STK335 Analisis Eksplorasi Data
Pertemuan 05

Pendugaan Bentuk Sebaran (Fungsi Kepekatan)

Bagus Sartono
2019

Outline

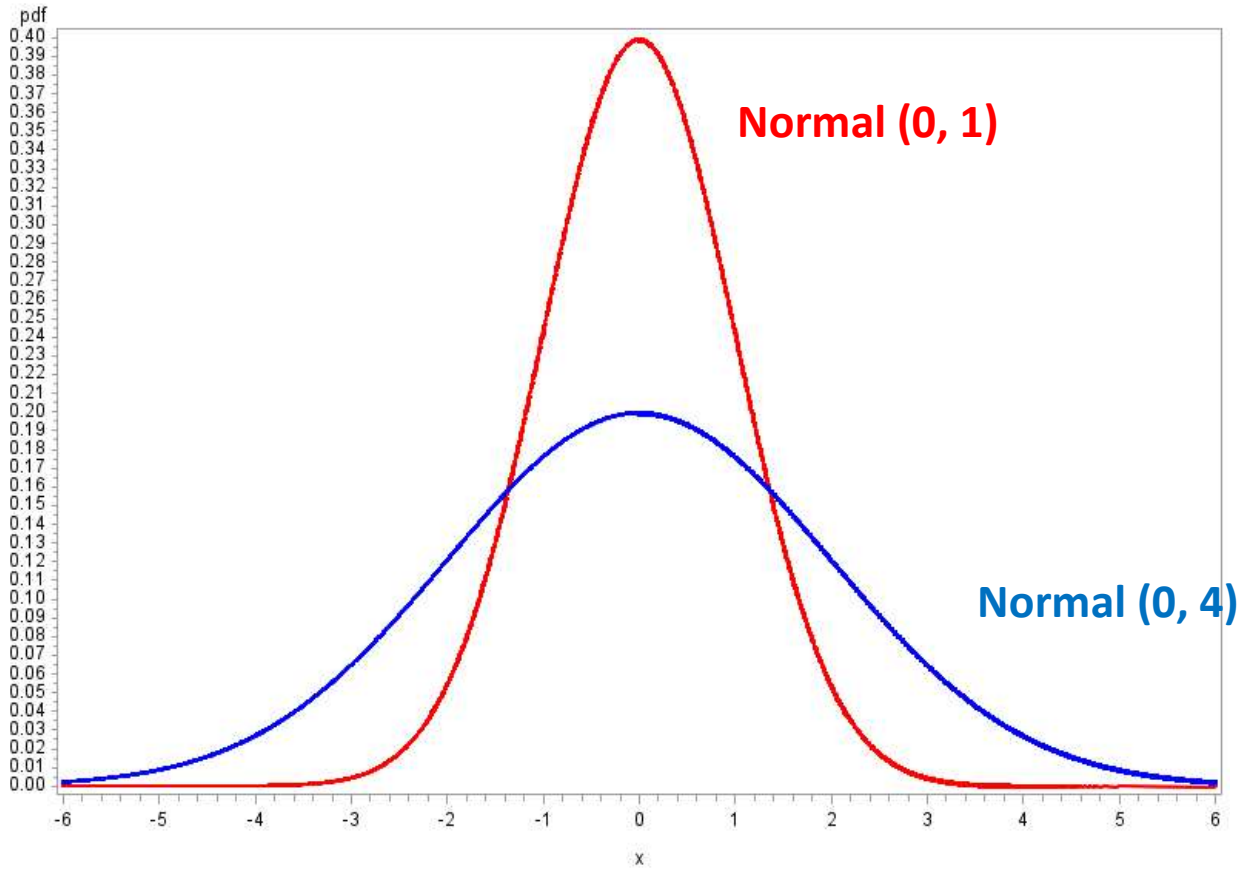
- **Fungsi kepekatan dari variabel kontinu**
- **Apa gunanya diduga?**
- **Metode Pendugaan**
  - Penduga Naïve
  - Penduga Kernel

Fungsi Kepekatan

- Masih ingat sifat-sifat fungsi kepekatan?

