Università di Firenze
Corso di laurea in Statistica
A.A. 2010/2011

STATISTICA
parte I, B

Carla Rampichini
rampichini@ds.unifi.it

Leonardo Grilli
grilli@ds.unifi.it

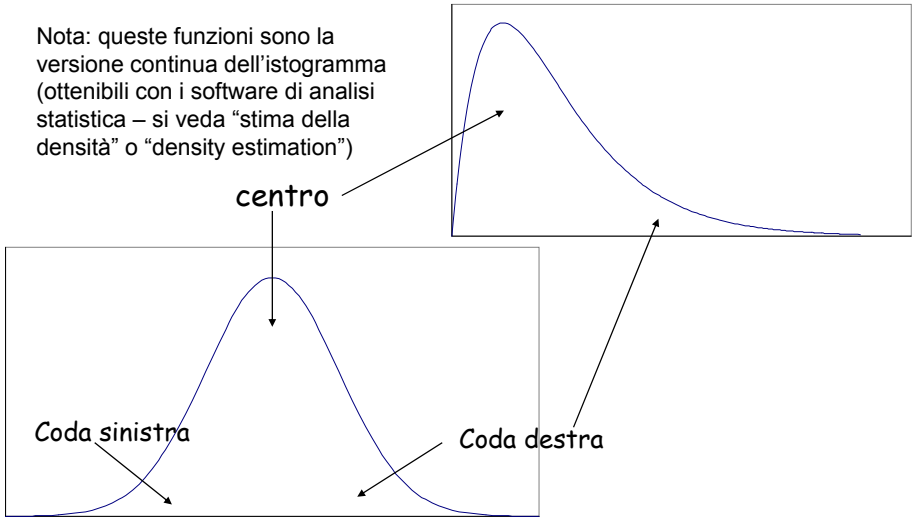
http://www.ds.unifi.it/rampichini/statistica2010_11.htm

Indici di posizione

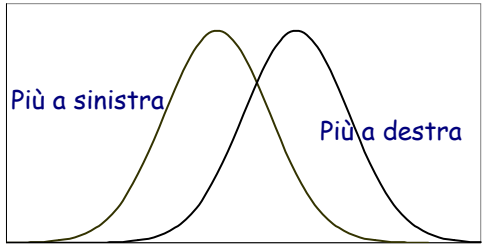
Cicchitelli Cap. 5

Descrivere le distribuzioni

Nota: queste funzioni sono la versione continua dell'istogramma (ottenibili con i software di analisi statistica – si veda "stima della densità" o "density estimation")

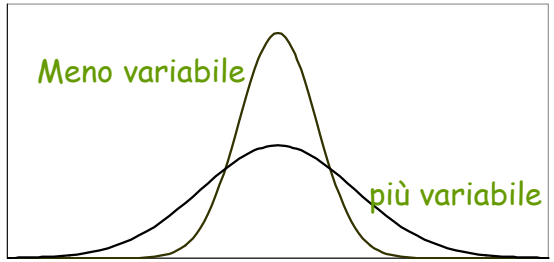


Aspetti caratterizzanti le distribuzioni



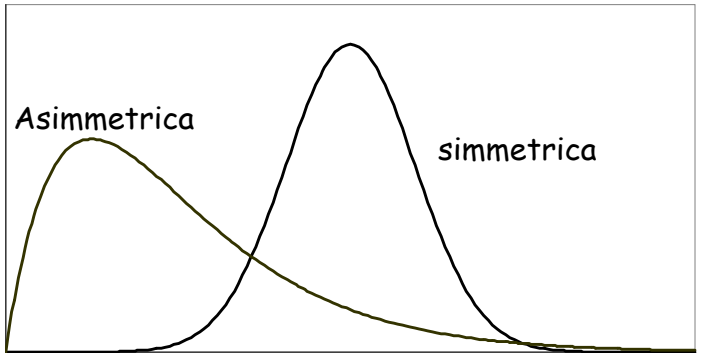
Posizione

Variabilità

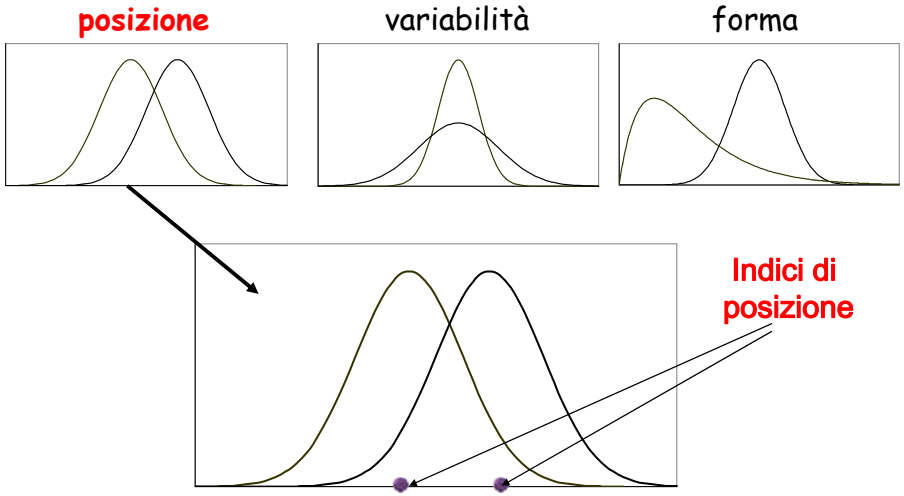


Aspetti caratterizzanti le distribuzioni

Forma



Indici di posizione (o di tendenza centrale)



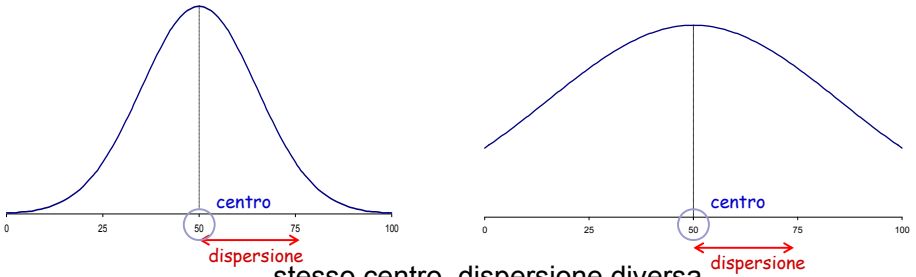
Gli indici di posizione: medie

Sintesi della distribuzione attraverso un valore rappresentativo.

Quali medie sono calcolabili dipende dal tipo di variabile:

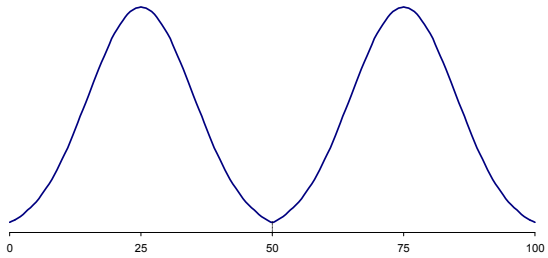
Tipo di variabile	***Moda***	***Mediana***	***Media aritmetica***
Qualitativa nominale	☑	☒	☒
Qualitativa ordinale	☑	☑	☒
Quantitativa	☑	☑	☑

Un indice di posizione è solo una sintesi



stesso centro, dispersione diversa

Se la distribuzione è bimodale il centro non è una buona sintesi della distribuzione!



