ECS 132 Term Project

Ryan Vakhshoori, Hannah Aldor, Parsa Bazargani

June 13, 2023

1 Overview

For any given sample X, we would like to give rise to plausible explanations for the density of X being "well-approximated" by the density of a continuous family of distributions which we already know about. In this project, we found a set of sample data for each of four distributions: normal, exponential, beta, and gamma. We started by plotting the data as a histogram using R's `hist()` function to determine the general shape of the data's distribution. Then, we used R's `density()` function to match the density of the data to the given distribution. To test our hypothesis that the data could be modeled using the chosen distribution, we used two techniques to estimate the parameters of the distribution: the Method of Moments Estimators (MME) and the Maximum Likelihood Estimators (MLE).

To estimate the parameters using MME, we generate equations that relate attributes of the sample data to the parameters. For distributions with one parameter, we related the mean of the sample data to the estimator of the parameter to solve for the estimator. For distributions with two parameters, we added a second equation that relates the variance of the sample data to the estimators of the parameters. Then, we solved the system of two equations to determine the estimators for both parameters. To estimate the parameters using MLE, we wrote a function in R to return the negative log likelihood function. In general, these functions followed the form:

```
ll <- function(parameters) {
    loglik <- sum(d(dist)(x, parameters, log = TRUE))
    -loglik
}
```

Then we called R's `mle()` function to perform the calculation:

```
    z <- mle(minuslogl = ll, start = list(parameters = 1))
```

We superimposed the density of the distribution with the estimated parameters over the original histogram and analyzed the suitability of the distribution. In the following sections, we perform this analysis for each of the four distributions.

2 Normal Distribution

2.1 LSAT Scores

Relating a sample of observed LSAT scores to the normal family of distributions, why might we assert that the probability distribution of LSAT scores seems to be inherently bell-shaped? Beyond interpreting the histogram of `lawschoolbrief$LSAT` as being "approximately" bell-shaped, something we can conclude simply by "eyeballing" the data, we could also consider the normal distribution as a target grade distribution for standardized tests. In other words, what properties of the normal family are ideal for performance across the entirety of students who take the LSAT?

One property of the normal family that seems ideal is that 68% of the data for a normally distributed random variable falls within 1 standard deviation of the mean. On the other hand, 26% of the data is 2 standard deviations from the mean, and the more extreme data lying 3 standard deviations away from the mean make up a measly 4.7% of the data. In plain English, the fact that most of the data for a normally distributed random variables lies within 1 standard deviation from the mean implies that most students perform, well, about average. On the other hand, fewer people perform above average or below average, but not exceptionally poor or exceptionally better than the mean. The exceptional cases, those people whose scores lie 3 standard deviations away from the mean (perhaps your average genius or someone who had a bad day or missed the exam) are the fewest in the sample population.

2.2 Calculation of Parameter Estimators

See `normal.r` for MLE code.
MME: M is an estimator for μ and S is an estimator for σ.

