Nilkamal School of Mathematics, Applied Statistics & Analytics, NMIMS

MSc. Statistics and Data Science (2023-25)

FROM BLEND TO BENEFIT

Fitting of mixture distributions on motor insurance claims

Research Mentor : Dr. Pradnya Khandeparkar.

GROUP MEMBERS

NAME	SAP-ID	DEPARTMENT
ROHAN SHAH	86062300001	MSc. STATISTICS AND DATA SCIENCE
AASTHA SHARMA	86062300064	MSc. STATISTICS AND DATA SCIENCE
VERNI SHARMA	86062300020	MSc. STATISTICS AND DATA SCIENCE
ANSHIKA SHARMA	86062300034	MSc. STATISTICS AND DATA SCIENCE
ANIRUDDHA SHELKE	86062300008	MSc. STATISTICS AND DATA SCIENCE

TABLE OF CONTENTS –

- ABSTRACT
- INTRODUCTION
- RATIONALE
- AIM AND OBJECTIVES
- DATA PREPARATION
- METHODOLOGY
- RESULTS AND DISCUSSION
- SUMMARY AND CONCLUSION
- REFERENCES

ABSTRACT

Problem statement: The modeling of claims is an important task of actuaries. Our problem is in modeling the actual motor insurance claim data set. In this study, we show that the actual motor insurance claim can be fitted by a finite mixture model. Approach: Firstly, we analyze the actual data set and then we choose the finite mixture of Lognormal distributions as our model. The estimated parameters of the model are obtained from the EM algorithm. Then, we use the K-S test to show how well the finite mixture Lognormal distributions fit the actual data set. Results: From the tests, we found that the finite mixture lognormal distributions fit the actual data set with a significant level of 0.10. Conclusion: The finite mixture Lognormal distributions can be fitted to motor insurance claims and this fitting is better when the number of components (k) is increasing.

INTRODUCTION

Finite mixtures of distributions have provided a mathematical approach to the statistical modeling of a wide variety of random phenomena. It is an extremely flexible method of modeling and has continued to receive increasing attention over the years from both practical and theoretical point of view. Areas in which mixture models have been successfully applied include astronomy, biology, genetics, medicine, psychiatry and economics. Very little literature is on the applications in general insurance setting. According to the motor insurance is an important branch of non-life insurance in many countries, with contributions amongst the total premium income category. It is a fact that, most insurance claims exhibit some level of clustering, and the usefulness of mixture distribution in modeling heterogeneity in a cluster analysis context is obvious. In practice, most motor insurance claims which occur with losses are modeled by unimodal loss models and . Motor insurance claims with multimodal loss distributions are more advance to apply common unimodal loss models. We therefore extend our knowledge on mixture distributions using finite mixtures of regression models to model such case. Finite mixtures of regression models are a popular method to model unobserved heterogeneity or to account for over dispersion in the claims data. They are flexible models and in theory it is easy to modify and extend them by using more complex models for the component distribution functions and estimate the corresponding parameters. Finite mixture models with a fixed number of components are usually estimated with the expectation-maximization (EM) algorithm within a maximum likelihood framework. Since there are many different modes for claim possibilities, a finite mixture model should work well, and compared (numerically) two approaches to the estimation of the parameters of the component densities in a univariate mixture of normal distributions; one approach is based on a constrained maximum likelihood (ML) algorithm; the other, is on the fuzzy c-means (FCM) clustering algorithm, [8]. Finite mixture models so far include components of the data structure. The purpose of this study is to determine an appropriate finite mixture model for the claims data. The results which can help us determine the expected reserves.

Rationale

Why did we go for Motor Insurance?

The insurance industry plays a crucial role in the economy by providing financial protection against risks. Research in this field can contribute to a better understanding of risk management, financial stability, and economic resilience. This focuses on improving the customer experience which includes studying consumer behavior, developing customer-centric products, and enhancing the claims processing experience.

The insurance industry generates vast amounts of data. Research in data science and analytics can lead to the development of predictive models, fraud detection techniques, and other data-driven innovations that benefit the insurance sector. It provides valuable insights and contributes to the development of expertise in a dynamic and growing industry. The motor insurance industry plays a crucial role in the overall functioning of the economy and society. Reasons being: Legal Requirements, Financial Protection, Protection Against Unforeseen Events Asset Protection, Medical Coverage, etc.

What is a Mixture Distribution?

A mixture distribution is a statistical distribution that is composed of a mixture of two or more component distributions. Each component distribution is associated with a certain probability, and the overall probability distribution is a weighted sum (or mixture) of these components.

$$g(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \ldots + \pi_k f_k(x)$$

- π_i's are the weights
- $f_i(x)$'s are the distributions of the individual components
- it goes from 1 to k.

Why Mixture Distribution on Claims?

1. Claims data are often skewed or non-normal distributions. Mixture models, especially GMMs, are capable of capturing a wide range of distribution shapes, making them suitable for handling diverse claims data
Handling Skewed or Non-Normal Distributions

2. Heterogeneity in Claims Data:
Claims data often exhibits heterogeneity, which means that it may come from different underlying distributions. Mixture models can capture this heterogeneity by representing the data as a combination of multiple distributions.

3. Modeling Complex Distributions:
Claims data may not be easily characterized by a single probability distribution. Mixture models provide a flexible framework for approximating complex data distributions by combining simpler component distributions

4. Fraud Detection
Mixture models can be used for fraud detection in insurance claims. By modeling the normal and abnormal behavior separately, anomalies can be detected more effectively.

5. Predictive Modeling:
Mixture models are useful for predictive modeling in insurance, helping companies anticipate future claim patterns and identify potential risks. This is particularly valuable for strategic planning and resource allocation.

AIM & OBJECTIVES

To provide financial coverage to policyholders in the event of accidental damage, theft, or loss of their vehicles and also to mitigate the financial risks associated with owning and operating a motor vehicle. The following objectives are :

1. Ensure that policyholders receive compensation to repair or replace their vehicles, minimizing the financial impact on them.
2. Help policyholders manage the financial consequences of unexpected events, promoting confidence and security in vehicle ownership.

