# Data Analysis and Modeling of Claim Amounts of Car Insurance using Big Data: A Study for Pakistan


## ABSTRACT

Modelling of data of claim amount is of paramount importance to manage risk reserve for payment of claims. Actuaries model uncertainty using probability distributions.

In this research paper claim amount distribution of the data of an insurance concern has been estimated and analysis was performed on big-data of claim amounts for better understanding and fitting of various probability distribution using R.

It was noticedthat the claim amounts distribution is highly positive skewed, therefore we have studied Exponential distribution, Gamma distribution and Weibull distribution as possible candidates for modelling the claim amount data. Chi Square test has been used as goodness of fit technique to decide suitable statistical model to representing the claim amounts under study.

Exponential distribution is found suitable for modelling the data under study.

Proposed model is usefulto estimate claim amount on aggregate for insurance concern when total loss is required to be computed to manage the risk reserve for the payments of claims.



**Keywords:** Probability distributions, Actuarial Modelling, Claim amounts, Maximum Likelihood Estimation.




## 1. INTRODUCTION

Modeling of claim amount is veryimportant and is of great interest for actuaries. Actuaries measure the degree of uncertainty on the basis of models. It is used to solve many problems in actuarial science as well as predicting insurance cost. The model could be used to decide when a claim be made and how much be paid [1]Therefore, modeling of claim amount is an important technique for actuaries to estimate parameters of the data foe the proposed model and making decision for loses and premium calculation’s[2]

Study of claim amount pattern using probability distribution approach when relevant data is available is an important technique foresting price of insurance policies, in order to estimate the liabilities of insurance companies. Modeling of claims amount and frequency can be used for better understanding of the implications of claims to the solvency of the company[3].

In order to study information about claim amount in insurancecompanies, which is very important for making decisions about premium levels, estimation of reservesobtained from premium and the profitability of insurance portfolios, loss modelling of claim amount plays an importantrole[4]. Claim amount collected in insurance arepositivelyskewed, therefore, probability distributions exhibit this characteristic are used for modeling[5].

Most commonlyused rightly skewed distributions in actuaries for modeling claim amount are Gamma, Lognormal Weibull and Pareto distribution,

Beta, Pareto, Burr, Weibull, Lognormal[6], Normal distribution and many other distributions[7].

## 2. Some Claim Amount Distributions

Below aregivenfew claim amount distributions used for modelling[8] of claim amounts and their cumulative density functions in Table 1. Commonly used measures of central tendency and dispersion of each distribution are also given for reference as these are required to study useful quantities in insurance in Table 2.



**Table 1 Probability Density Functions and Cumulative Density Functions**

| Probability distribution | Probability density function | Cumulative density function |
|---|---|---|
| **Exponential distribution** | f(x) = $\frac{1}{\beta} e^{-\frac{x}{\beta}}$ where $0 \leq x < \infty$, and $\beta > 0$ | F(x) = $\int_0^x \frac{1}{\beta} e^{\frac{-x}{\beta}}\, dx$<br>F(x) = $1 - e^{\frac{-x}{\beta}}$ where $x \geq 0$, $\beta > 0$ |

| | | |
|---|---|---|
| | | |
| **Gamma distribution** | $f(x)=\frac{\beta^{\alpha}}{\Gamma\alpha} x^{\alpha-1} e^{-\beta x}$<br>where $0 < x < \infty$ | $F(x) = \int_0^x \frac{\beta^{\alpha}}{\Gamma\alpha} x^{\alpha-1} e^{-\beta x}$ |
| **Weibull Distribution** | $f(x) = \frac{\gamma}{\alpha}\left(\frac{x-\mu}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{(x-\mu)}{\alpha}\right)^{\gamma}} \quad x \geq 0; \quad \mu, \gamma, \alpha > 0$ | $F(x) = \int_0^x \gamma x^{(\gamma-1)} e^{-x^{\gamma}} dx$<br>$F(x) = 1 - e^{-x^{\gamma}} \quad x \geq 0; \ \gamma > 0$ |

