altre misure di centro e dispersione

media geometrica

la radice n-esima del prodotto degli n dati

dati **G**

2, 8 ----> $\sqrt{2 \cdot 8} = 4$

4, 1, 1/32 -> $\sqrt[3]{4 \cdot 1 \cdot 1/32} = 1/2$

applicazioni

poco utilizzata nelle scienze sociali

è utile per rappresentare la tendenza centrale in distribuzioni non simmetriche

relazione con la media di log(x)

se logM è la media aritmetica di log(x)

allora G = antilog(logM)

il logaritmo di x è il numero a cui va elevata la base per ottenere x: log10(100) = 2

l’antilogaritmo di log è il numero che si ottiene elevando la base alla potenza log: antilog10(2) =10^2 = 100

media armonica

il reciproco della media dei reciproci dei dati

dati

1, 2, 4

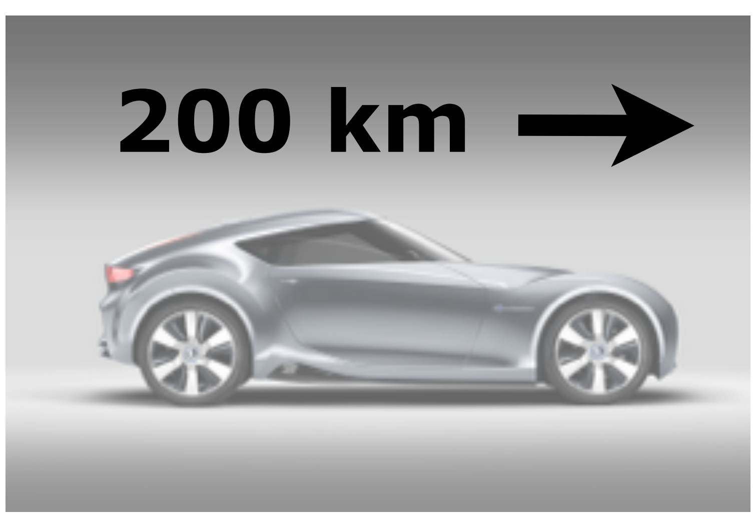
H

$$\frac{1}{\frac{1}{3}\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{4}\right)} = \frac{12}{7} = 1.\overline{714285}$$

applicazioni

poco utilizzata nelle scienze sociali

usata in fisica in situazioni in cui occorre mediare rapporti o tassi di crescita



S	V
100 km	200 km/h
100 km	100 km/h

S	V	T
100 km	200 km/h	0.5 h
100 km	100 km/h	1 h

V media sui 200 km = S/T

= 200/1.5

= 133 km/h

media aritmetica = (200 + 100)/2

= 150 km/h!

media armonica = 1/[(1/200 + 1/100)/2]

= 133 km/h

```
> d <- read.table("IQ.txt", header = TRUE)
> ma <- mean(d$Height)
> mg <- prod(d$Height)^(1/length(d$Height))
> mh <- 1/mean(1/d$Height)

> ma
[1] 68.5375

> mg
[1] 68.42843

> mh
[1] 68.32095
```

```
> mg
[1] 68.42843

> 10^mean(log10(d$Height))
[1] 68.42843

> exp(mean(log(d$Height)))
[1] 68.42843
```

deviazione mediana assoluta

mediana degli scarti non segnati dalla media

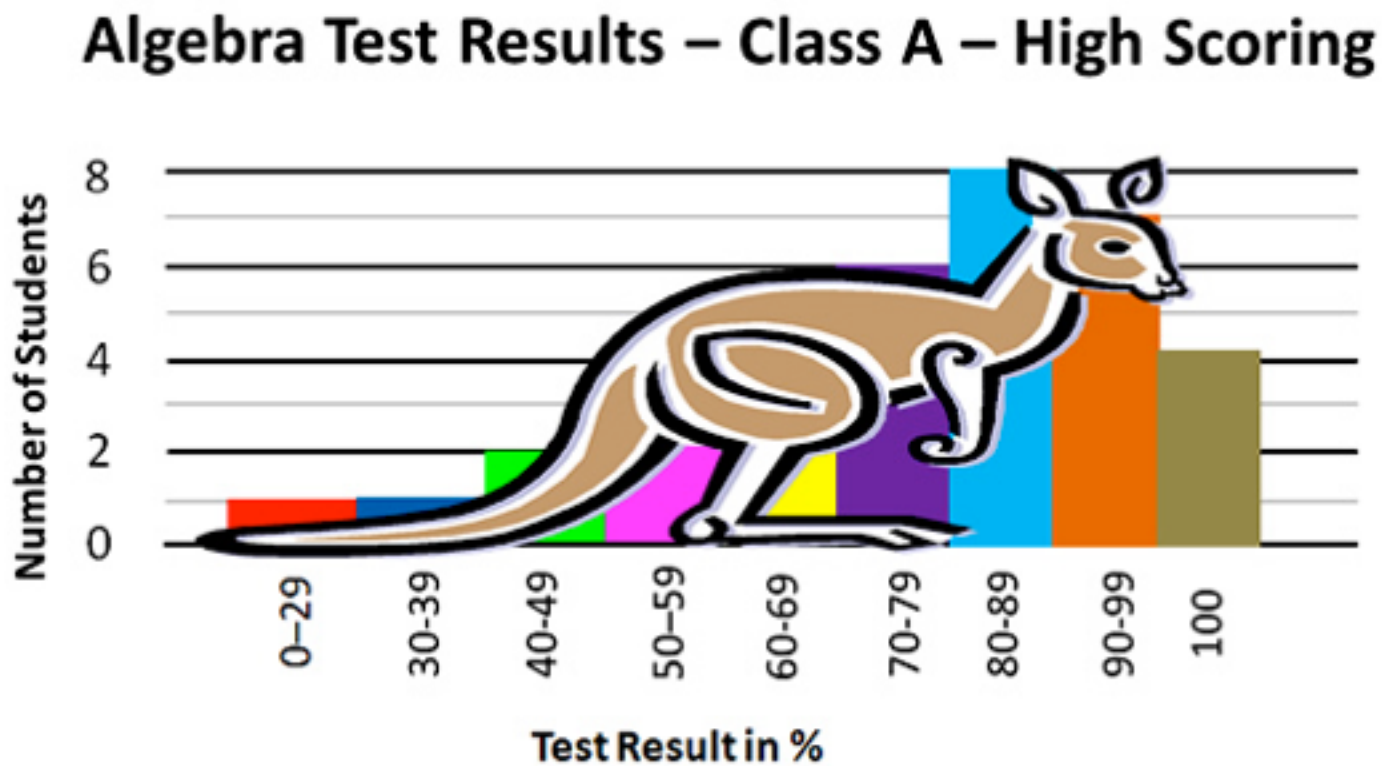
Median Absolute Deviation

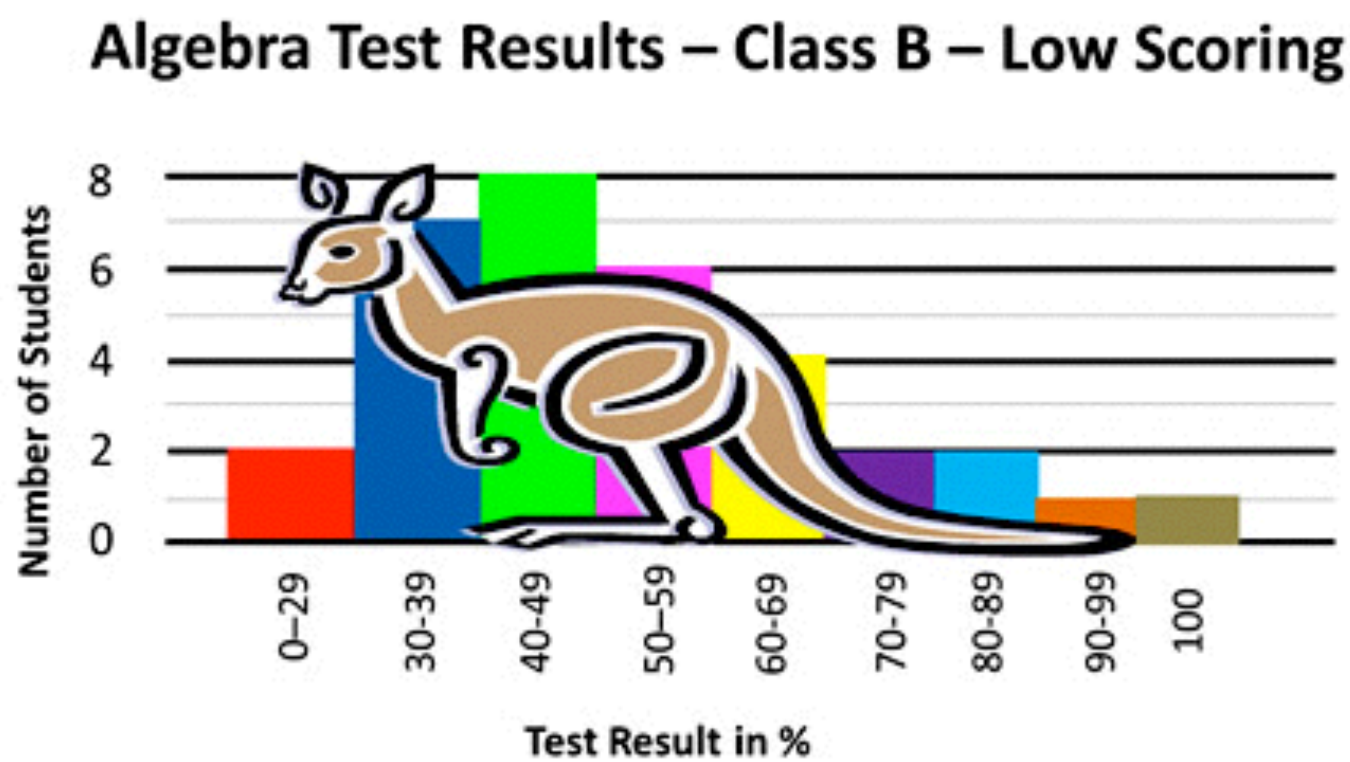
mad()

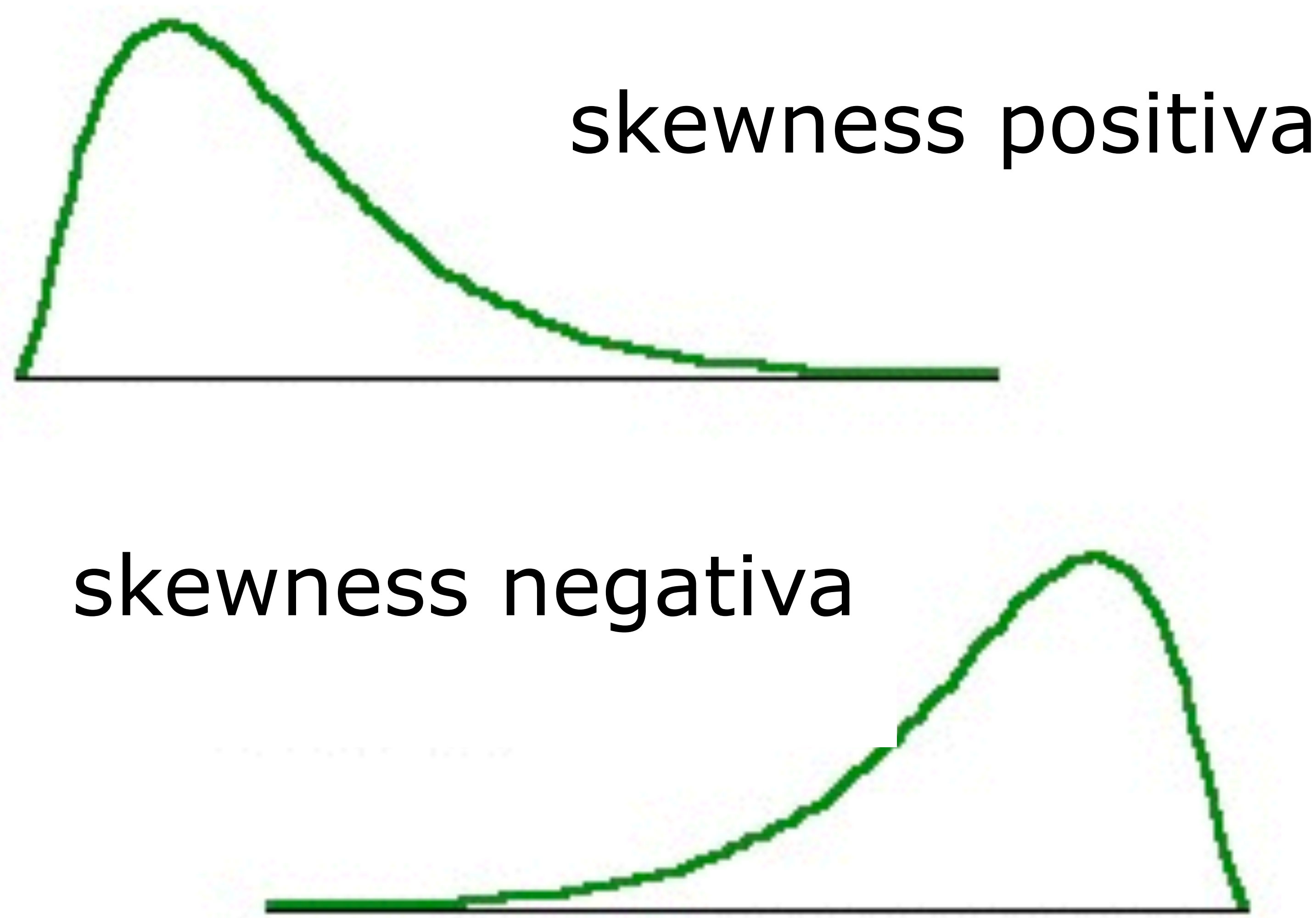
```
> mad(d$Height)
[1] 3.33585

> sd(d$Height)
[1] 3.943816
```