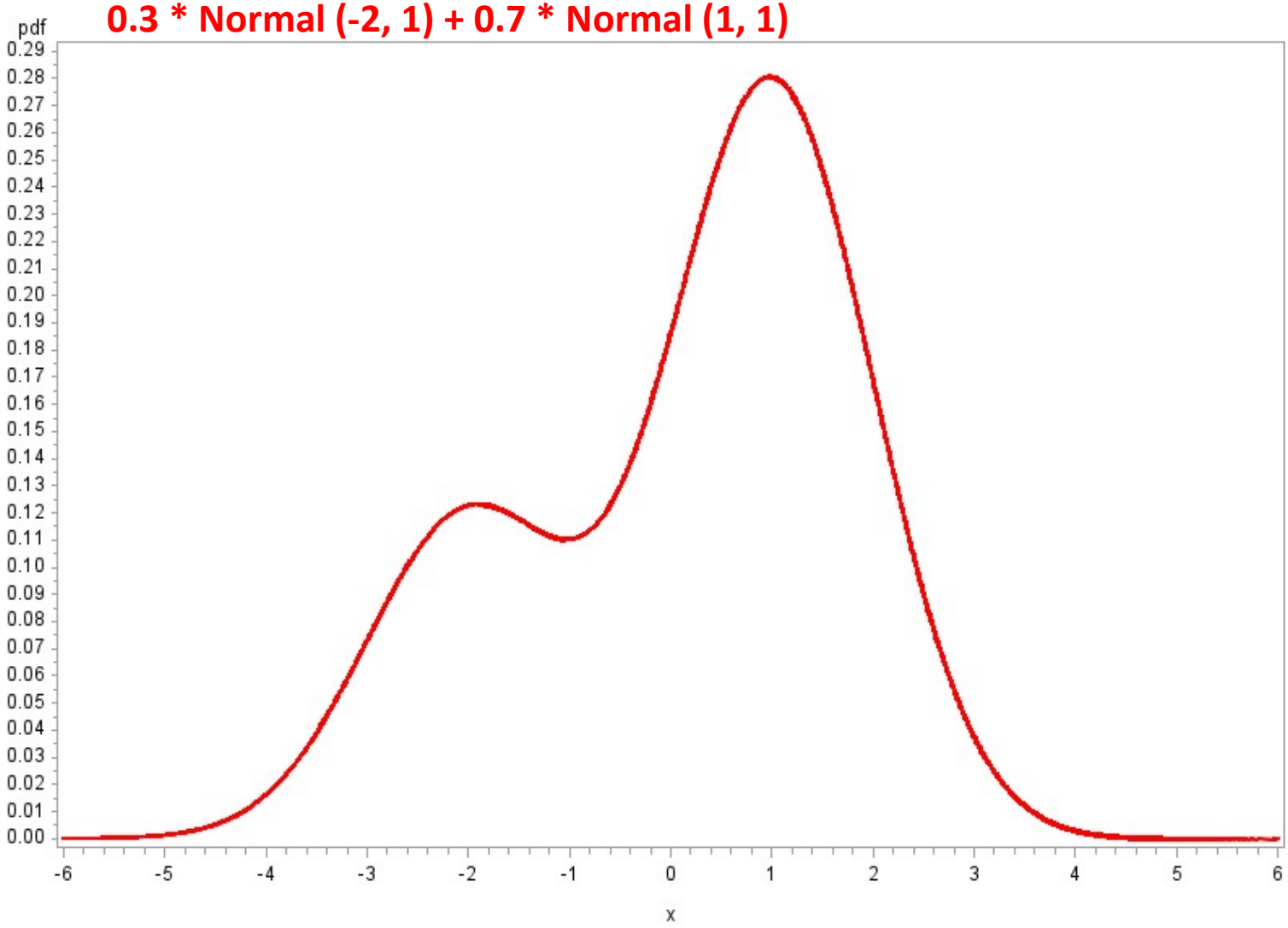
Fungsi Kepekatan

- Masih ingat sifat-sifat fungsi kepekatan?
  - Non Negatif
  - Integral (luas di bawah kurvanya) = 1


pdf
0.40
0.39
0.38
0.37
0.36
0.35
0.34
0.33
0.32
0.31
0.30
0.29
0.28
0.27
0.26
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
x
Normal (0, 1)
Normal (0, 4)

Normal Campuran

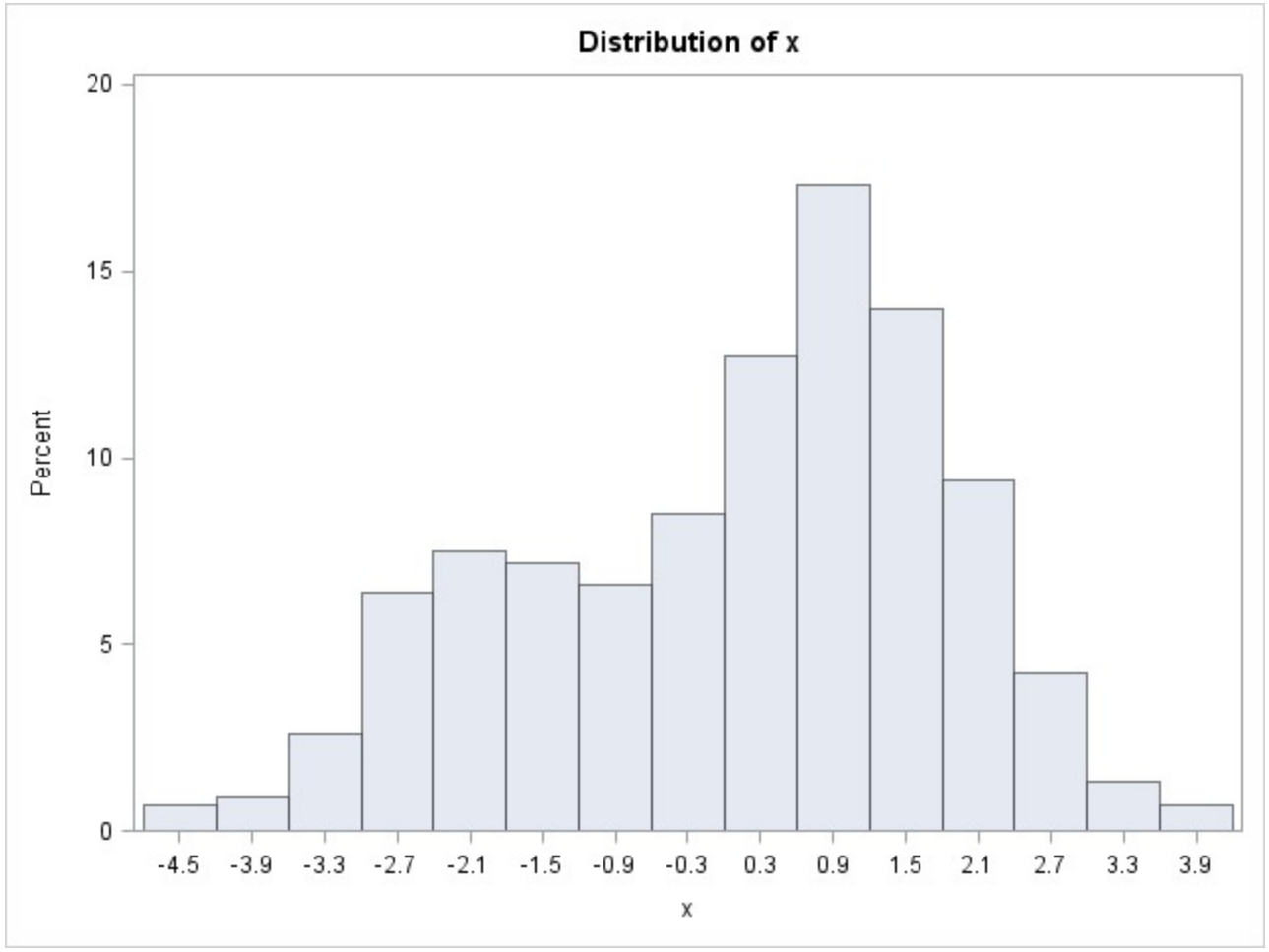
0.3 * Normal (-2, 1) + 0.7 * Normal (1, 1)

