ISTITUZIONI DI MATEMATICHE E FONDAMENTI DI BIOSTATISTICA

11. ELEMENTI DI STATISTICA

A. A. 2014-2015 *L. Doretti*

La **STATISTICA** si può definire come:

l'insieme dei metodi scientifici finalizzati alla raccolta, presentazione, organizzazione, analisi, interpretazione di dati e anche, in senso più lato, alla previsione basata su tali dati

Il termine statistica venne introdotto nel XVII secolo con il significato di ***"scienza dello stato"***, poiché aveva lo scopo di raccogliere dati per ottenere informazioni utili all'amministrazione pubblica:

- entità e composizione della popolazione

- movimenti migratori

- mutamenti anagrafici

- la mortalità determinata da particolari malattie o epidemie

- dati sui commerci, sui raccolti, sulla distribuzione della ricchezza, sull'istruzione e la sanità, …

Oggi, la **statistica** si rivolge ad ogni campo della conoscenza umana ed è legata a tutte le scienze nel momento in cui esse *osservano* e si *pongono interrogativi* sulla *realtà dei fatti*

Fra i campi di applicazione primeggiano fin dall'inizio la biologia e la medicina

I metodi statistici applicati alla ricerca in biologia e in medicina sono definiti come

BIOSTATISTICA o **STATISTICA MEDICA**

La statistica ha due scopi principali:

1. Ricavare da una molteplicità di dati individuali, troppo numerosi per poter essere presi in considerazione singolarmente, alcune informazioni significative secondo le esigenze del particolare problema che interessa

STATISTICA DESCRITTIVA:
insieme delle metodologie finalizzate a descrivere in forma sintetica le caratteristiche principali dell' insieme di dati raccolti

2. Fornire metodi che consentono, per quanto possibile, di trarre delle <u>conclusioni generali dall'osservazioni di casi particolari</u>

STATISTICA INDUTTIVA *o* ***INFERENZA STATISTICA****: insieme delle metodologie che consentono di stimare una o più caratteristiche di una "popolazione" o di prendere decisioni su di essa, sulla base dei risultati ottenuti su una sua parte, detta"campione"*

(linguaggio e fondamento della statistica induttiva e la *teoria della probabilità*)

STATISTICA DESCRITTIVA

Nelle indagini statistiche si prendono in considerazione sempre delle collettività, o come più comunemente si dice delle *popolazioni*

La grande varietà delle applicazioni della statistica rende però opportuna una generalizzazione del concetto di popolazione

Si definisce ***popolazione statistica***, o ***universo statistico***, *un insieme di elementi (esseri umani, animali, vegetali, oggetti, osservazioni, ecc,) che presentano tutti delle caratteristiche comuni*

Tali elementi si dicono ***individui*** o ***unità statistiche***

E’ essenziale che la popolazione oggetto di un’indagine statistica sia perfettamente individuata

*(se occorre fare particolari **convenzioni** per delimitare una popolazione, tali convenzioni dovranno essere esplicitate quando si comunicano i dati rilevati e i risultati dedotti da tali dati)*

Esempio

- Per compiere rilevamenti per un’indagine sull’occupazione giovanile in Italia, è necessario stabilire entro quali limiti di età un cittadino italiano deve essere preso in considerazione agli effetti dell’indagine

IL CAMPIONAMENTO

Spesso un'indagine statistica, riguardante una certa popolazione, viene compiuta su un **campione**, cioè su una parte opportunamente scelta della popolazione considerata, e ciò per vari motivi:

- *limitata disponibilità di tempo*
- *limitata disponibilità di mezzi finanziari*
- *particolare natura dell'indagine stessa*

  (per es., sperimentazione di una o più terapie,…)
- …..

La scelta del campione deve essere fatta in modo da poter soddisfare i due obiettivi seguenti:

- ottenere un campione *sufficientemente rappresentativo della popolazione*
- poter valutare l'entità dell'errore, detto *errore di campionamento*, che si commette quando si assumono come validi per tutta la popolazione i risultati dedotti dall'analisi del campione

Per ottenere le condizioni precedenti, si devono considerare campioni costituiti da:

- un numero elevato di elementi
- gli individui del campione devono essere prelevati a caso dalla popolazione *(per evitare l'uso di un criterio soggettivo di scelta)*

Campionamento casuale

Per *prelevare a caso* un campione da una determinata popolazione, si può procedere così:

- si *associa* un diverso numero d'ordine a ciascun individuo della popolazione

- si *selezionano* i numeri corrispondenti agli individui da includere nel campione *mediante estrazione a sorte* da una o più urne *(effettuate in modo da garantire ad ogni individuo la stessa probabilità di essere estratto)*

Esempio: **si vuole estrarre un campione di 100 individui da una popolazione di 3.000 persone, delle quali siano noti i nominativi**

- *Dopo aver abbinato a ciascun individuo un numero compreso tra 0 e 2999, si predispongono quattro urne A, B, C, D. L'urna A contiene i numeri 0, 1, 2 mentre le altre tre urne contengono ciascuna i dieci numeri da 0 a 9*
- Estraendo un numero da ciascuna urna nell'ordine, si ottiene il numero di un individuo della popolazione che entrerà a far parte del campione
- *Ripetendo il procedimento si individuano tutti gli elementi del campione (dovranno essere ripetute le estrazioni che abbiano per esito un numero d'ordine estratto in precedenza)*

26	96927	19931	36809	74192	77567	88741	48409	41903
27	43909	99477	25330	64359	40085	16925	85117	36071
28	15689	14227	06565	14374	13352	49367	81982	87209
29	36759	58984	68288	22913	18638	54303	00795	08727
30	69051	81717	40951	78453	4004	89872	87201	97245
31	05007	16632	81194	14873	04197	85576	45195	96565
32	68732	55259	84292	08796	43165	93739	31685	97150
33	45740	41807	65561	33302	07051	93623	18132	09547
34	27816	78416	18329	21337	35213	37741	04312	68508
35	66925	55658	39100	78458	11206	19876	87151	31260
36	08421	44753	77377	28744	75592	08563	79140	92454
37	53645	66812	61421	47836	12609	15373	98481	14592
38	66831	68908	40772	21558	47781	33586	79177	06928
39	55588	99404	70708	41098	43563	56934	48394	51719
40	12975	13258	13048	45144	72321	81940	00360	02428
41	96767	35964	23822	96012	94591	65194	50842	53372
42	72829	50232	97892	63408	77919	44575	24870	04178
43	88565	42628	17797	49376	61762	16953	88604	12724
44	62964	88145	83083	69453	46109	59505	69680	00900
45	19687	12633	57857	95806	09931	02150	43163	58636
46	37609	59057	66967	83401	60705	02384	90597	93600
47	54973	86278	88737	74351	47500	84552	19909	67181
48	00694	05977	19664	65441	20903	62371	22725	53340
49	71546	05233	53946	68743	72460	27601	45403	88692
50	07511	88915	41267	16853	84569	79367	32337	03316

In alternativa si può utilizzare una ***tabella di numeri casuali***, cioè di una *tabella di numeri ad una cifra che si susseguono in un ordine casuale*

Dopo aver abbinato un numero d'ordine a ciascun individuo della popolazione:

- si sceglie una riga a caso nella tabella
- si considerano i numeri ottenuti raggruppando a 4 a 4 le cifre
- si conservano solo i numeri compresi tra 0000 e 2999

Campionamento per stratificazione

- Si applica quando la popolazione, oggetto di un'indagine statistica, può essere suddivisa in **strati**, cioè in parti, ciascuna delle quali si possa ritenere omogenea ai fini dell'indagine stessa

- Per estrarre un campione da una tale popolazione, è preferibile estrarre a caso, da ciascuno strato, una parte del campione, in modo tale da riprodurre, in quest'ultimo, le stesse proporzioni tra i vari strati esistenti nella popolazione

Esempio

Si vuole stimare mediante un campione di 1500 individui l'altezza media di 2 milioni di abitanti di una regione suddivisa in quattro province A, B, C, D ciascuna contenente rispettivamente il 10%, il 20%, il 30%, il 40% degli abitanti.

Come ottenere il campione?

Detti X_A, X_B, X_C, X_D i numeri che indicano quanti individui del campione dovranno essere prelevati, rispettivamente, dalle province (o strati) A, B, C, D , si ricava:

$X_A = 150$ *(10% del totale)*

$X_B = 300$ *(20% del totale)*

$X_C = 450$ *(30% del totale)*

$X_D = 600$ *(40% del totale)*

L'estrazione in ogni strato avverrà poi in modo casuale

Campionamento a stadi successivi

Si usa quando la popolazione oggetto di un'indagine statistica è molto numerosa

In una *prima fase*, suddivisa la popolazione in varie parti, si estraggono a sorte un certo numero di queste

Successivamente, *suddivise a loro volta in porzioni più piccole le parti selezionate nella prima fase*, si estraggono a sorte un certo numero di tali porzioni, e così di seguito

Si individuano così porzioni sempre più esigue della popolazione, fino a determinare delle parti tanto piccole da consentire di avere, mediante un'estrazione a sorte finale, gli individui da includere nel ***campione***

CARATTERI DI UNA POPOLAZIONE E TIPOLOGIA DEI DATI

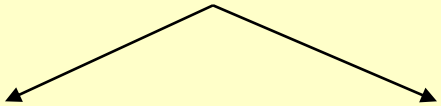
Una popolazione statistica può essere studiata in relazione alle diverse proprietà caratteristiche degli elementi che la compongono

Ogni caratteristica di una popolazione si chiama ***carattere*** o ***attributo*** ed il modo con cui il carattere si manifesta si dice ***modalità*** del carattere

Esempi

- *Gli impiegati dello Stato si possono studiare rispetto a vari caratteri: età, titolo di studio, stipendio annuo percepito, anzianità di servizio, mansioni che esplicano,....*
- *Le famiglie italiane possono essere studiate secondo il numero dei componenti, il numero dei figli, il reddito,...*

Un **carattere** può essere



qualitativo

quantitativo

Uno degli aspetti dell'indagine statistica è lo studio di
come si distribuisce una popolazione
rispetto al carattere considerato

CARATTERI QUALITATIVI E QUANTITATIVI

I ***caratteri qualitativi***

non sono esprimibili numericamente, ma possono essere indicati soltanto con espressioni verbali

(es.: stato civile, colore degli occhi, gruppo sanguigno, titolo di studio,…)

I ***caratteri quantitativi*** sono *esprimibili con un numero*

In questo caso si usa spesso il termine di ***variabile***, in luogo di carattere

(es.: numero di componenti di una famiglia, peso, altezza, lunghezza, area, …)

Carattere quantitativo (o variabile)

Discreto

i dati derivano da conteggi

(possono assumere solo valori espressi da numeri interi)

Continuo

i dati derivano da misurazioni

(possono assumere valori espressi da un qualunque numero reale compreso in un certo intervallo)

Esempi:

- *il numero dei petali di una certa varietà di fiori*
- *il numero dei vani di un'abitazione*
- *Il numero di componenti delle famiglie italiane*

Esempi:

- *i caratteri biometrici (statura, peso, perimetro toracico, pressione arteriosa,...)*
- *l'età*
- *i prezzi dei generi di consumo*
- *l'area delle superfici delle abitazioni*

PRESENTAZIONE DEI DATI

In ogni campo della scienza, l'informazione, per poter essere utile sia alla ***conoscenza*** che ***all'attività pratica*** deve essere resa compatta e significativa

La **statistica descrittiva** organizza e sintetizza un insieme di dati e consente una visione di insieme delle sue caratteristiche generali

I metodi statistici più appropriati per descrivere un insieme di dati sono: ***tabelle***, ***grafici*** e ***misure di sintesi numerica***

TABULAZIONE DEI DATI

- Una **tabella** è il modo più semplice per organizzare e sintetizzare un insieme di dati, ed è utilizzabile per dati sia *qualitativi* che *quantitativi*

Un tipo di tabella comunemente utilizzata per valutare i dati è la **tabella di distribuzione di frequenza**

DISTRIBUZIONE DI FREQUENZA ASSOLUTA

- Nello studio di una popolazione (o di un campione) secondo un carattere qualitativo o quantitativo, i dati raccolti vengono ripartiti in tanti sottoinsiemi (*classi*) quante sono le modalità che il carattere stesso può assumere

- Per ogni modalità del carattere si considera la **frequenza assoluta**, ovvero il numero dei dati che rispecchiano quella modalità del carattere

- Il risultato di questa organizzazione dei dati si presenta sottoforma di una ***tabella di distribuzione di frequenza assoluta***

Distribuzione di frequenza assoluta per un carattere qualitativo

Tabella 2.1-1

Tabella di frequenza che mostra le 10 cause di morte più comuni degli adolescenti statunitensi di età compresa tra 15 e 19 anni nel 1999.

Causa di morte	Frequenza
Incidenti	6688
Omicidio	2093
Suicidio	1615
Tumore maligno	745
Cardiopatia	463
Anomalie congenite	222
Malattia respiratoria cronica	107
Influenza e polmonite	73
Malattie cerebrovascolari	6
Altri tumori	52
Tutte le altre cause	1653
Totale	**13 778**

Distribuzione di frequenza assoluta per un carattere quantitativo

Tabella 1. Numero di foglie contate su 45 rami.

5	6	3	4	7	2	3	2	3	2	6	4	3	9	3
2	0	3	3	4	6	5	4	2	3	6	7	3	4	2
5	1	3	4	3	7	0	2	1	3	1	5	0	4	5

Il primo passaggio, quasi intuitivo in una distribuzione discreta, consiste nel **definire le classi**:

- è sufficiente identificare il **valore minimo (0,** nei dati della tabella) **e quello massimo (9)**,
- contando quante volte compare ogni **modalità di espressione** (cioè quanti sono i rami con un numero di foglie uguali).

Classe	**x**	0	1	2	3	4	5	6	7	8	9
Freq. Assoluta	**n**	3	3	7	12	7	5	4	3	0	1

DATI RAGGRUPPATI

Quando, nello studio di un carattere quantitativo, le *modalità* con cui il carattere si esprime (ovvero le classi) sono molto numerose, si ha la necessità di "condensare" i dati (***dati raggruppati***)

Per ***realizzare la tabella di distribuzione di frequenza con dati raggruppati*** si procede così:

- si suddivide l'intervallo di variabilità, *cioè l'intervallo tra il più piccolo ed il più grande dei valori osservati*, in un certo numero di intervalli distinti, consecutivi, non sovrapposti, e possibilmente della stessa ampiezza: **ciascun sottointervallo costituirà così una classe**
- si raggruppano in una stessa classe tutti i dati compresi tra gli estremi di ogni intervallo parziale

Tabella 3. Altezza in cm. di 40 giovani piante.

107	83	100	128	143	127	117	125	64	119
98	111	119	130	170	143	156	126	113	127
130	120	108	95	192	124	129	143	198	131
163	152	104	119	161	178	135	146	158	176

Distribuzione di frequenza assoluta dell'altezza di 40 giovani piante

Classe	X_i	60-79	80-99	100-19	120-39	140-59	160-79	180-99
Freq. Assoluta	n_i	1	3	10	12	7	5	2

- Per i dati raggruppati, gli intervalli che individuano le classi non devono essere *né troppi*, *né troppo pochi*

  (di solito il loro numero varia da 5 a 15)

  Uno dei metodi per calcolare il numero opportuno di classi è il seguente:

  Se N è il numero dei dati, il numero C di classi è dato dalla formula :

$$C = 1 + \frac{10}{3} LogN$$

- Una tabella di dati raggruppati fornisce un quadro d'insieme dei dati, indicando come essi sono "distribuiti"

  (ovviamente si perdono alcune osservazioni: dalla tabella non si possono più ricreare i valori originali dei dati)

Le tabelle forniscono maggiori informazioni quando *non sono troppo complesse*

Regola generale:

– ogni tabella e ogni colonna (o riga) al suo interno devono essere definite con chiarezza

– se sono utilizzate delle unità di misura, devono essere specificate

Frequenze assolute dei livelli di colesterolo sierico in 1067 soggetti della popolazione degli Stati uniti di età compresa tra 25 e 34 anni, 1976-1980

Livello di colesterolo (mg/100 ml)	Numero di soggetti
80-119	13
120-159	150
160-199	442
200-239	299
240-279	115
280-319	34
320-359	9
360-399	5
Totale	1.067

DISTRIBUZIONI DI FREQUENZA RELATIVA

E’ spesso utile conoscere la proporzione di dati che rientra in ogni classe anziché il *numero assoluto* di tali dati

- Si chiama **frequenza relativa** di ciascuna classe, il rapporto fra la *frequenza assoluta* della classe stessa ed il *numero totale* dei dati
- Nell’uso corrente la frequenza relativa è data generalmente come ***percentuale***

  (è quindi moltiplicato per 100 il rapporto tra il numero dei dati inclusi nella classe e il numero totale dei dati)
- La somma delle frequenze relative di tutte le classi è 1 (o 100, se le frequenze sono espresse in percentuale)

Una tabella nella quale sia rappresentato l'insieme delle classi, con l'indicazione della frequenza relativa di ciascuna di esse, è detta ***distribuzione di frequenza relativa***

Distribuzione di frequenza assoluta e relativa (in %) dell'altezza di 40 giovani piante (cfr. Tabella 3, p. 27)

Classe	X_i	60-79	80-99	100-19	120-39	140-59	160-79	180-99
Freq. Assoluta	n_i	1	3	10	12	7	5	2
Freq. Relativa %	f_i	2,5	7,5	25,0	30,0	17,5	12,5	5,0

- Le frequenze relative sono utili per confrontare serie di dati riferite a popolazioni con *diverso numero di individui*

Tabella 2.6 Frequenze assolute e relative dei livelli di colesterolo sierico in 2.294 soggetti della popolazione maschile degli Stati Uniti, 1976-1980

Livello di colesterolo (mg/100 ml)	Età 25-34		Età 55-64	
	Numero di soggetti	Frequenza relativa (%)	Numero di soggetti	Frequenza relativa (%)
80-119	13	1,2	5	0,4
120-159	150	14,1	48	3,9
160-199	442	41,4	265	21,6
200-239	299	28,0	458	37,3
240-279	115	10,8	281	22,9
280-319	34	3,2	128	10,4
320-359	9	0,8	35	2,9
360-399	5	0,5	7	0,6
Totale	1.067	100,0	1.227	100,0

DISTRIBUZIONI DI FREQUENZA CUMULATIVA

Per confrontare serie di dati sono utili anche le *frequenze relative cumulative*

Per ***frequenza relativa cumulativa*** di una classe si intende *la percentuale del numero totale dei dati che hanno un valore inferiore o uguale a quello della classe, o a quello che delimita superiormente la classe stessa se quest'ultima è espressa da un intervallo*

La *frequenza relativa cumulativa* di una classe è calcolata sommando la frequenza relativa della classe considerata alle frequenze relative delle classi che la precedono

Distribuzione di frequenza assoluta, relativa e cumulativa (in %) dell'altezza di 40 giovani piante (Tabella 3, p. 27)

Classe	X_i	60-79	80-99	100-19	120-39	140-59	160-79	180-99
Freq. Assoluta	n_i	1	3	10	12	7	5	2
Freq. Relativa %	f_i	2,5	7,5	25,0	30,0	17,5	12,5	5,0
Freq. Cumulata	---	2,5	10,0	35,0	65,0	82,5	95,0	100,0

Tabella 2.7 Frequenze relative e frequenze relative cumulative dei livelli di colesterolo sierico in 2.294 soggetti della popolazione maschile degli Stati Uniti, 1976-1980

Livello di colesterolo (mg/100 ml)	Età 25-34		Età 55-64	
	Frequenza relativa (%)	Frequenza relativa cumulativa (%)	Frequenza relativa (%)	Frequenza relativa cumulativa (%)
80-119	1,2	1,2	0,4	0,4
120-159	14,1	15,3	3,9	4,3
160-199	41,4	56,7	21,6	25,9
200-239	28,0	84,7	37,3	63,2
240-279	10,8	95,5	22,9	86,1
280-319	3,2	98,7	10,4	96,5
320-359	0,8	99,5	2,9	99,4
360-399	0,5	100,0	0,6	100,0

GRAFICI

- I dati possono essere sintetizzati ed illustrati anche attraverso l'uso di ***grafici***, cioè rappresentazione figurate di dati che derivano da *tabelle di distribuzione di frequenza*
- I grafici devono essere realizzati in modo tale da comunicare immediatamente l'andamento generale di una serie di dati
- Come le tabelle, anche i grafici devono essere *definiti con chiarezza* ed è necessario *indicare le unità di misura*, se sono utilizzate

Diagramma a barre rettangolari (a rettangoli distanziati)

Il ***diagramma a barre rettangolari*** è un tipo comune di grafico utilizzato per illustrare una distribuzione di frequenza per *dati qualitativi*:

- in un diagramma a barre le diverse modalità del carattere sono presentate lungo un *asse orizzontale*, con *segmenti di uguale ampiezza, successivi, ma distanziati tra loro*

- al di sopra di ogni modalità è tracciata una *barra rettangolare verticale* avente l'*altezza* uguale alla frequenza assoluta o relativa di tale modalità

- le *barre* devono avere tutte la *stessa ampiezza* ed essere *separate* l'una dall'altra per non implicare alcuna continuità

