UNIVERSITÀ
DI SIENA
1240

Prof. F. Ferretti

Statistica medica per le professioni sanitarie

INDICE

INTRODUZIONE

L'uso della statistica per la diffusione di notizie o per l'esame di fenomeni di varia natura trova nella società attuale un impiego sempre più largo. Tuttavia, se è vero che la statistica ci aiuta nelle scelte e nella scoperta di nuovi fattori, è altrettanto vero che questa scienza può essere utilizzata per interpretare in modo distorto la realtà.
Nell'arco di questo corso di statistica per Tecnici di Laboratorio Biomedico saranno forniti gli elementi di base che permetteranno al tecnico di possedere le conoscenze essenziali per un corretto utilizzo degli strumenti analitici messi a disposizione dalla scienza statistica.
In poche parole, la statistica può essere definita come quell'insieme di metodi di natura logica e matematica atti a raccogliere, analizzare ed interpretare dati numerici o numerabili. Dunque, l'applicazione della statistica alla metodologia della ricerca (anche in campo biomedico) permette una più corretta, oltre che più obiettiva, interpretazione dei fenomeni osservati ed esprime un diverso grado di credibilità dei risultati ottenuti.
All'interno della scienza statistica si possono individuare due argomenti principali che raggruppano specifiche metodologie, ovvero:

1. la *statistica descrittiva*, nella quale gli strumenti analitici hanno il compito di organizzare, elaborare e presentare i dati;
2. la *statistica inferenziale*, nella quale le metodologie servono a trarre conclusioni pertinenti riguardanti la popolazione a partire dai risultati forniti da un campione rappresentativo di essa.

Sulla base di queste prime definizioni è opportuno precisare anche la terminologia con la quale nella letteratura statistica si indicano gli elementi sopra citati. Si definisce *universo* o *popolazione* (indicati con il simbolo U) la massa di possibili individui od osservazioni, mentre il *campione* è formato dalle osservazioni effettivamente utilizzate, purché tratte dalla massa secondo criteri opportuni. Ogni singolo soggetto in esame prende il nome di *unità statistica*, il rilievo che si esegue da luogo all'*osservazione* e l'insieme delle osservazioni costituisce il *dato statistico*.

STATISTICA DESCRITTIVA

Variabili

Come abbiamo già precisato, la statistica descrittiva ha come primo scopo quello di riassumere le informazioni provenienti da una massa di dati, permettendo al lettore di individuare a prima vista le caratteristiche di un gruppo di valori.
Prima di esaminare in dettaglio i principali strumenti analitici è opportuno approfondire la conoscenza dell'oggetto sul quale vengono utilizzate queste metodologie: il *dato statistico*.
Quando si compie una rilevazione di dati, che in seguito saranno elaborati mediante tecniche descrittive, in realtà si rilevano i valori o le modalità con quali si presentano determinate variabili. Sesso, età, altezza, colore degli occhi, ed altre ancora non sono che variabili le cui determinazioni possono essere oggetto di una rilevazione statistica. Ognuna di queste presenta particolari caratteristiche che permettono di classificarla in una ben precisa categoria. Saper riconoscere e classificare le informazioni raccolte è molto importante, poiché la scelta del metodo di analisi appropriato dipende direttamente dal tipo di variabile sulla quale la metodologia deve essere utilizzata.
In funzione delle proprie caratteristiche, le variabili possono essere suddivise in :

1. *variabili qualitative*;
2. *variabili quantitative*.

Nel primo caso le caratteristiche possono essere espresse mediante attributi (p.e.: colore dei capelli, stato civile, sesso, ...), mentre nel secondo tali caratteristiche vengono indicate da numeri (età, altezza, peso, ...).
All'interno di queste due categorie è possibile individuare altri sottogruppi entro cui classificare i tipi di grandezze. Le variabili quantitative possono essere suddivise in :

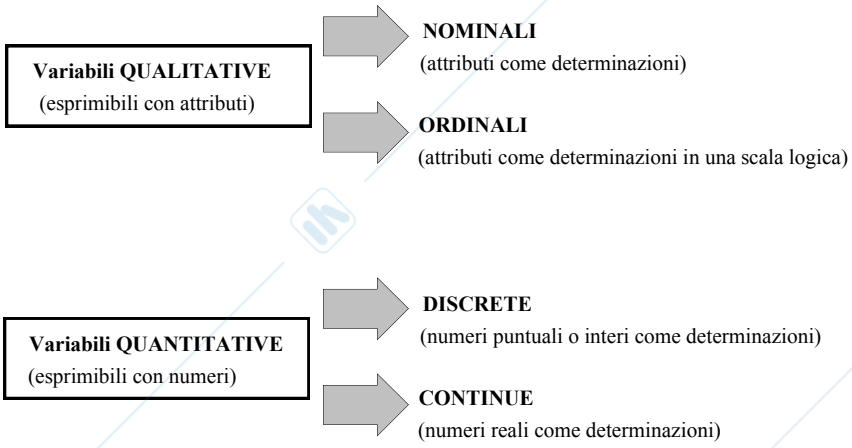
1.1 *variabili qualitative nominali*;
1.2 *variabili quantitative ordinali*.

Nel primo caso (nominali) l'attributo che indica la caratteristica ha il solo scopo di rappresentare la modalità con la quale si presenta il dato. Per la variabile "colore degli occhi" alcuni tra i possibili attributi, per esempio, potrebbero essere "celeste", "marrone", "nero". In questa classe si indicano con il termine *variabili dicotomiche* quelle grandezze che ammetteno solo due risposte possibili (p.e. sesso: maschio/femmina), mentre sono dette *variabili policotomiche* quelle con più risposte ammissibili (p.e. stato civile: celibe/coniugato/separato/ ....).

Nelle variabili qualitative ordinali l'attributo ha ancora lo scopo di rappresentare la modalità con la quale si presenta il dato, ma queste modalità sono organizzate in una scala di valori logici. Per esempio, la scolarità è una variabile che appartiene a questa classe poiché le sue determinazioni sono degli attributi, ma il loro ordine esprime una scala di importanza del titolo di studio posseduto. (elementare/media/superiore/laurea). Un altro esempio potrebbe essere la sensazione di dolore avvertita durante l'esecuzione di un esame, variabile che potrebbe essere catalogata come ordinale mediante le seguenti codifiche: occasionale, moderato, medio, intenso.
Le variabili quantitative, invece, possono essere suddivise in:

2.1 *variabili quantitative discrete*;
2.2 *variabili quantitative continue*.

Alla base della differenza vi è il tipo di numero che viene utilizzato per rappresentare la determinazione della variabile. Nel primo caso si tratta di numeri non necessariamente interi, collocabili in una scala di tipo ordinale, ma capaci di assumere solo certi valori puntuali. Alcuni esempi molto semplici sono: il numero dei membri delle famiglie, la frequenza cardiaca, il numero di gravidanze ed altre ancora.
Le variabili quantitative continue, invece, sono quelle che possono assumere tutti gli infiniti valori all'interno o meno di un determinato intervallo. Appartengono a questa classe grandezze quali l'altezza (p.e. 174,5 cm, 148,32 cm, ...), il peso (p.e. 58,7Kg, 63,25 Kg, ...) e qualsiasi altro dato misurabile e rappresentabile attraverso un numero reale, ovvero le classiche informazioni ricavabili da una misurazione.
Nel testo abbiamo utilizzato un termine che riassume le categorie elencate; tutte quante le tipologie individuate, infatti, possono essere indicate con il termine di *scale di misurazione*. Il seguente schema riassume la classificazione delle variabili enunciata.



Distribuzioni di frequenza

Da un punto di vista formale la distinzione tra tipi variabili obbliga l'operatore non solo ad un uso ragionato delle metodologie, ma anche ad utilizzare una terminologia appropriata in funzione dei dati di cui dispone. Per questo motivo si definiscono *serie statistiche* le successioni "delle intensità o delle frequenze corrispondenti alle modalità di un carattere qualitativo"[1]. Analoga è la definizione che lo stesso Autore fornisce della *seriazione*, con la differenza che quest'ultima si riferisce a caratteri quantitativi.

Nella pratica, a volte, possono sorgere dei dubbi nello stabilire se ci si trova di fronte ad una serie o ad una seriazione, e questo perché i concetti di "qualità" e "quantità", essendo intuitivi, non sempre sono di facile determinazione. Pertanto, per superare i casi di incertezza occorre fare molta attenzione alla natura del fenomeno in esame.

Le serie statistiche sono in genere successioni di dati ordinati all'interno di una tabella secondo un certo criterio in cui ad ogni modalità viene associata la frequenza con cui questa si presenta nel gruppo di soggetti sui quali è stata effettuata la rilevazione. Per quanto riguarda, invece, le seriazioni, bisogna notare che gli elementi di ordinamento quantitativo delle tavole sono costituiti dai valori della variabile e dalle frequenze (cioè dal numero delle volte che l'intensità quantitativa si presenta). In genere, si usa indicare con X la variabile (le cui determinazioni saranno x_1, x_2, ... , x_n) e con Y le frequenze corrispondenti (y_1, y_2, ... , y_n).

Per tutti quanti i tipi di variabili l'elaborazione più immediata per organizzare i dati in un set omogeneo ed efficace di informazioni è costituita dalla *distribuzione di frequenza*, ovvero da una rappresentazione tabellare nella quale vengono illustrate le modalità con cui si presenta la variabile e le corrispondenti intensità.

A titolo di esempio, supponiamo di considerare uno studio compiuto sul periodo d'incubazione di una certa malattia XYZ. nella tavola 1 alla pagina seguente vengono riassunti i dati raccolti su un gruppo di soggetti di numerosità pari a n=40. I valori rappresentano il numero di giorni d'incubazione della malattia.

Ricordando la terminologia specificata in precedenza, vogliamo sottolineare che le osservazioni su cui è stata effettuata la rilevazione sono 40, la variabile oggetto dell'analisi è X="numero di giorni d'incubazione delle malattia XYZ", le cui determinazioni, come si può evincere dalla tabella, variano in un intervallo di numeri interi compresi tra 1 e 9. Si tratta evidentemente di una variabile discreta, la cui rappresentazione tabellare è realizzata associando ad ogni modalità il numero di frequenze con cui questa si presenta nei 40 soggetti.

---

1 Boldrini M.: "Statistica: teoria e metodi"; Giuffré, Milano, 1950.

2 La densità media delle frequenze si ottiene dividendo la frequenza totale della classe per la corrispondente ampiezza.

Soggetto n.	*Periodo incubaz. giorni*	*Soggetto n.*	*Periodo incubaz. giorni*
1	5	21	4
2	6	22	4
3	5	23	3
4	7	24	5
5	1	25	6
6	8	26	2
7	5	27	7
8	4	28	9
9	3	29	5
10	6	30	4
11	2	31	5
12	7	32	7
13	2	33	6
14	5	34	4
15	6	35	5
16	4	36	6
17	8	37	4
18	6	38	5
19	5	39	6
20	3	40	5

Tavola 1: periodo d'incubazione espresso in giorni della malattia XYZ in 40 pazienti (dati fittizi).

La seguente tavola illustra il risultato dell'operazione descritta.

Giorni incubaz. X	*Frequenze* Y
1	1
2	3
3	3
4	7
5	11
6	8
7	4
8	2
9	1
Totale	*40*

Tavola 2: distribuzione di frequenza del periodo d'incubazione espresso in giorni della malattia XYZ in 40 pazienti (dati fittizi).

Come è facile osservare, nella prima colonna della tavola 2 vengono riportate le determinazioni della variabile, mentre nella seconda il numero dei soggetti corrispondenti alle singole modalità. Il totale in basso nella seconda colonna deve necessariamente coincidere con il numero di individui sui quali è stata compita l'osservazione.

Questa elaborazione ci permette già di avere una prima visualizzazione delle informazioni provenienti dai dati contenuti nella tavola 1. È facile individuare, infatti, come il maggior numero di persone ha un periodo d'incubazione che dura 5 giorni (11 soggetti), mentre sono ben 26 i pazienti per i quali lo stesso evento si è verificato tra i 4 e i 5 giorni.

Sulla base della distribuzione di frequenza appena mostrata si possono eseguire ulteriori elaborazioni che permettono di integrare le informazioni provenienti dal gruppo di dati. Stiamo facendo riferimento alle *distribuzioni percentuali di frequenza* e alle *distribuzioni percentuali cumulate di frequenza*. Nella tavola 3 abbiamo aggiunto due colonne alla precedente distribuzione per illustrare il significato di queste due elaborazioni.

Giorni incubaz. X	*Frequenze* Y	*Percentuali*	*Percentuali cumulate*
1	1	2,5%	2,5%
2	3	7,5%	10,0%
3	3	7,5%	17,5%
4	7	17,5%	35,0%
5	11	27,5%	62,5%
6	8	20,0%	82,5%
7	4	10,0%	92,5%
8	2	5,0%	97,5%
9	1	2,5%	100,0%
Totale	*40*	*100,0%*	

Tavola 3: distribuzione di frequenza, percentuale e percentuale cumulata del periodo d'incubazione espresso in giorni della malattia XYZ in 40 pazienti (dati fittizi).

Per quanto riguarda la distribuzione percentuale di frequenza, lo scopo è quello di fornire (oltre al valore assoluto delle frequenze descritto dalla prima colonna) la composizione della distribuzione rispetto all'intero gruppo di informazioni. I valori compresi in questa colonna sono stati ottenuti dividendo le frequenze per il totale di colonna e presentando il risultato in forma di precentuale.

Se con la precedente distribuzione eravamo giunti a sapere che il maggior numero di soggetti ha un periodo d'incubazione che dura 5 giorni, adesso possiamo affermnare che questo gruppo rappresenta il 27,5% degli individui sui quali è stata compiuta la rilevazione.

Lla colonna contenente i valori della distribuzione percentuale cumulata è stata ottenuta dalla somma progressiva delle frequenze percentuali della colonna precedente. Le informazioni provenienti da questa ulteriore elaborazione sono di facile lettura e di notevole utilità. Dalla tavola 3, per esempio, si può agevolmente evincere che oltre

l'82% dei soggetti ha un periodo d'incubazione che non oltrepassa il 6° giorno, mentre il 65% dell'intero gruppo sviluppa la malattia tra il 4° ed il 6° giorno (percentuale ottenuta sottraendo dalla frequenza percentuale cumulata del 6° giorno quella relativa al 3° giorno).

Sia la distribuzione di frequenza che quella percentuale possono essere eseguite su tutti i tipi di variabile precedentemente elencati. Al contrario, l'elaborazione della distribuzione percentuale cumulata non è indicata per le variabili qualitative di tipo nominale, per le quali ha poco senso una tale elaborazione dei dati.

Non sempre i dati in possesso del ricercatore sono facilmente rappresentabili in una forma tabellare come quella appena illustrata. Nel caso della variabile età, per esempio, sarebbe poco efficace rappresentare le osservazioni compiute su un gruppo consistente di soggetti, in quanto le modalità da riportare in tabella sarebbero molto numerose. In queste occasioni è più utile costruire delle classi entro cui vengono comprese più modalità assunte dalla stessa varabile. Più specificamente, per *classe* intendiamo gli inervalli in cui è suddivisa una variabile in una tavola di frequenza.

La scelta dell'ampiezza e del numero di classi è generalmente arbitraria e dipende dalla natura dei dati e dagli scopi dell'indagine. Tuttavia, per decidere quale suddivisione in classi adottare, può essere utile la cosiddetta *regola di Sturges*, che fornisce, in funzione degli n dati, il numero di classi in cui è possibile raggrupparli, senza perdere eccessive informazioni.

La formula da utilizzare è la seguente:

$$K = 1 + 3{,}3 \log_{10} n$$

dove K è il numero delle classi in cui viene suddiviso il campo di variazione della variabile (che, lo ricordiamo, equivale alla distanza esistente tra il valore massimo ed il valore minimo della variabile), ed n è il numero complessivo delle osservazioni effettuate, ovvero il totale delle frequenze.

Pur essendo poco numerosi i dati, possiamo applicare la suddetta regola sulla distribuzione di frequenza illustrata nella tavola 2 per ricodificare in classi la variabile. Il campo di variazione della variabile X contiene valori compresi in un intervallo che ha come valore massimo 9 e valore minimo 1. La numerosità, lo ricordiamo, è pari a n=40. La formula sopra indicata diventa:

$$K = 1 + 3{,}3 \log_{10} 9 \cong 4$$

Il numero delle classi in cui dovrebbe essere suddiviso il campo di variazione è dunque pari a 4. Ne segue che l'ampiezza di ogni classe risulta:

$$\frac{(X_{max}-X_{min})}{K} = \frac{(9-1)}{4} = \frac{8}{4} = 2$$

La prima di esse è definita dai limiti 1 e 1+2=3; le seguenti vengono definite sulla base dello stesso criterio. La tavola 4 illustra il risultato della procedura descritta sui dati della tavola 2.

Giorni incubaz. X	*Frequenze* Y
1 - 3	4
3 - 5	10
5 - 7	19
7 - 9	7
Totale	*40*

Tavola 4: distribuzione di frequenza del periodo d'incubazione espresso in giorni della malattia XYZ in 40 pazienti (variabile in classi) (dati fittizi).

Se prescindiamo da questo metodo automatico per la suddivisione in classi del campo di variazione relativo alle intensità di un certo fenomeno, si pone il problema della scelta ragionata dei limiti delle classi, ovvero della fissazione dei limiti in modo che non sorgano dubbi sulla collocazione dei casi che stanno proprio a cavallo di due classi.

La tavola 4 presenta l'evidenza di questo problema. Il limite superiore della prima classe, per esempio, coincide con quello inferiore della seconda. Questa situazione rende difficile collocare le frequenze relative a questa modalità in una delle due classi senza che sorgano dubbi ed ambiguità. Tale problema viene risolto utilizzando un simbolo (-) che indica l'esclusione del limite superiore dalle modalità della variabile. In questo modo nella prima classe della tabella sono contenute solo le frequenze corrispondenti alle modalità "1 o 2 giorni di incubazione".

Prima di concludere, volgiamo mettere in evidenza la diversa ampiezza dell'ultima classe della tabella. La regola di Sturges fornisce una divisione del campo di variazione che, a causa delle approssimazioni nei decimali, non sempre restituisce una classe finale di uguale ampiezza rispetto alle precedenti. In queste situazioni, pur di non lasciare un limite aperto (situazione da evitare, quando possibile), è opportuno fissare quale limite superiore della classe il valore massimo del campo di variazione.

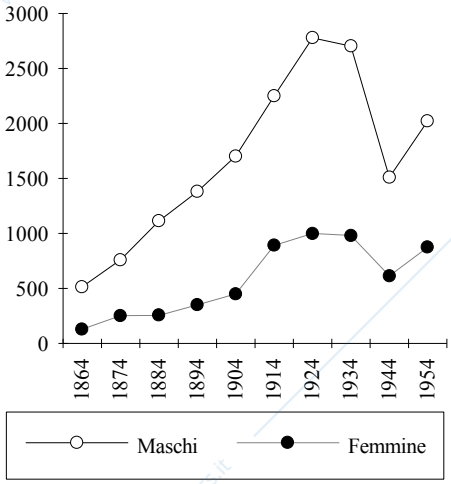
Rappresentazioni grafiche

Se il primo passo per eseguire le più elementari statistiche descrittive è quello di elaborare le tabulazioni di frequenza, il successivo è rappresentato senza dubbio dalle rappresentazioni grafiche, attraverso cui viene visualizzato l'andamento delle variabili esaminate. Anche in questo caso, non ci stanchiamo di ripetere che il tipo di variabile da analizzare determina la scelta della rappresentazione più opportuna.

Esaminiamo per primi gli strumenti grafici più frequenti per visualizzare i dati di una serie statistica. La rappresentazione grafica più semplice di una serie è fornita dal *diagramma a punti* che consiste nell'individuare sul piano cartesiano, i punti corrispondenti alle coppie di valori coordinati x,y. La seguente tabella ed il relativo grafico costituiscono un esempio di tale rappresentazione.

Anni	*Suicidi*	
	M	F
1864	516	130
1874	762	253
1884	1115	255
1894	1381	351
1904	1705	451
1914	2249	891
1924	2778	1000
1934	2703	983
1944	1510	615
1954	2023	874

Tavola 5: morti per suicidio secondo il sesso tra il 1984 ed il 1954 (dati fittizi).



