Problem Set 6: Visualizing Data (Kernel Density Estimation)

You may not plagiarize code or use any packages other than those preloaded by R (−10). For problems 1 and 2, if you would rather write than type equations, please write legibly or I will require all paper submissions to be typewritten for future homework. See the instructions in Problem Set 2 regarding the submission of R code for problem 3 (−2).

1. Consider the kernel density estimate of a probability density function f,

$$\hat{f}_h(a) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{a-x_i}{h}\right), \quad -\infty < a < \infty,$$

where $K(u) \geq 0$ for all u, $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$, and $h > 0$.

   a. [5 pts] Show that the function $\hat{f}_h$ is a probability density function. Explain the significance of this property.

   b. [5 pts] Let $m_1 = \int_{-\infty}^{\infty} u\hat{f}_h(u)\mathrm{d}u$. Find and interpret m_1.

2. **[5 pts]** Suppose the data comprise $x_1 = 1$, $x_2 = 1.2$, $x_3 = 1.5$, $x_4 = 2.8$, $x_5 = 3$. Let f denote the probability density function of the underlying distribution of the data. Compute the kernel density estimate of $f(2.25)$ using an Epanechnikov kernel,

$$K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{(0,1)}\left(|u|\right),$$

and a bandwidth of 0.8. Provide details of your calculations by producing a table similar to that on page 36 of the slides for topic D.

3. Let K denote the Epanechnikov kernel as defined in problem 2.

   a. [5 pts] Write a function named `epanechnikov` that accepts a required argument named `a`, two optional arguments named `x` and `h` with default values of 0 and 1, respectively, and returns

$$K_h(a-x) = \frac{1}{h}K\left(\frac{a-x}{h}\right).$$

   Examples of calling the function `epanechnikov`:

R Console

```
> epanechnikov(1.2)
[1] 0
> epanechnikov(1.2, x = 1, h = 0.5)
[1] 1.26
> epanechnikov(1.2, x = 2.8, h = 0.5)
[1] 0
```

b. [5 pts] Use the built-in function `curve` and the function `epanechnikov` created in (a) to plot the graph of $K_h(a-x)$ as a function of a, for (x,h) equal to $(0,1)$, $(0,0.5)$, $(0,2)$, $(1,0.75)$, all on the same axes. The horizontal axis should range from -3 to 3 and vertical axis from 0 to 2. Label the axes and provide a legend to identify the four curves.

For the following problems, use only vectorized operations and do not perform the computations separately for each observation.

c. [10 pts] Create a grid of 500 evenly spaced values from 0 to 4 and let a_j denote the jth value in the grid. Suppose the data comprise $x_1 = 1$, $x_2 = 1.2$, $x_3 = 1.5$, $x_4 = 2.8$, $x_5 = 3$. Use the function `epanechnikov` created in (a) to compute

$$\frac{1}{5}K_{0.75}(a_j - x_i)$$

for $i = 1, \ldots, 5$ and $j = 1, \ldots, 500$. The result should be a 500×5 matrix, where the (i, j)th element of the matrix gives

$$\frac{1}{5}K_{0.75}(a_j - x_i).$$

Name the matrix `sand.piles` and use it to graph the rescaled kernel for each observation on the same axes. Add a dot plot of the observations to the graph. An example of the graph is shown on page 46 of the slides for topic D (without the curve that represents the kernel density estimate).

d. [5 pts] Let f denote the probability density function of underlying distribution of the data. Compute the Epanechnikov kernel density estimate of f using a bandwidth of 0.75 at each point on the grid created in (c), as a sum of the five rescaled kernels at the corresponding point. The result should be a vector of length 500. Name the vector `piled.sand` and use it to add a plot of the kernel density estimate to the graph in (c).