STATISTICA MEDICA

Dott.ssa Marta Di Nicola

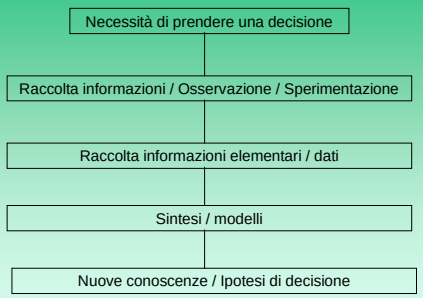
N.P.D. 3° Blocco 2° piano

0871-3554007

m.dinicola@unich.it

http://www.biostatistica.unich.it

Informazioni, nuove conoscenze, decisioni



LA STATISTICA

La Statistica ha come scopo la *conoscenza quantitativa dei fenomeni collettivi.*

L'analisi statistica mira ad individuare ***modelli*** di interpretazione della realtà, attraverso canoni e tecniche che sono astrazioni, semplificazioni di una moltitudine di aspetti e di manifestazioni del reale.

E' costituita da un insieme dei metodi che consentono dunque di raccogliere, ordinare, riassumere, presentare ed analizzare dati e informazioni, trarne valide conclusioni e prendere decisioni sulla base di tali analisi e risultati.

Funzione Descrittiva → Offre il metodo per riassumere le informazioni in modo da renderle utilizzabili più facilmente. Riduce i dati in forma maneggevole, sostituendo a molti numeri poche misure.

Funzione Inferenziale → Permette di generalizzare le informazioni, ricavando proprietà e leggi generali sulla base di dati rilevati solamente su una parte (campione) della popolazione.

GLOSSARIO

POPOLAZIONE: l'insieme di tutte le unità statistiche oggetto dell'osservazione (es.: medici, paramedici, studenti, diabetici, obesi, addetti all'agricoltura…).

CAMPIONE: la parte delle unità statistiche sottoposte all'osservazione, all'esperimento, etc.

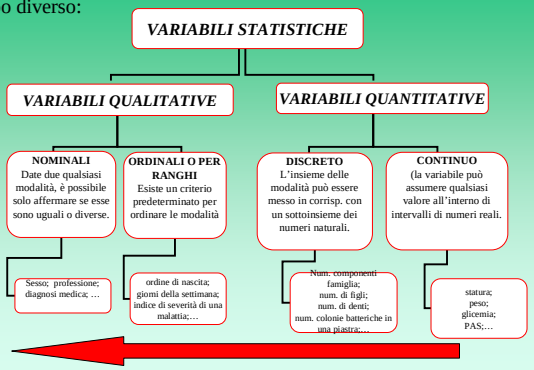
UNITA' STATISTICA: per ogni elemento o caso appartenente alla popolazione oggetto diretto della osservazione da cui si raccolgono i dati.

CARATTERE (O VARIABILE): la caratteristica (attributo o misura) osservata sulla unità statistica.

MODALITA': ogni diversa presentazione del carattere o variabile osservata su ciascuna unità statistica.

FREQUENZA: numero di volte che si presenta una data modalità.

I dati della statistica riguardano ***variabili***, cioè grandezze che possono assumere valori differenti. Le variabili possono essere di tipo diverso:



In una ricerca, si definisce ***variabile indipendente*** quella che viene manipolata direttamente dallo sperimentatore, o in alternativa selezionata attraverso il metodo di campionamento. Per esempio, il fatto che i pazienti siano trattati con un farmaco o con placebo è un esempio di variabile indipendente manipolata direttamente dallo sperimentatore. In alternativa, se viene selezionato un campione di maschi da confrontare con un campione di femmine, il sesso è una variabile indipendente controllata indirettamente attraverso il sistema di campionamento.

Al contrario, la ***variabile dipendente*** è quella che misuriamo per verificare la sua correlazione con la variabile indipendente. Nei due esempi precedenti, la variabile dipendente potrebbe essere la risposta alla terapia nel primo caso, e l'incidenza di una certa patologia nei due sessi nel secondo caso.

Informatizzazione dei dati

I dati possono essere inseriti in una tabella (**matrice dei dati**)

ID	sesso	età	n° figli	peso	altezza	DPB	SBP	…
Paul	M	57	1	65	158	90	150	…
Janet	F	70	3	100	175	75	138	…
Tim	M	45	0	71	162	72	141	…
David	M	38	2	58	164	81	160	…
John	M	25	1	81	170	69	135	…
…	…	…	…	…	…	…	…	…

Ciascuna riga rappresenta un'unità statistica.

Ciascuna colonna rappresenta una variabile

Statistica descrittiva:
riassunto e presentazione dei dati mediante tabelle (distribuzioni di frequenza) e grafici

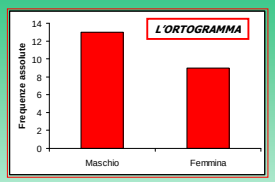
- Tabelle di frequenza (numero assoluto di casi per modalità)
- Tabelle percentuali (% di casi per modalità)
- Tabelle crociate (matrici 2 x 2, 2 x 3, ecc.)

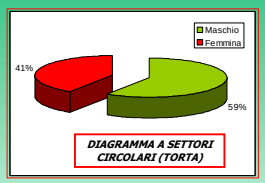
Distribuzione di frequenze della variabile età rilevata su un campione di 22 soggetti.

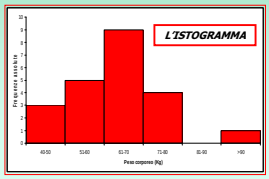
Età	Frequenze assolute	Frequenze assolute %	Frequenze assolute cumulate	Frequenze assolute cumulate %
17	3	13.6	3	13.6
18	6	27.3	9	40.9
19	12	54.6	21	95.5
20	1	4.5	22	100
Totale	22	100		

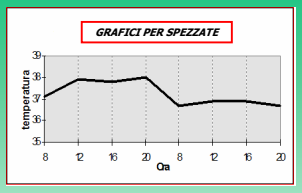
SESSO	STATO ALLA DIMISSIONE			
	Guarito	Malato	Morto	Totale
Maschi	20	12	4	36
Femmine	15	9	6	30
Totale	35	21	10	66

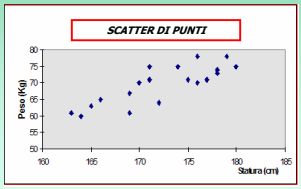
RAPPRESENTAZIONI GRAFICHE











Statistica descrittiva:

Individuare un indice che rappresenti significativamente un insieme di dati statistici.