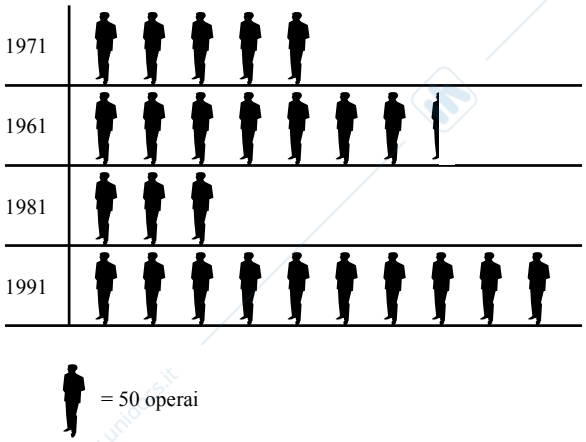
Come si può osservare, nell'asse orizzontale vengono riportate le modalità assunte dalla variabile, mentre in quello verticale sono indicate le frequenze con cui le modalità sono presenti, in relazione alle due serie di dati (maschi e femmine).

Un altro modo molto suggestivo di rappresentare i dati di una serie statistica è rappresentato dall'*ideogramma*. Questo strumento visivo si avvale di figure schematizzate rappresentanti direttamente o simbolicamente gli aspetti elementari dei fenomeni che si vogliono trattare. Per indicare l'ammontare complessivo di un fenomeno occorre riprodurre un numero adeguato di figure di uguali dimensioni, allineate in modo opportuno, ad una uguale distanza tra di loro. La seguente tabella ed il relativo grafico forniscono un esempio di rappresentazione mediante ideogramma. Per individuare l'unità di misura viene scelto un simbolo che in corrispondenza di ogni anno viene ripetuto un numero di volte pari al rapporto tra la frequenza corrispondente a quella classe ed il peso assegnato all'unità di misura. L'ideogramma ci fornisce

immediatamente l'impressione dell'andamento del fenomeno oltre ad una corretta impressione visiva circa la frequenza con cui i dati si presentano nella tabella.

Anni	*Frequenze*
1961	253
1971	387
1981	168
1991	508

Tavola 6: operai licenziati nel settore edile tra il 1961 ed il 1991 (dati fittizi).

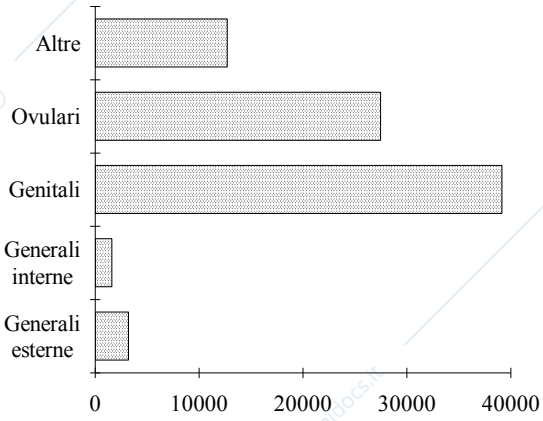


L'ideogramma è comunque una rappresentazione grafica da utilizzare quando non ci si rivolge ad un pubblico di "addetti ai lavori". In questi casi, è più opportuno ricorrere ad altri tipi di grafici.

Una di queste rappresentazioni è data dal *diagramma a nastri*, nel quale, al fine di valutare il significato statistico della lunghezza dei nastri, è necessario affiancare una scala di misura di riferimento. Occorre, innanzi tutto, organizzare una graduatoria arbitraria delle modalità della variabile lungo un'asse (preferibilmente quello verticale) che ci consentirà di individuare dei punti equidistanti in corrispondenza dei quali vengono indicate le diverse modalità del carattere in esame. Su un secondo asse, orizzontale e perpendicolare al primo, si rappresenterà un'opportuna scala metrica che indica le frequenze assunte dal fenomeno considerato.In corrispondenza dei punti del primo asse si tracciano dei nastri paralleli al secondo, la cui lunghezza è pari alla frequenza relativa a quella modalità. Nella seguente tabella viene riportato un esempio, a cui il grafico a fianco fa riferimento.

Cause	*Aborti*
Generali esterne	3282
Generali interne	1677
Genitali	39215
Ovulari	27536
Altre	12781

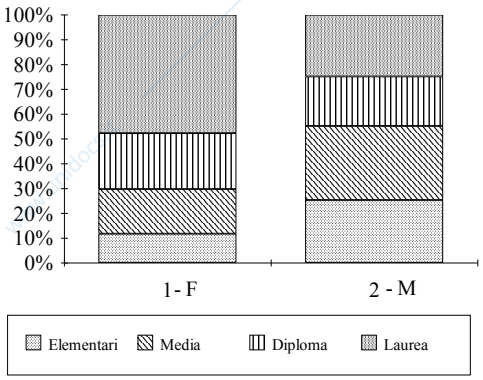
Tavola 7: distribuzione delle cause di aborto (dati fittizi).



Un ulteriore tipo di grafico è il *diagramma a colonne*, per il quale è necessario eseguire non solo le distribuzioni di frequenza, ma anche quelle percentuali. Sulla base dei dati contenuti nella seguente tabella sono stati costruiti dei rettangoli di uguale base e di altezza proporzionale all'intensità del fenomeno da esaminare. A volte, come viene mostrato nel grafico, al fine di evidenziare la composizione percentuale delle serie è possibile suddividere i rettangoli in parti proprozionali all'ammontare delle frequenze.

Scolarità	*M*		*F*	
	freq.	*%*	*freq.*	*%*
Elementari	159	47,7%	121	24,9%
Media	75	22,5%	97	20,0%
Diploma	60	18,0%	145	29,8%
Laurea	39	11,7%	123	25,3%
Totale	333	100%	486	100%

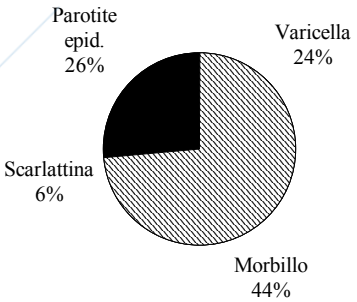
Tavola 8: distribuzione della scolarità secondo il sesso (dati fittizi).



Per concludere la rassegna degli strumenti grafici a disposizione del ricercatore per la presentazione delle serie statistiche, vogliamo ricordare il *diagramma a torta* con il quale è possibile evidenziare sia il valore assoluto di un certo fenomeno che la sua composizione percentuale. Non occorrendo ulteriori spiegazioni su questo strumento di larga diffusione, rimandiamo all'osservazione della seguente tabella per un esempio della metodologia.

Malattie	*Frequenze*
Varicella	36623
Morbillo	68246
Scarlattina	9848
Parotite epid.	40949

Tavola 9: distribuzione della malattie in gruppo di bambini minori di 10 anni (dati fittizi).



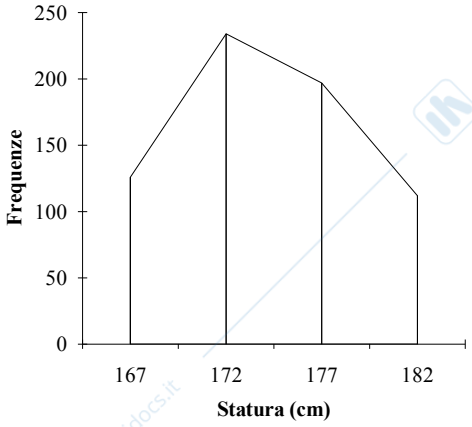
Per quanto riguarda le rappresentazioni grafiche più opportune per la visualizzazione di una seriazione statistica gli strumenti a disposizioni non sono molti e quasi tutte le elaborazioni grafiche possono essere fatte attraverso il cosiddetto *istogramma*. Nel seguente esempio abbiamo rappresentato una tipica seriazione la cui distribuzione è

divisa in classi. Tale esempio ci permetterà di illustrare i possibili strumenti grafici da utilizzare per visualizzarne i dati, a partire dalla rappresentazione più semplice che è data dal *diagramma a bastoni*.

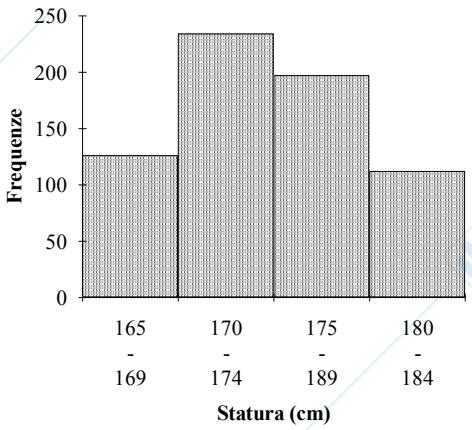
Classi di statura (cm.)	*Valori centrali*	*Frequenze*
165 - 169	167	126
170 - 174	172	234
175 - 189	177	197
180 - 184	182	112
	Totale	*669*

Tavola 10: distribuzione di frequenza dell'altezza rilevata in un gruppo di soggetti (variabile in classi) (dati fittizi).

Il suddetto grafico viene costruito innalzando, in corrispondenza di ogni valore della variabile o al valore centrale delle classi individuate (come nel caso del nostro esempio), parallelamente all'asse verticale un segmento che si innalza per una lunghezza pari alla frequenza ad esso associata. Se congiungiamo con una spezzata i vertici dei segmenti tracciati con il diagramma a bastoni, otteniamo il *poligono di frequenza*. Il grafico a fianco illustra un esempio degli strumenti citati sulla base dei dati della tavola 10.



Da questa rappresentazione non è difficile passare all'istogramma. In quest'ultimo tipo di grafico vengono utilizzati dei rettangoli, in corrispondenza di ogni classe, aventi una superficie proporzionale alle rispettive frequenze. È opportuno che le classi abbiano uguale ampiezza, per non incorrere in rappresentazioni poco coerenti. Qualora ciò non dovesse succedere, si dovrebbero utilizzare al posto delle frequenze le rispettive *densità medie delle frequenze*[2]. Il grafico a fianco illustra un esempio di istogramma ottenuto sui dati della tavola 10.



2 La densità media delle frequenze si ottiene dividendo la frequenza totale della classe per la corrispondente ampiezza.

Indici di tendenza centrale

Le distribuzioni di frequenza ed i grafici hanno lo scopo rispettivamente di organizzare i dati e visualizzarne l'andamento. Gli indicatori di tendenza centrale, di cui ci occuperemo in questo paragrafo, assolvono il compito di sintetizzare le informazioni provenienti da essi, riassumendo in pochi valori gli aspetti rappresentativi del gruppo di osservazioni.

Il più comune indicatore di tendenza centrale è il valore medio di una distribuzione. Accade spesso di identificare un intero fenomeno con un singolo indicatore, la media appunto, proprio perchè attorno ad esso si concentrano il maggior numero di frequenze. In questi casi sintetizziamo le informazioni relative al gruppo, interpretandole attraverso un valore che è rappresentativo di tutte quante.

È necessario essere consapevoli che il concetto di media è molto ampio e che gli indicatori a disposizione per rappresentare il valore "centrale" di una distribuzione sono molteplici. La loro scelta è nuovamente condizionata sia dal tipo di variabile su cui eseguire l'elaborazione (qualitativa/quantitativa), che dalla sua distribuzione all'interno del campo di variazione.

Il primo indicatore di tendenza centrale che vogliamo illustrare è la *moda* (il cui simbolo è Mo). Formalmente, se indichiamo con X un certo carattere e con x_i (i=1, 2, ... , n) le intensità che può assumere, la moda (o norma, o valore modale) di X è data dall'intensità x_i che si presenta con la massima frequenza. Se ad esempio consideriamo i valori 2, 2, 3, 4, 4, 5, 5, 6, 6, 6, 6, 7, 10 è facile vedere che la moda è pari a 6, poichè si tratta del valore che tra quelli esaminati si è presentato più volte.

Più problematica e meno precisa è la determinazione del valore modale nel caso in cui la variabile è continua e organizzata in classi. Occorre, innanzi tutto, disporre di una distribuzione caratterizzata da classi di ampiezza omogenea, alla quale, se così non fosse, si dovrebbe giungere attraverso opportuni artifici. La seguente tabella mostra un esempio in cui si evince facilmente che il valore modale è contenuto nella classe 200-299.

Classi	*Frequenze*
100 - 199	1971
200 - 299	4728
300 - 399	3280
400 - 499	1143
Totale	*11122*

Tavola 11: esempio di distribuzione di frequenza di una variabile in classi (dati fittizi).

Ma qual'è il valore nella distribuzione a cui corrisponde esattamente la moda ? Tale problema non si presenta, per esempio, nel caso della tavola 12, nella quale il valore modale viene facilmente individuato in corrispondenza della massima frequenza (3 stanze).

Stanze *X*	*Abitazioni* *Y*
1 stanza	533
2 stanze	857
3 stanze	4321
4 stanze	4127
5 stanze	2109
6 stanze	1864
7 stanze e oltre	901
Totale	*14712*

Tavola 12: distribuzione delle abitazioni in relazione al numero di stanze (dati fittizi).

Nel caso precedente è opportuno ricorrere ad una particolare formula che permetta di individuare all'interno della classe modale una stima del valore da assumere quale moda[3]. La moda è un efficace indicatore di tendenza centrale quando si verifica un'elevata concentrazione di frequenze nei pressi della classe modale. Nel caso della tavola 12 osserviamo come questo vincolo non sia del tutto rispettato, in quanto due classi hanno un volume di frequenze molto simile. In situazioni come questa l'uso della moda è sconsigliabile a favore di altri metodi che vedremo in seguito.

Una seconda media di posizione (termine utilizzato come sinonimo di valore di tendenza centrale) è la *mediana* (il cui simbolo è Me), anch'essa utilizzata per descrivere in forma sintetica un determinato fenomeno. Se disponiamo di n osservazioni x_1, x_2, ... , x_n di una variabile X e le ordiniamo in modo non decrescente, cioè tali che:

$$x_1 \leq x_2 \leq \ldots\ldots \leq x_n$$

3 Per individuare la moda all'interno della classe modale la formula è la seguente: $Md = x_0 + \frac{d}{2}\frac{y_1 - y_{-1}}{2y_0 - y_1 - y_{-1}}$

dove :
- x_0 = valore centrale della classe modale
- x_1 = valore centrale della classe successiva a quella modale
- x_{-1} = valore centrale della classe precedente a quella modale
- y_0 = frequenza corrispondente alla classe modale
- y_1 = frequenza corrispondente alla classe successiva a quella modale
- y_{-1} = frequenza corrispondente alla classe precedente a quella modale
- d = ampiezza della classe modale

possiamo definire mediana quell'indicatore che divide la graduatoria in due parti uguali. Occorrono, però, due differenti formula applicative, da adottare in relazione alla numerosità dei dati. Se n è dispari la mediana è rappresentata dal termine centrale della successione, individuato dalla formula (n+1)/2. Se invece n è pari allora non possimo determinare un valore centrale unico, ma piuttosto una coppia di valori mediani. Nella pratica è consuetudine accettare come mediana la semisomma dei termini della successione che occupano i posti n/2 e (n/2)+1. Se, per esempio, uno studente ha sostenuto sette esami, riportando una serie di voti pari a 23, 24, 26, 25, 24 ,27, 28, la mediana viene calcolata ordinando innanzi tutto i dati, per ottenere la successione:

23, 24, 24, 25, 26, 27, 28

Siccome n è dispari il voto mediano sarà quello che occupa il (n+1)/2=4° posto, corrispondente al voto 25. Se invece lo stesso studente avesse sostenuto dieci esami, riportando voti pari a 23, 24, 26, 25, 24, 27, 28, 27, 28, 26, la successione ordinata diverebbe:

23, 24, 24, 25, 26, 26, 27, 27, 28, 28

in cui, essendo n pari, la mediana risulterà dalla semisomma dei termini che occupano il 5° ed il 6° posto, cioè Me=26.

Il calcolo della mediana in una distribuzione di frequenza è lievemente più complesso, pur seguendo gli stessi criteri finora indicati[4], soprattutto se la variabile è di tipo quantitativo continuo.

Il valore di tendenza centrale più utilizzato è certamente la *media aritmetica* (il cui simbolo è m) che, assieme ad altri indicatori che in seguito illsutreremo, fa parte delle cosiddette *medie analitiche*. Se una certa variabile X assume n determinazioni x_1, x_2, ... , x_n, si può definire la media aritmetica secondo la seguente formula:

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

[4] Se la variabile è discreta, per esempio, occorre costruire la *distribuzione cumulata delle frequenze* (costruita in modo analogo a quella delle frequenze percentuali) e trovare con i criteri indicati il valore mediano, il quale indicherà la classe a cui corrisponde la mediana

Se invece disponiamo di una distribuzione di frequenza in cui alle intensità x_i si affiancano le frequenze y_i, nel calcolo della media è opportuno tenere in considerazione anche queste ultime ed utilizzare la formula della *media aritmetica ponderata*, cioè:

$$m = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sum_{i=1}^{n} y_i}$$

Qualora nella distribuzione di frequenza la variabile sia espressa in classi, al posto delle determinazioni x_i dobbiamo utilizzare per il calcolo della media i corrispondenti valori centrali delle classi. La seguente tavola riporta un esempio del calcolo della media ponderata su una distribuzione di frequenza con variabile in classi.

Classi di addetti	*Imprese*	*Valori centrali Xi*	*Xi*Yi*
3 - 5 addetti	15	4	60
6 - 8 addetti	21	7	147
9 - 11 addetti	8	10	80
12 - 14 addetti	3	13	39
Totale	*47*		*326*

Tavola 13: distribuzione delle aziende in relazione al numero di addetti (dati fittizi).

$$m = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sum_{i=1}^{n} y_i} = \frac{326}{47} = 6,94$$

In genere la media aritmetica viene utilizzata quando tra le grandezze in esame esiste un legame di natura additiva. Esistono altre medie (appartenenti ancora al gruppo delle medie analitiche) da utilizzare quando tra le grandezze esistono altri tipi di legami.
La *media geometrica* (che si indica con m_o) si applica di fronte a grandezze in cui vi è un comportamento di tipo moltiplicativo. La media geometrica si definisce come la radice ennesima (presa con segno positivo) del prodotto di n termini positivi, ovvero[5]:

$$m_o = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Nel cso delle distribuzioni di frequenza la media geometrica ponderata risulta dalla formula :

5 L'operatore aritmetico introdotto è noto con il nome di moltiplicatoria ed equivale al prodotto di n fattori, per cui:

$$\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdot \ldots \cdot a_n$$

$$m_o = \sum_{i=1}^{n} y_i \sqrt{\prod_{i=1}^{n} x_i^{y_i}}$$

In generale conviene ricorrere alla media geometrica quando occorre utilizzare uno strumento più sensibile ad una variazione assoluta dei termini bassi piuttosto che ad una eguale variazione dei termini alti (proprietà assente nella media aritmetica).

Come ultimo indicatore analitico del valore centrale di una distribuzione vogliamo ricordare la *media armonica*, il cui simbolo e m_{-1}, la quale, disponendo di n termini diversi da zero (x_1, x_2, ... , x_n), viene ottenuta dalla formula:

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

che nel caso delle distribuzioni di frequenza diventa :

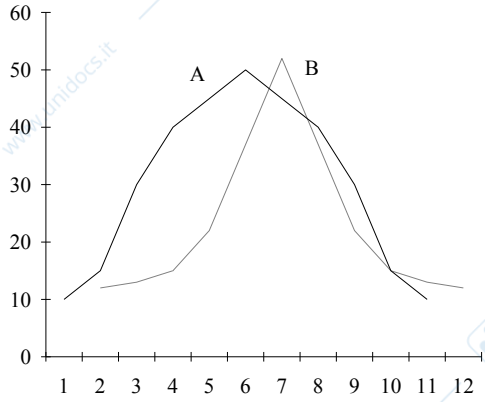
$$m_{-1} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \frac{y_i}{x_i}}$$

È opportuno utilizzare la media armonica quando si dispone di dati per i reciproci dei quali si dovrebbe calcolare la media.

