Psicometria 1 (023-PS)

Michele Grassi
mgrassi@units.it

Università di Trieste

Lezione 3-4-5

Piano della presentazione

Statistica descrittiva e inferenziale

- Rappresentazione numerica o grafica dei dati o di importanti caratteristiche dei dati.
- Viene eseguita per la maggior parte con l'ausilio di un software statistico.
- In questo corso useremo R, un programma simile a SPLUS. Altri programmi statistici sono Minitab, SAS, SPSS. Talvolta vengono usati fogli elettronici come Excel.

- A causa degli errori introdotti dal processo di campionamento, le caratteristiche del campione non rappresentano esattamente le caratteristiche della popolazione.
- L'inferenza statistica studia per l'appunto i procedimenti induttivi, di natura logica e matematica, con i quali si perviene a risultati e conclusioni di validità generale attraverso indagini condotte su un conveniente sottoinsieme (campione) delle manifestazioni del fenomeno di studio.

distribuzioni di frequenze

- Sia X una generica variabile qualitativa intorno al quale s'intende indagare.
- Fissata la popolazione P composta dalle N unità statistiche da analizzare e precisate le modalità (h) in cui si articola la variabile oggetto di studio si dice frequenza assoluta il numero n_j di unità statistiche che presentano la modalità $X = x_j$.
-

Distribuzioni di frequenze

L'elenco delle modalità x_j , insieme alle frequenze (assolute) n_j, viene detto distribuzione di frequenze.

X	n
x_1	n_1
x_2	n_2
$\dots$	$\dots$
x_j	n_j
$\dots$	$\dots$
x_h	n_h
Tot	N

Distribuzioni di frequenze

- La distribuzione delle unità statistiche della popolazione viene detta distribuzione della popolazione.
- Solitamente la distribuzione della popolazione non è conosciuta (non è stata osservata o misurata).
- La distribuzione delle unità statistiche di un campione estratto dalla popolazione viene detta distribuzione empirica (o del campione) e, ovviamente, può essere osservata.

Classificazione delle distribuzioni

X			
qualitativa		quantitativa	
nominale	ordinale	discreta	continua
serie sconnessa	serie ordinata	seriazione	
Modalità	Frequenze	Modalità	Frequenze
x_1	n_1	$x_0 - x_1$	n_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_j	n_j	$x_{j-1} - x_j$	n_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_h	n_h	$x_{h-1} - x_h$	n_h
Tot	N	Tot	N

Frequenze relative e frequenze cumulate

Si dice frequenza relativa di una modalità x_j , o di una classe di modalità $(x_{j-1}; x_j)$ e si indica con f_j , la frazione o proporzione di unità statistiche che presentano tale modalità:

$$f_j = \frac{n_j}{\sum_{j=1}^{h} n_j} = \frac{n_j}{n}$$

Proprietà

$$\sum_{j=1}^{h} f_j = \frac{n}{n} = 1$$

Si dice frequenza cumulata la somma delle frequenze (assolute) sino alla modalità considerata.

Funzione di ripartizione

La funzione di ripartizione $F(x_j)$ della variabile statistica quantitativa X è la proporzione di unità statistiche con valori di $X \leq x_j$:

$$F(x_j) = \sum_{i=1}^{j} f_i = Pr\left(X \leq x_j\right)$$

Illustrazione

illustrazione

Quale esempio di serie ordinata consideriamo i dati di Hout et al. (1987) riportati da Agresti (1990) che mosrtano le risposte di 91 coppie sposate al seguente *item* di un questionario: *"Sex is fun for me and my partner"* .

Le modalità della variabile risposta sono:

- never or occasionally
- fairly often
- very often
- almost always

Modalità	Never fun	Fairly often	Very often	Always fun
Frequenza assoluta	12	28	18	33

- Nel caso di una seriazione di frequenze non è possibile presentare in una tabella tutte le modalità della variabile analizzata, dato che la maggior parte di esse sono osservate solo una volta nel campione.
- È invece necessario raggruppare le osservazioni in classi (laddove una classe corrisponde ad un intervallo di valori) e contare la frequenza delle u.s. in ciascuna classe.
- Benchè le distribuzioni empiriche di variabili continue possano essere presentate in forma tabulare, le rappresentazione grafiche sono preferibili in quanto più facilmente interpretabili.

Illustrazione in R

illustrazione

Si considerino i dati della tabella 3.1 che riportano il tasso di omicidi (numero di omicidi per ogni 100,000 abitanti) nei diversi Stati americani nel 1993.

Le prime 5 righe della tabella 3.1 sono

```
library(smss); data(statewide.crime.2)
omicidi<-statewide.crime.2
omicidi[1:5,]
##   State   VR   MR    M    W    H    P    S
## 1    AK  761  9.0 41.8 75.2 86.6  9.1 14.3
## 2    AL  780 11.6 67.4 73.5 66.9 17.4 11.5
## 3    AR  593 10.2 44.7 82.9 66.3 20.0 10.7
## 4    AZ  715  8.6 84.7 88.6 78.7 15.4 12.1
## 5    CA 1078 13.1 96.7 79.3 76.2 18.2 12.5
```

Agresti e Finlay scelgono gli intervalli $\{0-2.9; 3-5.9; \ldots; 18-20.9\}$.

Illustrazione con R

Tabella: 3.1

Tasso di omicidi	Frequenza assoluta
$0.0 - 2.9$	5
$3.0 - 5.9$	16
$6.0 - 8.9$	12
$9.0 - 11.9$	12
$12.0 - 14.9$	4
$15.0 - 17.9$	0
$18.0 - 20.9$	1
Tot	50

Washington DC ha un Murder Rate fuori intervallo max (20.9):

```
 which(omicidi$State=="DC")
##  [1] 51
 omicidi[51,]
##    State   VR   MR   M    W    H    P    S
## 51    DC 2922 78.5 100 31.8 73.1 26.4 22.1
```

Illustrazione con R

Tabella: 3.1

Tasso di omicidi	Frequenza assoluta
$0.0 - 2.9$	5
$3.0 - 5.9$	16
$6.0 - 8.9$	12
$9.0 - 11.9$	12
$12.0 - 14.9$	4
$15.0 - 17.9$	0
$18.0 - 20.9$	1
Tot	50

Non viene conteggiato nella Tabella 3.1, possiamo quindi rimuoverlo:

```
tasso.omicidi <- omicidi[omicidi$State!="DC",3]

## oppure, in modo equivalente può essere ricavato come
## tasso.omicidi <- omicidi$MR[-c(51)]
```

Illustrazione con R

Tabella: 3.1

Tasso di omicidi	Frequenza assoluta
$0.0 - 2.9$	5
$3.0 - 5.9$	16
$6.0 - 8.9$	12
$9.0 - 11.9$	12
$12.0 - 14.9$	4
$15.0 - 17.9$	0
$18.0 - 20.9$	1
Tot	50

Ricaviamo le frequenze assolute della variabile $Tasso\ di\ omicidi$:

```
f.ass <- hist(tasso.omicidi, br=c(0, 2.9, 5.9, 8.9, 11.9,
                  14.9, 17.9, 20.9),freq=FALSE)

f.ass$counts
[1]  5 16 12 12  4  0  1
```

Illustrazione con R

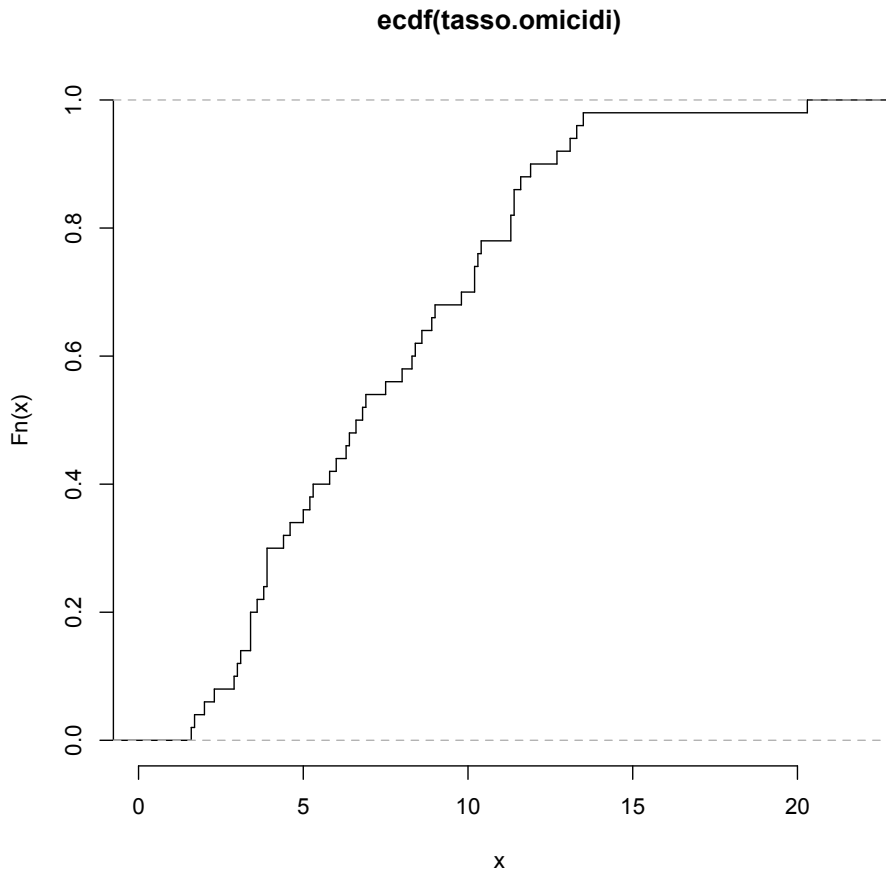
Per i dati relativi al $tasso\ di\ omicidi$ dell'esempio precedente la funzione ripartizione empirica, ovvero la proporzione dei dati aventi valore minore di x_j,

$$\hat{Pr}(X \leq x_j) = \frac{\sum_{i=1}^{n} (X_i \leq x_j)}{N}$$

(considerando tutte le modalità e non lori classi!)
si genera in R nel seguente modo

```
F.Rip<-ecdf(tasso.omicidi)
plot(F.Rip, verticals = TRUE, do.points = FALSE,bty="n")
```