**Table 2 Measures of Exponential, Gamma and Weibull Probability distributions**

| **Measures** | **Exponential** | **Gamma** | **Weibull** |
|---|---|---|---|
| **Mean** | $\beta$ | $\frac{\alpha}{\beta}$ | $\Gamma(\frac{\gamma+1}{\gamma})$ |
| **Mode** | 0 | $\frac{\alpha-1}{\beta}$ | $\left(1-\frac{1}{\gamma}\right)^{\frac{1}{\gamma}}$ |
| **Range** | 0 to ∞ | 0 to ∞ | 0 to ∞ |
| **Standard deviation** | $\beta$ | $\frac{\sqrt{\alpha}}{\beta}$ | $\sqrt{\Gamma\left(\frac{\gamma+2}{\gamma}\right) - \left(\Gamma\left(\frac{\gamma+1}{\gamma}\right)\right)^2}$ |
| **Coefficient of variation** | 1 | $\frac{1}{\sqrt{\alpha}}$ | $\sqrt{\frac{\Gamma\left(\frac{\gamma+2}{\gamma}\right)}{\left(\Gamma\left(\frac{\gamma+1}{\gamma}\right)\right)^2} - 1}$ |

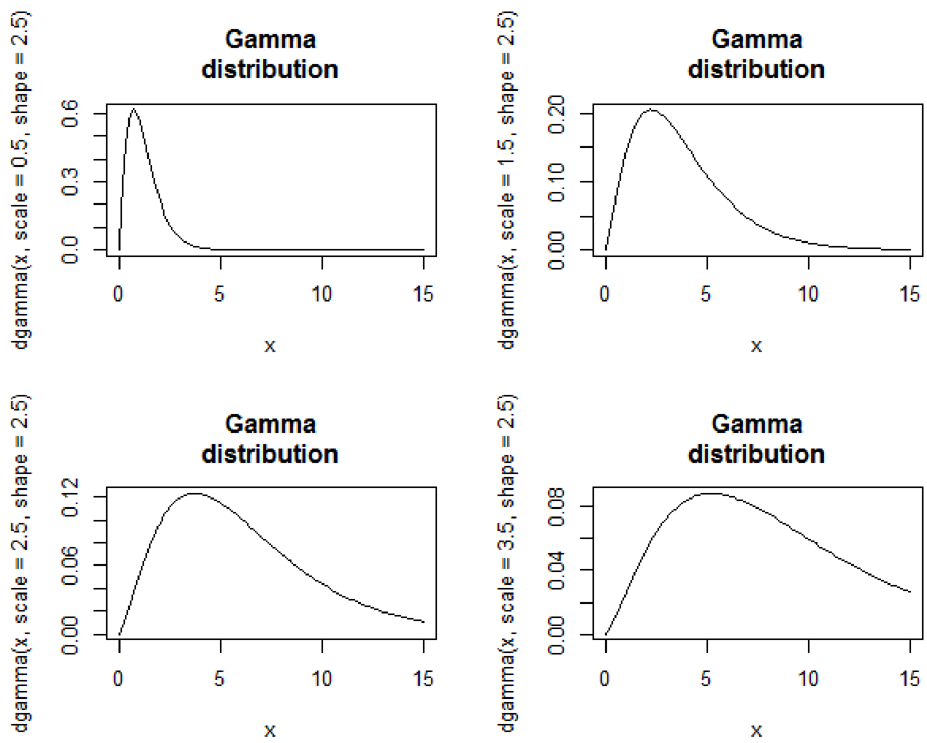



**Figure 1 Plot of Gamma distribution with different parameters**

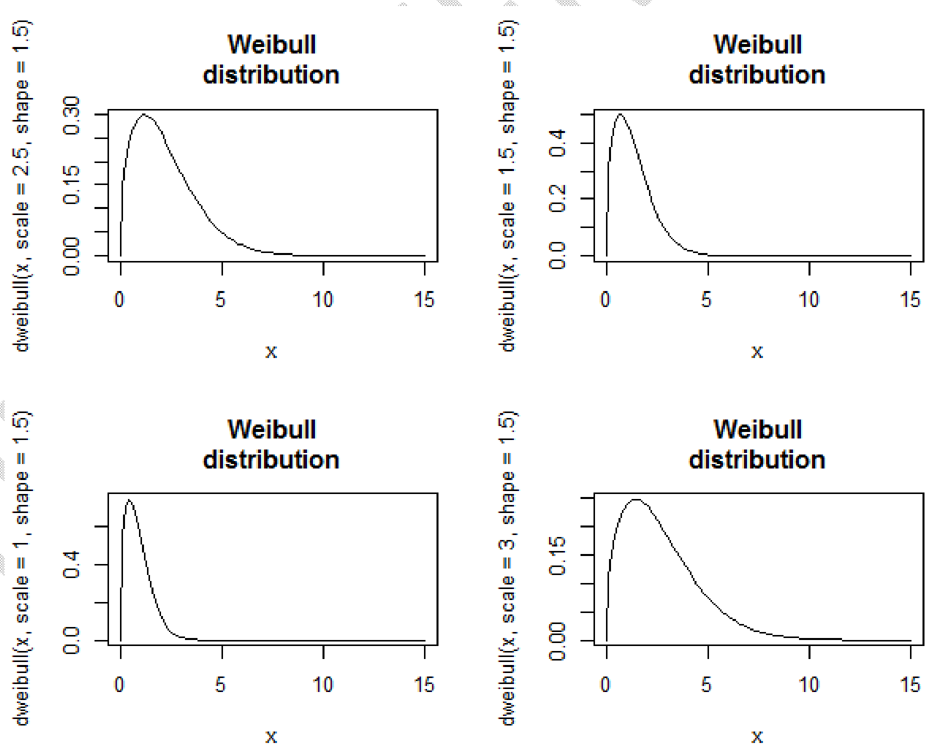



**Figure 2 Plot of Weibull distribution with different parameters**

## 3. Literature Review

Kazemi et al studied the claim amount data collected from financial records of a state owned major general insurance company in Iran. They fitted Skew Normal, Skew Laplace, generalized logistic, generalized hyperbolic, variance Gamma, normal inverse Gaussian, Marshal- Olkin, Log-Logistic, and Kumaraswamy Marshal-Olkinlog-Logistic distributions to the collected data and concluded that  Kumaraswamy Marshal-Olkin log-Logistic distributions is better for modelling the claim amount[9].

Cyprian et alpresented methodological frame work to select models for representing data of claim frequency and claim amount in insurance. They carried out this study on the basis of built-in data from R package insurance data. In this study researchers found lognormal distribution better model for representing claim amounts and negative binomial and geometric distribution as better distributions for modeling claim frequency[4].