Per concludereè opportuno ribadire che la scelta della misura di tendenza centrale da utilizzare, oltre che dipendere dalle proprietà dei dati, deve essere valutata principalmente in relazione al tipo delle variabili. Mentre su una grandezza di tipo quantitativo (sia continuo che discreto) possono essere utilizzati tutti gli strumenti analitici illustrati in questo paragrafo, per dati qualitativi ha poco senso calcolare, per esempio, una media o una mediana. Volendo ampliare il concetto di misura centrale, tra gli indicatori visti solo la moda potrebbe essere utilizzata su variabili di tipo qualitativo, limitandone l'interpretazione all'individuazione della classe in cui si concentrano il maggior numero delle frequenze.

Misure di variabilità

Le misure di variabilità, assieme a quelle di tendenza centrale, permettono di descrivere in modo più ampio e completo la distribuzione di un gruppo di valori. Osservando una qualsiasi distribuzione di frequenza, si può osservare come un buon numero di casi siano distanti in varia misura dal valore centrale che è stato adottato per rappresentarli. Una media, infatti, non indicherà mai in quale misura tali dati si distribuiscono attorno al valore centrale della distribuzione. Due campioni, per esempio, pur avendo una media molto simile, possono essere molto differenti nel modo in cui le frequenze si distribuiscono attorno ad esse. Il seguente grafico ci dimostra come, benché i due fenomeni analizzati abbiano un valore centrale abbastanza simile, i dati hanno in realtà distribuzioni molto differenti tra di loro. Nella curva A, ad esempio, si può osservare un folto gruppo di osservazioni distanti dal valore centrale m_1, mentre nella curva B la maggior parte delle determinazioni si concentra attorno alla media m_2. Un qualsiasi indicatore di tendenza centrale non sarebbe stato in grado di evidenziare questa differenza tra le due distribuzioni. Quindi per completare il set di strumenti per analizzare un gruppo di osservazioni occorrono anche degli indici che misurino la variabilità, ovvero la dispersione dei dati attorno al valore centrale.



L'indice di variabilità più semplice (che abbiamo già nominato precedentemente) è il *campo di variazione*, ottenibile sottraendo il valore minimo delle determinazioni della variabile X dal suo valore massimo. Tale indice, dipendendo solo dal primo e dall'ultimo termine della successione di osservazioni, non tiene conto dei termini intermedi e dunque della loro concentrazione attorno al valore medio.

Per superare questa grave carenza occorre focalizzare l'attenzione sulla distanza dei dati dal valore centrale individuato nella distribuzione. Una formula molto elementare che tiene conto di questo aspetto è quella che comunemente viene utilizzata per calcolare lo *scostamento semplice medio* dalla media aritmetica, che, disponendo di una successione $x_1, x_2, ... , x_n$ di n determinazioni della variabile X, risulta essere:

$$S_m = \frac{\sum_{i=1}^{n} |x_i - m|}{n}$$

In questa formula, come è facile osservare, si calcola la somma delle distanze in valore assoluto delle osservazioni dalla media aritmetica della successione, dividendo il risultato per la numerosità dei dati. Nel caso di distribuzioni di frequenza, la formula diventa:

$$S_m = \frac{\sum_{i=1}^{n} |x_i - m| y_i}{\sum_{i=1}^{n} y_i}$$

Se il concetto di base che viene utitlizzato per calcolare la variabilità è sempre quello di misurare la distanza delle singole osservazioni dal loro valore medio, la formula più utilizzata è senza dubbio la *deviazione standard* (o scostamento quadratico medio). Questo indicatore analitico è molto simile al precedente ma, al posto del valore assoluto, troviamo l'elevamento alla seconda potenza della differenza più volte citata. La formula da utilizzare per il suo calcolo è la seguente:

$$S = \sqrt[2]{\frac{\sum_{i=1}^{n} (x_i - m)^2}{n}}$$

L'estrazione della radice è stata inserita per ridimensionare l'effetto dell'elevamento dei dati sulla dimensione delle grandezze esaminate. Nel caso delle distribuzioni di frequenza la formula diviene:

$$S = \sqrt[2]{\frac{\sum_{i=1}^{n} (x_i - m)^2 y_i}{\sum_{i=1}^{n} y_i}}$$

dove lo ricordiamo, le y_i sono le frequenze associate alle determinazioni x_i della variabile X. Il quadrato della deviazione standard prende il nome di *varianza*, della quale presentiamo di seguito entrambe le formule.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n} \qquad S^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2 y_i}{\sum_{i=1}^{n} y_i}$$

È opportuno tenere presente che le formule della deviazione standard e della varianza sopra descritte, avendo al denominatore la quantità n, devono riferirsi a calcoli effettuati sulla popolazione intera delle possibili osservazioni. Quando, invece, si utilizzano gruppi di osservazioni o campioni di dimensione limitata è consigliabile porre al denominatore delle due formule la quantità (n-1). Infatti, se n è molto grande, per calcolare i suddetti indici dividere per n o per (n-1) fornirà risultati molto simili; al contrario, se n è piccolo tale approssimazione non è più accettabile, per cui occorre dividere le somme degli scarti per (n-1). Le suddette formule diventano :

$$S = \sqrt[2]{\frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}} \qquad S^2 = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}$$

Per concludere la rassegna degli strumenti analitici a disposizione per misurare la variabilità, illustriamo il *coefficiente di variazione*. Si ricorre a questo indice non solo quando vogliamo confrontare fenomeni che differiscono per l'ordine di grandezza, ma anche per rendere comparabili fenomeni espressi in differenti unità di misura. Il coefficiente di variazione è un indice di variabilità relativa poiché ottenuto rapportando un indice assoluto (la deviazione standard) con la consistenza media del fenomeno. Il risultato è un numero adimensionale che permette paragoni e confronti tra distribuzioni con caratteristiche differenti. La formula da utilizzare è la seguente:

$$V = \frac{S}{m}$$

in cui, come ormai è noto, S è la deviazione standard ed m è la relativa media aritmetica. Il seguente esempio riguarda un'applicazione dei principali indicatori di variabilità illustrati.

Supponiamo di aver effettuato un misurazione della statura (espressa in centimetri) su un gruppo di n=10 individui e di aver raccolto i seguenti valori:

$x_1 = 153$	$x_6 = 184$
$x_2 = 163$	$x_7 = 171$
$x_3 = 169$	$x_8 = 158$
$x_4 = 174$	$x_9 = 166$
$x_5 = 178$	$x_{10} = 181$

Per prima cosa si procede al calcolo della media aritmetica, che utilizzando la formula nota risulta essere pari a m=169,7 cm. Si calcola, poi, la somma degli scarti al quadrato delle osservazioni dalla media, secondo il seguente procedimento:

$$\sum_{i=1}^{10}(x_i - m)^2 = (x_1 - m)^2 + \ldots + (x_{10} - m)^2 = (153 - 169{,}7)^2 + \ldots + (181 - 169{,}7)^2 = 896{,}1$$

Trovata la suddetta quantità non rimane che sostituirla nelle formule relative alla deviazione standard ed alla varianza per ottenere il risultato cercato:

$$S = \sqrt[2]{\frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}} = \sqrt[2]{\frac{896{,}1}{9}} = 9{,}978 \qquad S^2 = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1} = \frac{896{,}1}{9} = 99{,}567$$

Sulla base dei valori ottenuti è possibile ricavare anche il valore del coefficiente di variazione, che risulta pari a:

$$V = \frac{S}{m} = \frac{9{,}978}{169{,}7} = 0{,}059$$

PROBABILITÀ E VARIABILI CASUALI

Nozioni elementari di calcolo delle probabilità

Prima di procedere oltre con l'illustrazione delle metodologie statistiche per l'analisi e l'interpretazione dei dati, occorre soffermarci su alcuni concetti fondamentali di calcolo delle probabilità. Parte delle nozioni contenute nelle prossime pagine costituiscono la base delle teorie statistiche di cui parleremo in seguito.
Sintetizzando, si può affermare che tutta la teoria probabilistica poggia su tre concetti elementari:

1. la *prova*, o esperimento;
2. l'*evento*, o risultato;
3. la *probabilità*.

Questi concetti sono strettamente legati tra di loro, a tal punto che possiamo riassumerli in un'unica frase: *la prova genera l'evento con una certa probabilità*. Tirare una moneta, in realtà, è una prova (esperimento); il fatto che il risultato sia "testa" costituisce un evento, a cui è associata una probabilità di realizzazione. A proposito di quest'ultimo concetto occorre precisare che la probabilità è rappresentata da un numero, il quale esprime il livello di fiducia che vogliamo assegnare alla realizzazione di un risultato, ovvero al verificarsi di un evento nel contesto di una prova.
Sempre considerando l'esempio della moneta, supponiamo di effettuare N lanci nelle medesime condizioni e supponiamo che durante queste prove l'evento "testa" si sia presentato un numero n di volte. Sotto queste condizioni è ovvio che:

$$0 \leq n \leq N$$

La quantità di eventi favorevoli (testa) viene anche detta *frequenza assoluta*. Se dividiamo la precedente disuguaglianza per il numero complessivo di prove effettuate otteniamo la cosiddetta *frequenza relativa* dell'evento considerato, cioè:

$$0 \leq \frac{n}{N} \leq 1$$

La frequenza relativa potrebbe essere considerata come una misura empirica della probabilità. Se aumentiamo in modo consistente il numero N delle prove, allora possiamo affermare che l'esattezza di questa misura tende a crescere.

In questo corso adotteremo l'*impostazione frequentista* delle teorie probabilistiche, secondo la quale il riprodursi di un evento ripetibile un numero di volte molto elevato, porta la frequenza relativa verso una stabilità che cresce con il numero delle prove, fino ad arrivare ad un valore costante. Questa definizione abbastanza approssimativa di probabilità può anche essere riformulata nel seguente modo:

$$\Pr(E) = \lim_{N\to\infty} \frac{n}{N}$$

dove E è l'evento favorevole che attendiamo dall'esperimento, N il numero complessivo di volte in cui viene ripetuta tale prova, n è il numero di volte in cui si verifica E ed infine n/N è la frequenza relativa dell'evento E. Dunque, facendo divergere il numero delle prove la frequenza relativa di un evento coincide con la sua probabilità di realizzarsi. Poiché, come abbiamo visto, la frequenza relativa restituisce un valore compreso tra 0 e 1, anche la probabilità gode di questa proprietà, cioè:

$$0 \le \Pr(E) \le 1$$

Agli estremi di questo intervallo la probabilità assume due significati ben precisi. Se Pr(E) è uguale a zero parleremo di evento "impossibile", mentre se Pr(E) è pari all'unità, allora ci troviamo di fronte ad un evento certo.
Se consideriamo tutti i possibili eventi E_i (i = 1, 2, ... , k) in un esperimento che si articola in N prove, avremo:

$$n_1 + n_2 + \ldots + n_k = N$$

e se dividiamo l'uguaglianza per N otteniamo:

$$\frac{n_1}{N} + \frac{n_2}{N} + \ldots + \frac{n_K}{N} = 1$$

Al divergere di N al posto delle frequenze relative possiamo sostituire la probabilità e scrivere:

$$\Pr(E_1) + \Pr(E_2) + \ldots + \Pr(E_k) = 1$$

L'insieme degli eventi possibili in un esperimento è indicato con il termine *spazio degli eventi*. In questo spazio ogni elemento viene messo in corrispondenza biunivoca con tutti o possibili esiti dell'esperimento stesso. Nel caso del lancio di due monete lo spazio degli eventi risulta essere il seguente, dove T indica l'evento "testa" e C l'evento "croce":

$$(T\ ,\ T) \qquad (T\ ,\ C)$$
$$(C\ ,\ T) \qquad (C\ ,\ C)$$

Come è possibile osservare, questi elementi rappresentano tutti i possibili risultati dell'esperimento consistente nel lancio di due monete.
Sulla base di questi concetti primitivi vengono definiti una serie di principi appartenenti alla teoria probabilistica.
Consideriamo un esperimento in cui si effettuano N prove. In questo esperimento si verificano due eventi A e B incompatibili[6] tra di loro, che si presentano rispettivamente n_A e n_B volte. Il *principio delle probabilità totali* riguarda l'esame dell'evento A+B, che, tradotto a parole, equivale ad un risultato che è costituito da A o da B. Tale risultato si presenterà un numero di volte pari a $n_A + n_B$ volte, per cui la sua frequenza relativa sarà uguale alla quantità $(n_A + n_B)/N$. Sulla base di queste considerazioni elementari possiamo scrivere:

$$\frac{(n_A + n_B)}{N} = \frac{n_A}{N} + \frac{n_B}{N}$$

Se facciamo divergere all'infinito il numero delle N prove, ricordando quanto scritto in precedenza, risulterà:

$$Pr(A + B) = Pr(A) + Pr(B)$$

Quest'ultima relazione esprime appunto il principio delle probabilità totali (o della somma) per due eventi incompatibili.

6 Due eventi si dicono incompatibili quando il verificarsi dell'uno esclude l'altro (p.e.: nel lancio di una moneta il risultato può essere o testa o croce).

Variabili casuali e curve di probabilità

Le nozioni elementari di teoria probabilistica contenute nelle precedenti pagine sono basilari per alcuni argomenti che tratteremo durante il corso di elementi di statistica per Tecnici di Laboratorio Biomedico. Tali nozioni, infatti, sono fondamentali per la definizione delle variabili casuali.
Si definisce *variabile casuale discreta* una grandezza che assume i valori $x_1, x_2, \ldots, x_n$ a cui vengono associate le rispettive probabilità $p_1, p_2, \ldots, p_n$. La somma di queste probabilità deve essere pari ad 1, ovvero:

$$\sum_{i=1}^{n} p_i = 1$$

In definitiva, una variabile casuale è rappresentata da un insieme di coppie di valori del tipo:

$$(x_1, p_1); (x_2, p_2); (x_3, p_3); \ldots (x_n, p_n)$$

dove il primo valore indica la determinazione della grandezza X, mentre il secondo la probabilità che viene associata al verificarsi dell'evento. Per chiarire questo concetto osserviamo il seguente esempio che illustra la distribuzione di una variabile casuale discreta rappresentata dai risultati ottenuti lanciando tre volte consecutivamente la stessa moneta. Se, dunque, l'esperimento è il lancio della moneta, l'evento atteso è il presentarsi della faccia corrispondente a "croce". La variabile X, che descrive l'esperimento, esprimerà il numero di "croci" che possono apparire nei tre lanci. Tale numero non è noto a priori, ma possiamo con certezza affermare che i possibili risultati dell'esperimento variano in uno spazio discreto compreso tra 0 e 3. In questo spazio si trovano le determinazioni x_i della variabile X. Per individuare le probabilità associate ad ogni singolo evento è opportuno costruire lo *spazio campionario originario* illustrato nella seguente tavola.

Probabilità	*Eventi*
(1/2)*(1/2)*(1/2)=1/8	C,C,C
(1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8	C,C,T C,T,C T,C,C
(1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8	C,T,T T,C,T T,T,C
(1/2)*(1/2)*(1/2)=1/8	T,T,T

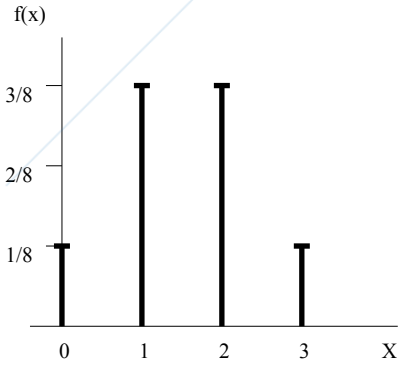
Nella prima colonna sono state calcolate le singole probabilità attraverso cui si realizzano i possibili aventi dell'esperimento, mentre nella seconda sono descritti i tipi di risultati che è possibile ottenere. Nel primo blocco di risultati (C,C,C) l'evento "croce" si presenta in tutti e tre i lanci, dunque, la determinazione della variabile X assumerà il suo valore massimo, cioè 3. Nel blocco successivo gli eventi "croce" nei tre lanci sono due e questo sarà il valore assunto dalla x_i. Procedendo secondo questo criterio si ottiene lo *spazio campionario ridotto*, nel quale sono contenuti i valori assunti dalla variabile e le probabilità ad essi associate.

Spazio campionario originario		*Spazio campionario ridotto*	
Probabilità	*Eventi*	*X*	*P*
(1/2)*(1/2)*(1/2)=1/8	C,C,C	3	(1/8)=0,125
(1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8	C,C,T C,T,C T,C,C	2	(3/8)=0,375
(1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8 (1/2)*(1/2)*(1/2)=1/8	C,T,T T,C,T T,T,C	1	(3/8)=0,375
(1/2)*(1/2)*(1/2)=1/8	T,T,T	0	(1/8)=0,125
		Totale	1

Dalla lettura della tavola si può leggere come la determinazione $x_1=3$ ha una probabilità di verificarsi pari a 0,125, la determinazione $x_2=2$ ha una probabilità $p_2=0,375$ e così via. La somma delle probabilità è pari ad uno, come richiede la definizione.

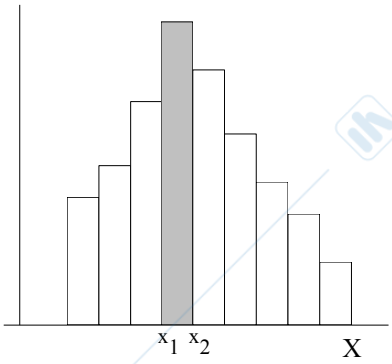
La variabile casuale appena descritta individua una *funzione di massa*, che graficamente è rappresentata da un diagramma a bastoni (vedi la figura seguente) in cui la somma delle asticelle che vengono innalzate in corrispondenza delle determinazioni della variabile deve essere pari ad uno. Siccome parliamo di funzione, al posto del simbolo che individua le probabilità (p_i) siamo autorizzati a sostituire quello di $f(x_i)$, in quanto valore associato alla variabile X. Il lancio delle tre monete genera la seguente funzione di massa , nel quale la somma delle ordinate è illustrata a fianco.

$$\sum_{i=1}^{4} f(x_i) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{8}{8} = 1$$



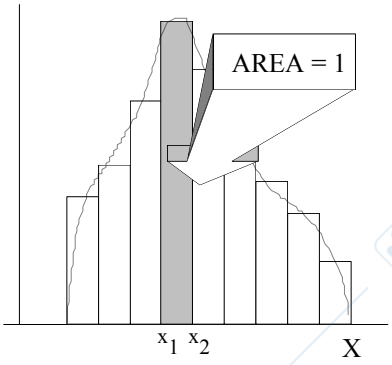
Finora abbiamo parlato di variabili casuali di tipo discreto che generano le cosiddette funzioni di massa. Se ampliamo il concetto alle variabili casuali continue allora dovremo definire la forma distributiva di queste variabili come *funzione di densità*.

Abbiamo visto nelle prime pagine di questa dispensa come l'assetto distributivo di un dato fenomeno possa essere rappresentato mediante un istogramma nel quale in corrispondenza delle frequenze rilevate per le classi definite sulla variabile vengono innalzati dei rettangoli ad esse proporzionali. Immaginiamo che, a determinati intervalli di una variabile casuale continua, venga associata una certa probabilità di realizzazione dei valori in essi contenuti. La situazione descritta potrebbe essere rappresentata dal grafico a fianco. Siccome ogni rettangolo contiene un certo numero di frequenze n_i, la probabilità che una determinazione della variabile X sia contenuta nell'intervallo (x_1, x_2) è data da:



$$\Pr(x_1 < X < x_2) = \frac{n_i}{N}$$

dove N, ovviamente, indica la somma totale delle frequenze. È importante sottolineare che l'area dei rettangoli rappresentati nell'istogramma deve essere uguale all'unità, anche perché come abbiamo visto, essi descrivono una probabilità. L'istogramma precedente potrebbe essere perequato da una curva, la cui area compresa tra l'asse orizzontale e gli estremi del campo di variazione deve essere unitaria.