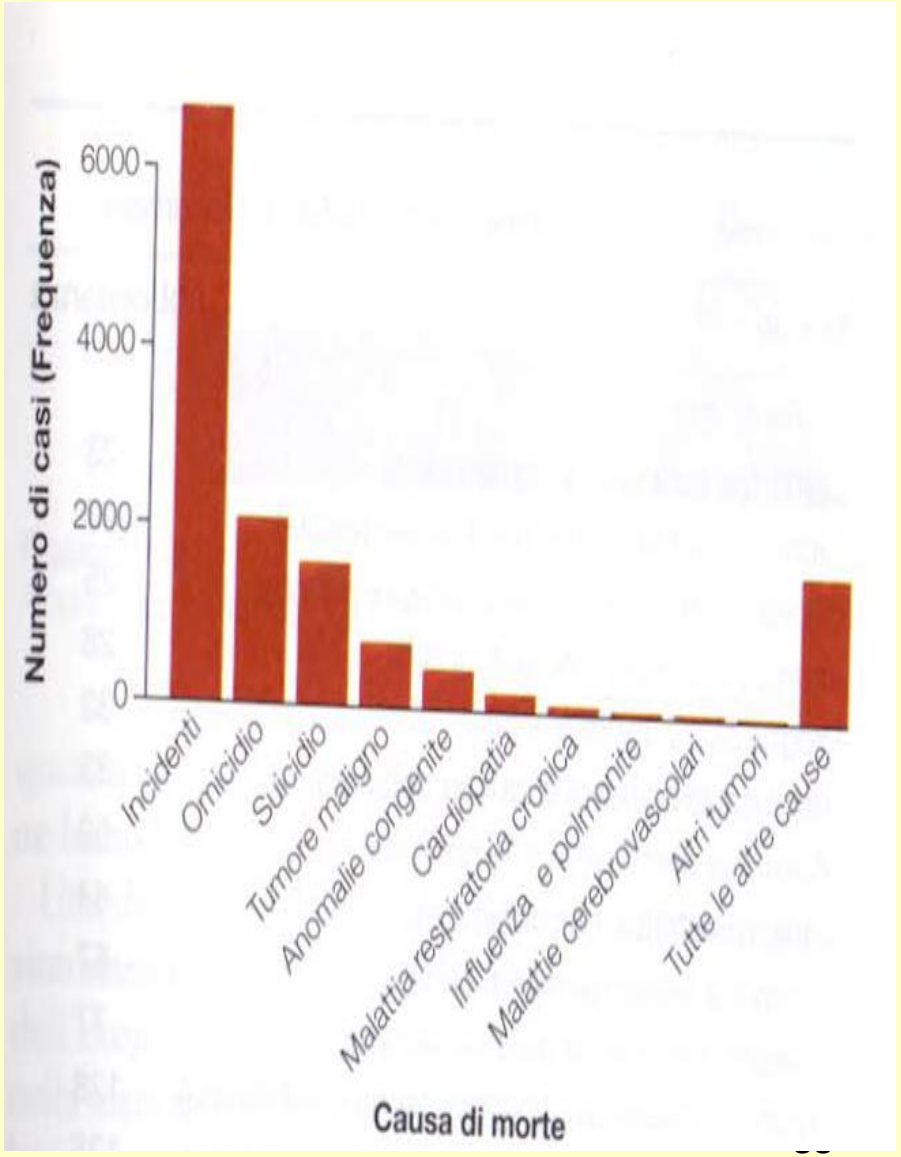
Distribuzione di frequenza assoluta

Tabella 2.1-1

Tabella di frequenza che mostra le 10 cause di morte più comuni degli adolescenti statunitensi di età compresa tra 15 e 19 anni nel 1999.

Causa di morte	Frequenza
Incidenti	6688
Omicidio	2093
Suicidio	1615
Tumore maligno	745
Cardiopatia	463
Anomalie congenite	222
Malattia respiratoria cronica	107
Influenza e polmonite	73
Malattie cerebrovascolari	6
Altri tumori	52
Tutte le altre cause	1653
Totale	**13778**

Diagramma a barre della distribuzione



Istogrammi

L'***istogramma*** *fornisce una rappresentazione grafica di una distribuzione di frequenza* per ***dati numerici discreti o continui***

Si ottiene da una *distribuzione di frequenza* (assoluta o relativa), procedendo nel modo seguente:

- *si fissa nel piano un riferimento cartesiano*
- si rappresentano sull'asse delle ascisse le diverse classi di distribuzione (singoli valori numerici o intervalli)
- *si riporta sull'asse delle ordinate (che deve iniziare sempre da 0) la frequenza assoluta o relativa di ogni classe*
- per ogni classe si *costruisce un rettangolo* con **area** *proporzionale alla frequenza della classe; se le classi hanno tutte la stessa ampiezza, le altezze dei singoli rettangoli indicano le frequenze delle classi corrispondenti (in tale caso, infatti, l'area è ovviamente proporzionale alla frequenza)*

Distribuzione di frequenza assoluta dell'altezza di 40 giovani piante

Classe	X_i	60-79	80-99	100-19	120-39	140-59	160-79	180-99
Freq. Assoluta	n_i	1	3	10	12	7	5	2

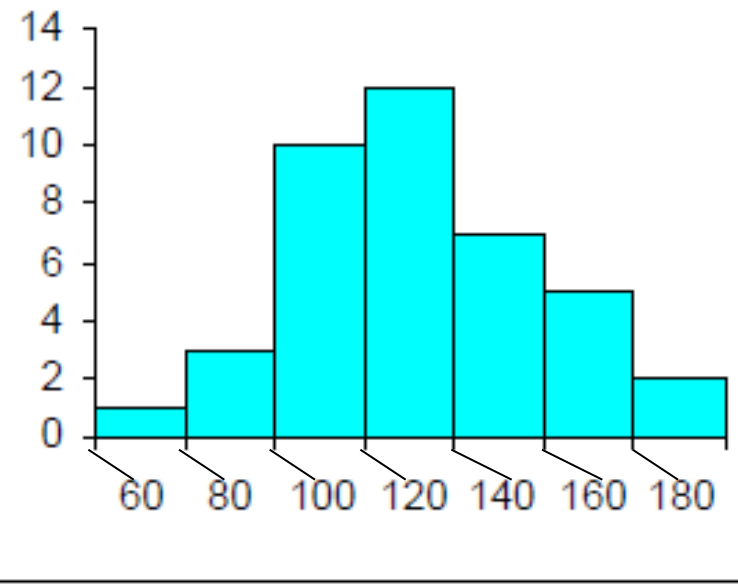


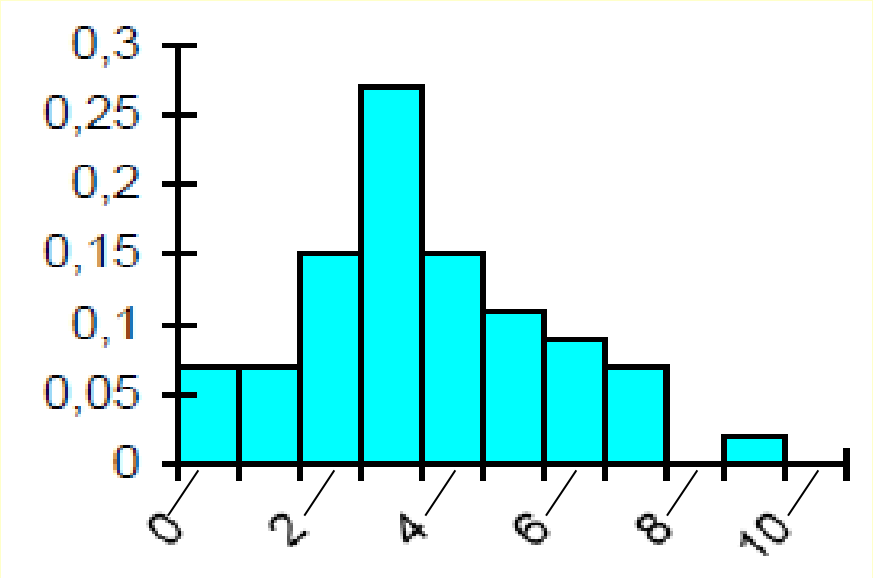
Figura 2. Istogramma dei dati di Tab. 4

(Valore iniz. =60; Valore finale =199; Passo =20; Classi=7)

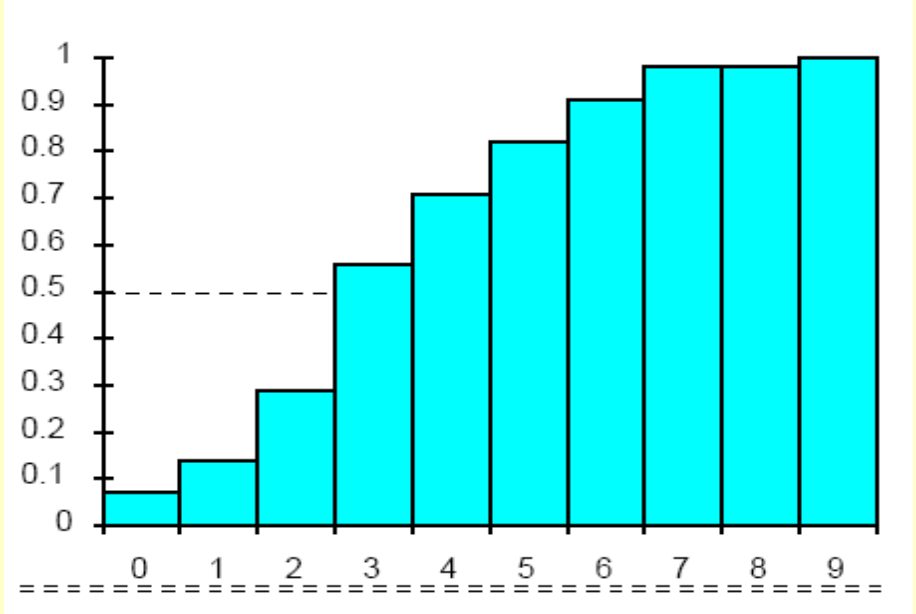
Distribuzioni di frequenza assoluta, relativa e cumulativa delle foglie in 45 rami

Classe	x	0	1	2	3	4	5	6	7	8	9
Freq. Assoluta	n	3	3	7	12	7	5	4	3	0	1
Freq. Relativa	f	0,07	0,07	0,15	0,27	0,15	0,11	0,09	0,07	0,00	0,02
Freq. Cumulata	---	0,07	0,14	0,29	0,56	0,71	0,82	0,91	0,98	0,98	1,00

Istogramma frequenza relativa



Istogramma frequenza cumulativa



Se le classi hanno tutte la stessa ampiezza:

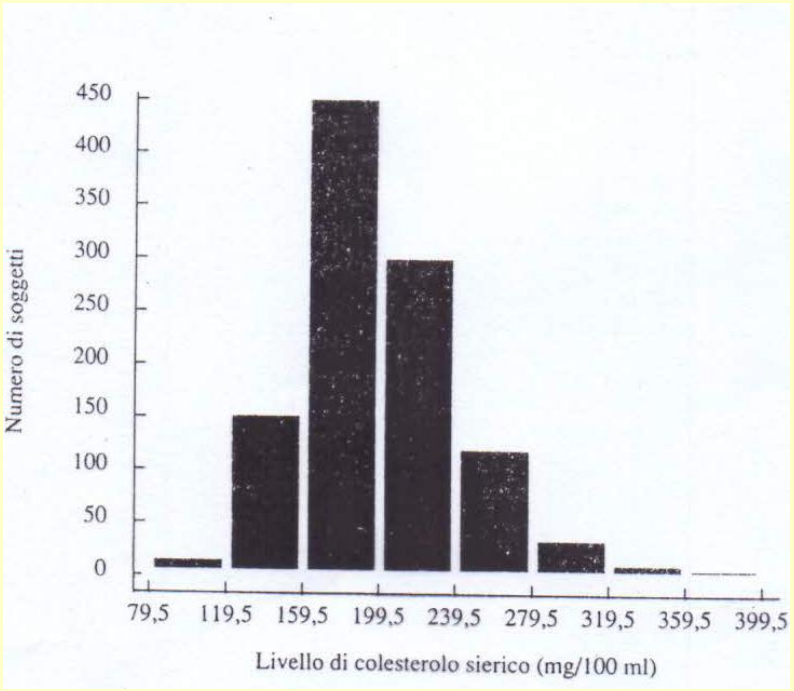
un istogramma che illustra le frequenze relative avrà la stessa forma di un istogramma che illustra le frequenze assolute

