Spring 4-24-2023

Using a Distributive Approach to Model Insurance Loss

Kayla Kippes

Using a Distributive Approach to Model Insurance Loss

Kayla Kippes

A thesis presented for the degree of
Bachelor of Mathematics

University of Mary Washington
Fredericksburg, Virginia
April 2023

This thesis by **Kayla Kippes** is accepted in its present form as satisfying the thesis requirement for Honors in Mathematics.

Date	Approved
_____	_____
	Melody B. Denhere, Ph.D. Thesis Advisor
_____	_____
	Julius N. Esunge, Ph.D. Committee Member
_____	_____
	Debra L. Hydorn, Ph.D. Committee Member

Contents

Abstract

Insurance loss is an unpredicted event that stands at the forefront of the insurance industry. Loss in insurance represents the costs or expenses incurred due to a claim. An insurance claim is a request for the insurance company to pay for damage caused to an individual's property. Loss can be measured by how much money (the dollar amount) has been paid out by the insurance company to repair the damage or it can be measured by the number of claims (claim count) made to the insurance company. Insured events include property damage due to fire, theft, flood, a car accident, etc. An actuary aims to calculate the probability of an insured event occurring. In this paper we take a set of existing auto insurance data and model it using the Gamma and Weibull distributions. We use method of moments and maximum likelihood estimation to obtain parameter estimates for each distribution. We divide the data into different attributes where we found that the 60+ age group has slightly different shape/scale parameters and there is no real difference between male and female drivers.

1 Introduction

Generally speaking, insurance loss data has a distribution that is often unimodal shaped and right-skewed with heavy tails. In the insurance industry, there is much debate as to which distribution models insurance loss most accurately. That being said, numerous heavy-tailed models have been proposed in literature such as Pareto, Lognormal, Weibull, and Gamma (Ahmad et al). For all of these distributions, their parameters need to be estimated in order to provide an accurate fit for the data.

Two common methods of parameter estimation are the method-of-moments estimation, MME, and the maximum-likelihood estimation, MLE. The method-of-moments estimator is based on the law of large numbers where the sample mean converges to the distributional mean as the number of observations increased. The maximum likelihood estimator is obtained by maximizing the log-likelihood function. Both of these estimators are consistent and asymptotically normal, meaning that they should produce very similar results for parameter estimation (Brazauskas and Kleefeld). Even though these two methods should produce similar results, it is argued that the MLE provides more accurate parameters in some cases. This statement is valid for any distribution belonging to the one-parameter exponential family and any linked function (Brouste et al).

When using these estimation methods, the distribution is assumed to be known. That is, the distribution has already been established. Therefore, hypothesis testing has to be done in order to see if the estimated distribution function is the proposed distribution function. Literature argues that a common way to test this, is to conduct the Kolmogorov-Smirnov Test, which assesses if a sample comes from a certain population distribution (Hogg and Klugman). In addition to this test, we will use two other goodness-of-fit tests to help determine whether the estimated distribution is a good fit to the data.

2 Preliminaries

Several preliminary topics must be covered in order to understand the entirety of this research. These topics include the main two distributions used, Gamma and Weibull, and how to find their parameter estimates using the method-of-moments estimation, MME, and maximum-likelihood estimation, MLE. Additionally, it is important to understand how the different types of goodness of fit tests work. R was used to analyze the data, and also simulate data to make comparisons to the existing data set. Corresponding R Code will be included in the appendix to show the process of how results were derived.

2.1 Gamma Distribution

The Gamma distribution is a two-parameter family of right-skewed, continuous probability distributions. The distribution has shape parameter α, and scale parameter β, which is equal to $1/rate$. The probability density function, pdf, is defined by,

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{\frac{-x}{\beta}}, x > 0 \tag{1}$$

where Γ represents the Gamma function. We also know that the mean and variance are

$$\mu = \frac{\alpha}{\beta}, \quad \sigma^2 = \frac{\alpha}{\beta^2} \tag{2}$$

First, using MME, we need to obtain the first and second moments. When $X \sim \text{Gamma}(\alpha, \beta)$, then $E[X] = \frac{\alpha}{\beta}$ and $E[X^2] = \frac{\alpha+\alpha^2}{\beta^2}$. Therefore, we can determine the method of moments estimators $\hat{\alpha}, \hat{\beta}$ by solving the equations (3) and (4).

$$E[X] = \frac{\hat{\alpha}}{\hat{\beta}} \tag{3}$$

$$E[X^2] = \frac{\hat{\alpha} + \hat{\alpha}^2}{\hat{\beta}^2} \tag{4}$$

Using substitution and solving for both $\hat{\alpha}$ and $\hat{\beta}$, we obtain the parameter estimates,

$$\hat{\alpha} = \frac{n\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{5}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n\overline{x}} \tag{6}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$. When using MLE to find our parameters, we look at the log-likelihood function (L) represented in equation 7.

$$ln[L] = \alpha \sum lnx_i - \beta \sum x_i + n\alpha(ln\beta) - nln\Gamma(\alpha) \tag{7}$$

When taking the derivative with respect to β, we can easily find our estimated $\hat{\beta}$ parameter from this:

$$\hat{\beta} = \frac{\overline{x}}{\hat{\alpha}} \tag{8}$$

When we take the derivative with respect to α, we obtain,

$$n * ln(\frac{\hat{\alpha}}{\overline{x}}) - n * digamma(\hat{\alpha}) + \sum lnx_i = 0 \tag{9}$$

However, to solve (9) for α, the Newton-Raphson method has to be used. Newton's method is a root-finding algorithm which produces successively better approximations to the roots (or zeroes) of a real-valued function. The approximation is a repeated process until a sufficiently precise value is reached. The process for finding our estimated α is

$$\hat{\alpha}_{n+1} = \hat{\alpha}_n - \frac{h(\hat{\alpha}_n)}{h'(\hat{\alpha}_n)} \tag{10}$$

In (10), (9) and it's derivative are taken to predict the next alpha term. Hence, $h(\hat{\alpha}_n)$ is the same as equation 9 and $h'(\hat{\alpha}_n)$, the derivative of (9), is equal to $\frac{n}{\alpha_n} - n * trigamma(\alpha_n)$. Note that (10) involves the digamma and trigamma functions.

2.2 Weibull Distribution

The Weibull distribution is a two-parameter, prominently heavy-tailed, continuous distribution. The distribution has shape parameter k and scale parameter λ. The pdf is defined by

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(\frac{x}{\lambda})^k}, x \geq 0 \tag{11}$$

We also know that the mean and variance are

$$\mu = \lambda\Gamma\left(1 + \frac{1}{k}\right), \ \ \sigma^2 = \lambda^2\left[\Gamma\left(1 + \frac{2}{k}\right) - \left(\Gamma\left(1 + \frac{1}{k}\right)\right)^2\right] \tag{12}$$

In order to find our parameter estimates using MME, we will need to use the t^{th} moments and the corresponding sample t^{th} moments. The t^{th} moments $m_t, t = 1, 2, 3, ...$ is given by

$$m_t = \lambda^t\Gamma\left(1 + \frac{t}{k}\right) \tag{13}$$

The sample t^{th} moments $M_t, t = 1, 2, 3, ...$ is given by

$$M_t = \frac{1}{n}\sum_{i=1}^{n} x_i^t \tag{14}$$

Now that we have these two equations, we need to analyze the first and second moment $(t = 1, 2)$ for each of these equations and set them equal to each other. By doing that we obtain,

$$\lambda\Gamma\left(1 + \frac{1}{k}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{15}$$

$$\lambda^2\Gamma\left(1+\frac{2}{k}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \tag{16}$$

Note that $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ and $\frac{1}{n}\sum_{i=1}^{n} x_i^2 = s^2 + \bar{x}^2$ where $\bar{x}$ is the sample mean and s^2 is the sample variance. Next we can solve (15) for λ and find that

$$\lambda = \frac{\bar{x}}{\Gamma\left(1+\frac{1}{k}\right)} \tag{17}$$

To solve for k however, we have to take a different approach. To do this, (16) is divided by the (15) squared. This gives us,

$$\frac{\lambda^2\Gamma\left(1+\frac{2}{k}\right)}{\lambda^2\Gamma^2\left(1+\frac{1}{k}\right)} = \frac{s^2}{\bar{x}^2} + 1 \tag{18}$$

With the λ's canceling and setting (18) equal to 0, we get,

$$\left(\frac{s^2}{\bar{x}^2}+1\right)\Gamma^2\left(1+\frac{1}{k}\right) - \Gamma\left(1+\frac{2}{k}\right) = 0 \tag{19}$$

Now we can use Newton's method on (19) to get our estimated k parameter.

