



6. Medidas de assimetria e curtose

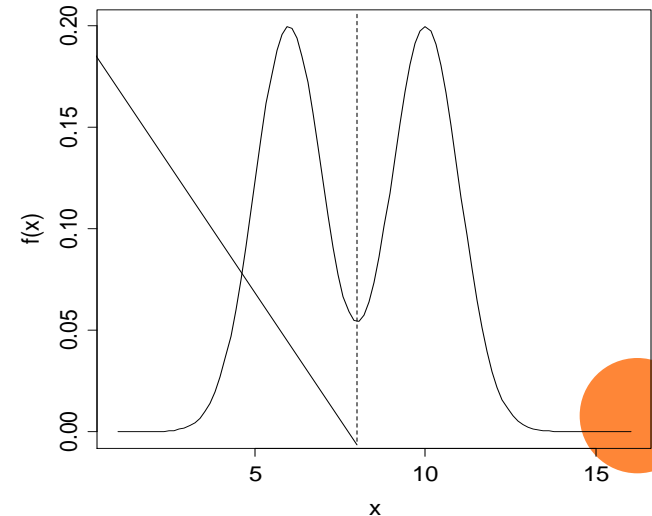
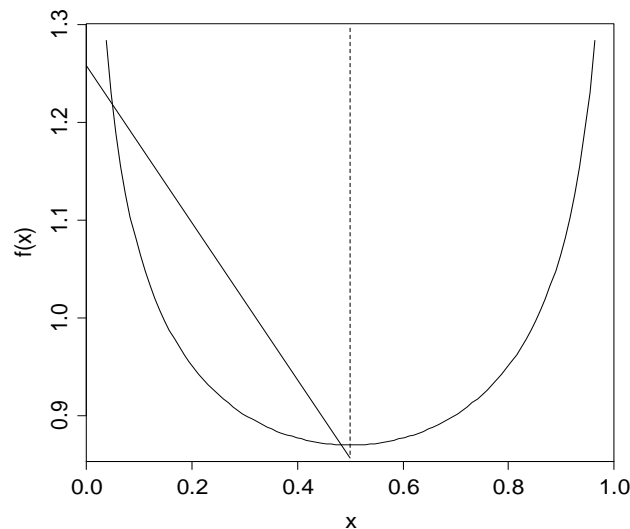
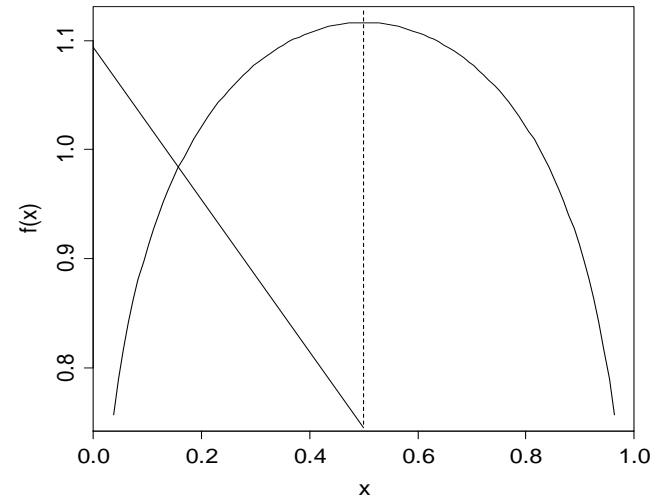
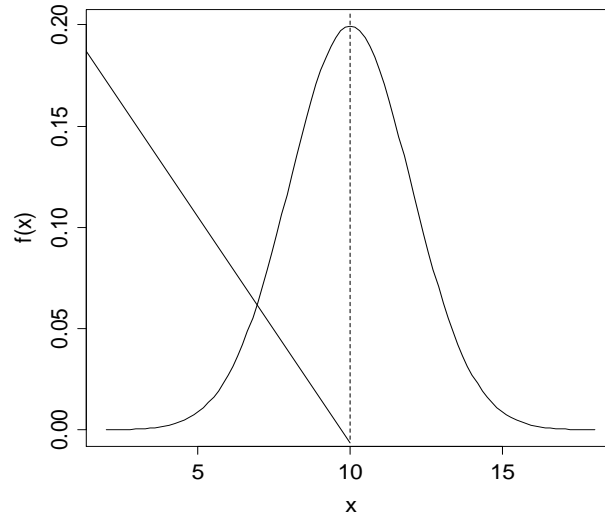
2011



6.1. Medidas de assimetria

Uma **variável aleatória contínua** X tem distribuição **simétrica** (*symmetric*) em relação a um valor x_0 se $f(x_0 - a) = f(x_0 + a)$, para todo a .

