

STATISTICA MODERNA

contenuti:

- raccolta dei dati
- elaborazione numerica delle informazioni
- presentazione dei risultati

finalità:

- agevolare l'analisi e i processi decisionali
- trarre conclusioni sull'intera popolazione, anche quando si conoscono solamente i dati di uno o più campioni

componenti:

STATISTICA DESCRITTIVA

insieme dei metodi che riguardano raccolta, presentazione e sintesi di un insieme di dati per descriverne le caratteristiche essenziali

STATISTICA INFERENZIALE

insieme dei metodi con cui si possono elaborare i dati dei campioni per dedurre omogeneità o differenze nelle caratteristiche analizzate

Supponiamo di voler conoscere la velocità d'accrescimento somatico di una determinata specie animale o vegetale; è ovvio che non è possibile prendere tutti gli individui esistenti di quella specie, la **POPOLAZIONE** od **UNIVERSO**, ma solamente alcuni di essi, un **CAMPIONE**.

Quando poi si trattasse di misurare rapporti tra organi interni di una specie animale, è ovvio che non si può pretendere di dissezionare tutti gli individui della specie.

Tuttavia le conclusioni devono non essere limitate ai pochi casi del campione utilizzato, ma estese a tutta la popolazione, per rivestire una effettiva importanza generale e contribuire alla costruzione di teorie scientifiche universalmente valide.

L'esigenza di metodi di statistica inferenziale deriva anche dalla necessità di ricorrere al **CAMPIONAMENTO** (*), affinché i dati analizzati in un numero relativamente ridotto di casi rappresentino in modo corretto le caratteristiche di tutta la popolazione.

La teoria della probabilità permette poi di verificare la **VEROSIMIGLIANZA** che i risultati del campione non si discostino dagli eventuali risultati che si sarebbero raggiunti analizzando tutta la popolazione o l'universo dei dati.

(*) Disegno sperimentale e campionamento sono le due fasi preliminari indispensabili ad una corretta impostazione degli esperimenti e della raccolta dei dati. Tuttavia la loro presentazione didattica richiede concetti complessi e metodologie sofisticate, che la limitata preparazione specifica delle persone non ancora esperte impone di affrontare in una fase successiva, allorché saranno più familiari terminologia statistica, concetti e metodi fondamentali dell'inferenza.

STATISTICA DESCRITTIVA PER DISTRIBUZIONI SEMPLICI

La conduzione dell'indagine (o ESPERIMENTO) è un percorso di ricerca scientifica articolabile in quattro fasi:

1 - disegno sperimentale

- osservazioni in natura e ripetizioni in laboratorio non raccolte ed attuate a caso, ma scelte e programmate in funzione della ricerca e delle ipotesi esplicative
- chiarire a priori la formulazione dell'**IPOTESI ESPLICATIVA** (alternativa all'**IPOTESI NULLA**)

Le eventuali differenze riscontrate dovranno essere imputate a

FATTORI CAUSALI SPECIFICI ?

o solamente a

FATTORI CASUALI IGNOTI ?

attribuibili alla naturale variabilità di misure e materiale utilizzato

2 - campionamento

- raccogliere i dati in funzione dello scopo della ricerca
- rispettare le caratteristiche della popolazione

Numero limitato di dati —> conclusioni generali —> tutta la popolazione
(UNIVERSO)

3 - descrizione dei dati raccolti per verificare l'adeguatezza di:

- disegno sperimentale
- campionamento
- analisi condotte
- risultati conseguiti

4 - utilizzo dei tests (programmati nel disegno sperimentale e in funzione dei quali è stato effettuato il campionamento)

processo logico-matematico che, mediante il calcolo di probabilità, porta alla conclusione di non poter respingere oppure di dover **respingere l'ipotesi nulla**

Soltanto con una corretta applicazione del campionamento e dei test di confronto statistico è possibile rispondere alla **DOMANDA INFERENZIALE** di verifica dell'ipotesi nulla:

**LE DIFFERENZE FRA LE OSSERVAZIONI EMPIRICHE
SONO DOVUTE A FATTORI PURAMENTE CASUALI**

? quale è la probabilità che, fra tutte le alternative possibili, si presenti proprio la situazione descritta dai dati raccolti ?

- probabilità alta (convenzionalmente \Rightarrow 5%) \longrightarrow **fattori casuali**
- probabilità bassa ($<$ 5%) \longrightarrow **fattori non casuali**
 cioé rientranti tra i criteri con cui i dati sono stati raggruppati

Analisi e conclusioni sono rese complesse fundamentalmente da tre aspetti:

- | | |
|--|--|
| errori nelle misurazioni | generati da strumenti e da differenti abilità degli sprimentatori |
| utilizzo di campioni | i dati utilizzati in una ricerca non sono mai identici a quelli rilevati nelle altre |
| fattori contingenti di disturbo | possono incidere in modo differente sul fenomeno indagato (es.: tempo, luogo, ...) |

TIPI DI DATI - SCALE DI MISURA

Ai due tipi fondamentali di variabili casuali sono associati due TIPI DI DATI:

- **QUALITATIVI** generati da risposte categoriali
- **QUANTITATIVI** generati da risposte numeriche e distinti in:
 - = **DISCRETI** derivano da un conteggio
 - = **CONTINUI** derivano da una misurazione

A proprietà formali differenti dei dati (che di conseguenza consentono operazioni differenti) sono associati quattro TIPI DI SCALE DI MISURA:

Scala **NOMINALE** (o **classificatoria**)

- livello più basso di misurazione
- utilizzata quando i dati possono essere raggruppati in categorie, eventualmente identificati con simboli
- gli individui attribuiti a classi diverse sono tra loro differenti; quelli della stessa classe sono tra loro equivalenti rispetto alla proprietà utilizzata nella classificazione
- l'attribuzione di numeri per identificare le varie categorie nominali (es.: i giocatori di squadre) non autorizza ad elaborare quei numeri come tali
- quesiti statistici: frequenze degli individui per categoria, per confronti tra loro o rispetto a valori attesi

Scala **ORDINALE** (o per **ranghi**)

- contiene una quantità di informazione superiore
- alla proprietà di equivalenza tra gli individui della stessa classe si aggiunge quella di gradazione tra le classi (es.: un reagente colora una serie di provette secondo la quantità di sostanza analizzata contenuta, consentendo di ordinare le provette in base all'intensità del colore)
- le risposte, apparentemente definite a livello nominale, possono venire espresse su scala ordinale (es.: giovane, adulto, anziano; insufficiente, sufficiente, discreto, buono, ottimo)
- eventuali rappresentazioni simboliche (es.: - -, -, =, +, ++)
- impossibilità di valutare la distanza tra livelli (es.: tra insufficiente e sufficiente c'è una distanza diversa che tra buono ed ottimo?)
- **SCALA MONOTONICA**: alle variabili è possibile applicare una serie di tests non parametrici, ma non quelli parametrici

Scala di **INTERVALLI**

- alle due caratteristiche della scala ordinale si aggiunge quella di poter misurare le distanze tra tutte le coppie di valori
- si fonda su una misura oggettiva e costante, anche se punto di origine e unità di misura sono arbitrari (es.: la temperatura misurata in gradi Celsius o Fahrenheit, i calendari)
- solo le differenze tra i numeri sono quantità continue ed ISOMORFICHE e possono essere tra loro sommate, elevate a potenza e divise, determinando quantità utilizzate nella statistica parametrica

Le misure della temperatura possono essere facilmente ordinate e le differenze tra loro sono direttamente confrontabili e quantificabili; le date con un calendario gregoriano, islamico, ebraico o cinese possono essere tra loro ordinate dalla più antica a quella più recente e le differenze temporali possono essere misurate con precisione oggettiva. Ma una temperatura di 40 gradi non è il doppio di 20 gradi e l'anno 2000 significa che è trascorso il doppio del tempo rispetto all'anno 1000 solamente con riferimento al punto di origine su cui ogni calendario si basa.

Scala di **RAPPORTI**

- alle tre proprietà della scala precedente aggiunge quella ad avere una origine reale
- è il tipo di misurazione più sofisticato e completo (es.: altezza, distanza, età, peso, reddito procapite)
- non solo le differenze ma gli stessi valori possono essere moltiplicati o divisi per quantità costanti senza che l'informazione ne risulti alterata
- 0 (zero) significa quantità nulla (a differenza di quanto avviene, per es., con la temperatura di 0 (zero) gradi Celsius)
- si possono usare la media geometrica ed il coefficiente di variazione, che richiedono che il punto 0 sia reale e non convenzionale
- può essere applicato qualsiasi test statistico

CLASSIFICAZIONE IN TABELLE

Prima di qualunque elaborazione, una serie di dati va ordinata e sintetizzata in

DISTRIBUZIONE DI FREQUENZA (o di intensità)

poichè una serie non ordinata non permette quasi mai di evidenziare le caratteristiche del fenomeno in esame.

ESEMPIO

Conteggio del numero di foglie (variabile discreta) spuntate su 45 rami di uguale lunghezza di una pianta in un dato intervallo di tempo :

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

Definire le classi:

- 1 - identificare il valore minimo (0 nell'esempio) e quello massimo (9 nell'esempio), contando quante volte compare ogni variabile
- 2 - dalla frequenza assoluta n_i si calcola la frequenza relativa f_i formata dal rapporto tra la frequenza assoluta n_i ed il numero totale di casi N

E' utile soprattutto quando si vogliono confrontare due o più distribuzioni dello stesso fenomeno, ognuna con un numero differente di osservazioni

ESEMPIO

Distribuzione di frequenze assolute e relative (arrotondate) delle foglie di 45 rami:

classe (x_i)	0	1	2	3	4	5	6	7	8	9
freq. assol. (n_i)	3	3	7	12	7	5	4	3	0	1
freq. rel. (f_i)	0,07	0,07	0,15	0,27	0,15	0,11	0,09	0,07	0,00	0,02
freq.cumulata	0,07	0,14	0,29	0,56	0,71	0,82	0,91	0,98	0,98	1,00

Quante classi di frequenza costruire?

- da un minimo di 4-5 ad un massimo di 15-20 (prassi abituale) in funzione del numero complessivo di osservazioni. Infatti:

- se il numero di classi è troppo basso: perdita d'informazione sulle caratteristiche della distribuzione e la rende non significativa
- se il numero di classi è troppo alto: disperde i valori e non manifesta con evidenza la forma della distribuzione

Non è necessario costruire intervalli uguali; ma la loro rappresentazione grafica ed il calcolo dei parametri fondamentali esigono alcune avvertenze non sempre intuitive

ESEMPIO - parte a

Raggruppamento in classi di una variabile continua: altezza (cm) di 40 piante:

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

Procedura:

- 1 - individuare il valore minimo e massimo (64 e 198)
- 2 - stabilire l'intervallo di variazione, che ovviamente deve comprendere l'intero campo di variazione (cm 140, da cm 60 a cm 199 compresi)
- 3 - sulla base di N (40) si decide il numero di classi (nel caso specifico potrebbero essere 7, con passo 20)

avvertenze:

- 4 - definire con precisione il valore minimo e quello massimo di ogni classe, per evitare incertezze nell'attribuzione di un singolo dato tra due classi contigue
- 5 - la determinazione dei valori estremi, del numero di classi e dell'intervallo di ogni classe è soggettiva
- 6 - la scelta di una particolare serie al posto di un'altra può tradursi in un'immagine completamente diversa dei dati:
 - per piccoli campioni, l'alterazione e le differenze possono essere sensibili
 - per grandi campioni, gli effetti delle scelte soggettive, purchè non siano estreme, incidono meno sulla concentrazione dei dati
- 7 - la classe iniziale e terminale non devono essere aperte (es.: < 80 quella iniziale; ≥ 180 quella finale), poichè:
 - si perderebbe l'informazione del loro valore minimo e massimo e quindi del valore centrale (indispensabili per calcolare la media e gli altri parametri da essa derivati)
 - verrebbe impedita o resa soggettiva anche la rappresentazione grafica, per la quale è indispensabile conoscere i valori iniziale e terminale

ESEMPIO - parte b

Distribuzione di frequenza assoluta e relativa (%) dell'altezza delle 40 piante:

classe (x_j)	60-79	80-99	100-119	120-139	140-159	160-179	180-199
freq. ass. (n_j)	1	3	10	12	7	5	2
freq. rel. (f_j)	2,5	7,5	25,0	30,0	17,5	12,5	5,0
freq. cumul.	2,5	10,0	35,0	65,0	82,5	95,0	100,0

Rispetto all'elenco grezzo, la tabella di distribuzione delle frequenze fornisce diversi vantaggi:

POSIZIONE (o dimensione)
TENDENZA CENTRALE
VARIABILITÀ (o dispersione)
FORMA: simmetria
 curtosi

... e uno svantaggio:

non poter conoscere come sono distribuiti i dati entro ogni classe (per questa ragione, quando è richiesta la conoscenza di tutti i dati compresi in un particolare intervallo, viene usato il valore centrale di ogni classe)

N.B. Le distribuzioni delle frequenze relative o percentuali sono indispensabili quando si confrontano due o più gruppi di misure, che quasi mai presentano lo stesso numero di osservazioni

RAPPRESENTAZIONI GRAFICHE DI DATI QUANTITATIVI

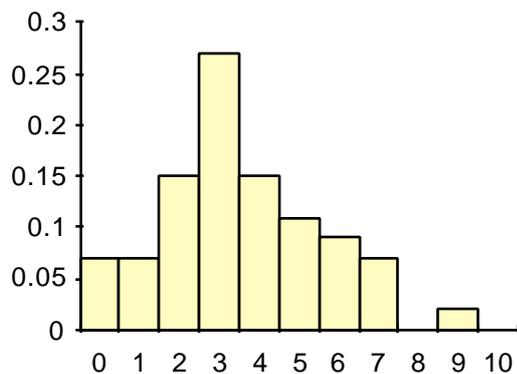
Le rappresentazioni grafiche forniscono:

- una sintesi visiva delle caratteristiche fondamentali delle distribuzioni
- impressioni percepite con maggiore facilità
- meno particolari.
- una descrizione espressa mediante una interpretazione soggettiva

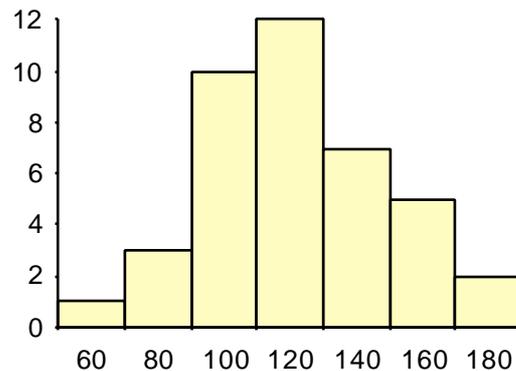
ISTOGRAMMI e POLIGONI

dati quantitativi raggruppati in distribuzioni di frequenza assoluta, o di frequenza relativa, o di percentuali

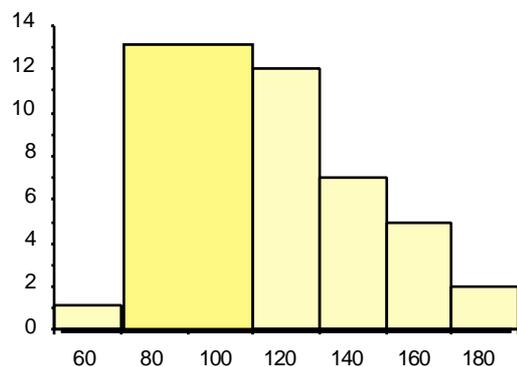
Istogrammi : grafici a barre verticali in cui i rettangoli vengono costruiti in corrispondenza degli estremi di ciascuna classe. La variabile casuale o il fenomeno di interesse è tracciato lungo l'asse x, mentre l'asse y rappresenta il numero assoluto (o la frequenza relativa o quella percentuale) con cui compaiono i singoli valori delle classi



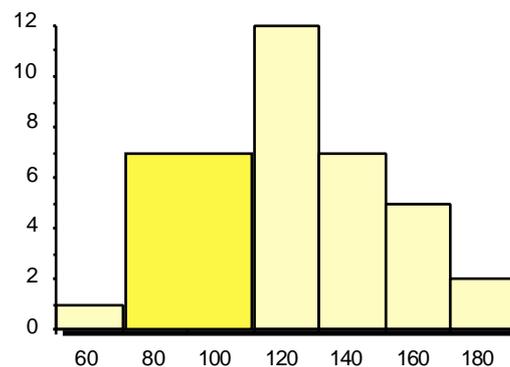
Dati di Tab. 2 (9 classi)



Dati di Tab. 4 (Val. iniz. = 60; Val. fin. = 199; Passo = 20; Classi = 7)



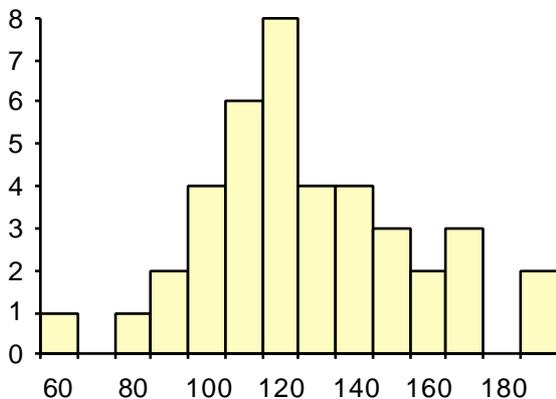
Somma errata di due classi



Somma corretta di due classi

Gli **ISTOGRAMMI** sono rappresentazioni grafiche di tipo areale

- aree dei rettangoli
proporzionali alle frequenze
- altezze dei rettangoli
proporzionali alle frequenze
- basi dei rettangoli :
ampiezze uguali → ragionare in termini di altezze o di aree è equivalente
ampiezze diverse → occorre rendere le altezze proporzionali dividendo il numero di osservazioni per il numero di classi contenute nella base
- asse verticale :
deve mostrare lo zero reale (o “origine”) al fine di non travisare le caratteristiche dei dati



(Valore iniziale = 60; Valore finale = 199; Passo = 10; Classi = 14)

Questa rappresentazione grafica non è significativa, a causa di una eccessiva suddivisione in classi

POLIGONI

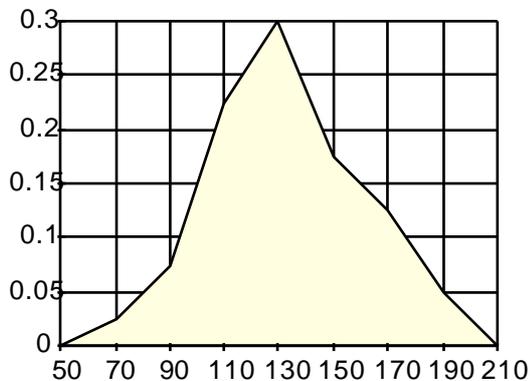
valori relativi o percentuali (simili agli istogrammi) ottenuti unendo con linea spezzata i punti centrali di ogni classe

- l'asse orizzontale rappresenta il fenomeno
- l'asse verticale rappresenta la proporzione o percentuale di ogni classe
- area sottesa : 1 per le frequenze relative; = 100 per le percentuali
- linea spezzata unita all'asse orizzontale all'inizio e alla fine (il valore centrale della 1.a classe con quello di una precedente classe fittizia di valore 0; il valore centrale dell'ultima classe con quello di una classe successiva fittizia di valore 0)

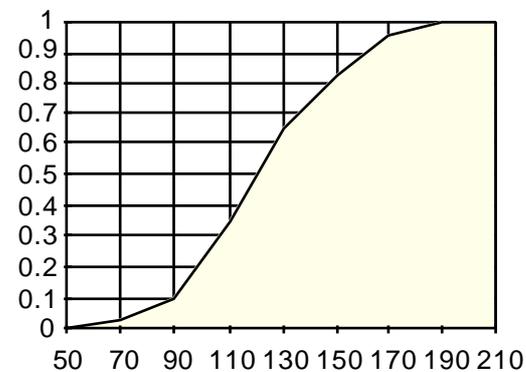
DISTRIBUZIONI CUMULATE E GRAFICI POLIGONALI

evidenziano quante sono in totale le misure inferiori o superiori ad un certo valore

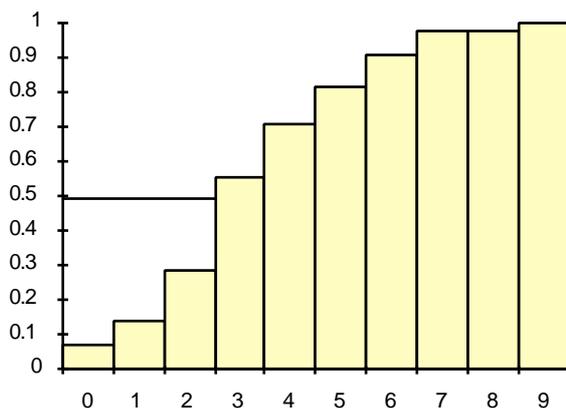
- il valore dell'asse orizzontale corrispondente al 50% dell'asse verticale identifica la mediana (importante quando la distribuzione dei dati è asimmetrica)



Poligono



Poligono cumulado



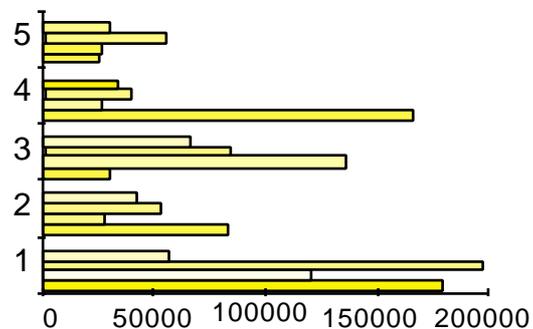
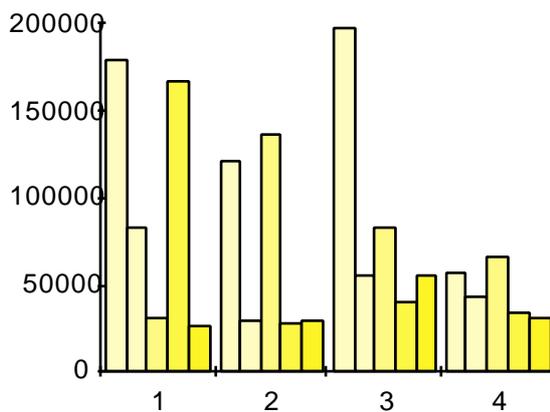
Istogramma cumulado

RAPPRESENTAZIONI GRAFICHE DI DATI QUALITATIVI

RETTANGOLI DISTANZIATI (o GRAFICI A COLONNE)

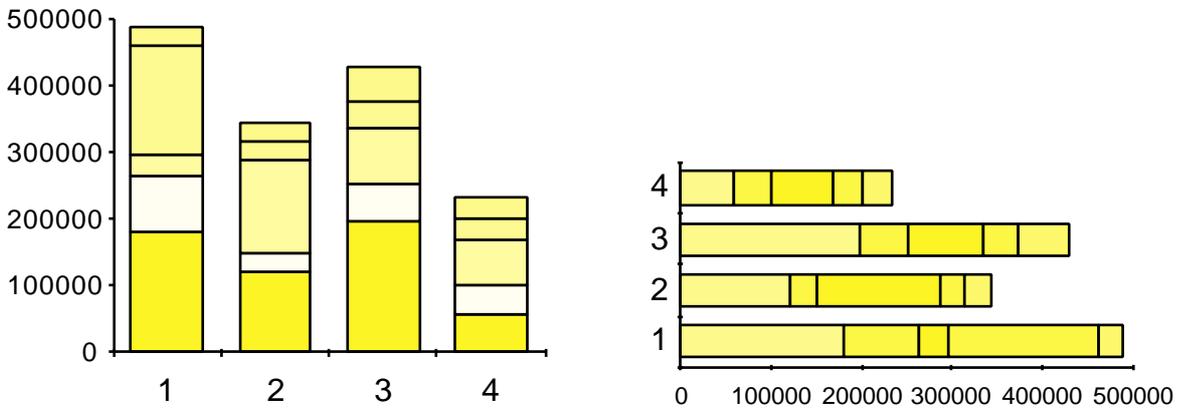
Si tratta di rettangoli con basi uguali ed altezze proporzionali alle intensità (o frequenze) corrispondenti ai vari gruppi considerati

- quando non esiste una logica specifica nell'ordine, i rettangoli o colonne vengono disposti dal maggiore al minore
- se le classi qualitative sono composte da sottoclassi, è possibile anche questa ulteriore rappresentazione grafica dividendo il rettangolo relativo in più parti, le cui altezze devono essere proporzionali alle frequenze delle sottoclassi
- avendo basi uguali, le aree sono proporzionali alle altezze, pertanto anche i diagrammi a rettangoli distanziati sono rappresentazioni areali



ORTOGRAMMI (o GRAFICI A NASTRI)

Sono simili ai rettangoli distanziati, ma con le classi di frequenza sequenziali sulla stessa barra per una migliore lettura



DIAGRAMMI A PUNTI

Si ottengono sostituendo ai rettangoli una linea punteggiata

- rappresentano molto bene le informazioni contenute in distribuzioni di frequenza di dati qualitativi

AREOGRAMMI

Sono superfici di figure piane (quadrati, rettangoli, cerchi o loro parti)

- utilizzati con frequenze o quantità di una distribuzione di variabile qualitativa
- la rappresentazione può avvenire:
 - con più figure dello stesso tipo aventi superfici proporzionali alle frequenze o quantità
 - con unica figura suddivisa in parti ugualmente proporzionali

AREOGRAMMI A TORTA (o CIRCOLARI)

E' un cerchio suddiviso in parti proporzionali alle classi di frequenza, come per i rettangoli.

MEDIANA

è il valore che occupa la posizione centrale in un insieme ordinato di dati

- Proprietà :

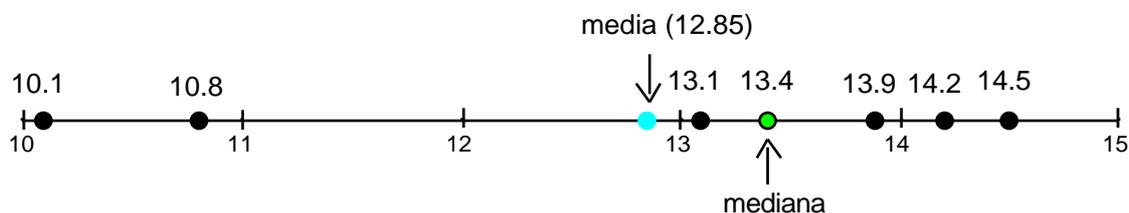
- non è influenzata dai valori estremi, ma solo dal numero delle osservazioni
- ogni osservazione estratta a caso ha la stessa probabilità di essere inferiore o superiore alla mediana

- Si usa :

- per attenuare l'effetto di valori estremi molto alti o bassi
- nel caso di scale ordinali o di ranghi

N.B. : Occorre ordinare i valori :

- se il campione ha un numero dispari di dati, la mediana è il valore del dato centrale, in posizione $(n+1)/2$
- se il campione ha un numero pari di dati, la mediana è la media aritmetica dei valori numerici dei due valori centrali (posizioni $n/2$ e $n/2+1$)



MODA

è il valore più frequente di una distribuzione

- Proprietà:

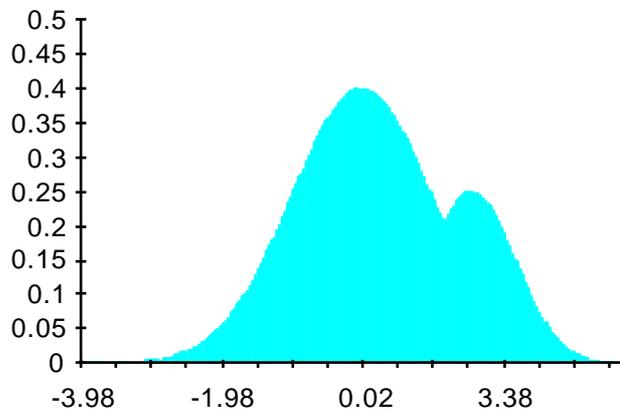
- non è influenzata dalla presenza di alcun valore estremo
- differisce quando con gli stessi dati si formano classi di ampiezza differente

- Si usa :

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

DISTRIBUZIONI DI FREQUENZA

- **UNIMODALI** → hanno un'unica moda
- **BIMODALI (e PLURIMODALI)** → hanno mode secondarie



Distribuzione bimodale

INTERVALLO MEDIO

E' la media aritmetica tra il valore più piccolo e quello più grande

- Proprietà : si calcola rapidamente anche con un elevato numero di dati
- Si usa :
 - quando non ci sono valori erratici (outliers) per evitare un valore dell'intervallo medio molto distorto
 - in meteorologia, nel caso di una serie di dati sulla temperatura o per il calcolo della precipitazione media mensile, essendo improbabile la presenza di valori estremi

MEDIA INTERQUARTILE

E' la media fra 1° e 3° interquartile (=mediane della 1ª e della 2ª metà dei dati)

- Proprietà : risente in misura molto più ridotta della presenza di valori estremi

TRIMEDIA (proposta da Tuckey) : $T = \frac{Q_1 + 2Q_2 + Q_3}{4}$

Q_2 = mediana; Q_1 e Q_3 = mediane della prima e seconda metà dei dati ordinati

- Si usa :
 - quando si dispone di materiale molto variabile
 - con una distribuzione molto asimmetrica (es.: misure d'inquinamento atmosferico con picchi anomali)

MISURE DI DISPERSIONE O VARIABILITA'

CAMPO DI VARIAZIONE (O INTERVALLO DI VARIAZIONE)

E' la differenza tra il valore massimo e il valore minimo

- Proprietà :
 - intuitivo e semplice, in particolare quando i dati sono ordinati
 - incapace di misurare come i dati sono distribuiti entro l'intervallo
 - risente della presenza di valori anomali
- Si usa : quando i valori delle osservazioni devono restare entro limiti prestabiliti

DIFFERENZA INTERQUARTILE

tra il 3° ed il 1° quartile (tra il valore della mediana della seconda metà e quello della mediana della prima metà della distribuzione)

$$\begin{array}{ccccccc} & \frac{1}{4} & & \frac{1}{2} & & \frac{3}{4} & & 1 \\ \text{---} & \frac{1}{4} & \text{---} & \frac{1}{2} & \text{---} & \frac{3}{4} & \text{---} & 1 \\ & Q_1 & & Q_2 & & Q_3 & & Q_4 \end{array}$$

QUANTILI (O FRATTILI) :

Sono misure di posizione non-centrale con esclusive finalità descrittive (ogni gruppo parziale contiene la stessa frazione di osservazioni)

- **DECILI** —> dividono i dati ordinati in decine
- **PERCENTILI** —> dividono i dati ordinati in centesimi
- Proprietà : individuano i valori che delimitano una % o frazione stabilita di valori estremi (es.: nel monitoraggio dell'inquinamento indicano i valori che rientrano nell'x% dei massimi o minimi)
- Si usano :
 - quando non si conosce la forma della distribuzione
 - quando la distribuzione è fortemente asimmetrica

SCARTO MEDIO ASSOLUTO (S_m) DALLA MEDIA (\bar{x})

$$S_m = \frac{\sum |x_i - \bar{x}|}{n} \quad \text{per dati semplici}$$

$$S_m = \frac{\sum |x_i - \bar{x}| \cdot n_i}{n} \quad \text{per dati ponderati con la frequenza } n_i \text{ di ogni classe}$$

x_i = valore del dato i-esimo in una distribuzione semplice e valore centrale della classe in una distribuzione di frequenza

n = n° totale di dati

n_i = n° di dati della classe i-esima in una distribuzione di frequenza

SCARTO MEDIO ASSOLUTO DALLA MEDIANA

E' la media degli scarti assoluti dei singoli dati dalla loro mediana e viene calcolato come sopra, sostituendo la mediana alla media

• Proprietà :

- rende minima la somma degli scarti assoluti
- è inferiore allo scarto medio assoluto dalla media (è uguale solo quando media e mediana coincidono)
- viene usato come misura di dispersione in alcuni test di statistica non parametrica

DEVIANZA (o SOMMA DEI QUADRATI degli scarti dalla media, SQ, SUM OF SQUARES, SS) **E' la base delle misure di dispersione dei dati**

Formule EURISTICHE :

$$\text{devianza (SQ)} = \sum (x_i - \bar{x})^2 \quad \text{per serie ordinate di dati}$$

$$\text{devianza (SQ)} = \sum (x_i - \bar{x})^2 n_i \quad \text{per dati in distribuzioni di frequenza}$$

FORMULA EMPIRICA (o ABBREVIATA) :

$$\text{devianza (SQ)} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$\sum x^2 = \text{sommatoria dei quadrati delle osservazioni}$$

$$(\sum x)^2 = \text{sommatoria totale quadrata}$$

$$n = \text{n° delle osservazioni}$$

ESERCIZIO

Calcolare la devianza (SQ) nei due modi descritti dei valori: 5 6 7 7 8 10

$$\bar{x} = \frac{5+6+7+7+8+10}{6} = \frac{43}{6} = 7,1\bar{6}$$

$$\begin{aligned} \text{devianza (SQ)} &= \sum (x_i - \bar{x})^2 = \\ &= (5 - 7,1\bar{6})^2 + (6 - 7,1\bar{6})^2 + (7 - 7,1\bar{6})^2 + (7 - 7,1\bar{6})^2 + (8 - 7,1\bar{6})^2 + (10 - 7,1\bar{6})^2 = \\ &= 4,665 + 1,3456 + 0,0256 + 0,0256 + 0,7056 + 8,0656 = 14,8356 \end{aligned}$$

$$\begin{aligned} \text{devianza (SQ)} &= \sum x^2 - \frac{(\sum x)^2}{n} = \\ &= (25 + 36 + 49 + 49 + 64 + 100) - \frac{43^2}{6} = 323 - \frac{1849}{6} = 323 - 308,1\bar{6} = 14,84 \end{aligned}$$

VARIANZA (o QUADRATO MEDIO, Mean Square, MS)

media dei quadrati degli scarti dei valori dalla loro media (devianza media)

V. DI UNA POPOLAZIONE devianza diviso il n° di osservazioni n

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

V. DI UN CAMPIONE devianza diviso n-1 (correzione di Student)

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

- nella statistica inferenziale, cioè quando si utilizzano i dati del campione per stimare le caratteristiche di una popolazione, si usa sempre la varianza campionaria
- n-1, n° di osservazioni indipendenti, è chiamato GRADI DI LIBERTÀ (gdl, df); poiché la somma degli scarti dalla media è uguale a zero, l'ultimo valore è fissato a priori e non è libero di assumere qualsiasi valore

DEVIAZIONE STANDARD (o **SCARTO QUADRATICO MEDIO**, σ per una popolazione; s per un campione)

E' la radice quadrata della varianza

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

- Proprietà :
 - è sempre un valore positivo
 - è una misura della dispersione della variabile casuale attorno alla media

COEFFICIENTE DI VARIAZIONE (CV)

Misura la dispersione percentuale relativa dei dati in rapporto alla media

$$cv = \left(\frac{\sigma}{\mu} \right) \cdot 100 \quad (\mu = \text{media}; \sigma = \text{deviazione standard})$$

- Proprietà :
 - è un numero puro svincolato da ogni scala di misura e dalla tendenza centrale del fenomeno studiato
 - in quanto rapporto, ha significato solo se calcolato per variabili misurate con una scala di rapporti
 - tende ad essere costante per ogni fenomeno (di solito oscilla tra il 5% e il 30%)
- Si usa per confrontare :
 - la variabilità di due o più gruppi con medie molto diverse
 - dati espressi in unità di misura diverse
 - popolazioni differenti per lo stesso carattere (es.: confronto tra la variabilità di specie animali di taglie diverse, come cani e cavalli)

N.B. Un C.V. molto basso (<5%) fa sospettare l'esistenza di un fattore limitante che abbassa notevolmente od elimina la variabilità; un C.V. molto alto (>50%) è indice di condizioni anomale (es.: quando in un gruppo animale gli individui mostrano grandi differenze nell'accrescimento, si può sospettare uno squilibrio alimentare).

