Homework 10

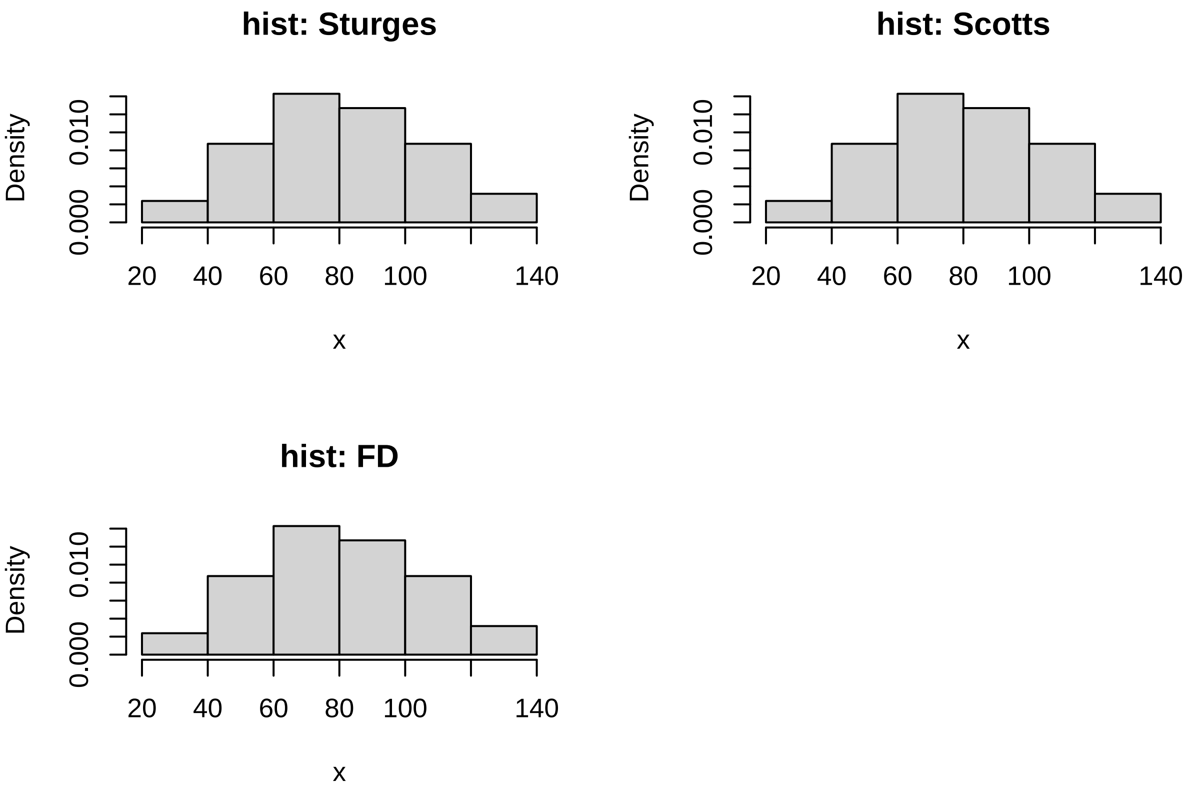
11/30/2021

Question 1

Construct a histogram estimate of density for the buffalo data set in the gas package using Sturges' Rule, Scott's Normal Reference Rule, and Freedman-Diaconis Rule. Which rule provides better density estimate? Compute the corresponding density estimates $f(x)$ when $x = 88$ from the three histogram estimates.

```
n <- 1000
data("buffalo")
x <- buffalo

par(mfrow=c(2,2))
h.sturges <- hist(x, breaks='Sturges', freq = F, main="hist: Sturges")
h.scotts <- hist(x, breaks='Scott', freq = F, main="hist: Scotts")
h.fd <- hist(x, breaks='FD', freq = F, main="hist: FD")
par(mfrow=c(1,1))
```



```
x0 <- 88
b <- which.min(h.sturges$breaks <= x0) -1
print(c(b,h.sturges$density[b]))
```

```
## [1] 4.00000000 0.01269841
```

```
b <- which.min(h.scotts$breaks <= x0) -1
print(c(b,h.scotts$density[b]))
```

```
## [1] 4.00000000 0.01269841
```

```
b <- which.min(h.fd$breaks <= x0) -1
print(c(b,h.fd$density[b]))
```

```
## [1] 4.00000000 0.01269841
```

