PSTAT 105 HW 4

Matthew Xu (5752811)

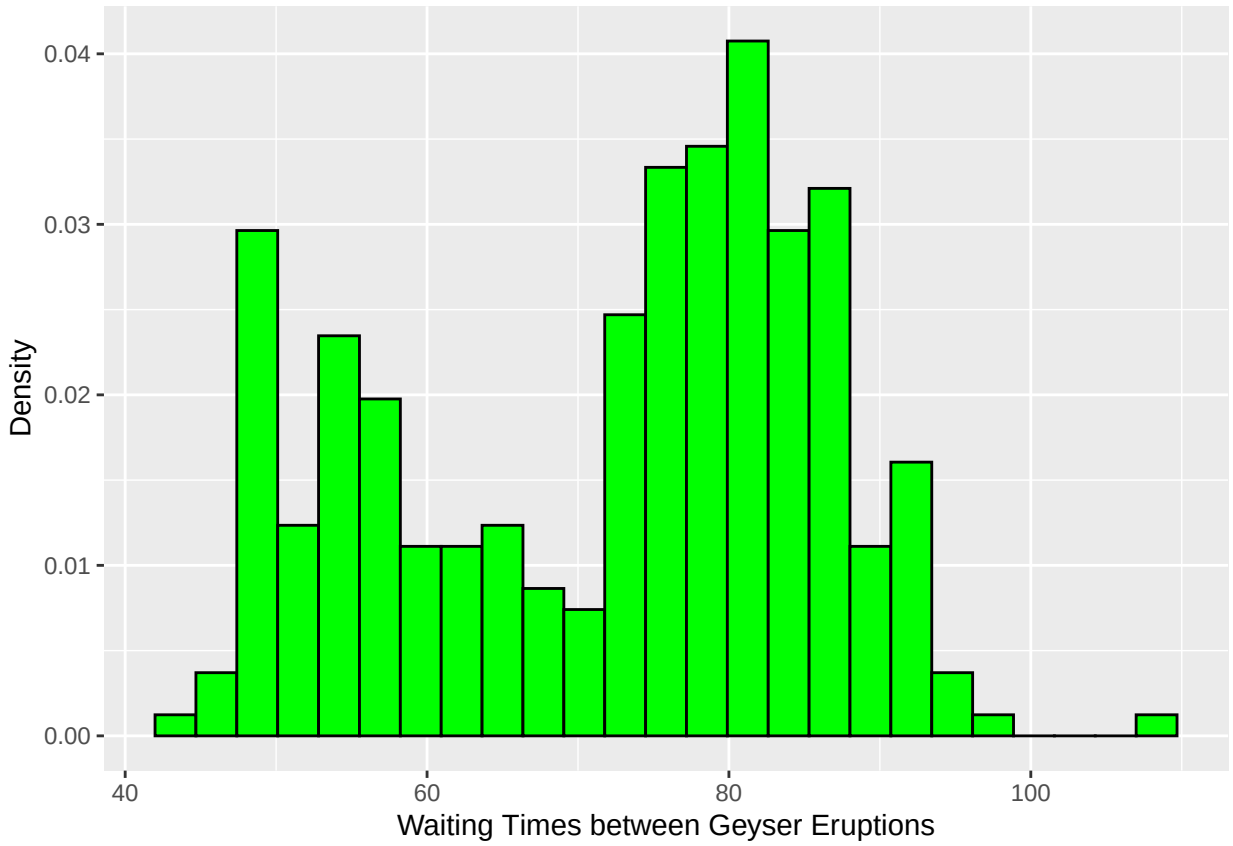
12 Feburary 2021

```
# libraries
library(MASS)
library(tidyverse)

# scan in snow data
snow <- scan("snow.txt")
```