To find our parameters using MLE, we take a look at the log-likelihood function of k and λ. This is represented by Equation 20.

$$ln[L(k,\lambda)] = nln[k] - nkln[\lambda] - \frac{1}{\lambda^k}\sum_{i=1}^{n}(x_i)^k + (k-1)\sum_{i=1}^{n}(x_i) \tag{20}$$

We need to take the derivative of (20) with respect to k and λ. Taking the derivative with respect to λ, we can simply find our estimated λ parameter, represented in (21).

$$\hat{\lambda} = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^k\right]^{\frac{1}{k}} \tag{21}$$

When we take the derivative with respect to k we obtain,

$$\frac{1}{k} - \frac{\sum_{i=1}^{n}(x_i)^k * ln(x_i)}{\sum_{i=1}^{n}(x_i)^k} + \frac{1}{n}\sum_{i=1}^{n}(x_i) = 0 \tag{22}$$

To solve for our estimated k parameter from (22), we again need to use the Newton-Raphson method.

2.3 Goodness-of-Fit Tests

Throughout this paper we will use three different types of goodness-of-fit tests to determine whether or not we have correctly estimated parameters and distributions for our data.

The Kolmogorov-Smirnov (KS) test is used to decide if a sample comes from a population with a specific distribution. If the p-value obtained is relatively small, then it is acceptable to agree that the estimated distribution function with the estimated parameters is not a good fit for the data. Therefore, the null hypothesis and alternative hypothesis are:

$$H_0 : \text{The data follows the specified distribution}$$
$$H_a : \text{The data does not follow the specified distribution}$$

The test statistic D is used to determine if we reject or fail to reject H_0. If D is greater than the critical value obtained from a KS table, then H_0 is rejected. For the sake of this paper, the significance level is $\alpha = 0.05$ and the critical value obtained from the KS table is

$$1.36 \cdot \sqrt{\frac{n_1 + n_2}{n_1 * n_2}}$$

where 1.36 is the coefficient for the corresponding significance level.

We can also use the Kullback-Leibler (KL) divergence, a statistical distance. This is a measure of how one probability distribution is different from a second probability distribution and how closely they align. KL divergence can be denoted as $D_{KL}(P \parallel Q)$ where P is the original or observed data and Q is the estimated distribution. Since all of our distributions throughout this paper are continuous, we define KL divergence as

$$D_{KL}(P \parallel Q) = \int p(x) ln \left(\frac{p(x)}{q(x)} \right) dx,$$

where p and q are the respective probability densities of P and Q. When D_{KL} is equal to 0, we can say that the two distributions are identical to each other. D_{KL} has no upper bound, so the closer our statistic is to 0, the more confident we can be that our distributions are a good fit for each other.

Additionally, Q-Q Plots are used to visually represent the differences between two distributions by using a scatter plot that is created by plotting two sets of quantiles against one another. Quantiles from the same distribution should form a line that's roughly straight. If there are outlier points or the data doesn't make a straight line, we can assume that the data represented from the quantiles do not fall into the same distribution.

3 Modeling Insurance Loss

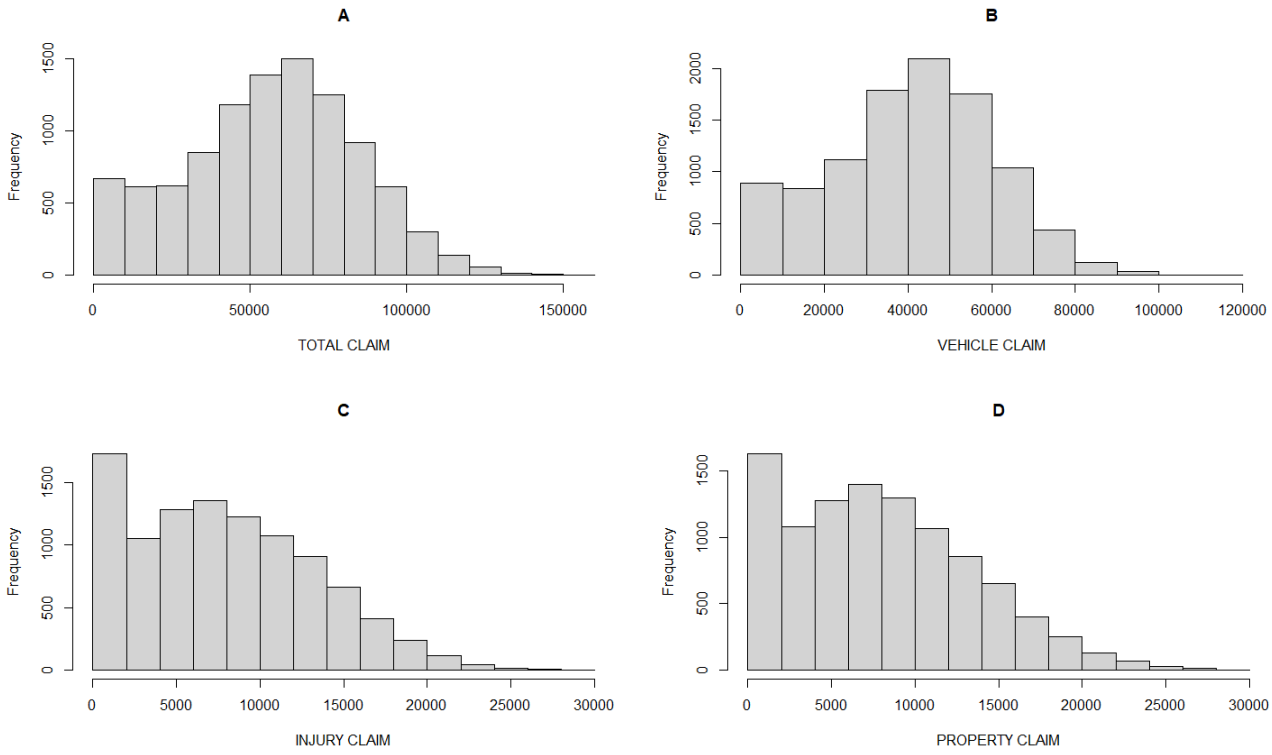

Figure 1: Four different types of claim histograms: Total (A), Vehicle (B), Injury (C), Property (D)

For the purpose of this research, the data set being analyzed includes private passenger automobile insurance-claims focusing on incidents from 31 Dec 2014 to 28 Feb 2015, with 10,211 records broken down into 39 different categories. Each record represents policy level claim information describing the policy, the policy holder, the insured person, the insured vehicle, the occurred accident, and the resulting claim size (i.e loss amount) in U.S. dollars. The loss amount for each individual claim is broken down into four components: total claim amount, injury claim, vehicle claim and property claim. Figure 1 shows the histogram of the four claim types. Additionally, Table 1 outlines the summary statistics for these four claim types. Note that total claim is made up of the other three claims. Injury claim and property claim share very similar shapes as they are more right-skewed. Also, total claim amount and vehicle claim share similar shapes as they are more unimodal with less skewness. The shapes of these data sets play a large role in trying to figure out which distribution is of best fit for them. Throughout the rest of this section, each claim will be represented with either the Gamma and Weibull distribution and we determine how they both fit or don't fit the data. There will also be an explanation why different distributions did not work. Section 4 will look at the different categories for each record and see how different attributes (age, gender, deductible amount, etc.) affect the distribution and then see if the distribution changes based on a certain attribute.

Summary Statistics				
	Total Claim	Injury Claim	Property Claim	Vehicle Claim
Mean	56586.94	7911.03	8027.50	40794.42
Median	58170	7455	7600	42155
Standard Deviation	27648.78	5460.94	5519.05	19665.25
Skewness	**-0.0639**	**0.4121**	**0.4694**	**-0.1033**
Minimum	100	0	0	10
Maximum	154740	30000	29700	110800

Table 1: Summary Statistics of the Claim Amounts ($) by Type

3.1 Injury and Property Claim

Initially, both Gamma and Weibull distributions were used to try to fit the **injury** and **property** claim data. However, when using both of the methods for parameter estimation, we found that the MME and MLE estimators were very similar to each other when using Gamma to fit these two claims. For the other two claim data sets, there was a significant difference between the parameters that were given for MME and MLE. Therefore, the Gamma distribution is being used for the injury and property claim data. Using the estimated parameters we solved for in section 2.2, we found the following parameter estimates for these two claims represented in Table 2.

	Injury Claim	Property Claim
Observations, n	9204	9239
Alpha, α	2.0977	2.1275
Beta, β	4150.697	4137.085

Table 2: Gamma Parameter Estimates

When the goodness-of-fit tests were conducted with our claim data and an estimated Gamma distribution using our parameters, we obtain the following test statistics shown in Table 3.

	Injury Claim	Property Claim
KS Statistic, D	0.0583	0.0669
KL Divergence	0.5818	0.4279

Table 3: Goodness-of-Fit Tests

Our critical value obtained from the KS table is, $1.36 \cdot \sqrt{\frac{n_1+n_2}{n_1 * n_2}} = 0.020$. These test statistics state that neither the injury claim data nor the property claim data come from a Gamma distribution with our estimated parameters. However, as we can see from the graphs from Figure 2, the right tails of the distributions seem to be similar but the peaks and left tails appear to be different. Also, the actual mean and variance from the data are not close enough to the estimated ones using the parameters that were determined. This leads us to believe that there could be potential bias or outliers in our data that are causing the goodness-of-fit tests to indicate that we do not have the correct distribution here. We can confirm this by

looking at our Q-Q Plots shown in Figure 3. The scatter plots are more of a curve and there are also numerous outlier points on the plot.