Oyugi studied secondary data of claim amounts obtained from certain Insurance company in Nairobi, regarding their motor comprehensive policy.These researchers fitted Exponential, Gamma, Weibull and lognormal probability distributions and concluded that lognormal distribution is suitable for modeling the data under study[10].

Mohamed et al studied a model for claim amounts based on simulation of claim amounts. They fitted different probability distributions for claim amounts and found thatParetodistribution is suitable. This model was used for estimating insurance premiums for retention limit[11].

Burney and Hashmi have discussed different claim amount distributions as well as selection methods of distribution functions  for claim amounts[12].

Talangtam, et al studied in order to model the data set of claim amounts of motor insurance using finite mixture lognormal distributions, and estimating parameters by EM algorithm. To decide best fitted model Kolmogorov Smirnoff (K-S) and A-D tests were used[1].

Meyres studied the data of 250 claims to decide suitable statistical probability distribution which could be used for modeling the data of claim amounts. The researcher fitted Gamma, Weibull and lognormal probability distributions to the data of claim amounts under study. The parameters of the fitted distribution were estimated by the method of maximum likelihood [13]

## 4. Methodology

### 4.1 Data Analysis

The data for this research based on 133, 255 claimed amounts in Pk. Rupees. for the period Jan. 2010 to Dec 2015 from a well-known car insurance company in Pakistan. The name of company is not being mentioned here for confidentiality. The claim amounts of all types of vehicle showedan average of 29238 with variance 10561577230respectively. (No need)

Histogram of the data under study is drawn to decide the suitable probability distribution likely to fit the data.