$$M = \frac{1}{n} * \sum_{i=1}^{n} x_i$$

$$S = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (x_i - \mu)^2}$$

2.3 Visuals

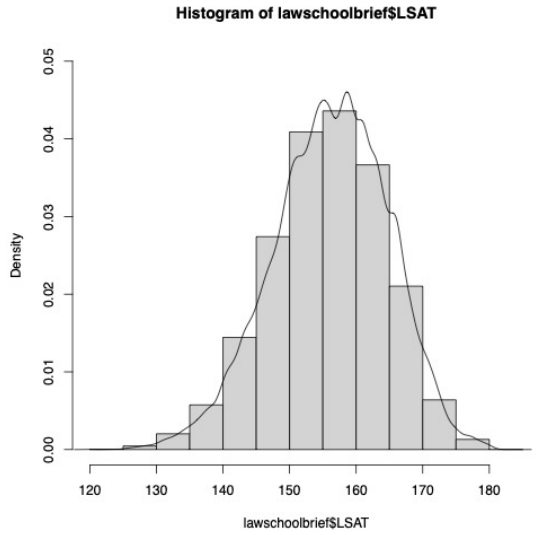

Figure 1: Sample Density Superimposed on Sample Histogram

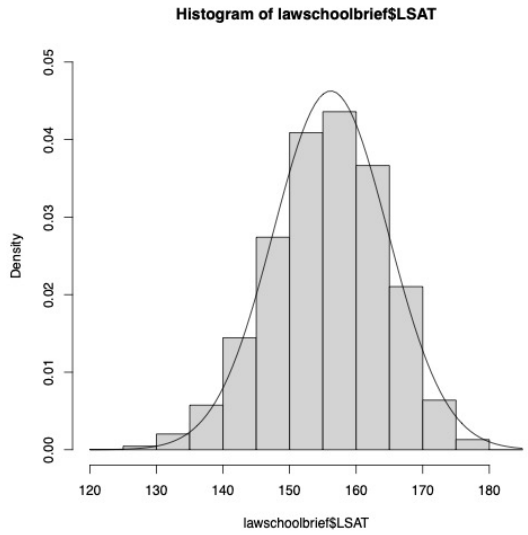

Figure 2: MME Density Superimposed on Sample Histogram

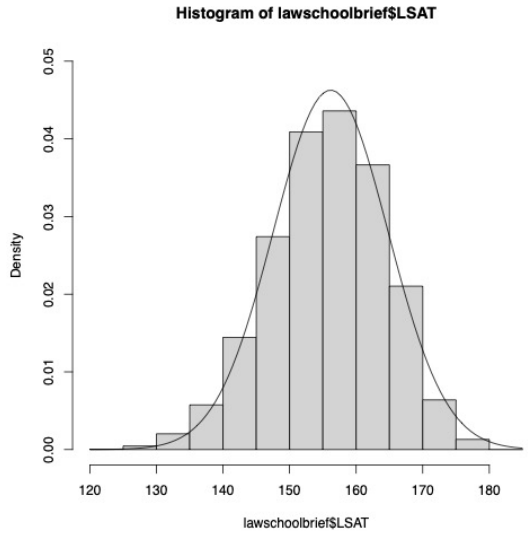

Figure 3: MLE Density Superimposed on Sample Histogram

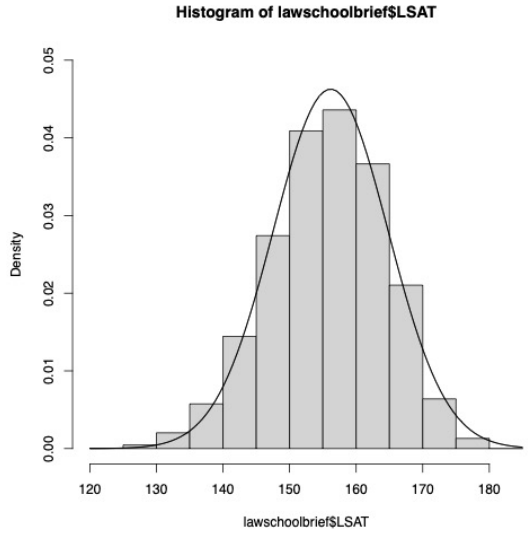

Figure 4: MLE Density and MME Density Comparison Superimposed on Sample Histogram

2.4 Analysis

Revisiting the possibility that the probability distribution of LSAT performance is approximately normal, we also observe the inherent bell shape of another column of LSAT grades observed in a different sample taken from the same database, `law.school.admissions$LSAT`. This information could lead us to wonder if perhaps all scores for LSAT offerings in a geographical region have the same distribution. Although we limit our scope of analysis to LSAT performance, it would not be surprising to see grade distributions for any other standardized test be well-approximated by the normal family.

Examining first the sample density superimposed over the sample histogram, we first note the shape of the density. The sample density itself has many bumps that disappear when raising the bandwidth. This is to be expected, as a finite sample is expected to have variability between

individual points. Although the visual after increasing the bandwidth was not included, the reader should make note that the bell shape is preserved, and generally should not affect the effectiveness of the estimate. While looking at both the MLE and MME density plots, the parameters produced by both estimators turn out to evaluate to the same numbers. This, also, is to be expected, since the estimated parameters to the family are simply the sample mean and sample variance. The fact that the MLE and MME density functions coincide is not at all surprising.