Se l'istogramma si riferisce ad una distribuzione di frequenza relativa, l'area dell'istogramma si può assumere uguale ad 1 *(poiché è 1 la somma delle frequenze relative di una tale distribuzione)*

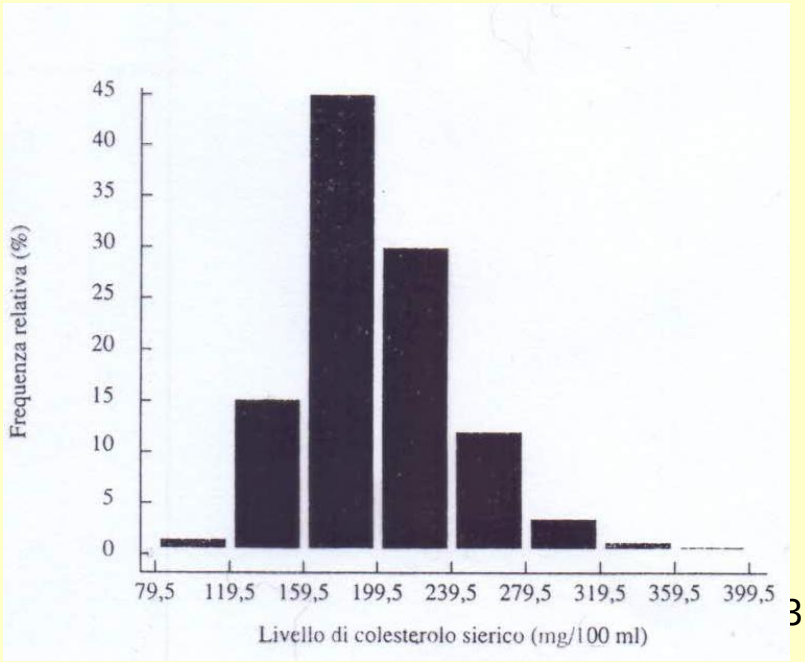
Frequenze assolute dei livelli di colesterolo sierico in 1067 soggetti della popolazione degli Stati uniti di età compresa tra 25 e 34 anni, 1976-1980

Livello di colesterolo (mg/100 ml)	Numero di soggetti
80-119	13
120-159	150
160-199	442
200-239	299
240-279	115
280-319	34
320-359	9
360-399	5
Totale	1.067

Istogramma frequenza assoluta



Istogramma frequenza relativa



Nota

Poiché l'area di ciascun rettangolo rappresenta la proporzione dei dati in un certo intervallo, occorre prestare molta attenzione nella *costruzione di un istogramma con ampiezze diverse degli intervalli:*

l'altezza deve variare con l'ampiezza in modo che l'area di ciascun rettangolo abbia la giusta proporzione

Poligoni di frequenza

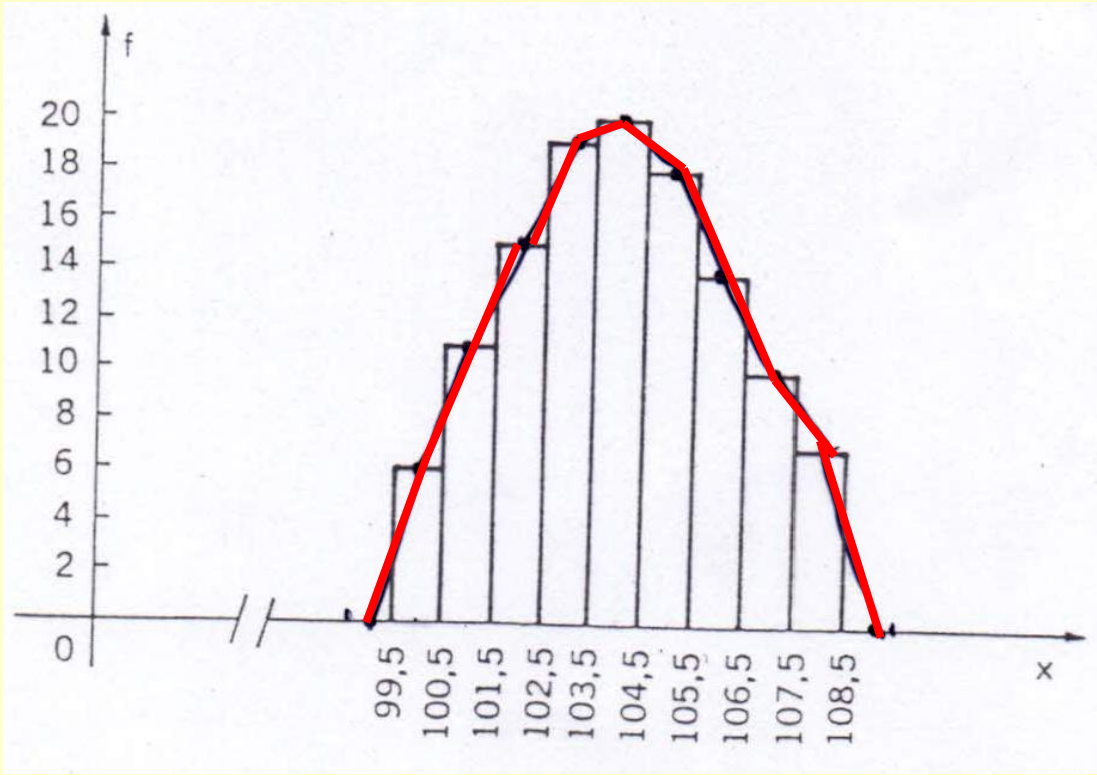
Un ***poligono di frequenza*** è un altro grafico comunemente utilizzato che si costruisce a partire da un istogramma:

- *per ogni classe si considerano il valore centrale della classe e la frequenza associata alla classe stessa*
- *si individuano nel piano i punti aventi come ascissa ed ordinata, rispettivamente, il valore centrale e la frequenza di ogni classe*
- *si aggiungono ai punti precedenti anche i due punti che hanno come ascisse, rispettivamente, i valori centrali delle due classi (vuote di dati) che immediatamente precedono e seguono le classi che contengono i dati, e come ordinate lo zero*
- *si uniscono con tratti di retta i punti individuati*

Dall'istogramma al poligono di frequenza

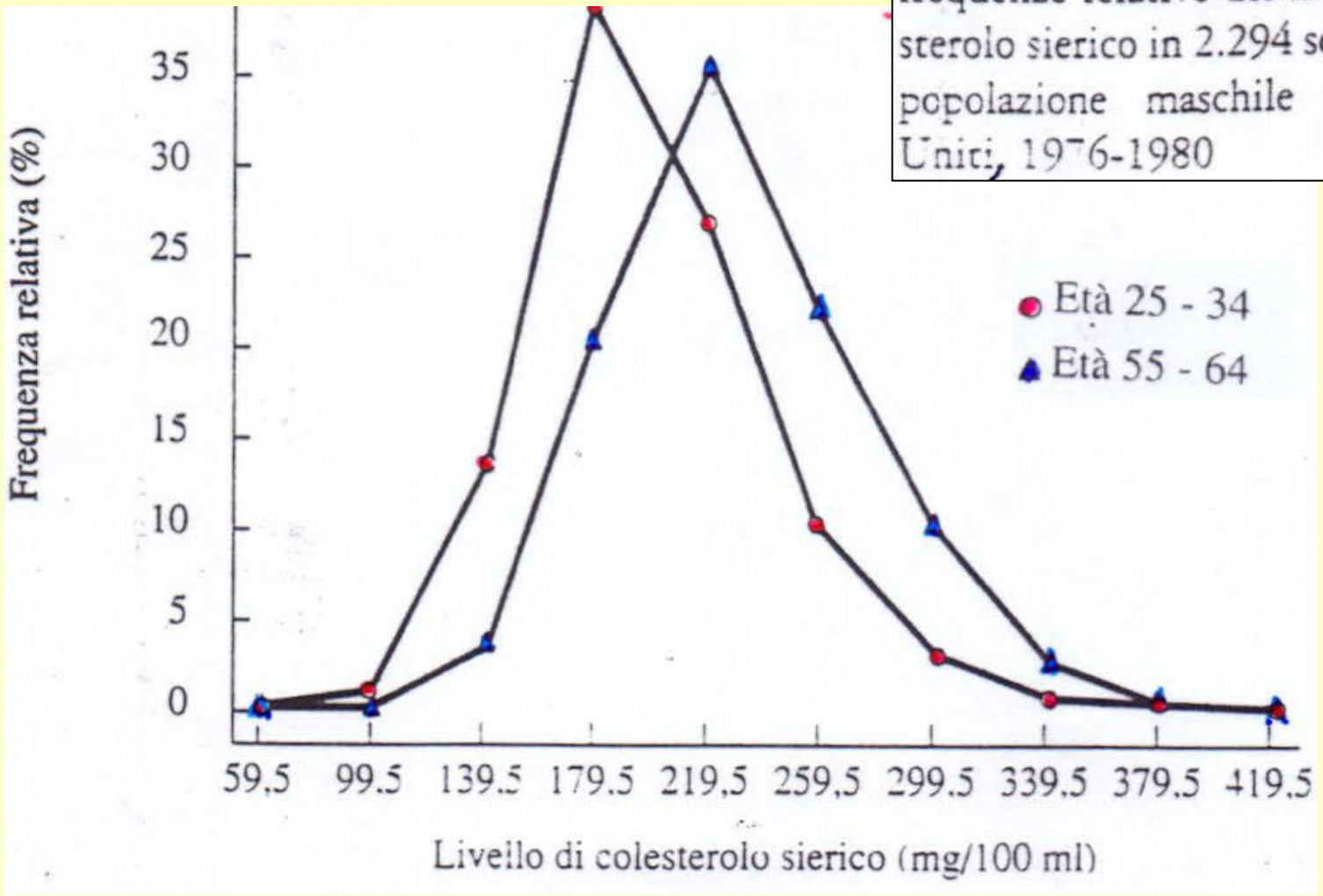
Tabella 1

Limiti classi	*Valore centrale*	*Frequenza*
99,5 — 100,5	100	6
100,5 — 101,5	101	11
101,5 — 102,5	102	15
102,5 — 103,5	103	19
103,5 — 104,5	104	20
104,5 — 105,5	105	18
105,5 — 106,5	106	14
106,5 — 107,5	107	10
107,5 — 108,5	108	7
		120



I poligoni di frequenza sono più idonei degli istogrammi per confrontare serie di dati poiché *si possono facilmente sovrapporre*

