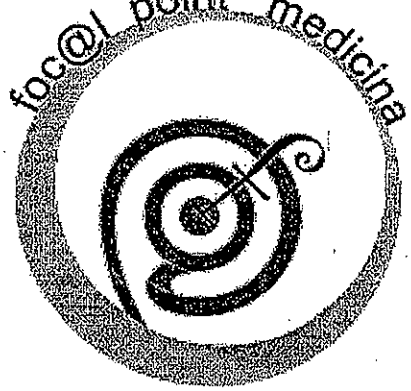
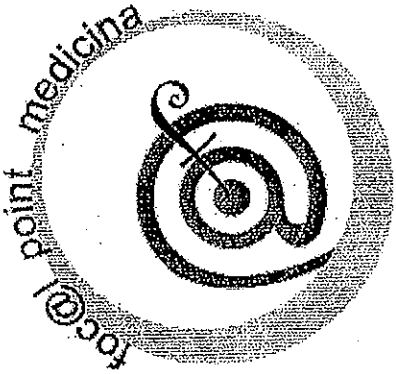


STATISTICA

(Prof.ssa Gloria)



Fulvia Gloria

APPUNTI ED ESERCIZI DI PROBABILITÀ E STATISTICA

Per i Corsi di

Statistica
Statistica Medica
Statistica Sociale

Anno Accademico 2008-2009

Capitolo 1

ELEMENTI DI INSIEMISTICA

Definizione di insieme

Un insieme è semplicemente una collezione di oggetti detti elementi dell'insieme. L'affermazione *s* è un elemento dell'insieme *S* si scrive:

$$s \in S$$

Se *A* e *B* sono insiemi, allora *A* è un sottoinsieme di *B* se e solo se ogni elemento di *A* è anche un elemento di *B*:

$$A \subseteq B \Leftrightarrow s \in A \Rightarrow s \in B$$

Per definizione, ogni insieme è completamente individuato dai suoi elementi. Pertanto, gli insiemi *A* e *B* sono uguali se e solo se hanno gli stessi elementi:

$$A = B \Leftrightarrow A \subseteq B \quad e \quad B \subseteq A$$

Operazioni tra insiemi

Unione

Dati due insiemi *A* e *B*, si dice unione di *A* e *B* l'insieme che contiene sia gli elementi di *A* che di *B*, cioè:

$$A \cup B = \{x:\ x \in A \quad e/o \quad x \in B\}$$

Esempio:

$$A = \{1, 2, 3\};\quad B = \{3, 5, 7\}$$

L'unione è:

$$A \cup B = \{1, 2, 3, 5, 7\}$$

Intersezione

Dati due insiemi A e B, si dice intersezione di A e B l'insieme che contiene gli elementi comuni ad A e B, cioè:

$$A \cap B = \{x:\ x \in A \ \ e \ \ x \in B\}$$

Esempio:

$G = \{2, 4, 3\}; \quad F = \{2, 6, 3\}$

l'intersezione è:

$$G \cap F = \{2,\ 3\}$$

L'intersezione tra due insiemi può essere anche vuota, ad esempio:

$$A = \{1, 2, 3\}; \quad B = \{4, 5\}$$

$$A \cap B = \emptyset$$

Prodotto Cartesiano

Dati due insiemi A e B, si dice prodotto cartesiano di A e B l'insieme che contiene le coppie ordinate il cui primo elemento sta in A ed il secondo in B, cioè:

$$A \times B = \{(x, y):\ x \in A \ \ e \ \ y \in B\}$$

Esempio:

$A = \{1, 2, 3\}; \quad B = \{5, 6, 3, 1\}$

$A \times B = \{(1,5), (1,6), (1,3), (1,1), (2,5), (2,6), (2,3), (2,1), (3,5), (3,6), (3,3), (3,1)\}$

Notiamo che il prodotto cartesiano non è commutativo, cioè $A \times B$ è diverso da $B \times A$, infatti:

$B \times A = \{(5,1), (5,2), (5,3), (6,1), (6,2), (6,3), (3,1), (3,2), (3,3), (1,1), (1,2), (1,3)\}$

Capitolo 2

CALCOLO COMBINATORIO

Il Calcolo Combinatorio è una branca della matematica ancillare al Calcolo delle Probabilità. Ha come scopo il calcolo dei modi con i quali possono essere raggruppati, secondo regole stabilite, gli elementi (sottoinsiemi) di uno o più insiemi.

