ECS 132 Term Project

Steven Alvarado, Russell Chien, and Ruth Hailu

University of California, Davis

June 2023

Chapter 1

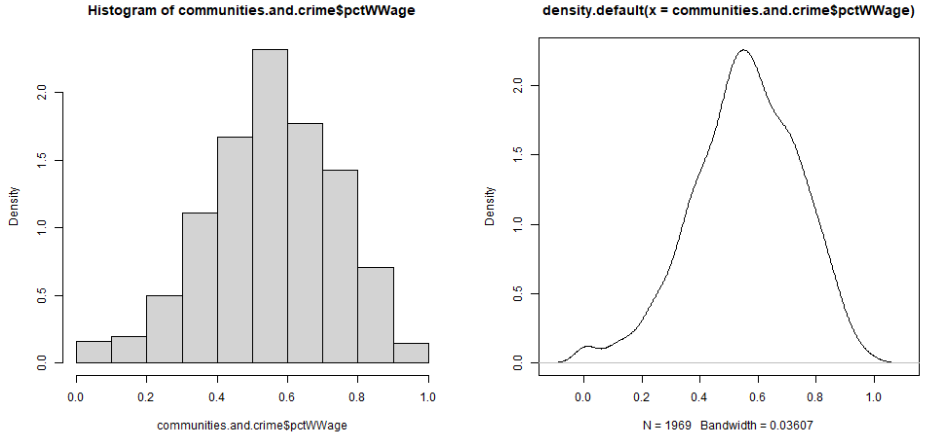
The Normal Family

1.1 Communities and Crime: pctWWage

Our group observed that the variable **pctWWage** of the Communities and Crime dataset seemed well-approximated by the normal family of continuous distributions. According to the UCI Machine Learning Repository, **pctWWage** is described as the percentage of households within the United States with wage or salary income in 1989.

1.2 Histogram and Density

Below are the histogram and density plots of pctWWage, generated using R's hist(), plot(), and density() functions.



1.3 Maximum Likelihood Estimation

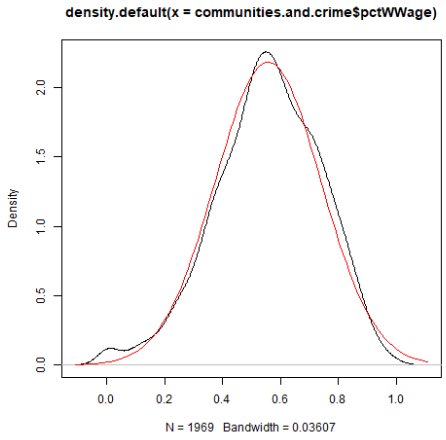
To find the MLE of pctWWage, we first defined our log-likelihood function as:

$$LL(\mu, \sigma^2) = -n \log(2\pi) + \frac{\log(\sigma^2)}{2} - \frac{\sum(x-\mu)^2}{2\sigma^2} \tag{1.1}$$

Using R's built-in mle() function, we utilized the *negative* log-likelihood function to find the normal parameters of pctWWage.

```
z <- mle(minuslogl = ll, start = c(list(mean = 1),
  list(var = 1)))
```

Superimposing the resulting density on pctWWage's kernel plot results in the graph:

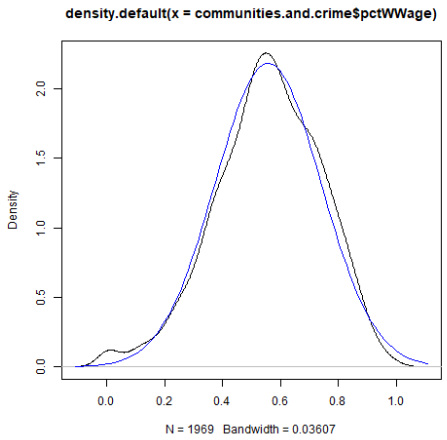


1.4 Method of Moments

To find the MM-estimated density of pctWWage, we used the following function to predict the values of μ and σ^2:

```
mm <- function(x) {
  mu <- mean(x)
  sigma <- sqrt(mean(x^2) - mu^2)
  return(c(mu, sigma))
}
```

Superimposing the resulting density results in the graph:



1.5 Analysis

The MLE and MM normal approximations appear to be a good fit for pctWWage's data. The density estimates obtained from both methods closely align with the density curve derived directly from the dataset. Thus, our group can confidently conclude that the variable pctWWage is well-approximated by the normal distribution family.