VARIANZA IN DATI RAGGRUPPATI (CORREZIONE DI SHEPPARD (o CORREZIONE PER LA CONTINUITÀ)

In una distribuzione di frequenza di misure continue, il raggruppamento in classi approssima tutti i valori compresi nell'intervallo al loro valore centrale, e il loro risultato non coincide con quello calcolato sui dati reali

Se la distribuzione è normale, per il calcolo della media le approssimazioni a sinistra della media compensano quelle a destra e, tra i due sistemi di calcolo, si hanno solo differenze casuali di entità ridotta

Per il calcolo della varianza, le approssimazioni di segno opposto sono elevate al quadrato e dunque si sommano: la varianza reale calcolata dai dati originari è inferiore a quella calcolata sui raggruppamenti in classi, e le differenze crescono all'aumentare dell'ampiezza dell'intervallo delle classi

Alla varianza è calcolata su una distribuzione di dati raggruppati in classi, si deve apportare la correzione :

$$\sigma^2_{\text{reale}} = \sigma^2_{\text{calcolata}} - \frac{h^2}{12} \quad (h=\text{ampiezza delle classi})$$

ESEMPIO

In una distribuzione di frequenza in cui le classi hanno ampiezza costante con intervallo $h=10$ è stata calcolata una varianza $\sigma^2=50$. La varianza corretta, che si sarebbe ottenuta utilizzando i singoli valori, secondo Sheppard dovrebbe essere:

$$\sigma^2_{\text{reale}} = 50 - \frac{10^2}{12} = 50 - 8,3\bar{3} = 41,6\bar{6}$$

La relazione è valida per le popolazioni, mentre con pochi dati campionari, è difficile sapere se la distribuzione rispetta le condizioni fissate da Sheppard (essere continua, limitata ad un intervallo di ampiezza finito, le due code della distribuzione tendere a zero in modo graduale)

Per piccoli campioni la correzione potrebbe essere sbagliata e determinare un errore maggiore, per cui molti sperimentatori preferiscono non applicare la correzione

INDICI DI FORMA DI UNA DISTRIBUZIONE

Riguardano due caratteristiche :

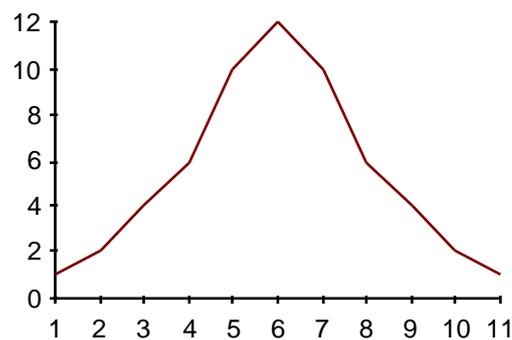
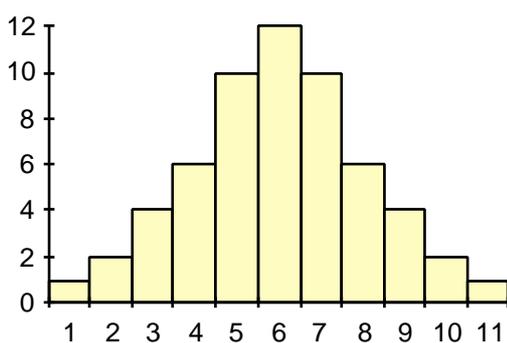
- **SIMMETRIA**
- **CURTOSI**

Caratteristiche:

- le misure sono ancora rudimentali
- le definizioni permangono equivoche

Si ha **SIMMETRIA** :

- nelle distribuzioni unimodali, quando:
media, moda e mediana coincidono
- nelle distribuzioni bimodali, quando :
solo media e mediana coincidono
- in qualunque distribuzione, quando :
i valori equidistanti dalla mediana
presentano la stessa frequenza
(questa è quindi una condizione che non
caratterizza la distribuzione in modo univoco)



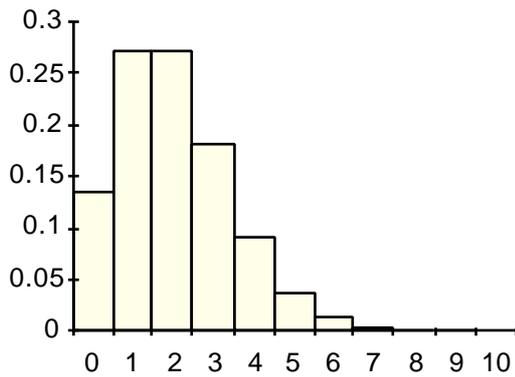
Distribuzioni simmetriche

Si ha **ASIMMETRIA A DESTRA** quando :

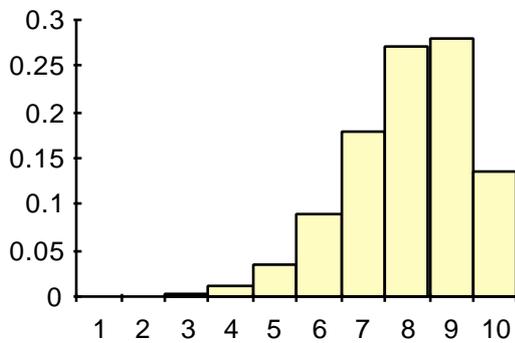
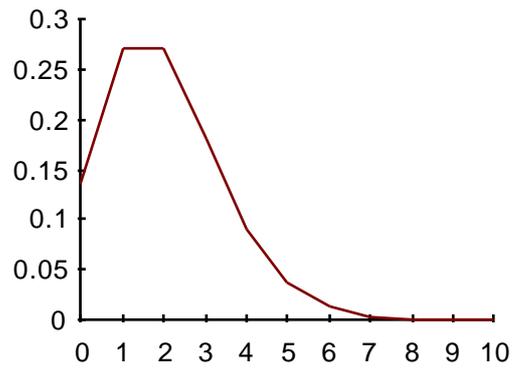
i valori maggiori sono più frequenti (la successione delle misure di tendenza centrale da sinistra a destra è: moda, mediana, media)

Si ha **ASIMMETRIA A SINISTRA** quando :

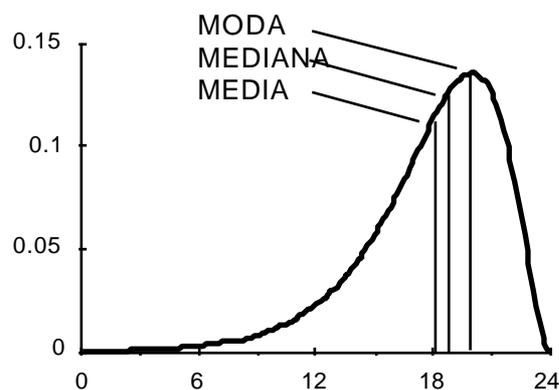
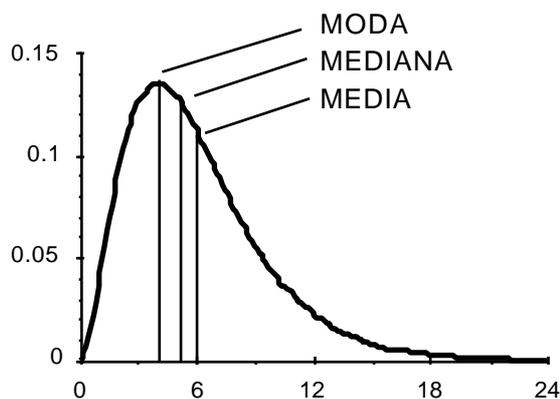
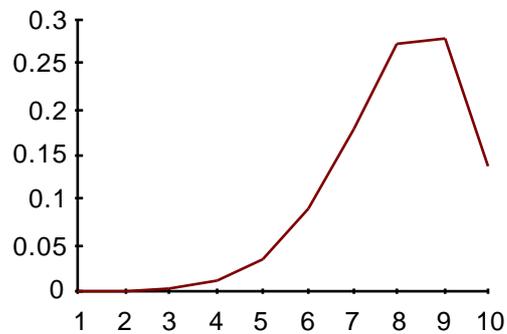
i valori minori sono più frequenti (la successione delle misure di tendenza centrale da sinistra a destra è: media, mediana, moda)



Distribuzione con asimmetria a sinistra



Distribuzione con asimmetria a destra



Attenzone alle

- **false simmetrie ...**

nella distribuzione 4 16 20 20 20 30 30 media, mediana e moda sono coincidenti (valore 20), ma la sua forma non è simmetrica

- **false asimmetrie ...**

analizzando la distribuzione dei dati di un campione, può capitare di rilevare un'asimmetria determinata dal ridotto numero di osservazioni, o da modalità inadeguate di raggruppamento in classi

In questi casi si parla di **ASIMMETRIA FALSA**, da distinguere dalla **ASIMMETRIA VERA** che esiste solo per le distribuzioni di popolazioni

INDICI DI ASIMMETRIA

- Dovrebbero essere = 0 se (e solo se) la distribuzione è simmetrica
- Non godono della stessa proprietà degli indici di variabilità o dispersione
 - quando la distribuzione è simmetrica sono nulli
 - quando la distribuzione è asimmetrica possono essere nulli

INDICI DI ASIMMETRIA ASSOLUTA

si esprimono con le distanze tra la media e la moda o la mediana

DIFFERENZA TRA MEDIA E MODA :

$$d = \text{media} - \text{moda}$$

$d = 0$ \longrightarrow la curva è simmetrica

$d > 0$ \longrightarrow la curva ha asimmetria positiva (o sinistra) :
media > mediana > moda

$d < 0$ \longrightarrow la curva ha asimmetria negativa (o destra) :
media < mediana < moda

INDICI DI ASIMMETRIA RELATIVA

Si utilizzano per confrontare l'asimmetria di più distribuzioni con valori differenti

SKEWNESS DI PEARSON (sk)

INDICE γ_1 DI FISHER

INDICE β_1 DI PEARSON

SKEWNESS DI PEARSON (sk)

E' la differenza (d) tra media e moda divisa per la deviazione standard (s)

$$sk = \frac{d}{s}$$

• Proprietà :

- sk può essere nullo, positivo o negativo secondo la forma della distribuzione
- essendo un rapporto, è misura adimensionale, e quindi può essere utilizzato per il confronto tra due o più distribuzioni

INDICE γ_1 DI FISHER

E' il momento standardizzato di terz'ordine

$$\gamma_1 = \frac{m_3}{\sigma^3}$$

INDICE β_1 DI PEARSON

$$\beta_1 = \left(\frac{m_3}{\sigma^3} \right)^2$$

Tra questi due ultimi indici vale la relazione: $\gamma_1 = \sqrt{\beta_1}$

N.B.

Nel caso di distribuzioni simmetriche gli indici sk, γ_1 , β_1 danno un risultato nullo; ma non sempre vale l'inverso, cioè non sempre l'indice di asimmetria uguale a zero esprime la perfetta simmetria di una distribuzione

MOMENTI DI ORDINE K rispetto ad un punto c :

$$m_k = \frac{\sum (x_i - c)^k}{n} \quad \text{per una serie di dati}$$

$$m_k = \frac{\sum (x_i - c)^k \cdot f_i}{n} \quad \text{per una distribuzione di frequenza divisa in classi}$$

c = **origine** (c = 0) --> **momento rispetto all'origine**,

oppure

c = **media** (c = media)--> **momento centrale**

Momento di ordine 1 rispetto all'origine (k=1; c=0) --> media

Momento centrale di ordine 1 (k=1; c=media) --> 0

(è la somma degli scarti dalla media)

Momento centrale di ordine 2 (k=2; c=media) --> varianza

$$m_1 = 0 \quad m_2 = s^2$$

Nello stesso modo si possono calcolare i momenti centrali di ordine terzo (m_3), quarto (m_4), quinto (m_5),...ennesimo (m_n).

I momenti centrali di ordine dispari (m_3, m_5, \dots) sono indici di simmetria :

- sono **nulli** per distribuzioni simmetriche
- sono **non-nulli** per distribuzioni asimmetriche (quanto maggiore è l'asimmetria, tanto più grande è il valore del momento centrale di ordine dispari)
- hanno valore positivo in distribuzioni con asimmetria destra
- hanno valore negativo in distribuzioni con asimmetria sinistra

N.B.

I valori dei momenti dipendono dalla scala utilizzata; per avere una **misura adimensionale**, che permetta i confronti tra più distribuzioni, **bisogna dividerli per la potenza n** (n=3 per il terz'ordine, n=4 per il quart'ordine, ecc.) **dello scarto quadratico medio**

CURTOSI (dal greco $\kappa\upsilon\rho\tau\omicron\sigma$, curvo o convesso)

E' il grado di appiattimento, rispetto alla curva normale (o gaussiana) delle **curve unimodali simmetriche**

MESO- : forma uguale alla distribuzione normale

LEPTO-: eccesso di frequenza delle classi centrali, frequenza minore delle classi intermedie e frequenza maggiore di quelle estreme

PLATI-: numero più ridotto dei valori centrali, frequenza maggiore di quelle intermedie e frequenza minore di quelle centrali ed estreme

INDICI DI CURTOSI

Si basano su rapporti, e sono pertanto **misure adimensionali**

Il rapporto $\frac{\mu_4}{\sigma^4} [= \frac{\mu_4}{\mu_2^2}]$ è una quantità adimensionale :

- distribuzione perfettamente normale $\longrightarrow 3$
- dati più addensati verso il centro (lepto) $\longrightarrow > 3$
- curva schiacciata (plati) $\longrightarrow < 3$

INDICE γ_2 DI FISHER

differenza tra il rapporto fra il momento centrale di quart'ordine e lo scarto quadratico medio (o deviazione standard) elevato alla quarta potenza e la costante 3

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3$$

- distribuzione mesocurtica o normale $\longrightarrow 0$
- distribuzione leptocurtica o ipernormale $\longrightarrow +$
- distribuzione platicurtica o iponormale $\longrightarrow -$

INDICE β_2 DI PEARSON

rapporto fra il momento centrale di quart'ordine e lo scarto quadratico medio (o deviazione standard) elevato alla quarta potenza :

$$\beta_2 = \frac{m_4}{\sigma^4} \quad [\text{Tra questi due indici vale la relazione } \beta_2 = \gamma_2 + 3]$$

N.B. Tutti gli indici presentati si applicano sia alle variabili discrete che alle continue, con l'ovvia approssimazione data dal raggruppamento in classi

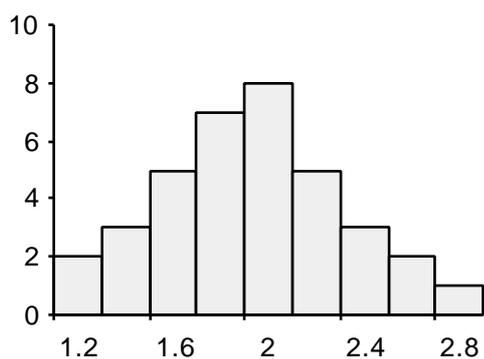
ESERCIZIO

Concentrazioni (mg/l) di sodio e cloruri in 36 laghi appenninici :

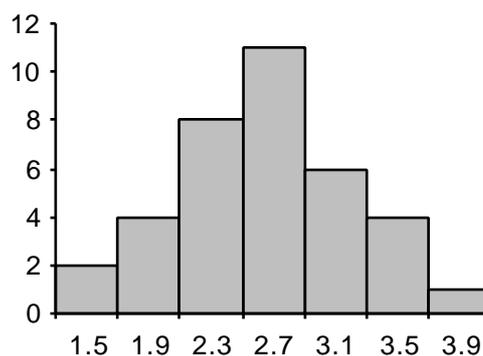
- rappresentare graficamente i dati e la loro distribuzione di frequenza
- calcolare le misure di tendenza centrale, di dispersione e gli indici di forma

Lago	Sodio	Cloruri
1	1,78	1,60
2	1,63	1,80
3	1,85	2,90
4	2,10	2,90
5	1,35	2,90
6	1,40	2,90
7	1,82	2,00
8	1,35	2,00
9	2,06	2,00
10	1,85	2,20
11	1,51	2,30
12	2,00	2,30
13	2,02	2,80
14	1,90	2,80
15	1,60	2,80
16	2,18	2,50
17	1,82	2,50
18	1,90	2,50
19	1,75	2,60
20	2,11	2,60
21	2,30	2,60
22	1,95	2,70
23	2,60	2,90
24	2,44	2,90
25	2,18	3,00
26	2,51	3,10
27	2,37	3,10
28	2,54	3,30
29	2,06	3,30
30	2,77	3,40
31	2,31	3,40
32	2,81	3,60
33	2,33	3,70
34	1,45	3,80
35	1,78	3,80
36	2,09	3,90

	Sodio	Cloruri
Numero di dati (Count, N. of data)	36	36
Somma (Sum)	72,87	101,4
Minimo (Minimum)	1,37	1,6
Massimo (Maximum)	2,81	3,9
Intervallo (Range)	1,46	2,3
Media aritmetica (Mean)	2,024	2,817
Media geometrica (Geometric mean)	1,987	2,756
Media armonica (Harmonic mean)	1,949	2,692
Devianza (Sum of squares)	152,785	297,38
Varianza (Variance, Mean square)	0,151	0,336
Deviazione standard (Standard deviation)	0,389	0,58
Errore standard (Standard error)	0,065	0,097
Curtosi (Kurtosis)	-0,655	-0,53
Asimmetria (Skewness)	0,084	-0,015



Concentrazioni sodio



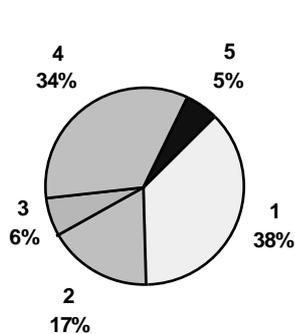
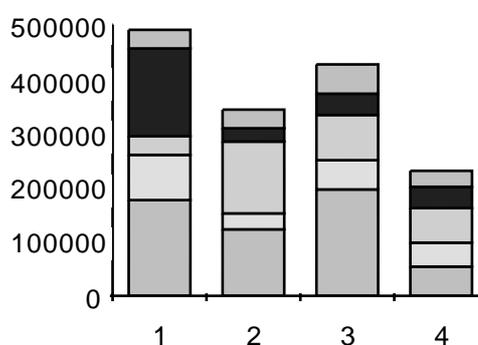
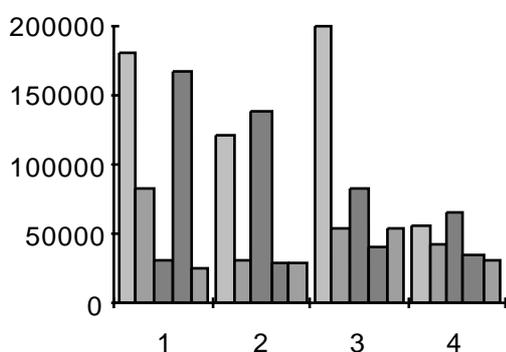
Concentrazioni cloruri

ESERCIZIO

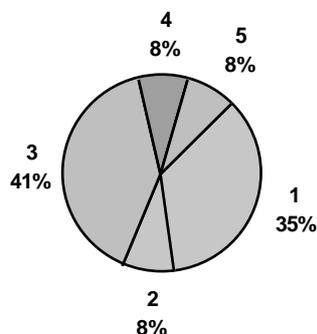
Densità dei principali taxa fitoplanctonici in 4 laghi appenninici

- rappresentare i dati in tabella nelle forme grafiche di uso più comune

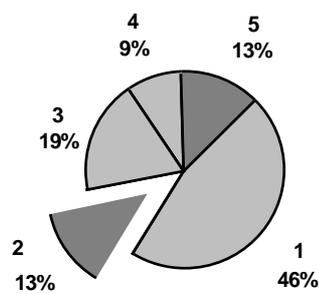
	Clorophyceae	Cryptophyceae	Crysophyceae	Diatomophyceae	Dinophyceae
Lago 1	179.857	83.497	30.891	166.861	25.600
Lago 2	120.893	29.000	136.791	27.500	28.000
Lago 3	198.043	54.454	82.770	38.712	54.734
Lago 4	57.496	42.980	66.440	34.356	31.270



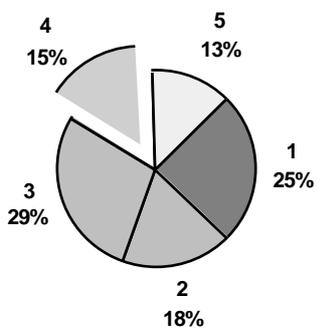
Lago 1



Lago 2



Lago 3



Lago 4

ESERCIZIO

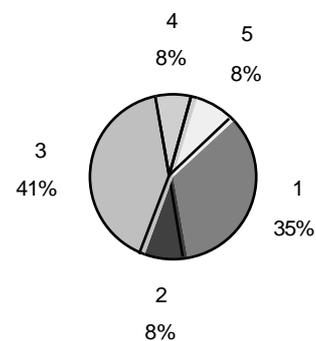
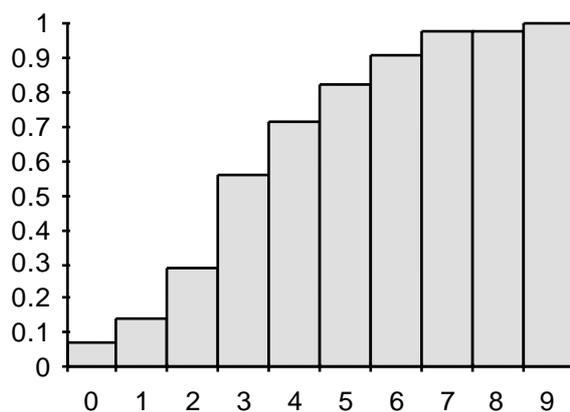
Delle due serie di dati:

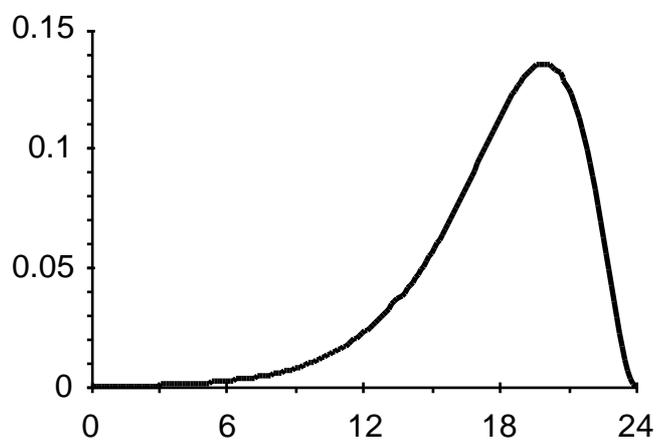
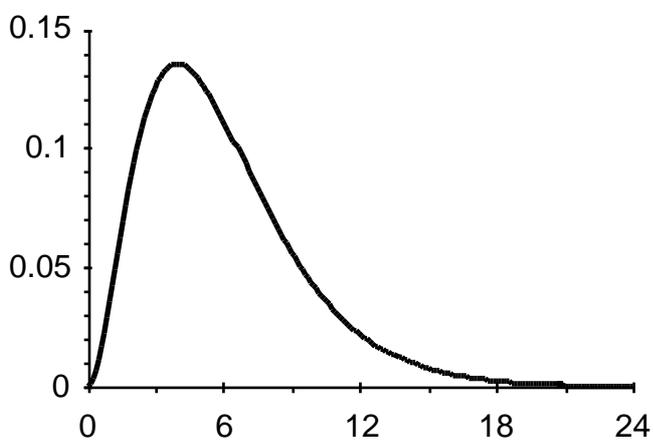
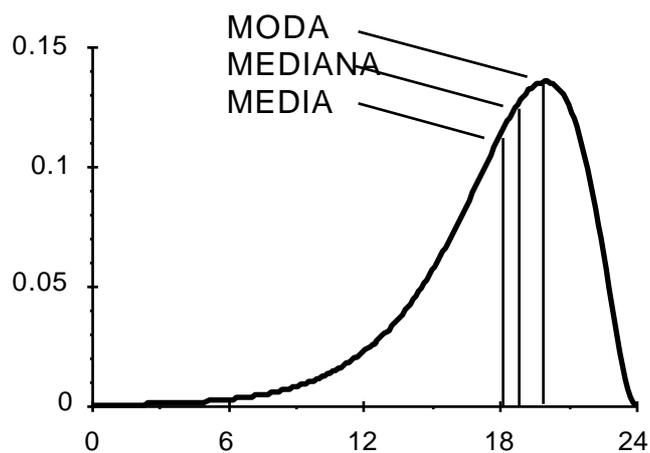
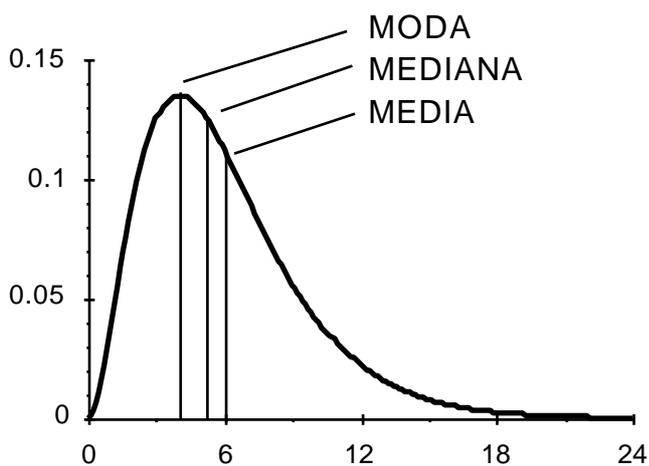
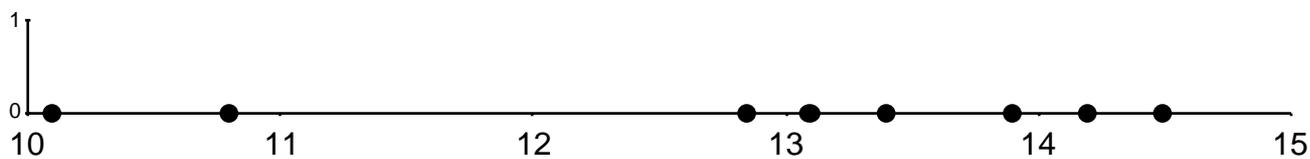
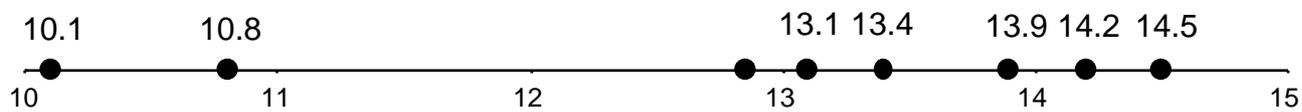
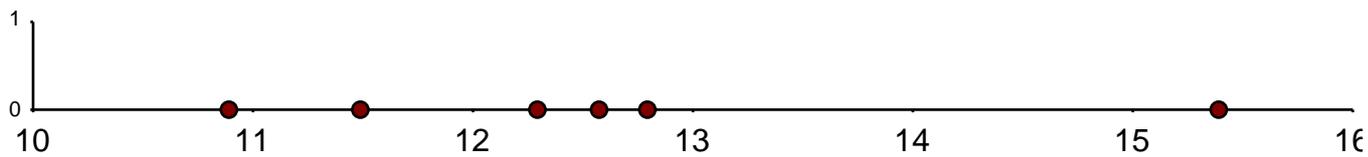
A: 5 7 2 4 3

B: 15 11 9 8 10 12

calcolare le misure di tendenza centrale, di dispersione e gli indici di forma

	A	B
Numero di dati (Count, No. of data)	5	6
Somma (Sum, Summation)	21	65
Minimo (Minimum)	2	8
Massimo (Maximum)	7	15
Intervallo (Range)	5	7
Media (Mean, Average)	4,2	10,833
Media geometrica (Geometric mean)	3,845	10,60
Media armonica (Harmonic mean)	3,506	10,398
Devianza (Sum of squares, SS)	103	735
Varianza (Variance, Mean square)	3,7	6,167
Deviazione standard (Stn. dev.)	1,924	2,483
Errore standard (Standard error)	0,86	1,014
Curtosi (Kurtosis)	-1,005	-0,605
Asimmetria (Skewness)	0,084	0,636





CONCETTO DI PROBABILITÀ'

Il **risultato** (o **esito**) di ogni singolo evento, in una sequenza fondata su processi

- casuali
- mutualmente esclusivi
- equiprobabili
- indipendenti

è **imprevedibile**

Se il numero di eventi (o osservazioni) è elevato, si stabiliscono delle “regolarità”, che renderanno l’esito prevedibile e calcolabile con precisione crescente all'aumentare delle osservazioni

PROBABILITÀ MATEMATICA (o A PRIORI O CLASSICA):

- peculiarità :

- non si richiede alcun dato sperimentale
- i risultati sono conosciuti a priori
- è basata sul solo ragionamento logico formalizzato nel

- **principio di Laplace :**

la probabilità di un evento è il rapporto tra il numero di casi favorevoli ed il numero di casi possibili, purchè tutti i casi siano ugualmente probabili

- esempi :

- lancio di una moneta
- lancio di un dado
- lotterie (la probabilità di fare ambo è superiore a quella di fare terno)
- ordini d'arrivo in una gara dove nessuno ha i favori del pronostico

- comporta limitazioni per la ricerca sperimentale poichè questa è basata su un approccio non teorico ma empirico :

- per valutare una probabilità sarebbe necessario conoscere preventivamente le diverse probabilità dei vari eventi
- non sarebbe possibile rispondere a quesiti che per loro natura richiedano osservazioni ripetute

PROBABILITÀ FREQUENTISTA

(o A POSTERIORI, o LEGGE EMPIRICA DEL CASO, o STATISTICA) :

- se in un insieme di prove la frequenza di un evento è all'incirca costante, questo valore di frequenza è assunto come probabilità
- si basa sul **principio di von Mises** (formulato nel 1920) :
la probabilità di un evento, in una serie di prove condotte nelle stesse condizioni, è il limite a cui essa tende al crescere del numero delle osservazioni
- si applica in tutti quei casi in cui **non sono note a priori** le leggi dei fenomeni studiati, ma possono essere determinate a posteriori; ovvero...

...per calcolare la probabilità attesa di trovare un numero stabilito di individui in un conteggio, deve essere nota la percentuale di presenza rilevata attraverso una precedente serie di osservazioni. Infatti, l'unico modo per rispondere ai quesiti empirici è condurre una serie di osservazioni od esperimenti, in condizioni controllate statisticamente, per rilevare la frequenza relativa del fenomeno

PROBABILITÀ SOGGETTIVISTICA (o "BAYESIANA")

Le probabilità classica e frequentista richiedono che gli eventi ripetuti si verificano in condizioni uniformi o presunte tali. Ma nella teoria della probabilità sono inclusi anche fenomeni che non possono essere ricondotti a queste condizioni, perchè sono considerati eventi unici od irripetibili

Ad esempio, determinare la probabilità che ...

- avvenga una catastrofe
- entro la fine dell'anno scoppi la terza guerra mondiale
- una specie animale o vegetale si estingua

... presuppone il giudizio di più individui o stime personali, e introduce un terzo tipo di probabilità: la **probabilità soggettiva** (o **bayesiana**)

- si fonda sul principio che la probabilità è una stima del grado di aspettativa di un evento, secondo l'esperienza personale di un individuo
- è una misura della convinzione circa l'esito o l'accadimento di un evento
- ha vaste ed interessanti applicazioni nelle scienze sociali ed economiche, dove l'attesa di un fenomeno o una convinzione possono influire sui fenomeni reali (svalutazione, prezzi di mercato, comportamenti sociali)
- aspetti controversi :
 - come misurare un grado di aspettativa, dato che sperimentatori diversi attribuiscono probabilità differenti allo stesso fenomeno ?
 - come modificare la probabilità soggettiva di partenza in dipendenza dei successivi avvenimenti oggettivi, in assenza di replicazioni ?
 - se il mondo esterno è realtà oggettiva indipendente, la conoscenza non può derivare da convinzioni personali o da preferenze individuali: l'approccio soggettivo non risulta attendibile, in quanto non permette la conoscenza oggettiva del reale

Nel contesto delle scienze sperimentali predominano i casi di eventi ripetibili, in condizioni almeno approssimativamente uguali o simili, ertanto di norma si fa ricorso all'impostazione frequentista, trascurando quella soggettivistica più utile in altre discipline

LEGGI DI PROBABILITÀ

CALCOLO COMBINATORIO DI AGGRUPPAMENTI SEMPLICI

- è strumento fondamentale nella statistica
- sebbene il risultato di ogni singolo tentativo sia imprevedibile, con un numero elevato di ripetizioni si stabiliscono regolarità che possono essere calcolate e, dunque, previste
- serve per collegare una scelta alla probabilità di attesa dell'evento desiderato, nel contesto di tutti gli eventi possibili
- il risultato è sempre un valore compreso tra 0 e 1

ESEMPIO

Gara di corsa tra 10 concorrenti

- quanti differenti ordini d'arrivo sono possibili ?
- quale è la probabilità di indovinare i primi tre :
 - nell'ordine ?
 - senza stabilire il loro ordine ?
- conviene scommettere 10.000 lire per guadagnarne 500.000 se si indovineranno i primi 2 :
 - nell'ordine ?
 - senza stabilire il loro ordine ?

Requisiti fondamentali degli eventi:

- si escludono a vicenda
- sono tutti ugualmente possibili
- vengono generati da eventi puramente casuali
- avvengono in modo indipendente

Gli aggruppamenti si distinguono in :

- **PERMUTAZIONI**
 - **DISPOSIZIONI**
 - **COMBINAZIONI**

PERMUTAZIONI SEMPLICI

I sottoinsiemi che si possono formare collocando n elementi differenti

$a_1 \quad a_2 \quad a_3 \quad \dots \quad a_n$ **in tutti gli ordini possibili**

Il numero di permutazioni di n elementi è : $P_n = n!$

dove : $n!$ (n fattoriale) = $1 \cdot 2 \cdot 3 \cdot \dots \cdot n$ (*)

ESEMPIO. Le permutazioni degli elementi a b c sono : [abc acb bca bac cba cab]

ESEMPIO. Le permutazioni degli elementi a b c d sono : $4! = 1 \cdot 2 \cdot 3 \cdot 4 = 24$
[abcd abdc acbd adcb cabd cdba dbac cbda]

(*) I primi 25 numeri fattoriali

1! =	1
2! =	2
3! =	6
4! =	24
5! =	120
6! =	720
7! =	5.040
8! =	40.320
9! =	362.880
10! =	3.628.800
11! =	39.916.800
12! =	479.001.600
13! =	6.227.020.800
14! =	87.178.291.200
15! =	1.307.674.368.000
16! =	20.922.789.888.000
17! =	355.687.428.096.000
18! =	6.402.373.705.728.000
19! =	121.645.100.408.832.000
20! =	2.432.902.008.176.640.000
21! =	51.090.942.171.709.440.000
22! =	1.124.000.727.777.607.680.000
23! =	25.852.016.738.884.976.640.000
24! =	620.448.401.733.239.439.360.000
25! =	15.511.210.043.330.985.984.000.000

Nel calcolo fattoriale, per definizione : $0! = 1$ e $1! = 1$

DISPOSIZIONI SEMPLICI

I sottoinsiemi di p elementi, tratti da un insieme di n oggetti differenti

$$a_1 \quad a_2 \quad a_3 \quad a_p \quad \dots \quad a_n$$

che si diversificano **per almeno un elemento o per il loro ordine**

Il numero di disposizioni semplici di n elementi presi p a p è :

$$D_n^p = \frac{n!}{(n-p)!}$$

ESEMPIO. Le disposizioni di 4 elementi **a b c d** presi 3 a 3 sono :

abc abd acd acb adb adc bac bad bcd bca bda bdc
cab cad cbd cba cda cdb dab dac dbc dba dca dcb

cioè :

$$D_4^3 = \frac{4!}{(4-3)!} = \frac{24}{1} = 24$$

Un metodo alternativo per calcolare le disposizioni semplici di n elementi presi p a p :

$$D_n^p = n(n-1)(n-2)\dots(n-p+1)$$

Questo metodo è più pratico e più rapido quando n e p sono quantità elevate.

Infatti, le disposizioni di 4 elementi presi 3 a 3 si possono calcolare come :

$$D_4^3 = 4(4-1)(4-2) = 4 \cdot 3 \cdot 2 = 24$$

ESEMPIO. Le disposizioni di 7 elementi presi 3 a 3 sono :

$$D_7^3 = 7(7-1)(7-2) = 7 \cdot 6 \cdot 5 = 210$$

COMBINAZIONI SEMPLICI

I sottoinsiemi di p elementi, tratti da un insieme di n oggetti differenti

$$a_1 \quad a_2 \quad a_3 \quad a_p \quad \dots \quad a_n$$

che si diversificano per **almeno un elemento, ma non per il loro ordine**

Il numero di combinazioni semplici di n elementi presi p a p è :

$$C_n^p = \frac{n!}{(n-p)! p!}$$

Corrisponde al numero di disposizioni di n elementi presi p a p , diviso il numero di permutazioni di p elementi

Il numero di combinazioni risulta sempre un numero intero indicato con $n \setminus p$
ed è chiamato **COEFFICIENTE BINOMIALE** (si legge: “n su p”)

La sequenza dei coefficienti binomiali è data dai coefficienti del **Triangolo di Tartaglia** ($n \setminus p$; $p \rightarrow$)

Ad es., le combinazioni di **a b c d** presi 3 a 3 sono abc abd acd bcd, cioè :

$$C_4^3 = \frac{4!}{(4-3)!3!} = 4$$

N.B.