Question 2 (12.4)

Construct a frequency polygon density estimate for the precip data set, using a bin width determined by substituting

$$\hat{\sigma} = IQR/1.348$$

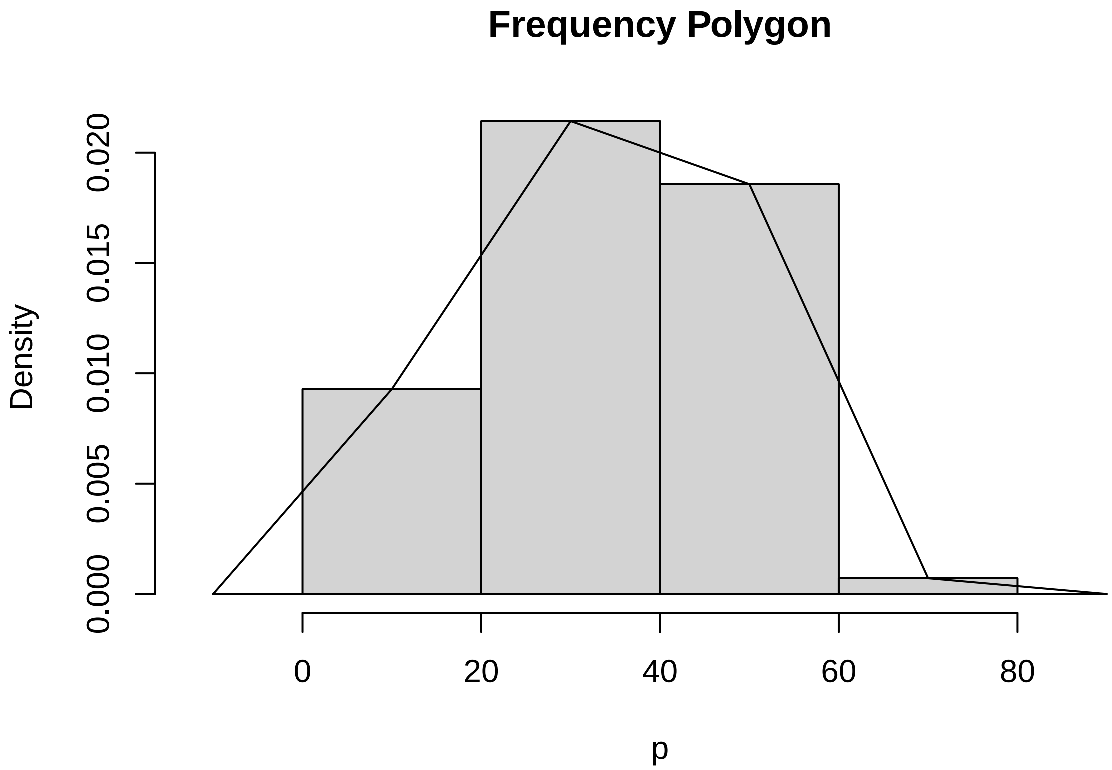
for standard deviation in the usual Normal Reference Rule for a frequency polygon.

```
p <- precip
n <- length(p)

# freq poly bin width using normal ref rule
h <- 2.15*(IQR(p))*n^(-1/5)

# calculate the sequence of breaks and histogram
br <- pretty(p, diff(range(p))/h)
brplus <- c(min(br)-h, max(br+h))
histg <- hist(p, breaks=br, freq=F, main="Frequency Polygon", xlim=brplus)

vx <- histg$mids # midpoints of each class interval
vy <- histg$density # density est at vertices of polygon
delta <- diff(vx)[1] # h after pretty is applied
k <- length(vx)
vx <- vx+delta
vx <- c(vx[1]-2*delta, vx[1]-delta, vx)
vy <- c(0, vy, 0)
# add the polygon to the histogram
polygon(vx,vy)
```



Question 3

Construct an *ASH* density estimate for the *faithful$eruptions* data set in R, using width h determined by the normal reference rule, set $m = 15$.

