Exploratory Statistical Data Analysis With R Software (ESDAR)

Swayam Prabha

Lecture 16

Kernel Density and Stem–Leaf Plots

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Kernel Density Plots

In histogram, the continuous data is artificially categorized.

Choice and width of class interval is crucial in the construction of histogram.

Other option is kernel density plot.

It is a smooth curve and represents data distribution.

Kernel Density Plots or Density Plots

Density plots are like smoothened histograms.

The smoothness is controlled by a parameter called bandwidth.

Density plot visualises the distribution of data over a continuous interval or time period.

Density plot is a variation of a histogram that uses kernel smoothing to smoothen the plots by smoothing out the noise.

Kernel Density Plots or Density Plots

Peaks of a density plot display where values are concentrated over the interval.

Density Plots are better to determine the distribution shape than histograms because they're not affected by the number of bins used.

Density plots use a *kernel density estimate*.

Kernel Density Plots

A kernel density plot is produced by the function

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h} \right), h > 0$$

n : **sample size**

h : **bandwidth**

K : **kernel function**

Different choices of *K* provides different estimates.

Kernel functions are not arbitrarily defined but they satisfy the conditions as of probability density function.

Kernel Density Plots

Example, rectangular kernel is

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } -1 \le x \le 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Similarly, Epanechnikov kernel is

$$K(x) = \begin{cases} \frac{3}{4}\left(1 - x^2\right) & \text{if } |x| < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Kernel based on normal distribution is called "Gaussian Kernel". This is the default kernel in R software.

Kernel Density Plots or Density Plots

Example: **Height of 50 persons in centimetres are recorded as follows:**

166,125,130,142,147,159,159,147,165,156,149,164,137,166,135,142,

133,136,127,143,165,121,142,148,158,146,154,157,124,125,158,159,

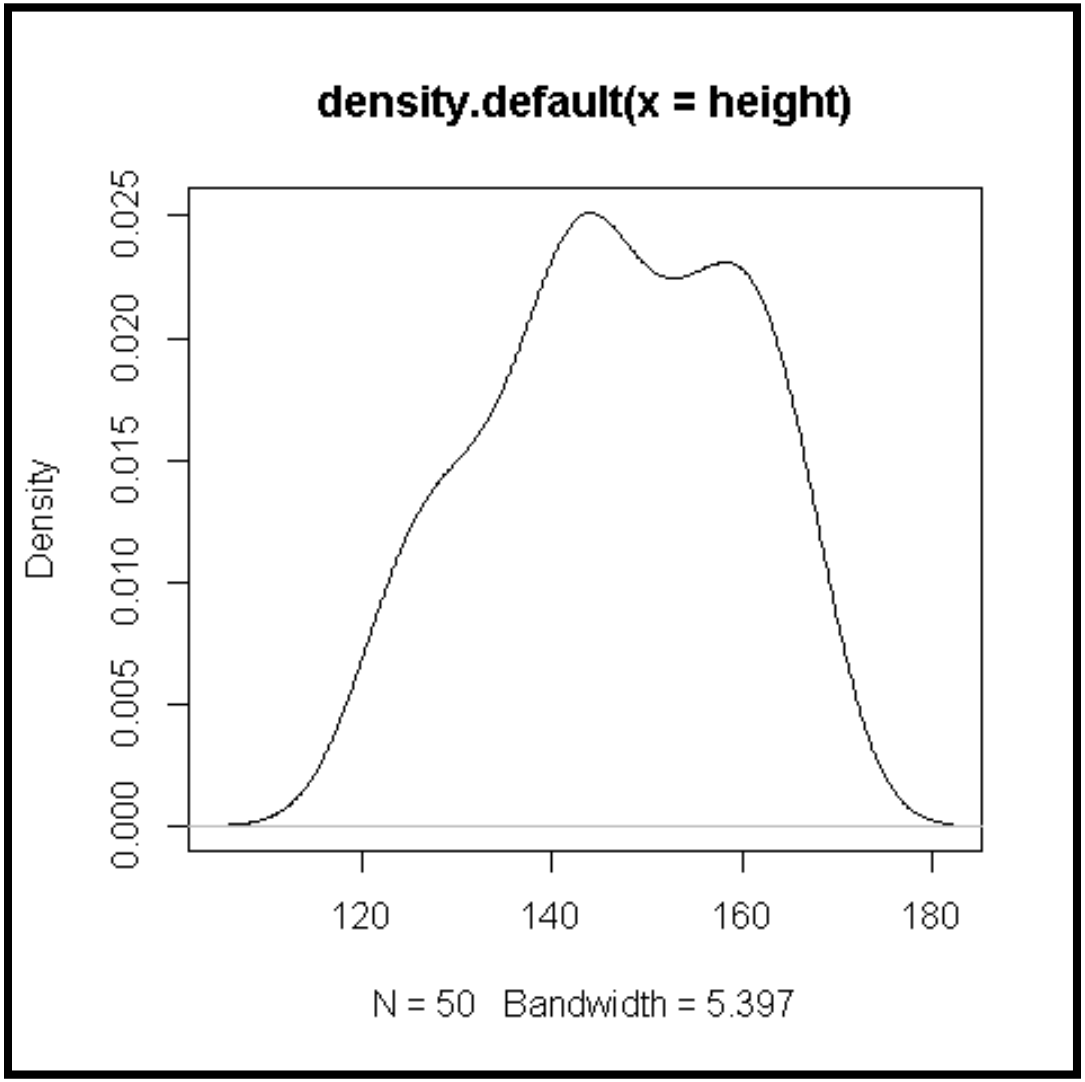
164,143,154,152,141,164,131,152,152,161,143,143,139,131,125,145,