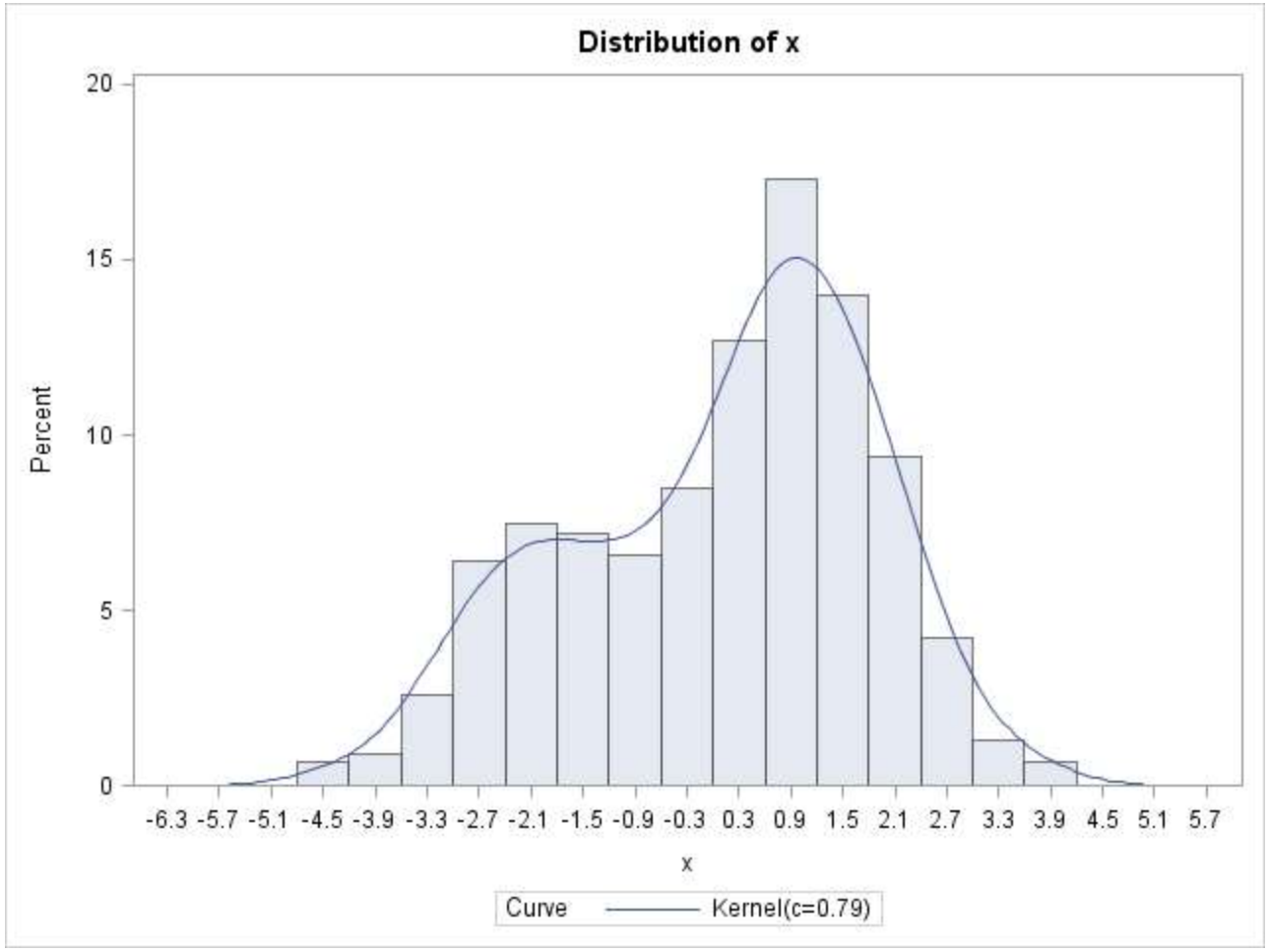


Kegunaan Menduga Fungsi Kepekatan

- Memahami bentuk sebaran dengan lebih baik
- Penting dalam proses simulasi statistika
- Berguna dalam beberapa analisis statistika

Histogram


Distribution of x
Percent
20
15
10
5
0
-4.5
-3.9
-3.3
-2.7
-2.1
-1.5
-0.9
-0.3
0.3
0.9
1.5
2.1
2.7
3.3
3.9
x


Distribution of x
Percent
20
15
10
5
0
-6.3 -5.7 -5.1 -4.5 -3.9 -3.3 -2.7 -2.1 -1.5 -0.9 -0.3 0.3 0.9 1.5 2.1 2.7 3.3 3.9 4.5 5.1 5.7
x
Curve
Kernel(c=0.79)

Penduga Naive

$$\hat{f}_h(x) = \frac{1}{2hn}(\text{banyaknya data pada selang } (x-h, x+h))$$

atau

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

dengan

$$w(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1; \\ 0 & \text{otherwise.} \end{cases}$$

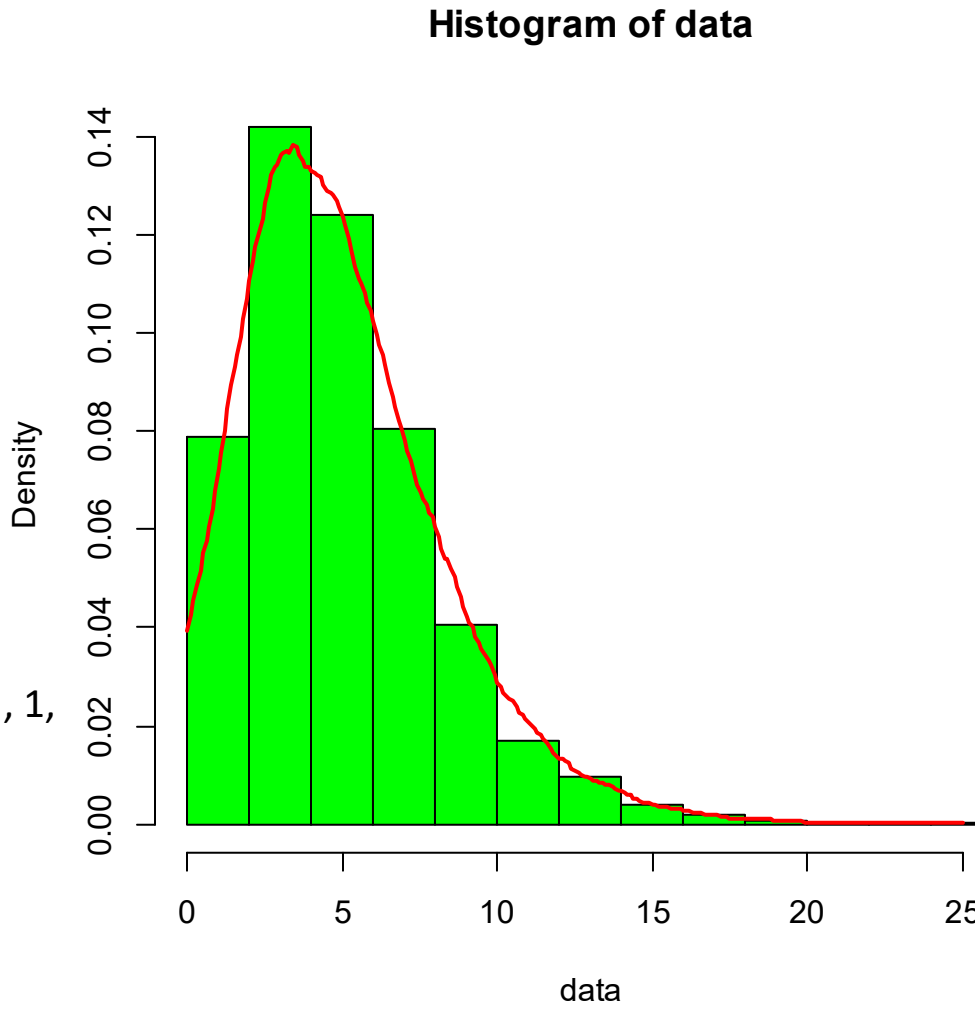
```
data <- rchisq(3000, 5)
hist(data, breaks=10, col="green",
freq=FALSE)

summary(data)

h = 2
x <- seq(0, 25, by=0.1)
f <- 1:length(x)


for (i in 1:length(x)){
f[i] = sum(ifelse( abs((data - x[i])/h) < 1 , 1,
0)) / (2*h*length(data))
}

lines(x, f, type="l", col="red", lwd=2)
```

