i quantili di una distribuzione

definizione di quantile

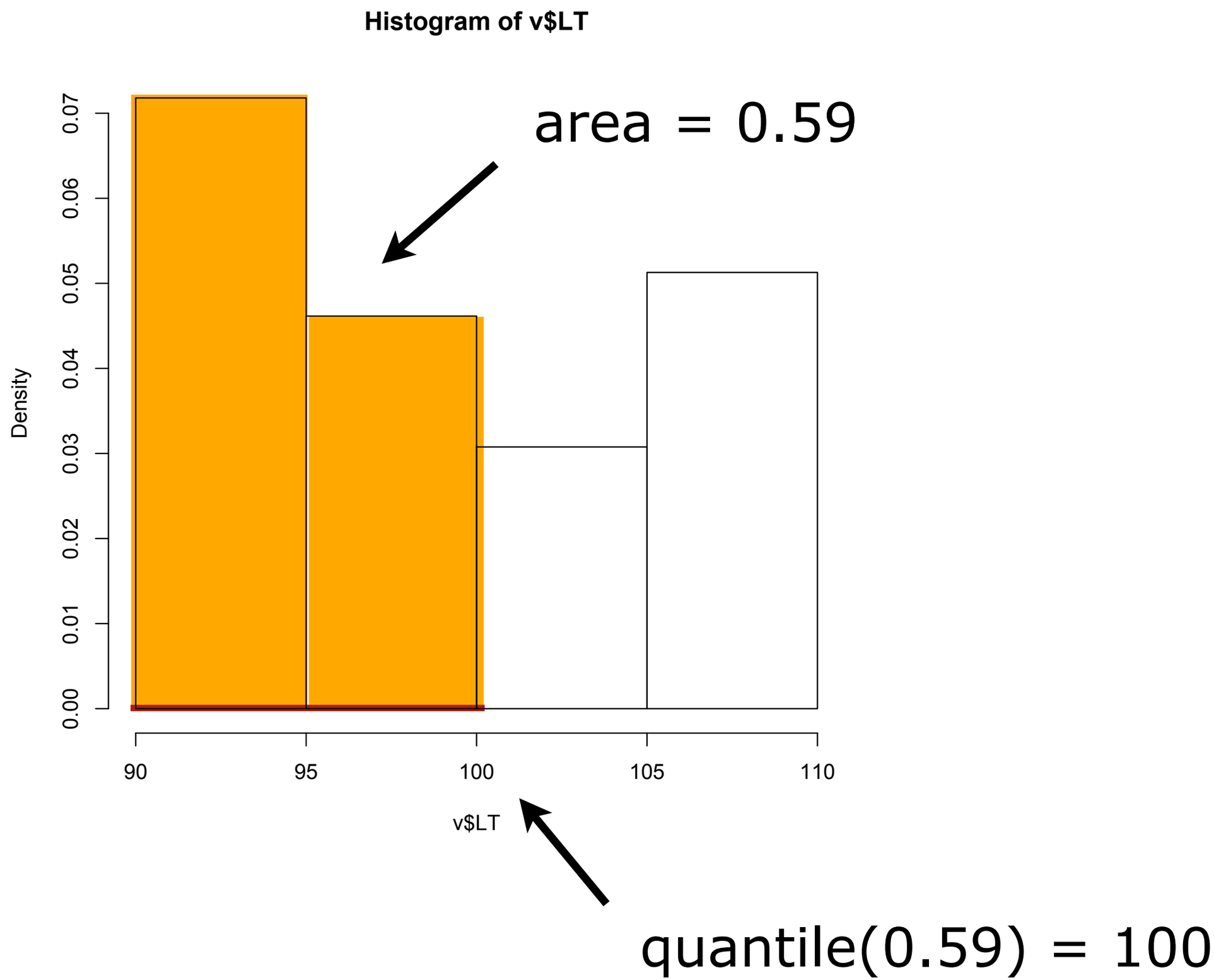
Il quantile q di ordine p $[0 < p < 1]$ è il valore di una distribuzione che ha la proprietà di dividerla in due parti le cui frequenze relative sono p e $1 - p$.

in parole povere

un valore x che sottende un'area relativa p nell'istogramma

p è l'area che corrisponde all'intervallo x- min(x)