Data Preparation

Data preparation is a crucial step in the data analysis process, and documenting it thoroughly is essential for the transparency and reproducibility of your work. When including a section on data preparation in your report, consider the following components:

DATA COLLECTION :

Data was obtained from Kaggle

age	policy_state	policy_annual_premium	insured_sex	insured_education_level	incident_dat	incident_type	incident_severity	incident_st	incident_hour_of_the_day	number_of_vehicles_involved	bodily_injuries	witnes	total_claim_amount	injury_claim	p_claim	vehicle_claim	auto_make	auto_year	fraud_reported
48	OH	1406.91	MALE	MD	25/01/15	Single Vehicle Co	Major Damage	SC	5	1	1	2	71610	6510	13020	52080	Saab	2004	Y
42	IN	1197.22	MALE	MD	21/01/15	Vehicle Theft	Minor Damage	VA	8	1	0	0	5070	780	780	3510	Mercedes	2007	Y
29	OH	1413.14	FEMALE	PhD	22/02/15	Multi-vehicle Coll	Minor Damage	NY	7	3	2	3	34650	7700	3850	23100	Dodge	2007	N
41	IL	1415.74	FEMALE	PhD	10/01/15	Single Vehicle Co	Major Damage	OH	5	1	1	2	63400	6340	6340	50720	Chevrolet	2014	Y
44	IL	1583.91	MALE	Associate	17/02/15	Vehicle Theft	Minor Damage	NY	20	1	0	1	6500	1300	650	4550	Accura	2009	N
39	OH	1351.1	FEMALE	PhD	02/01/15	Multi-vehicle Coll	Major Damage	SC	19	3	0	2	64100	6410	6410	51280	Saab	2003	Y
34	IN	1333.35	MALE	PhD	13/01/15	Multi-vehicle Coll	Minor Damage	NY	0	3	0	0	78650	21450	7150	50050	Nissan	2012	N
37	IL	1137.03	MALE	Associate	27/02/15	Multi-vehicle Coll	Total Loss	VA	23	3	2	2	51590	9380	9380	32830	Audi	2015	N
33	IL	1442.99	FEMALE	PhD	30/01/15	Single Vehicle Co	Total Loss	WV	21	1	1	1	27700	2770	2770	22160	Toyota	2012	N
42	IL	1315.68	MALE	PhD	05/01/15	Single Vehicle Co	Total Loss	NC	14	1	2	1	42300	4700	4700	32900	Saab	1996	N
42	OH	1253.12	FEMALE	Masters	06/01/15	Single Vehicle Co	Total Loss	NY	22	1	2	2	87010	7910	15820	63280	Ford	2002	N
61	OH	1137.16	FEMALE	High School	15/02/15	Multi-vehicle Coll	Major Damage	SC	21	3	1	2	114920	17680	17680	79560	Audi	2006	N
23	OH	1215.36	MALE	MD	22/01/15	Single Vehicle Co	Total Loss	SC	9	1	1	0	56520	4710	9420	42390	Saab	2000	N
34	OH	936.61	FEMALE	MD	08/01/15	Parked Car	Minor Damage	SC	5	1	1	1	7280	1120	1120	5040	Toyota	2010	N
38	OH	1301.13	FEMALE	College	15/01/15	Single Vehicle Co	Total Loss	SC	12	1	0	2	46200	4200	8400	33600	Dodge	2003	Y
58	IN	1131.4	FEMALE	MD	29/01/15	Multi-vehicle Coll	Major Damage	WV	12	4	0	0	63120	10520	10520	42080	Accura	1999	Y
26	OH	1199.44	MALE	College	22/02/15	Multi-vehicle Coll	Major Damage	NY	0	3	1	2	52110	5790	5790	40530	Nissan	2012	N
31	IN	708.64	MALE	High School	06/01/15	Single Vehicle Co	Total Loss	WV	9	1	0	2	77880	14160	7080	56640	Suburu	2015	N
37	OH	1374.22	FEMALE	MD	19/01/15	Single Vehicle Co	Total Loss	NY	19	1	1	0	72930	6630	13260	53040	Accura	2015	N
39	IN	1475.73	FEMALE	High School	22/02/15	Multi-vehicle Coll	Major Damage	VA	8	3	2	0	60400	6040	6040	48320	Nissan	2014	N
62	IN	1187.96	MALE	JD	01/01/15	Multi-vehicle Coll	Minor Damage	NY	20	3	1	0	47160	0	5240	41920	Suburu	2011	N
41	IL	875.15	FEMALE	Associate	10/02/15	Multi-vehicle Coll	Total Loss	SC	15	3	1	2	37840	0	4730	33110	Accura	1996	N
55	IL	972.18	MALE	High School	11/01/15	Multi-vehicle Coll	Major Damage	SC	20	3	0	0	71520	17880	5960	47680	Suburu	2000	Y
55	IN	1268.79	MALE	MD	19/01/15	Single Vehicle Co	Total Loss	WV	15	1	2	2	98160	8180	16360	73620	Dodge	2011	Y
40	IN	883.31	MALE	College	24/02/15	Single Vehicle Co	Minor Damage	VA	6	1	1	3	77880	7080	14160	56640	Ford	2005	N
35	OH	1266.92	MALE	Masters	09/01/15	Multi-vehicle Coll	Major Damage	OH	16	3	1	3	71500	16500	11000	44000	Ford	2006	Y
43	IN	1322.1	MALE	High School	28/01/15	Parked Car	Minor Damage	PA	4	1	1	3	9020	1640	820	6560	Toyota	2005	N
34	IN	848.07	MALE	JD	07/01/15	Vehicle Theft	Minor Damage	VA	5	1	2	1	5720	1040	520	4160	Suburu	2003	Y
40	OH	1291.7	FEMALE	JD	08/01/15	Single Vehicle Co	Minor Damage	SC	21	1	1	0	69840	7760	15520	46560	Dodge	2009	N
45	IL	1104.5	FEMALE	PhD	15/02/15	Single Vehicle Co	Minor Damage	SC	5	1	2	2	91650	14100	14100	63450	Accura	2011	N
25	IL	954.16	MALE	Masters	18/01/15	Multi-vehicle Coll	Major Damage	SC	22	4	0	0	75600	12600	12600	50400	Toyota	2005	N
37	IL	1337.28	MALE	JD	28/02/15	Multi-vehicle Coll	Major Damage	WV	10	3	2	2	67140	7460	7460	52220	Ford	2006	Y
35	IL	1088.34	FEMALE	Associate	24/02/15	Multi-vehicle Coll	Total Loss	NY	16	3	2	3	29790	3310	3310	23170	BMW	2008	N
30	IL	1558.29	MALE	High School	09/01/15	Multi-vehicle Coll	Major Damage	NY	1	3	1	2	77110	14020	14020	49070	Suburu	2015	N
37	IL	1415.68	MALE	PhD	12/02/15	Single Vehicle Co	Total Loss	WV	17	1	0	1	64800	10800	5400	48600	Audi	1999	N
33	OH	1334.15	MALE	High School	24/01/15	Single Vehicle Co	Major Damage	WV	15	1	2	0	53100	10620	5310	37170	Mercedes	1995	Y