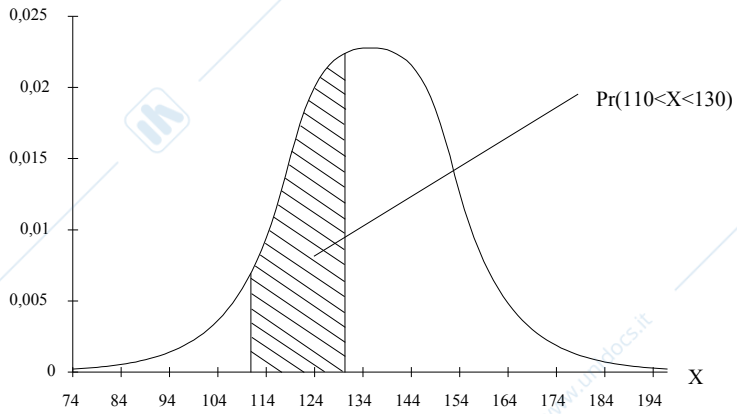


Se invece della curva perequatrice scegliamo di stringere l'intervallo (x_1, x_2) fino a dimensioni molto piccole, allora vedremmo ridurre l'ampiezza dei rettangoli

dell'istogramma fino a divenire un continuo che descrive non più una spezzata ma una vera e propria funzione di densità. In questa funzione la probabilità che un valore della variabile X sia contenuto in un intervallo (x_1, x_2) è data dall'area racchiusa dalla curva in quell'intorno, ovvero dall'integrale della forma funzionale calcolato nell'intervallo x_1,x_2.Secondo lo stesso criterio l'integrale della funzione calcolato per l'intero campo di variazione della variabile dovrà essere unitario, altrimenti la variabile continua che stiamo analizzando non è una variabile casuale.

In questo corso non approfondiremo ulteriormente le tematiche legate alle distribuzioni di variabili casuali continue, ma analizzeremo ancora gli aspetti di questo argomento dal punto di vista dell'uso delle distribuzioni probabilistiche.

Concentriamoci adesso sul concetto di area come equazione di probabilità. Supponiamo che la variabile X rappresenti la pressione sanguigna sistolica di una certa popolazione di soggetti. La seguente funzione rappresentata nel grafico esprime la funzione di densità (che d'ora in poi chiameremo anche con il nome di *curva di probabilità*) della suddetta variabile.

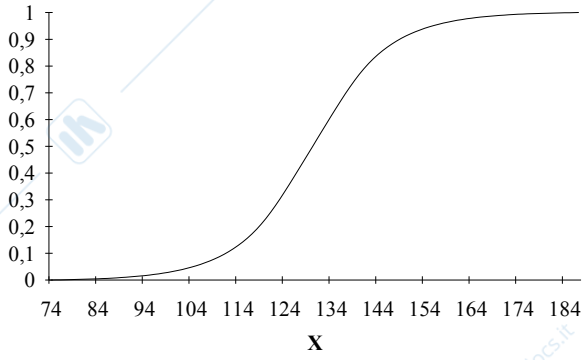


L'area sottostante alla curva presenta una parte tratteggiata in corrispondenza dei valori pressori compresi tra 110 e 130. Secondo quanto abbiamo detto in precedenza, l'area individuata dal tratteggio identifica la probabilità che la pressione sanguigna della popolazione sia compresa tra 110 e 130. Nell'esempio illustrato nel grafico il valore dell'area è pari a Pr(110<X<130)=0,323, valore che, tradotto in termini percentuali, indica come il 32,3% dei soggetti ha una pressione sistolica compresa in questo intervallo.

Se, invece, siamo interessati a conoscere la probabilità che la popolazione abbia dei valori pressori inferiori a 110, allora andremo a calcolare l'area della porzione di spazio al di sotto della curva nell'intervallo situata tra il valore minimo della distribuzione ed il

limite suddetto. Nel nostro esempio risulterà che Pr(X<110)=0,125, ovvero il 12,5% della popolazione ha una pressione sistolica inferiore a 110.

Per calcoli di questo genere la curva di probabilità è però scarsamente usata. Per semplicità si preferisce ricorrere ad una diversa forma distributiva che viene costruita sulla base della curva di probabilità. Si tratta della *curva cumulativa di probabilità* (o funzione ripartizione), nella quale l'asse orizzontale rappresenta i possibili valori della variabile casuale X, mentre l'asse verticale rappresenta la scala di probabilità con i valori che variano da 0 a 1. La curva cumulativa di probabilità indica, per ogni valore *a* della variabile preso sull'asse orizzontale, la probabilità di avere valori inferiori (o superiori) ad *a*. Questa probabilità, come ben sappiamo, è precisamente l'area che sta al di sotto della curva di probabilità per valori di X compresi nell'intervallo (-∞, a) (oppure (a, +∞) nel secondo caso). In particolare se a=∞, si avrà che Pr(-∞<X<+∞)=1 e ciò indica che X dovrà certamente assumere un valore nell'intervallo (-∞, +∞). Il seguente grafico illustra un esempio di curva cumulativa di probabilità.



La curva cumulativa di probabilità è molto utile. Ad esempio, visto che l'area sottostante alla curva è unitaria, ne segue che:

$$\Pr(X > a) = 1 - \Pr(X \leq a)$$

La suddetta curva può anche essere utilizzata per calcolare la probabilità corrispondente all'intervallo (a, b), dove *a* e *b* (con a<b) sono due valori specifici che la variabile X può assumere. In queste condizioni avremo che:

$$\Pr(a < X < b) = \Pr(X < b) - \Pr(X \leq a)$$

Come si può facilmente osservare, si tratta di una formula che individua la differenza tra il valore della curva cumulativa nel punto *b* ed il valore della curva cumulativa nel punto *a*.

Distribuzione normale e normale standardizzata

Nelle pagine precedenti abbiamo rappresentato la forma distributiva della nostra ipotetica variabile casuale X mediante una curva dal caratteristico andamento cmpanulare. Nella teoria statistica questa classe di funzioni prende il nome di *distribuzione normale*. In questo paragrafo ne illustreremo le principali caratteristiche.
La sua importanza è rilevante, anche se i test statistici che vedremo in seguito non prendono a modello tale distribuzione. La normale è usata per rappresentare un particolare gruppo di variazioni di una variabile casuale continua. Ha, come abbiamo già accennato, una caratteristica forma a campana, è simmetrica rispetto al valore medio della variabile ed in quel punto la funzione assume il suo massimo. Inoltre, media, moda e mediana hanno valori coincidenti. Si tratta, ovviamente, di una funzione di probabilità in cui l'area sottostante alla curva ha un valore unitario.
Nella simbologia statistica è possibile indicare sinteticamente le caratteristiche della variabile casuale normale X nel seguente modo:

$$X \sim N(\mu, \sigma)$$

ovvero, X ha una distribuzione normale con media μ e deviazione standard σ^2.
Per calcolare la probabilità che X assuma valori compresi in un intervallo (x_1, x_2) $[Pr(x_1<X<x_2)]$, oppure la probabilità che X sia inferiore a x_1 $[Pr(x_1<X)]$, il procedimento è del tutto analogo a quello che abbiamo illustrato nel precedente paragrafo, fatta eccezione per l'uso di particolari tavole, nelle quali per determinati valori di X è possibile calcolare l'area della curva che vi è compresa è dunque la corrispondente probabilità. Ma di ciò avremo modo di parlare in seguito.
Per l'uso delle suddette tavole si deve procedere alla *standardizzazione* della variabile casuale X. Questo procedimento è una trasformazione di variabile che, lasciando immutata la forma distributiva, ne trasla il valore centrale sullo zero e fissa la varianza su un valore unitario. La trasformazione avviene secondo la seguente formula:

$$Z = \frac{X - \mu}{\sigma}$$

Sottraendo dalle determinazioni della variabile X il valore medio e dividendo per la deviazione standard si ottengono le determinazioni di una nuova variabile Z (*variabile casuale normale standardizzata*) che ha ancora una distribuzione normale ma con media uguale a 0 e varianza uguale a 1, ossia:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

Campioni e distribuzioni campionarie

Da quanto abbiamo detto è ormai noto che il compito principale della statistica descrittiva è quello di esaminare una popolazione finita di individui qualsiasi, allo scopo di determinare la distribuzione di un certo loro carattere qualitativo o quantitativo, discreto o continuo.

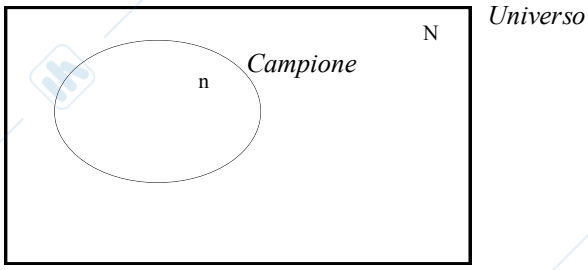
Per popolazione in questo caso intendiamo il gruppo complessivo di individui che possiedono l'attributo esaminato dalla ricerca. Per esempio, l'insieme degli studenti universitari, dei cittadini italiani con età superiore a 65 anni, degli operai impiegati nel settore metalmeccanico e così via.

In molti casi, però, risulta impossibile studiare il fenomeno nell'intera popolazione per tutta una serie di motivi (difficoltà di reperire i dati necessari, mezzi economici insufficienti, tempi di realizzazione della ricerca eccessivi, ecc..). Un esempio di rilevazione globale dei dati su un'intera popolazione sono i *censimenti*, nei quali si raccoglie il più ampio volume di informazioni possibile dalla popolazione italiana. A causa delle grosse difficoltà organizzative, delle risorse economiche impiegate e del tempo necessario per elaborare i dati, queste rilevazioni vengono svolte con cadenza decennale e, nonostante il censimento sia la forma in teoria più efficace per ottenere stime corrette, può accadere che le informazioni rilevate siano affette da errori di vario genere.

Per queste ragioni è molto più frequente l'uso di *campioni probabilistici*, che cercano di ottimizzare l'efficacia dell'indagine nel rispetto dei vincoli organizzativi, economici e temporali. Non si tratta di accontentarsi di una conoscenza approssimativa del fenomeno, basata su una rilevazione parziale delle informazioni che interessano, ma di studiare in modo opportuno una popolazione attraverso un gruppo ristretto di oggetti, affinché le stime provenienti da quest'ultimo possano essere estese, con un determinato margine di errore, alla popolazione di riferimento.

La rilevazione parziale di cui sopra consiste nell'esaminare un numero n do soggetti selezionati tra gli N individui appartenenti alla popolazione. Il seguente grafico mostra

come, in termini insiemistici, il campione sia un sotto-insieme dell'universo di oggetti che compongono la popolazione.



Secondo quanto abbiamo precisato nell'introduzione, al posto del termine popolazione possiamo usare il sinonimo *universo*, indicandolo con il simbolo U.

Dunque, gli n elementi estratti rappresentano quello che comunemente chiamiamo campione casuale della popolazione oggetto di studio. A seconda della numerosità campionaria n, dall'universo si possono estrarre un numero molto elevato di campioni differenti tra di loro, ognuno con una propria media ed una propria varianza. In teoria potremmo selezionare anche il solo campione che ci fornisce valori di media e varianza uguali a quelli della popolazione. Nella pratica, invece, i campioni tendono a fornire dei valori caratteristici (media e varianza) che si discostano in varia misura dai corrispondenti parametri della popolazione.

È opportuno ricordare che i campioni non sono in grado di rappresentare un'immagine esatta dell'universo da cui sono estratti, ma ci permettono di ottenere delle stime che possono essere più o meno vicine a quelle originarie. Ovviamente, come è facile immaginare, la vicinanza tra le stime ottenute dal campione ed il valore vero dei parametri nella popolazione aumenta quanto più è numeroso il gruppo di soggetti selezionati che costituisce il campione.

Se, per esempio, dobbiamo studiare un gruppo N=10.000 di individui sulla base di un campione formato da n=1.000 soggetti, allora andremo a stimare i parametri media e deviazione standard della popolazione, che rispettivamente si indicano con μ e σ, attraverso i valori campionari ricavati dalla media aritmetica e dalla deviazione standard, che rispettivamente si indicano con m e S.

A questo punto occorre fare una importante precisazione. Il suddetto procedimento di stima fornirà risultati corretti (ovvero i valori campionari potranno essere rappresentativi di quelli della popolazione, fissato un certo margine di errore) solo se il campione è stato scelto secondo metodologie opportune. I criteri di scelta, lo studio dei quali forma l'oggetto di una disciplina denominata *Teoria dei campioni*, sono fondamentali per la correttezza e l'efficienza del processo di stima.

Prima di illustrare le principali tecniche di campionamento, facciamo un passo indietro e torniamo a parlare di distribuzioni campionarie. Consideriamo la solita popolazione di N=10.000 individui dalla quale estraiamo un campione di numerosità n=1.000. Come abbiamo detto, sarà possibile estrarre un numero molto grande di campioni che hanno una numerosità n, ognuno con una propria media ed una propria varianza. Nella seguente tavola sono descritti gli ipotetici valori caratteristici relativi a 8 dei possibili campioni ed i corrispondenti valori dei parametri nella popolazione.
Se il numero di campioni è molto elevato possiamo immaginare le determinazioni della media e della varianza i valori di una vera e propria distribuzione. In generale, ogni funzione dei dati campionari (p.e.: la media) è a sua volta una variabile casuale ad una dimensione la cui funzione di densità dipende dalla funzione di densità della variabile originaria.

Valori caratteristici	*Campioni*								*Popolazione*
	C1	C2	C3	C4	C5	C6	C7	C8	
Media	9,2	10,8	10,2	8,7	11,1	10,3	9,5	9,8	10
Varianza	4,3	4,5	2,7	5,3	4,9	2,9	3,1	2,7	3,8

Se consideriamo, per esempio, tutte le medie calcolate su tutti i campioni di 1.000 elementi che è possibile estrarre dalla popolazione di 10.000 unità, possiamo affermare che queste rappresentano le determinazioni di una variabile casuale è quella delle medie campionarie. Indicando con X la variabile oggetto di studio, le cui determinazioni sono x_1, x_2, ... , x_n, la variabile casuale che indica la *distribuzione delle medie campionarie* sarà $\overline{X}$ le cui determinazioni sono m_1, m_2, ... , m_k, dove k indica il numero dei possibili campioni, composti ognuno da n elementi. Se X è normale con media μ e deviazione standard, anche la distribuzione delle medie campionarie $\overline{X}$ sarà nuovamente normale con media μ e deviazione standard $\sigma/\sqrt{n}$, ovvero:

$$\text{se} \quad X \sim N(\mu, \sigma) \quad \text{allora} \quad \overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

Da ciò si evince che il valore medio della distribuzione delle medie campionarie è identico a quello della distribuzione della popolazione. Riprendendo l'esempio illustrato nella tavola precedente e calcolando il valore medio delle medie degli 8 campioni possiamo osservare come questo quasi coincida con quello della popolazione (9,95 contro10), seppure i campioni estratti siano pochi.
In questo esempio abbiamo illustrato i concetti principali che guidano alla comprensione della distribuzione delle medie campionarie, concetti che riprenderemo in

seguito. Gli stessi criteri metodologici possono essere applicati per costruire la distribuzione della varianza campionaria, o di qualsiasi altro parametro descrittivo.

Tecniche di campionamento

Abbiamo appurato che più elevata è la numerosità del campione rispetto alla popolazione di riferimento, più esatte saranno le stime che se ne possono ricavare e quindi minore sarà la varianza delle distribuzioni campionarie.
La correttezza e la precisione delle stime sono influenzate in maniera determinante anche dal criterio di campionamento che viene utililizzato. Senza approfondire gli aspetti metodologici dei vari criteri, ci sembra più opportuno elencarne le tipologie e descriverne l'applicazione.
Il modo più semplice per estrarre un campione è rappresentato dal *campionamento casuale semplice*. Tale metodo consiste, data una popolazione P composta da N elementi, nella selezione in modo casuale di un campione di ordine n.
La selezione casuale in questione può essere ottenuta estraendo delle palline numerate da un'urna, per esempio, oppure utilizzando apposite tavole che forniscono successioni di numeri casuali. Per applicare il metodo descritto è necessario conoscere la numerosità della popolazione e numerare in una lista i suoi componenti, affinché l'estrazione sia in grado di individuare uno specifico soggetto. Secondo questo procedimento ogni elemento della popolazione ha la stessa probabilità di essere estratto, in quanto l'eventuale pallina utilizzata per la scelta casuale ogni volta viene reinserita nell'urna.
Una variante rispetto al campionamento casuale semplice è il *campionamento casuale semplice non ripetitivo*, che segue gli stessi criteri metodologici del precedente fatta eccezione per il fatto che le eventuali palline estratte non vengono reinserite nell'urna.
Una tecnica che può efficacemente sostituire il campionamento casuale semplice è il cosiddetto *campionamento sistematico.* I principali vantaggi rispetto ai metodi precedenti sono rappresentati da una maggiore semplicità che si traduce in minori errori nell'estrazione del campione ed un aumento della precisione che può essere ottenuta organizzando opportunamente la lista degli individui appartenenti alla popolazione. Se dunque disponiamo di un universo composto da N individui da cui si vuole estrarre un campione di numerosità n è opportuno disporre di una popolazione finita e ordinare tale lista secondo una caratteristica che possibilmente deve essere correlata con l'oggetto della ricerca. Se, per esempio, dobbiamo fare una ricerca sulle aziende artigiane in una determinata dimensione territoriale, sarà necessario ordinare la lista delle aziende in relazione al tipo di attività svolta. Metodologicamente parlando, il campionamento sistematico seleziona da questa lista un elemento ogni k. Se N=10.000 e n=1.000 avrò

un salto di campionamento k=10, ovvero estrarrò dalla lista un nominativo ogni 10. Il primo elemento del campionene sistematico deve essere estratto in modo casuale tra i primi k soggetti all'inizio della lista. Come abbiamo detto, questo tipo di tecnica presenta dei miglioramenti in termini di efficienza nei riguardi del campionamento casuale semplice e non ripetitivo.

Nel caso di popolazioni molto eterogenee può risultare necessario utilizzare un diverso tipo di tecnica, il cosiddetto *campionamento stratificato*. Secondo questa metodologia si procede ad una divisione della popolazione di N elementi in L sottopopolazioni, dette *strati*, ognuno contenente N_h (h=1, ... ,L) elementi omogenei rispetto ad una determinata caratteristica e tali che:

$$\sum_{h=1}^{L} N_h = N$$

Se ogni strato è omogeneo in modo che i suoi elementi variano di poco l'uno dall'altro, si può ottenere con piccoli campioni una stima precisa dei parametri della popolazione, stima più corretta di quella ottenibile con i precedenti metodi. Inoltre, l'efficacia del metodo aumenta se gli strati hanno una grossa omogeneità all'interno (bassa varianza interna agli strati) ed una grossa eterogeneità tra gli strati (alta varianza tra gli strati).

Per esempio un criterio di stratificazione potrebbe essere l'appartenenza a determinate aree geografiche (nord, centro, sud), a particolari classi d'età (< 14 anni, 14-21 anni, ecc..) e così via.

Supponiamo che gli strati costruiti sulla popolazione originaria N=10.000 siano 3: al primo appartengono N_1=6.000 individui, al secondo N_2=1.000, ed al terzo N_3=3.000. Se il campione è costituito da n=1.000 soggetti, la distribuzione dell'estrazione del campione può avvenire secondo vari criteri, tra cui il più semplice è quello dell'*assegnamento proporzionale*, secondo il quale si assegnerà una numerosità campionaria ad ogni strato proporzionale alla densità di popolazione. Così nel primo strato andremo ad estrarre un campione di n_1=600 individui, nel secondo di n_2=100, e nel terzo di N_3=300. L'estrazione del campionene all'interno degli strati può avvenire secondo uno dei criteri illustrati sopra.

Fin qui abbiamo esaminato tecniche di campionamento che possono essere applicate quando la popolazione ha una numerosità relativamente bassa. Quando invece l'universo è composto da un numero molto elevato di soggetti, può risultare molto difficile se non impossibile estrarre un campione con le tecniche viste. Ad esempi se volessimo eseguire un'indagine sul numero degli italiani colpiti da influenza nello scorso inverno è ovvio che non disporremo di una lista completa in cui tutti i cittadini siano numerati.

Risulterebbe impossibile, quindi, estrarre un campione secondo uno dei precedenti procedimenti. Un modo per superare queste difficoltà potrebbe essere quello di suddividere l'intera popolazione per comuni di residenza e di costruire un campione in due fasi distinte: nella prima si estraggono casualmente un gruppo di comuni, nella seconda si estrae un campione di popolazione da questi comuni. Così facendo sarà senz'altro più facile disporre di liste nominative ed al tempo stesso i costi dell'indagine possono essere contenuti in maniera rilevante. Questa tecnica prende il nome di campionamento a grappolo a due stadi. Il seguente grafico illustra una ipotetica realizzazione del procedimento descritto.