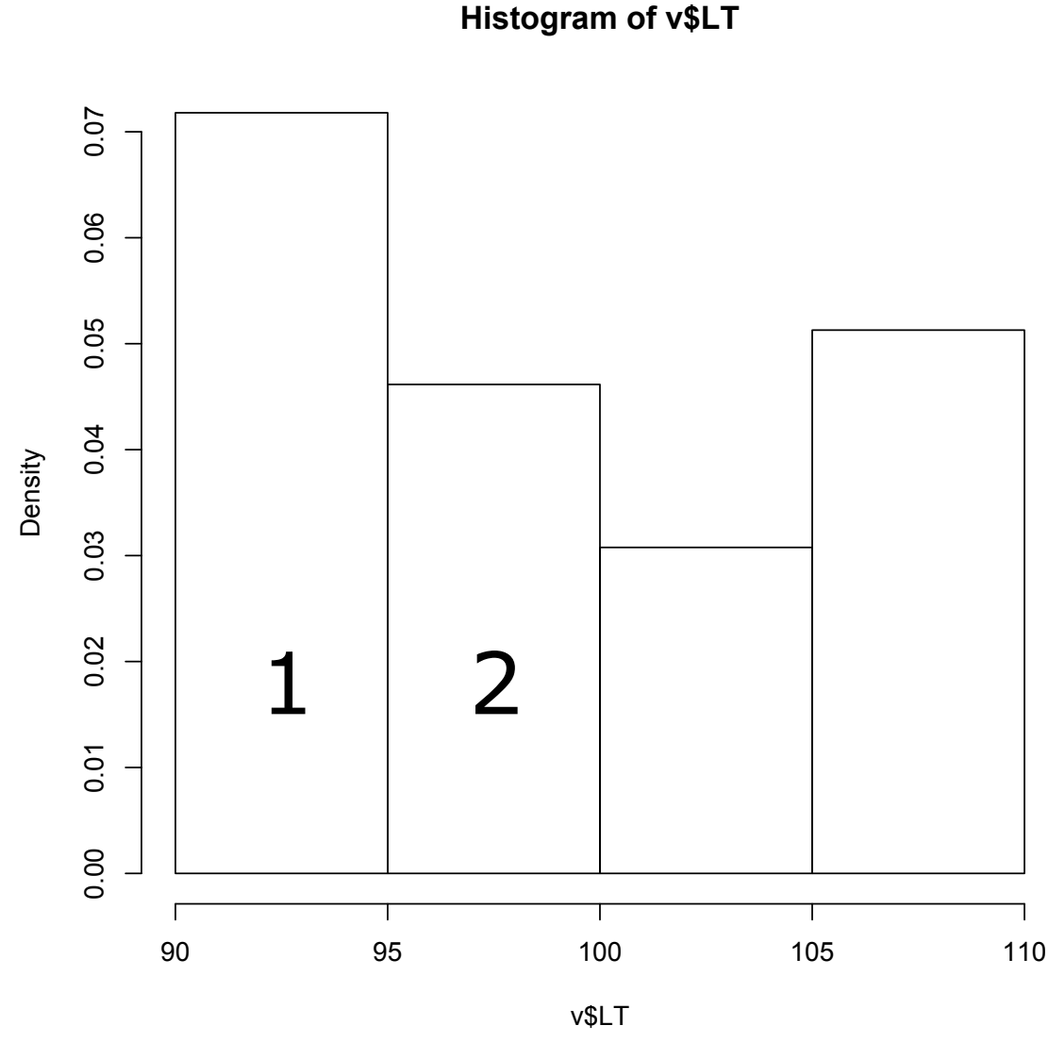
x = quantile di ordine p


Histogram of v$LT
area = 0.59
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
90
95
100
105
110
v$LT
quantile(0.59) = 100

```
v
         90          91          92          93          94
0.05128205 0.07692308 0.02564103 0.02564103 0.10256410
         95          96          97          98         100
0.07692308 0.02564103 0.05128205 0.07692308 0.07692308
        101         102         103         105         107
0.02564103 0.05128205 0.02564103 0.05128205 0.10256410
        109         110
0.02564103 0.12820513
```

```
> c12 <- sum(t[1:10])
> c12
```

```
[1] 0.5897436
```



quantile()

```
> quantile(v$LT, 0.5897436)

58.97436%
 100.4103

> quantile(v$LT, 1 - (60/n))

23.07692%
       94
```

primi cugini

percentile o centile = analogo a quantile ma numero intero, interpretabile come %

quintile, decile, ventile = come sopra ma dividono la distribuzione in 5, 10 o 20 parti uguali