Furthermore, we want to investigate how well the estimators fit the sample. The most obvious detail is that the sample density approximation itself is not perfectly bell-shaped, noting a slight asymmetry to the sample histogram. Rather than overlook this defect, we could note that the slight asymmetry does not in fact rule out the possibility of the data being normally distributed, as the sample is taken from a proper subset of our population. We are also simply looking for what appears to be a good estimate. We note that the discrepancies between the MLE and MME density estimators are small, and overall reflect well the shape of our sample data, without any needed treatment of modifying the bandwidth or binwidth parameters to the plots.

3 Exponential Distribution

3.1 Frequency of Having A Specific Number Priors

Denoting X_i as the frequency of observed people with i priors, $0 \leq i < 30$, we formally define "priors" as the number of previous arrests and/or convictions observed in an individual. Based on the shape of the histogram, the majority of people observed in the sample have 0-1 priors, which is about over five times the amount of people that have committed 2-3 priors. For every two additional priors, the frequency is cut approximately in half. In other words, the amount of people with 4 or 5 priors is twice as large of those who have committed 6 or 7.

The histogram of `compas$priors_count` is exponential due to the rapid decay of frequency following $X = 0$, and the subsequent, but slow, flattening out of the curve once we reach more priors. Why can we conclude that the probability distribution of priors is displayed exponentially? If we randomly select someone from a crime report, likely we select someone that has done their first or second crime. Thus, we also observe that the percentage of people committing over 10 crimes, over 20 crimes grows far less likely. We observe that the data is concentrated to the left, or "skewed-left". This means that the median of the data will likely belong to someone who has either committed no priors or committed 1, but the mean will represent a number of priors that a far majority of the population has not reached.

3.2 Calculation of Parameter Estimators

See `exponential.r` for MLE code.
MME: L is an estimator for λ.

