ECS132 Term Project Report

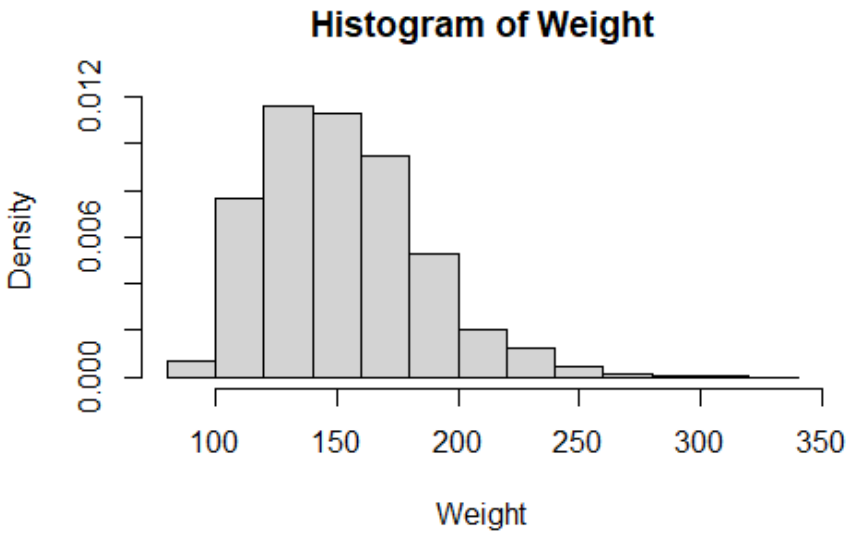
Jonathan Tran
Alex Din
Haosen Cao

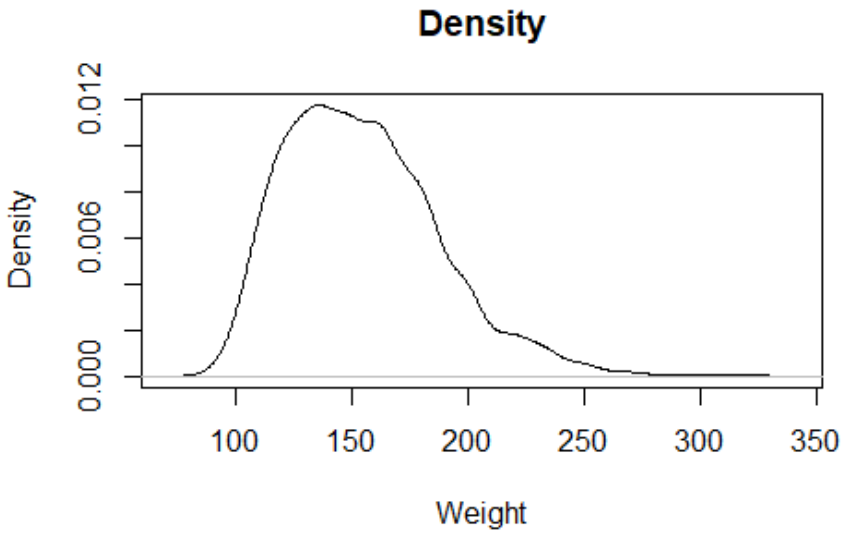
June 11, 2023

Contents

Normal Family

To model the normal family of distributions we decided to base our model on the weight column of the national.longitudinal.survey data, which contains data of 4908 individuals' statistics like age, race, height, and weight. The weight is distributed in the range from 87 to 325. When we plotted the histogram, most of the data are in the range from 100 to 200, and as we go to the left or right of that, the number of points decreases. But most importantly the data looked similar to a bell curve.