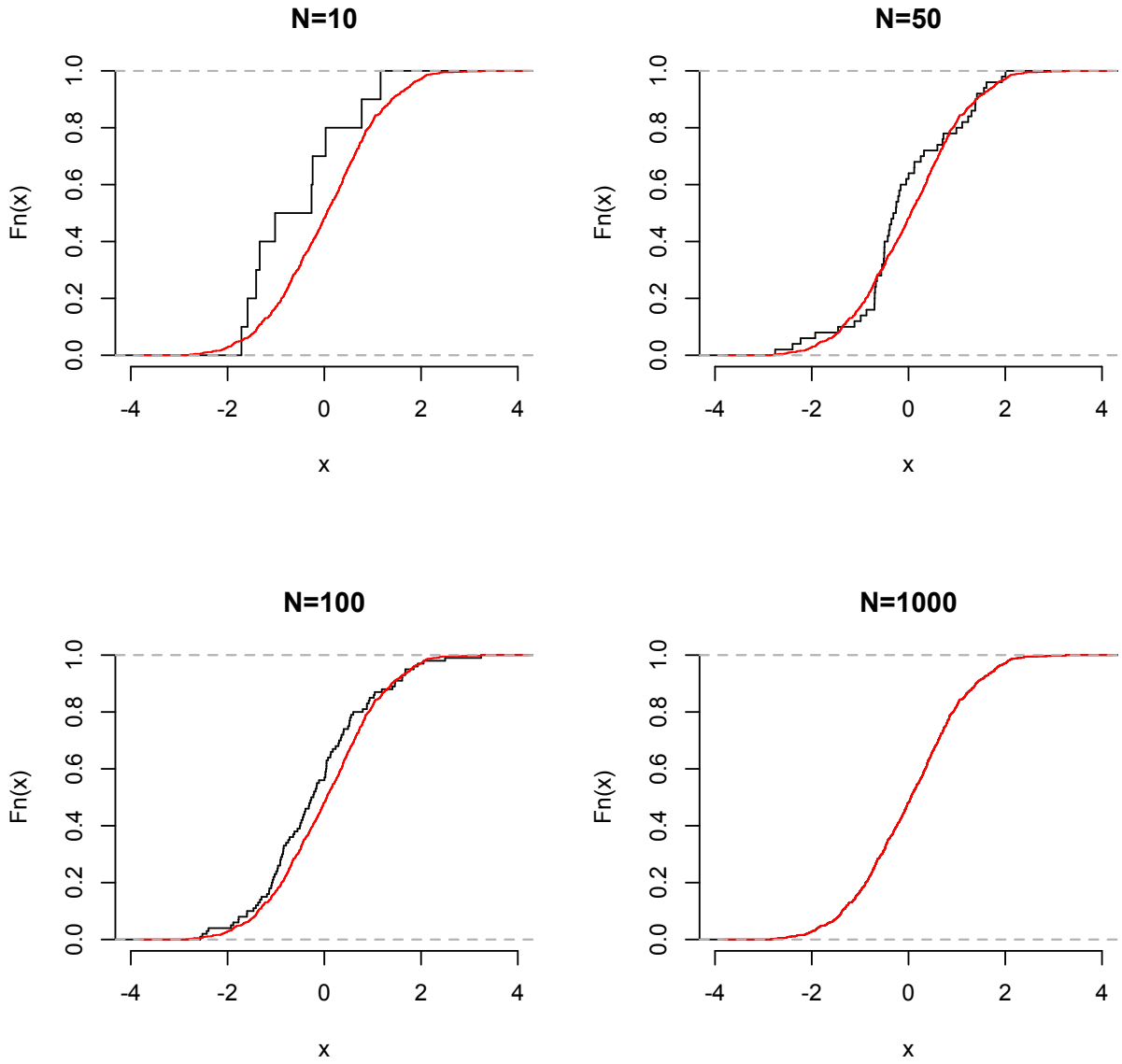
Illustrazione con R



Si noti che la funzione di ripartizione empirica è una funzione a scalini.

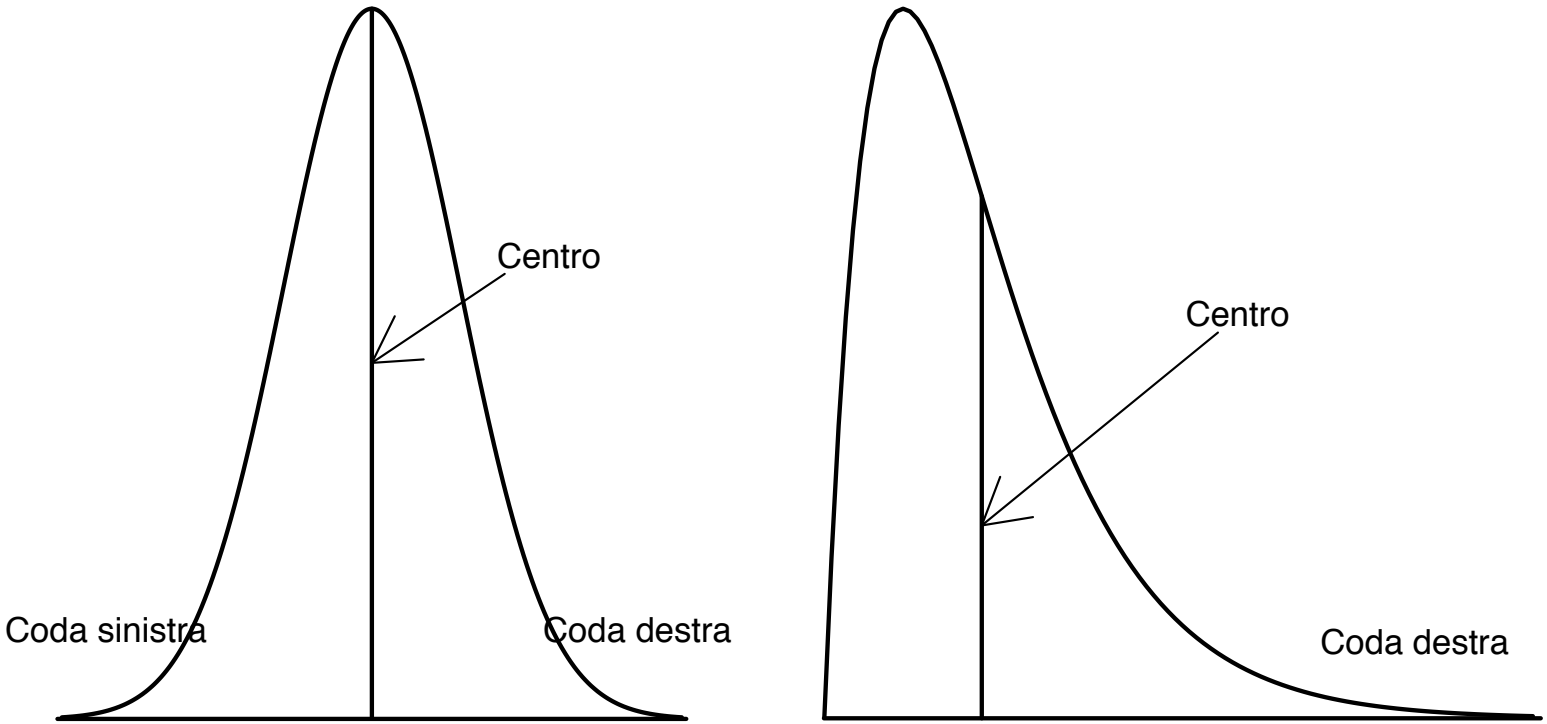
Nelle figure successive sono riportate le distribuzioni ripartizione empiriche e teoriche di quattro campioni casuali di $n = \{10; 50; 100; 1000\}$ osservazioni tratte dalla distribuzione normale.

Ciò che dobbiamo notare è che, al crescere del numero di osservazioni, la funzione ripartizione empirica si approssima sempre più ad una curva continua.

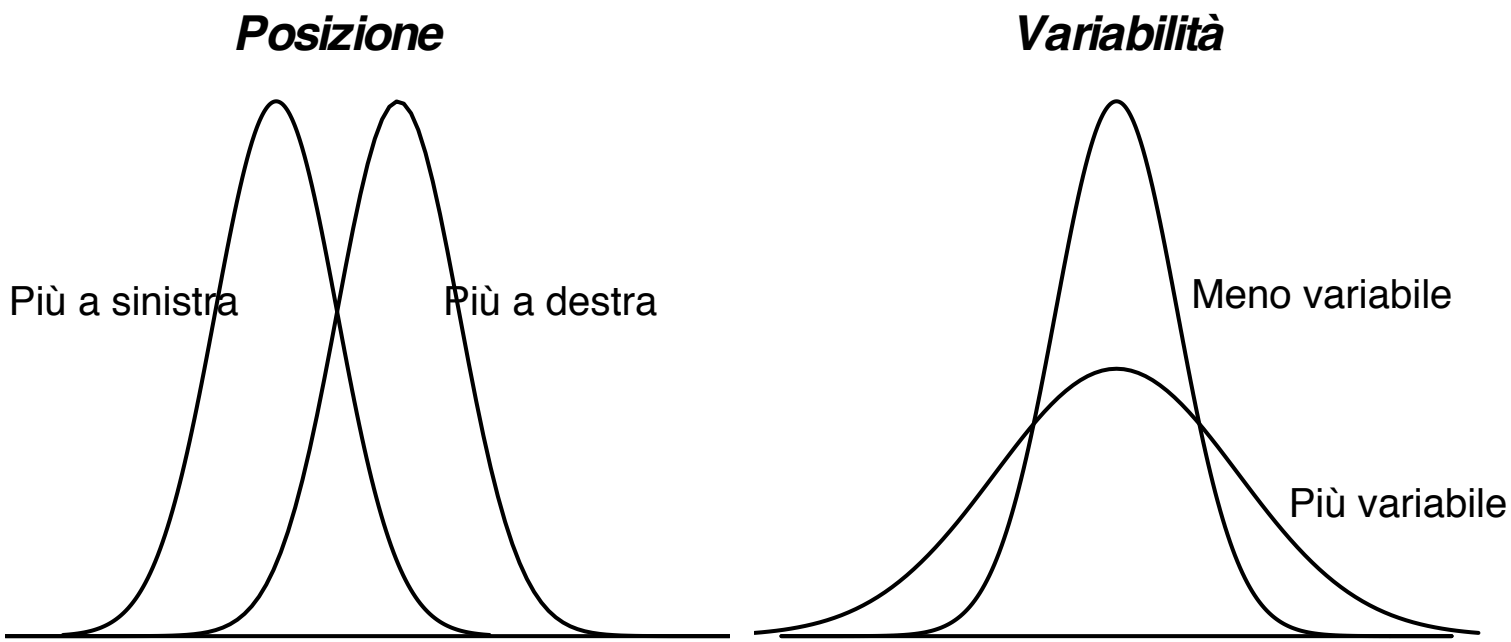

N=10
N=50
N=100
N=1000
Fn(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4

Aspetti notevoli delle distribuzioni

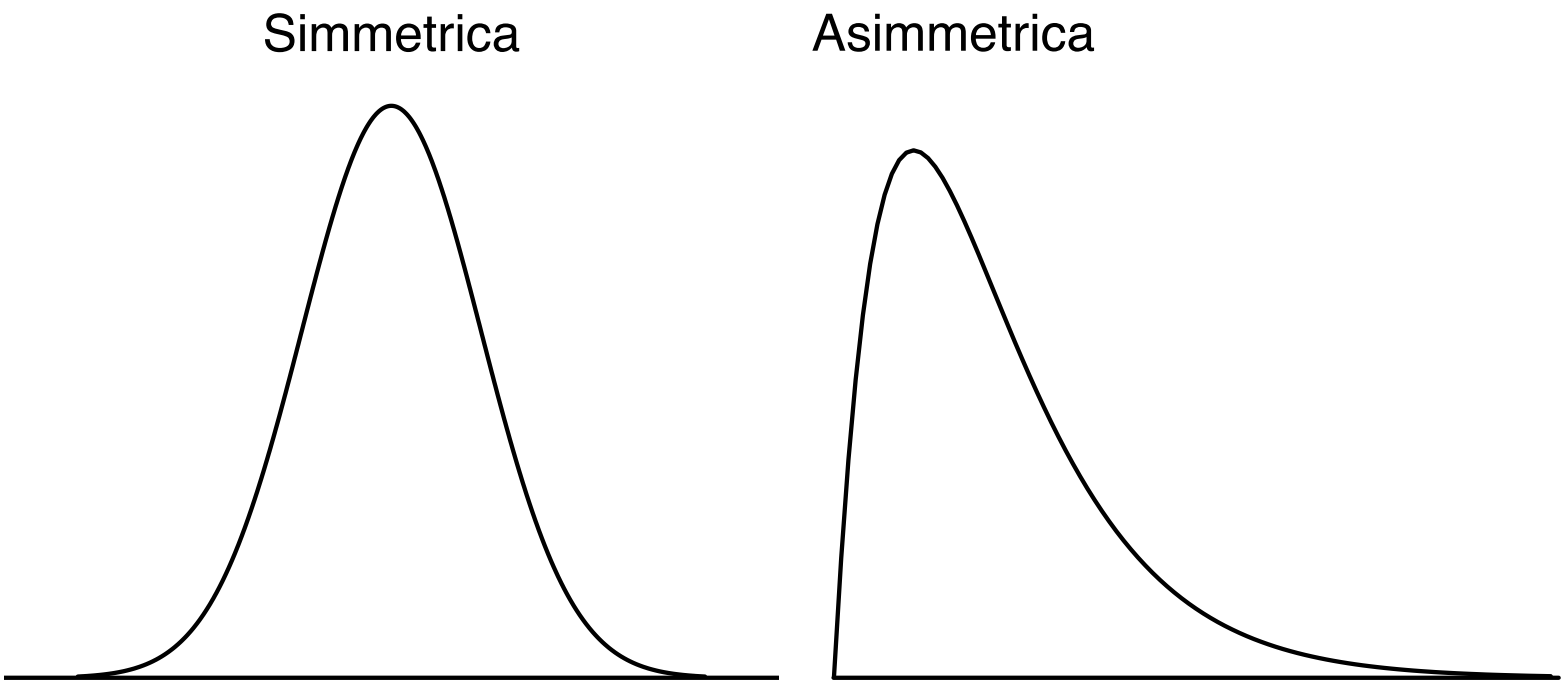
- Posizione
- Variabilità
- Forma



Posizione e variabilità



Forma



Misure di tendenza centrale e dispersione

- Il centro e la dispersione di una distribuzione possono essere valutati visivamente.
- È tuttavia possibile misurare queste proprietà di una distribuzione in maniera più precisa.
-

- Le misure di tendenza centrale sono particolarmente utili quando vogliamo confrontare due distribuzioni per rispondere a domande del tipo: "*Gli uomini sono più depressi delle donne?*" o "*Agli extra-comunitari arrestati per avere commesso un reato vengono inflitte pene più pesanti che ai cittadini italiani?*"
- Le misure di dispersione sono utili per rispondere a domande del tipo "*C'è una maggiore variabilità nel reddito pro-capite nel nord o nel sud Italia?*"

Indici di posizione

- Gli indici di posizione consentono una sintesi della distribuzione attraverso un valore rappresentativo.
- **Distribuzioni sconnesse**: moda
- **Distribuzioni ordinate**: mediana
- **Seriazioni**: media aritmetica

Media aritmetica

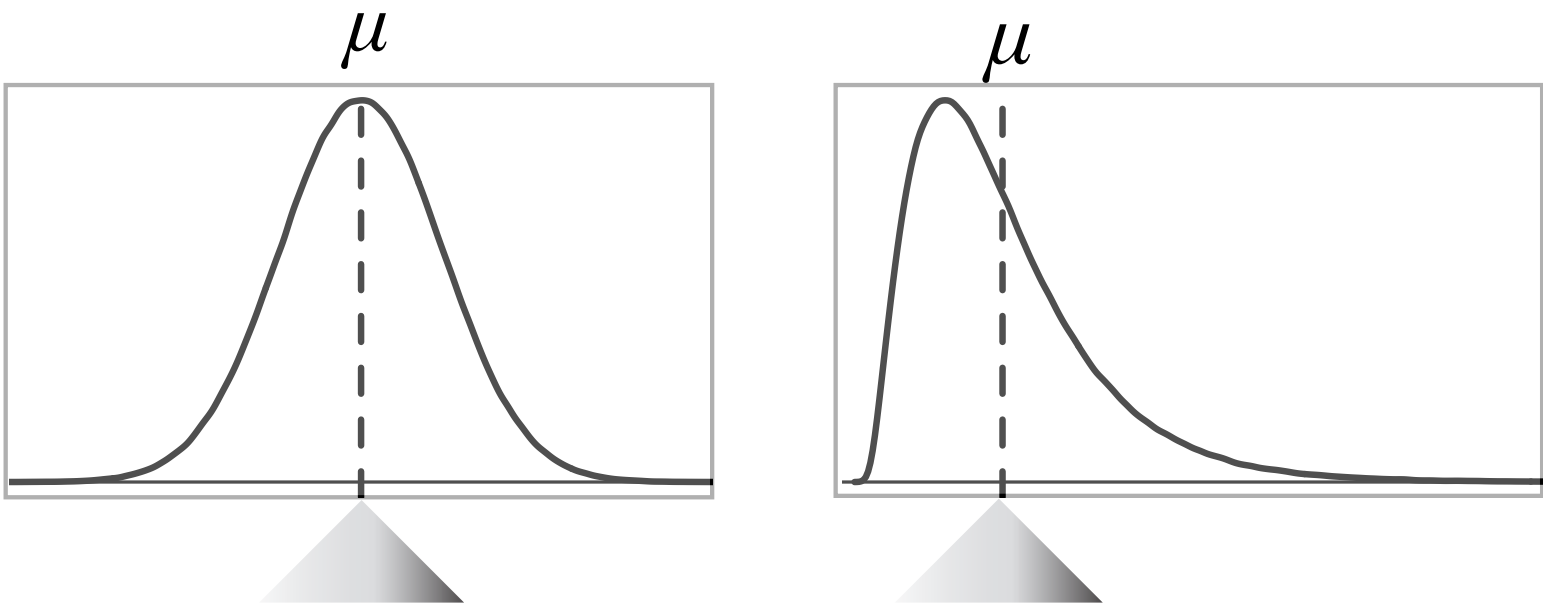
- La misura di tendenza centrale più comunemente usata nella statistica descrittiva univariata quantitativa è la media aritmetica.
- **Media del campione**: si indica con $\bar{x}$ la media aritmetica delle modalità che una v.s. quantitativa X ha esibito in un campione di n osservazioni;

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

- Se ciascun dato, x_i, fosse un punto sulla linea dei numeri reali, allora $\bar{x}$ rappresenterebbe il punto di equilibrio.

Media aritmetica

- **Media della popolazione**: si indica con μ la media aritmetica delle modalità che una v.s. quantitativa X *assume all'interno della popolazione.*



Proprietà della media aritmetica

Teorema della somma degli scarti

Data la variabile X che presenta le n modalità $x_1, x_2, \ldots, x_n$, la somma degli scarti di ciascuna modalità dalla propria media aritmetica vale zero.

Dimostrazione.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

Proprietà della media aritmetica

Teorema della media di una trasformazione lineare

Data la variabile X che presenta le n modalità $x_1, x_2, \ldots, x_n$, ed avente media $\bar{x}$, se consideriamo la trasformazione lineare $Y = a + bX$ si avrà allora che $\bar{y} = a + b\bar{x}$.

Dimostrazione.

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\sum (a + bx_i)}{n} = \frac{na}{n} + b\frac{\sum x_i}{n} = a + b\bar{x}$$

Proprietà della media aritmetica

Teorema della devianza

Data la variabile X che presenta le n modalità $x_1, x_2, \ldots, x_n$, la quantità $\sum (x_i - c)^2$ avrà il suo valore *minimo* se e solo se $c = \bar{x}$. Il valore $\sum (x_i - \bar{x})^2$ prende il nome di *devianza*.

Dimostrazione.

Riscriviamo la quantità

$$\sum (x_i - c)^2 ,$$

come

$$\sum \left[(x_i - \bar{x}) + (\bar{x} - c) \right]^2 ;$$

sviluppiamo il quadrato,

$$\sum (x_i - \bar{x})^2 + \sum (\bar{x} - c)^2 + 2 \sum (x_i - \bar{x}) (\bar{x} - c) .$$

□

Proprietà della media aritmetica

Teorema della devianza

Data la variabile X che presenta le n modalità $x_1, x_2, \ldots, x_n$, la quantità $\sum (x_i - c)^2$ avrà il suo valore *minimo* se e solo se $c = \bar{x}$. Il valore $\sum (x_i - \bar{x})^2$ prende il nome di *devianza*.

Dimostrazione.

Essendo la quantità $(\bar{x} - c)$ costante, e per il teorema della somma degli scarti, possiamo scrivere il terzo addendo come

$$2 (\bar{x} - c) \sum (x_i - \bar{x}) = 0,$$

ottenendo

$$\sum (x_i - \bar{x})^2 + \sum (\bar{x} - c)^2 ,$$

che avrà il valore più basso solo quando $c = \bar{x}$

□

Mediana

- La media aritmetica non è sempre l'indice che meglio rappresenta la tendenza centrale di una distribuzione: se la distribuzione è asimmetrica, la mediana è più adeguata della media quale misura di tendenza centrale.
- La mediana rappresenta la modalità che, una volta ordinate nel senso non decrescente le n unità di P rispetto alle modalità medesime, è posseduta da quella che occupa il posto centrale, ovvero che lascia alla sua destra ed alla sua sinistra un numero uguale di unità.

Mediana

- Se n è dispari la mediana sarà data dalla modalità a cui corrisponde l'unità statistica di posto $\frac{n+1}{2}$.
- Se n è pari non è detto che la mediana sia univocamente determinata in quanto essa sarà data dalle modalità a cui corrispondono le unità statistiche di posto $\frac{n}{2}$ e $\frac{n}{2}+1$. **Se queste due modalità sono diverse un procedimento per il calcolo della mediana sarà quello di effettuare la loro media aritmetica**.

Proprietà della Mediana

- La mediana gode di una importante proprietà che è quella di minimizzare la somma degli scarti assoluti dei valori ossia:

$$\sum_{i=1}^{n} |x_i - M_e| = min$$

- La mediana M_e resta invariata se si sostituiscono i termini $x < M_e$ o $x > M_e$: la mediana non risente di valori anomali.
- Applicabile anche per v.s. ordinali.

Proprietà della Mediana

La mediana come centro della distribuzione

$$\sum_{i=1}^{n} |x_i - M_e| = min$$

Dimostrazione.

```
 x<-rnorm(10001)
 pos.Me<-(10001+1)/2
 x.ordinata<-sort(x)
 x.ordinata[pos.Me]
## [1] -0.01425113

  median(x)
## [1] -0.01425113
```

□

Proprietà della Mediana

La mediana come centro della distribuzione

$$\sum_{i=1}^{n} |x_i - M_e| = min$$

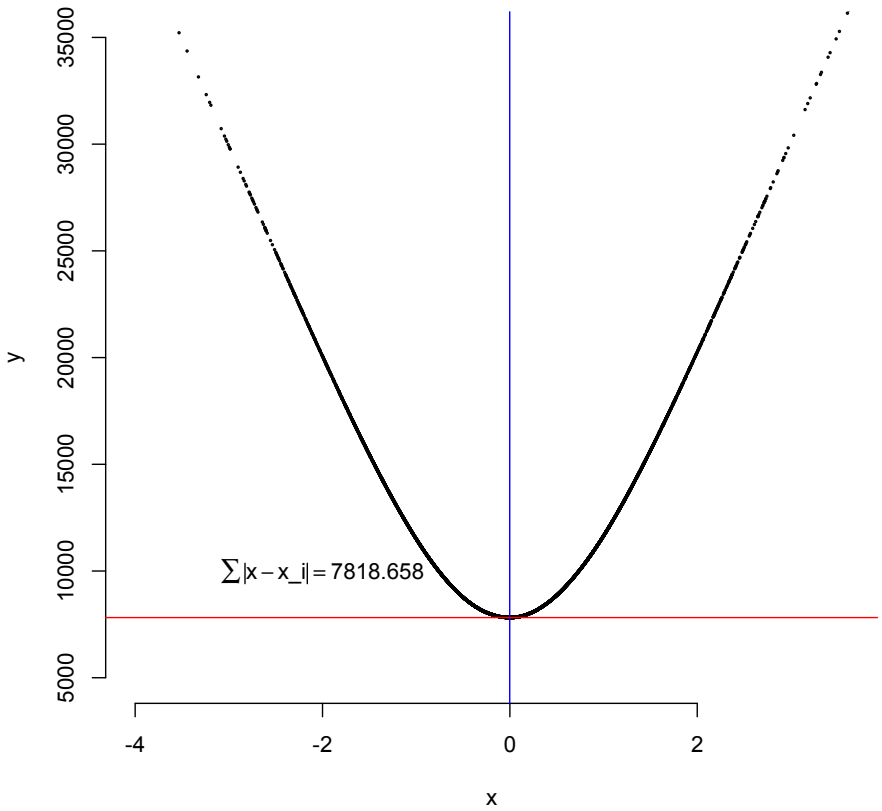
Dimostrazione.

```
sum( abs(x.ordinata-median(x)) )
##[1] 7818.658

abs.res.sum <- function(x,mediana){
res<-sum( abs(x-mediana) )
res}
abs.res.sum(x=x,mediana=median(x))
##[1] 7818.658
```

Proprietà della Mediana

Dimostrazione.

```
y<-rep(0,10001)

for(i in 1:10001){
y[i]<-abs.res.sum(x=x,
                  mediana=x[i])
}

plot(x=x,y=y, bty="n",
       ylim=c(5000,35000),
       cex=0.2)
 abline(v=median(x),
           col="blue")
 abline(h=7818.658,
           col="red")
```



Moda

- La moda: è data dalla modalità che presenta la frequenza relativa o assoluta più elevata e viene indicata con m_0.
- La moda è utile per distribuzioni unimodali.
  - Distribuzione *unimodale* possiede un unico massimo locale.
  - Distribuzione *bimodale* possiede più di un massimo locale.

Quantili

- La sintesi di una distribuzione operata attraverso un indice di posizione può risultare troppo drastica. Il calcolo dei quantili offre un possibile compromesso: sintetizziamo i dati con un numero molto limitato di valori corrispondenti a punti tipici della distribuzione.
- Possiamo interpretare la mediana come la modalità che lascia alla sua sinistra il 50% delle u.s. della distribuzione, cioè che divide la distribuzione a metà.

Quantili

- Si può effettuare lo stesso ragionamento cercando di individuare le modalità che dividono la distribuzione in un certo numero fissato di parti uguali, ciascuna delle quali con frequenza relativa p.
- In questo senso la mediana diventa il quantile di ordine $p = 1/2$.

Quantili

- I quantili più utilizzati sono i quartili, che dividono la distribuzione in 4 parti uguali e vengono di solito indicati con Q_1, $Q_2 = Me$ e Q_3, i decili, che dividono la distribuzione in 10 parti uguali, ed i percentili, che dividono la distribuzione in 100 parti uguali.
- Chiaramente il 25-esimo percentile è pari a Q_1, il 75-esimo è pari a Q_3, mentre il 50-esimo percentile ed il 5^o decile coincidono entrambi con la mediana $Me = Q_2$.

Quantili

- Il **primo quartile** Q_1 è quel valore tale che il 25% delle osservazioni ha un valore più piccolo, mentre il restante 75% ha un valore più grande. Cioè Q_1 è l'osservazione di posto $\frac{n+1}{4}$ nell'insieme ordinato dei dati osservati.
- Il **terzo quartile** Q_3 è quel valore tale che il 75% delle osservazioni ha un valore più piccolo, mentre il restante 25% ha un valore più grande. Cioè Q_3 è l'osservazione di posto $\frac{3(n+1)}{4}$ nell'insieme ordinato dei dati osservati.

Quantili

- Si possono utilizzare procedure diverse per il calcolo dei quantili.
- La procedura presentata di seguito si può utilizzare senza l'ausilio di un computer.
- Distinguiamo tre casi diversi.

Quantili: 3 situazioni di calcolo

1. Se la posizione $\frac{(n+1)}{4}$ nel caso di Q_1 e $\frac{3(n+1)}{4}$ nel caso di Q_3 è un numero intero, il quartile ha il valore dell'osservazione corrispondente.
2. Se la posizione $\frac{(n+1)}{4}$ nel caso di Q_1 e $\frac{3(n+1)}{4}$ nel caso di Q_3 è a metà tra due numeri interi, si adotta la convenzione di scegliere come quartile la media (delle modalità) delle osservazioni corrispondenti.
3. Se nel calcolo della posizione Se la posizione $\frac{(n+1)}{4}$ nel caso di Q_1 e $\frac{3(n+1)}{4}$ nel caso di Q_3 non cadiamo in uno dei casi precedenti, cioè la posizione non risulta essere né un numero intero né a metà tra due numeri interi, allora si adotta la convenzione di approssimarla per difetto o per eccesso all'intero più vicino e di scegliere come quartile il valore (della modalità) dell'osservazione corrispondente.

