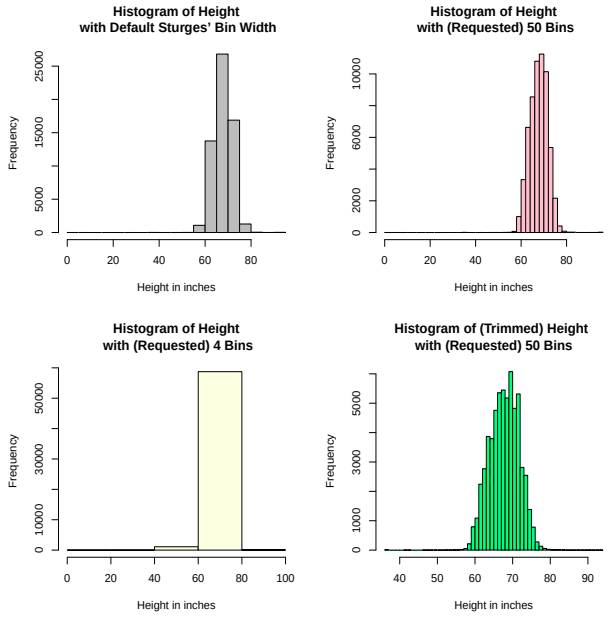
Visualization and Learning Structure: Problem Session 7/2/16 Solutions

1. In class today, we explored the age and smoke status of the OkCupid users; now we'll look at their height and drug use status.

   (a) *First, explore the effect of different histogram bin widths on the distribution of height. Note that the height variable has some questionable values in it. You can choose to include them or not. See* `help(hist)` *for how to change the number/location of the bins.*

   Specifically, you may want to compare using the Sturges' default rule to using smaller or larger bins. Does your height data look normal?



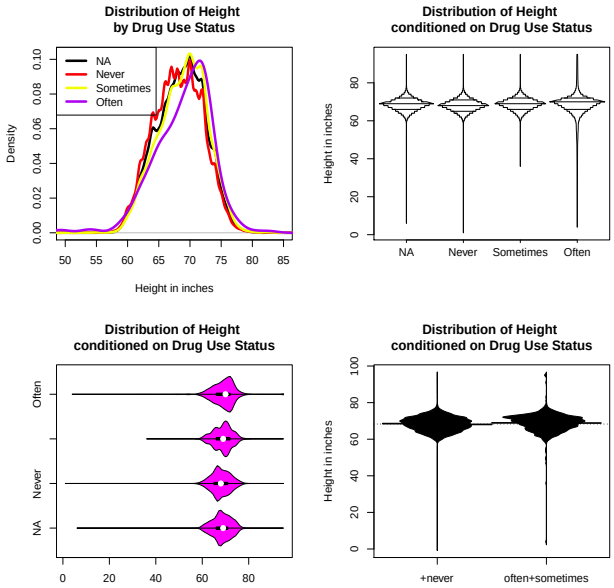
So here we've created four histograms for the OK Cupid height values using 1) default Sturges' bin width, 2) 50 bins, 3) 4 bins, and 4) 50 bins but only for height observations between 36in and 95in (recall in class we talked about the likely presence of typos and unknown codes in the height variable). While our distribution is fairly normal and so somewhat justifies the use of the default Sturges' bin width, I still prefer the upper right histogram that shows the same normality but at a finer grain resolution for the gradient drop off. The lower left histogram is way too coarse and shows almost no features. We created the "trimmed histogram" in order to zoom-in on the realistic height values to verify that their shape remained normal (the outliers in our original data can