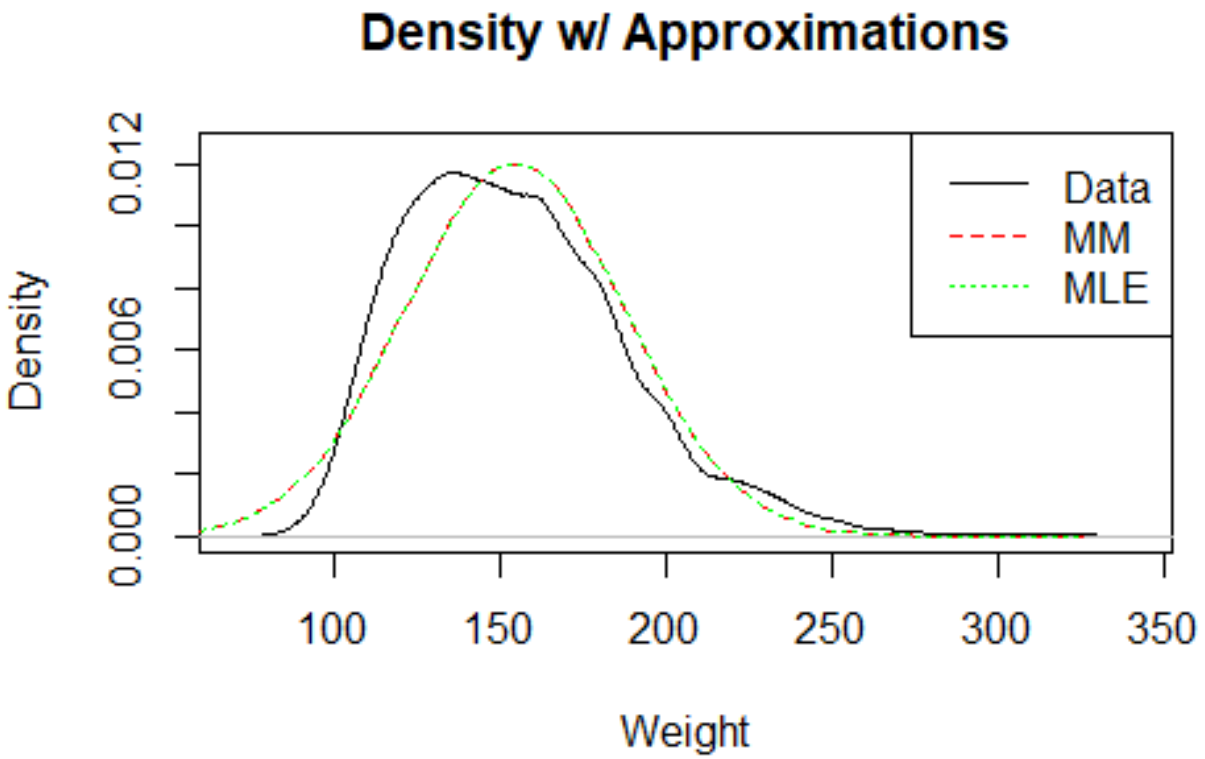
The normal family of distributions has two parameters, mean and standard deviation. We are assuming that the weights in this dataset can be modeled with a normal distribution. Therefore we use two methods to find estimates for the mean and standard deviation parameters, which we will call M and SD, respectively.

For maximum likelihood, we used R's `mle()` function to find M and SD, which takes a negative log-likelihood function. While the likelihood function is the product of density functions on the given data, the log-likelihood is the sum of the logarithm of those density functions. Therefore we use this line as our negative log-likelihood function:

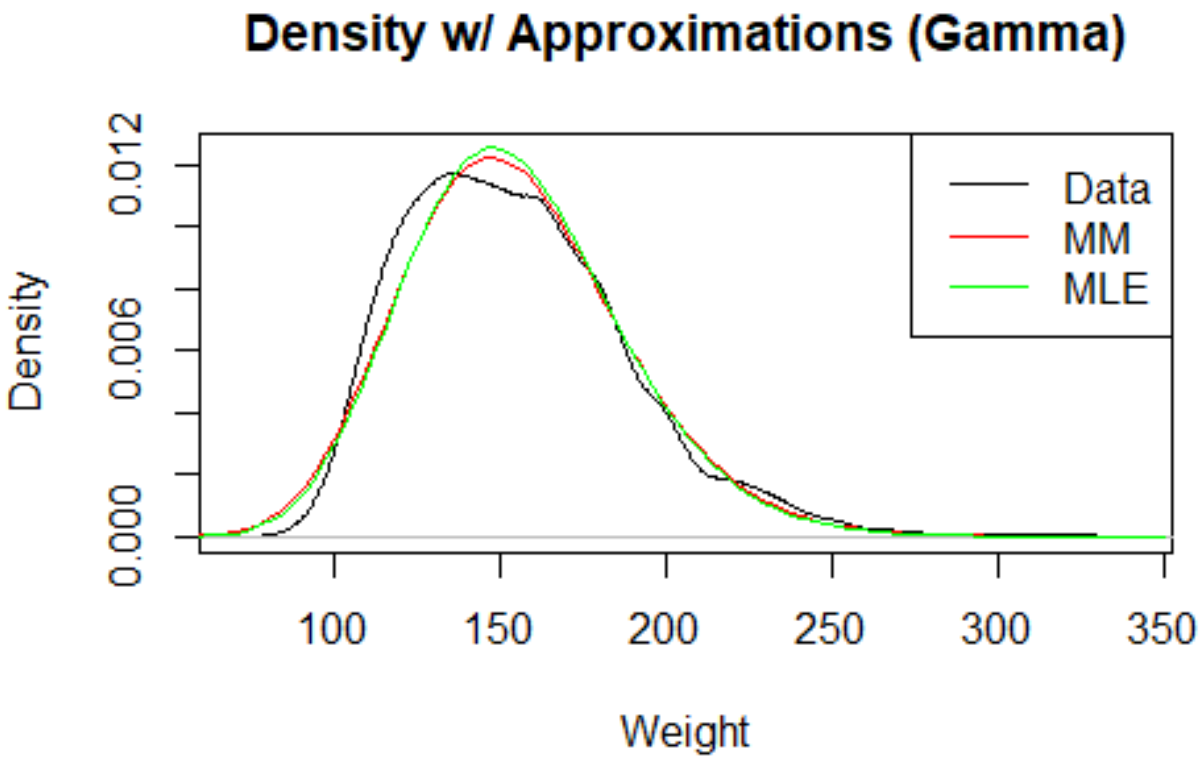
```
loglik <- sum(dnorm(weights, M, SD, log = TRUE))
```

