Capitolo 4

- La variabilità di una distribuzione
- Intervalli di variabilità
- Box-plot
- Indici basati sullo scostamento dalla media
- Confronti di variabilità
- Standardizzazione

La variabilità

- La **variabilità** di una distribuzione esprime la tendenza delle unità di un collettivo ad assumere diverse modalità della variabile

- Un **indice di variabilità** deve soddisfare almeno due requisiti:
  - deve assumere il valore minimo se e solo se tutte le unità della distribuzione presentano uguale modalità della variabile
  - Deve aumentare all'aumentare della "diversità" tra le modalità assunte dalle varie unità.

Osservazione

Esempio:

Consideriamo due gruppi di pazienti, ciascuno composto da 5 soggetti, nei quali viene osservato il numero di giorni di trattamento con un dato farmaco.

Dall'osservazione risultano i seguenti dati:

	Valori osservati di durata trattamento (in giorni)	**Media aritmetica**
Gruppo 1	1 4 5 1 4	(1+4+5+1+4)/5= 15/5=3
Gruppo 2	3 3 3 3 3	(3+3+3+3+3)/5=3 15/5=3

La media aritmetica non riflette la **variabilità** interna a ciascun gruppo in quanto dipende solo dalla somma complessiva dei valori osservati (uguale nei due gruppi e pari a 15 giorni di trattamento complessivo) e dal numero di osservazioni effettuate (5 pazienti per ogni gruppo) e coincide con il valore osservabile in una situazione di assenza di variabilità nei valori osservati (situazione rappresentata dal gruppo 2).

Osservazione

Compliance	N. pazienti	%
Buona	0	0
Discreta	30	100
Scarsa	0	0
Totale	30	100

Situazione 1: variabilità nulla

Compliance	N. pazienti	%
Buona	5	16,7
Discreta	17	56,7
Scarsa	8	26,6
Totale	30	100

Situazione 2: variabilità intermedia

Compliance	N. pazienti	%
Buona	10	33,3
Discreta	10	33,3
Scarsa	10	33,3
Totale	30	100

Situazione 3: variabilità massima

Compliance	N. pazienti	%
Buona	5	16,7
Discreta	17	56,7
Scarsa	8	26,6
Totale	30	100

Situazione a: variabilità intermedia

Compliance	N. pazienti	%
Buona	16	53,4
Discreta	7	23,3
Scarsa	7	23,3
Totale	30	100

Situazione b: variabilità intermedia

Intervalli di variabilità

- Il **campo di variazione**

Dati *n* valori e ordinati in senso crescente,

$$x_1 \leq x_2 \leq \cdots \leq x_n$$

si considera la differenza tra il più grande e il più piccolo valore:

$$R = x_n - x_1$$

- non fornisce indicazioni su come si distribuisce la variabile tra i due valori limite;
- i valori estremi che definiscono il campo di variazione possono essere influenzati da oscillazioni accidentali;
- dipende dal numero di osservazioni: tende ad aumentare al crescere del numero di osservazioni;
- è espresso nella stessa unità di misura della variabile.

Esempio:

N. patologie concomitanti	N. pazienti	
1-2	15	**minimo valore = 1**
3-4	5	**massimo valore = 6**
5-6	10	**campo di variazione 5 (1,6)**
Totale	**30**	

Intervalli di variabilità

La **differenza interquartilica**

Dati *n* valori consideriamo la differenza tra il terzo e il primo **quartile**:

$$W = Q_3 - Q_1$$

Oss: rappresenta il campo di variazione per il 50% delle unità più vicine alla mediana.

Box-plot

Un modo per rappresentare graficamente la variabilità di una distribuzione è dato dal box-plot.