stretch the tails and make everything else look like "the middle"). We have some minor multi-modality at the peak but it's more than likely attributable to noise in the data.

```
summary(height)
sort(height)[1:20]  #looking at the shortest heights
rev(sort(height))[1:20] #looking at the tallest heights
par(mfrow=c(2,2))
hist(height,col="grey",xlab="Height in inches",main="Histogram of Height\n with
Default Sturges' Bin Width")
hist(height,breaks=50,col="pink",xlab="Height in inches",main="Histogram of
Height\n with (Requested) 50 Bins")
hist(height,breaks=4,col="lightyellow",xlab="Height in inches",main="Histogra
of Height\n with (Requested) 4 Bins")
height.trim<-height[height>=36 & height < 95]
hist(height.trim, breaks=50,col="springgreen1",xlab="Height in inches",main=
"Histogram of (Trimmed) Height \nwith (Requested) 50 Bins")
```

(b) Now let's look at height by drug use status (" " the blank group, "never", "sometimes", "often"). Try at least three different methods of comparing the conditional distributions of height given the drug status.

- If you use kernel density estimates in some form, what kernel and bandwidth do you use?
- Describe any differences you see in the height distributions across subgroups.



In the previous figure, we have kernel density estimates (upper left), box-percentile plots (upper right), violin plots (lower left), and bean plots (lower right). The kernel density estimates (default kernel, default bandwidth) show primary modes in about the same location (68-70 inches) but the level of smoothness differs. We could supplement this graph with information about the sample size in each group (to help us determine if we're looking at signal or noise) and a "rug plot" which graphs the actual data values on the bottom of the graph, colored by subgroup. Including the actual data (even if the individual values aren't discernible) might help us better ascertain the cause of the bumps we see in the primary mode. Note that our ability to do that rests on our use of only one variable. The box-percentile plots and the violin plots show us roughly the same information. The "Other" group is slightly shifted older, and we see a noisier distribution for the "Never" group. The side-by-side bean plots indicate similar subgroup averages. The "Blank" group and the "Never" group seem more similar to each other than the pair "Often" and "Sometimes". Overall, without additional research questions asking for more specific information, I would just choose the KDEs. We could always add the averages, medians, etc on top of that graph if necessary.

Note that all these graphs included all of the height values, including the unreasonable ones. We could easily re-do this analysis using the trimmed group of heights.

```
par(mfrow=c(2,2))
de1<-density(height[drugs==""],na.rm=T,from=min(height,na.rm=T),to=max(height,
na.rm=T))
de2<-density(height[drugs=="never"],na.rm=T,from=min(height,na.rm=T),to=
max(height,na.rm=T))
de3<-density(height[drugs=="sometimes"],na.rm=T,from=min(height,na.rm=T),to=
max(height,na.rm=T))
de4<-density(height[drugs=="often"],na.rm=T,from=min(height,na.rm=T),to=
max(height,na.rm=T))
plot(de1,type="l",lwd=3,xlim=c(50,85),xlab="Height in inches",main="Distribution
of Height \nby Drug Use Status")
lines(de2,lwd=3,col="red"); lines(de3,lwd=3,col="yellow"); lines(de4,lwd=3,col="purple")
legend("topleft",c("NA","Never","Sometimes","Often"),col=c("black","red",
"yellow","purple"),lwd=3)
```