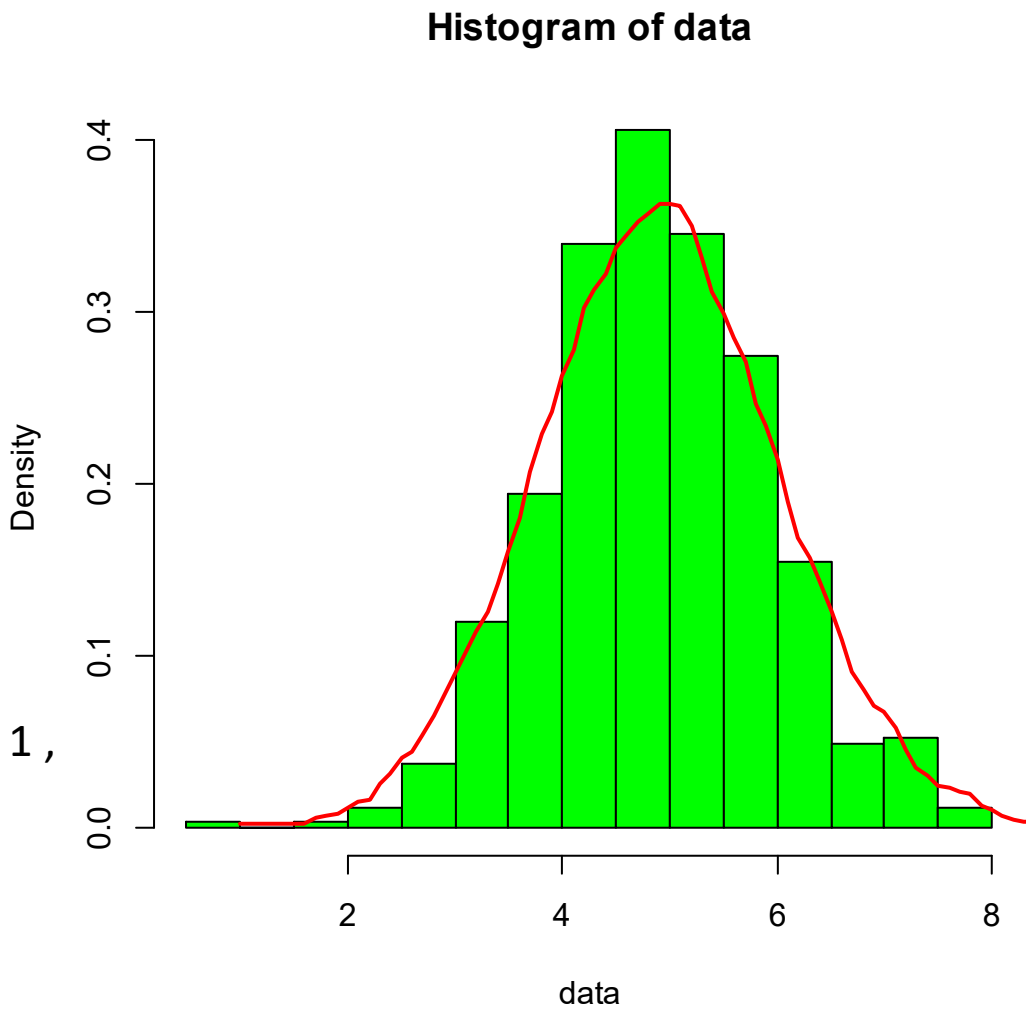


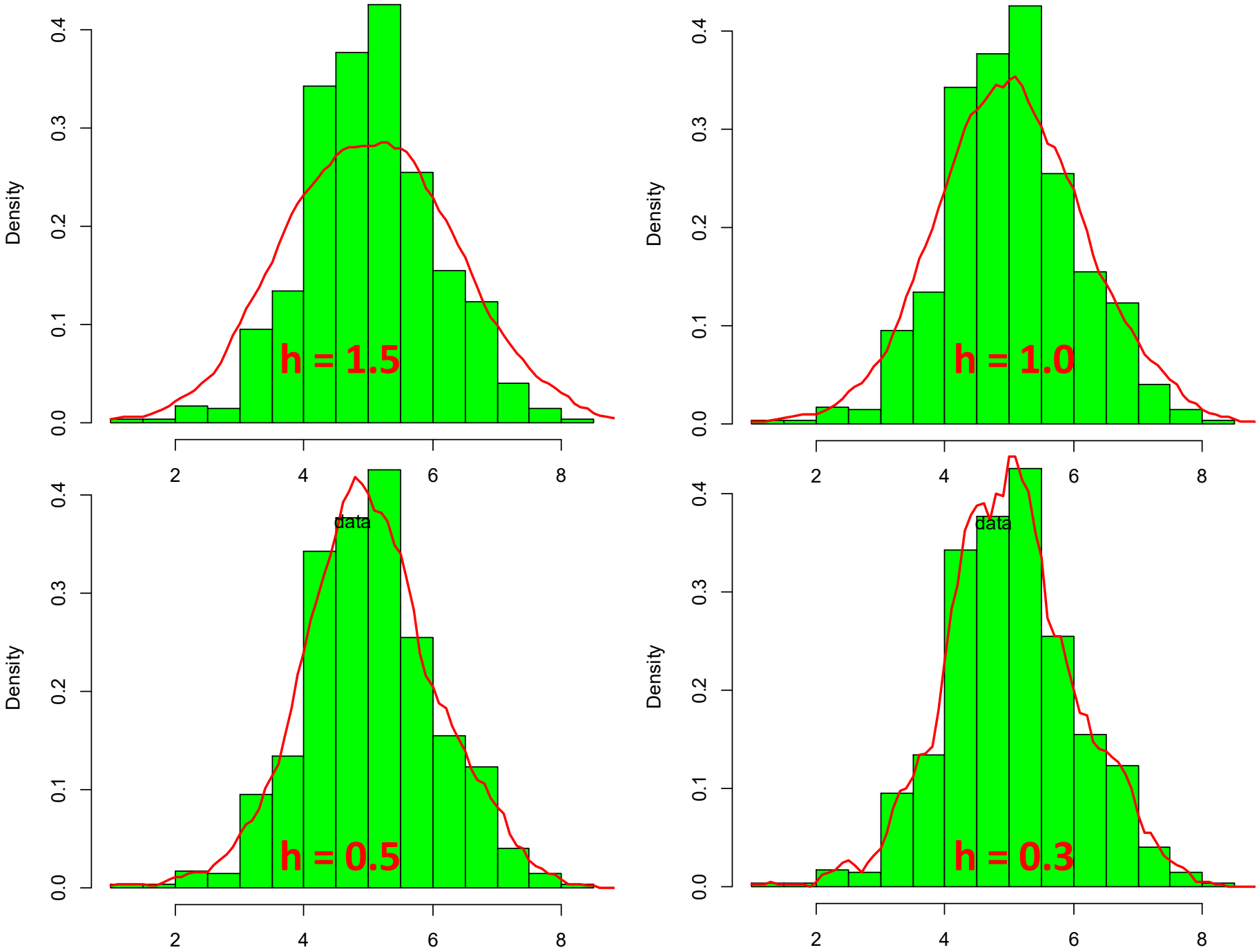
```
data <- rnorm(700, 5, 1)
hist(data, breaks=20, col="green",
freq=FALSE)

summary(data)

h = 0.8
x <- seq(1, 10, by=0.1)
f <- 1:length(x)


for (i in 1:length(x)){
f[i] = sum(ifelse( abs((data - x[i])/h) < 1 ,
1, 0)) / (2*h*length(data))
}
lines(x, f, type="l", col="red", lwd=2)
```




Density
0.0
0.1
0.2
0.3
0.4
2
4
6
8
h = 1.5
h = 1.0
h = 0.5
h = 0.3
data

Pengaruh Lebar Jendela (h)

- Apa yang Anda simpulkan?
- Berapa lebar jendela yang sebaiknya digunakan?

Penduga Kernel

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

- *K* adalah fungsi yang memenuhi $\int K(x)\,dx = 1$
- K adalah fungsi yang non-negatif