Il numero degli elementi contenuti in un insieme o in un sottoinsieme è detto "ordine" dell'insieme o del sottoinsieme.

I sottoinsiemi possono essere:

- sottoinsiemi con ripetizione
- sottoinsiemi senza ripetizione

per esempio, partendo da un insieme costituito da quattro elementi A, B, C, D (insieme di ordine 4) vogliamo contare il numero di sottoinsiemi di ordine 2 (ossia costituiti da 2 elementi) che possono essere formati a partire dai quattro elementi dell'insieme dato:

- se consideriamo sottoinsiemi con ripetizione, dagli elementi di partenza si formeranno i sottoinsiemi: AA, BB, CC, DD, AB, AC, AD, BC, BD, CD.
- se, invece, consideriamo sottoinsiemi senza ripetizione avremo soltanto i sottoinsiemi: AB, AC, AD, BC, BD, CD

- sottoinsiemi dotati di ordine (disposizioni)
- sottoinsiemi non dotati di ordine (combinazioni)

Il problema di formare sottoinsiemi da un insieme è banale se il numero degli elementi dell'insieme è piccolo, poiché in questo caso è sufficiente scrivere esplicitamente tutti i possibili raggruppamenti e contarli, ma quando il numero di elementi è elevato la difficoltà consiste proprio nel formare tutti i raggruppamenti senza tralasciarne alcuno e senza cadere in ripetizioni.

Il calcolo combinatorio riveste notevole importanza nella matematica del discreto, ossia quando gli elementi appartengono a N (insieme dei numeri naturali).

I raggruppamenti di tali elementi possono essere fatti tenendo conto dell'ordine con il quale gli elementi sono scelti oppure può interessare solo la natura degli elementi. È bene analizzare lo scopo che ci si prefigge di raggiungere ed esaminare attentamente i dati prima di applicare ad essi il calcolo combinatorio.

RAGGRUPPAMENTI DEGLI ELEMENTI

Disposizioni

- Dato un insieme A di n elementi, si definiscono disposizioni di classe k i raggruppamenti di $k \leq n$ elementi dell'insieme A tali che ogni raggruppamento differisca dagli altri raggruppamenti:
  - per la natura degli elementi
  - per l'ordine degli elementi.

Le disposizioni si dicono:

- semplici, se ogni raggruppamento contiene elementi distinti tra loro. Tali disposizioni si indicano con $D_{n,k}$
- con ripetizione, se nei raggruppamenti gli elementi di A possono comparire più di una volta. Tali disposizioni si indicano con $D'_{n,k}$

Tratteremo solo disposizioni semplici

- $D_{n,1} = n$ poiché sono n i raggruppamenti di un solo elemento;

- $D_{n,2} = n(n-1)$ perché per formare i raggruppamenti di due elementi distinti, a ogni elemento si può associare uno degli $n-1$ elementi rimanenti dell'insieme, diversi da quello già considerato;

- $D_{n,3} = n(n-1)(n-2)$ in quanto per formare i raggruppamenti di tre elementi distinti si deve associare a ognuna delle $n(n-1)$ coppie già ottenute, uno degli $(n-2)$ elementi rimanenti dell'insieme, diversi da quelli già considerati. Per k qualsiasi, purché $k \leq n$, si ha:

- $D_{n,k} = n(n-1)(n-2)\ldots[n-(k-1)]$

Perciò: *il numero delle disposizioni semplici di n elementi di classe k è uguale al prodotto di k fattori interi consecutivi decrescenti a partire da n.*