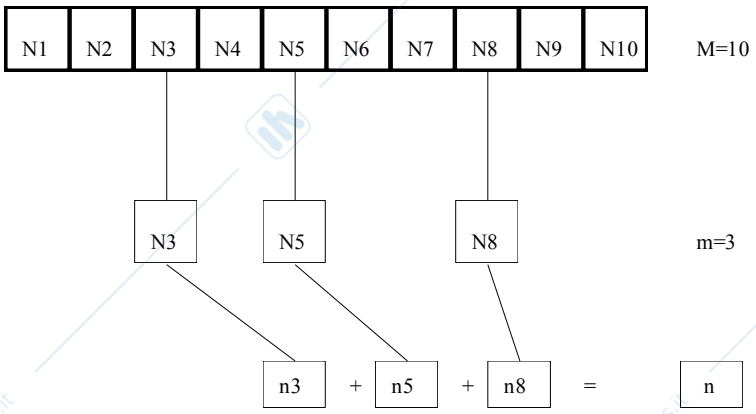
Supponiamo di suddividere la popolazione originaria N in un numero M di sottopopolazioni chiamate *grappoli*. Ogni grappolo deve contenere almeno un elemento ed un elemento deve appartenere solo ad un grappolo. Se indichiamo con N_i (i=1, ... ,M) il numero di elementi appartenenti a ciascun grappolo, avremo:

$$\sum_{h=1}^{L} N_h = N$$

Se estraiamo casualmente secondo una tecnica di quelle illustrate m degli M grappoli selezioniamo le *unità primarie di campionamento*, mentre se all'interno dei grappoli selezionati si estraggono i campioni di numerosità $n_i<N_i$ otteniamo le cosiddette *unità secondarie di campionamento*. Ovviamente:

$$\sum_{i=1}^{m} n_i = n$$

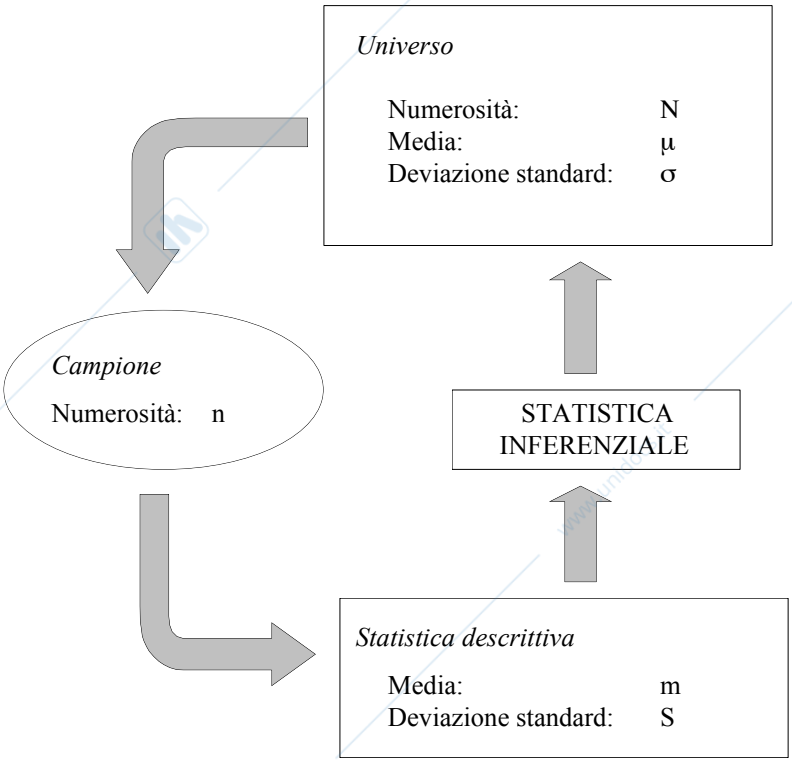
Ne seguente esempio si ipotizza la divisione in M=10 grappoli della popolazione N, con l'estrazione di un campione di numerosità n da un numero di m=3 grappoli.



Cenni di inferenza statistica

All'inizio di questa dispensa abbiamo stabilito una divisione tra i settori di applicazione delle metodologie statistiche, distinguendo tra statistica descrittiva ed inferenziale. Abbiamo già ampiamente trattato il significato e le modalità d'uso delle tecniche di statistica descrittiva. In queste ultime pagine ci siamo avvicinati, invece, alle tematiche trattate dalla *statistica inferenziale*.

Il primo problema lo abbiamo affrontato con gli intervalli di confidenza. Può accedere spesso, infatti, che, sulla base delle informazioni campionarie, siamo chiamati a formulare delle ipotesi sulle caratteristiche della popolazione da cui i campioni sono stati tratti. Oppure ci troviamo a confrontare due o più popolazioni e dobbiamo valutare se queste possiedono caratteristiche uguali.



Il precedente diagramma descrive il processo logico all'interno del quale si collocano le tecniche inferenziali nel contesto dell'analisi statistica dei dati.

Data, per ipotesi, una popolazione di numerosità finita N, caratterizzata da una media μ e da una deviazione standard σ. Si estrae un campione di oggetti di numerosità n. La statistica descrittiva rappresenta il primo passo per interpretare i dati. Essa ci permette, innanzi tutto, di individuare le caratteristiche principali della distribuzione oggetto di

studio, quali, per esempio, la media campionaria m e la deviazione standard S. In generale, come abbiamo visto, le tecniche descrittive trovano applicazione in tutte quelle operazioni inerenti l'organizzazione, l'elaborazione e la presentazione dei dati. Ma esse si limitano alla descrizione del gruppo di valori provenienti dal campione.

Le tecniche inferenziali hanno lo scopo di generalizzare questi risultati all'intera popolazione, fornendo delle stime dei parametri veri nell'universo. Il processo di generalizzazione può avvenire secondo diverse metodologie ed il concetto di stima assume una posizione centrale in questo problema.

Gli intervalli di confidenza esaminati nel precedente paragrafo rappresentano un momento del processo di inferenza statistica in quanto dai valori campionari si cerca di risalire al valore della popolazione. In questo caso particolare, però, ci troviamo di fronte ad una *stima per intervalli*, in quanto non viene determinato un valore puntuale quale parametro per l'universo, ma piuttosto un intervallo che contiene con una certa probabilità tale valore.

Molto più interessante è l'uso delle *stime puntuali*, ovvero di quelle metodologie che forniscono un valore preciso da assumere quale parametro per l'universo. A questo fine si giunge utilizzando i cosiddetti *stimatori*, che in pratica sono delle regole o metodi per stimare un parametro della popolazione sulla base dei valori campionari. Tali stimatori per fornire stime attendibili devono possedere le seguenti caratteristiche:

1. uno stimatore è *non distorto* se la sua distribuzione campionaria ha media uguale a quella del parametro da stimare nella popolazione;
2. uno stimatore è detto *migliore* se ha la varianza minore di altri stimatori;
3. uno stimatore è detto *efficiente* se possiede entrambe le precedenti proprietà.

Nelle pagine precedenti abbiamo già visto un esempio di stimatore, allorquando abbiamo sostituito al posto della varianza della popolazione il suo valore campionario nel calcolare gli intervalli di confidenza, ovvero:

$$\sigma^2 \cong \hat{S}^2 = \frac{n}{n-1} S^2$$

In questo caso il parametro $\hat{S}^2$ risulta essere uno stimatore corretto della varianza della popolazione σ^2. Come è facile si tratta di stime puntuali poiché queste formule ci forniscono un valore esatto da utilizzare quale stima del parametro nell'universo.

Intervalli di confidenza

Al momento in cui si estrae un campione per realizzare un determinato studio, questo ci fornisce una media campionaria e quindi una determinazione della variabile $\overline{X}$. Sulla base del valore campionario noi dobbiamo stimare la media della popolazione.

Date queste condizioni appare molto improbabile che il campione estratto sia in grado di fornirci una stima esatta del valore del parametro nell'universo, anche perché la probabilità che m (media campionaria) e μ (valore medio della popolazione) coincidano è praticamente nulla.

Per ridurre il grado di incertezza della stima è preferibile prendere in considerazione, piuttosto che un singolo valore della $\overline{X}$, un intervallo di valori di $\overline{X}$ e vedere se, dato un certo livello di probabilità, il valore vero del parametro ignoto μ è contenuto in esso. È stato inserito il concetto di probabilità poiché, innanzi tutto, non si tratta di una stima puntuale e secondariamente perchè il procedimento è assimiliabile alla verifica di una ipotesi sulla quale si corre il rischio di commettere errori.

Il valore di probabilità di cui sopra viene indicato con 1-α, dove α rappresenta un numero compreso tra 0 ed 1 e che generalmente è abbastanza piccolo. Di solito vengono adottati valori di α pari a 0,10, 0,05 e 0,01 a cu corrispondono valori di 1-α rispettivamente di 0,90, 0,95 e 0,99. Questi ultimi altro non sono che i livelli di probabilità che abbiamo stabilito affinché la nostra affermazione possa essere ritenuta corretta. In altre parole se α=0,05 e 1-α=0,95 significherà che ho il 95% di probabilità di fare un'affermazione corretta ed il rimanente 5% sarà il rischio che sono disposto a correre nella stima.

Tornando alla distribuzione della media campionaria $\overline{X}$, lo scopo degli *intervalli di confidenza* è quello di individuare, sulla base dei valori campionari media e deviazione standard, un intervallo all'interno del quale è contenuto il valore vero della popolazione m, dato un certo livello di probabilità 1-α.

Come già abbiamo visto, se X ha una distribuzione normale N(μ,σ) allora $\overline{X}$ si distribuisce come una normale $N(\mu, \sigma/\sqrt{n})$. Il fine è quello di individuare gli estremi dell'intervallo che contiene il valore di μ. Da ciò possiamo concludere che:

$$P\left[\mu - u_\alpha \frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + u_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

in cui si indica che la determinazione campionaria della media è contenuta in un intorno simmetrico della media campionaria e dove u_α e $-u_\alpha$ sono i due valori della distribuzione che isolano, il primo alla propria sinistra ed il secondo sulla sua destra,

due aree di probabilità pari ad $\alpha/2$. Sottraendo μ da entrambi i membri e procedendo con i calcoli otteniamo:

$$P\left[u_\alpha \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < u_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

$$P\left[-\overline{X} - u_\alpha \frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + u_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

da cui invertendo i segni e spostando i membri ricaviamo:

$$P\left[\overline{X} - u_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

Da quest'ultima relazione è facile verificare che il valore medio della popolazione è contenuto nell'intervallo costruito sulla base delle determinazioni campionarie.
Dunque, data una certa variabile X che si distribuisce come una normale standardizzata, fissato un valore di α=0,05, il corrispondente livello di significatività dell'intervallo di confidenza sarà del 95%. Il valore di $u_{0,025}$, che viene ricavato con semplicità da una qualsiasi delle tavole della normale standardizzata, è uguale a 1,96.
Supponiamo che da una popolazione che si distribuisce secondo una normale con media μ ignota e con varianza σ^2=4 sia stato estratto un campione di 16 elementi la cui media aritmetica è m=5. L'intervallo di confidenza per la media ad un livello di probabilità del 95% sarà dato da:

$$5 - 1,96\frac{2}{\sqrt{16}} < \mu < 5 + 1,96\frac{2}{\sqrt{16}}$$

da cui si ricava che: $4,02<\mu<5,98$.
Da ciò possiamo concludere affermando che l'intervallo (4,02; 5,98) contiene il valore vero della media con probabilità del 95%.
Un aumento dei della variabilità interna ai dati avrebbe comportato un allargamento dei limiti dell'intervallo. Nel precedente esempio, se la varianza fosse stata σ^2=16, allora l'intervallo sarebbe stato: $3,04<\mu<6,96$
Osservando la formula si può notare che l'intervallo di confidenza viene determinato in funzione delle determinazioni della media campionaria e della deviazione standard della popolazione (σ). In molti casi quest'ultimo parametro può essere ignoto. In tali situazioni è possibile utilizzare una stima corretta del valore della varianza della popolazione (σ^2) mediante il seguente stimatore:

$$\sigma^2 \cong \hat{S}^2 = \frac{n}{n-1} S^2$$

Seguendo l'impronta del precedente esempio se la media campionaria è m=5 e la varianza campionaria è S^2=3,8, allora l'intervallo di confidenza è il seguente:

$$5 - 1,96 \frac{2,01}{\sqrt{16}} < \mu < 5 + 1,96 \frac{2,01}{\sqrt{16}}$$

in cui gli estremi risultano essere: 4,01<μ<5,99.

I TEST DI SIGNIFICATIVITÀ

Nozioni generali sui test di significatività

Tutte le metodologie illustrate finora descrivono le caratteristiche principali delle variabili e ne studiano i rapporti con l'universo di provenienza. È necessario sottolineare che tutte quante si riferiscono ad analisi di tipo *univariato*, ovvero che prendono in considerazione una sola variabile. I test statistici che analizzeremo in questa sezione della dispensa ampliano le tecniche a disposizione del ricercatore, fornendo degli strumenti per studiare le possibili relazioni esistenti tra coppie di variabili.

Volendo dare una semplice definizione, un *test di significatività* è una procedura statistica che permette di valutare la conformità dei risultati ottenuti con una determinata ipotesi. Il risultato di questo processo di valutazione si chiama grado di significatività dei dati rispetto all'ipotesi, ed alla base di questa valutazione ci sono supposizioni e calcoli.

L'obiettivo di un test di significatività è quello di determinare il grado di accordo tra una specifica ipotesi, che indicheremo con H, ed un gruppo di osservazioni. Per ipotesi, supponiamo di aver estratto un campione da una popolazione di N individui (in cui la variabile X si distribuisce normalmente con media μ e deviazione standard σ) e di aver calcolato il valore medio delle osservazioni campionarie (m). Invece di utilizzare gli intervalli di confidenza, potremmo sottoporre a test il valore medio campionario per sapere se questo è uguale a quello della popolazione. Di solito è consuetudine associare all'*ipotesi nulla* (H_o) un'*ipotesi alternativa* (H_A), con la quale si indica il comportamento del test qualora l'ipotesi nulla dovesse essere rigettata. Dunque, la coppia di ipotesi che pongo è la seguente:

$$H_o : m=\mu$$
$$H_A : m\neq\mu$$

Secondo quanto abbiamo visto prima la media osservata m non è altro che una determinazione della variabile casuale media campionaria $\overline{X}$, la quale si distribuisce nuovamente come una normale $N(\mu, \sigma/\sqrt{n})$.

Nell'ipotesi H_o, come si può osservare, si confronta il valore campionario con quello della popolazione. Per sua natura (essendo un parametro relativo ad un campione), può assumere anche valori molto distanti da μ. È nostro compito individuare dei limiti oltre i quali la differenza tra m e μ è eccessivamente elevata. In queste aree, che rappresentano la *zona di rifiuto* dell'ipotesi H_o, si trovano tutti quei valori di $\overline{X}$ che hanno una bassa

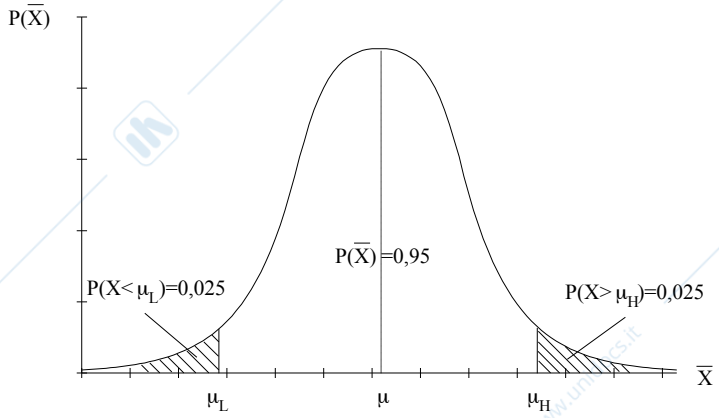
probabilità di manifestarsi. Questa probabilità è normalmente riferita a due livelli corrispondenti al 5% ed all'1%.
Supponendo di usare un livello di significatività del 5%, se indichiamo con μ_L e μ_H rispettivamente il limite inferiore e quello superiore della zona di rifiuto dell'ipotesi H_o (e dunque di accettazione dell'ipotesi H_A), dobbiamo scegliere i due valori μ_L e μ_H in modo che:

$$P\left(\overline{X} < \mu_L\right) + P\left(\overline{X} > \mu_H\right) = 0{,}05$$

Data la simmetria della distribuzione avremo che:

$$P\left(\overline{X} < \mu_L\right) = 0{,}025 \quad \text{e} \quad P\left(\overline{X} > \mu_H\right) = 0{,}025$$

Il seguente grafico illustra la situazione descritta fino ad ora.



Per effettuare il test è necessario ricorrere alle distribuzione della variabile Z, ovvero la normale standardizzata, sopponendo a test le ipotesi nel seguente modo:

$$H_o \text{ è rigettata se } \frac{(m-\mu)\sqrt{n}}{\sigma} < Z_L = -1{,}96$$

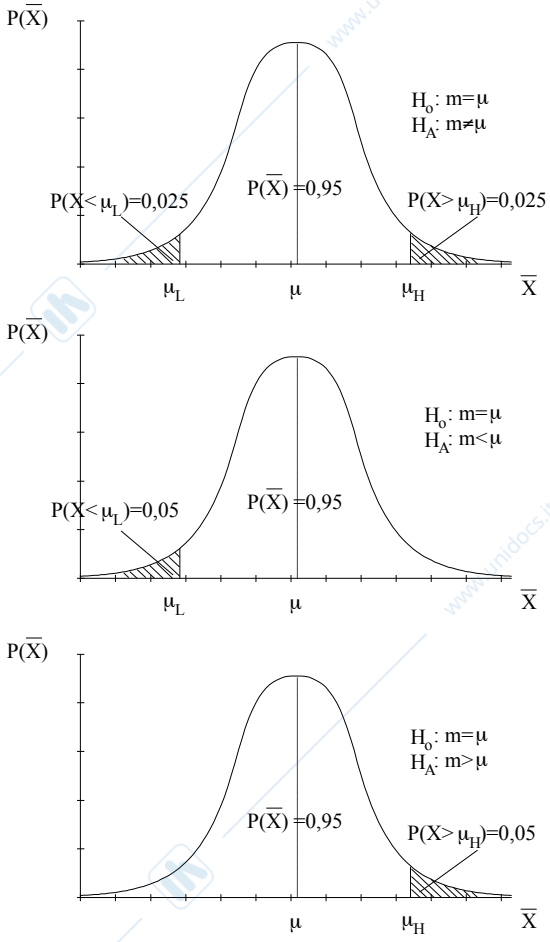
$$H_o \text{ è rigettata se } \frac{(m-\mu)\sqrt{n}}{\sigma} > Z_H = 1{,}96$$

Dove Z_L e Z_L sono equivalenti ai limiti μ_L e μ_H, traslati però nella variabile normale standardizzata. Se, invece:

$$Z_L < \frac{(m-\mu)\sqrt{n}}{\sigma} < Z_L$$

non posso rigettare l'ipotesi nulla. Occorre fare una precisazione molto importante. Come si può osservare, non possiamo parlare di accettazione dell'ipotesi nulla, ma piuttosto di non rigetto, poichè si tratta di una valutazione che, nel caso dell'esempio, ha una significatività del 95%, ovvero posso commettere un errore nell'accettare l'ipotesi nulla quando questa è da rigettare. Tutte le ipotesi sono sottoposte a test in termini probabilistici.

Il test che abbiamo appena osservato ha la particolarità di essere a *due code*. Tale caratteristica dipende dall'ipotesi alternativa H_A. Se in tale ipotesi, invece che indicare un qualsiasi valore alternativo (p.e.: H_A : $m \neq \mu$), si provvede a stabilire una direzione (p.e.: H_A : $m > \mu$, oppure H_A : $m < \mu$) allora il test diviene ad *una coda* secondo le modalità espresse nel seguente grafico.