- Fungsi *K* disebut sebagai fungsi *kernel*
- *h* adalah bilangan positif yang biasa disebut sebagai *lebar jendela (bandwidth / window width)*.

Some Commonly Used Kernel Functions

(a) Uniform Kernel $\mathrm{K(t)} = \frac{1}{2}\mathrm{I}(|\,\mathrm{t}\,| \le 1)$

(b) Gaussian kernel $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$

(c) Epanechnikov kernel

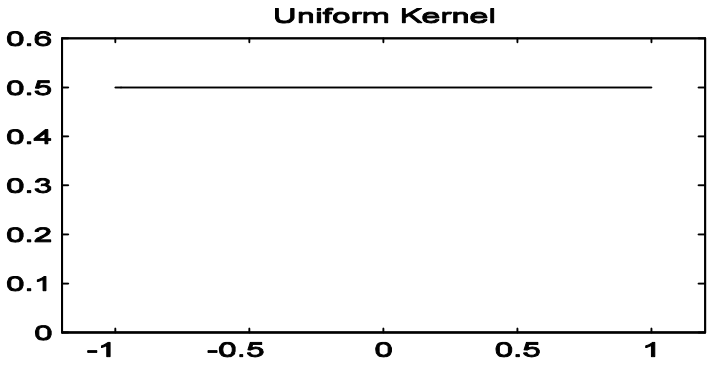
$$K(t) = \begin{cases} \frac{3}{4}(1-t^2), & |\,t\,| < 1 \\ 0 & \text{otherwise} \end{cases}$$

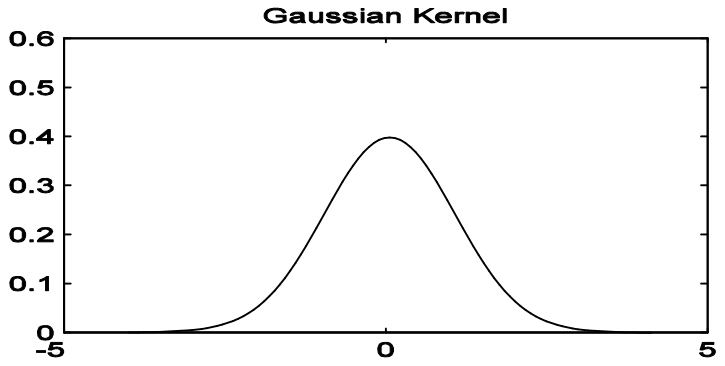
(d) Biweight kernel

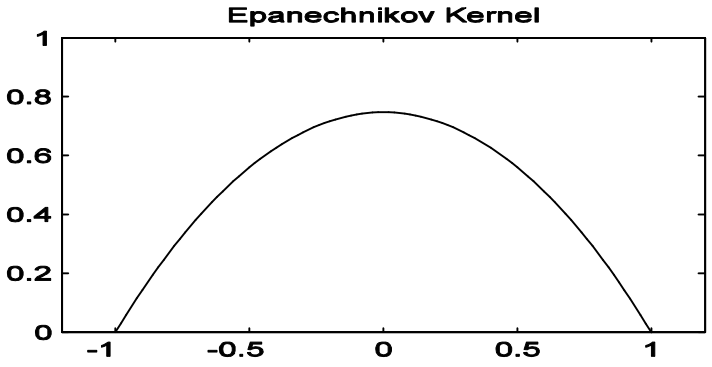
$$K(t) = \begin{cases} \frac{15}{16}(1-t^2)^2, & |\,t\,| < 1 \\ 0 & \text{otherwise} \end{cases}$$