3. The Old Faithful Data that we looked at in lecture is available in R as geyser in the MASS library

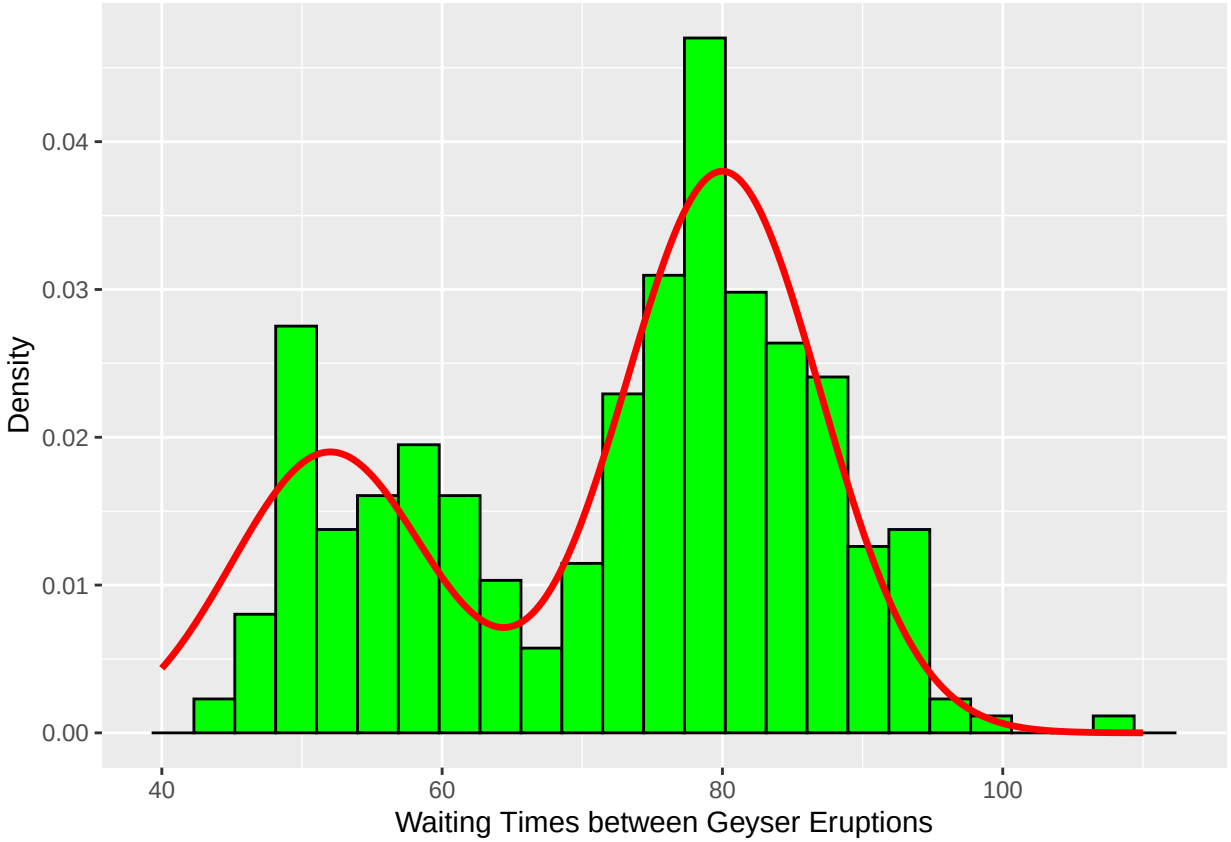
(a) Plot a histogram of this data with a reasonable number of bins.

```
# histrogrm with reasonable bins of 25
geyser_hist <- ggplot(geyser, aes(x = waiting, y = after_stat(density))) + geom_histogram(bins
    fill = "green", color = "black") + labs(x = "Waiting Times between Geyser Eruptions",
    y = "Density")

geyser_hist
```



(b) Plot the histogram again, and add onto it a curve representing the mixture density f(x) = 1/3phi1(x) + 2/3phi2(x) where where phi1(x) and phi2(x) are the normal densities with standard deviation 7 and means 52 and 80 respectively.

```
# approximate range of observations of waiting times
r <- seq(40, 110, length = 300)

# density curve representing mixture density phi are normal densities with
# paramaters above
density_line <- (dnorm(r, mean = 52, sd = 7)/3) + (2 * dnorm(r, mean = 80, sd = 7)/3)

# adding on density curve
geyser_hist + annotate(geom = "line", x = r, y = density_line, size = 1.2, color = "red")
```



(c) Does this mixture density look like it fits the data?