Media aritmetica

X v.s. quantitativa

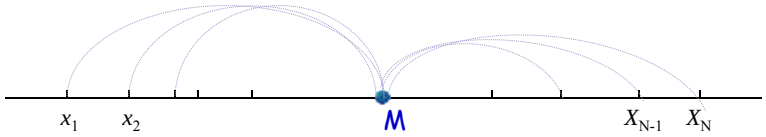
$\{x_1, x_2, \ldots, x_N\}$ successione

$$\mu = M = \frac{1}{N}\sum_{i=1}^{N} x_i$$

esempio X: 2 3 2 4 → M = (2+3+2+4)/4 = 2.75

Se il carattere è discreto la media potrebbe non appartenere all'insieme delle modalità

MEDIA: CENTRO DELL'INSIEME DEGLI *N* PUNTI



Proprietà della media aritmetica

- Internalità (propr. di Cauchy) $x_{\min} \le M \le x_{\max}$
- Baricentro $\sum_{i=1}^{N}(x_i - M) = 0$
- Lascia invariato l'ammontare complessivo: $NM = \sum_{i=1}^{N} x_i$
- Invarianza per trasformazioni lineari

$$Y = a + bX \Rightarrow M(Y) = a + bM(X)$$

- Centro di ordine 2 (minimi quadrati)

$$D(k) = \sum_{i=1}^{N}(x_i - k)^2 \quad \Rightarrow \quad D(k) \text{ è minimo quando } k = M$$

Media come 'centro'

- Distanza di ordine *r* tra l'insieme di punti $\{x_1, x_2, \ldots, x_N\}$ e il punto *k*

$$\sum_{i=1}^{N} |x_i - k|^r$$

- Il centro di ordine *r* dell'insieme di punti $\{x_1, x_2, \ldots, x_N\}$ è il valore che rende minima la distanza di ordine *r*

$$C_r : \arg\min_{C_r} \sum_{i=1}^{N} |x_i - C_r|^r$$

- Per $r = 2$ → C_2=M media aritmetica
- Per $r = 1$ → C_1=M_e mediana

Media aritmetica (distribuzione di frequenze)

Tabella di frequenza

Mod.tà	Freq.	Fr.rel.
x_1	n_1	f_1
x_2	n_2	f_2
…	…	…
x_j	n_j	f_j
…	…	…
x_k	n_k	f_k
Totale	N	1

$$M = \frac{1}{N}\sum_{j=1}^{k} x_j n_j = \sum_{j=1}^{k} x_j f_j$$

Distribuzione di freq: $\{(x_j, n_j)\}_{j=1,2,\ldots,k}$

Due modi di calcolare la media

Distribuzione disaggregata (successione di valori)

39 29 43 52 39 44
40 31 44 35

$$\bar{X} = \frac{39+29+43+\ldots}{10} = 39.6$$

Distribuzione di frequenze (serie)

29	31	35	39	40	43	44	52
1	1	1	2	1	1	2	1

$$\bar{X} = \frac{29\times1+31\times1+35\times1+39\times2+\ldots}{10} = 39.6$$

$$\bar{X} = 29\times\frac{1}{10}+31\times\frac{1}{10}+35\times\frac{1}{10}+39\times\frac{2}{10}+\ldots = 39.6$$

Media aritmetica (dati in classi)

Ipotesi istogramma:
equidistribuzione frequenze all'interno delle classi

Tabella di frequenza

Mod.tà	Freq.	Fr.rel.
x_0-x_1	n_1	f_1
x_1-x_2	n_2	f_2
…	…	…
x_{j-1}-x_j	n_j	f_j
…	…	…
x_{k-1}-x_k	n_k	f_k
Totale	N	1

Valore centrale di classe:

$$c_j = (x_j - x_{j-1})/2$$

$$M = \frac{1}{N}\sum_{j=1}^{k} c_j n_j = \sum_{j=1}^{k} c_j f_j$$

Seriazione: $\{(x_{j-1}; x_j), n_j)\}_{j=1,2,\ldots,k}$

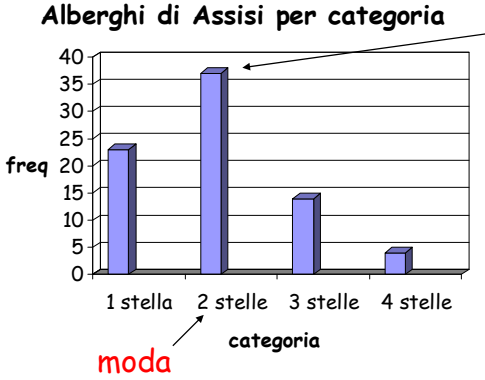
Media ponderata

- Come possiamo calcolare la media degli esami sostenuti, tenendo conto del fatto che gli insegnamenti hanno un numero di crediti diverso?
- Possiamo attribuire ad ogni voto x_i un **peso** w_i pari al numero di crediti dell'insegnamento corrispondente

$$M_w = \frac{1}{\sum_{i=1}^{N} w_i}\sum_{i=1}^{N} w_i x_i$$

La moda

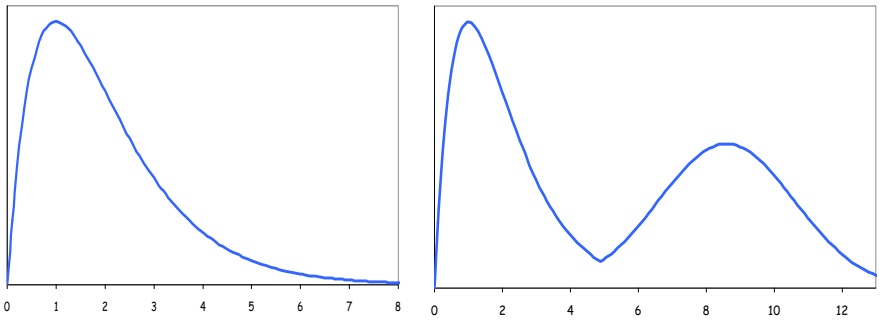
Moda: modalità cui corrisponde la frequenza più alta



Attenzione: nei dati raggruppati in classi la moda è la classe cui corrisponde la ***densità*** più alta (può essere diversa dalla classe con la ***frequenza*** più alta nel caso di classi con ampiezza variabile)

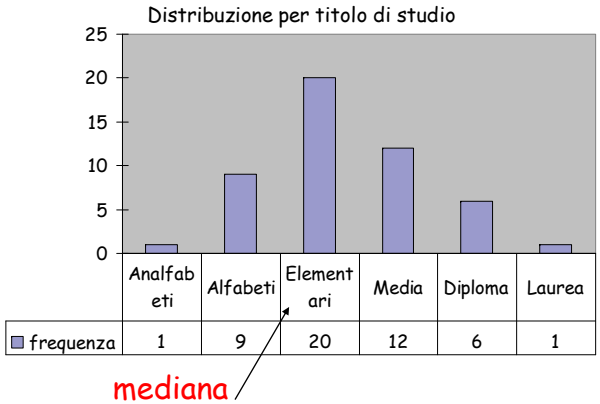
Moda e massimi locali

La moda può essere fuorviante se la distribuzione ha massimi locali



La mediana

Modalità centrale: 50% delle osservazioni stanno sotto e 50% sopra



La mediana (di una successione)

La mediana M_e di *n* numeri ordinati in senso non decrescente $\{y_1, \ldots, y_N\}$ è:

- per *N* dispari $M_e = y_{(N+1)/2}$
- per *N* pari $M_e \in [y_{N/2}; y_{(N/2)+1}]$

se X è quantitativa, $M_e = [y_{N/2} + y_{(N/2)+1}]/2$

Modalità centrale: 50% delle osservazioni stanno sotto e 50% sopra

Proprietà della mediana

- Internalità $x_{\min} \le M_e \le x_{\max}$
- Centro di ordine 1 $M_e : \sum_{i=1}^{N} |x_i - M_e| = \min$
- Applicabile anche a v.s. ordinali
- M_e non risente di valori anomali: resta invariata se si sostituiscono i termini $x < M_e$ o $x > M_e$