Where weights is our weight data, and M and SD are our parameter estimates. Plugging this into the `mle()` function gives us $M = 154.5827$ and $SD = 33.18112$.

The method of moments for a normal distribution is relatively simple. The two parameters of the normal distribution are the mean and standard deviation, which we can generate estimates for using our sample data directly from R. M is just $\bar{A}$, our sample mean from calling $mean(weights)$, and SD is just $\sqrt{s^2}$ which can be given by calling $sqrt(var(weights))$. The difference between using S^2 and s^2 is negligible. Using this method, we get $M = 154.5827$ and $SD = 33.18453$.



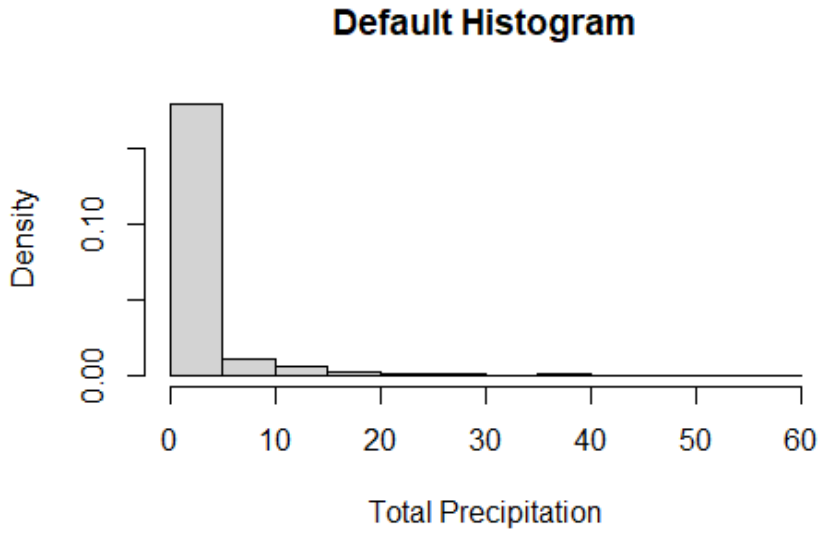
We would say that the normal family is a suitable estimator for the weights in this dataset. A normal distribution, according to the central limit theorem, is what results from a set of summations of random variables. The weight of a person can be thought of as a sum of the various parts of their body, whose sizes can be thought of as random. However, based on the graph, which has a slightly steeper slope on the left side than on the right, it can be argued that a gamma distribution would better model this data.

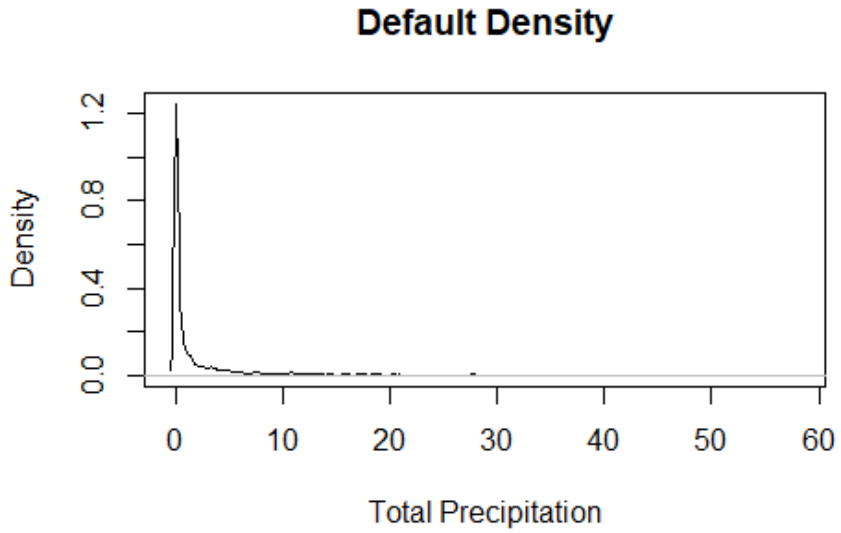


And from the graph, we must concede that the gamma family may model this data better than the normal family. But if the central limit theorem says that sums of random variables look normal, and the gamma distribution is a sum of exponentially distributed random variables, it's no wonder that the graphs look pretty similar. If we're summing a large enough number of exponential variables, it's possible that the gamma distribution could converge to a normal distribution. From this line of thinking, the normal and gamma distribution are likely both good estimators for the weights in the dataset.