Molto spesso la varianza della popolazione non è nota e per calcolare il valore del test è necessario utilizzare lo stimatore campionario $\hat{S}^2$ che abbiamo già visto nel caso degli intervalli di confidenza. In questo caso, però, l'applicazione del test necessita di alcune raccomandazioni.

Se, infatti, la numerosità del campione è n<30 il test deve essere eseguito utilizzando la *distribuzione t-Student* al posto della normale standardizzata. L'uso dello stimatore modifica la distribuzione campionaria di $\overline{X}$ e quindi la grandezza utilizzata per il calcolo del test è assimilabile ad una distribuzione *t*-Student, ovvero:

$$\frac{(m-\mu)\sqrt{n}}{S} \sim t_{n-1}$$

In appendice sono contenute le tavole relative a questa distribuzione, il cui utilizzo è esattamente uguale a quello osservato nel caso precedente.
Se la numerosità del campione è superiore a 30 le tavole della *t*-Student sono sovrapponibili a quelle della distribuzione normale standardizzata e quindi possiamo evitarne l'uso. Le due distribuzioni, infatti, sono molto simili; hanno la stessa forma simmetrica e campanulare, fatta eccezione per un maggiore appiattimento della curva nel caso della *t*-Student.

Test parametrici e non parametrici

Nei test statistici sopra descritti si pone quale condizione fondamentale per la loro applicazione la provenienza dei dati campionari da una popolazione in cui la variabile oggetto dello studio ha una distribuzione nota e l'ogetto del confronto è rappresentato da parametri distributivi. In questi casi indichiamo questi test con il termine di *test parametrici*.
Quando invece la distribuzione non è nota, oppure non è necessario precisare tra le ipotesi la presenza di una particolare forma distributiva, allora si parla di *test non parametrici*. In questo paragrafo illustreremo i più comuni strumenti utilizzati nelle applicazioni bio-mediche per studiare i legami esistenti tra le variabili.
Per sottoporre a test i dati organizzano le osservazioni in una apposita rappresentazione tabellare, cosiddetta *tavola di contingenza*. Dal momento che gran parte dei dati raccolti nella ricerca medica è di natura binaria, possiamo supporre di collocare le frequenze in una tabella a doppia entrata. Supponiamo, per esempio, di dover reppresentare in una tavola di contingenza le frequenze associate ad una coppia di variabili, in cui la prima rappresenta il tipo di farmaco che è stato utilizzato per curare una certa patologia (farmaco di tipo A e di tipo B) e la seconda riguarda l'esito della malattia (guarigione, non guarigione). I dati possono essere riportati in una tavola con due righe e due colonne (2x2), come ci mostra la seguente tabella. Si tratta, appunto di una tavola di contingenza, sulla quale l'applicazone dii test ha lo scopo di verificare l'esistenza e la forza di relazioni tra le variabili

Esito / *Farmaco*	*Guargione*	*Non guarigione*	Totale
Farmaco tipo A	7	3	10
Farmaco tipo B	2	7	9
Totale	9	10	*19*

La prima domanda che ci dobbiamo porre è se lo schema logico di classificazione (e dunque la relazione tra le variabili) corrispondono anche nella popolazione. I prossimi test analizzano dati disposti in tavole di contingenza 2x2, ne seguiranno altri utili ad esaminare tabella più ampie.

Test del χ^2 per tavole di contingenza

Come abbiamo detto poco sopra, la questione della corrispondenza tra lo schema di classificazione redatto nella tavola di contingenza e quello della popolazione è un aspetto centrale in tutti i tests statistici.

Se supponiamo che tale corrispondenza non esista allora dobbiamo ricorrere al *test del chi quadrato* (χ^2) per verificare il grado di coerenza tra questa ipotesi ed i dati-campione raccolti.

Supponiamo di effettuare un esperimento in cui si possono avere due soli risultati, che noi indicheremo con *successo* e *insuccesso*. Richiamando la precedente tavola la guarigione potrebbe essere interpretata come il successo e la non guarigione come l'insuccesso. Altri esempi potrebbero essere rigetto/non rigetto nel caso di trapianti, morte/sopravvivenza, ecc..

Ad ognuno di questi due casi è possibile associare una percentuale di frequenza, che, come ben sappiamo, può assumere valori compresi tra 0 ed 1, la cui somma è sempre pari all'unità poiché questi due sono gli unici risultati possibili dell'esperimento in questione. Questa percentuale di frequenza può essere interpretata come la probabilità del corrispondente risultato. Se, quindi, p è la probabilità dell'evento riguardante il successo, quella relativa all'insuccesso sarà uguale a 1-p.

Nella tavola di contingenza questa variabile dicotomica (successo/insuccesso) viene confrontata in due gruppi: il gruppo 1, al quale è stato somministrato il farmaco A, ed il gruppo 2, al quale è stato somministrato il farmaco B. Per fare ciò dobbiamo supporre che:

1. le probabilità per i gruppi 1 e 2 sono espresse da p_1 e p_2; all'interno di ciascun gruppo la probabilità di successo non deve variare da un individuo ad un altro;

2. il risultato ottenuto (successo o insuccesso) dai membri di ciascun gruppo non deve essere influenzato dagli altri individui.

Quando la numerosità delle osservazioni è piuttosto elevata è opportuno ricorrere ad un test che ne fornisce un valore approssimato, il *test del chi quadrato* (χ^2).
Nella seguente tavola di contingenza introduciamo i simboli O_{11}, O_{12}, O_{21} e O_{21} per indicare le *frequenze osservate* contenute in ogni cella. La tavola di contingenza può essere rappresentata come segue.

	Successi	*Insuccessi*	Totale
Gruppo 1 Gruppo 2	O 11 O 21	O 12 O 22	R1 R2
Totale	C1	C2	*N*

Gli indici associati ad ogni frequenza osservata O_{ij} (i=indice di riga; j=indice di colonna) ne individuano la posizione nella tavola. Per esempio, O_{11} individua la frequenza osservata relativa alla cella della prima colonna e della prima riga, mentre O_{21} quella situata nella seconda riga e nella prima colonna.
Lo scopo del test è quello di determinare il grado di significatività dei dati con l'ipotesi nulla H_o : $p_1=p_2$, ovvero la probabilità di appartenere alla categoria 1 (successi) è la stessa nei due gruppi.
Per applicare il test occorre innanzi tutto supporre che l'ipotesi nulla sia vera e costruire una tavola (la cosiddetta tavola delle *frequenze attese*) in cui R_1, R_2, C_1 e C_2 sono quelli del campione, ma i valori contenuti nelle celle sono i più coerenti con l'ipotesi nulla H_o, come ci mostra la seguente tavola, in cui le e_{ij} individuano frequenze attese relative alle celle in indice.

	Successi	*Insuccessi*	Totale
Gruppo 1 Gruppo 2	e 11 e 21	e 12 e 22	R1 R2
Totale	C1	C2	*N*

Se la tavola delle frequenze osservate sarà molto diversa da quella delle frequenze attese il grado di significatività dei dati rispetto ad H_o sarà basso, presentando valori pari a 0,05 o inferiori.

La seguente tavola riporta un esempio su cui eseguiremo il test del χ^2 valori. Ad un gruppo di pazienti sono stati trapiantati due dosaggi di midollo osseo e l'osservazione del successo/insuccesso si è basata sul verificarsi o meno di situazioni di rigetto.

Dosaggio \ *Esito*	*Nessun rigetto*	*Rigetto*	Totale
< 3,0	19	17	36
>= 3,0	28	4	32
Totale	47	21	*68*

Per calcolare i valori attesi si utilizza l seguente formula: $e_{ij} = \frac{(R_i \cdot C_j)}{N}$

I valori delle celle vengono ottenuti nel seguente modo:

$$e_{11} = \frac{(36 \cdot 47)}{68} = 24{,}882 \qquad e_{12} = \frac{(36 \cdot 21)}{68} = 11{,}118$$

$$e_{21} = \frac{(32 \cdot 47)}{68} = 22{,}118 \qquad e_{22} = \frac{(32 \cdot 21)}{68} = 9{,}882$$

Dunque la tavola delle frequenze attese sarà:

Dosaggio \ *Esito*	*Nessun rigetto*	*Rigetto*	Totale
< 3,0	24,882	11,118	36
>= 3,0	22,118	9,882	32
Totale	47	21	*68*

La formula per calcolare il valore del test è la seguente:

$$\chi^2 = \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

che applicata all'esempio illustrato è pari a:

$$\chi^2 = \frac{(19 - 24{,}882)^2}{24{,}882} + \frac{(17 - 11{,}118)^2}{11{,}118} + \frac{(28 - 22{,}118)^2}{22{,}118} + \frac{(4 - 9{,}882)^2}{9{,}882} = 9{,}568$$

Una formula che migliora l'approssimazione di questo test è quella che contiene la cosiddetta *correzione di continuità di Yates*, che viene illustrata di seguito:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(\left|o_{ij} - e_{ij}\right| - \frac{1}{2}N\right)^2}{e_{ij}}$$

Come si può osservare, la formula non subisce trasformazioni di rilevo, ma al valore assoluto delle differenze tra valori osservati e attesi viene sottratta la metà della numerosità dei soggetti su cui viene eseguito il test.

L'esempio ci ha mostrato che la procedura per calcolare il valore del test può essere abbastanza laboriosa. Esiste una terza formula che, contenendo il fattore di correzione di Yates, permette di calcolare in modo più rapido il valore del χ^2. Utilizzando questa relazione si può addirittura fare a meno di calcolare il valore atteso delle frequenze per la tadi contingenza 2x2. La formula da utilizzare è la seguente:

$$\chi^2 = \frac{N\left(\left|o_{11}\cdot o_{22} - o_{12}\cdot o_{21}\right| - \frac{1}{2}N\right)^2}{R_1 \cdot R_2 \cdot C_1 \cdot C_2}$$

Se utilizziamo quest'ultima relazione sull'esempio finora utilizzato otteniamo:

$$\chi^2 = \frac{68\left(\left|19\cdot 4 - 17\cdot 28\right| - 34\right)^2}{36\cdot 32\cdot 47\cdot 21} = 8{,}011$$

Finora abbiamo illustrato come si ottiene il valore del test del χ^2, ma quale significato ha il valore ottenuto ? In generale si può affermare che se χ^2 è basso, il test è in accordo con l'ipotesi nulla H_o, viceversa se χ^2 è alto possiamo rigettare questa ipotesi, che, lo ricordiamo, indicava che entrambi i gruppi hanno la stessa probabilità di avere un successo.

Per calcolare il *grado di significatività del test* per un certo gruppo di osservazioni, occorre sommare le probabilità che la distribuzione della variabile χ^2 (poiché anche questa grandezza è governata da una curva di probabilità) assuma valori più alti di quello calcolato sulla base delle osservazioni, che per comodità chiameremo χ_0^2. Per ottenere questa probabilità io devo conoscere la distribuzione di χ^2 e andare a calcolare il valore di:

$$\Pr\left(\chi^2 \geq \chi_0^2\right)$$

Esistono delle tavole, che riportiamo in appendice, nelle quali è possibile individuare il il livello di significatività del valore del test ottenuto. Osservando le tavole si notano nella prima riga i valori di α, ovvero i vari livelli di significatività; in colonna invece

troviamo i livelli di k, che individuano i cosiddetti *gradi di libertà* che tratteremo in seguito, per adesso basti sapere che nel caso delle tavole di contingenza 2x2 il valore di k per il test del χ^2 è pari ad 1. Se scorriamo la prima riga possiamo osservare che il valore osservato ($\chi^2 = 8{,}01$) si trova tra i livelli di significatività α=0,005 e α=0,001, ovvero:

$$0{,}005 \le \Pr\left(\chi^2 \ge \chi_0^2\right) \le 0{,}001$$

Questi due valori indicano che il test ha un'approssimazione accurata, la cui significatività è superiore al 99,5%, ma non supera il 99,9%. Questo fatto ci permette di concludere che i dati non sono in accordo con l'ipotesi e quindi, nel caso dell'esempio, la frequenza di rigetto non è uguale in entrambi i gruppi.

Per applicare in modo efficace questo test occorre tenere presente una regola empirica che assicura una accurata approssimazione, ovvero bisogna fare attenzione a verificare che tutti le frequenze attese e_{ij} siano maggiori di 5. Una regola più compiacente si accontenta che i valori attesi siano maggiori di 2. In tutti questi casi sarebbe più opportuno ricorrere al test di Fisher.

Finora ci siamo occupati di dati espressi in forma dicotomica o binaria. Nella realtà i risultati di uno studio possono essere ben più complessi per rappresentare i quali può non essere sufficiente una tavola 2x2. In queste circostanze si possono utilizzare tavole di contingenza più ampie con *a* righe e *b* colonne. Siccome non potremo più parlare di successi/insuccessi e di gruppi 1 e 2, il modello di base di queste tavole di contingenza prevede che ciascuna osservazione del campione possa essere classificata secondo due schemi concettuali o *categorie di classificazione*.

La seguente tavola riporta un esempio di come si possono creare due schemi di classificazione per rappresentare i dati relativi alla cura della tubercolosi polmonare.

Analisi escreato / *Trattamento*	*Striscio +*	*Striscio +* *Cultura -*	*Striscio -* *Cultura -*	Totale
Solo PAS	56	30	13	99
Solo streptomicina	46	18	20	84
Entrambi	37	18	35	90
Totale	139	66	68	*273*

Se lo schema di classificazione A riguarda il trattamento, in riga compaiono i vari tipi di terapia a cui sono stati somministrati i soggetti. B, invece è lo schema di classificazione

riguardante la risposta al trattamento valutata sulla base dello striscio e della cultura, che, come si può osservare, possono avere esito positivo e negativo.

Anche nel caso di tavole di contingenza più ampie la formula da utilizzare per calcolare il valore del test è la stessa delle tavole 2x2. Anche il risultato sarà interpretato allo stesso modo, fatta eccezione per il calcolo dei gradi di libertà che in questo caso diventano determinanti per valutare la significatività del test. Il valore dei *gradi di libertà* (k), infatti, deve essere calcolato sulla base della seguente formula, in cui *a* e *b* sono rispettivamente il numero di righe ed il numero di colonne:

$$k = (a-1)\cdot(b-1)$$

Nel caso della tavola precedente il valore dei gradi di libertà è pari a k=4. La significatività del test sarà ricercata nelle tavole apposite che abbiamo già utilizzato, scorrendo i valori di α nella riga corrispondente al livello dei gradi di libertà k.

Anche in questi casi vale la regola empirica secondo cui le frequenze attese di ogni cella dovrebbero essere superiori a 5, ma possiamo ottenere risultati con una buona approssimazione se ci accontentiamo di tavole in cui il numero delle celle la cui frequenza attesa è inferiore a 5 non supera il 20% del numero delle celle contenute nella tavola di contingenza.

MISURE DI CORRELAZIONE

Il concetto di correlazione

Nel precedente capitolo abbiamo esaminato l'argomento dei test statistici ed abbiamo presentato una gruppo di strumenti analitici attraverso cui stabilire se tra gli attributi di due variabili vi possa essere un certo legame, sottoponendo questa ipotesi ad una verifica. Questi procedimenti, però, forniscono una risposta univoca che permette di accettare o rigettare le ipotesi poste. Altri strumenti analitici permettono di approfondire la conoscenza del legame tra due variabili determinandone non solo la misura ma anche la direzione.

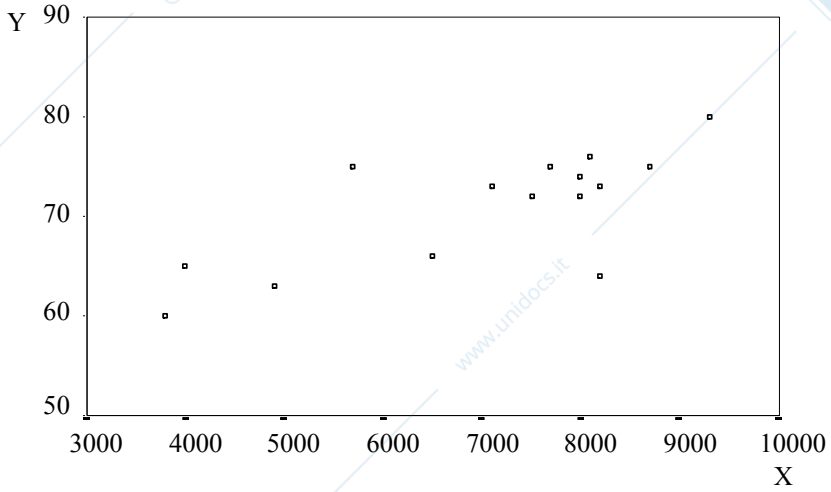
In questo capitolo parleremo di *misure di correlazione*, proponendo alcune metodologie che, applicate sui dati osservati, indicheranno se esista una relazione tra le variabili, quanto essa è forte ed in quale modo si esplica questa relazione. In altri termini correlare significa stabilire una reciproca relazione tra due ordini di grandezze, ovvero verificare se le variazioni che si verificano in una variabile possono o meno determinare delle corrispondenti variazioni nella seconda.

Nella seguente tavola riportamo un esempio che illustra in un gruppo di 10 soggetti i dati riguardanti la variabile X (reddito approssimato x 1.000) e la variabile Y (pressione diastolica).

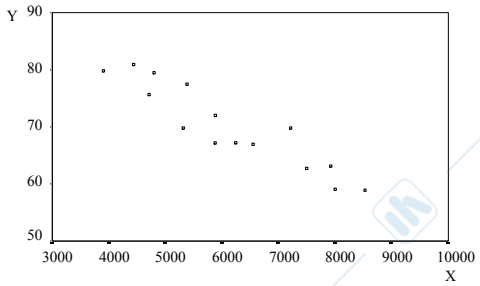
Soggetto	*Reddito* X	*Pressione* Y
A	3800	60
B	4000	65
C	4900	63
D	5700	75
E	6500	66
F	7100	73
G	7500	72
H	7700	75
I	8000	72
J	8000	74
K	8100	76
L	8200	64
M	8200	73
N	8700	75
O	9300	80

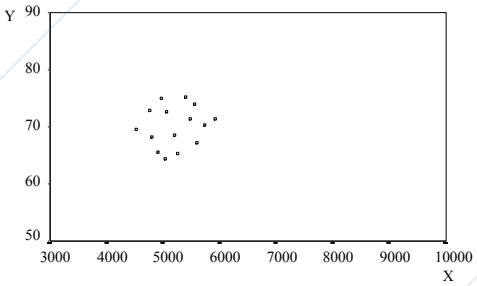
Il primo modo per verificare se possa esistere un legame tra le due variabili è quello di rappresentarle graficamente in un diagramma cartesiano, costruendo il *diagramma a punti* (scatterplot). In questo diagramma in ogni asse vengono riportati i valori di X e di Y, mentre nei quadranti vengono individuati i punti che corrispondono alle coppie di

dati relativi ad ogni osservazione. Osservando l'andamento dei punti che sono stati tracciati si può già evidenziare o meno la presenza di correlazione tra le due variabili. Sui dati riportati nella precedente tavola abbiamo costruito il seguente diagramma a punti.



Come si può osservare i punti si dispongono lungo una certa direttrice tanto da far ipotizzare che ad un aumento nella variabile X possa corrispondere un aumento anche nella variabile Y. Se denominiamo la X *variabile indipendente* e la Y *variabile dipendente*, allora potremmo affermare che vi è un legame diretto tra le due grandezze. Ciò, ma questo lo vedremo meglio in seguito, darà luogo ad una *correlazione positiva*.
Le osservazioni all'interno del diagramma apunti si possono disporre in modo anche molto differente. I seguenti diagrammi illustrano altre due situazioni con cui i dati si possono presentare.





Nel diagramma di sinistra le osservazioni assumono una tendenza decrescente e lasciano ipotizzare che al crescere della variabile X si verifichi una diminuzione del valore di Y secondo un rapporto inverso. Questa situazione dà luogo ad una *correlazione negativa*.

Nel diagramma di destra le osservazioni non presentano una direzione ben precisa, ma piuttosto tendono disporsi come una "nuvola di punti" attorno ad un valore centrale. In questo caso si potrebbe parlare di *assenza di correlazione* tra le due variabili in quanto il variare della prima non ha nessuna influenza sulle variazioni della seconda.
Se questo modo di procedere ci può fornire una prima approssimativa misura del legame esistente tra due variabili, i metodi che illustreremo ci offrono il valore di quanto sia forte e soprattutto se tale legame può essere ritenuto significativo.