statistiche per la <u>forma</u> di una distribuzione

asimmetria *(skewness)*

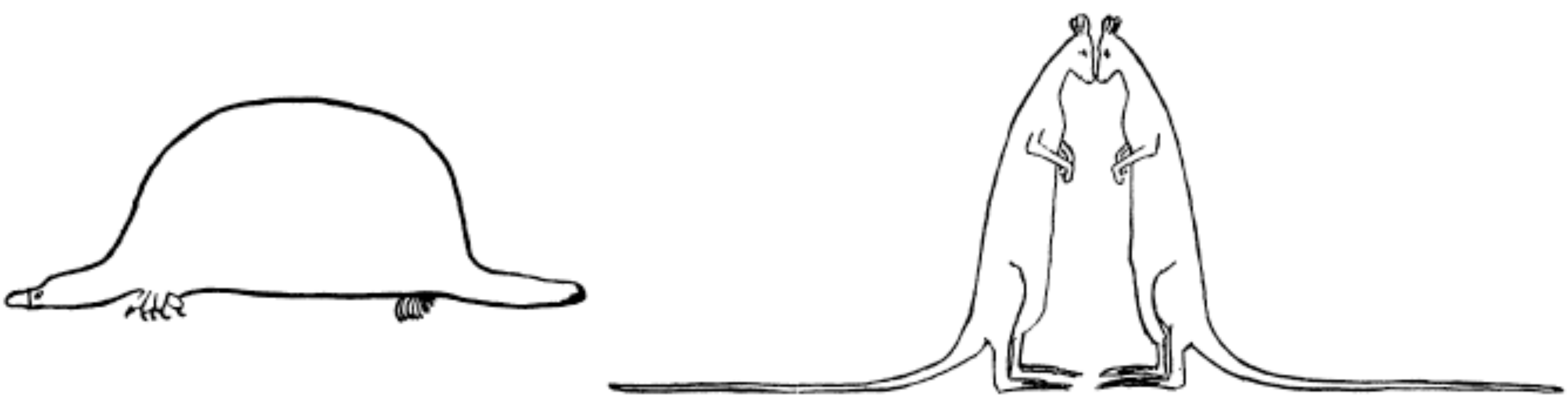





skewness positiva
skewness negativa

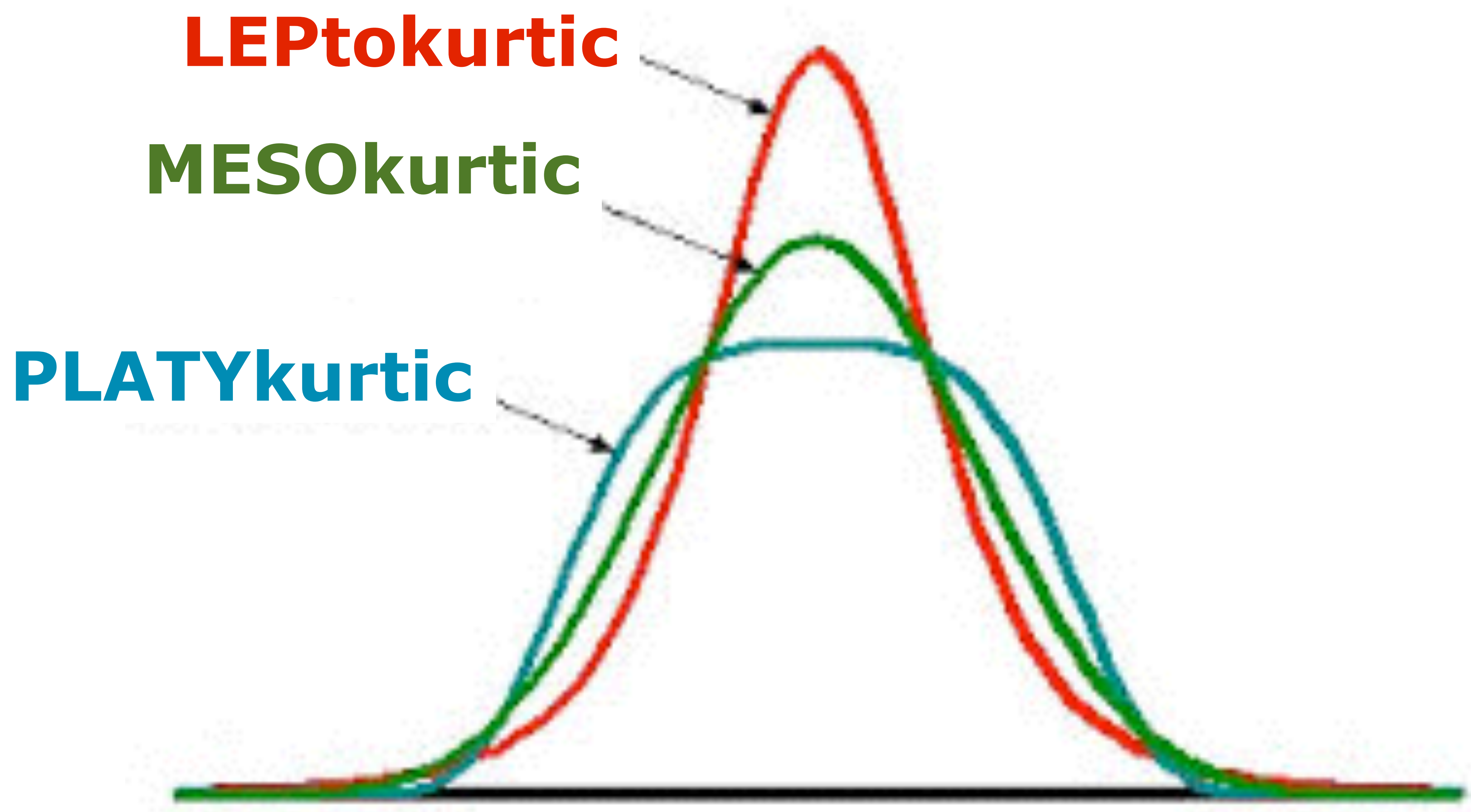
curtosi

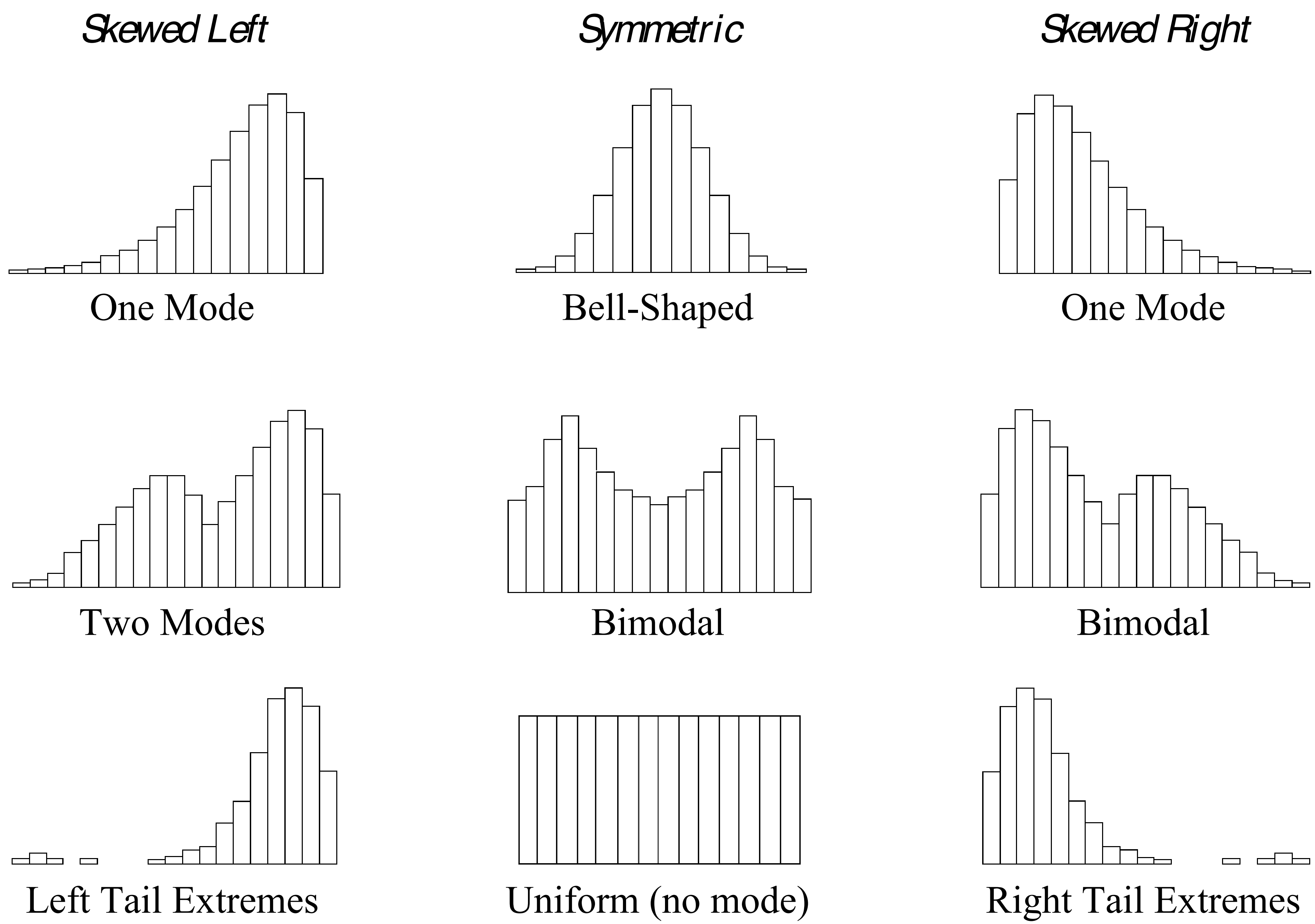
* In case any of my readers may be unfamiliar with the term "kurtosis" we may define mesokurtic as "having β_2 equal to 3," while platykurtic curves have $\beta_2 < 3$ and leptokurtic > 3. The important property which follows from this is that platykurtic curves have shorter "tails" than the



normal curve of error and leptokurtic longer "tails." I myself bear in mind the meaning of the words by the above *memoria technica*, where the first figure represents platypus, and the second kangaroos, noted for "lepping," though, perhaps, with equal reason they should be hares!

Student (1927) *Biometrika*, *19*, 160


LEPtokurtic
MESOkurtic
PLATYkurtic



Skewed Left
Symmetric
Skewed Right
One Mode
Bell-Shaped
One Mode
Two Modes
Bimodal
Bimodal
Left Tail Extremes
Uniform (no mode)
Right Tail Extremes

coefficiente di skewness di Fisher

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

Chiorri

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3=$$

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

correzione per n piccoli

Measuring Skewness: A Forgotten Statistic?

David P. Doane
Oakland University

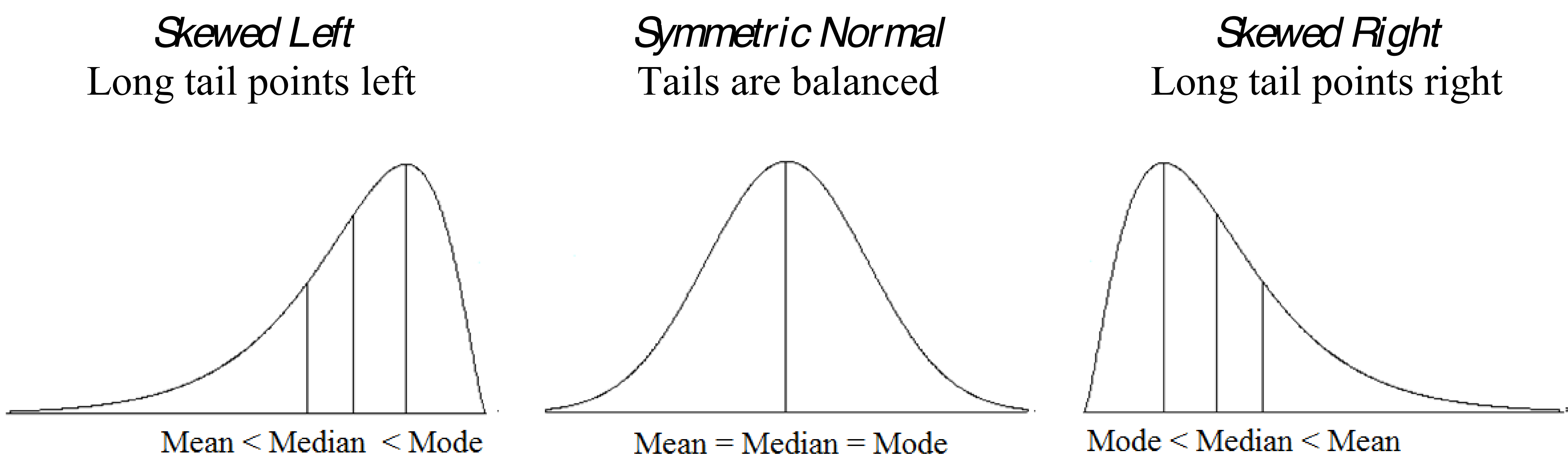
Lori E. Seward
University of Colorado

Abstract

This paper discusses common approaches to presenting the topic of skewness in the classroom, and explains why students need to know how to measure it. Two skewness statistics are examined: the Fisher-Pearson standardized third moment coefficient, and the Pearson 3 coefficient that compares the mean and median.

This paper suggests reviving the Pearson 3 skewness statistic for the introductory statistics course because it compares the mean to the median in a precise way that students can understand. The paper reiterates warnings about what any skewness statistic can actually tell us.

skewness media, mediana e moda



SK3 di Pearson

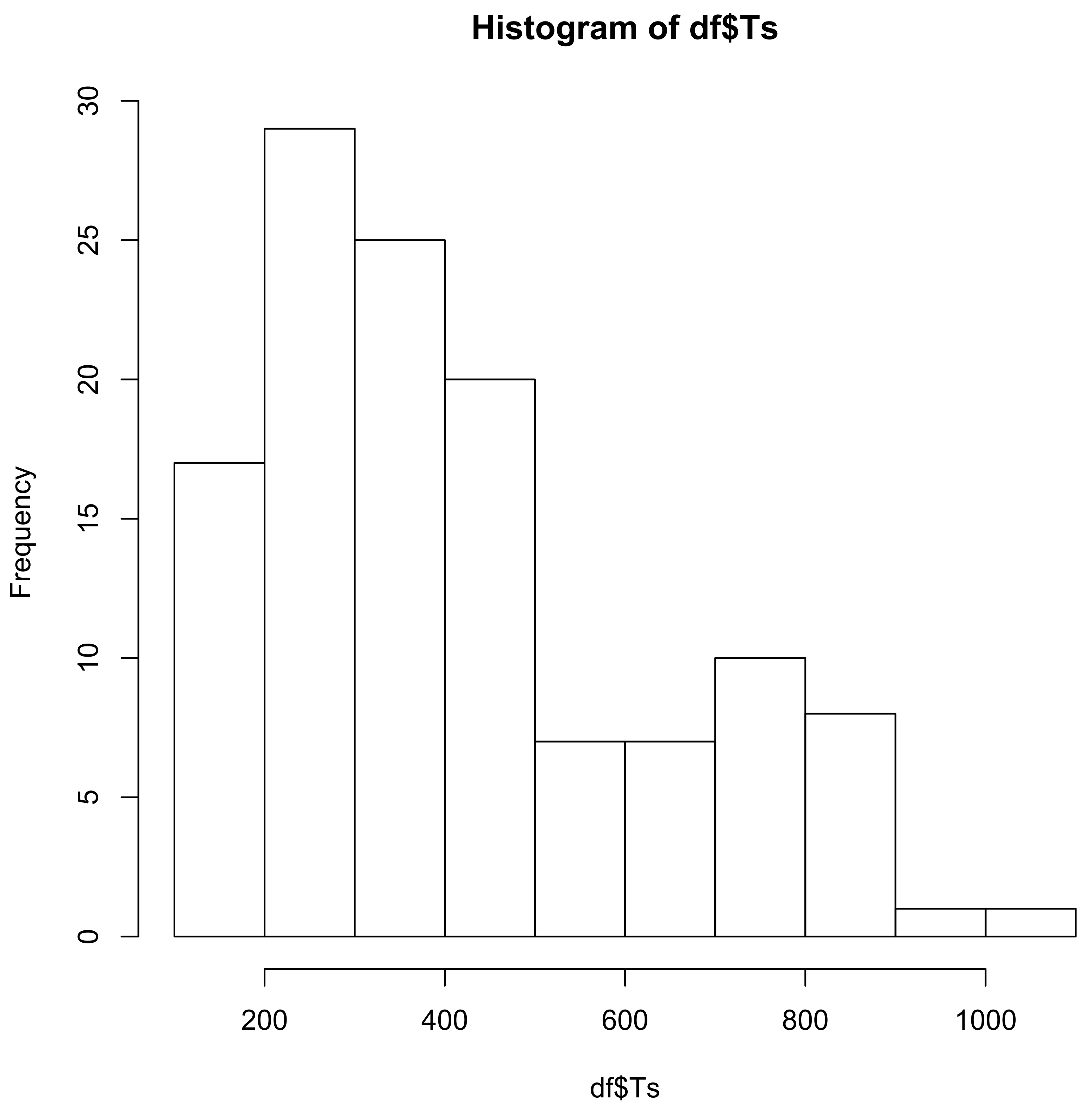
SK3 = 3(media - mediana) / DS

SK1 = (media - moda) / DS
SK2 = 3(media - moda) / DS

coefficiente di curtosi di Fisher

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

```
> df <- read.table("~/Desktop/dati completi.txt",
header = TRUE)

> head(df)

  OvsR Sex HAND        RVF     LIKF  Ts SPAF   Eng compito
1    O   f    dx -3.9873143 1.500000 190  1.0     1       c
2    O   f    dx -0.4353741 4.166667 413  5.0     2      cd
3    R   f    dx  3.6857530 3.500000 491  2.5     1      cd
4    O   m    dx -1.4842264 5.666667 277  6.0     1       c
5    R   m    dx -1.3211927 6.000000 480  7.0     2      cd
6    R   m    dx -0.2262290 4.166667 265  3.0     2       c

> hist(df$Ts)
```


Histogram of df$Ts
Frequency
0
5
10
15
20
25
30
200
400
600
800
1000
df$Ts

```r
df <- read.table("~/Desktop/dati
completi.txt", header = TRUE)

m <- mean(df$Ts)
n <- length(df$Ts)
s <- sqrt(sum((df$Ts - m)^2)/n)

num <- sum((df$Ts - m)^3)/n
den <- (sum((df$Ts - m)^2)/n)^1.5
sk <- num/den

num <- sum((df$Ts - m)^4)/n
den <- (sum((df$Ts - m)^2)/n)^2
ku <- num/den
```

```
> sk
[1] 0.911130

> ku
[1] 2.884776

> library(moments)

> skewness(df$Ts)
[1] 0.9111309

> kurtosis(df$Ts)
[1] 2.884776
```

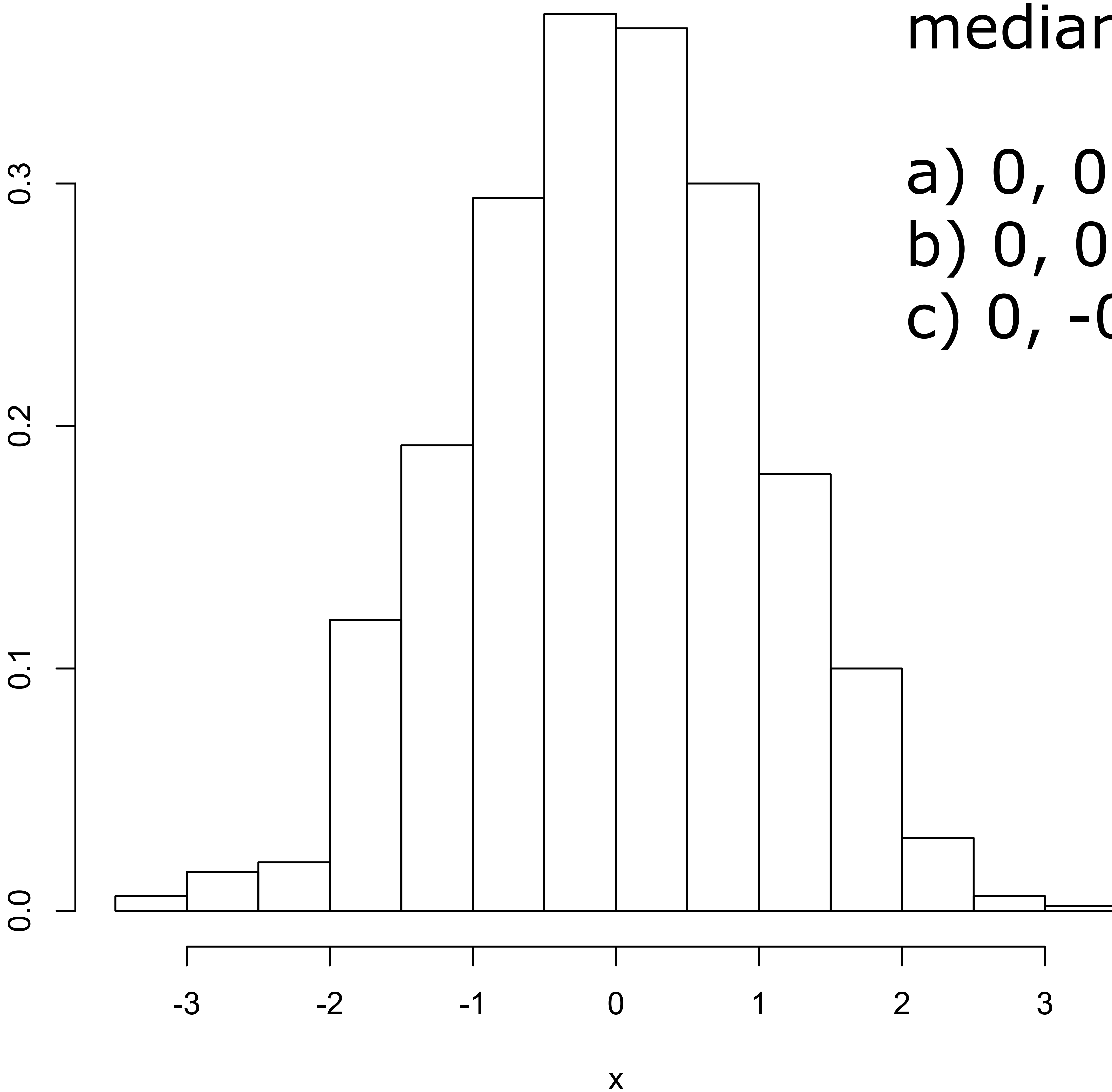
riassumendo

la *skewness* è 0 se la distribuzione è simmetrica, <0 se ha coda sinistra e >0 se ha coda a destra

la curtosi è 3 se la distribuzione è normale, >3 se è leptocurtica e <3 se è platicurtica

alcuni programmi calcolano il <u>coefficiente di eccesso</u> ku - 3, in tal caso la distribuzione normale ha curtosi 0 (R non lo fa)