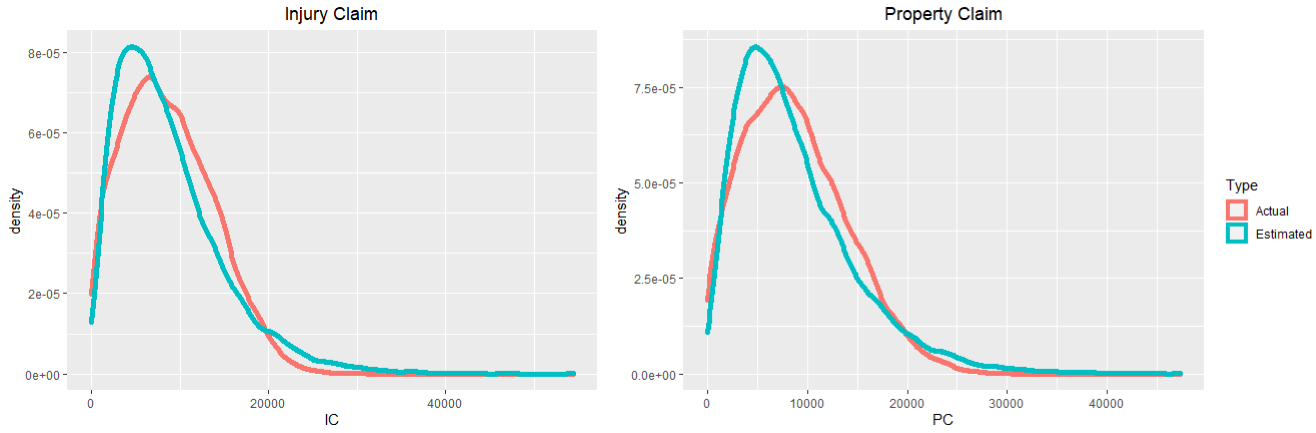

Figure 2: Gamma: Comparing Estimated vs. Actual Data

Theoretical vs. Sample Quantiles

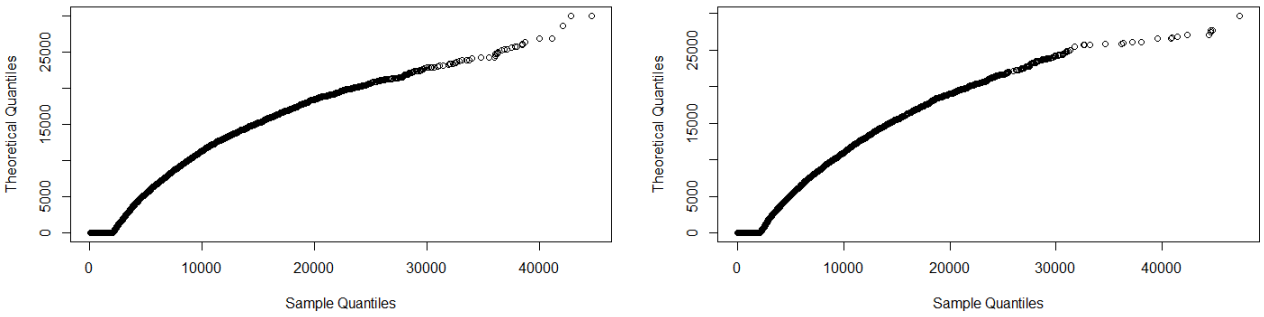

Figure 3: Q-Q Plots of Injury (L) and Property Claim (R)

In order to eliminate the bias or outliers, we have broken down the data into two different parts. To determine where we should break the data up, we used our initial histograms shown in Figure 1. There is a large dip in both the injury and property claim histograms that seems to be causing a problem in fitting our data, the break occurs at 4000 for each claim type.

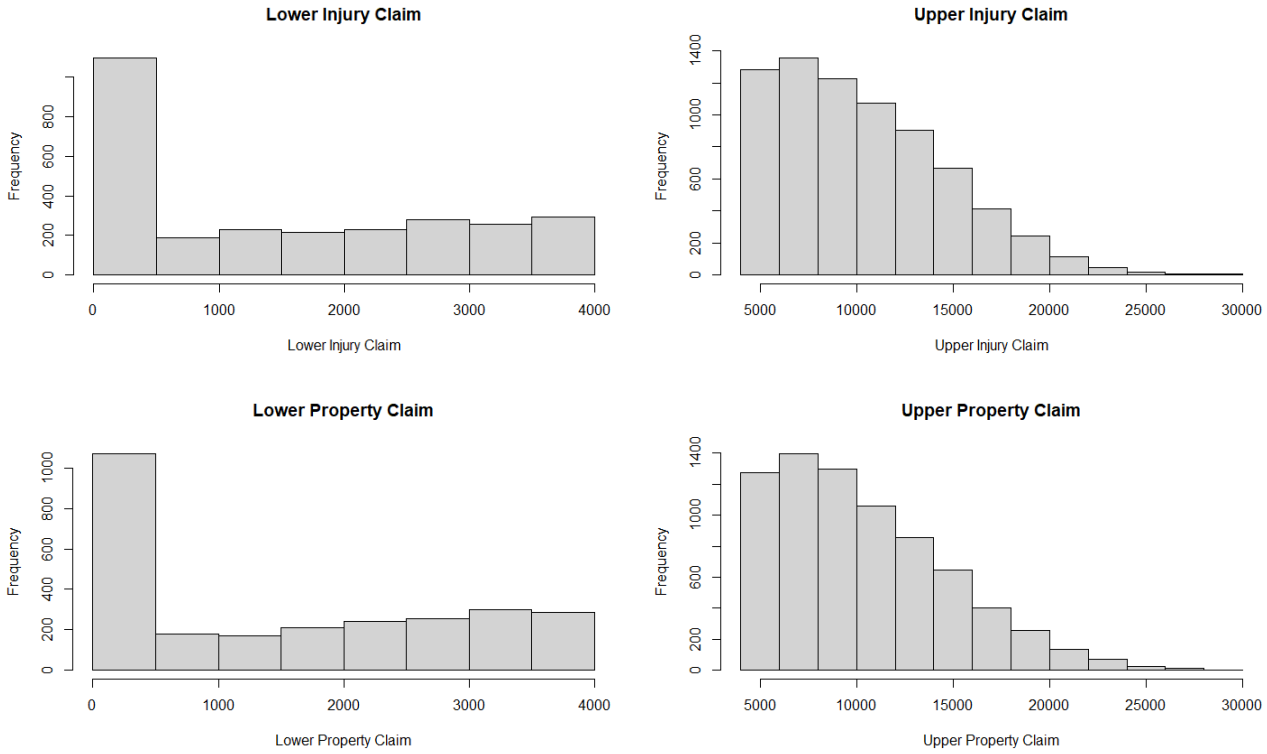

Figure 4: Separated Injury and Property Claim

Now we are analyzing claims from 0 to 4000 and claims greater than 4000. Our new histograms are represented in Figure 4. The histograms on the left represented claims from 0 to 4000 and the histograms on the right represent claims greater than 4000. To better understand these new data sets, a short summary statistics is shown in Table 4.

Summary Statistics				
	Lower IC	Upper IC	Lower PC	Upper PC
Observations, n	2785	7345	2706	7424
Mean	1474.86	10351.55	1509.21	10403.49
Standard Deviation	1394.02	4327.39	1415.06	4438.17
Skewness	0.3399	0.6899	0.2975	0.7872

Table 4: Summary Statistics

From the shape of the two lower tail claims, we can see that this data cannot be fit with a specific distribution. The Pareto distribution was considered to try and fit this shape. However, since the frequency of claims is not decreasing and is instead staying fairly uniform, the Pareto distribution was not a good fit. The reason there is such a high frequency in the first histogram bar, is due to the large amount of claims that were 0 for injury and property claim. Including these claims is reflective of real life insurance data because when a claim is filed, there may only be damage or a claim needed in one area leaving the other claims at a 0 value. Without those 0 value claims, we can assume that for claims under 4000 in both

injury and property claim, the distribution can be almost uniform.

The greater than 4000 claims are the claims that can be better represented by our Gamma distribution. Both the Upper IC and Lower IC have extreme skewness with heavy tails. Using these two new components of injury and property claim we have our new parameter estimates shown in Table 5.

	Upper IC	Upper PC
Alpha, α	5.7229	5.7235
Beta, β	1808.783	1817.681

Table 5: New Gamma Parameter Estimates

When rerunning the goodness-of-fit tests in R with our separated claim data, we obtain the following test statistics shown in Table 6.

	Injury Claim	Property Claim
KS Statistic, D	0.04153	0.03233
KL Divergence	0.4008	0.3603

Table 6: Goodness-of-Fit Tests

Our critical value obtained from the KS table based on our new number of observations for injury claim is 0.022 and for property claim is also 0.022. We now have test statistics that are significantly closer to our critical values and our KL Divergence values were cut in half are substantially closer to 0. Additionally, our estimated mean and standard deviation are almost the exact same as the true mean and standard deviation which indicates these estimators are a good fit. While these statistics might not be perfect, we can be confident that we have correctly represented our injury and property claim data using the Gamma distribution. This can be reflected in the graphs from Figure 5.