## Results and discussion

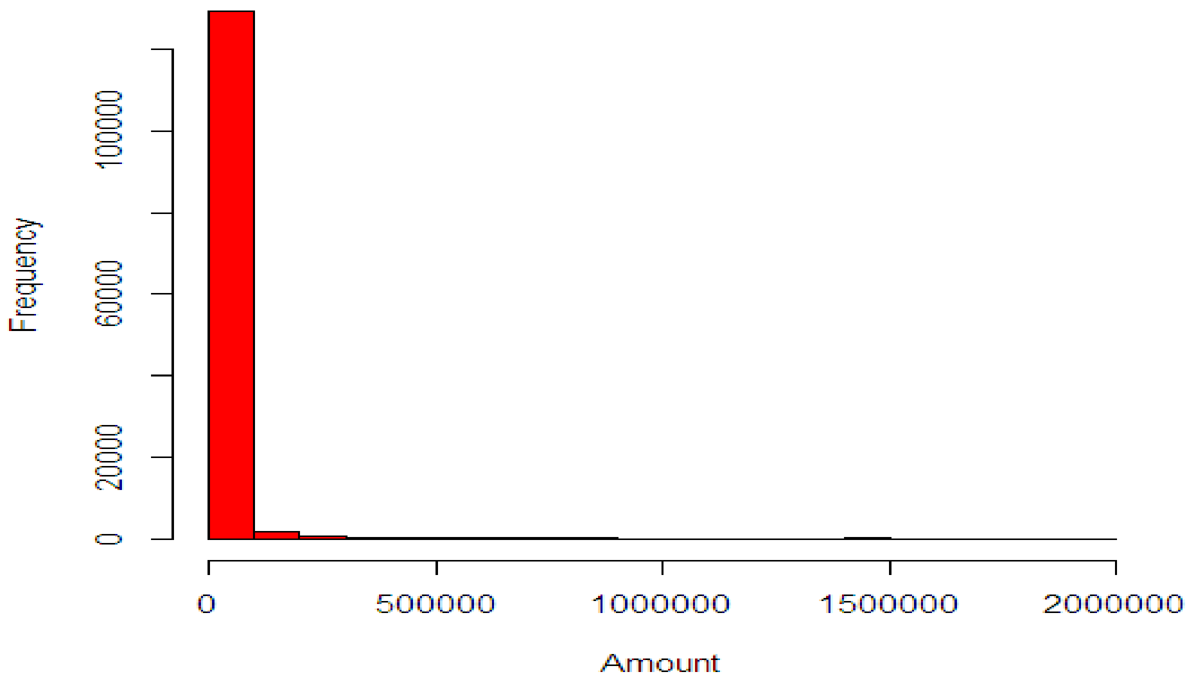



**Figure 3: Histogram of claim amounts for period Jan. 2010 to Dec 2015 in Pakistan**

Histogram shows highly skewed behavior. It can be seen that most of the claim amounts are small and there are few very high claims amounts. Therefore, it was decided to use highly skewed probability distributions for modeling the given data. Among rightly skewed continuous probability distributions, we used Exponential, Gammaand Weibull probability distributions for modeling the data under study using R package.

Following are the hypotheses to be tested about the modeling of claim amount data.

1. $H_0^*$: Exponential distribution is suitable candidate for modeling the data of claim amount under study.

2. $H_0^{**}$ : Gamma distribution is suitable candidate for modeling the data of claim amount under study.

3 $H_0^{***}$ : Weibull distribution is suitable candidate for modeling the data of claim amount under study.

Maximum Likelihood Estimates of the parameters of fitted distributions are given in Table 3

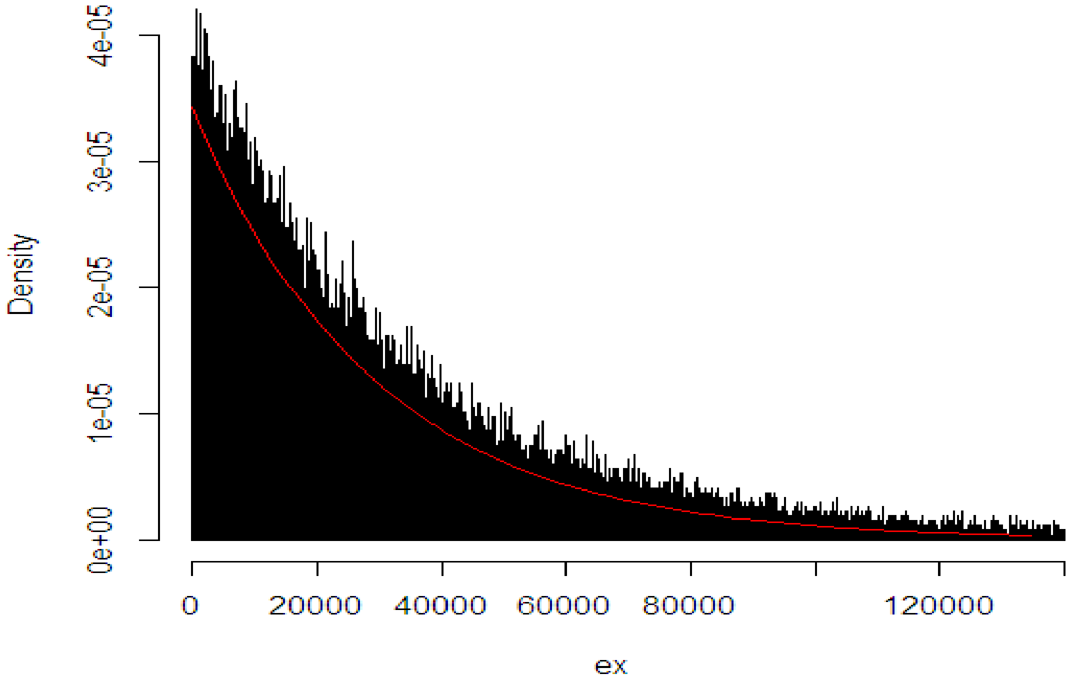
**Table 3 Estimates of parameters of Exponential and Gamma distribution**