The density line and histrogram are similiar at low points, particularily when waiting times is at about 65 and 100. The density line and histrogram differ at the peaks, particularily when waiting times is at about 50 and 80. The histogram at these peaks has a larger density than the computed density from the curve.

(d) Use a KS test to test to see if the mixture density f(x) fits the geyser data

```
# function that outputs cdf of mixture, input as function argument in ks.test by
# taking the integral of f(x), we get the integrals of the normal densities f1
# and f2, their cdfs

# q is for qunatile
mixtureCDF = function(q) {
    cdf = (pnorm(q, mean = 52, sd = 7)/3) + (2 * pnorm(q, mean = 80, sd = 7)/3)
    return(cdf)
}

ks.test(geyser$waiting, mixtureCDF)
```

```
##
```

```
##  One-sample Kolmogorov-Smirnov test
##
## data:  geyser$waiting
## D = 0.081275, p-value = 0.0385
## alternative hypothesis: two-sided
```

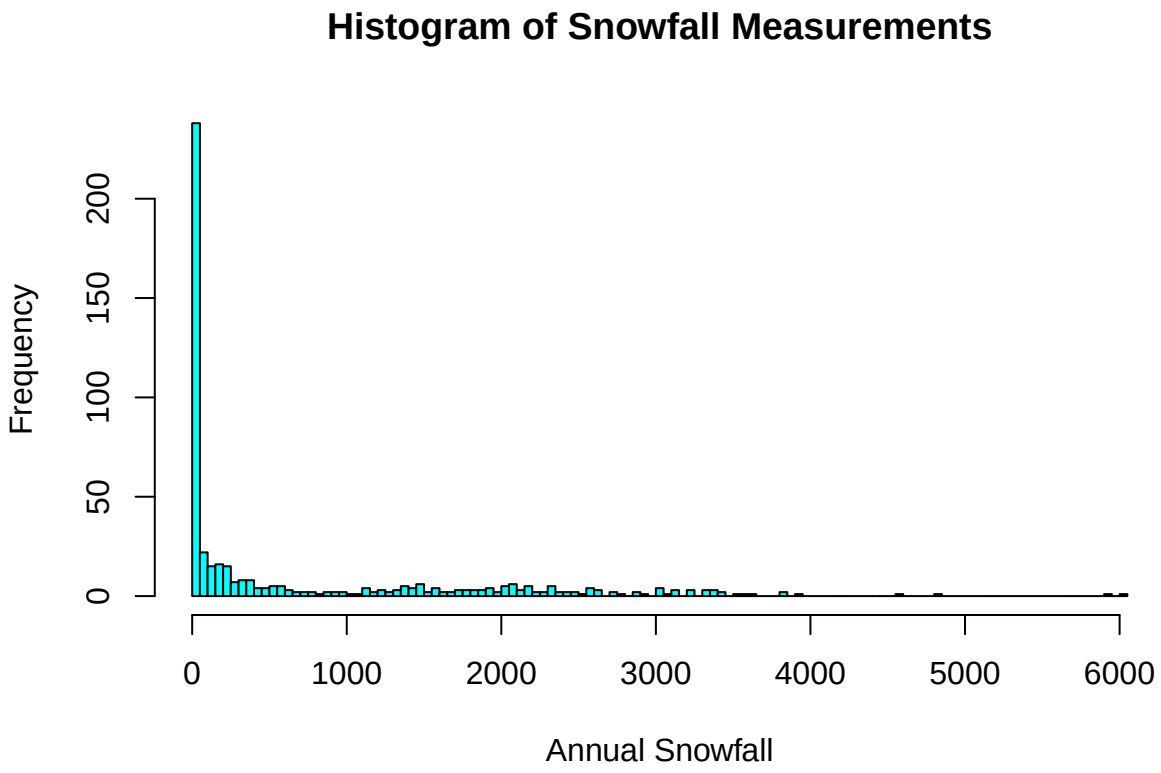
H0: Mixture density f(x) fits the geyser data Ha: Mixture density f(x) does not fit the geyser data