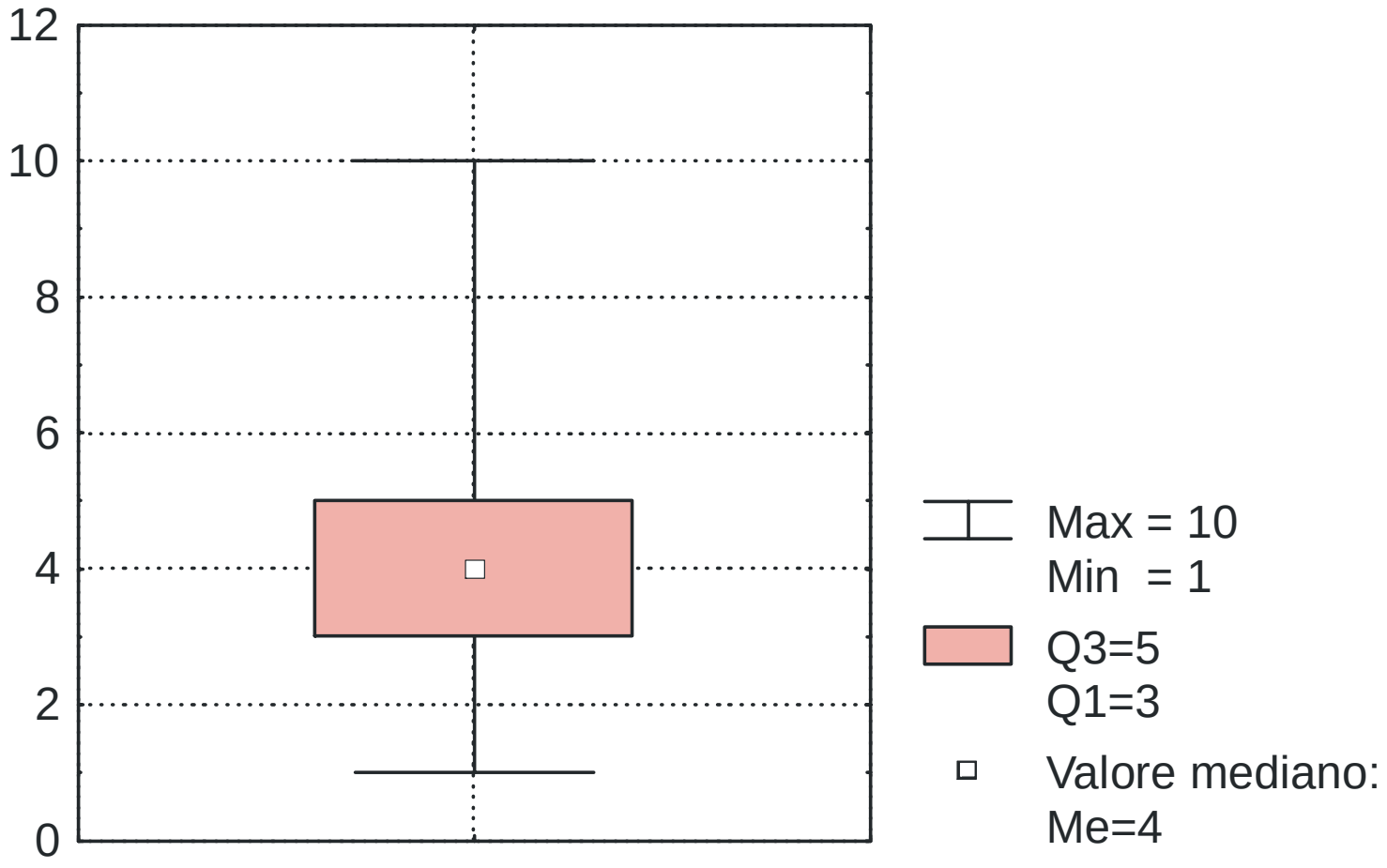
Il **box-plot** è un grafico caratterizzato da tre elementi:

- una linea o punto, che indicano la posizione della media della distribuzione;
- Un rettangolo (box) la cui altezza indica la variabilità dei valori "prossimi" alla media;
- Due segmenti che partono dal rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione.

Ad esempio, come media si può prendere la mediana, come altezza del box la distanza interquartile e come estremi dei segmenti il valore minimo e massimo osservati.