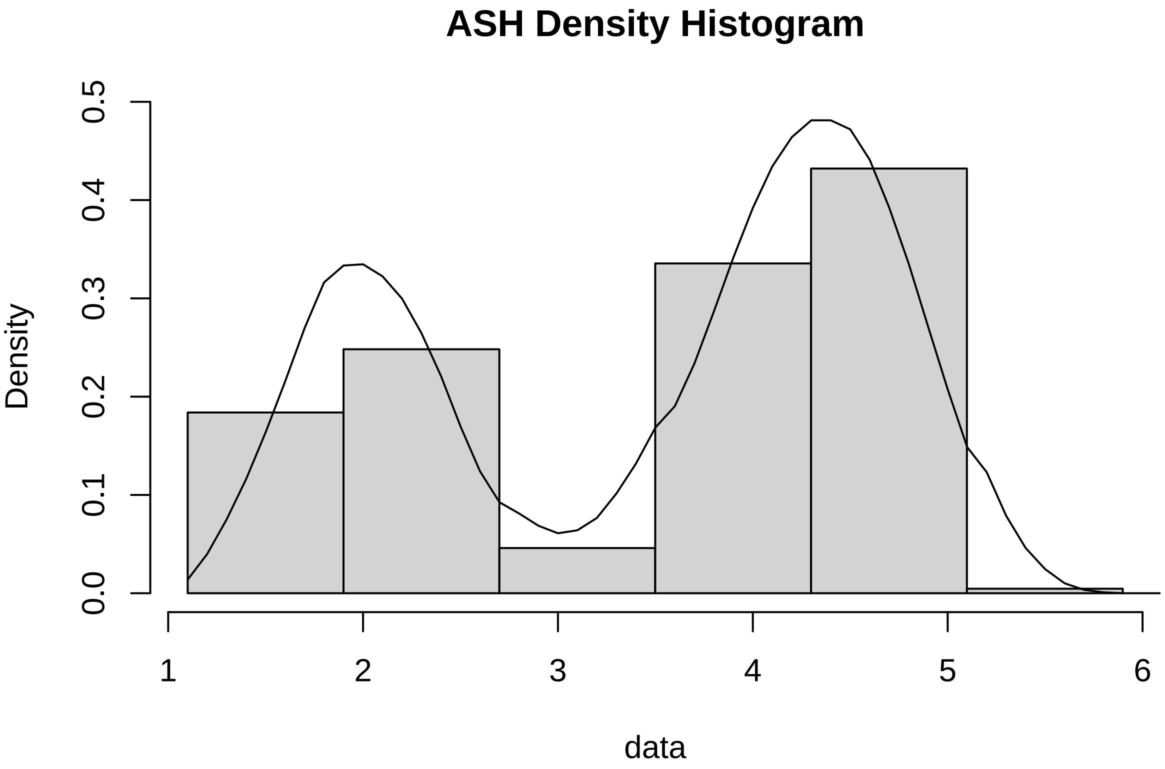
```r
data <- faithful$eruptions
n <- length(data)
m <- 15
a <- min(data) - 0.5
b <- max(data) + 0.5
h <- 2.15*sqrt(var(data))*n^(-1/5)
delta <- h/m

# get the bin counts on the delta-width mesh
br <- seq(a - delta*m, b + 2*delta*m, delta)
histg <- hist(data, breaks = br, plot = F)
nk <- histg$counts
K <- abs((1-m):(m-1))

fhat <- function(x){
  #locate the leftmost interval containing x
  i <- max(which(x>br))
  k <- (i-m+1):(i+m-1)
  #get the 2m-1 bin counts centered at x
  vk <- nk[k]
  sum((1-K/m)*vk)/(n*h) #f.hat
}
```

```r
# density can be computed at any points in range of data
z <- as.matrix(seq(a,b+h, .1))
f.ash <- apply(z, 1, fhat) #density estimates at midpts

# plot ASH density estimate over histogram
br2 <- seq(a, b+h, h)
hist(data, breaks = br2, freq=F, main="ASH Density Histogram", ylim=c(0, max(f.ash)))
lines(z, f.ash, xlab="")
```



Question 4 (12.8)

The buffalo data set in the *gss* package contains annual snowfall accumulations in Buffalo, New York from 1910 to 1973. The 64 observations are

