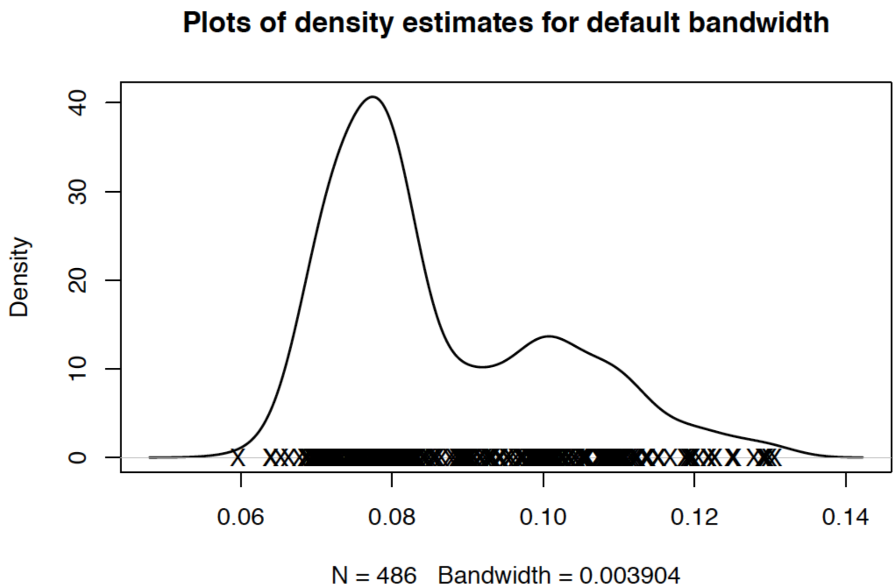
1A

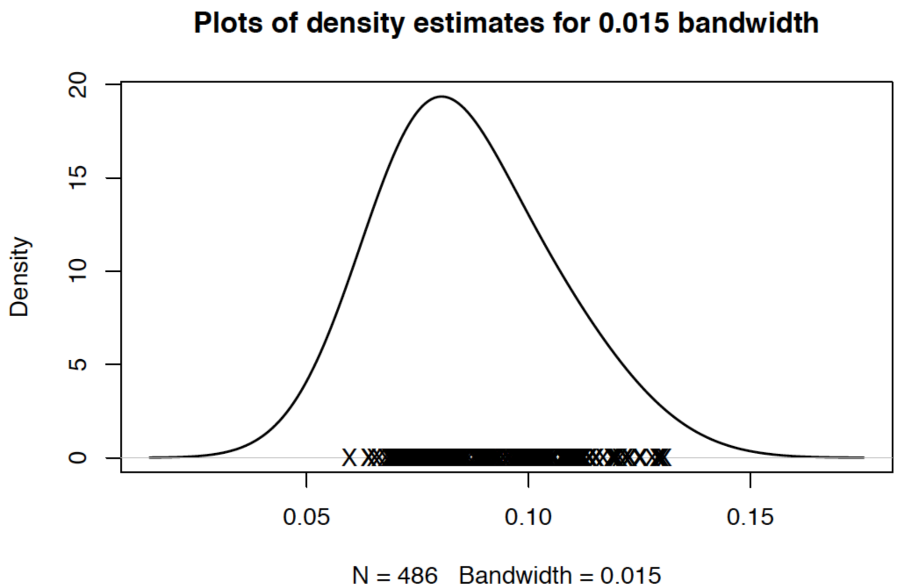
Plots of density estimates of varying bandwidth. We will be analyzing the modes for these plots.

```
# Load stamp data
stampdata <- scan("stamp.txt")
plot(density(stampdata), lwd=1.5, main = "Plots of density estimates for default bandwidth")
points(stampdata, rep(0,486), pch="X")
```



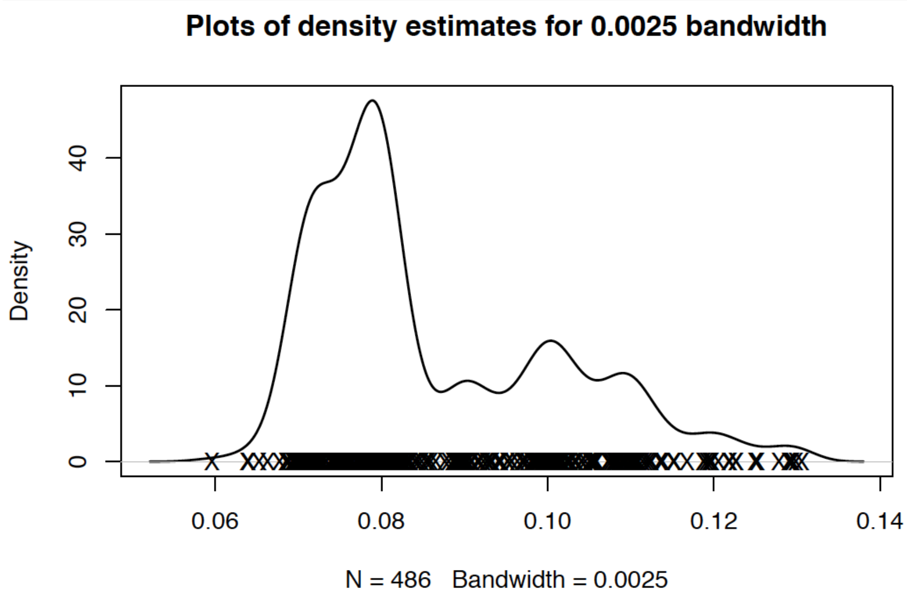
Starting off with the default bandwidth we see that the default bandwidth value is 0.003904. This density plot has 2 modes.

```
plot(density(stampdata, bw=0.015), lwd=1.5, main = "Plots of density estimates for 0.015 bandwidth")
points(stampdata, rep(0,486), pch="X")
```



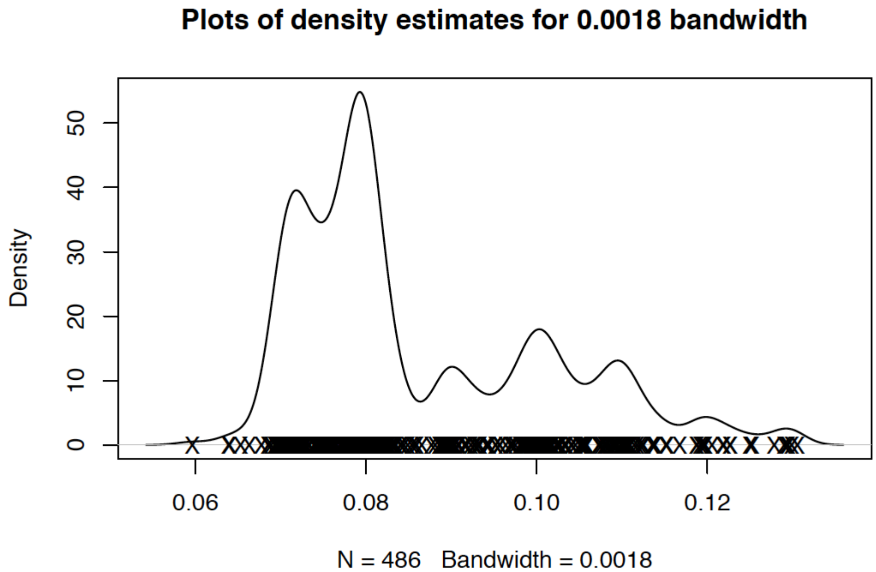
If we set the bandwidth to 0.015 then we see that the density estimate plot has one mode.

```
plot(density(stampdata, bw=0.0025), lwd=1.5, main = "Plots of density estimates for 0.0025 bandwidth")
points(stampdata, rep(0,486), pch="X")
```



If we set the bandwidth to 0.0025 then we see that the density estimate plot has five modes.

```
plot(density(stampdata, bw=0.0018), lwd=1.2, main = "Plots of density estimates for 0.0018 bandwidth")
points(stampdata, rep(0,486), pch="X")
```



If we set the bandwidth to 0.0018 then we see that the density estimate plot has seven modes.

Analysis:

As we have seen from the density estimate graph with a bandwidth of about 0.0025 produced five modes and a density estimate graph with a bandwidth of about 0.0018 produced seven modes. From these trends we notice that as we decrease the bandwidth value that the local modes become more pronounced. The greater the bandwidth the less modes we will see on the graph.

Sta355 q1 b)

i) want to Show

$\mathcal{L}(h) \uparrow \infty$ as $h \downarrow 0$

assume $w(0) > 0$ and for $x \neq 0$
$h^{-1}(w/h) \to 0$ as $h \downarrow 0$

We do not know if $X_1, \ldots, X_n$ are distinct or not So we handle each case individualy.

Non distinct
Some elements of $X_1, \ldots, X_n$ are equal meaning for some $i, j \in [1, n]$, $i \neq j$
$X_i = X_j$.

Can also be interpreted as atleast 2 different elements from $X_1, \ldots, X_n$ are the same Value.

from this we know $X_i - X_j = 0$ and from assumption we know $w(0) > 0$ thus.

$$\sum_{j=1}^{n} w\left(\frac{X_i - X_j}{h}\right) = \sum_{j=1}^{n} w(0) > 0$$

we know $h \downarrow 0$ so $\frac{1}{nh} \to \infty$

then,

$$\frac{1}{nh}\sum_{j=1}^{n} w\left(\frac{X_i - X_j}{h}\right) \to \infty$$

finally by adding the following terms we have.

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} \ln\left[\frac{1}{nh}\sum_{j=1}^{n} w\left(\frac{X_i - X_j}{h}\right)\right]}_{= L(h)} \to \infty$$

In the second part assume that $X_1, \ldots, X_n$ are distinct. Meaning $i, j \in [1, n]$, $i \neq j$, $X_i \neq X_j$.

This can also be be interpreted all no 2 differing elements of $X_1, \ldots, X_n$ have the same value.

then as $X_i - X_j \neq 0$ we have

$$h^{-1} w\left(\frac{x_i - x_j}{h}\right) \to 0 \quad \text{(from assumption)}$$

then

$$\frac{1}{nh}\sum_{j=1}^{n} w\left(\frac{X_i - X_j}{h}\right) \to 0$$

cont.

as $x \to 0$ $\ln(x) \to -\infty$

so

$$\ln\left[\frac{1}{nh}\sum_{j=1}^{n} w\left(\frac{x_i - x_j}{h}\right)\right] \to -\infty$$

finally

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} \ln\left[\frac{1}{nh}\sum_{j=1}^{n} w\left(\frac{x_i - x_j}{h}\right)\right]}_{L(h)} \to -\infty$$

thus from both cases we can conclude $L(h) \uparrow \infty$ as $h \downarrow 0$

ii) Show when $X_1, \ldots, X_n$ are distinct (no tied observations) $CV(h) \to -\infty$ as $h \downarrow 0$ and $h \uparrow \infty$.

assume $w(0) > 0$, $x \neq 0$, $h^{-1} w(x/h) \to 0$ as $h \downarrow 0$

use same definition of distinct as part i)

① case $CV(h) \to -\infty$ as $h \uparrow \infty$

as $X_1, \ldots, X_n$ are distinct we know $X_i - X_j \neq 0$ and if $h \uparrow \infty$ then we have

$$\frac{X_i - X_j}{h} \to 0 \quad \text{then,}$$

$$w\left(\frac{X_i - X_j}{h}\right) \to w(0)$$

from assumption we know $w(0) > 0$ thus,

$$w\left(\frac{X_i - X_j}{h}\right) > 0 \quad \text{So,}$$

$$\sum w\left(\frac{X_i - X_j}{h}\right) > 0 \quad [i \neq j]$$

also as $h \uparrow \infty$ then $\frac{1}{h} \to 0$ then $\frac{1}{(n-1)h} \to 0$

combing it together we have

$$\frac{1}{(n-1)h} \sum w\left(\frac{x_i - x_j}{h}\right) \to 0 \quad [i \neq j]$$

then,

$$\ln\left(\frac{1}{(n-1)h} \sum w\left(\frac{x_i - x_j}{h}\right)\right) \to -\infty \quad [i \neq j]$$

finally

$$\frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{1}{(n-1)h} \sum w\left(\frac{x_i - x_j}{h}\right)\right) \to -\infty \quad [i \neq j]$$

$$CV(h) \to -\infty \quad \text{as} \quad h \uparrow \infty$$

Case 2 $CV(h) \to -\infty$ as $h \downarrow 0$

as $X_1, \dots, X_n$ are distinct we know

$X_i - X_j \neq 0$

Using assumption we know

$$h^{-1} w\left(\frac{x_i - x_j}{h}\right) \to 0 \quad \text{as } h \downarrow 0$$

thus

$$\frac{1}{h} \sum w\left(\frac{x_i - x_j}{h}\right) \to 0 \quad [j \neq i]$$

then,

$$\frac{1}{h(n-1)} \sum w\left(\frac{x_i - x_j}{h}\right) \to 0 \quad [j \neq i]$$

so,

$$\ln\left(\frac{1}{h(n-1)} \sum w\left(\frac{x_i - x_j}{h}\right)\right) \to -\infty \quad [j \neq i]$$

finally

$$\frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{1}{h(n-1)} \sum w\left(\frac{x_i - x_j}{h}\right)\right) \to -\infty \, [j \neq i]$$

Q1B Pg 7/7

which means

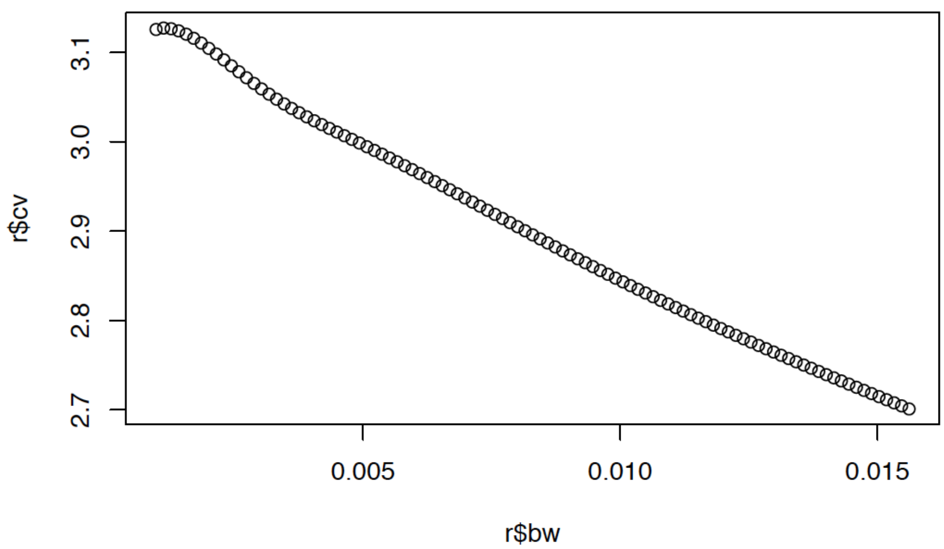
$CV(h) \to -\infty$ as $h \downarrow 0$

with both parts we have showed $CV(h) \to -\infty$ as $h \downarrow 0$ and $h \uparrow \infty$

1c

```
kde.cv <- function(x,h) {
              n <- length(x)
              if (missing(h)) {
                 r <- density(x)
                 h <- r$bw/4 + 3.75*c(0:100)*r$bw/100
                 }
              cv <- NULL
              for (j in h) {
                 cvj <- 0
                 for (i in 1:n) {
                    z <- dnorm(x[i]-x,0,sd=j)/(n-1)
                    cvj <- cvj + log(sum(z[-i]))
                    }
                 cv <- c(cv,cvj/n)
                 }
                r <- list(bw=h,cv=cv)
                r
}

r <- kde.cv(stampdata)
k <- plot(r$bw,r$cv) # plot of bandwidth versus CV
```



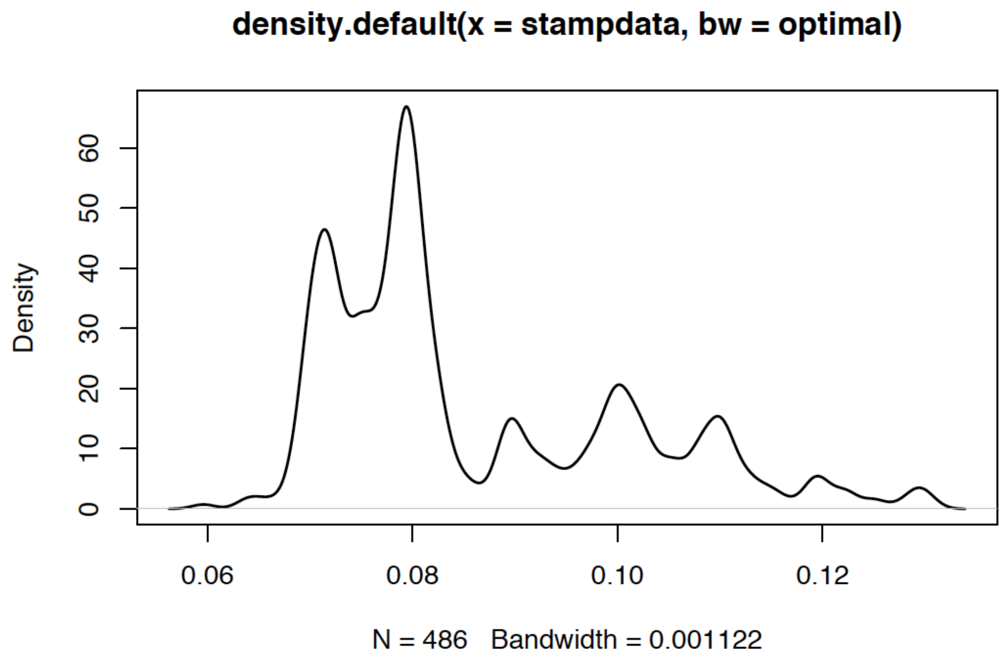
From above we were able to first plot plot CV versus.

```
optimal <- r$bw[r$cv==max(r$cv)] # bandwidth maximizing CV
optimal
```

```
## [1] 0.001122498
```