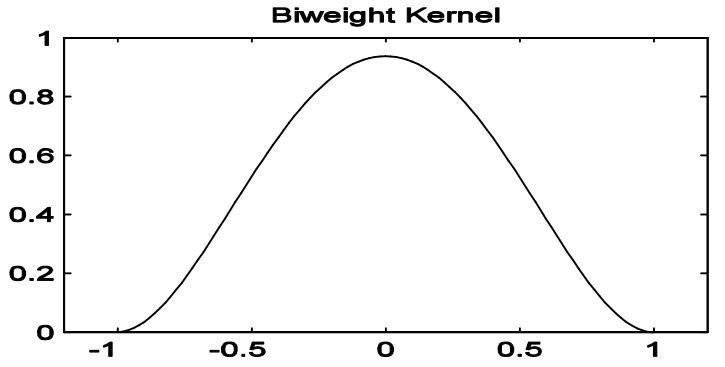
Examples of Kernels

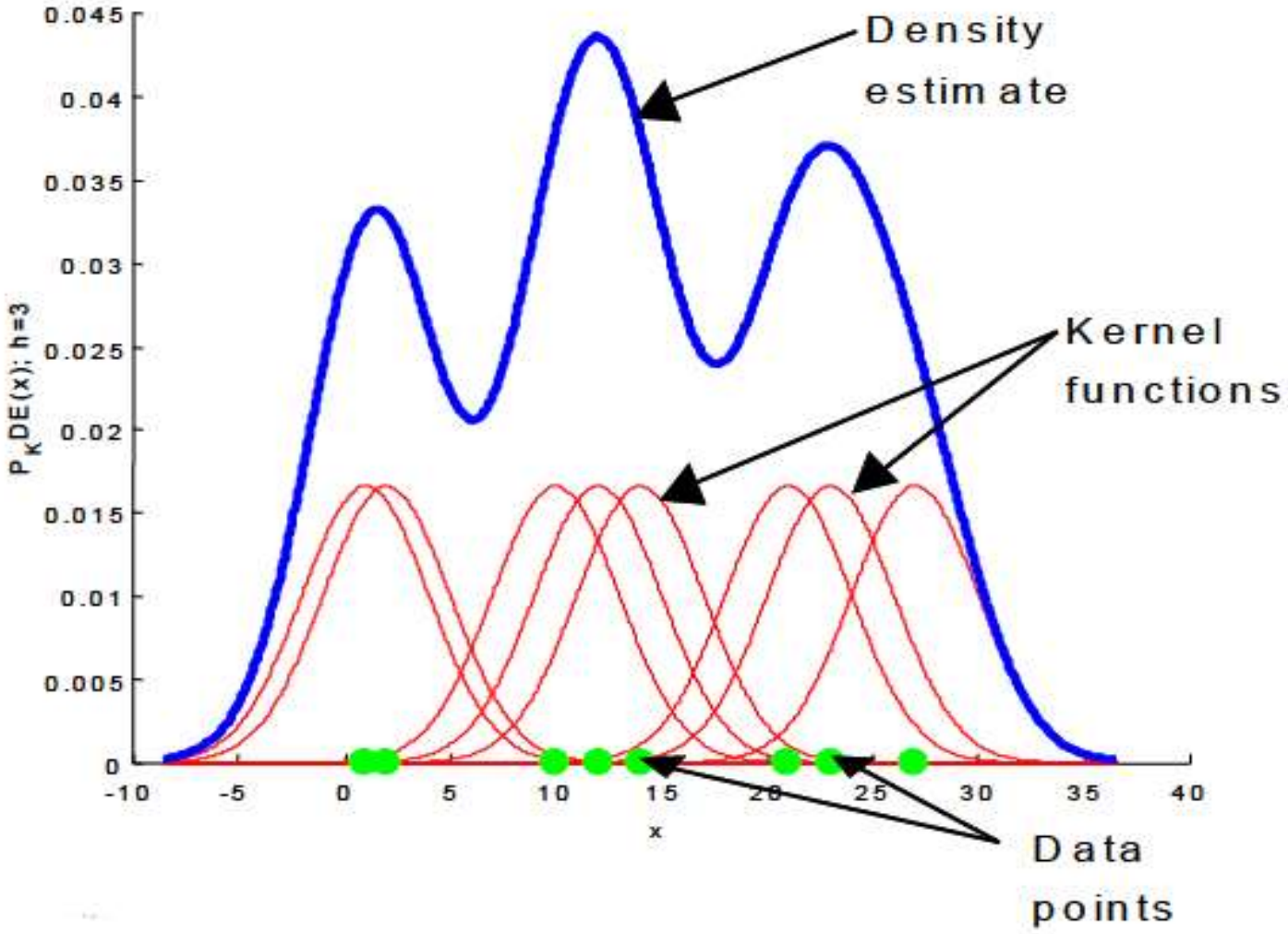
- 4 kernels

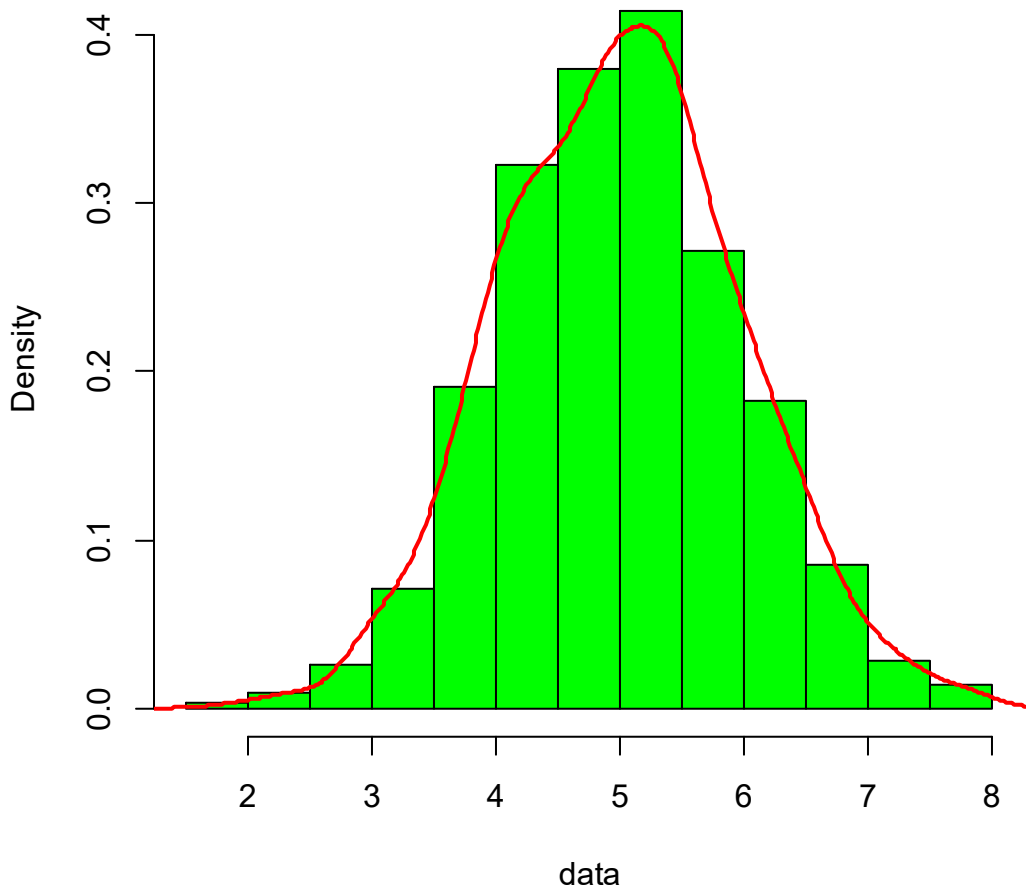









Density
estimate
Kernel
functions
Data
points
$P_K DE(x)$; h=3
x
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