Exponential Family

To model the exponential family of distributions we decided to base our model on the PRECTOT column of the weatherTS data, which looks like it took weather data from Eastern Australia for 10 years (1985–1995). Of this data, the one we were most interested in was the PRECTOT column, which we interpreted as the total precipitation on a given day. When we plotted the histogram, the data was heavily skewed left, tapering off more to the right, which resembled that of an exponential graph. The most important thing was that the peak of the data was at $x = 0$, which is where exponential graphs peak.





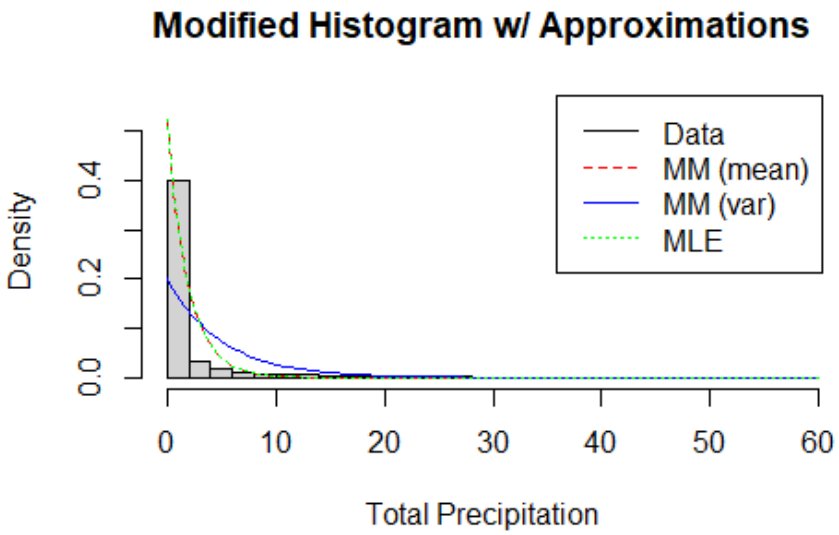
Just from looking at these graphs, while the density function doesn't look quite exponential, the histogram definitely does. We will discuss possible reasons why later on. But for now, we will find estimators for the parameters of the exponential family. The exponential family is a one-parameter family, that being lambda. Using the two methods, we are trying to find an estimator for lambda, which will be referred to as L. We use this line of code to get the negative log-likelihood function for our maximum likelihood.

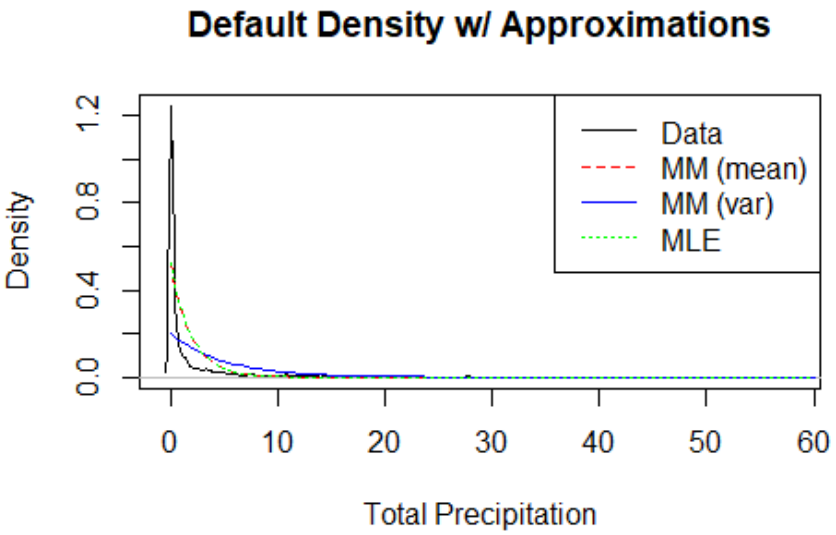
```
loglik <- -sum(dexp(prectot, L, log = TRUE))
```

Where prectot is our precipitation data and L is our estimator for lambda. Plugging this into the `mle()` function we get L to be 0.5230325.