facciamo un po’ di esercizio

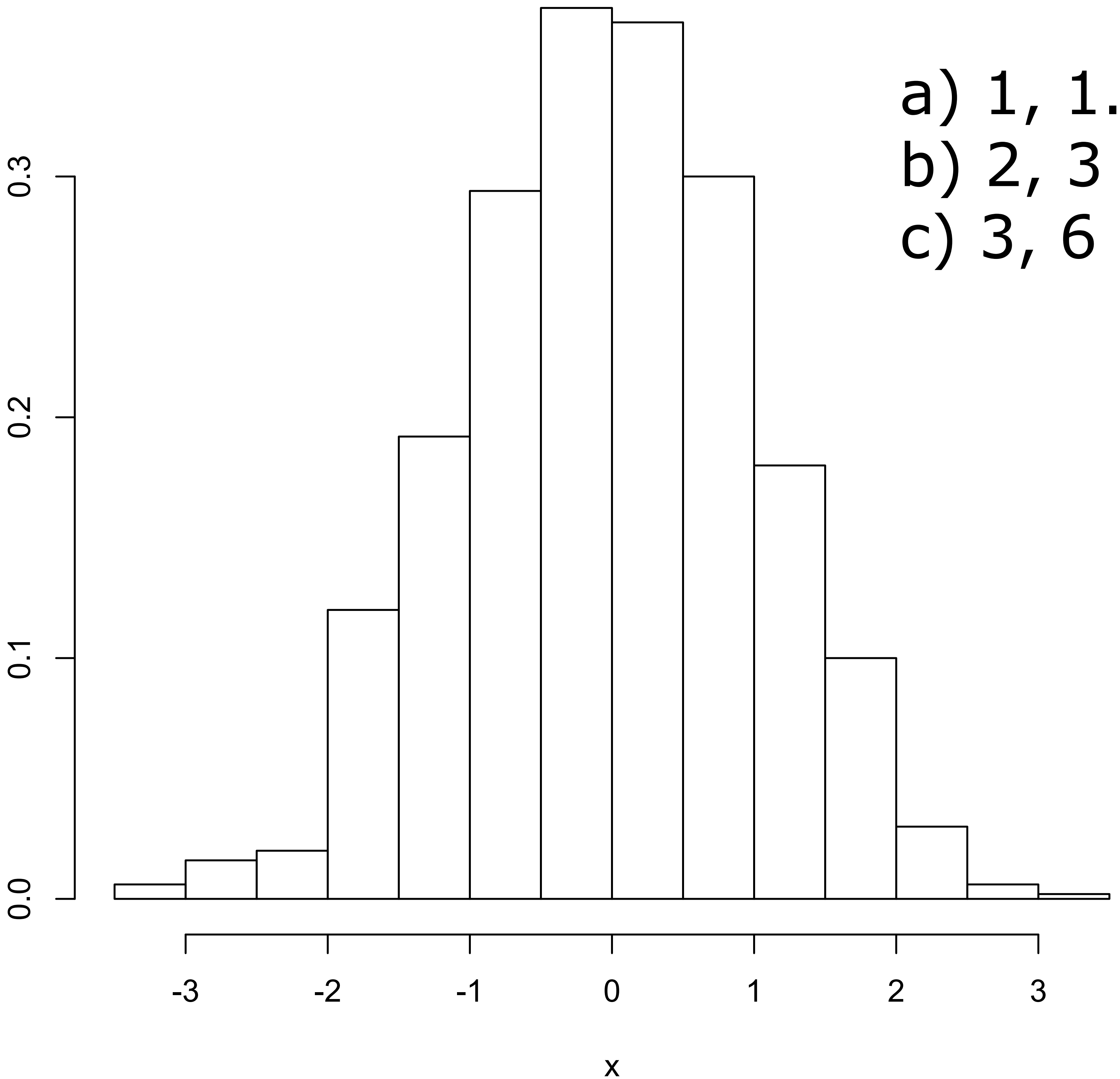


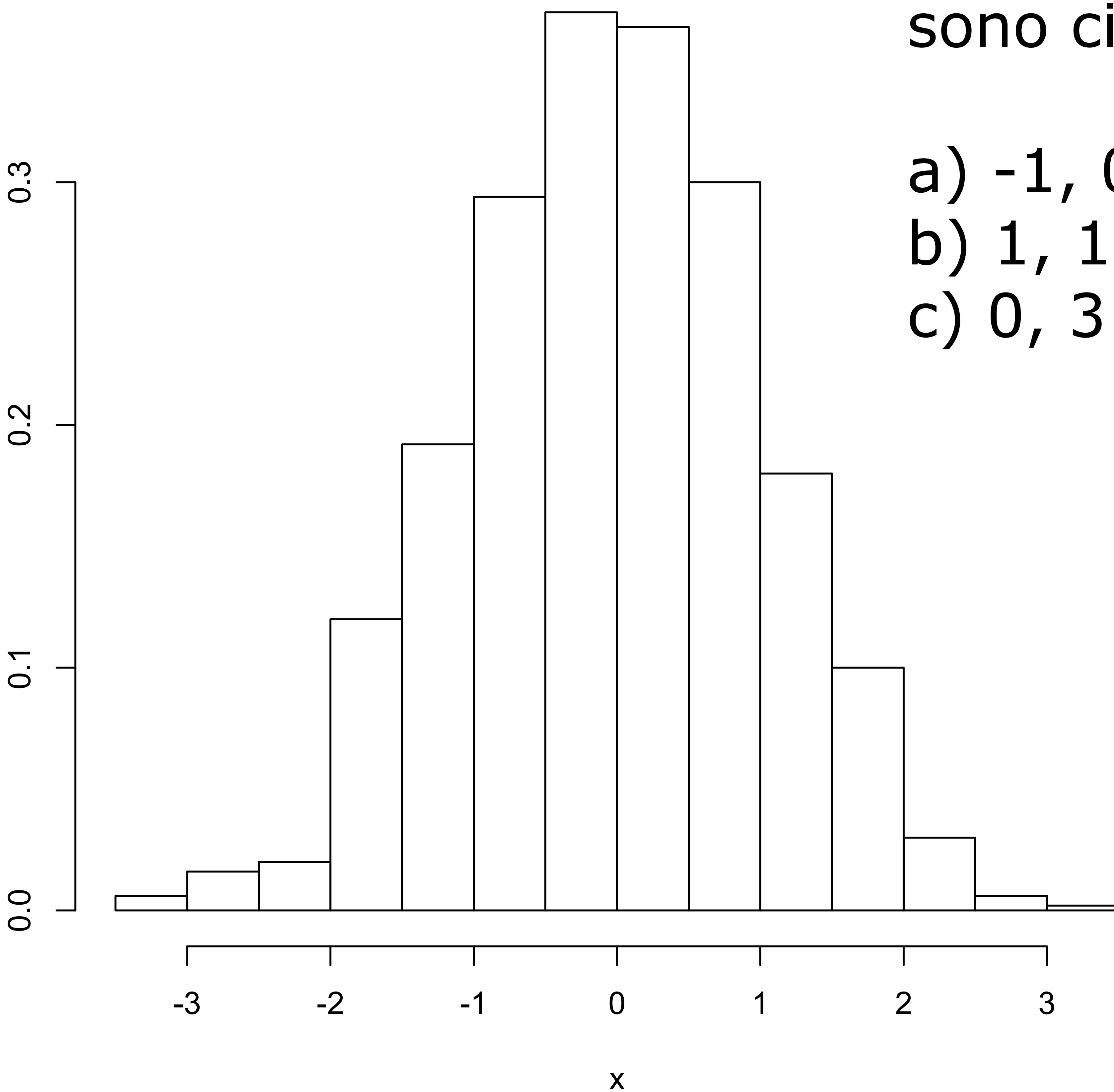
moda, media e
mediana sono circa:

a) 0, 0, 0
b) 0, 0.5, -0.5
c) 0, -0.5, 0.5

DS e GIQ sono circa:

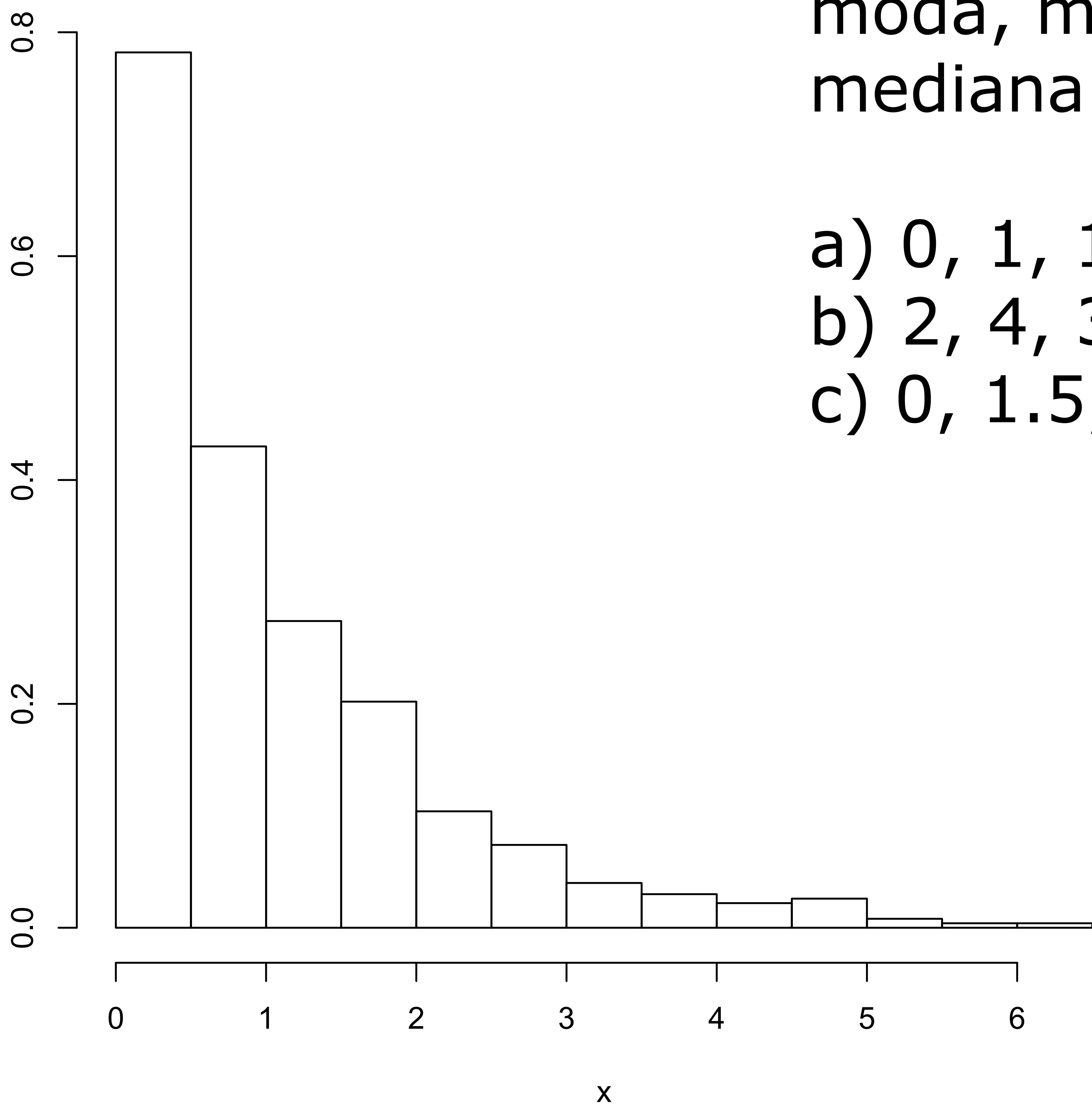
a) 1, 1.5
b) 2, 3
c) 3, 6





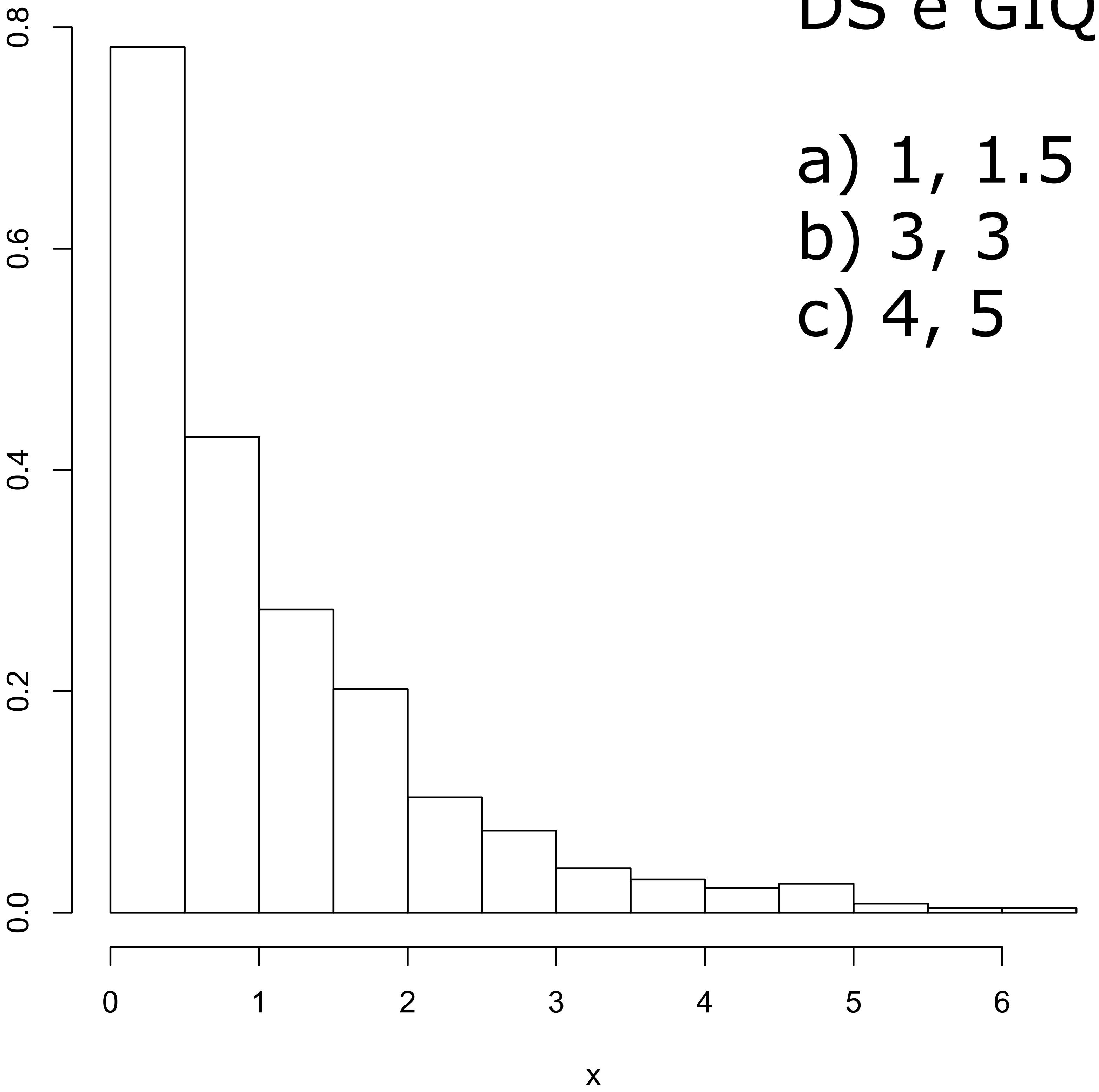
skewness e curtosi
sono circa:

a) -1, 0
b) 1, 1
c) 0, 3



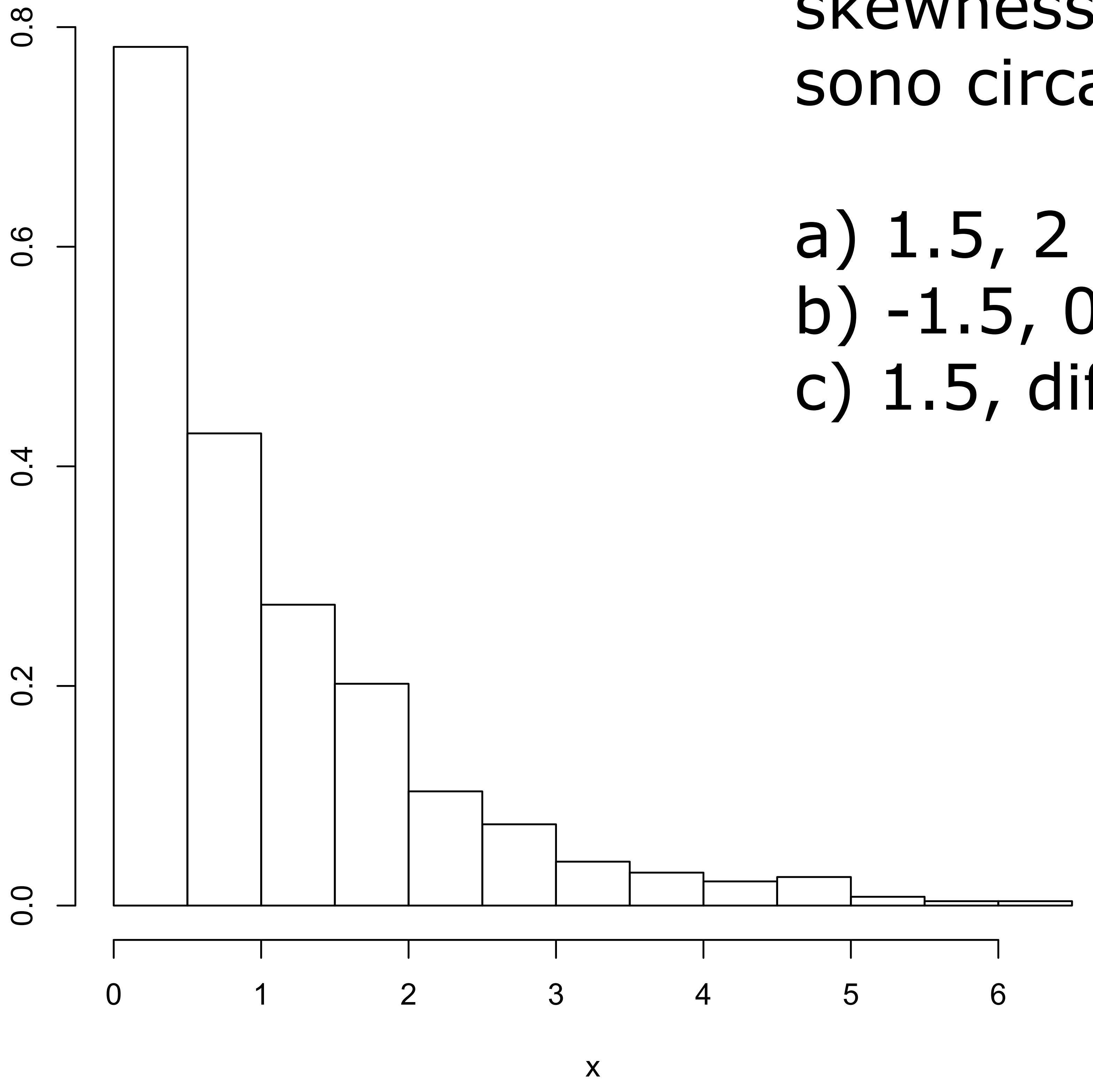
moda, media e
mediana sono circa:

a) 0, 1, 1.5
b) 2, 4, 3
c) 0, 1.5, 1



DS e GIQ sono circa:

a) 1, 1.5
b) 3, 3
c) 4, 5

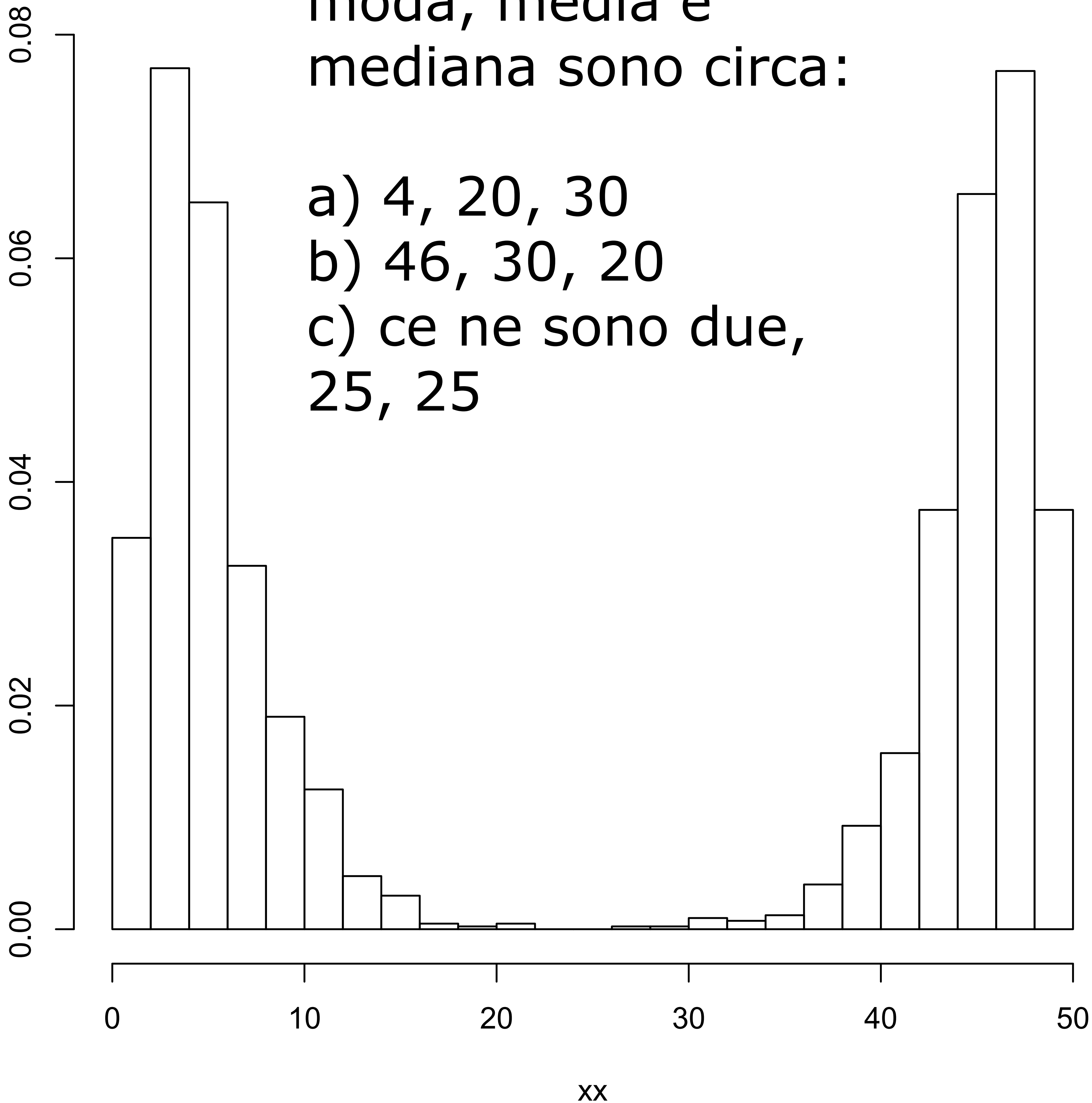


skewness e curtosi
sono circa:

a) 1.5, 2
b) -1.5, 0
c) 1.5, difficile dirlo

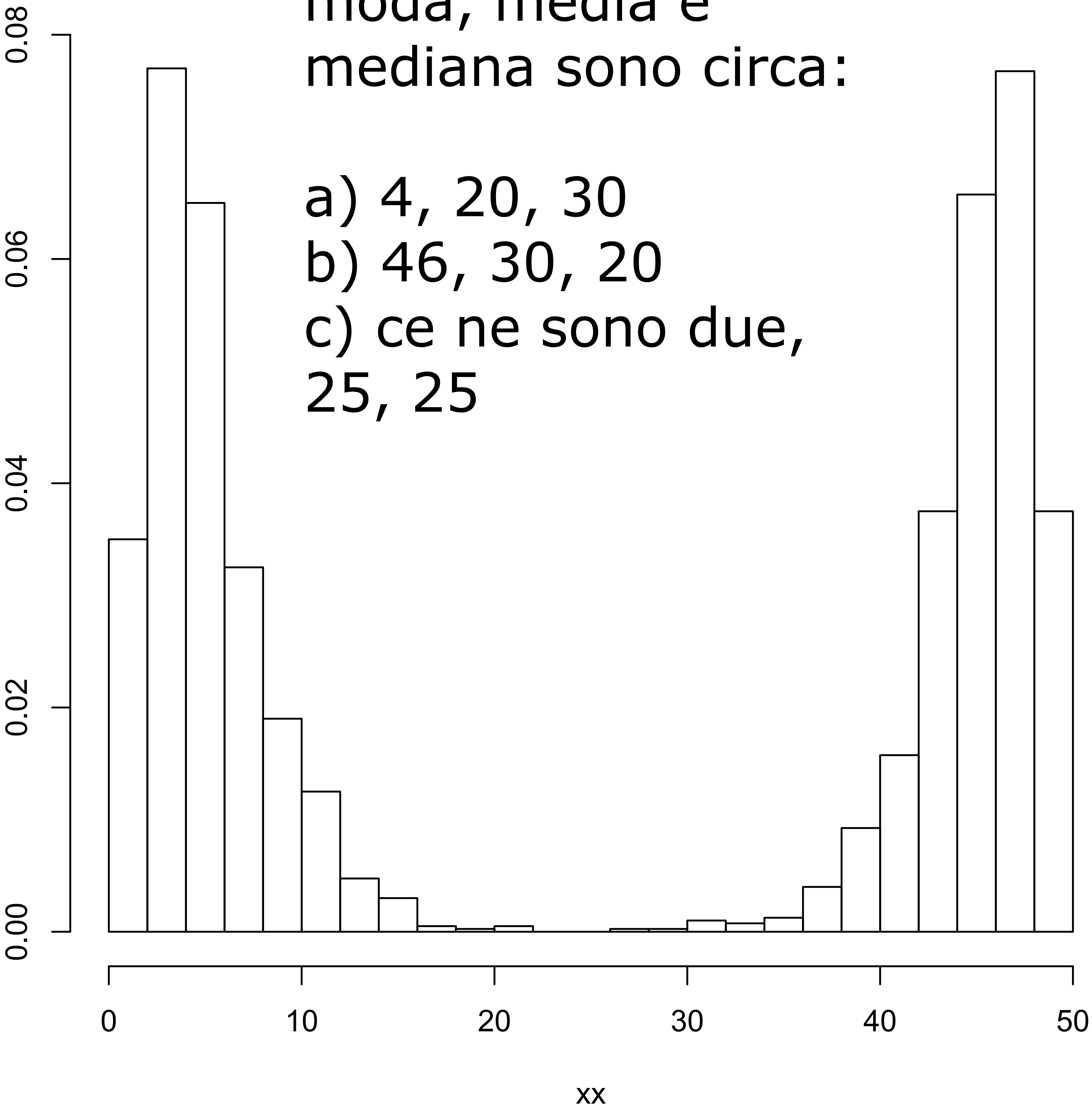
moda, media e
mediana sono circa:

a) 4, 20, 30
b) 46, 30, 20
c) ce ne sono due,
25, 25



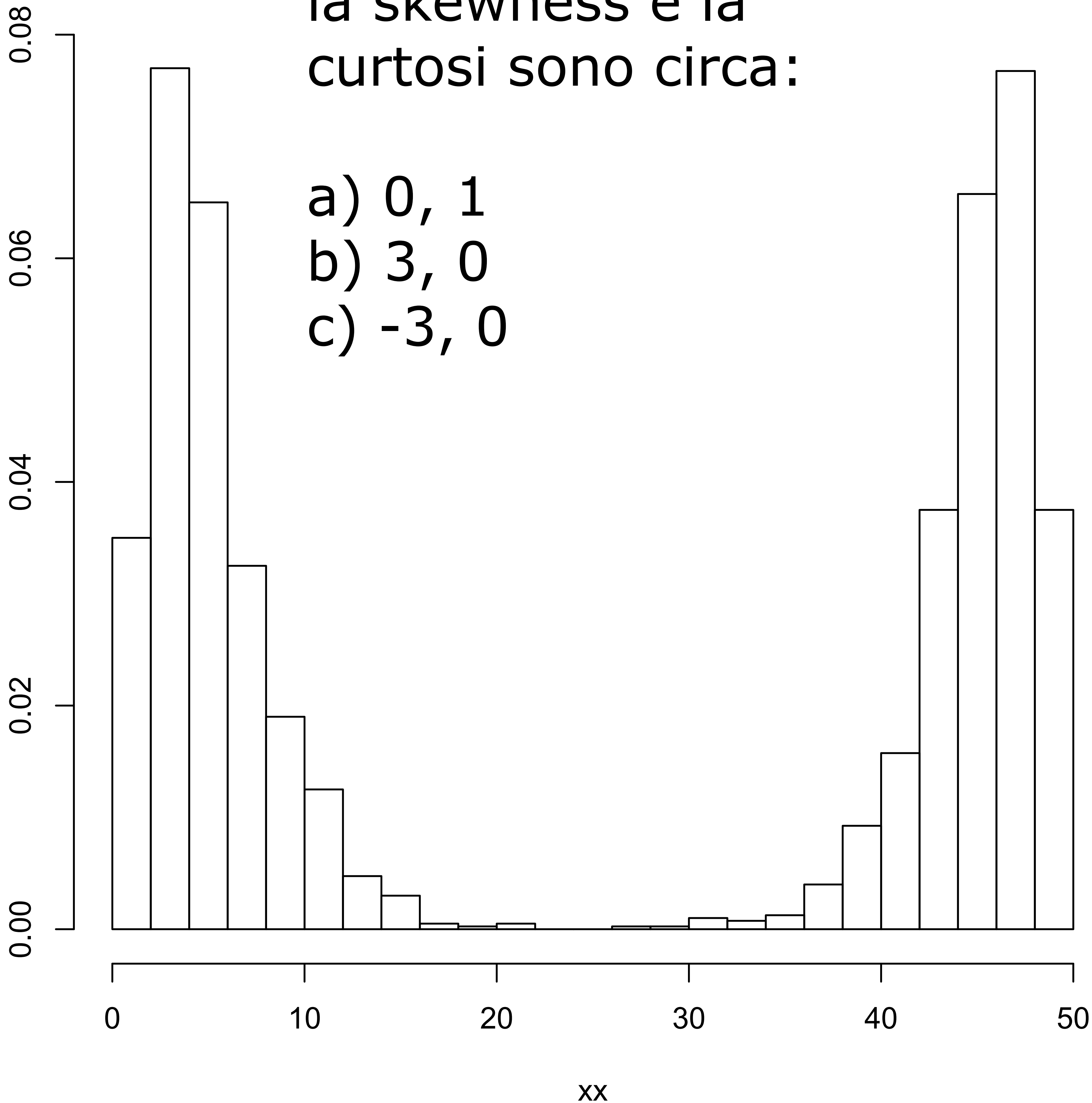
moda, media e
mediana sono circa:

a) 4, 20, 30
b) 46, 30, 20
c) ce ne sono due,
25, 25



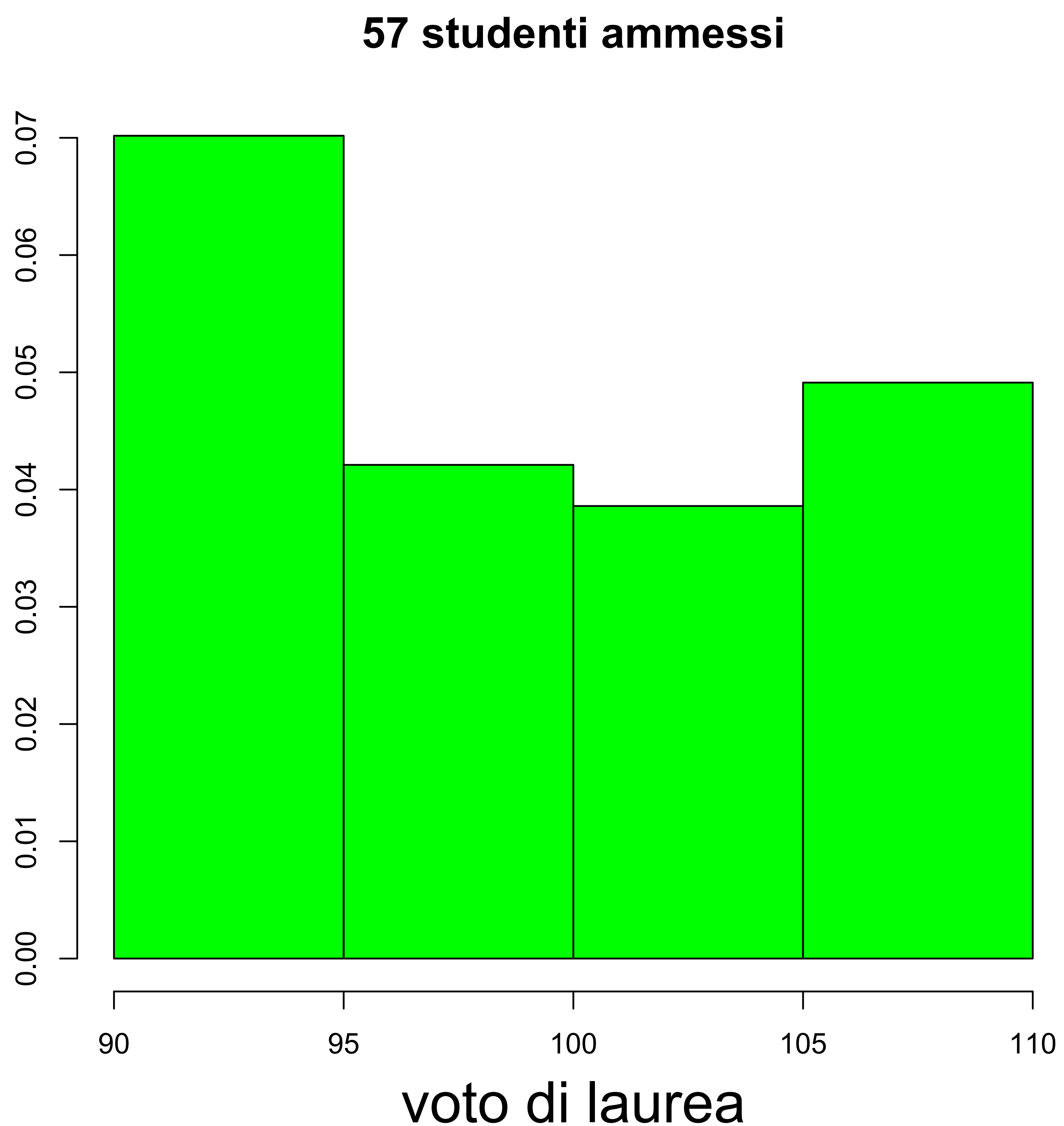
la skewness e la
curtosi sono circa:

a) 0, 1
b) 3, 0
c) -3, 0



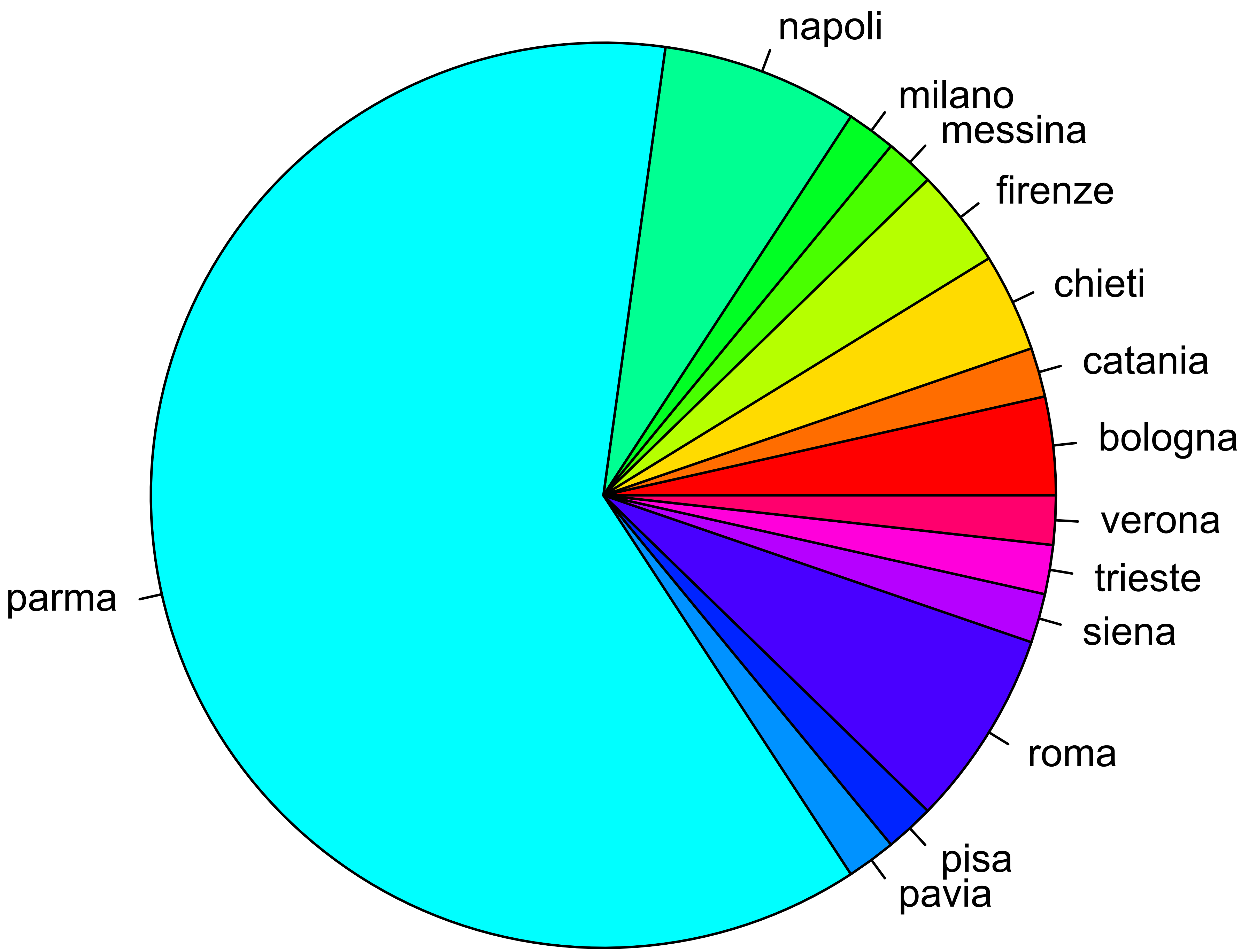
standardizzazione

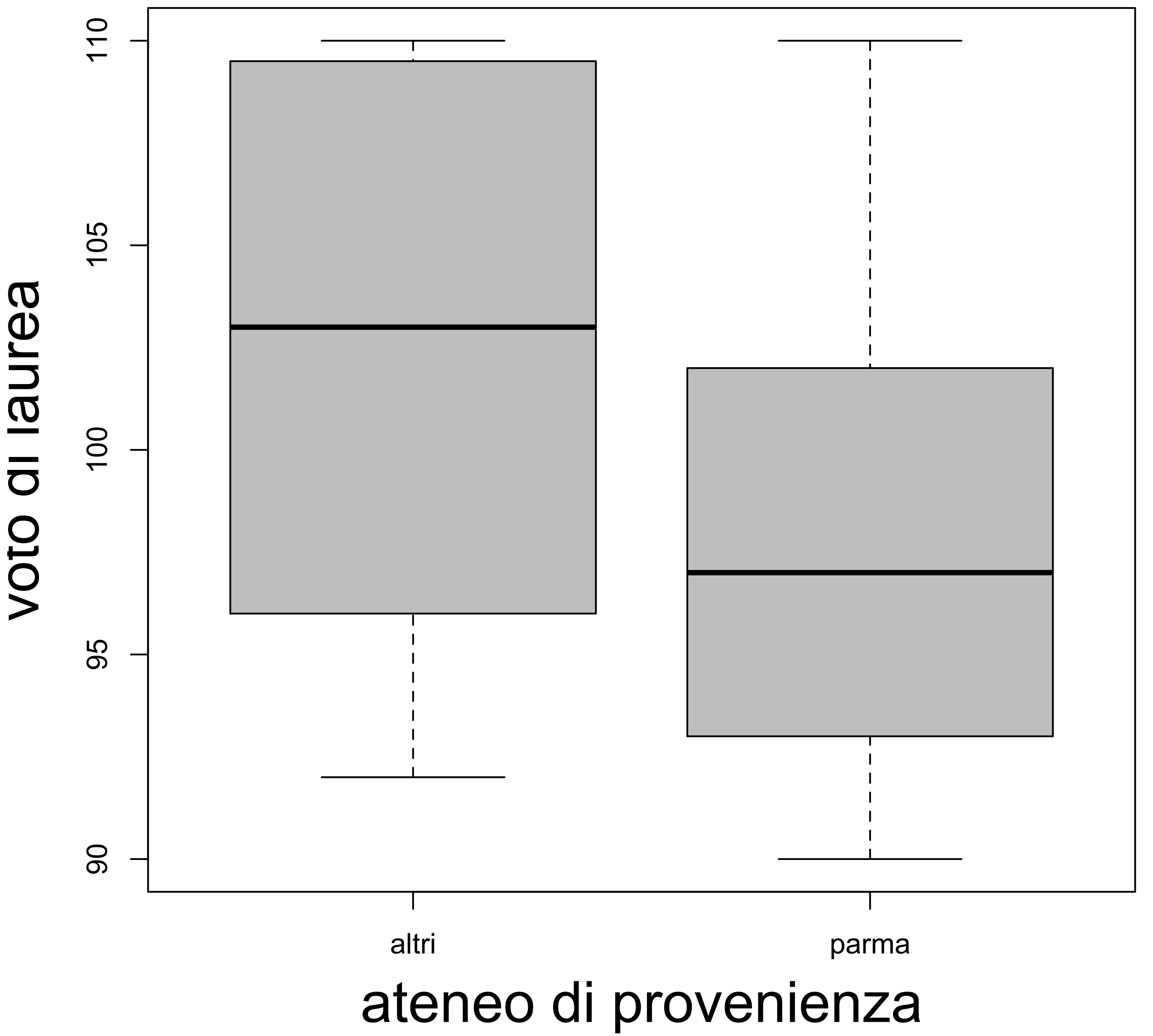
```
> d <- read.table("~/Desktop/VL.txt",
header = TRUE)

> head(d)
     UN     VL  PR
1 parma   92  1
2 parma   94  1
3 parma   93  1
4 parma   90  1
5  roma 109  0
6 parma 103  1
```



57 studenti ammessi
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
90
95
100
105
110
voto di laurea

ateneo di provenienza





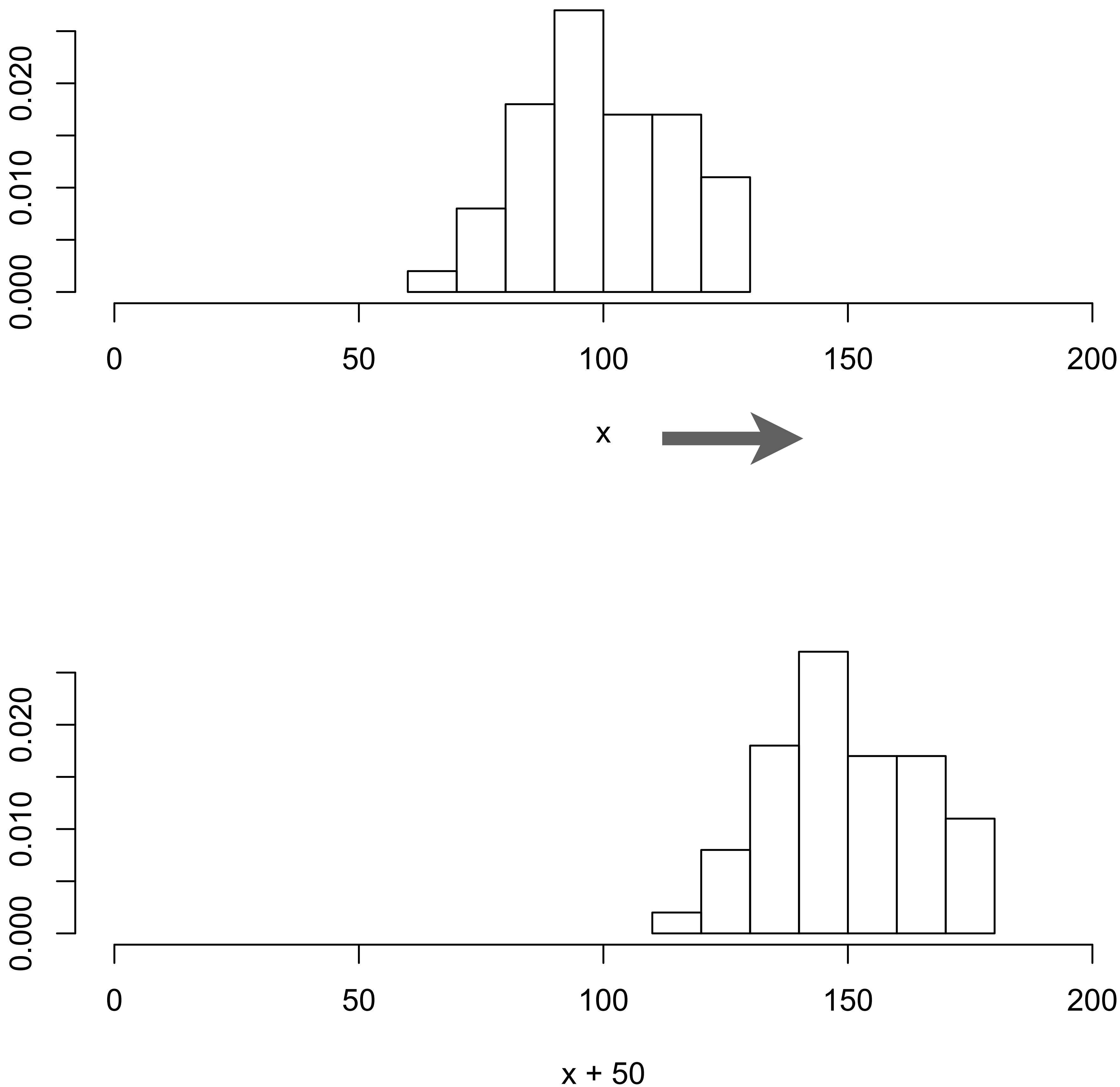
110
105
100
95
90
voto di laurea
altri
parma
ateneo di provenienza

standardizzazione

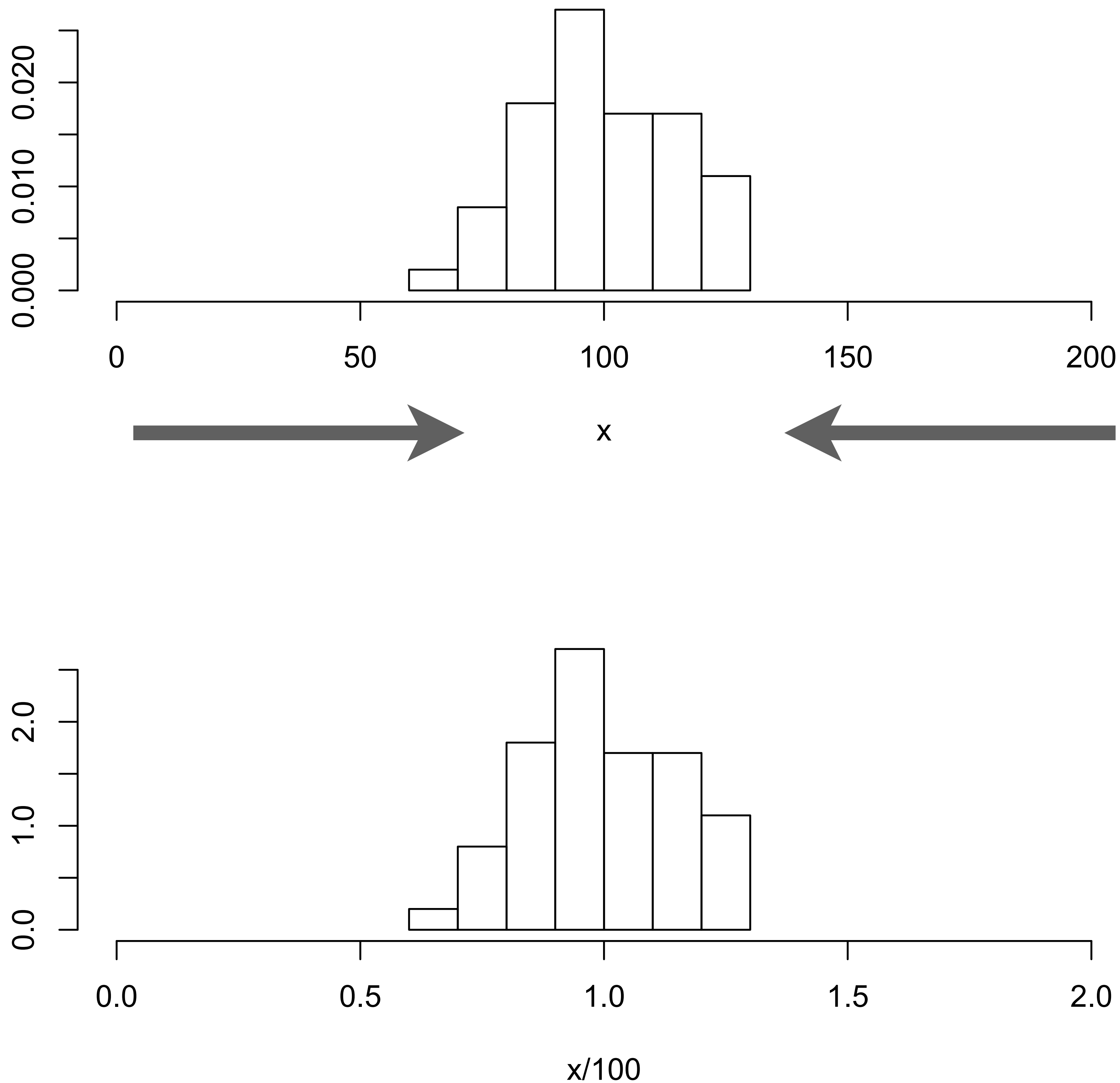
una traslazione

+

un cambio di scala



0.000
0.010
0.020
0
50
100
150
200
x
0.000
0.010
0.020
0
50
100
150
200
x + 50


0.000
0.010
0.020
0
50
100
150
200
x
0.0
1.0
2.0
0.0
0.5
1.0
1.5
2.0
x/100

esempio

Fahrenheit --> Celsius

$$C = (F - 32) / 1.8$$

traslazione

cambio di scala

```
> (10 - 32)/1.8
[1] -12.22222

> (32 - 32)/1.8
[1] 0

> (70 - 32)/1.8
[1] 21.11111

> (90 - 32)/1.8
[1] 32.22222
```

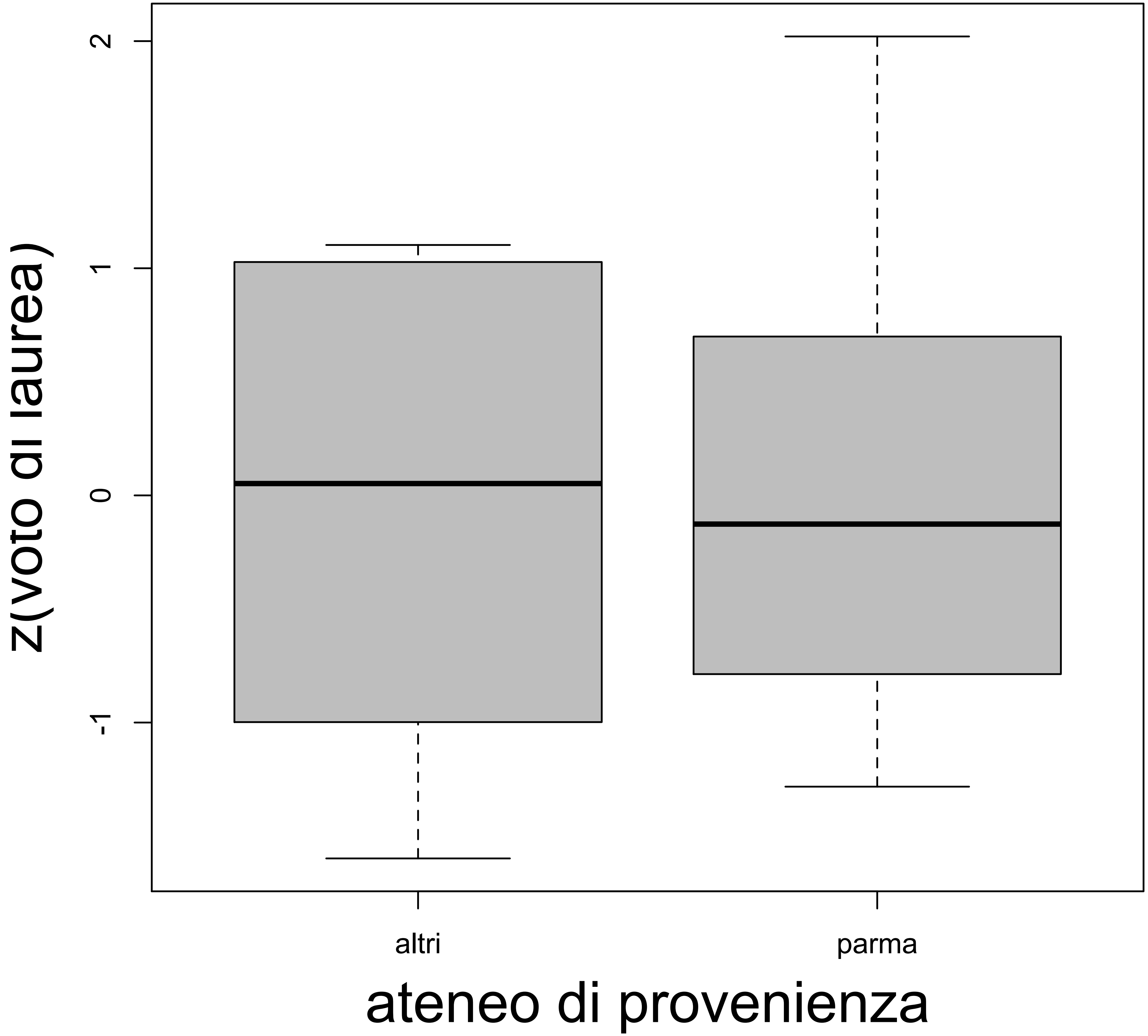
standardizzazione

traslazione: M --> 0

+

cambio di scala: DS = 1

```
> a <- d$VL[d$PR == 0]
> p <- d$VL[d$PR == 1]

> zp <- (p - mean(p))/sd(p)
> za <- (a - mean(a))/sd(a)

> boxplot(za, zp, names = c("altri",
"parma"), xlab = "ateneo di provenienza",
cex.lab = 2, col = "grey", ylab = "z(voto di
laurea)")
```