Calcolo della mediana tramite la funzione di ripartizione

- X: numero atti aggressivi in un'ora di gioco
- 138 bambini di 2/3 anni

x_j	1	2	3	4	5	6	7	8	9	10	tot
n_j	3	8	30	45	22	12	10	5	2	1	138
N_j	3	11	41	86	108	120	130	135	137	138	
$F(x_j)$	0.02	0.08	0.3	0.62	0.78	0.87	0.94	0.98	0.99	1.0	

Mediana: primo valore di x_j per cui vale $F(x_j) > 0.5$

Attenzione: se esiste x_j per cui vale $F(x_j) = 0.5$, allora la mediana è tra x_j e x_{j+1}

Calcolo della mediana per dati in classi (ipotesi dell'istogramma)

Per definizione:

Estremo inferiore della classe mediana

$$F(M_e) = F(x_{m-1}) + (M_e - x_{m-1})\, d_m = 0.5$$

Densità della classe mediana

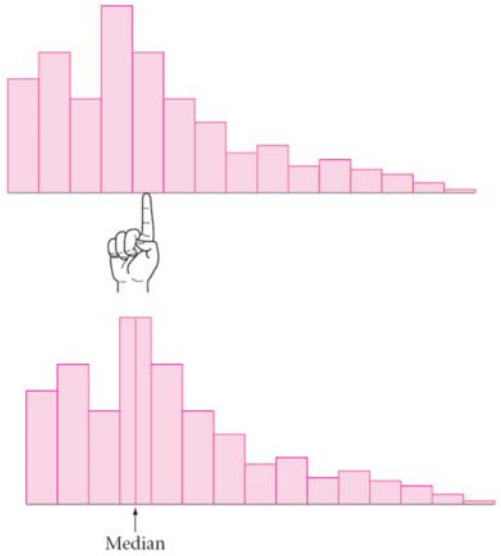
Quindi:
(1) Trovare la classe mediana
(2) Calcolare M_e come segue

$$M_e = x_{m-1} + \frac{0.5 - F(x_{m-1})}{d_m}$$

Media vs mediana

- Sono entrambi indici di posizione → indicano il **centro** della distribuzione
- La mediana divide la distribuzione in due parti uguali
- La media è il punto di equilibrio dell'istogramma, come una bilancia, si ottiene sommando i valori e dividendo per il numero di valori

Media vs mediana



Per trovare la media osservando un istogramma, trovate il punto in cui dovreste mettere un dito sotto l'asse orizzontale per tenere in equilibrio la distribuzione immaginando che i rettangoli abbiano un peso proporzionale alla loro area.

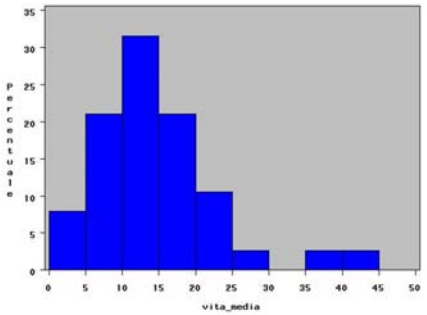
La mediana divide l'area dell'istogramma in due parti uguali (in termini di area)

Media e mediana: speranza di vita dei mammiferi

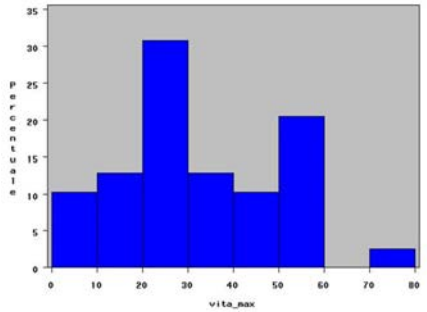
- Il valore in cui l'istogramma sta in equilibrio (media=13,1) è più grande del valore che divide l'area in due parti uguali (la mediana=12) (per il calcolo si veda il foglio excel)

⇩

la distribuzione non è simmetrica

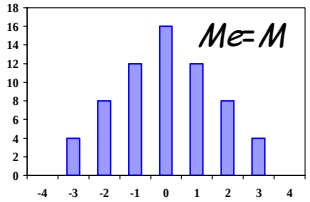


- Se la distribuzione fosse simmetrica media e mediana sarebbero uguali
- I valori anomali a destra tendono a far crescere il valore medio ma non hanno effetto sulla mediana
- Per esempio, se i valori della classe [35, 40) fossero *spostati* nella classe [45, 50) la mediana resterebbe uguale mentre la media sarebbe più grande!

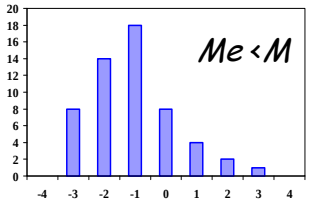


Media vs mediana

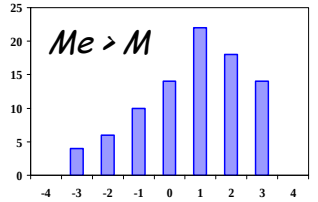
Simmetria



Asimmetria positiva



Asimmetria negativa



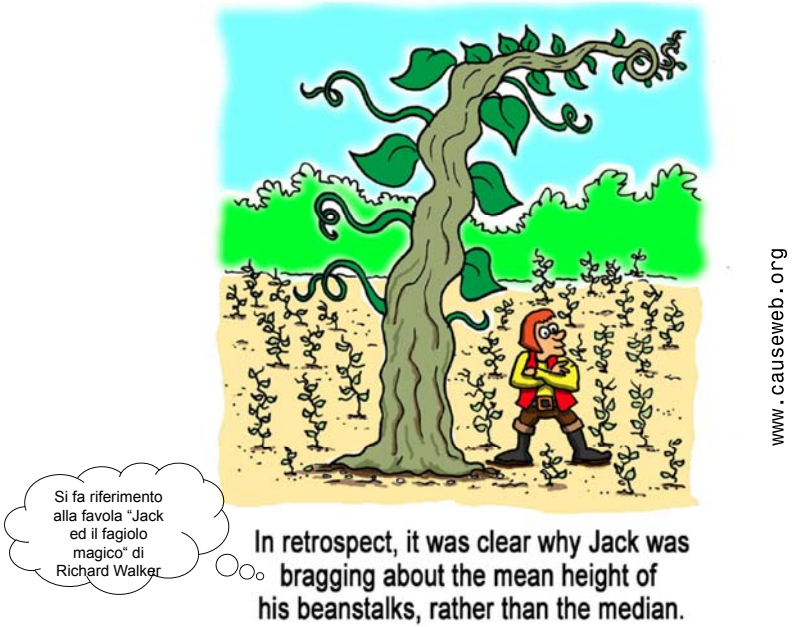
Pro e contro della mediana

- Usa solo in parte l'informazione contenuta nei dati (l'ordine ma non i valori)
  - ☹ dati diversi possono avere la stessa mediana
  - ☺ è un indice *robusto*, cioè non è influenzato dai valori estremi (outliers)

29 31 35 39 39 40 43 44 44 52 — Me = 39.5 M= 39.6

⇩

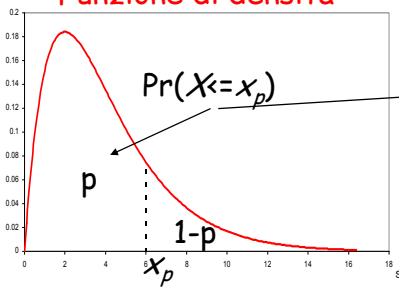
29 31 35 39 39 40 43 44 44 92 — Me = 39.5 M= 43.6



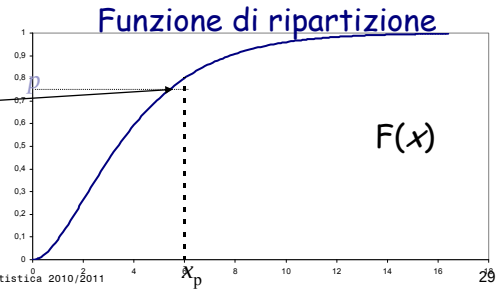
Quantili

- p=0.01,0.02,…,0.98,0.99 Percentili
- p =0.1,0.2, …,0.8,0.9 Decili
- p =0.25,0.50,0.75 Quartili
- p =0.5 Mediana