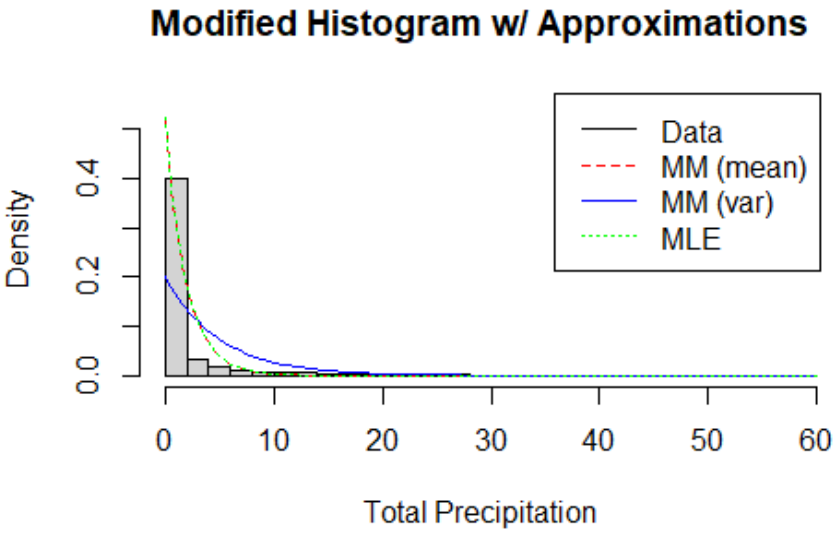
Next, for method of moments, we algebra to estimate L. For a given exponentially distributed variable X, the expected value is given by $E(X) = \frac{1}{\lambda}$. Assuming the mean of our sample data, $\bar{A}$, is an unbiased estimator for the true mean of the population data, $\bar{A} = \frac{1}{L}$, and therefore $L = \frac{1}{\bar{A}}$. Plugging this into R we get L to be 0.5230367.

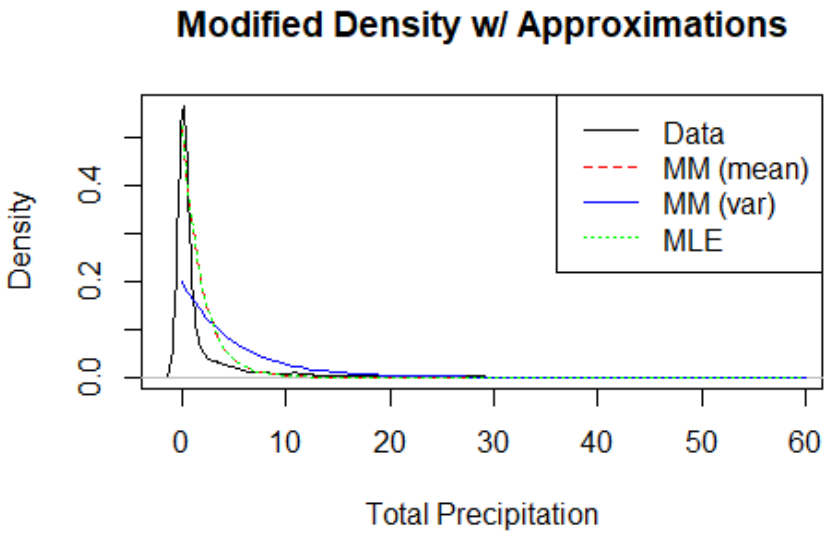
Because this is a one-parameter family, we can also derive L from our sample variance s^2. The variance of an exponentially distributed variable X is given by $Var(X) = 1/\lambda^2$. Therefore L is given by $L = \sqrt{1/s^2}$. With this method, L is shown to be 0.2002137, which is very different from the other two methods above. From the graphs below, this particular estimate for lambda isn't very good.



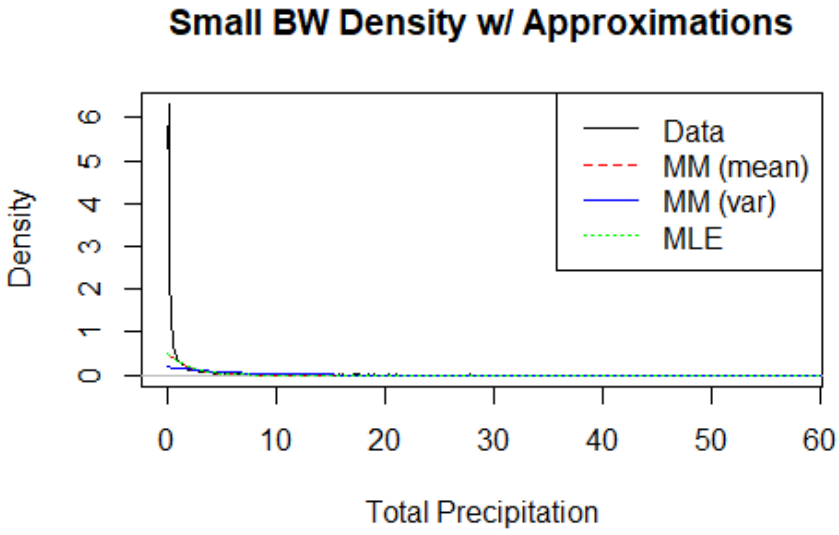


It is important to note that our exponential distribution plots don't quite match those of the `hist()` or `density()` functions on our sample data, at least with their default arguments. These discrepancies can be attributed to a bandwidth that doesn't accurately reflect our sample. By decreasing the bandwidth for the `density()` function and increasing it for the `density()` function, we get graphs that better match our plots.