Numero di combinazioni di n elementi :

- presi ad n ad n : $C_n^n = \frac{n!}{n!0!} = 1$ (un solo sottoinsieme formato da tutti gli elementi)

- presi ad 1 ad 1 : $C_n^1 = \frac{n!}{1!(n-1)!} = n$ (n è il numero di sottoinsiemi con un solo elemento)

- presi 0 a 0 : $C_n^0 = \frac{n!}{0!n!} = 1$ (c'è un solo sottoinsieme vuoto)

ESEMPIO

In un esperimento sulla fertilità di un terreno, si vogliono esaminare in modo sistematico gli equilibri binari tra : Ca, Mg, Na, N, P, K

- Quante coppie di elementi occorrerà prendere in considerazione ?
- Per valutare tutti gli equilibri ternari, quanti gruppi diversi si dovranno formare ?

$$(\text{ Risposta: } C_6^2 = \frac{6!}{(6-2)! 2!} = \frac{5 \cdot 6}{2} = 15)$$

$$(\text{ Risposta: } C_6^3 = \frac{6!}{(6-3)! 3!} = 20)$$

ESEMPIO

Risposte ai cinque quesiti introduttivi

1 - In una corsa con 10 concorrenti, i possibili ordini d'arrivo sono le **permutazioni** di 10 elementi : $P_{10} = 10! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 = 3.628.800$

2 - I possibili gruppi dei primi 3 concorrenti tra 10, tenendo conto dell'ordine d'arrivo, sono le **disposizioni** di 10 elementi presi 3 a 3 : $D_{10}^3 = \frac{10!}{(10-3)!} = 720$

Probabilità di indovinare : $1/720 = 0,001389$

3 - I possibili gruppi dei primi 3 concorrenti tra 10, senza distinzioni dell'ordine di arrivo, sono le **combinazioni** di 10 elementi presi 3 a 3 : $C_{10}^3 = \frac{10!}{(10-3)! 3!} = 120$

Probabilità di indovinare : $1/120 = 0,00833$ cioè 6(3!) volte più alta di quella in cui si vuole prevedere anche l'ordine

4 - La probabilità di indovinare i primi 2 tra 10, stabilendo chi sarà primo e chi secondo, è data dalle **disposizioni** di 10 elementi presi 2 a 2 : $D_{10}^2 = \frac{10!}{(10-2)!} = 90$

Probabilità di indovinare : $1/90$, meno favorevole del rapporto $1/50$ fissato nella scommessa (non conviene scommettere)

5 - La probabilità di indovinare i primi 2 tra 10, senza stabilire l'ordine, è data dalle **combinazioni** di 10 elementi presi 2 a 2 : $C_{10}^2 = \frac{10!}{(10-2)! 2!} = 45$

Probabilità di indovinare : $1/45$, più favorevole del rapporto $1/50$ fissato nella scommessa (conviene scommettere)

DISTRIBUZIONI DISCRETE

BINOMIALE tende alla gaussiana, per $n \rightarrow \infty$

MULTINOMIALE

POISSONIANA tende alla gaussiana per medie elevate

IPERGEOMETRICA

BINOMIALE NEGATIVA

UNIFORME

DISTRIBUZIONI CONTINUE

NORMALE (o GAUSSIANA)

PROPRIETÀ E USO DELLA NORMALE

NORMALE (o GAUSSIANA) STANDARDIZZATA

UTILIZZO DELLA NORMALE STANDARDIZZATA

CORREZIONI PER LA CONTINUITA' IN PROBABILITA' DISCRETE

RETTANGOLARE (o uniforme continua)

ESPOENZIALE NEGATIVA

Tendono alla "normale:

- la binomiale, per $n \rightarrow \infty$;

- la poissoniana, per $\bar{x} \gg 0$

DISTRIBUZIONI DISCRETE

DISTRIBUZIONE BINOMIALE (o di BERNOULLI)

- distribuzione **teorica discreta e finita**
- fornisce le probabilità che un evento, con probabilità (a priori o a posteriori) **p**, avvenga 0, 1, 2, ... **r**, ... **n** volte, nel corso di **n** prove identiche ed indipendenti ...
... che possono essere ripartite solo in due classi A e B

- con frequenze assolute **n_a** e **n_b**

- con frequenze relative

$$p = \frac{n_a}{n} \quad q = \frac{n_b}{n} \quad \text{tali che} \quad p + q = 1$$

- la probabilità di ottenere **r** volte l'evento A (**n-r** volte l'evento B) è :

$$P_r = C_n^r p^r q^{n-r}$$

$$\text{dove : } C_n^r = \frac{n!}{r! (n-r)!}$$

N.B. Le prove possono essere successive oppure simultanee, purchè non si influenzino reciprocamente

ESEMPIO

Nella specie umana nascono più maschi che femmine, con un rapporto di 105 maschi per 100 femmine

A posteriori, sulla base dei dati rilevati, si può affermare che la **probabilità frequentista** di un nato maschio è **p=0,52** e di un nato femmina è di **q=0,48 (=1-p)**

La distribuzione binomiale calcola le specifiche probabilità di 0, 1, 2, 3, 4 nascite di figli maschi nelle famiglie con 4 figli :

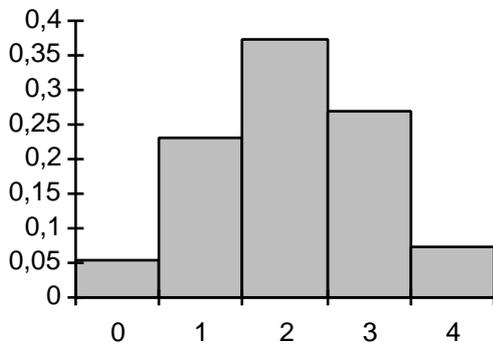
$$P_0 = C_4^0 p^0 q^4 = 1 \cdot 1 \cdot (0,48)^4 = 0,05$$

$$P_1 = C_4^1 p^1 q^3 = 4 \cdot (0,52) \cdot (0,48)^3 = 0,23$$

$$P_2 = C_4^2 p^2 q^2 = 6 \cdot (0,52)^2 \cdot (0,48)^2 = 0,37$$

$$P_3 = C_4^3 p^3 q^1 = 4 \cdot (0,52)^3 \cdot (0,48) = 0,28$$

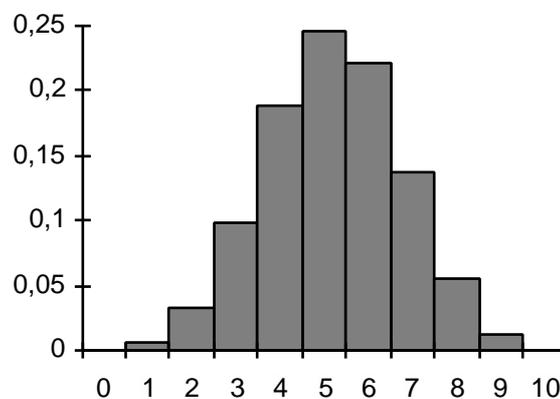
$$P_4 = C_4^4 p^4 q^0 = 1 \cdot (0,52)^4 \cdot 1 = 0,07$$



Probabilità del numero di nascite di maschi in famiglie con 4 figli

Probabilità del numero di nascite di figli maschi in famiglie con 10 figli

x	P
0	0.000649
1	0.007034
2	0.034289
3	0.099056
4	0.187793
5	0.244131
6	0.220396
7	0.136436
8	0.055427
9	0.013344
10	0.001446



La distribuzione binomiale:

- è leggermente asimmetrica, poichè le probabilità $p \neq q$
- tende ad essere simmetrica all'aumentare del numero di osservazioni, anche se $p \neq q$
- si utilizza anche quando le probabilità sono note a priori, come nel caso dei dadi (ovviamente bilanciati e non truccati)

ESEMPI

Probabilità di ottenere 3 volte il numero 1 lanciando un dado 5 volte ($n=5$ $r=3$ $p=1/6$ $q=5/6$):

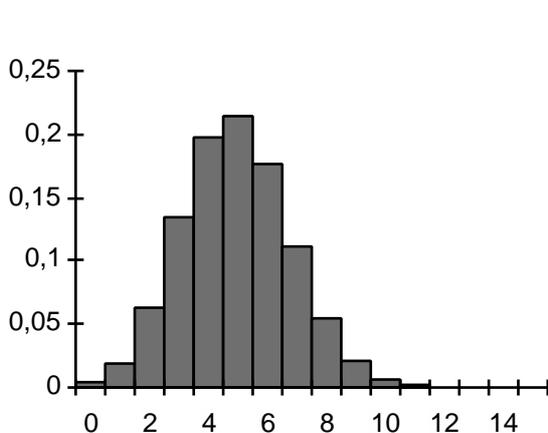
$$P_3 = C_5^3 p^3 q^2 = \frac{5!}{3!2!} \cdot \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^2 = 0,03215$$

Probabilità di estrarre 4 biglie tutte nere da un'urna contenente un elevato numero di biglie per il 70% nere e per il 30% bianche ($n=4$ $r=4$ $p=0,7$ $q=0,3$):

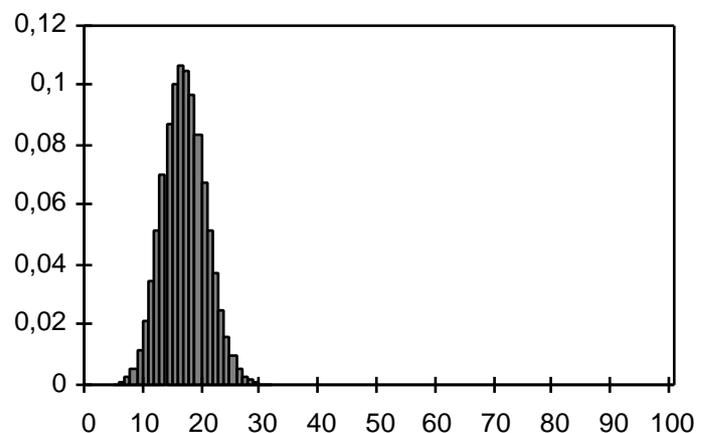
$$P_4 = C_4^4 p^4 q^0 = \frac{4!}{4!0!} \cdot 0,7^4 \cdot 0,3^0 = 0,2401$$

Probabilità che 9 esperimenti di laboratorio risultino positivi e 1 negativo, se di solito gli esperimenti sono positivi nel 20% dei casi ($n=10$ $r=9$ $p=0,2$ $q=0,8$):

$$P_9 = C_{10}^9 p^9 q^1 = \frac{10!}{9!1!} \cdot 0,2^9 \cdot 0,8^1 = 0,000004096$$



Distribuzione binomiale ($n=15$ $p=0,33$)



Distribuzione binomiale ($n=100$ $p=0,167$)

In una distribuzione binomiale :

- quando n è elevato, la forma è praticamente normale e quasi simmetrica anche se p è sensibilmente diverso da 0,5
- le probabilità associate ai diversi tipi di estrazione sono espresse dai termini dello **sviluppo del binomio** $(p + q)^n$ dove p e q sono le probabilità dei due diversi eventi semplici “A” e “B” (“A” e “non-A”), come nel caso dei numeri che possono comparire nel lancio dei dadi
- **la media è $n \cdot p$**
- **varianza è $\sigma^2 = n \cdot p \cdot q$**
- la varianza è inferiore alla media, poichè $q < 1$:
 $q = 1 - p$; $\sigma^2 = n \cdot p \cdot (1 - p)$

DISTRIBUZIONE MULTINOMIALE

- rappresenta una estensione di quella binomiale
- si applica a k eventi indipendenti di probabilità $p_1 p_2 \dots p_i \dots p_k$ ($\sum p_i = 1$) che possono comparire nel corso di N prove indipendenti (successive o simultanee)

ESEMPIO

In un'urna contenente moltissime biglie :

il 10% ($p_1 = 0,10$) sono bianche il 40% ($p_2 = 0,40$) sono rosse
il 20% ($p_3 = 0,20$) sono gialle il 30% ($p_4 = 0,30$) sono verdi

D.:

- su 10 biglie estratte, qual'è la probabilità che 2 siano bianche, 3 rosse, 2 gialle e 3 verdi ?
- su 8 biglie estratte, qual'è la probabilità di che 4 siano rosse e 4 verdi ?

R.:

Le probabilità sono determinate dallo **sviluppo del multinomio** :

$$P_{(n_1 \ n_2 \ \dots \ n_k)} = \frac{N!}{n_1! n_2! \ \dots \ n_k!} p_1^{n_1} p_2^{n_2} \ \dots \ p_k^{n_k}$$

$$P_{(2b, 3r, 2g, 3v)} = \frac{10!}{2!3!2!3!} (0,10)^2 (0,40)^3 (0,20)^2 (0,30)^3 = 0,011612$$

$$P_{(4r, 4v)} = \frac{8!}{0!4!0!4!} (0,10)^0 (0,40)^4 (0,20)^0 (0,30)^4 = 0,04587$$

DISTRIBUZIONE POISSONIANA

- è una distribuzione teorica discreta ed infinita, totalmente determinata da un solo parametro, la media μ
- è la distribuzione limite della binomiale per $p \rightarrow 0$

Se: $n \rightarrow \infty$ e $p \rightarrow 0$, in modo tale che $n \cdot p$ sia costante, Poisson nel 1837

$$\text{dimostrò che : } P_i = \frac{\mu^i}{i!} e^{-\mu} = \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} \binom{n}{i} p^n q^{n-i}$$

- la media attesa μ è uguale a c
- la varianza attesa σ^2 è uguale a μ

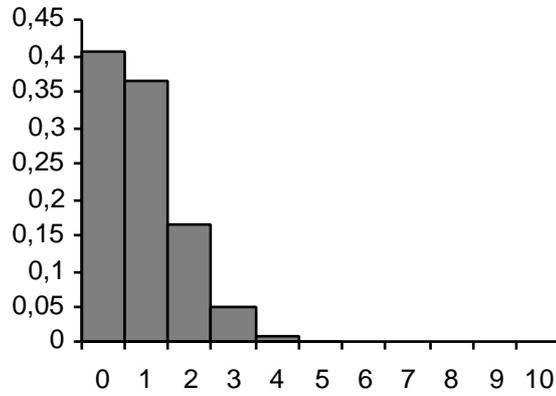
nella binomiale $\sigma_B^2 = npq$; applicando le condizioni su enunciate :

$$\sigma^2 = \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} npq = \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} (np)q = \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0}} c(1-p) = c = \mu$$

- è detta **legge degli eventi rari**, essendo assai più frequenti le classi con zero o pochi eventi rispetto alle classi con numerosi eventi
- è detta **legge dei piccoli numeri**, essendo la frequenza assoluta degli eventi espressa da un numero piccolo, anche con molte prove
- è molto asimmetrica per valori piccoli di μ (< 3)
- è quasi simmetrica già per $\mu \approx 7$ (si diversifica poco dalla gaussiana)

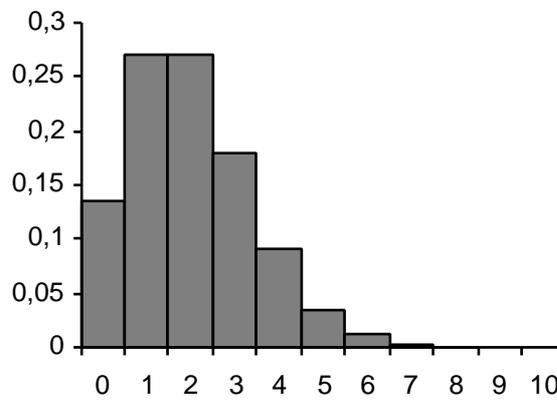
Distribuzione di Poisson, $\mu=0.9$

i	P
0	0.40657
1	0.365913
2	0.164661
3	0.049398
4	0.011115
5	0.002001
6	0.000300
7	0.000039
8	0.000004
9	0.000000
10	0.000000



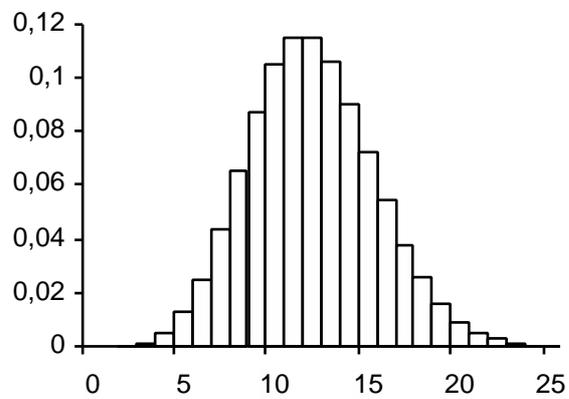
Distribuzione di Poisson, $\mu=2$

i	P
0	0.135335
1	0.270671
2	0.270671
3	0.180447
4	0.090224
5	0.036089
6	0.012030
7	0.003437
8	0.000859
9	0.000191
10	0.000038



Distribuzione di Poisson, $\mu = 12$

i	P
0	0.000006
1	0.000074
2	0.000442
3	0.00177
4	0.005309
5	0.012741
6	0.025481
7	0.043682
8	0.065523
9	0.087364
10	0.104837
11	0.114368
12	0.114368
13	0.10557
14	0.090489
15	0.072391
16	0.054293
17	0.038325
18	0.02555
19	0.016137
20	0.009682
21	0.005533
22	0.003018
23	0.001574
24	0.000787
25	0.000378



ESEMPIO

In letteratura, è famoso l'esempio di Bortkewitch, un veterinario dell'armata prussiana del XIX secolo che per 20 anni contò il numero di soldati di 10 corpi d'armata che ogni anno morivano a causa di un calcio di mulo

x : numero di decessi i	0	1	2	3	4
f : eventi osservati r	109	65	22	3	1

$$\text{media} = \frac{122}{200} = 0,6100 \quad \text{varianza} = 0,6079$$

Applicando la distribuzione di Poisson, si determinano le probabilità teoriche di osservare 0 1 2 3 4 decessi ogni anno

$$P_0 = \frac{0,61^0}{0!} \cdot \frac{1}{2,71^{0,61}} = \frac{1}{1} \cdot \frac{1}{1,837} = 0,5440$$

$$P_1 = \frac{0,61^1}{1!} \cdot \frac{1}{2,71^{0,61}} = \frac{0,61}{1} \cdot \frac{1}{1,837} = 0,3318$$

$$P_2 = \frac{0,61^2}{2!} \cdot \frac{1}{2,71^{0,61}} = \frac{0,3721}{2} \cdot \frac{1}{1,837} = 0,1010$$

$$P_3 = \frac{0,61^3}{3!} \cdot \frac{1}{2,71^{0,61}} = \frac{0,2270}{6} \cdot \frac{1}{1,837} = 0,0203$$

$$P_4 = \frac{0,61^4}{4!} \cdot \frac{1}{2,71^{0,61}} = \frac{0,1385}{6} \cdot \frac{1}{1,837} = 0,0029$$

numero di decessi	0	1	2	3	4
eventi <u>osservati</u>	109	65	22	3	1
frequenze relative attese	0,5440	0,3318	0,1010	0,0203	0,0029
eventi <u>attesi</u> (su 200)	108,80	66,36	20,20	4,06	0,58

Con $\mu=0,61$ la probabilità teorica di 0 morti è 0,544 (rapportata ai 200 eventi comporta una frequenza attesa di 108,8)

Si noti che lo scarto tra osservato ed atteso è molto piccolo

ESEMPIO

In una comunità planctonica la popolazione di *Eudiaptomus vulgaris* è presente col 2% degli individui

D.:

- campionando 200 individui quale è la probabilità di **non** trovare *Eudiaptomus* ?
- campionando 100 individui quale è la probabilità di trovarlo 4 volte ?
- con una presenza del 5%, come cambierebbero le probabilità precedenti ?

R. :

Campionando **200** individui:

- media della popolazione (presenza 2%) $m = n \cdot p = 200 \cdot 0,02 = 4$
- probabilità di non trovare individui (2%) $P_0 = \frac{4^0}{0!} e^{-4} = 0,0183$
- media della popolazione (presenza 5%) $m = n \cdot p = 200 \cdot 0,05 = 10$
- probabilità di non trovare individui (5%) $P_0 = \frac{10^0}{0!} e^{-10} = 0,0000454$

Campionando **100** individui:

- media della popolazione (presenza 2%) $m = n \cdot p = 100 \cdot 0,02 = 2$
- probabilità di trovare 4 individui (2%) $P_4 = \frac{2^4}{4!} e^{-2} = 0,0902$
- media della popolazione (presenza 5%) $m = n \cdot p = 100 \cdot 0,05 = 5$
- probabilità di trovare 4 individui (5%) $P_4 = \frac{5^4}{4!} e^{-5} = 0,1755$

DISTRIBUZIONE IPERGEOMETRICA

Quando nell'urna ci sono **moltissime** biglie ...

ogni estrazione non altera le probabilità di quelle successive, il che equivale a supporre che ogni biglia estratta sia reintrodotta (o che il numero di biglie sia praticamente infinito),

ma quando nell'urna ci sono **poche** biglie ...

senza reintroduzione, le probabilità di estrarre biglie di un dato colore non sono costanti, ma dipendono dagli eventi precedenti

... tali probabilità possono essere calcolate con la distribuzione ipergeometrica

ESEMPIO

Da un'urna con N biglie, delle quali n_1 bianche e $N-n_1$ nere, si estraggono n biglie ($n \leq N$) senza reintroduzione

Determinare la probabilità $P_{(r/n)}$ che delle n biglie estratte r siano bianche ($r \leq n$)

La distribuzione delle probabilità di tutti gli eventi possibili può essere determinata col calcolo combinatorio :

1. delle N biglie, n possono essere estratte in $\binom{N}{n}$ modi differenti
2. delle n_1 biglie bianche, r possono essere estratte in $\binom{n_1}{r}$ modi differenti
3. delle $N-n_1$ biglie nere, $n-r$ possono essere estratte in $\binom{N-n_1}{n-r}$ modi differenti
4. ognuna delle $\binom{n_1}{r}$ diverse possibilità di estrarre biglie bianche si combina con ognuna delle $\binom{N-n_1}{n-r}$ possibilità di estrarre biglie nere

Ne consegue che

$$P_{(r/n)} = \frac{C_n^r \cdot C_{N-n}^{n_1-r}}{C_N^{n_1}}$$

N intero positivo

n intero non negativo al massimo uguale a N

n_1 intero positivo al massimo uguale a N

ESEMPIO

In una piccola riserva naturale sono presenti 9 cinghiali: 3 femmine e 6 maschi; per ridurre il loro numero viene decisa una battuta di caccia, nella quale ne verranno catturati 5 senza attenzione al sesso

D.:

Stimare i possibili effetti secondo le probabilità :

- che vengano catturate tutte le 3 femmine
- che vengano catturate 2 femmine
- che venga catturata 1 femmina
- che non venga catturata alcuna femmina

animali presenti	$N = 9$				
animali catturati	$n = 5$				
femmine presenti	$n_1 = 3$				
femmine catturate	$r = 3$	$r = 2$	$r = 1$	$r = 0$	
animali non catturati	$N - n$				
femmine non catturate	$n_1 - r$				

R.:

$$\begin{aligned} a) P_{(3/5)} &= \frac{C_5^3 \cdot C_{9-5}^{3-3}}{C_9^5} = \frac{3!2! \cdot 0!4!}{9!} = 0,119 \text{ (11,9\%)} \\ b) P_{(2/5)} &= \frac{C_5^2 \cdot C_{9-5}^{3-2}}{C_9^5} = \frac{2!3! \cdot 1!3!}{9!} = 0,4762 \text{ (47,62\%)} \\ c) P_{(1/5)} &= \frac{C_5^1 \cdot C_{9-5}^{3-1}}{C_9^5} = \frac{1!4! \cdot 2!2!}{9!} = 0,3572 \text{ (35,72\%)} \\ d) P_{(0/5)} &= \frac{C_5^0 \cdot C_{9-5}^{3-0}}{C_9^5} = \frac{0!5! \cdot 3!1!}{9!} = 0,0476 \text{ (4,76\%)} \end{aligned}$$

Probabilità di catturare	3 femmine	11,9%
“	2 femmine	47,62%
“	1 femmina	35,72%
“	0 femmine	4,76%

DISTRIBUZIONE BINOMIALE NEGATIVA

La distribuzione binomiale positiva :

- $p + q = 1$
- con **n** prove, le probabilità dei diversi eventi sono determinate dallo sviluppo del binomio $(p + q)^n$
- presenta varianza **npq** inferiore alla media **np**, essendo **q** < 1

La distribuzione binomiale negativa :

- è impiegata soprattutto nei conteggi di popolazioni animali (foglie con 0, 1, 2, ... parassiti) e negli studi epidemiologici (periodi -giorni, settimane o mesi- con 0, 1, 2, ... morti)
- può essere intesa come un mix di distribuzioni poissoniane
- ha varianza **npq** superiore alla media **np**

Nei fenomeni semplici

- a media unica
- **n** grande
- **p** basso

le frequenze attese sono fornite dalla poissoniana

Nei fenomeni complessi

- la distribuzione è determinata da più fattori ognuno con media diversa
- la variabilità aumenta sicchè la varianza è superiore alla media
- la distribuzione delle frequenze può essere stimata in modo appropriato dalla distribuzione binomiale negativa

Se un fenomeno presenta una distribuzione binomiale negativa, la probabilità P_i che l'evento atteso si verifichi i volte ($0, 1, 2, \dots, k$) è :

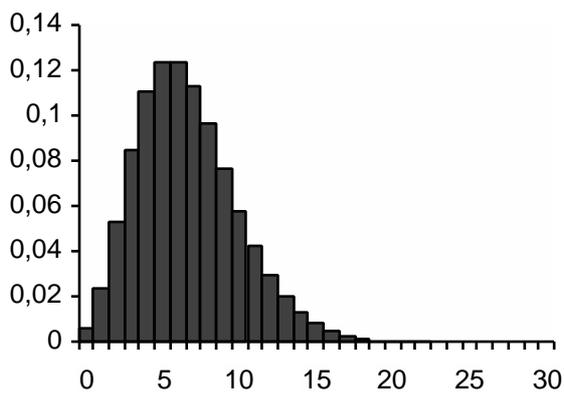
$$P_i = \frac{(k+i-1)! \left(\frac{p}{q}\right)^k}{i!(k-1)!q^k}$$

Parametri essenziali :

- media np
- esponente k ($-n$)

dove :

$$k = \frac{(n \cdot p)^2}{n \cdot p \cdot q - n \cdot p}$$



Distribuzione binomiale negativa ($\mu = 6.66, p = 0.6$)

DISTRIBUZIONE UNIFORME

- è la più semplice distribuzione discreta
- identica possibilità del verificarsi di tutti i possibili risultati (ad es., la probabilità di ottenere 1 ... 6 con un dado non truccato è uguale per ognuno dei risultati)
- l'impiego è limitato quasi esclusivamente all'analisi di probabilità a priori
- la probabilità del singolo evento in una variabile discreta X che segue questa distribuzione è :

$$P(x) = \frac{1}{(b - a) + 1}$$

b = risultato maggiore possibile di X

a = risultato minore possibile di X

Per i dadi (**b**=6 e **a**=1) è semplice verificare che $P(x) = \frac{1}{(6-1)+1} = \frac{1}{6}$

• media $\mu = \frac{a + b}{2}$

• varianza $\sigma = \sqrt{\frac{[(b - a) + 1]^2 - 1}{12}}$

DISTRIBUZIONI CONTINUE

DISTRIBUZIONE NORMALE o DISTRIBUZIONE DI GAUSS

- è la più importante distribuzione continua

- proposta da Gauss (1809) nell'ambito della teoria degli errori, è stata attribuita anche a Laplace (1812), che ne definì le proprietà principali in anticipo rispetto alla trattazione più completa di Gauss

- il nome deriva dalla convinzione che i fenomeni fisico-biologici solitamente si distribuiscono con frequenze più elevate nei valori centrali e frequenze progressivamente minori verso gli estremi

- è detta anche **CURVA DEGLI ERRORI ACCIDENTALI**, in quanto, soprattutto nelle discipline fisiche, la distribuzione degli errori commessi nel misurare ripetutamente la stessa grandezza, è molto bene approssimata da questa curva

- è considerata il limite della distribuzione binomiale per $n \rightarrow \infty$ (mentre né p né q tendono a 0 come per la poissoniana)

- la variabile considerata, quantificata per unità discrete con pochi dati, può essere espressa, in classi d'ampiezza sempre minore, come **grandezza continua**

- secondo il teorema di De Moivre (1833), quando $n \rightarrow \infty$ (a condizione che né p né q tendano a 0), la probabilità P_i della binomiale è approssimata da :

$$P(i) = \frac{1}{\sqrt{2\pi \cdot n \cdot p \cdot q}} e^{-\frac{(i-n \cdot p)^2}{2 \cdot n \cdot p \cdot q}}$$

Sostituendo

- **np** con la media sperimentale μ
- **npq** con la varianza calcolata σ^2
- il conteggio **i** con la misura **x**

si ottiene :

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

espressione della funzione di densità per le frequenze **f(x)** della normale

Principali **proprietà** della distribuzione :

- ha media μ e varianza σ^2 il cui variare comporta infinite curve normali
- **è indicata con N(μ, σ)**
- è simmetrica rispetto alla media
- ha media, moda e mediana coincidenti
- cresce da $-\infty$ a μ e decresce da μ a $+\infty$
- ha andamento asintotico rispetto all'asse x

DISTRIBUZIONE NORMALE STANDARDIZZATA

- consente di ricondurre alla stessa forma le infinite forme della distribuzione normale determinate dalle diverse medie e varianze
- è ottenuta mediante cambiamento di variabile $x \rightarrow X$

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

che consiste nel :

- rendere $\mu = 0$ sottraendo ad ogni valore la media
- prendere σ come unità della nuova variabile **X** e quindi costruire una distribuzione con $\sigma = 1$

Gli scarti **x - μ** si trasformano in **scarti ridotti** $\frac{x - \mu}{\sigma}$

- la nuova distribuzione viene indicata con **N(0,1)**

- dopo il cambiamento di variabile, la densità di probabilità è $y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$

(Si evidenzia l'assenza di dipendenza da media e varianza: la distribuzione è sempre la stessa, qualunque sia la distribuzione gaussiana considerata)

Tendono alla normale:

- la distribuzione binomiale $(p + q)^n$, quando 'n' $\rightarrow \infty$
 - la distribuzione poissoniana, quando la media è elevata (in pratica, con media $\approx 10-12$ la normalità della distribuzione è manifesta)
- Sono alla base della **LEGGE DEI GRANDI NUMERI** (o LEGGE DEL CASO o LEGGE DI BERNOULLI) che costituisce il teorema fondamentale della Statistica :
se si ripete 'n' volte (per 'n' $\rightarrow \infty$) una prova in cui la probabilità a priori di accadimento dell'evento A è 'p', la probabilità dello stesso evento A tende a 'p'
- Danno luogo al **TEOREMA DEL LIMITE CENTRALE** (Laplace nel 1812) utilizzato per la media di valori di un campione :
le MEDIE di campioni, di dimensioni 'n' sufficientemente grandi, estratti da una popolazione comunque distribuita, seguono la legge della distribuzione normale, con media 'm' e varianza 's²/n'

TRASFORMAZIONI

Quando una variabile è distribuita normalmente, l'applicazione di funzioni matematiche quali logaritmi, radici quadratiche o cubiche, funzioni esponenziali, reciproci, ecc. conduce a una variabile distribuita in modo approssimativamente normale

ESEMPIO

Il caso più frequente è quello di $x' = \log x$

dove :

x' è distribuita normalmente

in cui si dice che x è distribuito secondo la DISTRIBUZIONE LOG-NORMALE

Tale distribuzione è frequente in tutti quei fenomeni in cui i fattori hanno tra loro effetti moltiplicativi

PROPRIETÀ E USO DELLA DISTRIBUZIONE NORMALE

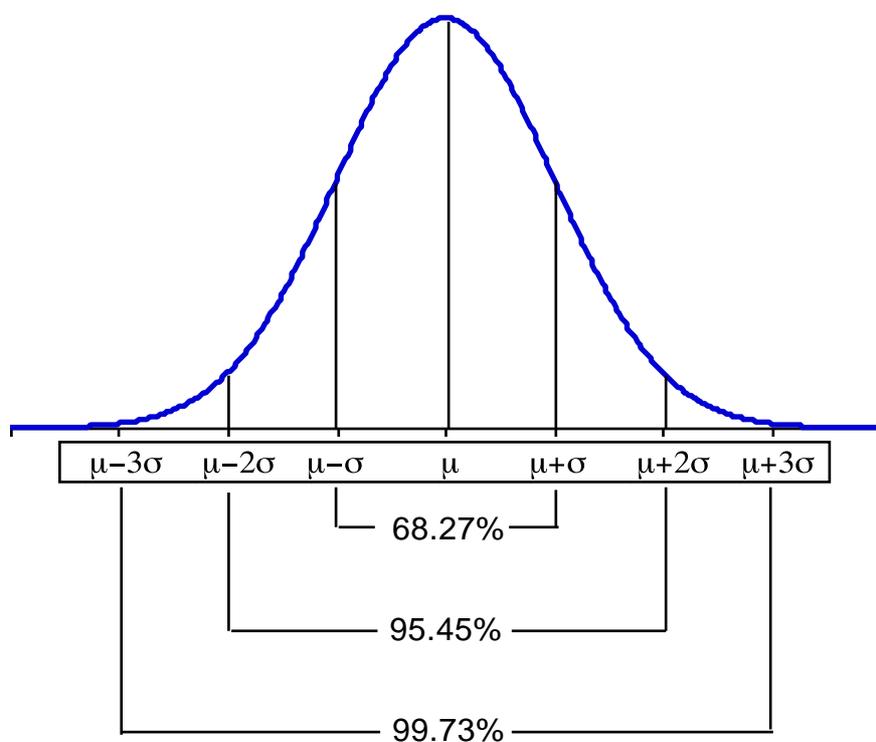
• relazioni tra la distanza dalla media (misurata in unità di deviazione standard) di un qualunque valore dell'asse x e la densità di probabilità sottesa dalla curva :

- frazione dei casi compresi nell'intervallo $\mu - \sigma \dots \mu + \sigma$ = 68,27%

- frazione dei casi compresi nell'intervallo $\mu - 2\sigma \dots \mu + 2\sigma$ = 95,45%

- frazione dei casi compresi nell'intervallo $\mu - 3\sigma \dots \mu + 3\sigma$ = 99,73%

In pratica la quasi totalità dei dati è compresa nell'intorno $\mu \pm 3\sigma$



E' pertanto possibile dedurre la distribuzione di dati quando siano noti μ e σ

UTILIZZO PRATICO DELLA DISTRIBUZIONE NORMALE STANDARDIZZATA

Le tabelle dei valori dell'integrale di probabilità della distribuzione normale standardizzata forniscono le probabilità di ottenere un valore dello scarto standardizzato

$$z = \frac{x - \mu}{\sigma}$$

maggiore di z (o minore di z , a seconda del tipo di tabella)

ESEMPIO

In una popolazione di pesci $\mu = 35$ (cm) e $\sigma = 5$ (cm)

D.:

calcolare le probabilità di pescare pesci di lunghezza :

- a) $l \geq 40$ (a destra di $z = +1$)
- b) $l < 40$ (tra media e $z = +1$)
- c) $l < 25$ (a sinistra di $z = -2$)
- d) $l \geq 40$ e $l \leq 50$ (tra $z = +1$ e $z = +3$)
- e) $l \geq 30$ e $l \leq 40$ (tra $z = -1$ e $z = +1$)

ricordando che :

probabilità area sottesa tra μ e $z = 1$	0,3413 (34,13%)
probabilità area sottesa a sinistra di $z = -2$	0,0228 (2,28%)

R.:

a) probabilità di pescare pesci di $l \geq 40$ cm	0,1587 (15,87%)
b) “ $l < 40$ cm	0,8413 (84,13%)
c) “ $l < 25$
d) “ $l \geq 40$ e $l \leq 50$ (differenza 0,49865-0,3413)	0,1573 (15,73%)
e) “ $l \geq 30$ e $l \leq 40$ (intervallo $z = -1$ e $z = 1$)	0,6826 (68,26%)

ESEMPIO

In una specie di roditori adulti, femmine e maschi si distinguono per le dimensioni :

femmina: $\mu = 37,5$ cm ; $\sigma = 3,8$ cm

maschio: $\mu = 34,5$ cm ; $\sigma = 3,2$ cm

D.:

- rispetto alle μ del loro sesso, sono più rari i maschi ≥ 40 cm o le femmine ≥ 41 cm ?
- quale è la lunghezza minima del 5% delle femmine di dimensioni $> \mu$?
- quale è la lunghezza massima del 5% dei maschi di dimensioni $< \mu$?
- tra il 30% delle femmine di dimensioni $> \mu$, quanti maschi è possibile trovare ?
- tra il 20% delle femmine di dimensioni $< \mu$, quanti maschi è possibile trovare ?

R.:

- per i maschi ≥ 40 cm $z=1,72$ che esclude a destra un'area equivalente al 5,26%
per le femmine ≥ 41 cm $z=0,92$ che esclude a destra un'area equivalente al 17,88%
 \therefore i maschi ≥ 40 cm sono molto più rari delle femmine ≥ 41 cm
- il 5% delle femmine di dimensioni $> \mu$ sono alla destra di $1,645 \cdot \sigma$ equivalente a
 $1,645 \cdot 3,8 = 6,251$ cm
 \therefore lunghezza minima del 5% delle femmine di dimensioni maggiori :
 $37,5 + 6,251 = 43,751$ cm
- il 5% dei maschi di dimensioni $< \mu$ sono alla sinistra di $1,645 \cdot \sigma$ equivalente a
 $1,645 \cdot 3,2 = 5,264$ cm
 \therefore lunghezza massima del 5% dei maschi di dimensioni minori :
 $34,5 - 5,264 = 29,236$ cm
- il valore di z che esclude il 30% della popolazione è 0,525 corrispondente alla destra della μ alle femmine di dimensioni $\mu + 0,525 \cdot \sigma$
pari a $37,5 + 0,525 \cdot 3,8 = 39,495$ cm
 \therefore ai maschi di tali dimensioni minime corrisponde
 $z = \frac{39,495-34,5}{3,2} = 1,56$ pari a una frequenza di probabilità del 5,94%
- il valore di z che esclude il 20% della popolazione è 0,842 corrispondente alla sinistra della μ alle femmine di dimensioni $\mu - 0,842 \cdot \sigma$
pari a $37,5 - 0,842 \cdot 3,8 = 34,3004$ cm
 \therefore ai maschi di tali dimensioni massime corrisponde
 $z = \frac{34,3004-34,5}{3,2} = -0,0623$ pari ad una frequenza di probabilità del 47,5%

CORREZIONI PER LA CONTINUITA' IN PROBABILITA' DISCRETE

Come già sottolineato, molte distribuzioni discrete (binomiale, ipergeometrica, ...) sono bene approssimate dalla distribuzione normale al crescere delle dimensioni del campione

Tuttavia mentre le prime forniscono le probabilità per **singoli valori** della variabile casuale, cioè la probabilità di ottenere **esattamente** il numero x , con le distribuzioni continue (tra cui la normale) si calcola l'area sottesa, cioè la densità di probabilità

Per calcolare la probabilità di verificarsi di un **singolo valore** x , con la distribuzione normale si deve calcolare l'area sottesa dall'intervallo $x \pm 0.5$

ESEMPIO

Si supponga che, da dati di letteratura, sia noto che in una popolazione zooplanctonica lacustre, gli individui di *Eudiaptomus vulgaris* assommino al 10% del totale individui. In un campionamento casuale di 120 individui quale è la probabilità di trovare:

D.:

Con un campione casuale di 120 individui, calcolare la probabilità di trovare

- a) **esattamente** 15 individui di *Eudiaptomus*
- b) **almeno** 15 individui di *Eudiaptomus*
- c) **meno** di 15 individui di *Eudiaptomus*

$$n = 120 \quad x = 15$$

$$\mu = np = 120 \cdot 0,10 = 12$$

$$\sigma^2 = npq = 120 \cdot 0,10 \cdot 0,90 = 10,8$$

Per valori discreti si deve aggiungere o togliere 0,5 al valore x (a seconda che il valore debba essere compreso od escluso), mentre per dati continui non si apporta alcuna correzione

R.:

- a) Probabilità di trovare **esattamente** 15 individui di *Eudiaptomus*: 7,90 %

$$z_1 = \frac{(x + 0,5) - \mu}{\sigma} = \frac{(15 + 0,5) - 12}{\sqrt{10,8}} = \frac{3,5}{3,29} = 1,06$$

per cui tra μ e $1,06 \cdot \sigma$ è compreso il 35,54% delle osservazioni

$$z_2 = \frac{(x - 0,5) - \mu}{\sigma} = \frac{(15 - 0,5) - 12}{\sqrt{10,8}} = \frac{2,5}{3,29} = 0,76$$

per cui tra μ e $0,76 \cdot \sigma$ è compreso il 27,64% delle osservazioni

$$\rightarrow = 35,54 - 27,64 = 7,90\%$$

[Il risultato si ottiene anche con la binomiale : $C_{120}^{15} (0,10)^{15} (0,90)^{105}$]

- b) Probabilità di trovare **almeno** 15 individui di *Eudiaptomus* : 22,36 %

$$z = \frac{(x + 0,5) - \mu}{\sigma} = \frac{15,5 - 12}{\sqrt{10,8}} = \frac{3,5}{3,29} = 1,06$$

per cui l'area a destra di $x = 15$ esprime una probabilità del 14,46% che, sommata al 7,90% del punto (a), porta alla probabilità complessiva del 22,36%

- c) Probabilità di trovare **meno** di 15 individui di *Eudiaptomus* : 77,64 %

$$z = \frac{(x - 0,5) - \mu}{\sigma} = \frac{14,5 - 12}{\sqrt{10,8}} = \frac{2,5}{3,29} = 0,76$$

per cui l'area tra $\bar{x} = 12$ e $x = 15$ esprime una probabilità del 27,64% che, sommata al 50% a sinistra della media (prob. di x da $x = 0$ a $\bar{x} = 12$), porta alla probabilità complessiva del 77,64%

[Il risultato si ottiene anche con la distribuzione binomiale, sommando le probabilità esatte di trovare 0, 1, 2, 3, ..., 14 individui di *Eudiaptomus*:

$$P(x, n) = \sum_{x=0}^n C_n^x \cdot p^x \cdot q^{n-x}$$

$$C_{120}^0 (0,1)^0 (0,9)^{120} + C_{120}^1 (0,1)^1 (0,9)^{119} + \dots + C_{120}^{14} (0,1)^{14} (0,9)^{106}]$$

DISTRIBUZIONE RETTANGOLARE

- come nelle distribuzioni discrete, anche tra le distribuzioni continue la più semplice è la distribuzione rettangolare o uniforme continua
- la densità di frequenze relativa all'intervallo $x_1 = a \dots x_2 = b$, è :

$$f(x) = \frac{1}{\beta - \alpha} \quad \text{con } (\alpha < x < \beta) \text{ costante in tutto l'intervallo } [a \dots b]$$

- nella rappresentazione grafica ha la forma di un rettangolo, da cui il nome
- media $\mu = \frac{\alpha + \beta}{2}$
- varianza $\sigma^2 = \frac{(\beta - \alpha)^2}{12}$
- è l'equivalente continuo della distribuzione rettangolare uniforme discreta

DISTRIBUZIONE ESPONENZIALE NEGATIVA

- la sua funzione è :

$$f(x) = \alpha e^{-\alpha x} \quad \text{con } \alpha > 0 \text{ e } x > 0$$

(prende il nome dall'esponente negativo che compare nella relazione)

- è una funzione positiva o nulla continuamente decrescente che tende a 0 per $x \longrightarrow \infty$
- nel discreto ha il suo equivalente nella D. GEOMETRICA DECRESCENTE
- media $\mu = \frac{1}{\alpha}$
- varianza $\sigma^2 = \frac{1}{\alpha^2} = \mu^2$ (N.B.: la varianza è il quadrato della media)

APPLICAZIONI DEI MODELLI DI DISTRIBUZIONE

Le applicazioni pratiche dei modelli di distribuzione teorica sono numerose; di particolare importanza sono quelle che riguardano la dispersione sul territorio di popolazioni animali e vegetali, dai micro-organismi a quelli di dimensioni maggiori. Il territorio è diviso in aree di dimensioni uguali ed entro ognuna di esse viene contato il numero di individui presenti

Trattandosi di conteggi, sono utili soprattutto le distribuzioni di variabili discrete, tra le quali si tratta di scegliere quella più appropriata a descrivere la distribuzione osservata. Il rapporto tra la media e la varianza è il primo e più immediato fra i criteri quantitativi di confronto o di valutazione, che permettono al ricercatore di individuare, seppure in via preliminare ed orientativa in attesa di verifiche ulteriori più approfondite, quale tipo di distribuzione si adatti meglio ai dati sperimentali raccolti

