Final Project

ECS 132 — Winter 2015

William Otwell (#########)
Rupali Saiya (#########)
Nicholas Layton(996933702)
Syeda Inamdar(#########)

March 14, 2015

Table of Contents

1 Problem 1 **3**
1.1 Part A: Comparison of Two Means . . . 3
1.2 Part B: Comparison of Two Proportions . . . 4

2 Problem 2 **5**
2.1 Part A . . . 5
2.2 Part B . . . 5
2.3 Part C . . . 5
2.4 Part D . . . 6
2.5 Part E . . . 6
2.6 Part F . . . 6
2.7 Part G . . . 6
2.8 Part H . . . 6

3 Problem 3 **7**
3.1 Part A: Nonparametric Density Estimation . . . 7
3.2 Part B: Fitting a Parametric Model . . . 7
3.2.1 Method of Moments . . . 7
3.2.2 Method of Maximum Likelihood . . . 7
3.3 Part C: Plotting Parametric-Model Curves . . . 8

A Problem 1 Code **9**

B Problem 2 Code **11**

C Problem 3 Code **13**

D Problem 3 Plots **15**
D.1 Part A . . . 15
D.2 Part C: Method of Moments . . . 16
D.3 Part C: Maximum Likelihood . . . 17
D.4 Part C: Combined . . . 18

E Group Member Contributions **19**

1 Problem 1

1.1 Part A: Comparison of Two Means

The bike sharing dataset provided by the UC Irvine (UCI) Machine Learning Repository allows us to analyze bike rental statistics during the years 2011 and 2012. We decided to analyze and compare the quantity of bike rentals on days during the months of March 2011 and March 2012 to gain insight as to how much the bike rental rates changed from one year to the next. In more formal terms, we estimated the difference between the average number of bikes rented in March 2011 and the average number of bikes rented in March 2012. To do so, we constructed a confidence interval from the sample provided by UCI to estimate the difference between the population means of the averages from each month. We first had to define our random variables before constructing our confidence interval. Below is a list of the random variables that we used:

- X: the number of bikes rented on any given day in March 2011
- Y: the number of bikes rented on any given day in March 2012
- $\bar{X}$: the sample mean of the number of bikes rented per day in March 2011
- $\bar{Y}$: the sample mean of the number of bikes rented per day in March 2012
- s_1: the standard deviation of the number of bikes rented per day in March 2011
- s_2: the standard deviation of the number of bikes rented per day in March 2012

An important aspect worth noting is that the random variables X and Y are independent of one another. After defining our random variables, we had to construct our model for constructing the confidence interval. Our model turned out to be the following:

$$\bar{X} - \bar{Y} \pm (1.96)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{1}$$

... where n_1 and n_2 were the number of days in March 2011 and March 2012 (respecfully) for which data was recorded. The R code that we developed to calculate this confidence interval can be found in Appendix A.

After running our R calculations, we found that the average number of bike rentals in March 2011 was 2065.968 bikes and that of March 2012 was 5318.548 bikes. The difference between the means of our two samples was -3252.581. The standard deviation of the number of bikes rented in March 2011 was 550.9717, and the standard deviation of that in March 2012 was 1251.1627. From these values, we found our confidence interval to be $(-5932.108, -573.0535)$. In formal terms, we are 95% confident that the difference between the mean number of bike rentals on days in March 2011 and days in March 2012 lies within the interval ranging from -5932.108 to -573.0535.

Although this is a very wide ranging interval, we still were able to make inferences about the bike rental trends in March 2011 versus the trends in March 2012. Being that we subtracted the sample mean of the bike rentals in March 2012 from that of the bike rentals in March 2011, we inferred that there were significantly more bike rentals in March 2012 than there were in March 2011. This is because the entirety of the interval lies below 0. In the next section, we used a proportion of warmer days in the same two months to see if temperature could have influenced the increase in the number of bike rentals from March 2011 to March 2012.

1.2 Part B: Comparison of Two Proportions

We wanted to compare the proportion of warmer than average days in March 2011 to those in March 2012. We constructed a confidence interval to estimate the difference between population proportions of the number of warmer days in March 2011 and March 2012. Again, we defined our random variables and then constructed the confidence interval.