z(voto di laurea)
2
1
0
-1
altri
parma
ateneo di provenienza

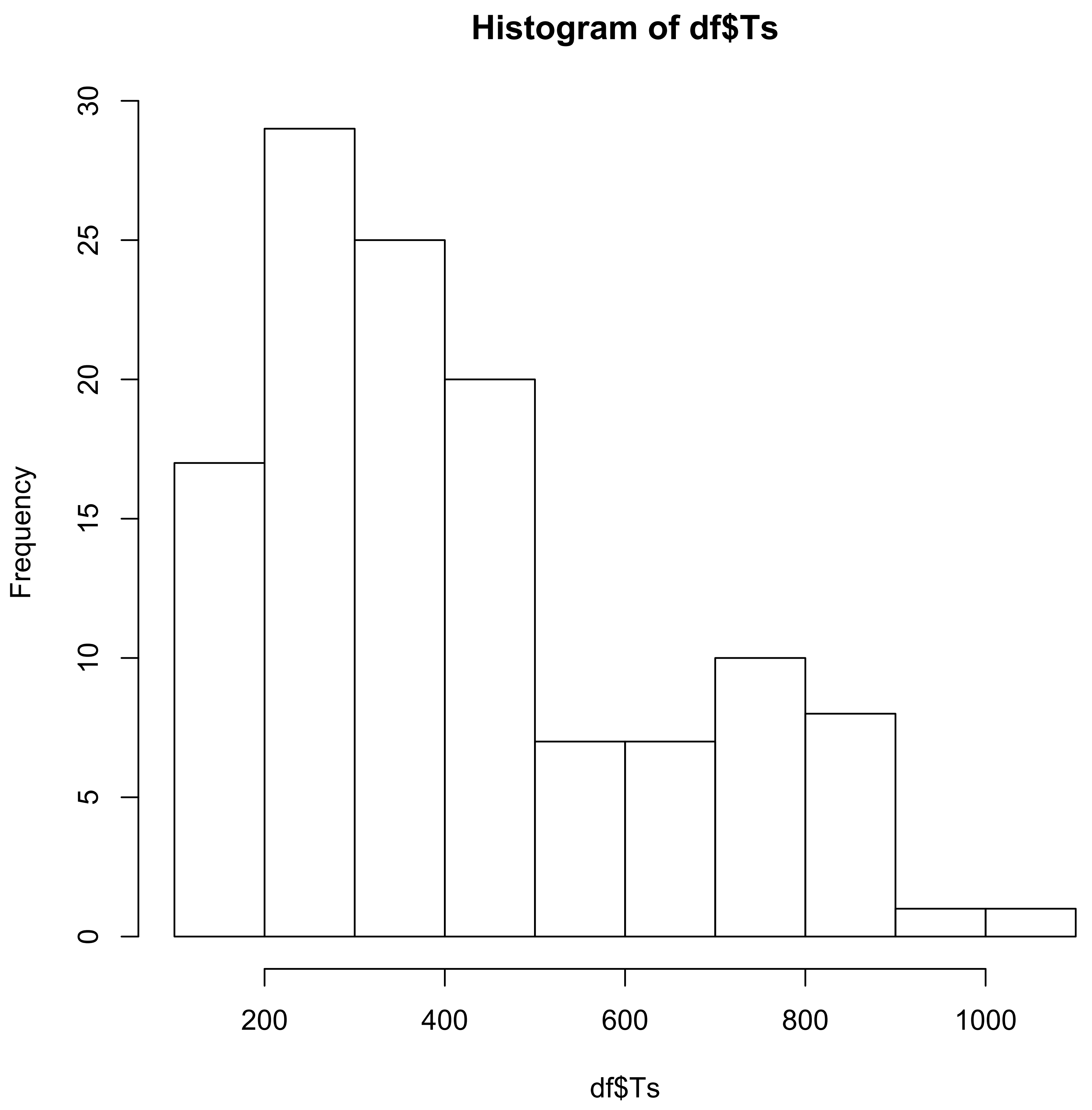
scorciatoia

```
> tapply(d$VL, d$PR, scale)
```

```
$`0`
           [,1]
 [1,]  0.95244689
 [2,] -1.29819815
 [3,]  0.05218887
 [4,]  0.65236088
 [5,]  1.10248989
 [6,]  0.65236088
 [7,]  1.10248989
 [8,] -1.29819815
 [9,] -0.84806915
[10,]  0.05218887
[11,]  1.10248989
[12,] -1.59828416
[13,] -0.69802614
[14,]  0.35227488
[15,]  1.10248989
[16,]  0.65236088
[17,] -0.39794014
[18,]  0.05218887
[19,] -1.29819815
[20,]  1.10248989
[21,] -1.44824116
[22,] -1.14815515
[23,]  1.10248989
attr(,"scaled:center")
[1] 102.6522
attr(,"scaled:scale")
[1] 6.664756

$`1`
           [,1]
 [1,] -0.95196357
 [2,] -0.62169049
 [3,] -0.78682703
 [4,] -1.28223664
 [5,]  0.86453834
 [6,] -1.28223664
 [7,]  2.02049410
 [8,]  1.19481141
 [9,]  0.03885566
[10,] -0.29141742
[11,]  0.69940180
[12,] -0.78682703
[13,]  0.53426527
[14,]  0.53426527
[15,] -1.11710010
[16,]  0.36912873
[17,] -0.95196357
[18,]  1.02967488
[19,] -0.29141742
[20,] -1.28223664
[21,]  0.69940180
[22,] -1.11710010
[23,] -1.11710010
[24,]  1.52508449
[25,]  0.36912873
[26,] -0.45655396
[27,] -0.12628088
[28,] -0.78682703
[29,] -0.12628088
[30,] -0.45655396
[31,]  2.02049410
[32,] -0.12628088
[33,]  0.03885566
[34,]  2.02049410
attr(,"scaled:center")
[1] 97.76471
attr(,"scaled:scale")
[1] 6.055595
```

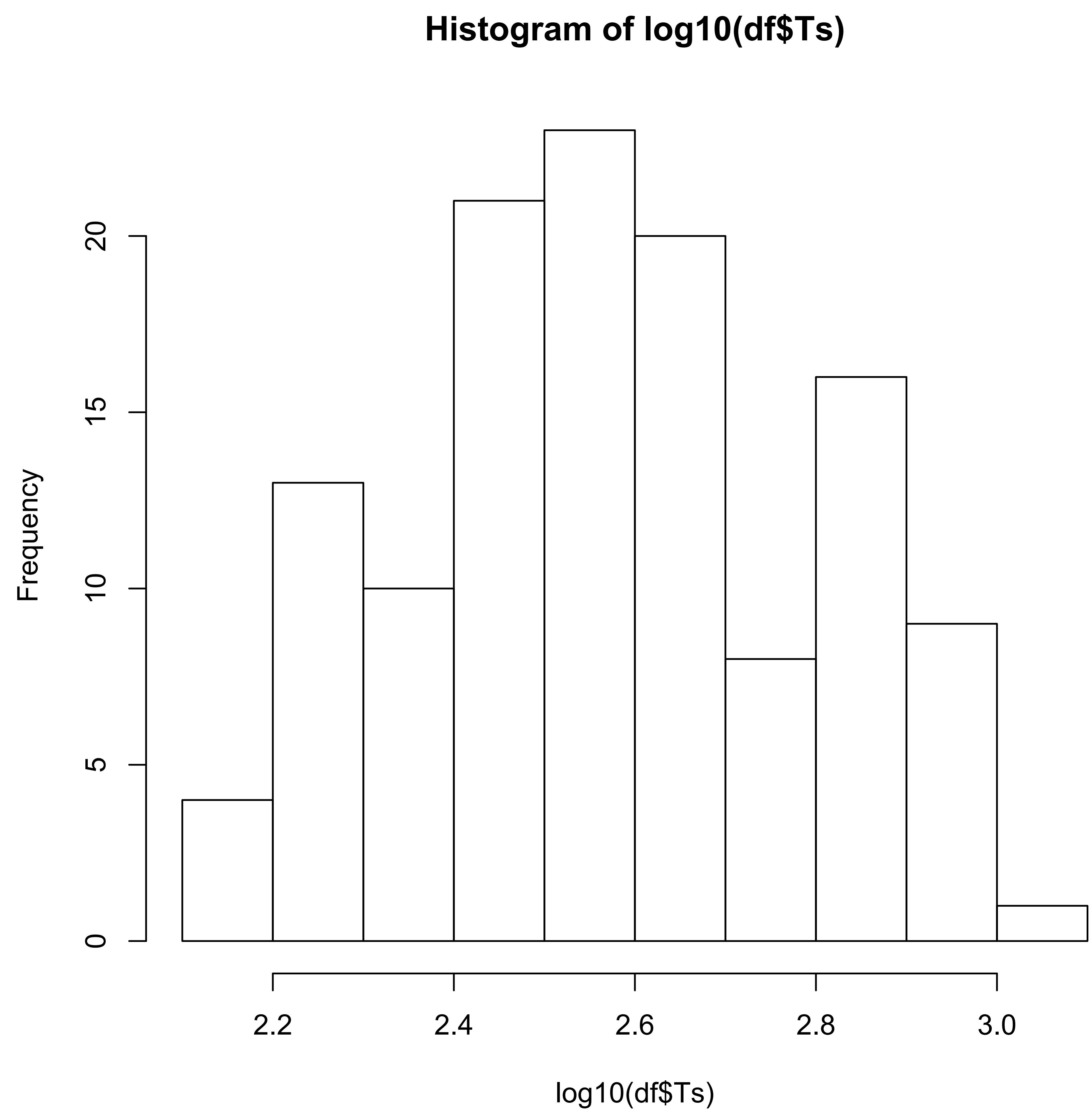
trasformazione logaritmica

```
> df <- read.table("~/Desktop/dati completi.txt",
header = TRUE)

> head(df)

  OvsR Sex HAND        RVF     LIKF  Ts SPAF  Eng compito
1    O   f   dx -3.9873143 1.500000 190  1.0    1       c
2    O   f   dx -0.4353741 4.166667 413  5.0    2      cd
3    R   f   dx  3.6857530 3.500000 491  2.5    1      cd
4    O   m   dx -1.4842264 5.666667 277  6.0    1       c
5    R   m   dx -1.3211927 6.000000 480  7.0    2      cd
6    R   m   dx -0.2262290 4.166667 265  3.0    2       c

> hist(df$Ts)
```

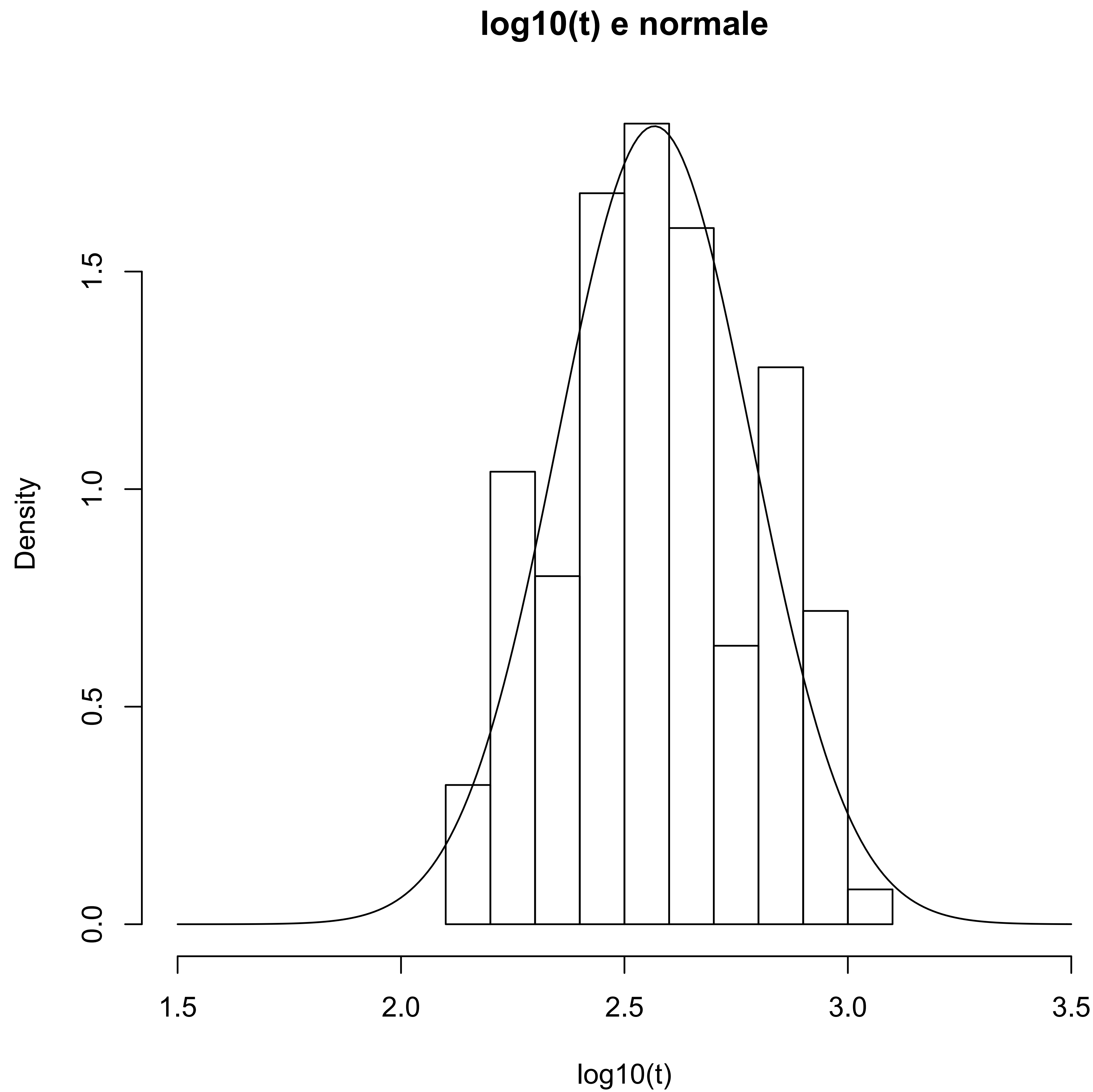


Histogram of df$Ts
Frequency
0
5
10
15
20
25
30
200
400
600
800
1000
df$Ts

```
> hist(log10(df$Ts))
```

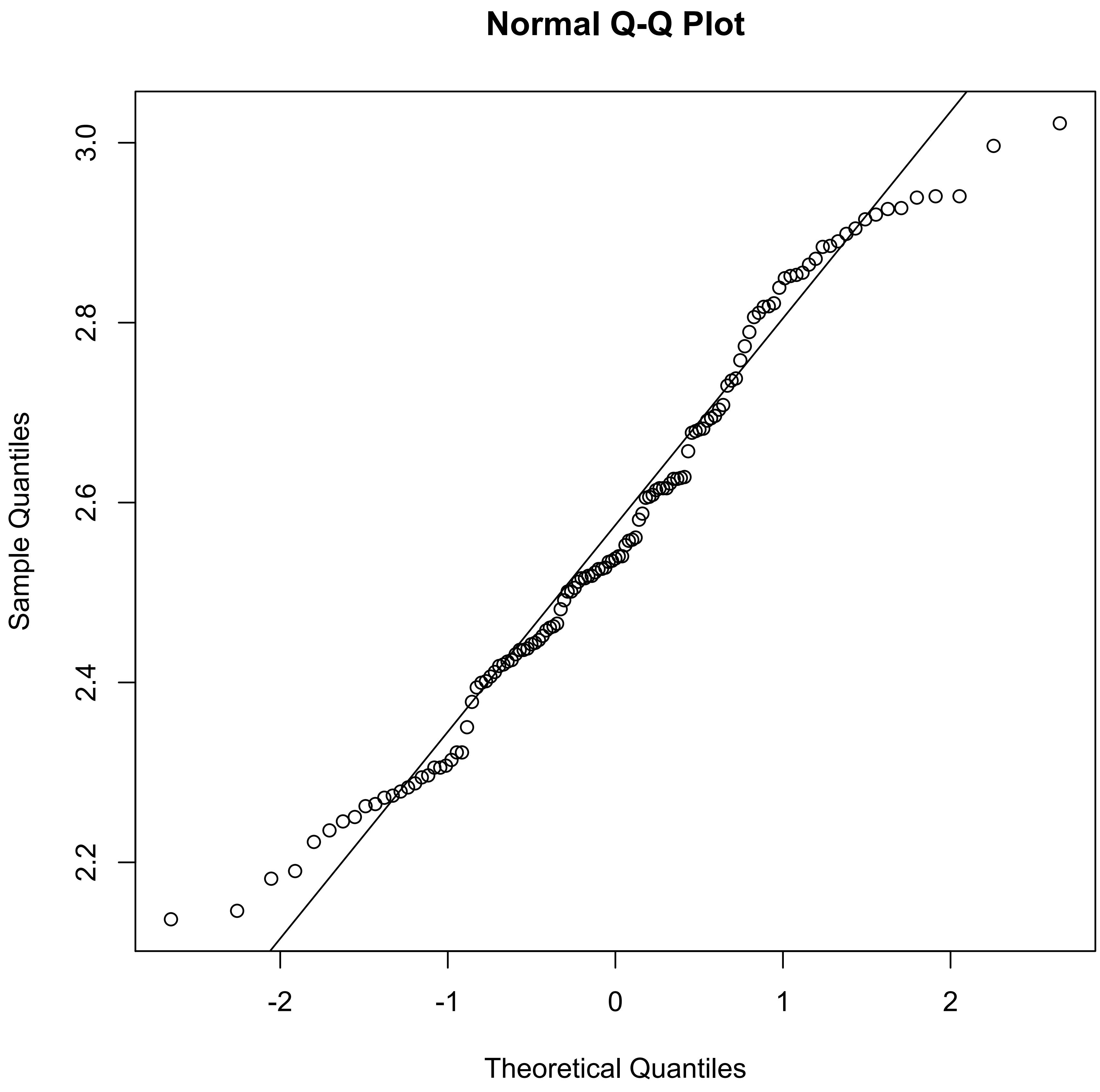

Histogram of log10(df$Ts)
Frequency
0
5
10
15
20
2.2
2.4
2.6
2.8
3.0
log10(df$Ts)

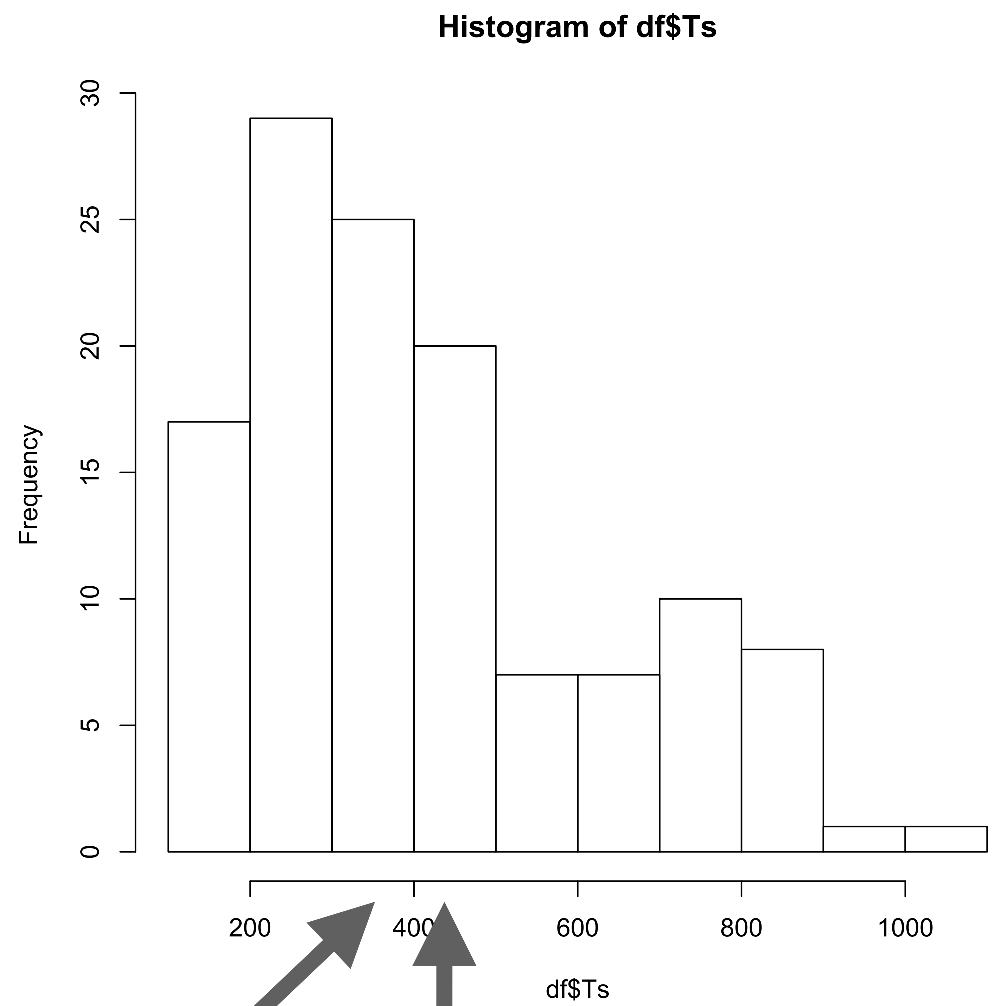
```
> library(moments)

> skewness(df$Ts)
[1] 0.9111309

> skewness(log10(df$Ts))
[1] 0.1454733

> kurtosis(df$Ts)
[1] 2.884776

> kurtosis(log10(df$Ts))
[1] 2.129199
```