Funzione di densità

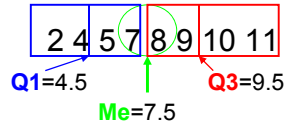


Funzione di ripartizione

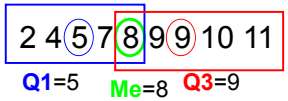


Trovare i quartili (distribuzione disaggregata)

- Ordinate i valori dal più piccolo al più grande
- Dividete i valori in due parti uguali
- Quindi dividete ciascuna metà ancora in due parti uguali (se N dispari → mediana inclusa in entrambe le parti)
- Esempio1: n. di film visti in un anno da 8 studenti



- Esempio2: n. di film visti in un anno da 9 studenti



Definizione di quartile

$x_1, x_2, \ldots, x_N$ distribuzione disaggregata

$y_1, y_2, \ldots, y_N$ distribuzione ordinata $(y_1 \le y_2 \le \ldots \le y_N)$

$\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}$ frequenze relative cumulate

y_{i-1} e y_i termini a cui corrispondono $\frac{i-1}{N}$ e $\frac{i}{N}$ tali che

$$\frac{i-1}{N} \le \frac{l}{4} \le \frac{i}{N} \qquad (l = 1,2,3)$$

Si chiama l-mo $(l = 1,2,3)$ quartile la quantità

$$q_l = \begin{cases} \frac{y_{i-1} + y_i}{2} & \text{se } \frac{i-1}{N} = \frac{l}{4} \\ y_i & \text{altrimenti} \end{cases}$$

Cicchitelli Def. 5.7

Calcolo dei quantili

- Possiamo dividere la distribuzione in 10 parti uguali considerando i decili, in 100 parti uguali considerando i centili, ecc.
- In generale, consideriamo la frazione $p \in (0,1)$.

y_{i-1} e y_i termini a cui corrispondono $\frac{i-1}{N}$ e $\frac{i}{N}$ tali che

$$\frac{i-1}{N} \le p \le \frac{i}{N} \qquad p \in (0,1)$$

Si chiama quantile di ordine p la quantità

$$q_p = \begin{cases} \frac{y_{i-1} + y_i}{2} & \text{se } \frac{i-1}{N} = p \\ y_i & \text{altrimenti} \end{cases}$$

(per il calcolo si veda il foglio excel)

Calcolo dei quartili tramite la funzione di ripartizione

- X: numero atti aggressivi in un'ora di gioco
- 138 bambini di 2/3 anni

x_j	1	2	3	4	5	6	7	8	9	10	tot
n_j	3	8	30	45	22	12	10	5	2	1	138
N_j	3	11	41	86	108	120	130	135	137	138	
$F(x_j)$	0.02	0.08	0.3	0.62	0.78	0.87	0.94	0.98	0.99	1.0	

Primo valore di x_j per cui vale $F(x_j) > p$, per p=0.25, 0.5, 0.75

Attenzione: se esiste x_j per cui vale $F(x_j) = p$, allora il corrispondente quartile è tra x_j e x_{j+1}

Calcolo dei quantili per dati raggruppati in classi (ipotesi dell'istogramma)

$$p \in (0,1)$$

$$x_{[p]} : pr\left\{X \le x_{[p]}\right\} = F(x_{[p]}) = p$$

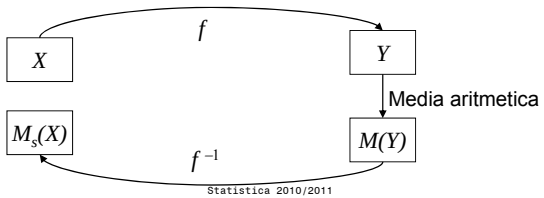
1) Trovare la classe (x_{j-1}, x_j) in cui F supera p

2) Calcolare $$x_{[p]} = x_{j-1} + \frac{p - F(x_{j-1})}{d_j}$$

Medie di potenze (momenti)

$$M_s = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^s \right)^{1/s}$$

- s= 1 $M_1=M$ media aritmetica
- s= 2 $M_2=M_q$ media quadratica
- s=–1 $M_{-1}=M_a$ media armonica
- s→0 $M_0=M_g$ media geometrica



Media quadratica

- $f(x)=x^2$

$$M_2 = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} = \left(\frac{\sum_{i=1}^{N} x_i^2}{N} \right)^{1/2}$$

M_2: valore che sostituito agli N termini della successione ne lascia invariata la somma dei quadrati

Media geometrica

$f(x)= \log x$ (logaritmo naturale)

$$M_g = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log x_i\right) = \exp\left(\sum_{i=1}^{N}\log x_i^{\frac{1}{N}}\right) = \prod_{i=1}^{N}\exp(\log x_i^{\frac{1}{N}})$$

$$M_g = \left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}}$$

- M_g valore che sostituito agli N termini della successione ne lascia invariato il prodotto
- M_g applicata ad una progressione geometrica (con N dispari) fornisce il termine centrale della progressione

Media geometrica: esempio

- La media geometrica consente di calcolare il tasso medio di crescita
- Esempio: un capitale investito per tre anni ha fatto registrare i seguenti rendimenti: 2%, 18%, 10%.

Qual è il tasso di rendimento medio?

$$C_{finale} = C_{iniziale}(1.02)(1.18)(1.10)$$

$$= C_{iniziale}(1+r)^3$$

Obiettivo: trovare r tale che

$$(1+r)^3 = (1.02)(1.18)(1.10)$$

$$\Rightarrow \quad 1+r = \left[(1.02)(1.18)(1.10)\right]^{\frac{1}{3}} = 1.098057$$

$$\Rightarrow \quad r = 0.098057 \text{ (ovvero } 9.8\%)$$

Media armonica

- $f(x)=1/x$

$$M_a = \frac{N}{\sum_{i=1}^{N}\frac{1}{x_i}}$$

Si usa quando il reciproco di x ha un significato e l'obiettivo è lasciare invariata la somma dei reciproci

Media armonica: esempio

- Tempo impiegato da tre falegnami per realizzare una sedia: 1h 2h 2h

x (ore per una sedia)	1/x (sedie in un'ora)
1	1
2	1/2
2	1/2

- In un'ora i 3 falegnami realizzano 2 sedie → mediamente ognuno realizza 2/3 di sedia in un'ora, ovvero per una sedia impiega 3/2 di ora (cioè un'ora e mezzo)

$$M_a = \frac{N}{\sum_{i=1}^{N}\frac{1}{x_i}} = \frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{2}} = \frac{3}{2} = 1.5$$

Quale media?

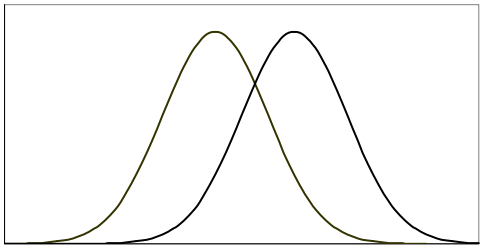
1. Le medie calcolabili dipendono dal tipo di variabile: se nominale si può calcolare solo la moda, se quantitativa si possono calcolare moda, mediana e medie analitiche
2. La scelta mediana vs medie analitiche dipende dalla asimmetria della distribuzione e dalla presenza di outliers
3. La media analitica standard è la media aritmetica
   - Tuttavia in alcuni casi la natura del fenomeno suggerisce l'uso di una media diversa da quella aritmetica: es. la *media armonica* dei tempi lascia invariata la produttività totale, oppure la *media geometrica* lascia invariato il montante finale di un investimento a interesse composto

"Sorry, we don't let people discard outliers without a good reason."

