**Chris Cirelli**

**HW1**

**09/14/2020**

**Instructions**

Below you will find a CSV file with 1,500 general liability claims randomly chosen from late settlement lags that were provided by Insurance Services Office, Inc.

Each claim consists of an indemnity payment (Loss) and an allocated loss adjustment expense (ALAE).

ALAE are types of insurance company expenses that are specifically attributable to the settlement of individual claims such as lawyers' fees and claims investigation expenses.

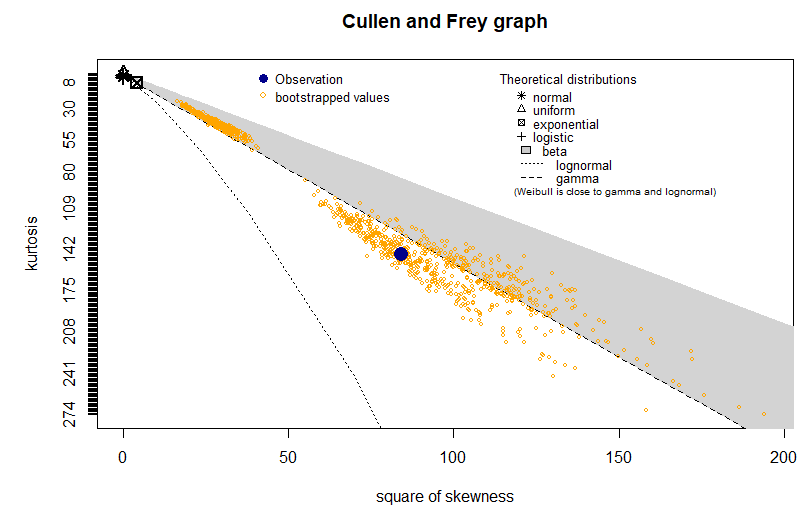
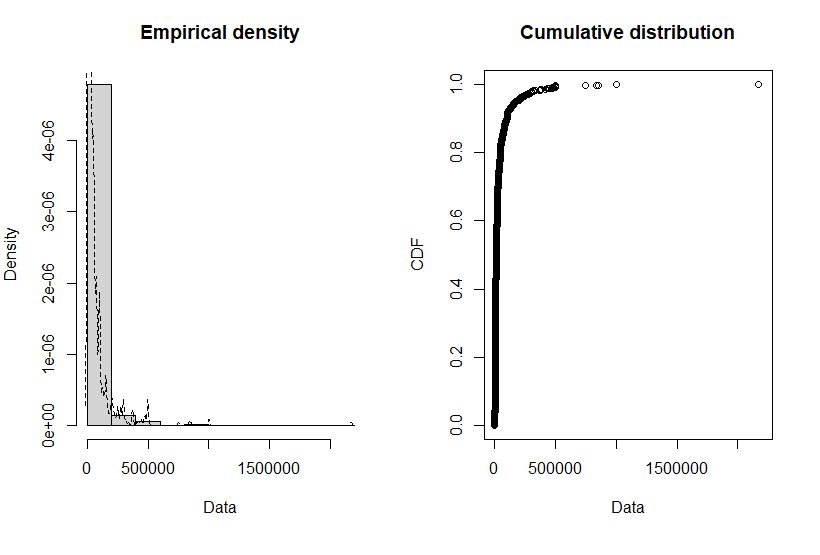
The dataset also has the limit of the policy (Limit) and an attribute called Censored which gives the row names of the indemnity payments that were capped at their policy limit.