Le disposizioni semplici di *n* elementi di classe *k* si possono esprimere per mezzo dei fattoriali. Infatti, dalla relazione:

$$D_{n,k} = n(n-1)..(n-k+1)$$

moltiplicando per $(n-k)!$ numeratore e denominatore si ricava:

$$D_{n,k} = \frac{n(n-1)..(n-k+1)(n-k)!}{(n-k)!} = \frac{n!}{(n-k)!}$$

La condizione $k \leq n$ per le disposizioni semplici è imposta dal fatto che si possono fare dei raggruppamenti formati con elementi tutti diversi solo se, al massimo, si prendono tutti gli elementi dell'insieme. Tale limitazione non esiste, ovviamente, per le disposizioni con ripetizione perché, in questo caso, gli elementi possono essere ripetuti quante volte si vuole.

Permutazioni

Partiamo da un esempio: "In un concorso sono stati selezionati tre concorrenti *A, B, C* per la graduatoria finale. Calcolare in quanti modi i tre concorrenti si possono presentare alla graduatoria finale". Al primo posto può esservi uno qualunque dei tre concorrenti, al secondo posto uno dei due rimanenti ed al terzo posto l'ultimo dei tre concorrenti.

Tutti i raggruppamenti possibili sono i seguenti: *ABC, ACB, BAC, BCA, CAB, CBA,* che sono in numero di $3 \cdot 2 \cdot 1 = 6$. In questo caso i raggruppamenti contengono *tutti* gli elementi dell'insieme e ogni raggruppamento differisce dagli altri *solo per l'ordine* secondo cui gli elementi sono presi. Raggruppamenti di questo tipo sono detti permutazioni.

- Dato un insieme *A* di *n* elementi, si definiscono permutazioni di *n* elementi (diversi fra loro) i raggruppamenti formati dagli *n* elementi di A presi in un ordine qualsiasi (le permutazioni sono disposizioni semplici di n elementi a n a n).

Quindi una permutazione differisce da un'altra *solo* per l'ordine degli elementi. Dalla definizione segue che le permutazioni coincidono con le disposizioni semplici di *n* elementi di classe *n*

$$P_n = D_{n,n} = n(n-1)(n-2)\ldots\ldots \quad 3 \cdot 2 \cdot 1$$

Il prodotto dei primi *n* numeri naturali s'indica con il simbolo *n!* (che si legge *n* fattoriale cioè si pone:

$$\cdot n(n-1) = n(n-1)\ldots \cdot 3 \cdot 2 \cdot 1 = n!$$

Il numero delle permutazioni di *n* elementi è:

$$P_n = n!$$

Osserviamo che *n!* è funzione di *n* e cresce rapidamente al crescere di *n*.

Combinazioni

"In una classe di 25 studenti si vogliono scegliere 2 allievi come rappresentanti di classe. In quanti modi è possibile effettuare la scelta?" Il problema chiede di scegliere due allievi fra 25, ma la scelta non implica un ordinamento. Una qualunque coppia è detta una combinazione, e differisce da un'altra coppia per almeno un elemento che la compone. Precisamente si dà la seguente definizione:

- Dato un insieme *A* di *n* elementi, si definiscono combinazioni semplici degli *n* elementi di classe *k* i raggruppamenti di *k* elementi, scelti fra gli *n* dell'insieme *A*, tali che ogni raggruppamento differisca dagli altri per la natura degli elementi (senza considerare l'ordine degli elementi).

Si deve notare la differenza fra disposizioni e combinazioni semplici: mentre nelle disposizioni si tiene conto dell'ordine, nelle combinazioni non se ne tiene conto.

Per determinare il numero di combinazioni di *n* elementi di classe *k* ricaviamo una formula che esprime il legame fra il numero delle combinazioni e quello delle disposizioni di *n* elementi a *k*, a *k*.

Detta formula si ottiene osservando che le disposizioni di *n* elementi di classe *k* si ottengono dalle combinazioni di *n* elementi di classe *k*, permutando fra loro i k elementi che costituiscono ciascun raggruppamento.