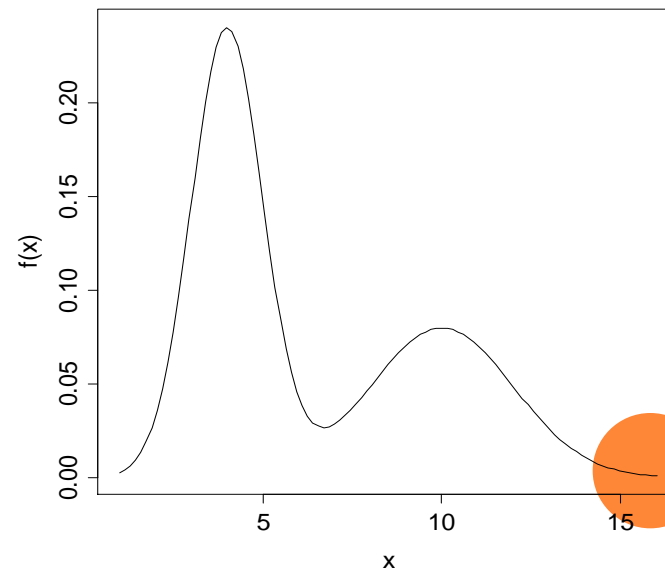
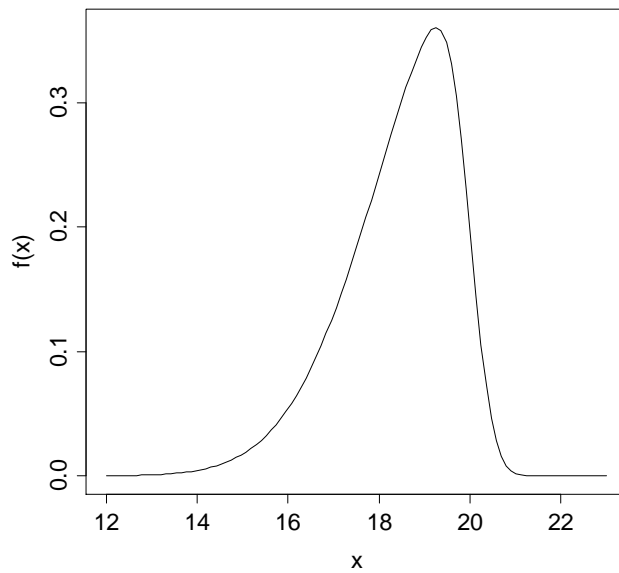
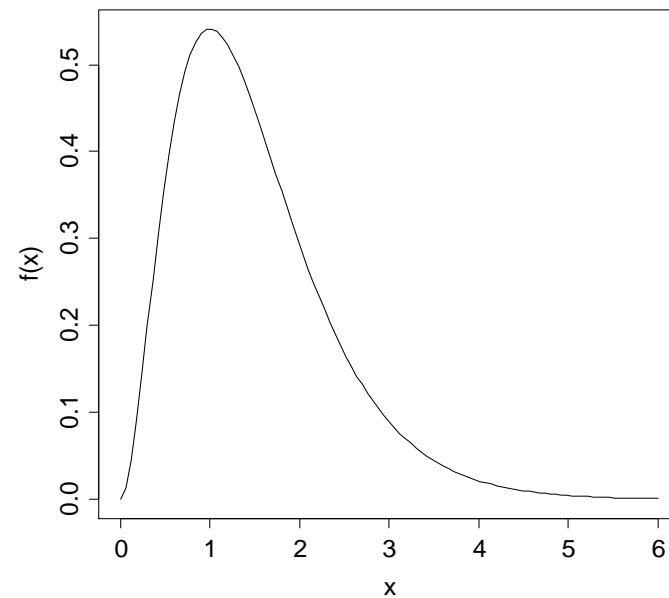
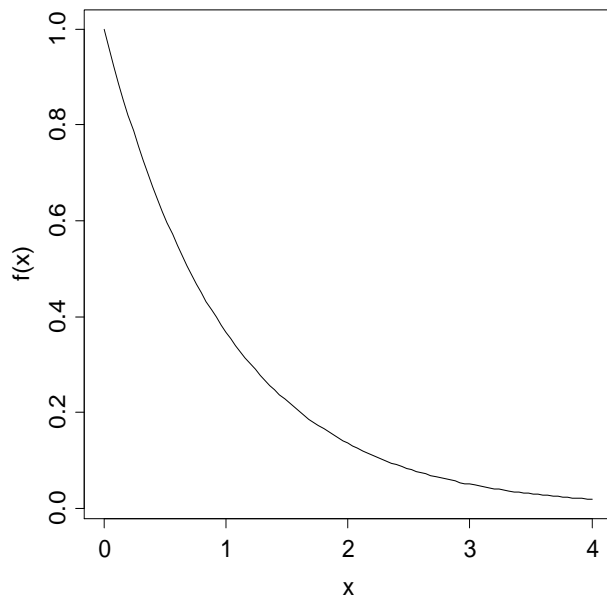
Distribuições
simétricas:



Obs. $M = x_0$.

Moda e média
não
necessariamente
são iguais a x_0 .

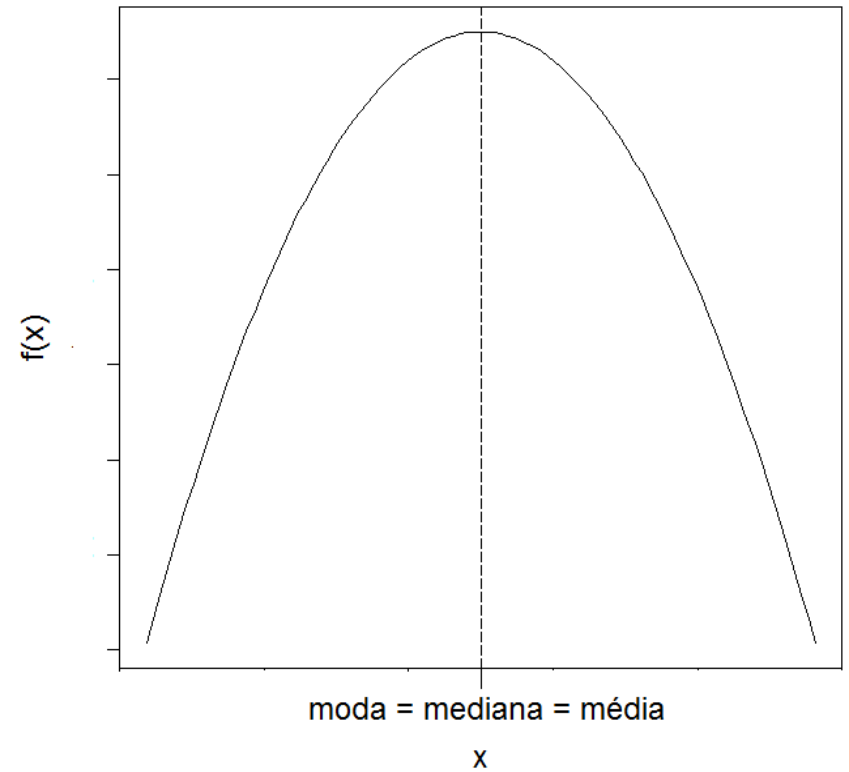
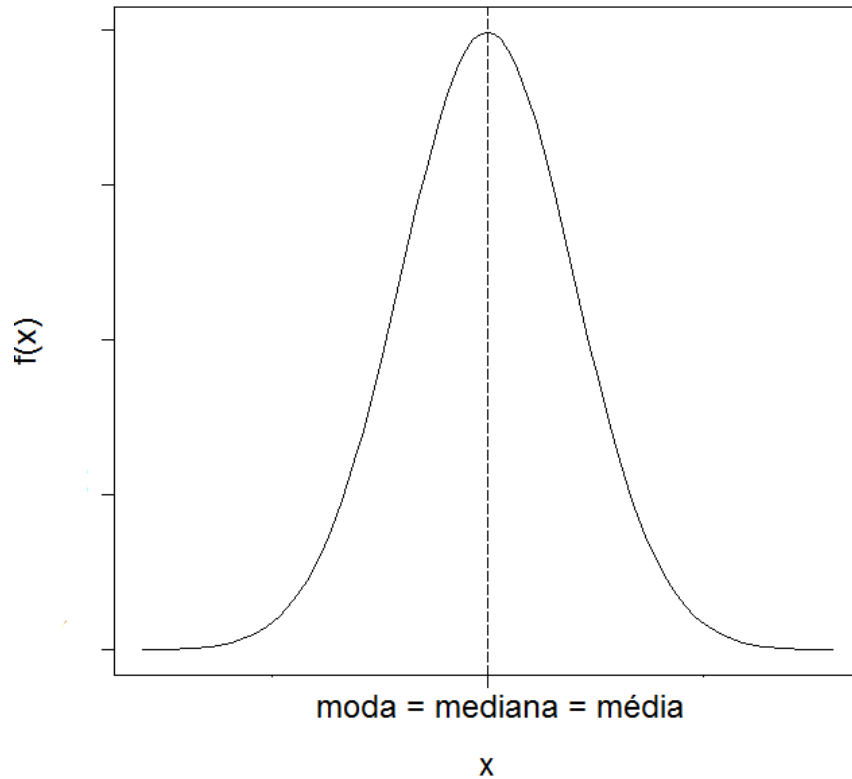
Distribuições assimétricas:



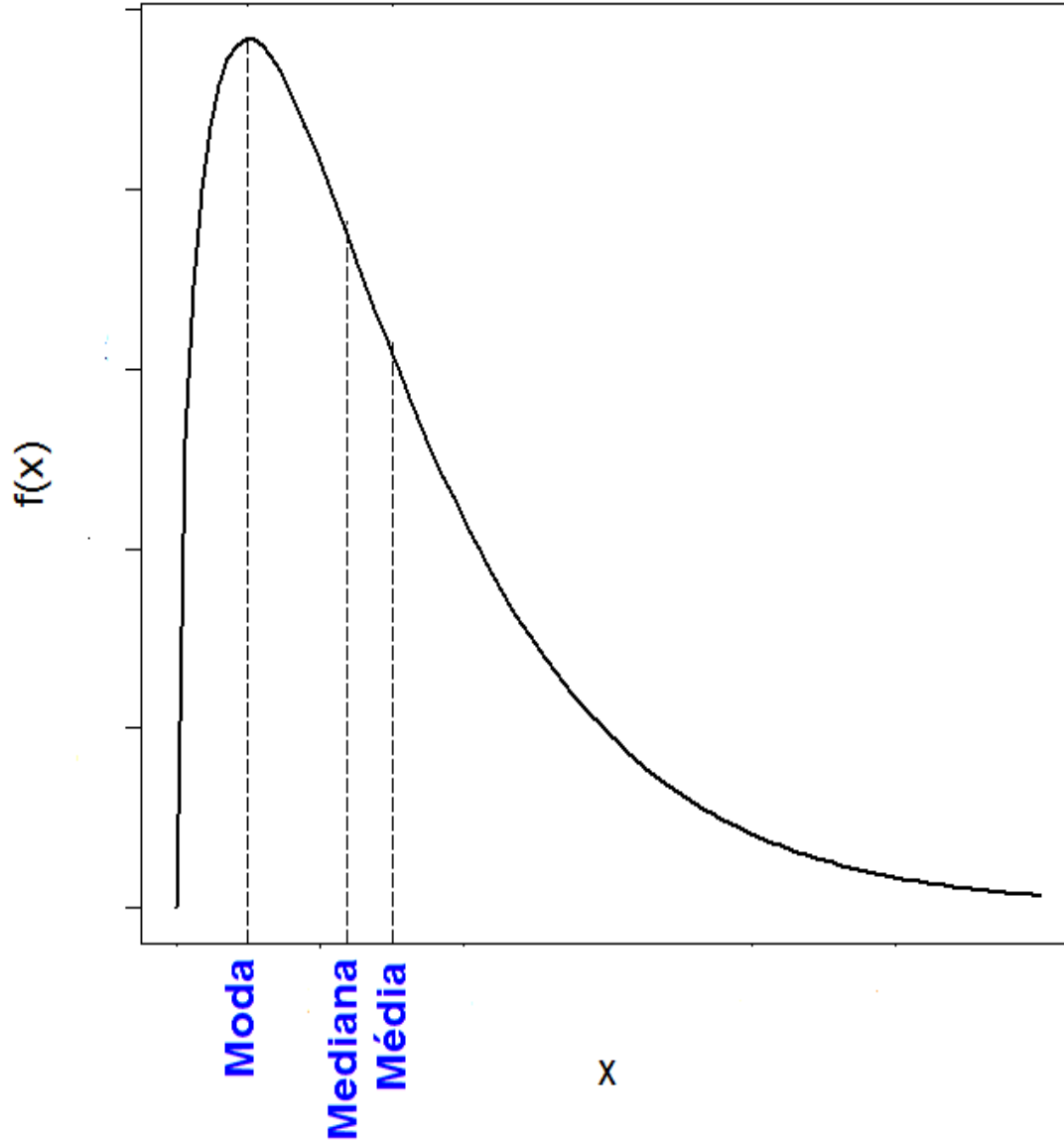
Relação entre moda, mediana e média

Supomos que a distribuição é **unimodal** e que a **média existe**.

Distribuição **simétrica**: moda = mediana = média.



Relação entre moda, mediana e média



Distribuição
assimétrica à direita ou
assimétrica positiva
(*right skewed* ou
positive skewed):
média > mediana >
moda.

Cauda direita (*right tail*) é
mais longa.

