kumaraswamy (1980), McDonald (1984)'s generalized Beta of the 1st kind, Armero and Bayarri (1994)'s Gauss Hypergeometric and also arcsine distributions. The second Beta-type distribution is the modified Beta distribution consisting of 5 parameters. It can also be related to the Beta distribution and other Beta-type distributions mentioned above. Meanwhile, the third proposed distribution is the mixture of 2 modified Log-Normal distributions with 7

parameters from the Skew Normal family distributions. The proposed mixture distribution could be reduced to the mixture of 2 Log-Normal, modified Log-Normal and Log-Normal distributions.

The main difference between the two Beta-type distributions and the mixture distribution is the constraint present within the x-variable. The Beta-type distributions have a variable constraint between 0 and 1 while Skew Normal family distributions require it to be greater than zero. These distributions were proposed due to the large number of parameters present which provides a lot of flexibility in fitting various shapes of data. The flexibility of the distributions will be discussed later.

The proposed distributions are applied to rainfall data from the Langat River, Selangor and Chuah (2016) has compared the fittings with some well known existing distributions. It is concluded that the mixture of 2 Log-Normal distribution has the best fit among the distributions in his study. Unfortunately, the fitting of the proposed mixture of 2 modified Log-Normal distributions was unable to be examined as Chuah (2016) failed to compute the maximum likelihood estimator for this distribution due to the large number of parameters present. It was then presumed that the proposed mixture distribution will be a good fit to the data as well because it is a general form to the mixture of 2 Log-Normal distribution. This presumption was supported through the fitting results of the proposed generalized Beta distribution where it tends to fit better than its sub-distributions.

Among the three distributions proposed, the general properties of the first Beta-type (generalized Beta) distribution and the mixture of 2 modified Log-Normal distributions are yet to be derived. Only the modified Beta distribution's properties was successfully derived by Chuah (2016). In addition, the failure of computing the maximum likelihood estimator for the mixture of 2 modified Log-Normal distributions is of concerned because it has the potential to be a suitable model to characterize rainfall data from the Langat River. Clearly, further work is needed in order to have a better understanding on the distributions in fitting the data.

The objective of our study is to explore the properties and the application of the two underived proposed (generalized Beta and mixture of 2 modified Log-Normal) distributions. The distributions' flexibility will also be discussed and properties such as the cumulative distribution function (CDF), expected value, moment generating function (MGF), variance, skewness and kurtosis will be derived. A conference proceeding regarding the study of properties on generalized Beta distribution which is about the derivations of the generalized Beta distribution's properties has been accepted for publication (Ng et al., 2018).

The popular maximum likelihood approach will be used for the parameter estimation due to its generality and asymptotic efficiency while model selection criteria such as Kolmogorov-Smirnov (K-S) test, Akaike information criterion (AIC) and root mean square error (RMSE) will be applied to compare the performance between the selected models. The problem of computing the maximum likelihood estimator for the proposed

mixture distribution faced by Chuah (2016) is solved using a certain statistical software. It will be presented in Chapter 4. The distributions will be fitted to rainfall volume dataset collected from the Langat river basin but from a different reservoir as compared to Chuah (2016). Besides that, comparisons and discussions will be done to see which model has a better fit and to check whether the result supports the conclusion made by Chuah (2016). A summary on the differences in methodologies used by Chuah (2016) and this research is shown in Table 1.1 below.

Table 1.1: Summary of Research Studies

Study	Chuah (2016)	This Study
Dataset	Data collected is on rainfall volume from Sg. Lui but from different reservoirs.	
Theoretical	Modified Beta distribution's properties is derived	An extended study on generalized Beta and mixture of 2 modified Log-Normal distributions' properties is conducted.
Empirical	Comparison of various models including the proposed distributions across the Beta and Skew Normal family were conducted using the same data massaging and transformation method for both families. Mixture of 2 Log-Normal distribution is concluded to be the best fitted model.	Comparison of 3 distributions within the Beta family and 3 distributions within the Skew Normal family were conducted. Two different data massaging and transformation methods were applied to both families separately.

In this research, aside from the introduction that was mentioned in Chapter 1, literature review will be discussed in Chapter 2. The proposed distributions will be introduced in Chapter 3 where some of the properties for the two distributions are derived. The data collected and methodology used in this research will be explained in Chapter 4. In Chapter 5, the empirical results will be discussed and comparisons between the various models selected will be presented using model selection criteria. Finally, conclusions and comments on future research work will be presented in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

2.1 Background

Although present distributions are commonly used in various field of researches, there is still a drawback when it comes to flexibility where they are unable to be a good representation as a model towards certain fields. To overcome this, the method of generalization is introduced. Tahir et al. (2015) mentioned that proposing generalized models attracted theoretical and applied statisticians due to their flexible properties and the generalization of the distributions are done either to provide a physical or statistical argument in order to explain a generated data mechanism, was an appropriate model that has been used successfully before or was a model that has been proved to fit empirically well towards a dataset.

In this chapter, the Beta family distribution will be discussed in Section 2.2 where distributions nested to the proposed generalized Beta distribution will be introduced and various forms of generalized Beta distributions will be briefly discussed. Whereas, the Skew Normal family continuous distributions will be discussed in Section 2.3. Distributions related to the proposed mixture of 2 modified Log-Normal distributions such as the Skew Normal distribution will be reviewed. In Section 2.4, the application of various distributions towards a wide range of studies will be presented. In addition, different parameter estimation methods will be

stated in Section 2.5 and finally various types of model evaluation metrics as well as the selection criteria will be presented in Section 2.6.

2.2 Beta Family Distributions

In probability theory and statistics, Beta family distributions are continuous distributions that are usually bounded by the interval (0,1) and the parameters that appear as exponents of the random variable actually control the shape of the distribution. Beta family distributions with finite range gives an advantage in fitting a dataset because Beta PDFs are versatile where it provides the advantage of modelling various forms of uncertainties since they can produce increasing, decreasing, unimodal, uniantimodal or even uniform shapes where it depends on it's γ and q parameters' values (Johnson et al., 1996).

Nonetheless, the 2-parameter Beta distribution is undesirable in certain ways because it's precision is limited and has not much flexibility in fitting certain types of data. Jacob (2013) even mentioned that the application of the classical Beta distribution is limited due to its inability to fit the data very well. He explained that the flexibility of the model could be enhanced by adding more parameters to it. Hence, it is preferred to have Beta models that are more flexible parametrically in order to provide richer empirical descriptions about the data and also providing more structures instead of a non-parametric estimator (Chuah, 2016).

Various forms of generalized Beta distributions such as McDonald and Xu (1995), Chotikapanich et al. (2007), Alexander et al. (2012) and other related studies were developed in many studies in order to increase the flexibility of the model in fitting the data. In Jacob (2013) research, the development of the classical distribution with (0,1) and $(0,\infty)$ domains to 3, 4 and 5-parameter generalized Beta distributions and other Beta-type models stated in the mentioned references through various methods were derived and compiled. For instance, distributions developed from the Beta generated distributions such as Beta Gamma, Beta Pareto and Beta Rayleigh distributions as well as Beta distributions constructed from special functions such as Beta Bessel, generalized Beta and Gauss Hypergeometric distributions were shown in his research.

Although additional parameters were added to the classical Beta distribution to increase it's flexibility, it might not be enough to fit certain shapes of data. During the fitting of the developed Beta distributions to family income data by McDonald and Xu (1995), it was found out that the 4-parameter generalized Beta was the better fit and not the 5-parameter model based on the likelihood and other model selection criteria. Therefore, Jacob (2013) mentioned that the development of the 5-parameter generalized Beta distribution does not provide additional flexibility compared to the 4-parameter distribution.

Due to the lack of additional flexibility in the 5-parameter generalized Beta distribution, Chuah (2016) then proposed a 6-parameter generalized Beta distribution. This generalized Beta distribution is expanded from the generalized Gauss Hypergeometric function where the development was

based on the special functions method mentioned above. The 6-parameter generalized Beta distribution is an extension of the 5-parameter model under the (0,1) domain with the addition of the ${}_2F_1$ and ${}_3F_2$ Hypergeometric functions. The presence of the Hypergeometric functions provide the advantage of modelling complex numbers as well. Hence, it would be of great interest to study the performance of this improved flexible proposed generalized Beta distribution consisting of 6 parameters ($\gamma, \rho, \beta, \alpha, \sigma$ and z) in fitting the data. In addition, the 6-parameter generalized Beta distribution can be reduced to various Beta family distributions by setting it's parameters with certain values as follows:

1. Kumaraswamy distribution (Kumaraswamy, 1980)
   $z = 0$; $\alpha = \gamma$; $\rho - \beta = q$; $b = 1$ and $\gamma = 1$

2. Generalized Beta of the 1st Kind distribution (McDonald, 1984)
   $z = 0$; $\alpha = \gamma$ and $\rho - \beta = q$

3. Gauss Hypergeometric distribution (Armero and Bayarri, 1994)
   $\alpha = \gamma$; $\rho = \beta + \theta$ and $z = -t$

4. Standard Arcsine distribution
   $z = 0$; $\alpha = 0$; $\beta = 0$; $\gamma = 0.5$ and $\rho = 0.5$

2.3 Skew Normal Family Distributions

Skew Normal distribution introduced by Azzalini (1985) is a strongly unimodal distribution with properties of "strict inclusion" for normal density and mathematical tractability. It was derived from the Normal distribution by multiplying the standard normal PDF with the CDF. Martínez-Flórez et al. (2013) mentioned that the proposal of the Skew

Normal distribution was to conform data that has a range of asymmetry and kurtosis that are out of the range allowed by the normal distribution. Thus, it was also mentioned that Lin and Stoyanov (2009) presented the Log-Skew Normal or modified Log-Normal distribution to conform data with asymmetry and kurtosis that are out of the range allowed by the Log-Normal distribution.

The presence of the logarithmic function, (ln) provides the advantage for the distribution to cater data of various shapes including a symmetrical one. It is suitable in modelling continuous random variables that are greater than zero or data that appears to be more or less skewed. Skewness occurs when average are low, variances are large and the values cannot be negative (Limpert et al., 2001). These concepts are then applied to the Skew Normal distribution which results in the modified Log-Normal distribution. The PDF of the distribution is as follows:

$$f(x) = \frac{2}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \times \int_0^{(\frac{x}{e^{\mu}})^{\frac{(c)}{\sigma}}} \frac{1}{\sqrt{2\pi}t} e^{-\frac{(\ln t)^2}{2}} dt$$

$$0 < x < \infty$$

$$\sigma > 0,\ \mu \,\epsilon\, \Re,\ -1 \leq c \leq 1$$

However, Wirjanto and Xu (2009) stated that Gaussian related distributions might exhibit substantial leptokurtosis, also known as fat tails and asymmetry around the mean. Thus, they suggested the usage of a more flexible model such as the mixture distribution to accommodate this stylized fact. This suggestion was proposed because McDonald and Butler (1987) explained that mixture distributions provide a framework for models where a random variable of a distribution has a particular form

which provide an approach to model randomized or heterogeneous data as well as a rationale for some thick-tailed distributions.

In regards to the concerns of skewness, asymmetry and kurtosis as well as flexibility to model positive value data, we are therefore interested to further study the mixture of 2 modified Log-Normal distributions proposed by Chuah (2016). The mixture distribution is closely related to the Skew Normal distribution introduced by Azzalini (1985) with the application of the Log-Normal properties. The distribution is modified with the logarithmic of the random variable and the mixture provides the flexibility for the distribution to be unimodal or bimodal where it is able to model heterogeneous data as well. In addition, the mixture of 2 modified Log-Normal distributions can be reduced to the following distributions by setting certain parameters of the proposed mixture distributions with other values such as follows:

1. mixture of 2 Log-Normal distribution when $c_1,\ c_2 = 0$
2. modified Log-Normal when $p = 1$
3. Log-Normal distribution when $c_1,\ c_2 = 0$ and $p = 1$

A relationship chart between the Beta and Skew Normal families including the proposed distributions is shown in Figure 2.1.

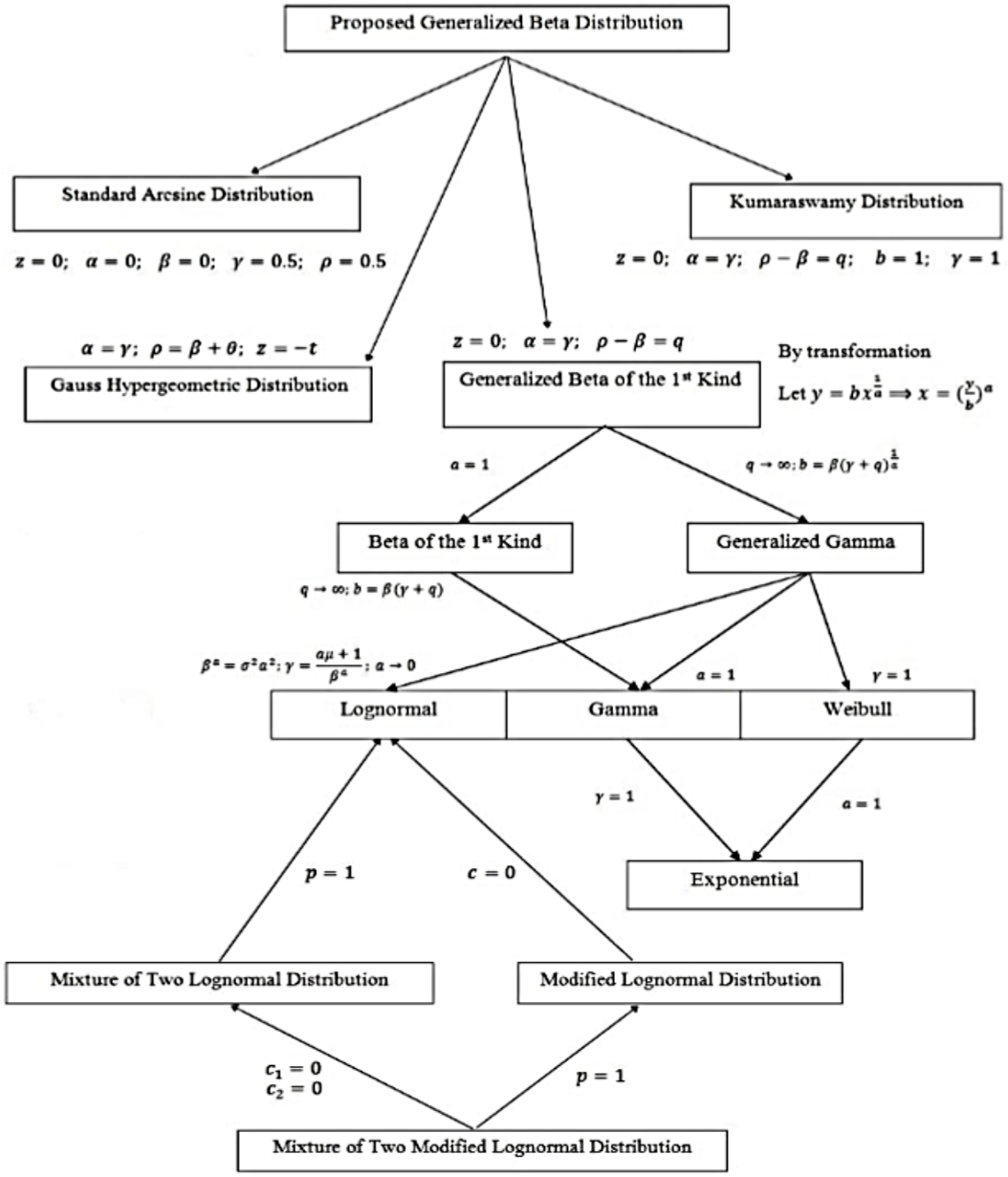

Figure 2.1: Relationship Chart of Various Distributions (Chuah, 2016)

2.4 Application to Various Field of Studies

Application of new models in distribution fitting can be done towards various field of studies. For example under the Beta family distributions, McDonald and Xu (1995) fitted the 4-parameter and 5-parameter Beta distributions to income distribution, stock returns and regression analysis and were compared with other Beta related distributions (i.e. generalized Beta of first and second kind, exponential generalized Beta and etc.). Chotikapanich et al. (2007) fitted a 3-parameter Beta-2 distribution, a generalization of the normalized Beta model to income and inequality data. Even Sarabia et al. (2014) fitted the bivariate Beta generated models to income, health and education data. Moreover, the generalized Beta distribution of the 1^{st} kind along with special models of the Beta-generated distributions such as generalized Beta Normal distribution were applied to financial and voltage data by Alexander et al. (2012).