IDENTIFYING THE VARIABLE :

Total claim Amount&Vehicle Claim Amount (here in dollars)

PREPROCESSING THE DATA:

1) No missing values were found in our data

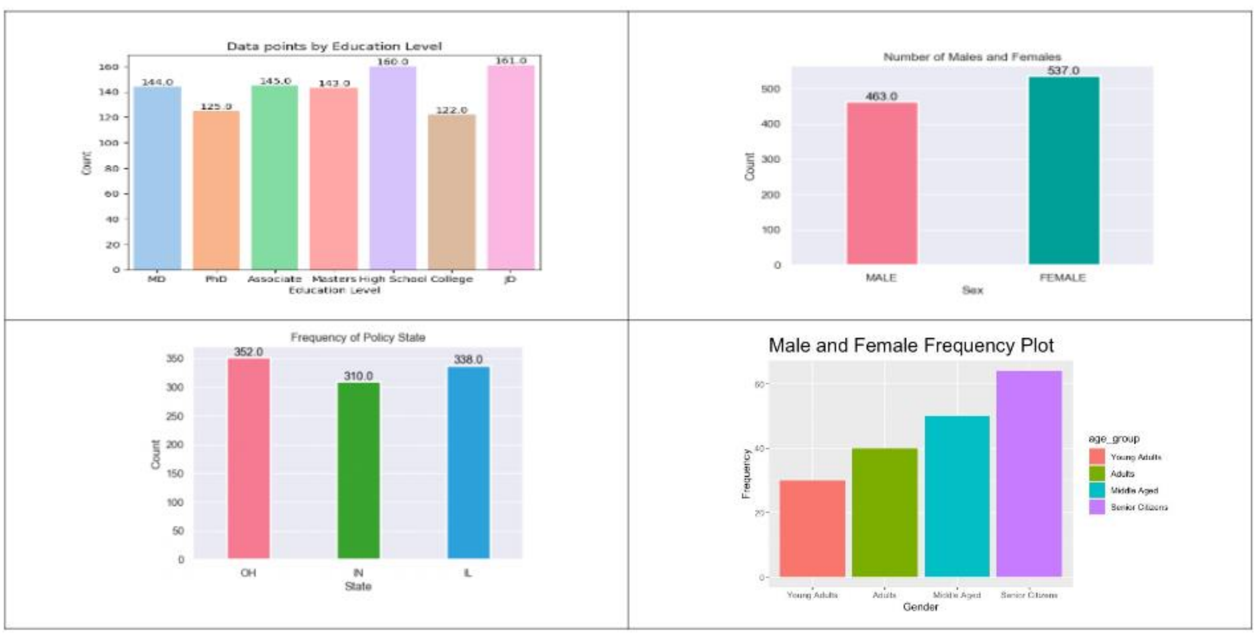
2) Some columns not used in our study were removed ("hobbies of insured" etc.)

3) Investigated bias: Data was not found to be biased (in terms of sex, education level or state variables, etc.)

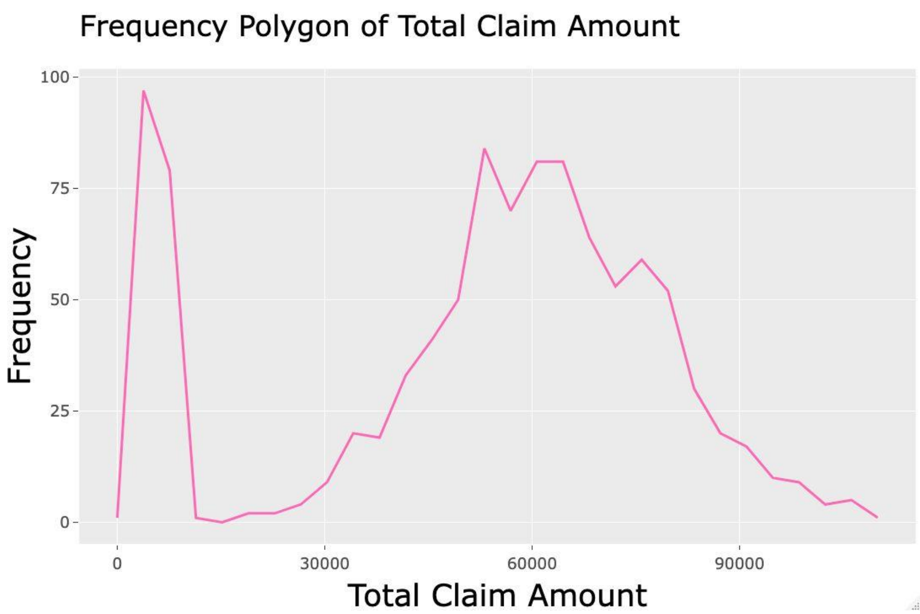
TOOLS AND SOFTWARE USED

R-Programming and Python.