$$L = \frac{1}{\sum_{i=1}^{n} x_i}$$

3.3 Visuals

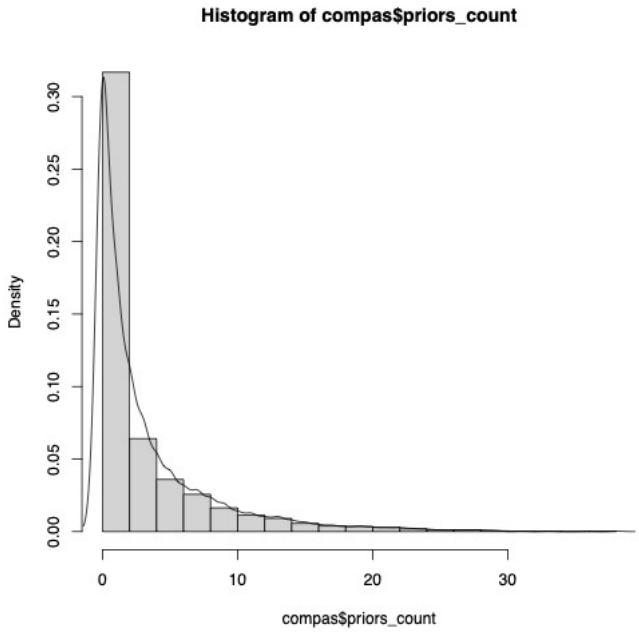

Figure 5: Sample Density Superimposed on Sample Histogram

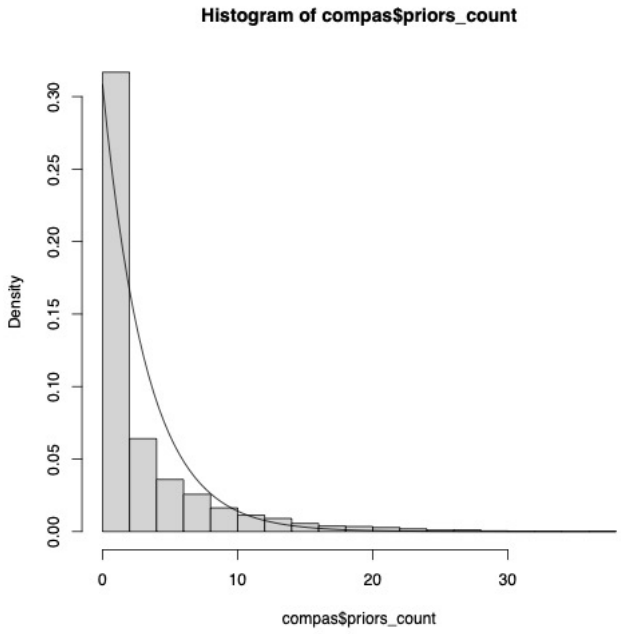

Figure 6: MME Density Superimposed on Sample Histogram

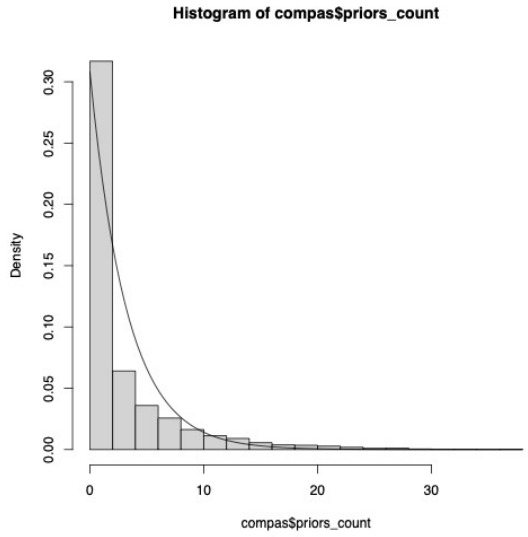

Figure 7: MLE Density Superimposed on Sample Histogram

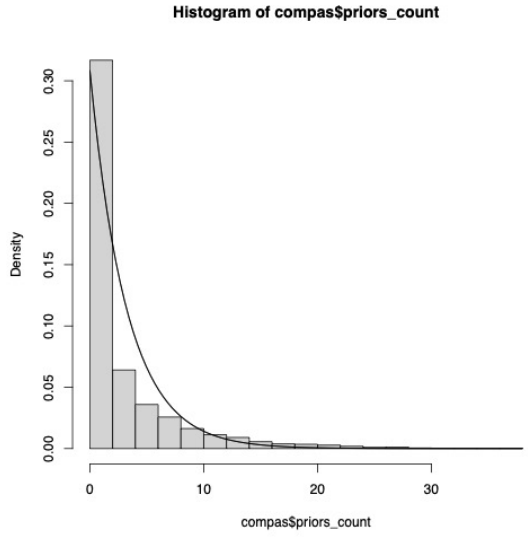

Figure 8: MLE Density and MME Density Comparison Superimposed on Sample Histogram

3.4 Analysis

Regarding the sample density superimposed on the histogram, the curve of the density follows the histogram accurately, as it only overestimates the density by about 0.05 at $X = 2$, and underestimates the density at $X = 1$ by about 0.1. However, this is due to the rapid decay of our exponential curve. If we were to decrease the bandwidth, we would see the curve represent the histogram data far more accurately, and subsequently follow the rapid drop in individuals more precisely. It would maintain an exponential shape, but we would encounter more bumps that would make the curve less smooth.

Figure 8 shows that the estimators are exactly the same in estimating our data set. Even if we zoom in, we cannot tell the estimators apart when they are superimposed on each other. When calculating lambda for our two estimators, our values were only 0.0000005 apart, which is

displayed in our graphs. However, the estimators seem to be overestimating some values while underestimating others during the rapid decay from $X = 0$ to $X = 2$. It measures the density to be lower than the sample shows at $X = 0$ and 1, while overestimating the sample at $X = 2$ and 3. Both MME and MLE continue to slightly overestimate the sample until we reach $X = 8$. From there, it effectively parallels the sample data. Despite the small irregularities between our data set and estimators, our estimators overall mimic the shape of our data. While there are slight differences between them during the rapid decay, the estimators still follow the overall message of the data, which indicates a significant decrease of individuals from having 0 or 1 prior counts to having 2 or 3. With the lack of deviation between MME and MLE, and the minimal discrepancies between our estimators and the data itself, we can establish that our method to determine our estimators through the exponential family for this data set is accurate.