Quando la varianza risulta di entità simile alla media, si può supporre che la distribuzione territoriale della popolazione segua la legge poissoniana; trattandosi di eventi casuali ed indipendenti, l'interpretazione ecologica suggerisce che ogni individuo tenda a collocarsi nell'ambiente in modo completamente casuale ed indipendente dal comportamento di tutti gli altri individui della stessa specie, che non esistano né fattori che tendono ad aggregare né quelli che tendono a disperdere in modo uniforme. Se la varianza osservata risulta minore di quella teorica, la omogeneità della distribuzione può essere significativa: la specie in esame ha una dispersione geografica regolare, tipica di organismi con territorialità, che è necessaria quando la ricerca del cibo o la sopravvivenza esigono uno spazio minimo individuale per la sopravvivenza, di dimensioni approssimativamente simili per ogni individuo

Quando la varianza è maggiore dell'atteso, si può supporre che la distribuzione territoriale sia di tipo aggregato o contagioso, come quella degli animali con struttura sociale o delle piante e dei microorganismi concentrati in colonie: esistono ampi spazi liberi e contemporaneamente zone con una elevata densità di presenze

Quando variano le condizioni ambientali od aumenta la densità della specie, possono essere applicate strategie diverse di distribuzione geografica degli individui. Dal punto di vista statistico, la prima conseguenza è un forte aumento della varianza: un buon adattamento dei dati sperimentali alla distribuzione binomiale negativa può essere una indicazione importante, per inferire in prima approssimazione i fattori che regolano la dispersione

Le tecniche del campionamento rivestono un ruolo importante nella comprensione di questi fenomeni di particolare rilevanza specifica sono le dimensioni del campione e soprattutto quelle dell'area unitaria entro la quale sono contati gli individui presenti. E' dimostrato che variazioni nelle dimensioni dell'area presa come unità di campionamento mutano sensibilmente la forma della distribuzione, inducendo nel ricercatore deduzioni spesso fuorvianti

VERIFICA DELLE IPOTESI

Poiché in statistica ogni ipotesi è fondata su un confronto tra una “verità, nota, a livello di campione” e una “verità, ignota, a livello di popolazione”, esiste sempre una possibilità, anche se remota, che la conclusione cui porta un test (inferenza) sia sbagliata

Ogni test è pertanto associato a quattro probabilità interdipendenti che “misurano” il rischio che si corre (o della sicurezza che si ha) nel formulare una conclusione :

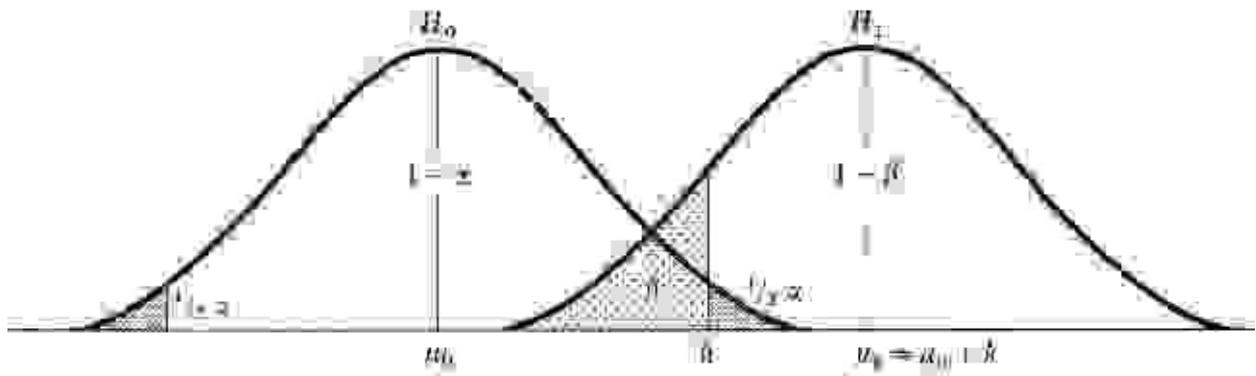
- **Errore di I[^] tipo (o rischio α):** [noto come “livello di significatività p”]
probabilità che esprime il rischio di rifiutare H_0 quando è vera
- **Errore di II[^] tipo (o rischio β):**
probabilità che esprime il rischio di accettare H_0 quando è falsa
- **Protezione del test $1-\alpha$ (complementare all'errore di I[^] tipo):**
probabilità, al livello prescelto, di accettare H_0 quando è vera
- **Potenza del test $1-\beta$ (complementare all'errore di II[^] tipo):**
probabilità, al livello prescelto, di rifiutare H_0 quando è falsa

		REALTA'	
CONCLUSIONE DEL TEST		H_0 vera	H_0 falsa
accetto H_0 statisticamente non significativo		Esatto $p = 1-\alpha$ PROTEZIONE	Errore β di II [^] tipo $p = \beta$
rifiuto H_0 statisticamente significativo		Errore α di I [^] tipo $p = \alpha$	Esatto $p = 1-\beta$ POTENZA

Il concetto di errore si comprende meglio ragionando sulle due distribuzioni di p legate alle due ipotesi H_0 e H_1 mutualmente esclusive

Anche le due distribuzioni sono mutualmente esclusive: una, quella legata all'ipotesi corretta, è vera; l'altra esiste solo in forma ipotetica

Il test statistico mette a confronto la stima campionaria con le distribuzioni H_0 e H_1



Distribuzioni definite da \bar{D}_0 e \bar{D}_1

L'“errore” si origina dal fatto che non è mai essere sicuri che il valore stimato dal test appartenga più di diritto all'una che all'altra delle due distribuzioni

L'area di sovrapposizione delle due curve, in relazione al valore campionario stimato, determina il rigetto o l'accettazione di H_0

Ne consegue la probabilità α di commettere un errore rispettivamente di I^o tipo (rigetto il vero) o di II^o tipo (accetto il falso)

Il valore x è determinato dall'area che rimane all'esterno del punto di stima rispetto al valore medio della distribuzione H_0

Nel confronto tra due frequenze, secondo H_0 non esiste differenza sostanziale, se non quella dovuta a fattori casuali. Occorre dunque stimare la probabilità p di trovare, con esperimenti ripetuti e nel caso che H_0 sia vera, un valore uguale o superiore a quello calcolato

Se p (riportata nelle tabelle) è inferiore al valore di significatività α prefissato ($\alpha=5\%$ o $\alpha=1\%$), si rifiuta H_0 ; ma se H_0 è vera, nel rifiutarla si sbaglia con probabilità $p < 5\%$ (errore di I^o tipo)

Per ridurre la probabilità di commettere errori di I^o tipo si abbassa il livello di significatività da $p=5\%$ a $p=1\%$

La probabilità calcolata dal test si riferisce al caso in cui H_0 è vera e stima la probabilità di commettere un errore di I^o tipo

C'è “concorrenza” tra l'errore di I^o tipo e l'errore di II^o tipo : se si abbassa il livello di significatività, cioè la probabilità di commettere un errore di I^o tipo, si accresce quella di commettere un errore di II^o tipo e viceversa

INTERVALLO di CONFIDENZA di una MEDIA CON σ nota

Volendo conoscere il valore più probabile di un parametro incognito, la statistica inferenziale fornisce due valori che determinano l'INTERVALLO DI CONFIDENZA (o LIMITI FIDUCIALI) entro cui si colloca il valore del parametro secondo la probabilità scelta

Specificare i limiti fiduciali è solamente un modo alternativo di inferire circa i parametri di una popolazione, sulla base di osservazioni campionarie

I limiti fiduciali della media della popolazione sono stimati dalla distribuzione normale standardizzata :

- il 95% dell'area sottesa dalla curva si trova tra -1,96 e +1,96 dell'ascissa
[$P(-1,96 \leq Z \leq +1,96) = 0,95$]

- il 99% dell'area sottesa dalla curva si trova tra -2,58 e +2,58 dell'ascissa
[$P(-2,58 \leq Z \leq +2,58) = 0,99$]

Così come s valuta la dispersione di campionamento delle osservazioni, l'ERRORE STANDARD (ES) valuta la dispersione delle medie campionarie :

$$ES = \frac{s}{\sqrt{n}}$$

La distribuzione di campionamento di medie con media m ed $ES = \frac{s}{\sqrt{n}}$ diventa

$P\left(-Z \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq +Z\right) = P(z)$ e può essere usata per determinare i limiti fiduciali :

- al 95% diventa $P\left(\bar{x} - 1,96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$

- al 99% sostituire, nella formula sopra scritta, 1,96 con 2,58

ESEMPIO

Da una popolazione con $\sigma=3$ è stato estratto un campione di 10 dati con $m=25$

D.: Calcolare l'intervallo di confidenza alla probabilità del 99%

R.: $25 \mp 2,58 \cdot \frac{3}{\sqrt{10}} = 25 \mp 2,58 \cdot 0,9487 = 25 \mp 2,45 = \begin{cases} 22,55 \\ 27,45 \end{cases}$

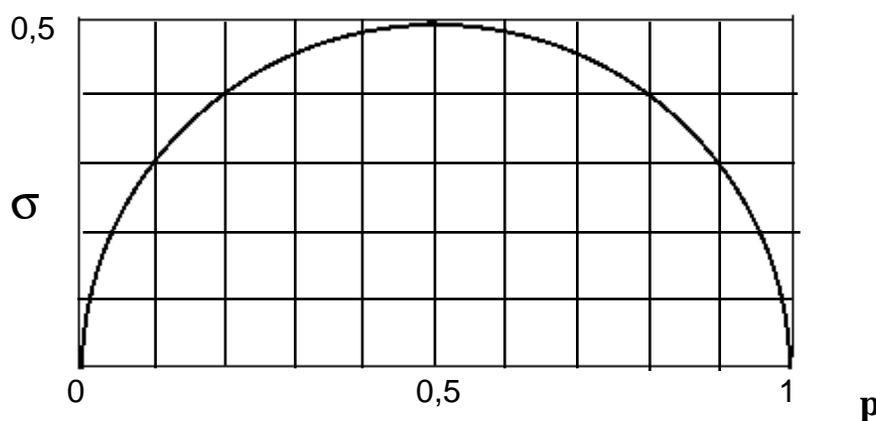
Secondo le informazioni fornite da una campione di 10 misure con $m=25$, con probabilità 99% μ si trova nell'intervallo compreso tra 22,55 e 27,45

[... rimane la probabilità dell'1% che μ si trovi fuori da questo intervallo ...]

Intervallo di confidenza di una proporzione

Per i limiti fiduciali di una proporzione si adotta l'approssimazione della normale alla binomiale

In una proporzione, il valore di σ è completamente determinato dal valore della media p , infatti con n costante $\sigma = \sqrt{p \cdot (1-p)}$



La σ di una proporzione si approssima a zero quando p è molto piccolo o molto grande e presenta valore massimo quando p è prossimo al valore centrale 0,5

L'intervallo di confidenza di una percentuale è dato da $p \pm Z \cdot \sqrt{\frac{p \cdot q}{n}}$

ESEMPIO

In un campione di 80 fumatori, il 35% ha presentato sintomi di polmonite

D.:

- Calcolare i limiti fiduciali della media al 95% e al 99% nella popolazione dei fumatori con sintomi di polmonite

- Calcolare gli stessi limiti fiduciali (95% e 99%) partendo da un campione di 100 fumatori anziché 80

R.:

Con un campione di 80 fumatori si ha :

$$\text{Per il 95\% : } 0,35 \pm 1,96 \cdot \sqrt{\frac{0,35 \cdot 0,65}{80}} = 0,35 \pm 0,1045 = \begin{cases} 0,2455 \\ 0,4545 \end{cases}$$

$$\text{Per il 99\% : } 0,35 \pm 2,58 \cdot \sqrt{\frac{0,35 \cdot 0,65}{80}} = 0,35 \pm 0,1376 = \begin{cases} 0,2124 \\ 0,4876 \end{cases}$$

Con un campione di 100 fumatori si avrebbe :

$$\text{Per il 95\% : } 0,35 \pm 1,96 \cdot \sqrt{\frac{0,35 \cdot 0,65}{100}} = 0,35 \pm 0,09349 = \begin{cases} 0,2565 \\ 0,4435 \end{cases}$$

$$\text{Per il 99\% : } 0,35 \pm 2,58 \cdot \sqrt{\frac{0,35 \cdot 0,65}{100}} = 0,35 \pm 0,1231 = \begin{cases} 0,2269 \\ 0,4731 \end{cases}$$

Si noti che con un campione di 100 individui gli intervalli sono più stretti rispetto a quelli prodotti dal campione di 80 individui

ANALISI DELLE FREQUENZE e CONFRONTI TRA DISTRIBUZIONI

DISTRIBUZIONI OSSERVATE e DISTRIBUZIONI ATTESE

Nella teoria statistica e nella pratica sperimentale, sia con dati qualitativi (classificati in categorie nominali) che con dati quantitativi (distribuiti in classi di intervallo), per verificare se esiste accordo tra una distribuzione osservata e la corrispondente distribuzione attesa si ricorre al

TEST PER LA BONTÀ DELL'ADATTAMENTO (goodness of fit)

ESEMPIO

distribuzioni di frequenze osservate di classi fenotipiche vs.
distribuzioni di frequenze attese secondo le leggi di segregazione mendeliana

D.:

Verificare se la distribuzione della progenie degli ibridi rispetta :

- la distribuzione teorica attesa di **3:1** per un carattere
- la distribuzione teorica attesa di **9:3:3:1** per due caratteri

N.B.

- tra distribuzioni osservate e distribuzioni attese non c'è mai perfetta coincidenza, ma valori molto simili
- le classi di una distribuzione osservata sono conteggi (numeri interi)
- le classi di una distribuzione attesa seguono una legge teorica (descritta da numeri frazionari)

R.:

Distribuzioni osservate ed attese di *Pisum sativum* (Mendel)

A - Segregazione di un ibrido			
<i>carattere</i>	<i>dominante</i>	<i>recessivo</i>	<i>totale</i>
colore del fiore (d. oss.)	rossi 705	bianchi 224	929
distribuzione attesa (3:1)	696,75	232,25	
lunghezza del fusto (d. oss.)	alti 787	bassi 277	1064
distribuzione attesa (3:1)	798	266	
colore del seme (d. oss.)	gialli 6022	verdi 2001	8023
distribuzione attesa (3:1)	6017,25	2005,75	
forma del seme (d. oss.)	lisci 5474	rugosi 1850	7324
distribuzione attesa (3:1)	5493	1831	

B - Segregazione di un diibrido		
<i>colore e forma del seme</i>	<i>distr. osservata</i>	<i>d. attesa (9:3:3:1)</i>
gialli-lisci	315	$9/16 = 312,75$
gialli-rugosi	101	$3/16 = 104,25$
verdi-lisci	108	$3/16 = 104,25$
verdi-rugosi	32	$1/16 = 34,75$
totale	556	556

Il problema statistico di capire se le differenze tra osservato e atteso sono trascurabili e quindi dovute al caso, oppure tali da fare supporre l'esistenza di fattori causanti una distribuzione realmente diversa da quella attesa...

... è problema di INFERENZA STATISTICA per verificare l'attendibilità dell'ipotesi nulla H_0 (differenze casuali) o dell'ipotesi alternativa H_1 (differenze dovute a fattori non casuali), mediante l'impiego di un test di significatività

IL TEST CHI QUADRO (CHI QUADRATO, o χ^2)

Proposto da Pearson nel 1900, utilizza le frequenze assolute

$$\chi_{(g.d.l.)}^2 = \sum_{i=1}^n \frac{(f_i^{oss} - f_i^{att})^2}{f_i^{att}}$$

f_i^{oss} = i-esima frequenza osservata

f_i^{att} = i-esima frequenza attesa

gdl = n° di gruppi meno uno **n-1** (in basso, tra parentesi: $\chi_{(g.d.l.)}^2$)

Σ estesa a tutti i gruppi (o classi) posti a confronto

I valori attesi, calcolati sul totale secondo la legge di distribuzione, possono assumere qualsiasi valore, eccetto l'ultimo, la cui frequenza sommata alle precedenti deve rispettare il totale

Procedimento logico nell'applicazione del χ^2 :

- 1: stabilire l'ipotesi nulla (H_0) e l'eventuale ipotesi alternativa (H_1)
- 2: individuare il test più appropriato per saggiare l'ipotesi nulla H_0
- 3: scegliere: livello di significatività, ampiezza del campione, gdl
- 4: trovare la distribuzione teorica del test statistico nell' H_0 (fornita dalle tabelle)
- 5: stabilire la zona di rifiuto (solitamente fissata al 5% oppure all' 1%)
- 6: calcolare il valore del test statistico sulla base dei dati sperimentali, stimando il valore di probabilità ad esso associato
- 7: se la probabilità è superiore a quella tabulata, non si può rifiutare H_0 ;
se la probabilità è inferiore a quella tabulata, si rifiuta H_0 (implicitamente si accetta H_1)

ESEMPIO

Calcolare il χ^2 con i dati sulla segregazione del di-ibrido colore / forma del seme :

$$\chi_{(3)}^2 = \frac{(315 - 312,75)^2}{312,75} + \frac{(101 - 104,25)^2}{104,25} + \frac{(108 - 104,25)^2}{104,25} + \frac{(32 - 34,75)^2}{34,75}$$

$$\chi_{(3)}^2 = \frac{(2,25)^2}{312,75} + \frac{(-3,25)^2}{104,25} + \frac{(3,75)^2}{104,25} + \frac{(-2,75)^2}{34,75} = 0,47$$

Attraverso le tavole è possibile stimare la probabilità di trovare differenze uguali o superiori a quelle riscontrate tra distribuzione osservata e distribuzione attesa, nell'ipotesi (H_0) che le differenze siano imputabili a fattori casuali

Nella tavola della distribuzione dei valori critici del χ^2 :

- per 3 gdl (rif. riga) e per $p=0,05$ (rif. colonna) --> $\chi^2 = 7,81$

Il valore calcolato (0,47) è molto minore di quello tabulato, dunque la probabilità che le differenze siano imputabili al caso è superiore al valore prefissato del 5% ($p > 0,05$), e non si può rifiutare H_0 (le differenze sono imputabili a fattori casuali)

Ipotesi nulla H_0 : le differenze tra distribuzione osservata e distribuzione attesa sono trascurabili e quindi non significative

Ipotesi alternativa H_1 : le differenze tra distr. osservata e distr. attesa sono rilevanti, non dovute al caso, ma ad un fattore che determina una segregazione diversa

Test più adatto : in base alle caratteristiche dei dati e alle ipotesi formulate, è il χ^2

Livello di **significatività** prescelto : 5%

Valori di riferimento del χ^2 (livello di significatività e gdl) : sono forniti dalla tabella

Zona di rifiuto è solo da una parte della distribuzione : si tratta di test ad una sola coda [il χ^2 tende a crescere per valori osservati sia inferiori che superiori ai valori attesi; inoltre non sono possibili valori negativi]

Confronto tra χ^2 calcolato e χ^2 tabulato (3 gdl con $\mathbf{p} = 5\%$): $0,47 \ll 7,81$

Probabilità di trovare scarti uguali o superiori a quello calcolato, nell' $H_0 : \mathbf{p} > 5\%$ (solo differenze imputabili al caso tra osservato ed atteso)

Non si può concludere che esista una differenza significativa tra la distribuzione osservata e quella attesa
Tale differenza potrebbe tuttavia esistere, ma con $\mathbf{p} < 5\%$ (la stessa probabilità con cui si può sbagliare affermando che la differenza non esiste)

ESEMPIO

In una popolazione lacustre di *Mixodiptomus Kupelwieseri* (copepode) sono state rilevate le frequenze di 4 alleli al locus MPI (mannoso fosfato isomerasi)

	freq. oss.		
allele 1	26		
allele 2	38		
allele 3	62		
allele 4	118	totale	244

D.:

Essendo la frequenza attesa per ogni allele, secondo l'ipotesi di pura casualità, $= \frac{244}{4} = 61$, le differenze riscontrate fra le frequenze dei vari alleli sono casuali ?

R.:

$$\chi_{(3)}^2 = \frac{(26-61)^2}{61} + \frac{(38-61)^2}{61} + \frac{(62-61)^2}{61} + \frac{(118-61)^2}{61} = \frac{1225}{61} + \frac{529}{61} + \frac{1}{61} + \frac{3249}{61} = 82,0328$$

• nella tabella del χ^2 per 3 gdl e significatività $\mathbf{p} = 0,001$ il χ^2 calcolato è molto più grande di quello tabulato

• la probabilità che le differenze tra i gruppi siano imputabili al caso è $\mathbf{p} < 0,001$, pertanto le differenze non possono essere considerate casuali

LA DISTRIBUZIONE χ^2

La distribuzione dei valori del χ^2 può essere studiata empiricamente mediante ripetuti lanci di una moneta. Ad esempio, per 100 lanci, si può ipotizzare di ottenere le seguenti frequenze assolute, che portano ai valori di χ^2 a fianco indicati :

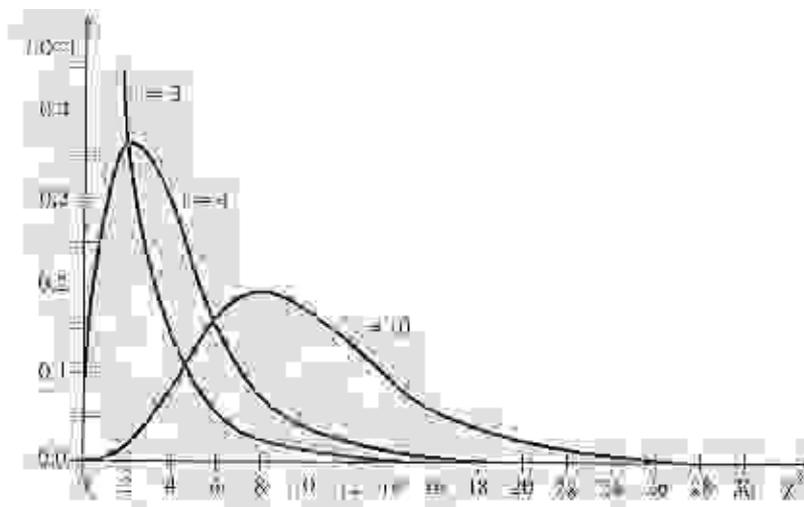
testa	croce	χ^2
51	49	0,04
47	53	0,36
49	51	0,04
50	50	0,00
42	58	2,56
48	52	0,16
53	47	0,36

La distribuzione di questi χ^2 empirici è simile a quella tabulata per **1** gdl

- **1** gdl :: distribuzione dei quadrati di **n** variabili casuali normali standardizzate indipendenti (in termini matematici: $z^2 \cong \chi_{(1)}^2$)
- **n** gdl :: distribuzione della somma dei quadrati di **n** variabili casuali normali standardizzate indipendenti ($\sum_{i=1}^n z_i^2 \cong \chi_{(n)}^2$)

[la standardizzazione è ottenuta dividendo la differenza tra osservato ed atteso per il valore atteso]

- il n° dei gdl è funzione dei vincoli fra le frequenze dei vari gruppi :quando tra **n** variabili casuali sussistono **k** vincoli lineari (relazioni che riducono il n° di osserv. indep.), i gdl del corrispondente χ^2 diminuiscono di **k**



Densità di frequenza della distribuzione di χ^2 per $v=1, 2, e 10$

CONDIZIONI DI VALIDITA' DEL χ^2

- solo per grandi campioni [non esiste concordanza generale su quando un campione può essere ritenuto di grandi dimensioni]
- il n° dei gdl dipende dal n° di gruppi
- il n° totale di osservazioni deve essere $N > 100$ [per alcuni $N > 200$ o $N > 500$]
 - richiede una correzione (**Yates**) quando $30 < N < 100$ che consiste nel :
 - sottrarre 0,5 al valore assoluto dello scarto maggiore (tra freq. oss. e freq. att.)
 - aggiungere 0,5 al valore assoluto dello scarto minore
- perde attendibilità quando $N < 30$ [per alcuni questo limite è 40, per altri 25-20]
 - poiché il n° totale di osservazioni è suddiviso in più classi, ogni gruppo o classe deve avere, per le frequenze attese, un n° minimo di 5 osservazioni

ESEMPIO In tre parcelle equivalenti sono stati contati 15, 21 e 24 individui di un vegetale

D.: Verificare se l'osservato si discosta in modo significativo dall'atteso teorico di 20, 20 e 20, secondo l'ipotesi di distribuzione uniforme

R.:

$$\chi^2 \text{ senza correzione : } \chi_{(2)}^2 = \frac{(5)^2}{20} + \frac{(1)^2}{20} + \frac{(4)^2}{20} = 2,100$$

$$\chi^2 \text{ con correzione di Yates : } \chi_{(2)}^2 = \frac{(4,5)^2}{20} + \frac{(1,5)^2}{20} + \frac{(4)^2}{20} = 1,925$$

La correzione di Yates riduce il χ^2 in modo tanto maggiore quanto più ridotto è il numero di osservazioni, infatti :

- quando il numero di osservazioni è piccolo, le variazioni casuali tendono ad aumentare la loro incidenza relativa: lo scarto tra osservato ed atteso non risente solamente delle differenze realmente esistenti tra i due fenomeni a confronto, ma anche delle variazioni casuali

- intuitivamente si comprende che $\chi^2 = 0$ quando il n° di osservazioni è molto basso, infatti le variazioni casuali tendono ad essere così elevate, da non permettere più di evidenziare in modo significativo l'esistenza di differenze reali tra osservato ed atteso, ovvero il "rumore" (le variazioni casuali) è così forte da non permettere di evidenziare le reali tendenze di fondo delle distribuzioni

CONFRONTO TRA FREQUENZE RELATIVE DI DUE POPOLAZIONI INDIPENDENTI

TEST a UNA CODA - TEST a DUE CODE

Nel confronto tra osservato e atteso sono possibili due diverse impostazioni concettuali :

test bilaterale (o test a due code) :

ci si chiede se esiste una differenza nelle frequenze relative tra i due gruppi, senza porre vincoli sul segno algebrico di tale differenza

test unilaterale (o test a una coda) :

ci si chiede se un gruppo abbia una frequenza relativa significativamente maggiore oppure minore, scartando a priori l'ipotesi alternativa

ESEMPIO

Confronto tra due differenti metodi di ricattura di animali

- Quando si vuole sapere se tra i due metodi c'è differenza significativa, ed è indifferente che risulti migliore il primo metodo oppure il secondo, si tratta di un test bilaterale a due code
- Quando ad un vecchio metodo si vuole sostituire un metodo nuovo ritenuto più efficace, e si vuole dimostrare la sua superiorità rispetto al precedente, si ha a che fare con un test unilaterale a una coda

Da tale distinzione dipende la distribuzione delle probabilità per rifiutare l'ipotesi nulla

Scegliendo la probabilità del 5% :

- in un test a due code si hanno due zone di rifiuto collocate ai due estremi, ognuna con un'area di 2,5%

- in un test a una coda si ha una sola zona di rifiuto, con un'area di 5%

CONFRONTO TRA DISTRIBUZIONI OSSERVATA E ATTESA IN PICCOLI CAMPIONI

Per stabilire la BONTÀ DELL'ADATTAMENTO (o BONTÀ DELLA CONFORMITÀ) di una distribuzione osservata a una distribuzione teorica, quando

- il n° di osservazioni è molto ridotto, convenzionalmente inferiore a 30
- le frequenze attese entro ogni gruppo sono inferiori a 5

si ricorre al **TEST DI KOLMOGOROV-SMIRNOV** anziché al χ^2

Requisiti :

- i gruppi devono essere ordinati secondo una scala ordinale (con il χ^2 l'ordine dei gruppi è ininfluente)
- il confronto viene attuato tra le due distribuzioni cumulative, tra le quali si determina il valore di massima divergenza
- la distribuzione di campionamento indicherà la probabilità di trovare una divergenza superiore a quella calcolata (H_0)

ESEMPIO

In dieci ore, dalle 7 alle 17, dal luogo di appostamento un osservatore avvista 15 uccelli della stessa specie :

Orario	7-8	9-10	11-12	13-14	15-16
Uccelli avvistati	0	1	1	9	4

Verificare se la distribuzione è casuale, cioè se le variazioni osservate rientrano nei limiti degli errori accidentali (H_0), oppure se è più attendibile pensare ad una incidenza dell'orario sul numero di avvistamenti (H_1)

Se l'ora non incidesse, l'osservatore avrebbe dovuto avvistare un numero fisso di uccelli pari alla media (15/5) ad intervalli costanti, 3 ogni 2 ore :

Ore	7-8	9-10	11-12	13-14	15-16
Distribuzione attesa	3	3	3	3	3

Il confronto a coppie tra le due distribuzioni cumulative permette di trovare la differenza massima assoluta (o scarto massimo assoluto) :

Ore	7-8	9-10	11-12	13-14	15-16
Distribuzione cumulativa <u>osservata</u>	0	1	2	11	15
Distribuzione cumulativa <u>attesa</u>	3	6	9	12	15
Scarti assoluti (differenze, Δ)	3	5	7	1	0

^^^

- è intuitivo pensare che lo scarto massimo assoluto sia tanto più grande quanto maggiori sono i singoli scarti tra osservato ed atteso e che questo valore dipenda anche dal numero totale di osservazioni

- per rendere lo scarto massimo assoluto indipendente dal numero totale di osservazioni si ricorre al rapporto

$$D[\text{deviazione massima}] = \frac{\text{scarto massimo}}{\text{numero totale di osservazioni}} \quad D = \frac{7}{15} = 0,466$$

- sulla tabella dei valori critici di D per un campione $N = 15$ al diminuire del livello di significatività da 0,20 a 0,01 il valore critico cresce da 0,266 a 0,404

- il valore $D = 0,466$ è superiore a quello tabulato sia per $p=0,05$ che per $p=0,01$

- si rifiuta H_0 e implicitamente si accetta H_1 (le variazioni del numero di osservazioni durante la giornata non siano casuali)

Utilizzando il χ^2 :

- occorrere raggruppare i dati per classi adiacenti
- si perdono informazioni sulle differenze tra le varie ore
- è implicita una elevata dose di soggettività
- è più difficile dimostrare che esiste una differenza significativa tra osservato ed atteso, quando fosse vera H_1

Utilizzando il test di Kolmogorov-Smirnov :

- aumenta la potenza rispetto al test χ^2
- si possono impiegare anche piccoli campioni
- non si perdono informazioni per formare gruppi
- si possono usare campioni di medie dimensioni suddivisi in gruppi

CONFRONTO TRA DUE DISTRIBUZIONI OSSERVATE

LE TABELLE 2×2 per il TEST DI INDIPENDENZA

Quando si confrontano le frequenze relative di risposte binarie (SÌ / NO) tratte da due popolazioni indipendenti, si può costruire una

TABELLA DI CONTINGENZA

 (a doppia entrata)

con il n° di successi e il n° di insuccessi in ognuno dei due gruppi, in modo da verificare se le proporzioni di successi e di insuccessi nei due gruppi sono indipendenti dal trattamento a cui sono sottoposti

Questo χ^2 è chiamato **TEST DI INDIPENDENZA** con le ipotesi :

- H_0 : c'è indipendenza tra l'appartenere al gruppo A o B e la risposta
- H_1 : non c'è indipendenza tra l'appartenere al gruppo e la risposta

Se H_0 non può essere respinta, poiché le frequenze tra i due gruppi sono simili, allora non esiste un rapporto tra le due variabili

Se H_0 viene respinta, allora esiste un rapporto tra le due variabili

N.B.

Sia che il χ^2 venga utilizzato per verificare la differenza tra due frequenze relative di due gruppi, sia che venga impiegato per saggiare l'indipendenza tra due variabili, i calcoli e i risultati sono gli stessi

ESEMPIO

Si vuole controllare l'effetto di due sostanze tossiche su due gruppi di animali :

- l'agente A, somministrato a 70 animali, ha causato la morte di 22 individui (48 sono sopravvissuti)
- l'agente B somministrato a 50 animali ha causato la morte di 24 individui (26 sono sopravvissuti)

D.:

Le due sostanze hanno gli stessi effetti sulla mortalità o sopravvivenza (H_1), oppure le differenze riscontrate debbono essere ritenute casuali (H_0) ?

Le frequenze osservate vengono poste in una tabella a due entrate :

osservati	morti	sopravvissuti	totale
agente A	22	48	70
agente B	24	26	50
totale	46	74	120

Le frequenze attese secondo H_0 possono essere calcolate dai totali marginali :

$$frequenza\ attesa = \frac{totale\ riga \cdot totale\ colonna}{totale\ generale}$$

attesi	morti	sopravvissuti	totale
agente A	26,83	43,17	70
agente B	19,17	30,83	50
totale	46	74	120

Calcolata la prima frequenza attesa ($26,83 = \frac{70 \cdot 46}{120}$), le altre si ottengono per differenza dai totali rispettivi (:: la tabella di contingenza 2×2 ha 1 gdl):

$$43,17 = \frac{70 \cdot 74}{120} \quad oppure \quad 43,17 = 70 - 26,83$$

$$19,17 = \frac{50 \cdot 46}{120} \quad oppure \quad 19,17 = 46 - 26,83$$

(30,83 può essere calcolata sia dai suoi due totali marginali che dal totale generale)

R.:

Per calcolare il valore del χ^2 :

- si può utilizzare la formula generale $\chi_{(1)}^2 = \sum_{i=1}^4 \frac{(f_i^{oss} - f_i^{att})^2}{f_i^{att}}$

$$\begin{aligned}\chi_{(1)}^2 &= \frac{(22 - 26,83)^2}{26,83} + \frac{(48 - 43,17)^2}{43,17} + \frac{(24 - 19,17)^2}{19,17} + \frac{(26 - 30,83)^2}{30,83} = \\ &= \frac{23,33}{26,83} + \frac{23,33}{43,17} + \frac{23,33}{19,17} + \frac{23,33}{30,83} = 0,87 + 0,55 + 1,24 + 0,76 = 3,42\end{aligned}$$

- si può utilizzare la formula per il **calcolo rapido** (più veloce e sempre corretta):

$$\chi_{(1)}^2 = \frac{(a \cdot d - b \cdot c)^2 \cdot N}{n_1 \cdot n_2 \cdot n_3 \cdot n_4}$$

a, b, c, d : frequenze osservate

n_1, n_2, n_3, n_4 : totali marginali

N : totale generale

	X	x	totale
Y	a	b	n₁
y	c	d	n₂
totale	n₃	n₄	N

$$\chi_{(1)}^2 = \frac{(22 \cdot 26 - 48 \cdot 24)^2 \cdot 120}{70 \cdot 50 \cdot 46 \cdot 74} = \frac{(572 - 1152)^2 \cdot 120}{11914000} = \frac{336400 \cdot 120}{11914000} = \frac{40368000}{11914000} = 3,389$$

Nella tabella dei valori critici della distribuzione χ^2 per 1 gdl 3,389 è inferiore a quello tabulato per la probabilità del 5% (3,84)

CORREZIONE PER LA CONTINUITÀ (CORREZIONE DI YATES)

Va apportata per piccoli campioni (n° totale di osservazioni tra 30 e 100) e consiste nel sottrarre $\frac{N}{2}$ a $|ad-bc|$:

$$\chi_{(1)}^2 = \frac{\left(|a \cdot d - b \cdot c| - \frac{N}{2} \right)^2 \cdot N}{n_1 \cdot n_2 \cdot n_3 \cdot n_4}$$

ESEMPIO

Per valutare gli effetti di due diserbanti, si conta il numero di piante cresciute e di quelle non cresciute nei rispettivi appezzamenti :

	piante cresciute	piante non cresciute	totale
diserbante A	12	6	18
diserbante B	26	9	35
totale	38	15	53

E' un confronto tra due campioni indipendenti con un numero di osservazioni sufficientemente grande per consentire l'uso del test χ^2 con la correzione di Yates :

$$\chi_{(1)}^2 = \frac{\left(|12 \cdot 9 - 6 \cdot 26| - \frac{53}{2} \right)^2 \cdot 53}{18 \cdot 35 \cdot 38 \cdot 15} = \frac{(|108 - 156| - 26,5)^2 \cdot 53}{359100} = \frac{462,25 \cdot 53}{359100} = \frac{24499,25}{359100} = 0,0945$$

Il risultato è inferiore a quello tabulato per $p=90\%$:

esiste una probabilità molto elevata di trovare scarti uguali a quelli attesi e di conseguenza non si può rifiutare H_0 (= le differenze riscontrate tra gli effetti dei due diserbanti sono solamente dovute a variazioni casuali)

ESEMPIO

Confronto tra due metodi di “cattura e ricattura” per la stima della dimensione di popolazioni animali :

	animali ricatturati	animali non ricatturati	totale
metodo A	40	160	200
metodo B	39	111	150
totale	79	271	350

D.:

Esiste una differenza significativa tra i due metodi ?

R.:

Trattandosi di un campione di grandi dimensioni è possibile usare la formula per il calcolo rapido :

$$\chi_{(1)}^2 = \frac{(40 \cdot 111 - 160 \cdot 39)^2 \cdot 350}{200 \cdot 150 \cdot 79 \cdot 271} = \frac{(4440 - 6240)^2 \cdot 350}{642270000} = \frac{3240000 \cdot 350}{642270000} = \frac{1134000000}{642270000} = 1,765$$

Nella tabella del χ^2 a 1,765 corrisponde una probabilità di ~ 20%

N.B.

Le tabelle di contingenza 2×2 :

- consentono di effettuare solo test a due code
- si possono usare anche per i confronti tra frequenze relative

METODO ESATTO (o DELLE PROBABILITÀ ESATTE) DI FISHER

- permette di stimare la specifica probabilità di ottenere una data risposta sperimentale tra tutte le possibili con il numero di dati a disposizione
- si usa quando il campione ha un basso numero di osservazioni ($N < 30$) e il χ^2 non può essere usato nemmeno nelle tabelle 2×2
- a condizione di mantenere costanti i totali marginali, la probabilità esatta di osservare una particolare serie di frequenze può essere calcolata dalla distribuzione ipergeometrica
- la probabilità di trovare un particolare insieme dei dati osservati è :

$$p = \frac{C_{a+c}^a \cdot C_{b+d}^b}{C_N^{a+b}} = \frac{\frac{(a+c)!}{a! \cdot c!} \cdot \frac{(b+d)!}{b! \cdot d!}}{N!} = \frac{(a+b)! \cdot (c+d)! \cdot a! \cdot b!}{N! \cdot a! \cdot b! \cdot c! \cdot d!}$$

oppure, più semplicemente,

$$p = \frac{n_1! \cdot n_2! \cdot n_3! \cdot n_4!}{a! \cdot b! \cdot c! \cdot d! \cdot N!}$$

ESEMPIO

Confronto tra gli effetti letali di due biocidi :

	animali sopravvissuti	animali morti	totale
pesticida A	7	1	8
pesticida B	3	6	9
totale	10	7	17

D.:

Tra i due biocidi esiste una differenza significativa ?

