CENNI DI STATISTICA DESCRITTIVA UNIVARIATA

Definizione. La statistica è lo studio quantitativo dei fenomeni collettivi.

- quantitativo poiché tale studio viene realizzato effettuando misure e rilevazioni numeriche
- fenomeni collettivi in quanto la statistica studia fenomeni che riguardano una pluralità di individui. E' interessante osservare che la statistica non fornisce informazioni e risultati sul singolo individuo o elemento, bensì una *sintesi* sulla collettività studiata.

- La statistica non è una scienza esatta
- Ci sono molte strade per risolvere un problema di statistica
- L'analisi degli stessi dati può portare a conclusioni diverse, se le ipotesi da cui si parte sono diverse
- Le conclusioni statistiche vanno sempre lette e interpretate con cautela e senso critico

Terminologia

Popolazione. L'insieme degli individui oggetto di un'indagine statistica si chiama popolazione e universo statistico o collettivo statistico.

Unità statistica. I singoli elementi che formano una popolazione sono detti unità statistiche.

Campione. Alcune indagini statistiche vengono svolte interpellando l'intera totalità di individui della popolazione, a volte però l'indagine si concentra solo su una parte della popolazione detta campione.

Numerosità. Il numero si unità statistiche si dice numerosità della popolazione.

Carattere. Si chiama carattere la proprietà o caratteristica oggetto di studio dell'unità statistica.

Modalità. Si chiama modalità ciascuna delle varianti con cui un carattere può presentarsi. Le modalità osservate sono dette *dati*.

Caratteri quantitativi e qualitativi. Un carattere le cui modalità sono espresse da numeri è un carattere quantitativo ed è detto *variabile*; in tal caso la modalità verrà anche detta *valore*. Un carattere le cui modalità non possono essere espresse da numeri è qualitativo ed è detto *mutabile*.

Variabili continue o discrete. Una variabile si dice discreta se può assumere solo un numero finito di valori (o un insieme di valori che può essere posto in corrispondenza biunivoca con l'insieme N dei numeri naturali che è un infinito numerabile). Una variabile si dice continua se può assumere tutti i possibili valori reali di un determinato intervallo.

Esempio.

Fenomeno studiato	Popolazione	Carattere	Modalità	Tipo di carattere
Il colore degli occhi degli italiani	Tutti gli italiani	Il colore degli occhi	Verdi, neri, castani, azzurri, …	Qualitativo
I bimbi e la televisione	I bambini italiani da 0 a 10 anni	➢ Ore trascorse davanti alla televisione ➢ Programmi seguiti ➢ Etc…	1 ore, 2 ore, 3 ore …. Cartoni, documentari, … …	Quantitativo discreto Qualitativo …
Altezza degli studenti di una classe	Tutti gli studenti della classe	La misura dell'altezza	1,72 m; 1,68m; 1,80m; …	Quantitativo continuo
Le caratteristiche degli abitanti di Busto Arsizio	Tutti i residenti a Busto Arsizio	Titolo di studio Sesso Numero di figli Anno di nascita …	Nessuno, licenza media, diploma, laurea, … Maschio, femmina 1,2,3,4,… 1970, 1956, … …	Qualitativo Qualitativo Quantitativo discreto Quantitativo discreto …

Tipi di statistica

Per studiare un fenomeno statistico si può decidere di interpellare l'intera popolazione, in tal caso si parla di **censimento**, oppure di analizzare solo una parte della popolazione, un campione, e di estendere poi i risultati ottenuti all'intera popolazione.
Nel primo caso si parla di **statistica descrittiva**, mentre nel secondo caso di **statistica inferenziale.**
Statistica descrittiva univariata: prende in esame un solo carattere dell'intera popolazione
Statistica descrittiva multivariata: prende in esame più caratteri dell'intera popolazione.
In particolare si parla di **statistica descrittiva bivariata** quando i caratteri sono solo due.

Il metodo statistico

- ***Individuazione del fenomeno collettivo da studiare e degli obiettivi***
- ***Individuazione della popolazione e delle unità statistiche***
- ***Scelta dei caratteri da analizzare e delle modalità con cui si pensa di rilevarli***
- ***Spoglio:*** classificazione in tabelle dei dati rilevati per renderli più leggibili e facilmente utilizzabili
- ***Elaborazione e rappresentazione***: è la fase che utilizza le tecniche matematiche che consentono di trasformare e rappresentare graficamente i dati (indici di posizione e variabilità, grafici …)
- ***Interpretazione dei risultati***

Dove cercare le informazioni?

- http://www.istat.it/
- http://www.censis.it
- http://www.doxa.it
- http://www.ilsole24ore.com

Dalla tabella dei dati grezzi alla tabella di frequenza

Mediante lo spoglio si rilevano i dati acquisiti durante un'indagine statistica e li si strutturano in tabelle.
Il risultato della rilevazione dei caratteri statistici, che descrivono il fenomeno oggetto di studio, sulle unità statistiche della popolazione, o di un campione, produce la matrice dei dati. La matrice dei dati è un prospetto che contiene tutte le informazioni rilevate sulle unità statistiche (i dati grezzi) e rappresenta il punto di partenza per le analisi successive.
La matrice dei dati è una matrice nxk, dove n è il numero di unità statistiche (dimensione della popolazione o l'ampiezza del campione) e k è il numero di variabili statistiche osservate sulle unità statistiche (matrice individui x variabili).

Esempio.

Indagine sui trasporti degli studenti del primo anno di STB di Busto Arsizio	
A piedi	X X X
In bicicletta	X X
Motorino	X X X X
Auto come autista	X X X X X X
Auto da passeggero	X X X X X X X X X
Pullman urbano	X X X
Treno	X X
Bus extraurbano	X

Popolazione formata da 19 studenti

Tali tabelle sono però ancora matrici di dati grezzi. Tali dati vanno organizzati in modo da essere sintetici e fruibili.

Al fine di interpretare i risultati di un'indagine statistica è necessario elaborare adeguatamente i dati grezzi contenuti nella matrice dei dati.

Un primo modo è quello di costruire delle tabelle di frequenza. Le tabelle di frequenza costituiscono una prima sintesi delle informazioni presenti nella matrice dei dati.

Indagine sui trasporti	
A piedi	3
In bicicletta	2
Motorino	4
Auto come autista	6
Auto da passeggero	9
Pullman urbano	3
Treno	2
Bus extraurbano	1

E' interessante notare come un carattere qualitativo (mezzo di trasporto) possa essere analizzato mediante dati numerici quantitativi , contando il numero di volte in cui ciascuna modalità si presenta. Inoltre la somma delle frequenze supera il numero di studenti, in quanto uno stesso studente può usare diversi mezzi di trasporto.

Si introduce il concetto di frequenza

Frequenza assoluta di una modalità statistica: numero di volte in cui la modalità è stata registrata, ossia il numero di unità statistiche che presentano la stessa modalità.

Frequenza relativa di una modalità: il rapporto tra la frequenza assoluta e il numero totale di rilevazioni, ossia la proporzione di unità statistiche che presentano la stessa modalità.

$$f_r = \frac{f_a}{n}$$

Frequenza percentuale di una modalità: è la rappresentazione percentuale della frequenza relativa; si ottiene moltiplicando la frequenza relativa per cento

$$f_r\% = \frac{f_a}{n} \cdot 100$$

Frequenza cumulata di una modalità di carattere quantitativo: la somma delle frequenze di tutte le modalità minori o uguali a quella considerata.

Esempio. In un gruppo di N=30 persone abbiamo 10 italiani, 15 inglesi, 3 tedeschi, 2 russi

$f_1 = 10 \quad f_2 = 15 \quad f_3 = 3 \quad f_4 = 2$

$f_{r1} = \frac{10}{30} = \frac{1}{3} = 0{,}33... \qquad f_{r2} = \frac{15}{30} = \frac{1}{2} = 0{,}5 \qquad f_{r3} = \frac{3}{30} = 0{,}1 \qquad f_{r4} = \frac{2}{30} = \frac{1}{15} = 0{,}06666...$

$f_{r1}\% = 33{,}3\% \qquad f_{r2}\% = 50\% \qquad f_{r3}\% = 10\% \qquad f_{r4}\% = 6{,}7\%$

Oss. $\sum_{i=1}^{n} f_r = N$ con N = numerosità della popolazione

$\sum_{i=1}^{n} f_{ri} = 1$ infatti $0 \le f_{ri} \le 1$ (condizione di normalizzazione)

Con Excel: (importanza del riferimento assoluto in Excel)

Popolazione	Frequenze assolute	Frequenze relative	Frequenze relative percentuali
Italiani	10	0,333333333	33,33333333
Inglesi	15	0,5	50
Tedeschi	3	0,1	10
Russi	2	0,066666667	6,666666667
	30	1	100

Esempio. Numero di incidenti sul lavoro per giorno della settimana

Giorni Lavorativi	Freq. Ass.	Freq. Rel.	Freq. %	Freq. Cum.
Lunedì	52	0,208	20,80%	20,80%
Martedì	48	0,192	19,20%	40,00%
Mercoledì	45	0,180	18,00%	58,00%
Giovedì	52	0,208	20,80%	78,80%
Venerdì	53	0,212	21,20%	100,00%
Totale	250	1,000	100,00%	

Si chiama distribuzione di frequenze semplice l'insieme delle coppie (modalità, frequenza), dove la frequenza può essere assoluta, relativa o percentuale;
Oss. La distribuzione di frequenza si dice semplice se è riferita ad un solo carattere; si dice doppia se è riferita a due caratteri congiuntamente considerati; si dice multipla se si riferisce a più di due caratteri.

Nota. La costruzione di tabelle di frequenza è utile sia nella fase iniziale che nella fase finale dell'analisi dei dati.
- Nella fase iniziale perché attraverso le tabelle di frequenza è possibile controllare la coerenza e la completezza dei dati osservati.
- Nella fase finale perché le tabelle di frequenza, accompagnate eventualmente da opportune rappresentazioni grafiche, permettono di rappresentare in modo efficace i risultati delle analisi.

Suddivisione delle modalità in classi (o intervalli)

In alcuni casi, soprattutto se i dati sono molti, può essere conveniente, al fine della costruzione della tabella di frequenza, determinare delle classi di modalità contigue a cui assegnare le unità statistiche. Le frequenze quindi si riferiscono alle classi e non alle singole modalità.

Le classi devono:
a) essere in numero abbastanza limitato per poter fornire una adeguata sintesi della distribuzione;
b) essere tra loro disgiunte;
c) comprendere tutte le possibili modalità della variabile;
d) avere, se possibile, tutte la stessa ampiezza.

Si definisce classe (o intervallo) l'insieme degli elementi compresi fra due valori, detti limiti (inferiore e superiore) della classe.

Qual è il numero ottimale di classi in cui suddividere le misure da sintetizzare?

