Final Project

ECS 132 — Winter 2015

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

March 10, 2015

Contents

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 ####, and the standard deviation of that in March 2012 was ####. 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 ####, and the standard deviation of the proportion of warmer days in March 2012 was ####. 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

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: Method of Moments and the Maximum Likelihood Fittings

3.3 Part C: Plotting Parametric-Model Curves

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
  tempMonthYear0 <- day$cnt[(day$yr == 0) & (day$mnth == month) &
     (day$temp >.3)]
  tempMonthYear1 <- day$cnt[(day$yr == 1) & (day$mnth == month) &
     (day$temp > .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)]

  # obtain total number of recordings taken
  # in the month of March 2011 and March 2012
  n0 <- length(monthYear0)
  n1 <- length(monthYear1)

  # calculate the proportion of days in the months of
  # March 2011 and March 2012 for which
  # the temperature was reater than 0.3
  p_hat0 <- length(tempMonthYear0) / n0
  p_hat1 <- length(tempMonthYear1) / n1

  # calculate the difference between the proportions
  Difference2 <- p_hat0 - p_hat1

  # calculate the standard error
  SE2 <- sqrt((p_hat0 * (1 - p_hat0) / n0) + (p_hat1 * (1 - p_hat1) / n1))

  # construct the confidence interval
  CI_2props <- c(Difference2 - 1.96 *SE2, Difference2 + 1.96 * SE2)

  # display results
  cat("Confidence Interval for difference between 2 population
     proportions\n")
  cat("of days in which the temperature in March \n")
  cat("was greater than .3 from 2011 to 2012:\n")
  cat(CI_2props)
} # PartB
```

B Problem 2 Code

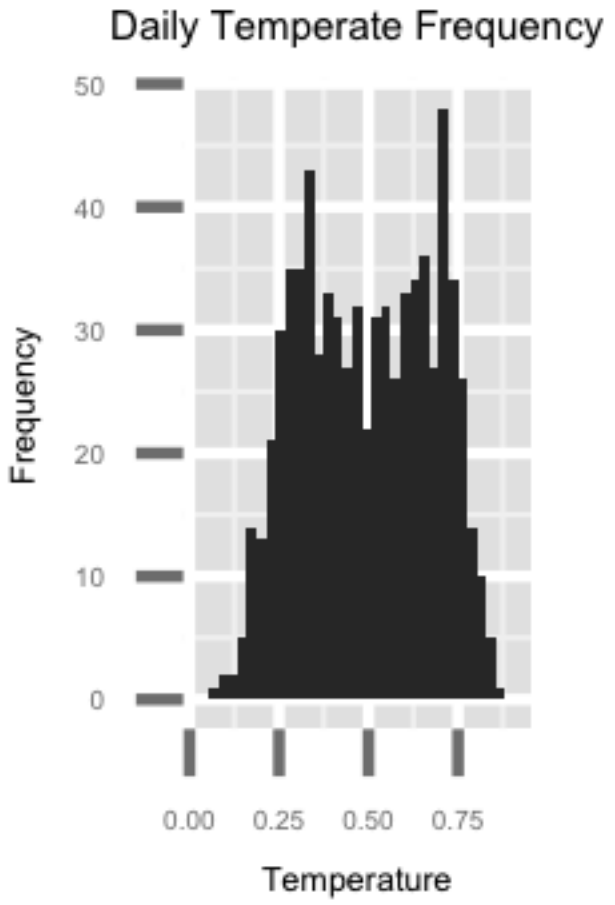
```
Problem2 <- function()
{
    day <- read.csv(file="day.csv",header=TRUE,sep=",")
    return (day)
} # end Problem2
```

C Problem 3 Code

```
library(ggplot2)

PartA <- function()
{
    day <- read.csv(file="day.csv",header=TRUE,sep=",")
    p <- ggplot(day) + geom_histogram(aes(temp))
    p <- p + labs(title="Daily Temperate Frequency",
       x="Temperature",y="Frequency")
    p <- p +
       theme(title=element_text(size=3),text=element_text(size=3),axis.ticks=element_l
    print(p)
    ggsave("Problem3A.pdf",width=1,height=1.5,plot=last_plot())
    dev.off()
} # PartA

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

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

D Problem 3 Plots

D.1 Part A



E Group Member Contributions

- William Otwell:
    1. Planned and conceptualized the statistics that were compared in Problem 1
    2. Helped develop and debug R code for Problem 1
- Rupali Saiya:
    1. Planned and conceptualized the statistics that were compared in Problem 1
    2. Helped develop the Problem 1 Write-Up
- Nicholas Layton:
    1. Developed R code for Problem 1 Part A
    2. Helped elaborate on Problem 1 Write Up
    3. Developed the R code for Problem 3 Part A
- Syeda Inamdar:
    1. Developed R code for Problem 1 Part B