Besides applying Beta family distributions to man-made data, the distributions were seen to be applied to environmental phenomena and compared with other family distributions. Lima et al. (2016) proposed the 4-parameter Beta distribution that was used to estimate the intensity duration frequency curve of rainfall. It was concluded to fit well towards the historical data. In addition, it could be further evident from Khaleel et al. (2017) research where a 1-parameter Beta Burr type X distribution was introduced and known to fit better than other distributions towards a certain rainfall dataset. Furthermore, Murshed et al. (2018) even modelled the Beta-P distribution for flood frequency analysis and it turns out to perform better than other well known distributions.

For Skew Normal family distributions, numerous field of research were explored to observe the applicability of their statistical distributions. Related distributions such as the Log-Normal distribution was developed to linear and non-linear Log-Normal models were applied to estimate the soil water retention curves by Hwang and Choi (2006) and they were found to be well performed. Besides that, application of Log-Skew Normal distribution to wireless communication was studied by Li et al. (2011). The distribution was transformed and fitted to overcome the drawback of restriction for the skewness in the samples where it was observed to be effective in providing accurate simulation results. The Skew Normal and Skew-student distributions were even fitted to insurance claims data by Eling (2012) and compared with 17 other distributions as well as transformation kernel where they are seen to be good models.

Towards the application of the Skew Normal family distributions on environmental studies, Adiku et al. (1997) did a study on Gamma distribution and concluded that the 2-parameter distribution is more appropriate for data with long tail distribution because it was found to be a good fit to daily rainfall at two sites in Ghana. In addition, Gamma and Log-Normal distributions were studied by Cho et al. (2004) and they mentioned that both of the distributions matched well to the PDF of rainfall data. In addition, Bartoletti and Loperfido (2010) fitted the Skew Normal distribution to the air pollution data where it was seen to be a very good fit.

On the other hand, there seems to be an increasing trend in applying mixed-distributions towards environmental studies such as rainfall as well.

Based on Woolhiser and Roldan (1982), they fitted the Exponential, Gamma and mixed Exponential distributions to rainfall data and it was concluded that the mixed distribution is superior compared to the others. This is futher supported by Suhaila and Jemain (2007a), Suhaila and Jemain (2007b), Suhaila and Jemain (2008) and Suhaila et al. (2011). Studies were done on several types of combined discrete and continuous mixed distributions towards rainfall data which included dry days. It was concluded that the mixed Log-Normal distribution has the best fit in most of the rain stations in Peninsular Malaysia. Thus, this shows that mixture distributions are favourable in modelling rainfall data.

Besides, parameters re-calibrations could also be considered where such practices are common in the banking industry. Parameter re-calibration is a method of parameter initialization before proceeding with the fitting process. For example, certain information about the data is known which help in setting certain parameters of a model to some fixed calculated values. It was stated in some studies that this method produce parameters that helped improved the estimation strength of the models studied.

In the study of Tong et al. (2016), it was mentioned that the zero-adjusted gamma model is more accurate in calibration than the benchmark models. In addition, Tong et al. (2013) presented that the zero adjusted gamma model presents a powerful alternative to existing loss given default approaches using a semi-parametric model. Even Misankova et al. (2015) mentioned that used structural models and reduced-form models as well as Loss Given Default models to become a part of credit risk is accepted not only by academic institutions but also by the banking

industries. Spuchl'akova and Cug (2015) also concluded that reduced-form models based on the assumption that the market price of defaultable financial instruments disclose the investors' expectations about credit risk parameters have proven useful in analysing the dynamics of credit spreads. Comparisons could be done to identify whether re-calibrated distributions would produce more accurate estimations for the proposed distributions for various field of studies.

Due to the common interest of applying statistical distributions to environmental phenomena such as rainfall volume for both the Beta and Skew Normal families, the proposed distributions (i.e. generalized Beta and mixture of 2 modified Log-Normal) will be applied to that particular field of study as well. Moreover, the fitting results of the proposed models could be compared with Chuah (2016) to observe whether the results are consistent or not. Parameter re-calibrations as mentioned above will be used to calculate the zero-inflated probability and will be explained in detail later.

2.5 Methods of Parameter Estimation

In literature, many parameter estimation techniques such as Method of Moments (MOM), Maximum Likelihood Estimators (MLE), Least Squares Method (LSM), and maximum goodness-of-fit estimator were introduced. Currently, many studies are ongoing to identify which method is more suitable to be used depending on the type of distributions under study or the type of data that is being fitted to. Such studies could be seen in

Zhang (1997), Nwobi and Ugomma (2014), Karakoca et al. (2015) and Kateregga et al. (2017). Even new methods of distribution fitting were proposed to fit certain type of complex distributions such as a combination of 2 methods in Fournier et al. (2007), MLE-Least Squares approach in George and Ramachandran (2011) and a newly proposed approach consisting of 5 various methods in Elmahdy and Aboutahoun (2013).

From the studies mentioned above, there are many different methods of parameter estimation. However, it can be seen that MLE is one of the most common method used in many research due to its efficiency as well as asymptotically unbiased properties (Teimouri et al., 2013). Fisher (1925) introduced the concepts of consistency, efficiency as well as sufficiency into statistical theory. They can be found in the MLE method that was well mentioned in Norden (1972). In addition, Myung (2003) mentioned that among the two general methods of parameter estimation, MLE and LSM, MLE is important in the theory of inference which possesses many optimal properties in parameter estimation mentioned above and inferential techniques in statistics. On the other hand, LSM is just primarily a descriptive tool. Aside from the advantages of the MLE mentioned above, it is also preferable as it could also be used to fit censored data as well (Natrella, 2010).

In many studies, it could be observed that MLE tends to be preferred for certain kind of distributions or will perform better than other estimation methods. For example under the Beta family distributions, it was mentioned in Bowman and Shenton (1992) that over a limited parameter space in the Beta distribution, MLE is preferred for the

estimation of the parameters. Furthermore, the classical Beta distribution was also fitted using various estimation method such as MLE, MOM, quantile estimator and etc. In Owen (2008), it was concluded that the MLE performed as well as the straightforward MOM and quantile estimation method. Besides that, Erick et al. (2016) manage to improve the MLE method using the Expectation-Maximization algorithm to cater for Type II censoring scheme data to be fitted to the Kumaraswamy (1980) distribution.

For the Skew Normal family distribution, Dey (2010) estimated the parameters of the Skew Normal distribution using the MLE method by approximating the ratio of the PDF and CDF by linear and non-linear functions. It was concluded that the linear approximation performance was quite satisfactory. Besides that, Karakoca et al. (2015) mentioned that the MLE was the best performed method in fitting its mixture distribution for data with large sample sizes. Hanevik (2016) fitted the Black Scholes model which is very well related to the Log-Normal distribution and mentioned that MLE is preferred as it uses more information of the likelihood function to estimate the parameters making the estimates produce smaller variance and being less biased. As the rainfall data is a censored type data and there are favourable results of using the MLE as the parameter estimation method in fitting complex as well as mixture models, hence it is chosen to fit the proposed distributions in this research.

2.6 Types of Model Selection Criteria and Evaluation Metrics

Model selection criteria are important and are also common practices in examining the performance of a different fitted models. The correct choice of models to detect the similarities and differences depend on the appropriate use of model selection strategies as they might favour different models (see Wong (1994)). Thus, choosing the correct model selection criteria is important in order to obtain reliable results. Here, several model selection criteria for examining the performance of a set of alternative models are reviewed. They are the K-S goodness of fit test, Chi-Square goodness of fit test, AIC, Bayesian Information Criteria (BIC), log-likelihood ratio test (LRT), RMSE and mean absolute error (MAE).

Goodness of fit test is a type of model selection criteria that are usually applied to test a hypothesis about a certain population's distribution. A few of the most commonly used tests are the K-S test and Chi-Square (χ^2) test. To choose an appropriate goodness of fit test among the two, Massey Jr (1951) discussed the advantage of the K-S test to the χ^2 test in their research. Firstly, it was mentioned that the K-S test will be able to detect smaller deviations in cumulative distributions compared to the χ^2 test. Secondly, it is also stated that the K-S test treats each observation individually and thus does not lose information by grouping which is needed by the χ^2 test especially when the sample size is small. Lilliefors (1967) even stated that the K-S test appears to be more powerful compared to the χ^2 test for any sample size. Furthermore, Mitchell (1971)'s research supported the advantage mentioned by explaining that the K-S test does not have the expected frequencies constraints which is

normally associated with χ^2. Therefore, with these advantages mentioned, the K-S test is chosen for this research.

In Wong (1994), it was explained that LRT is only useful for moderating sample sizes in conjunction with nested Chi-square difference test while the AIC are systematically biased toward models incorporating group differences in log-linear modelling. The BIC is known to be the most reliable in his research as it provides consistent results when the sample size is sufficiently large. Anderson and Burnham (1999) mentioned that Kullback-Leibler information criteria such as AIC, corrected AIC (AIC_c) and corrected Quasi-AIC ($QAIC_c$), attempt to select good approximating models for inference based on the principle of parsimony where it does not assume a true model exists. On the other hand, criteria such as BIC, minimum description length (MDL) and Hannan-Quinn (HQ) are "dimension consistent" that consistently estimate the dimension of the true model where it was assumed to exists which requires very large sample sizes.

It was also mentioned in Cavanaugh (2009)that AIC should be selected when the primary goal of modelling is *predictive* where a model is build to predict new outcomes. Whereas, BIC should be used when the goal is *descriptive*, where a model is developed to represent the most meaningful factors influencing the outcome based assessment of relative importance. As this research is to identify the most suitable model that could estimate the frequency of rainfall volume and a true model was not assumed to exist, AIC is chosen as a model selection criteria.

To choose between model evaluation metrics such as RMSE and MAE, Willmott and Matsuura (2005) mentioned that RMSE is inappropriate as it is a function of three characteristics for a set of errors and MAE is a more natural measure of unambiguous average error. However, Chai and Draxler (2014) managed to show that the RMSE which has been used as a standard metric to measure model performances in meteorology, air quality and climate research is not ambiguous and is more appropriate than MAE when the errors follow a normal distribution. They mentioned that RMSE is preferable as it avoids the use of absolute value, which is highly undesirable in many mathematical calculations especially on model error sensitivities or data assimilation applications. In addition, the sum of squared errors is often defined as the cost function to be minimized by adjusting model parameters where penalizing large errors through the defined least-square terms proves to be very effective in improving model performance. As this research is related to climate, RMSE is thus chosen.

The proposed distributions will be introduced in the next chapter. Their properties will be derived using various known theorems and concepts.

CHAPTER 3

PROPOSED DISTRIBUTIONS

From Chuah (2016), three new distributions were proposed which are the modified Beta, generalized Beta and mixture of 2 modified Log-Normal distributions. Studies were done on the modified Beta distribution where it's properties were developed. However, no further studies were done on the remaining two proposed models (i.e. generalized Beta and mixture of 2 modified Log-Normal) where some of their properties are derived in this research.

In this chapter, theoretical studies on the derivation of the proposed generalized Beta distribution will be shown in Section 3.1. Under this section, the theories needed to derive the properties of the generalized Beta distribution will be presented in Section 3.1.1 Related Theories. In addition, a basic introduction on the PDF and CDF including the flexibility of the PDF will be discussed in Section 3.1.2 Density Function. Under Section 3.1.3 Moment Generating Function (MGF), the expected value, general moment, variance, skewness and kurtosis properties will be derived in that respective order.

The derivation of the proposed mixture of 2 modified Log-Normal distributions will be shown in Section 3.2. The theories involved will be stated in Section 3.2.1 such as the related distributions and functions used together with the lemma needed. Under Section 3.2.2 Density Function,

an introduction on the PDF and also its flexibility as well as the CDF will be explained which includes the factorization and simplification of the mixture distribution. Section 3.2.3 Moment Generating Function (MGF) will present the derivations of the expected value, general moment, variance, skewness and kurtosis accordingly.

Detailed proves of the equations and functions can be found in the appendix.

3.1 Generalized Beta Distribution

3.1.1 Related Theories

To derive the properties of generalized Beta distribution, the ${}_2F_1$ Hypergeometric contiguous relation function is needed. It was found in Rakha et al. (2011), Equation (22) where two Hypergeometric function with the same argument z is said to be contiguous if their parameters a, b and c differ by integers as follows:

$$c\,{}_2F_1(a,b;c;z) - a\,{}_2F_1(a+1,b;c+1;z) + (a-c)\,{}_2F_1(a,b;c+1;z) = 0$$

$$\implies {}_2F_1(a,b;c;z) = \frac{a\,{}_2F_1(a+1,b;c+1;z) - (a-c)\,{}_2F_1(a,b;c+1;z)}{c}$$

3.1.2 Density Functions

Probability Density Function, PDF

$$f(x) = \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})B(\gamma,\rho)}, \quad (3.1)$$

where $\sigma > 0, \gamma > 0, z < 0.5, (\gamma+\rho-\alpha-\beta) > 0.$

${}_3F_2$ is a generalized Hypergeometric function defined by:

$${}_3F_2(\alpha_1,\alpha_2,\alpha_3;\beta_1,\beta_2;z) = \sum_{k=0}^{\infty}\frac{(\alpha_1)_k(\alpha_2)_k(\alpha_3)_k}{(\beta_1)_k(\beta_2)_k}\frac{z^k}{k!},$$

$$F(\alpha,\beta;\gamma;x) = {}_2F_1(\alpha,\beta;\gamma;x)$$

and ${}_2F_1$ is a Hypergeometric function defined by:

$${}_2F_1(a,b;c;z) = \textstyle\sum_{k=0}^{\infty}\frac{(a)_k(b)_k}{(c)_k}\frac{z^k}{k!}$$

Given that

$$\int_0^1 x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)dx = \frac{\Gamma(\gamma)\Gamma(\rho)\Gamma(\gamma+\rho-\alpha-\beta)}{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}(1-z)^{-\sigma}\times$$
$${}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1}),$$

where $\mathrm{Re}(\gamma) > 0, \mathrm{Re}(\rho) > 0, \mathrm{Re}(\gamma+\rho-\alpha-\beta) > 0,\ |arg(1-z) < \pi|$

(Gradshteyn and Ryzhik, 2014)

It can be easily shown that

$$\int_0^1 \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})B(\gamma,\rho)}\,dx = 1$$

Density Function Flexibility

From the generalized Beta distribution, it can be understood that the model is complex and is also a generalized version for many Beta family distributions. Therefore, this distribution has an advantage in terms of

flexibility and versatility where it could provide descriptions to numerous different data types which includes increasing, decreasing, bath-tub, uniantimodal or even unimodal shape distribution based on the x-variable that is within the range of 0 and 1 ($0 < x < 1$) as well as the value of its parameters. By identifying the shapes of the density functions based on the parameters, it will help in identifying the values of the parameters that needs to be initialized for an actual dataset empirical study. Their flexibility is illustrated graphically to identify their fitting capability as shown below. Figure 3.1 illustrates the various forms of PDF for the 6-parameter generalized Beta distribution set with different parameter values as shown in Table 3.1.

Table 3.1: Generalized Beta Distribution Parameters

Set (Colour)	γ	ρ	α	β	σ	z
A (Blue)	0.476	5.275	0.501	0.380	6.620	0.285
B (Green)	6	5	-2	2	2	1
C (Red)	0.5	1.5	1	0.5	0	0.4
D (Light Blue)	3	1	2	1	2	0.25
E (Purple)	3	1.5	-1.5	1	2	0.4
F (Dark Yellow)	2	3	-1	-1	1	0.3

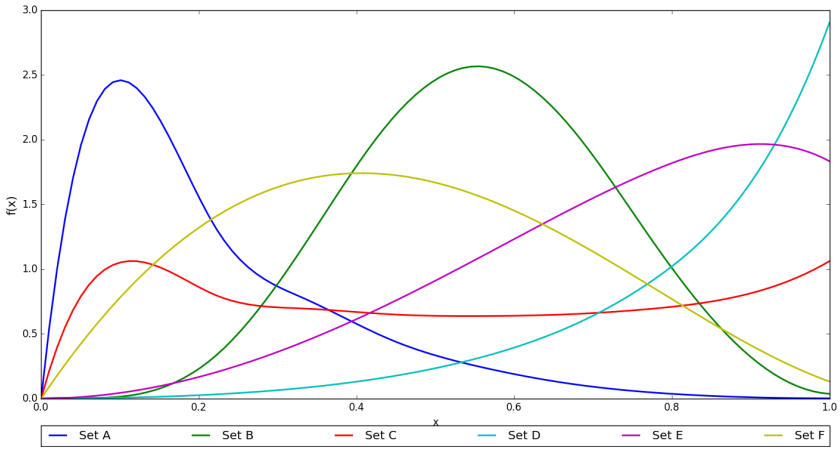

Figure 3.1: Generalized Beta Distribution Versatile PDF

Cumulative Distribution Function, CDF

In the e-handbook of Natrella (2010), it was mention that the CDF of the Beta distribution is also called the incomplete Beta function ratio denoted by I_x defined by:

$$F(x) = I_x(p,q) = \frac{\int_0^x t^{p-1}(1-t)^{q-1}\ dt}{\boldsymbol{B}(p,q)}, \quad 0 \leq x \leq 1;\ p,q > 0$$

where $\boldsymbol{B}$ is the Beta function.