quartili

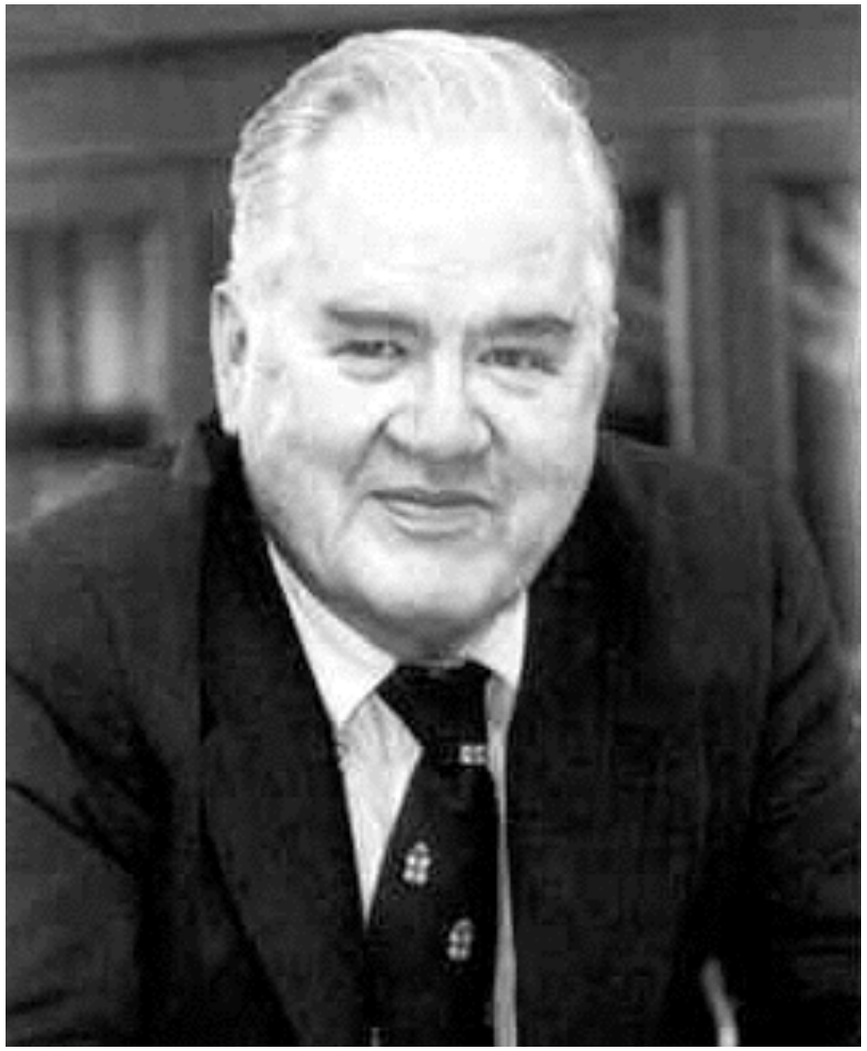
Q1 = quantile di ordine 0.25

Q2 = quantile di ordine 0.5
= mediana

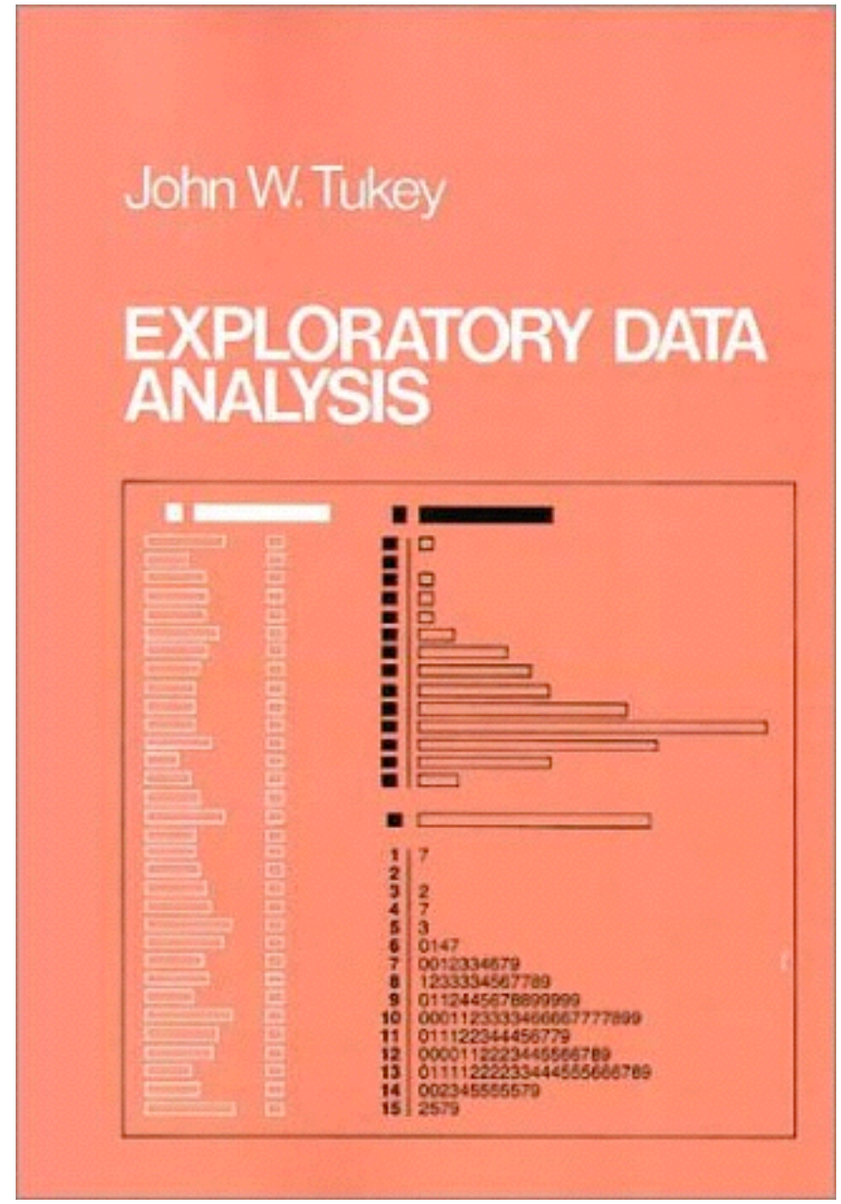
Q3 = quantile di ordine 0.75

summary()e fivenum()