-10
-5
0
5
10
15
20
25
30
35
40

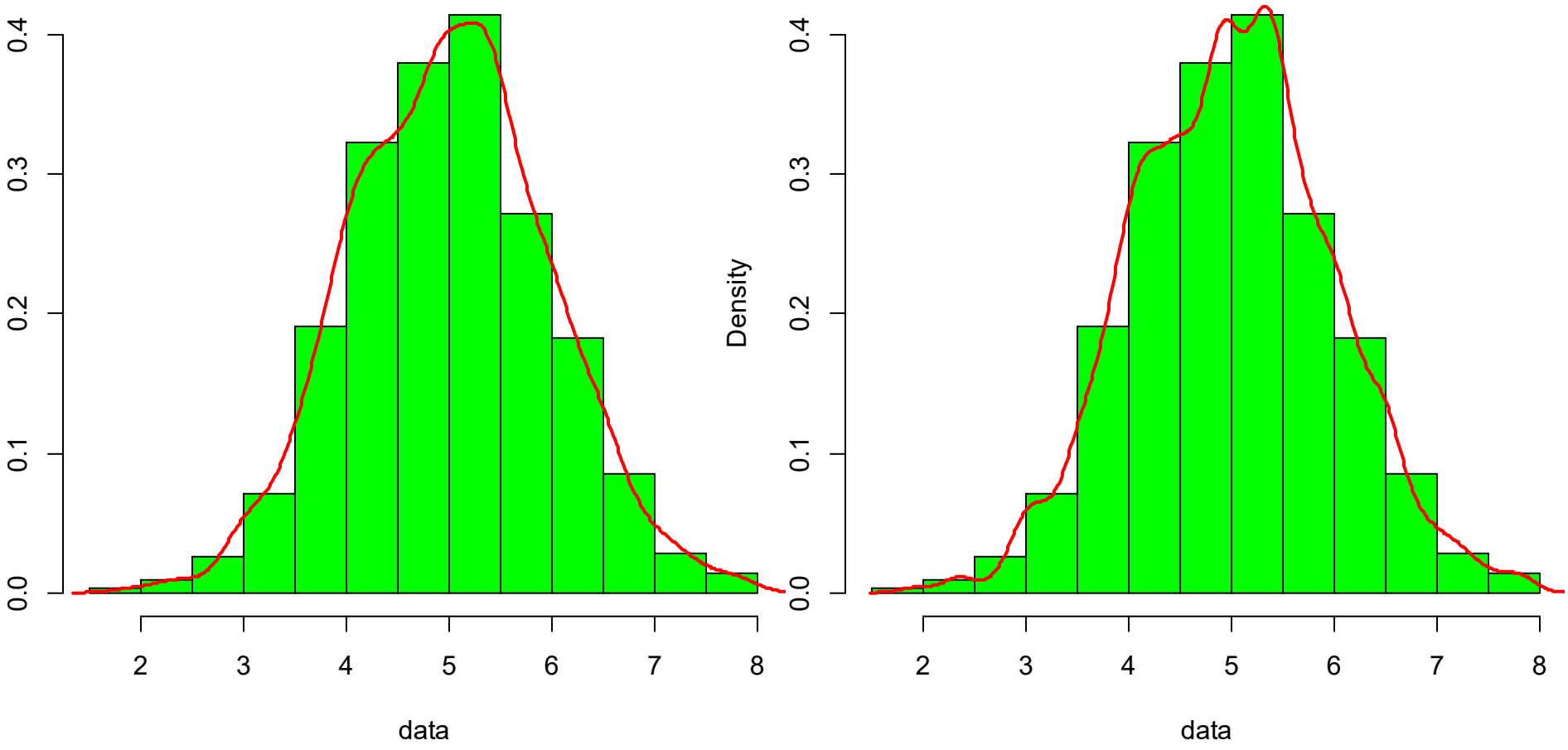



```
data <- rnorm(700, 5, 1)
hist(data, breaks=20, col="green", freq=FALSE, main="")

kepekatan <- density(data)
lines(kepekatan$x, kepekatan$y, type="l", col="red", lwd=2)
```