Box-plot: esempio

N° atti aggressivi	1	2	3	4	5	6	7	8	9	10
frequenza	3	8	30	45	22	12	10	5	2	1



Esempio

Con i dati dell'esempio relativo al peso di 20 studenti

48 54 56 57 65 66 68 68 69 70 70 71 72 73 75 76 76 78 84 85

1° quartile

3° quartile

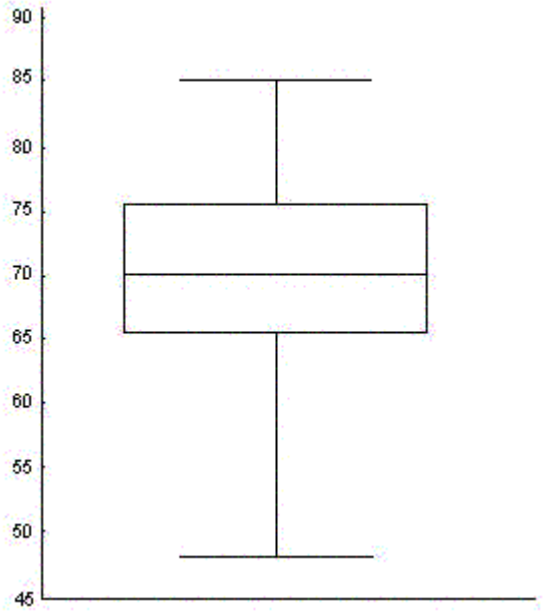
il primo quartile si trova tra il 5° e 6° posto e assume il valore kg 65,5.
Il terzo quartile si trova tra il 15° e il 16° posto e assume il valore kg 75,5.
Abbiamo, dunque:

a) (85-48)=37
b) 65,5 e 75,5
c) 70

Range

Range interquartile

Mediana



http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_2.htm

Esercizio

In una classe liceale di un istituto del Centro Italia sono stati rilevati i pesi di 10 alunni maschi. I pesi (in kg) sono i seguenti:

75 69 65 73 83 62 73 68 64 66.

Dopo aver calcolato i valori dei quartili (primo, secondo e terzo) e la differenza interquartile, costruire il box plot relativo e commentare il risultato.

Soluzione:

Mettiamo in ordine i dati:

62 64 65 66 68 69 73 73 75 83

Primo quartile (Q_1) = 64,5

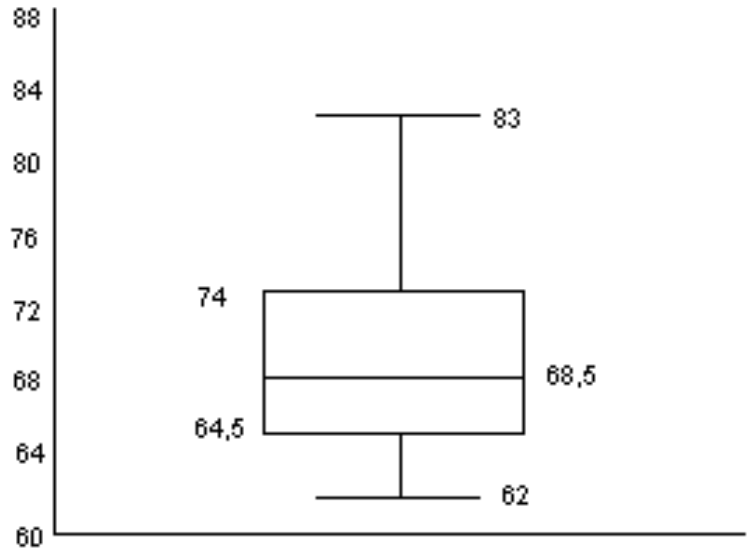
Secondo quartile (mediana) (Q_2) = 68,5

Terzo quartile (Q_3) = 74

Differenza interquartile = 74-64,5=9,5

Min = 62

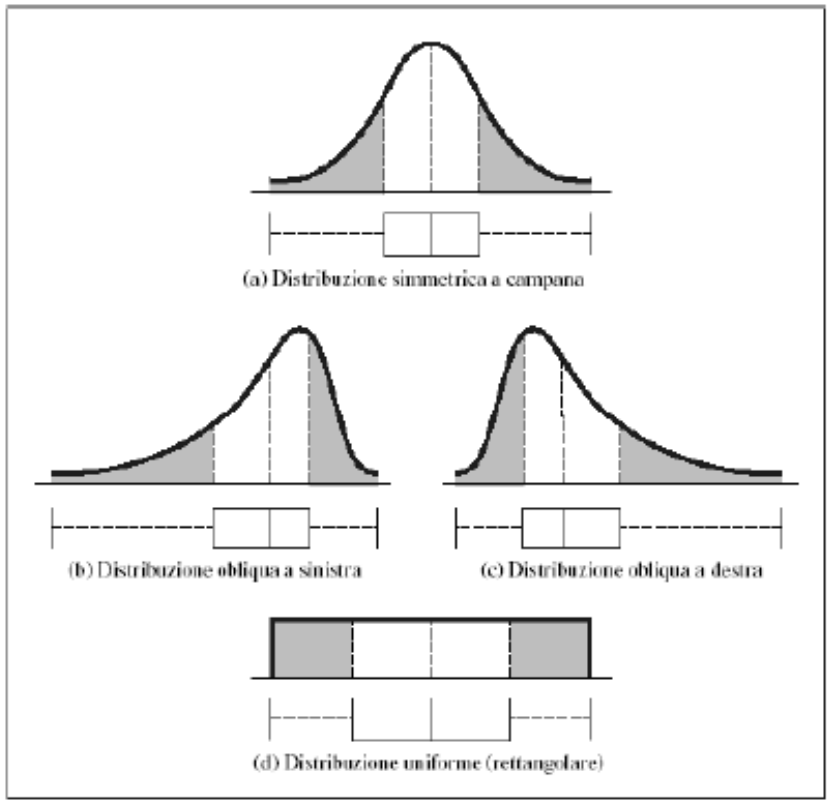
Max= 83



http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_2.htm

Il diagramma a “Scatola e Baffi” (o Boxplot)

Per valutare la relazione che sussiste tra i metodi di analisi esplorativa dei dati, come il diagramma scatola e baffi, e le rappresentazioni grafiche, come i poligoni, consideriamo la Figura, nella quale sono riportati i diagrammi scatola e baffi e i poligoni relativi a quattro ipotetiche distribuzioni.



NOTA: l’area sottostante a ciascuna curva è divisa nei quartili corrispondenti ai cinque numeri di sintesi su cui si basa il diagramma scatola e baffi.

La varianza

Un indice basato sugli scostamenti dalla media aritmetica è la **varianza**.

La varianza di n valori $x_1, x_2, \ldots, x_n$ di una variabile X con media aritmetica $\bar{x}$ è:

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

Il numeratore è detto **devianza**: $\sum_{i=1}^{n}(x_i - \bar{x})^2$

- **Calcolo semplificato della varianza**

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

Si vuol costruire una misura di variabilità con il concorso di tutti i valori osservati.

Lo **scarto** o **scostamento** è una quantità estremamente importante nello studio della variabilità.
E' definito come differenza di due valori osservati:

$$\textit{scarto} = \textit{(valore A} - \textit{valore B)}$$

o come la differenza tra un valore osservato e un valore medio:

$$\textit{scarto} = \textit{(valore A} - \textit{media aritmetica)}$$

sintetizziamo tali valori in una quantità, per esempio una media aritmetica:

$$\textit{scarto medio} = \frac{\textit{somma degli scarti dalla media aritmetica}}{\textit{n. totale degli scarti}}$$

essendo questa quantità sempre pari a zero, sintetizziamo gli scarti elevandoli al quadrato:

$$\frac{\textit{somma degli scarti dalla media aritmetica elevati al quadrato}}{\textit{n. totale degli scarti}}$$

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

La deviazione standard

Alcune considerazioni sulla varianza:

- è nulla in assenza di variabilità e aumenta all'aumentare della variabilità;
- svolge un ruolo rilevante in ambito inferenziale.

Osservazione: la varianza non possiede la stessa unità di misura dei valori della distribuzione (è espressa nel quadrato dell'unità di misura della variabile).

Si può utilizzare quindi come indice di variabilità la **deviazione standard** o **scarto quadratico medio** che è espresso nella stessa unità di misura della variabile:

$$\sigma = \sqrt{\sigma^2}$$

Alcune considerazioni sulla deviazione standard:

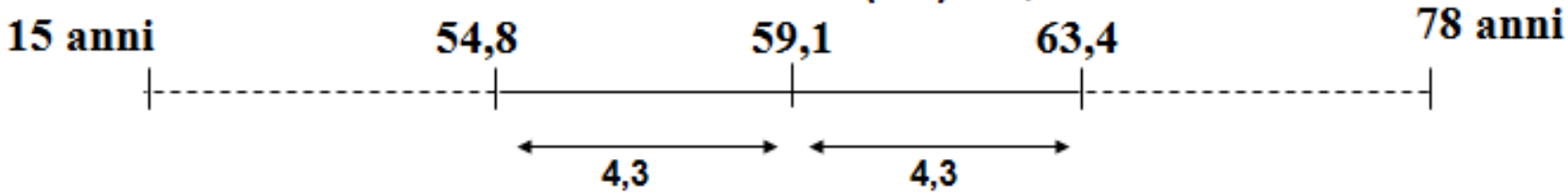
- è espressa nella stessa unità di misura della variabile;
- è nulla in assenza di variabilità e aumenta all'aumentare della variabilità;
- svolge un ruolo rilevante sia in ambito descrittivo che inferenziale.