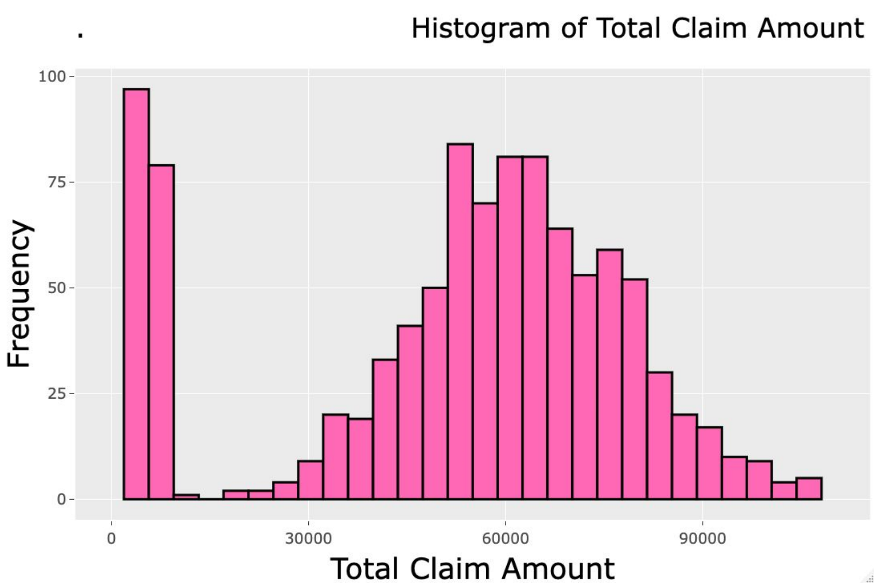
EXPLORATORY DATA ANALYSIS OF THE DATA



Observing the DATA to see if the Mixture Exists taking our target variable which is Total claim amount



.



METHODOLOGY

The principles, processes, and rules that guide our approach for designing and conducting our study are -

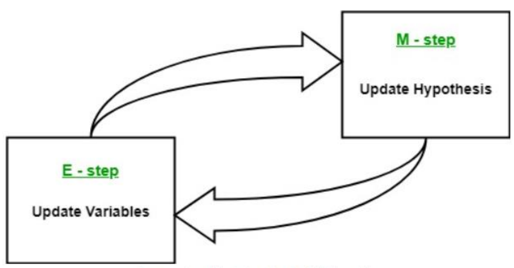
1)There are many algorithms available however literature suggests that the EM algorithm has been the most widely used for mixture distributions for its accuracy

2) Goodness of Fit: The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.

About EM algorithm-

The Expectation-Maximization (EM) algorithm is an iterative optimization method that combines different unsupervised machine learning algorithms to find maximum likelihood or maximum posterior estimates of parameters in statistical models that involve unobserved latent variables. The EM algorithm is commonly used for latent variable models and can handle missing data. It consists of an estimation step (E-step) and a maximization step (M-step), forming an iterative process to improve model fit.

- In the E step, the algorithm computes the latent variables i.e., the expectation of the log-likelihood using the current parameter estimates.
- In the M step, the algorithm determines the parameters that maximize the expected log-likelihood obtained in the E step, and corresponding model parameters are updated based on the estimated latent variables.



Expectation-Maximization in EM Algorithm

By iteratively repeating these steps, the EM algorithm seeks to maximize the likelihood of the observed data. It is commonly used for unsupervised learning tasks, such as clustering, where latent variables are inferred and has applications in various fields, including machine learning, computer vision, and natural language processing.

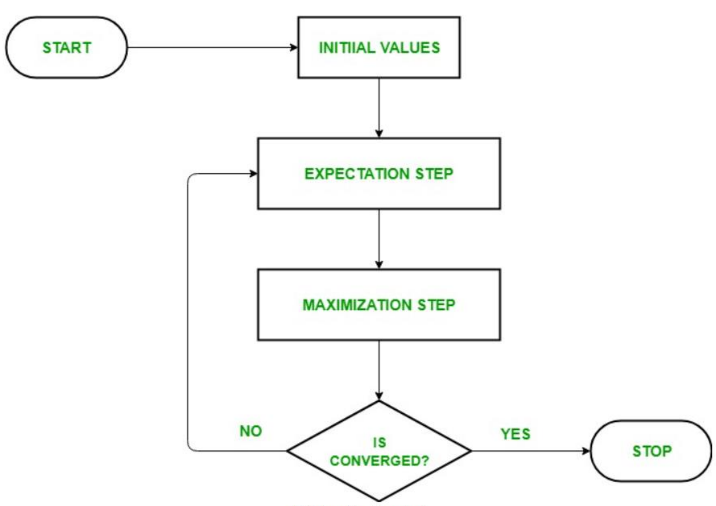
Key Terms in Expectation-Maximization (EM) Algorithm

Some of the most commonly used key terms in the Expectation-Maximization (EM) Algorithm are as follows:

- **Latent Variables:** Latent variables are unobserved variables in statistical models that can only be inferred indirectly through their effects on observable variables. They cannot be directly measured but can be detected by their impact on the observable variables.
- **Likelihood:** It is the probability of observing the given data given the parameters of the model. In the EM algorithm, the goal is to find the parameters that maximize the likelihood.
- **Log-Likelihood:** It is the logarithm of the likelihood function, which measures the goodness of fit between the observed data and the model. EM algorithm seeks to maximize the log-likelihood.