Illustrazione

- Supponiamo di misurare una variabile x (almeno su scala di modalità ordinale) su $n = 10$ unità. I dati ordinati di x sono:

$$x_{sort} = \{2.0; 2.9; 3.3; 4.9; 5.2; 6.4; 7.6; 8.1; 9.0; 11.5\}$$

- Dal momento che $\frac{(n+1)}{4} = \frac{(11)}{4} = 2.75$ e $\frac{3(n+1)}{4} = \frac{33}{4} = 8.25$, Q_1 assumerà il valore dell'osservazione di posto 3, mentre Q_3 quello dell'osservazione di posto 8:

$$Q_1 = x_{sort_3} = 3.3$$

e

$$Q_3 = x_{sort_8} = 8.1$$

Illustrazione con R

Con R tale risultato si ottiene utilizzando la funzione `quantile()` contenuta nel pacchetto base `stats` con l'opzione `type=2` oppure `type=5`:

```
x <-c(2.0, 2.9, 3.3, 4.9, 5.2, 6.4, 7.6, 8.1, 9.0, 11.5)
x
##  [1]  2.0  2.9  3.3  4.9  5.2  6.4  7.6  8.1  9.0 11.5
quantile(x, probs=c(25,50,75)/100, type=2)
## 25% 50% 75%
## 3.3 5.8 8.1
 quantile(x, probs=c(25,50,75)/100, type=5)
## 25% 50% 75%
## 3.3 5.8 8.1
```

Illustrazione con R

Per il calcolo dei quantili, la funzione `summary()` utilizza una procedura di interpolazione condivisa dall'opzione `type=7` nella funzione `quantile()`. Nel caso dei dati precedenti, per esempio, avremo

```
summary(x)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.700   5.800   6.090   7.975  11.500
quantile(x,prob=c(.25,.5,.75),type=7)
##    25%    50%    75%
## 3.700 5.800 7.975
```

Per l'esempio relativo ai tassi di criminalità negli Stati Uniti con `summary()` otteniamo

```
library(smss)
data(crime2005)
 # Eliminiamo il dato di Washington DC
 tasso.omicidi<-crime2005$VI2[-c(51)]

summary(tasso.omicidi)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   27.00   36.50   40.22   51.00   79.00
```

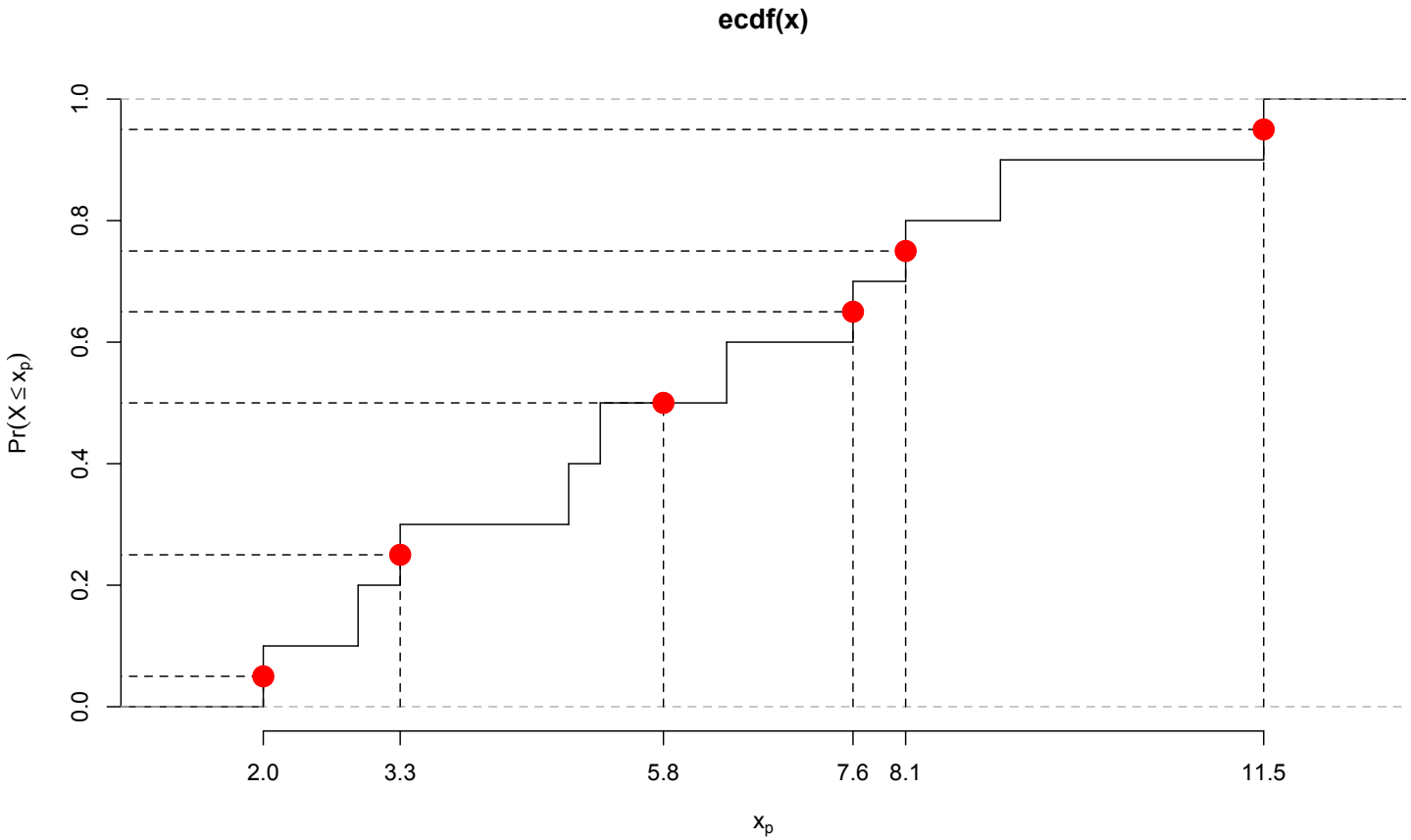
riproducendo così i risultati riportati a p. 55 del testo di Agresti e Finlay.

Quantili in R

In R, la funzione `quantile(x,type=c(2))` restituisce l'inversa della funzione ripartizione empirica `ecdf()`. In altri termini, il quantile di ordine p corrisponde al valore x_p tale per cui $Pr(X \leq x_p) = p$ nella funzione ripartizione empirica.

```
 quantile(x,prob=c(.05,.25,.50,.75,.65,.95),type=2)
##  5%  25%  50%  75%  65%  95%
## 2.0  3.3  5.8  8.1  7.6 11.5
```

Quantili in R

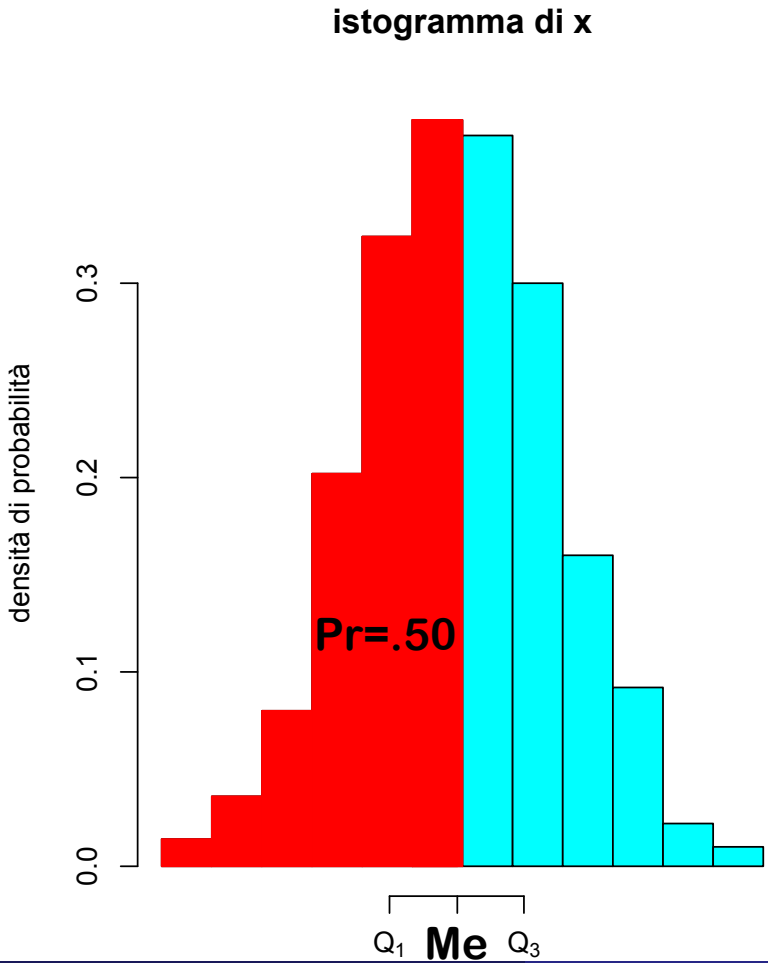


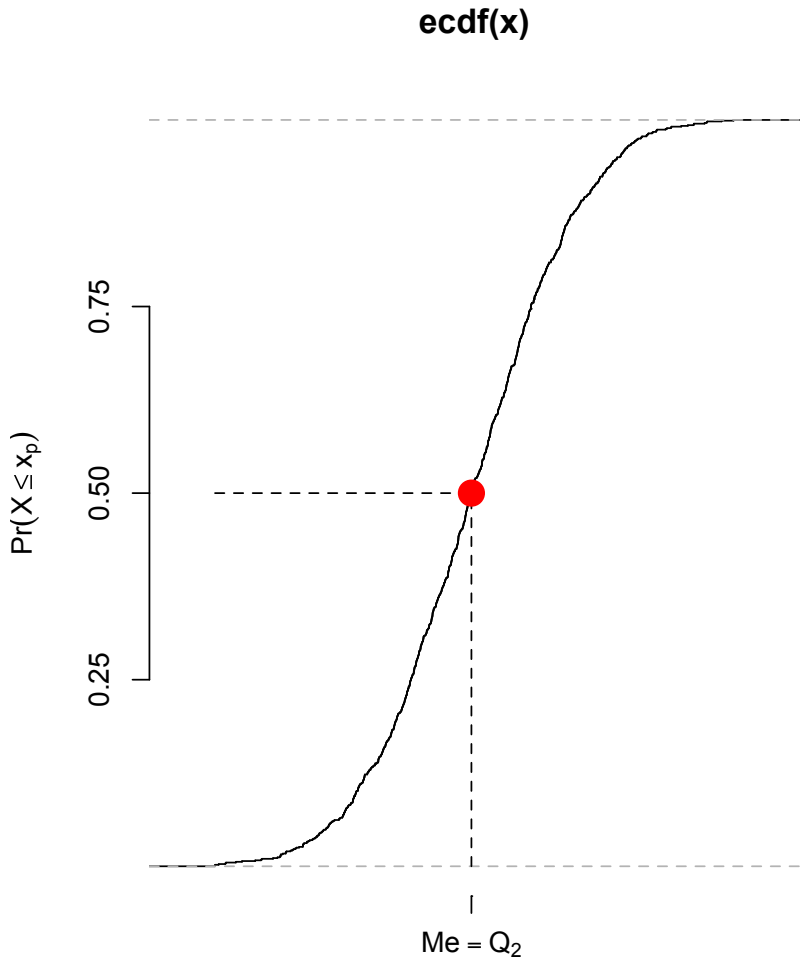
```
# seguendo le 3 regole di calcolo abbiamo che:
# il 5 percentile (p=.05) è il valore di x = 2
 0.2/4*11
## [1] 0.55 # (arrotondato ad 1)
 sort(x)[1]
## [1] 2

# il 25 percentile (p=.25) è il valore di x = 3.3
  1/4*11
## [1] 2.75 # (arrotondato a 3)
 sort(x)[3]
## [1] 3.3

# il 50 percentile (p=.50) è il valore di x = 5.8
  2/4*11
## [1] 5.5 # (posizione intermedia tra 5 e 6)
 sort(x)[c(5,6)]
## [1] 5.2 6.4
 (5.2+6.4)/2 # (media tra le due modalità)
## [1] 5.8
```

```
# il 65 percentile (p=.65) è il valore di x = 7.6
  2.6/4*11
## [1] 7.15 # (arrotondato a 7)
 sort(x)[7]
## [1] 7.6

# il 75 percentile (p=.75) è il valore di x = 8.1
  3/4*11
## [1] 8.25 # (arrotondato a 8)
 sort(x)[8]
## [1] 8.1

# il 95 percentile (p=.95) è il valore di x = 11.5
  3.8/4*11
## [1] 10.45 # (arrotondato a 10)
 sort(x)[10]
## [1] 11.5
```

Quantili in R

Questa procedura è più semplice nel caso di variabili continue, ovvero di campioni n più grandi, dove le posizioni seriali dei quantili non devono essere approssimate.