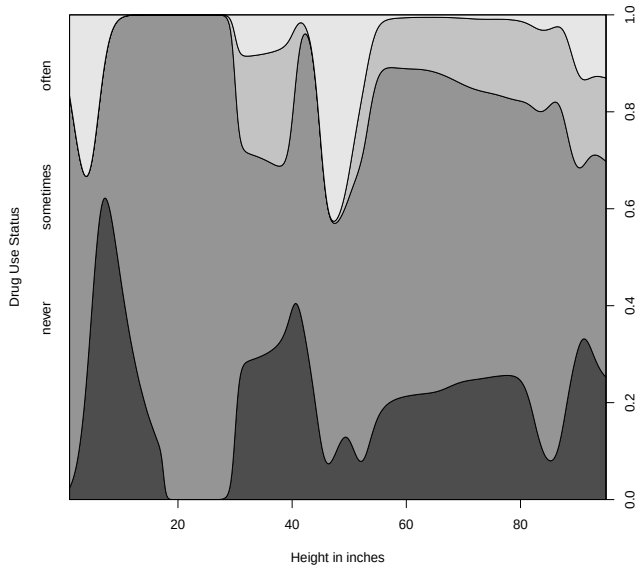
```
library(Hmisc)
bpplot(height[drugs==""],height[drugs=="never"],height[drugs=="sometimes"],
height[drugs=="often"],name=c("NA","Never","Sometimes","Often"),ylab="Height
in inches",main="Distribution of Height \nconditioned on Drug Use Status")

library(vioplot)
vioplot(height[drugs==""&!is.na(height)],height[drugs=="never"&!is.na(height)],
height[drugs=="sometimes"],height[drugs=="often"],names=c("NA","Never",
"Sometimes","Often"),h=1,horizontal=TRUE)
title("Distribution of Height \nconditioned on Drug Use Status")

library(beanplot)
beanplot(height~drugs,side="both",what=c(1,1,1,0),ylab="Height in inches")
title("Distribution of Height \nconditioned on Drug Use Status")
```

(c) Compare the conditional distributions of drug use status as a function of height (conditional density plot; `help(cdplot)`). Pick a reasonable bandwidth.

At what heights do we see big changes in the distribution of drug use status?



First note that we changed the order of the drug use categories so there would be an increase in usage along the axis (R will default into putting categories into alphabetical order - rarely helpful). One of the first features that jumps out is that the "never" category completely dominate around 20-30 inches.

(This doesn't exactly make sense; we would need to check the height values. It's likely also a sample size issue.) Shorter and taller people were more likely to leave the question blank. The "often" use of drugs is maximized in these conditional distributions about 45-50 inches. The "Sometimes" category appears in pockets of height (35-40, 50-80). As we approach the shortest and the tallest people, we'll need to pay close attention to sample size and not overstate the strength of the related conclusions.

Again, we could re-do this analysis removing the people with questionable height values.

```
library(graphics) ##to get the cdplot function
Drugs<-factor(drugs,levels=c("","never","sometimes","often"))  ##reordering categories
cdplot(Drugs~height,bw=2,xlab="Height in inches",ylab="Drug Use Status")
```

2. We can estimate the true density $f(x)$ using a kernel density estimate $\hat{f}(x)$:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where n is the number of observations, h is the bandwidth, and K is the choice of kernel function. Kernels are often chosen for ease of computation and to reduce bias as much as possible (i.e. $E[\hat{f}(x)] \approx f(x)$).

Three useful properties are:

$$\int K(t)dt = 1 \qquad \int tK(t)dt = 0 \qquad \int t^2K(t)dt = \sigma_K^2$$

These properties are more standard/well-known for the Gaussian "bell" kernel of $K_G(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$, but you can use any appropriate function for which they hold.

Verify that these properties hold for the Uniform [-1,1] kernel, $K_U(t) = \frac{1}{2}$, and the Epanechnikov "bubble" kernel, $K_E(t) = \frac{3}{4}(1 - t^2)$. Both compact on $t \in [-1, 1]$. (Note that $\sigma^2_{K_U} = \frac{1}{3}$ and $\sigma^2_{K_E} = \frac{1}{5}$.)