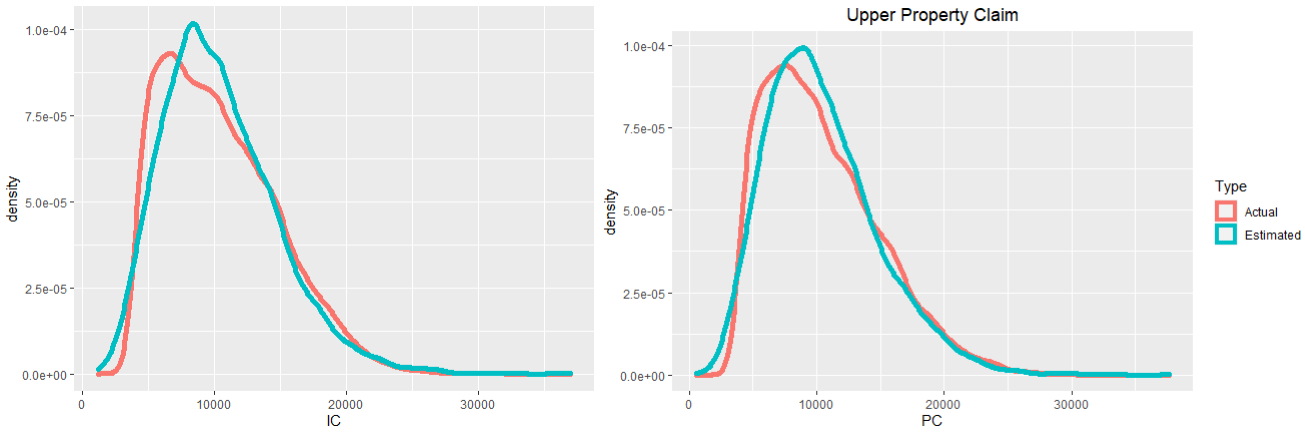

Figure 5: Upper Injury Claim & Property Claim Actual vs. Estimated

Because our data set is a company's real life data, getting a perfect match would be unrealistic but our estimations are believed to be as close as we can get. We did try to fit this separated

injury and property claim data with the Rayleigh distribution and the Weibull Distribution but neither were as close of fit as the Gamma distribution for these two claims.

3.2 Total and Vehicle Claim

Contrary to the other two claims, we found that the MLE and MME Weibull estimators were very similar to each other for **total** and **vehicle** claims. Therefore, the Weibull distribution is being used for the total claim amount and vehicle claim data. The parameter estimates for these two claims can be seen in Table 7.

	Total Claim Amount	Vehicle Claim
k	2.08172	2.11328
λ	63457.52	45744.65

Table 7: Weibull Parameter Estimates

When the goodness-of-fit tests were conducted with our claim data and an estimated Weibull distribution using our parameters, we obtain the test statistics shown in Table 8.

	Total Claim Amount	Vehicle Claim
Test Statistic, D	0.070089	0.067029
KL Divergence	0.4245	0.3549

Table 8: Goodness-of-Fit Tests

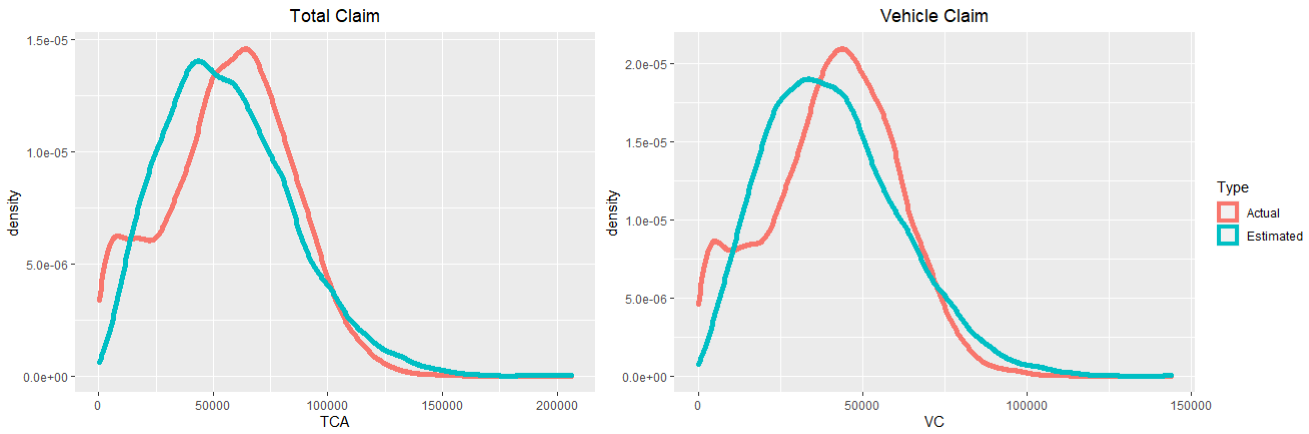

Figure 6: Weibull: Comparing Estimated vs. Actual Data

We have the same critical value as before equal to 0.019 because $n_1 = n_2 = 10130$. Again the KS test statistics state that neither the total claim amount data nor the vehicle claim data come from a Weibull distribution with our estimated parameters. However, these test statistics are better than the previous ones and suggest that Weibull might be the correct distribution for this data. Also note that our KL Divergence is much closer to 0 than it was for the Gamma estimates. Additionally, based off of these parameters, the estimated mean

and standard deviation are not as close as the true values. Again, as we can see from the graphs in Figure 6, the right tails of the distributions seem to be similar with only slightly different peaks and left tails. This can be confirmed by our Q-Q Plots shown in Figure 7. The scatter plots are curved to start but then they start to form a straight line as the claims get larger.

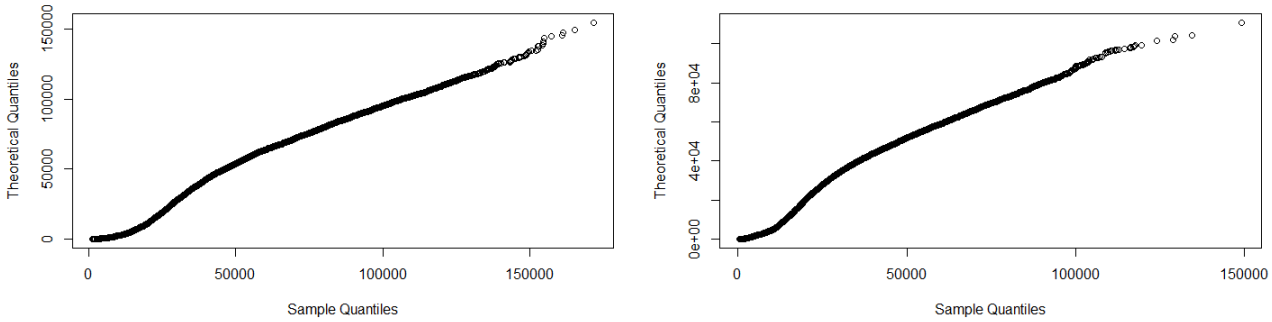

Figure 7: Q-Q Plots of Total (L) and Vehicle Claim (R)

To find a better fit for our data, we will segregate total and vehicle claims but at different points. Total claim will be separated at 30,000 so we have claims from 0 to 30,000 and claims greater than 30,000. Vehicle claim will be separated at 10,000, so we have claims from 0 to 10,000 and claims greater than 10,000. The summary statistics of the broken down claims are shown in Table 9.

Summary Statistics				
	Lower TC	Upper TC	Lower VC	Upper VC
Observations, n	1905	8225	892	9238
Mean	14857.07	66252.04	4823.49	44267.69
Standard Deviation	8812.213	20658.09	2801.64	16920.25
Skewness	0.0475	0.4692	0.1490	0.1423

Table 9: Summary Statistics

The new histograms are shown in Figure 8. The histograms on the left represent the two lower claims of the data and the right-side histograms show the upper portion of the claims.