Il coefficiente di correlazione per ranghi di Spearman ρ

Fra tutti gli stimatori basati sui ranghi, il *coefficiente di correlazione per ranghi di Spearman* è stato tra i primi ad essere elaborato e a tutt'oggi è forse il meglio conosciuto. Questo coefficiente, che noi indicheremo con *rho* (ρ) ma che spesso viene anche indicato con r_s, fornisce una misura dell'associazione la quale richiede che entrambe le variabili siano misurate almeno da una scala ordinale, in modo tale che gli oggetti o individui possano essere ordinati in una graduatoria, ovvero disposti nei *ranghi* di due serie ordinate.
Volendo studiare l'associazione tra concetti quali autoritarismo e pregiudizi sociali, la seguenta tavola illustra i punteggi assegnati ad affermazioni riguardanti i suddetti concetti in un gruppo di 12 studenti.

Soggetto	*Autoritarismo* X	*Preg. sociali* Y
A	82	42
B	98	46
C	87	39
D	40	37
E	116	65
F	113	88
G	111	86
H	83	56
I	85	62
J	126	92
K	106	54
L	117	81

Da questa tavola possiamo ottenere la tavola dei ranghi che riporta la posizione riportata in una graduatoria crescente dai valori delle osservazioni delle due variabili, tavola che riproduciamo di seguito. Per esempio il soggetto A presenta un punteggio che lo colloca al secondo posto nella graduatoria crescente dell'autoritarismo, mentre risulta terzo in quella dei pregiudizi sociali.

Soggetto	Ranghi *Autoritarismo* X	Ranghi *Preg. sociali* Y
A	2	3
B	6	4
C	5	2
D	1	1
E	10	8
F	9	11
G	8	10
H	3	6
I	4	7
J	12	12
K	7	5
L	11	9

Secondo quanto abbiamo detto sul legame tra le due variabili, se vi fosse correlazione perfetta tra autoritarismo e pregiudizi sociali i ranghi dovrebbero coincidere cioè indicando con x_i il rango di X e con y_i il rango di Y, dovrebbe risultare che $x_i=y_i$. Dunque, appare logico utilizzare la loro differenza quale elemento fondamentale per misurare la discordanza delle due serie di ranghi. Il test, infatti, è basato sulla differenza $d_i = x_i - y_i$ e la formula da applicare per il suo calcolo è la seguente, acui segue la tabella contenente i calcoli necessari per l'esempio proposto:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n \cdot \left(n^2 - 1\right)}$$

Soggetto	*Ranghi* *Autoritarismo* X	*Ranghi* *Preg. sociali* Y	d_i	d_i^2
A	2	3	-1	1
B	6	4	2	4
C	5	2	3	9
D	1	1	0	0
E	10	8	2	4
F	9	11	-2	4
G	8	10	-2	4
H	3	6	-3	9
I	4	7	-3	9
J	12	12	0	0
K	7	5	2	4
L	11	9	2	4
				$\sum d_i^2 = 52$

I valori della tavola ci permettono di completare il calcolo di ρ applicando la formula indicata, ovvero:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n \cdot \left(n^2 - 1\right)} = 1 - \frac{6 \cdot (52)}{12 \cdot \left(12^2 - 12\right)} = 0{,}82$$

Possiamo concludere che tra questi 12 studenti la correlazione fra autoritarismo e pregiudizi sociali è pari a ρ=0,82.

Siccome il valore massimo di ρ è pari ad 1 che si verifica nel caso in cui si ha correlazione perfetta ovvero $d_i = x_i - y_i = 0$, potremmo affermare che il valore ottenuto indica un certo legame tra le due variabili.

Il problema che ci dobbiamo porre a questo punto è quello di verificare se questa associazione si presenta anche nella popolazione. Se i soggetti su cui sono stati calcolati i punteggi appartengono ad un campione casuale della popolazione di riferimento, è necessario eseguire un *test di significatività*, stabilendo quale ipotesi nulla che le due variabili non sono assiciate nell'universo e che il valore osservato di ρ differisce da zero solo per effetto del caso. Per effettuare il test di significatività bisogna distinguere due casi: piccoli e grandi campioni.

Nel caso di *piccoli campioni* (ipotizzando vera l'ipotesi nulla) è possibile utilizzare la carta riportata in appendice nella quale in funzione del numero di coppie di osservazioni e del valore assunto dal coefficiente di correlazione per ranghi è possibile stabilire la significatività di quest'ultimo a due livelli di probabilità (α=0,05 e α=0,01). Nel caso dell'esempio sopra citato, avendo una numerosità di coppie pari a 12 ed un valore di ρ pari a 0,82 il coefficiente di correlazione per ranghi di Spearman risulta significativo ad un livello superiore a 0,01, ovvero superiore al 99%, e possiamo rigettare l'ipotesi nulla.

Se invece disponiamo di grandi campioni (per esempio una numerosità di coppie superiore a 20) la significatività in H_o di un valore osservato di ρ può essere determinata con una formula come la sguente:

$$\rho\sqrt{\frac{n-2}{1-\rho^2}} \sim t_{n-2}$$

Come si può osservare, la suddetta relazione si distribuisce come una t-Student con (n-2) gradi di libertà. Così la probabilità in H_o di un valore estremo come il valore osservato di ρ si può determinare utilizzando, allo stesso modo di come abbiamo già visto, le tavole della distribuzione t-Student.

La regressione ed il metodo dei minimi quadrati

Le metodologie statistiche illustrate finora riguardanti i test per la verifica delle ipotesi e più in generale inerenti lo studio dei legami esistenti tra coppie di variabili, muovono i passi da variabili a carattere qualitativo (sia nominale che ordinale) o da variabili quantitative che, a seguito di una successiva codifica sono state chiuse in classi e quindi sono assimilabili alla tipologia delle qualitative ordinali.

In questo paragrafo illustreremo una metodologia che, partendo dalla disponibilità di informazioni a carattere strettamente quantitativo, esamina tre aspetti dello studio dei legami tra coppie di variabili, ovvero:

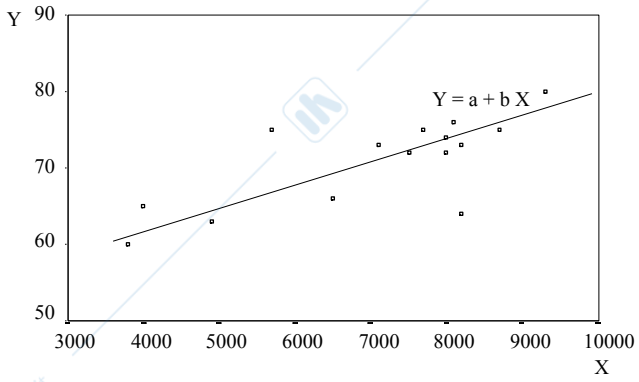
1. la *forma*;
2. l'*intensità*;
3. la possibilità di *prevedere* l'andamento di una variabile sulla base dei dati osservati per l'altra grandezza oggetto d'analisi.

All'inizio di questo capitolo abbiamo mostrato come uno dei primi criteri da utilizzare per esaminare l'esistenza di correlazioni tra coppie di variabili è quello di visualizzare mediante un diagramma a punti la disposizione delle osservazioni in un piano cartesiano. Abbiamo anche detto che quanto più è forte il legame tra le due grandezze, tanto maggiore sarà la tendenza dei punti a disporsi secondo una certa direzione (indipendentemente dal senso di quest'ultima).

Per descrivere la *forma del legame* potremmo scegliere di descrivere l''andamento delle osservazioni mediante una funzione matematica che, interpolando i dati, sia in grado di rappresentarne la tendenza e di conservarne le informazioni principali.

Il metodo della regressione lineare consiste appunto nell'individuare una retta che sia capace di rappresentare la distribuzione dei punti nel piano a due dimensioni. Come è facile immaginare, se i punti corrispondenti alle osservazioni si dispongono vicino alla retta, allora il modello scelto sarà in grado di descrivere con efficacia il legame tra le variabili.

In teoria esistono un numero infinito di rette che potrebbero essere tracciate per interpolare le osservazioni; ma in pratica vi è un solo modello matematico che ottimizza la rappresentazione dei dati. Il grafico alla pagina seguente illustra una ipotetica retta che è stata tracciata per interpolare i dati dell'esempio riportato nella tavola a fianco.



Soggetto	*Reddito* X	*Pressione* Y
A	3800	60
B	4000	65
C	4900	63
D	5700	75
E	6500	66
F	7100	73
G	7500	72
H	7700	75
I	8000	72
J	8000	74
K	8100	76
L	8200	64
M	8200	73
N	8700	75
O	9300	80

La forma analitica utilizzata è anche detta *retta di regressione*.
L'obiettivo da raggiungere è quello di individuare una retta, il cui modello generale è:

$$Y = a + b\,X$$

che rappresenti al meglio l'andamento dei dati. Nel modello precedente abbiamo utilizzato come *variabile indipendente* la X e come *indipendente* la Y. I parametri a e b, che rappresentano rispettivamente l'*intercetta* con l'asse delle Y ed il *coefficiente angolare* (ovvero la pendenza della retta), dovranno essere stimati sulla base delle osservazioni raccolte per le due variabili incluse nel modello. Una volta ottenuti i valori di questi parametri, attraverso una metodologia che illustreremo tra breve, il modello analitico è in grado di fornire un valore stimato per le osservazioni della variabile Y (y_i, con i=1, ... , n), stime che indicheremo con $\hat{Y}$, le cui determinazioni sono $\hat{y}_i$ (i=1, ... , n).
Tra le rette che si possono tracciare per interpolare i dati la più adatta a rappresentarne l'andamento sarà certamente quella che minimizza la somma delle distanze tra i valori osservati della variabile Y (y_i) ed i relativi valori stimati mediante il modello ($\hat{y}_i$). Il problema principale è quello di scegliere una coppia di valori numerici per i parametri a e b tale che sia minima la media dei quadrati degli scarti fra i valori empirici della Y ed i valori forniti dalla retta di regressione, ovvero deve essere minimizzata la seguente funzione:

$$\psi(a,b) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

dove n rappresenta, come al solito, la numerosità delle osservazioni.
La precedente relazione, che permette di individuare la retta di regressione, è basata sul cosiddetto criterio dei *minimi quadrati*. La formula può essere altrimenti esplicitata come segue, sostituendo al posto del valore stimato il modello analitico mediante il quale è stato ottenuto, ovvero:

$$\psi(a,b)=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-a-b\cdot x_i\right)^2$$

Abbiamo ricavato una funzione in due variabili (a e b): dobbiamo trovare adesso la coppia dei valori delle variabili che rendono minima la funzione $\psi(a,b)$. Utilizzando semplici nozioni di analisi matematica, possiamo individuare il valore minimo, trovando i valori di a e b per i quali si annullano le derivate parziali della funzione calcolate rispetto alle suddette variabili.
Sulla base di queste considerazioni, otteniamo il sistema di equazioni seguente:

$$\begin{cases}\dfrac{\psi(a,b)}{\partial a}=-\dfrac{2}{n}\displaystyle\sum_{i=1}^{n}\left(y_i-a-b\cdot x_i\right)=0\\[2ex]\dfrac{\psi(a,b)}{\partial b}=-\dfrac{2}{n}\displaystyle\sum_{i=1}^{n}\left(y_i-a-b\cdot x_i\right)\cdot x_i=0\end{cases}$$

Da cui, cambiando il segno, dividendo per 2 e sciogliendo le sommatorie, otteniamo:

$$\begin{cases}\dfrac{\psi(a,b)}{\partial a}=\dfrac{\displaystyle\sum_{i=1}^{n}y_i}{n}-\dfrac{n\cdot a}{n}-\dfrac{\displaystyle\sum_{i=1}^{n}b\cdot x_i}{n}=0\\[2ex]\dfrac{\psi(a,b)}{\partial b}=\dfrac{\displaystyle\sum_{i=1}^{n}x_i\cdot y_i}{n}-\dfrac{\displaystyle\sum_{i=1}^{n}a\cdot x_i}{n}-\dfrac{\displaystyle\sum_{i=1}^{n}b\cdot x_i^2}{n}=0\end{cases}$$

$$\begin{cases}\dfrac{\psi(a,b)}{\partial a}=\dfrac{\displaystyle\sum_{i=1}^{n}y_i}{n}-a-b\cdot\dfrac{\displaystyle\sum_{i=1}^{n}x_i}{n}=0\\[2ex]\dfrac{\psi(a,b)}{\partial b}=\dfrac{\displaystyle\sum_{i=1}^{n}x_i\cdot y_i}{n}-a\cdot\dfrac{\displaystyle\sum_{i=1}^{n}x_i}{n}-b\cdot\dfrac{\displaystyle\sum_{i=1}^{n}x_i^2}{n}=0\end{cases}$$

Sostituendo al posto delle sommatorie la relativa media aritmetica si ricava il sistema:

$$\begin{cases} \dfrac{\psi(a,b)}{\partial a} = m_y - a - b \cdot m_x = 0 \\ \dfrac{\psi(a,b)}{\partial b} = \dfrac{\sum_{i=1}^{n} x_i \cdot y_y}{n} - a \cdot m_x - b \cdot \dfrac{\sum_{i=1}^{n} x_i^2}{n} = 0 \end{cases}$$

Utilizzando il metodo della sostituzione per trovare la soluzione del sistema otteniamo:

$$\begin{cases} a = m_y - b \cdot m_x \\ \dfrac{\sum_{i=1}^{n} x_i \cdot y_y}{n} - (m_y - b \cdot m_x) \cdot m_x - b \cdot \dfrac{\sum_{i=1}^{n} x_i^2}{n} = 0 \end{cases}$$

$$\begin{cases} a = m_y - b \cdot m_x \\ \dfrac{\sum_{i=1}^{n} x_i \cdot y_y}{n} - m_x \cdot m_y + b \cdot m_x^2 - b \cdot \dfrac{\sum_{i=1}^{n} x_i^2}{n} = 0 \end{cases}$$

Da cui, dopo opportuni passaggi ricaviamo che:

$$\begin{cases} a = m_y - b \cdot m_x = m_y - \dfrac{S_{xy}}{S_x^2} \cdot m_x \\ b = \dfrac{S_{xy}}{S_x^2} \end{cases}$$

Le due soluzioni del sistema indicano i metodi di stima dei parametri a e b del modello di regressione lineare stimato con il metodo dei minimi quadrati. In essi compare una grandezza di cui finora non abbiamo mai parlato. Si tratta della cosiddetta *covarianza* (S_{xy}), la cui formula di calcolo è:

$$S_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - m_x) \cdot (y_i - m_y)$$

in cui m_x e m_y sono rispettivamente le medie aritmetiche delle determinazioni delle variabili X e Y. Il termine stesso "covarianza" sta ad indicare che questa quantità è adatta a misurare la "covariazione" di due variabili e che, quindi, è adatta a mettere in evidenza se ed in che misura due grandezze tendono a variare nello stesso senso o in senso contrario. Per esempio, un valor positivo di S_{xy} indica che al crescere di X anche la Y tende ad aumentare, mentre un valore negativo indica il contrario. La covarianza è vicina a zero quando il variare di X non da informaioni sul valore di y.

La seguente tabella fornisce gli elementi di calcolo per stimare i parametri a e b della retta di regressione sulla base dei dati relativi all'ultimo esempio.

Soggetto	Reddito X	Pressione Y	(x_i-m_x)	$(x_i-m_x)^2$	(y_i-m_y)	$(y_i-m_y)^2$	$(x_i-m_x)(y_i-m_y)$
A	3800	60	-3246,67	10540866,09	-10,87	118,16	35291,30
B	4000	65	-3046,67	9282198,09	-5,87	34,46	17883,95
C	4900	63	-2146,67	4608192,09	-7,87	61,94	16894,29
D	5700	75	-1346,67	1813520,09	4,13	17,06	-5561,75
E	6500	66	-546,67	298848,09	-4,87	23,72	2662,28
F	7100	73	53,33	2844,09	2,13	4,54	113,59
G	7500	72	453,33	205508,09	1,13	1,28	512,26
H	7700	75	653,33	426840,09	4,13	17,06	2698,25
I	8000	72	953,33	908838,09	1,13	1,28	1077,26
J	8000	74	953,33	908838,08	3,13	9,80	2983,92
K	8100	76	1053,33	1109504,09	5,13	26,01	5403,58
L	8200	64	1153,33	1130170,09	-6,87	47,20	-7923,38
M	8200	73	1153,33	1130170,09	2,13	4,54	2456,59
N	8700	75	1653,33	2733500,09	4,13	17,06	6828,25
O	9300	80	2253,33	5077496,09	9,13	83,36	20572,90
Totali	*105700*	*1063*		*40577333,35*		*467,47*	*101893,29*

$$n = 15$$

$$m_x = \frac{105700}{15} = 7046,67 \qquad m_y = \frac{1063}{15} = 70,87$$

$$S_x^2 = \frac{40577333,35}{15} = 2705155,56 \qquad S_{xy} = \frac{101893,29}{15} = 6792,89$$

I parametri della retta di regressione risultano essere:

$$b = \frac{S_{xy}}{S_x^2} = \frac{6792,89}{2705155,56} = 0,0025$$

$$a = m_y - b \cdot m_x = 70,87 - (0,0025 \cdot 7046,67) = 53,25$$

Da cui la retta di regressione è: $Y = 53,25 + 0,0025X$

Se questo è il modello generale che descrive il legame tra la coppia di variabili X e Y, i parametri ottenuti permettono di stimare i valori della Y sulla base delle determinazioni della X. Siccome il modello utilizzato non esprime una relazione lineare perfetta tra le osservazioni, ma determina la retta che minimizza gli scarti tra i valori osservati della Y e quelli stimati $\hat{Y}$, la differenza tra il valore assunto dalle y_i e quello stimato mediante il modello $\hat{y}_i = a + b \cdot x_i$ ci permette di determinare i *residui* della regressione che indicheremo con e_i. La seguente tavola contiene gli elementi di calcolo per la determinazione dei residui utilizzando la retta di regressione calcolata nell'esempio precedente.

Soggetto	Reddito X	Pressione Y	Press. stimata $\hat{Y}$	Residui e_i
A	3800	60	62,75	-2,75
B	4000	65	63,25	1,75
C	4900	63	65,5	-2,75
D	5700	75	67,5	7,5
E	6500	66	69,4	-3,5
F	7100	73	71	2
G	7500	72	72	0
H	7700	75	72,5	2,5
I	8000	72	73,25	-1,25
J	8000	74	73,25	0,75
K	8100	76	73,5	2,5
L	8200	64	73,75	-9,75
M	8200	73	73,75	-0,75
N	8700	75	75	0
O	9300	80	76,5	3,5
Totali	*105700*	*1063*		

L'argomento dei residui sarà ripreso in seguito ed approfondito per esaminare la signbificatività del modello di regressione. Adesso è più opportuno soffermarci ad osservare nel dettaglio il risultato ottenuto. Esaminiamo, innanzi tutto, il significato del parametro b. Come abbiamo visto, tale parametro individua la pendenza della retta. Nel caso del nostro esempio, l'osservazione empirica dell'andamento dei dati mediante il diagramma a punti ci ha fatto ipotizzare un legame diretto tra le due variabili, tendenza che è confermata dal valore positivo di b, dal quale deduciamo che la retta di regressione ha un andamento crescente, anche se la sua pendenza è molto bassa.

Per quanto riguarda il parametro a, abbiamo già detto che individua il punto di intersezione tra l'asse delle ordinate (Y) e la retta di regressione. Il modello calcolato ci permette di affermare che quando il valore del reddito è nullo (X=0), allora la pressione sistolica dovrebbe essere pari a 53,25. Per fare questa affermazione non ci siamo basati sui dati osservati, ma abbiamo estrapolato una "previsione" dal modello analitico. Anche per chi non è medico, è facile intuire che ci troviamo di fronte ad una osservazione che ha del paradossale.

Ma allora è colpa del nostro modello se non otteniamo dei dati plausibili ? Questa domanda ci ripropone un problema che abbiamo già accennato in precedenza e che tratteremo in linea generale, richiedendo uno spazio eccessivo rispetto alle attuali necessità del corso: la *previsione*.