Valori medi o medie algebriche

- ✓ media aritmetica;
- ✓ media armonica;
- ✓ media geometrica.

Indici di posizione o medie lasche

- ✓ mediana;
- ✓ moda;
- ✓ etc

LA MODA

Riportiamo i tempi di sopravvivenza (mesi) di 19 pazienti affetti da cancro dell'addome

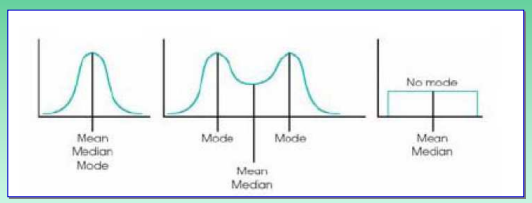
Mesi di sopravvivenza (x_i)	Frequenza (f_i)
8,5	2
9,2	4
7,3	8
6,8	2
10,1	3
Totale	19

Moda: la modalità che si presenta con la maggiore frequenza

per ogni tipo di dati

LA MODA

Molte distribuzioni di frequenza presentano più di un valore modale



LA MEDIANA

Mesi di sopravvivenza (x_i)	Frequenze (f_i)
8,5	2
9,2	4
7,3	8
6,8	2
10,1	3
Totale	19

6,8
6,8
7,3
7,3
7,3
7,3
7,3
7,3
7,3
7,3
8,5
8,5
9,2
9,2
9,2
9,2
10,1
10,1
10,1

Mediana: la modalità assunta dall'unità statistica che occupa il posto centrale nella distribuzione ordinata

per dati almeno ordinali

LA MEDIANA

Mesi di sopravvivenza (x_i)	Frequenze (f_i)	Frequenze cumulate
8,5	2	2
9,2	4	6
7,3	8	14
6,8	2	16
10,1	3	19
Totale	19	

n=19 (numerosità dispari)

(n+1)/2=10

In caso di numerosità pari

n/2 e (n/2)+1

Mediana: la modalità assunta dall'unità statistica che occupa il posto centrale nella distribuzione ordinata

per dati almeno ordinali

LA MEDIA ARITMETICA

Mesi di sopravvivenza (x_i)	Frequenza (f_i)	$x_i \cdot f_i$
8,5	2	17
9,2	4	36,8
7,3	8	58,4
6,8	2	13,6
10,1	3	30,3
Totale	19	156,1

Media aritmetica: è quel valore che avrebbero tutte le osservazioni se non ci fosse la variabilità (casuale o sistematica).

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = \frac{156,1}{19} = 8,2$$

La **media aritmetica** è la misura di posizione più usata ma. A volte, altre misure come la **mediana** e la **moda** si dimostrano più utili.

Si consideri un campione di valori di VES (*velocità di eritrosedimentazione*, mm/ora) misurati in 7 pazienti

{8, 5, 7, 6, 35, 5, 4}

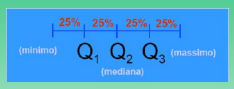
Media=10 mm/ora

Mediana=6 mm/ora

In questo caso, la media che è = 10 mm/ora non è un valore tipico della distribuzione: soltanto un valore su 7 è superiore alla media!

Limite della media aritmetica:
è notevolmente influenzata dai valori estremi della distribuzione.

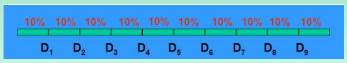
I **QUARTILI** dividono la distribuzione in quattro parti uguali. (Si osservi che il 2 quartile coincide con la mediana)



$IQR = Q_3 - Q_1$

Range interquartile

I **DECILI** dividono la distribuzione in dieci parti uguali



I **PERCENTILI** dividono la distribuzione in cento parti uguali

Statistica descrittiva:

Individuare un indice che possa misurare quanto una distribuzione sia "sparsa", ovvero quantificarne la variabilità (dispersione)

- ✓ Campo di variazione (range);
- ✓ Devianza;
- ✓ Varianza (S^2 o σ^2);
- ✓ Deviazione standard (S o σ);
- ✓ Coefficiente di variazione (indice di variabilità relativa).

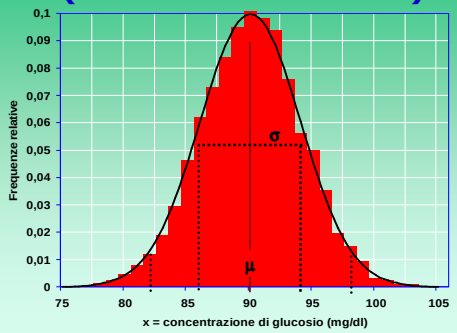
LA DEVIAZIONE STANDARD

La deviazione standard (σ) è la radice quadrata della varianza, un indicatore di dispersione che si ottiene sottraendo la media da ciascuna delle singole osservazioni, sommando i quadrati di queste differenze, e dividendo per il numero di osservazioni (meno uno).

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

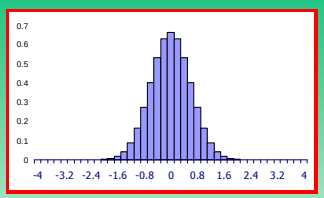
È espressa nella stessa unità di misura della variabile osservata

La curva di Gauss (distribuzione normale)

