9
Kernel Density Estimation

9.1 Introduction

The goal of density estimation is to approximate the probability density function of a random variable given a sample of observations. One of the most popular methods is to use *kernel density estimators*.

For example, suppose we observe ten values from $X \sim N(0,1)$, shown in Figure 9.1a. In general, we don't know the *true* underlying distribution. So how do we go about *estimating* this? How do kernel density estimators work?

9.2 Definition

A kernel is a non-negative, real-valued integrable function K that satisfies the following two requirements:

(9.1)

and

(9.2)

Expression 9.1 ensures that the kernel is a probability density function (pdf), whilst expression 9.2 makes the distribution symmetric about 0. Standard kernels are:

Epanechnikov:

Uniform:

Triangular:

Gaussian:

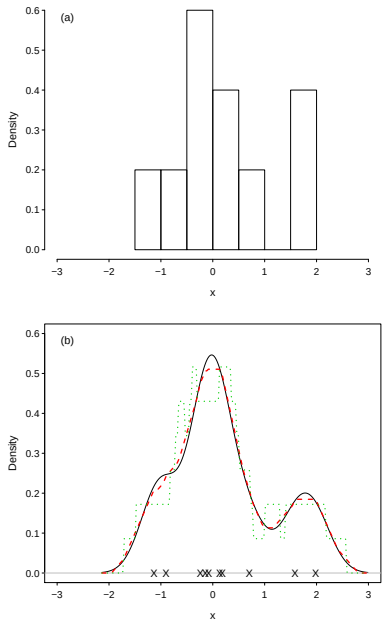


Figure 9.1: (a) Histogram of ten values sampled from a $N(0,1)$ distribution. (b) Three different Kernel density estimators. The data are the X's.

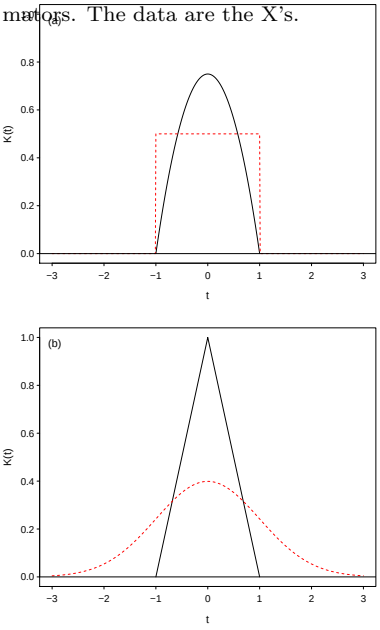


Figure 9.2: (a) The Epanechnikov and Uniform kernels. (b) The triangular and Gaussian kernel.

9.3 The general idea

Kernel density estimation can be summarised in four steps:

1. We have some sample data. In Figure 9.3a we have three points, so $n = 3$.

2. Around each of the data points, we draw a kernel. In Figure 9.3b we have used a Gaussian kernel. However, we could have used a Uniform, triangular, or Epanechnikov kernel.

3. Next we combine the kernels - the blue dashed line in Figure 9.3c.

4. The final step is to normalise the distribution. In our example, since we have three points, the total area under the blue dashed curve is 3. Hence, to recover a probability density we divide by 3 to get the black curve in Figure 9.3d.

Run `demo(kerneldensity)` to get figures 9.3 a-d.

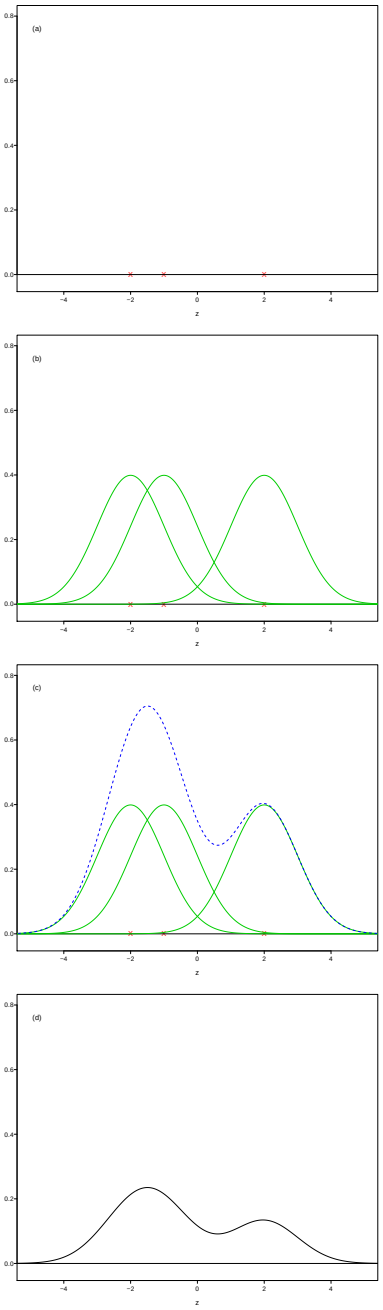


Figure 9.3: (a) Plot showing the data. (b) A Gaussian kernel is drawn round each point. (c) The kernels are combined. (d) The function is scaled an area of one.