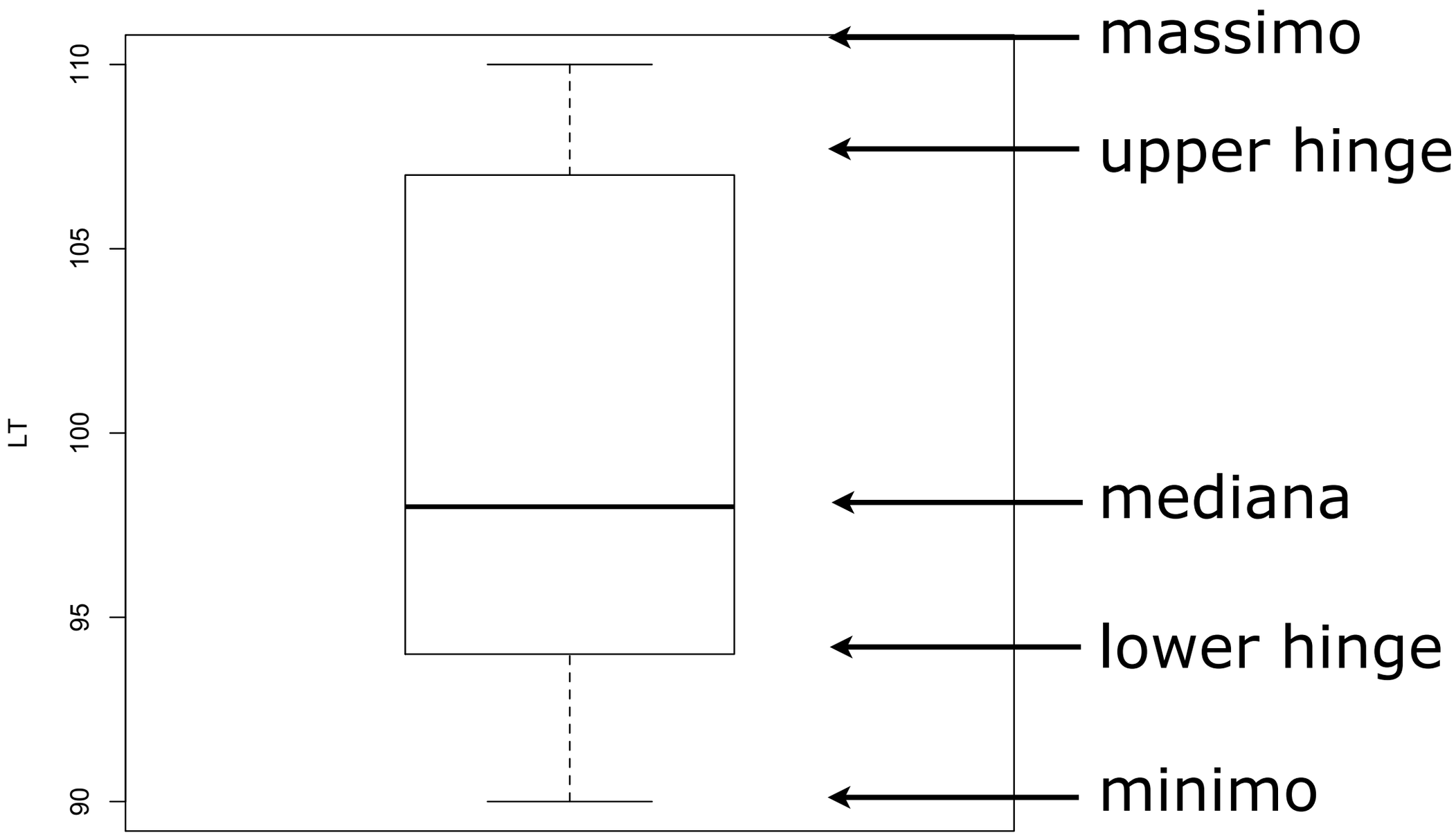
```
> summary(v$LT)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  90.00   94.00   98.00   99.69  106.50  110.00

> fivenum(v$LT)

[1]  90  94  98 107 110
```



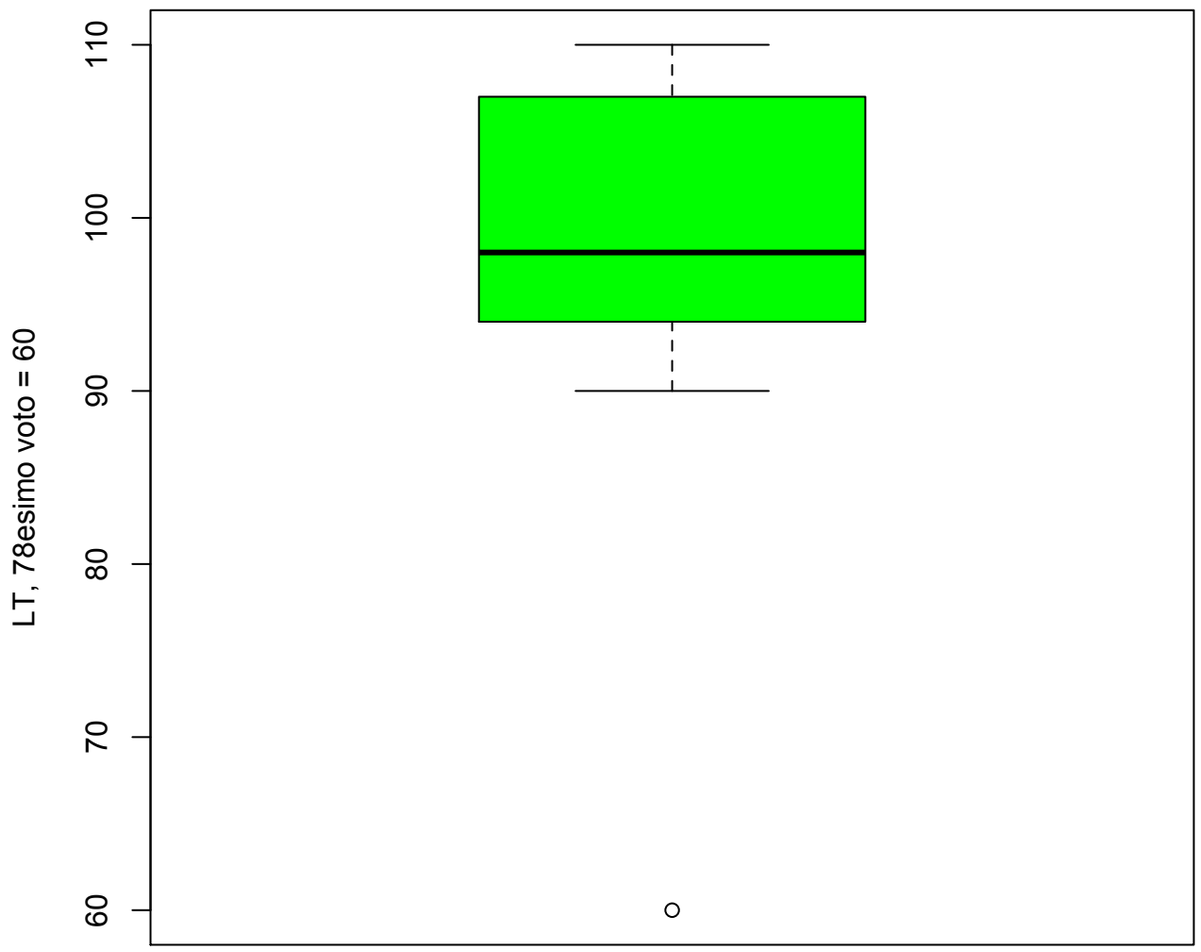
John Tukey



```
> boxplot(v$LT, ylab = "LT")
```