- **Maximum Likelihood Estimation (MLE)**: MLE is a method to estimate the parameters of a statistical model by finding the parameter values that maximize the likelihood function, which measures how well the model explains the observed data.
- **Posterior Probability**: In the context of Bayesian inference, the EM algorithm can be extended to estimate the maximum a posteriori (MAP) estimates, where the posterior probability of the parameters is calculated based on the prior distribution and the likelihood function.
- **Expectation (E) Step**: The E-step of the EM algorithm computes the expected value or posterior probability of the latent variables given the observed data and current parameter estimates. It involves calculating the probabilities of each latent variable for each data point.
- **Maximization (M) Step**: The M-step of the EM algorithm updates the parameter estimates by maximizing the expected log-likelihood obtained from the E-step. It involves finding the parameter values that optimize the likelihood function, typically through numerical optimization methods.
- **Convergence:** Convergence refers to the condition when the EM algorithm has reached a stable solution. It is typically determined by checking if the change in the log-likelihood or the parameter estimates falls below a predefined threshold.

- **How the Expectation-Maximization (EM) Algorithm Works:**
  The essence of the Expectation-Maximization algorithm is to use the available observed data of the dataset to estimate the missing data and then use that data to update the values of the parameters. Let us understand the EM algorithm in detail.



EM Algorithm Flowchart

1. **Initialization:**
   - Initially, a set of initial values of the parameters are considered. A set of incomplete observed data is given to the system with the assumption that the observed data comes from a specific model.
2. **E-Step (Expectation Step):** In this step, we use the observed data in order to estimate or guess the values of the missing or incomplete data. It is basically used to update the variables.

- Compute the posterior probability or responsibility of each latent variable given the observed data and current parameter estimates.
- Estimate the missing or incomplete data values using the current parameter estimates.
- Compute the log-likelihood of the observed data based on the current parameter estimates and estimated missing data.

E-step: compute

expectation of log of P(x|z)

$$E_{z|x,\theta^{(t)}}\left[\log\left(p(\mathbf{x},\mathbf{z}\mid\theta)\right)\right] = \sum_{\mathbf{z}} \log\left(p(\mathbf{x},\mathbf{z}\mid\theta)\right) p\left(\mathbf{z}\mid\mathbf{x},\theta^{(t)}\right)$$

3. **M-step (Maximization Step):** In this step, we use the complete data generated in the preceding "Expectation" step in order to update the values of the parameters. It is basically used to update the hypothesis.
   - Update the parameters of the model by maximizing the expected complete data log-likelihood obtained from the E-step.
   - This typically involves solving optimization problems to find the parameter values that maximize the log-likelihood.
   - The specific optimization technique used depends on the nature of the problem and the model being used.

M-step: solve

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} \sum_{\mathbf{z}} \log\left(p(\mathbf{x},\mathbf{z}\mid\theta)\right) p\left(\mathbf{z}\mid\mathbf{x},\theta^{(t)}\right)$$

4. **Convergence**: In this step, it is checked whether the values are converging or not, if yes, then stop otherwise repeat *step-2* and *step-3* i.e., "Expectation" – step and "Maximization" – step until the convergence occurs.
   - Check for convergence by comparing the change in log-likelihood or the parameter values between iterations.
   - If the change is below a predefined threshold, stop and consider the algorithm converged.
   - Otherwise, go back to the E-step and repeat the process until convergence is achieved.

Kolmogorov-Smirnov -

Kolmogorov–Smirnov Test is a completely efficient manner to determine if two samples are significantly one of a kind from each other. It is normally used to check the uniformity of random numbers. Uniformity is one of the maximum important properties of any random number generator and the Kolmogorov–Smirnov check can be used to check it. The Kolmogorov–Smirnov take a look at can also be used to check whether or not two underlying one-dimensional opportunity distributions differ. It is a totally green manner to determine if two samples are substantially distinct from each other. The Kolmogorov–Smirnov statistic quantifies the gap between the empirical distribution function of the pattern and the cumulative distribution feature of the reference distribution, or among the empirical distribution functions of samples.

How Kolmogorov-Smirnov test works?

To answer this first we need to discuss the purpose of using this test. The main idea behind using this test is to check whether the two samples that we are dealing with follow the same type of distribution or if the shape of the distribution is the same or not.

First of all, if we assume that the shape or the probability distribution of the two samples is the same then the maximum value of the absolute difference between the cumulative probability distribution difference between the two functions will be the same. And higher the value the difference between the shape of the distribution is high.
The hypothesis taken –

H0: The sample is drawn from the reference distribution. H1: The sample is not drawn from the reference distribution.

The formula for Test statistic for K-S test

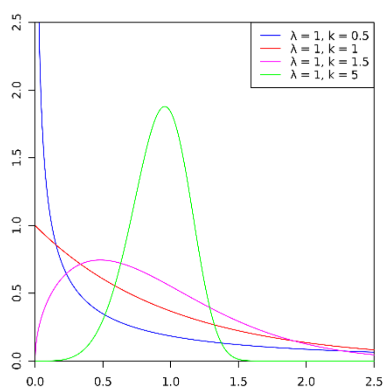
$$D = \max(|F_n(x) - F(x)|)$$

Distributions Investigated-

1) Weibull distribution

$$f(x) = \frac{\gamma}{\alpha}\left(\frac{(x-\mu)}{\alpha}\right)^{\gamma-1} exp^{(-(\frac{(x-\mu)}{\alpha})^{\gamma})} \quad x \geq \mu; \gamma, \alpha > 0$$

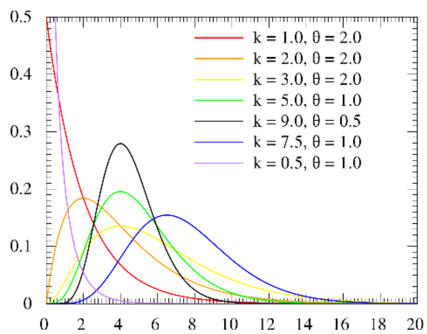
- γ is the **shape parameter**, also called as the Weibull slope or the threshold parameter.
- α is the **scale parameter**, also called the characteristic life parameter.
- μ is the **location parameter**, also called the waiting time parameter or sometimes the shift parameter.