Figura 2.5. Poligono di frequenza: frequenze relative dei livelli di colesterolo sierico in 2.294 soggetti della popolazione maschile degli Stati Uniti, 1976-1980



Poligono di frequenza cumulativa

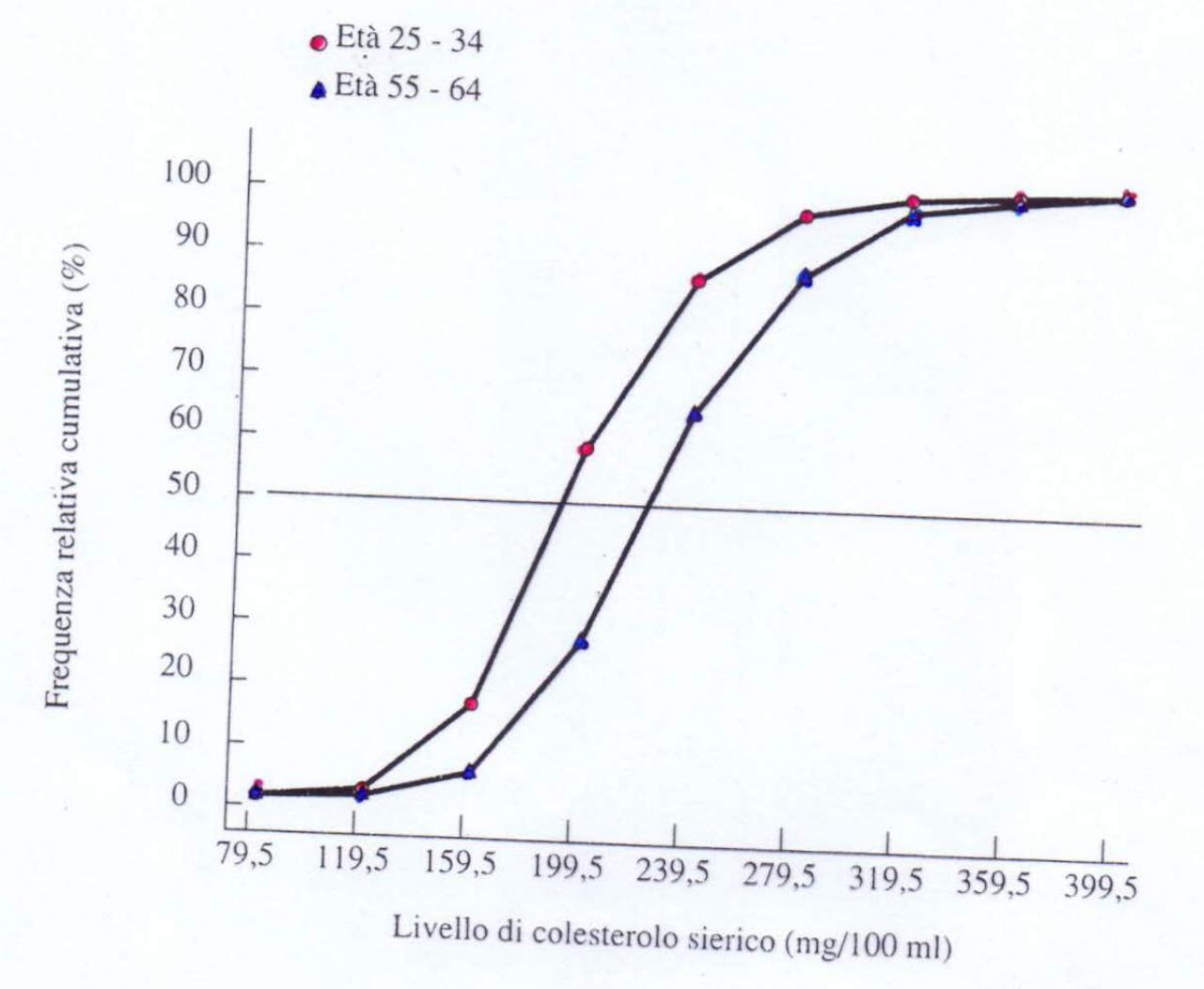
Per costruire un ***poligono di frequenza cumulativa*** si procede così:

- *si usa lo stesso asse orizzontale di un poligono di frequenza standard, ma sull'asse verticale si riporta la frequenza relativa cumulativa*

- si considerano i punti del piano che hanno come ***coordinate*** rispettivamente il limite superiore di ciascun intervallo e la frequenza relativa cumulativa associata allo stesso intervallo

- *si uniscono tra loro con segmenti i punti così ottenuti*

I poligoni di frequenza cumulativa possono essere utilizzati per confrontare dati, come mostrato nella figura seguente

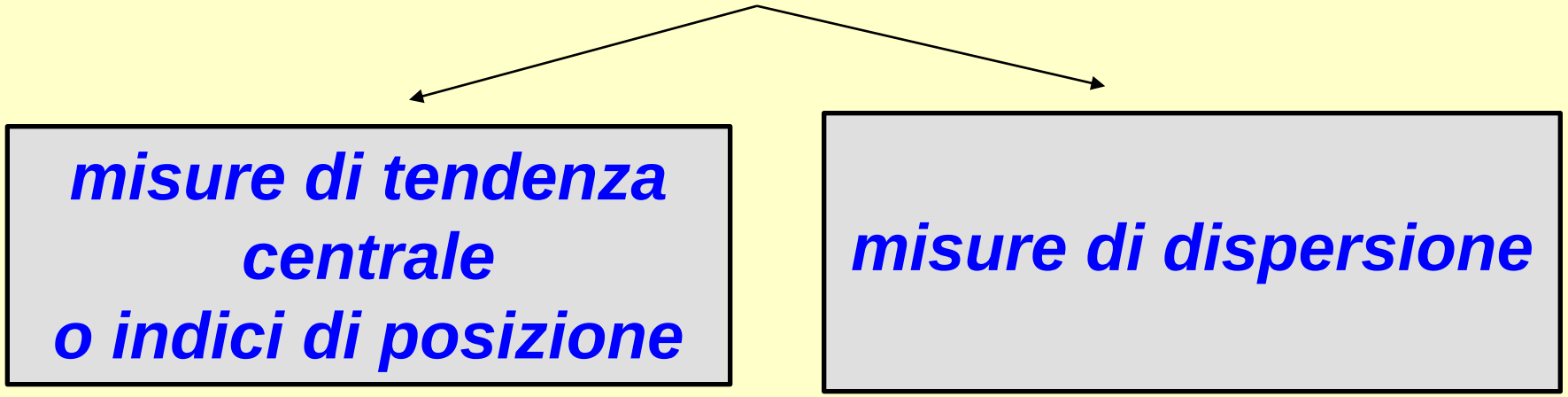


MISURE DI SINTESI NUMERICA

Le tabelle e i grafici non consentono di formulare *affermazioni sintetiche di tipo quantitativo* che caratterizzino una distribuzione nel suo insieme

Si utilizzano a tale scopo le

misure di sintesi numerica



Misure di tendenza centrale (o indici di posizione)

- Se un rilevamento statistico riguarda un *carattere quantitativo*, si pone il problema di determinare un *particolare valore* che possa essere ritenuto *tipico per la modalità del carattere* considerato

- Tale valore è detto anche ***indice di posizione*** poiché rappresenta una particolare "posizione" all'interno dell'insieme dei dati ordinati secondo l'ordine di grandezza

- Si usa anche l'espressione ***misura di tendenza centrale*** perché tale valore individua il centro dell'insieme di dati o il punto intorno a cui tendono a raccogliersi i dati

Misure di tendenza centrale:
media, mediana e moda

La scelta di uno di questi valori come rappresentante della modalità del carattere in esame dipenderà, in generale, dalla natura dei dati e dall'utilizzazione che se ne vuol fare

Media aritmetica

Si definisce ***media aritmetica***, o più semplicemente ***media*** o ***valor medio***, di un insieme di *n* dati numerici x_1 , x_2 ,…, x_n il numero che si ottiene dividendo per *n* la somma di tutti i dati

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

[per indicare la media, quando ci si riferisce all'intera popolazione, si usa il simbolo μ (mi)]

Vale la seguente proprietà

La somma algebrica degli scarti di ogni singolo dato dalla media è zero:

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

(per scarto di un dato dalla media si intende la differenza tra tale dato e la media)

Di più, è facile provare che:

"La media è l'unico valore la cui somma degli scarti da tale valore è nulla"

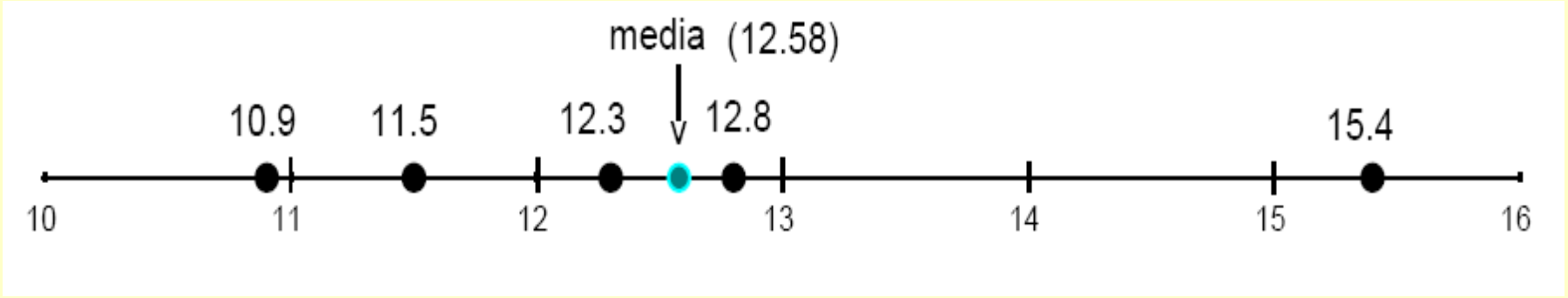
Per dimostrare graficamente che **la media aritmetica corrisponde al punto di bilanciamento o di equilibrio dei dati,** si supponga di avere 5 misure: 10,9 11,5 12,3 12,8 15,4.

La loro media

$$\overline{X} = \frac{10,9+11,5+12,3+12,8+15,4}{5} = 12,58$$

è uguale a 12,58.