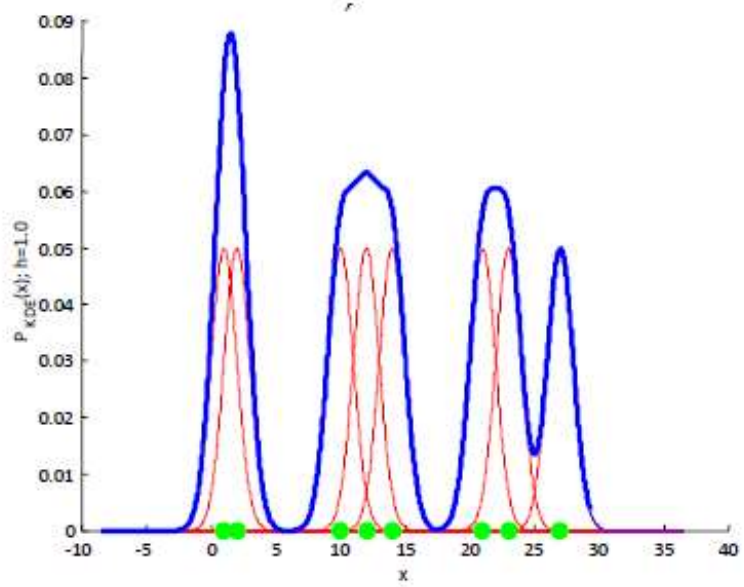
Fungsi di R
(package: stats)

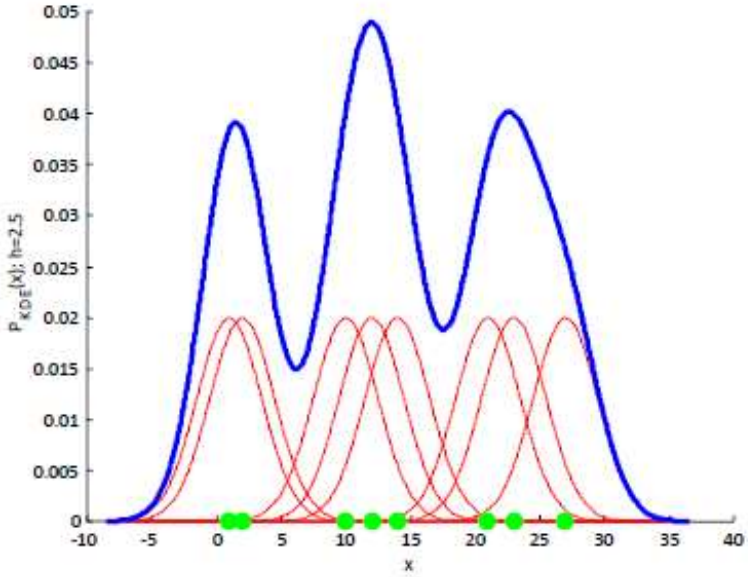
density(x, bw = "nrd0", adjust = 1, kernel = c("gaussian", "epanechnikov", "rectangular", "triangular", "biweight", "cosine", "optcosine"), weights = NULL, window = kernel, width, give.Rkern = FALSE, n = 512, from, to, cut = 3, na.rm = FALSE, ...)

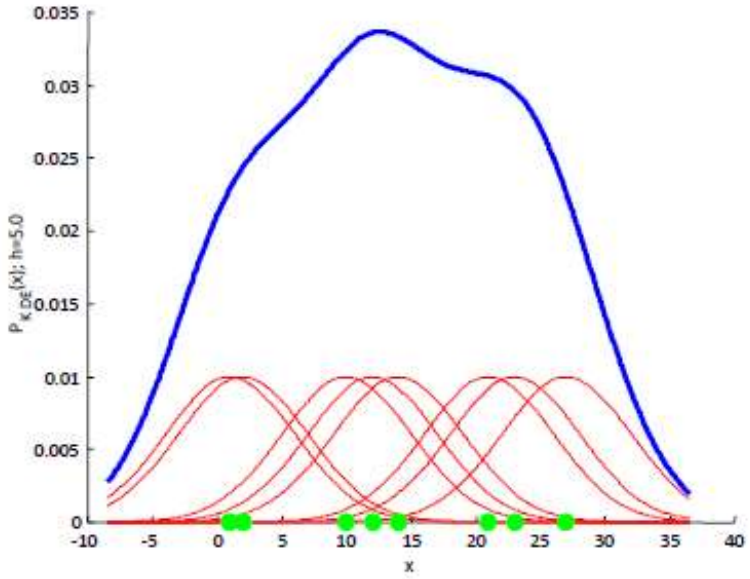


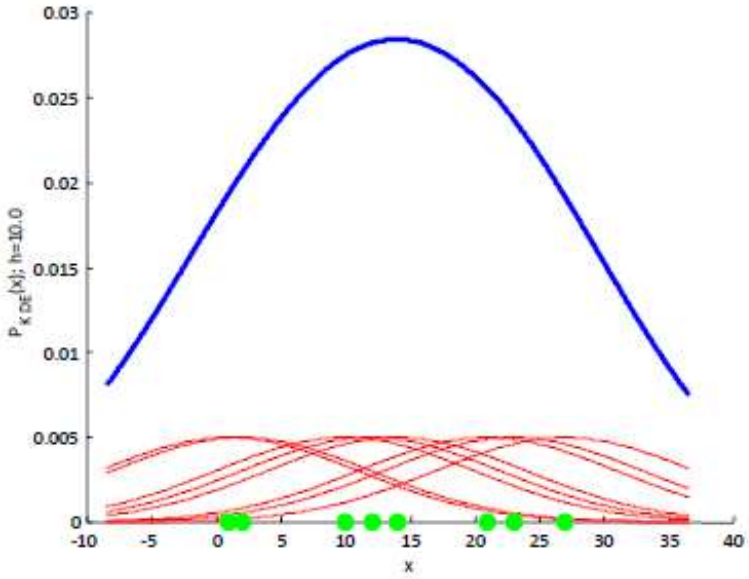
bw = 0.2

bw = 0.15


P_KDE(x); h=1.0
x


P_KDE(x); h=2.5
x


P_KDE(x); h=5.0
x


P_KDE(x); h=10.0
x

proc KDE data=data;
univar x / plots=histdensity unistats;
run;

Univariate Statistics	
	x
Mean	0.10
Variance	3.00
Standard Deviation	1.73
Range	8.82
Interquartile Range	2.61
Bandwidth	0.32