2) Gamma Distribution

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

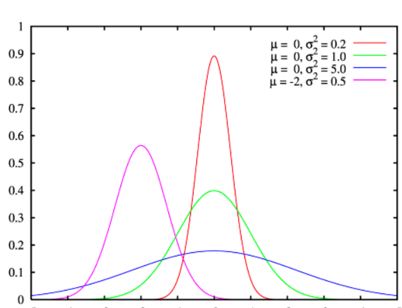
- x is the random variable.
- α is the shape parameter (also known as the "k shape parameter").
- β is the rate parameter (sometimes called the "theta scale parameter").
- $\Gamma(\alpha)$ is the gamma function evaluated at α.



3) Normal Distribution

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

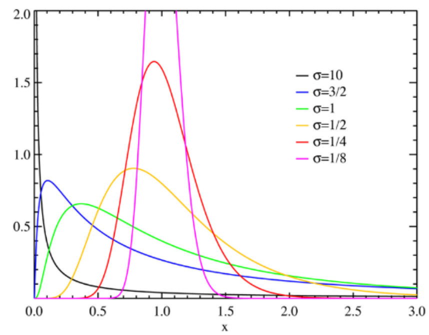
- x is the random variable.
- μ is the mean of the distribution.
- σ is the standard deviation of the distribution.
- π is the mathematical constant pi (approximately 3.14159).
- exp(·) represents the exponential function



4) Log normal distribution

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$$

- *x* is the random variable.
- μ is the mean of the natural logarithm of the distribution.
- σ is the standard deviation of the natural logarithm of the distribution.
- ln(*x*) denotes the natural logarithm of
- *x*.



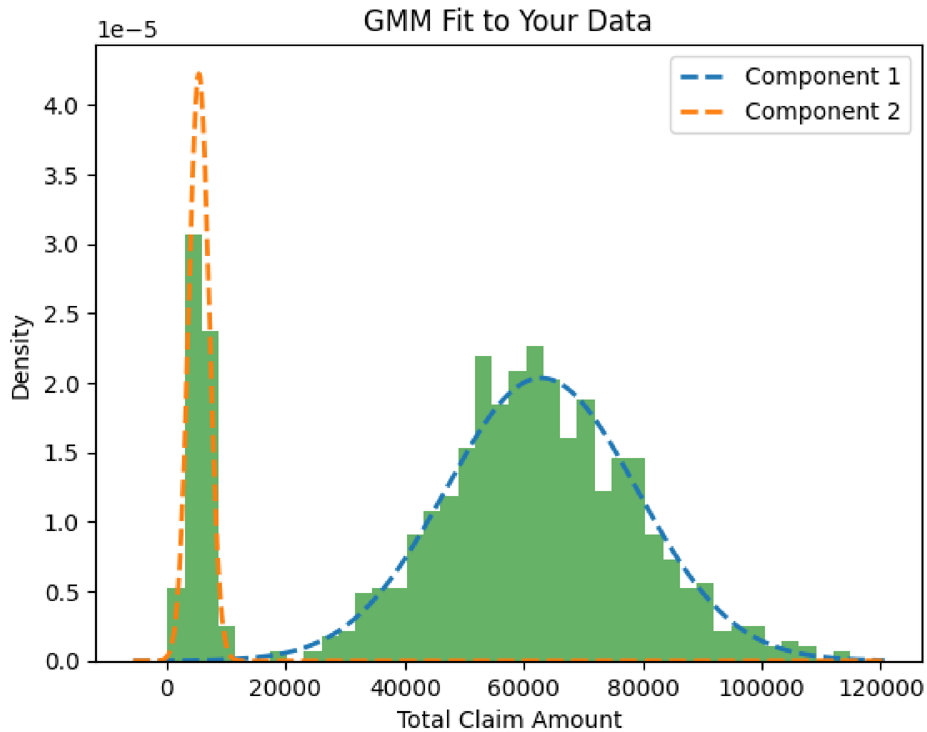
Results and Discussion

Theory suggests that the claim amount of data will tend to follow distributions that are positively skewed since the claim amount cannot take negative values

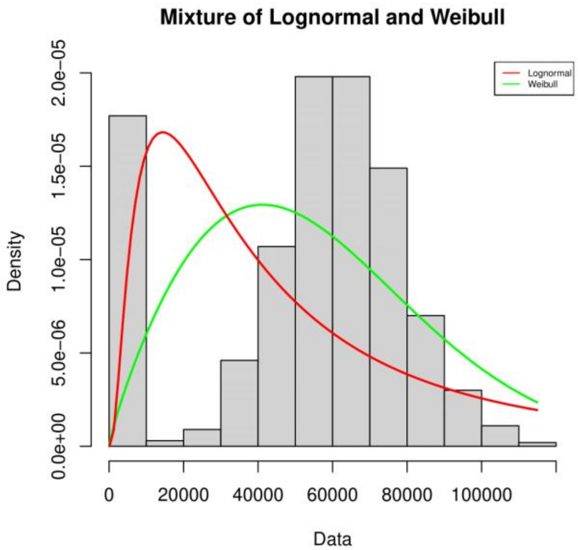
Hence, we have considered the distributions:

- Weibull
- Gamma
- Lognormal
- Normal (based on initial investigation)