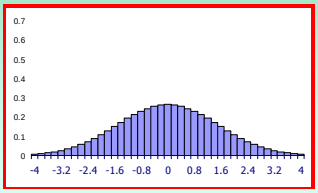


Le caratteristiche della distribuzione normale

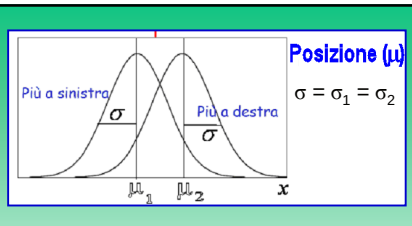
1. è simmetrica rispetto al valore medio
2. il valore di $x = \mu$ oltre che alla media aritmetica coincide con la moda e la mediana
3. è asintotica all'asse delle x da entrambi i lati
4. è crescente per $x<\mu$ e decrescente per $x>\mu$
5. possiede due punti di flesso per $x = \mu \pm \sigma$
6. l'area sotto la curva è = 1 (essendo la probabilità che si verifichi un qualsiasi valore di x)



media=0, DS=0.6



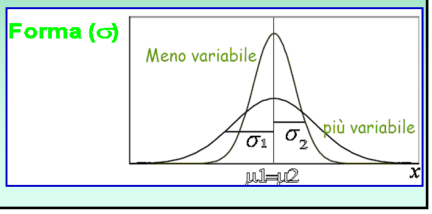
media=0, DS=1.5



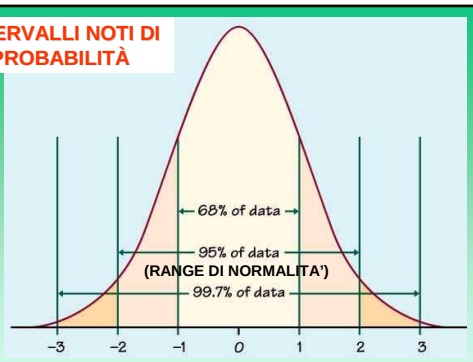
$\sigma = \sigma_1 = \sigma_2$

Al variare della media aritmetica (a parità di dev.standard) la curva trasla sull'asse delle x

Al variare della deviazione standard la curva modifica la sua forma

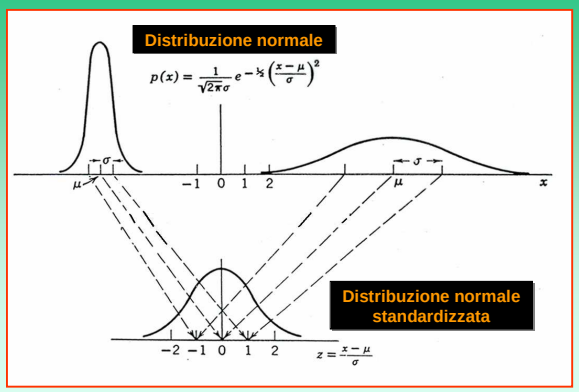


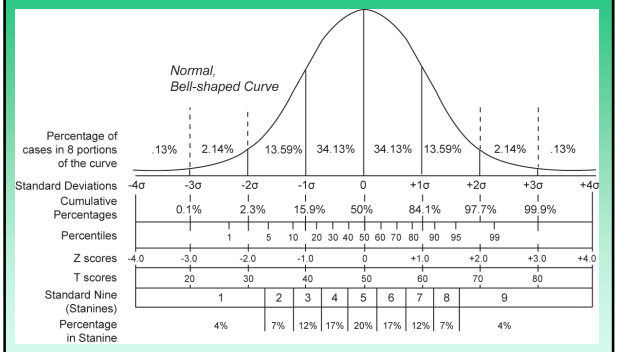
INTERVALLI NOTI DI PROBABILITÀ

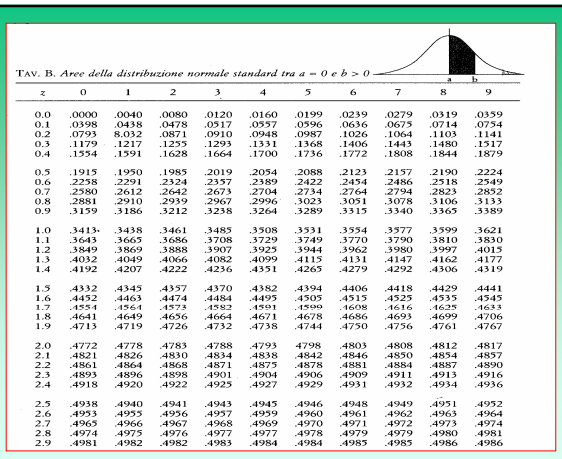


In una distribuzione normale perfetta:
68.26% dei casi sono compresi fra -**1** e +**1** DS attorno alla media
95.46% dei casi sono compresi fra -**2** e +**2** DS attorno alla media
99.74% dei casi sono compresi fra -**3** e +**3** DS attorno alla media

Distribuzione Normale Standardizzata







TAV. B. *Aree della distribuzione normale standard tra a = 0 e b > 0*

z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	8.032	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4351	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986

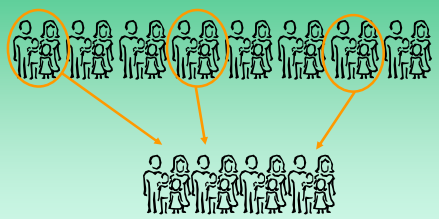
Campionamento statistico

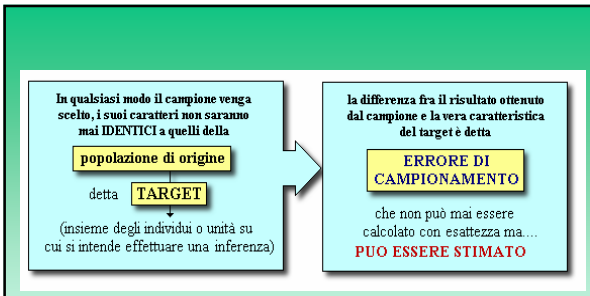
Nell'ambito della statistica descrittiva abbiamo finora considerato strumenti per descrivere un'intera popolazione quando siano noti tutti i dati ad essa relativi. Ma nella ricerca, in genere, non si conoscono i dati dell'intera popolazione, ma solo quelli di un campione.

Il campionamento si usa quando si vuole conoscere uno o più parametri di una popolazione, senza doverli misurare in ogni suo elemento. Il campionamento consiste nel selezionare un numero più piccolo di elementi fra tutti quelli che formano una popolazione. Può essere fatto in vari modi, ma deve sempre essere di tipo probabilistico (cioè garantire la casualità della selezione).

Parleremo allora di numerosità, media e deviazione standard del campione, e dobbiamo porci il problema di che rapporto esista fra questi valori e la numerosità, la media e la deviazione standard dell'intera popolazione.