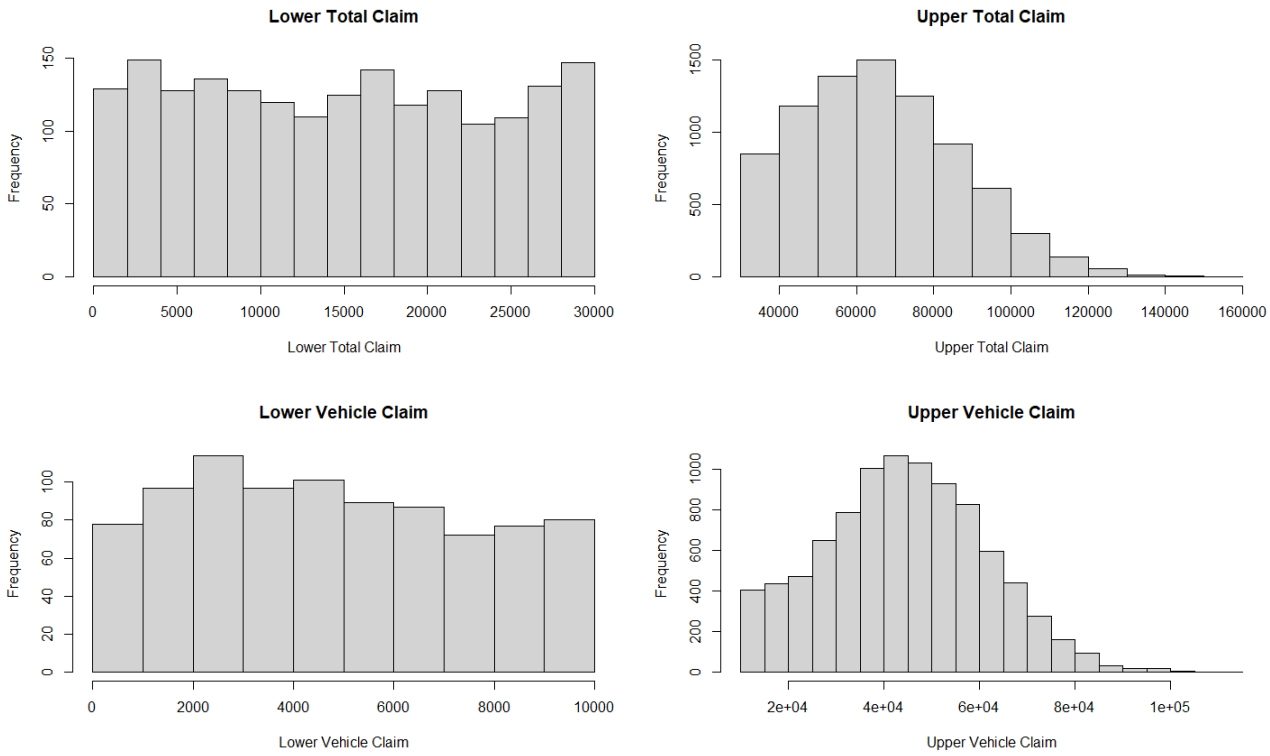

Figure 8: Separated Total and Vehicle Claim

Based on the histograms from the two lower half claim amounts, it can be concluded that the Uniform distribution would be the best fit because there is not significant fluctuation. The estimated Uniform parameters are represented in Table 10, where a is the minimum, and b is the maximum.

	Lower TC	Lower VC
a	100	10
b	30000	10000

Table 10: Uniform Parameter Estimates

For lower total claim, our critical value is 0.044 and we found a test statistic, D of 0.024. For lower vehicle claim, our critical value is 0.064 and we found a test statistic, D of 0.050. These statistics confirm that these lower claims do follow a Uniform distribution. This is helpful because it confirms that there was a portion of the whole data that was causing our initial tests to be thrown off.

We are now left with the upper portions of our total and vehicle claims. First to start with total claims, we note that these claims areskewed with a heavy tail. After analyzing both the Gamma and Weibull distributions are these upper portion of claims, we actually find that Gamma is the best distribution used to fit this data. Our parameters are shown in Table 11.

	Upper TC
Alpha, α	10.204
Beta, β	6493.065

Table 11: Gamma Parameters

To confirm that our new parameters and distribution are accurate for this upper portion of total claims, we ran our *ks.test* in R. We did not include the KL Divergence for these claims because the ranges of the actual and simulated data were too different causing KL to be invalid. With the number of observations being equal to 8225, our critical value from the KS Table is 0.021. The test statistic that we obtained from these new estimated parameters was 0.021. This means we have found the correct distribution for this portion of the total claim amounts. The graph of the estimated versus actual distribution and it's Q-Q Plot are shown in Figure 9 and 10.

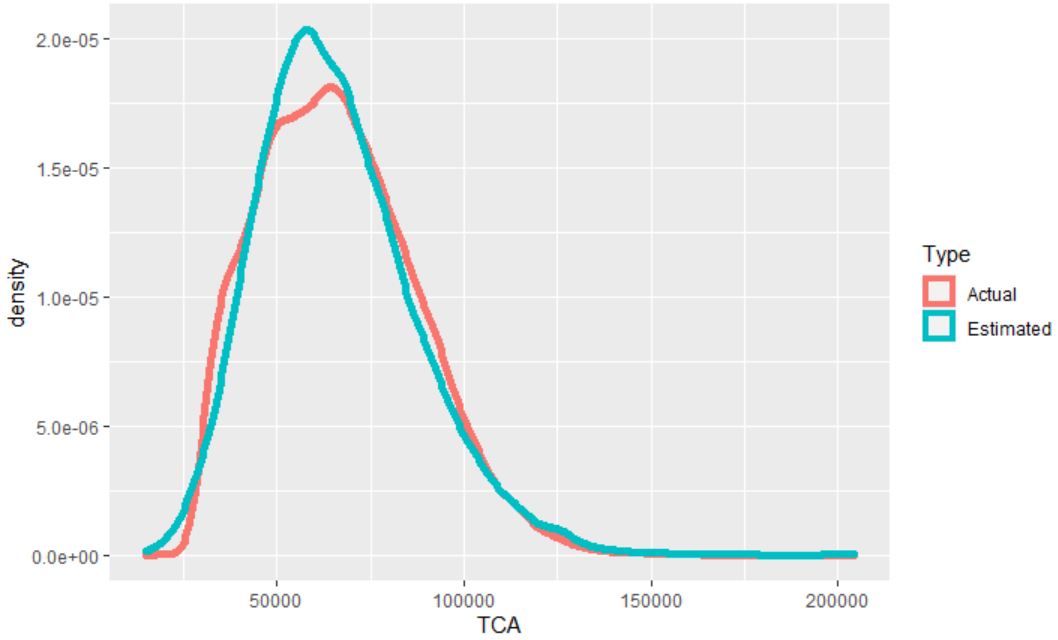

Figure 9: Upper Total Claim Actual vs. Estimated

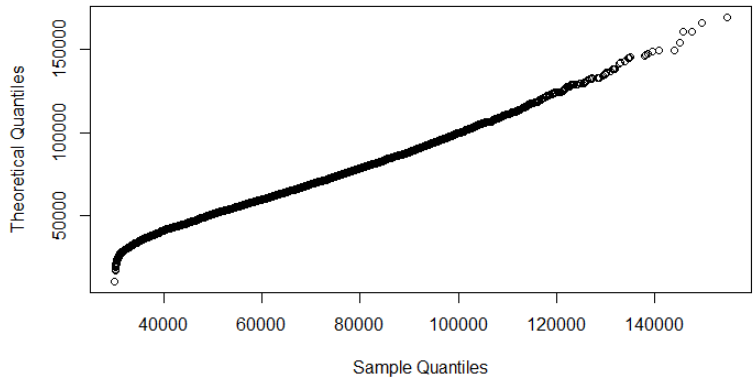

Figure 10: Upper Total Claim Q-Q Plot

Moving to the upper portion of vehicle claims, we notice that the histogram is not very skewed. While initially we considered using a Normal or Truncated Normal distribution to fit this data, the data does not include negative values for it to be Normal and there is a peak so it cannot be Truncated. After trying to fit the data with both the Gamma and Weibull distribution, we found that Weibull was of better fit for this portion of vehicle claims. Our parameters are shown in Table 12.

	Upper Vehicle Claim
k	2.856
λ	49689.191

Table 12: Upper Vehicle Claim Weibull Parameters

We again need to verify that our new parameters and Weibull distribution are accurate for this upper portion of vehicle claims so we will use the ks.test. With the number of observations being equal to 9238, our critical value from the KS Table is 0.020. The test statistic that we obtained from these new estimated parameters was 0.019. In addition to the total claims, we have now found the correct distribution for this portion of vehicle claims. The graph of the estimated versus actual distribution and it's Q-Q Plot are shown in Figure 11 and 12.

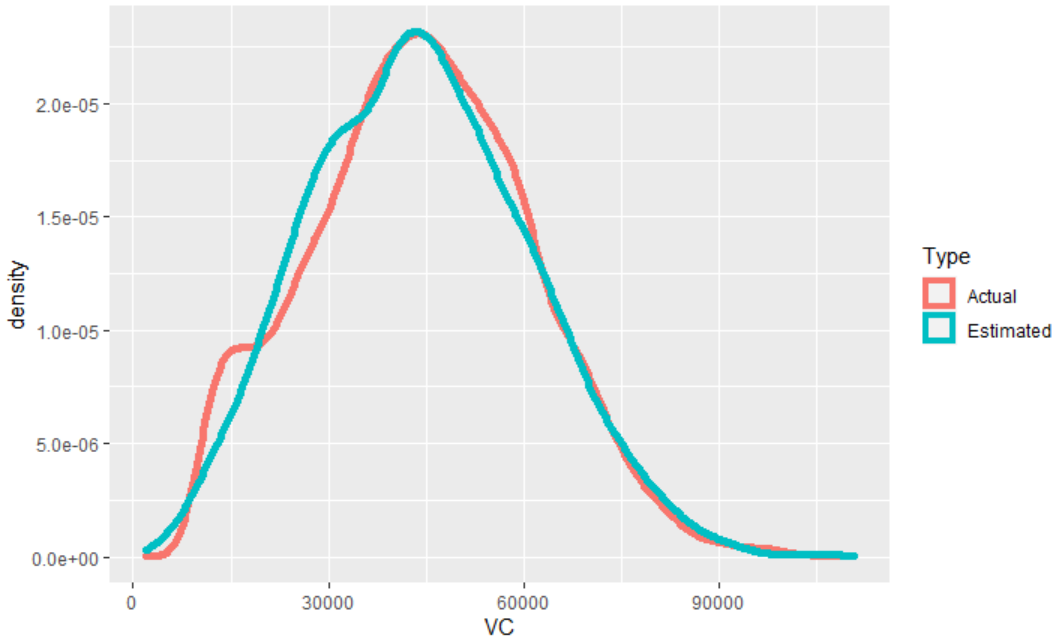

Figure 11: Upper Vehicle Claim Actual vs. Estimated

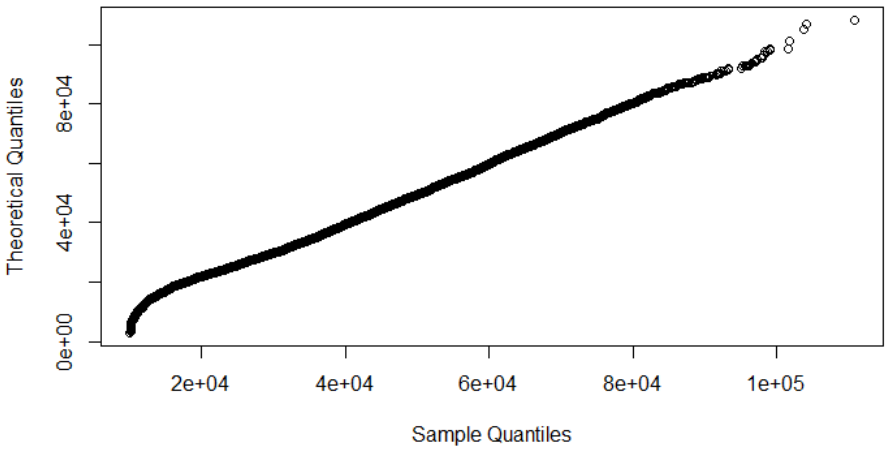

Figure 12: Upper Vehicle Claim Q-Q Plot

4 Analyzing Attributes

To better understand our data, we analyzed four different attributes that our data could be broken into: age, gender, incident time of day, and geographic region. For each of these attributes, we analyzed 'Total Claim Amount'(TCA), data for simplicity and due to the other

three claim types being a part of the total. Therefore, the attributes will be represented by the Weibull distribution.

4.1 Age

In this set of data, age ranged from 18 to 79 years. Since this was a larger range, we decided to break up the data into six smaller groups. The smaller groups were based on how insurance companies break up their age groups instead of making even mathematical breaks. The age groups are summarized in Table 13.

Age Group	Observations, n	TCA Mean
All	10130	56586.94
18 to 20	341	57072.17
21 to 25	752	56390.24
26 to 35	2898	55974.61
36 to 45	3297	55869.46
46 to 59	2350	57522.43
60+	492	60497.85

Table 13: Age Summary

To analyze if any of these groups were significantly different from each other or all the ages combined, we first used the Kruskal-Wallis test to compare each group. This is a non-parametic method for testing whether samples originate from the same distribution. If the p-value is less than the significance level, $\alpha = 0.05$, we can conclude that the samples come from different distributions. When running this test in R, we obtain a p-value of 0.012, meaning that there are differences between these 7 groups. To further analyze and find where the differences occur, we used a "pairwise t test" with the Bonferroni correction method. The t test compares the means of each individual group and the Bonferroni correction divides the critical p-value by the number of comparisons being made. We used this test because we wanted to see the effect of outliers on the mean. When administering this test in R, we found that there were significant differences (p-value ¡ 0.05) between all the ages and the age group of 60+. This suggests that individuals 60 years and older are more likely to have a larger total claim amount compared to younger age groups.

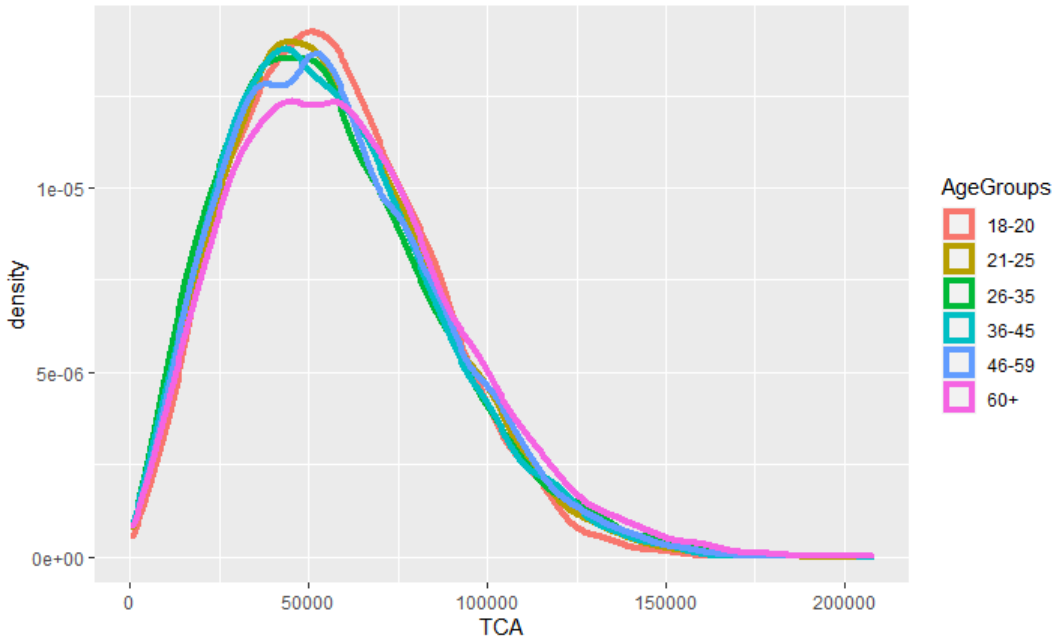

Figure 13: Comparing Age Group Distributions

We also estimated the parameters of each age group using the Weibull distribution to see if the distributions were similar to each other. In each estimation, the distribution has 10,130 observations. As shown in Figure 13, the 60+ age group is the group that is showing the most difference among the groups. It can also be said that there is a difference between the 18 to 20 age group. Our data and estimated parameters are consistent with insurance studies stating that there are differences in insurance claims for younger drivers and older drivers.

4.2 Gender

For this paper, gender is classified as male and female. The summary of these two variables are represented in Table 14. There are more females than males. Also, the means do not seem to differ too much from the whole TCA data set.

Gender	Observations, n	TCA Mean
All	10130	56586.94
Female	5405	57007.17
Male	4725	56106.24

Table 14: Gender Summary

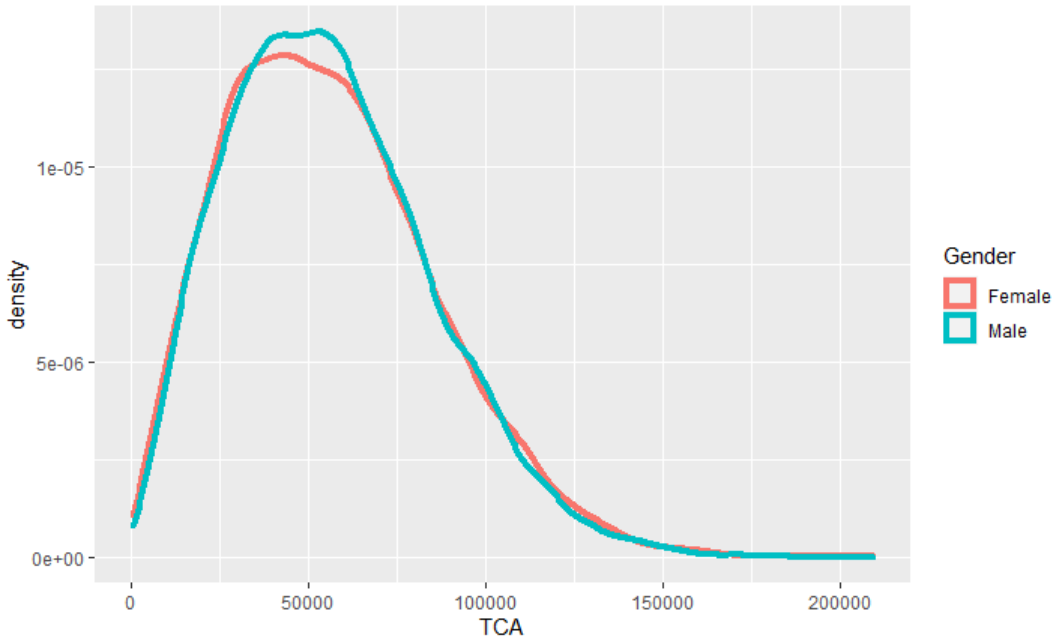

Figure 14: Comparing Gender Distributions

To analyze if these groups were significantly different from each other or all the TCA data combined, we again used the Kruskal-Wallis test to compare the groups. In this case, we obtained a p-value of 0.262. Therefore, we do not see a significant difference in total claim amounts between the two genders. However, as shown in Figure 14, the parameter estimates indicate that there is a difference between genders right around the peak of the distribution. The higher peak indicates that more males have the mean claim amount than females do. This could be explained by the fact that insurance claims do tend to be different for males in certain cases.