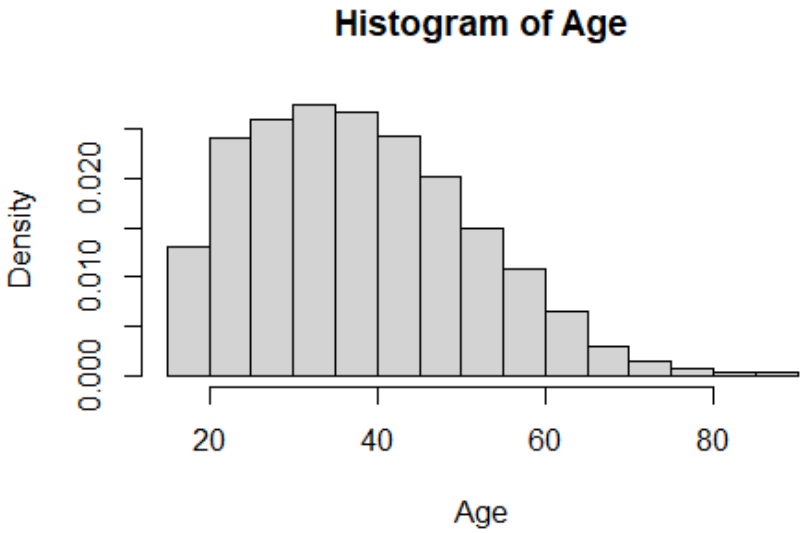
The histogram looks good, but the density function is arguably less exponential-looking as a result. Lowering the bandwidth of the density function does make it more exponential, but likely at the cost of being less accurate due to how few points can actually fit into an individual bandwidth.



But even with these discrepancies, given our plots of the exponential distribution, we would say this family is a suitable estimator for the total precipitation on a given day in Australia. The exponential distribution is said to be the continuous analog of the geometric distribution. How much precipitation you get in the day is pretty much the same as how long it rains on a given day. And how long it rains in a given day can be thought of as successive failures (the rain continuing) until a single success (the rain stopping), which is like a geometric distribution. Obviously, this is an oversimplification but it shows how an exponential distribution can model the density function of total precipitation.

Gamma Family

To model the gamma family of distributions, we decided to use the age column from the dataset adultfinal, which seems like it contains various data from specifically working individuals. We decided to use age because the histogram we observed was very gamma in nature. It shows a steep increase in the number of people in the beginning and more gradually decreases as we go up in age.



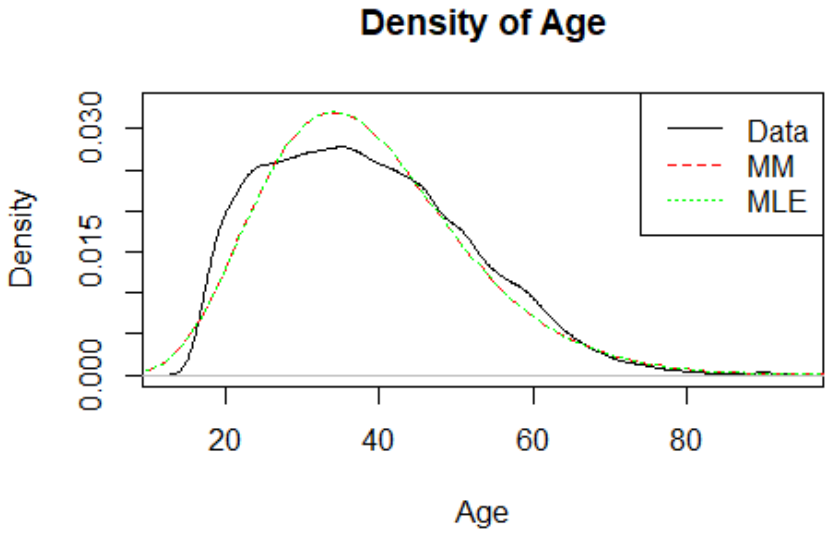
The gamma family can be described as a sum of exponentially distributed random variables. So the gamma family has two parameters, $lambda$, the same kind as those of the exponential family, and r, the number of exponential variables that are summed together. Assuming that this age data can be modeled with a gamma function, we are trying to find estimates for lambda and r, which will be called L and C, respectively. We use this line of code for the maximum likelihood estimator:

```
loglik <- -sum(dgamma(age, C, L, log = TRUE))
```

Pluging this into the `mle()` function gives us $C = 8.596535$ and $L = 0.2236146$. For the method of moments, we again use algebra to find C and L.