CAMPIONE STATISTICO





L'errore di campionamento

L'**errore di campionamento** è rappresentato dalla differenza tra i risultati ottenuti dal campione e la vera caratteristica della popolazione che vogliamo stimare.

L'errore di campionamento non può mai essere determinato con esattezza, in quanto la «vera» caratteristica della popolazione è (e resterà!) ignota.

Esso tuttavia può essere **contenuto entro limiti più o meno ristretti** adottando appropriati metodi di campionamento.

Inoltre, esso può essere ***stimato***; ciò significa che, con adatti metodi statistici, si possono **determinare i limiti probabili** della sua entità.

Dimensionamento del campione (sample size)

- La numerosità del campione dipende in modo critico dall'entità della differenza esistente fra le due popolazioni relativamente al parametro oggetto dello studio
- In uno studio RCT, quindi, è importante dimensionare in anticipo il campione, cioè decidere prima quanti soggetti dovranno essere arruolati per rispondere al quesito
- Il dimensionamento va fatto tenendo conto della differenza più piccola che si ha interesse a cogliere (grandezza del segnale minimo che si considera utile), e del livello di significatività statistica che si desidera raggiungere (cioè, della soglia fissata per il p)

Variabili quantitative

$$n = \frac{2\sigma^2}{(\mu_1 - \mu_2)^2} f(\alpha,\beta)$$

Variabili qualitative

$$n = \frac{p_1(100-p_1)+p_2(100-p_2)}{(p_1-p_2)^2} f(\alpha,\beta)$$

Esempio: Il farmaco di riferimento riduce la pressione sistolica di 25 mmHg, il nuovo farmaco per essere competitivo dovrebbe ridurre la pressione sistolica di almeno 30 mmHg (ovvero 5 mmHg in più). La deviazione standard della riduzione della pressione viene stimata in 10 mmHg da studi precedenti. Si adotta un alfa del 5% e una potenza del 90%, pertanto $f(\alpha,\beta)=10,5$

$$n = \frac{2(10)^2}{(5)^2} 10,5 = 84$$

Occorrono almeno 84 soggetti per gruppo.

Esempio: Nei pazienti affetti da tumore X in stadio avanzato, la sopravvivenza a 5 anni è del 30% con il trattamento standard. Dati preliminari suggeriscono che nei pazienti sottoposti ad un nuovo trattamento la sopravvivenza salga al 40%.
Si adotta un alfa del 5% e una potenza dell' 80%, pertanto $f(\alpha,\beta)=7,9$

$$n = \frac{30(100-30)+40(100-40)}{(10)^2} 7,9 = 355,5$$

Occorrono almeno 356 soggetti per gruppo.

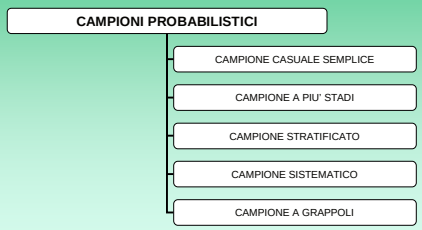
Definiamo *CAMPIONE RAPPRESENTATIVO* il sottoinsieme delle unità statistiche sottoposte all'osservazione che abbia:
una struttura rispecchiante quella della popolazione;
una numerosità adeguata alla popolazione di origine.

Come costruire un campione? **Si definisce piano di campionamento un metodo attraverso il quale si selezionano gli elementi che entrano a far parte del campione.**

Esistono diversi metodi di campionamento la scelta è legata ai costi, alla tempestività, alla precisione e alla disponibilità di una lista degli elementi della popolazione

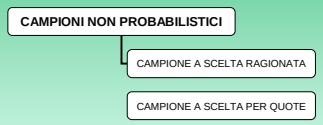
Tecniche di campionamento

Probabilistico quando ogni unità della popolazione ha la stessa probabilità nota di entrare a far parte del campione.



Tecniche di campionamento

Non probabilistico quando le unità non sono scelte in modo casuale ma attraverso scelte ragionate.



Campionamento casuale semplice

Si considerino N unità statistiche costituenti una popolazione e si assegni ad ogni unità un numero progressivo da 1 a N. Dalla lista così costruita vengono in successione estratte n unità statistiche (n<N) che vanno a costituire il campione.

La selezione delle unità statistiche che costituiscono il campione casuale semplice avviene attraverso le Tavole dei Numeri Casuali.

Il campionamento può essere:

Senza ripetizione: quando ogni unità statistica estratta viene poi esclusa dalla lista in modo che non possa essere estratta più di una volta. In questo caso la probabilità di estrazione di un unità statistica è n/N.

Con ripetizione: quando ogni unità statistica può essere estratta più di una volta. In questo caso la probabilità di estrazione di un unità statistica è $1/N^n$

TAVOLA DEI NUMERI ALEATORI

91	94	2	11	17	56	4	92	47	39	73	55	1	68	64
57	19	76	52	3	29	9	36	35	89	22	80	3	29	88
25	90	21	71	30	95	5	74	98	14	43	56	09	0	71
84	76	63	27	69	0	82	33	4	33	48	95	80	91	71
38	70	36	91	96	50	14	20	95	9	99	26	66	64	36
94	82	58	28	13	44	20	46	40	57	42	19	11	80	75
97	33	65	27	10	38	63	59	58	7	48	73	69	76	97
79	63	96	36	6	72	42	30	99	16	95	46	61	94	64
15	79	32	60	7	72	39	94	39	14	94	41	27	19	11
92	47	92	91	99	99	9	92	40	29	48	76	94	74	99
97	93	24	60	38	91	23	92	41	69	21	99	46	67	73
10	49	63	96	0	31	92	9	82	29	80	2	21	66	42
12	16	20	60	95	35	34	67	43	9	97	64	66	36	3
92	10	34	61	90	43	33	95	33	27	8	11	69	72	74
94	73	76	92	78	23	92	29	66	94	92	96	95	99	61
54	53	99	53	93	91	94	52	51	41	59	20	3	30	95
97	10	71	12	99	93	40	3	46	19	92	54	74	24	14
90	7	13	59	79	94	50	24	10	25	46	92	9	95	4
71	20	44	94	67	9	35	98	94	28	82	93	28	99	69
30	42	71	11	60	88	76	33	4	24	95	99	69	19	79
59	29	59	0	81	23	30	40	20	64	57	93	48	94	6
91	24	53	17	17	9	94	30	24	19	92	79	48	62	36
93	97	26	3	69	21	27	1	51	77	49	18	66	99	41
97	99	62	39	91	77	97	11	12	39	94	94	21	99	91
91	66	27	41	21	90	97	66	45	29	13	43	0	66	91