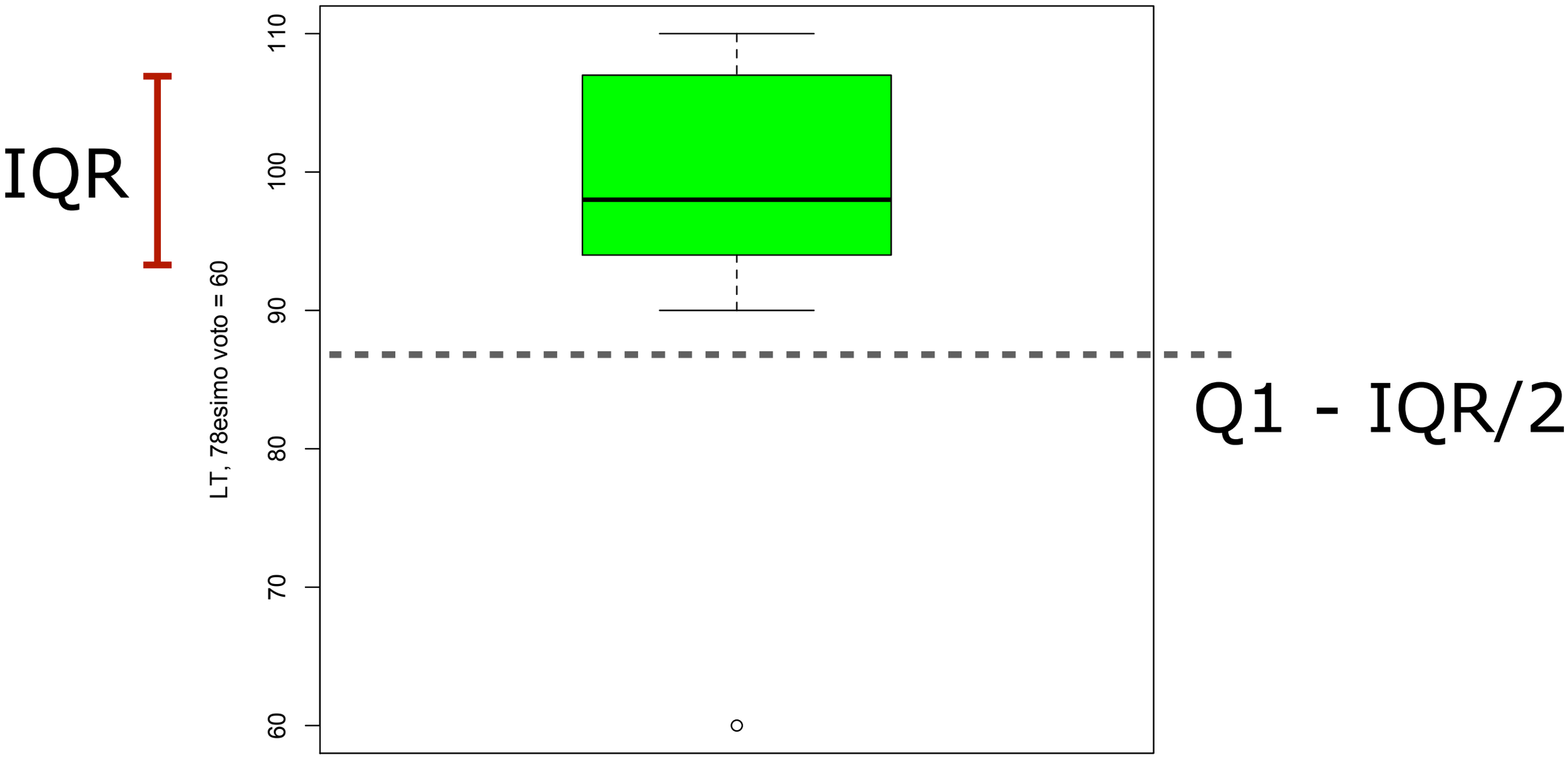


```
> v$LT[78] <- 60
> boxplot(v$LT, ylab = "LT, 78-esimo voto = 60")
```