R.:

La probabilità di avere, tra tutte le possibili risposte, quella osservata è :

$$p = \frac{8! \cdot 9! \cdot 10! \cdot 7!}{7! \cdot 1! \cdot 3! \cdot 6! \cdot 17!} = 0,03 \quad (\text{in percentuale, } 3\%)$$

Per valutare la significatività delle differenze riscontrate, occorre cumulare le risposte estreme, seguendo tre passaggi :

- individuare la frequenza osservata minore
- sostituire ad essa il valore 0 variando le altre 3 senza alterare i marginali
- aumentare di 1 tale valore finché compare 0 in un'altra casella

Con i dati dell'esempio, tenendo costanti i totali marginali, sono otto le risposte differenti che si sarebbero potute ottenere :

1)☞	2)☞	3)☞	4)☞				
8	0	7	1	6	2	5	3
2	7	3	6	4	5	5	4
☞	☞	☞	☞				
5)☞	6)☞	7)☞	8)☞				
4	4	3	5	2	6	1	7
6	3	7	2	8	1	9	0

[Non esistono altri valori che diano gli stessi totali di riga e di colonna]

Con il metodo esatto di Fisher :

- si calcola la probabilità di avere ognuna di queste risposte teoricamente possibili (totale: 1 se proporzione; 100 se percentuale)
- si passa da un estremo di un effetto più marcato per B (7 morti su 9, mentre con A sopravvivono tutti 8), all'altro estremo di un effetto più marcato per A (7 morti su 9, mentre con B sopravvivono tutti 9)
- per stabilire se esiste una differenza significativa, alla probabilità calcolata per la risposta 2 (che coincide con quella sperimentale) si somma la probabilità di ottenere le risposte più estreme nella stessa direzione (nell'esempio è una sola, la 1): se la somma supera il 5%, si accetta H_0
- le probabilità complessive calcolate possono essere estese in una sola direzione per test ad una coda; possono essere estese ad ambedue le direzioni per test a due code (nel qual caso la probabilità complessiva coincide con quanto è possibile calcolare con il test χ^2 , che è un test a due code)

TABELLE $M \times N$

Il metodo del χ^2 per verificare la differenza tra due proporzioni può essere esteso al caso generale del confronto tra M popolazioni indipendenti, per saggiare :

$$H_0 : p_1 = p_2 = p_3 = \dots = p_M$$

H_1 : almeno una frequenza relativa è diversa dalle altre

La tabella di contingenza $2 \times N$ ha $N-1$ gdl calcolati da $(N-1) \times (2-1)$, poiché i totali marginali sono invariabili

N.B. Evitare frequenze attese inferiori a 5, per non ridurre la potenza del test

ESEMPIO

Effetto di 5 biocidi sulla sopravvivenza di una specie animale :

DISTRIBUZIONE OSSERVATA

	biocida A	biocida B	biocida C	biocida D	biocida E	totale
morti	8	10	14	11	7	50
sopravvissuti	12	6	20	22	10	70
totale	20	16	34	33	17	120

Dai totali marginali e da quello generale si calcola la distribuzione attesa secondo H_0 (le percentuali di animali morti con i 5 biocidi sono uguali)

DISTRIBUZIONE ATTESA SECONDO H_0

	biocida A	biocida B	biocida C	biocida D	biocida E	totale
morti	8,33	6,67	14,17	13,75	7,08	50
sopravvissuti	11,67	9,33	19,83	19,25	9,92	70
totale	20	16	34	33	17	120

Il valore del χ^2 si calcola con la formula generale

$$\chi_{(g.d.l.)}^2 = \sum_{i=1}^{M \cdot N} \frac{(f_i^{oss} - f_i^{att})^2}{f_i^{att}}$$

$$\chi_{(4)}^2 = \frac{(8 - 8,33)^2}{8,33} + \frac{(10 - 6,67)^2}{6,67} + \dots + \frac{(10 - 9,92)^2}{9,92} = 3,9266$$

Il χ^2 è inferiore al valore critico del 5% e pertanto si accetta H_0 : le differenze riscontrate tra valori osservati e valori attesi sono imputabili solo a variazioni casuali di campionamento

Per una tabella di contingenza $M \times N$, il χ^2 può essere utilizzato come

test per l'indipendenza con $(M-1) \cdot (N-1)$ gdl

H_0 : non c'è associazione tra la variabile distribuita per righe e quella per colonna

N.B. Qualora comparissero frequenze attese inferiori a 5, occorrerebbe riunire due o più gruppi di variabili tra loro simili in un'unica categoria

ESEMPIO

- Verificare se in 4 diversi appezzamenti di terreno, con coltivazioni differenti,
- si ha la stessa distribuzione di 5 specie d'insetti (H_0)
- una o più specie sono più facilmente presenti in certe coltivazioni (H_1)

DISTRIBUZIONE OSSERVATA

	specie A	specie B	specie C	specie D	specie E	totale
coltivazione I	12	8	5	15	10	50
coltivazione II	15	10	5	20	8	58
coltivazione III	9	6	10	17	11	53
coltivazione IV	23	12	12	31	17	95
totale	59	36	32	83	46	256

DISTRIBUZIONE ATTESA SECONDO H_0

	specie A	specie B	specie C	specie D	specie E	totale
coltivazione I	11,5	7	6,3	16,2	9	50
coltivazione II	13,4	8,2	7,2	18,8	10,4	58
coltivazione III	12,2	7,5	6,6	17,2	9,5	53
coltivazione IV	21,9	13,3	11,9	30,8	17,1	95
totale	59	36	32	83	46	256

$$\chi^2_{(12)} = \frac{(12 - 11,5)^2}{11,5} + \frac{(8 - 7)^2}{7} + \dots + \frac{(17 - 17,1)^2}{17,1} = 5,5999$$

Il $\chi^2_{(12)}$ non è significativo, dunque in tutte le coltivazioni si ha una presenza equivalente delle 5 specie e **non esiste alcuna associazione tra tipo di coltivazione e specie**

SCOMPOSIZIONE DEI GDL

- si usa quando si vogliono individuare la causa di una deviazione da H_0
- fornisce informazioni dettagliate sugli effetti di ogni gruppo di dati
- è resa possibile dalla proprietà additiva del χ^2 e dei relativi gdl
- comporta la ripartizione di una tabella $M \times N$ in tante tavole 2×2 quanti sono i gdl disponibili

ESEMPIO

(tabella 3×3)

con $3 \cdot 3 = 9$ dati si ha un χ^2 con $(3-1) \cdot (3-1) = 4$ gdl

	TRATT. I	TRATT. II	TRATT. III	Totali
blocco A	a ₁	a ₂	a ₃	n ₁
blocco B	b ₁	b ₂	b ₃	n ₂
blocco C	c ₁	c ₂	c ₃	n ₃
totali	n ₄	n ₅	n ₆	N

Se il χ^2 risulta significativo, emerge il problema di conoscere a quali confronti singoli 2×2 sia da attribuire la differenza

- si possono fare solo 4 confronti
- la somma dei 4 χ^2 con 1 gdl deve risultare uguale al χ^2 complessivo
- la partizione dei 4 gdl è attuata secondo i seguenti confronti 2×2 :

1) Φ

a_1	a_2
b_1	b_2

2) Φ

$(a_1 + a_2)$	a_3
$(b_1 + b_2)$	b_3

3) Φ

$(a_1 + b_1)$	$(a_2 + b_2)$
c_1	c_2

4) Φ

$(a_1 + a_2 + b_1 + b_2)$	$(a_3 + b_3)$
$(c_1 + c_2)$	c_3

TEST DI OMOGENEITA'

I test di indipendenza forniscono implicitamente una misura dell'omogeneità tra le proporzioni e possono quindi servire per valutare se c'è eterogeneità tra le diverse proporzioni a confronto, rispetto ad un valore atteso generale

Con il test χ^2 si può saggiare H_0 per ogni singolo campione : ogni χ^2 con 1 gdl fornisce il grado di scostamento di ciascuna osservazione

Se esistono tante piccole differenze sistematiche e nessuna deviazione molto evidente, nessun test risulterà significativo e solo considerando simultaneamente l'insieme di tutti i dati, si potrà dimostrare uno scostamento non casuale

La somma dei singoli $\chi_{(1)}^2$ risulta più grande della deviazione media, quando le varie percentuali sono tra loro eterogenee, poiché è somma di 2 componenti :

- lo scostamento di ogni campione dal valore medio atteso
- la differenza tra le singole percentuali

Pertanto, sottraendo allo scostamento complessivo tra tutti i gruppi rispetto al valore atteso lo scostamento relativo a tutti i dati insieme, si determina l'eterogeneità tra le varie proporzioni

ESEMPIO

Verificare se, fra alcune popolazioni di vegetali, c'è omogeneità per quanto riguarda la frequenza del gene A, misurato in campioni di piccole dimensioni e valutata, in una data regione, al 22%

Saggiare se 5 campioni provenienti da aree diverse possono essere considerati appartenere alla stessa popolazione

campioni	A osservati	A attesi	non-A osservati	non-A attesi	totale	$\chi_{(1)}^2$
1	12	15,8	60	56,2	72	1,17085
2	15	17,2	63	60,8	78	0,36100
3	8	9,7	36	34,3	44	0,38219
4	17	20,2	75	71,8	92	0,64954
5	23	22,9	81	81,1	104	0,00055
totale	75	85,8	315	304,2	390	1,74287

Il $\chi_{(5)}^2$ determinato dalla somma dei 5 $\chi_{(1)}^2$ relativi a ogni campione, stima :

- la variabilità tra i campioni
- lo scostamento di ognuno di essi dalla frequenza allelica attesa (22%)

$$\chi_{(5)}^2 = 1,170 + 0,361 + 0,382 + 0,649 + 0,00055 = 2,56$$

Nel caso specifico, il valore non è significativo; pertanto i 5 campioni provengono da una stessa popolazione con frequenza del gene A del 22%

Sottraendo al $\chi_{(5)}^2$ il $\chi_{(1)}^2$ calcolato sulle frequenze totali osservate, si ottiene un $\chi_{(4)}^2$ che permette il confronto fra i 5 campioni e la verifica di omogeneità :

$$\chi_{(4)}^2 = \chi_{(5)}^2 - \chi_{(1)}^2 = 2,564 - 1,742 = 0,82$$

Il $\chi_{(4)}^2$ non è significativo e dunque i cinque campioni sono omogenei

Riassumendo, per misurare l'eterogeneità tra i 5 campioni, dopo aver rilevato le frequenze osservate in ogni campione ed in totale, calcolare :

- 1 i valori attesi per 5 campioni e per il totale, sulla base della frequenza generale attesa
- 2 la somma $\chi_{(5)}^2$ data dai χ^2 di ognuno dei cinque campioni
- 3 il $\chi_{(1)}^2$ per la frequenza totale
- 4 la differenza tra $\chi_{(5)}^2$ e $\chi_{(1)}^2$: il $\chi_{(4)}^2$ risultante misura l'eterogeneità tra i cinque campioni

ESEMPI PER TABELLE MxN

1 - Frequenze di tre alleli del marcatore 'ossidasi' in quattro popolazioni naturali di *Fagus sylvatica*

FREQUENZE OSSERVATE

	allele 1	allele 2	allele 3	totale
Abetone	7	244	49	300
Pisanino	8	156	24	188
Pradarena	22	231	31	284
Pradaccio	143	185	116	444
totale	180	816	220	1216

D.: Verificare se le frequenze dei tre alleli sono casuali

R.:

FREQUENZE ATTESE SECONDO H_0

	allele 1	allele 2	allele 3	totale
Abetone	44,4	201,3	54,3	300
Pisanino	27,8	126,2	34,0	188
Pradarena	42,0	190,6	51,4	284
Pradaccio	65,8	297,9	80,3	444
totale	180	816	220	1216

$$\chi^2_{(6)} = \frac{(7 - 44,4)^2}{44,4} + \frac{(244 - 201,3)^2}{201,3} + \frac{(49 - 54,3)^2}{54,3} + \frac{(8 - 27,8)^2}{27,8} + \frac{(156 - 126,2)^2}{126,2} +$$

$$+ \frac{(24 - 34)^2}{34} + \frac{(22 - 42)^2}{42} + \frac{(231 - 190,6)^2}{190,6} + \frac{(31 - 51,4)^2}{51,4} + \frac{(143 - 65,8)^2}{65,8} +$$

$$+ \frac{(185 - 297,9)^2}{297,9} + \frac{(116 - 80,3)^2}{80,3} = 240,571 \quad p > 0,001$$

Esiste una differenza altamente significativa rispetto alla media delle 4 zone :

- all'Abetone e al Pisanino : eccesso dell'allele 2 e carenza alleli 1 e 3
- al Pradaccio e a Pradarena : eccesso alleli 1 e 3 e carenza allele 2

2 - Cloni di *Daphnia magna* sono stati sottoposti a quattro diversi trattamenti alimentari e dopo 39 giorni si è controllato se il tasso di mortalità fosse uguale per i 4 diversi trattamenti

D.: Verificare se le differenze riscontrate sono dovute al caso o sono imputabili al diverso trattamento alimentare

FREQUENZE OSSERVATE

	cloni morti	cloni sopravvissuti	totale
Trattamento I	6	23	29
trattamento II	2	26	28
trattamento III	8	22	30
trattamento IV	3	20	23
totale	19	91	110

R.:

FREQUENZE ATTESE SECONDO H_0

	cloni morti	cloni sopravvissuti	totale
Trattamento I	5	24	29
trattamento II	4,8	23,2	28
trattamento III	5,2	24,8	30
trattamento IV	4,0	19	23
totale	19	91	110

$$\chi^2_{(3)} = \frac{(6-5)^2}{5} + \frac{(23-24)^2}{24} + \frac{(2-4,8)^2}{4,8} + \frac{(26-23,2)^2}{23,2} +$$

$p > 0,25$

$$+ \frac{(8-5,2)^2}{5,2} + \frac{(22-24,8)^2}{24,8} + \frac{(3-4)^2}{4} + \frac{(20-19)^2}{19} = 4,02325$$

La probabilità che sia vera H_0 è superiore al 25%, (molto elevata) e H_0 non può essere rifiutata

3 - Nella tabella sono riportati i risultati di un esperimento sulla schiusa di uova di *Heterocypris incongruens*, mantenute a diverse condizioni di temperatura

FREQUENZE OSSERVATE

	schiuso	non schiuso	totale
16°C	131	32	163
24°C	100	64	164
28°C	90	91	181
totale	320	188	508

D.: Si può affermare che le percentuali di uova schiuse alle 3 diverse temperature sono significativamente differenti, e che le uova mantenute a temperatura inferiore si schiudono con frequenza maggiore ?

R.:

FREQUENZE ATTESE SECONDO H_0

	schiuso	non schiuso	totale
16°C	102,7	60,3	163
24°C	103,3	60,7	164
28°C	114,0	67	181
totale	320	188	508

$$\chi^2_{(2)} = \frac{(131-102,7)^2}{102,7} + \frac{(32-60,3)^2}{60,3} + \frac{(100-103,3)^2}{103,3} +$$

$p > 0,0001$

$$+ \frac{(64-60,7)^2}{60,7} + \frac{(90-114)^2}{114} + \frac{(91-67)^2}{67} = 35,0145$$

Il valore del $\chi^2_{(2)}$ è alto e la probabilità che H_0 sia vera è molto bassa

LA DISTRIBUZIONE **t** DI STUDENT

Oltre alla media μ , anche la varianza σ^2 e, conseguentemente, la deviazione standard σ della popolazione sono ignote; la varianza del campione s^2 rappresenta la stima più logica ed attendibile della varianza della popolazione

Con σ ignota, la distribuzione delle probabilità non è fornita dalla distribuzione normale, bensì è fornita dalla distribuzione del test **t** di Student (pseudonimo di W.S. Gosset)

Per attuare una inferenza sulla media di una popolazione partendo da dati campionari, occorre pertanto considerare sia la variazione di \bar{x} come stima di μ , sia la variazione di s come stima di σ

Con **n** grande (grandi campioni) :

- **s** è la migliore stima di σ (oltre 100 gdl, **s** e σ sono praticamente identici)
- si ha convergenza dei valori della distribuzione **t** verso la distribuzione normale **z**

Con **n** piccolo (piccoli campioni) :

- la differenza tra **s** e σ è rilevante
- si deve utilizzare il test **t**

Gosset, usando campioni ridotti (**n** piccolo) studiò lo scarto tra la media dei campioni e la media dell'universo in rapporto all'ERRORE STANDARD e derivò una distribuzione ottenuta dalle variazioni determinate dal rapporto:

$t = \frac{\text{differenza fra due medie campionarie}}{\text{errore standard della differenza di due medie campionarie}}$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{s_d}{\sqrt{n}}}$$

Principale differenza tra la distribuzione normale e la distribuzione **t** :

- la distribuzione normale considera la variazione di campionamento solo della media
- la distribuzione **t** considera anche la variazione di campionamento della deviazione standard

Condizione di validità della distribuzione **t** :

- distribuzione dei dati normale
- osservazioni raccolte in modo indipendente

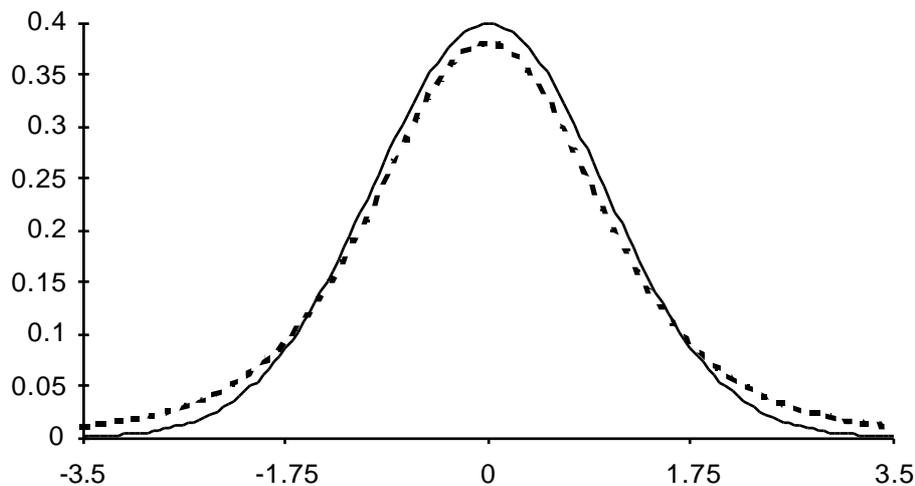
La distribuzione **t** è

La distribuzione **t** è :

- il rapporto tra la differenza della media campionaria \bar{x} con la media attesa μ ed il

suo errore standard $t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

- di area unitaria e di forma simmetrica (come la gaussiana degli Z)
- una famiglia di distribuzioni (una distribuzione per ogni gdl) a differenza di quanto avviene per la gaussiana
- coincidente con la gaussiana (cfr. le rispettive tabelle) per infiniti gdl (in pratica per $n > 100$)
- sempre più dispersa (platicurtica) al diminuire dei gdl
- ROBUSTA, cioè valida anche per distribuzioni di dati con marcate deviazioni dalla normalità, infatti UN TEST È ROBUSTO QUANDO I RISULTATI POSSONO ESSERE ACCETTATI ANCHE SE NON SI VERIFICANO RIGOROSAMENTE TUTTE LE ASSUNZIONI DI VALIDITÀ



Distribuzione normale standardizzata (linea continua) e distribuzione **t** per 65 gdl

Abitualmente nei testi di statistica sono riportate due differenti tabelle di valori critici della distribuzione **t** : quella per test unilaterali e quella per test bilaterali

In queste tabelle, la parte superiore di ogni colonna indica l'area sottesa dalle rispettive code della distribuzione, mentre le righe si riferiscono ai gdl

I valori critici per l'area in una coda al rischio α coincidono con quelli del rischio 2α nella distribuzione a due code (per esempio, i valori per $\alpha=0,05$ coincidono con la colonna di $\alpha=0,025$ nella tabella per test ad una coda)

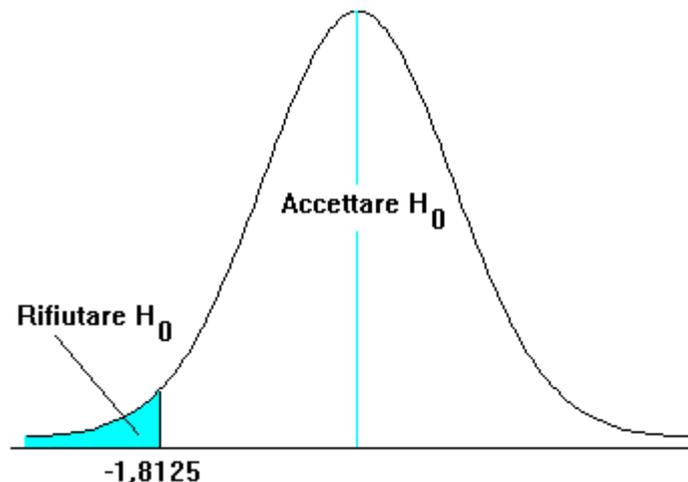
test	α	valore critico per 10 gdl
unilaterale	0,05	1,8125
bilaterale	0,05 (somma di $\alpha=0,025$ nelle due code)	2,228

ESEMPIO

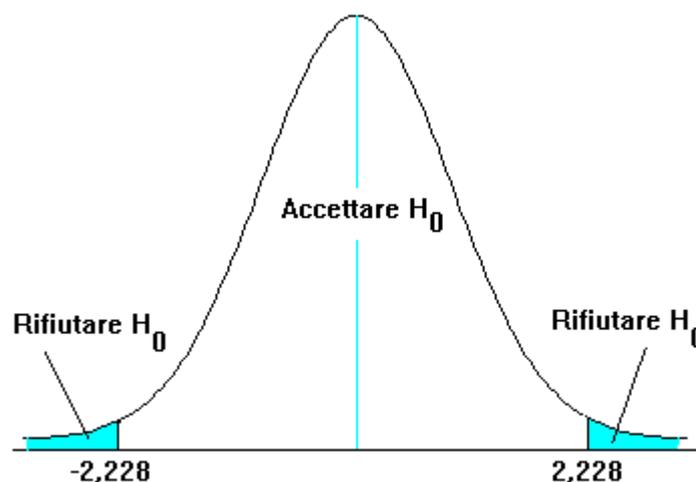
Nel confrontare gli effetti di due diversi inquinanti, in un test bilaterale si inferisce solo sulle due medie : effetti uguali (H_0) o effetti differenti (H_1) ?

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

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



Test unilaterale per la differenza appaiata al livello di significatività del 5% con 10 gdl



Test bilaterale per la differenza appaiata al livello di significatività del 5% con 10 gdl

INTERVALLO di CONFIDENZA DI UNA MEDIA CON σ NOTA

Volendo conoscere il valore più probabile di un parametro incognito, la statistica inferenziale fornisce due valori che determinano l'INTERVALLO DI CONFIDENZA (o LIMITI FIDUCIALI) entro cui si colloca il valore del parametro secondo la probabilità scelta

Specificare i limiti fiduciali è solamente un modo alternativo di inferire circa i parametri di una popolazione, sulla base di osservazioni campionarie

I limiti fiduciali della media della popolazione sono stimati dalla distribuzione normale standardizzata :

- il 95% dell'area sottesa dalla curva si trova tra -1,96 e +1,96 dell'ascissa
[$P(-1,96 \leq Z \leq +1,96) = 0,95$]

- il 99% dell'area sottesa dalla curva si trova tra -2,58 e +2,58 dell'ascissa
[$P(-2,58 \leq Z \leq +2,58) = 0,99$]

Così come σ valuta la dispersione di campionamento delle osservazioni, l'ERRORE STANDARD (ES) valuta la dispersione delle medie campionarie :

$$ES = \frac{\sigma}{\sqrt{n}}$$

La distribuzione di campionamento di medie con media μ ed $ES = \frac{\sigma}{\sqrt{n}}$ diventa

$$P\left(-Z \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq +Z\right) = P(z) \text{ e può essere usata per determinare i limiti fiduciali :}$$

- al 95% diventa $P\left(\bar{x} - 1,96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$

- al 99% sostituire 1,96 con 2,58

ESEMPIO

Da una popolazione con $\sigma = 3$ è stato estratto un campione di 10 dati con $m=25$

D.: Calcolare l'intervallo di confidenza alla probabilità del 99%

$$\underline{\underline{R.:}} \quad 25 \mp 2,58 \cdot \frac{3}{\sqrt{10}} = 25 \mp 2,58 \cdot 0,9487 = 25 \mp 2,45 = \begin{cases} 22,55 \\ 27,45 \end{cases}$$

Secondo le informazioni fornite da una campione di 10 misure con $m=25$, con probabilità 99% μ si trova nell'intervallo compreso tra 22,55 e 27,45 [... resta la probabilità dell'1% che μ si trovi fuori da questo intervallo ...]

INTERVALLO DI CONFIDENZA DI UNA MEDIA CON σ IGNOTA

Per stimare sia la varianza s^2 che la media \bar{x} dai dati campionari, la standardizzazione è ottenuta mediante :

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Da essa si deriva l'intervallo di confidenza di $\mu = \bar{x} \pm t_{\frac{\alpha}{2}; n-1} \cdot \frac{s}{\sqrt{n}}$

$t_{\frac{\alpha}{2}; n-1}$ valore della distribuzione per $n-1$ gdl al rischio $\frac{\alpha}{2}$

Un aumento del numero di dati campionari agisce sulla riduzione dell'intervallo di confidenza sia attraverso il valore del t , che diminuisce al crescere dei gdl, sia mediante la riduzione dell'errore standard $\left(\frac{s}{\sqrt{n}} \right)$

Con campioni provenienti da **popolazioni limitate** (il campione ne rappresenta una frazione non trascurabile), per ridurre l'errore standard nel calcolo dell'intervallo di confidenza si aggiunge il

FATTORE DI CORREZIONE PER LE POPOLAZIONI FINITE $\sqrt{\frac{(N-n)}{(N-1)}}$

N : dimensione della popolazione; **n** : dimensione del campione

ESEMPI

[1] Stimare, con probabilità 95%, l'intervallo di confidenza dell'altezza media di una varietà di pomodoro, attraverso esemplari alti 22, 25, 21, 23, 24, 25, 21 pollici

$$\bar{x} = 23 \quad s = 1,732 \quad t_{0,025;6} = 2,447 \quad n = 7$$

Il valore di t può essere scelto nella distribuzione ad una coda (con $\alpha=0,025$) o nella distribuzione a due code (con $\alpha=0,05$)

$$\mu = 23 \pm 2,447 \cdot \frac{1,732}{\sqrt{7}} = 23 \pm 1,602$$

I limiti risultano $l_1 = 21,398$ $l_2 = 24,602$

[2] Stimare con probabilità 99% l'intervallo di confidenza della lunghezza media di un campione di 13 individui del parassita *Aphis fabae*

1,21 1,39 1,21 1,21 1,21 1,21 1,20 1,18 1,23 1,21 1,23 1,24 1,33 mm

$$\bar{x} = 1,235 \quad s = 0,059 \quad t_{0,005;12} = 3,055 \quad n = 13$$

$$\mu = 1,235 \pm 3,055 \frac{0,059}{\sqrt{12}} = 1,235 \pm 0,05203$$

I limiti risultano $l_1 = 1,175$ $l_2 = 1,287$

[3a] In un campione di **tre** individui con altezze 1,70 1,80 1,90 m calcolare l'intervallo di confidenza della media al 95%

$$\bar{x} = 1,80 \quad s = 0,10 \quad t_{0,025;2} = 4,303 \quad n = 3$$

$$\mu = 1,80 \pm 4,303 \frac{0,10}{\sqrt{3}} = 1,80 \pm 0,2484$$

I limiti risultano $l_1 = 1,552$ $l_2 = 2,048$

[3b] In un campione di **sei** individui con altezze 1,70 1,80 1,90 1,70 1,80 1,90 m calcolare l'intervallo di confidenza della media al 95%

$$\bar{x} = 1,80 \quad s = 0,089 \quad t_{0,025;5} = 2,571 \quad n = 6$$

$$\mu = 1,80 \pm 2,571 \frac{0,089}{\sqrt{6}} = 1,80 \pm 0,0934$$

I limiti risultano $l_1 = 1,7066$ $l_2 = 1,8934$

Il significato di intervallo di confidenza ...

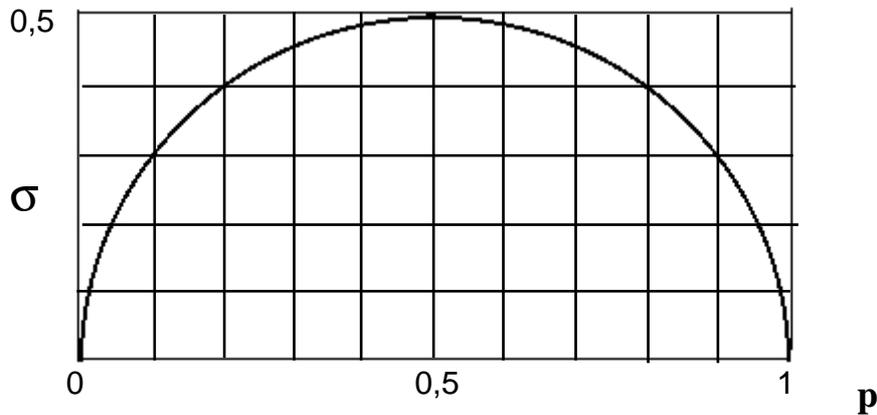
NON è : μ (o σ^2) hanno $p=1-\alpha$ di essere compresa nell'intervallo stimato, perchè il parametro della popolazione vi è o non vi è compreso

MA è : campionando 100 volte dalla stessa popolazione, si stima con $p=1-\alpha$ un intervallo che $(1-\alpha) \cdot 100$ volte conterrà μ (o σ^2) mentre $\alpha \cdot 100$ volte non la conterrà

INTERVALLO DI CONFIDENZA DI UNA PROPORZIONE

Si adotta l'approssimazione della normale alla binomiale

In una proporzione, il valore di σ è completamente determinato dal valore della media p , infatti con n costante $\sigma = \sqrt{p \cdot (1-p)}$



La σ di una proporzione si approssima a zero quando p è molto piccolo o molto grande e presenta valore massimo quando p è prossimo al valore centrale 0,5

L'intervallo di confidenza di una percentuale è dato da $p \pm Z \cdot \sqrt{\frac{p \cdot q}{n}}$

ESEMPIO

In un campione di 80 fumatori, il 35% ha presentato sintomi di polmonite

D.:

- Calcolare i limiti fiduciali della media al 95% e al 99% nella popolazione dei fumatori con sintomi di polmonite
- Calcolare gli stessi limiti fiduciali (95% e 99%) partendo da un campione di 100 fumatori anziché 80

R.: Con un campione di 80 fumatori si ha :

$$\text{Per il 95\% : } 0,35 \pm 1,96 \cdot \sqrt{\frac{0,35 \cdot 0,65}{80}} = 0,35 \pm 0,1045 = \begin{cases} 0,2455 \\ 0,4545 \end{cases}$$

$$\text{Per il 99\% : } 0,35 \pm 2,58 \cdot \sqrt{\frac{0,35 \cdot 0,65}{80}} = 0,35 \pm 0,1376 = \begin{cases} 0,2124 \\ 0,4876 \end{cases}$$

Con un campione di 100 fumatori si avrebbe :

$$\text{Per il 95\% : } 0,35 \pm 1,96 \cdot \sqrt{\frac{0,35 \cdot 0,65}{100}} = 0,35 \pm 0,9349 = \begin{cases} 0,2565 \\ 0,4435 \end{cases}$$

$$\text{Per il 99\% : } 0,35 \pm 2,58 \cdot \sqrt{\frac{0,35 \cdot 0,65}{100}} = 0,35 \pm 0,1231 = \begin{cases} 0,2269 \\ 0,4731 \end{cases}$$

N.B. Con un campione di 100 individui gli intervalli sono più stretti rispetto a quelli prodotti dal campione di 80 individui

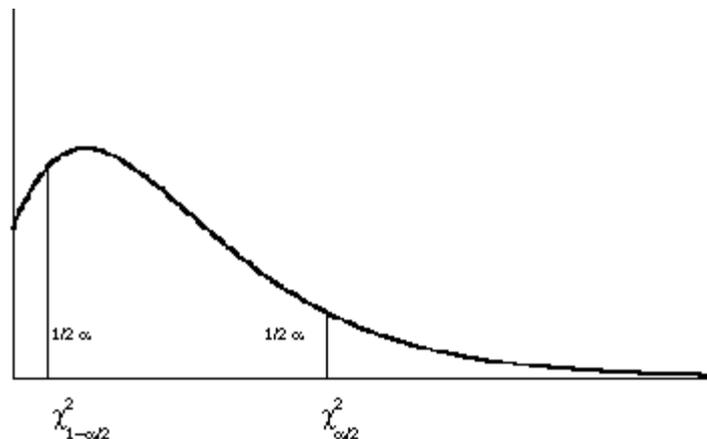
INTERVALLO DI CONFIDENZA DI UNA VARIANZA

E' possibile stimare la varianza dellapopolazione σ^2 partendo dai dati campionari, per verificare la precisione di uno strumento di misura, o per confrontare genotipi identici cresciuti in situazioni ambientali differenti

In popolazioni normalmente distribuite, il calcolo dell'intervallo di confidenza può essere ottenuto con la distribuzione χ^2 , poichè

$$\chi_{(n-1)}^2 = \frac{s^2 \cdot (n-1)}{\sigma^2} \quad \text{o in modo equivalente} \quad \frac{\sum (x - \bar{x})^2}{\sigma^2}$$

Per calcolare l'intervallo di confidenza a probabilità $p=1-\alpha$, occorre individuare i valori di χ^2 che escludono $\alpha/2$ da ciascuna delle due parti della distribuzione
 In una distribuzione non simmetrica è più laborioso scegliere i due valori di χ^2 che permettono di dividere equamente α tra le due code della distribuzione



Per un rischio $\alpha=0,05$, si scelgono i due valori di α tali che uno escluda 2,5% a sinistra e l'altro escluda 2,5% a destra

Intervallo di confidenza di :

$$\sigma^2 \rightarrow \frac{s^2 \cdot (n-1)}{\chi_{1-\frac{\alpha}{2}}^2} > \sigma^2 > \frac{s^2 \cdot (n-1)}{\chi_{\frac{\alpha}{2}}^2}$$

$$\sigma \rightarrow \sqrt{\frac{s^2(n-1)}{\chi_{1-\frac{\alpha}{2}}^2}} > \sigma > \sqrt{\frac{s^2(n-1)}{\chi_{\frac{\alpha}{2}}^2}}$$

N.B.

Requisito essenziale è che i dati siano distribuiti normalmente; questa assunzione è tanto più importante e difficile da rispettare quando n è piccolo

Quando la normalità della distribuzione campionaria non può essere dimostrata, i risultati del calcolo dell'intervallo fiduciale di una varianza vanno applicati con cautela

ESEMPIO

Determinare con $p=99\%$ l'intervallo di confidenza della varianza di composti clorurati totali ($\mu\text{g}/\text{m}^3$ a 0°C e 1013 mbar) nell'atmosfera di una metropoli sulla base di 16 prelievi con $s^2 = 8210,67$

$$\chi_{0,995; 15}^2 = 4,605 \quad \chi_{0,005; 15}^2 = 32,85$$

$$\frac{8210,67 \cdot 15}{32,85} < \sigma^2 < \frac{8210,67 \cdot 15}{4,605} \quad 3748,980 < \sigma^2 < 26743,540 = 15$$

CONFRONTO TRA DUE MEDIE

Le situazioni più ricorrenti non riguardano il confronto tra media campionaria e media della popolazione, bensì il confronto tra due medie campionarie

$H_0 : \mu_1 = \mu_2$ (oppure $H_0 : \mu_1 - \mu_2 = 0$), ovvero μ_1 e μ_2 sono :

- estratte dalla stessa popolazione
- diverse, nelle medie campionarie \bar{x}_1 e \bar{x}_2 , soltanto per differenze casuali
- identiche

Attraverso il test **t** si determina la probabilità **p** di ottenere differenze maggiori di quelle sperimentalmente osservate :

- se **p** risulta piccola (convenzionalmente $p < 5\%$), si rifiuta H_0
- se **p** risulta grande, si accetta H_0
- ➔ si inferisce che esiste una ragionevole evidenza per dubitare che sia vera, cioè esiste una differenza reale tra le due medie che appartengono a popolazioni diverse

N.B. Nel confronto tra un campione di soggetti sottoposti a trattamento ed un campione di soggetti:

- utilizzati come controllo : test unilaterale (test a una coda)
- sottoposti ad un altro trattamento : test bilaterale (test a due code)

• La direzionalità del confronto è insita nella natura dell'esperimento, ma va esplicitata, poichè da essa deriva la distribuzione delle probabilità alle quali è possibile rifiutare H_0 :

- test unilaterale : per dimostrare se una media è maggiore dell'altra, escludendo a priori che essa possa essere minore (esclude a priori che il confronto possa fornire una parte delle risposte teoricamente possibili, in quanto prive di significato nel caso specifico)
- test bilaterale : per dimostrare se una media è maggiore dell'altra, ma senza escludere a priori che essa possa essere minore

TEST t PER DUE CAMPIONI DIPENDENTI (DATI APPAIATI)

Caratteristica distintiva :

- poter accoppiare ogni osservazione di un campione con una e una sola osservazione dell'altro campione
- necessariamente i due gruppi hanno sempre lo stesso numero di dati

Scopo principale dell'appaiamento dei dati:

- creare il massimo di omogeneità entro ogni coppia
- creare il massimo di eterogeneità tra le coppie

Situazione A : AUTO-ACCOPPIAMENTO (dati auto-appaiati)

ogni soggetto serve come controllo di se stesso e i dati vengono ricavati dagli stessi individui in momenti diversi

Per esempio:

- confronto tra i livelli di pressione rilevati nello stesso gruppo di individui sia in condizioni normali che dopo uno stress
- confronti prima-e-dopo riferiti agli stessi individui

Situazione B : OSSERVAZIONI NATURALMENTE APPAIATE

non sono tratte dagli stessi individui, ma da coppie di individui scelti appositamente

Per esempio:

- misure rilevate in coppie di animali tratti dalla stessa nidiata e sottoposti a situazioni ambientali differenti
- confronto tra il comportamento materno e paterno nella cura alla prole, quando si dispone di dati relativi a coppie

Situazione C : APPAIAMENTO ARTIFICIALE

- studi di confronto con molte variabili, dove si rileva un parametro in una situazione ambientale compromessa e lo stesso parametro nella situazione naturale

Il confronto tra trattamento e controllo sugli stessi individui o tra situazioni simili si propone di eliminare alcune sorgenti di variabilità che potrebbero nascondere le reali differenze tra le due serie di misure: esaminare le differenze fra due misurazioni riduce l'effetto della variabilità intrinseca degli individui

Tecnicamente il confronto è semplice: l'analisi è ridotta alla sola serie risultante dalle differenze tra gli elementi di ciascuna coppia

H_0 : la media dell'universo delle differenze è 0 ($\delta = 0$)

H_1 è diversa nei due tipi di test :

- test bilaterale : la differenza media non è 0 ($\delta \neq 0$)

- test unilaterale : la differenza è maggiore oppure minore di 0 ($\delta > 0$; $\delta < 0$)

Il test della differenza media è $t_{n-1} = \frac{\bar{d} - d}{\frac{s}{\sqrt{n}}}$

\bar{d} media della colonna delle differenze,

δ differenza attesa, spesso ma non necessariamente 0

s deviazione standard calcolata sulla colonna delle differenze

n n° di paia di dati, corrispondente al numero delle differenze

$\frac{s}{\sqrt{n}}$ errore standard della media delle differenze

ESEMPI

[1] Ad 8 individui adulti è stata misurata la pressione (a) in condizioni normali e (b) dopo l'apprendimento di una notizia capace renderli ansiosi

Individuo	normale	ansia	differenza d
A	140	180	40
B	145	175	30
C	140	165	25
D	160	195	35
E	150	180	30
F	145	180	35
G	160	200	40
H	145	190	45

d media = 35

D.:

Verificare se gli individui in condizioni di ansia manifestano un aumento della pressione sistolica sanguigna mediamente superiore ai 30 mm Hg

La formulazione del problema fa capire che si tratta di un test ad una coda, con

$H_0 : \delta = 30$ e $H_1 : \delta > 30$

R.:

$$\bar{d} = \frac{280}{8} = 35 \quad s = \sqrt{\frac{300}{7}} = 6,55 \quad n = 8$$

$$t_7 = \frac{35 - 30}{\frac{6,55}{\sqrt{8}}} = 2,16$$

Valore critico per 7 gdl ; test ad una coda ; $\alpha = 0,05$ $t = 1,895$

Il valore calcolato è superiore a quello tabulato e quindi la probabilità che la differenza tra media osservata e media attesa sia casuale è $\alpha < 0,05$

➔ si rifiuta H_0 e si accetta H_1 (l'aumento di pressione in condizioni di stress supera 30 mm Hg)

[2] Con i dati dell'esempio precedente ci si sarebbe potuti anche chiedere, più semplicemente, se in condizioni di stress la pressione subisce un aumento

Anche in questo caso si tratta di un test ad una coda, ma varia la differenza attesa in $H_0: \delta = 0$ e $H_1: \delta > 0$

$$t_7 = \frac{35 - 0}{\frac{6,55}{\sqrt{8}}} = 15,15$$

Il t calcolato è molto superiore a quello tabulato sia per $\alpha = 0,01$ che per $\alpha = 0,005$ per cui la differenza è altamente significativa

➔ si rifiuta H_0 con un rischio bassissimo di commettere un errore di I^o tipo

[3] Un gruppo di 10 cavie è stato sottoposto ad una dieta diversa per cui ogni soggetto è stato pesato prima e dopo la nuova dieta

cavia	prima	dopo	differenza d	$(d - \bar{d})^2$
1	180	190	10	1
2	175	170	- 5	196
3	150	175	25	256
4	158	164	6	9
5	174	185	9	0
6	187	184	- 3	144
7	172	185	13	16
8	157	168	11	4
9	164	180	16	49
10	165	173	8	1

D.: La nuova dieta determina una differenza ponderale ?