140,163

```
> height <-c(166,125,130,142,147,159,159,147,
165,156,149,164,137,166,135,142,133,136,127,143,
165,121,142,148,158,146,154,157,124,125,158,159,
164,143,154,152,141,164,131,152,152,161,143,143,
139,131,125,145,140,163)
```

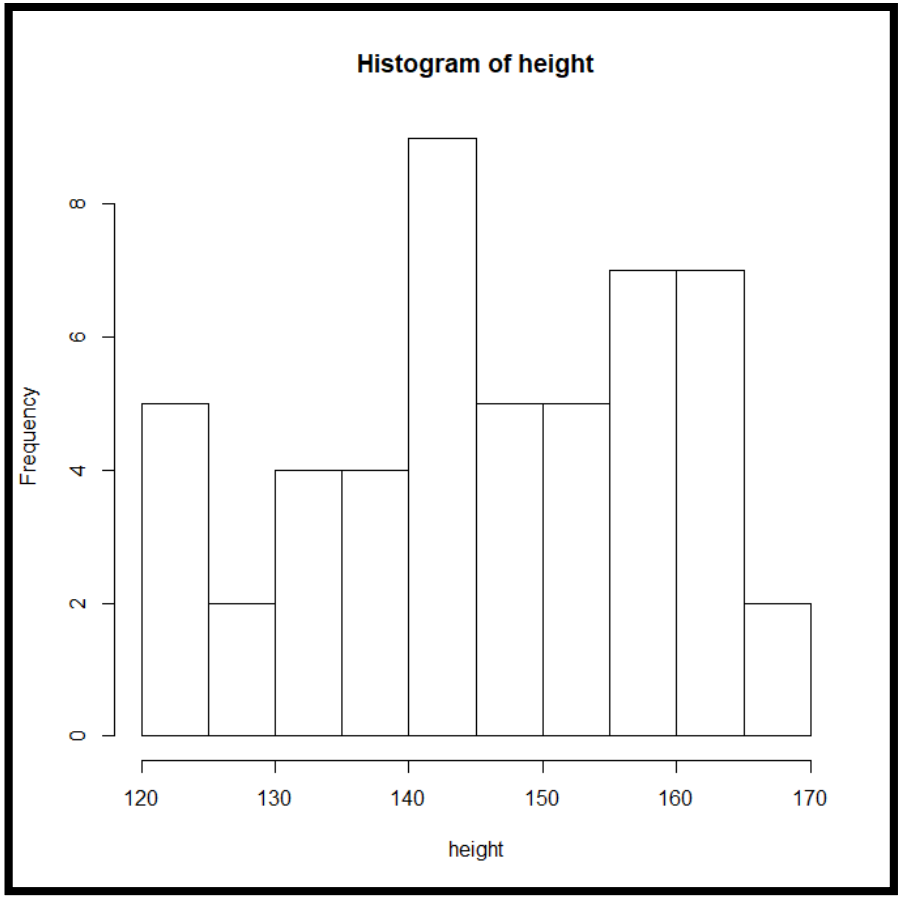
Kernel Density Plots or Density Plots

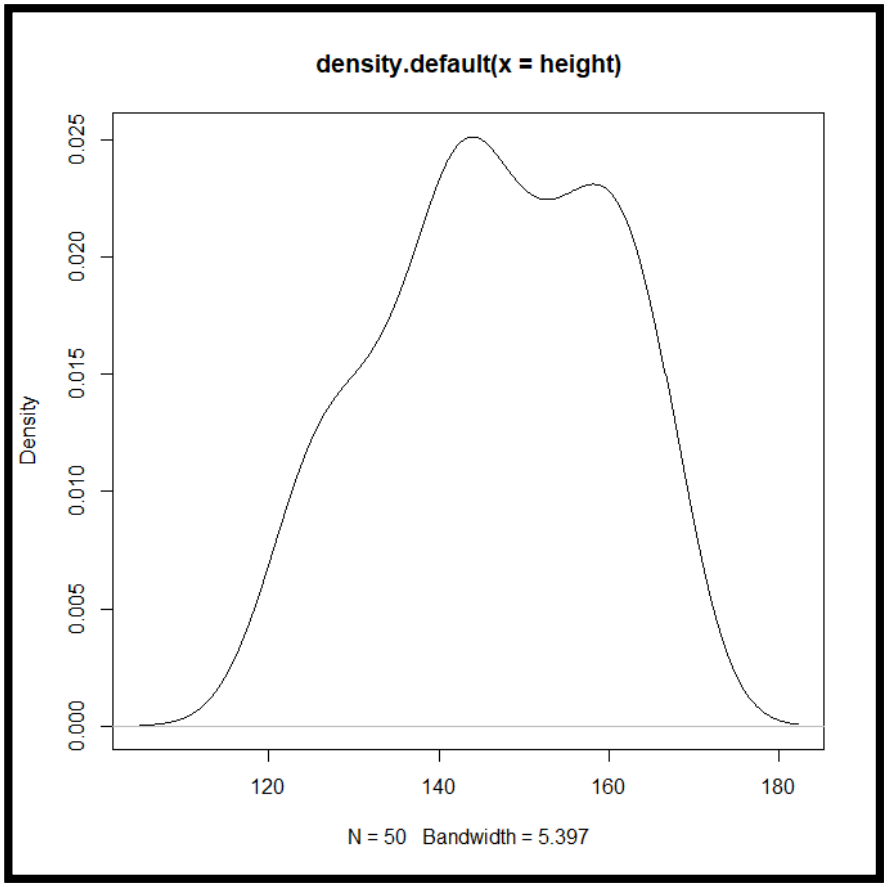
```
> plot(density(height)) #Default Gaussian kernel
```