Indicato con $C_{n,k}$ il numero delle combinazioni semplici di n elementi di classe k, si ha: $C_{n,k} \cdot P_k = D_{n,k}$

Cioè: $C_{n,k} = \frac{D_{n,k}}{P_k} = \frac{n(n-1)(n-2)...[n-(k-1)]}{k!} = \frac{n!}{(n-k)!k!}$

Solitamente si scrive: $C_{n,k} = \binom{n}{k}$

Il simbolo $\binom{n}{k}$ è detto coefficiente binomiale per il suo uso nello sviluppo delle potenze del binomio.

Riassumendo:

- Nel caso di sottoinsiemi ordinati il sottoinsieme AB è differente dal sottoinsieme BA e si parla di DISPOSIZIONI,
- Nel caso di sottoinsiemi non ordinati il sottoinsieme AB è uguale al sottoinsieme BA e si parla di COMBINAZIONI.

È chiaro che a parità di ordine dei sottoinsiemi il numero delle disposizioni è maggiore di quello delle combinazioni; infatti, a partire da una combinazione possiamo ottenere diverse disposizioni che pur avendo gli stessi elementi, si presentano con un ordine differente.

Capitolo 3

ELEMENTI DI CALCOLO DELLE PROBABILITÀ

SPAZIO CAMPIONARIO ED EVENTI

Lo spazio campionario è l'insieme S di tutti i possibili esiti di un esperimento.

Per esempio, se l'esperimento consiste nel lanciare un dado a sei facce e nel registrare il risultato, lo spazio campionario è $S = \{1, 2, 3, 4, 5, 6\}$ (l'insieme dei possibili esiti).

Calcolo delle probabilità

Il calcolo delle probabilità ci aiuta a studiare eventi casuali e non deterministici cioè quegli eventi che possono o non possono verificarsi, che dipendono unicamente dal caso (come ad esempio quale faccia esce dal lancio di un dado?).

Annotazione storica

Nel Seicento il calcolo delle probabilità nasce per risolvere alcuni problemi sui giochi d'azzardo (dadi).

Uno dei primi trattati di calcolo delle probabilità risale a J. Bernoulli (Ars conjectandi, 1713). Laplace da la prima impostazione sistematica della concezione classica.

Nell'Ottocento, J. Venn e A. Cournot presentano la concezione frequentista basata su prove ripetute di un sperimento oggetto di studio

Nel Novecento sorge la concezione soggettiva, che stabilisce la probabilità di un evento in base alle informazioni che un individuo ha sull'evento.

Nel nostro secolo si afferma l'impostazione assiomatica, dovuta a A. N. Kolmogorov e altri, che sviluppa tutta la teoria della probabilità partendo da due concetti primitivi: evento e probabilità, e assegnando alcuni assiomi

La probabilità P (E) di un evento E è il rapporto fra il numero *m* dei casi favorevoli (al verificarsi di E) e il numero *n* dei casi possibili, giudicati egualmente probabili.

$$P(E) = m/n \quad \text{con} \quad n \neq 0 \quad e \quad 0 \leq P(E) \leq 1$$

Esempio: Qual è la probabilità che lanciando un dado non truccato esca la faccia 6?: P(6)=1/6 (tutte le facce del dado hanno la stessa probabilità di uscire)

Vi sono state diverse impostazioni del calcolo delle probabilità:

-L'impostazione classica: $P(E) = m/n$ con $n \neq 0$ e $0 \leq P(E) \leq 1$ con la limitazione che i casi possibili devono essere tutti ugualmente probabili.

-L'impostazione frequentista: $P(E) = \lim m/n$ per $n \to \infty$ con la limitazione che nella realtà è impossibile fare infinite prove

-L'impostazione soggettiva: P(E)= grado di fiducia che un individuo attribuisce al verificarsi dell'evento E in base alle informazioni che ha, con la limitazione che varia da individuo ad individuo.

-L'impostazione assiomatica:P(E)= numero reale tale che $P(E) \geq 0$, P(S) =1 (S è lo spazio degli eventi) e se due eventi E_1 e E_2 hanno intersezione vuota, $P(E_1 \cap E_2) = P(E_1) + P(E_2)$.