The incomplete Beta function is also found in the generalized Beta distribution at integration part as follows:

$$\frac{\int_0^1 x^{\gamma-1}(1-x)^{\rho-1}dx}{\boldsymbol{B}(\gamma,\rho)}$$

Therefore, the CDF general form of the generalized Beta distribution could not be derived.

3.1.3 Moment Generating Function, MGF

To ensure that the derivations are systematically shown, a constant h is assigned to represent the parameters that do not contain the x-variable as they will not be affected by the integration.

Let

$$h = \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})B(\gamma,\rho)}$$

$$\implies f(x) = hx^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)$$

Expected Value, E[X]

$$
\begin{aligned}
E[X] &= \int_0^1 hx \cdot x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)\ dx \\
&= \int_0^1 hx^{(\gamma-1)+1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)\ dx
\end{aligned} \tag{3.2}
$$

From Equation (3.2), it can be seen that the γ parameter is affected when the x-variable is multiplied per expected value definition. Under the modified Beta distribution from Chuah (2016), the properties were derived based on a table of equations by Gradshteyn and Ryzhik (2014). Direct comparison could be made on the modified Beta distribution to derive its properties. However, the same approach could not be followed for the generalized Beta distribution as the γ parameter also exists in the ${}_2F_1$ Hypergeometric function, $F(\alpha,\beta;\gamma;x)$. In order to derive the properties, the contiguous relation function mentioned in Section 3.1.1 is needed where the following is obtained:

$$
\begin{aligned}
E[X] &= \int_0^1 hx^{(\gamma-1)+1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)dx \\
&= \int_0^1 hx^{(\gamma-1)+1}(1-x)^{\rho-1}(1-zx)^{-\sigma}\frac{\alpha\,{}_2F_1(\alpha+1,\beta;\gamma+1;x)}{\gamma} - \\
&\quad \frac{(\alpha-\gamma)\,{}_2F_1(\alpha,\beta;\gamma+1;x)}{\gamma}dx \\
&= h\int_0^1 \frac{\alpha}{\gamma}x^{(\gamma-1)+1}(1-x)^{\rho-1}(1-zx)^{-\sigma}\,{}_2F_1(\alpha+1,\beta;\gamma+1;x)dx - \\
&\quad h\int_0^1 \frac{\alpha-\gamma}{\gamma}x^{(\gamma-1)+1}(1-x)^{\rho-1}(1-zx)^{-\sigma}\,{}_2F_1(\alpha,\beta;\gamma+1;x)dx
\end{aligned}
$$

From here, it can be seen that adjustments have been made to the γ parameter of the ${}_2F_1$ Hypergeometric Function. With this, the table of equations by Gradshteyn and Ryzhik (2014) and direct comparison method mentioned above can be used to obtain the properties of the generalized Beta distribution as follows:

$$E[X] = \frac{h\alpha}{\gamma}[\frac{\Gamma(\gamma+1)\Gamma(\rho)\Gamma(\gamma+\rho-\ \alpha-\beta)}{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+1+\rho-\beta)}(1-z)^{-\sigma}$$
$${}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+1+\rho-\beta;\frac{z}{z-1})]-$$
$$\frac{h(\alpha-\gamma)}{\gamma}[\frac{\Gamma(\gamma+1)\Gamma(\rho)\Gamma(\gamma+1+\rho-\alpha-\beta)}{\Gamma(\gamma+1+\rho-\alpha)\Gamma(\gamma+1+\rho-\beta)}(1-z)^{-\sigma}$$
$${}_3F_2(\rho,\sigma,\gamma+1+\rho-\alpha-\beta;\gamma+1+\rho-\alpha,\gamma+1+\rho-\beta;\frac{z}{z-1})]$$

Through the substitution of h, the following equation is expanded:

$$E[X] = \frac{\alpha}{\gamma}[\frac{\Gamma(\gamma+1)\Gamma(\gamma+\rho-\beta)\Gamma(\rho)}{\Gamma(\gamma+1+\rho-\beta)\Gamma(\gamma+\rho)B(\rho,\gamma)}$$
$$\frac{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+1+\rho-\beta;\frac{z}{z-1})}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})}]$$
$$-\frac{\alpha-\gamma}{\gamma}[\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)\Gamma(\gamma+1)\Gamma(\gamma+1+\rho-\alpha-\beta)\Gamma(\rho)}{\Gamma(\gamma+1+\rho-\alpha)\Gamma(\gamma+1+\rho-\beta)\Gamma(\gamma+\rho-\alpha-\beta)\Gamma(\gamma+\rho)B(\gamma,\rho)}$$
$$\frac{{}_3F_2(\rho,\sigma,\gamma+1+\rho-\alpha-\beta;\gamma+1+\rho-\alpha,\gamma+1+\rho-\beta;\frac{z}{z-1})}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})}] \tag{3.3}$$

General Moment

To obtain the 2nd, 3rd, 4th moments and etc., multiple recursions are needed. As it only affects the ${}_2F_1$ Hypergeometric function, a generalized form could be derived. From the expansion derived as shown in Appendix A, a Binomial expansion pattern (1,1 ; 1,2,1 and 1,3,3,1) could be seen. After some derivations, a generalized version is formed:

$${}_2F_1(a,b;c;z) = \sum_{k=1}^{n+1}\binom{n}{k-1}(-1)^{n-k+1}\frac{(a)_{k-1}(a-c-n+k)_{n-k+1}}{(c)_n}$$
$${}_2F_1(a+k-1,b;c+n;z) \tag{3.4}$$

By placing Equation (3.4) into the general moment function, the following equation is obtained:

$$
\begin{aligned}
E[X^n] = & \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})B(\gamma,\rho)} \\
& \int_0^1 x^{\gamma-1+n}(1-x)^{\rho-1}(1-zx)^{-\sigma} \\
& \sum_{k=1}^{n+1}\binom{n}{k-1}(-1)^{n-k+1}\frac{(\alpha)_{k-1}(\alpha-\gamma-n+k)_{n-k+1}}{(\gamma)_n} \\
& {}_2F_1(\alpha+k-1,\beta;\gamma+n;x)\ dx
\end{aligned} \tag{3.5}
$$

Thus, the general moment of the generalized Beta distribution is derived.

Variance,Var[X]; Skewness,skew(x) and Kurtosis, kurt(x)

The general formula of the variance, skewness and kurtosis are as follows:

$$Var[X] = E[X^2] - (E[X])^2 \tag{3.6}$$

$$Skew(x) = E[(\frac{x-\mu}{\sigma})^3] = \frac{E[X^3] - 3E[X]E[X^2] + 2E[X]^3}{(E[X^2] - E[X]^2)^{\frac{3}{2}}} \tag{3.7}$$

$$Kurt(x) = E[(\frac{x-\mu}{\sigma})^4] = \frac{E[X^4] - 4E[X]E[X^3] + 6E[X]^2E[X^2] - 3E[X]^4}{(E[X^2] - E[X]^2)^2} \tag{3.8}$$

To obtain the general form of Equations (3.6), (3.7) and (3.8), the 2nd, 3rd and 4th moments are needed respectively. They can be obtained by substituting n with the nth moment needed to the Equation (3.5). As the function will be long and complex, it will not be shown in this dissertation.

Table 3.2 summarizes the proposed generalized Beta distribution's properties.

Table 3.2: Properties of generalized Beta distribution

Parameter(s)	$\gamma > 0, \sigma > 0,\ z < 0.5$ $(\gamma + \rho - \alpha - \beta) > 0$
PDF	$\frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}x^{\gamma-1}(1-x)^{\rho-1}(1-zx)^{-\sigma}F(\alpha,\beta;\gamma;x)}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})B(\gamma,\rho)}$, $0 < x < 1$ where $F(a,b;c;z)$ is the ${}_2F_1$ Hypergeometric function, ${}_3F_2(a,b,c;y;z)$ is the ${}_3F_2$ Hypergeometric function Γ is the gamma function and $\boldsymbol{B}$ is the Beta function
CDF	$I_x(\gamma,\rho,\alpha,\beta,\sigma,z)$ $= \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})B(\gamma,\rho)}$ $\int_0^x t^{\gamma-1}(1-t)^{\rho-1}(1-zt)^{-\sigma}F(\alpha,\beta;\gamma;t)\ dt,$ $0 < x < 1$ where $I_x(\gamma,\rho,\alpha,\beta,\sigma,z)$ is the incomplete Beta function of the generalized Beta distribution
E[X]	$\frac{\alpha}{\gamma}\big[\frac{\Gamma(\gamma+1)\Gamma(\gamma+\rho-\beta)\Gamma(\rho)}{\Gamma(\gamma+1+\rho-\beta)\Gamma(\gamma+\rho)B(\rho,\gamma)}$ $\frac{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+1+\rho-\beta;\frac{z}{z-1})}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})}\big]$ $-\frac{\alpha-\gamma}{\gamma}\big[\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)\Gamma(\gamma+1)\Gamma(\gamma+1+\rho-\alpha-\beta)\Gamma(\rho)}{\Gamma(\gamma+1+\rho-\alpha)\Gamma(\gamma+1+\rho-\beta)\Gamma(\gamma+\rho-\alpha-\beta)\Gamma(\gamma+\rho)B(\gamma,\rho)}$ $\frac{{}_3F_2(\rho,\sigma,\gamma+1+\rho-\alpha-\beta;\gamma+1+\rho-\alpha,\gamma+1+\rho-\beta;\frac{z}{z-1})}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})}\big]$
VAR[X]	$E[X^2] - (E[X])^2$
General Moment	$E[X^n] = \frac{\frac{\Gamma(\gamma+\rho-\alpha)\Gamma(\gamma+\rho-\beta)}{\Gamma(\gamma+\rho)\Gamma(\gamma+\rho-\alpha-\beta)}(1-z)^{\sigma}}{{}_3F_2(\rho,\sigma,\gamma+\rho-\alpha-\beta;\gamma+\rho-\alpha,\gamma+\rho-\beta;\frac{z}{z-1})B(\gamma,\rho)}$ $\int_0^1 x^{\gamma-1+n}(1-x)^{\rho-1}(1-zx)^{-\sigma}$ $\sum_{k=1}^{n+1}\binom{n}{k-1}(-1)^{n-k+1}\frac{(\alpha)_{k-1}(\alpha-\gamma-n+k)_{n-k+1}}{(\gamma)_n}$ ${}_2F_1(\alpha+k-1,\beta;\gamma+n;x)dx$

The Hypergeometric functions can be defined by ${}_3F_2(\alpha_1,\alpha_2,\alpha_3;\beta_1,\beta_2;z) = \sum_{k=0}^{\infty}\frac{(\alpha_1)_k(\alpha_2)_k(\alpha_3)_k}{(\beta_1)_k(\beta_2)_k}\frac{z^k}{k!}$ and $F(\alpha,\beta;\gamma;x) = {}_2F_1(\alpha,\beta;\gamma;x) = \sum_{k=0}^{\infty}\frac{(a)_k(b)_k}{(c)_k}\frac{z^k}{k!}$

3.2 Mixture of 2 Modified Log-Normal Distributions

3.2.1 Related Theories

To derive the properties of the mixture of 2 modified Log-Normal distributions, three important references are needed which are the Owen's T-function, Skew Normal distribution's properties and Lemma 2.1 (Brown, 2001). The Owen's T-function and Skew Normal distribution's properties are needed to derive the CDF. Lemma 2.1, a summarized function derived by Brown (2001) from page 5 to page 7 of the thesis is needed to derive the expected value, the general moment, variance, skewness and kurtosis.

I) Owen's T-Function

$$T(x,a) = \frac{1}{2\pi}\int_0^a \frac{e^{-\frac{x^2(1+t^2)}{2}}}{1+t^2}dt = \frac{1}{8}\text{erfc}(-\frac{x}{\sqrt{2}})\text{erfc}(\frac{x}{\sqrt{2}}),$$

$$\text{where erf} = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\ dt,\ \text{erfc}(x) = 1-\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\ dt$$

II) Skew-Normal Distribution

$$f(x) = \frac{2}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\int_{-\infty}^{\alpha\frac{x-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt, \qquad F(X) = \Phi(\frac{x-\mu}{\sigma}) - 2T(\frac{x-\mu}{\sigma},\alpha)$$

$$E[X] = \mu + \sigma\delta\sqrt{\frac{2}{\pi}},\ \text{where } \delta = \frac{\alpha}{\sqrt{1+\alpha^2}}, \qquad Var[X] = \sigma^2(1-2\frac{\delta^2}{\pi})$$

$$\text{skew} = \frac{4-\pi}{2}\frac{(\delta\sqrt{\frac{2}{\pi}})^3}{(1-2\frac{\delta^2}{\pi})^{\frac{3}{2}}} \qquad \text{kurtosis} = 2(\pi-3)\frac{(\delta\sqrt{\frac{2}{\pi}})^4}{(1-2\frac{\delta^2}{\pi})^2}$$

III) Lemma 2.1

$$E\{\Phi(hY+k)\} = \Phi(\frac{k}{\sqrt{1+h^2}})$$

3.2.2 Density Functions

Probability Density Function, PDF

$$f(x) = 2[\frac{p}{\sqrt{2\pi}\sigma_1 x} e^{-\frac{(\ln x-\mu_1)^2}{2\sigma_1^2}} \int_0^{(\frac{x}{e^{\mu_1}})^{\frac{(c_1)}{\sigma_1}}} \frac{1}{\sqrt{2\pi}t} e^{-\frac{(lnt)^2}{2}} dt + \frac{1-p}{\sqrt{2\pi}\sigma_2 x} e^{-\frac{(\ln x-\mu_2)^2}{2\sigma_2^2}} \int_0^{(\frac{x}{e^{\mu_2}})^{\frac{(c_2)}{\sigma_2}}} \frac{1}{\sqrt{2\pi}t} e^{-\frac{(lnt)^2}{2}} dt] \tag{3.9}$$

First, the integral part of equation (3.9) will be integrated. Let :

$u = \ln t \implies du = \frac{1}{t}dt$

When $t \to 0 \Rightarrow u \to -\infty$

When $t = \left(\frac{x}{e^{\mu_i}}\right)^{\frac{c_i}{\sigma_i}} \implies u = \frac{c_i}{\sigma_i}(\ln x - \mu_i)$, where *i=1,2*

Therefore,

$$\begin{aligned}
&\int_0^{(\frac{x}{e^{\mu_1}})^{\frac{(c_1)}{\sigma_1}}} \frac{1}{\sqrt{2\pi}t} e^{-\frac{(\ln t)^2}{2}} dt \\
&= \int_{-\infty}^{\frac{c_i}{\sigma_i}(\ln x-\mu_i)} \frac{1}{\sqrt{2\pi}t} e^{-\frac{u^2}{2}}\ t\ du \\
&= \int_{-\infty}^{\frac{c_i}{\sigma_i}(\ln x-\mu_i)} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\ du \sim Norm(\mu = 0, \sigma = 1) \\
&= \Phi(\frac{c_i}{\sigma_i}(\ln x - \mu_i)) \\
&= \Phi(z_i), \text{ where } z = \tfrac{c_i}{\sigma_i}(\ln x - \mu_i) \text{ and } i = 1, 2
\end{aligned}$$

$$\implies f(x) = 2[\frac{p}{\sqrt{2\pi}\sigma_1 x} e^{-\frac{(\ln x-\mu_1)^2}{2\sigma_1^2}} \Phi(\frac{c_1}{\sigma_1}(\ln x - \mu_1)) + \frac{1-p}{\sqrt{2\pi}\sigma_2 x} e^{-\frac{(\ln x-\mu_2)^2}{2\sigma_2^2}} \Phi(\frac{c_2}{\sigma_2}(\ln x - \mu_2))] \tag{3.10}$$

$$\implies f(x) = 2[\frac{p}{\sqrt{2\pi}\sigma_1 x} e^{-\frac{z_1^2}{2}} \Phi(c_1 z_1) + \frac{1-p}{\sqrt{2\pi}\sigma_2 x} e^{-\frac{z_2^2}{2}} \Phi(c_2 z_2)], \tag{3.11}$$

where $z_i = \frac{\ln x-\mu_i}{\sigma_i} \quad \forall \quad i = 1, 2$

Density Function Flexibility

From the mixture of 2 modified Log-Normal distributions PDF, it can be seen as a flexible unimodal or bimodal distribution with various peaks' magnitude due to the mixture properties present. For example, the first peak can be higher than the second or vice versa, only one peak is graphed which is approximately a Log-Normal distribution, both peaks are almost equivalent in height and etc. Such forms could be illustrated as the $x-$variable only accepts values greater than zero ($x > 0$) which provides a lot of flexibility from the diversified range of x. Figure 3.2 illustrates the various PDF forms given different parameter values as tabled in Table 3.3.