For the uniform:

$$\int K(t)dt = \int_{-1}^{1} \frac{1}{2}dt = \frac{1}{2}[1 - (-1)] = 1 \qquad \int tK(t)dt = \frac{1}{2}\frac{t^2}{2}|_{-1}^{1} = 0$$

$$\int t^2K(t)dt = \frac{1}{2}\frac{t^3}{3}|_{-1}^{1} = \frac{1}{2}\left[\frac{1}{3} + \frac{1}{3}\right] = \frac{1}{3}$$

For the Epanechnikov:

$$\int K(t)dt = \frac{3}{4}\int_{-1}^{1}(1-t^2)dt = \frac{3}{4}(t-t^3/3)|_{-1}^{1} = \frac{3}{4}[(1-1/3)-(-1+1/3)] = \frac{3}{4}\frac{4}{3} = 1$$

$$\int tK(t)dt = \frac{3}{4}\int_{-1}^{1}(t-t^3)dt = \frac{3}{4}(t^2/2-t^4/2)|_{-1}^{1} = \frac{3}{4}[(1-1)-(1-1)] = 0$$

$$\int t^2K(t)dt = \frac{3}{4}\int_{-1}^{1}(t^2-t^4)dt = \frac{3}{4}(t^3/3-t^5/5)|_{-1}^{1} = \frac{3}{4}[(t^3/3-t^5/5)-(t^3/3-t^5/5)]$$
$$= \frac{3}{4}[(1/3-1/5)-(-1/3+1/5)] = \frac{3}{4}\frac{4}{15} = \frac{1}{5}$$

3. We can measure how well our density estimate $\hat{f}(x)$ approximates the truth $f(x)$ by the asymptotic mean integrated square error (ASIME) for $\hat{f}(x)$ as an approximation of $f(x)$ is:

$$AMISE = MISE(h)_{n\to\infty} = E_{n\to\infty}\int(\hat{f}_h - f)^2 = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 R(f^{''})$$

The first term is the dominating term of the integrated variance, the second the integrated square bias. Also note that $R()$ refers to the "roughness" of the function; we can measure it by the integral of its square, i.e. $R(\psi(t)) = \int \psi(t)^2 dt$

(a) Show that the optimal bandwidth, h^*, that minimizes the $AMISE$ is

$$h^* = \left[\frac{R(K)}{\sigma_K^4 R(f'')}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

Our h^* has to satisfy the first order condition (i.e. taking the derivative wrt h):

$$\sigma_K^4 h^3 R(f^{''}) - \frac{R(K)}{nh^2} = 0$$

Rearranging we get:

$$\sigma_K^4 h^5 R(f^{''}) = \frac{R(K)}{n}$$

and then

$$h^* = \left(\frac{R(k)}{n\sigma_K^4 R(f'')}\right)^{1/5} = \left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{1/5} \cdot n^{-1/5}$$

We can see that as h approaches the endpoints of $(0,\infty)$, the AMISE would blow up to ∞. Our h^* is a minimizer.

(b) Given this h^*, find an expression for the optimal AMISE (i.e. $AMISE^*$) as a function of the kernel choice, the sample size, and the true unknown density.

Note that this calculation will also verify that the kernel's contribution to the AMISE is a function of $\sigma_K \cdot R(K)$, i.e. the variance and roughness of the kernel, and allows us to compute relative efficiencies across different kernels.
If we plug in our h^* to AMISE, we get

$$\frac{R(K)}{n\left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{1/5} n^{-1/5}} + \frac{1}{4}\sigma_K^4 \left(\left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{1/5} n^{-1/5}\right)^4 R(f'')$$

$$= \frac{(\sigma_K R(K))^{4/5} R(f'')^{1/5}}{n^{4/5}} + \frac{1}{4}\sigma_K^4 \left(\frac{R(K)}{\sigma_K^4 R(f'')n}\right)^{4/5} R(f'')$$

$$= \left(\frac{R(K)^4 R(f'')\sigma_K^4}{n^4}\right)^{1/5} \left(1 + \frac{1}{4}\right) = \frac{5}{4}\left(\frac{(\sigma_K R(K))^4 R(f'')}{n^4}\right)^{1/5}$$

So only $\sigma_k R(K)$ depends on the choice of kernel K. The kernel that minimizes this expression is the Epanechikov kernel (looks like bubble) with a value of $\sigma_k R(K) = 0.2683$. We could then find the relative efficiency of any other kernel as $\frac{\sigma_K R(K)}{0.2683}$.