Valgono alcuni principi del calcolo delle probabilità (impostazione classica):

PRINCIPIO DELLA PROBABILITÀ COMPOSTA

Se un evento E è costituito dal verificarsi simultaneo o in successione di due o più eventi E_1, E_2, ... E_n tra loro indipendenti $P(E)=P(E_1) \times P(E_2) \times .....$ $P(E_n)$, ossia P(E) è uguale al prodotto delle probabilità relative ai singoli eventi che compongono l'evento stesso.

PRINCIPIO DELLA PROBABILITÀ TOTALE

Se un evento E si può manifestare secondo due o più modalità E_1, E_2,.... E_n tra loro mutualmente escludentesi (gli eventi con intersezione vuota sono detti disgiunti o incompatibili o escludentesi) $P(E)=P(E_1)+P(E_2)+...+P(E_n)$,ossia P(E)= somma delle probabilità relative alle singole modalità E_1, E_2,.... E_n secondo le quali l'evento E può manifestarsi.

PRINCIPIO DI INDIPENDENZA (SECONDO L'IMPOSTAZIONE ASSIOMATICA)

Due eventi, A e B, sono indipendenti tra loro se e solo se la probabilità della loro intersezione è uguale al prodotto delle probabilità relative ai due eventi,

A e B sono indipendenti tra loro <=> $P(A \cap B) = P(A) \times P(B)$.

- Due eventi A e B disgiunti non sono tra loro indipendenti, infatti: $P(A \cap B) = 0$ mentre $P(A) \neq 0$ e $P(B) \neq 0$ e quindi $P(A) \times P(B) > 0$.

-Due eventi con intersezione non vuota possono essere dipendenti o indipendenti tra loro. La presenza di una intersezione non vuota è condizione necessaria ma non sufficiente perchè due eventi siano fra loro indipendenti.

ODDS

Nel campo delle scommesse vengono usati gli odds (odds in inglese significa pronostico). Se,per esempio, diciamo che gli odds a favore di una certa squadra sono 4:1 diciamo che tale squadra ha l'80% di probabilità di vincere.

In **statistica** Odds equivale al rapporto tra la probabilità p che l'evento accada sulla probabilita complementare (1-p) che l'evento non accada

Odds (a favore di E) = p/1-p ossia Odds (a favore di E)=m/n-m

Esempio: Su 1000 donne sottoposte a taglio cesareo (TC) negli Stati Uniti tra 1998 e 2001 si è rivelato una mortalità perinatale di 1.77 per 1000.

Odds (a favore di TC)= (1.77/1000) / 1- (998.23/1000) = 1.77/998.23=0.001773

L'odds è leggermente maggiore della probabilità. Per piccoli valori della probabilità la differenza tra odds e probabilità è minima, ma se la probabilità è alta la differenza tra odds e probabilità diventa consistente.

RISCHIO ASSOLUTO (RA)

La probabilità di un evento negativo può essere vista come Rischio (R) che accada quell'evento per cui

R(di morte per TC)= 1.77/1000=0.00177

RISCHIO RELATIVO e ODDS RATIO (R.R. e O.R.)

Il rischio relativo è il rapporto tra due rischi assoluti. Ad esempio, un uomo di 55 anni, senza diabete, con colesterolemia di 180 mg/dl, HDL pari a 45 mg/dl e pressione sistolica di 120 mmHg, se fuma ha un Rischio Assoluto di coronaropatie pari a 12.6 % mentre se non fuma ha un Rischio Assoluto di 7.7%, il Rischio Relativo di coronaropatie nei fumatori è pari a 12.6/7.7= 1.64.

I fattori di rischio possono essere di diversa natura,tossine, agenti infettivi, medicine, fattori comportamentali, fattori ereditari.

Ci sono due tipi di esperimenti per valutare il rischio relativo: uno in cui si esamina se un gruppo di persone all'inizio sane si ammala dopo esposizione ad un fattore di rischio (la durata dell'esposizione può variare a seconda del tipo di malattia) a paragone di un gruppo di controllo non esposto al fattore di rischio ed un altro in cui si esamina un gruppo di malati (casi) per vedere se l'esposizione ad un fattore di rischio sia associato alla malattia. In questo secondo tipo di esperimento il risultato sui "casi" viene confrontato con il risultato sul gruppo di "controllo" non esposto al fattore di rischio. Il campione dei casi deve essere abbastanza grande e i due gruppi (casi e controlli) devono essere confrontabili eccetto che per il fattore di rischio.