La rappresentazione grafica mostra visivamente come la somma delle distanze dalla media dei valori collocati prima di essa sia uguale alla somma delle distanze dalla media dei valori collocati dopo



- Quando alcuni dei dati x_1 , x_2 ,…, x_n coincidono, indicati con x'_1, x'_2 ,…, x'_r i valori distinti tra gli x_i e con m_1, m_2 ,…, m_r, rispettivamente, le loro frequenze assolute, allora la media aritmetica è anche espressa da:

$$\bar{x} = \frac{\sum_{i=1}^{r} m_i x'_i}{\sum_{i=1}^{r} m_i} = \frac{\sum_{i=1}^{r} m_i x'_i}{n} = \sum_{i=1}^{r} \frac{m_i}{n} x'_i$$

- In questo caso si usa dire che la media aritmetica è la ***media ponderata* dei numeri x'_1, x'_2,…,x'_r** *(ogni dato è considerato con il suo **peso**, cioè moltiplicato per la sua frequenza assoluta)*

- Dalla formula precedente segue anche che la media aritmetica è data dalla *somma dei prodotti degli x'_i per le rispettive frequenze relative*

Calcolo della media per dati raggruppati

Per il calcolo della *media* nel caso di *dati raggruppati* si procede così:

- nella tabella di distribuzione di frequenza, si sostituiscono i dati di ogni classe con un ugual numero di dati coincidenti con il valore centrale della classe stessa *(media aritmetica dei valori estremi della classe)*

- si calcola poi la media di tali valori *(come media ponderata)*

Esempio

Tabella 1

Limiti classi	*Valore centrale*	*Frequenza*
99,5 ⊢ 100,5	100	6
100,5 ⊢ 101,5	101	11
101,5 ⊢ 102,5	102	15
102,5 ⊢ 103,5	103	19
103,5 ⊢ 104,5	104	20
104,5 ⊢ 105,5	105	18
105,5 ⊢ 106,5	106	14
106,5 ⊢ 107,5	107	10
107,5 ⊢ 108,5	108	7
		120

$$\bar{x} = \frac{100 \cdot 6 + 101 \cdot 11 + 102 \cdot 15 + \cdots + 108 \cdot 7}{120} \cong 103{,}98$$

Il risultato che segue è molto intuitivo

Siano $x_1, x_2, \ldots, x_n$ numeri e y un ulteriore numero.

Allora detta $\bar{x}$ la media aritmetica degli x_i e

$\bar{x}'$ la media aritmetica di $x_1, x_2, \ldots, x_n, y$ si ha :

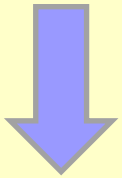
se $y < \bar{x}$ allora $\bar{x}' < \bar{x}$ e viceversa

se $y = \bar{x}$ allora $\bar{x}' = \bar{x}$ e viceversa

se $y > \bar{x}$ allora $\bar{x}' > \bar{x}$ e viceversa

(se aggiungiamo un numero ad un insieme di numeri, allora la media diminuirà, o rimarrà costante, o aumenterà a seconda che il numero che si aggiunge sia minore, uguale o maggiore della media preesistente)

- *La media aritmetica è il più comunemente usato tra gli indici di posizione*
- *Per come è definita, però, la media è estremamente sensibile a valori "insoliti", cioè a valori eccezionalmente grandi o eccezionalmente piccoli*



potrebbe essere necessario preferire una misura di sintesi che non sia così sensibile ad ogni singolo dato

Media geometrica

Dati n numeri reali positivi x_1 , x_2 ,…, x_n *non necessariamente distinti, la loro* ***media geometrica*** *è data da:*

$$m_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

• Per calcolare m_g si fa ricorso generalmente ai logaritmi (la scelta della base è arbitraria)

• Dalle proprietà dei logaritmi si ricava che:

il logaritmo della media geometrica è uguale alla media aritmetica dei logaritmi dei dati

• Determinato il logaritmo della media geometrica, si risale poi alla media geometrica stessa

Esempio

Determinare la media geometrica dei numeri:

3, 2, 4, 3, 5, 3, 6, 3, 7, 9, 3

Svolgimento

Dati i numeri:

3, 2, 4, 3, 5, 3, 6, 3, 7, 9, 3

si ha: $$m_g = \sqrt[11]{3^5 \cdot 2 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 9}$$

Passando ai logaritmi decimali (ma si può scegliere anche una base diversa da 10), si ottiene:

$$Log\, m_g = \frac{5Log3 + Log2 + Log4 + Log5 + Log6 + Log7 + Log9}{11} = \frac{6,5651}{11} \approx 0,5968$$

da cui si ricava:

$$m_g = 10^{0,5968} \approx 3.95$$

Mediana

- Dato un insieme di dati, non necessariamente distinti, x_1 , x_2 ,…, x_n, ma ordinati in ordine crescente o decrescente, la ***mediana*** è quel numero che divide tale sequenza *in due parti di uguale numerosità*:

  il 50% dei dati è maggiore o uguale al valore della mediana
  il 50% dei dati è minore o uguale al valore della mediana

  Se il numero dei dati è dispari, la mediana è il *valore centrale*

  es.: 1 2 **4** 6 9 **me = 4**

  Se il numero di dati è pari, la mediana è *la media dei due valori centrali*

  es.: 1 1 **1** **2** 5 8 **me = 1,5**

Definizione di mediana

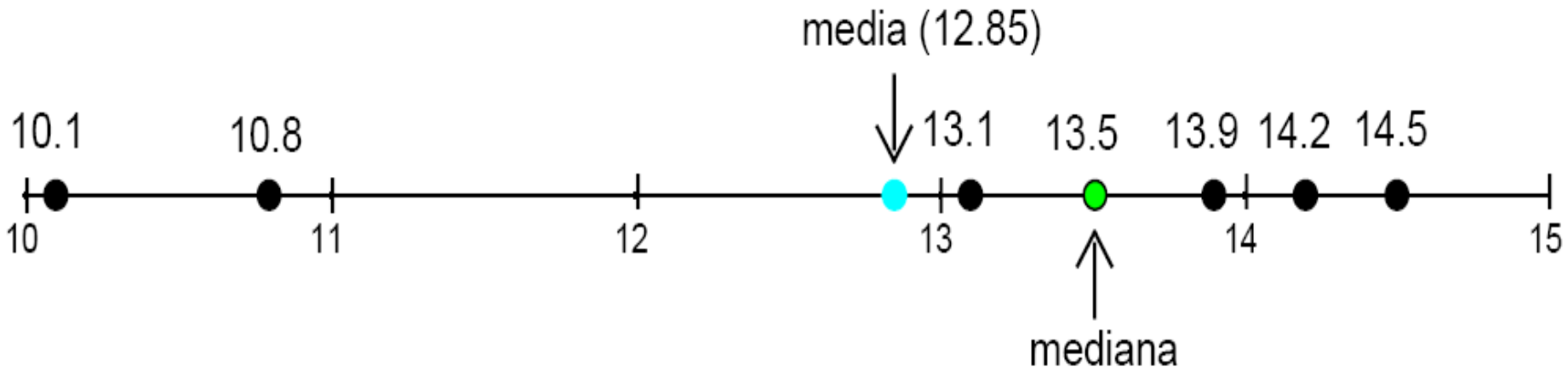
$$m_e = \begin{cases} x_{\frac{n+1}{2}} & \text{se n è dispari} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{se n è pari} \end{cases}$$

La mediana è una misura di tendenza centrale molto meno sensibile a valori "insoliti" rispetto alla media

Per meglio comprendere le differenze tra media aritmetica e mediana, con la stessa serie di 6 dati (10,1 10,8 13,1 13,9 14,2 14,5) in cui

- la media è 12,85 e
- la mediana 13,5

la rappresentazione grafica evidenzia come la media sia il baricentro della distribuzione e la mediana sia collocata tra i valori più addensati.



Moda

- Un'ulteriore misura di tendenza centrale è la moda che può essere usata come misura di sintesi per tutti i tipi di dati *(quindi anche qualitativi)*
- *La **moda** di un insieme di dati è la modalità del carattere che si verifica con maggiore frequenza*
- Nel caso di dati numerici raggruppati, si parla di ***classe modale*** riferendosi alla classe con il maggior numero di dati

- Non è escluso che esistano due o più classi ugualmente numerose, nel qual caso si parla di ***classi modali***

- Ovviamente, il ricorso alla moda ha senso se la concentrazione delle frequenze nella classe modale è abbastanza pronunciata

Esempi

a) *L'insieme costituito dai seguenti dati:*

10, 11, 6, 5, 10, 6, 14, 10, 6, 11, 12

ha due mode: 6 e 10, e viene detto bimodale

b) *L'insieme:*

3, 2, 4, 3, 5, 3, 6, 3, 7, 9, 3

ha moda 3 e viene detto unimodale

c) *L'insieme dei valori:*

5, 6, 7, 9, 3, 10, 4, 11

non ha moda

In definitiva, per la moda si può dire che:

- Proprietà:
  - non è influenzata dalla presenza di alcun valore estremo
  - differisce quando con gli stessi dati si formano classi di ampiezza differente

- Si usa :
  - solo a scopi descrittivi, essendo più variabile delle altre misure di tendenza centrale

Come regolarsi?

- *Si preferisce usare la media aritmetica*:

  in tutte quelle situazioni dove i dati in esame risultino distribuiti in modo abbastanza simmetrico a sinistra e a destra della media aritmetica, con un addensamento dei dati stessi in prossimità di tale valore

- *Si preferisce usare la media geometrica*:

  nelle situazioni in cui i logaritmi dei numeri dati si distribuiscono in modo abbastanza simmetrico a sinistra e a destra del logaritmo di m_g

- *Si preferisce usare la mediana*:

  in quelle situazioni in cui non interessano tanto i valori numerici in gioco, quanto piuttosto il loro ordinamento

- *Si usa la moda o classe modale*:

- *nei casi in cui non sarebbe possibile usare altri tipi di medie, perché i dati raccolti sono di tipo qualitativo*

- *nei casi numerici, quando interessa considerare un campione dal comportamento "tipico" o "normale", identificabile con il valore più frequente, non influenzato dall'eventuale presenza di elementi anomali*

Misure di dispersione (o indici di dispersione)

- La sola conoscenza di un indice di posizione *(sia esso media, mediana o moda)* non è sufficiente per descrivere in che modo i dati di partenza sono *distribuiti* intorno a quel valore *(cioè per descrivere la variabilità nell'insieme dei dati)*

- Per esempio, una stessa media aritmetica può derivare da insieme di dati molti dissimili fra loro: *piuttosto "addensati" vicini al valore medio o assai più "dispersi" rispetto a tale valore*

Esempio

Si scopre che Marte è abitato da una specie intelligente simile alla nostra. Misurate le altezze di sette marziani adulti si trova che media, moda e mediana coincidono e valgono 170 cm. Esaminiamo qualche possibile sequenza di dati che soddisfa tali condizioni. Cosa si può dedurre?

(a) 167, 169, 170, 170, 170, 172, 172
La variabilità è piccola e pare che l'altezza dei marziani sia quasi costante.

(b) 161, 163, 170, 170, 173, 175, 178
La variabilità riscontrata è vicina a quella della statura umana.

(c) 80, 100, 120, 170, 170, 250, 300
La variabilità è notevole: su Marte vi sono nani e giganti.