Abbiamo precisato che la regressione, e di conseguenza la correlazione, oltre a descrivere la forma e l'intensità del legame esistente tra le due variabili, determinano un modello che può essere utilizzato anche per prevedere i valori di una variabile sulla base delle osservazioni dell'altra. Il precedente esempio sembrerebbe smentire quanto

abbiamo appena detto, ma in relatà il modello di regressione lineare è un ottimo strumento previsivo purché utilizzato con opportuni criteri ed accorgimenti.

Innanzi tutto, è opportuno precisare che nel calcolare delle previsioni non dobbiamo mai cercare valori molto distanti da quelli osservati. Nel nostro caso, il modello fornirà previsioni attendibili, per esempio, se volessimo sapere quale dovrebbe essere la pressione se il reddito fosse pari a x_1=10.000, oppure x_2=3.300, oppure ancora x_3=10.700. In questi casi otterremmo valori plausibili di Y, che rispettivamente sono pari a y_1=78,25, y_2=61,5 e y_3=80. Se invece vorremmo prevedere i valori della Y per livelli di X molto lontani da quelli che individuano gli estremi del suo campo di variazione osservato, allora la stima ottenuta perderebbe di significatività, essendo affetta da un livello di errore molto elevato. Per esempio, in corrispondenza di un valore del reddito pari a x_4=40.000, otterremmo un valore della Y pari a y_4=153,25, il quale, rappresentando come è noto la pressione sistolica, ha un valore scarsamente plausibile.

Ovviamente l'attendibilitàdelle previsioni è ottimale se si considerano valori interni al campo di variazione della variabile dipendente.

Dunque, come abbiamo illustrato, la qualità delle previsioni è legata allo scarto tra il valore osservato della Y ed il suo valore stimato. Questi scarti altro non sono che i residui visti in precedenza.

Rimane da esaminare un aspetto del modello di regressione lineare, cioè quello che analizza l'*intensità* del legame tra le variabili. A tale scopo introduciamo il *coefficiente di correlazione lineare r di Pearson*, che misura la correlazione tra le variabili X e Y contenute nel modello. La formula di calcolo è la seguente:

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

in cui S_{xy} indica la covarianza e S_x e S_y sono rispettivamente le deviazioni standard di X e di Y. Sulla base dei dati contenuti nella tavola utilizzata quale esempio per il calcolo della regressione, calcoliamo adesso il valore deò coefficiente di correlazione lineare, che sarà:

$$S_{xy} = 6792,89 \qquad S_x = 1644,74 \qquad S_y = 5,56$$

$$r = 0,7428$$

Il coefficiente r ha un campo di variazione che varia tra -1 e 1. Se r=0 ci troviamo di fronte ad un caso in cui non vi è legame tra X e Y, ovvero le variazioni della Y non possono essere spiegate dalla X. Quanto più il suo valore si differenzia da 0 e si avvicina agli estremi del campo di variazione, tanto più è evidente la presenza di un

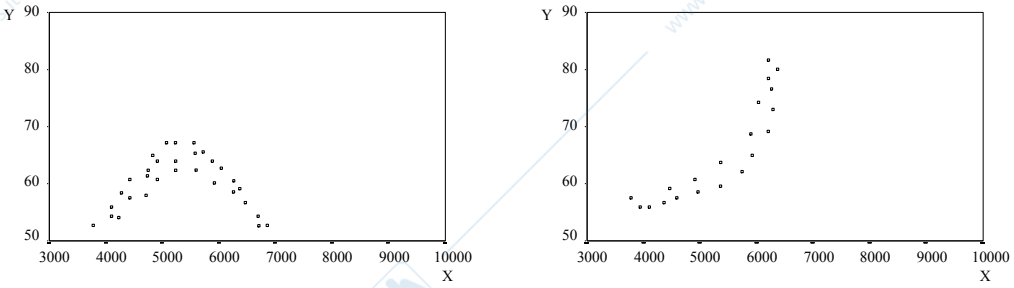
legame tra le suddette grandezze. In particolare, se r tende verso 1 (p.e.: r=0,85, r=0,92 o superiori), allora vi è una relazione diretta tra le variabili, cioè al crescere dell'una si verifica un aumneto anche nell'altra. Se invece r tende verso -1, allora la relazione è di tipo inverso, cioè al crescere di X la Y decresce. Nel primo caso la retta di regressione è crescente, mentre nel secondo caso l'andamento è decrescente.

Inoltre, e questo è l'aspetto più rilevante nell'analisi di r, valori molto vicini a 1 o -1 indicano una prossimità dei valori osservati alla curva di regressione, Se r fosse uguale a 1 o -1 allora arriveremmo al caso limite in cui ogni osservazione è allineata alla precedente e dunque i loro valori possono essere individuati con esattezza dalla retta interpolante.

Il valore di r che abbiamo ricavato nell'esempio (r=0,7428) indica la presenza di un certo legame tra le variabili, legame che per essere validato deve essere sottoposto all'analisi della varianza, ovvero ad un particolare test che permette di stabilire con quale significatività la X è in grado di spiegare la Y.

Per concludere questo paragrafo è opportuno sottolineare alcuni aspetti che riguardano la sua applicabilità. Finora abbiamo indicato la metodologia della regressione con aggettivandola con il termine *lineare*, ed abbiamo adottato un modello nel quale a fronte di aumenti unitari della X corrispondono variazioni proporzionali della Y.

Abbiamo suggerito questo modello portando ad esempio particolari distribuzioni dei dati all'interno del diagramma a punti. Se, invece, ci trovassimo di fronte ad altre forme con le quali i dati tendono a disporsi, come ad esempio i seguenti grafici illustrano, è opportuno chiedersi se il metodo della regressione lineare mantenga inalterato il suo valore interpretativo.



Come è facile osservare, un qualunque modello lineare, pur cogliendo un certo legame tra le variabili, non sarebbe in grado di individuare una funzione interpolante che si adatti sufficientemente all'andamento delle osservazioni.

Nei casi come quelli illustrati nell'esempio precedente è più indicato utilizzare una funzione diversa da quelle di tipo lineare, per esempio una parabola, adottando quindi il seguente modello:

$$Y = a \cdot X^2 + b \cdot X + c$$

Omettiamo il procedimento analitico attraverso cui si perviene alla stima dei parametri a, b e c del modello sopra descritto, limitandoci a sottolineare come, ancora una volta, l'applicazione della regressione debba essere precedetuta da un'analisi empirica delle osservazioni su un piano cartesiano, al fine di verificare che il modello lineare sia la forma interpretativa più adatta per l'interpolazione dei dati.

L'analisi della varianza

Come abbiamo detto, è possibile sottoporre atest l'ipotesi del legame tra le variabili X e Y che sono state utilizzate nel precedente modello della regressione lineare. Il metodo che utilizzeremo è denominato *analisi della varianza* ed è basato sulla scoposizione della variabilità totale di una variabile in varie componenti attribuibili a diversi fattori.
Con il termine *variabilità totale* di una variabile (VT) intendiamo indicare, nel caso della regressione, la variazione di una grandezza rispetto alsuo valore medio. Un primo indice sommario della variazione di Y è dato, come abbiamo visto in precedenza, dalla sommatoria:

$$\sum_{i=1}^{n}\left(y_i - m_y\right)^2$$

Se consideriamo le determinazioni della variabile Y e le stime $\hat{Y}$ ottenute mediante la regressione, queste due differiscono di un certo errore e_i, a meno che non vi sia una correlazione perfetta tra le variabili X e Y (siamo nel caso in cui r=1 e tutti le osservazioni sono allineate tra di loro. Dunque, possiamo scrivere che:

$$y_i = \hat{y}_i + e_i \qquad \text{ovvero} \qquad e_i = y_i - \hat{y}_i$$

Sottraendo da entrambi i membri la quantità m_y ottengo:

$$\left(y_i - m_y\right) = \left(\hat{y}_i - m_y\right) + e_i$$

Secondo quanto abbiamo detto finora, la prima parentesi esprime la variazione (o distanza) della variabile osservata dalla media; la parentesi a secondo membro indica la variazione della variabilità della variabile stimata dalla media ed infine troviamo i residui. Se facciamo la sommatoria dei quadrati di entrambi i membri otteniamo:

$$\sum_{i=1}^{n}\left(y_i - m_y\right)^2 = \sum_{i=1}^{n}\left[\left(\hat{y}_i - m_y\right) + e_i\right]^2 = \sum_{i=1}^{n}\left(\hat{y}_i - m_y\right)^2 + \sum_{i=1}^{n} e_i^2 + 2 \cdot \sum_{i=1}^{n}\left(\hat{y}_i - m_y\right) \cdot e_i$$

Sviluppando l'ultimo termine a secondo membro otteniamo:

$$2\cdot\sum_{i=1}^{n}\left(\hat{y}_i - m_y\right)\cdot e_i = 2\cdot\sum_{i=1}^{n}\left(a + b\cdot x_i - m_y\right)\cdot e_i = 2\cdot\sum_{i=1}^{n}\left(m_y - b\cdot m_x + b\cdot x_i - m_y\right)\cdot e_i =$$
$$= 2\cdot\sum_{i=1}^{n}\left(b\cdot x_i - b\cdot m_x\right)\cdot e_i = 2b\cdot\sum_{i=1}^{n}\left(x_i - m_x\right)\cdot e_i = 0$$

poiché la sommatoria degli scarti delle osservazioni di una variabile dalla sua media è nulla, per le proprietà della media aritmetica. Da ciò ricaviamo che:

$$\sum_{i=1}^{n}\left(y_i - m_y\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - m_y\right)^2 + \sum_{i=1}^{n} e_i^2$$

Nella precedente formula la sommatoria a primo membro indicac la variazione totale di Y (VT), la prima sommatoria a secondo membro la variazione dei valori sulla retta di regressione (VR) e l'ultima la somma dei residui al quadrato (VE).
È possibile scomporre ulteriormente il secondo membro nel seguente modo:

$$\sum_{i=1}^{n}\left(\hat{y}_i - m_y\right)^2 = \sum_{i=1}^{n}\left(a + b\cdot x_i - m_y\right)^2 = \sum_{i=1}^{n}\left(m_y - b\cdot m_x + b\cdot x_i - m_y\right)^2 =$$
$$= \sum_{i=1}^{n} b^2\cdot\left(x_i - m_x\right)^2 = b^2\cdot\sum_{i=1}^{n}\left(x_i - m_x\right)^2$$

Mediante questa serie di passaggi abbiamo scomposto la variabilità totale in una serie di componenti attribuibili a diversi fattori. Infatto possiamo riscrivere la relazione:

$$\sum_{i=1}^{n}\left(y_i - m_y\right)^2 = b^2\cdot\sum_{i=1}^{n}\left(x_i - m_x\right)^2 + \sum_{i=1}^{n} e_i^2 \qquad \text{come} \qquad VT = VR + VE$$

dove VT indica la *variazione totale*, VR la *variazione dei valori interpolati* delle y_i e VE la *variazione del residui* della regressione. In altre parole VR rappresenta l'effetto stimato di X sulla variazione di Y e VE l'effetto stimato della cmponente stocastica, ovvero quella attribuibile agli errori. Se dividiamo per VT l'ultima relazione otteniamo:

$$1 = \frac{VR}{VT} + \frac{VE}{VT}$$

in cui la prima frazione indica la quota della variazione totale di Y individuata dalla regressione e la seconda la quota residua. La prima frazione è anche nota come *coefficiente di determinazione* (R^2), ovvero:

$$R^2 = \frac{VR}{VT} = 1 - \frac{VE}{VT}$$

Da quanto detto ne consegue che: $0 \le R^2 \le 1$

Se R^2 è nullo ciò significa che VR=0 e cioè la X non riesce a spiegare le variazioni della Y, che dunque sono interamente dovute agli errori e_i (che altrimenti sono anche chiamati *componente stocastica*). Viceversa, quanto più R^2 si avvicina ad 1, tanto minore è l'influsso di questi ultimi.

Inoltre è anche dimostrabile che: $r^2 = R^2$

Per sottoporre a test l'ipotesi che la variazione della Y possa essere attribuita significativamente alla variabile X equivale a sottoporre a test la diversità o l'uguaglianza del parametro b della retta di regressione da zero.

Supponendo quindi di fissare quale ipotesi nulla H_o : b=0, il che significa ipotizzare che la variazione della Y non è imputabile alla variabile X ma a fattori casuali rappresentati dagli errori e_i, la formula da utilizzare per il test, denominato *F di Fisher*, è la seguente:

$$F = \frac{b^2 \cdot \sum_{i=1}^{n} (x_i - m_x)^2}{\sum_{i=1}^{n} e_i^2 \Big/ (n-2)} = \frac{VR/1}{VE/(n-2)}$$

Il valore della F calcolato con la suddetta formula è pari al rapporto tra la varizione causata dalla X, divisa per i suoi gradi di liberta (1), e la variazione residua della Y divisa anch'essa per i suoi gradi di libertà (n-2) (dove n esprime nuovamente la numerosità). Se questo rapporto non è significativamente diverso dall'unità, ciò significa che la variazione della Y attribuita alla regressione non è significativamente diversa da quella attribuita alla componente stocastica, per cui si può concludere che lipotesi nulla H_o : b=0 non può essere rigettata.

Per stabilire se la differenza tra le due variazioni sia significativa si fa ricorso alla statistica F di Fischer di cui riportiamo la tavola in appendice, cioè si confrontano le quantità:

$$\frac{VR/1}{VE/(n-2)} \quad \text{e} \quad F^{\alpha}_{1;n-2}$$

dove α è il livelllo di significatività prescelto (α=0,05; α=0,01; ....) e 1 e n-2 sono i gradi di libertà. Se il valore ottenuto dalle tavole risulta:

$$\frac{VR/1}{VE/(n-2)} \leq F^{\alpha}_{1;n-2}$$

allora l'ipotesi nulla H_o : b=0 (che indicava scarsa influenza della X sulle variazioni della Y) non può essere respinta. Se, invece:

$$\frac{VR/1}{VE/(n-2)} \geq F^{\alpha}_{1;n-2}$$

allora l'ipotesi nulla può essere rigettata e concludere che la variazione della Y può essere attribuita alla X con certo livello di probabilità α.

Applicando l'analisi della varianza all'esempio che abbiamo illustrato per calcolare la retta di regressione otteniamo:

$$b = 0,0025 \qquad \sum_{i=1}^{n}(x_i - m_x)^2 = 40577333,35 \qquad \sum_{i=1}^{n} e_i^2 = 213,19$$

$$F = \frac{b^2 \cdot \sum_{i=1}^{n}(x_i - m_x)^2}{\sum_{i=1}^{n} e_i^2 / (n-2)} = \frac{0,00000625 \cdot 40577333,35}{213,19/13} = 15,46$$

Segliendo un livello di significatività pari ad a=0,01 e con il numero di gradi di libertà osservati, possiamo ricavare dalle tavole della statistica F il seguente valore:

$$F^{0,01}_{1;13} = 9,07$$

e siccome: $F > F^{0,01}_{1;13}$

possiamo rigettare l'ipotesi nulla affermando che le variazioni della Y possono essere attribuite alla variabile X con una probabilità del 99%.

APPENDICE

TAVOLE PER TEST DI SIGNIFICATIVITÀ

Tavola 1: area della distribuzione normale standardizzata

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

Tavola 2: valori di $\chi^2_{\alpha,k}$

ν	0.20	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.0001
1	1.6424	2.7055	3.8415	5.0239	6.6349	7.8794	10.8274	12.1153	15.1343
2	3.2189	4.6052	5.9915	7.3778	9.2104	10.5965	13.8150	15.2014	18.4247
3	4.6416	6.2514	7.8147	9.3484	11.3449	12.8381	16.2660	17.7311	21.1040
4	5.9886	7.7794	9.4877	11.1433	13.2767	14.8602	18.4662	19.9977	23.5064
5	7.2893	9.2363	11.0705	12.8325	15.0863	16.7496	20.5147	22.1057	25.7507
6	8.5581	10.6446	12.5916	14.4494	16.8119	18.5475	22.4575	24.1016	27.8527
7	9.8032	12.0170	14.0671	16.0128	18.4753	20.2777	24.3213	26.0179	29.8814
8	11.0301	13.3616	15.5073	17.5345	20.0902	21.9549	26.1239	27.8674	31.8268
9	12.2421	14.6837	16.9190	19.0228	21.6660	23.5893	27.8767	29.6669	33.7247
10	13.4420	15.9872	18.3070	20.4832	23.2093	25.1881	29.5879	31.4195	35.5572
11	14.6314	17.2750	19.6752	21.9200	24.7250	26.7569	31.2635	33.1382	37.3647
12	15.8120	18.5493	21.0261	23.3367	26.2170	28.2997	32.9092	34.8211	39.1306
13	16.9848	19.8119	22.3620	24.7356	27.6882	29.8193	34.5274	36.4768	40.8735
14	18.1508	21.0641	23.6848	26.1189	29.1412	31.3194	36.1239	38.1085	42.5752
15	19.3107	22.3071	24.9958	27.4884	30.5780	32.8015	37.6978	39.7173	44.2596
16	20.4651	23.5418	26.2962	28.8453	31.9999	34.2671	39.2518	41.3077	45.9255
17	21.6146	24.7690	27.5871	30.1910	33.4087	35.7184	40.7911	42.8808	47.5591
18	22.7595	25.9894	28.8693	31.5264	34.8052	37.1564	42.3119	44.4337	49.1853
19	23.9004	27.2036	30.1435	32.8523	36.1908	38.5821	43.8194	45.9738	50.7873
20	25.0375	28.4120	31.4104	34.1696	37.5663	39.9969	45.3142	47.4977	52.3832
21	26.1711	29.6151	32.6706	35.4789	38.9322	41.4009	46.7963	49.0096	53.9599
22	27.3015	30.8133	33.9245	36.7807	40.2894	42.7957	48.2676	50.5105	55.5244
23	28.4288	32.0069	35.1725	38.0756	41.6383	44.1814	49.7276	51.9995	57.0668
24	29.5533	33.1962	36.4150	39.3641	42.9798	45.5584	51.1790	53.4776	58.6071
25	30.6752	34.3816	37.6525	40.6465	44.3140	46.9280	52.6187	54.9475	60.1360
26	31.7946	35.5632	38.8851	41.9231	45.6416	48.2898	54.0511	56.4068	61.6666
27	32.9117	36.7412	40.1133	43.1945	46.9628	49.6450	55.4751	57.8556	63.1660
28	34.0266	37.9159	41.3372	44.4608	48.2782	50.9936	56.8918	59.2990	64.6561
29	35.1394	39.0875	42.5569	45.7223	49.5878	52.3355	58.3006	60.7342	66.1524
30	36.2502	40.2560	43.7730	46.9792	50.8922	53.6719	59.7022	62.1600	67.6230
35	41.7780	46.0588	49.8018	53.2033	57.3420	60.2746	66.6192	69.1975	74.9253
40	47.2685	51.8050	55.7585	59.3417	63.6908	66.7660	73.4029	76.0963	82.0551
45	52.7288	57.5053	61.6562	65.4101	69.9569	73.1660	80.0776	82.8734	89.0704
50	58.1638	63.1671	67.5048	71.4202	76.1538	79.4898	86.6603	89.5597	95.9713
55	63.5772	68.7962	73.3115	77.3804	82.2920	85.7491	93.1671	96.1607	102.7735
60	68.9721	74.3970	79.0820	83.2977	88.3794	91.9518	99.6078	102.6971	109.4967
70	79.7147	85.5270	90.5313	95.0231	100.4251	104.2148	112.3167	115.5766	122.7443
80	90.4053	96.5782	101.8795	106.6285	112.3288	116.3209	124.8389	128.2636	135.7728
90	101.0537	107.5650	113.1452	118.1359	124.1162	128.2987	137.2082	140.7804	148.6198
100	111.6667	118.4980	124.3421	129.5613	135.8069	140.1697	149.4488	153.1638	161.3297

Tavola 3: valori di $t_{\alpha,k}$

	Area nella Coda di Destra sotto la Curva di Distribuzione *t*					
df	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356
34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340
36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313
40	1.303	1.684	2.021	2.423	2.704	3.307
∞	1.282	1.645	1.960	2.326	2.576	3.090