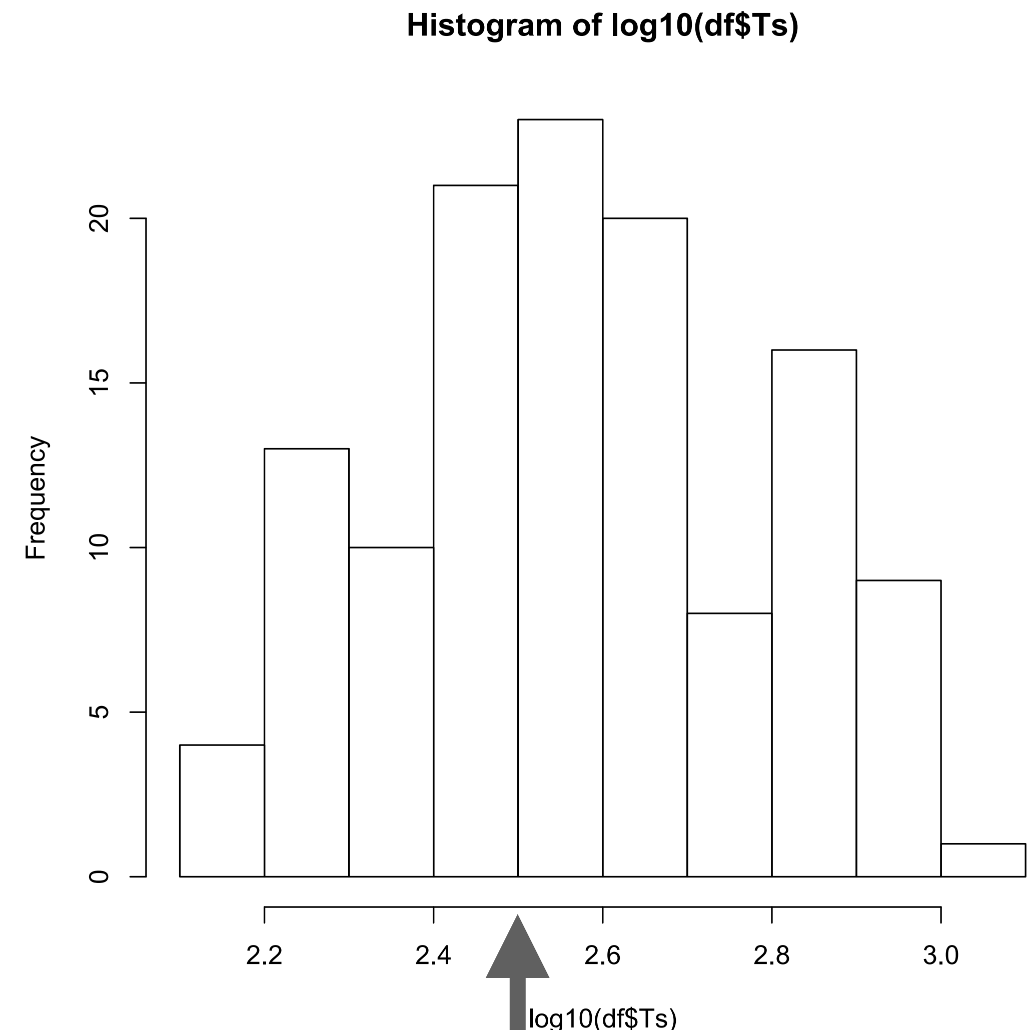

log10(t) e normale
Density
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
3.5
log10(t)

plot quantile-quantile normale

```
> qqnorm(log10(df$Ts))
> qqline(log10(df$Ts))
```


Normal Q-Q Plot
Sample Quantiles
3.0
2.8
2.6
2.4
2.2
-2
-1
0
1
2
Theoretical Quantiles






```
> mean(df$Ts)
[1] 418.3663
```

```
> mean(log10(df$Ts))
[1] 2.567262
```

```
> mg <- 10^(mean(log10(df$Ts)))
> mg
[1] 369.1999
```

grafici in R (elementi di)

menu

vedi:

Packages & Data
- **Package Manager**
- **Package Installer**
- **Data Manager**

graphics

funzioni di basso livello per gli elementi grafici

funzioni di alto livello per grafici preconfezionati

tipica maniera di procedere

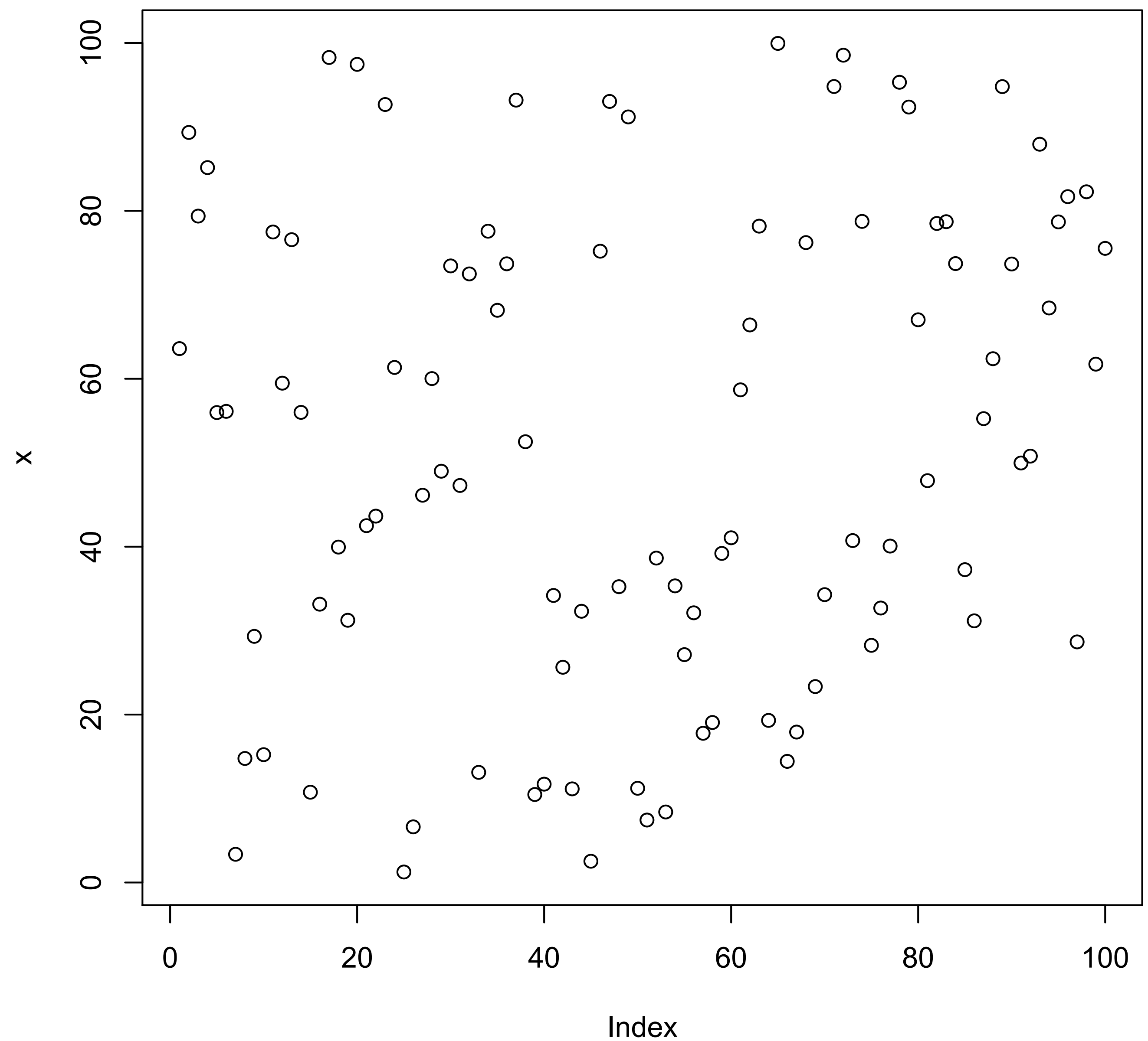
generare i grafici che mi servono con funzioni di alto livello

“annotare” usando ulteriori funzioni di basso livello

esempio

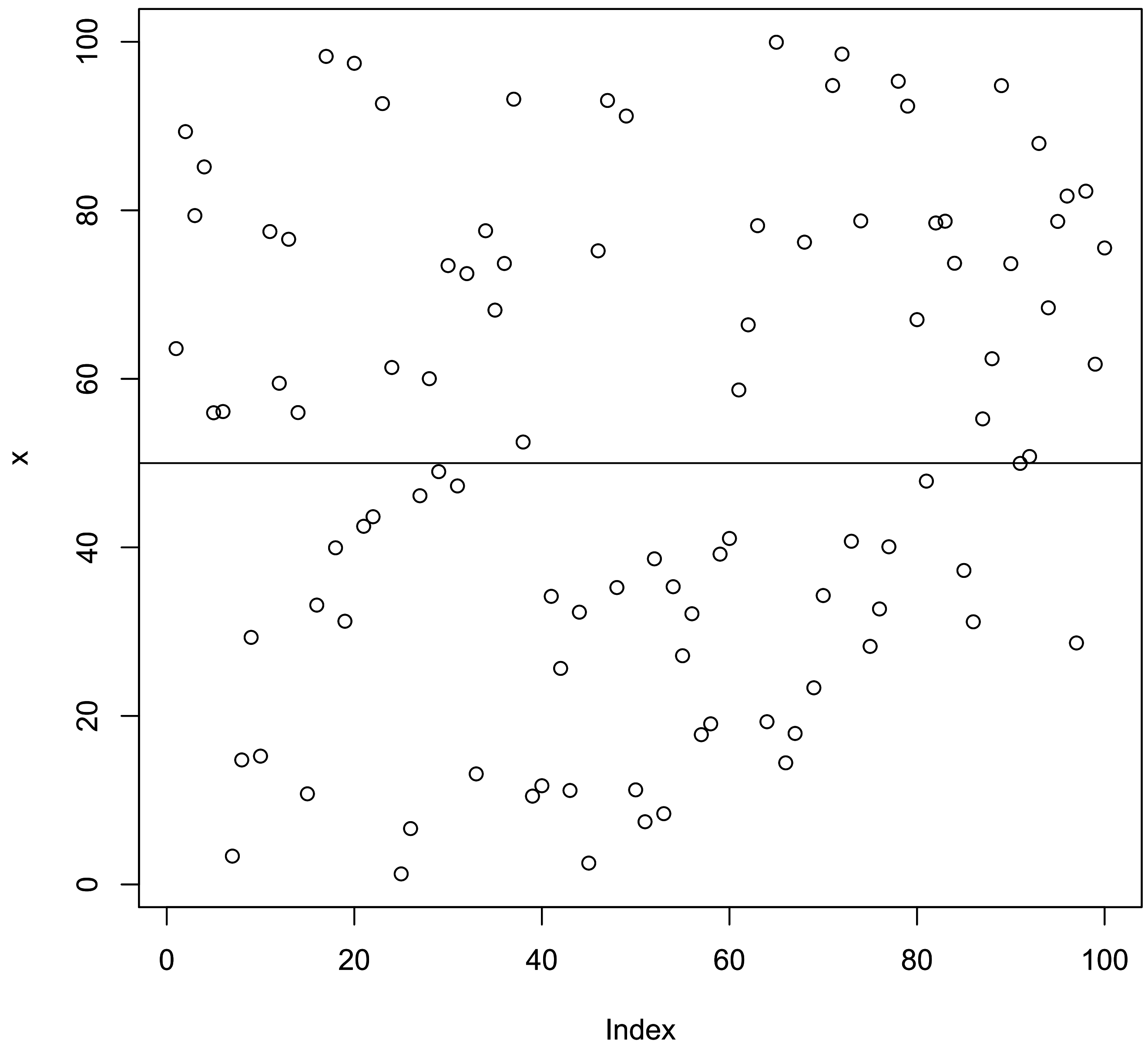
```
> x <- runif(100, 0, 100)
```

```
> plot(x)
```



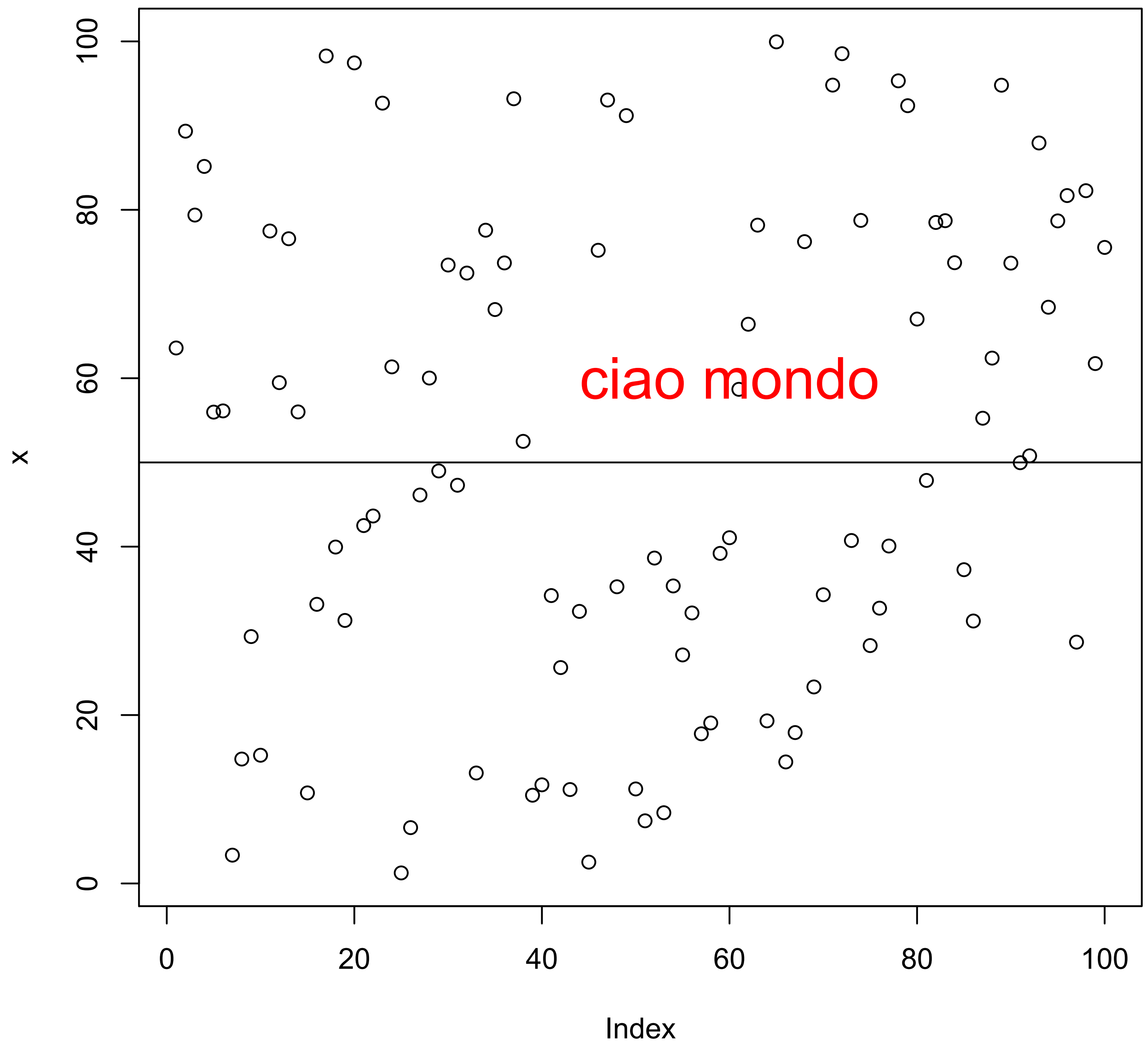
esempio

```
> x <- runif(100, 0, 100)

> plot(x)

> abline(h = 50)
```