Si tratta di un test a due code, con $H_0: \delta = 0$ $H_1: \delta \neq 0$

R.: $\bar{d} = \frac{90}{10} = 9$ $s = \sqrt{\frac{676}{9}} = 8,66$ $n = 10$

Per un test a due code il valore critico della distribuzione per 9 gdl e $\alpha = 0,05$ è $t = 2,262$

Il valore calcolato è superiore al valore critico e quindi la probabilità che la differenza riscontrata sia casuale è $\alpha < 0,05$ | \rightarrow si rifiuta H_0 e si accetta H_1 (la nuova dieta determina una differenza ponderale nelle cavie)

Si possono ottenere le medesime conclusioni attraverso la **STIMA DELL'INTERVALLO FIDUCIALE DELLA DIFFERENZA MEDIA** che per due campioni dipendenti, in analogia a quanto già visto, è

$$d = \bar{d} \pm t_{\frac{\alpha}{2}; n-1} \cdot \frac{s}{\sqrt{n}}$$

d per $\alpha = 0,05$ con $t_{9; 0,025}$ diventa

$$9 \pm 2,262 \cdot \frac{7,57}{\sqrt{10}} = 9 \pm 5,42$$

$$d_1 = 3,58 \quad d_2 = 14,42$$

La differenza media campionaria è $\bar{d} = 9$

L'intervallo entro cui, i con $\alpha = 0,05$, si trova μ (media reale della popolazione) è compreso tra 3,58 e 14,42

Si osservi che δ , espresso in termini di H_0 ($H_0 : \delta = 0$), risulta esterno all'intervallo di confidenza calcolato e quindi si discosta significativamente dal valore medio sperimentale

TEST **t** PER CAMPIONI INDIPENDENTI (DATI NON APPAIATI)

In molti casi non è fattibile o conveniente formare due campioni dipendenti, poiché non si possono misurare gli effetti di due differenti trattamenti sugli stessi individui :

- misure di accrescimento somatico alla stessa età in animali o piante sottoposte a condizioni ambientali differenti
- confronto tra parametri chimici, fisici, biologici di ambienti naturali

Due gruppi di osservazioni ottenute in modo indipendente hanno il vantaggio di:

- potere avere un numero differente di osservazioni ($n_1 \neq n_2$)
- essere più facilmente espressivi della variabilità casuale
- consentire i calcoli direttamente sulle due serie di osservazioni
(con i due campioni dipendenti i calcoli venivano effettuati sulla sola
colonna delle differenze)

Il test **t** pone la stessa domanda dei dati appaiati, ma la forma è diversa :

$$t_{n_1+n_2-2} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

dove :

- \bar{x}_1 e \bar{x}_2 medie dei due campioni
- μ_1 e μ_2 medie attese (la loro differenza è il valore atteso in H_0)
- n_1 e n_2 n° di osservazioni nei due campioni
- s_p^2 varianza associata (*POOLED*) dei due gruppi :

rapporto tra la somma delle due devianze e la somma dei rispettivi gdl (il procedimento è indispensabile quando $n_1 \neq n_2$)

$$H_0 : \mu_1 = \mu_2 \quad \text{oppure} \quad \mu_1 - \mu_2 = 0$$

$$H_1 \text{ per un test ad una coda : } H_1 : \mu_1 > \mu_2 \quad \text{oppure} \quad \mu_1 < \mu_2$$

$$[\text{o anche } H_1 : \mu_1 - \mu_2 > 0 \quad \text{oppure} \quad \mu_1 - \mu_2 < 0]$$

$$H_1 \text{ per un test a due code : } H_1 : \mu_1 \neq \mu_2 \quad \text{oppure} \quad \mu_1 - \mu_2 \neq 0$$

Condizioni di validità del test **t** :

- dati distribuiti normalmente (questa ipotesi di normalità può essere, sebbene non marcatamente, violata senza gravi effetti sulla potenza del test)
- osservazioni raccolte in modo indipendente (per due campioni dipendenti)
- varianze statisticamente uguali (per calcolare S^2 POOLED) (l'eguaglianza delle varianze delle due popolazioni indipendenti deve essere rispettata)

Se i dati delle due popolazioni sono distribuiti normalmente, il rapporto tra le due varianze si avvicina alla distribuzione F

La verifica dell'ipotesi $H_0: s_1^2 = s_2^2$ $H_1: s_1^2 > s_2^2$
 utilizza il rapporto $F_{(n_1-1); (n_2-1)} = \frac{s_1^2}{s_2^2}$

s_1^2 e s_2^2 varianza maggiore e varianza minore

n_1 e n_2 n° dati del gruppo a varianza maggiore e a varianza minore

I valori critici della distribuzione F dipendono dai gdl del numeratore, riportati nella prima riga della tabella, e da quelli del denominatore, riportati nella prima colonna

N.B. Se le varianze risultano statisticamente differenti, si ricorre a test di statistica non parametrica come l'approssimazione di Cochran o a test di statistica non parametrica per due campioni indipendenti

Intervallo fiduciale della differenza tra le due medie campionarie $(\bar{x}_1 - \bar{x}_2)$ con varianze statisticamente uguali :

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t\left(\frac{\alpha}{2}; (n_1+n_2-2)\right) \cdot s_p \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{gdl : } n_1+n_2-2$$

$$es_d = \sqrt{s_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

ESEMPI

[1] Saggiare se la concentrazione algale influisce positivamente sulla crescita (valori in mm) di *Daphnia magna*.

In laboratorio si sono allevati 40 individui dello stesso ceppo:

- 20 in una soluzione con concentrazione algale 120.000 cellule / ml
- 20 in una soluzione con concentrazione algale 24.000 celle / ml

120.000/ml(x 1)	24.000/ml (x2)
4,290	3,120
3,900	3,112
3,783	3,120
3,900	2,847
4,095	3,081
4,056	3,042
4,173	3,042
4,095	3,198
4,095	3,081
4,056	2,964
3,939	3,120
3,978	2,964
4,017	3,003
4,251	3,081
4,017	3,042
3,900	2,925
4,095	3,198
4,173	3,120
3,978	2,964
4,095	3,003

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 > \mu_2$$

	x1	x2
n	20	20
Media \bar{x}	4,0443	3,04335
Devianza SQ	0,30075	0,15326
Varianza s^2	0,015828	0,008066

Controllare se le due varianze, attraverso il rapporto fra quella maggiore e quella minore, non sono statisticamente diverse :

$$\frac{0,015828}{0,008066} = 1,962$$

e confrontare il risultato con il valore critico, per $\alpha = 0,05$, $F_{(20-1);(20-1)} = 2,16$

Essendo $1,962 < 2,16$ le due varianze sono statisticamente uguali, e si possono quindi confrontare le due medie

$$s_p^2 = \frac{0,30075 + 0,15326}{20 - 1 + 20 - 1} = \frac{0,45401}{38} = 0,01194$$

Errore standard della differenza fra medie :

$$es_d = \sqrt{0,01198 \cdot \left(\frac{1}{20} + \frac{1}{20} \right)} = 0,034554$$

$$t_{20+20-2} = \frac{4,0443 - 3,04355}{0,034554} = 29,157$$

Si tratta di test ad una coda poiché interessa valutare solo se la maggiore concentrazione algale produce una maggiore crescita di *Daphnia*

Valore critico per $\alpha = 0,01$ e 38 gdl : $t = 2,329$ [$\ll 29,157$]

➔ La maggior concentrazione algale influisce in modo altamente significativo sulla crescita di *Daphnia*

Il calcolo dell'intervallo fiduciale della differenza fra le due medie è un modo alternativo per verificare H_0 :

$$\text{per } \alpha = 0,05 \quad \rightarrow (\bar{x}_1 - \bar{x}_2) \pm t_{0,05 ; (n_1+n_2-2)} \cdot es_d = 1,00095 \pm 1,686 \cdot 0,034554$$

$$l_1 = 0,94269 \quad l_2 = 1,059208$$

$$\text{per } \alpha = 0,01 \quad \rightarrow (\bar{x}_1 - \bar{x}_2) \pm t_{0,05 ; (n_1+n_2-2)} \cdot es_d = 1,00095 \pm 2,429 \cdot 0,034554$$

$$l_1 = 0,91701 \quad l_2 = 1,08488$$

[2] Si è misurata la produzione di muffe (in termini di tempo trascorso prima della loro comparsa) in due formaggi da tavola di composizione simile :

- 12 trattati con polifosfati durante il confezionamento
- 13 trattati con derivati dell'acido salicilico

D : La differenza media osservata dello sviluppo di colonie nei due gruppi di formaggi è statisticamente significativa ?

polifosfati	ac. salicilico
x₁	x₂
7,94	7,30
8,03	7,26
8,18	6,82
8,03	7,08
8,19	7,13
8,01	7,37
8,16	7,42
8,16	7,16
8,18	6,89
8,29	6,96
7,94	7,13
8,29	7,08
	7,17

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

	x₁	x₂
n	12	13
Media \bar{x}	8,117	7,136
Devianza SQ	0,16656	0,37690
Varianza s^2	0,015	0,0314

Controllare che le due varianze non siano statisticamente diverse $\frac{0,0314}{0,015} = 2,093$

Valore critico per $\alpha=5\%$ $F_{(13-1);(12-1)} = 2,79 > 2,093$ e dunque le due varianze sono statisticamente uguali: si possono quindi confrontare le due medie

$$s_p^2 = \frac{0,16656 + 0,37690}{12 - 1 + 13 - 1} = \frac{0,54346}{23} = 0,02362$$

Errore standard della differenza fra le medie: $es_d = \sqrt{0,02362 \cdot \left(\frac{1}{12} + \frac{1}{13}\right)} = 0,06152$

$$t_{12+13-2} = \frac{8,117 - 7,136}{0,06152} = 15,946$$

Si tratta di test a due code poichè interessa valutare la significatività della differenza fra le medie dei tempi-muffa sui due gruppi di formaggi

Valore critico per $\alpha = 0,01$ e 23 gdl : $t = 2,807$ [$\ll 15,946$]

➔ I due tipi di formaggio hanno una resistenza allo sviluppo di muffe statisticamente molto diverso

Intervallo fiduciale della differenza fra le due medie :

$$\begin{aligned} \text{per } \alpha = 0,05 \quad \text{-->} \quad & (\bar{x}_1 - \bar{x}_2) \pm t_{0,05 ; (n_1+n_2-2)} \cdot es_d = 0,981 \pm 2,069 \cdot 0,06152 \\ & l_1 = 0,85083 \quad \quad \quad l_2 = 1,11116 \end{aligned}$$

$$\begin{aligned} \text{per } \alpha = 0,01 \quad \text{-->} \quad & (\bar{x}_1 - \bar{x}_2) \pm t_{0,001 ; (n_1+n_2-2)} \cdot es_d = 0,981 \pm 2,807 \cdot 0,06152 \\ & l_1 = 0,80441 \quad \quad \quad l_2 = 1,15758 \end{aligned}$$

DIMENSIONI DEL CAMPIONE

Una domanda che spesso si pone al ricercatore è di quale dimensione, cioè di quante osservazioni, deve essere composto il campione

Il test **t** per un campione fornisce già, se si analizzano i valori critici all'aumentare dei gdl, una prima risposta: alla probabilità di 0,05 per un test a due code, il valore di **t** da 12,7 per 1 gdl scende a 4,3 per 2 gdl; poi a 3,1 per 3 gdl e a 2,7 per 4 gdl. Successivamente, il valore di **t** diminuisce molto più lentamente all'aumentare del numero di dati. Rispetto a due soli dati (un gdl), un campione di 4-6 dati permette di rendere significativa una differenza nettamente minore: quattro dati (tre gdl) permettono di rendere significativa una differenza almeno quattro volte più piccola di quanto sia possibile con due soli dati (un gdl)

Per ottenere indicazioni meno vaghe, occorre conoscere alcune informazioni indispensabili, che la stessa formula per il test **t** indica:

- il valore della differenza minima di cui si intende saggiare la significatività
- la varianza del fenomeno (σ^2)
- il livello di significatività (α)

Quando è **noto** σ , si ricorre alla distribuzione normale $z = \frac{\bar{d}}{\frac{\sigma}{\sqrt{n}}}$ dalla quale si può

ricavare $n = \frac{z^2 \cdot \sigma^2}{\bar{d}^2}$

ESEMPI

[1] I limiti di legge di una sostanza inquinante A sono fissati a 50 mg / litro; è dimostrato che la strumentazione utilizzata ha una varianza (σ^2) uguale a 80

Quante osservazioni occorrono per dimostrare che la concentrazione della sostanza A è significativamente maggiore - alla probabilità $\alpha = 0,05$ - se essa è presente con media doppia (100 mg / litro) rispetto ai limiti definiti della norma di legge ?

$$z_{0,05} = 1,645 \quad \sigma^2 = 80 \quad \bar{d} = 50$$

Si richiede un test ad una coda $n = \frac{(1,645)^2 \cdot (80)^2}{(50)^2} = \frac{2,7060 \cdot 6400}{2500} = 6,92$

Con tale risultato si deduce che servono almeno 7 osservazioni

[2] Il primo esercizio era fondato su un test ad una coda. Se si fosse trattato di un test a due code, nel quale veniva richiesto di dimostrare una differenza significativa tra una media di 50 e una di 100, con la stessa varianza e alla stessa probabilità, occorre scegliere un valore di $z = 0,025$ sui due lati :

$$z_{0,05} = 1,96 \quad \sigma^2 = 80 \quad \bar{d} = 50$$

$$n = \frac{(1,96)^2 \cdot (80)^2}{(50)^2} = \frac{3,8416 \cdot 6400}{2500} = 9,83$$

Per un test a due code, servirebbero dunque almeno 10 osservazioni

(Si sottolinea la maggiore potenza del test ad una coda: rispetto al test ad una coda quello a due code in questo caso ha una potenza di $7 / 10 = 0,7$ ovvero del 70%)

[3] Un secondo ricercatore dispone di una strumentazione migliore, che nella misurazione dimostra una varianza $\sigma^2 = 60$

Quante osservazioni deve effettuare, per dimostrare che rispetto ad un valore medio di 50 è significativamente maggiore alla probabilità 0,05 una media di 75 mg/l ?

E' un test ad una coda, dove $z_{0,05} = 1,645$ $\sigma^2 = 60$ $\bar{d} = 25$

$$n = \frac{(1,645)^2 \cdot (60)^2}{(25)^2} = \frac{2,7060 \cdot 3600}{625} = 15,58$$

Occorrono dunque almeno 16 misurazioni

[4] Con i dati del secondo esercizio, quante osservazioni occorrono per dimostrare una differenza significativa per un test a 2 code alla probabilità 0,01 ?

$$z_{0,01} = 2,58 \quad \sigma^2 = 80 \quad \bar{d} = 50$$

$$n = \frac{(2,58)^2 \cdot (80)^2}{(50)^2} = \frac{6,6564 \cdot 6400}{2500} = 17,04$$

Non servono almeno 10 come nell'esercizio 1, ma almeno 18 dati

Nel caso di **frequenze relative** (percentuali), la formula per verificare la significatività di una differenza è uguale alla precedente, ricordando che σ^2 è uguale a **p(1-p)**, essendo totalmente determinato dal valore medio

$$z = \frac{\bar{p}}{\sqrt{\frac{p \cdot (1-p)}{n}}} \quad \bar{p} : \text{differenza media } (p_1 - p_2) \text{ che si vuole significativa}$$

Risolviendo per **n**, si ottiene $n = \frac{z^2 \cdot p \cdot (1-p)}{\bar{p}^2}$

Poichè la varianza di una percentuale o frequenza relativa ($p \cdot q$) è determinata direttamente dalla frequenza media, il numero di dati necessari per dimostrare la significatività di un differenza dipende dalle medie (p_1 e p_2) a confronto ($p_1 - p_2 = \bar{p}$); esso diminuisce in modo simmetrico, quanto più ci si allontana dal 50%

p media	p (1-p) σ^2
0,5	0,5 · 0,5 = 0,25
0,4	0,4 · 0,6 = 0,24
0,3	0,3 · 0,7 = 0,21
0,2	0,2 · 0,8 = 0,16
0,1	0,1 · 0,9 = 0,09
0,05	0,05 · 0,95 = 0,0475
0,04	0,04 · 0,96 = 0,0384
0,03	0,03 · 0,97 = 0,0291
0,02	0,02 · 0,98 = 0,0196
0,01	0,01 · 0,99 = 0,0099

ESEMPI

[1] In una popolazione animale arrivano in media all'età della riproduzione il 60% degli individui; con una nuova tecnica d'allevamento, si vuole dimostrare un miglioramento di almeno il 7%

Quanti individui servono perchè questa differenza risulti significativa alla probabilità 0,05 ?

E' un test ad una coda, dove $z_{0,05} = 1,645$ $\sigma^2 = p \cdot (1-p) = 0,6 \cdot 0,4 = 0,24$ $\bar{p} = 0,07$

$$n = \frac{(1,645)^2 \cdot 0,24}{(0,07)^2} = \frac{0,6494}{0,0049} = 132,5$$

Sono necessarie almeno 133 osservazioni

[2] Se la sopravvivenza è 90%, quanti dati si richiedono per valutare come statisticamente significativo alla stessa probabilità un miglioramento del 7% ?

$z_{0,05} = 1,645$ $\sigma^2 = 0,9 \cdot 0,1 = 0,09$ $\bar{p} = 0,07$

$$n = \frac{(1,645)^2 \cdot 0,09}{(0,07)^2} = \frac{0,2435}{0,0049} = 49,7$$

Si richiedono almeno 50 osservazioni.

Quando la varianza della popolazione σ^2 è ignota e si deve utilizzare la varianza del campione s^2 , si ricorre alla distribuzione **t**

Poichè il valore di **t** varia al variare dei gdl, e quindi delle dimensioni del campione, il calcolo di **n** richiede un procedimento di iterazione

$$n = \frac{t_{n-1}^2 \cdot s^2}{d^2} \quad \text{dove } d \text{ è la differenza media che si vuole sia significativa}$$

[3] In 5 campioni di acqua è stata misurata la concentrazione di una sostanza: la media è risultata pari a 39 grammi per litro e la varianza s^2 è risultata pari a 800

La differenza con il valore di 25 grammi/litro, indicato come il limite massimo tollerabile non risulta significativo

$$t_4 = \frac{39 - 25}{\sqrt{\frac{800}{5}}} = \frac{14}{12,65} = 1,107$$

Per un test ad una coda con, 4 gdl alla probabilità 0,05 il valore critico di t è pari a 2,1318; il valore calcolato è inferiore anche a quello tabulato alla probabilità 0,10 che è uguale a 1,5332

La probabilità di ottenere casualmente scarti uguali o maggiori di quello riscontrato tra la media rilevata e quella di legge è molto elevata

↳ si accetta H_0

Ma la media osservata è superiore a quella massima tollerabile; è ragionevole supporre che la differenza non sia risultata significativa a causa delle ridotte dimensioni del campione

Quanti dati sono necessari, a parità di media e di varianza, perchè quella differenza media risulti significativa alla probabilità 0,05 ?

Ricordando che all'aumentare dei gdl l'errore standard tende a diminuire, si può tentativamente scegliere t con 15 gdl alla probabilità 0,05 per un test unilaterale ($t_{15} = 1,7531$)

$$n = \frac{800 \cdot 1,7531^2}{14^2} = \frac{2458,72}{196} = 12,54$$

Sono pertanto necessari almeno 13 dati.

Il numero di osservazioni stimato si è dimostrato molto vicino a quello scelto a priori. Nel caso che tra i due risultati vi fosse stata una differenza rilevante, si sarebbero dovuti rifare i calcoli utilizzando il valore di un t con un numero di gdl intermedio, ripetendo il procedimento fino al valore esatto

ANALISI DELLA VARIANZA

Per il confronto tra le medie aritmetiche di più gruppi, non è possibile ricorrere al test **t**, suddividendo l'analisi in tanti confronti a coppie quante sono le combinazioni degli **n** gruppi 2 a 2.

Se i gruppi sono numerosi, la probabilità complessiva che almeno uno di essi sia significativo per caso aumenta proporzionalmente (ad es., con $\alpha=0,05$ e 20 confronti, mediamente uno risulterà significativo per caso, pur essendo vera H_0)

Nel confronto tra più medie, H_0 e H_1 assumono la formulazione :

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

le medie delle popolazioni dalle quali sono estratti casualmente i campioni sono tra loro uguali

H_1 : non tutte le medie aritmetiche sono uguali

si possono realizzare varie situazioni, e le più estreme sono:

- le medie sono tutte differenti tra loro
- una sola media è diversa dalle altre, tra loro uguali

Per verificare la significatività delle differenze tra le medie aritmetiche di vari gruppi si conduce un'ANALISI DELLA VARIANZA (sintetizzato in ANOVA, acronimo di **AN**alysis **O**f **V**ariance)

La distribuzione utilizzata è la distribuzione F in onore di Sir Ronald Aylmer Fisher (1890-1962), il più eminente statistico contemporaneo padre della statistica moderna

La metodologia attuale del test F è dovuta a Snedecor, un allievo di Fisher che ne perfezionò il metodo e ne semplificò la forma

Nel 1925 Fisher completò il metodo di Student per il confronto tra due medie, elaborando nel contempo il concetto di gdl: è suo il metodo attualmente utilizzato

ANOVA è la metodologia alla base della statistica moderna : gli stessi principi si applicano dalle analisi più semplici a quelle più complesse dell'analisi multivariata

IN ANOVA :

- si possono scomporre e misurare con precisione le fonti di variazioni sui valori osservati di due o più gruppi
- la fonte di variazione è detta FATTORE SPERIMENTALE (o TRATTAMENTO) e può essere a più livelli
- ogni unità od osservazione del fattore sperimentale è detta REPLICAZIONE

ANOVA AD UN CRITERIO DI CLASSIFICAZIONE (COMPLETAMENTE RANDOMIZZATA)

E' il modello più semplice di ANOVA

E' così chiamato in quanto si confrontano due o più livelli dello stesso fattore

E' detto anche MODELLO COMPLETAMENTE RANDOMIZZATO :

- prevede un campionamento in cui gli n individui omogenei (o repliche) sono assegnati casualmente ai vari livelli del fattore (o trattamenti)
- nel gruppo di soggetti da sottoporre ai diversi trattamenti per confrontarne gli effetti, l'attribuzione di ogni soggetto ad uno specifico trattamento va effettuato per estrazione casuale
- tutto il gruppo deve essere completamente randomizzato
- i vari gruppi possono non avere lo stesso n° di osservazioni o repliche (n_1, n_2, \dots, n_p sono in generale diversi tra loro)
- i dati sperimentali vanno riportati secondo la tabella sottostante

	LIVELLI DEL FATTORE SPERIMENTALE O TRATTAMENTI				
	T_1	T_2	T_3	...	T_p
UNITà' SPERIMENTALI (o REPLICAZIONI)	X_{11}	X_{12}	X_{13}	...	X_{1p}
	X_{21}	X_{22}	X_{23}	...	X_{2p}
	X_{31}	X_{32}	X_{33}	...	X_{3p}

	$X_{n_1,1}$	$X_{n_2,2}$	$X_{n_3,3}$...	$X_{n_p,p}$
	n_1	n_2	n_3	...	n_p
medie dei trattamenti	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}_{.3}$...	$\bar{X}_{.p}$
media generale	$\bar{X}_{..}$				

Secondo questo semplice modello di ANOVA, ogni singola osservazione X_{ij}

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

è composta da

- MEDIA GENERALE μ
- FATTORE α_j dovuto all'EFFETTO del TRATTAMENTO j-esimo misurato come

$$\alpha_j = \mu_j - \mu \quad \text{con :}$$

μ_j media del trattamento

μ media generale

- un FATTORE CASUALE ε_{ij} detto RESIDUO o ERRORE SPERIMENTALE (ricordiamo che “errore” non è sinonimo di sbaglio, ma di fattore sconosciuto, o non valutato, o non controllato nell'esperimento)

Gli errori ε_{ij} devono :

- ESSERE TRA LORO INDIPENDENTI: la variazione casuale di ogni replica non deve essere influenzata da quella di un'altra (è una indipendenza che può essere ottenuta solamente con una corretta distribuzione casuale delle repliche e quindi di una loro distribuzione secondo la normale)
- DARE VARIANZE OMOGENEE tra loro entro ogni trattamento
- ESSERE DISTRIBUITI NORMALMENTE

La metodologia di ANOVA prevede il calcolo di :

- devianza TOTALE scomposta in :
 - devianza TRA TRATTAMENTI (o BETWEEN) con i suoi gdl e la varianza relativa
 - devianza ENTRO TRATTAMENTI (o WITHIN o ERRORE) con i suoi gdl e la varianza relativa

Queste quantità abitualmente vengono presentate in uno specchietto :

devianza totale	gdl = n-1 (n = n° dati)	
devianza tra trattamenti	gdl = p-1 (p = n° gruppi)	“varianza tra”
devianza entro trattamenti	gdl = n-p	“varianza entro”

Devianza **TOTALE** (o SQ o Somma dei Quadrati degli scarti, o *Sum of Squares*) :

$$SQ_{\text{tot}} = \sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = \sum_{j=1}^p \sum_{i=1}^{n_j} X_{ij}^2 - \frac{(\sum_{j=1}^p \sum_{i=1}^{n_j} X_{ij})^2}{n}$$

- la prima formula, EURISTICA, definisce il significato di devianza totale
- la seconda formula, ABBREVIATA, è matematicamente equivalente alla prima, ma rende più semplici e rapidi i calcoli necessari

Devianza **TRA** TRATTAMENTI :

$$SQ_{\text{tra}} = \sum_{j=1}^p n_j \cdot (\bar{X}_j - \bar{X})^2 = \sum_{i=1}^{n_j} \sum_{j=1}^p (X_{ij}^2 / n_j) - \frac{(\sum_{i=1}^{n_j} \sum_{j=1}^p X_{ij})^2}{n}$$

Devianza **ENTRO** TRATTAMENTI :

$$SQ_{\text{entro}} = \sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Dividendo “devianza tra” e “devianza entro” per i rispettivi gdl si ottengono “varianza tra” e “varianza entro” :

- la “varianza tra” misura le differenze esistenti tra un gruppo e l'altro
- la “varianza entro” misura la variabilità esistente attorno alla media aritmetica di ogni gruppo

“Varianza tra” e “varianza entro” dipendono dalla variabilità esistente nei dati ed essendo due misure della stessa variabilità, dovrebbero avere lo stesso valore

La **DEVIANZA TOTALE** è data dalla somma dei quadrati degli scarti di ognuna delle 15 osservazioni rispetto alla media totale

A	B	C
$(2,71 - 2,478)^2$	$(1,75 - 2,478)^2$	$(2,22 - 2,478)^2$
$(2,06 - 2,478)^2$	$(2,19 - 2,478)^2$	$(2,38 - 2,478)^2$
$(2,84 - 2,478)^2$	$(2,09 - 2,478)^2$	$(2,56 - 2,478)^2$
$(2,97 - 2,478)^2$	$(2,75 - 2,478)^2$	$(2,60 - 2,478)^2$
$(2,55 - 2,478)^2$		$(2,72 - 2,478)^2$
$(2,78 - 2,478)^2$		

Quindi, svolgendo i calcoli e sommando i risultati

A	B	C
0,053824	0,529984	0,066564
0,174724	0,082944	0,009604
0,131044	0,150544	0,006724
0,242064	0,073984	0,014884
0,005184		0,058564
0,091204		
0,698040	0,837456	0,156340

$$\text{Devianza totale} = 0,698040 + 0,837456 + 0,156340 = 1,691836$$

Il metodo è lungo e produce stime non precise quando la media sia approssimata; per il calcolo manuale è conveniente utilizzare la formula abbreviata che comporta la somma dei quadrati di ogni replicazione

	A	B	C	
	7,3441	3,0625	4,9284	
	4,2436	4,7961	5,6644	
	8,0656	4,3681	6,5536	
	8,8209	7,5625	6,7600	
	6,5025		7,3984	
	7,7284			
Σx^2	42,7051	19,7892	31,3048	93,7991

$$\text{Devianza}_{tot} = Sx^2 - \frac{(Sx)^2}{n} = 93,7991 - \frac{(37,17)^2}{15} = 1,69184$$

“DEVIANZA TRA” :

- misura la variabilità esistente tra la media aritmetica di ogni gruppo e la media aritmetica generale, ponderata per il **n° di osservazioni** presenti in ciascun gruppo
- è la somma degli scarti di ogni media di gruppo rispetto alla media generale, ponderata per il **n° di replicazioni**
- ipotizza che, in assenza di variabilità d'errore, i dati sperimentali assumano i valori

A	B	C
2,652	2,195	2,496
2,652	2,195	2,496
2,652	2,195	2,496
2,652	2,195	2,496
2,652		2,496
2,652		
media totale		
2,478		

Pertanto con la formula euristica il calcolo diventa :

$$\text{Devianza}_{\text{tra}} = \sum_{j=1}^P n_j (\bar{X}_j - \bar{\bar{X}})^2$$

$$\begin{aligned} \text{Devianza}_{\text{tra}} &= 6 \cdot (2,652 - 2,478)^2 + 4 \cdot (2,195 - 2,478)^2 + 5 \cdot (2,496 - 2,478)^2 = \\ &= 6 \cdot 0,030276 + 4 \cdot 0,080089 + 5 \cdot 0,000324 = \\ &= 0,181656 + 0,320356 + 0,00162 = 0,503632 \end{aligned}$$

La formula abbreviata è più rapida e precisa :

$$\text{Devianza}_{\text{tra}} = \sum \frac{(Sx)_j^2}{n_j} - \frac{(Sx)^2}{n}$$

$$\text{Devianza}_{\text{tra}} = \frac{(15,91)^2}{6} + \frac{(8,78)^2}{4} + \frac{(12,48)^2}{5} - \frac{(37,17)^2}{15} = 92,610196 - 92,10726 = 0,502936$$

“DEVIANZA ENTRO” :

- misura la variazione tra il valore di ciascuna replicazione e la media aritmetica del proprio gruppo

- è la somma di queste differenze elevate al quadrato per ogni gruppo

A	B	C
$(2,71 - 2,652)^2$	$(1,75 - 2,195)^2$	$(2,22 - 2,496)^2$
$(2,06 - 2,652)^2$	$(2,19 - 2,195)^2$	$(2,38 - 2,496)^2$
$(2,84 - 2,652)^2$	$(2,09 - 2,195)^2$	$(2,56 - 2,496)^2$
$(2,97 - 2,652)^2$	$(2,75 - 2,195)^2$	$(2,60 - 2,496)^2$
$(2,55 - 2,652)^2$		$(2,72 - 2,496)^2$
$(2,78 - 2,652)^2$		

Sviluppando i calcoli e sommando si ottiene

	A	B	C
	0,003364	0,198025	0,076176
	0,350464	0,000025	0,013456
	0,035344	0,011025	0,004096
	0,101124	0,308025	0,010816
	0,010404		0,050176
	0,015376		
Devianza _{entro}	0,516076	0,517100	0,154720

- con la formula euristica (somma degli scarti al quadrato) risulta

$$\text{Devianza}_{\text{entro}} = 0,516076 + 0,517100 + 0,154720 = 1,187896$$

- può essere ottenuta sottraendo la “devianza tra” dalla devianza totale

$$\text{Devianza}_{\text{entro}} = \text{Devianza}_{\text{totale}} - \text{Devianza}_{\text{tra}} = 1,69184 - 0,502936 = 1,188904$$

Per riassumere i calcoli effettuati, si imposta una tabella che riporta le tre devianze con i rispettivi gdl :

- totale : n° di replicazioni meno 1
- “tra” : n° di trattamenti meno 1
- “entro” : n° di replicazioni meno il n° di trattamenti, equivalente ai gdl della devianza totale meno quelli della “devianza tra”

	DEVIANZE	GDL	VARIANZE
totale	1,69184	14	
“devianza tra” (between)	0,502936	2	0,251468
“devianza entro” (within)	1,188904	12	0,0990753

Dividendo “varianza tra” con “varianza entro”, si calcola il rapporto $F_{(2, 12)}$

$$F_{(2,12)} = \frac{0,251468}{0,0990753} = 2,538$$

- il valore critico di **F** (2 gdl al numeratore; 12 gdl al denominatore) per $\alpha=0,05$ è 3,89
- il valore calcolato di **F** è inferiore a quello tabulato: la probabilità che H_0 sia vera è $p>5\%$ e di conseguenza si accetta H_0 (i tre campioni sono stati estratti dalla stessa popolazione)

CONFRONTO TRA ANOVA CON DUE TRATTAMENTI E TEST t PER DUE CAMPIONI INDIPENDENTI

ANOVA può essere applicata anche a due soli trattamenti, in alternativa alla metodologia mediante test t

Test t e test F sono due modi solo apparentemente differenti per fare la stessa cosa: il test t è un caso speciale di ANOVA applicata a due gruppi

Tra t ed F esiste la precisa relazione matematica :

$$F_{(1, n)} = t_{(n)}^2$$

ovvero, il valore F (**un** gdl al numeratore e **n** gdl al denominatore) è uguale al quadrato di t con **n** gdl

ESEMPIO

Due gruppi di 10 uova di *Daphnia magna*, estratte casualmente dallo stesso clone, sono stati allevati in due diverse concentrazioni di cromo esavalente

Dopo un mese sono stati misurati gli individui sopravvissuti: 7 nel gruppo A e 8 nel gruppo B

A	B
2,7	2,2
2,8	2,1
2,9	2,2
2,5	2,3
2,6	2,1
2,7	2,2
2,8	2,3
	2,6

D.:

Verificare se le loro dimensioni sono statisticamente diverse

1- Medie:

media del gruppo A = 2,714

media del gruppo B = 2,250

2- Verifica di omogeneità delle due varianze, mediante il calcolo di devianze, gdl e rapporto **F** tra varianza maggiore e varianza minore

	A	B
devianze	0,10857	0,18000
gdl	6	7
varianze	0,018095	0,02571

$$F_{(7,6)} = \frac{0,02571}{0,018095} = 1,42$$

Con 7 gdl della varianza maggiore e 6 della varianza minore, per $\alpha=0,05$ l'**F** critico è 4,21 > 1,42 (**F** calcolato): dunque le varianze sono omogenee

3 - Varianza "pooled" $s_p^2 = \frac{0,10825 + 0,18000}{6 + 7} = 0,022173$

4 - **t** con 13 gdl $t_{13} = \frac{2,714 - 2,250}{\sqrt{0,022173 \cdot \left(\frac{1}{7} + \frac{1}{8}\right)}} = 6,02$

5 - Controllo della probabilità sulle tabelle dei valori critici: **p** << 0,001

6 - Prospetto di ANOVA

	devianze	gdl	varianze
totale	1,093333	14	
tra	0,804762	1	0,804761
entro	0,288571	13	0,022198

7 - **F** con 1 e 13 gdl $F_{(1,13)} = \frac{0,804761}{0,022198} = 36,25$

8 - Verifica che a tale valore corrisponde alla stessa probabilità, inferiore a 0,001

9 - Verifica che $t^2 = F$ infatti $t^2 = 6,02^2 = 36,24$

ANOVA A DUE CRITERI DI CLASSIFICAZIONE (BLOCCHI RANDOMIZZATI)

Nella pratica sperimentale, spesso è utile prendere in considerazione più di un fattore di variabilità quando si intende analizzare gli effetti di due o più cause contemporaneamente, oppure ridurre la varianza d'errore isolando gli effetti dovuti ad altre cause note

L'estensione più semplice è rappresentata da due criteri di classificazione, una struttura che si evidenzia nel disegno sperimentale a blocchi randomizzati, dove una classificazione riguarda i trattamenti e l'altra i blocchi

	p TRATTAMENTI					
k BLOCCHI	1	2	3	...	p	medie
1	X_{11}	X_{12}	X_{13}	...	X_{1p}	$\bar{X}_{1\cdot}$
2	X_{21}	X_{22}	X_{23}	...	X_{2p}	$\bar{X}_{2\cdot}$
...
k	X_{k1}	X_{k2}	X_{k3}	...	X_{kp}	$\bar{X}_{k\cdot}$
medie	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$	$\bar{X}_{\cdot 3}$...	$\bar{X}_{\cdot p}$	$\bar{X}_{\cdot\cdot}$

Nel caso più semplice si ha con una sola osservazione x_{ij} ad ogni intersezione della i-esima riga (blocco) per la j-esima colonna (trattamento)

Il modello lineare additivo, che considera l'effetto del trattamento e del blocco su ogni osservazione, è rappresentato da

$$X_{ij} = \mu + \alpha_j + \beta_i + R_{ij} \quad \text{con}$$

- μ media generale
- α_j effetto del trattamento stimato come differenza della sua
media rispetto alla media generale $\alpha_j = \bar{X}_{\cdot j} - \bar{X}$
- β_i effetto del blocco stimato come differenza della sua media
rispetto alla media generale $\beta_i = \bar{X}_{i\cdot} - \bar{X}$
- R_{ij} quota residua che ingloba, oltre a quelli considerati nei
blocchi e nei trattamenti, altri fattori non considerati e la loro
interazione insieme con gli effetti di campionamento o di
errore ϵ_{ij}

La metodologia ANOVA a due criteri di classificazione (**p** fattori e **k** blocchi) con una sola osservazione per casella prevede il calcolo delle seguenti quantità:

- devianza totale, con $p \cdot k - 1 = n - 1$ gdl
- devianza tra trattamenti, con $p - 1$ gdl, e rispettiva varianza
- devianza tra blocchi, con $k - 1$ gdl, e rispettiva varianza
- devianza d'errore, con $(p-1) \cdot (k-1) = (n-1) - (p-1) - (k-1) = p \cdot k - p - k + 1$ gdl, e rispettiva varianza

Devianze e gdl godono della proprietà additiva :

$$\begin{aligned}
 - \text{Devianza}_{\text{tot}} &= \text{Devianza}_{\text{tra tratt}} + \text{Devianza}_{\text{tra blocchi}} + \text{Devianza}_{\text{errore}} \\
 - \text{gdl}_{\text{tot}} &= \text{gdl}_{\text{tra tratt}} + \text{gdl}_{\text{tra blocchi}} + \text{gdl}_{\text{errore}}
 \end{aligned}$$

devianza totale	gdl: $n - 1 = p \cdot k - 1$	
devianza tra trattamenti	gdl: $p - 1$	varianza tra trattamenti
devianza tra blocchi	gdl: $k - 1$	varianza tra blocchi
devianza d'errore	gdl: $(p - 1) \cdot (k - 1)$	varianza d'errore

DEVIANZA TOTALE : variazione totale tra le osservazioni

$$\sum_{j=1}^p \sum_{i=1}^k (X_{ij} - \bar{\bar{X}})^2 = \sum_{j=1}^p \sum_{i=1}^k X_{ij}^2 - \frac{(\sum_{j=1}^p \sum_{i=1}^k X_{ij})^2}{n}$$

DEVIANZA TRA TRATTAMENTI : variazione tra le medie dei trattamenti

$$\sum_{j=1}^p k(\bar{X}_j - \bar{\bar{X}})^2 = \sum_{j=1}^p \left(\frac{\sum_{i=1}^k X_{ij}^2}{k} \right) - \frac{(\sum_{i=1}^k \sum_{j=1}^p X_{ij})^2}{n}$$

DEVIANZA TRA BLOCCHI : variazione tra le medie dei blocchi

$$\sum_{i=1}^k p(\bar{X}_i - \bar{\bar{X}})^2 = \sum_{i=1}^k \left(\frac{\sum_{j=1}^p X_{ij}^2}{p} \right) - \frac{(\sum_{i=1}^k \sum_{j=1}^p X_{ij})^2}{n}$$

DEVIANZA D'ERRORE (**RESIDUO**) : variazione di ogni osservazione dopo avere tolto l'effetto dovuto alla media generale, alla media del trattamento e alla media del blocco

$$\text{Devianza}_{\text{err}} = \text{Devianza}_{\text{tot}} - \text{Devianza}_{\text{tra tratt}} - \text{Devianza}_{\text{tra blocchi}}$$

Le varianze (tra trattamenti, tra blocchi, errore) si ottengono dividendo le rispettive devianze per i loro gdl

Il test **F** consiste nel confrontare sia la varianza tra trattamenti che quella tra blocchi separatamente con la varianza d'errore

• tra trattamenti : $F_{(p-1), (p-1) \cdot (k-1)} = \frac{\text{varianza tra tratt}}{\text{varianza d'errore}}$

• tra blocchi : $F_{(k-1), (p-1) \cdot (k-1)} = \frac{\text{varianza tra blocchi}}{\text{varianza d'errore}}$

ESEMPIO

Confrontare la quantità di Pb in sospensione nell'aria di 5 zone urbane, sapendo che esistono differenze durante la giornata; a distanza di 6 ore (alle 6, 12, 18 e 24) è stata fatta una rilevazione in ogni zona

D.:

C'è differenza tra ore e tra zone considerando i due fattori contemporaneamente ?