Table 3.3: Mixture of 2 Modified Log-Normal Distributions Parameters

Set (Colour)	p	μ_1	σ_1	c_1	μ_2	σ_2	c_2
A (Blue)	0.387	3.232	0.623	0.188	1.419	1.267	0.065
B (Green)	0.8	1	1	-1	1	1	-1
C (Red)	0.85	0	1	0.2	1.5	0.1	-0.1
D (Light Blue)	0.2	0	0.5	-0.1	2	0.1	0.1
E (Purple)	0.7	1	0.2	0.5	0.1	0.2	1
F (Dark Yellow)	0.65	1	0.4	0.5	0.5	0.35	1

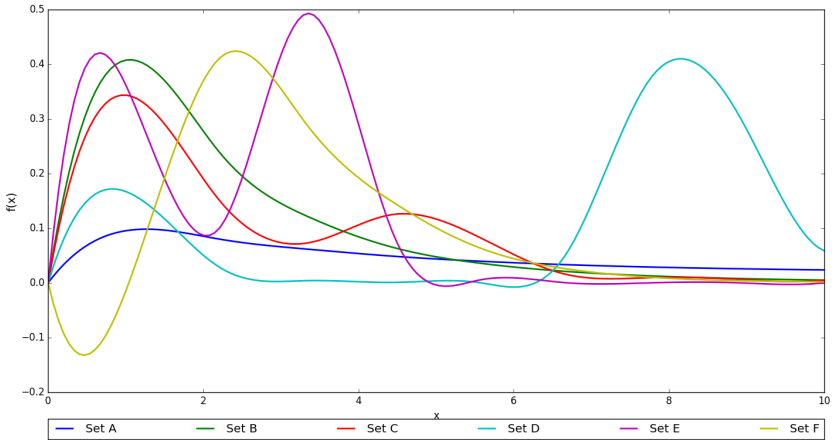

Figure 3.2: Mixture of 2 Modified Log-Normal Distributions Versatile PDF

Cumulative Distribution Function, CDF

$$F(X) = \int_0^x f(y)\ dy$$

$$F(X) = \int_0^x 2[\frac{p}{\sqrt{2\pi}\sigma_1 y} e^{-\frac{(\ln y-\mu_1)^2}{2\sigma_1^2}} \Phi(\frac{c_1}{\sigma_1}(\ln y-\mu_1)) + \frac{1-p}{\sqrt{2\pi}\sigma_2 y} e^{-\frac{(\ln y-\mu_2)^2}{2\sigma_2^2}} \Phi(\frac{c_2}{\sigma_2}(\ln y-\mu_2))]\ dy$$

By letting $\ln y = a$, $y = e^a \implies dy = e^a\ da$

When $y = x \implies a = \ln x$

When y approaches 0, $\ln 0$ is indefinite. Therefore, when $\ln y$ is limit to 0, the value tends to approach towards negative infinity.

$$F(X) = 2\lim_{k\to 0}\int_k^{\ln x} [\frac{p}{\sqrt{2\pi}\sigma_1} e^{-\frac{(a-\mu_1)^2}{2\sigma_1^2}} \Phi(\frac{c_1}{\sigma_1}(a-\mu_1)) + \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-\frac{(a-\mu_2)^2}{2\sigma_2^2}} \Phi(\frac{c_2}{\sigma_2}(a-\mu_2))]\ da$$

$$= 2\int_{-\infty}^{\ln x} [\frac{p}{\sqrt{2\pi}\sigma_1} e^{-\frac{(a-\mu_1)^2}{2\sigma_1^2}} \Phi(\frac{c_1}{\sigma_1}(a-\mu_1)) + \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-\frac{(a-\mu_2)^2}{2\sigma_2^2}} \Phi(\frac{c_2}{\sigma_2}(a-\mu_2))]\ da$$

$$= 2[\int_{-\infty}^{\ln x} \frac{p}{\sqrt{2\pi}\sigma_1} e^{-\frac{(a-\mu_1)^2}{2\sigma_1^2}} \Phi(\frac{c_1}{\sigma_1}(a-\mu_1))\ da + \int_{-\infty}^{\ln x} \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-\frac{(a-\mu_2)^2}{2\sigma_2^2}} \Phi(\frac{c_2}{\sigma_2}(a-\mu_2))\ da] \tag{3.12}$$

By comparing the Skew-Normal Distribution with Equation (3.12), the following is obtained:

$$F(X) = 2\{p[\frac{1}{2}\Phi(\frac{\ln x-\mu_1}{\sigma_1}) - T(\frac{\ln x-\mu_1}{\sigma_1}, c_1)] + (1-p)[\frac{1}{2}\Phi(\frac{\ln x-\mu_2}{\sigma_2}) - T(\frac{\ln x-\mu_2}{\sigma_2}, c_2)]\}$$

$$= p[\Phi(\frac{\ln x-\mu_1}{\sigma_1}) - 2T(\frac{\ln x-\mu_1}{\sigma_1}, c_1)] + (1-p)[\Phi(\frac{\ln x-\mu_2}{\sigma_2}) - 2T(\frac{\ln x-\mu_2}{\sigma_2}, c_2)], \tag{3.13}$$

where $\Phi(x') \sim N(\mu' = 0, \sigma' = 1)$ and $T(h, a) \sim Owen's\ T - function$

3.2.3 Moment Generating Function, MGF

Expected of X, (E[X])

$$
\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} x f(x) dx \\
&= \int_{0}^{\infty} x(2)[\frac{p}{\sqrt{2\pi}\sigma_1 x} e^{-\frac{(\ln x-\mu_1)^2}{2\sigma_1^2}} \Phi(\frac{c_1}{\sigma_1}(\ln x-\mu_1)) + \\
&\quad \frac{1-p}{\sqrt{2\pi}\sigma_2 x} e^{-\frac{(\ln x-\mu_2)^2}{2\sigma_2^2}} \Phi(\frac{c_2}{\sigma_2}(\ln x-\mu_2))]\ dx
\end{aligned}
$$

After some derivations done as shown in Appendix B, the following is obtained.

$$E[X] = 2pe^{\mu_1+\frac{1}{2}\sigma_1^2} E\{\Phi(c_1(b_1+\sigma_1))\} + 2(1-p)e^{\mu_2+\frac{1}{2}\sigma_2^2} E\{\Phi(c_2(b_2+\sigma_2))\}$$

By applying *Lemma* 2.1, $E\{\Phi(hY+k)\} = \Phi(\frac{k}{\sqrt{1+h^2}})$, where h and k are constants, the following equation is obtained:

$$E[X] = 2[pe^{\mu_1+\frac{1}{2}\sigma_1^2}\Phi(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{\mu_2+\frac{1}{2}\sigma_2^2}\Phi(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}})] \tag{3.14}$$

General Moment

The general moment for the mixture of 2 modified Log-Normal distributions is similar to the Skew-Normal distribution with some minor differences. Through the comparison of the Log-Normal distribution and Normal distribution's moment, it can be seen that $\mathrm{E}[Y^t]$ of the Log-Normal distribution is equivalent to $M_x(t)$ of the Normal distribution. This is proved mathematically as shown in Appendix C where the n differentiation of $M_x^n(t)$ of the Normal is equal to $\mathrm{E}[X^n]$ of the Log-Normal.

The same concept can be applied to the mixture of 2 modified Log-Normal distribution sas the proposed distribution is actually the 2 mixture, log of the Skew Normal distribution. For example, the mean of the mixture of 2 modified Log-Normal distribution could be obtained by assigning the value 1 to the t-variable of the Skew Normal distribution's general moment as derived in Appendix D. The Log-Normal and Normal distributions' concept in Appendix C as well as Appendix D are applied to the mixture of 2 modified Log-Normal distributions. Therefore, the general moment of the proposed mixture is as follows:

$$\begin{aligned} M_y(t) &= E[e^{tY}] = E[X^t] \\ &= 2[pe^{\mu_1 t+\frac{1}{2}\sigma_1^2 t^2}\Phi(\frac{c_1\sigma_1 t}{\sqrt{1+c_1^2}}) + (1-p)e^{\mu_2 t+\frac{1}{2}\sigma_2^2 t^2}\Phi(\frac{c_2\sigma_2 t}{\sqrt{1+c_2^2}})], \end{aligned} \tag{3.15}$$

where $Y = \ln X$

Variance, Var[X]

From the equation derived above, multiple moments can be obtained. By using the same concept mentioned, the 2nd moment, E$[X^2]$ can be found by substituting $t = 2$ to the moment function as follows:

$$E[X^2] = M_x(2) = 2[pe^{2\mu_1+2\sigma_1^2}\Phi(\frac{2c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{2\mu_2+2\sigma_2^2}\Phi(\frac{2c_2\sigma_2}{\sqrt{1+c_2^2}})] \tag{3.16}$$

With E$[X^2]$, the variance could be calculated as follows:

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 \\ &= 2[pe^{2\mu_1+2\sigma_1^2}\Phi(\frac{2c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{2\mu_2+2\sigma_2^2}\Phi(\frac{2c_2\sigma_2}{\sqrt{1+c_2^2}})] - \\ &\quad 2^2(pe^{\mu_1+\frac{1}{2}\sigma_1^2}\Phi(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{\mu_2+\frac{1}{2}\sigma_2^2}\Phi(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}}))^2 \end{aligned}$$

$$= 2pe^{2\mu_1+2\sigma_1^2}\Phi(\frac{2c_1\sigma_1}{\sqrt{1+c_1^2}}) + 2(1-p)e^{2\mu_2+2\sigma_2^2}\Phi(\frac{2c_2\sigma_2}{\sqrt{1+c_2^2}}) -$$
$$4[p^2e^{2\mu_1+\sigma_1^2}\Phi^2(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)^2e^{2\mu_2+\sigma_2^2}\Phi^2(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}})] +$$
$$2p(1-p)e^{\mu_1+\mu_2+\frac{1}{2}\sigma_1^2+\frac{1}{2}\sigma_2^2}\Phi(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}})\Phi(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}})]$$

$$Var[X] = 2pe^{2\mu_1+\sigma_1^2}[e^{\sigma_1^2}\Phi(\frac{2c_1\sigma_1}{\sqrt{1+c_1^2}}) - 2p\Phi^2(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}})] +$$
$$2(1-p)e^{2\mu_2+\sigma_2^2}[e^{\sigma_2^2}\Phi(\frac{2c_2\sigma_2}{\sqrt{1+c_2^2}}) - 2(1-p)\Phi^2(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}})] -$$
$$8p(1-p)e^{\mu_1+\mu_2+\frac{1}{2}\sigma_1^2+\frac{1}{2}\sigma_2^2}\Phi(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}})\Phi(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}}) \quad (3.17)$$

Skewness, Skew(x) and Kurtosis, Kurt(x)

In order to derive the skewness and kurtosis, the 1st, 2nd, 3rd and 4th moments are needed (i.e. E[X], $E[X^2], E[X^3]$ and $E[X^4]$). The moments mentioned above can be derived by substituting $t = 1, 2, 3$ and 4 respectively to the general moment of the mixture of 2 modified Log-Normal distributions. The four moment equations needed are available in Appendix E.

Skewness formula:

$$Skew(x) = E[(\frac{x-\mu}{\sigma})^3] = \frac{E[X^3] - 3E[X]E[X^2] + 2E^3[X]}{(E[X^2] - E^2[X])^{\frac{3}{2}}} \quad (3.18)$$

The Kurtosis Formula is as follows:

$$Kurt(x) = E[(\frac{x-\mu}{\sigma})^4] = \frac{E[X^4] - 4E[X]E[X^3] + 6E^2[X]E[X^2] - 3E^4[X]}{(E[X^2] - E^2[X])^2} \quad (3.19)$$

To derive the general form of the skewness and kurtosis for the mixture of 2 modified Log-Normal distributions, the moments obtained in Appendix E need to be substituted to equation (3.18) and (3.19) accordingly. Due to the complexity of the equation, it is will not be shown.

Table 3.4 below summarizes the properties of the proposed mixture of 2 modified Log-Normal distributions derived in this research.

Table 3.4: Properties of mixture of 2 modified Log-Normal distributions

Parameter(s)	$\mu_1 \ \epsilon\Re,\ \sigma_1 > 0,\ 0 \leq c_1 \leq 1$ $\mu_2 \ \epsilon\Re,\ \sigma_2 > 0,\ 0 \leq c_2 \leq 1$ $0 \leq p \leq 1$
PDF	$2[\frac{p}{\sqrt{2\pi}\sigma_1 x} e^{-\frac{z_1^2}{2}} \Phi(c_1 z_1) + \frac{1-p}{\sqrt{2\pi}\sigma_2 x} e^{-\frac{z_2^2}{2}} \Phi(c_2 z_2)]$, where $z_i = \frac{\ln x - \mu_i}{\sigma_i}$ and $\Phi(cz_i)$ is the CDF of the normal distribution $\forall\ i = 1, 2$
CDF	$p[\Phi(\frac{lnx-\mu_1}{\sigma_1}) - 2T(\frac{lnx-\mu_1}{\sigma_1}, c_1)] +$ $(1-p)[\Phi(\frac{lnx-\mu_2}{\sigma_2}) - 2T(\frac{lnx-\mu_2}{\sigma_2}, c_2)]$, where $\Phi(x') \sim N(\mu' = 0, \sigma' = 1)$ and $T(h, a) \sim$ Owen's T-function
E[X]	$2[pe^{\mu_1 + \frac{1}{2}\sigma_1^2} \Phi(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{\mu_2 + \frac{1}{2}\sigma_2^2} \Phi(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}})]$
VAR[X]	$2pe^{2\mu_1+\sigma_1^2}[e^{\sigma_1^2}\Phi(\frac{2c_1\sigma_1}{\sqrt{1+c_1^2}}) - 2p\Phi^2(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}})] +$ $2(1-p)e^{2\mu_2+\sigma_2^2}[e^{\sigma_2^2}\Phi(\frac{2c_2\sigma_2}{\sqrt{1+c_2^2}}) - 2(1-p)\Phi^2(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}})] -$ $8p(1-p)e^{\mu_1+\mu_2+\frac{1}{2}\sigma_1^2+\frac{1}{2}\sigma_2^2}\Phi(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}})\Phi(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}})$
General Moment	$M_y(t) = E[e^{tY}] = E[X^t]$ $E[X^t] = 2[pe^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} \Phi(\frac{c_1\sigma_1 t}{\sqrt{1+c_1^2}}) +$ $(1-p)e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2} \Phi(\frac{c_2\sigma_2 t}{\sqrt{1+c_2^2}})]$, where $Y = \ln X$

The research methodology used for the empirical studies where the proposed distributions and other selected models are fitted to a certain area of study will be discussed in the next chapter.

CHAPTER 4

RESEARCH METHODOLOGY

In this chapter, the procedures used for the empirical studies will be elaborated. In Section 4.1, a detailed explanation on data massaging, splitting and transformation will be shown while derivations of the maximum likelihood estimation will be presented in Section 4.2. The simulation algorithm will be discussed in Section 4.3 and the model selection criteria in Section 4.4.