Interpretazione dei quantili

Leinhardt e Wasserman (1979) riprendono i dati relativi ai tassi di mortalità infantile riportati dal quotidiano The New York Times il 28 settembre 1975. Ciascuna osservazione (riga) è una nazione. Le variabili (colonne) sono:

- il reddito pro-capite in dollari,
- la mortalità infantile per 1000 nati vivi,
- il continente (variabile qualitativa),
- una variabile qualitativa e che riguarda il fatto che quella nazione esporti (modalità $oil = yes$) o meno ($oil = no$) il petrolio.

Interpretazione dei quantili

```
library(car); data(Leinhardt)
europa <- Leinhardt[Leinhardt$region=="Europe" , ];
europa[1:10]
```

	income	infant	region	oil
Austria	3350	23.7	Europe	no
Belgium	3346	17.0	Europe	no
Denmark	5029	13.5	Europe	no
Finland	3312	10.1	Europe	no
France	3403	12.9	Europe	no
West.Germany	5040	20.4	Europe	no
Ireland	2009	17.8	Europe	no
Italy	2298	25.7	Europe	no
Japan	3292	11.7	Europe	no
Netherlands	4103	11.6	Europe	no
Norway	4102	11.3	Europe	no
Portugal	956	44.8	Europe	no
Sweden	5596	9.6	Europe	no
Switzerland	2963	12.8	Europe	no
Britain	2503	17.5	Europe	no

Interpretazione dei quantili

```
 america <- Leinhardt[Leinhardt$region=="Americas" , ]
 america [1:17]
```

	income	infant	region	oil
Canada	4751	16.8	Americas	no
United.States	5523	17.6	Americas	no
Ecuador	250	78.5	Americas	yes
Venezuela	1240	51.7	Americas	yes
Argentina	1191	59.6	Americas	no
Brazil	425	170.0	Americas	no
Chile	590	78.0	Americas	no
Colombia	426	62.8	Americas	no
Costa.Rica	725	54.4	Americas	no
Dominican.Republic	406	48.8	Americas	no
Guatemala	302	79.1	Americas	no
Jamaica	727	26.2	Americas	no
Mexico	684	60.9	Americas	no
Nicaragua	507	46.0	Americas	no
Panama	754	34.1	Americas	no
Peru	335	65.1	Americas	no
Trinidad.and.Tobago	732	26.2	Americas	no

Interpretazione dei quantili

- Ordiniamo i $n = 18$ dati *europei*

```
sort(europa$infant)
## [1]  9.6 10.1 11.3 11.6 11.7 12.8 12.9 13.5 15.1
## [10] 17.0 17.5 17.8 20.4 23.7 25.7 27.8 43.3 44.8
```

- Il primo e il terzo quartile si trovano rispettivamente nella posizione $(18+1)1/4 = 4.75 \approx 5$ e $(18+1)3/4 = 14.25 \approx 14$:

$$Q_1 = x_5 = 11.7$$

$$Q_3 = x_{14 \approx 14} = 23.7$$

- utilizzando R troviamo

```
 quantile(europa$infant, probs=c(25,50,75)/100, type=2)
##   25%   50%   75%
## 11.70 16.05 23.70
```

Interpretazione dei quantili

- Con la funzione `summary()` otteniamo risultati leggermente diversi:

```
summary(europa$infant)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.60   11.97   16.05   19.26   22.88   44.80
```

- In conclusione, nel 1975 la metà centrale delle nazioni europee (Giappone incluso) rivela una mortalità infantile compresa tra i 12 ed i 23 casi, per ogni 1000 nati vivi.
- Un quarto delle nazioni europee ha una mortalità infantile minore di 12.
- Un quarto delle nazioni europee ha una mortalità infantile maggiore di 23.

- Per i dati americani (con un valore mancante `Na` per Haiti) abbiamo:

```
summary(america$infant)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   16.80   38.78   53.05   55.12   62.32  170.00       1
```

- la metà centrale delle nazioni del continente rivela una mortalità infantile compresa tra i 39 ed i 62 casi, per ogni 1000 nati vivi.
- Un quarto delle nazioni ha una mortalità infantile minore di 39.
- Un quarto delle nazioni ha una mortalità infantile maggiore di 62.

Il sommario dei 5 numeri

- nel continente africano la metà delle nazioni ha una mortalità infantile superiore a 140 ogni 1000 nati vivi:

```
 summary(africa$infant)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0   105.4   143.2   142.3   168.1   300.0
```

I

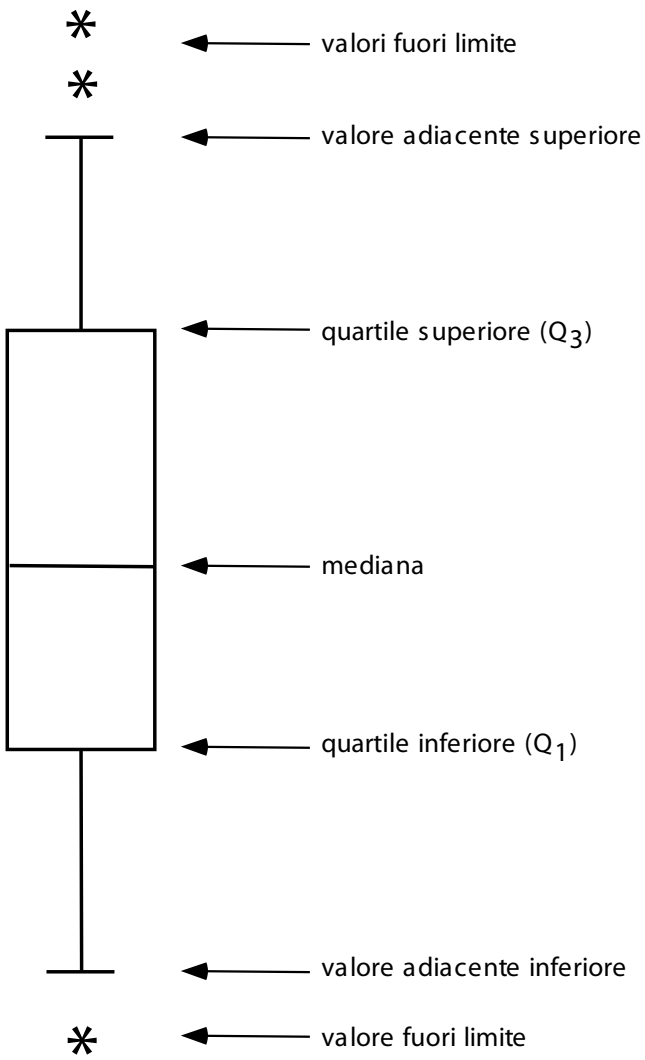
I sommario dei cinque numeri di una distribuzione riporta gli estremi, il primo e terzo quartile e la mediana:

$$Minimo \quad Q_1 \quad Mediana \quad Media \quad Q_3 \quad Massimo$$

- Il grafico a scatola, detto *box plot*, è un metodo grafico proposta dallo statistico americano J. W. Tukey per rappresentare visivamente tre caratteristiche fondamentali di una distribuzione statistica:
  - il grado di dispersione dei dati;
  - la simmetria;
  - la presenza di valori anomali.

- Il grafico è costruito nel modo seguente.
  - Si calcolano i tre quartili della distribuzione.
  - Le due linee esterne orizzontali che delimitano la scatola rappresentano il primo (Q_1) e il terzo quartile (Q_3).
  - La linea orizzontale, interna alla scatola, rappresenta la mediana.
    - Tra Q_1 e Q_3 per definizione sono compresi il 50% delle osservazioni.
    - La distanza interquartilica fornisce informazioni sulla forma della distribuzione (simmetria): se la linea inferiore e superiore hanno distanze differenti dalla mediana, la distribuzione è asimmetrica.

Grafico a scatola



- Le linee che si allungano dai bordi della scatola si concludono con due altre linee orizzontali (baffi).
- Questi punti estremi, evidenziati dai baffi, sono i valori adiacenti:
  - **valore adiacente inferiore**: il valore osservato più piccolo che sia maggiore o uguale a $Q_1 - 1.5\times$ differenza interquartile;
  - **valore adiacente superiore**: il valore osservato più grande che risulta minore o uguale a $Q_3 + 1.5\times$ differenza interquartile.

Grafici a scatola

- I valori esterni a questi limiti sono definiti valori anomali.
  - Nella rappresentazione grafica del *box-plot*, sono segnalati individualmente, poiché costituiscono una anomalia rispetto agli altri dati della distribuzione.