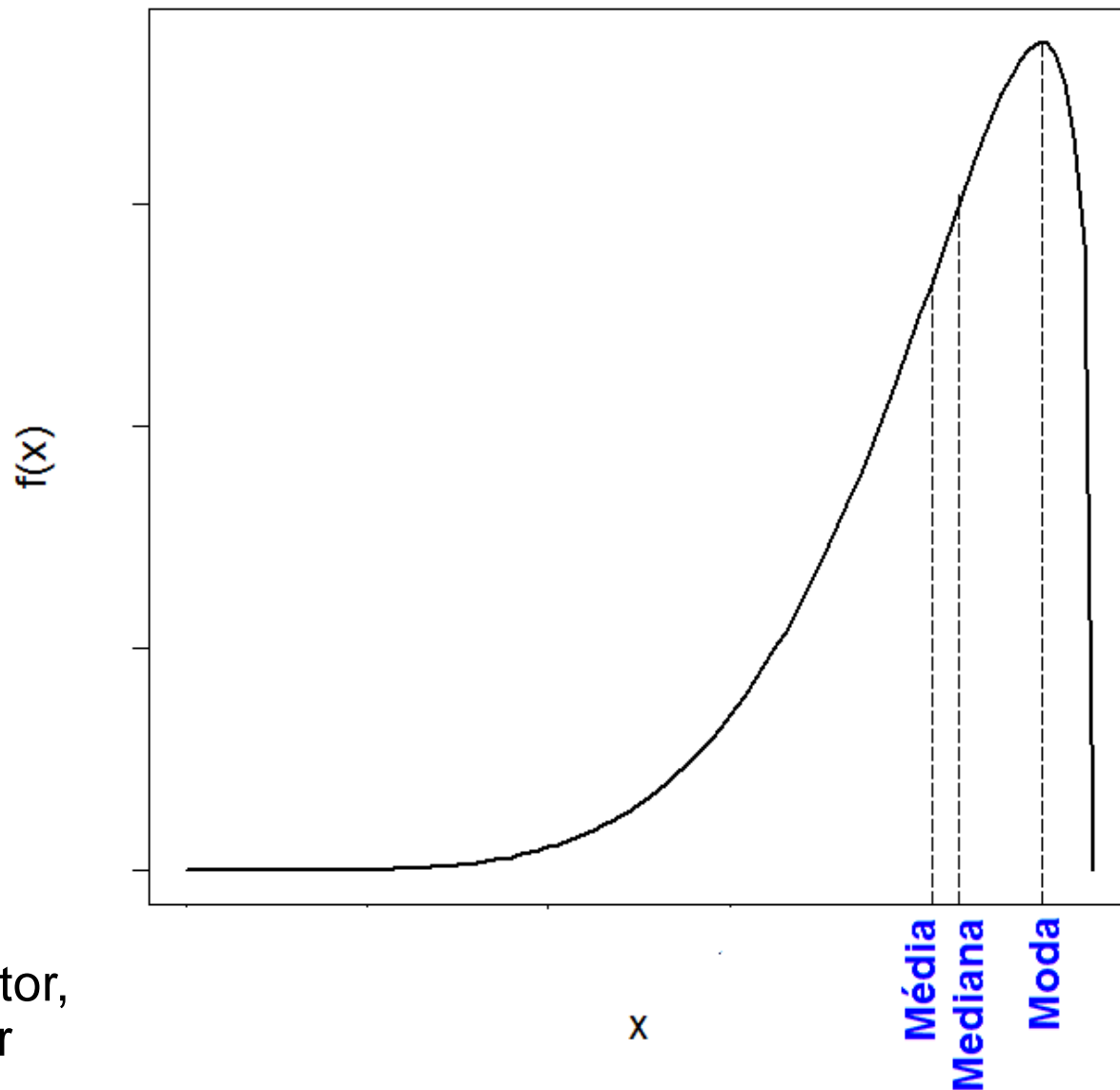


Relação entre moda, mediana e média

Distribuição
assimétrica à
esquerda ou
assimétrica negativa
(*left skewed* ou
negative skewed):
média < mediana <
moda.

Cauda esquerda (*left
tail*) é mais longa.

Obs. Dependendo do autor,
há troca de esquerda por
direita.



Um conjunto de dados x_1, x_2, \dots, x_n é **simétrico** em relação a x_0 se para todo x_j existe x_k tal que $x_j - x_0 = -(x_k - x_0)$, $j \neq k$.

Desta forma,

$$\sum_{i=1}^n (x_i - x_0) = 0. \quad \text{Logo, } x_0 = \bar{x}.$$

Pela simetria, $M = x_0$.

Exemplo. (24, 26, 26, 32, 40, 60, 68, 74, 74, 76) é simétrico em relação a $x_0 = 50$.

Medidas de assimetria. Quantificação da **falta de simetria** de x_1, x_2, \dots, x_n .

Redução (**drástica**) de n observações a **um só** valor.

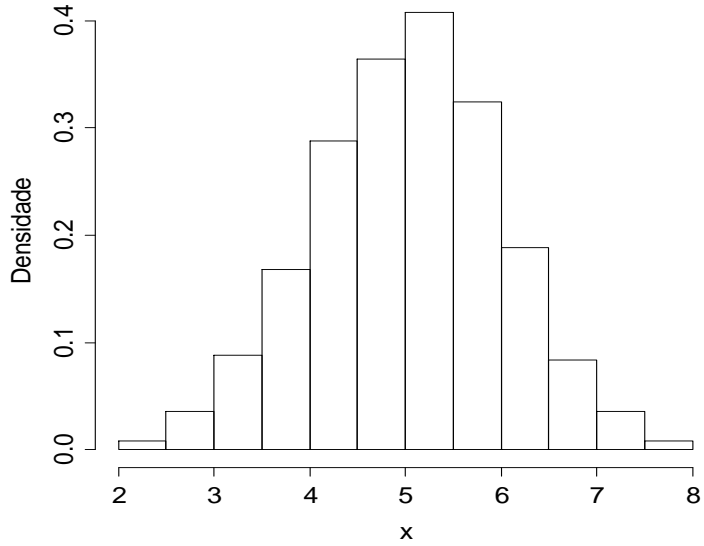
Assume valores reais. Medida = 0: **indicativo** de simetria.

