4.1 Find the order of each of the following expressions as $n \to \infty$

(a) $\frac{2}{n} + \frac{3\log(n)}{n^2}$

$\frac{2}{n}$

(b) $\frac{4n^2+1}{(n+1)^2}$

$\frac{4n^2}{n^2}$ will approach 4

(c) $O(n^2)O(\frac{1}{n}) + 4$

n

(d) $O(\frac{1}{n^2}) + O(\frac{1}{\sqrt{n}}) + O(1)$

$\frac{1}{\sqrt{n}}$

4.2

(a) Let X denote a random variable with a Poisson distribution with mean λ; that is, X has a discrete distribution with frequency function

$\frac{\lambda^x exp(-\lambda)}{x!}, x = 0, 1, 2, \ldots.$

Find the order of

$\frac{Pr(X=n+1)}{Pr(X=n)}$ as $n \to \infty$

$\frac{1}{x}$ will approach zero

(b) Let Y denote a random variable with a Poisson distribution with mean nλ; for instance, Y might be the sum of n independent Poisson random variables each with mean λ Find the order of

$\frac{Pr(Y=n+1)}{Pr(Y=n)}$ as $n \to \infty$

$\frac{\lambda e^1}{x}$ therefor $\frac{1}{x}$ will approach 0

4.3 Let $X_1, X_2, \ldots. X_{100}$ denote independent random variables, each with a continuous distribution with density function

$px(x) = \frac{3}{2}(x-1)^2, 0 < x < 2$

and let $Y_1, Y_2, \ldots. Y_{100}$ denote independent random variables each with a continuous distribution with density function

$py(y) = \frac{3}{8}(y)^2, 0 < y < 2$

Suppose that a histogram is constructed for each set of data, using a bin width of h=0 in both cases (that is, the bins are (0,.1), (.1,.2),..) and let $\hat{p}XH(.)$ and $\hat{p}YH(.)$ denote the corresponding histogram density estimators

(a) Use the AIMSE to determine which estimator do you expect to be more accurate? That is, do you expect $\hat{p}XH(.)$ to be a better estimate of $px(x)$ and $\hat{p}YH(.)$ to be a better estimate of $py(y)$? Or do you expect the reverse to be true?

px(x) and py(y) both have $h^3 > \frac{3}{n}$ suggesting bias makes a larger contribution to the AIMSE than does the variance. Within the limit $\hat{p}H(.)$ is a perfectly accurate estimate of p(,); the results from this suggest that we see a large bias.

(b) Now suppose that we are primarily interested in estimating the density at the argument 1; that is, we want to estimate either $px(1)$ or $py(1)$. Using the leading terms in the expansions for the MSE of $\hat{p}XH(.)$ and $\hat{p}YH(.)$, do you expect $\hat{p}XH(1)$ to be a better estimator of $px(1)$ than $\hat{p}YH(1)$ is of $py(1)$? Or do you expect the reverse to be true? Why?

I would expect px(x) to be a better estimate of px(1) because the variance increases greatly for py(y)

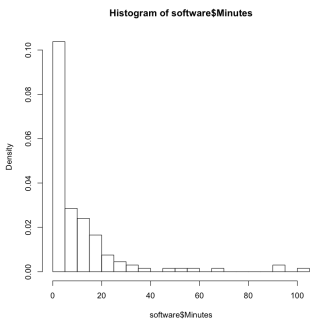
4.4 Consider the failure data for a software system. Construct a histogram for the data, choosing the bin width using the Freedman-Diaconis method. Compare the number of bins corresponding to the Freedman-Diaconis bin width to the number of bins you chose in Exercise 3.5

Using FD-bin width of 5

In exercise 3.5-bin width of 15

In [35]:
```
data_loc<-'/Users/Alexis/Documents/Spring2020/nonparametrics/data/software.csv'
software<-read.csv(data_loc, header=FALSE)
colnames(software) <- c("Minutes")
hist(software$Minutes, breaks='FD', freq=F)$breaks
```

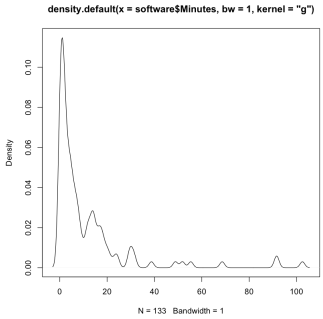
```
0  5  10  15  20  25  30  35  40  45  50  55  60  65  70  75  80  85  90  95  100  105
```