- $\hat{p_1}$: the sample proportion of warmer days in March 2011
- $\hat{p_2}$: the sample proportion of warmer days in March 2012
- s_1^2: the standard deviation of the proportion of warmer days in March 2011
  Mathematical expression: $s_1^2 = \hat{p_1}(1 - \hat{p_1})$
- s_2^2: the standard deviation of the proportion of warmer days in March 2012
  Mathematical expression: $s_2^2 = \hat{p_2}(1 - \hat{p_2})$

The random variables $\hat{p_1}$ and $\hat{p_2}$ are independent of one another. The model that we used to calculate the confidence interval for the difference in the two proportions was:

$$\hat{p_1} - \hat{p_2} \pm (1.96)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{2}$$

$$= \hat{p_1} - \hat{p_2} \pm (1.96)\sqrt{\frac{\hat{p_1}(1 - \hat{p_1})}{n_1} + \frac{\hat{p_1}(1 - \hat{p_1})}{n_2}} \tag{3}$$

. . . where n_1 and n_2 are the number of days in March 2011 and March 2012 (respectfull) during which the temperature was warm. We defined a "warmer day" to be one in which the temperature rose above 12.5 degrees Celsius, which translated to be 0.3 in the UCI dataset. The R code that was developed to evaluate the above expression can be found in Appendix A.

In our calculation we found that the proportion of warmer days in March 2011 was 0.5806452 and that the proportion of warmer days in March 2012 was 0.9032258. The difference between the proportions of our two samples was found to be 0.3225833. The standard deviation of the proportion of warmer days in March 2011 was 0.4934535, and the standard deviation of the proportion of warmer days in March 2012 was 0.2956500. The resulting confidence interval was $(-0.5250816, -0.1200797)$. We can say that we are 95% confident that the difference between the population proportions for warmer days in the months of March 2011 and March 2012 lies within the interval ranging from -0.5250816 to -0.1200797.

These two confidence intervals show that there could be a relationship between the proportion of warmer days and the number of bike rentals in the month of March. We can see that both the proportion of warmer days and the number of bike rentals in March 2012 were greater than that of March 2011, which suggests that warmer weather could have influenced more people to rent bikes. Although it makes intuitive sense that there is a relationship between warmer weather and an increase in bike rentals, our calculations cannot ensure that warmer weather was, in fact, the reason why there were more people renting bikes. We can only say that there is a potential relationship, given that the confidence intervals indicate as such.

2 Problem 2

2.1 Part A

2.2 Part B

2.3 Part C

When we fit a linear model on our training data set, the output returned is our sample estimate (estimate based on our sample data). Below is the output R returned to us.

```
Call:
lm(formula = training$cnt ~ training$season + training$yr + training$mnth +
    training$holiday + training$weekday + training$workingday +
    training$weathersit + training$temp + training$atemp + training$hum +
    training$windspeed)

Residuals:
     Min       1Q   Median       3Q      Max
-2893.64  -426.50    42.07   452.25  2912.31

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1344.734    248.963   5.401 1.05e-07 ***
training$season      380.727     62.636   6.078 2.49e-09 ***
training$yr         2018.880     90.428  22.326  < 2e-16 ***
training$mnth          3.964     19.918   0.199  0.84232
training$holiday    -380.192    210.128  -1.809  0.07103 .
training$weekday      43.618     16.938   2.575  0.01032 *
training$workingday   66.072     74.855   0.883  0.37786
training$weathersit -529.315     79.198  -6.683 6.54e-11 ***
training$temp        487.505   2327.181   0.209  0.83416
training$atemp      5276.362   2610.425   2.021  0.04381 *
training$hum        -926.507    305.937  -3.028  0.00259 **
training$windspeed -2265.539    476.813  -4.751 2.68e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 739.9 on 476 degrees of freedom
Multiple R-squared:  0.7631,    Adjusted R-squared:  0.7576
F-statistic: 139.4 on 11 and 476 DF,  p-value: < 2.2e-16
```

The first coefficient returned is the intercept value, which we can call β_0. The estimates for our next 11 coefficients returned are our predictors, which we can call β_i, where i is a number 1 through 11. These values can help us estimate average bike ridership from our sample data. The linear model of our estimated regression would be:

$$\text{mean count} = \beta_0 + \beta_1\text{season} + \beta_2\text{yr} + \beta_3\text{mnth} + \beta_4\text{holiday} + \beta_5\text{weekday} + \beta_6\text{workingday} + \ldots \quad (4)$$

$$\ldots \beta_7\text{weathersit} + \beta_8\text{temp} + \beta_9\text{atemp} + \beta_{10}\text{hum} + \beta_{11}\text{windspeed} \quad (5)$$

The above just gives a sample estimate, it is not representative of the population bike ridership function. Looking at the estimate values for the coefficients one can see that we have positive and

negative values. The coefficients that have negative values indicate attributes that result in a decrease of bike ridership. For example, the estimate for holidays is -380.192. This makes sense because bike ridership would decrease if it is a holiday as most people would be out of town or with their families in our sample. The coefficients that have positive values indicate attributes that result in an increase of bike ridership. For example, the estimate for temp is 487.505, which indicates that as the temperature increases (gets warmer) there is more bike ridership on average in our sample.