Chapter 2

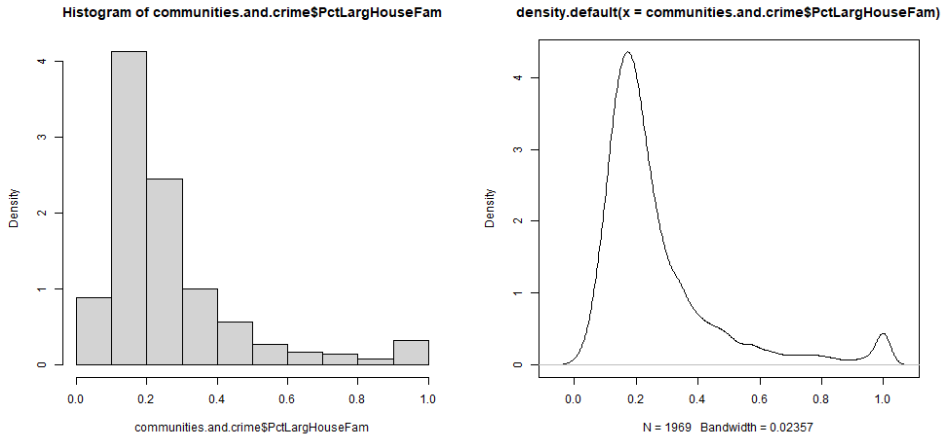
The Exponential Family

2.1 Communities and Crime: PctLargHouseFam

For the exponential family of continuous distributions, we observed that the variable **PctLargHouseFam** could potentially be a suitable approximation. The UCI ML Repo describes **PctLargHouseFam** as the percentage of family households with six or more family members.

2.2 Histogram and Density

Below are the histogram and density plots of PctLargHouseFam, using similar methods of generation as the previous section.



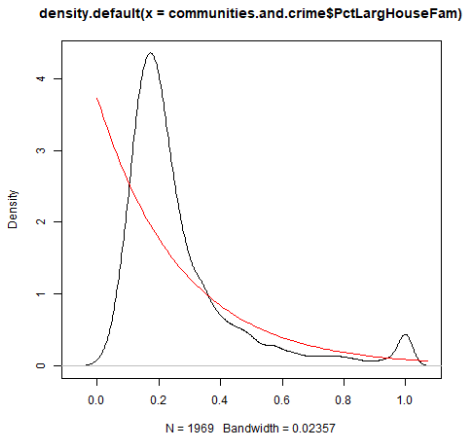
2.3 Maximum Likelihood Estimation

To find the MLE of PctLargHouseFam, we defined our log-likelihood function as:

$$LL(\lambda) = n \log \lambda - \lambda \sum x \tag{2.1}$$

Using R's built-in mle() function, we utilize the *negative* log-likelihood function to find the exponential parameters of PctLargHouseFam:

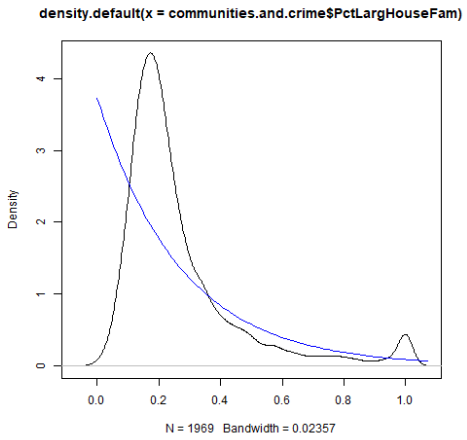
```
1  z <- mle(minuslogl = ll, start = c(list(lambda =
       1)))
```



2.4 Method of Moments

To find the MM-estimated density of PctLargHouseFam, we used the following function to predict λ:

```
  mm <- function(x) {
    lambda <- 1 / mean(x)
    lambda
}
```



2.5 Analysis

The exponential approximation shows a reasonable fit at the tail end of the density distribution, but it exhibits noticeable deviations from the actual data around $x = 0.2$. The exponential distribution assumes a constant and consistent decay rate, which does not accurately capture the characteristics of the data around its peak.