- In genere si suole utilizzare classi di ampiezza costante
- Il numero k delle classi dipende dal valore N che indica la numerosità della popolazione

Per determinare k non esiste una formula generale universalmente valida. Vanno analizzati i singoli casi. Ciò nonostante esistono delle utili regole pratiche:

- Prima regola pratica: $k \cong \sqrt{N}$
- Regola di Sturges: K=1+3.3*Log(N) (utile per valori di N molto grandi)

Divisione in classi di un carattere continuo

1. Si individuano il minimo e il massimo dei nostri dati
2. Si sceglie un adeguato numero di classi
3. Si calcola il range= massimo-minimo
4. Si determina l'ampiezza della classe a=Range/k.
5. La prima classe avrà estremi [minimo;minimo+a)

Voto in Matematica	Alunni
3	1
4	2
5	5
6	13
7	4
8	2
Totale	27

Variabili discrete (con un numero finito di modalità)

Statura (in cm)	Freq.
[120-130)	3
[130-140)	14
[140-150)	138
[150-160)	2334
[160-170)	5659
[170-180)	1739
[180-190)	111
[190-200]	2
Totale	10000

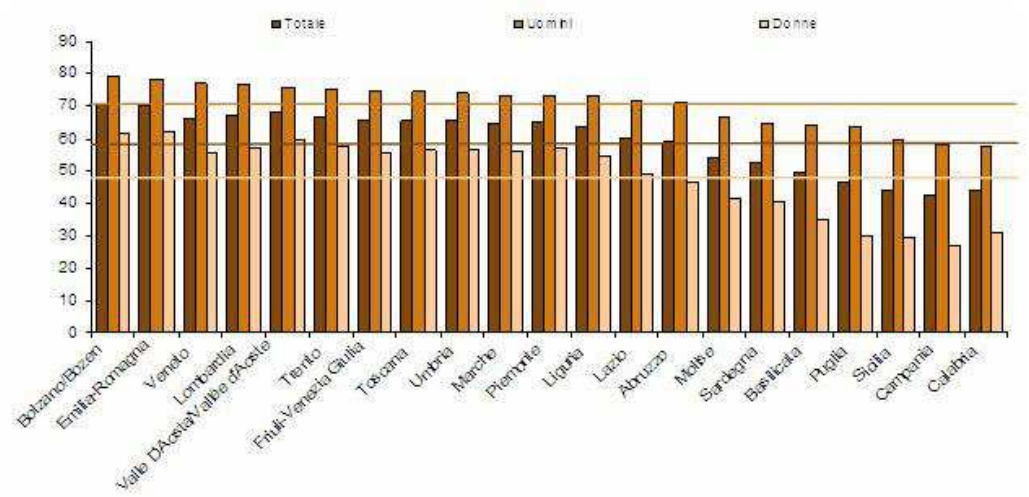
Variabili continue (con raggruppamento in classi)

Rappresentazioni grafiche

- **Per caratteri qualitativi**

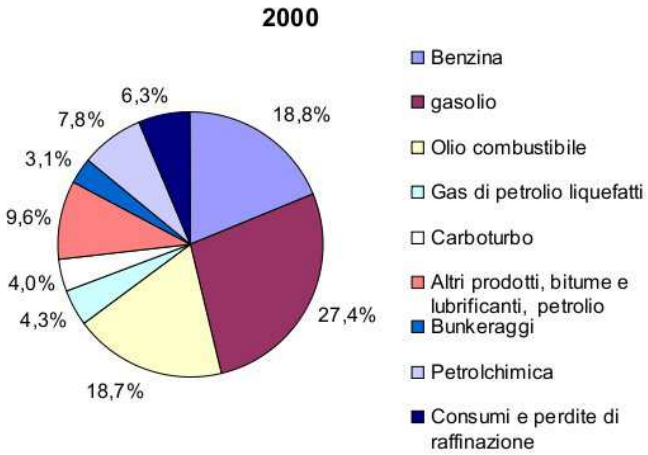
  - **Grafici a nastri**

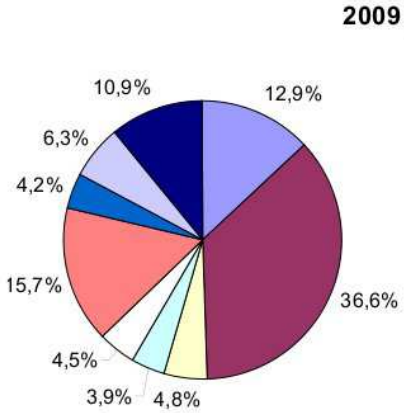
Altezza proporzionale alla frequenza; permettono di rappresentare contemporaneamente più fenomeni o il modo in cui un carattere si ripartisce in collettivi diversi (es. uomini, donne, etc…)



Tasso di occupazione della popolazione in età 15-64 anni per sesso nei paesi Ue - Anno 2008 (valori percentuali). Fonte: Istat 2010

  - **Grafici a torta**





Consumi dei principali prodotti petroliferi - Anni 2000, 2009 (a) - Fonte: Istat 2010

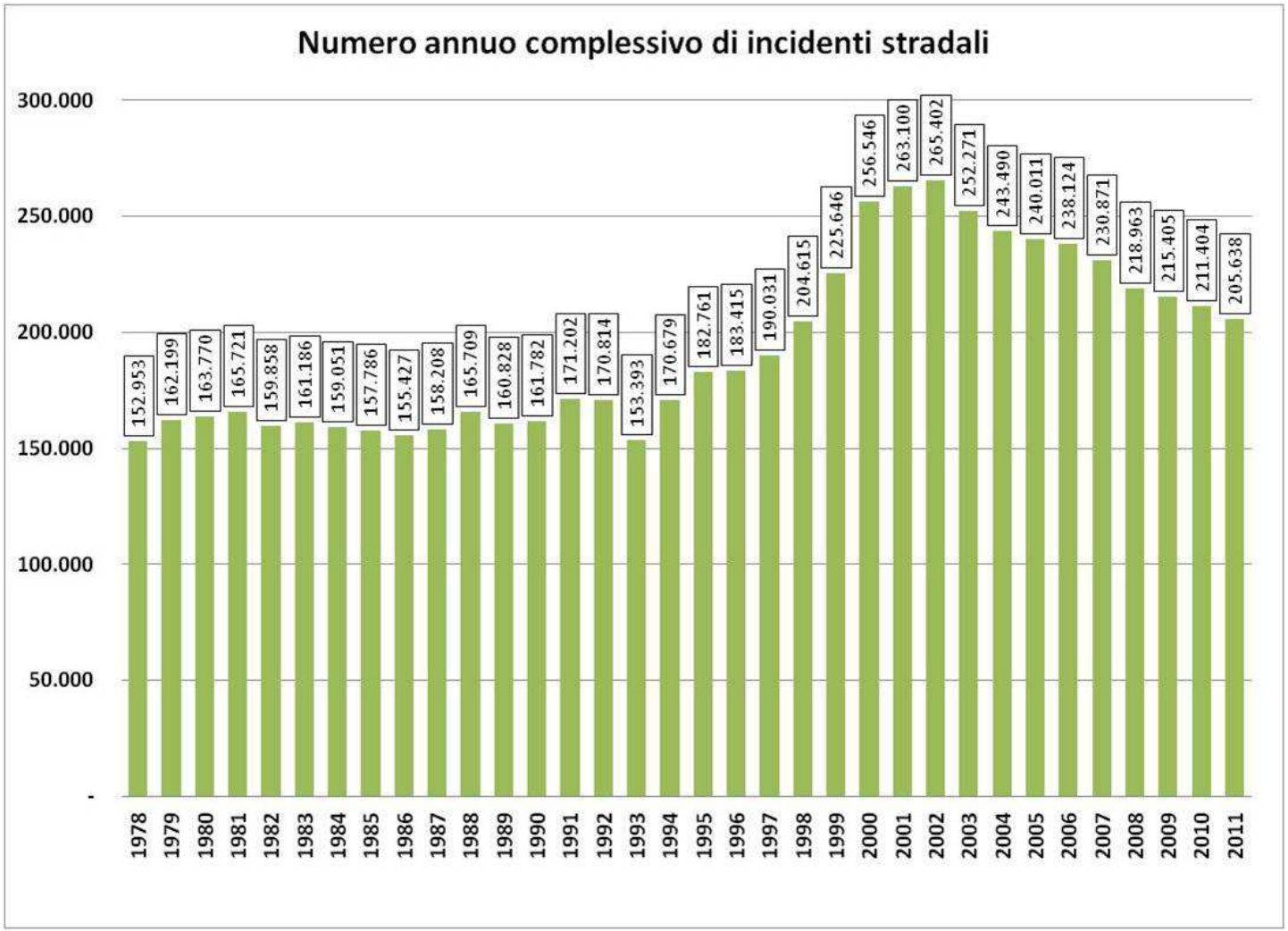
In un diagramma a torta l'area di un settore circolare è direttamente proporzionale alla frequenza

Per ottenere aree direttamente proporzionali alla frequenza si ricorda che in un cerchio, l'area di un settore circolare è direttamente proporzionale all'ampiezza, pertanto basterà ottenere ampiezze proporzionali alle frequenze mediante la seguente proporzione:

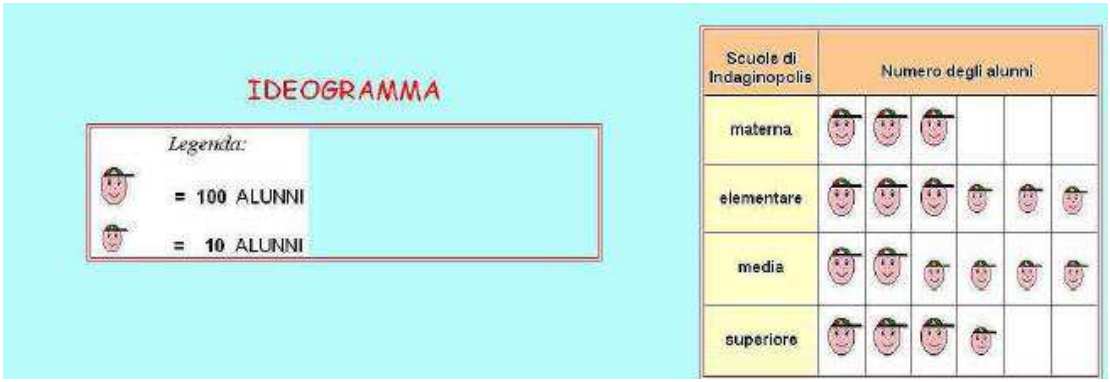
$$x : f_a = 360° : n \Rightarrow x = \frac{f_a}{n} \cdot 360° = f_r \cdot 360°$$

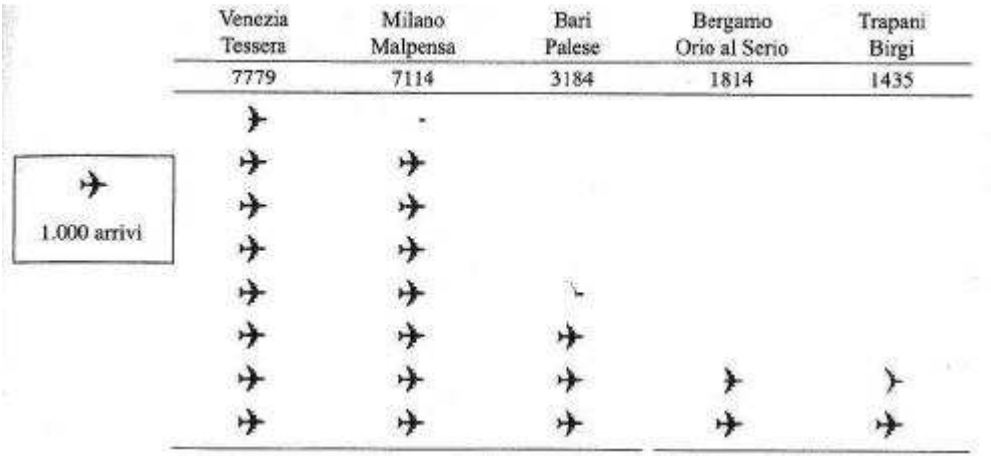
- **Per caratteri quantitativi discreti**
  - **Grafici a barre**