The empirical studies are conducted through distribution fitting to estimate the parameters of the proposed distributions and their respective families. Furthermore, data simulation will be done and evaluated as well. In this research, the considered dataset will be on rainfall volume collected from Sg Lui, Hulu Langat, Selangor from year 2002 to 2012. Before the estimation and fitting were done, certain adjustments need to be made. Based on the proposed model family distributions, the x-variable constraint for the Beta family and proposed generalized Beta distribution's is $0 < x < 1$ while the Skew Normal continuous family and proposed mixture of 2 modified Log-Normal distributions' variable constraint is $x > 0$. The minimum value of the data is 0 mm and and the maximum rainfall volume is 136.1 mm. Thus, the data need to be rescaled to extend the x-variable's range in order to proceed with the distribution fitting process. The procedures to fit and simulate the data for Beta family and Skew Normal continuous family distributions are slightly different. They will be explained in the Section 4.1.1 and Section 4.1.2 respectively.

The parameters will be estimated using the MLE method. The Python software will be used to undergo this procedure with the help of the Non-Linear Least-Squares Minimization package due to its consideration of the parameters' constraints (Newville et al., 2014). The distributions are "ln" and then squared which transforms the minimization process to maximization. Further elaborations on its derivatives will be shown in Section 4.2. With the estimated parameters, simulation is proceeded using the Accept-Reject algorithm (Casella et al., 2004). The performance of the proposed distributions will be compared with a few related classical continuous distributions for each of the statistical distribution families. Among the models compared, the best performed model will be selected based on the model evaluation metrics and selection criteria results.

Under the Beta family distributions, two models are chosen to be compared with the proposed generalized Beta distribution which are as follows:

1. Beta distribution,
2. Gauss Hypergeometric distribution.

The properties of the two Beta-family distributions mentioned are listed below.

Table 4.1: Properties of Beta distribution

Parameter(s)	$a > 0,\ b > 0$
PDF	$\frac{x^{a-1}(1-x)^{b-1}}{\boldsymbol{B}(a,b)}$, where $0 < x < 1$
CDF	$I_x(a,b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}\ dt}{\boldsymbol{B}(a,b)}$, (Natrella, 2010) where $0 < x < 1;\ a, b > 0$ and $I_x(a,b)$ is the incomplete Beta function
E[X]	$\frac{a}{a+b}$
VAR[X]	$\frac{ab}{(a+b)^2(a+b+1)}$
General Moment	$\frac{\Gamma(a+b)\Gamma(a+n)}{\Gamma(a)\Gamma(a+b+n)}$

$\boldsymbol{B}(a,b)$ is the Beta function defined by $\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

Table 4.2: Properties of Gauss Hypergeometric distribution

Parameter(s)	$\theta > 0,\ \gamma > 0,\ -\infty < \sigma < \infty$
PDF	$\frac{x^{\gamma-1}(1-x)^{\theta-1}(1+tx)^{-\sigma}}{{}_2F_1(\gamma,\sigma;\gamma+\theta;-t)\boldsymbol{B}(\gamma,\theta)}$, $0 < x < 1$ where ${}_2F_1$ is the Hypergeometric function
CDF	$\frac{x^{\gamma}}{\gamma\boldsymbol{B}(\gamma,\theta)}\frac{F_1(\gamma,1-\theta,\sigma;\gamma+1;x,-tx)}{{}_2F_1(\gamma,\sigma;\gamma+\theta;-t)}$, $0 < x < 1$ where ${}_2F_1(a,b;c;z)$ is the ${}_2F_1$ and $F_1(a,b_1,b_2;c;z_1,z_2)$ is Appell's 1^{st} Hypergeometric functions
E$[X)$	$\frac{\gamma}{\gamma+\theta}\frac{{}_2F_1(\sigma,\gamma+1;\gamma+\theta+1;-t)}{{}_2F_1(\sigma,\gamma;\gamma+\theta;-t)}$
VAR$[X]$	$\frac{\gamma}{\gamma+\theta}\Big[\frac{\gamma+1}{\gamma+\theta+1}\frac{{}_2F_1(\sigma,\gamma+2;\gamma+\theta+2;-t)}{{}_2F_1(\sigma,\gamma;\gamma+\theta;-t)} - \frac{\gamma}{\gamma+\theta}\Big\{\frac{{}_2F_1(\sigma,\gamma+1;\gamma+\theta+1;-t)}{{}_2F_1(\sigma,\gamma;\gamma+\theta;-t)}\Big\}^2\Big]$
General Moment	$\frac{\Gamma(\gamma+n)\Gamma(\gamma+\theta)}{\Gamma(\gamma)\Gamma(\gamma+\theta+n)}\frac{{}_2F_1(\sigma,\gamma+n;\gamma+\theta+n;-t)}{{}_2F_1(\sigma,\gamma;\gamma+\theta;-t)}$

The ${}_2F_1$ Hypergeometric functions can be defined by $F(\alpha,\beta;\gamma;x) = {}_2F_1(\alpha,\beta;\gamma;x)$ $= \sum_{k=0}^{\infty}\frac{(a)_k(b)_k}{(c)_k}\frac{z^k}{k!}$ and

$F_1(a,b_1,b_2;c;z_1,z_2)$ is Appell's first Hypergeometric function defined by

$\sum_{r_1,r_2=0}^{\infty}\frac{(a)_{r_1+r_2}(b_1)_{r_1}(b_2)_{r_2}}{(c)_{r_1+r_2}}\frac{z_1^{r_1}z_2^{r_2}}{r_1!r_2!}$

(Nagar and Bedoya Valencia, 2011)

For the Skew Normal family distributions, two models are chosen aside from the proposed mixture of 2 modified Log-Normal distributions. They are as follows:

1. Exponential distribution,

2. Gamma distribution.

The properties of the mentioned distributions are presented in the tables below.

Table 4.3: Properties of Exponential distribution

Parameter(s)	$\beta > 0$
PDF	$\frac{1}{\beta} e^{\frac{-x}{\beta}}$ $0 < x < \infty$
CDF	$1 - e^{\frac{-x}{\beta}}$ $0 < x < \infty$
E[X]	β
VAR[X]	β^2
General Moment	$n!\beta^n$

Table 4.4: Properties of Gamma distribution

Parameter(s)	$\alpha > 0, \theta > 0$
PDF	$\frac{1}{\Gamma(\alpha)\theta^\alpha}\ x^{\alpha-1} e^{\frac{-x}{\theta}}$, $0 < x < \infty$ where $\Gamma(\alpha)$ is the Gamma function
CDF	$\frac{1}{\Gamma(\alpha)} \gamma(\alpha, \frac{x}{\theta})$, $0 < x < \infty$ where $\gamma(\alpha, \frac{x}{\theta})$ is the lower incomplete gamma function
E[X]	$\alpha\theta$
VAR[X]	$\alpha\theta^2$
General Moment	$\frac{\theta^n \Gamma(\alpha+n)}{\Gamma(\alpha)}$

The Gamma function can be defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ and the lower incomplete gamma function can be defined by $\gamma(\alpha, \frac{x}{\theta}) = \int_0^{\frac{x}{\theta}} t^{\alpha-1} e^{-t} dt$

4.1 Data Massaging, Splitting and Distribution Transformation

4.1.1 Beta Family Distributions

For Beta family distributions, the following steps are considered for the empirical studies:

1. Data are massaged and distributions are transformed.
2. The massaged data are split into 80% and 20% with the first 80% being the train data and the remaining 20% being the test data, [80:20].
3. The selected distributions will be fitted using the train data.
4. The estimated parameters will be used to simulate a set of data which has the same count as the test data.
5. The distributions will be tested via K-S test to identify whether the data follows the specified distribution.
6. The performance of the fitted models will be tested using AIC.
7. The accuracy of the simulated data will be measured using RMSE.

Based on the generalized Beta and Beta family distributions, they do not accept values less than or equal to zero and values greater than or equal to one $(0 < x < 1)$. To transform, the method introduced by Smithson and Verkuilen (2006) is used where the data is linearly transformed from their original scale to an open unit interval $(0, 1)$. The method starts off with a normalizing process as follows:

$$x' = \frac{x-a}{b-a},$$

where x is the original data value;

x' is the newly normalized value;

a and b are the smallest and highest value in the dataset.

Then, the range of the data is compressed to avoid zeros and ones by:

$$x'' = \frac{x'(N-1)+0.5}{N},$$

where x'' is the newly transformed data;

N is the data sample size.

Although the dataset is now within the $x-$variable constraint, a transformation process needs to be done on the distributions as well. To transform, let

$$x = \frac{y}{b-a}\frac{N-1}{N} - \frac{a}{b-a}\frac{N-1}{N} + \frac{1}{2N} \implies dx = \frac{N-1}{N(b-a)}\, dy,$$

$$\int_0^x f(x)\, dx = \int_0^{\frac{y}{b-a}\frac{N-1}{N} - \frac{a}{b-a}\frac{N-1}{N} + \frac{1}{2N}} \frac{N-1}{N(b-a)} f\left(\frac{y}{b-a}\frac{N-1}{N} - \frac{a}{b-a}\frac{N-1}{N} + \frac{1}{2N}\right) dy,$$

where $0 < y < 1$

With the transformation process completed, the distribution fitting and simulation processes for the Beta family distribution can be proceeded.

4.1.2 Skew Normal Continuous Family Distributions

For Skew Normal Continuous family distributions, the following are the procedures done in this empirical study:

1. The collected data are split into 80% and 20% with the first 80% of the dataset being the train data and the remaining 20% being the test data, [80:20].
2. The train data is then split to zero and non-zero data.
3. The parameters will be estimated using the non-zero data and the zero data will be used to calculated the zero inflated probability.
4. The estimated parameters and zero inflated probability will be used to simulate a set of data which has the same count as the test data.
5. The distributions will be tested via K-S test to identify whether the data follows the specified distribution.
6. The performance of the models fitted will be tested using AIC.
7. The accuracy of the simulated data will be measured using RMSE.

For the mixture of 2 modified Log-Normal distributions and Skew Normal Continuous family distributions, they do not accept zero-values data where the zero-value represents no rainfall due to the presence of the 'ln' function. In order to solve this, the zero-inflation method discussed by Mullahy (1986) and Pakoksung and Takagi (2017) is introduced. This method splits the train data to zero and non-zero values where the general formula of the zero-inflated method is as follows:

$$P(Y = y) = \omega + (1 - \omega) \cdot f(y),$$

where Y is the count data;

ω is the zero-inflation probability;

$f(y)$ is the probability density function of the fitted distribution.

After the splitting is completed, the non-zero train data is used to fit the distributions to estimate their parameters while the zero-value data is used for parameter re-calibration to calculate the zero-inflation probability, ω with the following formula:

$$\omega = \frac{n_{0train}}{N_{train}},$$

where n_{0train} is the number of zero-value data in the train dataset;

N_{train} is the number of train data.

With the splitting of data completed, the distribution fitting and simulation processes can be continued.

4.2 Maximum Likelihood Estimation Derivations

The MLE method is defined as follows:
Suppose X_1, $X_2, \ldots$, X_n are n independent and identically distributed (IID) samples random variables with joint probability density denoted as

$$f_\theta(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n | \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is the vector of k parameters $(\theta_1, \theta_2, \ldots, \theta_k)$ for a particular distribution.

Given the observed IID values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, the likelihood of $\boldsymbol{\theta}$ which is the probability of observing a given set of data as a function of $\boldsymbol{\theta}$ represented by:

$$\begin{aligned} L(\boldsymbol{\theta}; x_1, x_2, \ldots, x_n) &= f(x_1, x_2, \ldots, x_n | \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} f(x_i | \boldsymbol{\theta}), \end{aligned}$$

where $L(\boldsymbol{\theta}; x_1, x_2, \ldots, x_n)$ is the likelihood function of $\boldsymbol{\theta}$.

The maximum likelihood estimator of $\boldsymbol{\theta}$ are the values of k parameters estimated that maximizes the likelihood function $L(\boldsymbol{\theta}; x_1, x_2, \ldots, x_n)$ making the observed data the "most likely or probable" to occur.

In some ways, maximizing the product of the density functions can be quite tedious. Therefore, it is often assumed the fact that logarithm is an increasing function making the maximizing of the log-likelihood function equivalent to the maximizing of the likelihood function which is as follows:

$$l(\boldsymbol{\theta}; x_1, x_2, \ldots, x_n) = \log(\prod_{i=1}^{n} f(x_i | \boldsymbol{\theta})),$$

where $l(\boldsymbol{\theta}; x_1, x_2, \ldots, x_n)$ is the log-likelihood function.

The maximizing procedure is done through the 1^{st} differentiation process and then equating it to zero which is as follows:

$$\frac{\partial}{\partial \boldsymbol{\theta}} (\sum_{i=1}^{n} \log(f(x_i | \boldsymbol{\theta}))) = 0$$

In the Non-Linear Least-Squares Minimization package written by Newville et al. (2014) from the Python software, it was stated that the main task of fitting a model to a set of data using non-linear least-squares is to write an *objective function* that takes the values of the fitting variables and calculates an array of values to be minimized. The *objective*

function returns an array of values ($\text{data}_{measured}$ - data_{model}) scaled by some weighing factor such as the inverse of the uncertainty in the data where the chi-square statistics is defined as follows:

$$\chi^2 = \sum_i^N \frac{[y_i^{meas} - y_i^{model}(\boldsymbol{v})]^2}{\epsilon_i^2},$$

where y_i^{meas} is the set of measured data or raw data;

$y_i^{model}(\boldsymbol{v})$ is the model calculated value;

$\boldsymbol{v}$ is a set or array of x-variables in the model to be fitted;

ϵ_i is the estimated uncertainty in the data;

N is the number of observed data. (Newville et al., 2014)

However, as the data obtained does not consider or assume to have any uncertainties the base *objective function* is to minimize the residual array, $y_i^{meas} - y_i^{model}(\boldsymbol{v})$. In addition, the data obtained contains only the x-variable which is rainfall volume and the measured data, y_i^{meas} does not exist. Therefore, the final form of the *objective function* is to minimize the following model:

$$\chi^2 = \sum_i^N [y_i^{model}]^2$$

As minimizing the function starting from different initialized parameter values will estimate many different parameter values, the y_i^{model} is then logged ("ln"). Once the "ln" function is placed, the *objective function* is now very similar to the log-likelihood function in the MLE as shown below:

$$\chi^2 = \sum_i^N [\ln(y_i^{model})]^2$$

Maximizing the log-likelihood function is now the main objective in this research. As y_i^{model} is actually the PDF where it is within 0 and 1 ($0 <$

$f(x) < 1$) in this research, the log of it would be less than zero ($\ln f(x) < 0$). Maximizing the log-likelihood of the model is to calculate the value that is closest to zero. However by squaring the log-density function, all the values are now positive making the largest value in the log-density function to be the smallest in the square of the log-density function. Thus, minimizing the sum of square of the log-density function is maximizing the sum of log-density function. Therefore, the final form of the *objective function* is now as follows:

$$\chi^2 = \sum_{i}^{N} [\ln(f^{model}(x_i))]^2$$

4.3 Simulation Algorithm

To simulate, the Accept-Reject algorithm by Casella et al. (2004) will be used. The following steps are the standard algorithm of the simulation:

At iteration $i (i \geq 1)$

1. Generate $X_i \sim g_i$ and $U_i \sim \mathcal{U}[0,1]$ independently.
2. $U_i \leq \epsilon_i f(X_i)/g_i(X_i)$, accept $X_i \sim f$;
3. otherwise, move to iteration $i+1$.

The algorithm mentioned can be used directly for the generalized Beta and Beta family distribution. However, a slight modification needs to be made for the mixture of 2 modified Log-Normal and Skew Normal family distribution. A randomized probability of the Uniform distribution needs to be generated before continuing with the Accept-Reject algorithm. If the probability randomized is less than or equal to the zero-inflated probability ($\omega_i \leq \omega$), a zero-value is returned or else it will proceed on to

the Accept-Reject algorithm simulation. The modified procedure is now as follows:

At iteration $j(j \geq 1)$ and $k(k \geq 1)$

1. Generate $X_k \sim \mathcal{U}[0,1]$.
2. If $X_k < \omega$, then return 0. Else proceed to Accept-Reject algorithm.
3. Generate $X_j \sim g_j$ and $U_j \sim \mathcal{U}[0,1]$ independently.
4. $U_j \leq \epsilon_j f(X_j)/g_j(X_j)$, accept $X_j \sim f$;
5. otherwise, move to iteration $j+1$.