Chapter 3

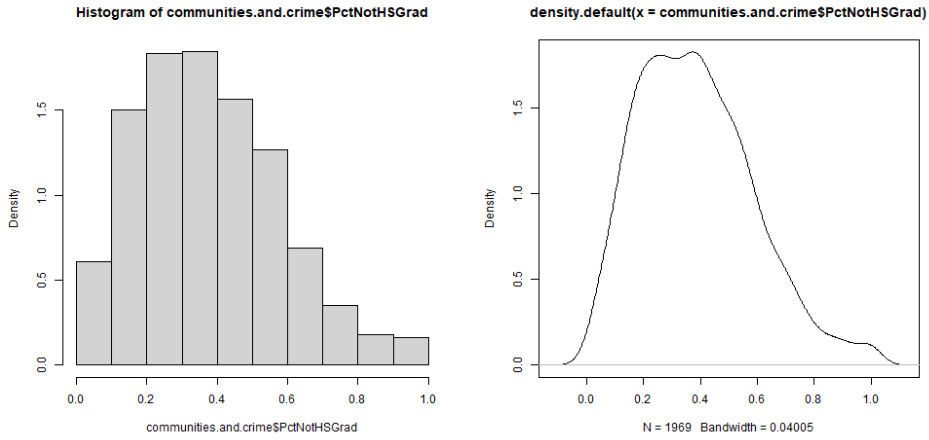
The Gamma Family

3.1 Communities and Crime: PctNotHsGrad

Our group observed that the gamma family of continuous distributions seemed to well-approximate the variable **PctNotHsGrad**. On the UCI ML Repo, **PctNotHsGrad** is defined as the percentage of people 25 and over that are not high school graduates.

3.2 Histogram and Density

Below are the histogram and density plots of PctNotHsGrad.



3.3 Maximum Likelihood Estimation

To find the MLE of PctNotHsGrad, we first defined our log-likelihood function as:

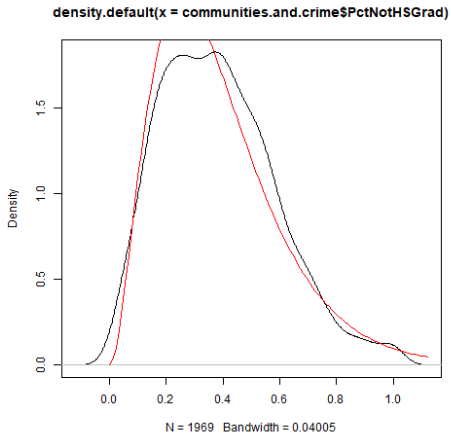
$$LL(k, \theta) = (k-1)\sum(\log x) - \sum(\frac{x}{\theta}) - nk\log(\theta) - n\log(\Gamma(k)) \quad (3.1)$$

We then had to scale our data to ensure the log-likelihood function remained finite:

```
1  x[which(x == 0)] <- 0.0001
```

Using R's built-in mle() function, we utilized the *negative* log-likelihood function to find the gamma parameters of PctNotHsGrad.

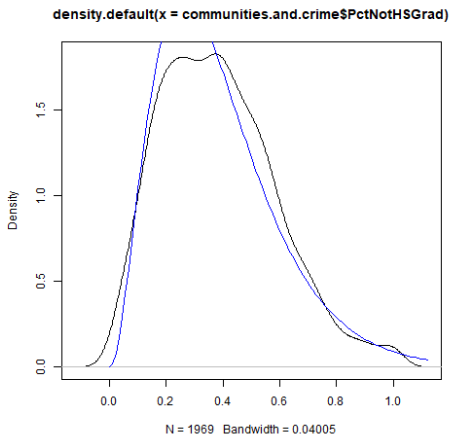
```
1  z <- mle(minuslogl = ll, start = c(list(k = 1),
     list(theta = 1)))
```



3.4 Method of Moments

To find the MM-estimated density of PctNotHsGrad, we used the following function to predict the values of k and θ:

```
  mm <- function(x) {
    mu <- mean(x)
    theta <- mean(x * log(x)) - mu * mean(log(x))
    k <- mu / theta
    return(c(k, theta))
}
```



3.5 Analysis

Both the MLE and MM gamma approximations exhibit a good overall fit to the PctNotHsGrad data, accurately capturing its general shape. However, both approximations tend to overestimate the height of the actual data peak. Despite this slight overestimation, the gamma approximations successfully capture the overall pattern of the data and seem to be suitable for approximating the population density.