You should analyze the data in Excel (or Numbers or R etc.) and write a report in Word or as a PDF. Your analysis should address:

**Question 1:**

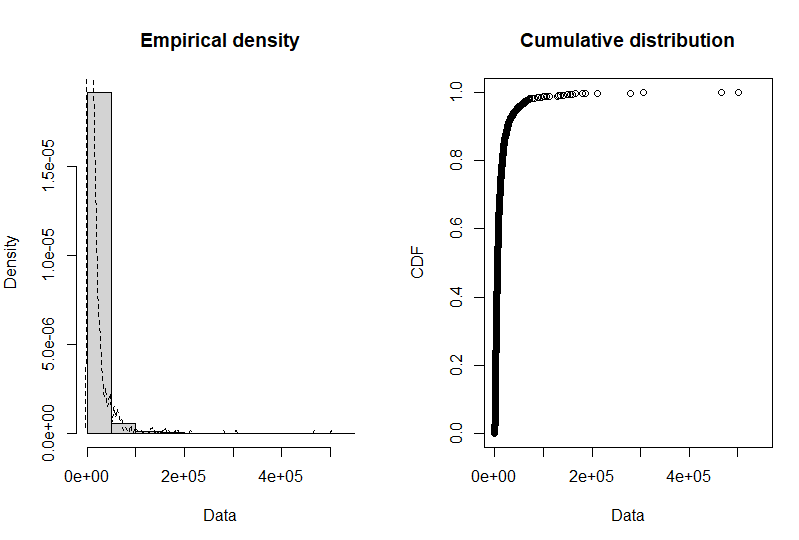
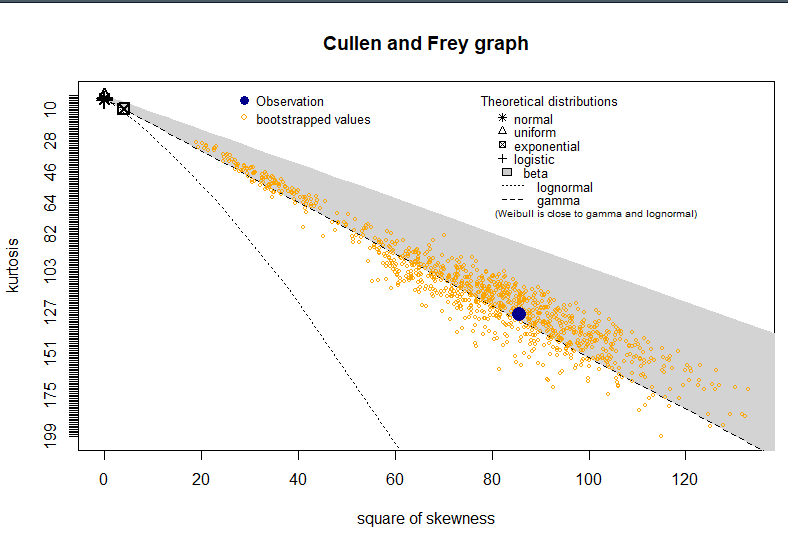
**Stylistic features of the distribution of the indemnity payments. What probability distribution does it most resemble? Is it skewed and/or leptokurtic? You should include histograms or other appropriate graphical elements.**

This distribution of this data is clearly right skewed with the majority of the observations taking place close to zero. Based on the density and cumulative distribution plots this distribution appears to resemble a gamma distribution with k=0.5 and *θ* = 1.0 <https://en.wikipedia.org/wiki/Gamma_distribution>. Based on the Cullen and Frey graph we can see that the kurtosis and square of skewness of this data most resembles a gamma distribution.

****

**Question 2: Sum of Indemnity Payments and ALAE**

ALAE presents a very similar distribution to that of Loss. In fact, based a visual inspection of the Cullen and Frey graph the distribution is more closely resembles a gamma distribution versus loss.