| Probability distribution | Estimates | |
|---|---|---|
| Exponential | $\lambda = 0.0000342$ | |
| Gamma | $\alpha = 0.00789$ | $\beta = 0.0000027$ |

**Table 4: Expected frequencies for Exponential and Gamma Distribution**

| Claim Amount In 100,000 (PKR) | Observed Frequency | Expected Frequency (Exponential Distribution) | Expected Frequency (Gamma Distribution) |
|---|---|---|---|
| 0 – 1 | 129378 | 66257 | 122702 |
| 1 - 2 | 1618 | 33249 | 4879 |
| 2 – 3 | 514 | 16659 | 2276 |
| 3 – 4 | 341 | 8419 | 1257 |
| 4 – 5 | 229 | 4374 | 790 |
| 5 – 6 | 242 | 2165 | 486 |
| 6 – 7 | 231 | 1018 | 269 |
| 7 – 8 | 102 | 566 | 184 |
| 8 – 9 | 81 | 274 | 132 |
| 9 – 10 | 62 | 145 | 76 |
| 10 -11 | 54 | 62 | 65 |
| 11 – 12 | 43 | !42 | 44 |
| 12 – 13 | 62 | 12 | 24 |
| 13 – 14 | 64 | 9 | 32 |
| 14 – 15 | 76 | 2 | 15 |
| 15 – 16 | 48 | 0 | 11 |
| 16 – 17 | 51 | 1 | 11 |
| 17 – 18 | 36 | 0 | 3 |
| 18 - 19 | 23 | 1 | 1 |

**Figure 4: Graph of original data along with fitted distribution**




### 4.2 Goodness of fit test for claim amount distribution.

Chi Square test was usedfor goodness of fit of the assumed probability distributions. Relevant commands of R are given in Appendix A. Chi Square values and p values are presented in Table 3. Observed frequencies and expected frequencies of Exponential and Gamma are presented in Table 3.

**Table5: Chi Square Test Statistic**

| | Chi Square statistic | p-values |
|---|---|---|
| **Exponential Distribution** | $\chi^2$ = **18.75859,** | **0.2813803** |
| **Gamma Distribution** | $\chi^2$ = **43.02885** | **0.006871283** |

For Table5, it is observed that the p value for Gamma distribution is very small ($p < 0.01$) which indicate that Gamma distribution is not a suitable candidate for modelling the data of claim amount under study. Exponential distribution on the other hand has large p value. Therefore, Exponential distribution is better candidate for modeling the claim amount data under study. We also fitted Weibull distribution and the relevant estimates are given in table 6, analysis of the fitting is as under:

**Table 6: Fitting of the distribution of Weibull distribution by the method of MLE**

| | Estimate | Standard Error |
|---|---|---|
| Shape | 7.704193e-01 | 0.001263357 |
| Scale | 2.273621e+04 | 60.443552362 |
| Log likelihood: -1488209 | AIC: 2976422 | BIC: 2976442 |

Large value of Log Likelihood and small values of AIC and BIC indicates better fit(see Achieng, O. M., TRACK: ASTIN). But in our case value of LL is very small and values of AIC and BIC are very large which indicates that Weibull distribution is not suitable for the data of claim amounts under study.

Hence it is concluded that Exponential distribution as compare to Gamma and Weibull distributions is suitable probability distribution for modeling claim amount data.