Si ha conferma del fatto che la sola conoscenza dell'indice di posizione non è sufficiente a caratterizzare una distribuzione di dati

- Un insieme di dati deve essere caratterizzato, quindi, oltre che da un valore medio, anche da un valore numerico che ne esprima la *maggiore o minore dispersione*, cioè il *maggiore o minore addensarsi dei dati stessi attorno al valore medio stesso*

- Questi ulteriori indicatori numerici sono detti **indici di dispersione** e misurano appunto il grado di dispersione dell'insieme di dati

Intervallo di variazione

- L'***intervallo di variazione*** (o ***range***) di un insieme di dati x_1 , x_2 ,…, x_n è espresso dalla differenza fra il valore massimo ed il valore minimo dell'insieme stesso:

$$I_V = x_{max} - x_{min}$$

- L'intervallo di variazione è il più semplice indicatore di dispersione e quello più facilmente determinabile

Caratteristiche dell'intervallo di variazione

- è *puramente descrittivo* e tiene conto solo di due dati di un'intera serie
- *non dà informazioni* su come i dati sono distribuiti tra il valore minimo e quello massimo
- *dipende* sia dal *numero di dati rilevati* che dalla *presenza di un solo valore molto diverso dagli altri*

E’ giustificata allora l’introduzione di altri indici di dispersione, meno influenzati dai valori estremi:

- varianza

- deviazione standard

- distanza interquartile

Varianza

Si definisce ***varianza*** di un insieme di dati

x_1 , x_2 ,…, x_n , in simboli *Var*, la media aritmetica del quadrato degli scarti:

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

Si ricorre alla varianza soprattutto quando non è essenziale esprimere l'indice di dispersione nelle stesse unità di misura in cui sono espressi i valori x_i

- Il principale inconveniente della varianza è infatti proprio la sua *disomogeneità dal punto di vista "dimensionale"* con l'insieme dei dati di partenza

- Si rende allora necessaria un'ulteriore modifica che consiste nell'annullare l'effetto degli elevamenti al quadrato mediante un'estrazione di radice quadrata

Deviazione standard o scarto quadratico medio

- Siano x_1 , x_2 ,…, x_n dati numerici rilevati su una popolazione. Si chiama ***deviazione standard*** o ***scarto quadratico medio*** di tali dati (in inglese, *standard deviation*), e si denota abitualmente con σ il numero reale che esprime la radice quadrata della varianza:

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

Tenuto conto del legame tra varianza e scarto quadratico medio, si scrive spesso σ^2 in luogo di Var

La somma dei quadrati degli scarti dalla media prende il nome di ***devianza***:

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

Nota

Nel caso di **dati raggruppati**, per il calcolo di varianza e deviazione standard, si procede come per il calcolo della media nell'analoga situazione *(si considera che gli elementi di ogni classi siano tutti coincidenti con il valore centrale della classe e che siano tanti quanto la frequenza assoluta della classe)*

Se i dati x_i , si riferiscono ad un ***campione*** e non all'intera popolazione, si parla di:

- ***deviazione standard stimata*** o di ***scarto quadratico medio stimato***

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

- ***varianza stimata***

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

OSSERVAZIONE

La varianza è anche uguale alle differenza tra la media dei quadrati degli x_i e il quadrato della media degli x_i:

$$Var = media\,(x_i^2) - \bar{x}^2$$

Infatti:

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

Distanza interquartile

- Considerati n dati numerici ordinati in ordine crescente, si fissa l'attenzione sulla mediana che permette di suddividere l'insieme dei dati in due parti ugualmente numerose

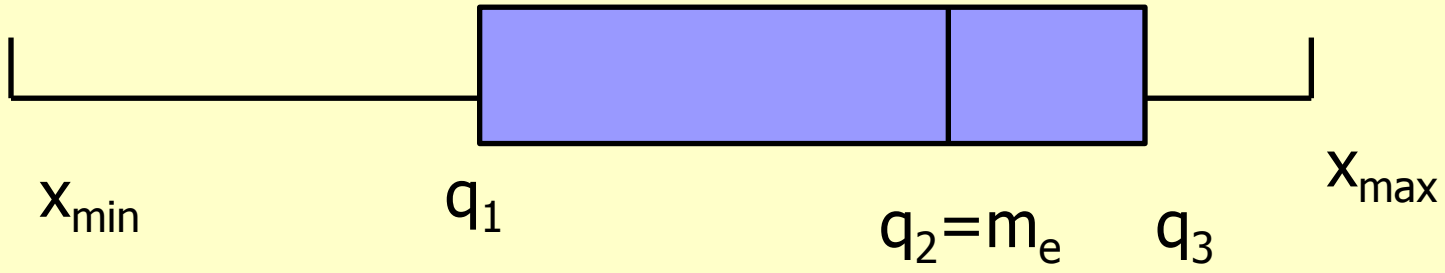
- Il passo successivo è quello di suddividere lo stesso insieme ordinato in quattro parti ugualmente numerose, considerando la mediana q_1 del primo sottoinsieme e la mediana q_3 del secondo sottoinsieme

- Si definisce ***distanza interquartile*** il numero:

$$\Delta = q_3 - q_1$$

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

Graficamente la dispersione di una serie di dati può essere visualizzata efficacemente con uno schema come quello in figura:



- *il rettangolo centrale racchiude i dati compresi tra il primo ed il terzo quartile*
- *il segmento sulla sinistra rappresenta l'intervallo entro cui variano i dati inferiori al primo quartile*
- *il segmento sulla destra rappresenta l'intervallo entro cui variano i dati superiori al terzo quartile*

Disuguaglianza di Chebychev

La ***disuguaglianza di Chebychev*** si applica alla media e alla deviazione standard di *qualsiasi distribuzione*, indipendentemente dalla forma di quest'ultima, e consente di descrivere l'intero insieme di dati attraverso questi due soli numeri:

"Per qualunque numero k > 1, almeno $[1-(1/k)^2]$ degli elementi di un insieme di dati è all'interno dell'intervallo centrato sulla media ed avente raggio uguale a k volte la deviazione standard"

Esempio

- *Se k = 2, almeno $1-(1/2)^2 = 1 - (1/4) = 3/4$*
  dei valori rientrano all'interno di due deviazioni standard dalla media, ovvero l'intervallo di estremi $\bar{x} \pm 2s$
  comprende almeno il 75% dei dati

- *Se k =3, almeno $1-(1/3)^2 = 1 - (1/9) = 8/9$*
  delle osservazioni rientrano all'interno di tre deviazioni standard dalla media, ovvero l'intervallo di estremi $\bar{x} \pm 3s$
  contiene almeno l'88,9% dei dati

DISTRIBUZIONI A DUE CARATTERI

Le situazioni studiate finora si riferiscono sempre ad un'*unica caratteristica* della popolazione in esame

Spesso interessa invece considerare simultaneamente *due caratteristiche* quantitative degli individui di una stessa popolazione (ovvero due variabili), per stabilire se esiste una qualche relazione tra l'una e l'altra

Capita di frequente che le due variabili siano effettivamente *correlate* fra loro, nel senso che i loro valori mostrano di essere in qualche modo legati l'uno all'altro

Quando vi è correlazione tra due variabili, l'obiettivo è quello di trovare una qualche *legge esprimibile in forma matematica*, capace di definire esattamente questi legami

Il problema presenta due *aspetti* distinti:

- verificare se fra le due variabili esiste un grado di associazione dei loro valori che sia significativo: questo aspetto porta al calcolo del ***coefficiente di correlazione***

- cercare di esprimere in forma matematica la relazione che esiste tra le variabili, ovvero, trovare la ***legge di regressione*** di una variabile sull'altra

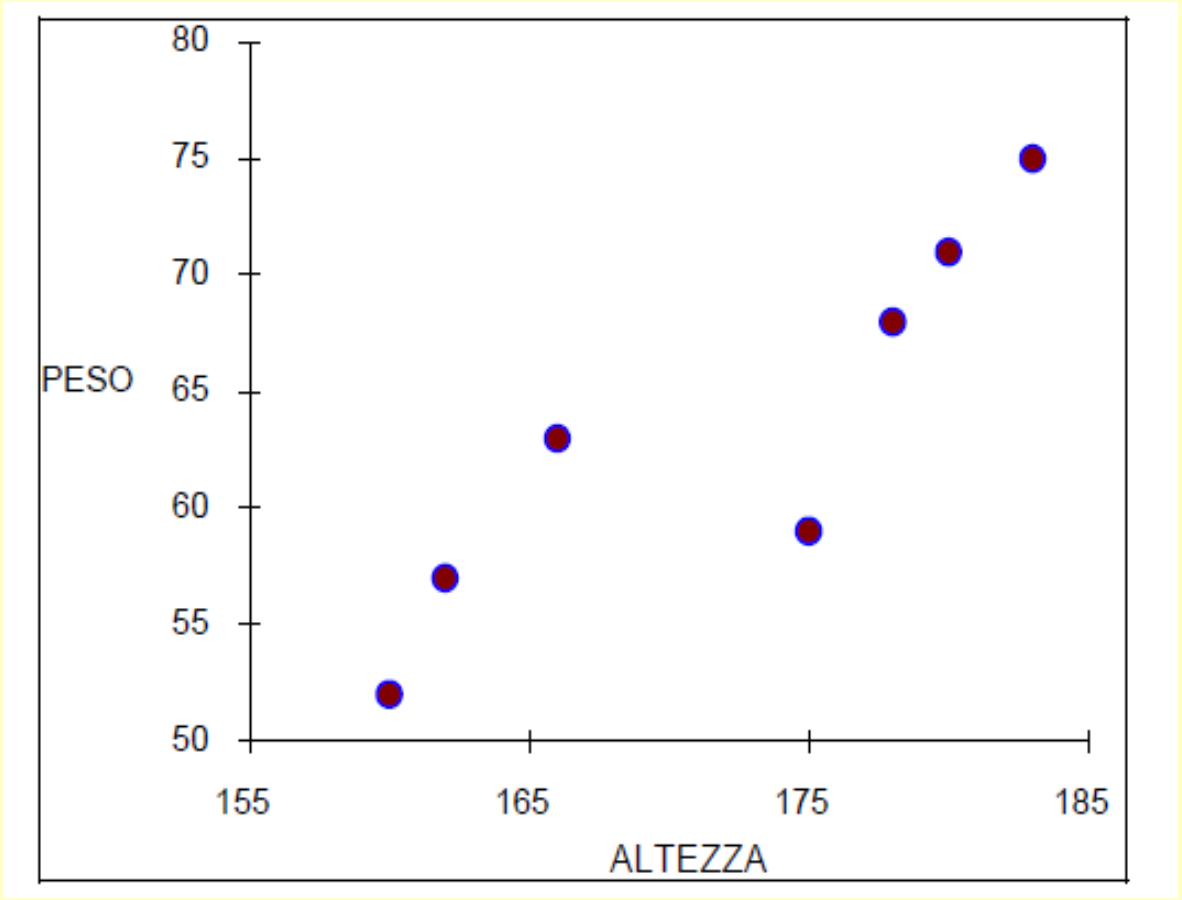
Per *studiare la correlazione* tra due variabili x e y si procede così:

- il primo passo è la raccolta di dati che mostrino valori corrispondenti delle variabili considerate
- poi si numerano da 1 ad n gli individui della popolazione e si associa a ciascuno di essi una coppia ordinata di numeri che esprimono i valori che le variabili x e y assumono in relazione all'individuo considerato *(la coppia ordinata associata all'i-esimo individuo sarà indicata con* (x_i, y_i)*)*
- il passo successivo ancora è di riportare ogni coppia (x_i, y_i) in un sistema di coordinate cartesiane del piano: il risultato è un insieme di punti detto ***diagramma a dispersione***
- dal diagramma a dispersione è spesso possibile stabilire il tipo di curva che si "adatta" meglio all'andamento dei punti sperimentali, e quindi che meglio approssima i dati, che si dirà ***curva interpolante***