4.3 Incident Time of Day

Another attribute that we thought would be relevant to look at was the time of day in which the incident occurred. We broke the 24 hour day into 8-hour increments. The first group is representative of the early hours of the day, from midnight to 7am. The next is representative of the morning to mid-afternoon from 8am to 3pm (i.e. 1500). The last group includes the afternoon to night time from 4pm (i.e. 1600) to 11pm (i.e. 2300). The summary of these groups are represented in Table 15. The mean is slightly higher as the hours of the day increase, especially in the later hours of the table.

Time	Observations, n	TCA Mean
All	10130	56586.94
Early	3272	52527.94
Mid	2801	57271.35
Late	4057	60337.18

Table 15: Time Summary

To test if any of these groups are significantly different, we again use the Kruskal-Wallis test. The p-value we obtained is so infinitely small that it is basically 0, indicating a significant difference. Therefore, we will use the 'pairwise t test' to asses which groups are different from each other. Again we got p-values that were so infinitely small, meaning that each individual group is significantly different from each other. This is shown in Figure 15. There are very little to no similarities across these three distributions.

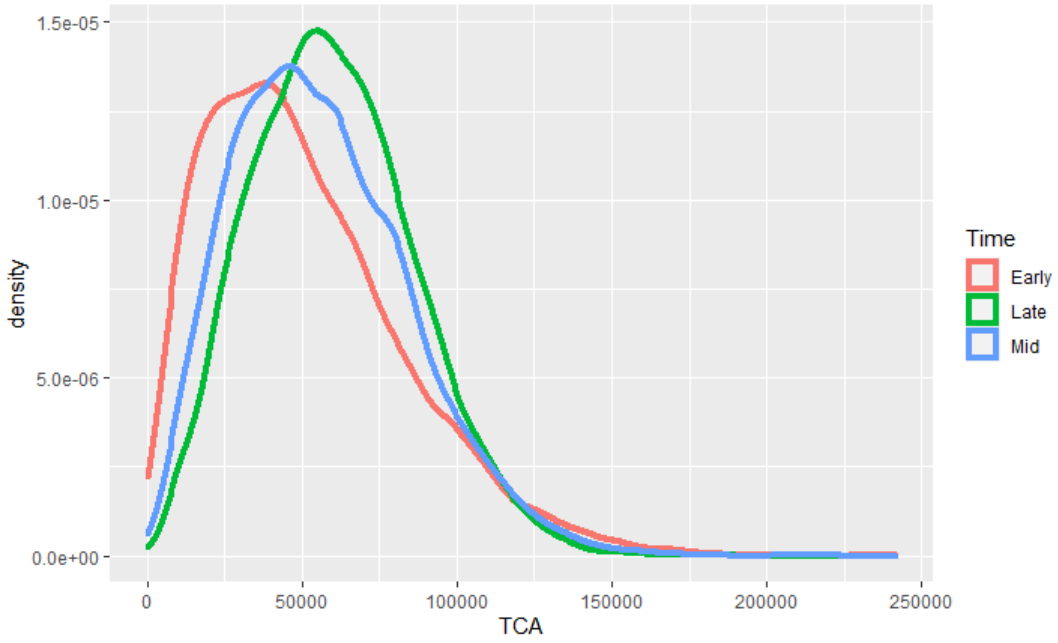

Figure 15: Comparing Time of Day Distributions

4.4 Region

The last attribute that we decided to look at within our data was region in which the incident occurred. Insurance differs by what state the insured lives in so we were curious to see if there was differences between two generalized groups. Seven different states were represented in this data. Ohio, Pennsylvania, and New York are being represented as northeast (NE), and Virginia, West Virginia, North Carolina, and South Carolina are being represented as southeast (SE). There is not much difference between these groups. The summary of these groups are shown in Table 16.

Region	Observations, n	TCA Mean
All	10130	56586.94
SE	7040	56385.52
NE	3090	57045.85

Table 16: Region Summary

We did not find a significant difference between these two groups when running our Kruskal-Wallis test. The p-value obtained was 0.542, meaning these two groups have the highest similarities among all of the attributes. We can see the similarities in the estimated distributions of the two groups shown in Figure 16.

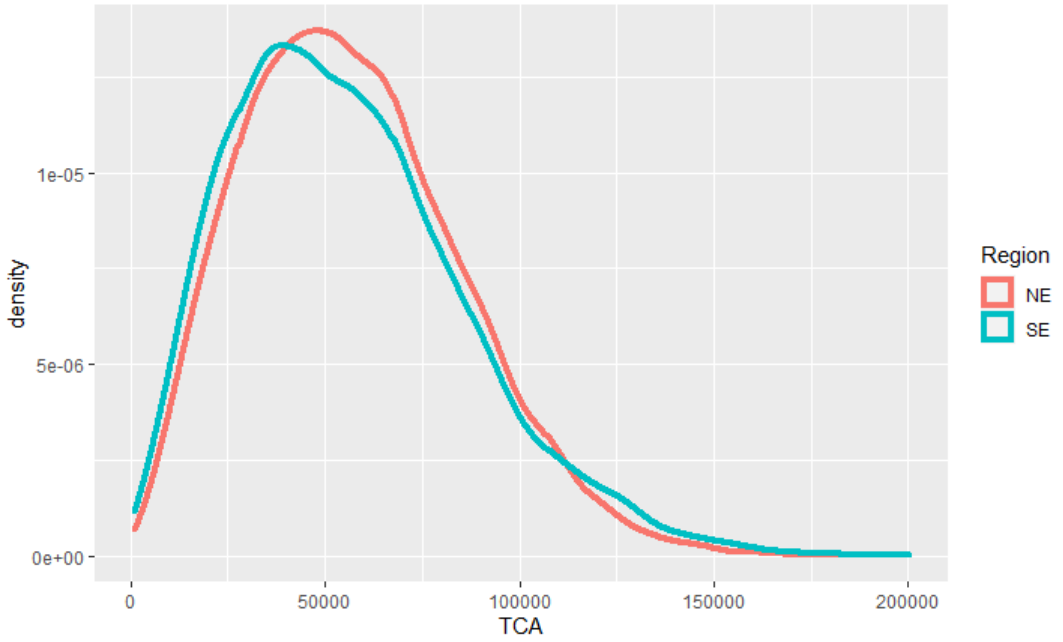

Figure 16: Comparing Region Distributions

If there are any differences, one could say that the NE has more incidents with some higher claim amounts compared the the SE. Otherwise, we cannot detect too many differences. This implies that even though insurance varies by state, it does not largely vary from region to region.

5 Conclusion

Overall, this research extensively examined the Gamma and Weibull distributions with respect to automobile insurance-claim data. The specific automobile insurance-claim data examined included four different types of claims. At first we used the Gamma distribution to estimate the parameters for two of these claims, injury and property, and used the Weibull distribution for the other two claims, total and vehicle. This gave us results that suggested the estimated distributions were not a good fit for the data. However, we noticed that the

data had specific areas of differences and outliers that led us to look closer at our data. Because we then separated the data into two different portions, we were able to eliminate the outliers and find the most accurate fit for the data that could be represented by the Gamma or Weibull distribution. If this research were to be done differently, a data science approach could be taken. This would divide the data into two or more sets and use one set to run analysis on to see if it can predict the other sets. Overall, we can conclude and concur with recent literature, that depending on the shape of the data, the Gamma and Weibull distribution's are the best distributions when fitting insurance data. We also found that there are instances where insurance data cannot be fit or shown in a distribution model.

References

[1] Ahmad, Zubair, Eisa Mahmoudi, and G. G. Hamedani. "A Family of Loss Distributions with an Application to the Vehicle Insurance Loss Data." *Pakistan journal of statistics and operation research* (2019): 731–744.

[2] Akash, Mervyn. *Insurance-Claim*. V1. (2019) Distributed by Kaggle Data Science Company, www.kaggle.com/datasets/mervynakash/insurance-claim?resource=download

[3] Brazauskas, Vytaras, and Andreas Kleefeld. "Folded and Log-Folded-t Distributions as Models for Insurance Loss Data." *Scandinavian Actuarial Journal*, no. 1 (2011): 59–74.

[4] Brouste, Alexandre, Christophe Dutang, and Tom Rohmer. "Closed-Form Maximum Likelihood Estimator for Generalized Linear Models in the Case of Categorical Explanatory Variables: Application to Insurance Loss Modeling." *Computational Statistics* 35, no. 2 (2019): 689–724.

[5] Hogg, Robert V., and Stuart A. Klugman. *Loss Distributions*. New York: Wiley, (1984).