Illustrazione in R

- Consideriamo i dati relativi alla variabile `infant` nel dataframe `Leinhardt` nella libreria `car`.

```
library(car); data(Leinhardt);
europa<-Leinhardt[Leinhardt$region=="Europe",]
```

- Un diagramma a scatola per questi dati può essere generato in R con il comando `boxplot()`.

```
boxplot(europa$infant, horizontal=TRUE,
     col="bisque", xlab="mortalità infantile ogni 1000 nati vivi")
```

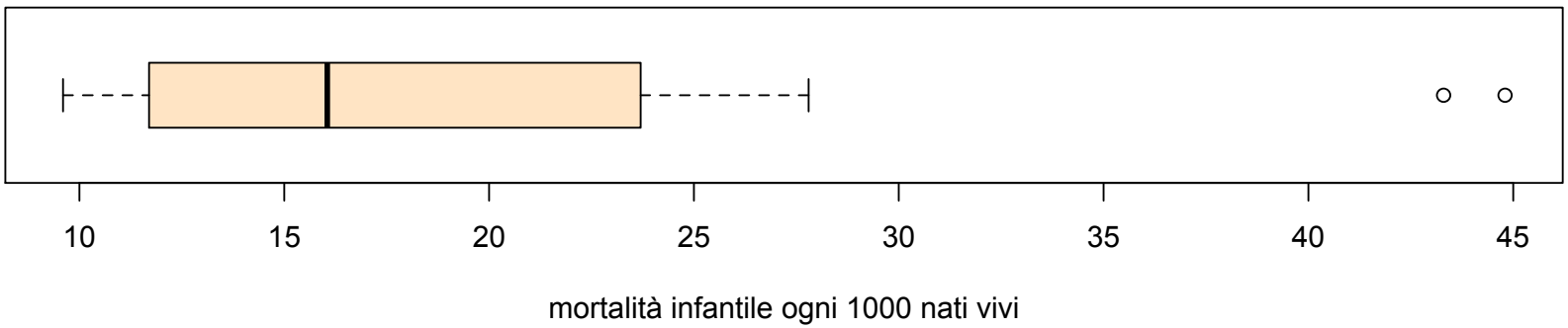


Illustrazione in R

I valori utilizzati da R per generare il diagramma si possono esaminare nel modo seguente:

```
 valori<-boxplot(europa$infant)
 valori$stats
##       [,1]
## [1,]  9.60
## [2,] 11.70
## [3,] 16.05
## [4,] 23.70
## [5,] 27.80
 valori$out
## [1] 44.8 43.3
```

La descrizione di questi valori è fornita da `?boxplot`:

```
stats: a matrix, each column contains the extreme
of the lower whisker, the lower hinge, the
median, the upper hinge and the extreme of the
upper whisker for one group/plot.
```

Illustrazione in R

Il valore adiacente inferiore (*lower whisker*) è il minimo della distribuzione (9.60) – non ci sono valori anomali nella coda sinistra della distribuzione empirica.

```
 summary(europa$infant)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.60   11.97   16.05   19.26   22.88   44.80
```

Illustrazione in R

- Il primo quartile (*lower hinge*) è `11.70`. Questo valore è diverso da quello fornito dalla funzione `summary()` dato che tale funzione calcola il valore che ha sotto di sè *esattamente* il 25% dei dati (anche se tale valore non è presente nella distribuzione). Come già detto si tratta di procedure di interpolazione che assumono una distribuzione continua nei dati.
- Il *lower hinge* corrisponde quindi al valore della distribuzione empirica *più simile al primo quartile "teorico"*.
- Il terzo quartile (*upper hinge*) è `23.7`.

Illustrazione in R

- Il valore adiacente superiore (*upper whisker*) è il valore osservato più grande (`27.8`) che risulta minore o uguale a

$$Q_3 + 1.5 \times \text{differenzainterquartile}$$

.

- Nel caso presente, $Q_3 + 1.5 \times (Q_3 - Q_1)$ è

```
23.70+1.5*(23.7-11.7)
## [1] 41.7
```

Illustrazione in R

- Gli ultimi 4 dati della distribuzione ordinata in ordine crescente sono:

```
sort(europa$infant)[15:18]
## [1] 25.7 27.8 43.3 44.8
```

- Dato che $Q_3 + 1.5 \times (Q_3 - Q_1)$ viene approssimato al valore osservato 27.8, ci sono due valori anomali:

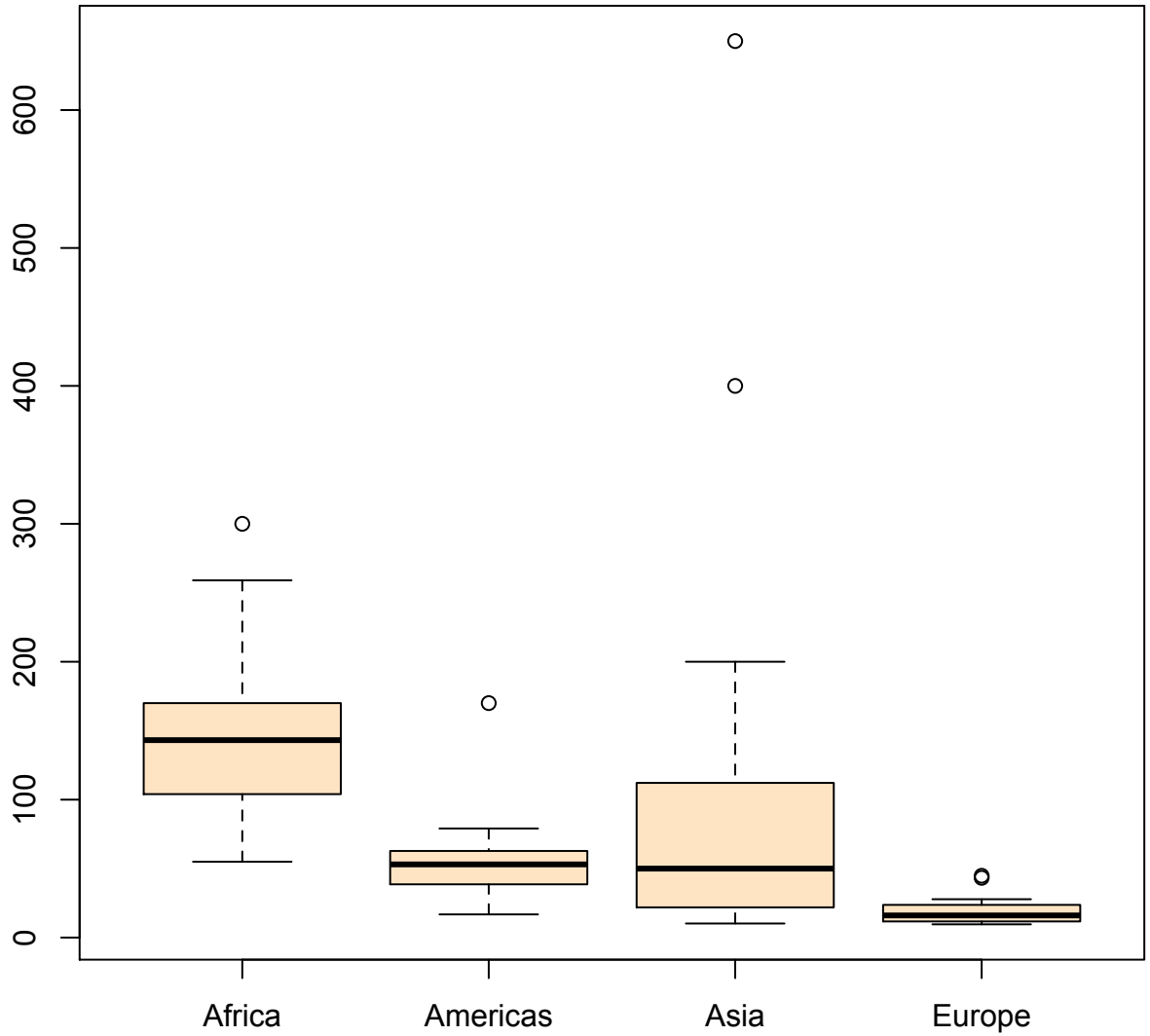
```
valori$out
## [1] 44.8 43.3
```

Illustrazione con R

I box plots sono utili per confrontare distribuzioni diverse. Nel caso dei dati relativi alla mortalità infantile, per esempio, possiamo confrontare le distribuzioni relative ai quattro continenti considerati.

```
boxplot(Leinhardt$infant~Leinhardt$region,col="bisque")
```

Illustrazione in R



Valori anomali e *Trimmed mean*

- La media campionaria è fortemente influenzata dai dati anomali o outliers. La mediana è robusta nei confronti dei dati anomali ma non fa uso di tutte le informazioni fornite dai dati. (La varianza campionaria della mediana è maggiore della varianza campionaria della media)
- Un compromesso che viene talvolta usato è la trimmed mean calcolata eliminando i valori estremi della distribuzione e facendo la media aritmetica dei dati rimanenti.
- Si veda l'argomento `trim` di `?mean` in R.

Valori anomali e *Trimmed mean*

```
 summary(Leinhardt$infant)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    9.60   26.20   60.60   89.05  129.40  650.00       4
 mean(Leinhardt$infant,trim=0,na.rm=TRUE)
## [1] 89.04752
 mean(Leinhardt$infant,trim=0.001,na.rm=TRUE)
## [1] 89.04752
 mean(Leinhardt$infant,trim=0.025,na.rm=TRUE)
## [1] 81.69175
 mean(Leinhardt$infant,trim=0.05,na.rm=TRUE)
## [1] 78.1978
 mean(Leinhardt$infant,trim=0.10,na.rm=TRUE)
## [1] 75.24074
 mean(Leinhardt$infant,trim=0.50,na.rm=TRUE)
## [1] 60.6
```

Varianza

- Benché il sommario dei cinque numeri fornisca un'utile descrizione numerica di una distribuzione, è più comune usare la media quale misura di tendenza centrale e la varianza quale misura di dispersione.

La varianza ci fornisce una misura della dispersione della distribuzione:

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

Deviazione standard

La deviazione standard è la radice quadrata della varianza. È espressa nella stessa unità di misura dei dati originari:

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

Deviazione standard

- La deviazione standard s misura la dispersione attorno alla media e dovrebbe essere usata soltanto quando la media è adeguata per misurare il centro della distribuzione (ovvero, nel caso di distribuzioni simmetriche).
- Quando tutte le osservazioni sono uguali, $s = 0$, altrimenti $s > 0$.
- Come nel caso della media $\bar{x}$, anche la deviazione standard è fortemente influenzata dai valori anomali.

Illustrazione in R

Si considerino i dati del famoso quartetto di Anscombe.