4 Beta Distribution

4.1 Wind Speed on Different Days

Let X_i be the frequency of observed days with the specific wind speed i on a certain day, $0.0 \leq i < 0.55$. These values have been normalized, since the creators of the data set divided the true observed wind speed values by 67. For example, the mean normalized speed, 0.19, multiplied by 67 would give us 12.73 km/h, which accurately matches the average wind speed calculated by **weatherspark.com**. Also, note that the beta family has support (0,1). In this situation, our data is between 0 and 0.55, which is in between the bounds for beta. We maintain the natural bounds by representing all of the registered data in the data set on our graphs.

4.2 Calculation of Parameter Estimators

See `beta.r` for MLE code.
MME: A is an estimator for α and B is an estimator for β.

$$\sum_{i=1}^{n} x_i = \frac{A}{A+B}$$

$$\frac{1}{n} * \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{AB}{(A+B)^2(A+B+1)}$$

4.3 Visuals

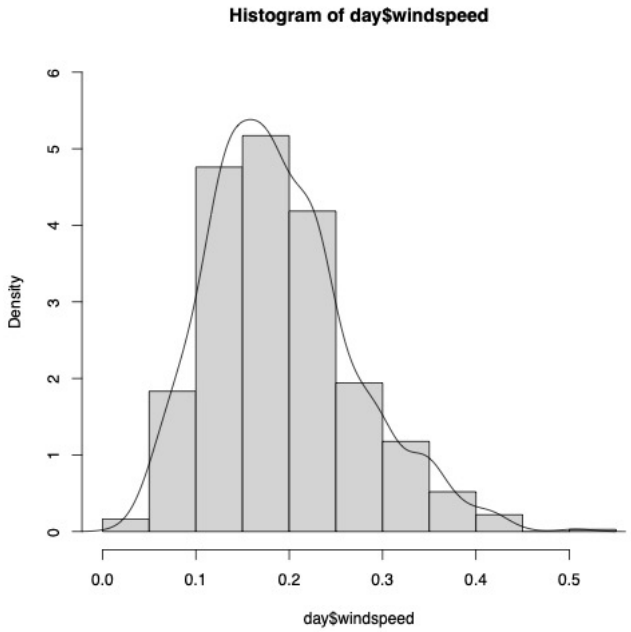

Figure 9: Sample Density Superimposed on Sample Histogram

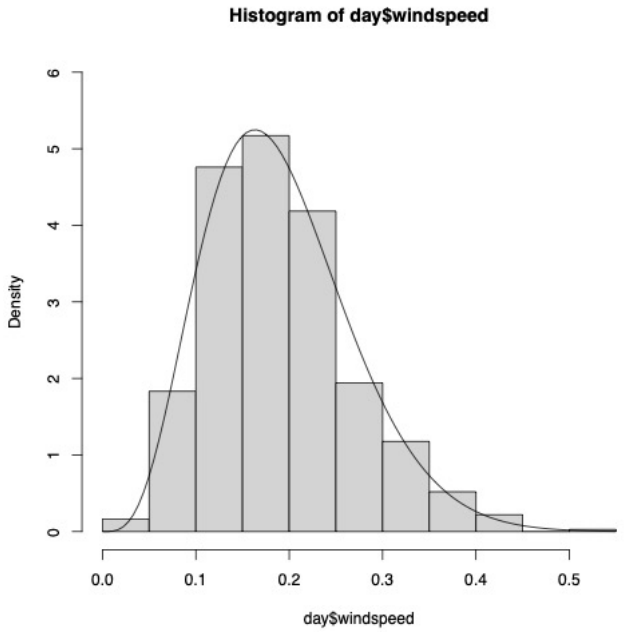

Figure 10: MME Density Superimposed on Sample Histogram

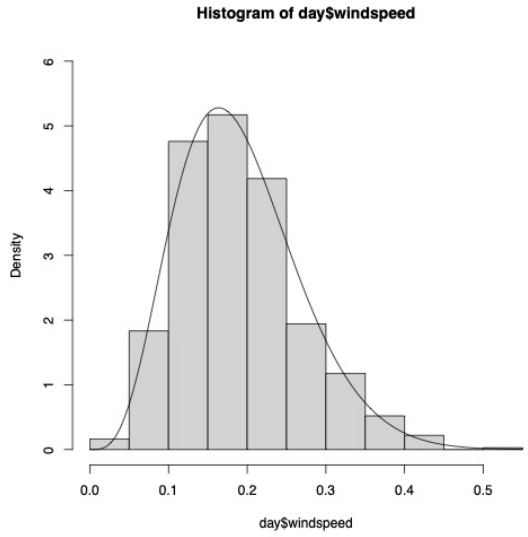

Figure 11: MLE Density Superimposed on Sample Histogram

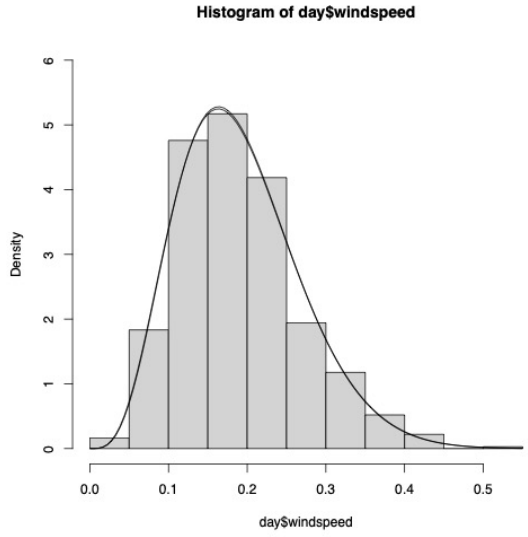

Figure 12: MLE Density and MME Density Comparison Superimposed on Sample Histogram

4.4 Analysis

Converting our histogram into a density curve, the curve very closely matches the growth and decline pattern in our histogram. Remarkably, the difference between these two lines is negligent; the increase in bandwidth would decrease the number of bumps in our graph, but it would modify the curve to fit our histogram worse. If we decrease the bandwidth, it would increase our number of bumps, but visually it wouldn't be as much of a curve.