To get an idea of how accurate our estimate is, we can use the standard error of each of β_i to calculate a 95% confidence interval. This will tell us that we are 95% confident that the true predictor for that particular coefficient is in that interval. The R-squared value tells us that we can predict bike ridership from these eleven coefficients about 76.31% of the time.

2.4 Part D

2.5 Part E

2.6 Part F

2.7 Part G

2.8 Part H

3 Problem 3

3.1 Part A: Nonparametric Density Estimation

3.2 Part B: Fitting a Parametric Model

In the previous section, we formed a nonparametric estimate of the density of the daily temperature data by constructing a histogram. After analyzing the histogram, we chose to model its density with a normal parametric distribution. The histogram showed two peaks: one peak among the values that represented colder weather and another among those of warmer temperatures. It also showed a dip among values that are somewhere between colder and warmer temperatures. These observations suggested that the daily temperature from 2011 to 2012 in Washington D.C. matched that of a bimodal distribution, however, we theorized that the population temperature, in fact, could have been normally distributed.

We thought that the frequency of days in which temperature was either extremely cold or extremely warm should be less than that of days in which the temperature was somewhere in between. Our theory came from the idea that climates seem to rarely shift from hot to cold (or vice-versa) in short periods of time, meaning there should be an extended period of time between these periods in which the weather is of moderate temperature. For this reason, we thought there should be a larger number of days of moderate temperature, and that the data should have represented a more normal distribution.

In the following sections we explain how we fitted our data to a normal parametric model using two different approaches: the Method of Moments approach, and the Maximum Likelihood approach.

3.2.1 Method of Moments

The method of moments involves equation population parameters for a particular parametric distribution to its population moments. In the case of the normal distribution these parameters are the mean and the variance. To accurately fit the normal parametric model to our data, we had to determine the method of moments estimators for the mean and variance. We found the first and second theoretical moments to be:

$$E(X_i) = \mu \quad \& \quad E(X_i^2) = \sigma^2 + \mu^2 \tag{6}$$

The second equation came from manipulating the formula for variance. We cam up with two equations to estimate the moments based on our sample data:

$$E(X_i) = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7}$$

$$E(X_i^2) = \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \tag{8}$$

We were able to calculate these values from our sample data of daily temperatures using R. The code that we developed to calculate these moments can be found in Appendix C in the function labeled "PartB()". From our code we determined the method of moments estimator for $E(X)$ to be .4953848 and the estimator for $E(X_i^2)$ to be .2788679 and therefore calculated the variance, σ^2, to be 0.030473 and the standard deviation, σ, to be 0.174565. We then used these estimated parameters in our function labeled "PartC()" to get an accurate fit to the data.

3.2.2 Method of Maximum Likelihood

The method of maximum likelihood is a more commonly used method to find the likelihood (the probability) of our data values occurring. In the case of our normal distribution, the parameters are

mean and standard deviation. To accurately fit the normal parametric model to our data, we had to determine the method of maximum likelihood estimators for the mean and standard deviation. The first step to find the estimators was to determine the likelihood function in terms of the product of the density values. We determined this to be:

$$L = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/(2\sigma^2)} \tag{9}$$

We can simplify this equation into:

$$L = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2} \tag{10}$$

However, it is usually easier to maximize the log likelihood of a normal distribution. We determined this to be:

$$l(\mu, \sigma) = -\frac{n}{2} log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu^2) \tag{11}$$

From here we were able to calculate the values from our sample data of daily temperature using R's mle() function. The code that we developed to calculate these moments can be found in Appendix C in the function labeled "PartB()". From our code we determined the method of maximum likelihood estimator for the mean, μ, to be .4953848 and the estimator for the standard deviation, σ, to be .1829285.

3.3 Part C: Plotting Parametric-Model Curves

We plotted our two parametric model curves on top of the nonparametric density estimation (histogram), and found them to both be very similar. The calculated values of μ and σ were very similar between the Method of Moments and the Method of Maximum Likelihood. In Appendix D you can find our plots for Method of Moments and Method of Maximum Likelihood superimposed on our plot from "Part A()". The parametric curve for Method of Moments in Appendix D.2 is drawn in red and the curve for Method of Maximum Likelihood in Appendix D.3 is drawn in blue. The graph of Appendix D.4 shows a curve in purple, which includes the plot of Method of Moments and Method of Maximum Likelihood and shows us that both these methods calculated the same mean and standard deviations for the normal distribution.