9.4 Estimation

Let K be a kernel and suppose our sample contents n values: $x_1, \ldots, x_n$. Then our estimate of the true pdf $f(x)$ is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x - x_i) \ . \tag{9.3}$$

Notice we use $K(x - x_i)$, since we draw a kernel around *each* x_i. It's fairly straightforward to see that $\hat{f}(x)$ is also a pdf, namely:

Example: River Monsters

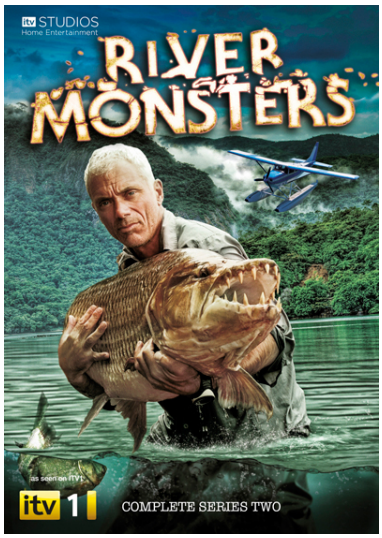
The data shown in Table 9.1 are the lengths (to the nearest cm) of 10 Giant Groupers caught by expert angler Jeremy Wade for the TV series *River Monsters*.

Catch *River Monsters* at 7:30pm, ITV1, every Tuesday.

101	97	99	104	103	94	102	94	102	106

Table 9.1: Length, in cm, of 10 *Giant Groupers* caught in the Everglades National Park, Florida.

This sample was taken in 2013 in a lake near the Chernobyl nuclear disaster of 1986. These fish usually grow to around 75cm in length, but genetic mutations caused by the nuclear explosions are thought to have increased the size of this species.



Produce a density plot for these data using the Gaussian kernel.

Solution

Let X represent the length of a fish. We have:

where the Gaussian kernel gives:

The range of our data is 94$\longrightarrow$106, so let's plot over the range 90$\longrightarrow$110. For example,

Similarly:

$$\hat{f}(\mathbf{94}) = \frac{1}{10}\left[\frac{1}{\sqrt{2\pi}}e^{-(\mathbf{94}-101)^2/2} + \ldots + \frac{1}{\sqrt{2\pi}}e^{-(\mathbf{94}-106)^2/2}\right]$$

$$= 0.08023.$$

Also:

$$\hat{f}(\mathbf{102}) = \frac{1}{10}\left[\frac{1}{\sqrt{2\pi}}e^{-(\mathbf{102}-101)^2/2} + \ldots + \frac{1}{\sqrt{2\pi}}e^{-(\mathbf{102}-106)^2/2}\right]$$

$$= 0.13404.$$

Similar calculations give the results shown in Table 9.2, for values in the range $90 \longrightarrow 110$ (increasing in steps of 1).

x	$\hat{f}(x)$	x	$\hat{f}(x)$
90	0.0000267	101	0.0995432
91	0.0000886	102	0.1340384
92	0.0107983	103	0.1183411
93	0.0484075	104	0.0807319
94	0.0802317	105	0.0546929
95	0.0538066	106	0.0457634
96	0.0354386	107	0.0246539
97	0.0461933	108	0.0054126
98	0.0488910	109	0.0004433
99	0.0515926	110	0.0000138
100	0.0600920		

Table 9.2: Kernel density estimation for the *River Monsters* dataset.

Plotting $\hat{f}(x)$ against x gives the curve shown in Figure 9.4, after smoothing between the points considered. In practice, we would use R to produce such plots, and R would plot over a much finer range for X. Notice how the kernel density estimate of the probability density function is a smoothed version of the histogram. Of course, any of the kernels considered could be used in place of the Gaussian kernel.

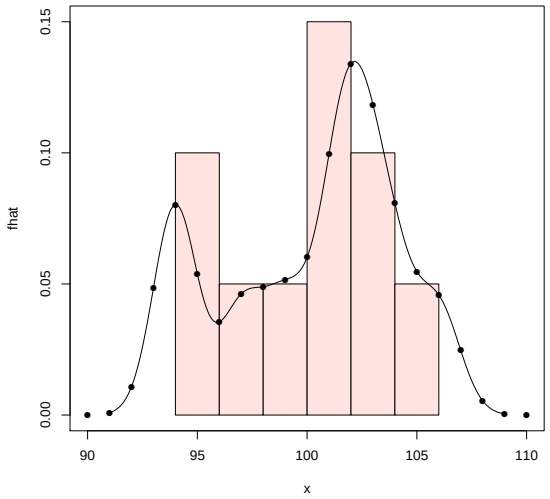


Figure 9.4: Gaussian kernel density estimate of the pdf for the *River Monsters* data, with histogram overlaid.