Looking into our MME and MLE density curves, we note not only the notion that they are the exact same, but also that there are minimal differences between them and our histogram. There is a slight overestimation by our density curve at $0.05 \leq X \leq 0.1$ and $0.25 \leq X \leq 0.3$ of 0.5. The rapid ascent of the curve following $X = 0.1$ and rapid descent following $X = 0.25$ lead to these slight over estimations. In the context of the data set, the city has very little days of extremely low or high

wind speed; most days the wind speed is in between 6.7 and 16.75 km/h, which is $X = 0.1$ and X 0.25, respectively. Acknowledging that our MME and MLE are vastly similar to our histogram and density curves, we no longer need to ponder if they accurately calculate the population density. Consequently, this data set is accurately represented best by the beta distribution family.

5 Gamma Distribution

5.1 The Frequency of Different Ages Being Held In Jail

Denote X_i as the frequency of observed people with the specific age i of incarcerated people in the sample, $15 \leq i < 100$. Examining our histogram, the highest frequency of people fall into the age category of 20-24, with almost 1400 individuals, with the second highest falling in the age category of 25-29, which is just lower at around 1250 individuals. For each further age group (of 5 years), the number of individuals is around 30 percent lower than the previous age group (eg. from 850 to 600, around a 30 percent decrease), until we reach the age of 80. After 80, there is low frequency among all age categories, which means that the percentage of those 80 and above among those in jail or prison is significantly low. On the other side of the graph, the age category of 15-19 consists of around 200 individuals.