$$E(X) = \frac{r}{\lambda} \qquad \text{Var}(X) = \frac{r}{\lambda^2}$$
$$\bar{A} = \frac{C}{L} \qquad s^2 = \frac{C}{L^2}$$
$$C = \bar{A}L \qquad s^2 = \frac{A}{L}$$
$$L = \frac{A}{s^2}$$

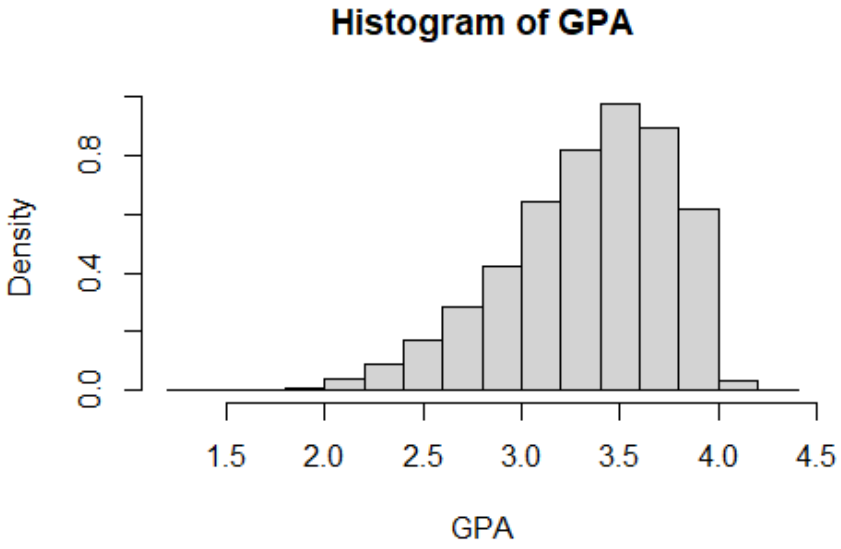
Substituting s^2 with $var(age)$ and $\bar{A}$ with $mean(age)$ we get $C = 8.587819$ and $L = 0.2233877$.

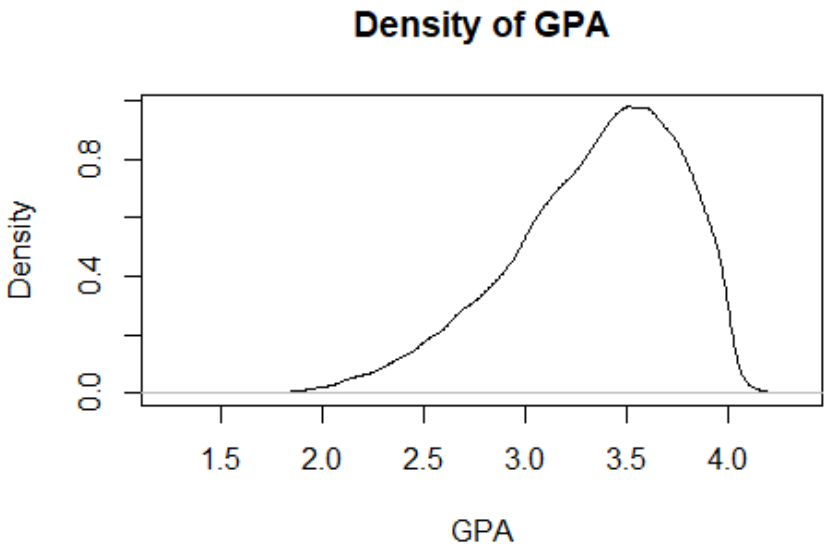


Judging from the graph, the gamma distribution is a decent estimator for the ages in this dataset. It works especially well for estimating the number of people aged 45 and above. It doesn't work as well for the rest of the data, but it does the job well enough. In any case, it would be hard to find a distribution family that better fits this data. It is important to note that the ages from this sample don't exactly reflect the age of everyone in a given population. There is some sampling bias here in that the data is only taken from people who have an occupation, and notably, there is no data from anyone aged 16 or below. Because of this, it is hard to justify why this dataset's age can be described as a gamma-distributed variable since there are a couple of factors going on here. But in terms of simply fitting the graph of the data, the gamma distribution does the job just fine.

Beta Family

To model the beta family of distributions we decided to use the GPA column of the lawschoolbrief dataset. GPA has a lower and upper bound, which is the kind of data a beta distribution can model. Judging that the dataset consists of statistics of law school applicants, meaning that the GPA is a college GPA, and that the maximum GPA found in this dataset is 4.23, we decided to use 4.3 (what some schools give for an A+) as the upper bound for our beta distribution. And of course, we're using 0 as our lower bound.