Indici di variabilità

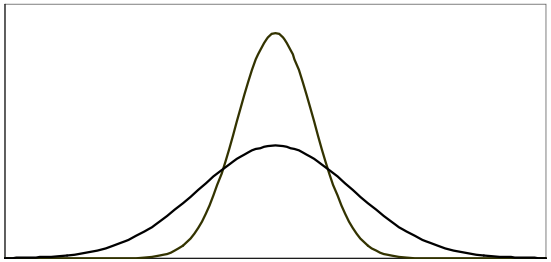
Cicchitelli Cap. 6

Variabilità (o dispersione)



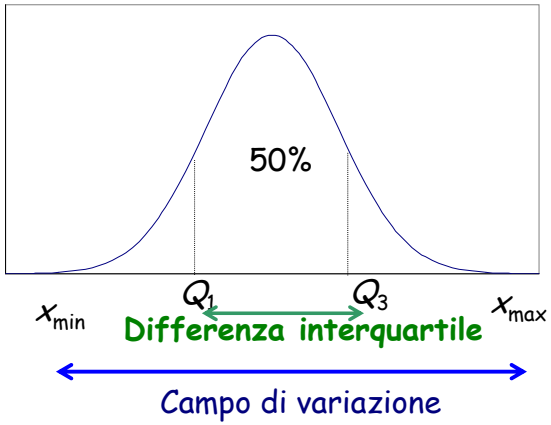
media diversa,
stessa variabilità

stessa media,
variabilità diversa



Indicatori elementari di variabilità

- Campo di variazione (range): R= x_{max}-x_{min}
- Differenza interquartile: DI=Q_3-Q_1



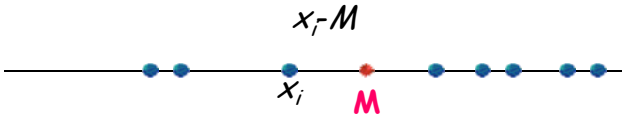
Varianza e deviazione standard

Scostamento dalla media $x_i - \mu$

Devianza $D = \sum_{i=1}^{N} (x_i - \mu)^2$

Varianza $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$

Deviazione standard $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$



Calcolo della varianza (distribuzione disaggregata)

alimento	*energia kcal*	*xi-M*	(xi-M)^2
pane	276	-90.25	8145.06
grissini	433	66.75	4455.56
crackers	428	61.75	3813.06
fette	410	43.75	1914.06
biscotti	418	51.75	2678.06
pasta	356	-10.25	105.06
riso	362	-4.25	18.06
pizza	247	-119.25	14220.56
Totale	2930	0.00	35349.50

$\mu = 366.25$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 = \frac{1}{8} 35349.5 = 4418.688$$

Calcolo della varianza (distribuzione di frequenze)

xj	nj	fj	xj*fj	xj-M	(xj-m)^2	fj(xj-m)^2
1	5	0.083	0.083	-2.5	6.25	0.520833
2	10	0.167	0.333	-1.5	2.25	0.375
3	15	0.25	0.75	-0.5	0.25	0.0625
4	15	0.25	1	0.5	0.25	0.0625
5	10	0.167	0.833	1.5	2.25	0.375
6	5	0.083	0.5	2.5	6.25	0.520833
totale	60	1	3.5	M	σ^2	1.916667
ds					σ	1.384437

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{J} (x_j - \mu)^2 n_j = \sum_{j=1}^{J} (x_j - \mu)^2 f_j$$

Media e varianza con dati raggruppati

Ipotesi istogramma:
equidistribuzione frequenze all'interno delle classi

Tabella di frequenza

Mod.tà	Freq.	Fr.rel.
x_0-x_1	n_1	f_1
x_1-x_2	n_2	f_2
…	…	…
x_{j-1}-x_j	n_j	f_j
…	…	…
x_{k-1}-x_k	n_k	f_k
Totale	N	1

Valore centrale di classe: $c_j = (x_j - x_{j-1})/2$

$$\mu = \sum_{j=1}^{k} c_j f_j$$

Approssima la vera media, a volte per difetto, a volte per eccesso

$$\sigma^2 = \sum_{j=1}^{J} (c_j - \mu)^2 f_j$$

Approssima la vera varianza, quasi sempre per difetto

Calcolo della varianza: formula alternativa

$$\sigma^2 = M_2^2 - M^2 = M(X^2) - [M(X)]^2$$

Varianza = (media quadratica al quadrato) – (media aritmetica al quadrato)

Nell'esempio delle kcal degli alimenti

$$M_2^2 = 138557.8$$

$$M = 366.25$$

$$\sigma^2 = 138557.8 - (366.25)^2 = 4418.688$$

Vedremo più avanti che nell'ambito dell'inferenza statistica il divisore della varianza non è N ma N-1

divisore N → varianza della popolazione

divisore N-1 → varianza campionaria

Attenzione: in molti software la varianza di default è quella campionaria

Es. in Excel

VAR() → divisore N-1

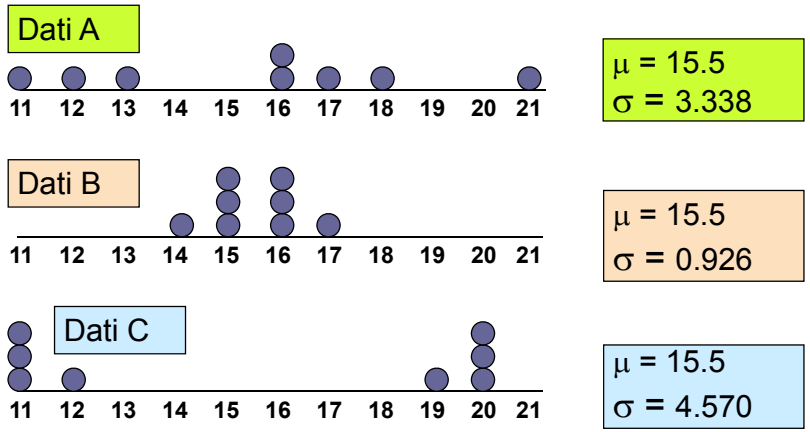
VAR.POP() → divisore N

www.causeweb.org

The tattoo parlor near campus got busy when the professor required hand calculation of the standard deviation.

Interpretare la deviazione standard

Deviazione standard: media quadratica degli scostamenti dalla media



Proprietà della deviazione standard

1. Stessa unità di misura di X
2. Non negatività

$$\sigma(X) \geq 0, \quad \text{con } \sigma(X) = 0 \Leftrightarrow X \text{ degenere}$$

3. Invarianza rispetto a traslazioni

$$\sigma(a + X) = \sigma(X)$$

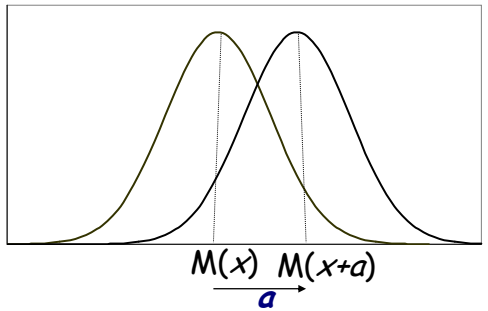
4. Omogeneità

$$\sigma(bX) = b\sigma(X)$$

Proprietà della deviazione standard

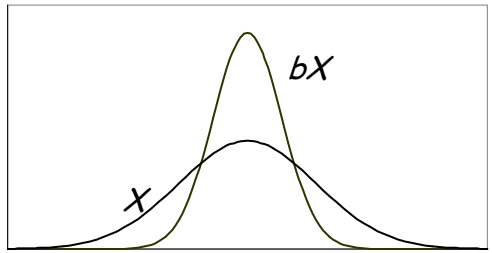
Invarianza rispetto a traslazioni

$$\sigma(a + X) = \sigma(X)$$



Omogeneità

$$\sigma(bX) = b\sigma(X)$$



in questo esempio $0<b<1$

Proprietà della deviazione standard

- La deviazione standard è molto sensibile ai valori anomali
- Alternativa robusta: lo scarto interquartile
- In termini di robustezza …

la deviazione standard sta allo scarto interquartile
come
la media aritmetica sta alla mediana