All together, this density distribution and histogram of `compas$age` falls into a gamma distribution due to its middling decay following its peak at age range 20-24, along with the notion that the prior age range of 15-19 also contains a noteworthy amount of individuals. Considering the question we routinely ask, age is a statistic that is found to maintain a gamma distribution among a plethora of data sets. In jail, it is reasonable that a majority of individuals fall into the 20-29 range, and then into the 30-39 age because these are the ages in which people are most prone to legal missteps resulting in imprisonment, along with the notion that this is when people can be the most physically active and potentially be involved in gang activity, legal mishaps, and other factors that are due to them being more "out and about." We also see this age range being the most frequent at banks in `bank$age` and being active in playing sports according to **npr.org** because they are the most active age group and because they have reduced mortality rates compared to an older age group. In both the `bank$age` study and **NPR's Sports and Health Poll**, the activity quota, following the age of 30, follows a very similar path to the `compas$age` data. Moving backward, while the percentage of individuals between the ages of 20-29 is the highest in `compas$age`, the other two data sets previously mentioned experience peaks at the same time. Also, ages that are lower than 20 do not need to go to the bank at the same rate older folk do, and this frequency is remarkably close to the `compas$age` data. However, in the **Sports and Health Poll**, we do see individuals younger than 20 with a significantly higher frequency of sports activity compared to the age range of 20-29, but this is due to the wide access and importance put on exercise by our education system.

5.2 Performing the Estimations

See `gamma.r` for MLE code.
MME: L is an estimator for λ and C is an estimator for r.

$$L = \frac{\sum_{i=1}^{n} x_i}{\frac{1}{n} * \sum_{i=1}^{n} (x_i - \sum_{i=1}^{n} x_i)^2}$$