Obs. Dados obtidos de uma **variável aleatória** X **simétrica** não necessariamente **são simétricos**.

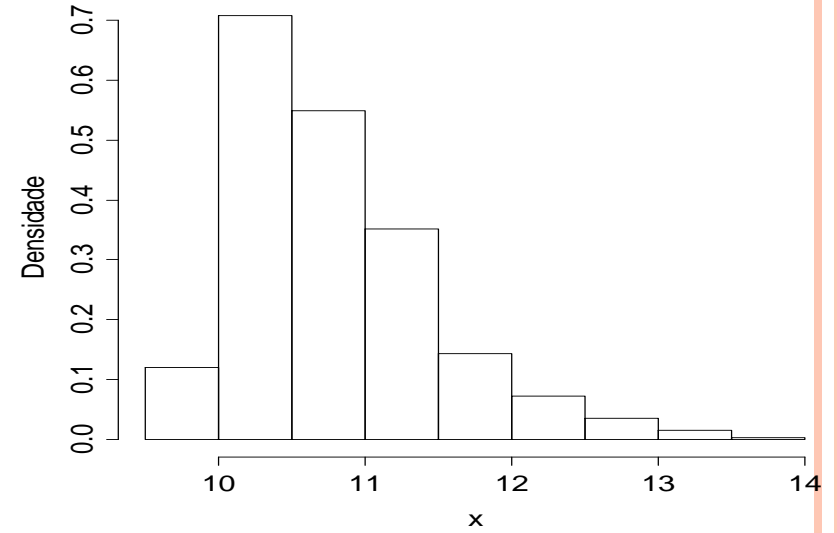


Assimetria em conjuntos de dados

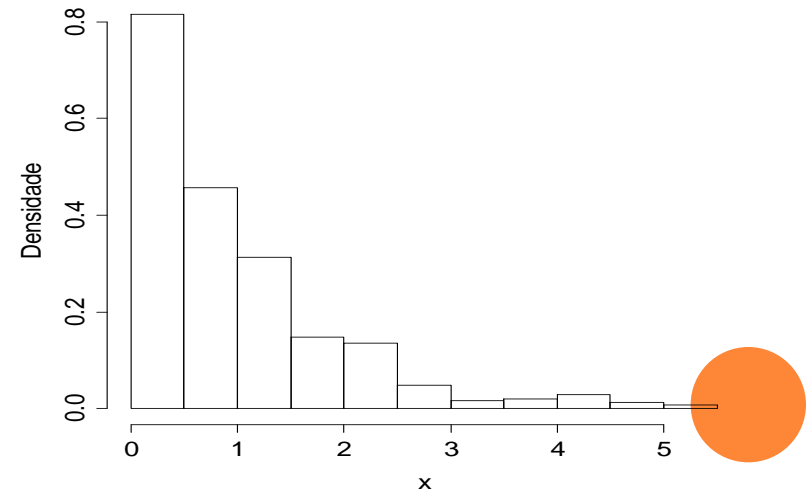
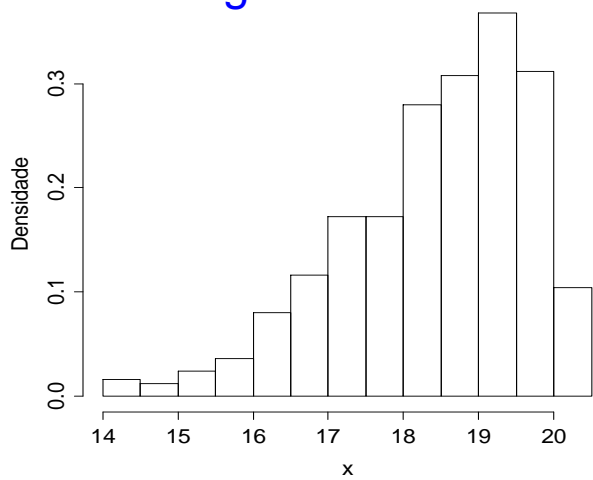
Aproximadamente simétrico



Assimetria positiva



Assimetria negativa



6.1.1. Medida de assimetria de Pearson

Utiliza a relação entre moda, mediana e média em distribuições unimodais.

$$g = \frac{\bar{x} - \text{Mo}}{s}.$$


Requer o cálculo da moda (Mo). É adimensional.

Momentos. Dados: x_1, x_2, \dots, x_n .

Momento de ordem k em relação à origem ou momento não central (*raw* ou *crude moment*):

$$m_k^* = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad k = 0, 1, 2, \dots, \text{ sendo que } m_0^* = 1 \text{ e } m_1^* = \bar{x}$$

Momento central de ordem k (*central moment*):

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, \quad k = 0, 1, 2, \dots, \text{ sendo que } m_0 = 1 \text{ e } m_1 = 0.$$


Obs. $m_2 = (n - 1) s^2 / n$.

Relação entre os momentos:

$$m_2 = m_2^* - \bar{x}^{-2},$$

$$m_2^* = m_2 + \bar{x}^{-2},$$

$$m_3 = m_3^* - 3m_2^* \bar{x} + 2\bar{x}^{-3} \quad \text{e}$$

$$m_3^* = m_3 + 3m_2 \bar{x} + \bar{x}^{-3} \quad \text{e}$$

$$m_4 = m_4^* - 4m_3^* \bar{x} + 6m_2^* \bar{x}^{-2} - 3\bar{x}^{-4}.$$

$$m_4^* = m_4 + 4m_3 \bar{x} + 6m_2 \bar{x}^{-2} + \bar{x}^{-4}.$$

Se uma **variável aleatória** X tem distribuição **simétrica**, seus **momentos centrais** de ordem k **ímpar**, se **existirem**, são todos nulos:

$E[(X - \mu)]^k = 0$, sendo que $\mu = E(X)$.

Sendo assim, podemos utilizar o **3º** momento **central** para propor uma medida de assimetria.

Obs. $E[(X - \mu)]^3 = 0$ **não implica** que a distribuição de X é **simétrica**.