**4.3Conclusion**

The objective of this study was to decide suitable probability distribution among the three skewed probability distributions. Findings on the basis of empirical analysis of the data indicate that exponential distribution is a suitable statistical model for claim amounts.

**4.4 Recommendations**

Insurance companies need an accurate pricing system which could make sufficient space for estimation of contingencies, expenses, losses and profits. In result of occurrence of claim there is loss on part of insurance company, therefore estimation of losses which are likely to occur in future is very important for insurance companies. Estimation of such losses is not possible without modeling of the data of claim amount or losses.

This research provides a basis for car insurance companies to develop suitable models for the data of claim amount of respective companies. It is suggested to the insurance companies should make necessary adjustments in the probability distributions on the basis of their own claim amount data.

## REFERENCES


[1] R., “Fitting of Finite Mixture Distributions to Motor Insurance Claims,” *J. Math. Stat.*, vol. 8, no. 1, pp. 49–56, Jan. 2012, doi: 10.3844/jmssp.2012.49.56.

[2] J. Rafal, “An empirical comparison of alternate regime-switching models or electricity spot prices,” *Munich Pers. RePEc Arch.*, Feb. 2010, [Online]. Available: https://mpra.ub.uni-muenchen.de/id/eprint/20546

[3] E. W. Frees and E. A. Valdez, “Hierarchical Insurance Claims Modeling,” *J. Am. Stat. Assoc.*, vol. 103, no. 484, pp. 1457–1469, Dec. 2008, doi: 10.1198/016214508000000823.

[4] C. O. Omari, S. G. Nyambura, and J. M. W. Mwangi, “Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions,” *J. Math. Finance*, vol. 08, no. 01, pp. 137–160, 2018, doi: 10.4236/jmf.2018.81012.

[5] *COTOR Challenge Round 2*. [Online Video]. Available: www.casact.org/cotor

[6] S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models: From Data to Decisions*, 1st ed. Wiley, 2008. doi: 10.1002/9780470391341.

[7] S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss models: from data to decisions*, Fifth edition. Hoboken, NJ: Wiley, 2019.

[8] Boland P. J., *STATISTICAL METHODS IN GENERAL INSURANCE*. Ireland, Dublin: Boland National University of Ireland, Dublin, 2006. [Online]. Available: https://www.stat.auckland.ac.nz/~iase/publications/17/5G1_BOLA.pdf

[9] R. K. Akram Kohansa, “Fitting Skew Distributions to Iranian Auto Insurance Claim Data,” *Appl. Appl. Math.*, vol. 12, no. 2, pp. 790–803, Jun. 2017.

[10] O. M. A. NAIROBI, KENYA, “ACTUARIAL MODELING FOR INSURANCE CLAIM SEVERITY IN MOTOR COMPREHENSIVE POLICY USING INDUSTRIAL STATISTICAL DISTRIBUTIONS”.

[11] Mohamed, “Approximation of Aggregate Losses Using Simulation,” *J. Math. Stat.*, vol. 6, no. 3, pp. 233–239, Aug. 2010, doi: 10.3844/jmssp.2010.233.239.

[12] S.M.Aqil Burney, “Risk Theory and Insurance an Invited Lecture.” [Online]. Available: https://www.researchgate.net/publication/342624357_Risk_Theory_and_Insurance_an_Invited_Lecture

[13] Vito Ricci, *Fitting Distribution with R*. Free Software Foundation, 2005. [Online]. Available: https://www.researchgate.net/publication/228791072_Fitting_Distributions_with_R

## APPENDIX

#########**Upload Original Data** ##########

```
setwd("D:/Insurance")

Claim_Amount_data<- read.table("Excel-Data.csv",header=TRUE,sep=",")

amount_scores<- Claim_Amount_data$Amount

bins<-seq(1,2000000,by=100000)

Scores<-cut(amount_scores,bins)

table(Scores)

transform(table(Scores))

hist(amount_scores,breaks=20, main="amount scores",xlab="Amount",col="red")
```

############**Summary, Mean & Variance of Original Data**###########

```
summary(amount_scores)

mean_amount_scores<-mean(amount_scores)

var_amount_scores<-var(amount_scores)
```