4.4 Model Selection Criteria

Model evaluation metrics and selection criteria such as the K-S test, RMSE and AIC are used to identify whether the rainfall volume follow a certain distribution as well as to measure the accuracy level of the simulated values and also the performance of the model respectively based on the test data. The K-S test starts off by identifying the null hypothesis and alternative hypothesis which are as follows:

H_0 : The frequency of rainfall volume data follow a specified distribution.
H_1 : The frequency of rainfall volume data do not follow a specified distribution.

The specified distributions mentioned are the distributions where the data were fitted to. Thus, the K-S test is defined as follows:

$$D = \max_{1 \leq i \leq N} (|F_0(x_i) - S_N(x_{i-1})|, |S_N(x_i) - F_0(x_i)|),$$

where N is the sample size,

$F_0(x_i)$ is the theoretical cumulative distribution and

$S_N(x_i)$ is the cumulative step-function of a sample

(i.e. $S_N(x_i) = \frac{k}{N}$, where k is the number of observations less than or equal to x_i).

(Massey Jr, 1951)

The test statistics critical point, $D_\alpha(N)$ such that $Pr[D > D_\alpha(N)] = \alpha$ is obtained from the K-S table which can found in Massey Jr (1951). If $D > D_\alpha(N)$, then the null hypothesis, H_0 is rejected else it is accepted at $\alpha\%$ level of significance.

After identifying whether the distributions follow a specified distribution, it is important to determine which model is the better fit. Therefore, the AIC is chosen to measure the performance of the model where it has the following formula:

$$AIC = 2k - 2\hat{l},$$

where k is the number of parameters present in the distribution;

$\hat{l}$ is the log-likelihood function.

In order to evaluate the accuracy and estimation strength of the models, the RMSE is calculated with the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}},$$

where y_i is the simulated value;

x_i is the actual value from the test dataset;

n is the sample size.

Empirical results will be discussed in the next chapter. The results of the parameter estimation as well as the calculated values of the model selection criteria will be tabled and compared with various models that was mentioned earlier in this chapter.

CHAPTER 5

EMPIRICAL RESEARCH FINDINGS AND DISCUSSION

With the theoretical studies done on both of the proposed distribution, it is important to apply the theoretical concepts towards practical situations. In order to understand the versatility of the proposed distributions in fitting a real dataset, they will be applied to rainfall volume in this research. In addition, the results could also be checked to identify whether do they support the conclusion made by Chuah (2016).

In this chapter, a description of the rainfall data collected will be described in Section 5.1. The results of the parameter estimation as well as the model selection criteria will be organized in a table and discussed in Section 5.2.

5.1 Description of Rainfall Volume Data

As mentioned, the data used in this research is on daily rainfall volume collected from Sg Lui, Hulu Langat, Selangor from the year 2002 to 2012. The trend of rainfall volume is illustrated in Figure 5.1.

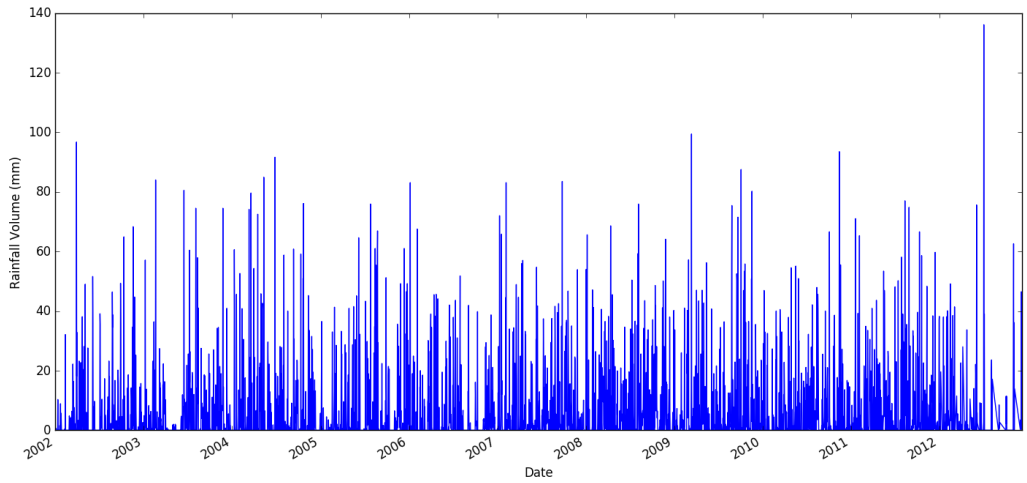

Figure 5.1: Rainfall Trend (2002-2012)

Figure 5.1 shows that in most days, Sg. Lui experienced no rainfall and on average the volume for a rainy day is approximately less than 20 mm per day. It is also observed that the highest volume of rainfall would reach to an amount that is close to almost 100 mm per day for at least once a year. However in year 2012, the highest amount of rainfall volume detected is close to 140 mm. A summarized information of the data is shown in Table 5.1.

Table 5.1: Statistical Properties of Collected Rainfall Volume Dataset

Dataset	E[X]	Std. Dev (X)	Count	Min	Max
Original (mm)	6.4889	13.42	3846	0	136.10
Transformed (mm)	0.04780	0.09857	3846	0.00013	0.99987

Overall, there are 3846 days worth of rainfall volume data in millimetre (mm) with a minimum rainfall of 0 mm which implies no rainfall and a maximum of 136.1 mm. The average rainfall volume is 6.4889 mm with a standard deviation of 13.4186 mm. However after transforming the data,

it has now a minimum value of 0.00013 mm and a maximum value of 0.99987 mm with an average of 0.04780 mm along with a standard deviation of 0.09857 mm.

From the statistical properties of the data calculated, it can be seen that the low average amount of rainfall obtained is because of the high occurrence of zero rainfall which is more than the number of rainy days during the span of 10 years. Besides that, there are also days when the rainfall volumes are higher than 100 mm based on Figure 5.1 which are a lot higher than the average rainfall calculated. Such fluctuations in rainfall greatly affects the standard deviation causing it to be relatively high due to the high volatility of the dataset. Hence, it can be concluded that in the region of Sg Lui, Hulu Langat, Selangor, it experiences higher number of non-rainy days and the average volume of rainfall collected is relatively low. Furthermore, the high volatility of the rainfall volume will cause difficulty in estimating the rainfall amount on a daily basis.

As the data is split to train and test data at 80% and 20% respectively, there are 3077 data to fit the models and another 769 data to be tested using the model selection criteria mentioned. However, in order to test the distribution using the K-S test, the cumulative step-up function needs to be calculated for each values and the data needs to be re-arranged with duplicates being removed. Thus, there are only 209 values left after the re-arrangement. The K-S test critical point,$D_\alpha(N)$ at $\alpha = 5\%$ significance level is obtained from the following formula (see Massey Jr (1951)):

$$D_{0.05}(N) = \frac{1.36}{\sqrt{N}} \ ,$$

where N is the number of samples.

Therefore, the critical point at 5% level of significance with 209 sample size is 0.093941 (i.e. $D_{0.05}(209) = 0.093941$).

Under the empirical studies, the parameter estimation of the 2 proposed distributions (generalized Beta and mixture of 2 modified Log-Normal) including the other models (Beta, Gauss Hyopergeometric, Exponential and Gamma) mentioned as well as simulation processes were conducted. The discussion of the results will be presented in the next section.

5.2 Results and Discussion

The empirical studies in this research were carried out in two groups which are the Beta family and Skew Normal family. As mentioned above, generalized Beta, Beta and Gauss Hypergeometric distributions will be compared under the Beta family distributions. Meanwhile, the mixture of 2 modified Log-Normal, Exponential and Gamma distributions will be compared under the Skew Normal family distributions. In Section 5.2.1, the Beta family distributions (Distr.) results consisting of the estimated parameters (Par.), K-S test, AIC, RMSE together with VaR values will be summarized in Table 5.2 and illustrated graphically in Figure 5.2. In Section 5.2.2, the Skew Normal family distributions results will be summarized in Table 5.3 and graphed in Figure 5.3.

5.2.1 Beta Family Results and Discussion

Table 5.2: Summary of Beta Family Fittings and Selection Criteria

Distr.	Generalized Beta	***Beta**	Gauss Hypergeometric
Par.	$\gamma = 0.4756, \rho = 5.2754,$ $\alpha = 0.5010, \beta = 0.3801,$ $\sigma = 6.6201, z = 0.2850$	$\boldsymbol{a = 0.4853,}$ $\boldsymbol{b = 3.5419}$	$a = 0.4615,\ b = 4.2182,$ $c = -0.3754,\ z = 8.4294$
K-S	0.4631	**0.4632**	0.4630
AIC	3257.17	**3235.68**	3239.81
RMSE (mm)	0.1914	**0.1822**	0.1894
VaR (mm)	0.6172	**0.6333**	0.6238

* represents the best performed model

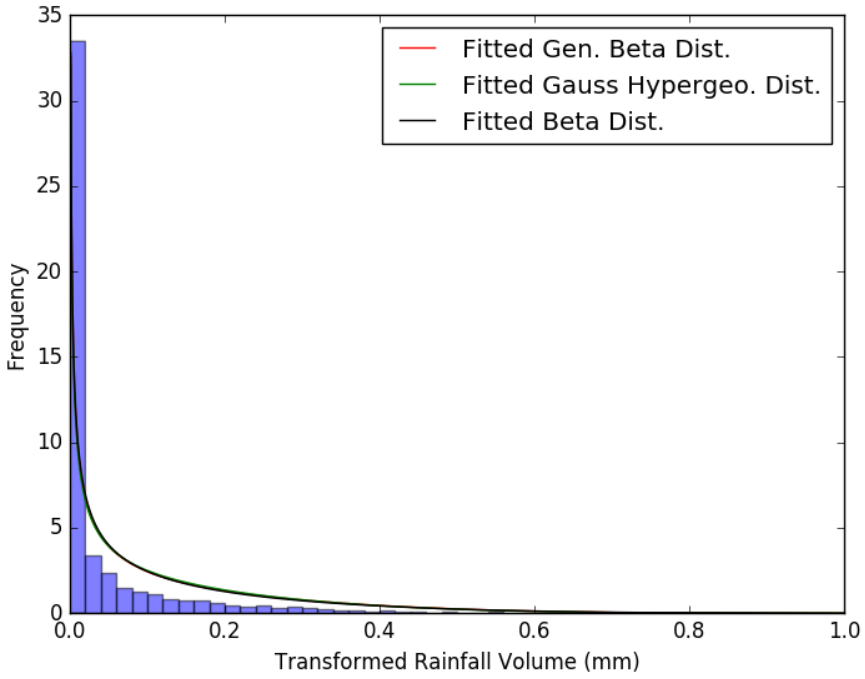

Figure 5.2: Fitting of Beta Family Distributions to Rainfall Volume

From Figure 5.2, it is observed that all the selected Beta family distributions fit do not touch most of the mid-points of the rainfall volume histogram. The three Beta family distributions seem to fit poorly towards rainfall volume data. Nonetheless, the overlapping lines in the graph show that the distributions have comparable estimation strength and fitting capability. Unfortunately, the performance of each distribution could not be clearly identified from the graph alone due to insufficient statistical evidence as well as their similar estimation strength. Therefore, the model selection criteria results calculated in Table 5.2 needs to be referred.

Based on Table 5.2, the K-S test statistic values for the generalized Beta, Beta and Gauss Hypergeometric distributions are 0.4631, 0.4632 and 0.4630 respectively. The values stated are all higher than the critical point, $(D_{0.05}(209) = 0.093941741)$ which shows that there is significant evidence to conclude that the null hypothesis is rejected. This mean the rainfall volume data do not follow the three specified statistical distributions. Although the results are unfavourable, it is still important to identify which model among the Beta family distributions has the better fit by referring to the AIC values.

The AIC of the generalized Beta, Beta and Gauss Hypergeometric distributions are 3257.17, 3235.68 and 3239.81 respectively. Based on the AIC values, it is concluded that the Beta distribution is the best performed model as it has the lowest AIC among the models compared. This is followed by the Gauss Hypergeometric distribution and lastly the generalized Beta distribution.

Meanwhile, the calculated RMSE values for the generalized Beta, Beta and Gauss Hypergeometric distributions are 0.1914 mm, 0.1822 mm and 0.1894 mm respectively which is approximately 0.19 mm on average. These values indicate that the distributions have comparable estimation strength towards rainfall volume which is consistent with the illustration made in Figure 5.2. From the Beta family's properties, it is known that x-variable's constraint is between 0 and 1 ($0 < x < 1$) and the minimum as well as maximum values after the data is scaled is 0.00013 mm and 0.999987 mm respectively. An approximate 0.19 mm RMSE implies that the simulated values differ from the actual values by almost 20% of the data range which is relatively quite large.

From the graph and results of the statistical tests, it is concluded that the selected Beta family distributions perform poorly in fitting rainfall volume dataset with the Beta distribution being the better fitted model among the models compared.

5.2.2 Skew Normal Family Results and Discussion

Table 5.3: Summary of Skew Normal Family Fittings and Selection Criteria

Distr.	***Mixture of 2 modified Log-Normal**	Exponential	Gamma
Par.	$\boldsymbol{p = 0.3871}$ $\boldsymbol{\mu_1 = 3.2317, \mu_2 = 1.419,}$ $\boldsymbol{\sigma_1 = 0.6226, \sigma_2 = 1.2671,}$ $\boldsymbol{c_1 = 0.1881, c_2 = 0.0648}$	$\beta = 18.0063$	$\alpha = 0.8003,$ $\theta = 22.7681$
K-S	**0.07967**	0.07674	0.08268
AIC	**3973.37**	4037.35	4011.84
RMSE (mm)	**15.92**	16.00	16.26
VaR (mm)	**43.85**	40.30	44.10

* represents the best performed model

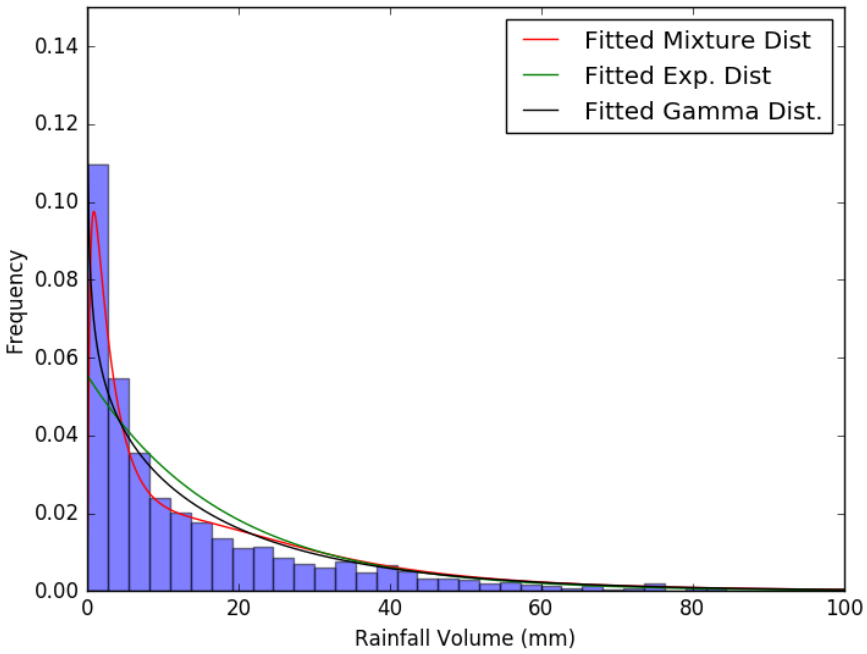

Figure 5.3: Fitting of Skew Normal Continuous Family Distributions to Rainfall Volume

From Figure 5.3, it could be seen that all the selected fitted distributions pass through most of the histogram's mid-points. This indicates a good performance in fitting rainfall volume data by the selected Skew Normal family models. Furthermore, it could also be observed that the estimation strength and fitting capability of the three distributions are quite similar as the lines of the distributions are close to each other. However, the performance of the distributions could not be concluded from the graph and the model selection criteria are needed.

Based on Table 5.3, the calculated K-S test statistics for the mixture of 2 modified Log-Normal, Exponential and Gamma distributions are 0.07967, 0.07674 and 0.08268 respectively. The values are lesser than the critical point of the K-S test statistic at 5% significant level ($D_{0.05}(209) = 0.093941741$). Hence, the null hypothesis is not rejected. There is insufficient evidence to conclude that the rainfall volume dataset do not follow the three distributions specified at 5% level of significance.