6.1.2. Assimetria (*skewness*)

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}.$$

Propriedades. (1) g_1 é **adimensional** e (2) g_1 é um número **real**.

6.2. Medidas de curtose

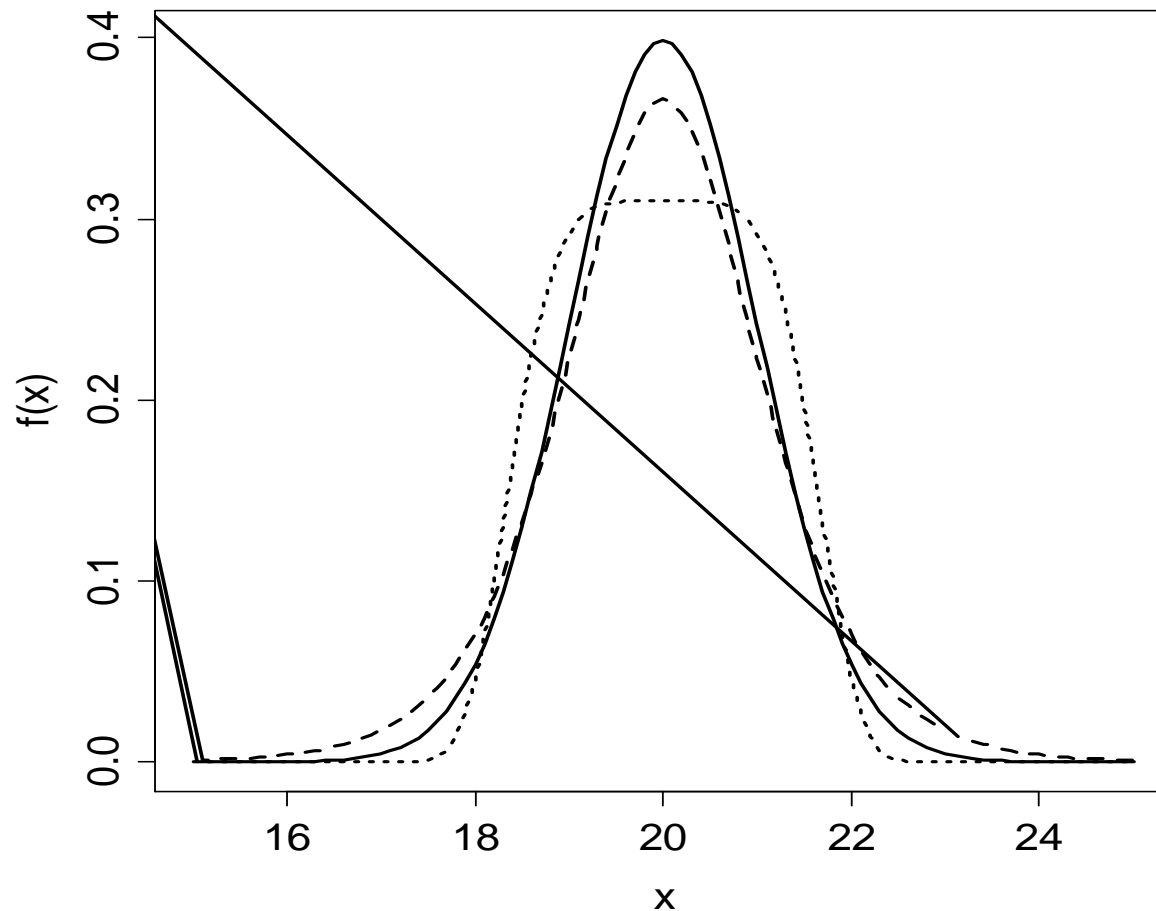
Em distribuições **unimodais**, a curtose (*kurtosis*) está associada ao **achatamento** da distribuição.

Característica do “pico” da distribuição (*degree of peakedness*).

No caso de uma **variável aleatória contínua unimodal**, a curtose diz respeito à forma de $f(x)$ em torno de sua **moda** (que pode coincidir com a média e a mediana).



Distribuições **simétricas** com **médias** e **variâncias** iguais:



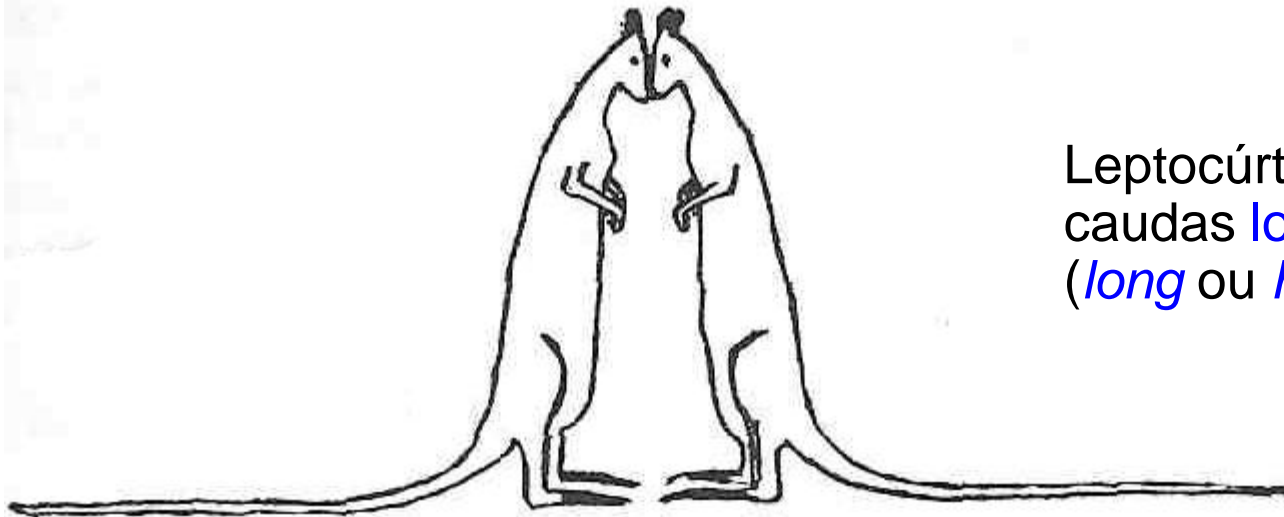
Diferenças quanto ao **afastamento** em relação à **média**, mas que **não** são captadas pela **variância**.