IQR()

```
> IQR(v$LT)

[1] 12.5

> summary(v$LT)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  90.00   94.00   98.00   99.69  106.50  110.00 

> 60 < quantile(v$LT, 0.25) - IQR(v$LT)/2

TRUE
```


IQR
LT, 78esimo voto = 60
110
100
90
80
70
60
Q1 - IQR/2

tendenza centrale

tendenza centrale

La tendenza dei dati quantitativi a raggrupparsi attorno a un valore tipico. Controintuitivamente, questo valore non rappresenta necessariamente il centro geometrico o di massa della distribuzione, ma dipende invece dalla definizione di “tipico” che viene adottata.

in parole povere

qual è il valore che meglio rappresenta le osservazioni?

cosa è tipico osservare?

misure diverse a seconda di cosa si intende per “tipico”

indicatori di t.c.

mediana = Q2
moda = caso o classe più frequente
media aritmetica = somma dei casi, diviso numero dei casi
media quadratica = radice quadrata della media dei quadrati dei casi
altre medie (geometrica, armonica...)

Trovare la mediana

casi: 9 1 2 2 0

ordina i casi

0

Trovare la mediana

casi: 9 1 2 2

ordina
i casi

0 1

Trovare la mediana

casi: 9 2 2

ordina
i casi

0 1 2

Trovare la mediana

casi: 9

2

ordina
i casi

0 1 2 2

Trovare la mediana

casi: 9

ordina
i casi

0 1 2 2 9

MEDIANA

```
> sv <- sort(d$LT)

> length(d$LT)

[1] 78

> med <- (sv[38] +  sv[39])/2
> med

[1] 98

> median(d$LT)

[1] 98
```

```
> t <- table(d$LT)
> mod <- t[t == max(t)]
> mod

110
 10
```

```
> sum(d$LT)/length(d$LT)

[1] 99.69231

> mean(d$LT)

[1] 99.69231
```

```
> sqrt(sum(d$LT^2)/length(d$LT))

[1] 99.90585

> mq <- function(x) sqrt(sum(x^2)/length(x))
> mq(d$LT)

[1] 99.90585

> ls()

[1] "mq" .......
```

statistiche “robuste”