Esempio

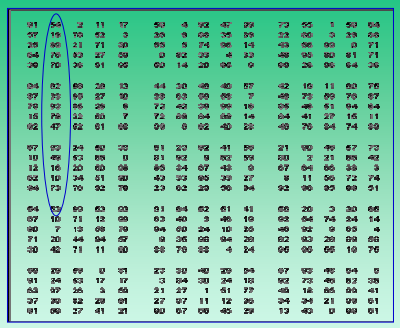
Supponiamo di voler controllare il tempo di disaggregazione di un campione di 100 compresse da estrarre con randomizzazione semplice da un lotto di 4000 compresse.

La procedura richiederà la numerazione da 1 a 4000 di tutte le compresse (per esempio con una matita), l'estrazione dei 100 numeri casuali e la selezione delle 100 compresse corrispondenti.

Esempio

Supponiamo di voler "costruire" un campione casuale semplice senza ripetizione di aziende ospedaliere allo scopo di valutare la degenza media di ciascun presidio.
Siano N=80 le aziende ospedaliere operanti sul territorio nazionale e supponiamo si decida che un campione composto da n=10 aziende sia sufficiente per rappresentare l'intera popolazione degli ospedali italiani.

Dato che stiamo trattando numeri a due cifre prendiamo le prime due colonne della tavola dei numeri aleatori selezionando i numeri ≤ 80.



Le aziende da campionare saranno quelle corrispondenti ai numeri:
54, 19, 76, 70, 33, 79, 47, 49, 16, 10, 73, 53

In questo modo vengono selezionate 12 aziende, due più del numero stabilito in modo da cautelarsi nel caso in cui qualche azienda non sia analizzabile.

LIMITE

Ovviamente vi sono situazioni in cui il campionamento per randomizzazione semplice risulta poco pratico se non addirittura inapplicabile. Infatti, il principale svantaggio è quello di richiedere la preventiva numerazione di tutti i soggetti; successivamente è necessario individuare nella popolazione le unità statistiche corrispondenti ai numeri estratti.

Campionamento sistematico

Le unità campionarie vengono estratte selezionandole sistematicamente una ogni dato intervallo (k=N/n).

Esempio: Campione sistematico di 2.000 soggetti ricoverati nel 2004 presso l'ospedale di Pescara

(k=N/n) 31.695/2.000=15.85

Campionamento stratificato

Si suddivide la popolazione in k classi (detti strati) ciascuna con elementi il più possibile omogenei tra loro e si estrae un campione casuale di opportune dimensioni da ciascuna classe.

Esempio Supponiamo di voler effettuare un campionamento per randomizzazione stratificata dei degenti di un reparto ospedaliero.

La stratificazione viene effettuata sulla base di un fattore che influenza il livello del carattere da studiare

Campionamento a grappoli

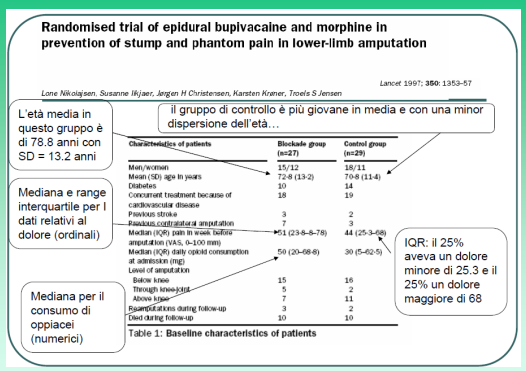
La popolazione viene suddivisa in sottogruppi detti grappoli (clusters) composti da unità il più possibile **eterogenee** tra loro e successivamente viene effettuato un campionamento sui grappoli.

Rispetto alla randomizzazione semplice, sistematica o stratificata, il campionamento a grappolo offre il vantaggio di facilitare notevolmente il reclutamento dei soggetti; di conseguenza si abbassano costi e tempi dell'indagine. Tuttavia, l'errore di campionamento può essere più elevato rispetto ai suddetti metodi di randomizzazione.

Esempio

Un reparto ospedaliero è composto da 15 stanze e ospita complessivamente 60 pazienti 4 soggetti in ogni stanza.
E' necessario prelevare un campione di sangue da un campione di 20 pazienti.

Effettuiamo un campionamento a grappolo: l'unità di studio non è più il *paziente* bensì la *stanza*. Si procede perciò a selezionare, ad esempio per randomizzazione sistematica, 5 stanze e si effettuano i prelievi dai 4 soggetti presenti in ciascuna di esse.



Randomised trial of epidural bupivacaine and morphine in prevention of stump and phantom pain in lower-limb amputation

Lancet 1997; **350**: 1353–57

Lone Nikolajsen, Susanne Ilkjaer, Jørgen H Christensen, Karsten Kramer, Troels S Jensen

L'età media in questo gruppo è di 78.8 anni con SD = 13.2 anni

il gruppo di controllo è più giovane in media e con una minor dispersione dell'età…

Mediana e range interquartile per i dati relativi al dolore (ordinali)

Mediana per il consumo di oppiacei (numerici)

IQR: il 25% aveva un dolore minore di 25.3 e il 25% un dolore maggiore di 68

Characteristics of patients	Blockade group (n=27)	Control group (n=29)
Men/women	15/12	18/11
Mean (SD) age in years	72·8 (13·2)	70·8 (11·4)
Diabetes	10	14
Concurrent treatment because of cardiovascular disease	18	19
Previous stroke	3	2
Previous contralateral amputation	7	3
Median (IQR) pain in week before amputation (VAS, 0–100 mm)	51 (23·6–8·78)	44 (25·3–68)
Median (IQR) daily opioid consumption at admission (mg)	50 (20–68·8)	30 (5–62·5)
Level of amputation		
Below knee	15	16
Through knee-joint	5	2
Above knee	7	11
Reamputations during follow-up	3	2
Died during follow-up	10	10

Table 1: **Baseline characteristics of patients**

Statistica inferenziale

Tutte le misure fino ad ora calcolate sono statistiche campionarie. L'inferenza statistica è il processo che permette di trarre delle conclusioni sull'intera popolazione a partire dalle statistiche campionarie.