La DEVIAZIONE STANDARD è una misura della variabilità e definisce la distanza media delle osservazioni dalla loro media aritmetica o intervallo medio di valori.

Maggiore è il valore della deviazione standard, maggiore è la variabilità. E' nulla in assenza di variabilità e aumenta all'aumentare della variabilità.

Età media = 59,1 anni

Deviazione Standard (DS) = 4,3 anni



In media i pazienti presentano un'età compresa tra 54,8 e 63,4 anni.

Calcolo della varianza e deviazione standard per una lista di valori individuali

Sono stati osservati i seguenti valori dell'età per 6 soggetti:

50 63 56 47 58 61

$$\text{media aritmetica} = \frac{50+63+56+47+58+61}{6} = 55.8$$

$$varianza = \frac{(50-55.8)^2 + (63-55.8)^2 + (56-55.8)^2 + (47-55.8)^2 + (58-55.8)^2 + (61-55.8)^2}{6}$$

$$= 39$$

$$deviazione\,standard = \sqrt{39} = 6.2 \text{ anni}$$

Intervallo medio dei valori: (55.8-6.2, 55.8+6.2) = (49.6, 62.0)

In media i 6 soggetti presentano un'età compresa tra 49.6 e 62 anni.

Calcolo della varianza e deviazione standard per una distribuzione di frequenza

Nell'ambito di una indagine viene chiesto a 53 pazienti: "Nell'ambito di quelle che sono le sue conoscenze della malattia, che voto (da 0 a 3) darebbe all'assistenza che le viene offerta?". I risultati sono riportati nella tabella seguente:

Punteggi	N. pazienti	%
0	7	11.7
1	16	26.7
2	25	41.6
3	12	20.0
Totale	**60**	**100**

$$\textit{media aritmetica} = \left(11.7 \times 0 + 26.7 \times 1 + 41.6 \times 2 + 20.0 \times 3\right) / 100$$

$$= 1.7$$

$$\textit{varianza} = \frac{\left(0-1.7\right)^2 \times 7 + \left(1-1.7\right)^2 \times 16 + \left(2-1.7\right)^2 \times 25 + \left(3-1.7\right)^2 \times 12}{60}$$

$$= 0.5$$

$$\textit{deviazione standard} = \sqrt{0.5} = 0.7 \qquad \text{intervallo medio} = (1, 2.4)$$

I pazienti in media danno voti da 1 a 2.4.

Esempio

Numero di imprese (migliaia) nel 1991 in cinque regioni italiane: 268, 106, 76, 238, 88

- $\overline{x} = 155,2$
- $\sigma^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - 155,2)^2 =$

  $= 6557,76$
- $\sigma = \sqrt{6557,76} = 80,98$

Esercizio

Sono state rilevate le seguenti lunghezze in cm di 10 grilli della stessa varietà:

6,5 6,9 6,1 5,8 7,2 6,4 6,8 6,9 7,4 5,6.

Calcolare il campo di variazione e lo scarto quadratico medio

Soluzione:
Per calcolare il campo di variazione (range) è opportuno mettere per prima cosa in ordine in ordine i dati:
5,6 5,8 6,1 6,4 6,5 6,8 6,9 6,9 7,2 7,4.
Il range è:
7,4 - 5,6 = 1,8.

La media aritmetica è:

$$M = \frac{5{,}6 + 5{,}8 + 6{,}1 + 6{,}4 + 6{,}5 + 6{,}8 + 6{,}9 + 6{,}9 + 7{,}2 + 7{,}4}{10} = 6{,}6.$$

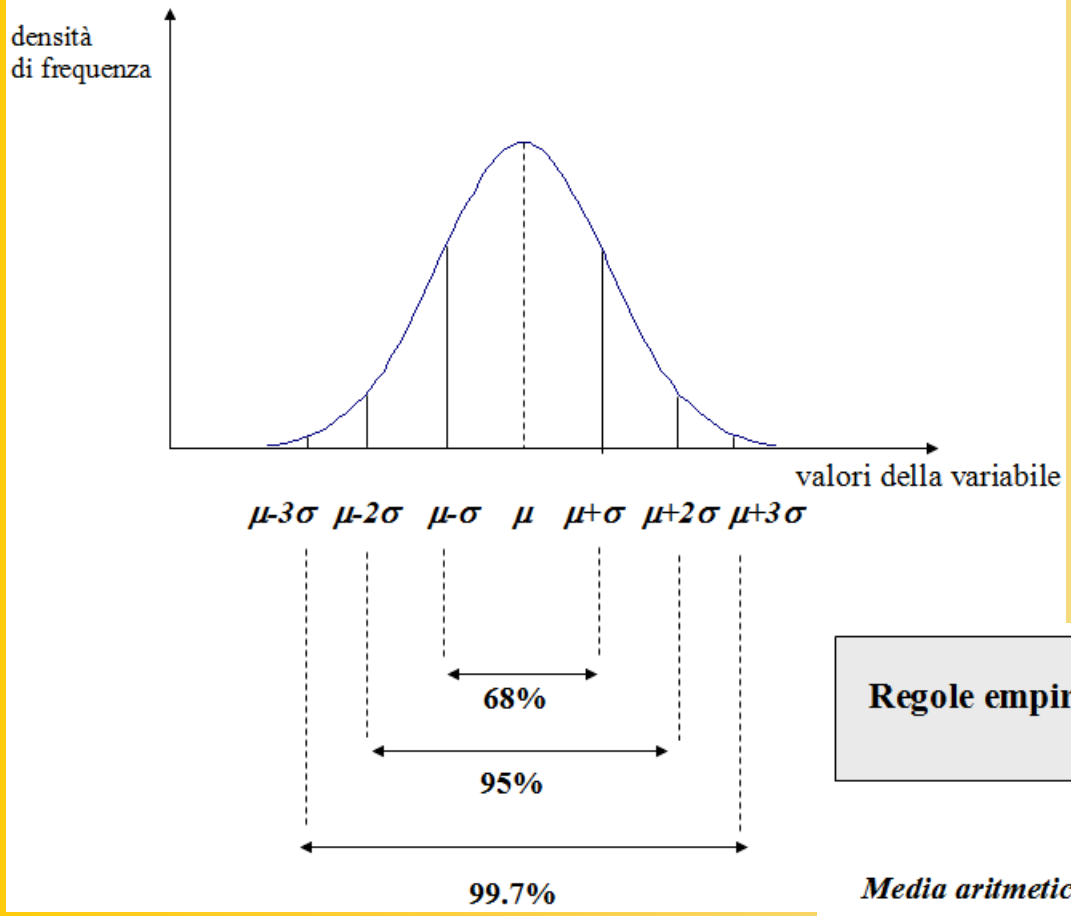
Per lo s.q.m. abbiamo

$$sqm = \sqrt{\frac{(5{,}6-6{,}6)^2 + (5{,}8-6{,}6)^2 + (6{,}1-6{,}6)^2 + (6{,}4-6{,}6)^2 + (6{,}5-6{,}6)^2 + (6{,}8-6{,}6)^2 + (6{,}9-6{,}6)^2 + (6{,}9-6{,}6)^2 + (7{,}2-6{,}6)^2 + (7{,}4-6{,}6)^2}{10}} = 0{,}6.$$

Possiamo concludere che la lunghezza dei grilli esaminati presenta un range di 1,8 s.q.m. di 0,6 cm.

http://www.istat.it/servizi/studenti/valoredati/Cap4/Cap4_5_2.htm

Distribuzione teorica normale



Regole empiriche per distribuzioni tendenzialmente normali e per grandi campioni

Media aritmetica ± deviazione standard	*~ 68% delle osservazioni*
Media aritmetica ± 2× deviazione standard	*~ 95% delle osservazioni*
Media aritmetica ± 3× deviazione standard	*~ 99.7% delle osservazioni*

Esempio:

In una popolazione di 150 pazienti sono state calcolate la media aritmetica e la deviazione standard per la variabile età distribuita in modo tendenzialmente normale:

media aritmetica = 55 anni

deviazione standard = 8 anni

Basandoci sulle regole empiriche possiamo dedurre:

circa il 68% dei soggetti ha un'età compresa tra 47 e 63 anni **[(55-8, 55+8])**

circa il 95% dei soggetti ha un'età compresa tra 39 e 71 anni **[(55-16,55+16)]**

circa il 99.7% dei soggetti ha un'età compresa tra 31 e 79 anni **[(55-24,55+24)]**

Altri indici di variabilità

- Lo **scostamento semplice medio dalla media aritmetica**:

$$S_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

- Lo **scostamento semplice medio dalla mediana**:

$$S_{M_e} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - M_e\right|$$

Le misure di variabilità per le variabili qualitative

Frequenza relativa della moda

molte osservazioni concentrate sulla moda → ridotta variabilità

dispersione delle osservazioni tra le varie modalità → maggiore variabilità

Densità relativa di frequenza della moda

Ottenuta dividendo la frequenza relativa della moda alla frequenza media delle altre modalità.