$$C = L * \sum_{i=1}^{n} x_i$$

5.3 Visuals

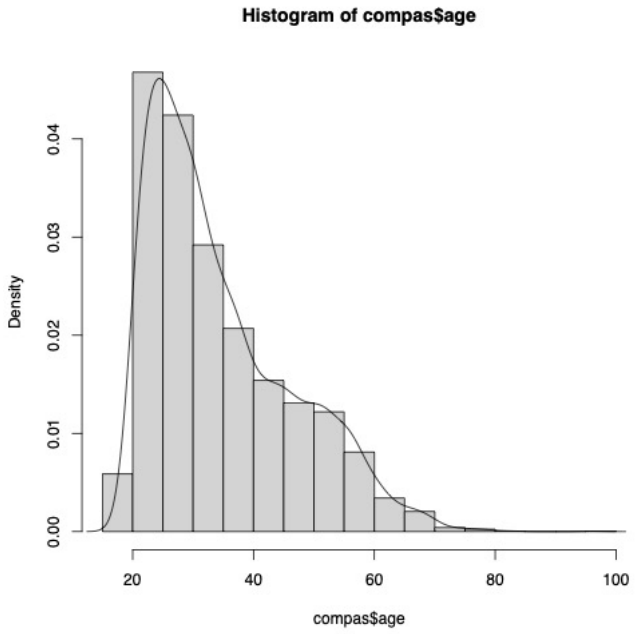

Figure 13: Sample Density Superimposed on Sample Histogram

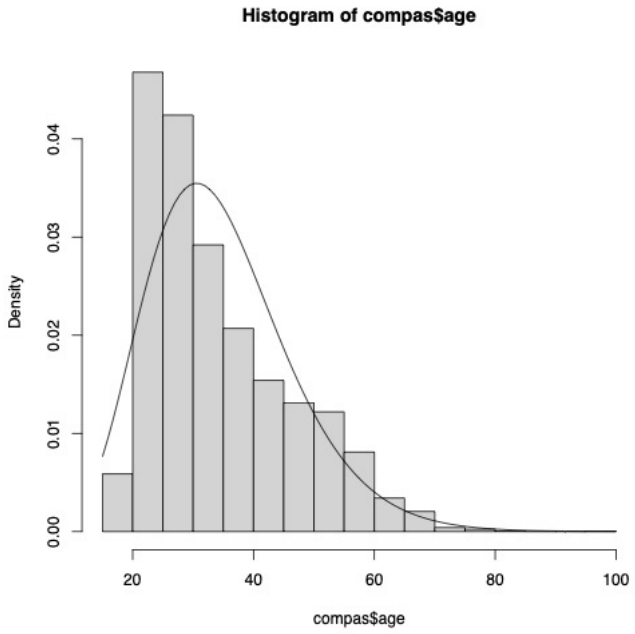

Figure 14: MME Density Superimposed on Sample Histogram

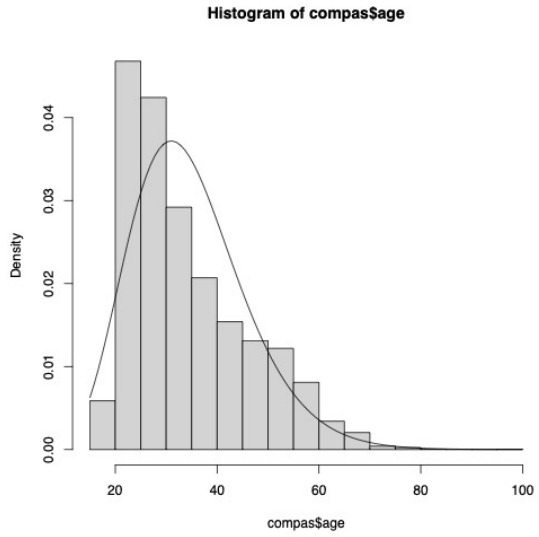

Figure 15: MLE Density Superimposed on Sample Histogram

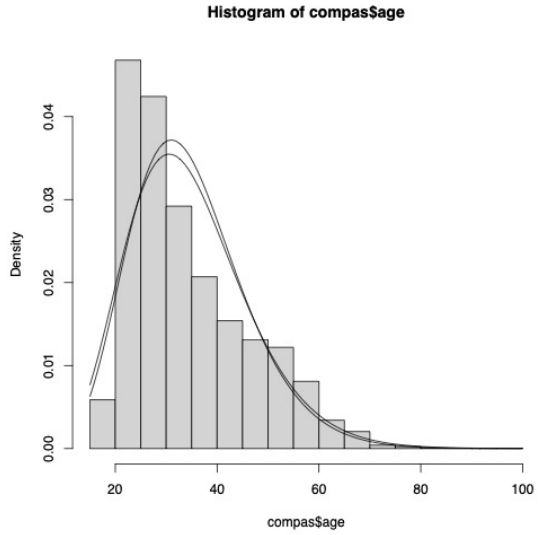

Figure 16: MLE Density and MME Density Comparison Superimposed on Sample Histogram

5.4 Analysis

Continuing our discussion of analyzing similarities and differences, we notice that our density curve accurately represents our data despite the slight overestimate of 0.004 for the age group of 30-35. This is due to a 30% decrease occurs from the prior age range; if we were to decrease the bandwidth then the density curve would follow the histogram more accurately, but once again, we want to limit the bumps that occur on the curve. It is difficult for a curve to represent the histogram with 100% accuracy when there are rapid changes in the histogram. That said, our density curve still successfully demonstrates the histogram information, finding the right balance of accuracy and bump reduction.

While both the MME and MLE successfully measure the density of our data set, the slight higher peak of the MLE allows it to be a better estimate. Both estimators follow the overall shape of the

histogram while also following the density plot of the histogram. Asides from the slight difference at the peak of both estimators, the two curves are nearly identical. However, both estimators slightly overestimate the sample data before the peak and following the peak, as the curves are not able to keep up with the rapid ascent of individuals at the age group of 20-24. It overestimates the data at $X < 20$, while also overestimating the data from $30 \leq X \leq 44$. Both curves also underestimate the peak of the data set; they underestimate the age range of $20 \leq X \leq 24$ and $25 \leq X \leq 29$ by 0.02 and .005, respectively. While both estimators do not necessarily reach the peak of the histogram or density curve, the MLE is closer to the the peak, and thus the better suitor for our gamma distribution. Nonetheless, both our MLE and MME measurements for gamma accurately calculate the population density, which results in the gamma family being the suitor for this data set.

6 Contributions

Ryan:
Lead Programmer, Overview Section, Graph Implementation, Reviser

Parsa:
Exponential, Beta, and Gamma Distribution Section, Programmer, Reviser

Hannah:
Normal Distribution Section, Overview Section, Latex File Research, Calculations, Reviser