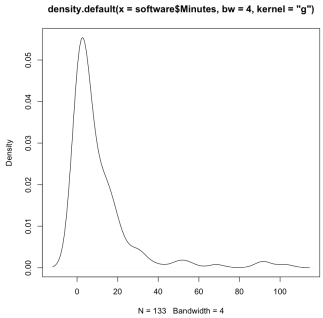


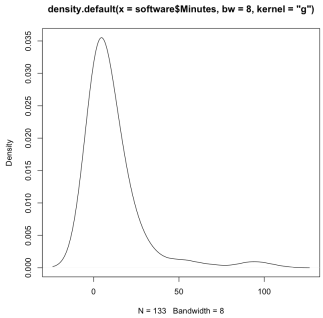
4.5 Consider the failure data. Construct four kernel density estimates for these data, using values of the smoothing parameter of 1,4,8, and 12, respectively and a Gaussian kernel. Which of the estimates appears to give the bets summarization of the data? In answering this question keep in mind the fact that the data are necessarily non-negative

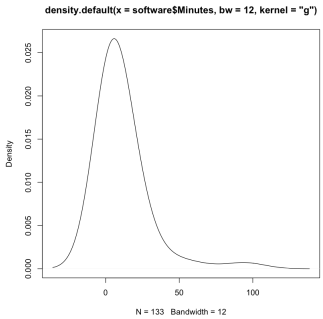
Based on the literature the plot with the smoothing parameter bs=4 produces the best summarization of the data as anything beyond that value looks marginally the same.

In [38]:
```
plot(density(software$Minutes, bw=1, kernel='g'))
plot(density(software$Minutes, bw=4, kernel='g'))
plot(density(software$Minutes, bw=8, kernel='g'))
plot(density(software$Minutes, bw=12, kernel='g'))
```









4.6 Consider the scores on the peabody dataset.

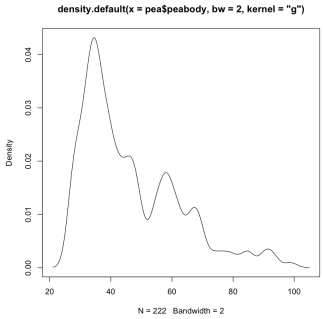
(a) Construct a kernel density estimate for these data, using a Gaussian kernel. Consider four possible values for the smoothing parameter h, 2, 3, 4, 5. Choose the best value of the smoothing parameter h subjectively and plot the estimate corresponding to your choice. Which value of h did you choose?

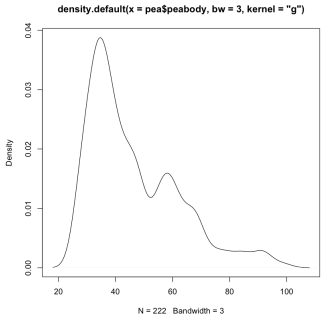
I would assume the highest smoothing parameter to yield the best value. Based on the results I would actually choose bw=3 as it keeps some stochastic nature of the data while smoothing.

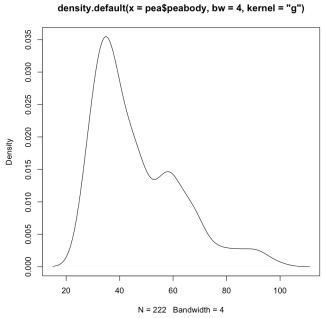
(b) For the density plot constructed in part a include a plot of the normal density function with mean μ and standard deviation σ with the value μ and σ taken to be the estimate based on the peabody data. Based on this plot do the peabody scores appear for the chidlren in the study appear to be normally distributed?

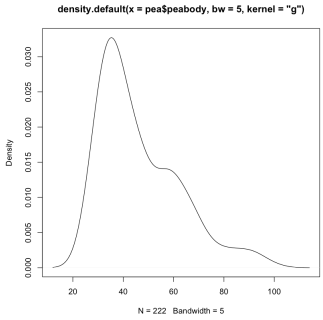
Based on this plot they don't appear to be normally distributed but are skewed to the right.

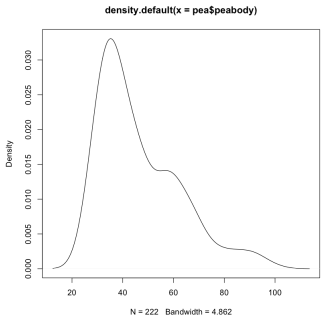
In [43]:
```
data_loc<-'/Users/Alexis/Documents/Spring2020/nonparametrics/data/peabody.csv'
#data_loc<-'/Users/aporter1350/Documents/Courses/Spring2020/nonparametrics/data/peabody.csv'
pea<-read.csv(data_loc)
plot(density(pea$peabody, bw=2, kernel='g'))
plot(density(pea$peabody, bw=3, kernel='g'))
plot(density(pea$peabody, bw=4, kernel='g'))
plot(density(pea$peabody, bw=5, kernel='g'))

plot(density(pea$peabody))
```











In [ ]: