

Concetti base della statistica e analisi descrittiva

1

Rosolen Valentina

Dirigente Statistico

SC Epidemiologia clinica e Ricerca sui servizi sanitari - IRCCS materno infantile
Burlo Garofolo

Fasi del processo di ricerca

Osservazione della realtà



Formulazione di ipotesi



Predisposizione di un piano
d'indagine



Esecuzione del piano d'indagine



Verifica dell'ipotesi

Che cos'è la statistica?

- La Statistica fornisce regole, strumenti e tecniche che consentono la raccolta e l'analisi quantitativa delle caratteristiche dei fenomeni collettivi



- La statistica può essere d'aiuto per descrivere dati, progettare esperimenti e verificare ipotesi
- È utile per rispondere a domande quali "Se le mie idee vengono contestate, sarò in grado di difenderle"

Statistica

In base alle sue funzioni, la statistica si può distinguere in:

- **statistica descrittiva**: descrive, rappresenta e sintetizza attraverso strumenti e tecniche opportune una popolazione o un campione di dati relativamente al fenomeno in studio
- **statistica inferenziale**: il cui obiettivo è quello di ricavare dallo studio condotto, leggi generali che si intendono estendere alla popolazione e/o ad altre situazioni simili

Come sono fatti i dati?

Che tipo di dati si possono rilevare su un soggetto/oggetto?

- ➡ presenza/assenza di un attributo
(fumatore? si/no)
- ➡ grado di una caratteristica (livello d'istruzione? Nessuno/elementare/medio/superiore/universitario)
- ➡ intensità di una caratteristica
(temperatura? 15°C/30°C...)

Definizioni

6

- **Variabile (X):** caratteristica o attributo osservabile sulle unità statistiche
- **Modalità (x):** valori che la variabile può assumere
- **Osservazione (x_i):** il valore assunto dalla variabile in una data unità statistica lo chiameremo osservazione

Tipi di variabili

Qualitative

Categoriche

Classificazione: genere, colore degli occhi

Ordinali

Ordinamento: titolo di studio, gravità del trauma

Quantitative

Discrete

Conteggio: solo valori interi

Continue

Misurazione: valori reali infiniti

Matrice di dati

I dati codificati di una rilevazione statistica effettuata su **N** unità statistiche con riferimento a **p** variabili, vengono raccolti in una tabella che viene chiamata “**matrice dei dati**”

Ogni riga rappresenta l'unità statistica

Ogni colonna rappresenta una variabile

Ogni cella rappresenta un'osservazione

Identificativo	Genere	Titolo di studio	Età	N° Ricoveri
1	F	Laurea	30	1
2	M	Diploma	55	4
...	...		...	...
N	F	Licenza media	48	2

Attenzione a.....

Imputazione dei dati:

- Scelta di un supporto informatico adeguato che possa prevedere la creazione di maschere di inserimento del dato in modo da ridurre eventuali errori di input (campi di variazioni di una variabile, relazioni (filtri) tra le variabili).
- È buona norma fare una pre-codifica delle domande aperte (ICD9, ATC) e creare delle istruzioni per l'inserimento dei dati
- Training all'operatore che si occuperà dell'inserimento

Controllo dei dati:

- I dati imputati (sul supporto informatico) sono verosimili, sono coerenti (range)? Se no, controllo sul supporto cartaceo (necessità di risalire al cartaceo id).
- Verificare se i filtri son stati applicati correttamente

Analisi dei dati

- La matrice dei dati contiene tutte le informazioni analitiche di ciascuna unità statistica
- Quando i dati sono molti, l'analisi diretta della matrice non consente di cogliere in via immediata gli aspetti salienti del fenomeno
- Occorre perciò ottenere una sintesi attraverso un'elaborazione statistica dei dati

Distribuzione di frequenza

- **FREQUENZA ASSOLUTA:** è il numero/conteggio di unità che cadano in una determinata classe (intervallo). La frequenza assoluta somma ad n
- **FREQUENZA RELATIVA:** è la frazione (o la percentuale) di unità (rispetto al numero totale di unità) che cadono in una determinata classe (intervallo). La frequenza relativa somma ad 1 (o 100%)
- **FREQUENZA CUMULATA:** corrisponde al numero di unità che sono inferiori ad una data soglia (appartengono all'intervallo $[-\infty, \text{soglia}[$). **La frequenza cumulata può essere sia assoluta sia relativa**

DISTRIBUZIONE DI FREQUENZA				
Titolo di studio	Frequenze assolute	Frequenze relative*	Frequenze assolute cumulate	Frequenze relative cumulate
x_i	f_i	p_i	F_i	P_i
Nessuno	1	0,02	1	0,02
Licenza elementare	4	0,08	5	0,10
Licenza media	10	0,20	15	0,30
Diploma superiore	25	0,50	40	0,80
Laurea	10	0,20	50	1
Totale	50	1	50	1

Frequenza relativa percentuale: $p_i \cdot 100$, il totale corrisponde a 100 anziché ad 1

Dati grezzi: peso alla nascita (grammi)

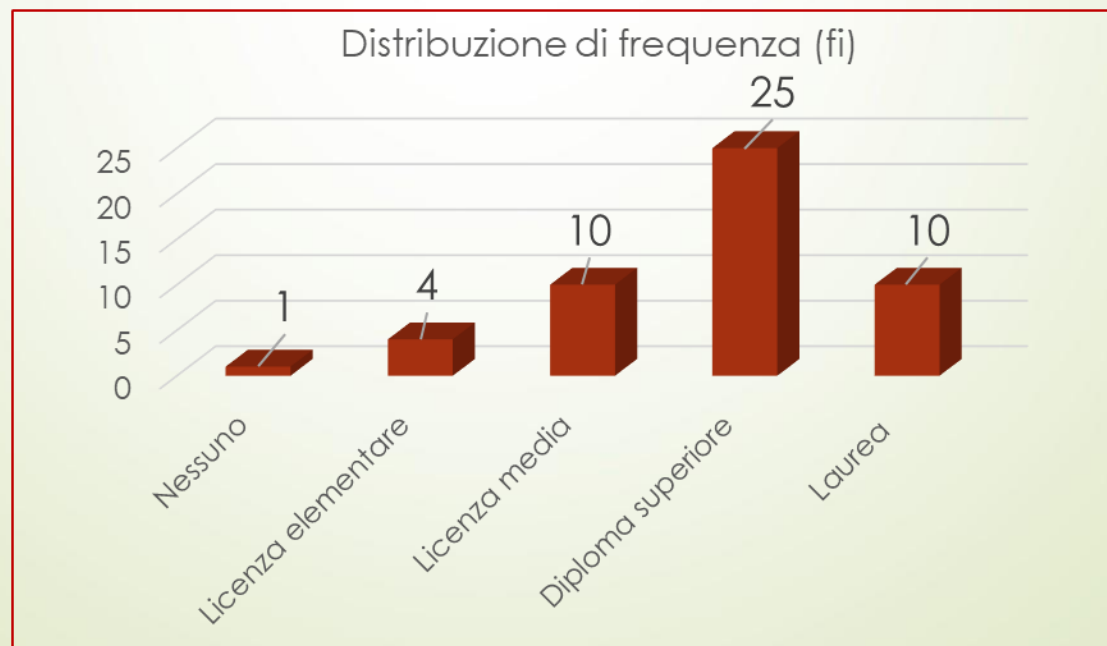
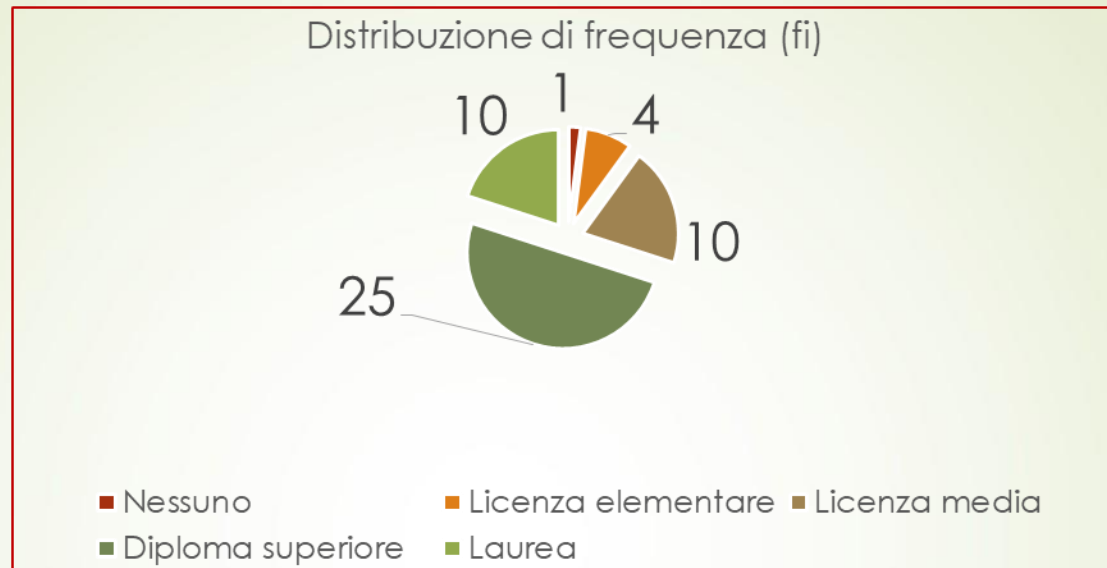
3200	2950	3060	3350	2750	3000	2880
2720	3090	3100	2700	3300	2650	3310
2620	3160	3130	2820	3200	2900	3260

**distribuzione
di frequenza**

peso alla nascita (g)	frequenza
2600-2799	5
2800-2999	4
3000-3199	6
3200-3399	6
totale	21

Rappresentazione grafica

14



Indici statistici

Costruire distribuzioni di frequenze ci permette di fare una prima sintesi dei dati e ci consente di comprendere come il carattere in esame si manifesti nel collettivo statistico oggetto della rilevazione

INDICI
STATISTICI

```
graph LR; A[INDICI STATISTICI] --> B[Per sintetizzare una certa caratteristica]; A --> C[Per confrontare situazioni differenti];
```

Per **sintetizzare** una
certa caratteristica

Per **confrontare**
situazioni differenti

Importante

- ➡ La conoscenza di quali siano le caratteristiche delle variabili consente di scegliere gli strumenti corretti per analizzarle!

Indici statistici

- **INDICI DI POSIZIONE (o di Tendenza Centrale):** esprimono sinteticamente il **centro ideale** della distribuzione, ovvero il valore intorno al quale tendono a gravitare i dati
- **INDICI DI VARIABILITA' (o di Dispersione):** consentono di valutare il **grado di diversità** delle modalità del carattere, ovvero forniscono informazioni sul **grado di dispersione** dei dati intorno al loro centro ideale.

Indici di posizione

Le misure di posizione più utilizzate:

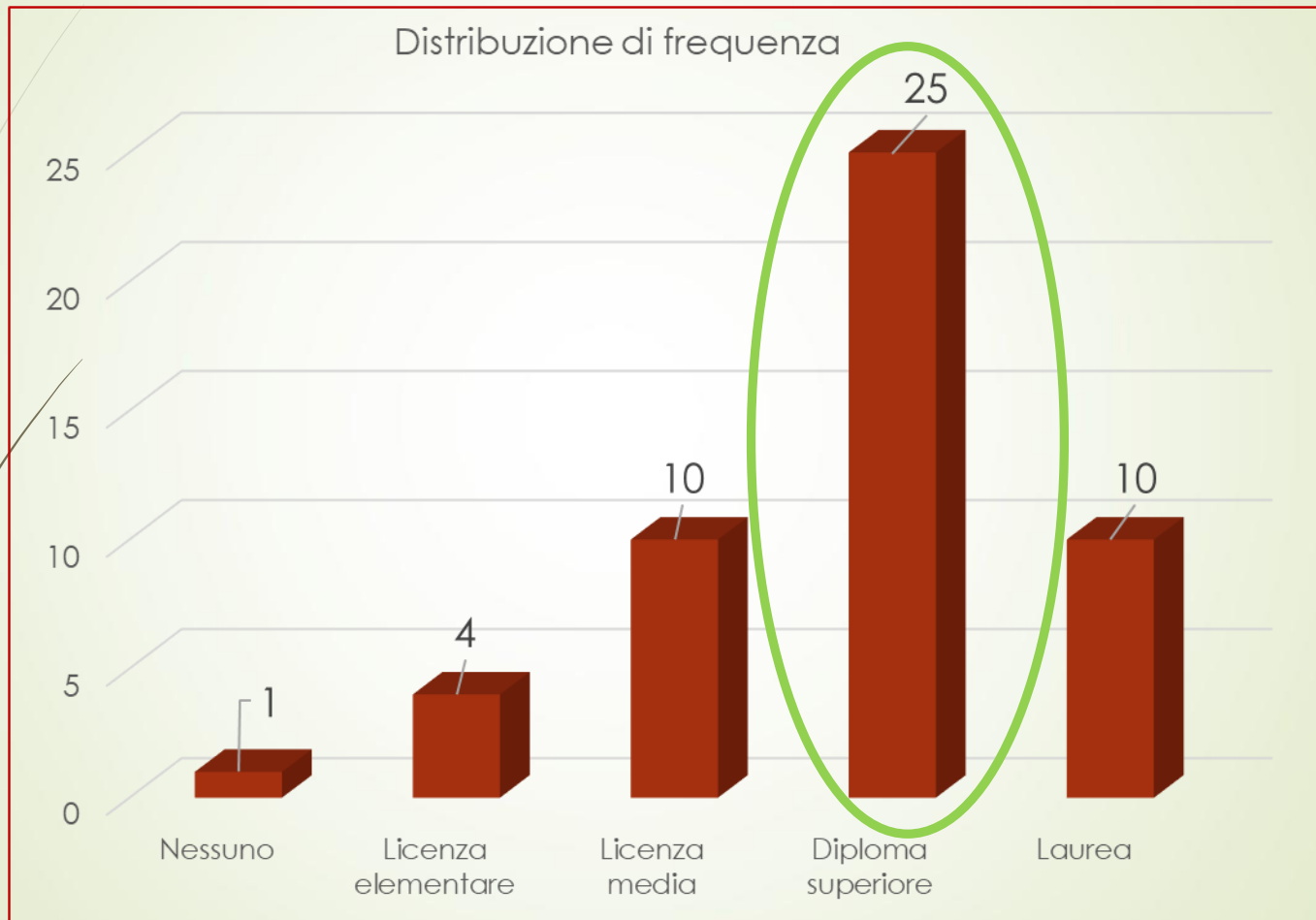
- moda
- mediana
- quantili
- media

Moda

La MODA corrisponde alla modalità con la frequenza assoluta (o relativa) più alta.

- E' un indice di posizione facilmente calcolabile
- Può essere determinato per qualsiasi tipo di variabile
- Non è influenzata dalla presenza di nessun valore estremo

Titolo di studio	Frequenze assolute	Frequenze relative
Nessuno	1	0,02
Licenza elementare	4	0,08
Licenza media	10	0,20
Diploma superiore	25	0,50
Laurea	10	0,20



Mediana

- La mediana è il valore che occupa la posizione centrale in un insieme ordinato di dati
- E' una misura robusta, in quanto poco influenzata dalla presenza di dati anomali, si ricorre quindi all'uso della mediana quando si vuole attenuare l'effetto di valori estremi

Per calcolare la mediana di un gruppo di dati, bisogna:

1. disporre i valori in ordine crescente oppure decrescente e contare il numero totale n di dati;
2. **se il numero (n) di dati è dispari**, la mediana corrisponde al valore numerico del dato centrale, quello che occupa la posizione $(n+1)/2$;
3. **se il numero (n) di dati è pari**, la mediana è stimata utilizzando i due valori centrali che occupano le posizioni $n/2$ e $n/2+1$

Consideriamo il seguente campione:

96	78	90	62	73	89	92	84	76	86
----	----	----	----	----	----	----	----	----	----

1. Ordiniamo le misurazioni in ordine crescente:

62	73	76	78	84	86	89	92	92	95
----	----	----	----	-----------	-----------	----	----	----	----

2. Dal momento che il numero di misurazioni è pari ($n=10$) la mediana è calcolata come la media dei due elementi centrali:

$$\text{Mediana} = \frac{84+86}{2} = 85$$

Quantili

- Estendendo il discorso fatto sulla mediana, possiamo immaginare di suddividere la distribuzione in p parti, ognuna delle quali contenente lo stesso numero di unità
- Il quantile p -esimo è la modalità che occupa nella serie ordinata dei dati quella posizione che lascia al di sotto la percentuale p dei dati e al di sopra la percentuale $1-p$

Quantili

Se si divide la distribuzione ordinata dei dati in 4 parti di uguali numerosità



$p=4$



Quartili (1° , 2° , 3°)

....

Se si divide la distribuzione ordinata dei dati in 100 parti di uguali numerosità



$p=100$



Percentili (1° , 2° , 3° , ..., 99°)

....

....

Media

- La **media aritmetica semplice** è la misura di posizione (o di tendenza centrale) più comunemente utilizzata per variabili quantitative e mostra il valore centrale dei dati (non necessariamente una modalità osservata)

È definita come la somma del valore di tutte le osservazioni diviso il numero di unità:

$$\mu = \frac{x_1 + x_2 + \dots + x_n}{n} \qquad \sum_{i=1}^n = \frac{x_i}{n}$$

- Ogni variabile quantitativa discreta o continua ha una media
- Un insieme di dati ha una sola media
- Nel calcolare la media vengono inclusi tutti i valori
- La media non è una statistica robusta, ossia risente dei valori molto alti o molto bassi

Riassumendo

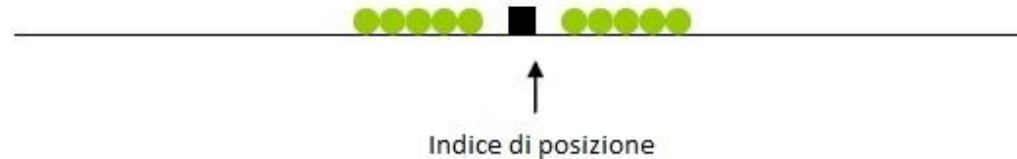
Tipo di variabile	MODA	MEDIANA	MEDIA
Qualitativa categorica	SI	NO	NO
Qualitativa ordinale	SI	SI	NO
Quantitativa	SI	SI	SI

La variabilità

Popolazione A



Popolazione B



Subito è evidente che l'indice di posizione è lo stesso ma la sintesi ottenuta è migliore nella popolazione B, perché le osservazioni sono maggiormente addensate intorno alla modalità centrale.

Indici di variabilità

Le misure di variabilità si distinguono in:

1) INDICI BASATI SULL'ORDINAMENTO DEI DATI

- Campo di Variazione (Range)
- Scarto Interquartile

2) INDICI DI DISPERSIONE

- Devianza
- Varianza
- Deviazione standard (o scarto quadratico medio)

N.B. gli indici di variabilità possono essere calcolati soltanto per caratteri quantitativi non per i qualitativi (nemmeno ordinabili)

Campo di variazione (range)

- Il campo di variazione (o Range) è dato dalla differenza tra il valore massimo e il valore minimo osservato:

$$\text{Range} = x_{\max} - x_{\min}$$

- Rappresenta l'ampiezza dell'intervallo dei dati
- È espresso nella stessa unità di misura dei dati
- Si basa soltanto sui valori estremi della distribuzione e non tiene conto dei valori intermedi
- È molto influenzato da osservazioni anomale (outliers)
- L'indice non è robusto

	1 ^a Prova	2 ^a Prova	3 ^a Prova
1° studente	3	2	6
2° studente	5	7	7
3° studente	8	8	6
4° studente	9	8	6
range	6	6	1

Range 1^a prova = 6 $\Rightarrow$ dati più dispersi,
risultati più eterogenei

Range 3^a prova = 1 $\Rightarrow$ dati più concentrati,
risultati più omogenei

Range 1^a prova = Range 2^a prova = 6 Stessa Distribuzione?

Scarto interquartile

- Lo Scarto (o distanza) interquartile è dato dalla differenza tra il 3° quartile Q_3 e il 1° quartile Q_1 :

$$S_{IQ} = Q_3 - Q_1$$

- In questo intervallo ricade la metà dei valori osservati, posti esattamente al centro della distribuzione (intorno alla mediana),
- non è influenzato da osservazioni anomale o estreme.

STATURA	f_i	p_i	P_i
162	1	0,03	0,03
168	1	0,03	0,06
169	1	0,03	0,09
170	3	0,09	0,18
172	2	0,06	0,24
174	2	0,06	0,29
175	5	0,15	0,44
176	3	0,09	0,53
177	3	0,09	0,62
178	3	0,09	0,71
179	1	0,03	0,74
181	1	0,03	0,76
182	2	0,06	0,82
183	2	0,06	0,88
184	1	0,03	0,91
188	1	0,03	0,94
192	1	0,03	0,97
193	1	0,03	1,00
Somma	34	1	

ESEMPIO:

$$\text{Range} = 193 - 162 = 31 \text{ cm}$$

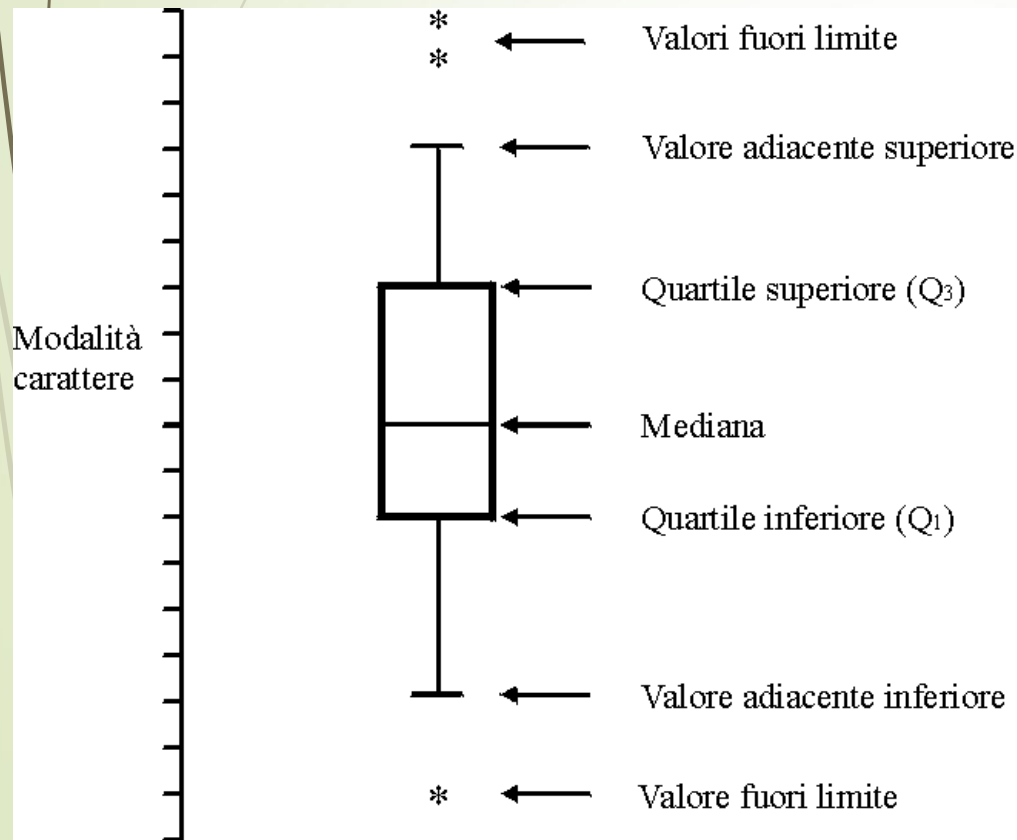
$$Q_1 = x_{0,25} = 174 \text{ cm}$$

$$Q_3 = x_{0,75} = 181 \text{ cm}$$

$$S_{IQ} = 181 - 174 = 7 \text{ cm}$$

Rappresentazione grafica

Il box plot (o grafico a scatola o a baffi) è un grafico che permette:



➤ di descrivere in modo stilizzato la distribuzione dei dati

➤ di evidenziare eventuali valori estremi

Misure di dispersione

- Tutti gli indici che presenteremo, si basano sul concetto di scarto rispetto la media.

$$(x_i - \mu)$$

- Gli scarti sono positivi per tutte le modalità più grandi della media e negativi per le modalità più piccole della media, inoltre, per la proprietà della media aritmetica, se si sommano gli scarti, questi si annullano.

Devianza

Lo Devianza è un indice di dispersione che misura quanto ciascun dato $\mathbf{x}_i$ si discosta dal valor medio μ :

$$\sum_{i=1}^N (\mathbf{x}_i - \mu)^2$$

- ▶ Questo indice presenta l'inconveniente di aumentare con l'aumentare della numerosità e ciò lo rende un valore non facilmente interpretabile
- ▶ È la base, però, delle misure di dispersione per variabili quantitative, infatti da questo indice discendono la Varianza e la Deviazione standard

Varianza

- La varianza è la media aritmetica degli scarti al quadrato delle modalità dalla loro media aritmetica

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

- È una devianza media ossia la devianza rapportata al numero delle osservazioni della popolazione (N)
- Tiene conto di tutte le osservazioni ed è dunque influenzata da eventuali osservazioni anomale
- Non è confrontabile con la media o con altri indici di posizione in quanto l'unità di misura è elevata al quadrato
- Assume solo valori positivi

	1 ^a Prova	scarti da M	scarti ²
1° studente	3	-3.25	10.5625
2° studente	5	-1.25	1.5625
3° studente	8	1.75	3.0625
4° studente	9	2.75	7.5625
media	6.25	0.00	5.6875

$$\sigma^2 = \frac{10,5625+1,5625+3,0625+7,5625}{4} = 5,6875$$

	1 ^a Prova	2 ^a Prova	3 ^a Prova
1° studente	3	2	6
2° studente	5	7	7
3° studente	8	8	6
4° studente	9	8	6
media	6.25	6.25	6.25
varianza	5.69	6.19	0.19

Deviazione standard

- Per trovare una misura espressa nella stessa unità di misura della variabile di partenza è sufficiente estrarre la radice quadrata della varianza:

$$\sigma = \sqrt{\Sigma(x_i - \mu)^2 / N}$$

- La deviazione standard (o scarto quadratico medio) è una misura di distanza dalla media e quindi ha sempre un valore positivo
- E' una misura della dispersione della variabile intorno alla media

	1 ^a Prova	2 ^a Prova	3 ^a Prova
1° studente	3	2	6
2° studente	5	7	7
3° studente	8	8	6
4° studente	9	8	6
media	6.25	6.25	6.25
varianza	5.69	6.19	0.19
deviazione standard	2.38	2.49	0.43

1^a prova $\Rightarrow$ **$\mu=6,25$ dev.std= $\pm 2,38$**

2^a prova $\Rightarrow$ **$\mu=6,25$ dev.std= $\pm 2,49$**

3^a prova $\Rightarrow$ **$\mu=6,25$ dev.std= $\pm 0,43$**