Altezza proporzionale alla frequenza

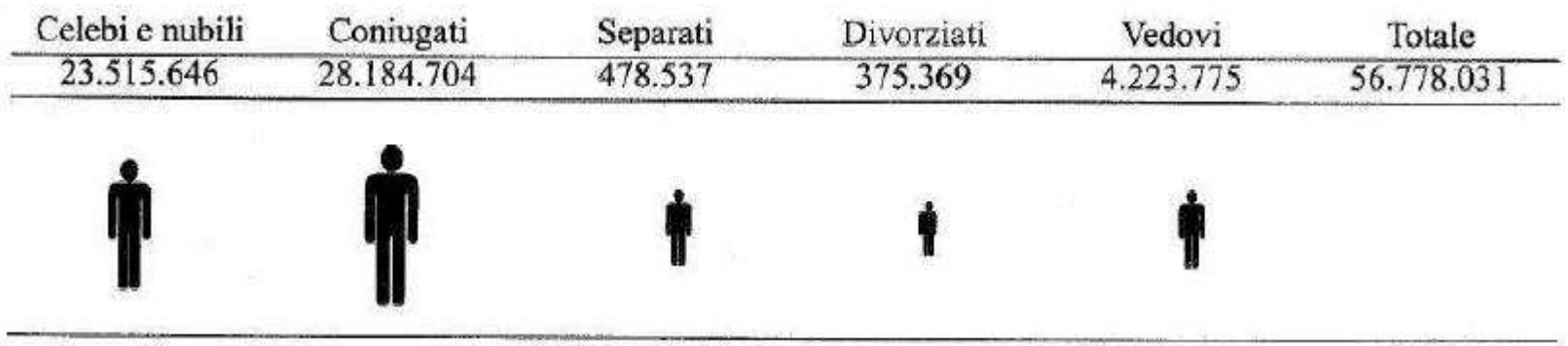


- **Pictogrammi o ideogrammi**





(fanno uso di figure stilizzate proporzionali alla frequenza)



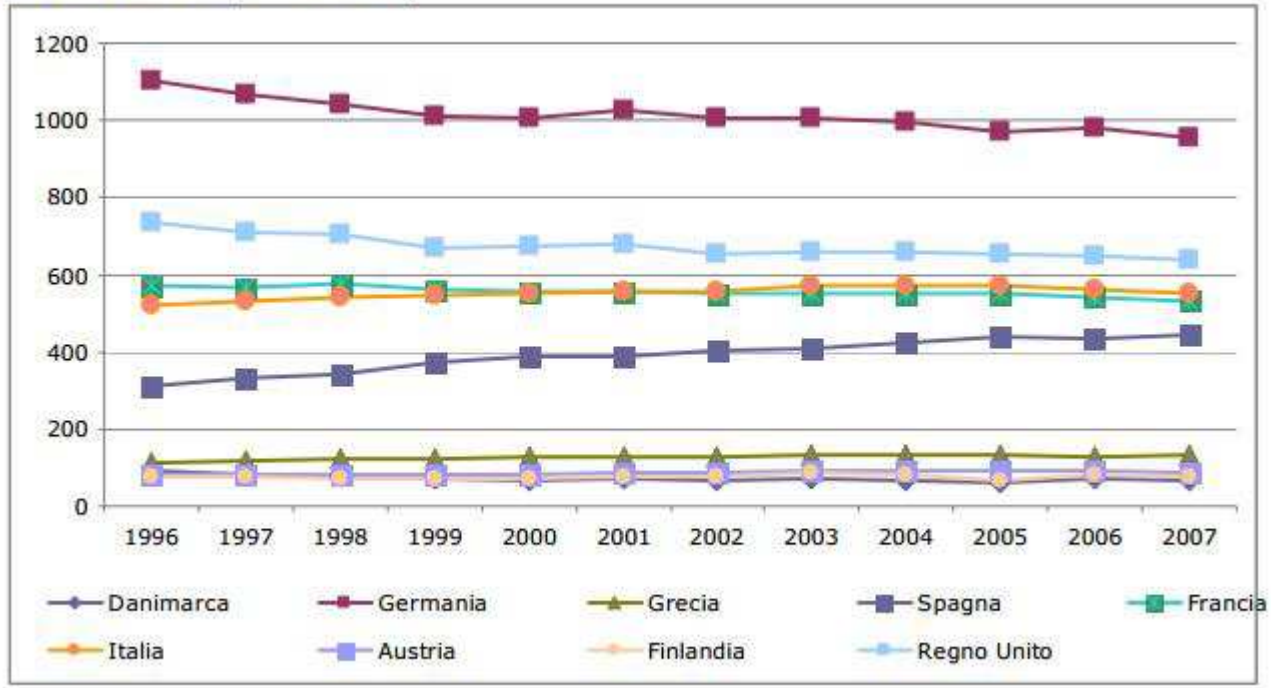
- **Cartogrammi**

ABRUZZO	7	MOLISE	2
BASILICATA	3	PIEMONTE	11
CALABRIA	13	PUGLIA	18
CAMPANIA	27	SARDEGNA	7
EMILIA ROMAGNA	13	SICILIA	19
FRIULI-VENEZIA GIULIA	4	TOSCANA	13
LAZIO	34	TRENTINO-ALTO ADIGE	2
LIGURIA	5	UMBRIA	4
LOMBARDIA	28	VALLE D'AOSTA	0,1
MARCHE	6	VENETO	14

- **Diagrammi cartesiani**

Emissioni totali nazionali di gas serra per alcuni Paesi dell'Unione europea

(milioni di tonnellate equivalenti di CO^2)



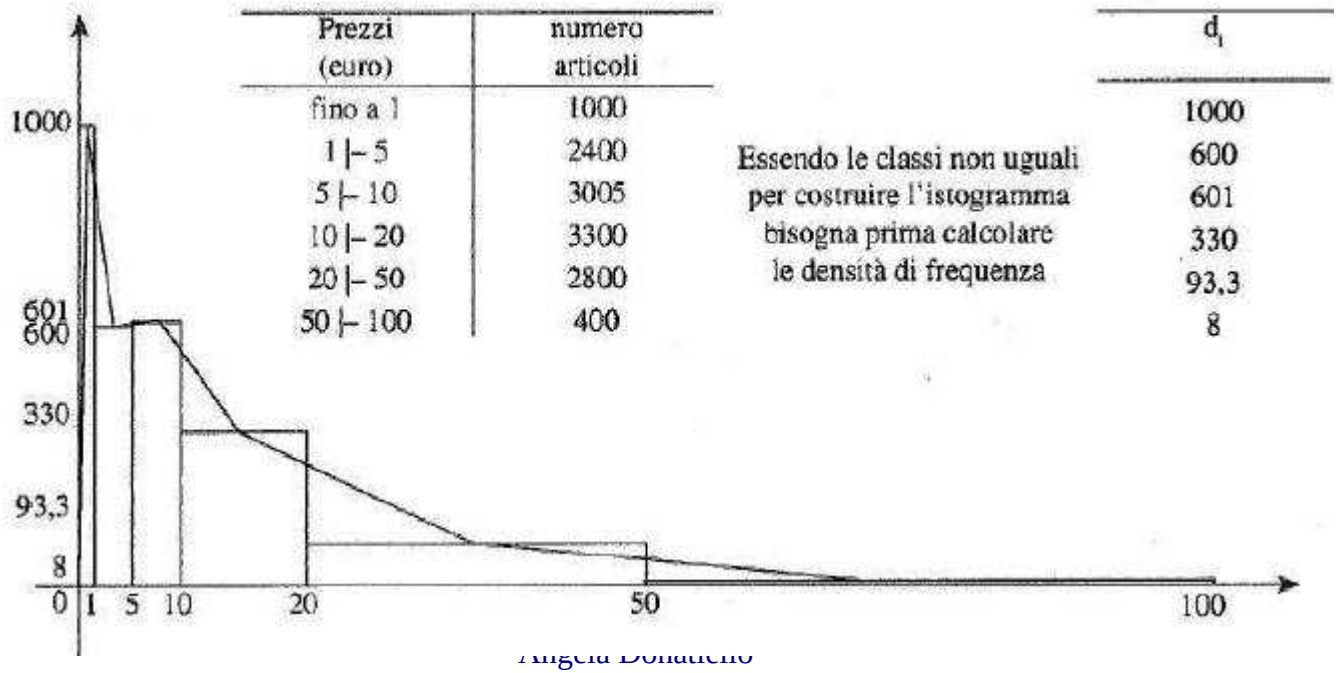
Fonte: Eurostat, Energy statistics

➢ Per caratteri continui raggruppati in classi

- **Istogramma:** è costituito da barre non distanziate, con basi che possono avere ampiezza diversa (dipende da come sono state costruite le classi).

L'area di ogni barra è proporzionale alla frequenza della classe (frequenza assoluta, relativa o percentuale). L'altezza del rettangolo (barra) è data dal rapporto fra la frequenza di classe(assoluta, relativa o percentuale) e l'ampiezza A della classe $h = \frac{f_a}{A}$ densità di frequenza

Nel caso di classi aventi stessa ampiezza, allora l'altezza sarà proporzionale alla frequenza.



Prezzi (euro)	numero articoli	d_i
fino a 1	1000	1000
1 |– 5	2400	600
5 |– 10	3005	601
10 |– 20	3300	330
20 |– 50	2800	93,3
50 |– 100	400	8

Essendo le classi non uguali per costruire l'istogramma bisogna prima calcolare le densità di frequenza

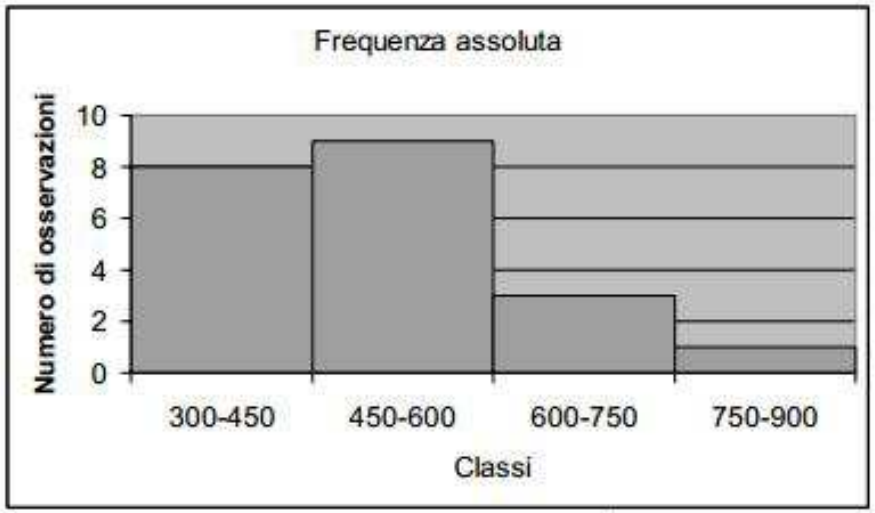
Esempio: serie di precipitazione annue osservate nella stazione di Caltanissetta nel periodo 1980-2000 (21 anni)