The beta family of distributions has two parameters, alpha and beta. Assuming that the GPA data can be modeled with a beta distribution, we are trying to find estimates for alpha and beta, which we will call a and b, respectively. We use this line of code for the maximum likelihood estimation:

```
loglik <- -sum(dbeta(gpa/fac, a, b, log = TRUE))
```

Where gpa is our data column. Since the beta distribution is by default bounded between 0 and 1, we have to scale our data that's bounded between 0 and 4.3 in order to fit those bounds. So we divide every value in our dataset by fac, 4.3, and then rescale the graph after we've gotten our parameters. From the `mle()` function those parameters are $a = 13.68352$ and $b = 3.932735$.

For the method of moments, we use algebra to derive a and b.

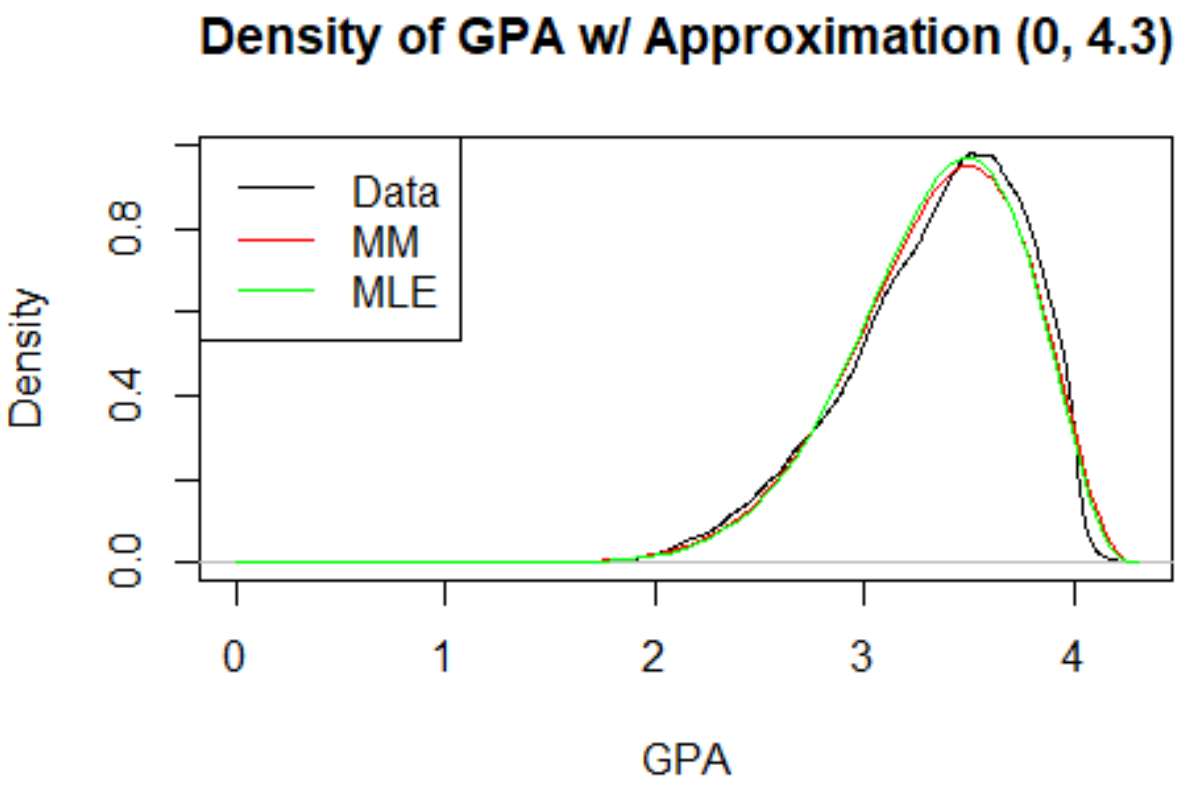
$$\begin{aligned}
\mathrm{E}(X) &= \frac{\alpha}{\alpha+\beta} \\
\bar{A} &= \frac{a}{a+b} \\
(a+b)\bar{A} &= a \\
a\bar{A} + b\bar{A} &= a \\
b\bar{A} &= a - a\bar{A} \\
b &= a/\bar{A} - a \\
b &= a\left(\frac{1}{\bar{A}} - 1\right)
\end{aligned}$$

The derivation for the other variable is tedious and the details aren't important, so we took this from The Book of Statistical Proofs. [1]

$$\begin{aligned}
a &= \bar{A}(\frac{\bar{A}(1-\bar{A})}{s^2}) - 1) \\
\mathrm{Var}(X) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
s^2 &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}$$

Plugging these into R gives us $a = 13.03444$ and $b = 3.7384$.

[1] https://statproofbook.github.io/P/beta-mome.html



Of all the other distributions we've graphed, this beta distribution created the best-fitting graph for the data we used. So we would say the beta distribution is a suitable estimator for the GPA of applicants to this specific law school. The beta distribution excels at modeling data in a finite range, so it makes sense that GPA, which has a finite range of values, is well modeled by this distribution family.