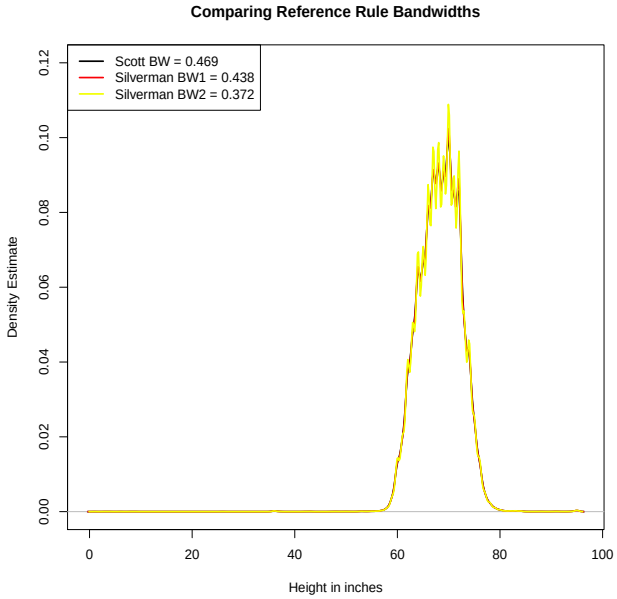
4. Unfortunately, we don't usually know the true $f(x)$ which means we would have to approximate $R(f'')$ to actually find the best bandwidth. Scott and Silverman derived some rules of thumb to help us approximate h:

   $h_{Sc} = 1.06 \cdot \hat{\sigma} \cdot n^{-\frac{1}{5}}$ $\quad$ $h_{Si1} = 0.79 \cdot IQR \cdot n^{-\frac{1}{5}}$ $\quad$ $h_{Si2} = 0.9 \cdot \min(\hat{\sigma}, \frac{IQR}{1.34}) n^{-\frac{1}{5}}$

   Calculate the three above bandwidths for our OkCupid heights.
   Find and graph three kernel density estimates using these bandwidths (using a Gaussian kernel). How different are they? What effects did the different bandwidths have on your density estimates (if any)?

   Useful R functions: `stdev(heights); summary(heights)` (for the IQR)
   The actual bandwidth values (0.469, 0.438, 0.372) are fairly similar but lead to different behavior at the peak of our primary mode (around 70 inches). As expected, a decrease in bandwidth corresponds to an increase in smaller "bumpy" modes. I would probably stick with the Scott bandwidth option. Although it's the largest, it's actually not that conservative and does pick up some minor variation at the top of the mode. The smaller Silverman bandwidths, to me, seem to pick up a little too much noise.

   It would also be useful to see what the suggested bandwidths would be if the tails of the height values were trimmed.




```
height.complete<-na.omit(height)  ##removing the missing values
n<-length(height.complete)
##using sd() in place of stdev() which may not be part of your R
s.height<-sd(height.complete)
h.sc<-1.06*s.height*n^(-1/5)
summary(height.complete);iqr.height<-71.0-66.0
h.si1<-0.79*iqr.height*n^(-1/5)
h.si2<-0.9*min(s.height,iqr.height/1.34)*n^(-1/5)

de.sc<-density(height.complete,bw=h.sc)
de.si1<-density(height.complete,bw=h.si1)
de.si2<-density(height.complete,bw=h.si2)
plot(de.sc,type="l",lwd=2,xlab="Height in inches",ylab="Density Estimate",
ylim=c(0,0.12),main="Comparing Reference Rule Bandwidths")
lines(de.si1,type="l",lwd=2,col="red")
lines(de.si2,type="l",lwd=2,col="yellow")
legend("topleft",c("Scott BW = 0.469", "Silverman BW1 = 0.438",
"Silverman BW2 = 0.372"),lwd=2,col=c("black","red","yellow"))
```