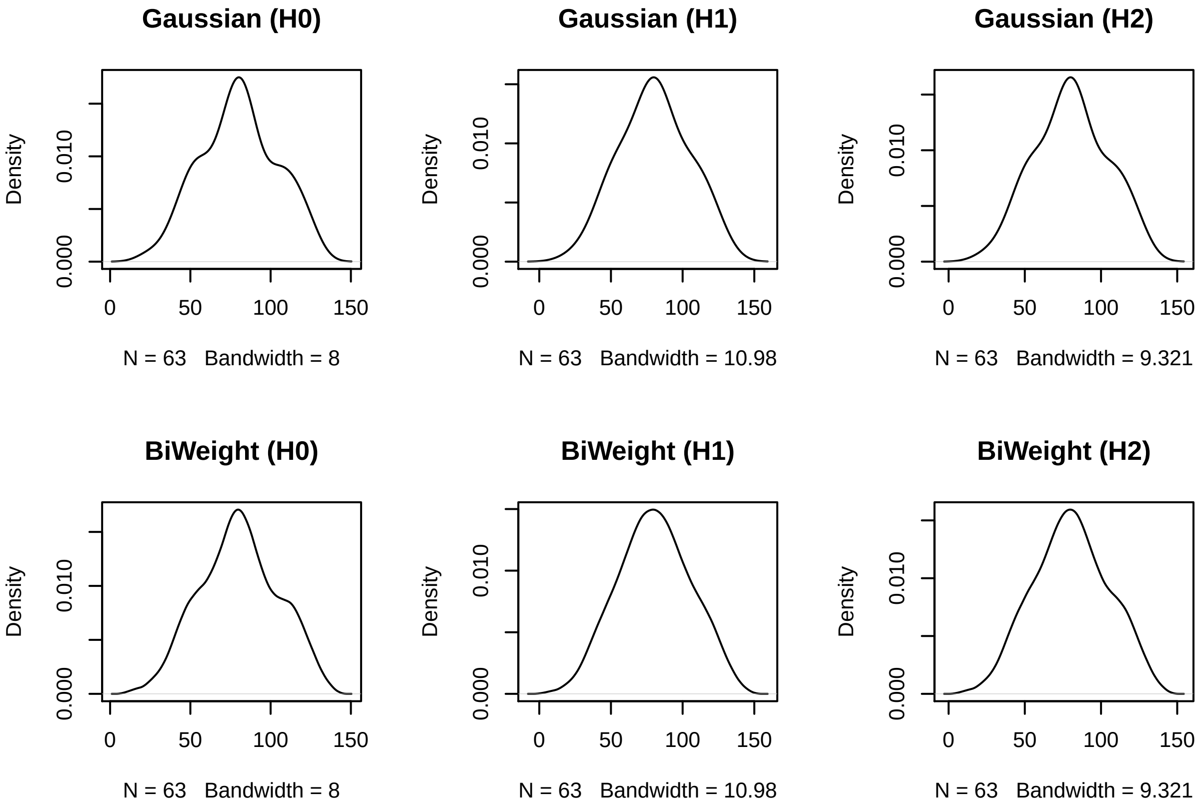
```
 126.4  82.4  78.1  51.1  90.9  76.2 104.5  87.4 110.5  25.0  69.3  53.5
 39.8   63.6  46.7  72.9  79.6  83.6  80.7  60.3  79.0  74.4  49.6  54.7
 71.8  49.1  103.9  51.6  82.4  83.6  77.8  79.3  89.6  85.5  58.0 120.7
 110.5  65.4  39.9  40.1  88.7  71.4  83.0  55.9  89.9  84.8 105.2 113.7
 124.7 114.5 115.6 102.4  101.4  89.8  71.5  70.9  98.3  55.5
 66.1  78.4 120.5  97.0 110.0
```

This data was analyzed by Scott [262]. Construct kernel density estimates of the data using Gaussian and biweight kernels. Compare the estimates for different choices of bandwidth. Is the estimate more influenced by the type of kernel or the bandwidth?

```r
data("buffalo")
x <- buffalo
n <- length(x)
```

```
h1 <- 1.06*sd(x)*n^(-1/5)
h2 <- 0.9*min(c(IQR(x)/1.34,sd(x)))*n^(-1/5)
h0 <- 8

par(mfrow=c(2,3))
plot(density(x,bw=h0, kernel="gaussian"), main="Gaussian (H0)")
plot(density(x,bw=h1, kernel="gaussian"), main="Gaussian (H1)")
plot(density(x,bw=h2, kernel="gaussian"), main="Gaussian (H2)")
plot(density(x,bw=h0, kernel="biweight"), main="BiWeight (H0)")
plot(density(x,bw=h1, kernel="biweight"), main="BiWeight (H1)")
plot(density(x,bw=h2, kernel="biweight"), main="BiWeight (H2)")
```



```
par(mfrow=c(1,1))
```

-Ans: The estimate looks to be more influenced by the bandwidth.