Esempio

Con le misure di peso (in Kg) e di altezza (in cm) di 7 giovani, come riportato in tabella, è possibile costruire il diagramma a dispersione*: si evidenzia una correlazione di tipo lineare*

Individui	1	2	3	4	5	6	7
Peso (Y)	52	68	75	71	63	59	57
Altezza (X)	160	178	183	180	166	175	162



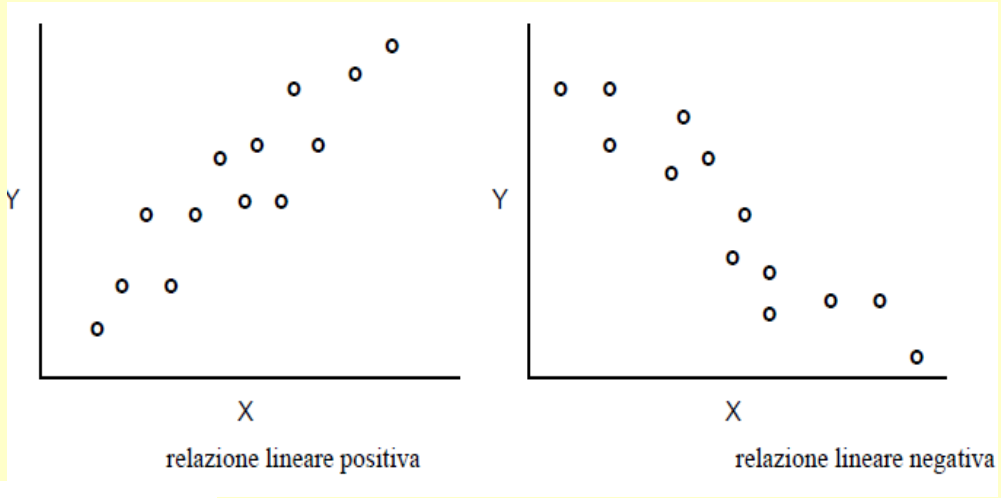
- Se i dati possono essere ben interpolati da una retta, la ***correlazione*** tra le variabili è detta ***lineare*** e, per gli scopi della regressione, è appropriata un'*equazione lineare*

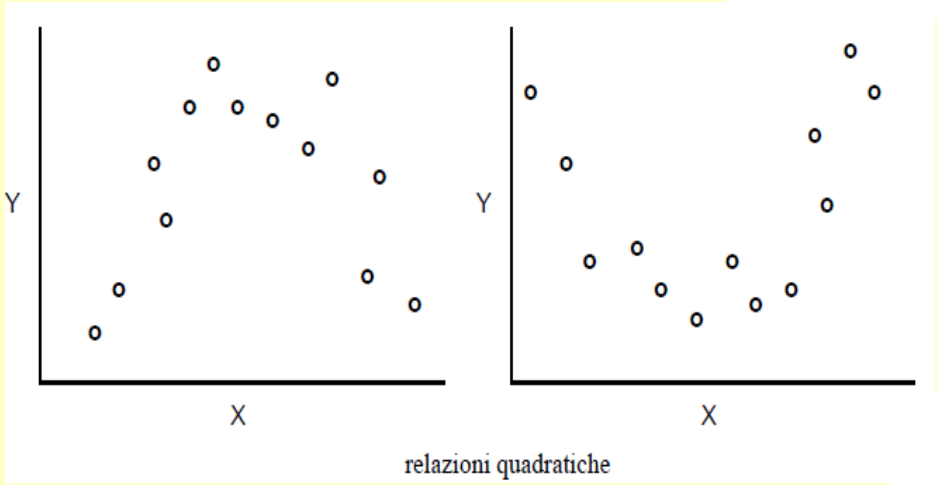
- Se tutti i punti sembrano giacere in prossimità di una curva, la ***correlazione*** è detta ***non-lineare*** e per la regressione si usa un'*equazione non lineare*

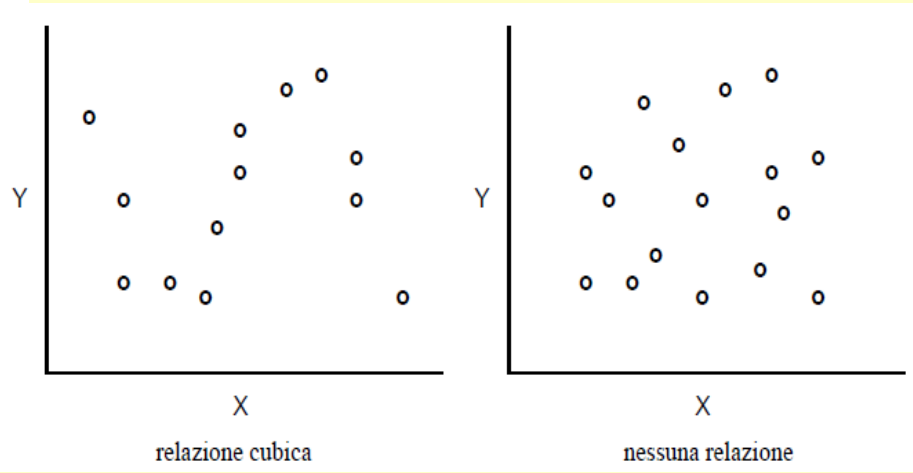
- Nel caso di correlazione tra le variabili, se y tende a crescere al crescere di x, la ***correlazione*** è detta ***positiva*** o ***diretta***, se y tende a decrescere al crescere di x, la ***correlazione*** è detta ***negativa*** o ***inversa***

- Se non c'è alcuna relazione tra le variabili, si dice che esse sono ***incorrelate***

Esempi di correlazioni





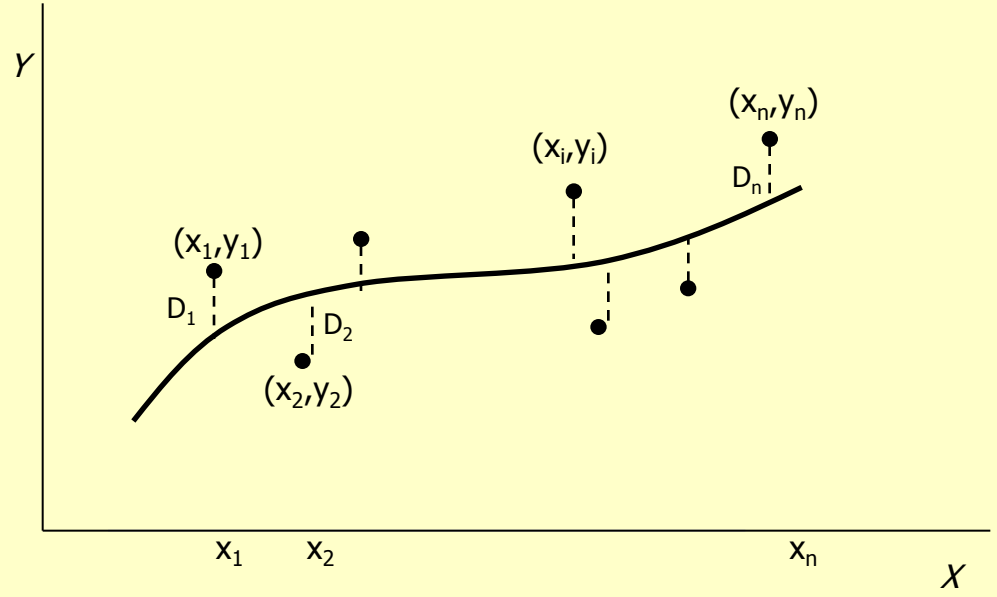


Il metodo dei minimi quadrati

Il problema generale di trovare l'equazione di una curva che interpoli certi dati è detto ***interpolazione***

Per evitare l'intervento del giudizio individuale nel costruire le curve interpolanti, è necessario accordarsi sulla definizione di ***migliore curva interpolante***

Per motivare una possibile definizione, consideriamo la figura seguente in cui sono indicati i punti dati (x_1, y_1), (x_2, y_2),…, (x_n, y_n) e una curva interpolante C



Per ogni valore x_i, i=1,…,n, ci potrà essere una differenza tra il valore di y_i e il corrispondente valore determinato sulla curva C. Tale differenza, indicata in figura con D_i, è denominata *deviazione* od *errore* e può essere positiva, negativa o nulla

Una misura della *"bontà dell'interpolazione"* effettuata per mezzo della curva C è fornita dalla somma:

$$D_1^2 + D_2^2 + \ldots + D_n^2$$

L'interpolazione è tanto migliore quanto più piccola è tale somma

Fra tutte le curve interpolanti un dato insieme di punti, si dice ***migliore interpolante*** la curva che rende minima "*la somma dei quadrati delle distanze dei punti sperimentali dai punti corrispondenti della curva teorica*"

Una curva avente questa proprietà è detta ***curva dei minimi quadrati*** (una retta avente tale proprietà è detta *retta dei minimi quadrati*, una parabola è detta *parabola dei minimi quadrati*, ecc.).

Il metodo matematico che permette di ottenere l'equazione della curva dei minimi quadrati è detto ***metodo dei minimi quadrati***

Retta dei minimi quadrati

Se la curva interpolante l'insieme dei punti

$$(x_1, y_1),(x_2,y_2),\ldots,(x_n, y_n)$$

è una retta, l'***equazione della retta dei minimi quadrati*** è **y=mx+q** , dove:

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

o equivalentemente

$$m = \frac{\bar{x}\cdot\bar{y} - \frac{1}{n}\sum_{i=1}^{n} x_i y_i}{\bar{x}^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}}$$

e $q = \bar{y} - m\bar{x}$, poiché le retta dei minimi quadrati passa per

il punto $(\bar{x}, \bar{y})$ (detto ***baricentro dei dati***)

Coefficiente di correlazione

Per vedere il grado di adattamento di un'equazione lineare ai dati sperimentali si usa il cosiddetto ***coefficiente di correlazione*** (di ***Pearson***):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Si prova che i valori che tale numero r può assumere sono tutti compresi nell'intervallo [-1, 1]

- Se ***r = 1***, significa che vi è una *relazione lineare* (della forma y=mx+q) nella quale le due variabili sono *correlate positivamente*
- Se ***r = -1*** vi è ancora una *relazione lineare* in cui però le variabili sono *correlate negativamente*
- Se ***r è sufficientemente vicino a +1 o a -1*** *(ciò significa di solito almeno 0,9 in valore assoluto),* allora l'interpolazione fornita dalla retta di regressione è buona
- Quanto più r si discosta dai valori 1 e –1, tanto meno preciso risulta l'allineamento dei punti: ***per r prossimo a 0***, *non sussiste alcuna correlazione lineare* fra le due variabili in esame *(può darsi comunque che esista un altro tipo di correlazione, di natura più complessa e quindi non esprimibile mediante funzioni lineari)*

Testo di riferimento

M Pagano, K Gauvreau – *Biostatistica* –
Ed. Gnocchi, Napoli 1994