la media aritmetica

è il “baricentro” della distribuzione

la somma degli scarti = 0

la somma degli scarti al quadrato
= MIN

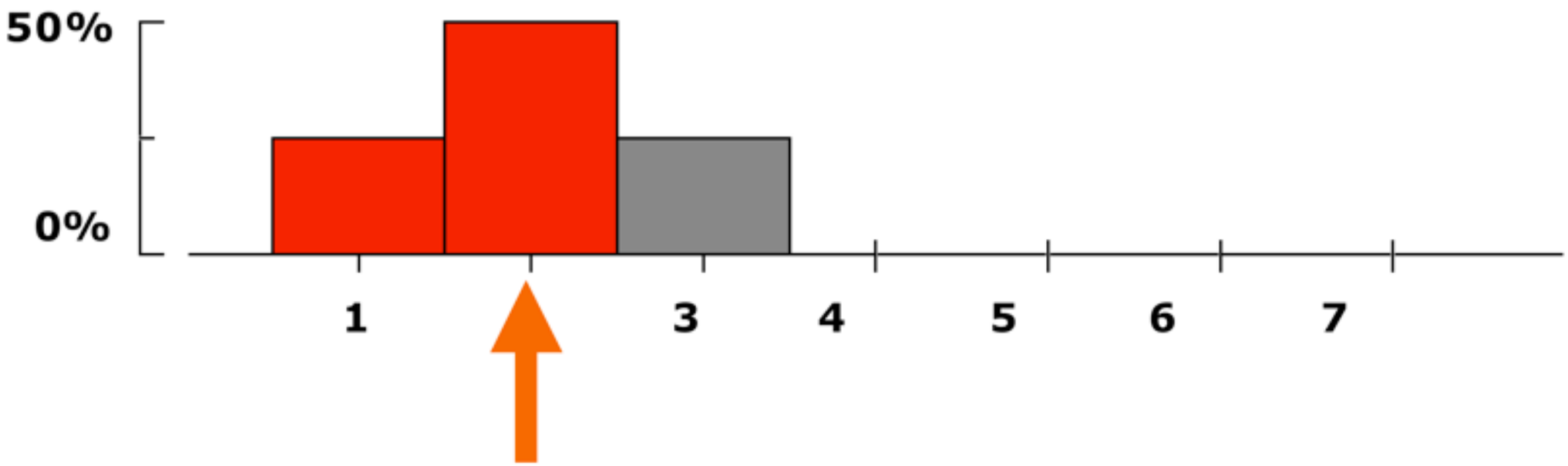
la media

è molto sensibile ai valori estremi

per questo non è sempre la migliore maniera di caratterizzare il valore tipico

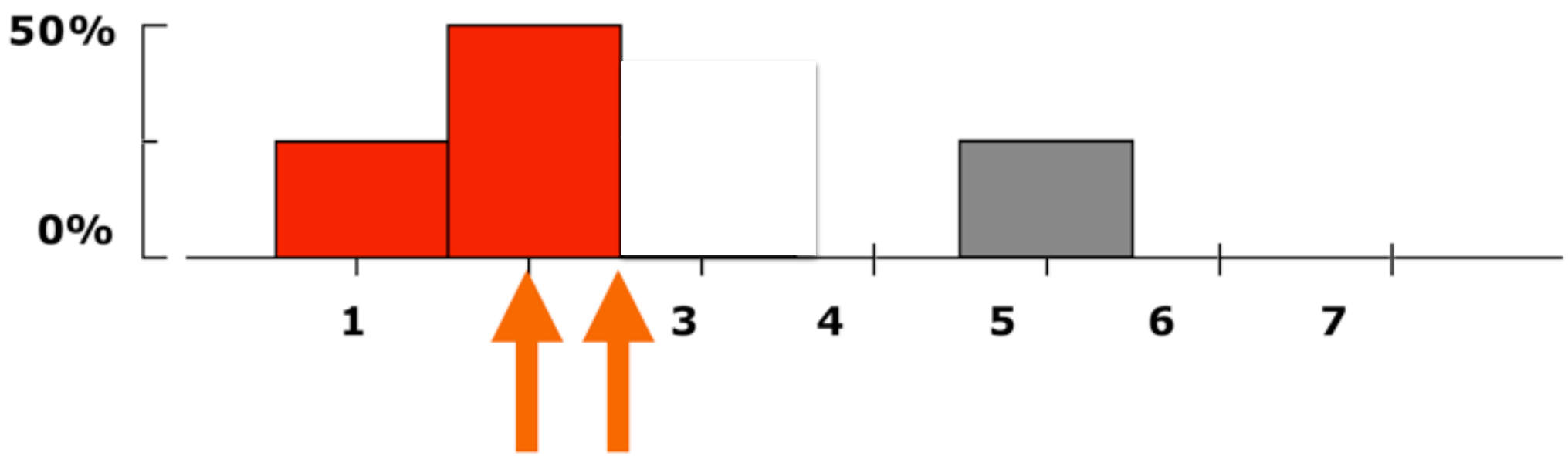
Esempio

casi: 1, 2, 2, 3



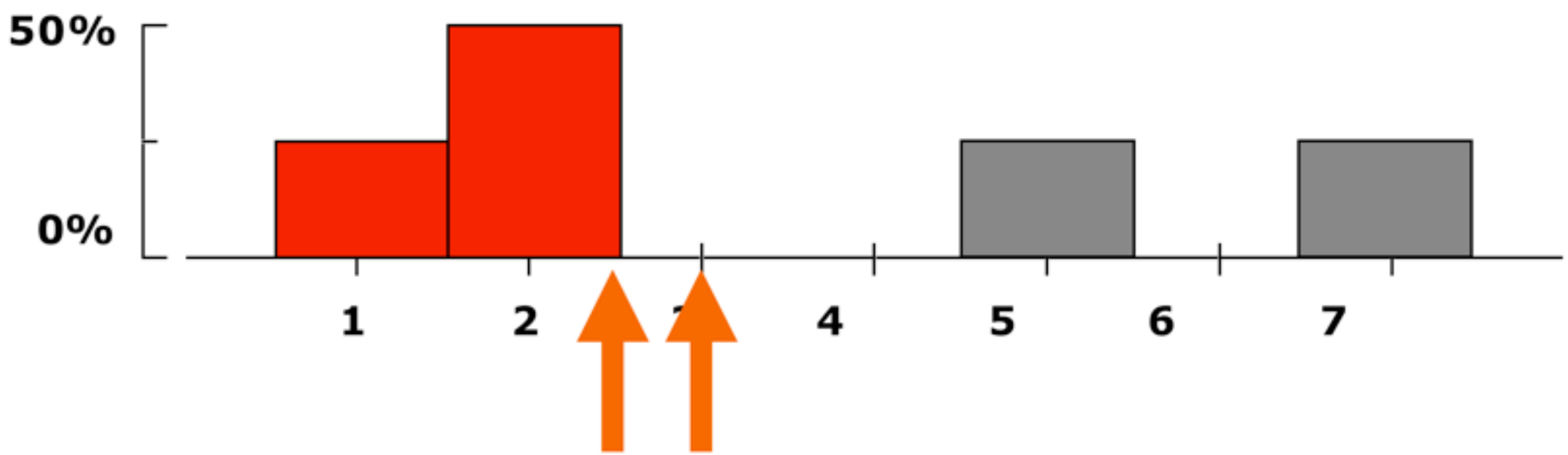
Esempio

casi: 1, 2, 2, 5



Esempio

casi: 1, 2, 2, 7



dispersione

dispersione

La tendenza dei dati quantitativi ad allontanarsi dal centro della distribuzione, ossia ad assumere valori in eccesso o in difetto rispetto alla posizione identificata come rappresentativa della tendenza centrale.

in parole povere

quanto sono diversi o disomogenei i dati?

di quanto, tipicamente, i dati si allontanano dal centro?

una distanza, non una posizione

indicatori di dispersione

gamma interquartile = Q3 - Q1

deviazione standard = media quadratica degli scarti dalla media aritmetica

Definizione

DS = RMS degli scarti

scarto = (caso - media)

RMS (*root mean square*) = radice quadrata della media dei quadrati

$$(x - M)$$

scarto

$$\sqrt{\frac{\sum x^2}{N}}$$

RMS dei casi

$$\sqrt{\frac{\sum (x - M)^2}{N}}$$

RMS degli scarti

Calcolare la DS

casi: 9, 1, 2, 2, 0

trova la media

M = 2.8

calcola gli scarti

9 - 2.8, 1 - 2.8 .....