Because the p value is less than alpha of 0.05, we reject the null hypothesis. There is sufficient information that the mixture density f(x) does not fit the geyser data and that it is statistcally significant.

4. The data file snow.txt contains measurements on annual snowfall from different weather stations around Lake Tahoe from the last 50 years. There are 499 measurements.

(a) Draw a histogram that shows the details of the distribution of the data.

```
# histrogram with reasonable breaks of 150
hist(snow, breaks = 150, col = 5, main = "Histogram of Snowfall Measurements", xlab = "Annual S
    ylab = "Frequency")
```



(b) Estimate the density of the distribution at the value of 2000 inches by counting the number of observations in a small interval around that value. Give a 95% confidence interval for this measurement. Justify the bandwidth that you use.

```r
# bandwith
h <- 100
# number of observations
n <- 499
hat.p <- mean(snow > (2000 - h/2) & snow < (2000 + h/2))
hat.f <- hat.p/(h)

# computing confidence inteval using binomial approximation to normal
error = sqrt((hat.p * (1 - hat.p))/(n * h^2))
# 95% confidence intevral z= 1.96
confidence_interval = hat.f + base::c(-1.96, 1.96) * error
confidence_interval
```

```
## [1] 3.709075e-05 2.434704e-04
```

The bandwidth appears to be reasonable with the 95% CI of the densities, which are consistent with the chart in part(d)

The 95% confidence interval for the number of observations in the interval around 2000 inches is between 3.709075e-05 and 2.434704e-04.

(c) Estimate the probability that a station sees no snow in a year.

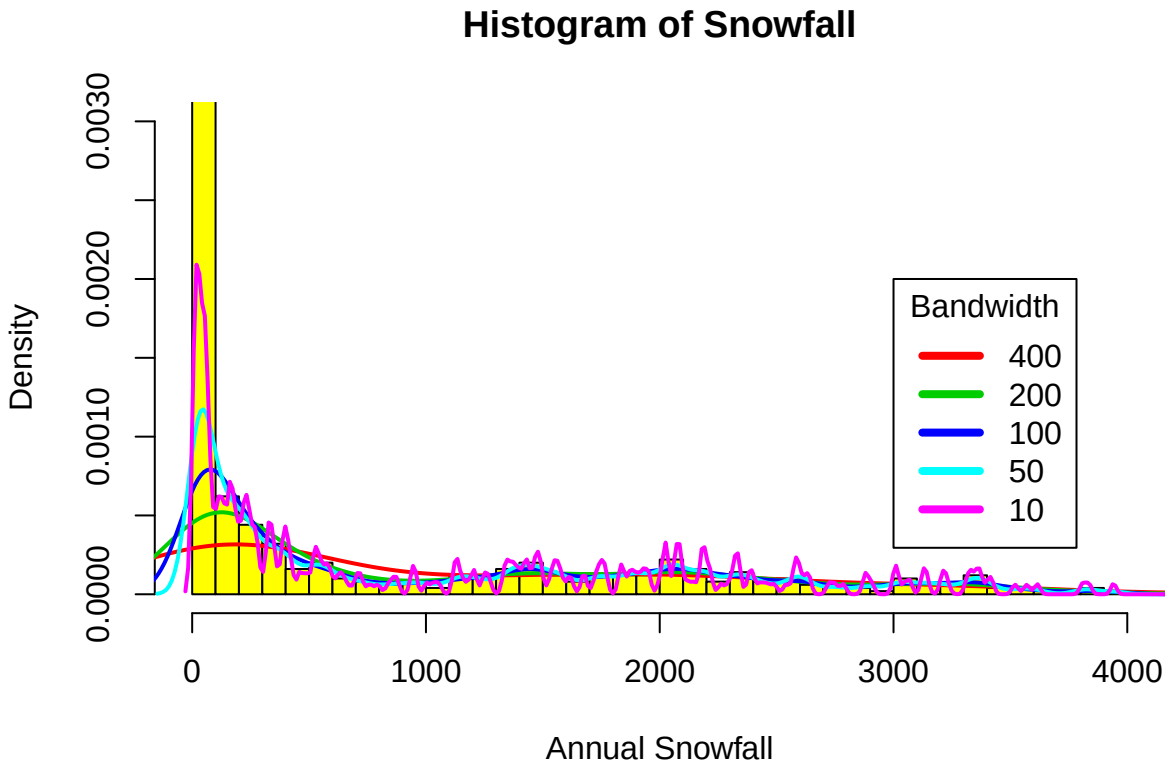
```r
# seeing no snow is when snow = 0
mean(snow == 0)
```

```
## [1] 0.3667335
```