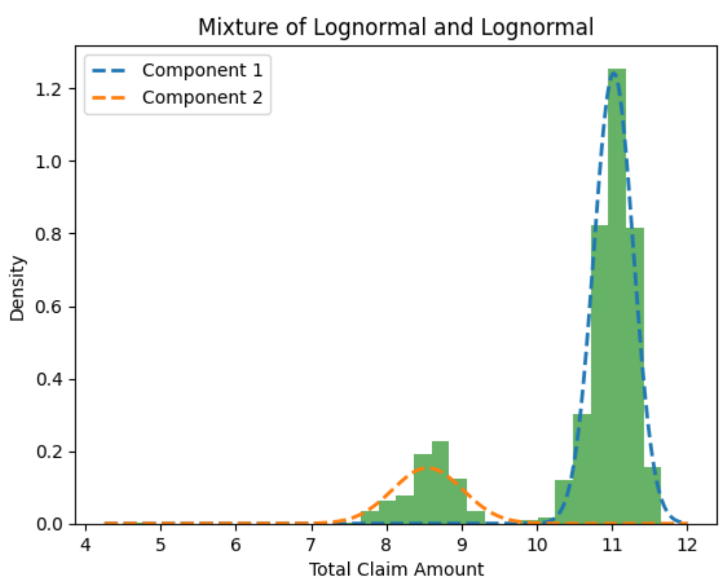
After plotting the following Fit plots –



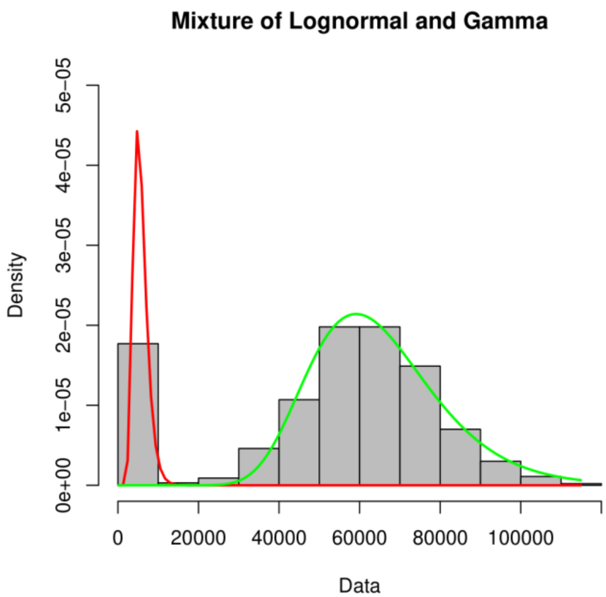
Mixture Model	Parameters	Values	Proportion(%)		K-S Test (p value)	AIC Score
			w1	w2		
Normal + Normal	μ_1	5408.94	17.738	82.26	0.0002	32.3397
	σ_1	1703.72				
	μ_2	62972.73				
	σ_2	16065.48				



Mixture Model	Parameters	Values	Proportion(%)		K-S Test (p value)	AIC Score
			w1	w2		
Log normal+Weibull	shape	1.84E+00	18.76	81.24	0.280001	22498.87
	scale(σ)	6.30E+07				
	μ	10.57375				
	σ	0.022336763				



Mixture Model	Parameters	Values	Proportion(%)		K-S Test (p value)	AIC Score
			w1	w2		
Log normal + Log normal	μ	8.5474	18.04	81.954	0.64781	11.3175
	σ	0.469				
	μ	11.01995				
	σ	0.26323				



Mixture Model	Parameters	Values	Proportion(%)		K-S Test (p value)	AIC Score
			w1	w2		
Lognormal+Gamma	shape	21365.48	16.95	83.08	0.3616896	45528.4
	scale(σ)	1039.003				
	μ	19.4821				
	σ	0.21195				

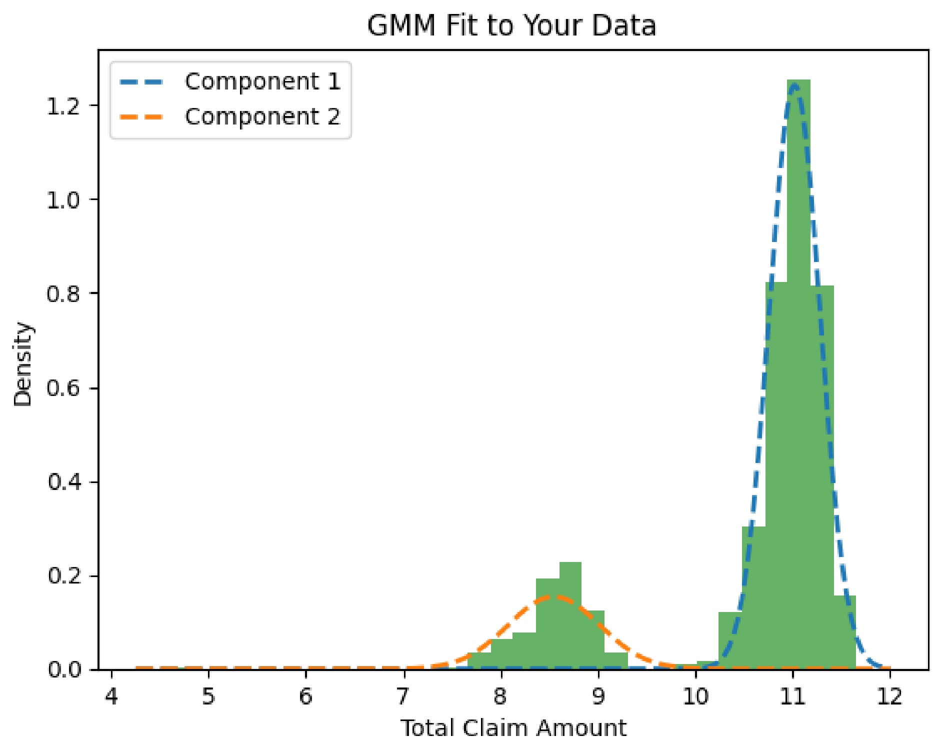
ESTIMATION RESULTS

Here we have the following table for the estimation results

Mixture Model	Parameters	Values	Proportion(%)		K-S Test (p value)	AIC Score
			w1	w2		
Normal + Normal	μ_1	5408.94	17.738	82.26	0.0002	32.3397
	σ_1	1703.72				
	μ_2	62972.73				
	σ_2	16065.48				
Log normal + Log normal	**μ**	**8.5474**	**18.04**	**81.954**	**0.64781**	**11.3175**
	σ	**0.469**				
	μ	**11.01995**				
	σ	**0.26323**				
Gamma+Lognormal	shape	21365.48	16.95	83.08	0.3616896	45528.4
	scale(σ)	1039.003				
	μ	19.4821				
	σ	0.21195				
Weibull + Log normal	shape	1.84E+00	18.76	81.24	0.280001	22498.87
	scale(σ)	6.30E+07				
	μ	10.57375				
	σ	0.022336763				

AIC - Akaike Information Criterion (AIC) is a single number score that can be used to determine which of multiple models is most likely to be the best model for a given data set. It estimates the models relatively.