A Problem 1 Code

```
PartA <- function()
{
  # import daily data
  day <- read.csv(file="day.csv",header=TRUE,sep=",")
  month <- 3

  # obtain vectors of the cnt values for which
  # the month is 3 in the years 2011 and 2012
  monthYear0 <- day$cnt[(day$yr == 0) & (day$mnth == month)]
  monthYear1 <- day$cnt[(day$yr == 1) & (day$mnth == month)]

  # calculate the sample mean cnt for
  # March 2011 and March 2012 from the above vectors
  X_bar <- mean(monthYear0)
  Y_bar <- mean(monthYear1)

  # calculate the difference between the sample mean cnt values
  Difference <- X_bar - Y_bar

  # calculate the sample standard deviation
  # for cnt in March 2011 and March 2012
  s1 <- sd(monthYear0)
  s2 <- sd(monthYear1)

  # calculate the standard error
  SE <- sqrt(((s1 ^2) / length(s1)) + ((s2 ^ 2) / length(s2)))

  # construct the confidence interval
  CI_2means <- c(Difference - 1.96 * SE, Difference + 1.96 * SE)

  # display results
  cat("Confidence Interval for Difference between 2 population
     means\n")
  cat("of bike rentals in March from 2011 to 2012:\n")
  cat(CI_2means)
} # PartA

PartB <- function()
{
  # import daily data
  day <- read.csv(file="day.csv",header=TRUE,sep=",")
  month <- 3

  # obtain vectors of values for which the daily temperatures in the
  # months of March 2011 and March 2012 were above 0.3
```

```
45    tempMonthYear0 <- day$cnt[(day$yr == 0) & (day$mnth == month) &
         (day$temp >.3)]
46    tempMonthYear1 <- day$cnt[(day$yr == 1) & (day$mnth == month) &
         (day$temp > .3)]
47
48    # obtain vectors of the cnt values for which
49    # the month is 3 in the years 2011 and 2012
50    monthYear0 <- day$cnt[(day$yr == 0) & (day$mnth == month)]
51    monthYear1 <- day$cnt[(day$yr == 1) & (day$mnth == month)]
52
53    # obtain total number of recordings taken
54    # in the month of March 2011 and March 2012
55    n0 <- length(monthYear0)
56    n1 <- length(monthYear1)
57
58    # calculate the proportion of days in the months of
59    # March 2011 and March 2012 for which
60    # the temperature was reater than 0.3
61    p_hat0 <- length(tempMonthYear0) / n0
62    p_hat1 <- length(tempMonthYear1) / n1
63
64    # calculate the difference between the proportions
65    Difference2 <- p_hat0 - p_hat1
66
67    #calculate variances
68    var0 <- p_hat0 * (1 - p_hat0)
69    var1 <- p_hat1 * (1 - p_hat1)
70
71    # calculate the standard error
72    SE2b <- sqrt((var0 / n0) + (var1 / n1))
73
74    # construct the confidence interval
75    CI_2props <- c(Difference2 - 1.96 *SE2, Difference2 + 1.96 * SE2)
76
77    # display results
78    cat("Confidence Interval for difference between 2 population
         proportions\n")
79    cat("of days in which the temperature in March \n")
80    cat("was greater than .3 from 2011 to 2012:\n")
81    cat(CI_2props)
82  } # PartB
```

B Problem 2 Code

```
PartA <- function()
{
    day <- read.csv(file="day.csv",header=TRUE,sep=",")
    summary(lm(day$cnt ~ day$season + day$yr + day$mnth +
       day$holiday + day$weekday + day$workingday + day$weathersit
       + day$temp + day$atemp + day$hum + day$windspeed))
} # PartA()

PartB <- function()
{
  day <- read.csv(file="day.csv",header=TRUE,sep=",")

  # get 2/3 of original data set for training set
  n1 <- ceiling(2 * nrow(day) / 3)
  training <- day[sample(n1, replace=FALSE),]

  # get 1/3 of original data set for validation set
  n2 <- nrow(day) - n1
  validation <- day[sample(n2, replace=FALSE),]
  validation <-
     cbind(validation['season'],validation['yr'],validation['mnth'],validation[

  # model the training set
  trainingModel <- lm(training$cnt ~ training$season + training$yr
     + training$mnth + training$holiday + training$weekday +
     training$workingday + training$weathersit + training$temp +
     training$atemp + training$hum + training$windspeed)
  summary(trainingModel)
  predictions <- predict(trainingModel,validation)
  return (predictions)
} # PartB()

PartC <- function()
{
  day <- read.csv(file="day.csv",header=TRUE,sep=",")

  # get 2/3 of original data set for training set
  n1 <- ceiling(2 * nrow(day) / 3)
  training <- day[sample(n1, replace=FALSE),]
  trainingModel <- lm(training$cnt ~ training$season + training$yr
     + training$mnth + training$holiday + training$weekday +
     training$workingday + training$weathersit + training$temp +
     training$atemp + training$hum + training$windspeed)
  summary(trainingModel)
} # PartC()
```

```
37
38 PartD <- function()
39 {
40
41 } # PartD()
42
43 PartE <- function()
44 {
45
46 } # PartE()
47
48 PartF <- function()
49 {
50
51 } # PartF()
```