**Fit gamma distribution to data using MLE.**

Fitting of the distribution ' gamma ' by maximum likelihood

Parameters:

estimate Std. Error

shape 0.5061027 0.01527287

rate 12.2813895 0.57969356

> summary(d.gamma)

Fitting of the distribution ' gamma ' by maximum likelihood

Parameters :

estimate Std. Error

shape 0.5061027 0.01527287

rate 12.2813895 0.57969356

Loglikelihood: 3595.047 AIC: -7186.095 BIC: -7175.468

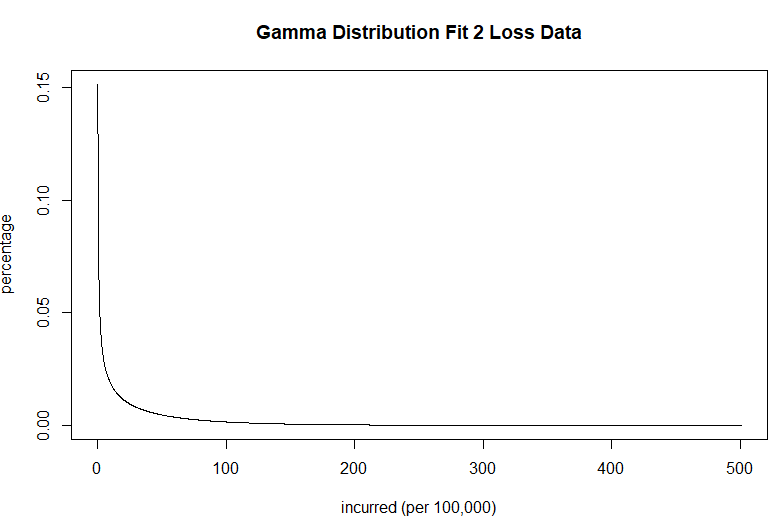
Correlation matrix:

shape rate

shape 1.0000000 0.6393388

rate 0.6393388 1.0000000

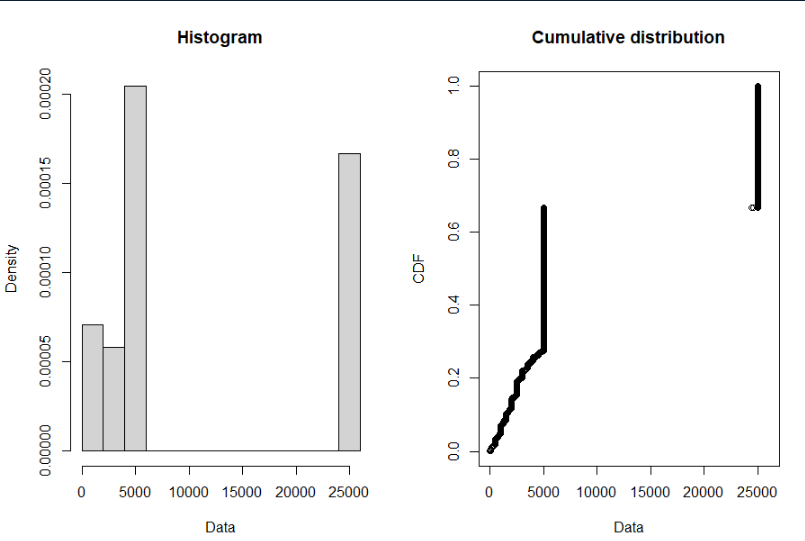
**Plot of Fit Gamm Distribution using above shape & rate parameters**



**Question 3**

What if we imposed a 5000 policy limit on the first 1000 indemnity payments and a 25000 policy limit on the rest? How would the distribution change? How about the mean and variance?

If we impose a policy limit of $5,000 on the first 1000 indemnity payments and $25,000 on the remaining 500, we get a bimodal distribution. This likely occurs as there are a number of claims that exceed each of the thresholds and are simply combined within a single bin.



**Question 4**:

So far, we paid no attention to records with a missing policy limit, as if it meant there is no policy limit. Is this assumption reasonable? How would you deal with a situation like this?



According to the above data dictionary, roughly 10% of the policy limits are missing. We could take a number of approaches to this issue. If we were working for the company who generated this data, we could simply try to rectify the issue with trying to figure out what is generating the error as it is rare for insurance systems to be missing something as important as the insurance limit.

Alternatively, and given that only 10% of the observations are missing, we could eliminate this data from the dataset.

Lastly, we could try to approximate the limits using the average or fit a regression line to limit ~ premium in order to generate predictions for each missing limit.