Riepilogo: effetto di una traslazione

$$(x_1, x_2, \ldots, x_n) \rightarrow \mu \quad \sigma^2 \quad \sigma$$

$$(x_1 + a, x_2 + a, \ldots, x_n + a) \rightarrow \mu + a \quad \sigma^2 \quad \sigma$$

Es. Reddito in euro, media 950 e Dev.Std. 70

Prelievo 30 euro ognuno ($a = -30$) → media 920 e Dev.Std. 70

Riepilogo: effetto di un cambiamento di scala

$$(x_1, x_2, \ldots, x_n) \rightarrow \mu \quad \sigma^2 \quad \sigma$$

$$(bx_1, bx_2, \ldots, bx_n) \rightarrow b\mu \quad b^2\sigma^2 \quad b\sigma$$

Es. Altezze in cm, media 172 e Dev.Std. 8

Trasformazione in metri (b=1/100) → media 1.72 e Dev.Std. 0.08

Quale coppia di indici?

- Quale indice di posizione e dispersione utilizzare dipende anche dall'obiettivo con cui si calcolano questi indici.
- Se l'obiettivo è meramente descrittivo, e la variabile è quantitativa, gli indici più informativi sono:
  - la media aritmetica e la deviazione standard se la distribuzione è simmetrica unimodale
  - la mediana e lo scarto interquartile se la distribuzione è asimmetrica o presenta valori anomali

Indici di variabilità relativi

- Utili per confrontare la variabilità di due distribuzioni quando:
  - Unità di misura diverse senza alcuna relazione
  - Stessa unità di misura, ma intensità media diversa
- Possibili soluzioni
  - relativizzare rispetto a una media (es. il CV)
  - relativizzare rispetto a un valore massimo
- Sono numeri puri, cioè senza unità di misura

Coefficiente di variazione

$$CV = \frac{\sigma}{\mu} 100 \qquad (x_i \geq 0,\ \mu \neq 0)$$

- È un numero puro (espresso in % ma non ha massimo)
- È definito solo per variabili con media diversa da 0, ed è utile per variabili che assumono valori solo positivi
- Consente il confronto tra la variabilità di fenomeni:
  - in unità di misura non omogenee (es. in una popolazione di bambini c'è più variabilità nel peso o nell'altezza?)
  - con diverso ordine di grandezza (es. riguardo al peso, c'è più variabilità tra i neonati o tra le madri?)

Utilizzo del CV

- Per $\mu \to 0$ il CV $\to \infty$: non usare quando la media è piccola!
- Il CV dovrebbe essere calcolato solo per variabili misurate su scala di rapporti

Scala di intervalli: esempio temperatura

							media	*ds*	*cv*
C	20	18	21	23	15	20	19,50	2,50	0,13
K	293,2	291,2	294,2	296,2	288,2	293,2	292,65	2,50	0,01
F	68	64,4	69,8	73,4	59	68	67,10	4,50	0,07

°K=°C − 273.15

°F =°C × 9/5 + 32

Scala di rapporti: esempio massa

peso							*media*	*ds*	*cv*
grammi	50	48	52	49	45	58	50,33	4,03	0,08
libbre	0,11	0,106	0,115	0,108	0,099	0,128	0,11	0,01	0,08
kg	0,5	0,48	0,52	0,49	0,45	0,58	0,50	0,04	0,08

1lb	453,6 gr
1gr	0,002205

Indici di eterogeneità

Carattere di qualunque natura: si usano solo le frequenze

Minima eterogeneità (= massima omogeneità)

Modalità	x_1	x_2	…	x_i	…	x_k	**Totale**
Frequenza	**0**	**0**	…	N	…	**0**	N

Massima eterogeneità

Modalità	x_1	x_2	…	x_i	…	x_k	**Totale**
Frequenza	N/k	N/k	…	N/k	…	N/k	N

Indici di eterogeneità

Indice di Gini

$$G = 1 - \sum_{i=1}^{k} f_i^2 \qquad G \in \left[0, \frac{k-1}{k}\right]$$

Indice di entropia

$$H = -\sum_{i=1}^{k} f_i \log f_i \qquad H \in [0, \log k]$$

Dividendo per il massimo si ottengono le versioni normalizzate

Calcolo indici di eterogeneità

Indici di eterogeneità per la valutazione di tre corsi di Ingegneria

a.a. 1999/2000 II sem

Distribuzione di frequenza relativa per corso

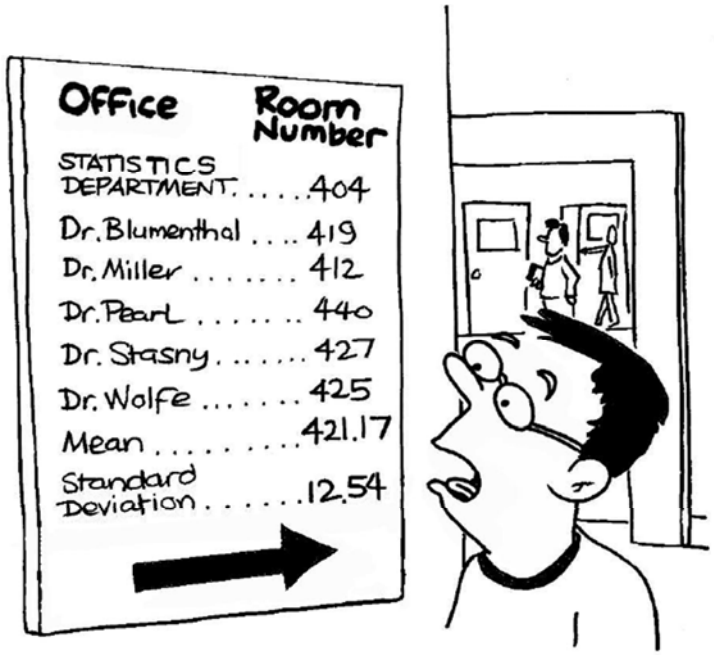
xj	A	B	C
dec no	0.030	0.200	0.056
+no/sì	0.194	0.311	0.361
+sì/no	0.463	0.356	0.569
dec sì	0.313	0.133	0.014
TOT	1.000	1.000	1.000

INDICE DI GINI: fJ^2

	A	B	C
	0.001	0.040	0.003
	0.038	0.097	0.130
	0.214	0.126	0.324
	0.098	0.018	0.000
	0.351	0.281	0.458
G	0.649	0.719	0.542
G'	0.866	0.959	0.723

INDICE DI ENTROPIA: fj*logfj (log base e)

A	B	C
-0.105	-0.322	-0.161
-0.318	-0.363	-0.368
-0.357	-0.368	-0.321
-0.364	-0.269	-0.059
-1.143	-1.321	-0.909
1.143	1.321	0.909
0.825	0.953	0.655



Indici di forma

Cicchitelli Cap. 7

La forma della distribuzione

- *Forme tipiche:*
  - rettangolare o uniforme 
  - simmetrica
  - a campana 
    - la più nota curva a campana simmetrica è la Normale
  - asimmetrica (a destra o a sinistra) 
  - bimodale 

Distribuzioni simmetriche

- Una distribuzione è simmetrica quando
  - le modalità a sinistra e a destra della mediana sono equidistanti dalla mediana
  - e ogni coppia di modalità equidistanti ha la stessa frequenza

Modalità	3	5	6	7	9	Totale
Frequenza	4	2	10	2	4	22

Distribuzioni simmetriche

- Proprietà
  - $M = M_e$ (= Moda, se unimodale)
  - $\sum_{i=1}^{N}(x_i - M)^r = 0$ per ogni r dispari
  - $|Q_1 - M_e| = |Q_3 - M_e|$