Mixture of Lognormal-



	Claim Amount Data	Log Transformed Data	Probability for distribution 1	Probability for distribution 2
0	71610.0	11.178990	1.000000e+00	2.168782e-08
1	5070.0	8.531096	3.137189e-19	1.000000e+00
2	34650.0	10.453053	9.996728e-01	3.271791e-04
3	63400.0	11.057219	9.999999e-01	7.572314e-08
4	6500.0	8.779557	1.705038e-15	1.000000e+00

So, with the help of GMM model in Python we observed that X~Normal and Y follows lognormal
820 points follow normal and 120 points follow log-normal.

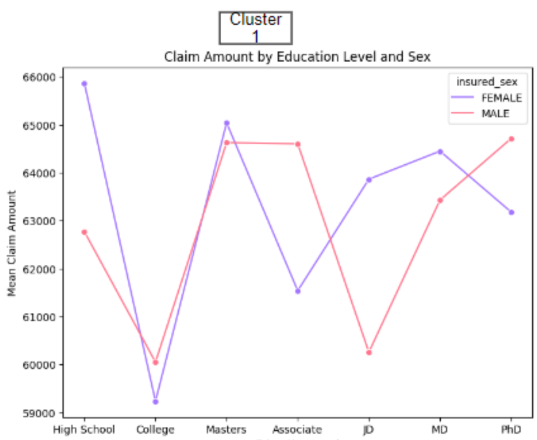
The obtained results –

VARIABLES	CLUSTER 1	CLUSTER 2
No. of data points	820	180
Weights	0.81	0.19
Parameters (μ, σ)	μ=11.0199 σ=0.0692931	μ=8.547433 σ=0.22011
Mean	63173	5752
Variance	16921.5	2853
Median	62290	5500
Mode	114920	19080
Percentage of females	54.14634%	51.66667%
Percentage of males	45.85366%	48.33333%
Age Category:		
Young Adults (0-30)	161~20%	36~20%
Adults (30-40)	**314~38%**	**82~46%**
Middle Aged (40-50)	246~30%	46~26%
Seniors (50-60)	99~12%	15~8%

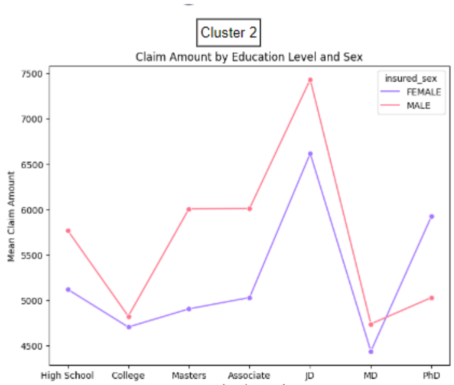
VARIABLES	CLUSTER 1	CLUSTER 2
Average Premium	1258.761	1245.68
Average Premium by sex:		
By female	1247.307	1249.006
By male	1272.285	1242.125
Percentage Education level:		
High School	15.48780	18.333333
College	12.80488	9.444444
Masters	14.39024	13.888889
Associate Education level	**13.41463**	**19.444444**
JD	15.85366	17.222222
MD	15.00000	11.666667
PhD	13.04878	10.000
Average Vehicle Claim	45374.54	4010.17

Weighted average claim: $ 52,263.01

Some of the additional Insights from the data used –



- Insurance claim amounts are predominantly made by females in high school.
- Males Master's, PhD degrees are likely to claim higher amount than females.



- Individuals holding a Professional degree, are likely to claim higher amount regardless of gender.
- Males with a Ph.D. degree and females with an MD degree appear to claim lesser amount

References

RESEARCH PAPERS
1)Fitting Finite Mixtures of Generalized Linear Regressions on Motor Insurance Claims January 2017International Journal of Statistical Distributions and Applications 3(4):124
DOI:10.11648/j.ijsd.20170304.19
Authors:

Nana Kena Frempong (Kwame Nkrumah University Of Science and Technology)

2) Fitting of Finite Mixture Distributions to Motor Insurance Claims December 2011Journal of Mathematics and Statistics 8(1):49 DOI:10.3844/jmssp.2012.49.56
Authors: P. Sattayatham (Suranaree University of Technology)

3) Modeling of Motor Insurance Extreme Claims through Appropriate Statistical Distributions

V SELVAKUMAR, DIPAK KUMAR SATPATHI, P T V PRAVEEN KUMAR, V V HARAGOPAL ,Department of Mathematics, Birla Institute of Technology and Science – Pilani, Hyderabad, India. BHAVAN'S VIVEKANANDA COLLEGE OF SCIENCE, HUMANITIES, AND COMMERCE, HYDERABAD, INDIA.

Email: vskselva79@gmail.com

For EM Algorithm:

https://onlinelibrary.wiley.com/doi/book/10.1002/9780470191613

For Mixture Distributions:

https://medium.com/@smallfishbigsea/an-explanation-of-discretized-logistic-mixture-likelihood-bdfe531751f0#:~:text=Mixture%20of%20Distributions&text=A%20mixture%20of%20distriction%20is,power%20by%20introducing%20more%20parameters.

For Lognormal Distributions:

https://www.sciencedirect.com/science/article/abs/pii/S2211692318300663