(d) Use the function density to estimate a density over the data that is greater than 0, and plot the results as a smooth line on top of a histogram of the data. Use a Gaussian kernel. Try a number of different bandwidths until you find one that looks right.

```r
hist(snow, breaks = 60, col = 7, prob = TRUE, main = "Histogram of Snowfall", xlab = "Annual Sn
    ylab = "Density", ylim = c(0, 0.003), xlim = c(0, 4000))

band400 <- density(snow[snow > 0], bw = 400, kern = "gaussian")
lines(band400$x, band400$y * (1 - mean(snow == 0)), lwd = 2, col = 2)

band200 <- density(snow[snow > 0], bw = 200, kern = "gaussian")
lines(band200$x, band200$y * (1 - mean(snow == 0)), lwd = 2, col = 3)

band100 <- density(snow[snow > 0], bw = 100, kern = "gaussian")
lines(band100$x, band100$y * (1 - mean(snow == 0)), lwd = 2, col = 4)

band50 <- density(snow[snow > 0], bw = 50, kern = "gaussian")
```

```
lines(band50$x, band50$y * (1 - mean(snow == 0)), lwd = 2, col = 5)

band10 <- density(snow[snow > 0], bw = 10, kern = "gaussian")
lines(band10$x, band10$y * (1 - mean(snow == 0)), lwd = 2, col = 6)

legend(3000, 0.002, legend = c("400", "200", "100", "50", "10"), col = c(2, 3, 4,
    5, 6), lwd = rep(4, 4), title = "Bandwidth")
```



(e) Can you give an educated guess as to how much bias is present in your answer to part (b)?

```
# elpison half the length of the confidence interval
#+elipson, -elpison represent the upper and lower intervals
elipson = (3.709075e-05 + 0.0002434704)/2

# sigmas best choice is n^(-1/5)
sigma = n^-0.2

# calculauting using bias formula
second_derv = (elipson^-2) * ((0.0002434704 - hat.f) - (hat.f - 3.709075e-05))
bias = (sigma^2 * second_derv)/6
```

Using the equation for bias and the second derivative with best choice of sigma = n^(-1/5), we calculated bias to be 1.958676e-05. It appears as you raise bandwidth, the densitites tend to increase. Underestimations typically occur at the large outleir extreme, at the beginning where

snowfall is approximately 100, for all the bandwidths that are tested. Overestimations occur when bandwiths are too large, making them smaller than the denstiy curve.