La statistica descrittiva, pur aiutandoci a capire le proprietà dei dati in nostro possesso, non aggiunge nulla alle informazioni che già abbiamo. Le sue affermazioni, essendo relative a dati certi, sono certe.

La statistica inferenziale, invece, si propone di fare nuove affermazioni a proposito di dati che non possediamo, le sue affermazioni, quindi, sono probabilistiche.

Statistica inferenziale

I problemi che la statistica inferenziale cerca di risolvere sono essenzialmente di due tipi:

1) Problema della stima puntuale o intervallare (per esempio stima di una media):

- fornisce informazioni sulla media di una popolazione quando sono note media e deviazione standard di un campione della stessa.

2) Problema della verifica di ipotesi (per esempio confronto fra due o più campioni):

- calcola la probabilità che due campioni, di cui siano note media e deviazione standard, siano campioni derivati da una stessa popolazione oppure da due popolazioni diverse.

Media del campione e media della popolazione

Immaginiamo di avere una popolazione rappresentata da mille persone.

Se conoscessimo la statura di ciascuno dei mille abitanti, potremmo descrivere la popolazione con assoluta precisione in termini di media e deviazione standard.

Se però non abbiamo le risorse per misurare la statura di mille abitanti, possiamo scegliere un campione casuale, per esempio di 30 abitanti. Avremo allora una media e una deviazione standard del campione.

Che rapporto c'è fra questi valori e quelli dell'intera popolazione di mille abitanti?

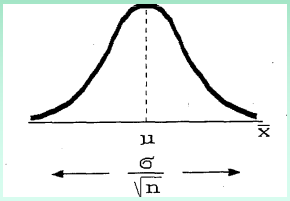
Immaginiamo di ripetere l'operazione di campionamento 20 volte, ogni volta con un diverso campione casuale di 30 abitanti. Otterremo 20 medie diverse, e 20 DS diverse.

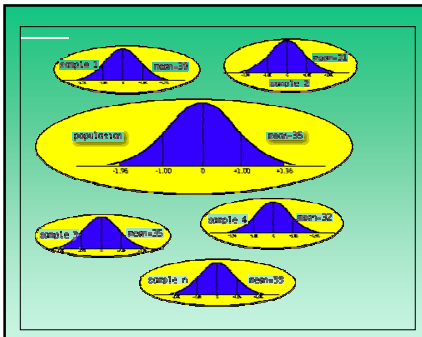
Un concetto importante è che l'insieme di queste medie dei campioni tende ad assumere una distribuzione normale, anche se la popolazione di origine non è distribuita normalmente.

In altre parole, il processo di campionamento casuale è di per sé un fenomeno che si distribuisce normalmente.

Teorema del limite centrale

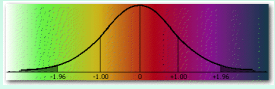
Il teorema del limite centrale afferma appunto che, **data una certa popolazione con media μ e deviazione standard σ**, da cui si estrae un numero infinito di campioni random e di numerosità n, man mano che n aumenta **la distribuzione delle medie dei campioni tende a una distribuzione normale**, con media μ (uguale a quella della popolazione di origine) e DS = σ/√n.





Campioni diversi di una popolazione.

Le medie dei vari campioni…



Distribution of Sample Means

…tendono a distribuirsi normalmente.

Stima Intervallare (Confidence Interval)

Confidence interval = intervallo all'interno del quale con una certa probabilità cade il parametro (ad es. media aritmetica) della popolazione

Per esempio, per un confidence level della media del 95%

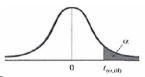
Se è nota la DS σ della popolazione generale:

$$\left\{\bar{x}-1{,}96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1{,}96\frac{\sigma}{\sqrt{n}}\right\}=1-\alpha$$

Se non è nota la DS σ della popolazione generale:

$$\left\{\bar{x}-t_{\alpha,gl}\frac{s}{\sqrt{n}} \le \mu \le \bar{x}+t_{\alpha,gl}\frac{s}{\sqrt{n}}\right\}=1-\alpha$$

Tavola della distribuzione T di Student



Gradi di libertà	Area nella coda di destra								
	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	3.078	6.314	12.706	15.894	31.821	63.656	127.321	318.289	636.578
2	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.328	31.600
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.214	12.924
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073

Se la **media del campione** è, per esempio, **25**, e il CI calcolato per un CL del 95% va da 22 a 28 (media ±3), allora si può dire che:

Secondo i dati a nostra disposizione, l'affermazione che
la media della popolazione di origine è compresa fra 22 e 28

ha il 95% di probabilità di essere vera.

NB: E' assolutamente sbagliato, dire che, con il 95% di probabilità, la media della popolazione di origine è uguale a 25

Stima Intervallare (Confidence Interval)

Per le variabili categoriche, in maniera assolutamente analoga, è possibile stimare la percentuale di una variabile nella popolazione generale a partire da quella nel campione, calcolando un CI.

Per esempio, per un confidence level della proporzione del 95%

$$\left\{ p - t_{\alpha,df}\sqrt{\frac{p(1-p)}{n}} \le \pi \le p + t_{\alpha,df}\sqrt{\frac{p(1-p)}{n}} \right\} = 1-\alpha$$

Verifica di ipotesi

La verifica di ipotesi è il secondo tipo di problema affrontato dalla statistica inferenziale.

L'ipotesi da verificare in questo caso è la cosiddetta "ipotesi nulla" (null hypothesis)

Ipotesi nulla

L'ipotesi nulla (H_0) è un'ipotesi che il ricercatore fa riguardo a un parametro della popolazione oggetto della ricerca e che viene confutata o non confutata dai dati sperimentali. Nel caso più comune, del confronto fra due campioni, la forma dell'ipotesi nulla è la seguente:

$$H_0: \mu_1 = \mu_2$$

Dove μ_1 e μ_2 sono le medie delle due popolazioni da cui sono stati tratti i due campioni.