Un indice di asimmetria (skewness)

$$\alpha_1 = \frac{1}{\sigma^3}\left[\frac{1}{N}\sum_{i=1}^{k}(x_i - M)^3 n_i\right]$$

- $\alpha_1 > 0$ → asimmetria positiva
- $\alpha_1 < 0$ → asimmetria negativa

Distribuzione simmetrica ⇄ $\alpha_1 = 0$

Un indice di asimmetria: esempio

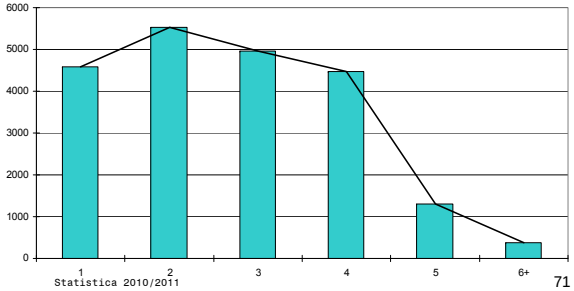
n. comp.	famiglie	xj*nj	(xj-M)^2*nj	(xj-M)^3*nj
1	4594	4594	13449.41	-23011.94
2	5528	11056	5588.85	-3973.67
3	4955	14865	1241.51	358.80
4	4467	17867	29686.79	38266.28
5	1294	6472	33910.70	77621.60
6+	382	2674	49184.47	210952.20
totale	21220	57527	133061.74	300213.26

Famiglie italiane (migliaia) per numero di componenti – 1998 (Fonte: ISTAT)

Asimmetria positiva

Indice di asimmetria

$\alpha_1 = 0.901$



$\alpha_1=0$ non implica simmetria

x_i	$(x_i - \mu)^3$
1	-729.00
5	-125.00
9	-1.00
10	0.00
10	0.00
10	0.00
17	343.00
18	512.00
Somma =	0.00

$\mu =$ 10.00

$\sigma =$ 5.24

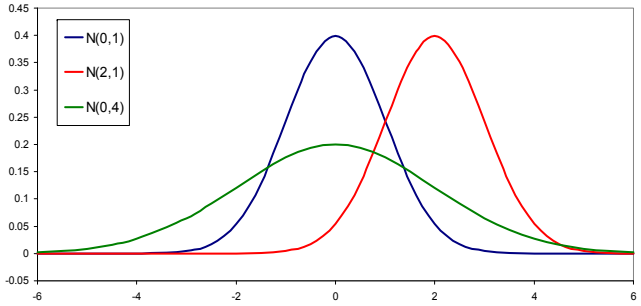
Distribuzione Normale o di Gauss

Parametri: $\mu, \sigma \qquad \mu \in \Re, \sigma \in \Re_+$

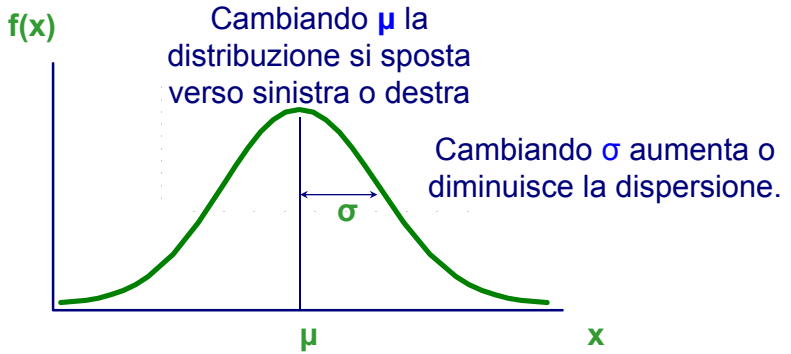
Funzione di densità:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad x \in \Re$$

μ	0	2	0
σ	1	1	2
		f(x)	
x	N(0,1)	N(2,1)	N(0,4)
-3	0.004432	1.49E-06	0.064759
-2.5	0.017528	1.6E-05	0.091325
-2	0.053991	0.000134	0.120985
-1.5	0.129518	0.000873	0.150569
-1	0.241971	0.004432	0.176033
-0.5	0.352065	0.017528	0.193334
0	0.398942	0.053991	0.199471
0.5	0.352065	0.129518	0.193334
1	0.241971	0.241971	0.176033
1.5	0.129518	0.352065	0.150569
2	0.053991	0.398942	0.120985
2.5	0.017528	0.352065	0.091325
3	0.004432	0.241971	0.064759



Forma della distribuzione Normale



Date la media μ e la varianza σ identifichiamo la distribuzione normale con la notazione

$X \sim N(\mu, \sigma^2)$

Curtosi [dal gr. kyrtós 'curvo, arcuato']

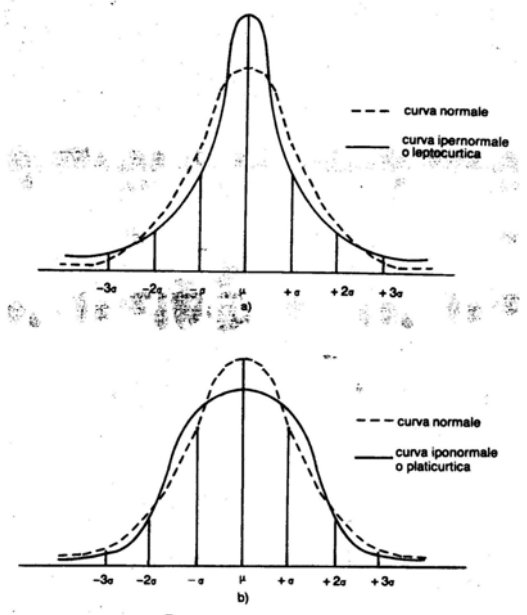


Fig. 23. Curve ipernormali e iponormali

Per distribuzioni simmetriche la curtosi valuta la frequenza nelle code, e il corrispondente appuntimento al centro, rispetto alla distribuzione normale con medesima media e deviazione std

Indice di curtosi (kurtosis)

$$\gamma = \frac{1}{\sigma^4}\left[\frac{1}{N}\sum_{i=1}^{k}\left(x_i - M\right)^4 n_i\right] - 3$$

- $\gamma > 0$ → ipernormale (code pesanti)
- $\gamma < 0$ → iponormale (code leggere)

Distribuzione normale ⟷ $\gamma = 0$

Boxplot

Disuguaglianza di Chebychev e regola empirica

Cicchitelli Cap. 8

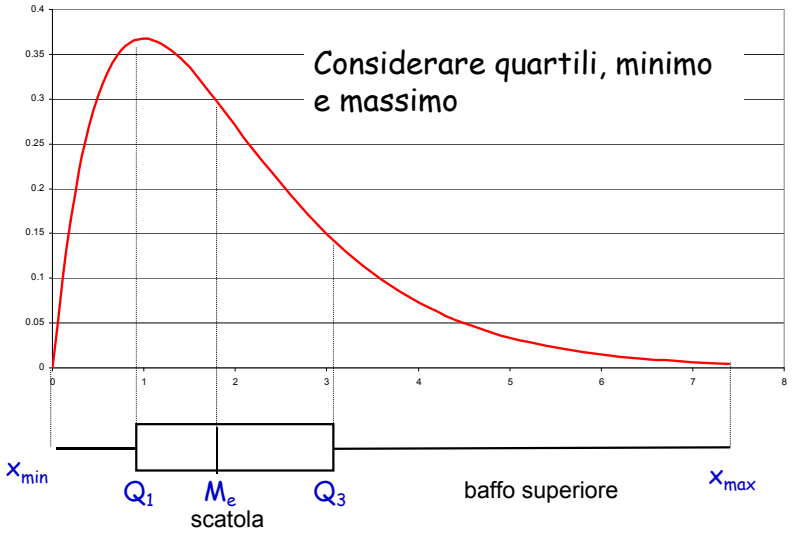
Sintetizzare la distribuzione con 5 numeri