Esempio:

Grado d'istruzione	**N. pazienti**	**%**
media	7	18
superiore	23	59
universitaria	9	23
Totale	**39**	**100**

Grado d'istruzione	**N. pazienti**	**%**
media	10	28
superiore	12	33
universitaria	14	39
Totale	**36**	**100**

moda = **superiore**

frequenza % della moda = **59%**

densità di freq. della moda = 59/33 = **1.8**

moda = **universitaria**

frequenza % della moda = **39%**

densità di freq. della moda = 39/33 = **1.2**

100/3=33

Il coefficiente di variazione

- **Osservazione**: la varianza e la deviazione standard sono indici che risentono dell'unità di misura e dell'ordine di grandezza dei dati. Pertanto il **confronto della variabilità tra due distribuzioni** risulta compromesso per:

  - variabili diverse per natura (es. peso e statura);
  - variabili diverse per ordine di grandezza (es. peso degli adulti e peso dei neonati);
  - una stessa variabile espressa secondo unità di misura diverse (es. dose del farmaco espressa in milligrammi o in grammi).

Per confrontare la variabilità di due distribuzioni per la variabile con $\bar{x} > 0$ può essere utilizzato il **coefficiente di variazione**:

$$CV = \frac{\sigma}{\bar{x}} 100$$

Un esempio di confronto della variabilità

9 industrie con dispositivo anti-inquinante di tipo A e 9 di tipo B.

Tipo	Quantità di pulviscolo								
A	69	80	44	52	54	54	86	77	66
B	35	62	43	23	30	28	22	40	25

$$\bar{x}_A = 64{,}67$$
$$\bar{x}_B = 34{,}22$$

$$\sigma_A = 13{,}65$$
$$\sigma_B = 12{,}02$$

$$\Rightarrow \quad CV_A = 21\%$$
$$CV_B = 35\%$$

Si può concludere che è la distribuzione B ad essere più variabile della distribuzione A.

Esercizio

In una gara di atletica leggera sono stati rilevati i seguenti 5 migliori risultati di salto in alto (in metri) e di corsa sui 100 m (in secondi):
Salto in alto

1,85 1,92 1,95 1,94 1,94

100 m

11,7 11,3 11,4 11,2 11,6.

Indicare quale delle due serie di risultati presenta maggiore variabilità.

Soluzione:
Per prima cosa calcoliamo la media e lo s.q.m. delle due distribuzioni:

$$M_s = \frac{1,85 + 1,92 + 1,95 + 1,94 + 1,94}{5} = 1,92m$$

$$s.q.m._s = \sqrt{\frac{\sum(_s x_i - M_s)^2}{N}} = 0,036m$$

$$M_c = \frac{11,7+11,3+11,4+11,2+11,6}{5} = 11,44 \text{ secondi}$$

$$s.q.m._c = \sqrt{\frac{\sum(_c x_i - M_s)^2}{N}} = 0,185 \text{ secondi.}$$

Siccome le due serie di dati sono espresse in unità di misura diverse (metri e secondi), per confrontare la loro variabilità si ricorre al coefficiente di variazione, il cui valore è un numero puro, svincolato cioè dall'unità di misura. I coefficienti di variazione sono:

$$CV_s = \frac{s.q.m._s}{M_s} x100 = 1,89$$

$$CV_c = \frac{s.q.m._c}{M_c} x100 = 1,62.$$

I risultati del salto in alto presentano quindi maggiore variabilità.

La standardizzazione

La **standardizzazione** è una particolare trasformazione lineare che applicata ai dati originali riconduce qualsiasi variabile X con media $\overline{x}$ e deviazione standard σ a una nuova variabile con **media nulla** e **varianza unitaria**.

Ogni osservazione x_i viene trasformata in un nuovo valore:

$$y_i = \frac{x_i - \overline{x}}{\sigma}$$

La distribuzione risultante ha media nulla e varianza unitaria.