Per esempio, se i due campioni si riferiscono a neonati a termine oppure a neonati pretermine, e la variabile misurata è il valore della glicemia a un'ora di vita, allora l'ipotesi nulla dice che:

non c'è differenza fra la media dei valori glicemia a un'ora di vita nelle due popolazioni.

L'ipotesi alternativa, cioè che la differenza esiste, prende il nome di H_1

Il t test per campioni indipendenti (unpaired t test)

Si considerino 2 campioni costituiti da soggetti caratterizzati da diverse abitudini alimentari

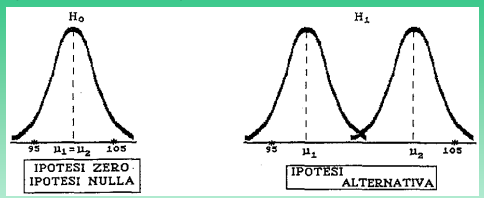
- si analizzano i livelli di glicemia di ciascun soggetto appartenente ai 2 campioni e si calcolano le medie aritmetiche e le deviazioni standard:

I campione	$n_1 = 41$	$\bar{x}_1 = 95$	$s_1 = 13.32$
II campione	$n_2 = 51$	$\bar{x}_2 = 105$	$s_2 = 16.94$

- l'alimentazione condiziona i livelli glicemici?

- le glicemie medie nei 2 campioni differiscono per le diverse abitudini alimentari o per effetto dello errore di campionamento?

- è possibile avanzare due ipotesi:



I due campioni sono stati estratti da popolazioni con medie uguali (μ 1= μ 2)

I due campioni sono stati estratti da popolazioni con medie diverse (μ 1? μ 2)

Il test del t di Student consente di saggiare la veridicità dell'ipotesi nulla

STATISTICA TEST ⇒

$$t_{g.l.} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{95 - 105}{\sqrt{\frac{13.32^2}{41} + \frac{16.94^2}{51}}} = -3.1696$$

	α/2			
g.l.	0.05	0.025	0.01	0.005
90	1.6620	1.9867	2.3685	2.6316

Il t test per campioni dipendenti (paired t test)

- si consideri un campione di 10 pazienti ipertesi cui viene somministrato un farmaco antiipertensivo; a questi pazienti viene misurata la pressione sistolica prima della somministrazione del farmaco e alcune ore dopo la somministrazione stessa:

Soggetto	PAS prima (mmHg)	PAS dopo (mmHg)
1	211	181
2	210	210
3	210	196
4	203	200
5	196	167
6	191	161
7	190	178
8	177	180
9	173	149
10	170	119
MEDIA	**193,1**	**174,1**

La pressione arteriosa è diminuita per l'errore di campionamento (H_0) o per effetto del farmaco (H_1)?

- in questo caso i 2 campioni (PAS prima e PAS dopo la somministrazione) sono appaiati (ovvero ciascuna osservazione di un campione si accoppia con una osservazione dell'altro campione)

- per saggiare l'ipotesi nulla si utilizza sempre il test del t di Student per campioni appaiati.

Soggetto	PAS prima (mmHg)	PAS dopo (mmHg)	di	(di-dm)	(di-dm)2
1	211	181	30	11	121
2	210	210	0	0	0
3	210	196	14	14	196
4	203	200	3	3	9
5	196	167	29	29	841
6	191	161	30	30	900
7	190	178	12	12	144
8	177	180	-3	-3	9
9	173	149	24	24	576
10	170	119	51	51	2601

$$dm = \frac{\sum d_i}{n} = \frac{190}{10} = 19$$

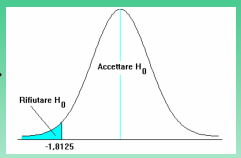
$$S_d = \sqrt{\frac{\sum (d_i - dm)^2}{n-1}} = \sqrt{\frac{2566}{9}} = 16{,}9$$

STATISTICA TEST ⇨ $t = \frac{dm}{S_d/\sqrt{n}} = \frac{19}{16{,}9/3{,}2} = 3{,}6$

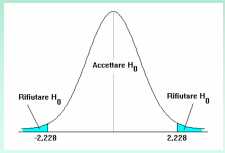
	α/2			
g.l.	0.05	0:025	0.01	0.005
90	1.6620	1.9867	2.3685	2.6316

Varietà di "t-test"

nel test ad una coda, la zona di rifiuto è solamente da una parte della distribuzione (a sinistra quando il segno è negativo, a destra quando è positivo)



nel test a due code, la zona di rifiuto è distribuita dalle due parti



Il test a due code è più conservativo (vi si ricorre quando non si ha alcuna idea sui possibili risultati) mentre il test ad una coda è più potente

		Realtà	
		Ho vera	Ho falsa
Conclusioni del Test	Accetto Ho	Conclusione corretta	Errore di II tipo (β)
	Rifiuto Ho	Errore di I tipo (α)	Conclusione corretta

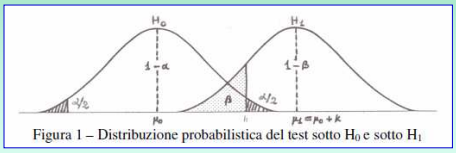


Figura 1 – Distribuzione probabilistica del test sotto H_0 e sotto H_1

VARIABILI STATISTICHE DOPPIE: CONFRONTO DI 2 CAMPIONI

(Confronto tra due percentuali)

Esempio 3. Si abbia un campione di 1020 soggetti diviso in
Fumatori (A): n_A=400 ➡ Prevalenza BCO 30%
Non fumatori (B): n_B=620 ➡ Prevalenza BCO 15%

Il fumo è "causa" (o fattore di rischio) per la bronchite? ossia il Δ (+15%) è statisticamente significativo?

	BRONCHITE		
FUMO	**SI**	**NO**	**TOT**
SI	120	280	400
NO	93	527	620
TOTALE	213	807	1020

La prevalenza di bronchite risulta statisticamente ≠ tra i fumatori e i non fumatori?
H0: La bronchite si sviluppa **indipendentemente** dal fumo;
H1: I fumatori sviluppano bronchite più dei non fumatori.