- **minimo**: il più piccolo valore osservato
- **Q1**: la mediana della prima metà dei valori
- **Mediana**: il valore che divide i dati in due parti
- **Q3**: la mediana della metà superiore dei valori
- **massimo**: il valore più grande osservato

```
1 | 1 2                    min      11
2 | 0 5                    Q1       30
3 | 0 0 0 2 5 9            median   37
4 | 0 0 0 2 5 8            Q3       42
5 | 0                      max      70
6 |
```

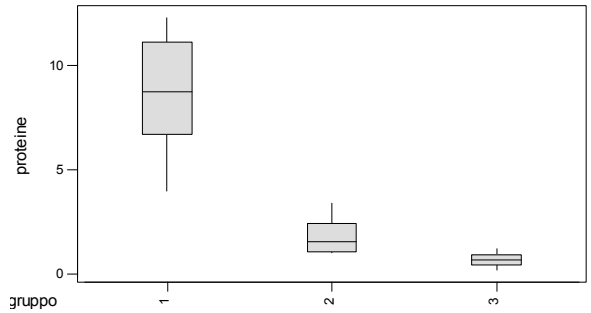
Boxplot (diagramma a scatola)

Boxplot (versione A)



Boxplot (versione A)

contenuto proteico 20 alimenti per gruppo



Descriptive Statistics: proteine by gruppo

Variable	gruppo	N	Mean	Median	TrMean	StDev
proteine	1	8	8.688	8.750	8.688	2.787
	2	6	1.783	1.550	1.783	0.898
	3	6	0.667	0.650	0.667	0.333

Variable	gruppo	SE Mean	Minimum	Maximum	Q1	Q3
proteine	1	0.985	4.000	12.300	6.700	11.175
	2	0.366	1.000	3.400	1.075	2.425
	3	0.136	0.200	1.200	0.425	0.900

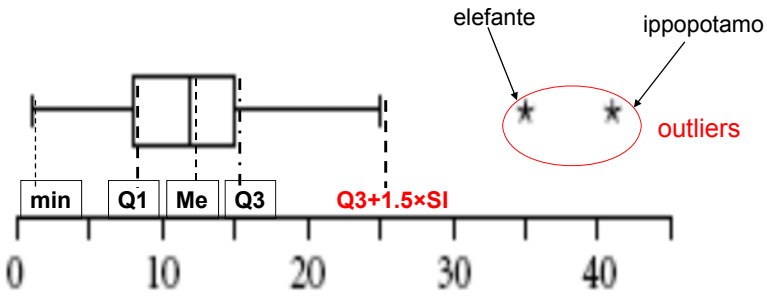
Boxplot (versione B)

- minimo: 1
- Quartile inferiore (*Q*1): **8** posizione 38*(1/4)=9.5➔10
- mediana: **12** posizione 38*(1/2)=19➔19 e 20
- Quartile superiore (*Q*3): **15** posizione 38*(3/4)=28.5➔29
- Massimo: **41**

Esempio:
Speranza di vita
di N=38 mammiferi

```
0 | 1 3 4
  | 5 5 5 6 7 7 8 8
1 | 0 0 0 2 2 2 2 2 2 2 2
  | 5 5 5 5 5 5 5 6
2 | 0 0 0 0
  | 5
3 |
  | 5
4 | 1
```


Boxplot (versione B)



Lunghezza del baffo: 1.5*SI (Scarto Interquartile, ovvero Q3-Q1)

Nota: il baffo viene troncato se supera il min o il max

Le osservazioni al di fuori dei baffi sono indicate con un simbolo

Boxplot senza baffi!

- Sì. Possono esserci boxplot senza baffi!
- Per esempio, in questo insieme di 12 dati

{1, 1, 1, 1, 2, 3, 5, 6, 7, 12, 14, 16}

il minimo e il primo quartile sono uguali

Disuguaglianza di Chebyshev

Per una distribuzione qualunque con

media μ

deviazione standard σ

si scelga arbitrariamente un valore $\delta > 0$

Allora, posto $Freq\{I\}$ = frequenza relativa complessiva dei termini che si trovano nell'intervallo I, si ha

$$Freq\{\mu - \delta < x < \mu + \delta\} \geq 1 - \frac{\sigma^2}{\delta^2}$$

Disuguaglianza di Chebyshev

Versione alternativa con $\delta = k\sigma$

Per una distribuzione qualunque con

media μ

deviazione standard σ

si scelga arbitrariamente un valore $k \geq 1$

Allora, posto $Freq\{I\}$ = frequenza relativa complessiva dei termini che si trovano nell'intervallo I, si ha

$$Freq\{\mu - k\sigma < x < \mu + k\sigma\} \geq 1 - \frac{1}{k^2}$$

Disuguaglianza di Chebyshev: esempi

Indipendentemente da come i dati sono distribuiti, almeno $(1 - 1/k^2)$ dei valori cadranno entro k deviazioni standard dalla media (per $k \geq 1$)

- Esempi:

k=1 $(1 - 1/1^2) = 0\%$ ..... $(\mu \pm 1\sigma)$

k=2 $(1 - 1/2^2) = 75\%$ ..... $(\mu \pm 2\sigma)$

k=3 $(1 - 1/3^2) = 89\%$ ..... $(\mu \pm 3\sigma)$

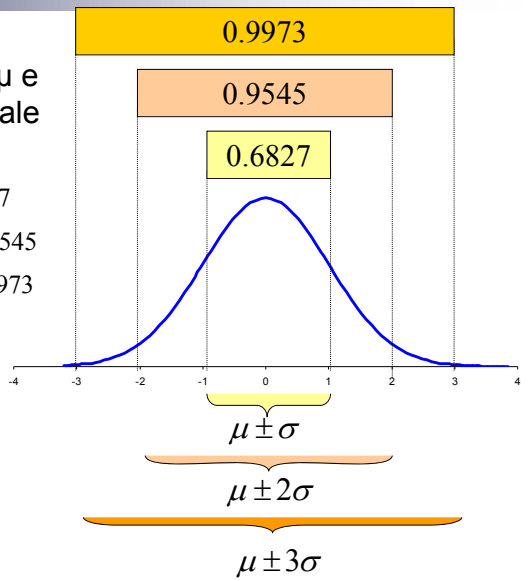
Distribuzione Normale

Se i dati seguono una distribuzione con media μ e deviazione standard σ, vale

$Freq\{\mu - \sigma < x < \mu + \sigma\} = 0.6827$

$Freq\{\mu - 2\sigma < x < \mu + 2\sigma\} = 0.9545$

$Freq\{\mu - 3\sigma < x < \mu + 3\sigma\} = 0.9973$



Regola empirica

- La distribuzione normale è un modello teorico: i dati sono discreti!
- Tuttavia, se l'istogramma ha una forma campanulare i dati hanno una distribuzione approssimativamente normale
- In tal caso, le frequenze 68%, 95% e 99.7% della normale valgono approssimativamente per i dati

regola empirica

se i dati hanno una distribuzione di forma campanulare, circa il 68% dei valori si trova nell'intervallo $\mu \pm 1\sigma$, circa il 95% nell'intervallo $\mu \pm 2\sigma$ e circa il 99.7% nell'intervallo $\mu \pm 3\sigma$

Disuguagl. di Chebyshev vs regola empirica

k	intervallo	Disuguagl. Chebyshev	Regola empirica
1	$\mu \pm \sigma$	$\geq 0\%$	$\cong 68\%$
2	$\mu \pm 2\sigma$	$\geq 75\%$	$\cong 95\%$
3	$\mu \pm 3\sigma$	$\geq 89\%$	$\cong 99.7\%$

La regola empirica

- è più informativa (è in termini di $\cong$ invece che $\geq$)
- però si applica solo alle distribuzioni campanulari