Anno	Precipitazione (mm)
1980	454.0
1981	356.2
1982	645.4
1983	409.8
1984	458.6
1985	487.2
1986	387.6
1987	452.6
1988	565.8
1989	332.8
1990	475.0
1991	687.6
1992	533.2
1993	376.2
1994	453.7
1995	357.2
1996	822.2
1997	618.4
1998	385.0
1999	390.2
2000	473.6

Consideriamo 4 classi di ampiezza 150 mm: 300-450, 450-600, 600-750, 750-900

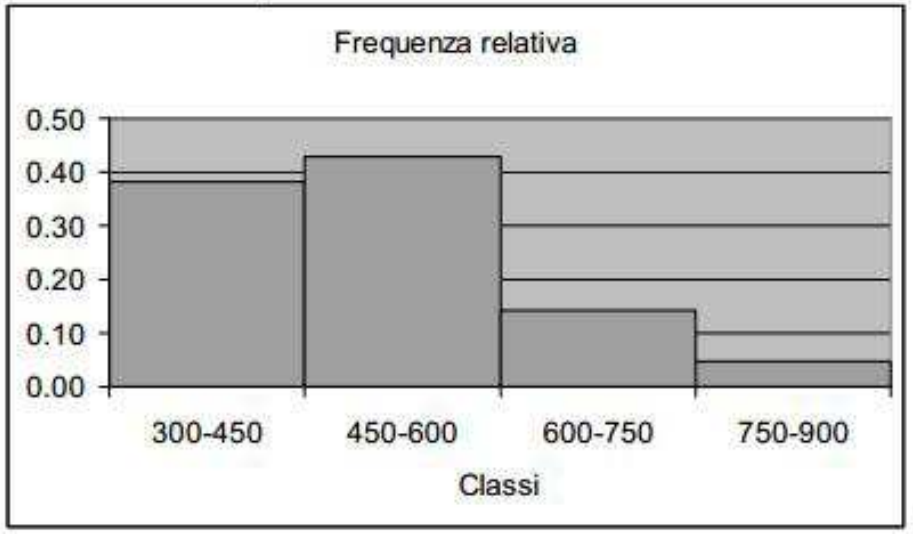
300-450	332.8, 356.2, 357.2, 376.2, 385.0, 387.6, 390.2, 409.8	8
450-600	452.6, 453.7, 454.0, 458.6, 473.6, 475.0, 487.2, 533.2, 565.8	9
600-750	618.4, 645.4, 687.6	3
750-900	822.2	1

Istogramma di frequenza assoluta e relativa



L'istogramma di frequenza assoluta riporta il numero di osservazioni che ricadono in ciascuna classe

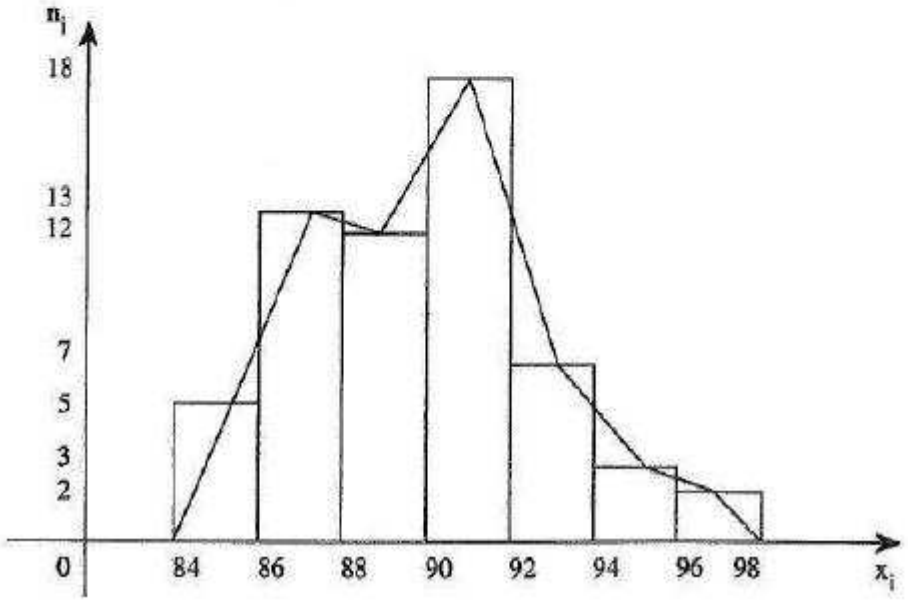
L'istogramma di frequenza relativa riporta il numero di osservazioni che ricadono in ciascuna classe in rapporto al numero totale di osservazioni



- **Poligono di frequenza:**

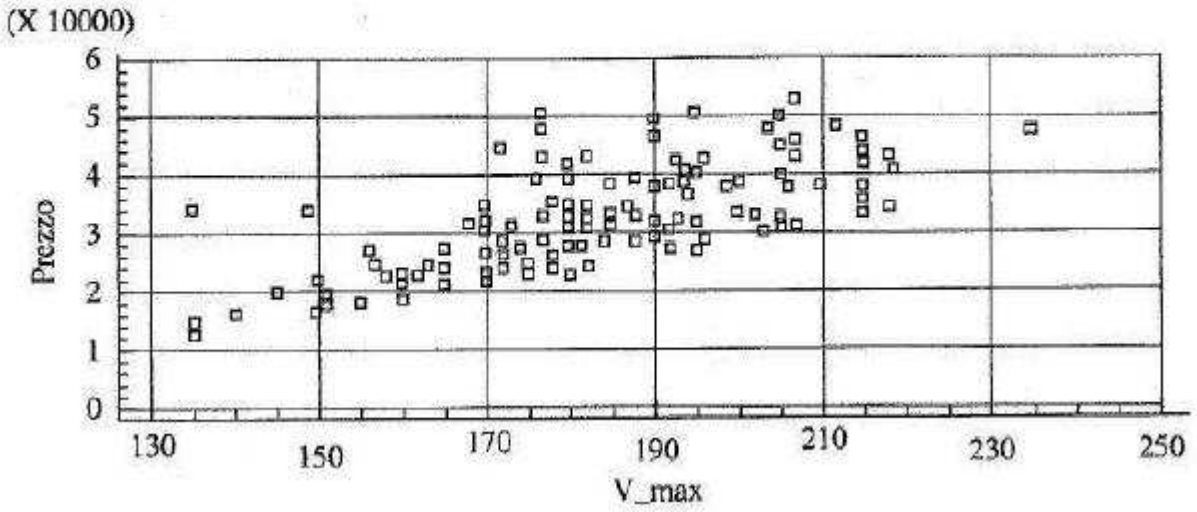
E' uno "smussamento"localmente lineare dell'istogramma. Per costruire il poligono di frequenza si segnano i punti medi dei lati superiori dei rettangoli dell'istogramma che vengono uniti con una spezzata di retta. Si ottiene così una spezzata sovrapposta all'istogramma

Classi di valori $h_i = x_{i-1} - x_i$	Frequenze n_i
84 |– 86	5
86 |– 88	13
88 |– 90	12
90 |– 92	18
92 |– 94	7
94 |– 96	3
96 |– 98	2
	60



- **Diagramma a dispersione**

Sono state rilevate per 132 autovetture le seguenti variabili, prezzo di listino e velocità massima (v. max) ed è stata effettuata una rappresentazione grafica bivariata del prezzo con la velocità mediante il diagramma di dispersione:



Utile nella statistica bivariata per valutare la correlazione tra le variabili

Riassumendo:

Studiare un carattere significa vedere come esso si distribuisce nella popolazione. Ciò si traduce nello studio della sua distribuzione di frequenza.

1) Costruire una tabelle di frequenza che ci permetta di comprendere come sono distribuite le frequenze tra le varie classi di misura;
2) Realizzare un grafico che rappresenti tale funzione;
3) Calcolare alcuni valori che forniscono un'indicazione riassuntiva della distribuzione, informando su dove è posizionata: **indici di posizione**;
4) Osservare come i dati si dispongono intorno agli indici di posizione e **misurare la variabilità**;
5) Studiare la **forma** della funzione distribuzione di frequenza (indici di forma)

Indici di posizione

Le sintesi ottenute con il calcolo delle frequenze relative sono di portata limitata, anche se molto utili sia per meglio valutare l'addensamento delle modalità, sia per fare confronti.
Un radicale processo di sintesi potrebbe portare invece a sostituire tutte le modalità della distribuzione con una "modalità" che le rappresenti. A tale modalità si dà il nome di media.

Definizione di media secondo Cauchy: si dice media di un insieme di dati qualsiasi valore compreso tra essi.

Definizione di media secondo Chisini: data una serie di dati $x_1,\dots,x_n$ si fissa una quantità che dipende dai dati (funzione obiettivo $f(x_1,\dots,x_n)$). Si definisce media dell'insieme di dati il valore che sostituito ad essi lasci invariata la funzione obiettivo.

$$f(x_1,\dots,x_n) = f(M,\dots,M)$$

Proprietà delle medie:

- ✓ Internalità: la media è sempre compresa tra il minimo e massimo valore della serie di dati
- ✓ Consistenza: la media di grandezze tutte uguali ad un valore k vale k
- ✓ Monotonia: se si calcola la media di due gruppi di dati tali che quelle del primo gruppo siano tutte minori o uguali a quelle del secondo gruppo, allora anche la media del primo gruppo di dati è minore o uguale alla media del secondo gruppo di dati

Osservazioni:

- ✓ La media potrebbe non coincidere con nessuno dei dati rilevati
- ✓ La media potrebbe non essere un dato sensato, ossia previsto come dato possibile. Esempio: nel documento L'Italia in cifre del 2004 (Fonte Istat) si ha che il numero medio di componenti delle famiglie italiane era 2,6 nel 2001 e 2,8 nel 1991.

OSS. Nel caso di distribuzione suddivisa in classi si assume come media il valore ottenuto sostituendo ogni classe con il suo valore centrale.

Media aritmetica semplice

Funzione obiettivo: $f(x_1,\ldots,x_n)=\sum_{i=1}^{n} x_i$

$$M=\frac{x_1+\ldots+x_n}{n}=\frac{\sum_{i=1}^{n} x_i}{n}$$

Media aritmetica ponderata

Data una serie di dati statistici $x_1,\ldots,x_n$ si considerino i "pesi" ossia le frequenze con cui si presentano tali dati.

$$M=\frac{x_1f_1+\ldots+x_nf_n}{f_1+..+f_n}=\frac{\sum_{i=1}^{n} x_if_i}{\sum_{i=1}^{n} f_i}$$

Funzione obiettivo: $f(x_1,\ldots,x_n)=\sum_{i=1}^{n} x_if_i$

Oss. La media aritmetica mantiene inalterata la somma

$$m=\frac{x_1+\ldots+x_n}{n}\Rightarrow n\cdot m=x_1+\ldots+x_n$$

Media geometrica semplice:

$$M_g = \sqrt[n]{x_1 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Funzione obiettivo: $f(x_1,\ldots,x_n) = \prod_{i=1}^{n} x_i$

Media geometrica ponderata:

Data una serie di dati statistici $x_1,\ldots,x_n$ si considerino i "pesi" ossia le frequenze con cui si presentano tali dati.

$$M_g = \sqrt[F]{x_1^{f_1} \cdot \ldots \cdot x_n^{f_n}} = \sqrt[F]{\prod_{i=1}^{n} x_i^{f_i}} \qquad \text{con } F = \sum_{i=1}^{n} f_i$$

Funzione obiettivo: $f(x_1,\ldots,x_n) = \prod_{i=1}^{n} x_i^{f_i}$

Oss: la media geometrica mantiene inalterato il prodotto

Media armonica semplice

Funzione obiettivo: $f(x_1,\ldots,x_n)=\sum_{i=1}^{n}\frac{1}{x_i}$

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

Media armonica ponderata

Data una serie di dati statistici $x_1,\ldots,x_n$ si considerino i "pesi" ossia le frequenze con cui si presentano tali dati.

Funzione obiettivo: $f(x_1,\ldots,x_n)=\sum_{i=1}^{n}\frac{f_i}{x_i}$

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

Media quadratica semplice

Funzione obiettivo: $f(x_1,\dots,x_n)=\sum_{i=1}^{n} x_i^2$

$$M_2=\sqrt[2]{\frac{x_1^2+\dots+x_n^2}{n}}=\sqrt[2]{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

Media quadratica ponderata

Data una serie di dati statistici $x_1,\dots,x_n$ si considerino i "pesi" ossia le frequenze con cui si presentano tali dati.

Funzione obiettivo: $f(x_1,\dots,x_n)=\sum_{i=1}^{n} f_i x_i^2$

$$M_2=\sqrt[2]{\frac{f_1x_1^2+\dots+f_2x_n^2}{f_1+\dots+f_n}}=\sqrt[2]{\frac{\sum_{i=1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i}}$$

Medie a confronto: $M_{arm} \le M_{geo} \le M_{arit} \le M_{quad}$

Oss. *In aula si è dimostrato per ogni tipo di media che essa effettivamente lascia invariata la funzione obiettivo indicata accanto.*

Altri valori di sintesi

La **moda** di una distribuzione di frequenza è la modalità a cui è associata la massima frequenza assoluta o relativa. Corrisponde quindi al valore "più rappresentativo" della distribuzione, quello che si è verificato più spesso. Si calcola per caratteri qualitativi sconnessi, caratteri qualitativi ordinali e per caratteri quantitativi discreti e continui organizzati in classi (classe modale). In tal caso se le classi sono di ampiezza diversa si fa riferimento alla densità di frequenza e non alla frequenza di ciascuna classe.

Classi	Freq	Ampiezza	Freq/ampiezza
0-100	5000	100	50
100-200	6500	100	65
200-400	12300	200	61,5
400-600	14200	200	71
600-1000	18400	400	46

classe Modale 400-600

Oss. Nel caso di carattere qualitativo, la moda è l'unico valore di sintesi considerabile

La **mediana** è la modalità dell' unità statistica che occupa il posto centrale nella distribuzione ordinata delle osservazioni. Si calcola per variabili qualitative ordinali e per variabili quantitative discrete e continue organizzate in classi.

Ordinati in senso crescente n numeri, si definisce mediana quel valore che, se n è dispari, coincide con il valore centrale, se n è pari, è ottenibile come media dei due valori centrali dell'ordinamento.

Esempio: la mediana dei valori 0; 3; **4**; 5; 8 è 4
la mediana dei valori 3; 3; **4; 6**; 7; 10 è 5 (media aritmetica di 4 e 6)

Nella successione di dati: 6; 11; 9; 2; 1; 3; 8; 13 è necessario prima ordinare la successione ottenendo 1; 2; 3; **6; 8**; 9; 11; 13. Poi si determina la mediana facendo la media dei due termini centrali, quindi 7.

Nel caso in cui la distribuzione dei dati sia fornita come distribuzione di frequenze, allora

- ✓ è necessario prima determinare le frequenze cumulate
- ✓ si valuta a quale valore o classe appartengono i due valori centrali
- ✓ si sceglie come mediana il valore trovato o il valore centrale della classe

Termini	Freq	Freq. Cum.
20	12	12
21	20	32
22	18	50
23	7	57
26	2	59
30	1	60

totale	60

Mediana **21**

Esempio. Distribuzione di frequenze

Numero di figli per famiglia	Frequenza	Frequenza cumulata
0	9	9
1	27	36
2	40	76
3	20	96
4	3	99
5	1	100
	100	

Determiniamo la media aritmetica, la moda e la mediana.

- ✓ La media aritmetica è una media ponderata:

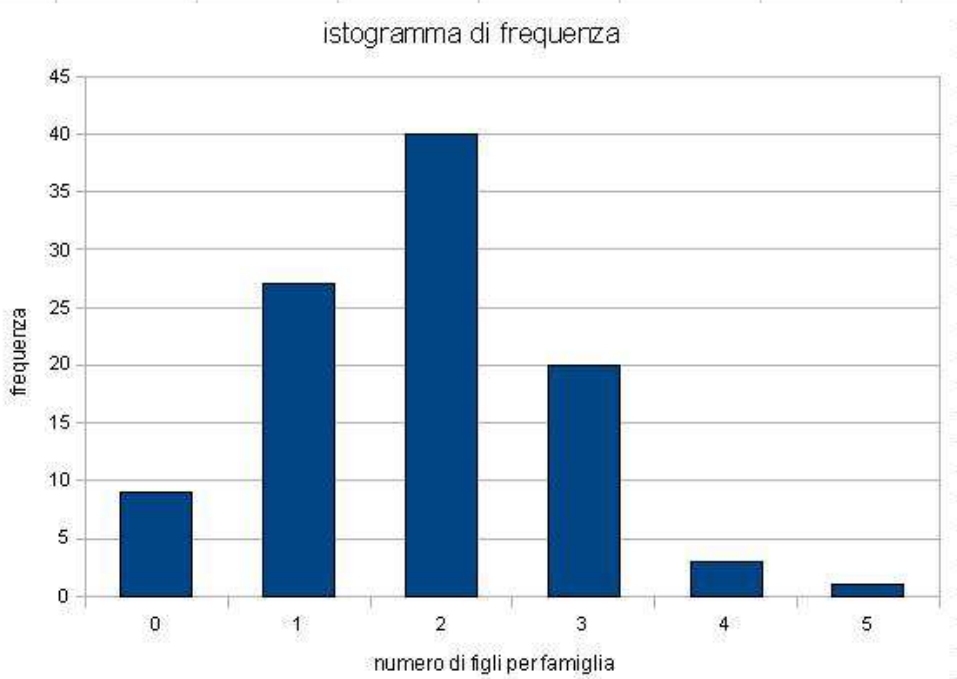
$$m = \frac{x_1 f_1 + \ldots + x_n f_n}{f_1 + \ldots + f_n} = \frac{0 \cdot 9 + 1 \cdot 27 + 2 \cdot 40 + 3 \cdot 20 + 4 \cdot 3 + 5 \cdot 1}{9 + 27 + 40 + 20 + 3 + 1} = \frac{184}{100} = 1{,}84$$

- ✓ La moda è il dato di maggior frequenza: moda = 2 (con frequenza 40)

- ✓ La mediana si determina considerando le frequenze cumulate. Il numero totale di famiglie intervistate è 100 (n pari). I due valori centrali sono pertanto 50 e 51. Si può osservare che questi due dati rientrano nella frequenza cumulata che indica il valore 76, pertanto la mediana è il valore corrispondente a tale frequenza cumulata. La mediana è quindi 2.

Oss. In questo caso la moda e la mediana coincidono, ma ciò non rappresenta una regola

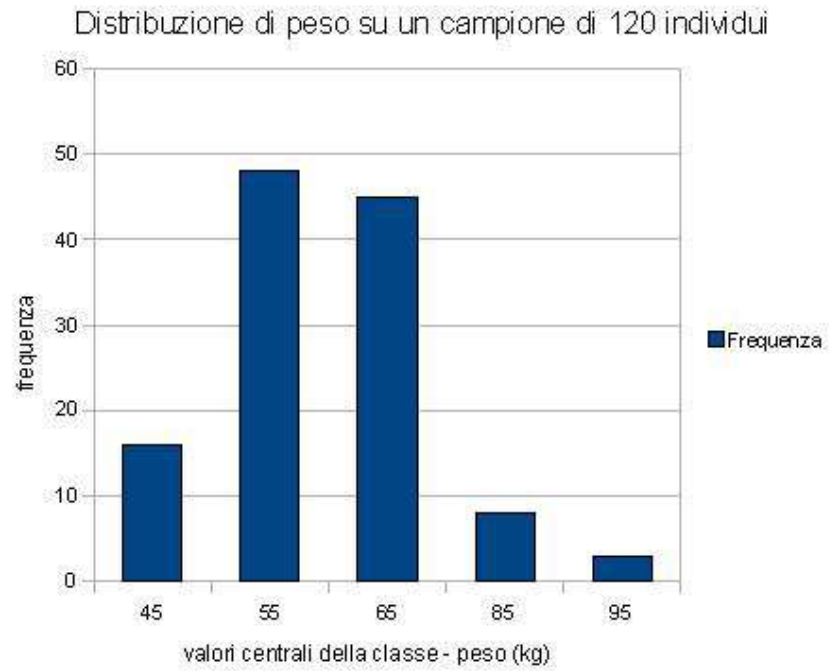
	A	B	C
1	Numero di figli per fam	Frequenza	Frequenza cumulata
2	0	9	9
3	1	27	36
4	2	40	76
5	3	20	96
6	4	3	99
7	5	1	100
8		100	



Esempio. Distribuzione suddivisa per classi

Un'indagine su un campione di individui ha prodotto la seguente distribuzione di frequenze.

	A	B	C	D	E	F
1	Peso (kg) – classi	Valore centrale della classe	Frequenza	Frequenza cumulata		
2	[40;50)	45	16	16		
3	[50;60)	55	48	64		
4	[60;70)	65	45	109		
5	[80;90)	85	8	117		
6	[90;100)	95	3	120		



- ✓ Si assume come media della distribuzione il valore che si ottiene sostituendo ad ogni classe il suo valore centrale e calcolando la media ponderata della distribuzione di frequenze

$$m = \frac{x_1 f_1 + \ldots + x_n f_n}{f_1 + \ldots + f_n} = \frac{(45 \cdot 16 + 55 \cdot 48 + 65 \cdot 45 + 85 \cdot 8 + 95 \cdot 3)\text{kg}}{16 + 48 + 45 + 8 + 3} = \frac{7250}{120} \cong 60{,}4\text{kg}$$

- ✓ Si assume come mediana il valore centrale della classe che contiene la mediana

In tal caso su 120 individui (n pari), i valori centrali sono due, 60 e 61. Osservando la colonna delle frequenze cumulate, si vede che il sessantesimo e il sessantunesimo peso osservato cadono nella classe $50 \le p < 60$, pertanto la mediana cade nella medesima classe e si può assumere come mediana il valore centrale di tale classe

$$\text{mediana} = \frac{50 + 60}{2} = 55$$

o in maniera più precisa si determina la mediana mediante la seguente proporzione

$$x : \text{ampiezza} = \left(\frac{\text{sommafrequenze}}{2} - \text{freq.cumulata classe prec.} \right) : \text{freq.classe mediana}$$

$$\text{mediana} = 50 + x = 50 + \frac{(60 - 16) \cdot 10}{48} = 59{,}2$$

In generale vale la formula:

$$\text{mediana} = l_1 + \frac{\frac{N}{2} + F}{f} a$$

Con:

l_1	= limite inferiore della classe mediana
N	= frequenza cumulata complessiva
F	= frequenza cumulata fino alla classe mediana
f	= frequenza (non cumulata) della classe mediana
a	= ampiezza della classe mediana

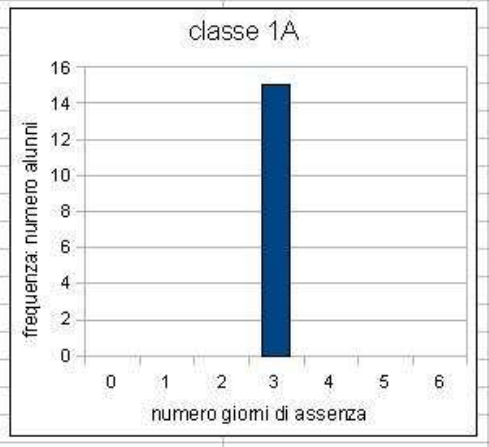
- ✓ Si assume come classe modale quella che ha maggiore frequenza se le classi hanno stessa ampiezza, o quella con maggiore densità di frequenza se le classi non hanno la stessa ampiezza.
  In tal caso l'ampiezza delle classi è la stessa A = 10, pertanto la classe modale è quella con frequenza maggiore, ossia $50 \leq p < 60$

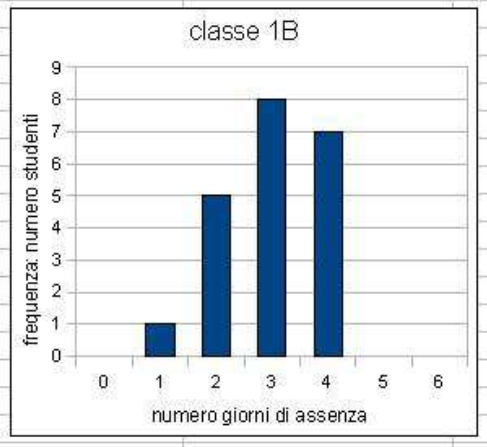
Oss. Il concetto di mediana può essere generalizzato introducendo il concetto di quartile. Si definiscono quartili i tre valori Q_1, Q_2, Q_3, Q_4 che dividono la distribuzione di dati in 4 parti uguali; sono indici di posizione calcolabili per caratteri qualitativi ordinali, per caratteri quantitativi discreti e quantitativi continui organizzati in classi. Il secondo quartile Q_2 coincide con la mediana.

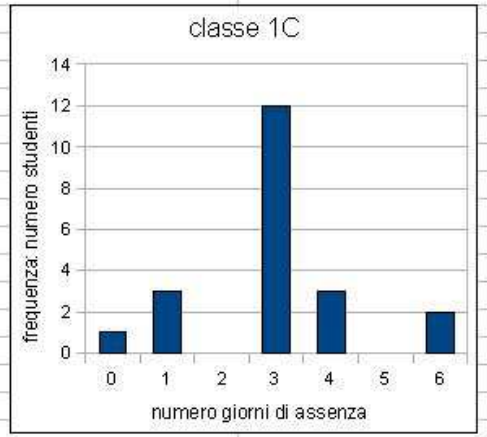
Gli indici di variabilità

Si considerino le seguenti distribuzioni di frequenze relativi al numero di assenze in una classe di scuole superiori

	A	B	C	D	E	F	G	H
1	Numero giorni di assenze	Frequenza: numero studenti		Numero giorni di assenze	Frequenza: numero studenti		Numero giorni di assenze	Frequenza: numero studenti
2	0	0		0	0		0	1
3	1	0		1	1		1	3
4	2	0		2	5		2	0
5	3	15		3	8		3	12
6	4	0		4	7		4	3
7	5	0		5	0		5	0
8	6	0		6	0		6	2







Si può osservare che tali distribuzioni hanno tutte stessa media, stessa moda e stessa mediana e valgono in tutti e tre i casi 3, ma le distribuzioni sono molto diverse l'una dall'altra.

Ciò fa capire che non basta la conoscenza di quale è la posizione media dei dati statistici.
La sola conoscenza di una media (sia essa la media aritmetica, o quella geometrica, o la mediana, o la moda) non è cioè sufficiente per descrivere in che modo i dati di partenza risultano distribuiti intorno a quel valore medio. Infatti gli esempi precedenti fanno vedere che una medesima media aritmetica può scaturire da insiemi di dati molto dissimili tra loro o che sono diversamente "addensati" vicino alla media aritmetica x.
Per misurare questo grado di dispersione, si introducono degli ulteriori indicatori numerici, detti appunto indici di dispersione. E' necessario dunque conoscere quale è la variabilità dei dati raccolti attorno al valore medio.
Allo scopo si introducono gli **indici di variabilità**. Essi devono possedere le seguenti caratteristiche:

- **Essere nulli in caso di variabilità nulla (tutti i dati statistici costanti)**
- **Essere positivi in caso di variabilità**
- **Essere crescenti all'aumentare della variabilità dei dati (con dati ordinati)**

Campo di variabilità o intervallo di variazione: $\text{range} = x_{max} - x_{min}$

La nozione di intervallo di variazione presenta un grave inconveniente: la sua ampiezza dipende in maniera determinante dalla presenza anche di un solo valore estremo molto diverso dagli altri, valore che il più delle volte è scarsamente significativo ai fini statistici (per es. può essere frutto della lettura errata di uno strumento, o di un errore di trascrizione, o simili). Ciò giustifica l'introduzione di altri indici di dispersione, meno influenzati dai valori estremi.

Semidispersione massima: $\frac{x_{max} - x_{min}}{2}$

Si dice scarto dalla media la quantità $s_i = x_i - M$. Tale quantità può essere sia positiva che negativa e la somma degli scarti dalla media è sempre nulla.

$$\sum_{i=1}^{n} s_i = \sum_{i=1}^{n} (x_i - M) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} M = nM - nM = 0$$

Per definire lo scarto semplice medio assoluto è dunque necessario introdurre il modulo.

Def. Scarto Semplice Medio Assoluto.
E' la media aritmetica dei valori assoluti degli scarti dalla media

$$S_M = \frac{\sum_{i=1}^{n} |s_i|}{n} = \frac{\sum_{i=1}^{n} |x_i - M|}{n}$$

$$S_M = \frac{\sum_{i=1}^{k} f_i |s_i|}{\sum_{i=1}^{k} f_i}$$

Devianza: la somma dei quadrati degli scarti

$$Dev(X) = \sum_{i=1}^{n} (x_i - M)^2 = \sum_{i=1}^{n} s_i^2$$

Def. Varianza sulla Popolazione

E' la media aritmetica degli scarti (dalla media aritmetica) al quadrato

$$\sigma^2(X) = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n} = \frac{\sum_{i=1}^{n} s_i^{\,2}}{n}$$

$$\sigma^2(X) = \frac{\sum_{i=1}^{k} f_i (x_i - M)^2}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i s_i^{\,2}}{\sum_{i=1}^{k} f_i}$$

Ma neppure la varianza è esente da inconvenienti. Infatti dal punto di vista "dimensionale" essa non è omogenea con i dati di partenza (se per es. gli xi sono lunghezze, la varianza rappresenta una lunghezza al quadrato; se gli xi sono temperature, o pressioni, la varianza rappresenta una temperatura al quadrato, una pressione al quadrato). Con un'ulteriore modifica si passa allora ad un nuovo indice, che di solito risulta preferibile alla varianza: la modifica consiste nell'annullare l'effetto degli elevamenti al quadrato mediante un'estrazione di radice quadrata.

Def. Scarto Quadratico Medio (Deviazione Standard) sulla Popolazione

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} s_i^{\,2}}{n}}$$

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - M)^2}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{\sum_{i=1}^{n} f_i s_i^{\,2}}{\sum_{i=1}^{k} f_i}}$$

Spesso le tecniche statistiche qui esposte vengono applicate non all'intera popolazione, ma solo ad un suo campione. Si cerca poi di stimare nel miglior modo possibile le caratteristiche dell'intera popolazione a partire dalle informazioni desunte dal campione. Quando si opera in questo modo, conviene modificare leggermente le formule, ponendo a denominatore il numero n - 1 in luogo del numero n. Si parla allora di varianza stimata e di scarto quadratico medio stimato o di deviazione standard stimata. Il motivo di questa modifica trova la sua giustificazione sulla base dei cosiddetti “gradi di libertà", che però in questa sede non approfondiremo. Per n abbastanza grande, la diversità tra varianza e varianza stimata, come pure tra scarto quadratico medio e scarto quadratico medio stimato, diventa trascurabile.

Varianza sul campione

$$\sigma^2(X) = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n-1} = \frac{\sum_{i=1}^{n} s_i^{\,2}}{n-1}$$

$$\sigma^2(X) = \frac{\sum_{i=1}^{k} f_i (x_i - M)^2}{\left(\sum_{i=1}^{k} f_i\right) - 1} = \frac{\sum_{i=1}^{k} f_i s_i^{\,2}}{\left(\sum_{i=1}^{k} f_i\right) - 1}$$

Deviazione standard sul campione

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} s_i^{\,2}}{n-1}}$$

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - M)^2}{\left(\sum_{i=1}^{k} f_i\right) - 1}} = \sqrt{\frac{\sum_{i=1}^{n} f_i s_i^{\,2}}{\left(\sum_{i=1}^{k} f_i\right) - 1}}$$

Distanza interquartile

Definiamo ora un altro indice di dispersione, che si ricollega alla nozione di mediana. Ricordiamo preliminarmente che, dopo avere riordinato gli n numeri xi per valori crescenti, la mediana Me suddivide questo insieme di numeri in due parti ugualmente numerose. Nulla vieta di suddividere lo stesso insieme ordinato di numeri in quattro parti ugualmente numerose.

Se per es. n = 27, si comincia col determinare la mediana: Me = elemento di posto centrale nell'insieme ordinato dei 27 valori x_i, ossia x_{14}. Si determina poi l'elemento di posto centrale nel sottoinsieme ordinato, formato dai 13 valori x_i che precedono Me, ossia x_7; analogamente si determina l'elemento di posto centrale nel sottoinsieme ordinato, formato dai 13 valori x_i che seguono Me, ossia x_{21}. I tre valori così ottenuti: $q_l = x_7$ $q_2 = Me = x_{14}$ $q_3 = x_{21}$ vengono detti **quartili** e più precisamente, nell'ordine, primo, secondo, terzo quartile. Naturalmente, se si applica il procedimento or ora descritto ad un insieme ordinato costituito da un numero qualsiasi n di valori x_i, può capitare che qualcuno dei sottoinsiemi da suddividere in due parti ugualmente numerose sia formato da un numero pari di elementi; in tal caso, come valore del corrispondente quartile si assume, al solito, la semisomma dei due valori più prossimi al posto centrale. Con queste notazioni, si considera come ulteriore indice di dispersione la cosiddetta **distanza interquartile**, definita da $\Delta = q_3 - q_1$

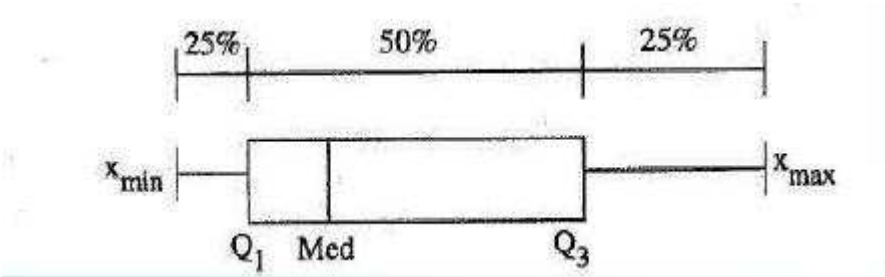
Per definizione, quindi, la distanza interquartile “taglia via" il 25% dei valori più bassi e il 25% dei valori più alti.

Un'efficace modalità di rappresentazione della distribuzione dei dati mediante l'utilizzo dei quartili è il cosiddetto **BOX PLOT**

Per determinare un box-plot servono: $x_{min}, Q_1, \text{Mediana}, Q_3, x_{max}$

Esso è così costituito da:

- ✓ Retta su cui situare i valori
- ✓ Box con estremi Q1 e Q3 (Differenze InterQuartile): all'interno del box sono contenute il 50% delle informazioni
- ✓ Una linea verticale all'interno del box che indica il valore della mediana
- ✓ Segmento estrema sinistra con lunghezza da xmin a Q1 : da xmin a Q1 sono contenute il 25% delle informazioni
- ✓ Segmento estrema destra con lunghezza da Q3 a x_{max} : da Q3 a x_{max} sono contenute il restante 25% delle informazioni

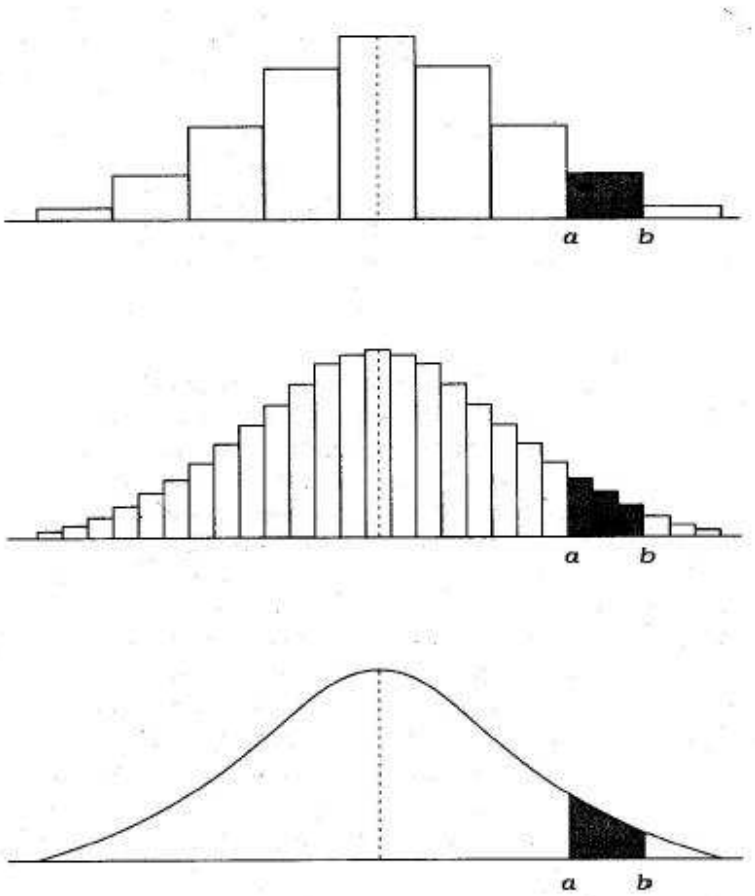


La distribuzione normale

Abbiamo visto come costruire un istogramma delle frequenze nel caso di grandezza che varia con continuità. Si suddivide l'intero intervallo delle misure in un numero finito n di intervallini (di solito tutti della stessa ampiezza). Si assume poi ciascun intervallino come base di un rettangolo dell'istogramma, facendo in modo che la *corrispondente area risulti proporzionale al numero delle misure che cadono entro l'intervallino considerato*. Facciamo ora l'ulteriore ipotesi, che la popolazione considerata sia molto numerosa (costituita da una quantità praticamente illimitata di individui). In tal caso il numero n degli intervallini può essere aumentato a piacere, diminuendone corrispondentemente le ampiezze. Si ottengono rettangoli via via più sottili e istogrammi via via più regolari, che in genere tendono a stabilizzarsi intorno ad una forma limite, approssimabile con una curva continua, detta curva di distribuzione delle frequenze relative con la classica forma "a campana".

Una funzione continua f(x) costituisce una distribuzione di frequenze relative su un insieme A se si verificano le due condizioni:

$$f(x) \geq 0 \qquad \int_A f(x)dx = 1$$



In questo esempio la curva limite appartiene alla famiglia di curve aventi equazioni del tipo:

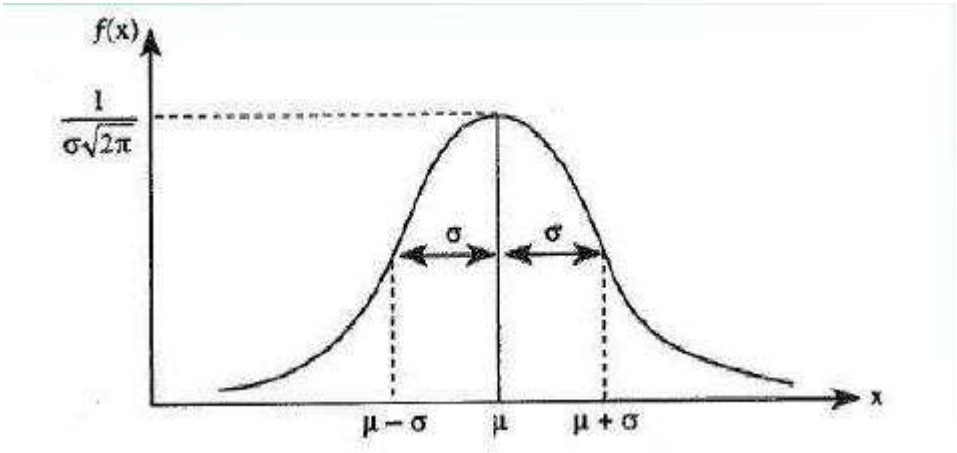
$$y = Ae^{-B(x-C)^2}$$

con A, B, C parametri opportuni.

Una siffatta distribuzione delle frequenze si chiama **distribuzione normale o distribuzione gaussiana**.

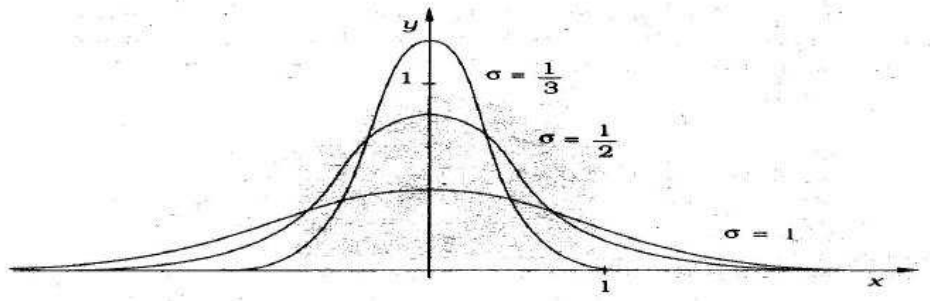
Se si sa già che la distribuzione è di tipo gaussiano, la determinazione dei valori numerici di A, B, C può essere ricondotta al solo calcolo della media aritmetica, che in questo contesto si denota tradizionalmente con μ, e dello scarto quadratico medio, che in questo contesto si denota tradizionalmente con σ. Risulta infatti:

$$A = \frac{1}{\sigma\sqrt{2\pi}} \qquad B = \frac{1}{2\sigma^2} \qquad C = \mu$$



$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Cerchiamo ora di interpretare il significato dei tre parametri A, B, C. Il valore di C si spiega facilmente: la distribuzione gaussiana è simmetrica e i valori delle singole misure si addensano intorno alla loro media aritmetica. Quindi la curva gaussiana teorica deve avere un massimo proprio in corrispondenza al valore $C = \mu$. Il valore assunto da B determina la maggiore o minore "ripidità" della curva gaussiana, e dipende quindi dalla maggiore o minore dispersione dei dati: quanto più σ è piccolo, tanto più la curva è "ripida", quanto più σ è grande, tanto più la curva è "piatta" Infine, il valore attribuito ad A serve a fare sì che l'area complessiva racchiusa tra la curva gaussiana e l'asse delle ascisse abbia misura unitaria.



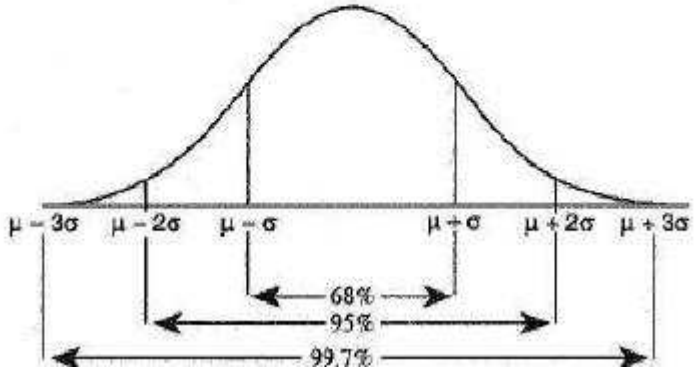
Proprietà della funzione $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$:

- ✓ È simmetrica rispetto alla retta $x=\mu$
- ✓ Ha punto di massimo in $x=\mu$, in tal caso $f(\mu)=\frac{1}{\sigma\sqrt{2\pi}}$
- ✓ Presenta due punti di flesso simmetrici in $x=\mu\pm\sigma$
- ✓ Tende asintoticamente a zero per x che tende a infinito $\lim_{x\to\pm\infty} f(x)=0$
- ✓ Media = Moda = Mediana
- ✓ Se si calcolano le frequenze cumulate si ottiene l'integrale improprio $F(y)=\int_{-\infty}^{y} f(x)dx$
- ✓ l'area sottesa dalla curva normale è pari a 1, in quanto rappresenta la somma di tutte le frequenze cumulate: $\int_{-\infty}^{+\infty} f(x)dx=\int_{R} f(x)dx=1$

Indicato con $[\mu - h\sigma; \mu + h\sigma]$ un intorno della media aritmetica μ, risulta che:

$$h=1 \Rightarrow \text{Freq}(\mu-\sigma \le x \le \mu+\sigma) = 0,68 = 68\%$$
$$h=2 \Rightarrow \text{Freq}(\mu-2\sigma \le x \le \mu+2\sigma) = 0,95 = 95\%$$
$$h=3 \Rightarrow \text{Freq}(\mu-3\sigma \le x \le +3\sigma) = 0,99 = 99\%.$$

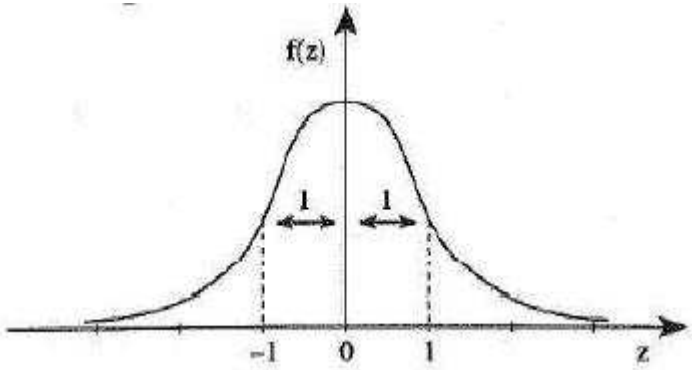
(Legge dei tre sigma) In una distribuzione normale la (quasi) totalità dei casi osservati è compresa in un intorno completo di μ di ampiezza 6σ.