Este fato sugere utilizarmos o **4º momento central** para quantificar estas diferenças.

Distribuições **platicúrticas**, **mesocúrticas** e **leptocúrticas**



Platicúrtica (*platykurtic*):
caudas **curtas** ou **leves** (*short*
ou *light* ou *thin*).



Leptocúrtica (*leptokurtic*):
caudas **longas** ou **pesadas**
(*long* ou *heavy* ou *thick* ou *fat*).

Fonte. Bulmer, M. G. (1979), *Principles of Statistics*, Dover: New York.

Mesocúrtica (*mesokurtic*): caudas **neutras** (nem curtas e nem longas).

Curtose (*kurtosis*)

$$g_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^2}.$$

Propriedades. (1) g_2 é adimensional e (2) $g_2 > 0$.

Obs. Se $X \sim N(\mu, \sigma^2)$, então $E[(X - \mu)^4] / \{E[(X - \mu)^2]\}^2 = E[(X - \mu)^4] / \sigma^4 = 3$.

Muitas vezes a curtose é expressa como $g_{2e} = g_2 - 3$ (**excess**).

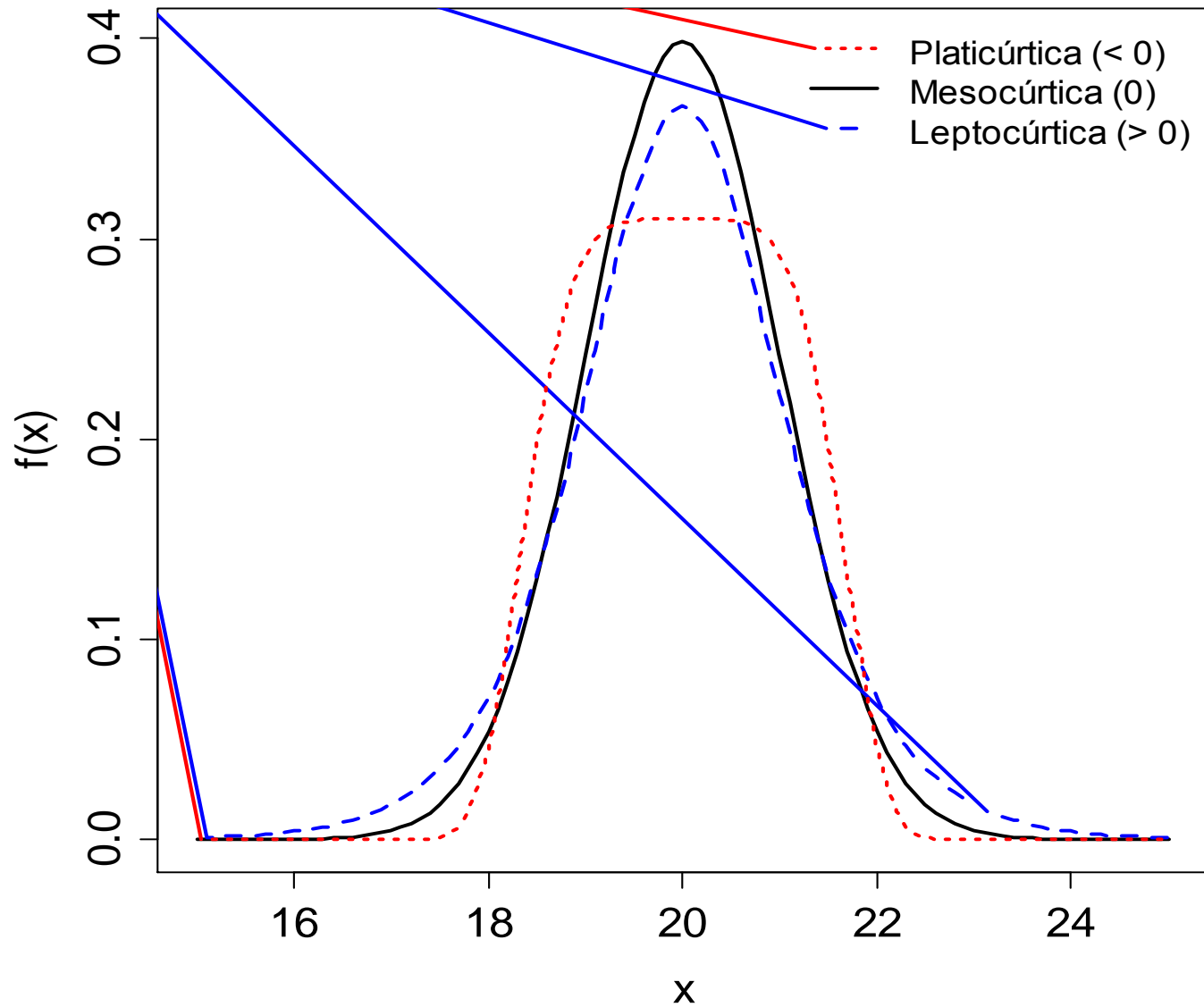
Platicúrtica: $g_{2e} < 0$.

Mesocúrtica: $g_{2e} = 0$.

Leptocúrtica: $g_{2e} > 0$.



