In Class Workup 1-29-14

Form into working groups. Each person should be sure to understand how to solve each question. Keep your record on your own paper and study it as part of preparing for exam 1.

Your group leader will submit one paper (not their own) with all participating students names on it. You may solicit my help or the help of anyone in class.

1-9 Kernel density smoothing and bootstrap.

1. Express the kernel shown below mathematically as a function \( f \) on real line \( \mathbb{R} \) and also as a function \( f \) defined on \( \mathbb{R} \) in computer language \( \mathbb{R} \).

\[
\begin{align*}
Y &= \begin{cases} 
1 - |x| & -1 \leq x \leq 1 \\
0 & x \in \mathbb{R} \setminus [-1,1] \\
1 - \text{abs}(x) & 
\end{cases} \\
\text{function}(x) &= \begin{cases} 
1 & x > -1 & \text{and} & x < 1 \\
0 & \text{else} 
\end{cases}
\end{align*}
\]

2. Refer to (1). Calculate the variance \( \int_{-1}^{1} x^2 f(x) \, dx \) (twice \( \int_{0}^{1} x^2 f(x) \, dx \)) by hand.

\[
\begin{align*}
\int_{0}^{1} x^2 (1-x) \, dx &= 2 \int_{0}^{1} x^2 \, dx - 2 \int_{0}^{1} x^3 \, dx \\
&= 2 \left( \frac{x^3}{3} - \frac{x^4}{4} \right) \bigg|_{0}^{1} = \frac{1}{6}
\end{align*}
\]

3. Using facts proven in this course (do not prove them) determine \( h > 0 \) for which the function \( g(x) = \frac{1}{h} f\left( \frac{x}{h} \right) \) is a probability density having mean 0 and variance 1.

\[
\begin{align*}
\text{Var}(g(x)) &= \int_{-h}^{h} [g(x)]^2 \, dx = \frac{1}{h} \int_{-1}^{1} \left( \frac{x}{h} \right)^2 \, dx = \frac{1}{6} \\
\text{so } h &= \sqrt{6}.
\end{align*}
\]

4. For \( g \) determined in (2) and data \( X_1 = 3.1, 3.6 \) plot the weighted kernels \( \frac{1}{n} g(x - X_i) \) and plot their sum. The sum is the kernel density smoothing of the data.

Two kernels are:
\[
\begin{align*}
\left\{ \begin{array}{ll}
\frac{1}{2h} \left( 1 - \frac{|X - 3.1|}{h} \right) & \text{for } 3.1 - h \leq X \leq 3.1 + h \\
0 & \text{else}
\end{array} \right.
\end{align*}
\]

\[
\begin{align*}
\left\{ \begin{array}{ll}
\frac{1}{2h} \left( 1 - \frac{|X - 3.6|}{h} \right) & \text{for } 3.6 - h \leq X \leq 3.6 + h \\
0 & \text{else}
\end{array} \right.
\end{align*}
\]

they sum up to be:
\[
\begin{align*}
\left\{ \begin{array}{ll}
\frac{1}{2h} \left( 1 - \frac{|X - 3.1|}{h} \right) & \text{for } 3.1 - h \leq X \leq 3.1 + h \\
\frac{1}{2h} \left( 1 - \frac{|X - 3.6|}{h} \right) & \text{for } 3.6 - h \leq X \leq 3.6 + h \\
\frac{1}{2h} \left( 1 - \frac{|X - 3.6|}{h} \right) - \frac{1}{2h} \left( 1 - \frac{|X - 3.1|}{h} \right) & \text{for } 3.1 - h \leq X \leq 3.6 + h \\
\frac{1}{2h} \left( 1 - \frac{|X - 3.1|}{h} \right) - \frac{1}{2h} \left( 1 - \frac{|X - 3.6|}{h} \right) & \text{for } 3.1 + h \leq X \leq 3.6 + h \\
0 & \text{else}
\end{array} \right.
\end{align*}
\]
5. If \( f \) is any kernel (i.e. pr density having mean 0 and variance 1) the bandwidth \( h \) define a density smoothing of data \( X_i \) by kernel \( f \) by \( p(x) = \frac{1}{n} \sum_{i=1}^{n} f\left(\frac{x-x_i}{h}\right). \)

5a. What is the variance of density \( p \)? Do not prove it, just use properties we covered.

\[
\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 f\left(\frac{X_i - \bar{X}}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 f\left(\frac{X_i - \bar{X}}{h}\right) + \frac{2}{n} \sum_{i=1}^{n} (X_i - \bar{X}) f\left(\frac{X_i - \bar{X}}{h}\right) + \frac{1}{n^2} \sum_{i=1}^{n} f\left(\frac{X_i - \bar{X}}{h}\right) + \frac{1}{n^2} \sum_{i=1}^{n} f\left(\frac{X_i - \bar{X}}{h}\right)^2
\]

5b. If \( f \) is the standard normal density how many derivatives has density \( p \) at every \( x \)?

\[
\infty = \frac{k^2}{N^2} + \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 f\left(\frac{X_i - \bar{X}}{h}\right)}{N^2}
\]

6. We smooth an iid sample \( X \) from a population \( \text{pop} \). Our bandwidth is \( h = 0.03 \).

6a. Is our sample density estimating the population density? If not, what is it estimating?

No. It is the density of the smoothed population which is a smoothened distribution similar to \( \text{pop} \).

6b. If the variance of the population is 200 what is the variance of the population smoothed by the same method we used on our sample?

\[
200 + \frac{0.03^2}{N} = 200,000?
\]

6c. What is the approximate variance of the sample density?

\[
200 + 0.03^2 = 200,030
\]

7. On a line use hashmarks \( \# \) to represent data for which \underline{variable bandwidth} density smoothing is strongly indicated.

7a. Here we have almost no data so we estimate the behavior of other points.

7b. In (7a) indicate an x location at which small bandwidth \( h \) is indicated and another location where larger bandwidth is indicated.

7c. Briefly explain the reasoning behind variable bandwidth density estimation.

For an accurate estimation we want a smaller bandwidth, that's why we want small bw at those points. However, for plotted we have barely no sample points. We need to give the density so large bw is chosen to cover the area. Plus, big bw is in only few pts with