**For Simulation**##############

#########Calculate Rate & Shape of Gamma Distribution From Original Data###########

```
rate.a<-mean_amount_scores/var_amount_scores

shape.a<-((mean_amount_scores)^2)/var_amount_scores
```

```
############Generate Random Numbers From Gamma Distribution#######

x.gam<-rgamma(133267,rate=rate.a,shape=shape.a)

summary(x.gam)

####check Minimum Maximum Number and #######

x.gam.cut<-
cut(x.gam,breaks=c(0,100000,200000,300000,400000,500000,600000,700000,800000,900000,
1000000,

1100000,1200000,1300000,1400000,1500000,1600000,1700000,1800000,1900000,2000000,21
00000,2200000,2300000,

2400000,2500000,2600000)) ##binning data

############Frequency of Gamma Random Numbers######

transform(table(x.gam.cut))

############For Expected Frequeny##############

mean_x.gam<-mean(x.gam)

var_x.gam<-var(x.gam)

l.est<-mean_x.gam/var_x.gam

a.est<-((mean_x.gam)^2)/var_x.gam

f1=(pgamma(100000,shape=a.est,rate=l.est)-pgamma(0,shape=a.est,rate=l.est))*133267

f2=(pgamma(200000,shape=a.est,rate=l.est)-pgamma(100000,shape=a.est,rate=l.est))*133267

f3=(pgamma(300000,shape=a.est,rate=l.est)-pgamma(200000,shape=a.est,rate=l.est))*133267

f4=(pgamma(400000,shape=a.est,rate=l.est)-pgamma(300000,shape=a.est,rate=l.est))*133267

f5=(pgamma(500000,shape=a.est,rate=l.est)-pgamma(400000,shape=a.est,rate=l.est))*133267
```

```
f6=(pgamma(600000,shape=a.est,rate=l.est)-pgamma(500000,shape=a.est,rate=l.est))*133267

f7=(pgamma(700000,shape=a.est,rate=l.est)-pgamma(600000,shape=a.est,rate=l.est))*133267

f8=(pgamma(800000,shape=a.est,rate=l.est)-pgamma(700000,shape=a.est,rate=l.est))*133267

f9=(pgamma(900000,shape=a.est,rate=l.est)-pgamma(800000,shape=a.est,rate=l.est))*133267

f10=(pgamma(1000000,shape=a.est,rate=l.est)-
pgamma(900000,shape=a.est,rate=l.est))*133267

f11=(pgamma(1100000,shape=a.est,rate=l.est)-
pgamma(1000000,shape=a.est,rate=l.est))*133267

f12=(pgamma(1200000,shape=a.est,rate=l.est)-
pgamma(1100000,shape=a.est,rate=l.est))*133267

f13=(pgamma(1300000,shape=a.est,rate=l.est)-
pgamma(1200000,shape=a.est,rate=l.est))*133267

f14=(pgamma(1400000,shape=a.est,rate=l.est)-
pgamma(1300000,shape=a.est,rate=l.est))*133267

f15=(pgamma(1500000,shape=a.est,rate=l.est)-
pgamma(1400000,shape=a.est,rate=l.est))*133267

f16=(pgamma(1600000,shape=a.est,rate=l.est)-
pgamma(1500000,shape=a.est,rate=l.est))*133267

f17=(pgamma(1700000,shape=a.est,rate=l.est)-
pgamma(1600000,shape=a.est,rate=l.est))*133267

f18=(pgamma(1800000,shape=a.est,rate=l.est)-
pgamma(1700000,shape=a.est,rate=l.est))*133267

f19=(pgamma(1900000,shape=a.est,rate=l.est)-
pgamma(1800000,shape=a.est,rate=l.est))*133267
```

```
f20=(pgamma(2000000,shape=a.est,rate=l.est)-
pgamma(1900000,shape=a.est,rate=l.est))*133267

f21=(pgamma(2100000,shape=a.est,rate=l.est)-
pgamma(2000000,shape=a.est,rate=l.est))*133267

f22=(pgamma(2200000,shape=a.est,rate=l.est)-
pgamma(2100000,shape=a.est,rate=l.est))*133267

f23=(pgamma(2300000,shape=a.est,rate=l.est)-
pgamma(2200000,shape=a.est,rate=l.est))*133267

f24=(pgamma(2400000,shape=a.est,rate=l.est)-
pgamma(2300000,shape=a.est,rate=l.est))*133267

f25=(pgamma(2500000,shape=a.est,rate=l.est)-
pgamma(2400000,shape=a.est,rate=l.est))*133267

f26=(pgamma(2600000,shape=a.est,rate=l.est)-
pgamma(2500000,shape=a.est,rate=l.est))*133267


f.ex<-
c(f1,f2,f3,f4,f5,f6,f7,f8,f9,f10,f11,f12,f13,f14,f15,f16,f17,f18,f19,f20,f21,f22,f23,f24,f25,f26)

f.os<-vector()

for(i in 1:26) f.os[i]<- table(x.gam.cut)[[i]] ## empirical frequencies

X2<-sum(((f.os-f.ex)^2)/f.ex) ## chi-square statistic

gdl<-26-2-1 ## degrees of freedom

1-pchisq(X2,gdl) ## p-value

############Generate Random Numbers From Exponential Distribution#######

ex <- rexp(133267, rate = rate.a)
```