Chapter 4

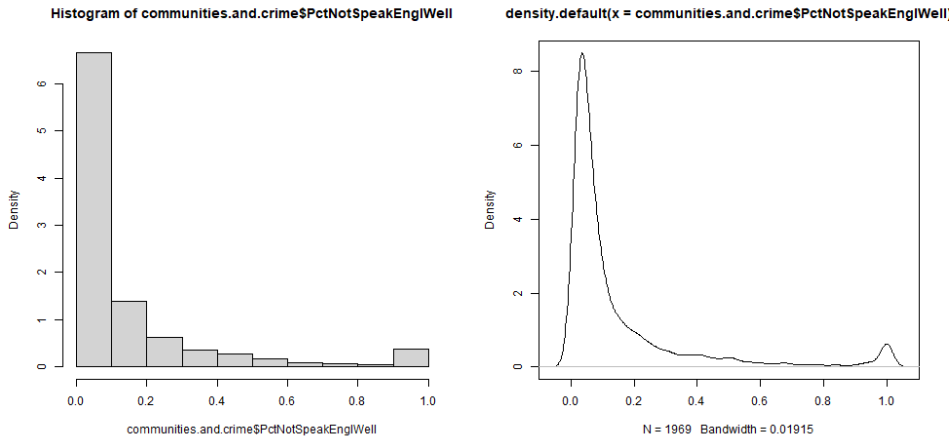
The Beta Family

4.1 Communities and Crime: PctNotSpeakEnglWell

Finally, our group observed that **PctNotSpeakEnglWell** may be a suitable approximation for the beta family of continuous distributions. According to the UCI ML Repo, **PctNotSpeakEnglWell** is described as the percentage of people who do not speak English well.

4.2 Histogram and Density

Below are the histogram and density plots of PctNotSpeakEnglWell.



4.3 Maximum Likelihood Estimation

To find the MLE of PctNotSpeakEnglWell, we first had to scale our data so it was within the beta family's support of $(0, 1)$:

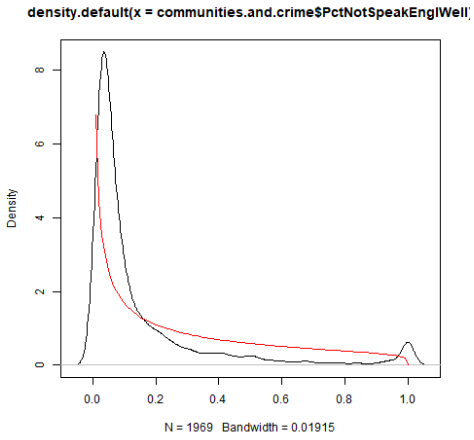
```
x[which(x == 0)] <- 0.0001
x[which(x == 1)] <- 0.9999
```

We then defined our log-likelihood function as:

$$LL(\alpha, \beta) = n \log\left(\Gamma(\alpha + \beta)\right) - n \log\left(\Gamma(\alpha)\right) - n \log\left(\Gamma(\beta)\right) + (\alpha - 1) \sum(\log(x)) + (\beta - 1) \sum(\log\left(1 - x\right)) \quad (4.1)$$

Using R's built-in mle() function, we utilized the *negative* log-likelihood function to find the beta parameters of PctNotSpeakEnglWell.

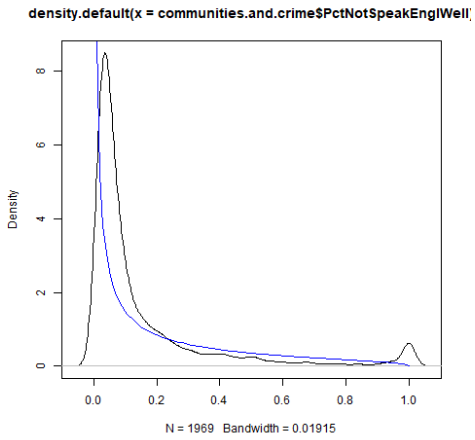
```
z <- mle(minuslogl = ll, start = c(list(alpha = 1),
  list(beta = 1)))
```



4.4 Method of Moments

To find the MM-estimated density of PctNotSpeakEnglWell, we used the following function to predict the α and β values:

```
mm <- function(x) {
  mu <- mean(x)
  var <- var(x)
  alpha <- mu * (mu * (1 - mu) / var - 1)
  beta <- (1 - mu) * (mu * (1 - mu) / var - 1)
  return(c(alpha, beta))
}
```



4.5 Analysis

The MLE and MM beta approximations of PctNotSpeakEnglWell exhibit notable differences from each other. The MLE approximation decays slower than the actual data, while the MM approximation matches the tail of the data well, suggesting that it captures the decay behavior of the population density more accurately.

Chapter 5

Contributions

- **Steven Alvarado**: Report formatting, MLE of beta and MM of gamma and beta, finding potential distribution variables from datasets
- **Russell Chien**: Finding suitable variables from datasets, MLE and MM of exponential, transferring code and plots into latex report
- **Ruth Hailu**: Data analysis