Using this function we were able to find the optimal bandwidth and using the bandwidth we are able to plot the estimate the density of the Hidalgo stamp data.

```r
finplot <- plot(density(stampdata, bw=optimal), lwd=1.5)
```



From this density estimate graph we are able to see that there are about 7 modes on this graph.

Sta 3ss Q2 a)

a) Show: $g(t) = t - L_f(t)$ is maximized at t satasfiying $F^{-1}(t) = \mu(F)$

To maximize $g(t)$ we solve

$g'(t) = 0.$

$$g'(t) = \frac{d}{dt}\left(t - L_f(t)\right) = 0$$

$$= 1 - \frac{d}{dt}\left(L_f(t)\right) = 0$$

$$= -\frac{d}{dt}\left(\frac{1}{\mu(F)}\int_0^t F^{-1}(s)\,ds\right) = -1$$

$$= \frac{-1}{\mu(F)}\left(\frac{d}{dt}\int_0^t F^{-1}(s)\,ds\right) = -1$$

$$= -F^{-1}(t) = -\mu(F)$$

$$= F^{-1}(t) = \mu(F)$$

Thus we have shown

$g(t) = t - L_f(t)$ is maximized at t

satasfiying $F^{-1}(t) = \mu(F)$

Pg 2/2

Sta355 q2 b)

b) Show using result from part a) that

$$P(F) = \frac{E_F[\,|X - \mu(F)|\,]}{2\mu(F)}$$

[assume F has a density f]

$$E(X - \mu(F)) = \int_0^{\mu(F)} (\mu(F) - x) f(x)\,dx +$$

$$\int_{\mu(F)}^{\infty} (x - \mu(F)) f(x)\,dx$$

$$= 2\int_0^{\mu(F)} (\mu(F) - x) f(x)\,dx = E(x - \mu(F))$$

thus

$$\int_0^{\mu(F)} (\mu(F) - x) f(x)\,dx = \frac{E(x - \mu(F))}{2}$$

furthermore,

$$\frac{E(x - \mu(F))}{2\,\mu(F)} = \left(\int_0^{\mu(F)} (\mu(F) - x) f(x)\,dx\right)\left(\frac{1}{\mu(F)}\right)$$

q2b pg 2/2

require change of variables
new bounds → $[0, F(u(F))]$

$F(x) \to F^{-1}(y)$ [from part a]

$dx = \frac{1}{f(x)} dy$

$$\left(\int_0^{F(u(F))} (u(F) - F^{-1}(y)) f(x) \frac{1}{f(x)} dy \right) \frac{1}{u(F)}$$

(from part a)

$$= \int_0^{t} \frac{u(F)}{u(F)} dy - \int_0^{t} \frac{F^{-1}(y)}{u(F)} dy$$

$$= t - \frac{1}{u(F)} \int_0^{t} F^{-1}(y) dy$$

$$= t - L_F(t)$$

from part a we know that because

$t = F(u(F))$ its at max

$= P(F)$

2c

```r
incomedata <- scan("incomes.txt")

pietra_fcn <- function(data){
  return (mean(abs(data-mean(data)))/(2*mean(data)))

}


pietra_fcn(incomedata)
```

```
## [1] 0.1875046
```

```r
theta_count <- NULL

for(i in 1:200){
  theta_count[i] <- pietra_fcn(incomedata[-i])
}

theta_dot <- mean(theta_count)

jackknife <- (sqrt((length(incomedata) - 1)/length(incomedata) * sum((theta_count - theta_dot)^2)))
jackknife
```

```
## [1] 0.009046412
```

Using the code above we found that the pietra value is 0.1875046 and the jackknife value is 0.009046412