In order to identify which model is the better fit, the AIC values are computed. By referring to the AIC values in Table 5.3, the best performed model among the three distributions for this dataset is the proposed mixture of 2 modified Log-Normal distributions with an AIC value of 3973.37 followed by Gamma with 4011.84 and lastly Exponential with 4037.35 AIC values.

In the meantime, the RMSE values calculated in Table 5.3 for the mixture of 2 modified Log-Normal, Exponential and Gamma distributions

are 15.9191 mm, 16.0464 mm and 16.2641 mm respectively which are approximately 16 mm. This suggests that the three distributions have similar estimation strength and fitting capability which supports Figure 5.3 illustration. For the Skew Normal family distributions, their x-variable constraint is $x > 0$ while the minimum and maximum rainfall volume are 0 mm and 136.1 mm respectively. As the RMSE values of the distributions are close to 16 mm, it shows that the simulated values differ from the actual values by about 12% of the data range which is reasonably minimal.

In a nutshell, from the graph illustrated as well as test statistics and model selection criteria calculated, the Skew Normal family distributions performed well in fitting rainfall volume data with the proposed mixture distribution being the better fitted model in this study.

5.2.3 Summary of Results and Discussion

To summarize, the three distributions (generalized Beta, Beta and Gauss Hypergeometric) under the Beta family seem to fit poorly to rainfall volume data with the proposed distribution being the worst performed. On the other hand, the Skew Normal family distributions (mixture of 2 modified Log-Normal, Exponential and Gamma) seem to fit well to rainfall data with the proposed mixture distribution being the best performed model among the three distributions. The comparisons of the models are done within their family distribution as different transformation methods were used on both the data and models due to the different x-variable constraints. Therefore, comparison among the two family distributions could not be done.

Although it was mentioned that the proposed distributions are very flexible and versatile that could provide a good description to different types of data, the generalized Beta distribution showed otherwise. This conclusion occurred might due to the complication in fitting the ${}_3F_2$ and ${}_2F_1$ Hypergeometric function that is part of the generalized Beta together with the Gauss Hypergeometric distributions as they are able to model complex numbers too. Other parameter estimation methods might need to be studied and higher level of computing power is needed to be considered in order to fit such complex distributions in this field of study.

In addition to the various model selection criteria used to analyse the fitting capability and estimation strength of the models towards rainfall volume data, the VaR at 95% confidence level was also calculated. The VaR for the Beta family generalized Beta, Beta and Gauss Hypergeometric distributions are 0.6172 mm, 0.6333 mm and 0.6238 mm respectively. On the other hand, the VaR for the Skew Normal family mixture of 2 modified Log-Normal, Exponential and Gamma distributions are 43.85 mm, 40.30 mm and 44.10 mm respectively. The values represents that at 95% confidence level, the daily rainfall volume will not exceed the respective values calculated within a certain period of time. If the daily rainfall volume were to exceed the amount stated, this indicates that there is a chance of flood or heavy downpour occurrences at 5% confidence level. The results discussed can be extended further for a more in-depth analysis which might be useful in other field of studies.

The conclusion as well as future research work that could be proceeded on in this study will be discussed in the following chapter.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this research, the properties of the two proposed models, the generalized Beta and also the mixture of 2 modified Log-Normal distributions are derived. The properties for the 6-parameter generalized Beta distribution are developed with a lot of emphasis being placed on the contiguous relation function by Rakha et al. (2011) and the table of equations by Gradshteyn and Ryzhik (2014). However, the general form of the CDF for the generalized Beta distribution is unable to be derived as it contains the incomplete Beta function ratio (Natrella, 2010). On the other hand, the mixture of 2 modified Log-Normal distributions' properties were derived using multiple substitutions and reference towards various theorems such as the Owen's T-function, Skew Normal and Log-Normal distributions' concepts along with Lemma 2.1 from Brown (2001).

For the empirical studies, they are done using simulation and distribution fitting that is applied to rainfall volume data. The simulations were done using Accept-Reject algorithm while the distributions were fitted using MLE method and the results were concluded based on K-S test, AIC and RMSE model selection criteria. The results concluded that Beta family distributions seem to perform poorly in fitting rainfall volume data with Beta distribution being the better fit, followed by Gauss Hypergeometric distribution and the proposed generalized Beta distribution. However, the Skew Normal family distributions seem to fit well to rainfall volume with the mixture of 2 modified Log-Normal

distributions being the best performed model followed by Gamma and then Exponential distributions for this dataset.

From the conclusion stated, the proposed mixture distribution fitted better than the other two classical Skew Normal family distributions is inline with the conlusion made by Chuah (2016). Nonetheless, the poor fitting of the 6-parameter generalized Beta distribution among the three Beta family distributions contradicts with the conclusion made by Chuah (2016). It was mentioned that the generalized Beta distribution fit better than it's sub-distributions. The contradiction occurred might be due to the different data massaging and distribution transformation method, statistical software as well as parameter estimation method used. Although the conclusions are formed through the fitting of the distribution to only one rainfall dataset, they could be fitted to more datasets from different areas for a better support towards the conclusion made.

For future research work, empirical studies could be done on other fields aside from rainfall analysis to observe how well does the proposed distributions fit to different areas containing different shapes of data through parameter estimation, simulation as well as model evaluation metrics and selection criteria. The distributions could be applied to areas such as income, stock returns and regression analysis (McDonald and Xu, 1995), health and education data (Sarabia et al., 2014) or even insurance claims data (Eling, 2012). As the applied distributions in the mentioned studies relate to the proposed distributions, it became of great interest to study how well does a general form (6-parameter generalized Beta and mixture of 2 modified Log-Normal) fit to the field of studies mentioned.

Furthermore, other parameter estimation methods such as MOM or even LSM could be used and compared with the MLE to identify the most suitable method that can produce the most accurate estimates for the parameters depending on the type of data. Besides that, the VaR values calculated can be extended for future work where more analysis could be done to improve the accuracy of the estimation on rainfall volume frequency. The results might be important in other field of studies such as flood insurance coverage, geographical or even marine life research.

LIST OF REFERENCES

Adiku, S., Dayananda, P., Rose, C. and Dowuona, G., 1997. An analysis of the within-season rainfall characteristics and simulation of the daily rainfall in two savanna zones in Ghana. *Agricultural and Forest Meteorology*, 86(1-2), pp. 51–62.

Alexander, C., Cordeiro, G. M., Ortega, E. M. and Sarabia, J. M., 2012. Generalized beta-generated distributions. *Computational Statistics & Data Analysis*, 56(6), pp. 1880–1897.

Anderson, D. R. and Burnham, K. P., 1999. Understanding information criteria for selection among capture-recapture or ring recovery models. *Bird Study*, 46(sup1), pp. S14–S21.

Armero, C. and Bayarri, M., 1994. Prior assessments for prediction in queues. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43(1), pp. 139–153.

Azzalini, A., 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, pp. 171–178.

Bartoletti, S. and Loperfido, N., 2010. Modelling air pollution data by the Skew-Normal distribution. *Stochastic Environmental Research and Risk Assessment*, 24(4), pp. 513–517.

Bowman, K. and Shenton, L., 1992. Parameter estimation for the Beta distribution. *Journal of Statistical Computation and Simulation*, 43(3-4), pp. 217–228.

Brown, N. D., 2001. *Reliability studies of the Skew Normal distribution.* Master's thesis. University of Maine.

Casella, G., Robert, C. P. and Wells, M. T., 2004. Generalized accept-reject sampling schemes. *Lecture Notes-Monograph Series*, pp. 342–347.

Cavanaugh, J. E., 2009. Model selection: Bayesian information criterion. *Wiley StatsRef: Statistics Reference Online*, .

Chai, T. and Draxler, R. R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), pp. 1247–1250.

Cho, H.-K., Bowman, K. P. and North, G. R., 2004. A comparison of Gamma and Lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied Meteorology*, 43(11), pp. 1586–1597.

Chotikapanich, D., Rao, D. S. P. and Tang, K. K., 2007. Estimating income inequality in China using grouped data and the generalized Beta distribution. *Review of Income and Wealth*, 53(1), pp. 127–147.

Chuah, H. L., 2016. *Statistical models for daily rainfall data: A case study in Selangor, Malaysia*. Master's thesis. Universiti Tunku Abdul Rahman.

Dey, D., 2010. *Estimation of the parameters of Skew Normal distribution by approximating the ratio of the Normal density and distribution functions*. PhD thesis. UC Riverside.

Eling, M., 2012. Fitting insurance claims to skewed distributions: Are the Skew-Normal and Skew-Student good models?. *Insurance: Mathematics and Economics*, 51(2), pp. 239–248.

Elmahdy, E. E. and Aboutahoun, A. W., 2013. A new approach for parameter estimation of finite Weibull mixture distributions for reliability modeling. *Applied Mathematical Modelling*, 37(4), pp. 1800–1810.

Erick, W. M., Kimutai, K. A. and Njenga, E. G., 2016. Parameter estimation of Kumaraswamy distribution based on progressive Type II Censoring Scheme using expectation-maximization algorithm. *American Journal of Theoretical and Applied Statistics*, 5(3), pp. 154–161.

Fisher, R. A., 1925. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 22. Cambridge University Press, pp. 700–725.

Fournier, B., Rupin, N., Bigerelle, M., Najjar, D., Iost, A. and Wilcox, R., 2007. Estimating the parameters of a generalized Lambda distribution. *Computational Statistics & Data Analysis*, 51(6), pp. 2813–2835.

George, F. and Ramachandran, K., 2011. Estimation of parameters of Johnson's system of distributions. *Journal of Modern Applied Statistical Methods*, 10(2), pp. 9.

Gradshteyn, I. S. and Ryzhik, I. M., 2014. *Table of integrals, series, and products*. Academic Press.

Hanevik, S., 2016. *Comparing maximum likelihood and generalized method of moments in short term interest rate models*. Master's thesis. The University of Bergen.

Hwang, S. I. and Choi, S. I., 2006. Use of a Lognormal distribution model for estimating soil water retention curves from particle-size distribution data. *Journal of Hydrology*, 323(1-4), pp. 325–334.

Jacob, O., 2013. *From the classical Beta distribution to generalized Beta distributions*. PhD thesis. University of Nairobi.

Johnson, N. L., Kotz, S. and Balakrishnan, N., 1994. *Continuous Univariate Probability Distributions, (Vol. 1)*. John Wiley & Sons.

Johnson, N. L., Kotz, S. and Balakrishnan, N., 1995. *Continuous Univariate Distributions, (Vol. 2)*. John Wiley & Sons.

Johnson, N. L., Kotz, S. and Balakrishnan, N., 1996. *Continuous univariate distributions, Computational Statistics & Data Analysis*. Vol. 36(3-4)1. Wiley.

Karakoca, A., Erisoglu, U. and Erisoglu, M., 2015. A comparison of the parameter estimation methods for bimodal mixture Weibull distribution with complete data. *Journal of Applied Statistics*, 42(7), pp. 1472–1489.

Kateregga, M., Mataramvura, S. and Taylor, D., 2017. Parameter estimation for stable distributions with application to commodity futures log-returns. *Cogent Economics & Finance*, 5(1), pp. 1318813.

Khaleel, M., Ibrahim, N., Shitan, M., Merovci, F. and Rehman, E., 2017. Beta Burr type-x with application to rainfall data. *Malaysian Journal of Mathematical Sciences*, 11, pp. 73–86.

Kumaraswamy, P., 1980. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1-2), pp. 79–88.

Li, X., Wu, Z., Chakravarthy, V. D. and Wu, Z., 2011. A low-complexity approximation to Lognormal sum distributions via transformed Log Skew Normal distribution. *IEEE Transactions on Vehicular Technology*, 60(8), pp. 4040–4045.

Lilliefors, H. W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318), pp. 399–402.

Lima, C. H., Kwon, H.-H. and Kim, J.-Y., 2016. A Bayesian Beta distribution model for estimating rainfall IDF curves in a changing climate. *Journal of Hydrology*, 540, pp. 744–756.

Limpert, E., Stahel, W. A. and Abbt, M., 2001. Log-Normal distributions across the sciences: Keys and clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize Log-Normal distributions, which can provide deeper insight into variability and probability—Normal or Log-Normal: That is the question. *AIBS Bulletin*, 51(5), pp. 341–352.

Lin, G. D. and Stoyanov, J., 2009. The logarithmic Skew-Normal distributions are moment-indeterminate. *Journal of Applied Probability*, 46(3), pp. 909–916.

Martínez-Flórez, G., Vergara-Cardozo, S. and González, L. M., 2013. The family of Log Skew-Normal Alpha-Power distributions using precipitation data. *Revista Colombiana de Estadística*, 36(1), pp. 43–57.

Massey Jr, F. J., 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253), pp. 68–78.

McDonald, J. B., 1984. Some generalized functions for the size distribution of income. *Econometrica: Journal of the Econometric Society*, 52(3), pp. 647–663.

McDonald, J. B. and Butler, R. J., 1987. Some generalized mixture distributions with an application to unemployment duration. *The Review of Economics and Statistics*, 69(2), pp. 232–240.

McDonald, J. B. and Xu, Y. J., 1995. A generalization of the Beta distribution with applications. *Journal of Econometrics*, 66(1-2), pp. 133–152.

Misankova, M., Spuchl'akova, E. and Frajtova-Michalikova, K., 2015. Determination of default probability by loss given default. *Procedia Economics and Finance*, 26, pp. 411–417.

Mitchell, B., 1971. A comparison of Chi-square and Kolmogorov-Smirnov tests. *Area*, pp. 237–241.

Mullahy, J., 1986. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3), pp. 341–365.

Murshed, M. S., Am Seo, Y., Park, J.-S. and Lee, Y., 2018. Use of Beta-P distribution for modeling hydrologic events. *Communications for Statistical Applications and Methods*, 25(1), pp. 15–27.

Myung, I. J., 2003. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1), pp. 90–100.

Nagar, D. K. and Bedoya Valencia, D., 2011. Product and quotient of independent Gauss hypergeometric variables. *Ingeniería y Ciencia*, 7(14), pp. 29–48.

Natrella, M., 2010. *NIST/SEMATECH e-handbook of statistical methods: Beta Distribution*. NIST/SEMATECH.

Newville, M., Stensitzki, T., Allen, D. B. and Ingargiola, A., 2014. LMFIT: Non-Linear Least Square Minimization and Curve-Fitting for Python.

Ng, D., Koh, S., Sim, S. and Lee, M., 2018. The study of properties on generalized Beta distribution. *Journal of Physics: Conference Series*. Vol. 1132. IOP Publishing, p. 012080.

Norden, R., 1972. A survey of maximum likelihood estimation. *International Statistical Review/Revue Internationale de Statistique*, pp. 329–354.

Nwobi, F. N. and Ugomma, C. A., 2014. A comparison of methods for the estimation of Weibull distribution parameters. *Metodoloski Zvezki*, 11(1), pp. 65.

Owen, C. E. B., 2008. *Parameter estimation for the Beta distribution*. Master's thesis. Brigham Young University.

Pakoksung, K. and Takagi, M., 2017. Mixed zero-inflation method and probability distribution in fitting daily rainfal data. *Engineering Journal (Eng. J.)*, 21(2), pp. 63– 80.

Rakha, M. A., Rathie, A. K. and Chopra, P., 2011. On some new contiguous relations for the Gauss hypergeometric function with applications. *Computers & Mathematics with Applications*, 61(3), pp. 620–629.

Sarabia, J. M., Prieto, F. and Jordá, V., 2014. Bivariate beta-generated distributions with applications to well-being data. *Journal of Statistical Distributions and Applications*, 1(1), pp. 15.

Smithson, M. and Verkuilen, J., 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variable.. *Psychological Methods*, 11(1), pp. 54.

Spuchl'akova, E. and Cug, J., 2015. Credit risk and LGD modelling. *Procedia Economics and Finance*, 23, pp. 439–444.

Suhaila, J., Ching-Yee, K., Fadhilah, Y. and Hui-Mean, F., 2011. Introducing the mixed distribution in fitting rainfall data. *Open Journal of Modern Hydrology*, 1(02), pp. 11–22.

Suhaila, J. and Jemain, A., 2008. Fitting the statistical distribution for daily rainfall in peninsular Malaysia based on AIC criterion. *Journal of Applied Sciences Research*, 4(12), pp. 1846–1857.

Suhaila, J. and Jemain, A. A., 2007a. Fitting daily rainfall amount in Malaysia using the Normal transform distribution. *Journal of Applied Sciences*, 7(14), pp. 1880–1886.