Histogram vs. Density Plot

Example: Comparison of histogram and density plot for the same data



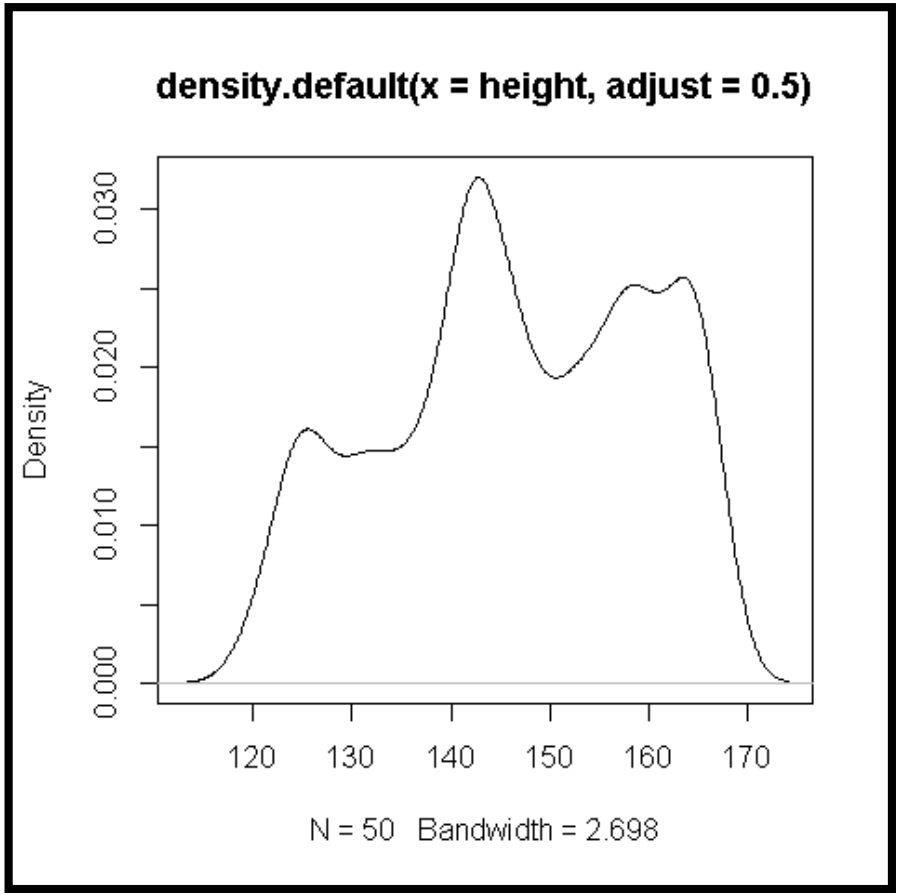


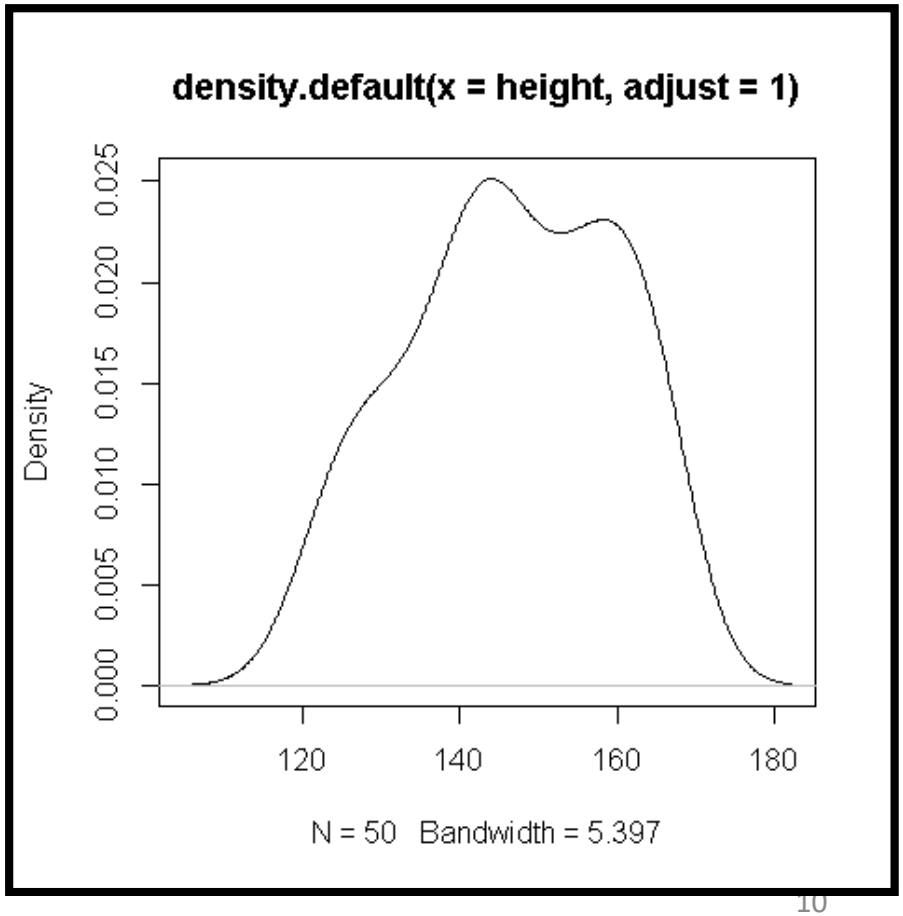
Kernel Density Plots or Density Plots

Example: Use of adjust

```
> plot(density(height,
adjust=0.5))
```

```
> plot(density(height,
adjust=1))
```