C Problem 3 Code

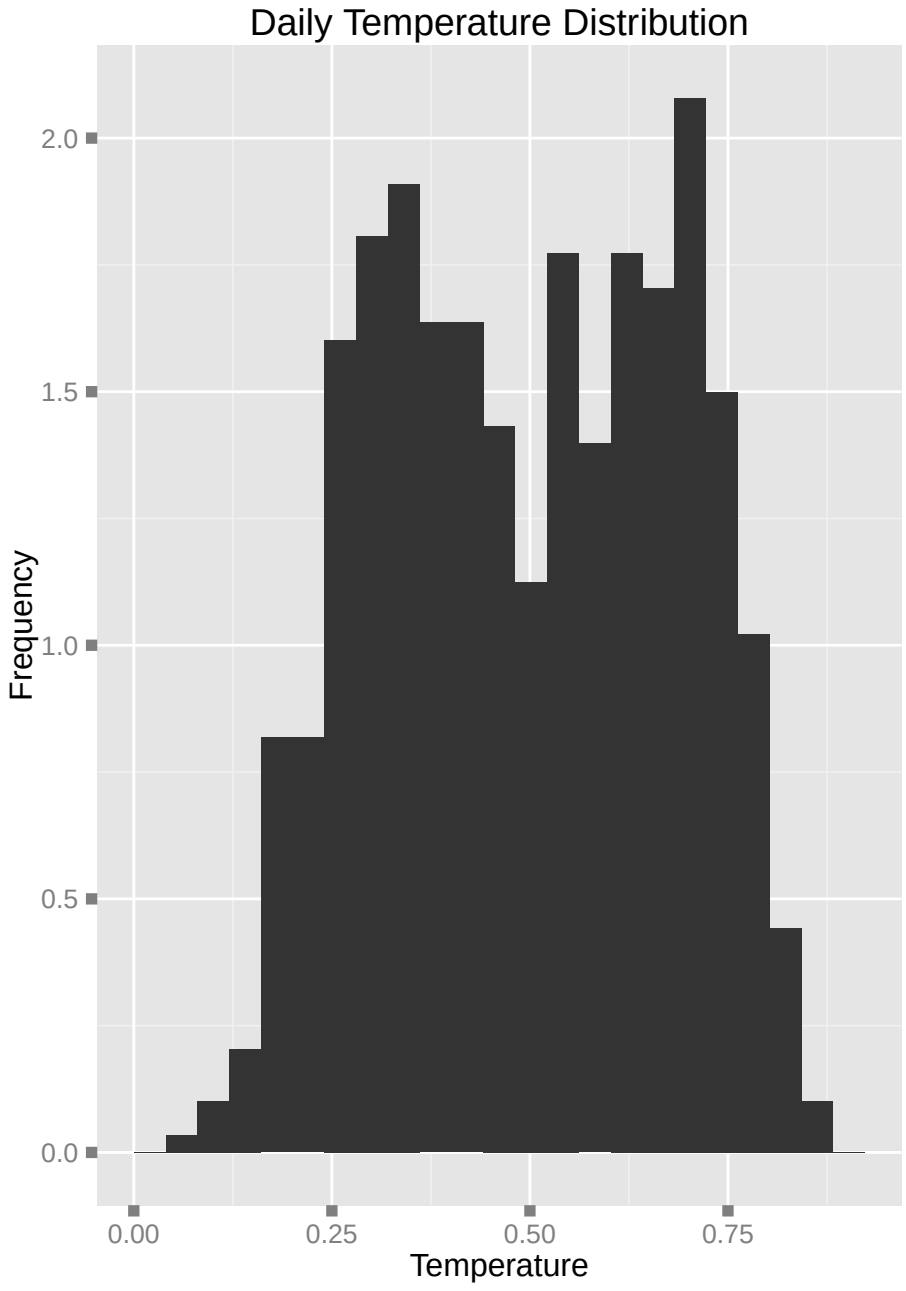
```
library(ggplot2)
library(stats4)

PartA <- function()
{
    day <- read.csv(file="day.csv",header=TRUE,sep=",")
    r <- range(day$temp)
    width <- (r[2] - r[1]) / 20
    p <- ggplot(day)
    p <- p + geom_histogram(aes(temp,y =..density..),binwidth=width)
    p <- p + xlab("Temperature") + ylab("Frequency") +
       ggtitle("Daily Temperature Distribution")
    p <- p + theme(title=element_text(size=12),
                   text=element_text(size=12),axis.ticks=element_line(size=2))
    print(p)
    ggsave("Problem3A.pdf",width=5,height=7,plot=last_plot())
    return (p)
} # PartA

PartB <- function()
{
  # read data file
  day <- read.csv(file="day.csv",header=TRUE,sep=",")

  ########## METHOD OF MOMENTS ##########
  # moment1 = E(X)
  moment1 <- mean(day$temp)

  # moment2 = E(X^2)
  moment2 <- mean((day$temp) ^ 2)

  # construct vector of mu and sigma
  moments <-c(moment1,sqrt(moment2 - (moment1 ^ 2)))

  ########## MAXIMUM LIKELIHOOD ##########

  # define function to be passed to mle
  x <- day$temp
  n <- length(day$temp)
  ll <- function(mu,sigma)
  {
    loglik <- ((-n / 2) * log(2 * pi)) - (n * log(sigma)) - ((1 /
       (2 * (sigma ^ 2))) * sum((x - mu) ^ 2))
    return (-loglik)
  }
```

```
44
45    # calc mle and get estimates for parameters
46    ests <- mle(minuslogl=ll,start=list(mu=0.5,sigma=0.5))@coef
47
48    return(cbind(moments,ests))
49  } # PartB
50
51  PartC <- function()
52  {
53    p <- PartA()
54
55    moments <- PartB()
56
57    ########## METHOD OF MOMENTS ##########
58    p1 <- p + stat_function(fun=dnorm,color="red",
59                            args=list(mean=moments[1,1],sd=moments[2,1]))
60    print(p1)
61    ggsave("Problem3CMoments.pdf",width=5,height=7,plot=last_plot())
62
63    ########## MAXIMUM LIKELIHOOD ##########
64    p2 <- p + stat_function(fun=dnorm,color="blue",
65                            args=list(mean=moments[1,2],sd=moments[2,2]))
66    print(p2)
67    ggsave("Problem3CMLE.pdf",width=5,height=7,plot=last_plot())
68
69    p3 <- p1 + stat_function(fun=dnorm,color="blue",
70                             args=list(mean=moments[1,2],sd=moments[2,2]))
71    print(p3)
72    ggsave("Problem3C.pdf",width=5,height=7,plot=last_plot())
73  } # PartC
```