BLOCCHI (ORE)	TRATTAMENTI (ZONE)					X_{ij}	
	1	2	3	4	5	totali	medie
ore 6	28	25	30	22	26	131	26,2
ore 12	34	32	37	31	30	164	32,8
ore 19	22	21	24	20	19	106	21,2
ore 24	36	31	40	33	29	169	33,8
totali	120	109	131	106	104	570	
medie	30,00	27,25	32,75	26,50	26,00		28,50

DEVIANZA TOTALE con 19 gdl :

$$(28 - 28,5)^2 + (34 - 28,5)^2 + (22 - 28,5)^2 + \dots + (29 - 28,5)^2 = 683,0$$

oppure $(28^2 + 34^2 + 22^2 + 36^2 + 25^2 + 32^2 + \dots + 29^2) - \frac{570^2}{20} = 683,0$

La quantità $\frac{(SX)^2}{n} = \frac{570^2}{20}$ che compare sia nel calcolo della “devianza tot” che nelle due “devianze tra”, è detta **TERMINE DI CORREZIONE GENERALE (TCG)**

DEVIANZA TRA TRATTAMENTI (zone) con 4 gdl :

$$4 \cdot (30,00 - 28,5)^2 + 4 \cdot (27,25 - 28,5)^2 + \dots + 4 \cdot (26,00 - 28,5)^2 = 128,5$$

oppure $\frac{120^2}{4} + \frac{109^2}{4} + \frac{131^2}{4} + \frac{106^2}{4} + \frac{104^2}{4} - \frac{570^2}{20} = 128,5$

DEVIANZA TRA BLOCCHI (ore) con 3 gdl :

$$5 \cdot (26,2 - 28,5)^2 + 5 \cdot (32,8 - 28,5)^2 + \dots + 5 \cdot (33,8 - 28,5)^2 = 525,8$$

oppure $\frac{131^2}{5} + \frac{164^2}{5} + \frac{106^2}{5} + \frac{169^2}{5} - \frac{570^2}{20} = 525,8$

DEVIANZA D'ERRORE e relativi gdl : ottenuti per differenza

$$683,0 - 128,5 - 525,8 = 28,7 \quad \text{con} \quad 19 - 4 - 3 = 12 \text{ gdl}$$

	DEVIANZE	GDL	VARIANZE
totale	683,0	19	
tra trattamenti (zone)	128,5	4	32,125
tra blocchi (ore)	525,8	3	175,266
errore	28,7	12	2,39

La significatività della differenza tra zone è verificata con $F_{4,12} = \frac{32,125}{2,39} = 13,44$

La significatività della differenza tra ore è verificata con $F_{3,12} = \frac{175,266}{2,39} = 73,33$

Poiché i valori ottenuti superano quelli critici per $\alpha=0,05$

$$[F_{4,12} = 3,26 \quad F_{3,12} = 3,49]$$

le differenze tra le zone e le differenze tra le ore sono significative

Per comprenderne più esattamente il significato, è utile vedere quanto di ogni osservazione sia imputabile agli effetti congiunti [media generale, media di riga, media di colonna] e quanto ai rimanenti effetti espressi dal residuo

Conoscendo le medie marginali e totale, è possibile calcolare per ogni casella quale sarebbe il valore atteso se agissero solo i tre effetti noti :

$$\text{media di riga} + \text{media di colonna} - \text{media generale}$$

BLOCCHI	TRATTAMENTI					medie
	1	2	3	4	5	
I	27,70	24,95	30,45	24,20	23,70	26,20
II	34,30	31,55	37,05	30,80	30,30	32,80
III	22,70	19,95	25,45	19,20	18,70	21,20
IV	35,30	32,55	38,05	31,80	31,30	33,80
medie	30,00	27,25	32,75	26,50	26,00	28,50

Utilizzando questi dati per calcolare le devianze, si avrebbero valori identici a quelli dell'esempio per la devianza totale, per quella tra trattamenti e per quella tra blocchi, mentre la devianza d'errore risulterebbe 0, infatti ...

... la devianza d'errore calcolata precedentemente è la somma dei quadrati degli scarti tra questi valori stimati e quelli precedenti osservati

In questa tabella, ogni valore è la somma degli effetti $\mu + \alpha_j + \beta_i$ mentre è privo dell'effetto R_{ij} determinato da fattori di interazione e da variazioni casuali

QUADRATI LATINI

- TRE CRITERI DI CLASSIFICAZIONE
- DOPPIO DISEGNO A BLOCCHI

Analizzare contemporaneamente due fattori di variazione a **p** livelli nel disegno a blocchi randomizzati richiede **p²** osservazioni, mentre, con le stesse modalità di programmazione, un esperimento con tre fattori di variazione a **p** livelli ne richiederebbe **p³**

I quadrati latini furono applicati per la prima volta in esperimenti di agraria, dove la suddivisione in righe e colonne di un appezzamento di terreno erano visualizzate in strisce di terreno tra loro perpendicolari; da qui il nome, per la somiglianza del frazionamento dell'area in una figura tipica dell'accampamento romano

Il disegno a quadrati latini permette di analizzare contemporaneamente tre fattori a **p** livelli con **p²** osservazioni solamente

Al vantaggio di un risparmio di materiale si contrappone lo svantaggio di una notevole rigidità, infatti tutti i tre criteri (“trattamenti”, “blocchi”, “fattore principale”) devono avere lo stesso n° di livelli

In un esperimento con 3 criteri, due sono rappresentati da righe e da colonne (i fattori secondari), mentre il terzo (il fattore principale) è distribuito entro lo schema della tabella in modo casuale ma bilanciato, e compare una volta sola sia in ogni riga e in ogni colonna

Indicando con **A, B, C, D** i 4 livelli di un fattore principale, la rappresentazione grafica bidimensionale dell'esperimento può essere :

	COLONNE			
RIGHE	1	2	3	4
1	D	B	C	A
2	C	D	A	B
3	B	A	D	C
4	A	C	B	D

Così come in un disegno a due criteri di classificazione, la randomizzazione è attuata assegnando a caso i livelli dei trattamenti entro ciascun blocco, in un quadrato latino, la randomizzazione è attuata permutando i diversi livelli del fattore principale nello schema ordinato di righe e colonne

Sono state costruite tabelle di distribuzione casuale, da utilizzare nel caso di più esperimenti a quadrati latini con schemi differenti

Il limite più pesante a questo modo di programmare l'esperimento è dato dalla sua rigidità: ad esempio, volendo analizzare un fattore a 5 livelli, occorrerà un n° uguale di livelli anche negli altri due criteri organizzati per righe e colonne

Il modello additivo lineare di ANOVA in un disegno sperimentale a quadrato latino richiede che la generica osservazione X_{ijk} appartenente al i-esimo “blocco”, al j-esimo “trattamento” e al k-esimo fattore, sia data da

$$X_{ijk} = \mu + \alpha_j + \beta_i + \gamma_k + \varepsilon_{ijk} \quad \text{con:}$$

- μ media generale
- α_j effetto medio del “trattamento” i-esimo
- β_i effetto medio del “blocco” j-esimo
- γ_k effetto medio del fattore k-esimo
- ε_{ijk} variabilità residua

Il calcolo delle devianze è semplice: la devianza totale, la devianza tra righe e quella tra colonne sono calcolate con la stessa metodologia utilizzata nel disegno a blocchi randomizzati; la devianza tra trattamenti viene calcolata rispetto alla somma e alla media dei vari trattamenti

ESEMPIO

Confrontare la produttività di 5 (A, B, C, D, E) varietà di sementi in rapporto al tipo di concime (1,2,3,4,5) e ad un diverso trattamento del terreno (I, II, III, IV, V)

Si è diviso l'appezzamento in 5 strisce equivalenti e in ognuna è stata condotta un'aratura di profondità differente; perpendicolarmente a queste strisce sono state tracciate altre 5 strisce concimate in modo diverso; nei 25 quadrati sono state seminate le 5 varietà di sementi secondo lo schema

TRATTAMENTO DEL TERRENO							
CONCIME	I	II	III	IV	V	totali	medie
1	A 42	C 47	B 55	D 51	E 44	239	47,8
2	E 45	B 54	C 52	A 44	D 50	245	49,0
3	C 41	A 46	D 57	E 47	B 48	239	47,8
4	B 56	D 52	E 49	C 50	A 43	250	50,0
5	D 47	E 49	A 45	B 54	C 46	241	48,2
totali	231	248	258	246	231	1214	
medie	46,2	49,6	51,6	49,2	46,2		48,56

sementi	A	B	C	D	E
totali	220	267	236	257	234
medie	44,0	53,4	47,2	51,4	46,8

I risultati di ANOVA sono

	DEVIANZE	GDL	VARIANZE
totale	480,16	24	
tra sementi	286,16	4	71,54
tra concimi	109,36	4	27,34
tra arature	17,76	4	4,44
errore	66,88	12	5,57

Si possono calcolare tre F, tutti con 4 e 12 gdl :

• tra sementi: $F_{4, 12} = \frac{71,54}{5,57} = 12,84$

• tra concimi: $F_{4, 12} = \frac{27,34}{5,57} = 4,91$

• tra arature: questa varianza è minore della varianza d'errore e pertanto è inutile calcolare il rapporto F per verificare se gli sia significativamente superiore

Per 4 e 12 gdl la tabella dei valori critici per $\alpha = 0,05$ fornisce il valore 3,26, per cui risultano significative :

- la differenza tra sementi ($F = 12,84$)
 - la differenza tra concimi ($F = 4,91$)
- ma non quella tra i diversi tipi di aratura ($F < 1$)

ESEMPIO

Tra le numerose applicazioni, con i quadrati latini si possono analizzare gli effetti di diversi farmaci (“fattore principale”) da somministrare ad alcune persone (“blocchi”) in giorni diversi (“trattamenti”), e accertare se l'effetto di un farmaco dipenda anche dal tempo in cui è somministrato

Si sperimentano gli effetti di 4 diversi farmaci (A, B, C, D) somministrati in 4 giorni diversi a 4 diverse persone :

	GIORNI			
PERSONE	1	2	3	4
I	A 48	C 35	D 40	B 51
II	D 37	B 50	C 33	A 45
III	B 42	D 64	A 53	C 39
IV	C 31	A 40	B 42	D 37

I risultati di ANOVA sono

	DEVIANZE	GDL	VARIANZE
totale	1098	15	
tra farmaci	389	3	129,7
tra giorni	125	3	41,7
tra persone	303	3	101,0
errore	281	6	46,8

Il disegno sperimentale a quadrati latini impone che le sue dimensioni non possano essere nè troppo piccole, né troppo grandi :

- il limite **minimo** è imposto dai gdl della varianza d'errore [= $n^2 - (n-1) \cdot 3 - 1$]:

- un quadrato latino **2x2** avrebbe in totale 3 gdl: 1 per il fattore principale, 1 per le colonne e 1 per le righe, senza più gdl per la varianza d'errore

- un quadrato latino **3x3**, avrebbe la varianza d'errore con solo 2 gdl, troppo pochi per rendere significative differenze tra medie non molto grandi

- il limite massimo è determinato dalla complessità dell'esperimento e viene abitualmente fissato per un quadrato **12x12**

La replica di un esperimento a quadrati latini determina i quadrati greco-latini, che sono la sovrapposizione di due quadrati latini; con più repliche si parla di QUADRATI CON PIÙ ALFABETI

A parte la crescente complessità dell'esperimento, un punto importante da ricordare è il n° di gdl della varianza d'errore che diminuisce proporzionalmente

DATI MANCANTI

Nel disegno a blocchi randomizzati e in quello a quadrati latini :

- la mancanza di una osservazione pone un problema di elaborazione dei dati
- si richiede un numero prefissato di osservazioni, a differenza di quanto avviene nel disegno sperimentale ad un criterio di classificazione, dove la validità di ANOVA non dipende dall'eguaglianza del n° di repliche

I dati possono mancare per :

- selezioni contro determinati valori (ad es. quelli molto grandi o molto piccoli) che uno strumento può non registrare perché troppo differenti dalla scala su cui è tarato

--> il campione raccolto è viziato in modo irrimediabile

- cause accidentali

--> è possibile rimpiazzare i dati mancanti

In una tabella a due fattori o a due entrate senza replicazioni (come nel disegno a blocchi randomizzati) il dato da stimare dipende dagli effetti di riga e di colonna calcolati dagli altri dati

Si stima un valore X'_{ij} che dipende dalla media generale μ , dall'effetto riga α_i e dall'effetto colonna β_j , che sarà privo della variazione casuale ε_{ij}

$$X'_{ij} = \mu + \alpha_j + \beta_i$$

In un disegno con r righe e c colonne, indicando con R_i il totale di riga, con C_j il totale di colonna e con T il totale generale, calcolati senza il dato mancante, X_{ij} può essere sostituito con X'_{ij}

$$X'_{ij} = \frac{r \cdot R_i + c \cdot C_j - T}{(r-1) \cdot (c-1)}$$

ESEMPIO: se manca l'osservazione del trattamento ZONA 3 e blocco ORA II

TRATTAMENTI						
BLOCCHI	ZONA 1	ZONA 2	ZONA 3	ZONA 4	ZONA 5	TOTALI
ORA I	28	25	30	22	26	131
ORA II	34	32	?	31	30	127
ORA III	22	21	24	20	19	106
ORA IV	36	31	40	33	29	169
TOTALI	120	109	94	106	104	533

il valore con cui sostituire tale osservazione è $X'_{ij} = \frac{4 \cdot 127 + 5 \cdot 94 - 533}{(4-1) \cdot (5-1)} = 37,08$

La sostituzione del dato mancante permette di eseguire i calcoli di ANOVA nel modo abituale: cambiano i gdl totale e i gdl della varianza d'errore, che saranno diminuiti di 1 (restano immutati quelli tra trattamenti e tra blocchi)

L'operazione di sostituzione ha il solo scopo di permettere di effettuare i calcoli di ANOVA in modo corretto, ma non aggiunge alcuna informazione che non fosse già contenuta nei dati osservati

Se manca più di un dato:

- si sostituiscono i dati mancanti meno uno con dati inventati, anche se logici
- il dato non sostituito viene stimato con la modalità su illustrata
- si stima un altro dato in sostituzione di un dato inventato
- si itera il procedimento per tutti i dati originariamente mancanti, finchè i valori stimati restano stabili

Nel caso di un disegno a quadrati latini $N \times N$, il dato mancante nella i -esima riga, j -esima colonna e k -esimo fattore può essere sostituito da

$$X'_{ijk} = \frac{n \cdot (R_i + C_j + T_k) - 2G}{(n-1) \cdot (n-2)} \quad \text{con :}$$

n : dimensione del quadrato latino
 $R_i C_j T_k$: totali riga, colonna, trattamento cui appartiene il dato mancante
 G : totale generale

Anche qui la varianza tra trattamenti e l'errore standard della differenza tra due trattamenti andrebbero ridotti

REGRESSIONE LINEARE SEMPLICE

Nell'analisi della varianza a due o a più criteri di classificazione sono considerati contemporaneamente più fattori, come i vari trattamenti e blocchi con le loro interazioni, ma relativi sempre alla medesima ed unica variabile

Quando si considerano due o più variabili quantitative oltre alle precedenti analisi su ognuna di esse, si possono esaminare anche il tipo e l'intensità delle relazioni che sussistono tra loro

Nel caso in cui per ogni individuo si rilevino congiuntamente due variabili, è possibile verificare se esse variano simultaneamente e quale relazione matematica sussiste tra queste due variabili. Allora è possibile ricorrere all'analisi della regressione e a quella della correlazione, di norma considerate tra loro alternative

- **analisi della regressione** : per sviluppare un modello statistico che può essere usato per prevedere i valori di una variabile, detta dipendente o più raramente predetta ed individuata come l'effetto, sulla base dei valori dell'altra variabile, detta indipendente o esplicativa, individuata come la causa

- **analisi della correlazione** : per misurare l'intensità dell'associazione tra due variabili quantitative, di norma non legate direttamente da causa-effetto, facilmente mediate da almeno una terza variabile, ma che comunque variano congiuntamente

Quando per ciascuna unità di un campione o di una popolazione si rilevano due caratteristiche, si ha una DISTRIBUZIONE DOPPIA e i dati possono essere riportati in forma tabellare o grafica :

unità	carattere X	carattere Y
1	X_1	Y_1
2	X_2	Y_2
3	X_3	Y_3
...
n	X_n	Y_n

- se il numero di dati è ridotto, la distribuzione doppia può riguardare una tabella che riporta tutte le variabili relative ad ogni unità od individuo misurato

- se il numero di dati è grande, si ricorre ad una sintesi tabellare chiamata **DISTRIBUZIONE DOPPIA DI FREQUENZE** in cui si suddividono le unità del collettivo in classi per i due caratteri (X_i e Y_j) e poi

- si riporta la prima (X) nella TESTATA
- si riporta la seconda (Y) nella COLONNA MADRE
- si contano le unità che hanno contestualmente entrambe le MODALITÀ (n_{ij})

	X_1	X_2	X_3	...	X_i	...	X_n	Totali
Y_1	a_{11}	a_{12}	a_{13}	...	a_{1i}	...	a_{1n}	N_1
Y_2	a_{21}	a_{22}	a_{23}	...	a_{2i}	...	a_{2n}	N_2
Y_3	a_{31}	a_{32}	a_{33}	...	a_{3i}	...	a_{3n}	N_3
...
Y_j	a_{j1}	a_{j2}	a_{j3}	...	a_{ji}	...	a_{jn}	N_j
...
Y_m	a_{m1}	a_{m2}	a_{m3}	...	a_{mi}	...	a_{mn}	N_m
Totali	M_1	M_2	M_3	...	M_i	...	M_n	T

I totali delle righe e delle colonne rappresentano due distribuzioni semplici e sono dette **DISTRIBUZIONI MARGINALI** della distribuzione doppia

Le frequenze riportate in una colonna o in una riga sono dette **DISTRIBUZIONI PARZIALI** della doppia distribuzione : ad esempio, nello schema tabellare qui sopra sono presenti due distribuzioni marginali e 10 distribuzioni parziali (5 per riga e 5 per colonna)

Una distribuzione doppia può essere rappresentata graficamente con :

- **ISTOGRAMMI** : si riportano le frequenze dei raggruppamenti in classi come nelle distribuzioni di conteggi con dati qualitativi (tabelle $m \times n$)
- **DIAGRAMMI DI DISPERSIONE** : si riportano le singole coppie di misure osservate considerando ogni coppia della distribuzione come coordinate cartesiane di un punto del piano, sicché :
 - è possibile rappresentare ogni distribuzione doppia nel piano cartesiano
 - si ottiene una **NUVOLA DI PUNTI**, che descrive in modo visivo la relazione tra le due variabili

ESEMPIO

Lo studio e la classificazione tassonomica di specie di Macrobiotidi si fonda sia su aspetti qualitativi sia sui rapporti tra gli arti e di loro segmenti e, di norma, si ha una bassa variabilità intraspecifica e una forte variabilità interspecie

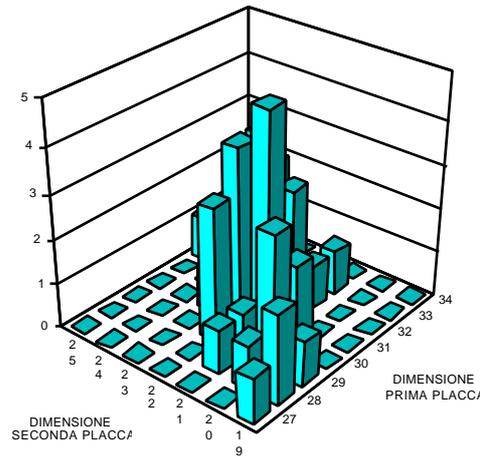
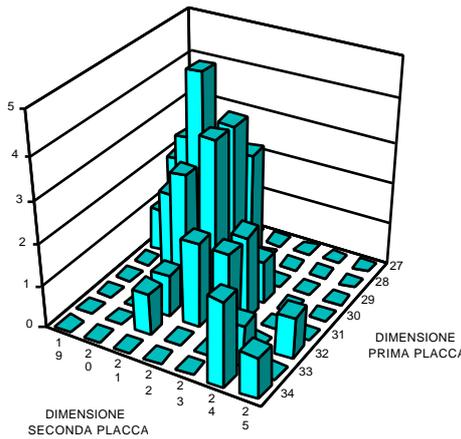
Per 45 animali della stesso gruppo *Macrobiotus hufelandi*, ma con forti dubbi sull'attribuzione della specie a causa delle difficoltà di classificazione dovute alla compresenza di giovani ed adulti, sono state misurate al microscopio le dimensioni (in μm) di parti dello scheletro, tra cui le dimensioni di prima e seconda placca

animali	prima placca	seconda placca
1	31	22
2	31	21
3	28	20
4	33	24
...
45	32	23

Per evitare pagine di numeri di difficile interpretazione, l'elevato numero di osservazioni impone il ricorso ad una rappresentazione più sintetica, ottenuta con una tabella

Per ogni coppia di valori diversi della prima variabile (testata) e della seconda variabile (colonna madre), si formano le distribuzioni di frequenza, con modalità analoghe a quelle della statistica univariata

		dimensione prima placca								totali
		27	28	29	30	31	32	33	34	
dimen- sione seconda placca	19	1	2	1	0	0	0	0	0	4
	20	0	1	3	2	0	0	0	0	6
	21	0	1	1	5	3	1	1	0	12
	22	0	0	3	4	4	2	0	0	13
	23	0	0	0	1	2	2	0	0	5
	24	0	0	0	0	0	0	1	2	3
	25	0	0	0	0	0	1	0	1	2
	totali	1	4	8	12	9	6	2	3	45



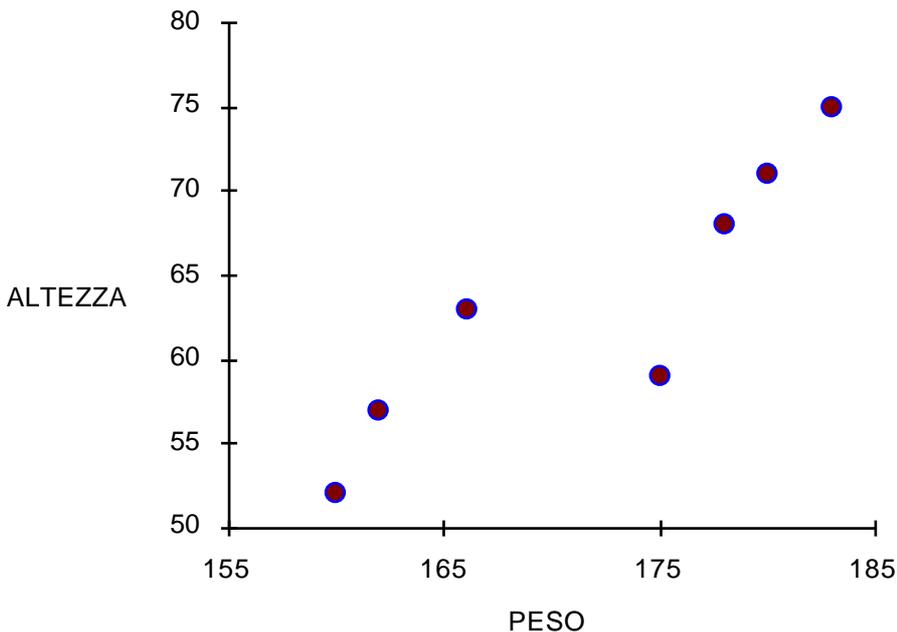
Quando le caselle sono troppe per essere riportate in una tabella di dimensioni medie, si ricorre al raggruppamento in classi di una sola o di entrambe le variabili

Quando i dati sono espressi in una scala continua, conviene darne una rappresentazione grafica mediante DIAGRAMMA DI DISPERSIONE :

- i dati di ogni individuo sono riportati su un diagramma bidimensionale ed indicati da un punto, le cui coordinate corrispondono ai valori X sull'asse delle ascisse e ai valori Y sull'asse delle ordinate

- più ricorrenze sono espresse da punti di dimensioni maggiori

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



MODELLI DI REGRESSIONE

Il diagramma di dispersione fornisce una descrizione visiva espressa in modo soggettivo, per quanto precisa, della relazione esistente tra le due variabili

La funzione matematica che la può esprimere in modo oggettivo è detta EQUAZIONE DI REGRESSIONE o FUNZIONE DI REGRESSIONE della variabile Y sulla variabile X

Il termine REGRESSIONE fu introdotto verso la metà dell'ottocento da Galton nei suoi studi di eugenica in cui si prefisse di verificare se la statura dei genitori influisse sulla statura dei figli e se questa corrispondenza potesse essere tradotta in una legge matematica

Galton confrontò anche l'altezza dei padri con quella dei figli ventenni e osservò che padri molto alti hanno figli alti, ma più vicini alla media dei loro genitori; parimenti egli osservò che i padri più bassi hanno figli maschi bassi, ma un pò più alti, più vicini alla media del gruppo, rispetto ai loro genitori (se egli avesse osservato l'altezza dei padri in rapporto ai figli avrebbe ugualmente trovato che i figli più bassi e quelli più alti hanno genitori con un'altezza più vicina alla media dei genitori)

Galton fu colpito da questo fenomeno, è affermò che la statura tende a "regredire" da valori estremi verso la media; nacque così il termine, che dal suo significato originario di "ritornare indietro" assunse quella della funzione che esprime matematicamente la relazione esistente tra la variabile attesa (o predetta o teorica) e la variabile empirica (o attuale)

La forma più generale di una equazione di regressione è

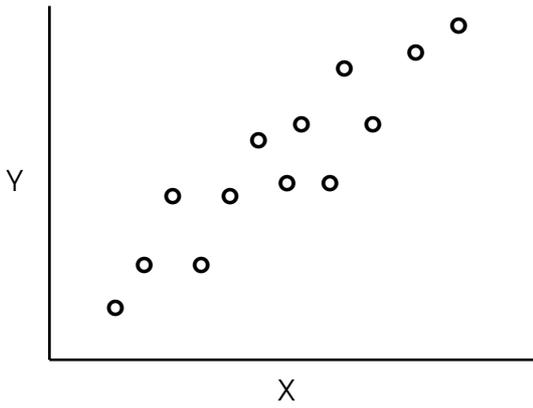
$$Y = a + b \cdot X + c \cdot X^2 + d \cdot X^3 + \dots$$

dove il secondo membro è un polinomio intero di X

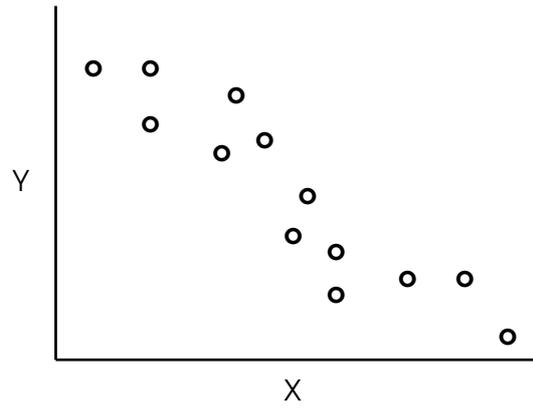
L'approssimazione della curva teorica ai dati sperimentali è tanto maggiore quanto più elevato è il numero di termini del polinomio :

- è frequente il caso di teorie che spiegano come, all'aumentare della variabile indipendente, si abbia una diminuzione o un aumento della variabile dipendente

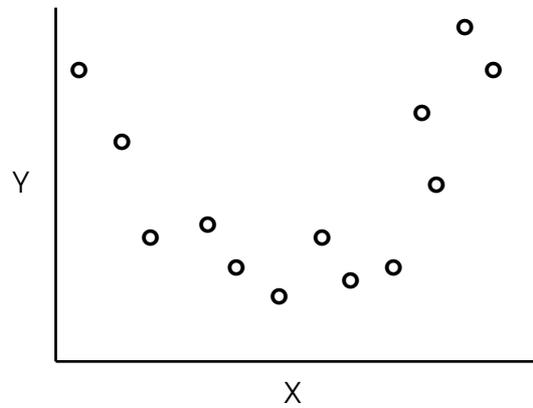
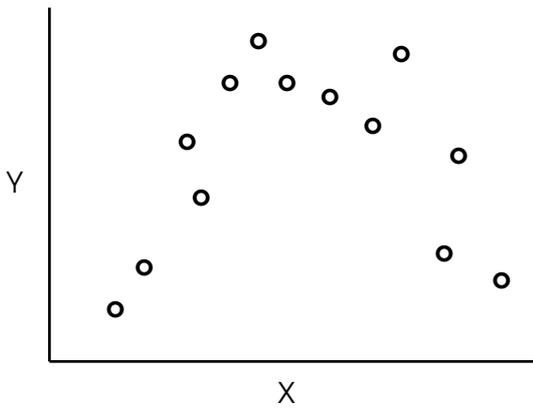
- è raro il caso in cui si può definire una teoria biologica o ambientale che spieghi una relazione più complessa (curva di terzo ordine o di ordine superiore)



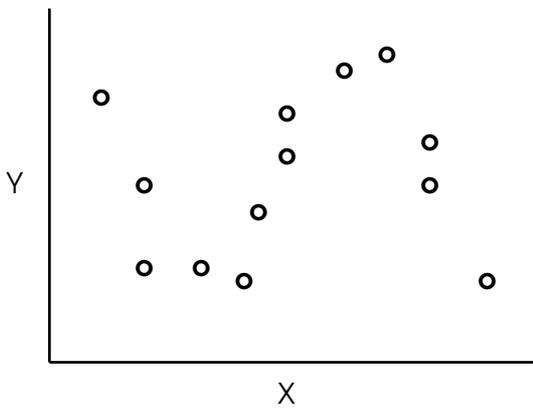
relazione lineare positiva



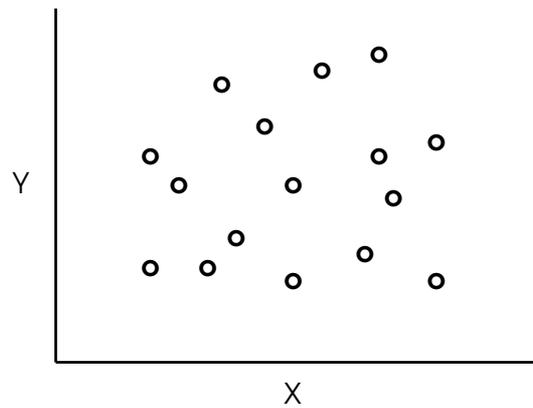
relazione lineare negativa



relazioni quadratiche



relazione cubica



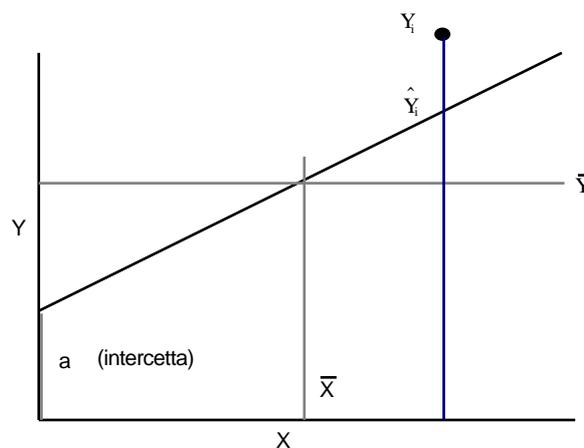
nessuna relazione

REGRESSIONE LINEARE SEMPLICE

La forma di relazione matematica più semplice tra due variabili è la regressione lineare semplice, rappresentata dalla retta di regressione

$$\hat{Y}_i = a + b \cdot X_i \quad \text{dove :}$$

- \hat{Y}_i valore stimato di Y per l'osservazione i-esima
- X_i valore empirico di X per l'osservazione i-esima
- a intercetta della retta di regressione
- b coefficiente angolare della retta di regressione



L'unica reale incognita è il valore del coefficiente angolare **b**, essendo l'intercetta **a** stimata da **b** e dai valori medi di Y e di X

$$a = \bar{Y} - b \cdot \bar{X}$$

Per calcolare la retta che meglio approssima la distribuzione dei punti, si può partire considerando che ogni punto osservato Y_i si discosta dalla retta di una certa quantità ε_i detta errore o RESIDUO

$$Y_i = a + b \cdot X_i + \varepsilon_i$$

Ogni valore ε_i può essere positivo o negativo:

- positivo quando il punto Y sperimentale è sopra la retta
- negativo quando il punto Y sperimentale è sotto la retta

La retta migliore per rappresentare la distribuzione dei punti nel diagramma di dispersione è quella stimata con il METODO DEI MINIMI QUADRATI (**V. PAGINA A FINE CAPITOLO**)

Indicando con Y_i i valori osservati (o empirici) e con \hat{Y}_i i corrispondenti valori stimati sulla retta, con un metodo analogo al calcolo della devianza si stima la migliore retta interpolante, cioè quella che minimizza la somma dei quadrati degli scarti dei valori osservati Y_i rispetto a quelli stimati \hat{Y}_i

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Essendo

$$\varepsilon_i = Y_i - (a + b \cdot X_i)$$

per il principio dei minimi quadrati si stimano **a** e **b** in modo che

$$\sum \varepsilon_i^2 = \sum (Y_i - (a + b \cdot X_i))^2 = \text{minimo}$$

Eguagliando a zero le derivate parziali rispetto ad **a** e a **b**, si trova che **b** è uguale al rapporto della codevianza XY con la devianza di X

$$b = \frac{\text{Codev}_{xy}}{\text{Dev}_x}$$

La CODEVIANZA :

- stima come X e Y variano congiuntamente, rispetto al loro valore medio
- è definita come la sommatoria dei prodotti degli scarti di X rispetto alla sua media e di Y rispetto alla sua media :

$$\text{Codev}_{xy} = \sum ((X - \bar{X}) \cdot (Y - \bar{Y}))$$

- si può esprimere con una formula empirica per un calcolo più rapido

$$\text{Codev}_{xy} = \sum (x \cdot y) - \frac{\sum x \cdot \sum y}{n}$$

In conclusione, il coefficiente angolare **b** è calcolato dalle coppie dei dati sperimentali X e Y come

$$b = \frac{\sum ((X - \bar{X}) \cdot (Y - \bar{Y}))}{\sum (X - \bar{X})^2}$$

oppure con l'equivalente formula rapida o empirica

$$b = \frac{\sum (x \cdot y) - \frac{\sum x \cdot \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

L'intercetta **a** si calcola come $a = \bar{Y} - b \cdot \bar{X}$

e poi si procede alla rappresentazione grafica, ricordando che :

- la retta passa sempre dal baricentro del grafico, individuato dal punto d'incontro delle due medie campionarie \bar{X} e \bar{Y}
- è sufficiente calcolare il valore di \hat{Y} corrispondente ad un qualsiasi valore di X per tracciare la retta che passa per questo punto calcolato e per il punto d'incontro tra le due medie

ESEMPIO

Per sette giovani è stato misurato il peso (Y) e l'altezza (X), allo scopo di stimare la retta che definisce la relazione media tra le due variabili