Tornando all'esempio dell'associazione tra BCO e fumo si ha la tabella delle frequenze attese:

	BRONCHITE CRONICA		
FUMO	SI	NO	TOT
SI	84	316	400
NO	129	491	620
TOTALE	213	807	1020

Es. (620x213)/1020 = 129; per differenza si calcolano le altre tre frequenze interne.

$$\chi^2 = \frac{(120-84)^2}{84} + \frac{(280-316)^2}{316} +$$

$$+ \frac{(93-129)^2}{129} + \frac{(527-491)^2}{491} = 32.21$$

LA FORMULA PER CALCOLARE L'INDICE-TEST CHI-QUADRATO

$$\sum_i \frac{(O_i - A_i)^2}{A_i}$$

Nel caso di tabelle 2x2 si può calcolare il valore del test χ^2 anche direttamente attraverso la formula seguente:

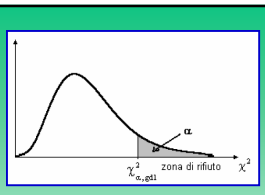
$$\chi^2\text{-test} = \frac{(ad - cb)^2 N}{N_1 N_2 N_A N_B}$$

FORMULA PER IL CALCOLO DEL χ^2 VALIDA SOLO NEL CASO DI TABELLE TETRACORICHE

$$\chi^2 = \frac{((120 \cdot 527) - (93 \cdot 280))^2 \cdot 1020}{213 \cdot 807 \cdot 400 \cdot 620} = 32.21$$

Valore quasi coincidente a quello calcolato con la precedente formula, quindi

LE DUE FORMULE FORNISCONO RISULTATI EQUIVALENTI



d.f.	α=0.250	0.100	0.050	0.025	0.010	0.005
1	1.3233	2.7055	3.8415	5.0239	6.6349	7.8794
2	2.7726	4.6052	5.9915	7.3778	9.2104	10.5965
3	4.1083	6.2514	7.8147	9.3484	11.3449	12.8381
4	5.3853	7.7794	9.4877	11.1433	13.2767	14.8602
5	6.6257	9.2363	11.0705	12.8325	15.0863	16.7496
6	7.8408	10.6446	12.5916	14.4494	16.8119	18.5475
7	9.0371	12.0170	14.0671	16.0128	18.4753	20.2777

Regressione

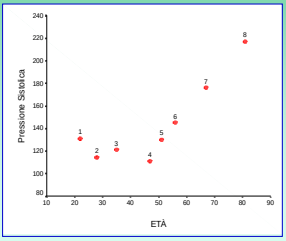
Quando si considerano due o più caratteri (variabili) si possono esaminare anche il tipo e l'intensità delle relazioni che sussistono tra loro.

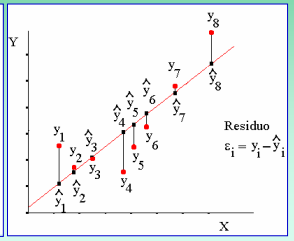
L'analisi di regressione consente di sviluppare un modello statistico che possa essere usato per prevedere i valori di una variabile, detta dipendente ed individuata come l'effetto, sulla base dei valori dell'altra variabile, detta indipendente o esplicativa, individuata come la causa.

soggetto	ETA' (anni)	PAS (mm Hg)
1	22	131
2	28	114
3	35	121
4	47	111
5	51	130
6	56	145
7	67	176
8	81	217

RETTA DI REGRESSIONE

$$\hat{Y} = a + bX$$





Secondo il principio dei minimi quadrati si stimano matematicamente a e b:

$$b = \frac{\text{CODEV(X,Y)}}{\text{DEV(X)}} \qquad a = \overline{y} - b \cdot \overline{x}$$

l'aumento medio della pressione è di circa **b=1.5 mmHg** per l'aumento di un anno di età.

alla nascita il valore della pressione *sarebbe* (!) di a**=68.56 mmHg**, ***ma*** questa è una indicazione teorica perché non è possibile stimare il valore della pressione arteriosa per età fuori del range considerato (22- 81 aa).

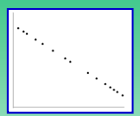
Coefficiente di correlazione

Il **coefficiente di correlazione** esprime quanto due variabili sono correlate fra loro, anche se non sussiste necessariamente un rapporto diretto di causalità. La correlazione può essere lineare o di altro tipo (quadratica, ecc.)

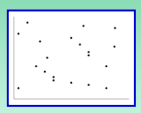
Un coefficiente di correlazione va da -1 (correlazione negativa) a 1 (correlazione positiva). I valori intorni allo 0 esprimono l'assenza di correlazione.

Il più semplice coefficiente di correlazione è quello di **Pearson**, detto r, che misura la **correlazione lineare fra due variabili** in un campione.

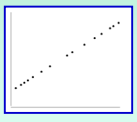
r = -1



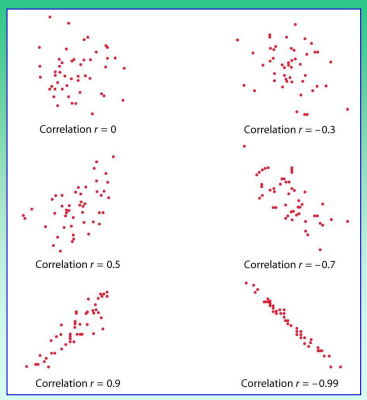
r = 0



r = +1



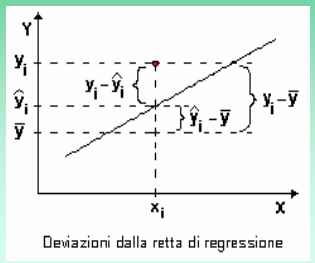
Altri esempi di r



Coefficiente di determinazione r^2

E' il **quadrato del coefficiente di correlazione,** ed esprime la percentuale della variazione dei valori di y che è spiegata dal modello di regressione ($0 \leq r^2 \leq 1$).

$$r^2 = \frac{\text{Devianza di Regressione}}{\text{Devianza Totale}}$$



Regressione multipla

I test di regressione multipla valutano la maniera in cui **molte variabili indipendenti** influenzano **una singola variabile dipendente**: per esempio, come vari fattori prognostici influenzano la sopravvivenza in una patologia neoplastica.