esempio

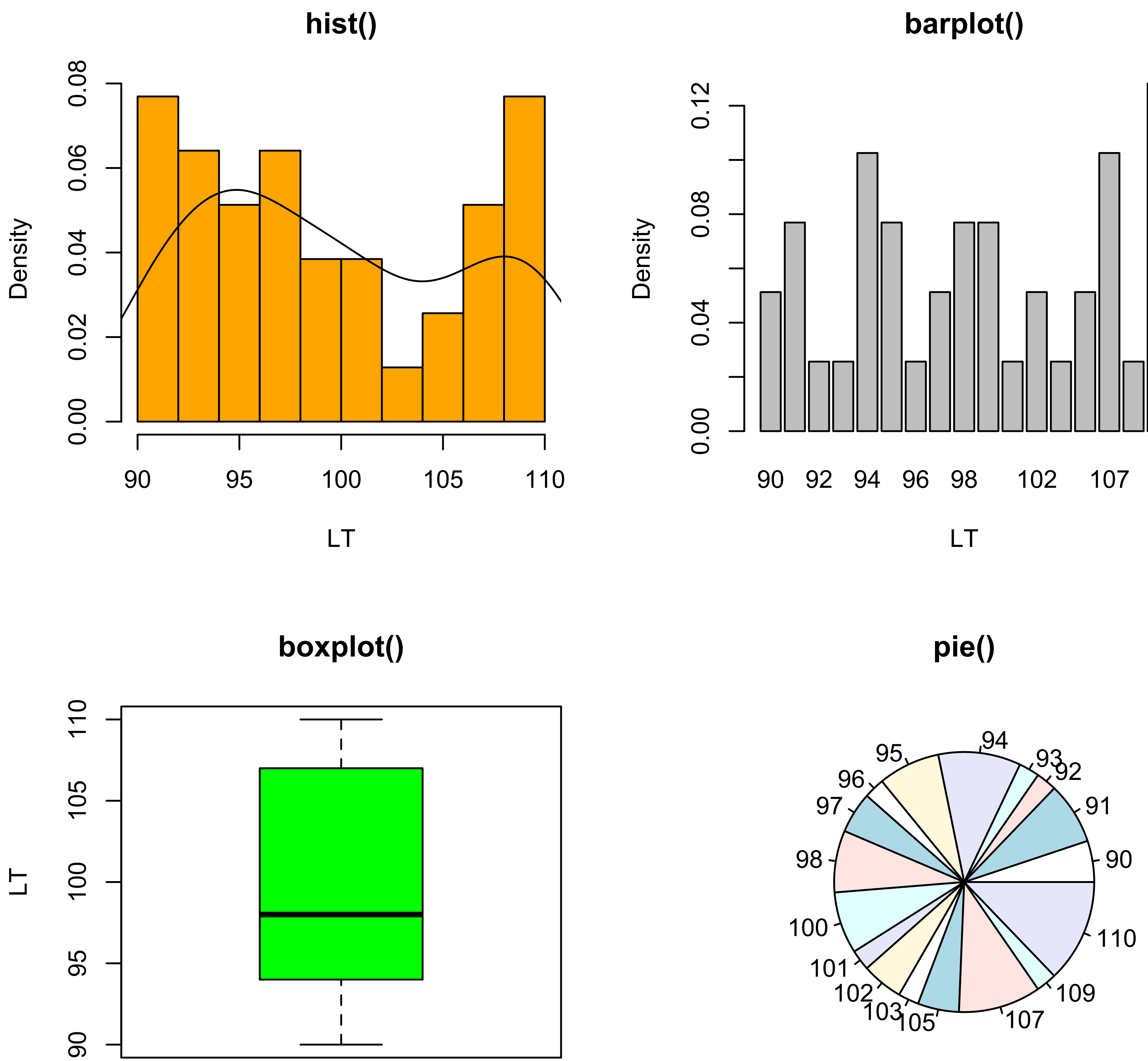
```
> plot(x)
> abline(h = 50)
> text(60, 60,
"ciao mondo",
col = "red",
cex = 2)
```



farsi un'idea

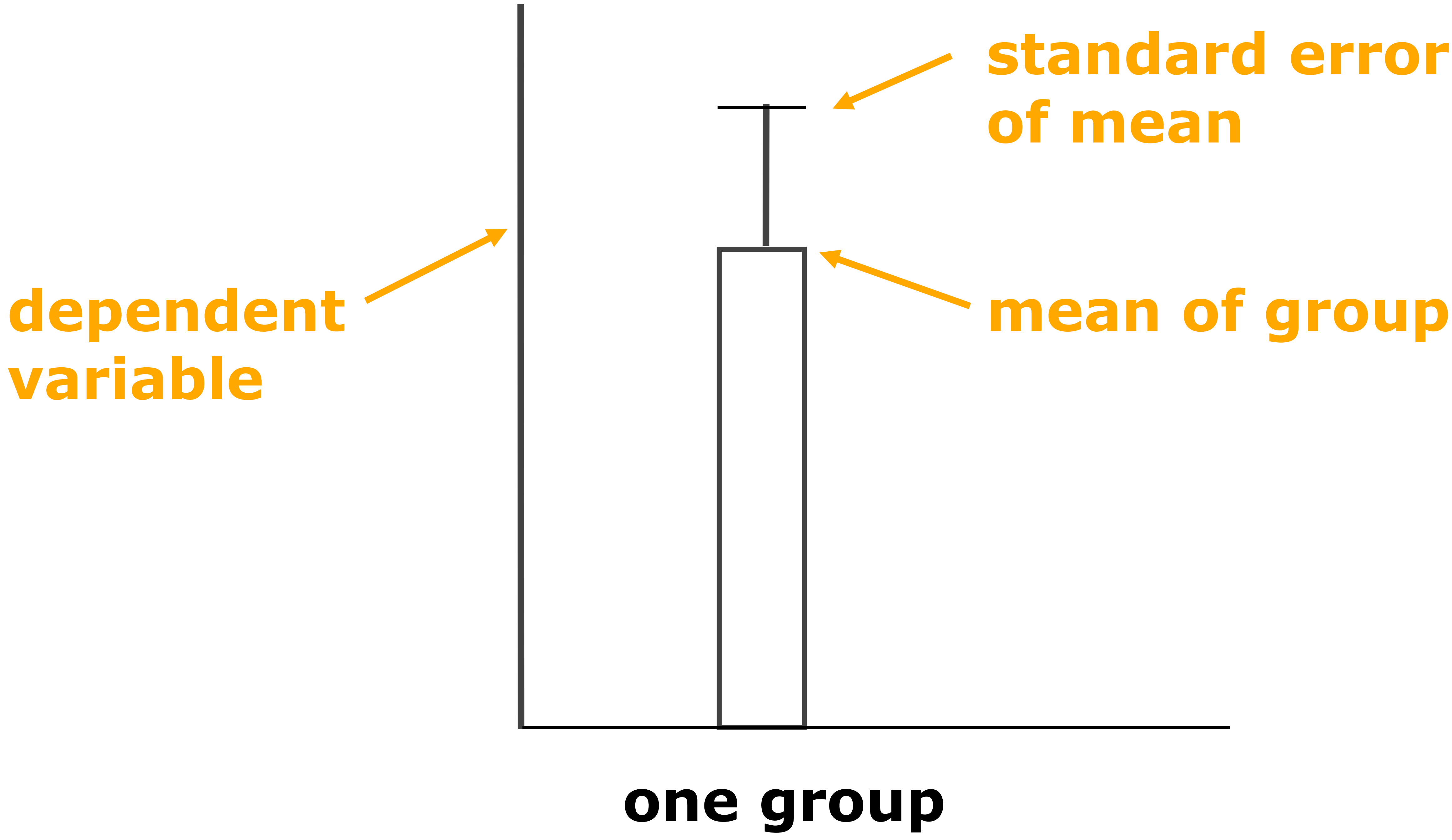
```
> demo(graphics)
```

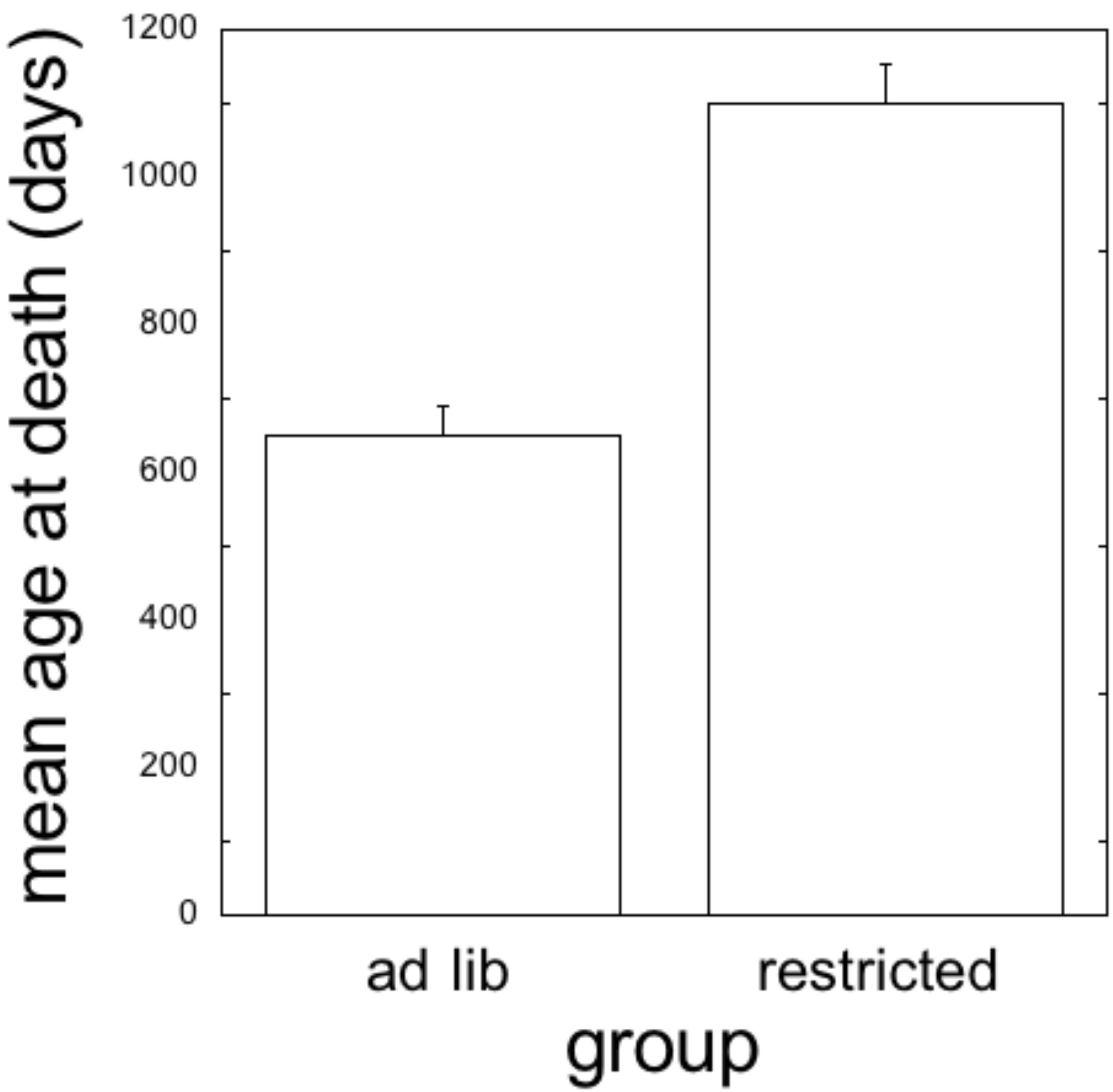
```
d <- read.table("~/Desktop/LT.txt", header = TRUE)
op <- par(mfrow = c(2, 2))
hist(d$LT, prob = TRUE, main = "hist()", xlab =
"LT", col = "orange")
lines(density(d$LT, col = "blue"))
barplot(table(d$LT)/length(d$LT), main =
"barplot()", xlab = "LT", ylab= "Density")
boxplot(d$LT, ylab = "LT", main = "boxplot()", col =
"green")
pie(table(d$LT)/length(d$LT), main = "pie()")
par(op)
```



hist()
Density
0.00
0.02
0.04
0.06
0.08
90
95
100
105
110
LT
barplot()
Density
0.00
0.04
0.08
0.12
90
92
94
96
98
102
107
LT
boxplot()
LT
90
95
100
105
110
pie()
94
93
92
91
90
95
96
97
98
100
101
102
103
105
107
109
110

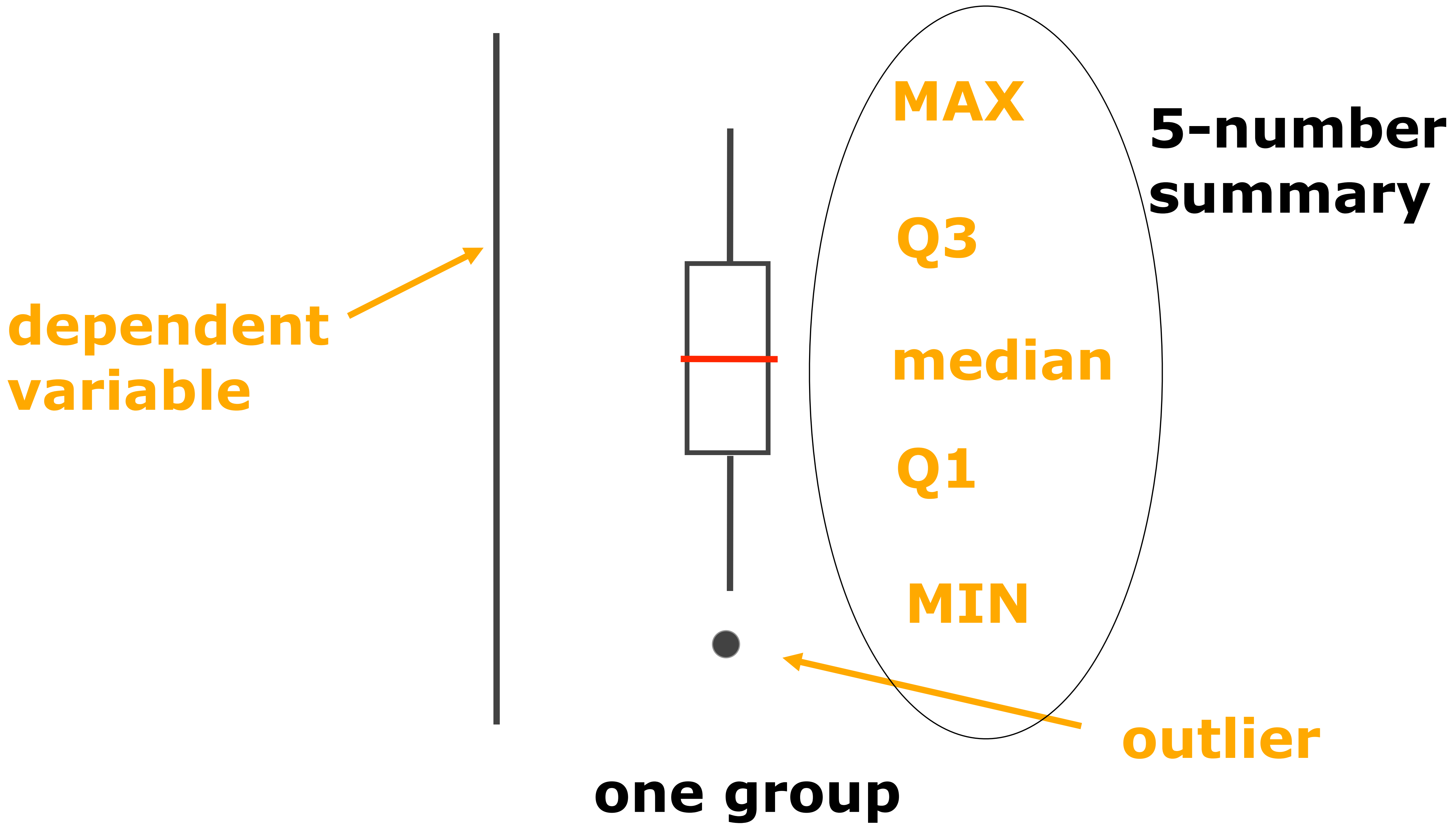
diagramma a barre

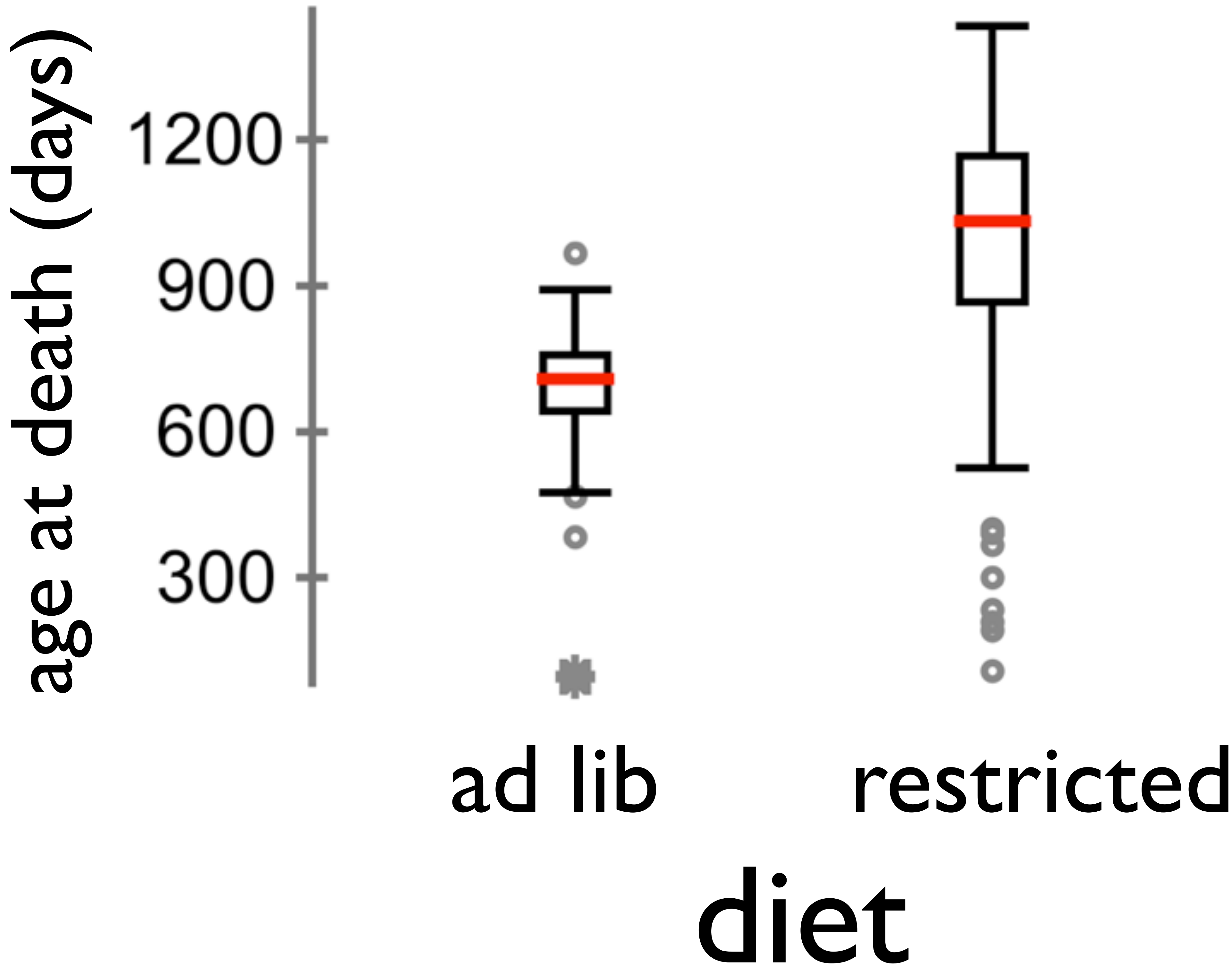




mean age at death (days)
1200
1000
800
600
400
200
0
ad lib
restricted
group

box-plot


dependent
variable
MAX
Q3
median
Q1
MIN
5-number
summary
outlier
one group


age at death (days)
1200
900
600
300
ad lib
restricted
diet