D Problem 3 Plots

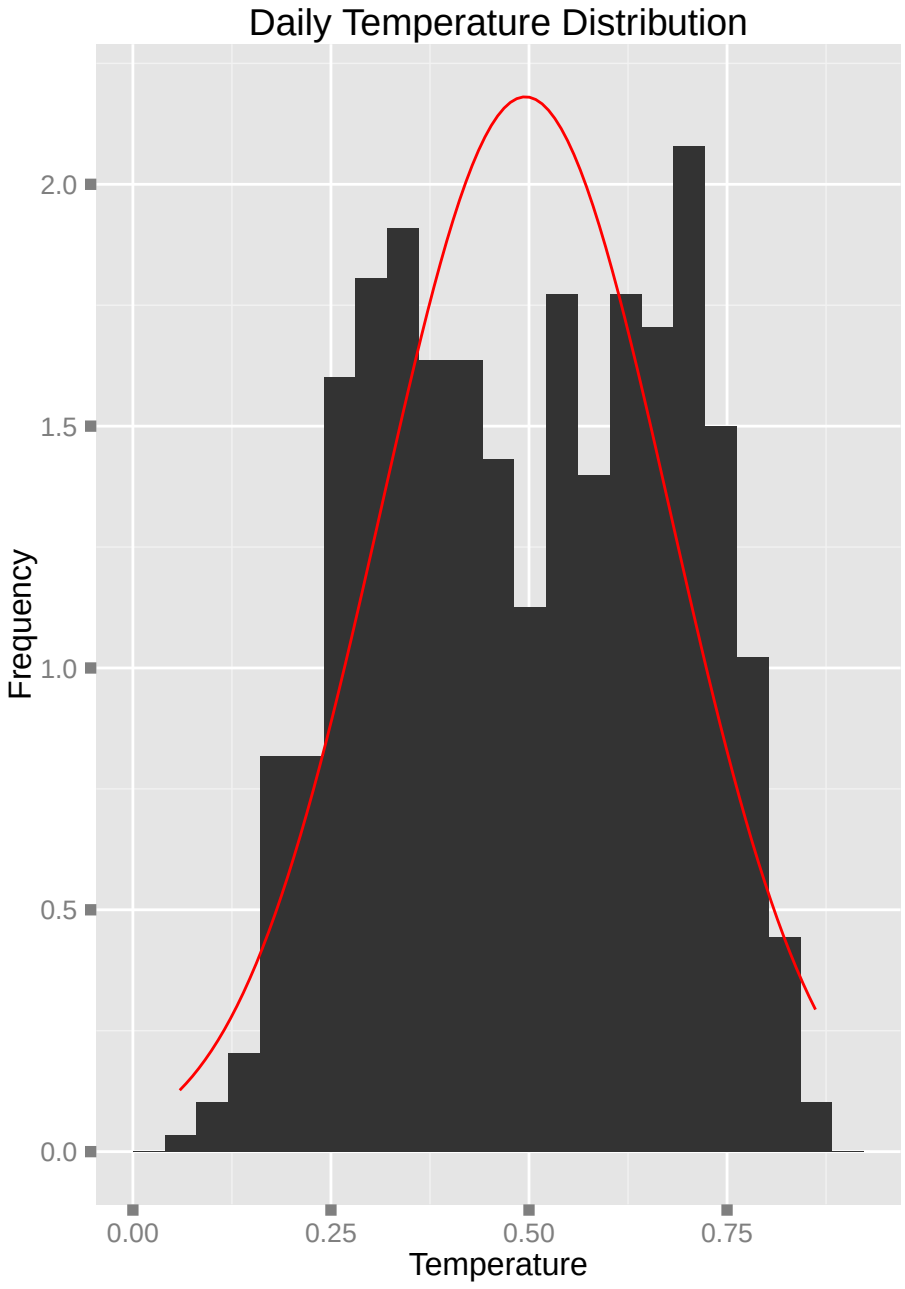
D.1 Part A

The plot below is the nonparametric estimate of our data set for daily temperatures.



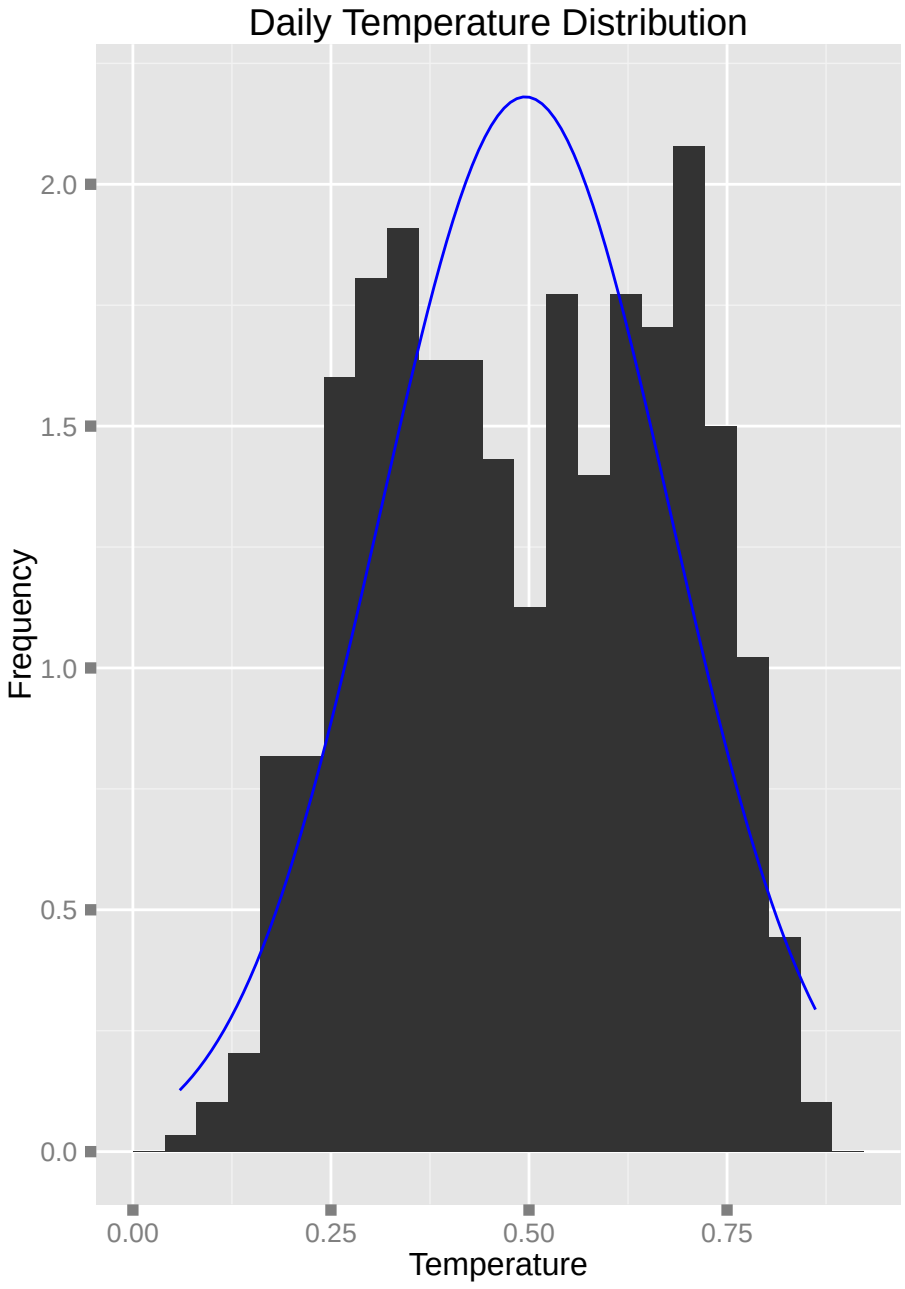
D.2 Part C: Method of Moments

The plot below is the parametric estimate of our data set for daily temperatures while using the Method of Moments approach to estimating the population parameters for a Normal Distribution.