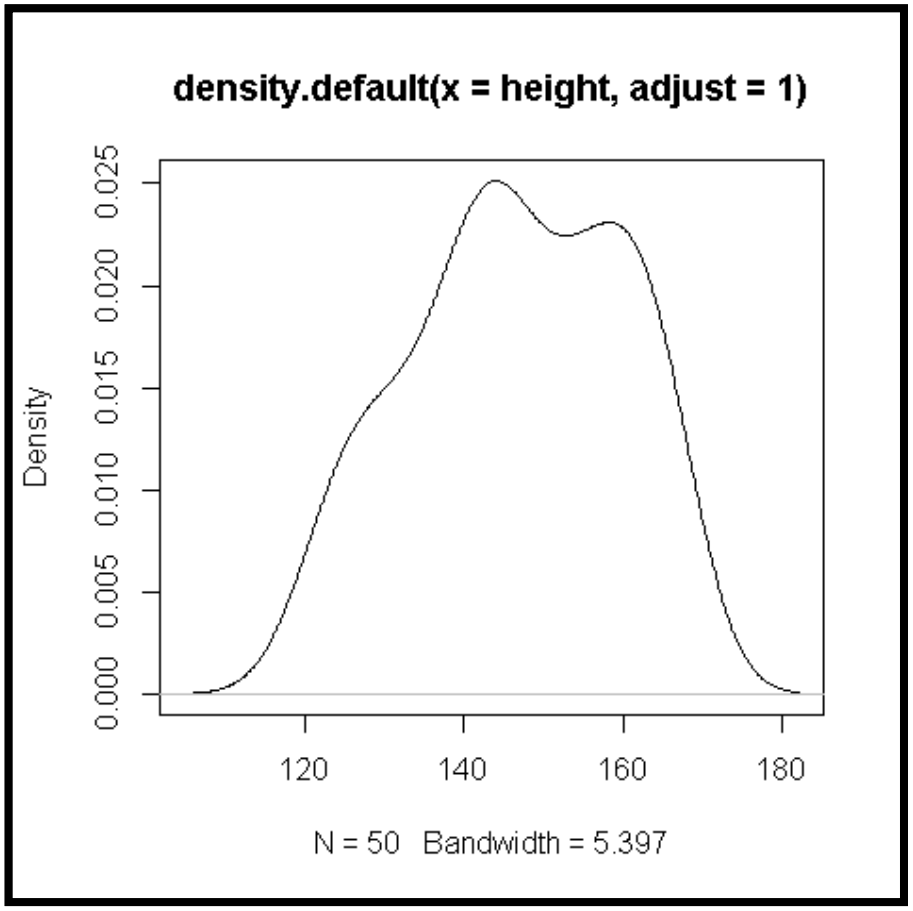




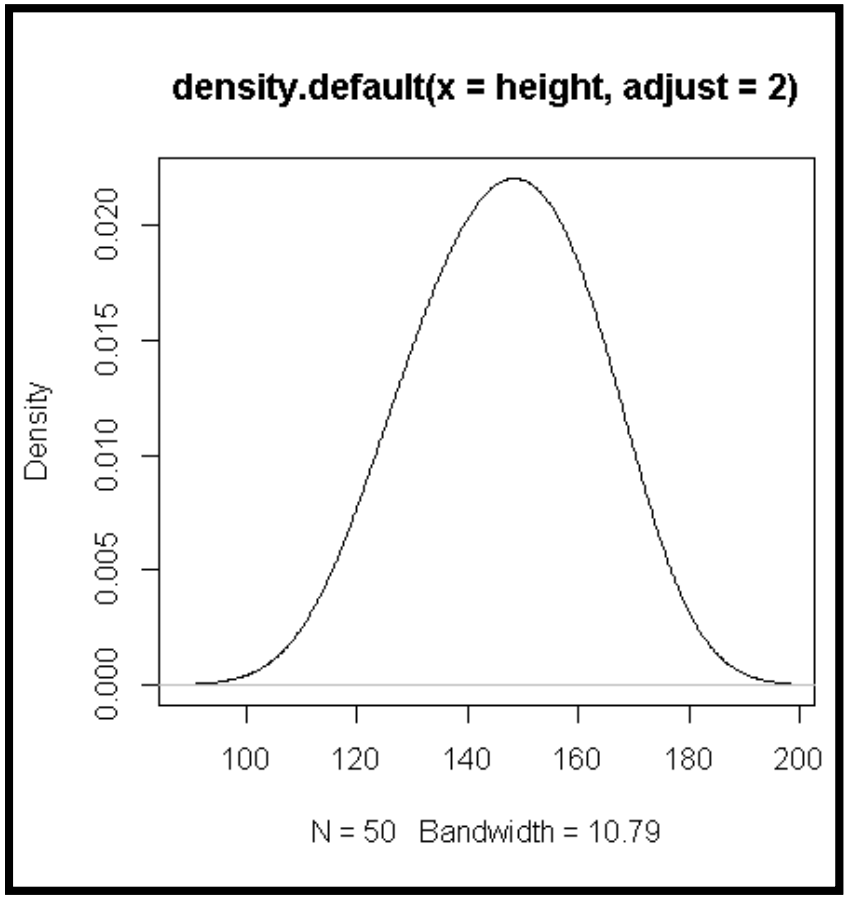
Kernel Density Plots or Density Plots

Example: Use of adjust

```
> plot(density(height,