summary(ex)

####**check Minimum and Maximum Number and bBreak the data**#######

x.exp.cut<-

cut(ex,breaks=c(0,20000,40000,60000,80000,100000,120000,140000,160000,180000,200000,

220000,240000,260000,280000,300000,320000,340000,360000,380000)) ##binning data

############**Frequency of Exponential Random Numbers**######

transform(table(x.exp.cut))

############**For Expected Frequency**##############

mean_ex<-mean(ex)

e.rate<-1/mean_ex

e.rate

f1=(pexp(20000,rate=e.rate)-pexp(0,rate=e.rate))*133267

f2=(pexp(40000,rate=e.rate)-pexp(20000,rate=e.rate))*133267

f3=(pexp(60000,rate=e.rate)-pexp(40000,rate=e.rate))*133267

f4=(pexp(80000,rate=e.rate)-pexp(60000,rate=e.rate))*1332675=(pexp(100000,rate=e.rate)-pexp(80000,rate=e.rate))*133267

f6=(pexp(120000,rate=e.rate)-pexp(100000,rate=e.rate))*133267

f7=(pexp(140000,rate=e.rate)-pexp(120000,rate=e.rate))*133267

f8=(pexp(160000,rate=e.rate)-pexp(140000,rate=e.rate))*133267

f9=(pexp(180000,rate=e.rate)-pexp(160000,rate=e.rate))*133267

f10=(pexp(200000,rate=e.rate)-pexp(180000,rate=e.rate))*133267

f11=(pexp(220000,rate=e.rate)-pexp(200000,rate=e.rate))*133267

```
f12=(pexp(240000,rate=e.rate)-pexp(220000,rate=e.rate))*133267

f13=(pexp(260000,rate=e.rate)-pexp(240000,rate=e.rate))*133267

f14=(pexp(280000,rate=e.rate)-pexp(260000,rate=e.rate))*133267

f15=(pexp(300000,rate=e.rate)-pexp(280000,rate=e.rate))*133267

f16=(pexp(320000,rate=e.rate)-pexp(300000,rate=e.rate))*133267

f17=(pexp(340000,rate=e.rate)-pexp(320000,rate=e.rate))*133267

f18=(pexp(360000,rate=e.rate)-pexp(340000,rate=e.rate))*133267

f19=(pexp(380000,rate=e.rate)-pexp(360000,rate=e.rate))*133267


#############For Chi square Test###########

f.ex<-c(f1,f2,f3,f4,f5,f6,f7,f8,f9,f10,f11,f12,f13,f14,f15,f16,f17,f18,f19)

f.os<-vector()

for(i in 1:19) f.os[i]<- table(x.exp.cut)[[i]] ## empirical frequencies

X2<-sum(((f.os-f.ex)^2)/f.ex) ## chi-square statistic

gdl<-19-2-1 ## degrees of freedom

1-pchisq(X2,gdl) ## p-value

##########Test#########

fit1 <- fitdistr(ex, "exponential")

hist(ex, freq = FALSE, breaks = 20000, xlim = c(0, quantile(ex, 0.99)))

curve(dexp(x, rate = fit1$estimate), col = "red", add = TRUE)
```

Conference name: 11th International Conference on ,Mathematics and Statistics Computer Science Actuarial Science Oct 27-28 2017At: IoBM Karachi

Available link: -

https://www.researchgate.net/publication/359199984_Data_Analysis_and_Modeling_of_Claim_Amounts_of_Car_Insurance_using_Big_Data_A_Study_for_Pakistan