6.2 -1.8 -0.8 -0.8 -2.8

Calcolare la DS

eleva al quadrato gli scarti

$$6.2^2 \quad -1.8^2 \quad -0.8^2 \quad -0.8^2 \quad -2.8^2$$

$$38.44 \quad 3.24 \quad 0.64 \quad 0.64 \quad 7.84$$

Calcolare la DS

calcola la media degli scarti al quadrato

$$6.2^2 \quad -1.8^2 \quad -0.8^2 \quad -0.8^2 \quad 2.8^2$$

$$38.44 + 3.24 + 0.64 + 0.64 + 7.84$$

$$50.80/5 = 10.16$$

varianza

Calcolare la DS

50.80/5 =10.16 ← **varianza**

trova la radice quadrata

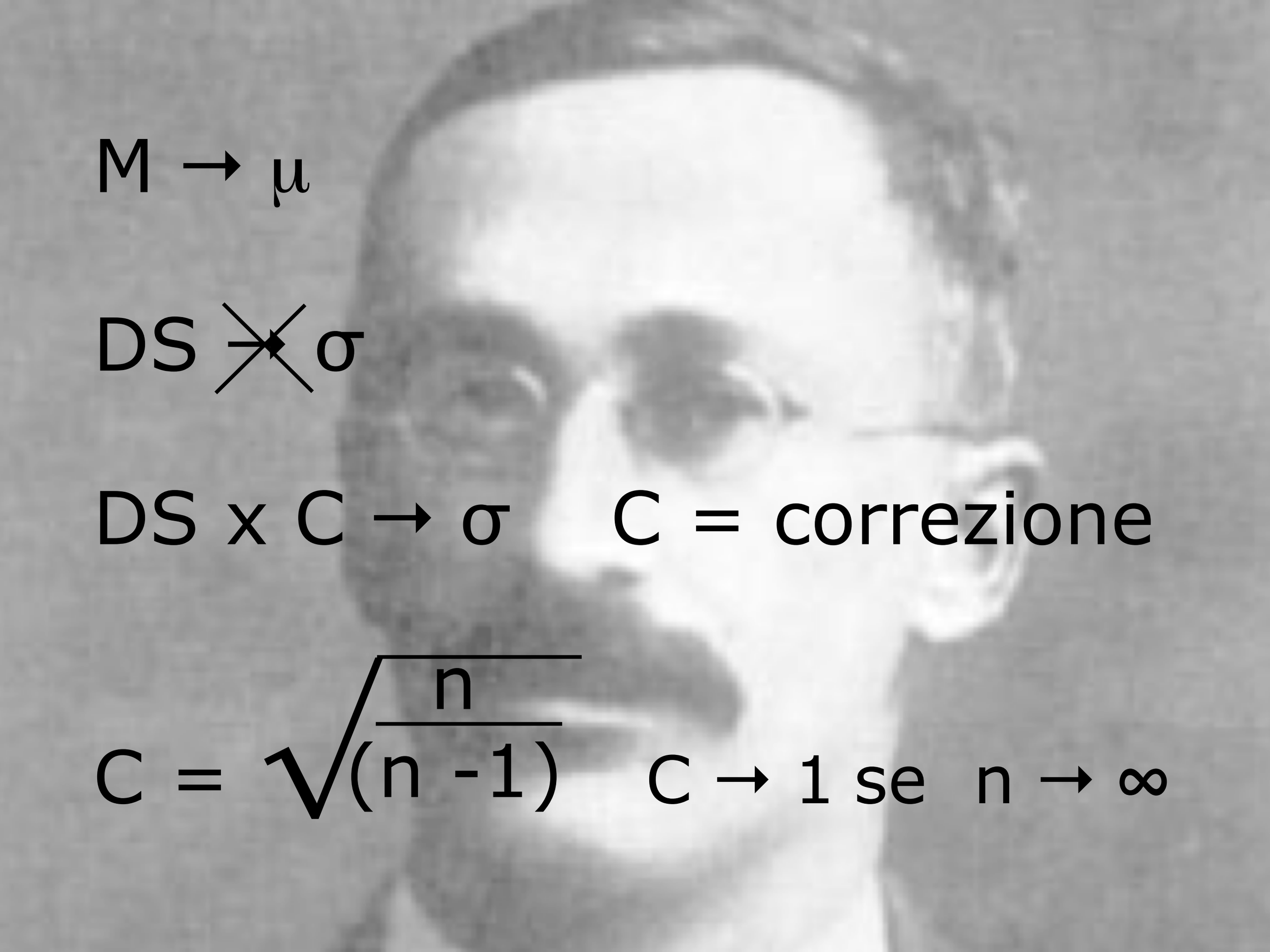
√10.16 ≈ 3.2 ← **DS**

la DS

non può mai essere negativa

misura la dispersione nella stessa unità di misura dei dati

si calcola in due modi diversi per la popolazione e per un campione



$$M \rightarrow \mu$$

$$DS \not\rightarrow \sigma$$

$$DS \times C \rightarrow \sigma \qquad C = \text{correzione}$$

$$C = \sqrt{\frac{n}{(n-1)}} \qquad C \rightarrow 1 \text{ se } n \rightarrow \infty$$

```
> m <- mean(d$LT)
> s <- d$LT - m
> ds <- sqrt(mean(s^2))
> ds

[1] 6.52865

> sd(d$LT)

[1] 6.570907
```

```
> m <- mean(d$LT)
> s <- d$LT - m
> n <- length(d$LT)
> dsc <- sqrt(sum(s^2)/(n -1))

[1] 6.570907

> sd(d$LT)

[1] 6.570907
```