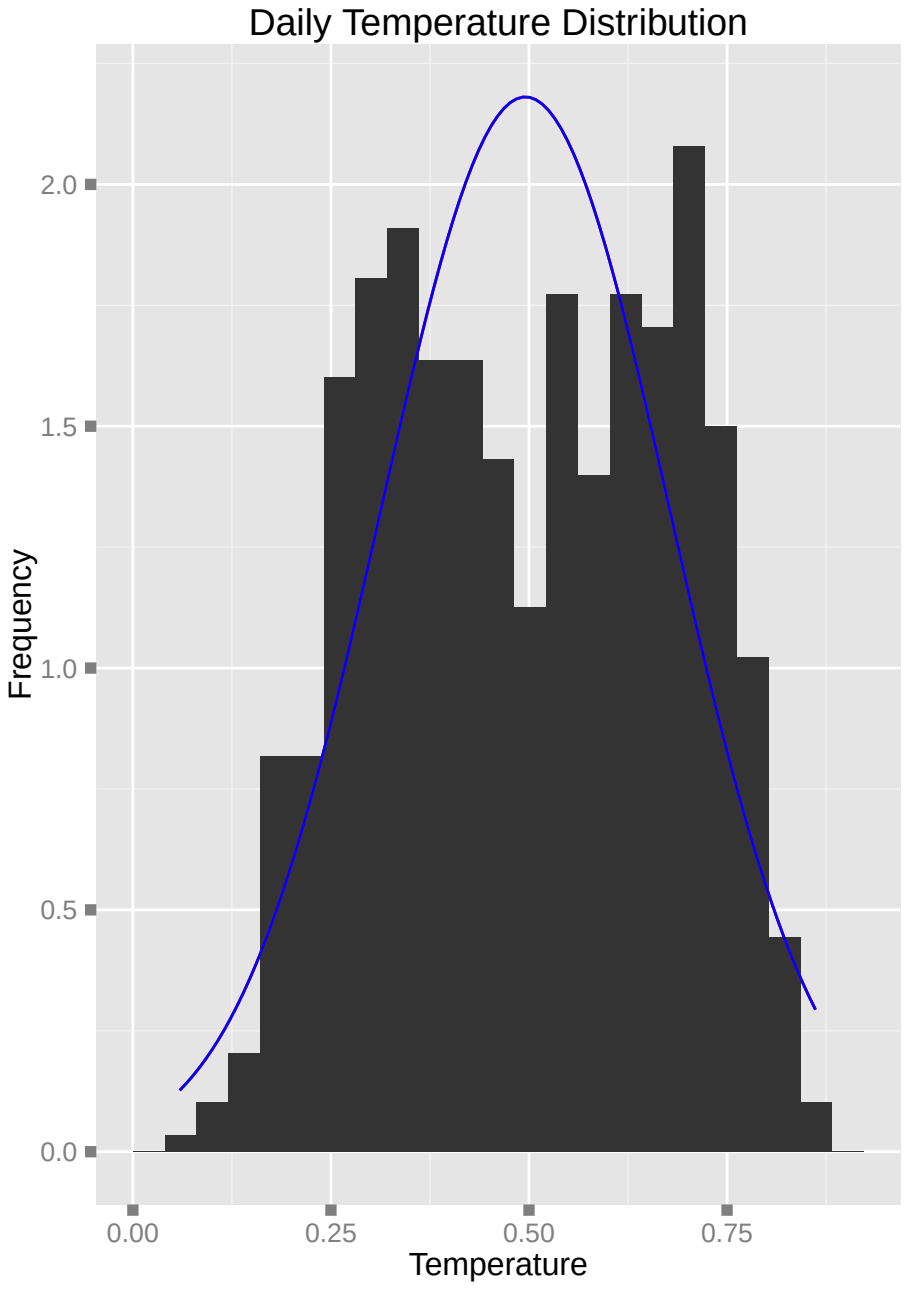


D.3 Part C: Maximum Likelihood

The plot below is the parametric estimate of our data set for daily temperatures while using the Method of Maximum Likelihood to estimate the population parameters for a Normal Distribution.



D.4 Part C: Combined



E Group Member Contributions

- William Otwell:
  - –
- Rupali Saiya:
  - –
- Nicholas Layton:
  - –
- Syeda Inamdar:
  - –