Curva normale standardizzata

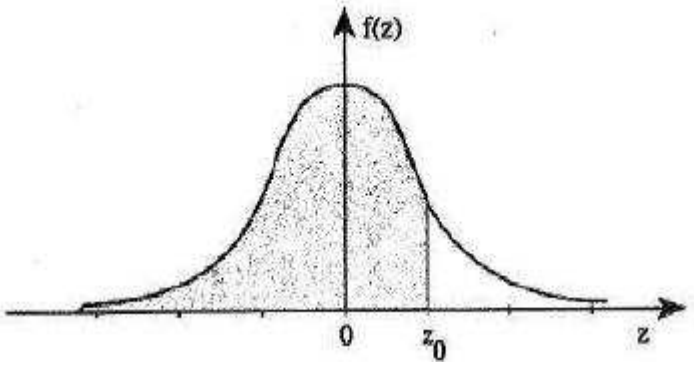
Se μ = 0 e σ = 1, l'equazione assume una forma particolarmente semplice, detta curva normale standardizzata

$$z = \frac{x-\mu}{\sigma} \qquad y = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$



Distribuzione di frequenza cumulata:

$$\Phi(z_0) = \int_{-\infty}^{z_0} f(t)dt$$

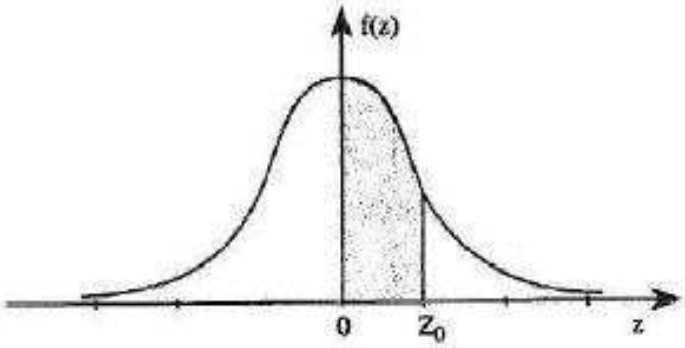


L'area sottesa dalla distribuzione normale standard nell'intervallo [a;b] è data dalla relazione

$$\text{Freq}(a \le z \le b) = \Phi(b) - \Phi(a)$$

$$\text{Freq}(0 \le z \le z_0) = \Phi(z_0) - \Phi(0) = \Phi(z_0) - \frac{1}{2}$$

$$\Phi(0) = \frac{1}{2}$$



$$\text{Freq}(a < x < b) = \text{Freq}\left(\frac{a-\mu}{\sigma} < \frac{x-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = \text{Freq}\left(\frac{a-\mu}{\sigma} < z < \frac{b-\mu}{\sigma}\right).$$

La distribuzione di frequenza della normale standard è tabulata.

Nella prima colonna sono indicati i valori di z fino alla prima cifra decimale. Nella prima riga sono indicati i valori della seconda cifra decimale.

I valori negativi non sono tabulati, in quanto per simmetria della curva si ottengono da quelli positivi.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

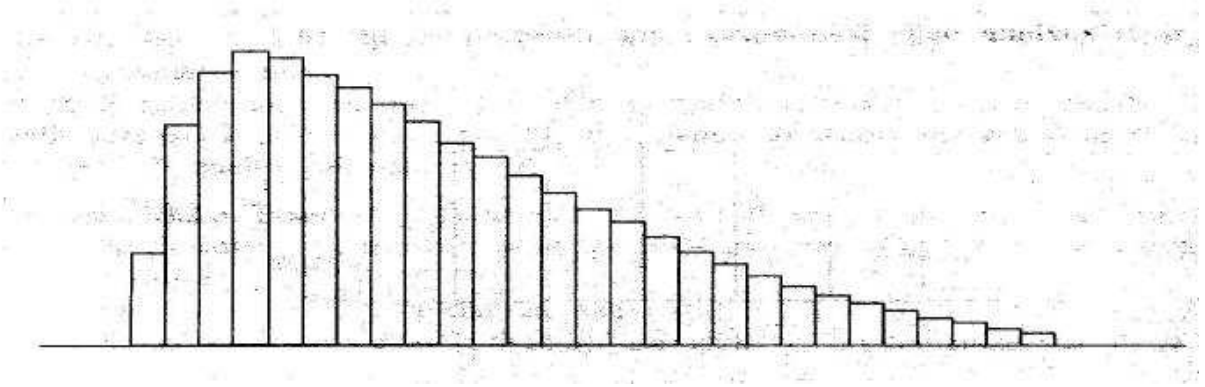
Esempio. I voti di un compito di matematica sono stati dall'1 al 10. Ipotizzando che si distribuiscano secondo la legge normale continua con valor medio 6,7 e scarto quadratico medio 1,2 determinare la percentuale di studenti che ha preso il voto tra 5,5 e 6,5.

$$x = 5{,}5 \Rightarrow z = \frac{5{,}5 - 6{,}7}{1{,}2} = -1 \qquad x = 6{,}5 \Rightarrow z = \frac{6{,}5 - 6{,}7}{1{,}2} = -0{,}1667 \cong -0{,}17$$

$\text{Freq}(-1 \le z \le -0{,}17) = \text{Freq}(0{,}17 \le z \le 1)$ per simmetria della distribuzione

$$\Rightarrow \text{Freq}(-1 \le z \le -0{,}17) = \text{Freq}(0{,}17 \le z \le 1) = \Phi(1) - \Phi(0{,}17) = 0{,}3413 - 0{,}0675 = 0{,}2738 = 27{,}38\%$$

Non sempre un insieme di misure tende a disporsi secondo una distribuzione gaussiana.
La constatazione se un insieme di misure sperimentali sia approssimabile o meno con una distribuzione gaussiana è un fatto di natura sperimentale.



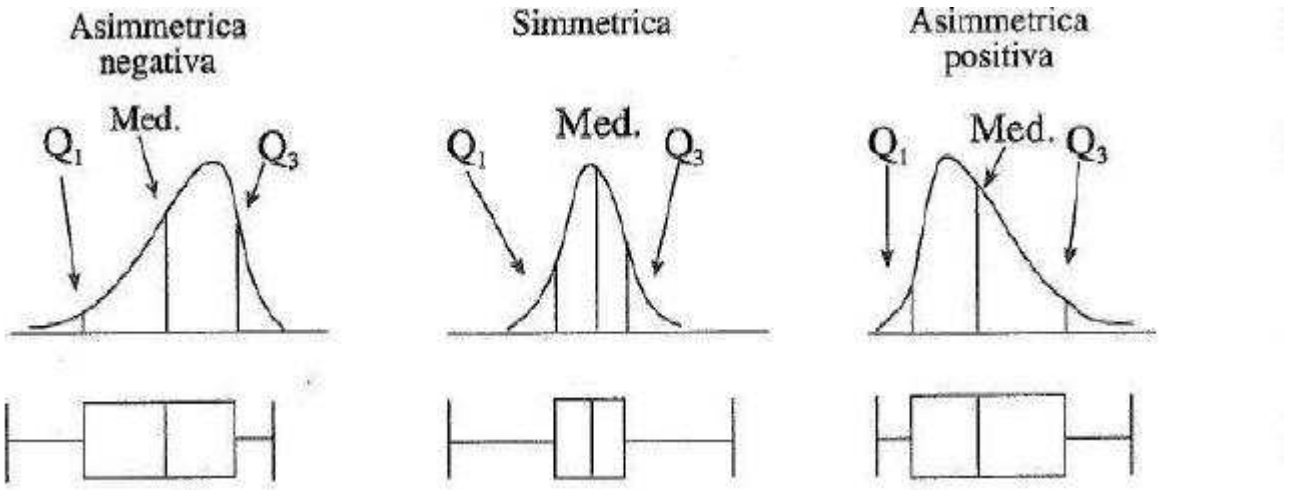
E’ ben noto però che se uno stesso sperimentatore, o sperimentatori diversi, ripetono più volte la misura di una medesima grandezza i risultati delle singole misure in generale non coincidono tra loro, per effetto della presenza di numerosi piccoli errori casuali. Le misure tendono però ad addensarsi in prossimità di un valore centrale, identificabile con la loro media aritmetica, dando luogo ad una distribuzione di tipo gaussiano. Se le misure non sono affette da errori sistematici (dovuti per es. ad un'errata taratura degli strumenti) è ragionevole assumere tale valore centrale come misura più probabile o misura attendibile della grandezza in esame. La distribuzione gaussiana ricopre pertanto un ruolo fondamentale nell’ambito della teoria degli errori per le scienze sperimentali. Essa ha una diffuso utilizzo anche nelle scienze sociali quali la psicologia o la sociologia.

Indici di forma

1) **Asimmetria** (skew)

$$SKEW = \frac{\sum_{i=1}^{n} s_i^3}{\left[\sum_{i=1}^{n} s_i^2\right]^{3/2}} = \frac{\sum_{i=1}^{n} (x_i - M)^3}{\left[\sum_{i=1}^{n} (x_i - M)^2\right]^{3/2}}$$

- ✓ Skew = 0 distribuzione simmetrica
- ✓ Skew < 0 distribuzione asimmetrica negativa: maggior contributo dei dati statistici minori della media rispetto alla distribuzione simmetrica
- ✓ Skew >0 distribuzione asimmetrica positiva: maggior contributo dei dati statistici maggiori della media rispetto alla distribuzione simmetrica

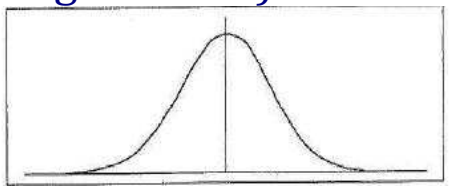


2) Curtosi

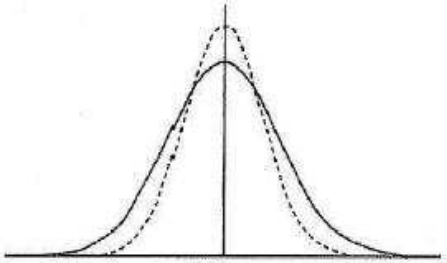
$$Curtosi = \frac{\sum_{i=1}^{n} s_i^4}{\left[\sum_{i=1}^{n} s_i^2\right]^2} = \frac{\sum_{i=1}^{n} (x_i - M)^4}{\left[\sum_{i=1}^{n} (x_i - M)^2\right]^2}$$

La Curtosi misura il peso relativo delle "code" della distribuzione rispetto alla parte centrale.
(il confronto avviene relativamente ad una distribuzione gaussiana)

Curtosi = 3 distribuzione normale



Curtosi < 3 distribuzione ipernormale
(più appuntita di una gaussiana;
code leggere)



Curtosi > 3 distribuzione iponormale
(distribuzione meno appuntita
di una gaussiana; code
pesanti)