Distribuições simétricas com médias e variâncias iguais:



Momentos em R

Pacote `moments`

```
> library(moments)
```

Exemplo. 1, 0, 2,7, -1,4 e 13 são os momentos **centrais** até ordem 4 de um conjunto de dados com **média = 3,8**.

```
> momc = c(1, 0, 2.7, -1.4, 13)
```

```
> xb = 3.8
```

```
> (mom0 = central2raw(momc, xb))
```

```
[1] 1.0000 3.8000 17.1400 84.2520 434.1616 (momentos não centrais)
```

Função `raw2central`: converte momentos não centrais em centrais.

```
> mom0 = c(1.0, 1.5, 3.1, 8.4, 27.0)
```

```
> (momc = raw2central(mom0))
```

```
[1] 1.0000 0.0000 0.8500 1.2000 3.2625 (momentos centrais)
```



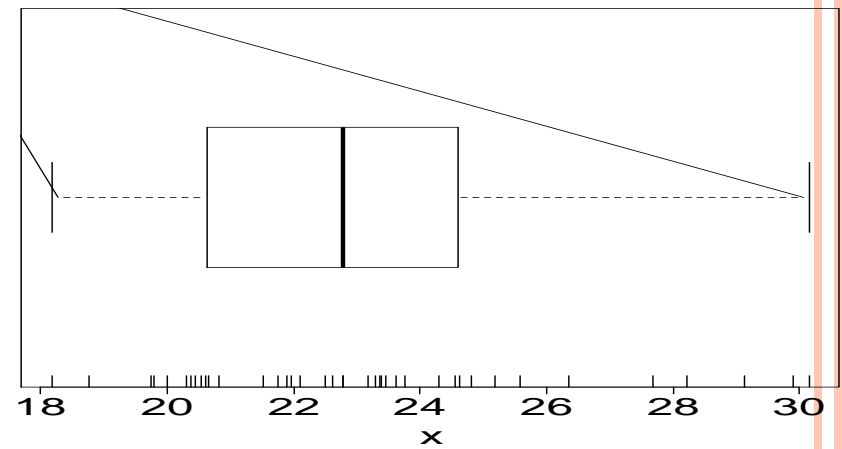
Momentos em R

40 observações

```
x = c(21.88, 22.61, 23.28, 25.19, 19.73, 22.50, 26.35, 22.09, 24.63,  
19.79, 23.61, 29.91, 29.12, 23.17, 23.38, 21.75, 18.77, 20.38,  
20.45, 28.23, 18.17, 30.15, 20.60, 19.99, 24.56, 25.59, 20.66,  
23.76, 23.35, 22.77, 21.52, 20.54, 20.30, 23.45, 27.69, 24.82,  
24.30, 21.97, 20.82, 22.78)
```

```
> boxplot(x, horizontal = TRUE,  
xlab = "X")
```

```
> rug(x)
```



Momento **não central** de ordem 3:

```
> moment(x, order = 3, central = FALSE)
```

```
[1] 12960.19
```

Qual o resultado de
 $\text{mean}(x^3)$?

Momentos **centrais** até ordem 4:

```
> (mom4 = all.moments(x, order.max = 4, central = TRUE))
```

```
[1] 1.000000e+00 4.440892e-16 8.521595e+00 1.843078e+01 2.206282e+02  
k = 0 1 2 3 4
```



Assimetria e curtose em R

Pacote `moments`

```
> skewness(x)
```

```
[1] 0.7409046
```

Qual o resultado de `mom4[4] / mom4[3]^1.5`?

```
> kurtosis(x)
```

```
[1] 3.03822
```

Qual o resultado de `mom4[5] / mom4[3]^2`?

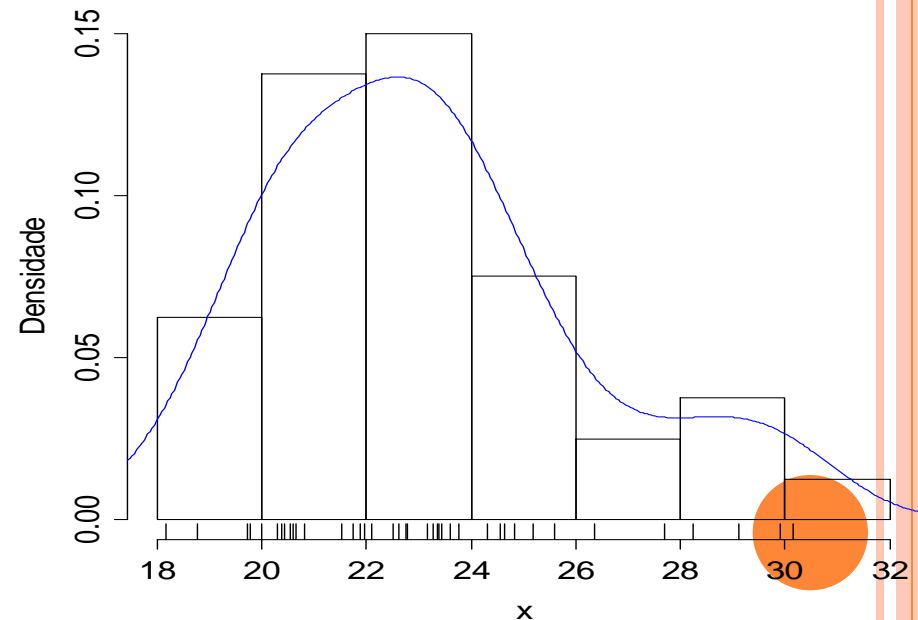
```
> kurtosis(x) - 3
```

```
[1] 0.03822
```

```
> hist(x, main = "", xlab =  
"x", freq = FALSE, ylab =  
"Densidade")
```

```
> rug(x)
```

```
> lines(density(x), col =  
"blue")
```



Dados faltantes (*missing data*) em R

Observações **indisponíveis** por algum motivo.

Em R: **NA** (*not available*).

40 (?) observações

```
x = c(21.88, NA, 23.28, 25.19, 19.73, 22.50, 26.35, 22.09, 24.63,  
19.79, 23.61, 29.91, 29.12, 23.17, 23.38, 21.75, NA, 20.38, 20.45,  
28.23, 18.17, 30.15, 20.60, 19.99, 24.56, 25.59, 20.66, 23.76,  
23.35, NA, 21.52, 20.54, 20.30, 23.45, 27.69, 24.82, 24.30, 21.97,  
20.82, 22.78)
```

```
> summary(x)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      NA's  
18.17  20.66   23.17   23.26  24.63   30.15      3.00
```

```
> mean(x)           > mean(x, na.rm = TRUE)
```

```
[1] NA                [1] 23.25568
```

```
> kurtosis(x)       > kurtosis(x, na.rm = TRUE)
```

```
[1] NA                [1] 2.891447
```

```
> which(is.na(x))
```

```
[1]  2 17 30
```

Substituição dos faltantes pela média:

```
> is.na(x) = mean(x, na.rm = TRUE)
```