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

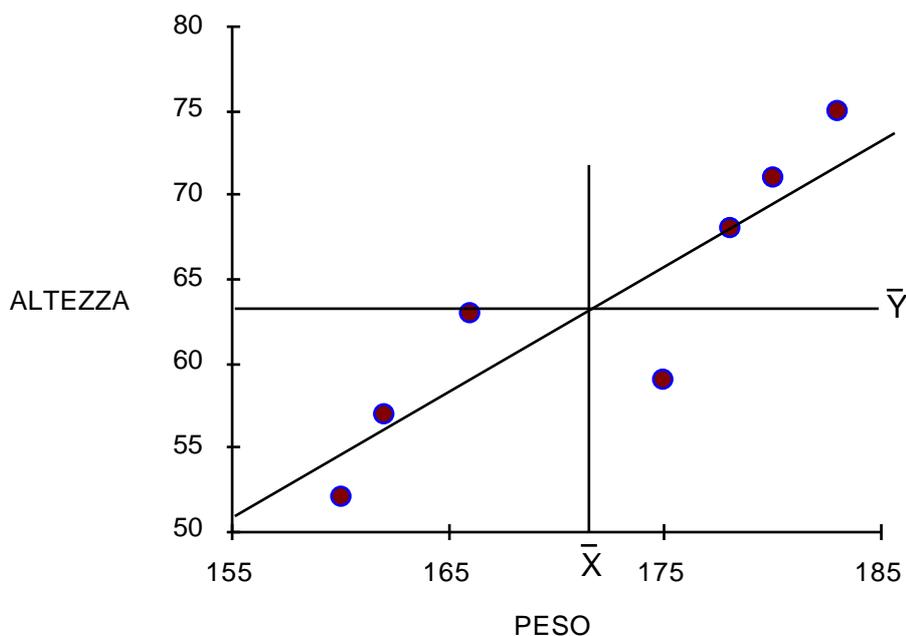
variabile indipendente (DETERMINISTICA) : altezza

variabile dipendente (STOCASTICA) : peso

$$\sum (X \cdot Y) = 76945 \quad \sum X = 1204 \quad \sum Y = 445 \quad \sum X^2 = 207598 \quad n = 7$$

$$b = \frac{\sum(x \cdot y) - \frac{\sum x \cdot \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{76945 - \frac{1204 \cdot 445}{7}}{207598 - \frac{1204^2}{7}} = 0,796$$

$$a = \bar{Y} - b \cdot \bar{X} = 63,571 - 0,796 \cdot 172 = -73,354$$



VALORE PREDITTIVO DELL'ANALISI DELLA REGRESSIONE

La semplice rappresentazione grafica dei valori osservati e della retta di regressione fornisce alcune indicazioni importanti per l'interpretazione delle relazioni esistenti tra le due variabili

Il valore del coefficiente angolare indica quanto aumenta in media la variabile dipendente Y all'aumento di una unità della variabile indipendente X

Se si cambia la scala della variabile indipendente o predittiva X (per esempio l'altezza misurata in mm o in m e non più in cm) lasciando invariata quella della variabile dipendente o predetta Y, muta proporzionalmente anche il valore del coefficiente angolare **b**

Nell'analisi della regressione :

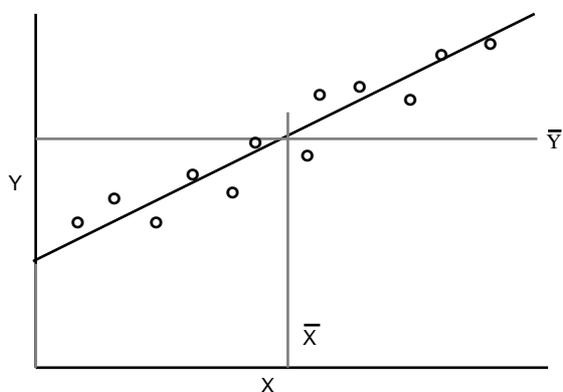
- è frequente, specialmente negli utilizzi predittivi, il ricorso al tempo come variabile indipendente
- viene spesso dimenticato che qualsiasi previsione o stima di Y derivata dalla retta è valida solo entro il campo di variazione della variabile indipendente X
- non è dimostrato che la relazione esistente tra le due variabili sia dello stesso tipo anche per valori minori o maggiori di quelli sperimentali rilevati

SIGNIFICATIVITÀ' DELLA RETTA DI REGRESSIONE

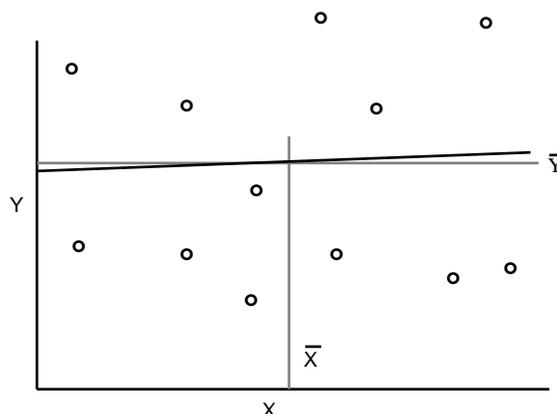
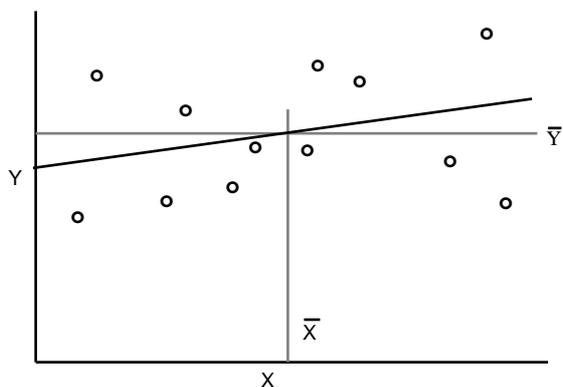
Il metodo dei minimi quadrati permette di avere sempre la retta che meglio si adatta ai dati rilevati, indipendentemente dalla loro dispersione intorno alla retta

Tuttavia la retta potrebbe indicare :

- sia l'esistenza di una relazione reale tra le due variabili, se il valore di **b** è alto e la dispersione dei punti intorno ad essa è ridotto
- sia di una relazione inesistente o non significativa, se i punti intorno ad essa sono dispersi in modo non differente rispetto alla media



(A) reale cambiamento di Y al variare di X



(B) caso incerto

(C) non c'è alcuna regressione

Il coefficiente angolare **b** della retta di regressione, che determina la quantità di variazione di Y per ogni unità aggiuntiva di X, è calcolato da osservazioni sperimentali

Ma ciò che interessa al ricercatore è la relazione esistente nella popolazione, e sebbene il valore di **b** sia differente da zero, non è detto che nella popolazione al variare di X si abbia una variazione di Y

La significatività del coefficiente di regressione nella popolazione (β) può essere saggiata mediante la verifica dell' $H_0: \beta = 0$

Accettando H_0 si assume che il valore reale del coefficiente angolare sia $\beta = 0$

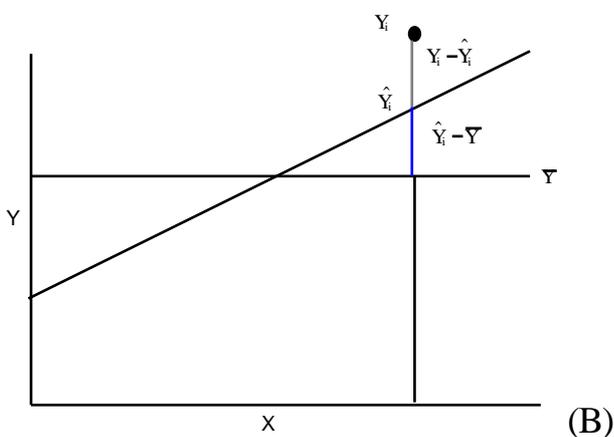
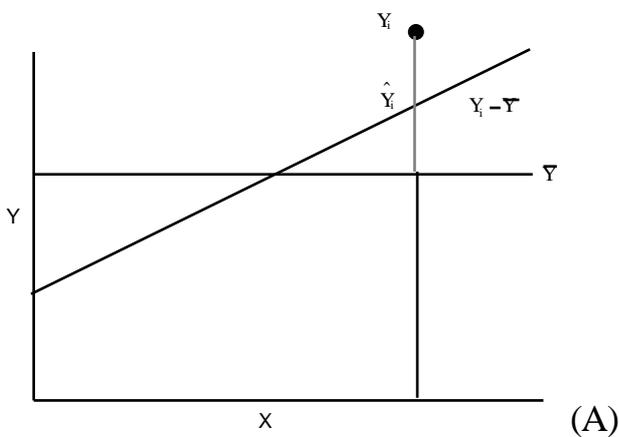
--> al variare di X, Y resta costante e uguale al valore dell'intercetta **a**

--> non esiste alcun legame tra X e Y

Rifiutando H_0 , si accetta H_1

--> al variare di X si ha una corrispondente variazione sistematica di Y

Un metodo per la verifica della significatività della retta calcolata è il **test F**, che si basa sulla scomposizione delle devianze



La somma dei quadrati delle distanze tra i tre punti Y , \hat{Y} e \bar{Y} definiscono le tre devianze: devianza totale, devianza della regressione o devianza dovuta alla regressione, devianza d'errore o devianza dalla regressione o residui:

$$\begin{aligned} \text{Devianza totale} &= \sum (Y - \bar{Y})^2 && \text{con gdl } n-1 \text{ (A)} \\ \text{Devianza della regressione} &= \sum (\hat{Y} - \bar{Y})^2 && \text{con gdl } 1 \text{ (B, parte inferiore)} \\ \text{Devianza d'errore} &= \sum (Y - \hat{Y})^2 && \text{con gdl } n-2 \text{ (B, parte superiore)} \end{aligned}$$

Queste formule richiedono calcoli lunghi e danno risultati approssimati quando i valori delle tre Y sono arrotondati, per cui si utilizzano le formule seguenti :

$$\text{Devianza totale} = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$\text{Devianza dalla regressione} = \frac{\text{Codev}_{xy}^2}{\text{Dev}_x}$$

$$\text{ricordando che} \quad \text{Cod}_{(x,y)} = \sum (x \cdot y) - \frac{\sum x \cdot \sum y}{n} \quad \text{Dev}_x = \sum X^2 - \frac{(\sum X)^2}{n}$$

Devianza d'errore (per differenza)

$$\text{Devianza d'errore} = \text{Devianza totale} - \text{Devianza della regressione}$$

Dal rapporto della devianza dovuta alla regressione e quella d'errore con i rispettivi gdl si stimano la varianza dovuta alla regressione e la varianza d'errore il cui rapporto determina il valore del test F con 1 e n-2 gdl

$$F_{(1, n-1)} = \frac{\text{Varianza dalla regressione}}{\text{Varianza d'errore}}$$

Se l'F calcolato è inferiore a quello tabulato per la probabilità prefissata e i gdl corrispondenti, si accetta H_0 (non esiste regressione lineare statisticamente significativa)

Se l'F calcolato supera quello tabulato si rifiuta l' H_0 e si accetta H_1 (la regressione lineare tra le due variabili è significativa)

Se $\beta = 0$, la varianza dovuta alla regressione e quella della regressione o d'errore sono stime indipendenti e non viziate della variabilità dei dati

Se $\beta \neq 0$, la varianza d'errore è una stima non viziosa della variabilità dei dati, mentre la varianza dovuta alla regressione è stima di una grandezza maggiore

Di conseguenza, il rapporto tra le varianze con rispettivamente 1 e n-2 gdl è da ritenersi utile alla verifica dell'ipotesi $\beta = 0$

Rifiutare H_0 :

- non significa che non esiste relazione tra le due variabili, ma solamente che non esiste una relazione di tipo lineare
- significa che potrebbe esistere una relazione di tipo differente, come quella curvilinea di secondo grado o di grado superiore

La TRASFORMAZIONE di uno o di entrambi gli assi è spesso sufficiente per ricondurre una relazione di tipo curvilineo a quella lineare

- la crescita esponenziale di una popolazione nel tempo, generata da tassi costanti, diviene lineare con la trasformazione logaritmica del tempo, di norma riportato sull'asse delle ascisse

- la relazione curvilinea tra lunghezza e peso di individui della stessa specie diviene lineare con la trasformazione mediante radice cubica del peso, correlato linearmente al volume

- l'analisi statistica permette qualsiasi tipo di trasformazione che determini una relazione lineare tra due variabili

ESEMPIO

Con le misure di peso ed altezza rilevati su 7 individui è stata calcolata la retta di regressione $\hat{Y} = -73,354 + 0,796 X$

Dopo aver costruito il diagramma di dispersione delle 7 coppie di osservazioni è stata rappresentata la retta :

- non è quella che passa più vicino ai punti, ma quella che rende minima la somma dei quadrati delle distanze tra la retta e i punti
- una retta con tale proprietà può essere sempre calcolata per qualsiasi gruppo di dati
- non è detto che tale retta sia rappresentativa o indice della reale esistenza di un rapporto lineare tra le due serie di dati

Pertanto, con le tecniche dell'inferenza, occorre verificare :

- se la retta può essere assunta come rappresentativa di un rapporto lineare tra le due variabili
- se è corretto affermare che, nella popolazione dei soggetti dalla quale è stato estratto il campione, ad una variazione in altezza corrisponde un cambiamento lineare nel peso
- se, mediante test F, $H_0: \beta = 0$ oppure $H_1: \beta \neq 0$

$$\sum(X \cdot Y) = 76945 \quad \sum X = 1204 \quad \sum X^2 = 207598 \quad \sum Y = 445 \quad \sum Y^2 = 28693$$

$$Devianza\ totale = 28693 - \frac{445^2}{7} = 28693 - 28289,285 = 403,715$$

$$Devianza\ della\ regr. = \frac{(76945 - \frac{1204 \cdot 445}{7})^2}{207598 - \frac{1204^2}{7}} = \frac{(76945 - 76540)^2}{207598 - 207088} = \frac{164025}{510} = 321,618$$

$$Devianza\ d'\ errore = 403,715 - 321,618 = 82,097$$

Tabella riassuntiva

	Devianze	gdl	Varianze
totale	403,715	6	321,62
regressione	321,618	1	16,42
errore	82,097	5	

$$F_{(1,5)} = \frac{321,62}{16,42} = 19,59$$

- i valori critici riportati nelle tavole degli F per 1 e 5 gdl sono: 6,61 per $\alpha = 0,05$ e 16,26 per $\alpha = 0,01$
- il valore calcolato è superiore a quello tabulato per $\alpha=0,01$
- con $p < 0,01$ (di commettere un errore di I° tipo, si rifiuta H_0 : esiste un rapporto lineare tra le variazioni in altezza e quelle in peso

La stima della significatività della retta o verifica dell'esistenza di una relazione lineare tra le due variabili può essere condotta anche con il test t, con risultati equivalenti al test F

Analogamente all'analisi della varianza ad uno e a due criteri di classificazione, il t con n-2 gdl (n = n° di osservazioni o coppie di dati) è

$$t_{(n-2)} = \sqrt{F_{(1, n-2)}}$$

Il test t è :

- fondato su calcoli didatticamente meno evidenti di quelli del test F, ma offre il vantaggio di poter essere applicato sia in test unilaterali ($\beta > 0$? oppure $\beta < 0$?) che in test bilaterali ($\beta \neq 0$?)

- fondato sul rapporto tra il valore del coefficiente angolare **b** (che rappresenta la risposta media di Y ai diversi valori di X entro il suo intervallo di variazione) ed il suo errore standard **s_b**

- $t_{(n-2)} = \frac{b - \beta}{S_b}$ dove β : valore atteso

Nella verifica della significatività della regressione si ha

$$\beta = 0$$

ma la formula può essere utilizzata per verificare la significatività dello scostamento da qualunque valore atteso

Un test relativamente frequente consiste nel verificare se **b** si discosta significativamente da 1, quando è atteso che all'aumentare di una unità di X si abbia un corrispondente aumento di una unità anche nel valore di Y, qualunque siano le unità di misura delle due variabili

Il valore di **S_b** è determinato dalla radice quadrata del rapporto tra la dispersione dei dati sperimentali Y intorno alla retta \hat{Y} e la devianza totale di X

$$s_b = \sqrt{s_b^2} \quad \text{dove: } s_b^2 = \frac{\text{Varianza d'errore della retta}}{\text{Devianza totale della X}} = \frac{s_e^2}{\sum (X_i - \bar{X})^2}$$

La varianza d'errore di **b** (s_b^2) diminuisce, e quindi il suo grado di precisione cresce, all'aumentare della devianza degli X

La varianza d'errore della retta s_e^2 chiamata anche ERRORE STANDARD DELLA STIMA è data da $s_e^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2}$

dove la devianza d'errore (al numeratore) è ottenuta in modo rapido per differenza dopo il calcolo della devianza totale e di quella dovuta alla regressione

$$s_e^2 = \frac{\text{Devianza totale di Y} - \text{Devianza dalla regressione}}{n - 2}$$

Per la devianza dovuta alla regressione sono state proposte anche altre formule che permettono calcoli più rapidi

Un metodo al quale si ricorre con frequenza utilizza parte dei calcoli necessari alla stima della retta

$$\text{Devianza dalla regressione} = \sum Y_i^2 - a \cdot \sum Y_i - b \cdot \sum (X_i \cdot Y_i)$$

ESEMPIO

Con le stesse 7 misure di peso ed altezza degli esercizi precedenti, si vuole stimare la significatività della regressione mediante il test t

In questo caso :

- si potrebbe ricorrere ad un test unilaterale (verificare solamente se il peso aumenti, oppure diminuisca, in modo significativo al crescere dell'altezza)

$$H_0: \beta = 0; \quad H_1: \beta > 0 \text{ oppure } H_1: \beta < 0$$

- si dovrebbe ricorrere ad un test bilaterale (verificare l'esistenza di una relazione lineare tra le due variabili senza indicarne il segno)

Ricordando che

$$b = 0,796 \quad \text{Varianza d'errore} = 16,42 \quad n = 7 \quad \text{Devianza di X} = 510$$

$$S_b^2 = \frac{16,42}{510} \quad s_b = 0,1794$$

si ha

$$t_5 = \frac{0,796}{0,1794} = 4,437$$

$$F_{1,5} = 19,59 \text{ corrisponde a } t_5 = \sqrt{19,59} = 4,426$$

La pendenza della retta è l'informazione più importante sulla relazione tra X e Y: fornisce la quantità di variazione media di Y per unità di variazione di X

Il test di significatività risponde solamente al quesito se essa si discosta da 0

Un caso che ricorre con frequenza è quando X e Y sono il risultato di due metodi differenti per stimare la stessa quantità di una sostanza, per cui al valore nullo di Y dovrebbe corrispondere un valore nullo anche per X

- per $X = 0$ si dovrebbe avere una risposta media che non si discosta significativamente da $Y = 0$
- la significatività dell'intercetta ($H_0: \alpha = 0$) può essere verificata sia con un test unilaterale che con un test bilaterale

$$t_{(n-2)} = \frac{a - \alpha}{s_a} \quad \text{con } S_a \text{ errore standard dell'intercetta } \mathbf{a} \text{ stimato come}$$

$$s_a = \sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right)}$$

Se non è possibile rifiutare H_0 relativa a \mathbf{b} (la retta campionaria non può essere assunta come significativa di una relazione lineare tra le due variabili), può essere richiesta la conoscenza della varianza $s_{\bar{Y}}^2$ e della deviazione standard $s_{\bar{Y}}$ della media \bar{Y} , che sono

$$s_{\bar{Y}}^2 = \frac{s_e^2}{n} \quad \text{e} \quad s_{\bar{Y}} = \frac{s_e}{\sqrt{n}}$$

ESEMPIO

Utilizzando i dati degli esempi precedenti, si stimi la significatività di \mathbf{a}

Con $a = -73,357$ var. err. : $s_e^2 = 16,101$ $n = 7$ dev. X : 510 $\bar{X} = 172$

- errore standard di \mathbf{a}
$$s_a = \sqrt{16,101 \cdot \left(\frac{1}{7} + \frac{172^2}{510} \right)} = 30,599$$

- t
$$t_5 = \frac{-73,357}{30,599} = -2,397$$

inferiore sia a $t_{5, 025}$ (2,571) che a $t_{5, 005}$ (4,032)

--> l'intercetta \mathbf{a} non è significativamente diversa da zero né all'1% né al 5%

LIMITI DI CONFIDENZA DI RETTA E INTERCETTA

Per verificare l'esistenza di una relazione lineare tra le variabili un altro metodo, equivalente al test t, è calcolare una stima per intervalli di confidenza di β : si rifiuta H_0 se il valore atteso (di solito, ma non obbligatoriamente come nel test per la media, $\beta = 0$) è compreso nell'intervallo di confidenza

stima per l'intervallo di confidenza di β : $b \pm t_{(n-2, \alpha/2)} \cdot s_b$

stima per l'intervallo di confidenza di α : $a \pm t_{(n-2, \alpha/2)} \cdot s_a$

dove s_a è l'errore standard dell'intercetta α

ESEMPIO

Ricorrendo agli stessi dati degli esercizi in cui sono stati calcolati la retta e la sua significatività, si ha

$$b = 0,796; \quad s_b = 0,1794; \quad t_{5, 0,025} = 2,571; \quad t_{5, 0,005} = 4,032$$
$$a = -73,357 \quad s_a = 30,599$$

Stima dell' intervallo di confidenza per il coefficiente angolare β

con $p = 95\%$

$$0,796 - 2,571 \cdot 0,1794 \leq \beta \leq 0,796 + 2,571 \cdot 0,1794 \qquad 0,335 \leq \beta \leq 1,257$$

con $p = 99\%$

$$0,796 - 4,032 \cdot 0,1794 \leq \beta \leq 0,796 + 4,032 \cdot 0,1794 \qquad 0,727 \leq \beta \leq 1.519$$

Stima dell'intervallo di confidenza per l'intercetta α

con $p = 95\%$

$$-73,357 - 2,571 \cdot 30,599 \leq \alpha \leq -73,357 + 2,571 \cdot 30,599 \qquad -152,027 \leq \alpha \leq 5,313$$

con $p = 99\%$

$$-73,357 - 4,032 \cdot 30,599 \leq \alpha \leq -73,357 + 4,032 \cdot 30,599 \qquad -196,732 \leq \alpha \leq 50,018$$

LIMITI DI CONFIDENZA PER I VALORI MEDI DEGLI Y STIMATI

La retta di regressione può essere utilizzata anche per previsioni sul valore medio di Y, corrispondente ad valore di X prescelto

E' una stima puntuale del valore medio effettivo del campione; anche in questo caso, può essere applicato il concetto di intervallo di confidenza quale stima del valore reale della popolazione

L'intervallo di confidenza per il valore previsto \hat{Y}_1 è dato da

$$\hat{Y}_1 \pm t_{(n-2, \alpha/2)} \cdot s_b \cdot \sqrt{\frac{1}{n} + \frac{(X_1 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

dove

\hat{Y}_1	valore previsto di Y per un dato valore di X
s_b	errore standard della retta b
n	dimensione del campione
X_1	valore dato di X a cui corrisponde \hat{Y}_1
$\sum (X_i - \bar{X})^2$	devianza di X

La lettura dell'equazione spiega come l'ampiezza dell'intervallo di confidenza dipenda da vari fattori

Per una data probabilità:

- aumenta al crescere della varianza d'errore;
- diminuisce all'aumentare del numero n di coppie di osservazioni per l'effetto congiunto del valore di $t_{n-2, \alpha/2}$ e del il rapporto $1/n$;
- varia secondo i valori di X, con valori minimi quando X_1 è vicino alla sua media e massimi quando X_1 ha distanza massima,
- diminuisce al crescere della devianza di X

L'intervallo di stima della vera media aritmetica varia come una funzione iperbolica della vicinanza di X alla sua media

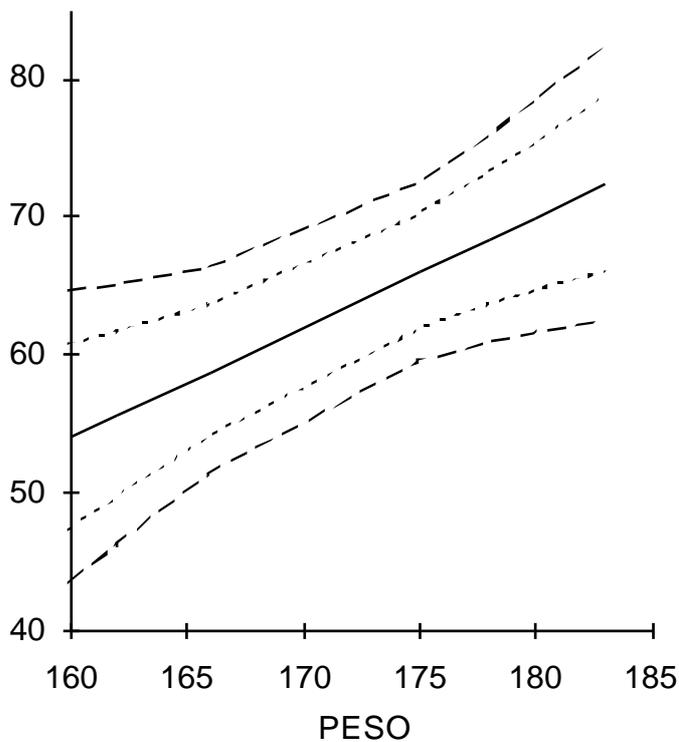
Quando si fanno previsioni su valori di X molto distanti dalla media, si stima un intervallo di confidenza molto più grande

Di conseguenza, i limiti della zona di confidenza non sono paralleli alla retta di regressione, ma se ne discostano progressivamente avvicinandosi agli estremi del valore di X

ESEMPIO

Consideriamo i 7 dati dell'esempio precedente; nella tabella sono riportati gli intervalli di confidenza degli Y stimati

Altezza X	Peso Y	Valori attesi di Y con il loro intervallo di confidenza	
		($\alpha = 0.05$)	($\alpha = 0.01$)
160	52	47,291 ≤ 54,018 ≤ 60,495	43,468 ≤ 54,018 ≤ 64,568
178	68	63,582 ≤ 68,348 ≤ 73,114	60,873 ≤ 68,348 ≤ 75,823
183	75	65,968 ≤ 72,328 ≤ 78,688	62,353 ≤ 72,328 ≤ 82,303
180	71	64,596 ≤ 69,940 ≤ 75,284	61,560 ≤ 69,940 ≤ 78,321
166	63	54,029 ≤ 58,795 ≤ 63,561	51,320 ≤ 58,795 ≤ 66,270
175	59	61,827 ≤ 65,960 ≤ 70,093	59,478 ≤ 65,960 ≤ 72,442
162	57	49,605 ≤ 55,611 ≤ 61,617	46,192 ≤ 55,611 ≤ 65,030



LIMITI DI CONFIDENZA PER SINGOLI VALORI DI Y STIMATI

Un'altra esigenza presente nella ricerca è la previsione dell'intervallo di confidenza per una singola risposta di Y

L'intervallo di confidenza ha una forma simile a quella del valore medio, ma è molto più ampio; ha infatti lo scopo di stimare un valore individuale e non un parametro

I valori stimati di Y per i singoli valori individuali di X, rispetto al valore medio che condidera tutta la retta, sono soggetti ad una sorgente addizionale d'errore, cioè alla dispersione intorno alla retta di regressione

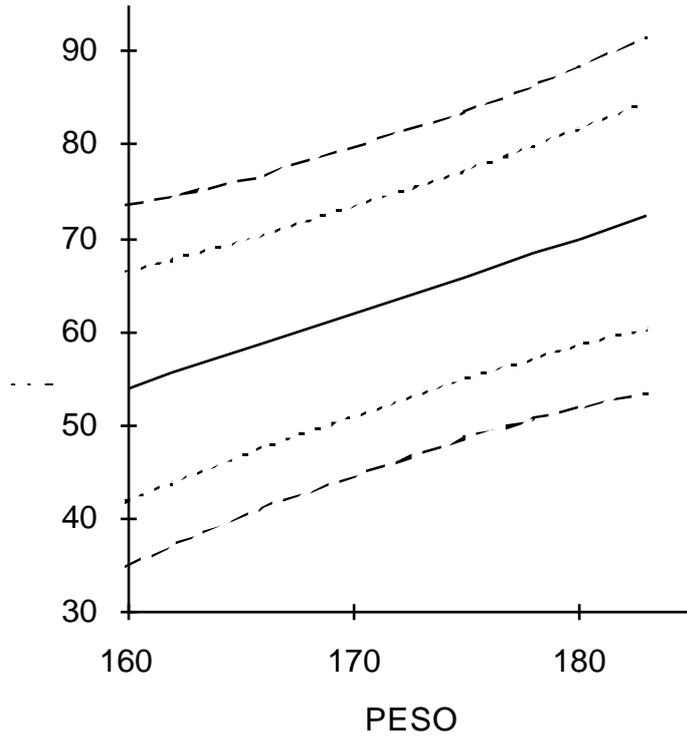
I limiti della zona di confidenza per singoli valori di X sono paralleli ai limiti della zona di confidenza della retta di regressione e sono più esterni ai precedenti

L'intervallo di previsione per un singolo valore di Y^{\wedge}_i per un dato valore x_i è dato da

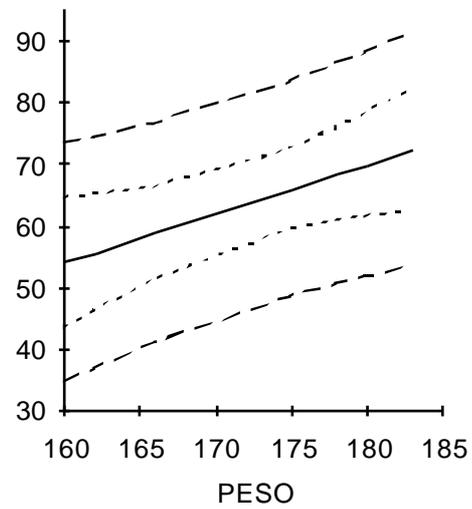
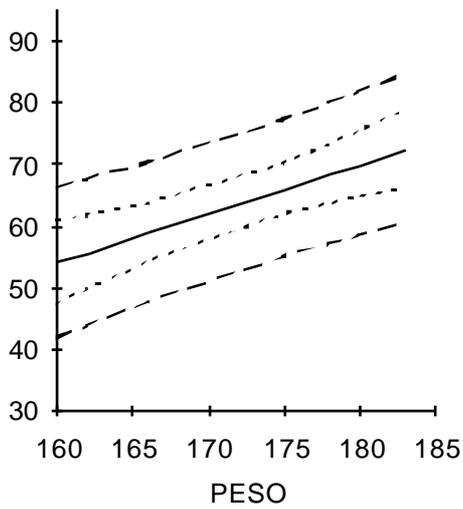
$$\hat{Y}_i \pm t_{(n-2, \alpha/2)} \cdot s_b \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

E' una espressione simile alla precedente; l'unica differenza è un 1 sommato all'argomento della radice quadrata

Altezza X	Peso Y	Valori attesi di Y con il loro intervallo di confidenza	
		($\alpha = 0.05$)	($\alpha = 0.01$)
160	52	41,702 ≤ 54,018 ≤ 66,334	34,703 ≤ 54,018 ≤ 73,332
178	68	56,984 ≤ 68,348 ≤ 79,712	50,526 ≤ 68,348 ≤ 86,170
183	75	60,208 ≤ 72,328 ≤ 84,447	53,321 ≤ 72,328 ≤ 91,335
180	71	58,322 ≤ 69,940 ≤ 81,558	51,719 ≤ 69,940 ≤ 88,161
166	63	47,431 ≤ 58,795 ≤ 70,159	40,973 ≤ 58,795 ≤ 73,617
175	59	54,846 ≤ 65,960 ≤ 77,074	48,531 ≤ 65,960 ≤ 83,389
162	57	43,674 ≤ 55,611 ≤ 67,548	36,890 ≤ 55,611 ≤ 74,332



Intervalli di confidenza per gli Y stimati al 5% (punteggiato) e all'1% (tratteggiato)



Intervalli di confidenza delle medie (linee punteggiate) e dei singoli valori di Y stimato (linee tratteggiate), per $\alpha = 0.05$ (a sinistra) e per $\alpha = 0.01$ (a destra)

COEFFICIENTE DI DETERMINAZIONE

Per una regressione lineare semplice, ma più in generale per qualsiasi regressione da quella curvilinea a quella lineare multipla, il coefficiente di determinazione r^2 è la proporzione di variazione spiegata dalla variabile dipendente sulla variazione totale:

$$r^2 = \frac{\text{Devianza dalla regressione}}{\text{Devianza totale}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

Espresso a volte in percentuale ed indicato in alcuni testi con \mathbf{R} oppure R^2 serve per misurare quanto la variabile indipendente X in media preveda la variabile dipendente Y

E' una misura che ha scopi prevalentemente descrittivi

La sua accettabilità non è legata ad inferenze statistiche, ma soprattutto agli scopi pratici, specifici dell'uso della regressione come metodo per prevedere Y conoscendo X

Il suo valore è tanto più elevato quanto più la retta passa vicino ai punti, fino a raggiungere 1 (oppure 100 se espressa in percentuale) quando i punti sperimentali sono collocati esattamente sulla retta e quindi ogni Y_i può essere predetto con precisione totale, senza alcun margine d'errore, quando sia noto il corrispondente valore di X_i

Nell'esempio con le 7 osservazioni su peso e altezza, è

$$r^2 = \frac{321,618}{403,715} = 0,797$$

Ciò significa che, noto il valore dell'altezza, quello del peso è stimato mediante la retta di regressione con una approssimazione di circa l'80 per cento; il restante 0,2 (rapportato a 1) oppure 20% è determinato dalla variabilità individuale di scostamento dalla retta

IPOTESI PER LA REGRESSIONE E LA CORRELAZIONE

Le ipotesi necessarie o condizioni di validità per l'analisi della regressione e della correlazione, che verrà trattata di seguito, sono analoghe a quelle già evidenziate per l'analisi della varianza e del test t di Student: normalità, omoschedasticità, indipendenza dall'errore

La condizione di **normalità** richiede che il valore di Y sia normalmente distribuito per ogni valore di X

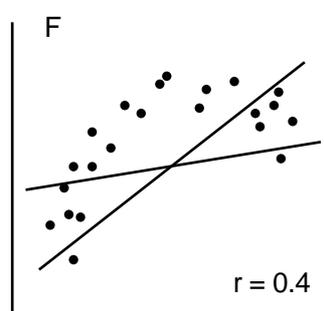
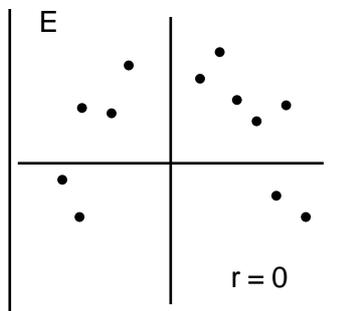
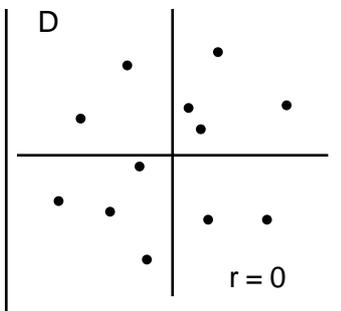
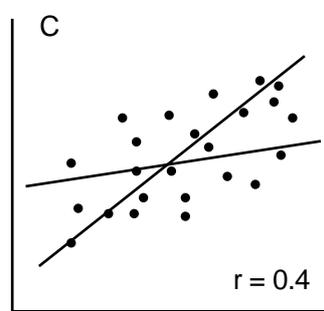
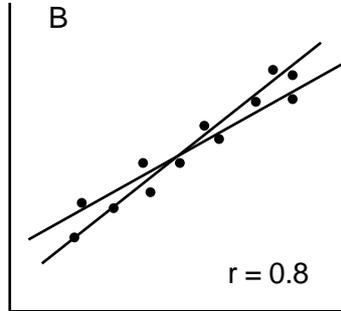
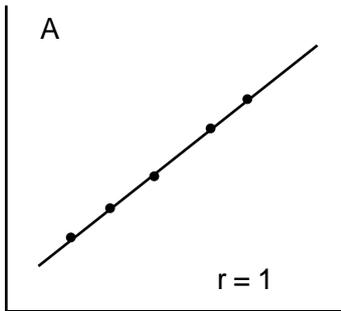
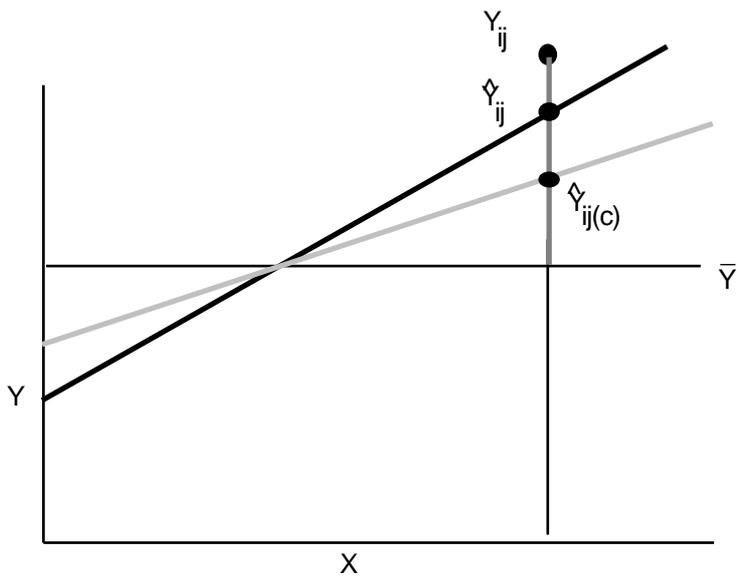
È una ipotesi facilmente comprensibile nel caso delle Y ripetute per lo stesso valore di X

Anche **l'analisi della regressione è robusta**, nel caso di deviazione dalla normalità: fino a quando la distribuzione dei valori di Y per lo stesso valore di X non si differenzia in modo estremo dalla normale, sia l'inferenza sulla regressione che quella sulla correlazione non sono eccessivamente distorte

La condizione di **omoschedasticità** richiede che le varianze delle disposizioni siano costanti per tutti i valori di X: i valori di Y devono variare nello stesso modo per qualunque valore di X

Sovente succede che all'aumentare delle X si abbia un aumento della varianza delle Y; come già esposto nell'analisi della varianza, le trasformazioni dei dati possono ricostruire questa ipotesi necessaria all'inferenza

La condizione di **indipendenza dell'errore** richiede che la distanza tra Y osservato ed Y previsto dalla regressione sia costante su tutto il campo di variazione della X



Metodo dei minimi quadrati - Impianto analitico

$$\begin{aligned}
 Q &= \sum (y_i - \mu)^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2 = \\
 &= \sum (y_i^2 + \beta_0^2 + \beta_1^2 x_i^2 - 2\beta_0 y_i - 2\beta_1 y_i x_i + 2\beta_0 \beta_1 x_i) = \\
 &= \sum y_i^2 + n\beta_0^2 + \beta_1^2 \sum x_i^2 - \\
 &\quad - 2\beta_0 \sum y_i - 2\beta_1 \sum x_i y_i + 2\beta_0 \beta_1 \sum x_i
 \end{aligned}$$

$$\frac{\partial Q}{\partial \beta_0} = 2n\beta_0 + 2\beta_1 \sum x_i - 2\sum y_i$$

$$\frac{\partial Q}{\partial \beta_1} = 2\beta_1 \sum x_i^2 + 2\beta_0 \sum x_i - 2\sum x_i y_i$$

uguagliando a zero i due risultati si ottiene un sistema di due equazioni a due incognite ...

$$\begin{cases}
 \beta_0 n + \beta_1 \sum x_i = \sum y_i \\
 \beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i
 \end{cases}$$

dalla prima equazione del sistema si ricava β_0 come :

$$\beta_0 = \left(\frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n} \right) = \bar{y} - \beta_1 \bar{x}$$

e per sostituzione nella seconda equazione si ricava β_1 come :

$$\left(\frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n} \right) \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \quad \dots$$

$$\beta_1 \left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \quad \dots = \frac{\text{codev}(xy)}{\text{dev}(x)}$$