Suhaila, J. and Jemain, A. A., 2007b. Fitting daily rainfall amount in Peninsular Malaysia using several types of Exponential distributions. *Journal of Applied Sciences Research*, 3(10), pp. 1027–1036.

Tahir, M. H., Cordeiro, G. M., Alizadeh, M., Mansoor, M., Zubair, M. and Hamedani, G. G., 2015. The odd generalized Exponential family of distributions with applications. *Journal of Statistical Distributions and Applications*, 2(1), pp. 1.

Teimouri, M., Hoseini, S. M. and Nadarajah, S., 2013. Comparison of estimation methods for Weibull distribution. *Statistics*, 47(1), pp. 93–109.

Tong, E. N. C., Mues, C., Brown, I. and Thomas, L. C., 2016. Exposure at default models with and without the credit conversion factor. *European Journal of Operational Research*, 252(3), pp. 910–920.

Tong, E. N., Mues, C. and Thomas, L., 2013. A zero-adjusted Gamma model for mortgage loan loss given default. *International Journal of Forecasting*, 29(4), pp. 548–562.

Willmott, C. J. and Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), pp. 79–82.

Wirjanto, T. S. and Xu, D., 2009. The applications of mixtures of Normal distributions in empirical finance: A selected survey. *Technical report*. Department of Statistics, University of Waterloo, Canada. Working paper.

Wong, R. S.-K., 1994. Model selection strategies and the use of association models to detect group differences. *Sociological Methods & Research*, 22(4), pp. 460–491.

Woolhiser, D. A. and Roldan, J., 1982. Stochastic daily precipitation models: 2. a comparison of distributions of amounts. *Water Resources Research*, 18(5), pp. 1461–1468.

Zhang, Z., 1997. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing*, 15(1), pp. 59–76.

LIST OF PUBLICATIONS

Ng, D., Koh, S., Sim, S. and Lee, M., 2018. The study of properties on generalized beta distribution. *Journal of Physics: Conference Series*. Vol. 1132. IOP Publishing, p. 012080.

APPENDICES

In this section, there are 5 appendices titled Appendix A, B, C, D and E which explains the derivations of the distributions' properties that was shown above in detail.

APPENDIX A

This section expands the ${}_3F_2$ Hypergeometric contiguous relation function which shows a Binomial expansion pattern of 1,1; 1,2,1 and 1,3,3,1. It will be used to derive the generalized form of n recursive functions for the ${}_3F_2$ Hypergeometric function.

$$
\begin{aligned}
{}_2F_1(a,b;c;z) &= \frac{a\,{}_2F_1(a+1,b;c+1;z)-(a-c)\,{}_2F_1(a,b;c+1;z)}{c} \\
&= \frac{a}{c}\Big[\frac{(a+1)\,{}_2F_1(a+2,b;c+2;z)-(a-c)\,{}_2F_1(a+1,b;c+2;z)}{c+1}\Big]- \\
&\quad \frac{a-c}{c}\Big[\frac{a\,{}_2F_1(a+1,b;c+2;z)-(a-c-1)\,{}_2F_1(a,b;c+2;z)}{c+1}\Big] \\
&= \frac{a(a+1)}{c(c+1)}\,{}_2F_1(a+2,b;c+2;z)-2\frac{a(a-c)}{c(c+1)}\,{}_2F_1(a+1,b;c+2;z)+ \\
&\quad \frac{(a-c)(a-c-1)}{c(c+1)}\,{}_2F_1(a,b;c+2;z) \\
&= \frac{a(a+1)}{c(c+1)}\Big[\frac{(a+2)\,{}_2F_1(a+3,b;c+3;z)-(a-c)\,{}_2F_1(a+2,b;c+3;z)}{c+2}\Big]- \\
&\quad \frac{2a(a-c)}{c(c+1)}\Big[\frac{(a+1)\,{}_2F_1(a+2,b;c+3;z)-(a-c-1)\,{}_2F_1(a+1,b;c+3;z)}{c+2}\Big]+ \\
&\quad \frac{(a-c)(a-c-1)}{c(c+1)}\Big[\frac{(a)\,{}_2F_1(a+1,b;c+3;z)-(a-c-2)\,{}_2F_1(a,b;c+3;z)}{c+2}\Big] \\
&= \frac{a(a+1)(a+2)}{c(c+1)(c+2)}\,{}_2F_1(a+3,b;c+3;z)- \\
&\quad 3\frac{a(a+1)(a-c)}{c(c+1)(c+2)}\,{}_2F_1(a+2,b;c+3;z)+ \\
&\quad 3\frac{a(a-c)(a-c-1)}{c(c+1)(c+2)}\,{}_2F_1(a+1,b;c+3;z)- \\
&\quad \frac{(a-c)(a-c-1)(a-c-2)}{c(c+1)(c+2)}\,{}_2F_1(a,b;c+3;z)
\end{aligned}
$$

APPENDIX B

This section explains a detailed step by step derivation of the expected value of X, (E[X]) for the mixture of 2 modified Log-Normal distributions. This includes applying *Lemma* 2.1 which is one of the related theories mentioned in Section 3.2.1.

Expected of X, (E[X])

$$
\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} x f(x) dx \\
&= \int_{0}^{\infty} x(2)[\frac{p}{\sqrt{2\pi}\sigma_1 x} e^{-\frac{(\ln x-\mu_1)^2}{2\sigma_1^2}} \Phi(\frac{c_1}{\sigma_1}(\ln x-\mu_1))+ \\
&\quad \frac{1-p}{\sqrt{2\pi}\sigma_2 x} e^{-\frac{(\ln x-\mu_2)^2}{2\sigma_2^2}} \Phi(\frac{c_2}{\sigma_2}(\ln x-\mu_2))]\ dx
\end{aligned}
$$

By letting $y = \ln x$, then

$dy = \frac{1}{x}dx \quad ; \quad dx = e^y dy$

When x is approaching to infinity, y is approaching to infinity as well. When x is approaches to 0, $\ln 0$ is indefinite. Therefore, as $\ln x$ approaches 0, the value tends to approach to a larger negative number towards negative infinity.

$$
\begin{aligned}
E[X] &= \lim_{m\to\infty}\lim_{n\to 0} 2\int_{n}^{m} e^y[\frac{p}{\sqrt{2\pi}\sigma_1} e^{-\frac{(y-\mu_1)^2}{2\sigma_1^2}} \Phi(\frac{c_1}{\sigma_1}(y-\mu_1))+ \\
&\quad \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-\frac{(lny-\mu_2)^2}{2\sigma_2^2}} \Phi(\frac{c_2}{\sigma_2}(lny-\mu_2))]\ dy \\
&= 2\int_{-\infty}^{\infty} e^y[\frac{p}{\sqrt{2\pi}\sigma_1} e^{-\frac{(y-\mu_1)^2}{2\sigma_1^2}} \Phi(\frac{c_1}{\sigma_1}(y-\mu_1))+ \\
&\quad \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-\frac{(lny-\mu_2)^2}{2\sigma_2^2}} \Phi(\frac{c_2}{\sigma_2}(lny-\mu_2))]\ dy
\end{aligned}
$$

Then, let $a_1 = \frac{y-\mu_1}{\sigma_1}$ and $a_2 = \frac{y-\mu_2}{\sigma_2}$. the following is obtained:

$da_1 = \frac{1}{\sigma_1} dy \quad ; \quad dy = \sigma_1 da_1$ and $da_2 = \frac{1}{\sigma_2} dy \quad ; \quad dy = \sigma_2 da_2$

$$E[X] = 2\int_{-\infty}^{\infty} e^{a_1\sigma_1+\mu_1} \frac{p}{\sqrt{2\pi}\sigma_1} e^{-\frac{a_1^2}{2}} \Phi(c_1 a_1)\sigma_1 \; da_1 + 2\int_{-\infty}^{\infty} e^{a_2\sigma_2+\mu_2} \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-\frac{a^2}{2}} \Phi(c_2 a_2)\sigma_2 \; da_2$$

$$= 2\int_{-\infty}^{\infty} e^{\mu_1} \frac{p}{\sqrt{2\pi}\sigma_1} e^{\frac{-a_1^2+2a_1\sigma_1}{2}} \Phi(c_1 a_1)\sigma_1 \; da_1 + 2\int_{-\infty}^{\infty} e^{\mu_2} \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{\frac{-a^2+2a_2\sigma_2}{2}} \Phi(c_2 a_2)\sigma_2 \; da_2$$

$$= 2\int_{-\infty}^{\infty} e^{\mu_1} \frac{p}{\sqrt{2\pi}\sigma_1} e^{\frac{-(a_1-\sigma_1)^2+\sigma_1^2}{2}} \Phi(c_1 a_1)\sigma_1 \; da_1 + 2\int_{-\infty}^{\infty} e^{\mu_2} \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{\frac{-(a_2-\sigma_2)^2+\sigma_2^2}{2}} \Phi(c_2 a_2)\sigma_2 \; da_2$$

$$= 2[pe^{\mu_1+\frac{1}{2}\sigma_1^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{\frac{-(a_1-\sigma_1)^2}{2}} \Phi(c_1 a_1)\sigma_1 da_1 + (1-p)e^{\mu_2+\frac{1}{2}\sigma_2^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_2} e^{\frac{-(a_2-\sigma_2)^2}{2}} \Phi(c_2 a_2)\sigma_2] \; da_2$$

Now, let b_1=a_1-σ_1 ; where db_1=da_1 and b_2=a_2-σ_2 ; where db_2=da_2

$$E[X] = 2pe^{\mu_1+\frac{1}{2}\sigma_1^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{\frac{-b_1^2}{2}} \Phi(c_1(b_1+\sigma_1))\sigma_1 \; db_1 + 2(1-p)e^{\mu_2+\frac{1}{2}\sigma_2^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_2} e^{\frac{-b_2^2}{2}} \Phi(c_2(b_2+\sigma_2))\sigma_2 \; db_2$$

$$E[X] = 2pe^{\mu_1+\frac{1}{2}\sigma_1^2} E\{\Phi(c_1(b_1+\sigma_1))\} + 2(1-p)e^{\mu_2+\frac{1}{2}\sigma_2^2} E\{\Phi(c_2(b_2+\sigma_2))\}$$

By applying *Lemma* 2.1 , $E\{\Phi(hY+k)\} = \Phi(\frac{k}{\sqrt{1+h^2}})$, where h and k are constants, the following equation is found:

$$E[X] = 2[pe^{\mu_1+\frac{1}{2}\sigma_1^2} \Phi(\frac{c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{\mu_2+\frac{1}{2}\sigma_2^2} \Phi(\frac{c_2\sigma_2}{\sqrt{1+c_2^2}})] \qquad \text{(B.1)}$$

APPENDIX C

This section explains about the relationship of the general moment between the Normal distribution and the Log-Normal distribution.

Moment Functions of Normal and Log-Normal Distributions

The Normal distribution has the following moment generating function:

$$M_x(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

As the moment generating function is calculated using the following formula, $M_x(t) = E[e^{tx}]$ and the Log-Normal distribution as well as the Normal distribution can be related with, $Y = \ln X \implies X = e^Y$, the following could be seen:

$$M_x(t) = E[e^{tX}], \text{ where } X \sim N(\mu, \sigma)$$

$$M_x^{(n)}(t) = E[X^n e^t],$$

$$M_x^{(n)}(0) = E[X^n]$$

$$M_y(t) = E[e^{tY}], \text{ where } \quad Y \sim LN(\mu, \sigma)$$

From $Y = \ln X \implies X = e^Y$ then,

$$M_y(t) = E[X^t]$$

$$\therefore M_x^{(n)}(0) = M_y(t)$$

APPENDIX D

This section shows a detailed explanation on the derivation of the general moment for the Skew Normal distribution which includes the application of*Lemma* 2.1. The first derivation of the moment is also presented in order to derive the expected value of X, E[X].

Moment function of Skew Normal Distribution

$$M_x(t) = E[e^{tx}]$$

$$E[e^{tx}] = 2\int_{-\infty}^{\infty} e^{tx}\phi(\frac{x-\mu}{\sigma})\Phi(\lambda(\frac{x-\mu}{\sigma}))\ dx$$

$$= 2\int_{-\infty}^{\infty} e^{tx}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Phi(\frac{\lambda(x-\mu)}{\sigma})\ dx$$

By letting y = $\frac{x-\mu}{\sigma}$, then

$$dy = \frac{1}{\sigma}\ dx \quad ; \quad dx = \sigma\ dy$$

$$E[e^{tx}] = 2\int_{-\infty}^{\infty} e^{t(y\sigma+\mu)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2}} \Phi(\lambda y)\sigma \; dy$$

$$= 2e^{\mu t} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2+2y\sigma t}{2}} \Phi(\lambda y) \; dy$$

$$= 2e^{\mu t} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-\sigma t)^2+\sigma^2 t^2}{2}} \Phi(\lambda y) \; dy$$

$$= 2e^{\mu t+\frac{1}{2}\sigma^2 t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-\sigma t)^2}{2}} \Phi(\lambda y) \; dy$$

Now, let $z = y - \sigma t$; where $dz = dy$

$$E[e^{tx}] = 2e^{\mu t+\frac{1}{2}\sigma^2 t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \Phi(\lambda(z+\sigma t)) \; dz$$

$$= 2e^{\mu t+\frac{1}{2}\sigma^2 t^2} E\{\Phi(\lambda(z+\sigma t))\}$$

By applying Lemma 2.1, $E\{\Phi(hY+k)\} = \Phi(\frac{k}{\sqrt{1+h^2}})$, where h and k are constants, the following equation is found:

$$M_x(t) = E[e^{tx}] = 2e^{\mu t+\frac{1}{2}\sigma^2 t^2} \Phi(\frac{\lambda\sigma t}{\sqrt{1+\lambda^2}})$$

To calculate E[X], the first differentiation of the moment is needed.

$$M_x'(0) = 2e^{\mu t+\frac{1}{2}\sigma^2 t^2} \phi(\frac{\lambda\sigma t}{\sqrt{1+\lambda^2}})(\frac{\lambda\sigma}{\sqrt{1+\lambda^2}}) + \Phi(\frac{\lambda\sigma t}{\sqrt{1+\lambda^2}})(2e^{\mu t+\frac{1}{2}\sigma^2 t^2})(\mu+\sigma^2 t)|_{t=0}$$

$$\implies E[X] = \mu + \sigma(\frac{\lambda}{\sqrt{1+\lambda^2}})(\sqrt{\frac{2}{\pi}})$$

APPENDIX E

This section shows the derivations to obtain the 1st, 2nd, 3rd and 4th moment of the mixture of 2 modified Log-Normal distributions from its' general moment through their respective substitution.

$$\begin{aligned} M_y(t) &= E[e^{tY}] \\ &= E[X^t] \\ &= 2[pe^{\mu_1 t+\frac{1}{2}\sigma_1^2 t^2}\Phi(\frac{c_1\sigma_1 t}{\sqrt{1+c_1^2}}) + (1-p)e^{\mu_2 t+\frac{1}{2}\sigma_2^2 t^2}\Phi(\frac{c_2\sigma_2 t}{\sqrt{1+c_2^2}})], \end{aligned}$$

where $Y = \ln X$

$$\text{t=1} \implies E[X] = 2[pe^{\mu_1+\frac{1}{2}\sigma_1^2}\Phi(\tfrac{c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{\mu_2+\frac{1}{2}\sigma_2^2}\Phi(\tfrac{c_2\sigma_2}{\sqrt{1+c_2^2}})]$$

$$\text{t=2} \implies E[X^2] = 2[pe^{2\mu_1+2\sigma_1^2}\Phi(\tfrac{2c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{2\mu_2+2\sigma_2^2}\Phi(\tfrac{2c_2\sigma_2}{\sqrt{1+c_2^2}})]$$

$$\text{t=3} \implies E[X^3] = 2[pe^{3\mu_1+\frac{9}{2}\sigma_1^2}\Phi(\tfrac{3c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{3\mu_2+\frac{9}{2}\sigma_2^2}\Phi(\tfrac{3c_2\sigma_2}{\sqrt{1+c_2^2}})]$$

$$\text{t=4} \implies E[X^4] = 2[pe^{4\mu_1+8\sigma_1^2}\Phi(\tfrac{4c_1\sigma_1}{\sqrt{1+c_1^2}}) + (1-p)e^{4\mu_2+8\sigma_2^2}\Phi(\tfrac{4c_2\sigma_2}{\sqrt{1+c_2^2}})]$$