[6] Ramírez, P, J. A. Carta, "Influence of the data sampling interval in the estimation of parameters of the Weibull wind speed probability density distribution a case study." *Energy Conversion and Management*, 46 (2005), 2419–2438.

Appendix

This appendix includes the R Code used throughout the research. Each separate code is specific to the individual processes used the estimate the parameters or metrics we were looking for.

1. *Gamma MME*

```
gamma_MME <- function(CLAIM){
  n <- length(CLAIM)
  mean_x <- mean(CLAIM)
  alpha <- n*(mean_x^2)/sum((CLAIM-mean_x)^2)
  beta <- sum((CLAIM-mean_x)^2)/n/mean_x
  estimate_MME <- data.frame(alpha,beta)
  return(estimate_MME) }
 gamma_MME(CLAIM)
```

2. *Gamma MLE*

```
gamma_MLE <- function(CLAIM){
  n <- length(CLAIM)
  mean_x <- mean(CLAIM)
  # initiate the convergence and alpha value
  converg <- 1000
  alpha_initial <- INITIAL ALPHA ESTIMATE
  # initiate two vectors to store alpha and beta in each step
  alpha_est <- alpha_initial
  beta_est <- mean_x/alpha_initial
  # Newton-Raphson
  while(converg>0.0000001){
    #equation
    eq <- n*log(alpha_initial/mean_x)-n*digamma(alpha_initial)
    +sum(log(CLAIM))
    #first derivative
    der1 <- n/alpha_initial-n*trigamma(alpha_initial)
    #calculate next alpha
    alpha_next <- alpha_initial-(eq/der1)
    # get the convergence value
    converg <- abs(alpha_next-alpha_initial)
    # store estimators in each step
    alpha_est <- c(alpha_est, alpha_next)
    beta_est <- c(beta_est, mean_x/alpha_next)
    # go to next alpha
    alpha_initial <- alpha_next }
  alpha <- alpha_next
  beta <- mean_x/alpha_next
  estimate_MLE <- data.frame(alpha,beta)
```

```
return(estimate_MLE) }
```

Both the MME and MLE can be confirmed by using the preexisitng R Package *EnvStats*. This allows us to use the *egamma* function which provides parameter estimates using each individual claim data.

3. *Weibull MME*

```
weibull_MME<-function(CLAIM){
  n<-length(CLAIM)
  mean_x<-mean(CLAIM)
  s <- sd(CLAIM)
  k_initial<-(mean_x/sqrt((s^2)))^1.086
  # initiate the convergence
  converg<-10000
  # initiate two vectors to store alpha and beta in each step
  k_est<-k_initial
  lambda_est<-mean_x/(gamma(1+(1/k_initial)))
  # Newton-Raphson
  while(converg>0.0000001){
    #equation
    eq<-(((s^2)/(mean_x)^2)+1)*gamma(1+(1/k_initial))
    *gamma(1+(1/k_initial))-gamma(1+(2/k_initial))
    #first derivative
    der1<-2*(((s^2)/(mean_x)^2)+1)*(digamma(1+(1/k_initial))
    *(-1/(k_initial)^2)*gamma(1+(1/k_initial)))
    -(digamma(1+(2/k_initial))*(-2/(k_initial)^2))
    #calculate next k
    k_next<-k_initial-eq/der1
    # get the convergence value
    converg<-abs(k_next-k_initial)
    # store estimators in each step
    k_est<-c(k_est, k_next)
    lambda_est<-c(lambda_est, mean_x/(gamma(1+(1/k_initial))))
    # go to next k
    k_initial<-k_next}
  k<-k_next
  lambda<-lambda_est<-mean_x/(gamma(1+(1/k_next)))
  estimate_MLE<-data.frame(k,lambda)
  return(estimate_MLE)}
weibull_MME(CLAIM)
```

4. *Weibull MLE*

```
weibull_MLE<-function(CLAIM){
  n<-length(CLAIM)
  mean_x<-mean(CLAIM)
```

```
  k_initial<- Previous estimated k
  # initiate the convergence
  converg<-10000
  # initiate two vectors to store k and lambda in each step
  k_est<-k_initial
  lambda_est<-((1/n)
  *sum((CLAIM)^k_initial))^(1/k_initial)
  # Newton-Raphson
  while(converg>0.0000001){
    #first derivative
    eq<-(1/k_initial)+(sum(log(CLAIM))/n)
    -(sum(((CLAIM)^k_initial)
    *(log(CLAIM)))
    /sum((CLAIM)^k_initial))
    #second derivative
    der1<--(1/(k_initial)^2)
    +(((sum(((CLAIM)^k_initial)
    *(log(CLAIM))))^2)
    /((sum((CLAIM)^k_initial))^2))
    -(sum(((CLAIM)^k_initial)
    *(log(CLAIM))^2)
    /sum((CLAIM)^k_initial))
    #calculate next k
    k_next<-k_initial-eq/der1
    # get the convergence value
    converg<-abs(k_next-k_initial)
    # store estimators in each step
    k_est<-c(k_est, k_next)
    lambda_est<-c(lambda_est,
    ((1/n)* sum((CLAIM)^k_initial))^(1/k_next))
    # go to next k
    k_initial<-k_next  }
  k<-k_next
  lambda<-lambda_est<-((1/n)
  *sum((CLAIM)^k_initial))^(1/k_next)
  estimate_MLE<-data.frame(k,lambda)
  return(estimate_MLE) }
weibull_MLE(CLAIM)
```

Both the MME and MLE for Weibull can be confirmed by using the preexisting R Package *EnvStats*. This allows us to use the *eweibull* function which provides parameter estimates using each individual claim data.

5. *Ages ANOVA/t-test*

```
AgesTCA <- data.frame(c(rep('All',dim(All)[1]),
```

```
rep('Age1',dim(Age1)[1]),
rep('Age2',dim(Age2)[1]),rep('Age3',dim(Age3)[1]),
rep('Age4',dim(Age4)[1]),rep('Age5',dim(Age5)[1]),
rep('Age6',dim(Age6)[1])),
rbind(All[,2],Age1[,2],Age2[,2],
Age3[,2],Age4[,2],Age5[,2],Age6[,2]))
colnames(AgesTCA)<-c('Age','TCA')
boxplot(TCA~Age,data=AgesTCA,horizontal = TRUE)
test1 <- aov(TCA~Age,data=AgesTCA)
summary(test1)
pairwise.t.test(AgesTCA$TCA,AgesTCA$Age,
p.adjust.method="bonferroni")
```

6. *Gender ANOVA*

```
GenderTCA <- data.frame(c(rep('All',dim(All)[1]),
rep('Female',dim(Female)[1]),rep('Male',dim(Male)[1])),
rbind(All[,2],Female[,2],Male[,2]))
colnames(GenderTCA)<-c('Gender','TCA')
boxplot(TCA~Gender,data=GenderTCA,horizontal = TRUE)
test1 <- aov(TCA~Gender,data=GenderTCA)
summary(test1)
```

7. *Time ANOVA/t-test*

```
TimeTCA <- data.frame(c(rep('All',dim(All)[1]),
rep('Early',dim(Early)[1]),rep('Mid',dim(Mid)[1]),
rep('Late',dim(Late)[1])),
rbind(All[,2],Early[,2],Mid[,2],Late[,2]))
colnames(TimeTCA)<-c('Time','TCA')
boxplot(TCA~Time,data=TimeTCA,horizontal = TRUE)
test1 <- aov(TCA~Time,data=TimeTCA)
summary(test1)
pairwise.t.test(TimeTCA$TCA,TimeTCA$Time,
p.adjust.method="bonferroni")
```

8. *Region ANOVA*

```
RegionTCA <- data.frame(c(rep('All',dim(All)[1]),
rep('SE',dim(SE)[1]),rep('NE',dim(NE)[1])),
rbind(All[,2],SE[,2],NE[,2]))
colnames(RegionTCA)<-c('Region','TCA')
boxplot(TCA~Region,data=RegionTCA,horizontal = TRUE)
test1 <- aov(TCA~Region,data=RegionTCA)
summary(test1)
```

9. *Generalized Plot*

```
a <- c(actual_data, simulated_data)
b <- c(rep('Actual', 10130), rep('Estimated', 10130))
df <- data.frame(a,b)
colnames(df)<-c('CLAIM', 'Type')
ggplot(df, aes(x=CLAIM, color=Type))+geom_density(lwd=2)
```

10. *Generalized KS Test*

```
data1 <- CLAIM
data2 <- rgamma(n, shape, scale)
ks.test(data1, data2)
```

11. *Generalized KL Divergence*

```
range(CLAIM)
seq1 <- seq(0,30000,2000)
seq2 <- seq(0, 60000, length.out = 16)
h1 <- hist(CLAIM, breaks = seq1)
h2 <- hist(ESTIMATED, breaks = seq2)
x <- rbind(h1$counts, h2$counts)
KL(x, est.prob = "empirical")
```

12. *Generalized Q-Q Plot*

```
data1 <- CLAIM
data2 <- rgamma(n, shape, scale)
qqplot(data1, data2)
```