adjust=1))
```

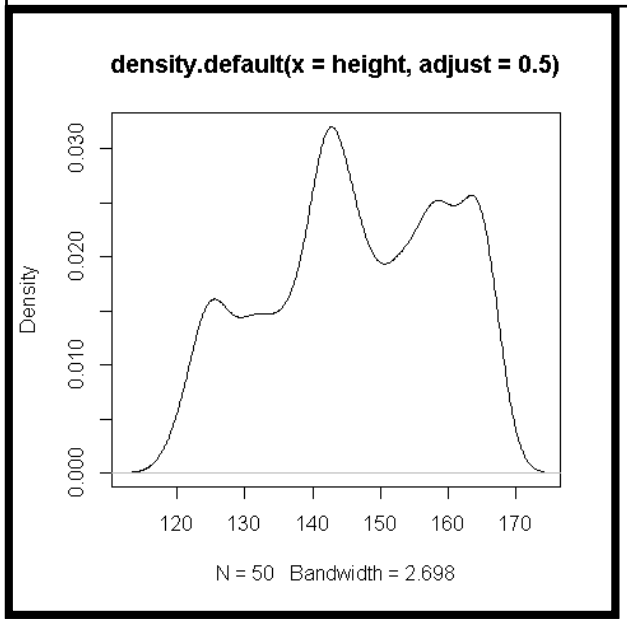
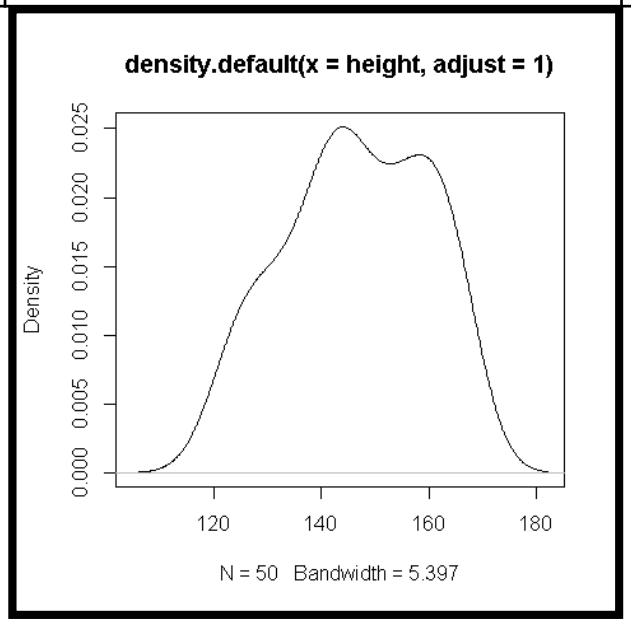
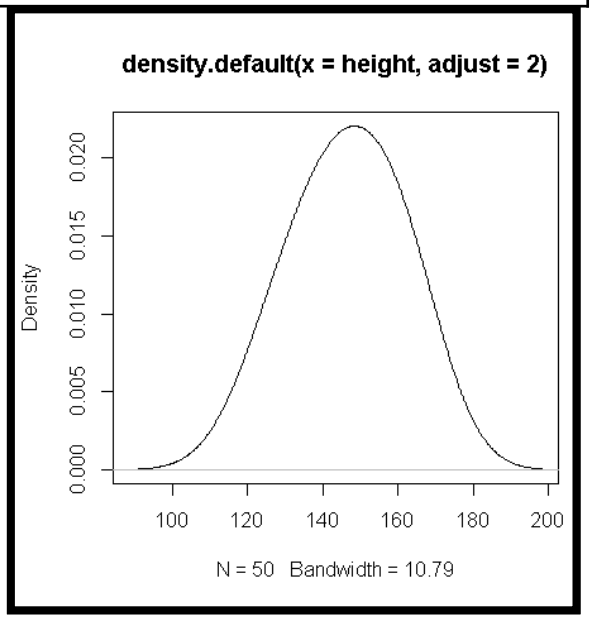



```
> plot(density(height,
adjust=2))
```