Data set		1-3	1	2	3		4	4
Variable		x	y	y	y		x	y
Obs. no.	1	10.0	8.04	9.14	7.46		8.0	6.58
	2	8.0	6.95	8.14	6.77		8.0	5.76
	3	13.0	7.58	8.74	12.74		8.0	7.71
	4	9.0	8.81	8.77	7.11		8.0	8.84
	5	11.0	8.33	9.26	7.81		8.0	8.47
	6	14.0	9.96	8.10	8.84		8.0	7.04
	7	6.0	7.24	6.13	6.08		8.0	5.25
	8	4.0	4.26	3.10	5.39		19.0	12.50
	9	12.0	10.84	9.13	8.15		8.0	5.56
	10	7.0	4.82	7.26	6.42		8.0	7.91
	11	5.0	5.68	4.74	5.73		8.0	6.89

TABLE. Four data sets, each comprising 11 (x, y) pairs.

Anscombe, Francis J. (1973) Graphs in statistical analysis. *American Statistician*, 27, pp.17-21.

Illustrazione in R

Il *quartetto* è disponibile in R, con il nome di `anscombe`

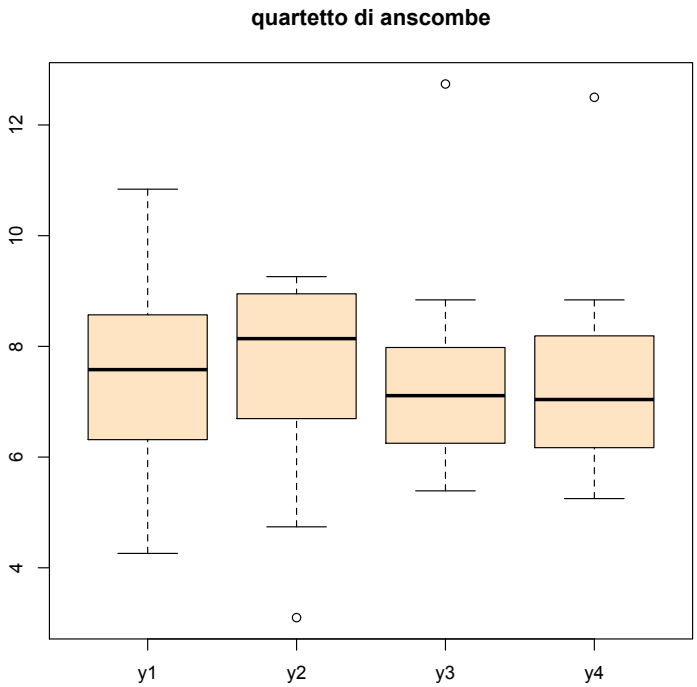
```
data(anscombe)
anscombe
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

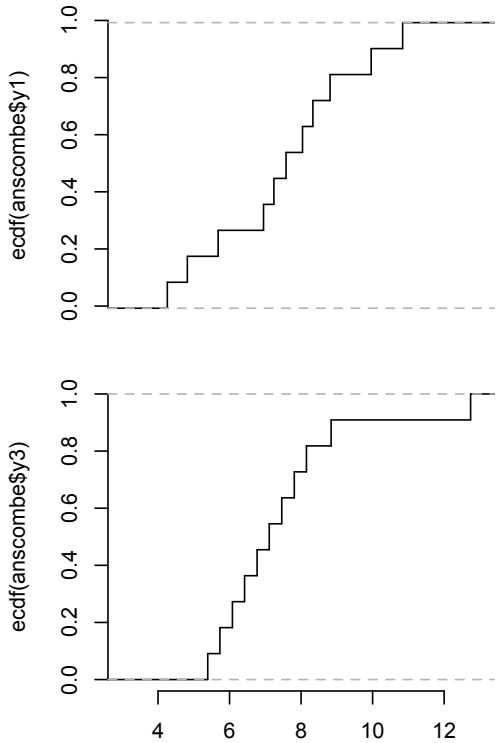
Illustrazione in R

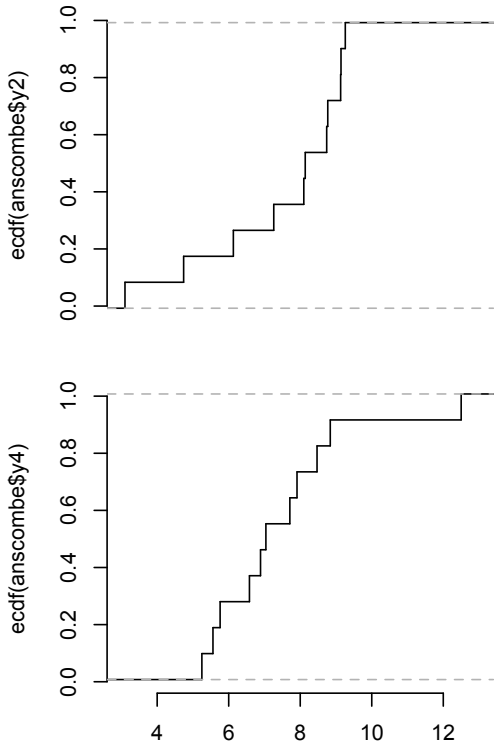
Le due distribuzioni hanno la stessa media e la stessa varianza (e quindi deviazione standard) ma due forme molto diverse

```
 var(anscombe$y1);mean(anscombe$y1)
## [1] 4.127269
## [1] 7.500909
  var(anscombe$y2);mean(anscombe$y1)
## [1] 4.127629
## [1] 7.500909
  var(anscombe$y3);mean(anscombe$y1)
## [1] 4.12262
## [1] 7.500909
  var(anscombe$y4);mean(anscombe$y1)
## [1] 4.123249
## [1] 7.500909
```

Illsutrazione in R







Proprietà della varianza

- La varianza e la deviazione standard non mutano se i dati vengono traslati sommando (o sottraendo) una costante.
- si considerino di dati $x_1, x_2, \ldots, x_n$ e una costante c. Se

$$y_1 = x_1 + c; y_2 = x_2 + c; \ldots; y_n = x_n + c,$$

allora

$$\begin{aligned} s^2_Y &= \sum \frac{\left((x_i+c)-\sum (x_i+c)/n\right)^2}{n-1} \\ &= \sum \frac{\left(x_i+c-\sum x_i/n-nc/n\right)^2}{n-1} \\ &= \sum \frac{\left(x_i+\cancel{c}-\bar{x}-\cancel{c}\right)^2}{n-1} = s^2_X \end{aligned}$$

- Varianza e la deviazione standard sono invece influenzate da un cambiamento della scala di misura.
- Se

$$y_1 = cx_1; y_2 = cx_2; \ldots; y_n = cx_n,$$

allora

$$\begin{aligned} s_Y^2 &= \sum \frac{\left(cx_i - \sum (cx_i)/n\right)^2}{n-1} \\ &= \sum \frac{\left(cx_i - c\sum x_i/n\right)^2}{n-1} \\ &= \sum \frac{c^2\left(x_i - \sum x_i/n\right)^2}{n-1} = c^2 s_X^2 \end{aligned}$$

e $s_Y = |c| s_X$

- Si noti il valore assoluto nell'ultima espressione. Non sono possibili varianze o deviazioni standard negative.

Gradi di libertà

- Nella definizione di varianza, la somma dei quadrati degli scarti dalla media viene divisa per $n - 1$, non per n come avverrebbe per una semplice media aritmetica.
- La divisione per $n - 1$ trova la sua giustificazione nella teoria degli stimatori ed è legata alla nozione di gradi di libertà.

Gradi di libertà

- Nel caso di n dati, $x_1; x_2; \ldots; x_n$, la conoscenza dei primi $n-1$ valori non ci aiuta a conoscere il valore dell'ultimo dato, x_n.
- Se però i dati vengono espressi come scarti dalla media $(x_1 - \bar{x}; x_2 - \bar{x}; \ldots; x_n - \bar{x})$, allora la loro somma deve essere uguale a zero. $\left[\text{Teorema della somma degli scarti}\right]$
- Di conseguenza, l'ultimo scarto dalla media sarà uguale al negativo della somma degli scarti degli altri dati dalla media:

$$(x_n - \bar{x}) = -\left[(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_{n-1} - \bar{x})\right]$$

Illustrazione in R

Per i 18 tassi di mortalità dei paesi europei nel file `Leinhardt`, avremo

```
 scarti<-europa$infant-mean(europa$infant)
 round(scarti,2)
##  [1]  4.44 -2.26 -5.76 -9.16 -6.36  1.14
##  [7] -1.46  6.44 -7.56 -7.66 -7.96 25.54
## [13] -9.66 -6.46 -1.76  8.54 -4.16 24.04
 sum(scarti[1:17])
## [1] -24.04444
 scarti[18]
## [1] 24.04444
```

Diciamo dunque che i $n = 18$ tassi di mortalità hanno 18 gradi di libertà, ma i 18 scarti dalla media hanno solo $n - 1 = 17$ gradi di libertà.

Illustrazione in R

Possiamo inoltra verificare le due proprietà della varianza:

```
 var(europa$infant)
## [1] 109.7085
 var(europa$infant+10) # somma di costante
## [1] 109.7085
 var(europa$infant*10) # cambio di scala
## [1] 10970.85
 var(europa$infant)*100
## [1] 10970.85
 var(europa$infant*(-10))
## [1] 10970.85
```

Da notare inoltre che una variabile X moltiplicata per una costante $c = \frac{1}{s_X}$, avrà varianza pari ad 1; l'unità di misura nuova è quindi la variabilità attorno alla media.

```
 dev.st<-sd(europa$infant)
 var(europa$infant*1/dev.st)
## [1] 1
```

Coefficiente di variazione

Unità di misura La varianza e la deviazione standard dipendono dall'unità di misura in cui la variabile è misurata.

Esempio Una prima variabile che assume valori che si addensano strettamente attorno ad una media di $\bar{x} = 1000$, diciamo, avrà una varianza maggiore di una seconda variabile i cui valori variano grandemente attorno ad una media più piccola, diciamo $\bar{x} = 50$.

Coefficiente di variazione

Per confrontare la dispersione di variabili aventi unità di misura diverse possiamo calcolare il coefficiente di variazione.

$$CV = \frac{s}{\bar{x}}$$

Il coefficiente di variazione è un numero puro (privo di unità di misura) e, dunque, consente il confronto della dispersione di variabili non commensurabili.

Conclusioni

- Operatori di tendenza centrale
  - Media e *Trimmed mean*
  - Mediana
  - Quartili, Percentili e Quantili
- Operatori di dispersione
  - Varianza e deviazione standard
  - Coefficiente di variazione
  -
- Metodi grafici di rappresentazione dei dati
  - Boxplot
  - nella prossima lezione
    - Diagrammi di dispersione
    - Istogramma