Kernel Density Plots or Density Plots

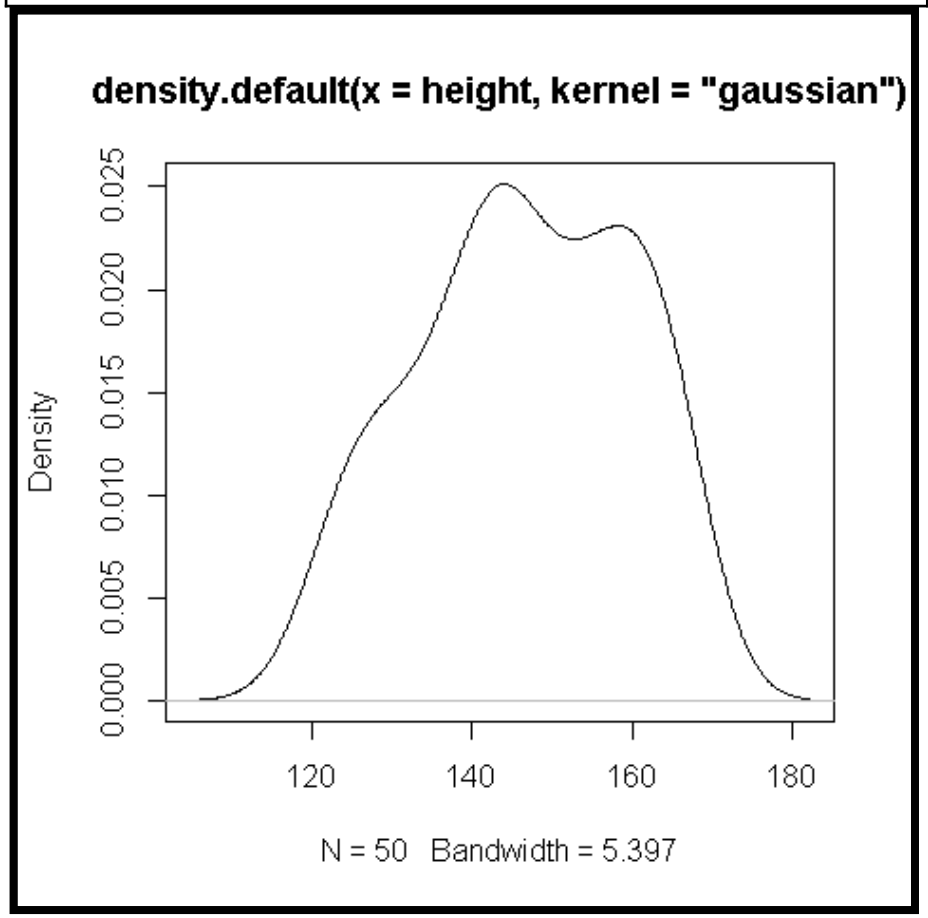
Example: Use of adjust

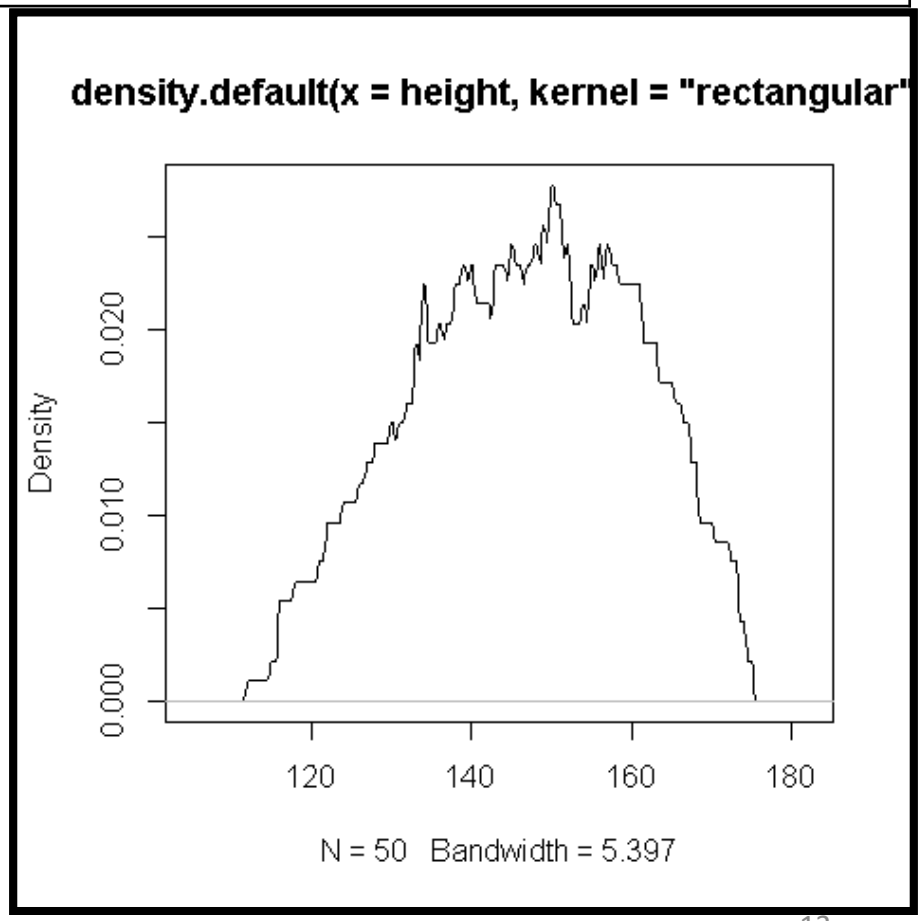
`plot(density(height, adjust=0.5))`	`plot(density(height, adjust=1))`	`plot(density(height, adjust=2))`
		

Kernel Density Plots or Density Plots

Example: Use of different kernels

`>plot(density(height,kernel ='gaussian'))`	`>plot(density(height,kernel ='rectangular'))`

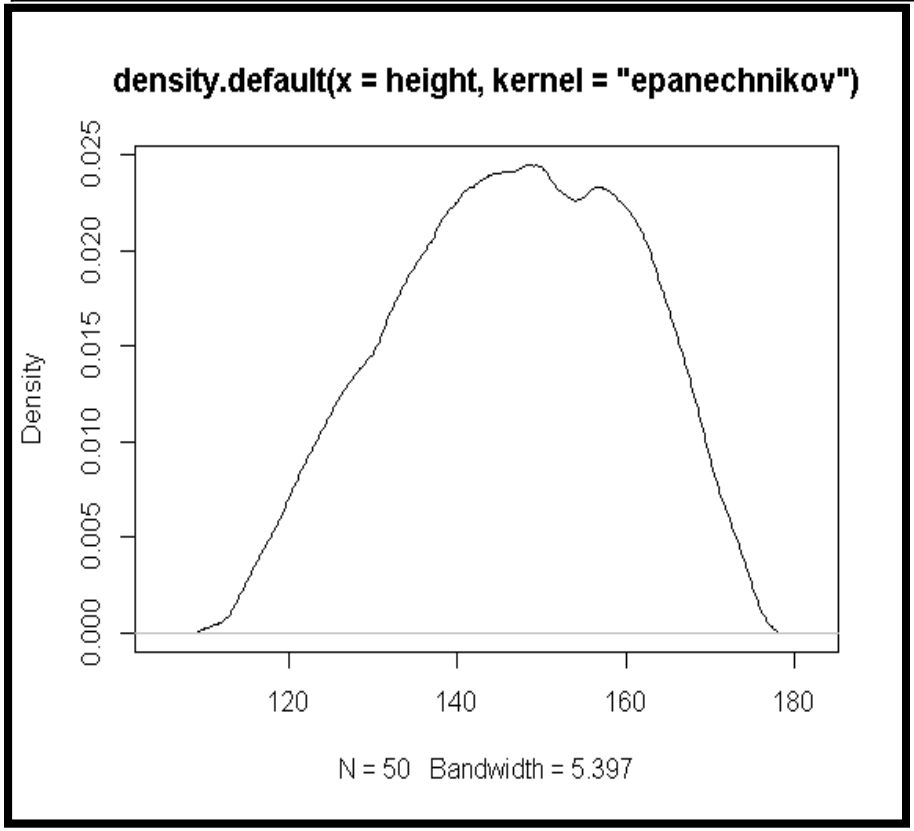




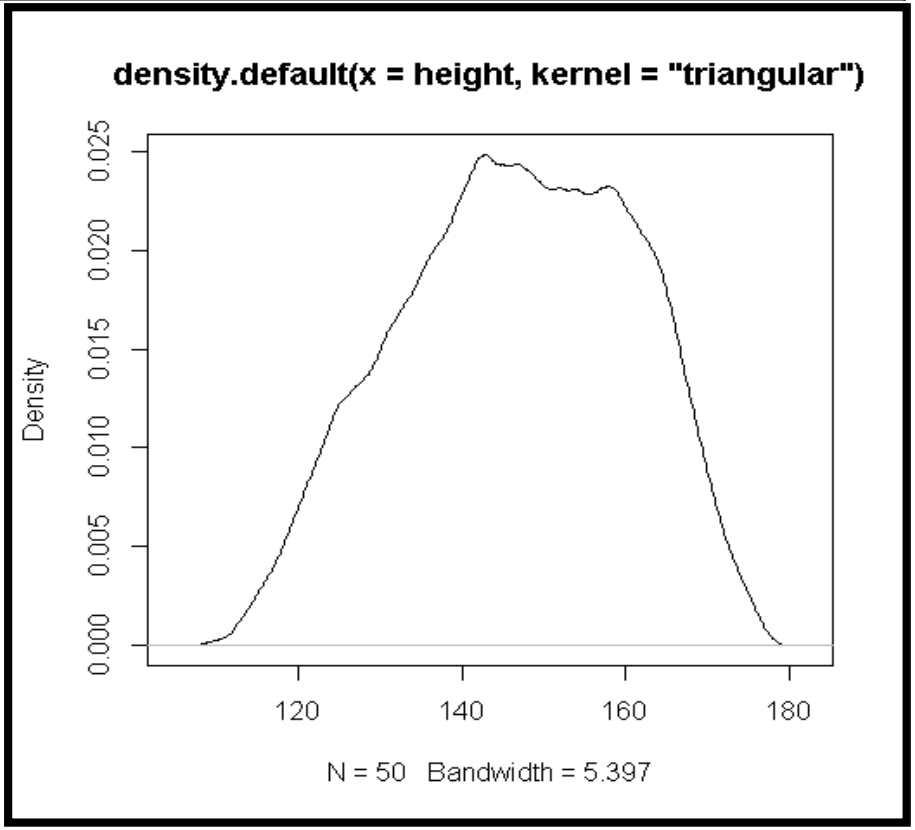
Kernel Density Plots or Density Plots

Example: Use of different kernels

```
> plot(density(height,
kernel='epanechnikov'))
```




```
> plot(density(height,
kernel='triangular'))
```



Stem-and-Leaf Plots

Stem-and-leaf plots show the absolute frequency in different classes like frequency distribution table or a histogram.

Stem-and-leaf plot also present the same information.

Stem-and-leaf plot of a quantitative variable is a textual graph that presents the data according to their most significant numeric digit.

More suitable for small datasets.

Stem-and-Leaf Plots

Stem-and-leaf plot is a sort of tabular presentation where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit).

Interpretations:

“56" is split into "5" (stem) and "6" (leaf)

Stem “2" Leaf “8" means 28

Stem-and-Leaf Plots

To make a stem-and-leaf plot,

1. **separate each observation into a stem consisting of all but the final (rightmost) digit and a leaf , the final digit.**
2. **Stem may have as many digits as needed but each leaf contains only a single digit.**
3. **Write the stem in vertical column with the smallest at the top, and draw a vertical line at the right of this column.**
4. **Write each leaf in the row to the right of its stem, in increasing order out from the stem.**

Stem-and-Leaf Plots

`stem` produces a stem-and-leaf plot of the values in `x`. The parameter scale can be used to expand the scale of the plot.

A value of scale = 2 will cause the plot to be roughly twice as long as the default.

Usage

```
stem(x, scale = 1, width = 80)
```

`x`	a numeric vector.
`scale`	Controls the plot length.
`width`	Controls desired width of plot.

Stem-and-Leaf Plots

Example

Number of defective items in 15 lots are found to be as follows:

46, 24, 53, 44, 18, 34, 65, 54, 66, 35, 48, 56, 73, 38, 49

```
> defective = c(46, 24, 53, 44, 18, 34, 65, 54,
66, 35, 48, 56, 73, 38, 49)

> defective
 [1] 46 24 53 44 18 34 65 54 66 35 48 56 73 38 49
```

Stem-and-Leaf Plots

Example

```
> defective
 [1] 46 24 53 44 18 34 65 54 66 35 48 56 73 38 49

> stem(defective)
  The decimal point is 1 digit(s) to the right
of the |

  0 | 8

  2 | 4458

  4 | 4689346

  6 | 563
```

Stem-and-Leaf Plots:

Example:

R Console

```
> stem(defective)

  The decimal point is 1 digit(s) to the right of the |

  0 | 8
  2 | 4458
  4 | 4689346
  6 | 563
```

Stem-and-Leaf Plots

Example: Role of scale

```
defective = c(46, 24, 53, 44, 18, 34, 65, 54,
66, 35, 48, 56, 73, 38, 49)
```

`> stem(defective, scale=2)`	`> stem(defective, scale=1)`

`The decimal point is 1 digit(s) to the right of the |`
 	
`1 | 8`
`2 | 4`
`3 | 458`
`4 | 4689`
`5 | 346`
`6 | 56`
`7 | 3`
	`The decimal point is 1 digit(s) to the right of the |`
	
	`0 | 8`
	`2 | 4458`
	`4 | 4689346`
	`6 | 563`

Stem-and-Leaf Plots

Example: Role of scale

R Console

```
> stem(defective, scale = 1)

  The decimal point is 1 digit(s) to the right of the |

  0 | 8
  2 | 4458
  4 | 4689346
  6 | 563
```

R Console

```
> stem(defective, scale = 2)

  The decimal point is 1 digit(s) to the right of the |

  1 | 8
  2 | 4
  3 | 458
  4 | 4689
  5 | 346
  6 | 56
  7 | 3
```

Stem-and-Leaf Plots

Example: Comparison with histogram

```
> stem(defective, scale=2)
```

```
 The decimal point is
1 digit(s) to the
right of the |

  1 | 8
  2 | 4
  3 | 458
  4 | 4689
  5 | 346
  6 | 56
  7 | 3
```

```
> hist(defective)
```