Nel primo tipo di esperimento lo studio viene condotto per coorte. Si usa il termine "coorte" per descrivere un gruppo di persone che hanno qualche aspetto in comune al momento della raccolta (per esempio sono persone sane) e che vengono osservate per un periodo di tempo. In questo tipo di esperimento si preferisce usare come indice di malattia il Rischio Relativo (R.R.) come rapporto di probabilità. R.R. ci dice proprio quante volte è più probabile una certa malattia nelle persone esposte ad un dato fattore di rischio rispetto alle persone non esposte.

Data la tabella

		Malattia	
		si	no
Fattore di rischio	si	a	b
	no	c	d

$$R.R. = \frac{\text{probabilità di malattia negli esposti}}{\text{probabilità di malattia nei non esposti}} = \frac{a/(a+b)}{c/(c+d)}$$

R.R. varia tra 0 e $+\infty$

R.R.=1 [a/(a+b)=c/(c+d)] l'evento è ugualmente probabile in ambedue i gruppi

R.R.>1 [a/(a+b)>c/(c+d)] l'evento è più probabile nel gruppo degli esposti

R.R.<1 [a/(a+b)<c/(c+d)] l'evento è più probabile nel gruppo dei non esposti

Nel secondo tipo di esperimento, ossia in uno studio caso-controllo, si preferisce usare come indice del rischio relativo l'ODDS RATIO (O.R.)

La tabella 2x2 sarà di questo tipo:

		casi	controlli
Fattore di rischio	si	a	b
	no	c	d

$$O.R. = \frac{\text{odds a favore del fattore di rischio nei casi}}{\text{odds a favore del fattore di rischio nei controlli}}$$

$$= (a/c) / (b/d) = ad/bc$$

O.R. varia tra 0 e + ∞

O.R.=1 (a/c = b/d) indica mancanza di associazione tra causa ed effetto

O.R.>1 (a/c > b/d) indica l'esistenza di un'associazione positiva tra causa ed effetto (l'esposizione favorisce la malattia)

O.R.<1 (a/c < b/d) indica l'esistenza di una associazione negativa tra causa ed effetto(l'esposizione protegge dalla malattia)

L'odds ratio di un trattamento è il rapporto tra la frequenza con la quale un evento si verifica in un gruppo di pazienti e la frequenza con la quale lo stesso evento si verifica in un gruppo di controllo.

Dal punto di vista concettuale l' odds ratio ed il rischio relativo sono relativamente simili ma dal punto di vista quantitativo (valore numerico) lo sono solo per eventi rari.

Capitolo 4

SPAZI DI PROBABILITA'

Gli elementi costitutivi di uno spazio di probabilità sono lo *spazio degli eventi S* ed *una funzione che assegna ad ogni evento la sua probabilità.*

Lo spazio degli eventi può essere di natura numerica o non numerica, discreto o continuo. A sua volta uno spazio di probabilità discreto può essere finito o infinito.

Media, Varianza e deviazione standard di una variabile casuale

Se lo spazio degli eventi è di natura numerica possono essere definite la media, la varianza e la deviazione standard.

Spazi discreti e finiti

Uno spazio di probabilità discreto e finito può essere rappresentato da questa tabella:

x_1	x_2	.....	x_n
$f\,x_1$	$f\,x_2$	.....	$f\,x_n$

(eventi numerici)

(probabilità)

La media μ della distribuzione è uguale alla somma dei prodotti di ogni valore della variabile casuale X per la probabilità di ogni valore, ossia

$$\mu=\sum_{i=1}^{n} x_i f(x_i)$$

La varianza ($Var=\sigma^2$) della distribuzione di probabilità è uguale alla somma dei prodotti di ogni differenza al quadrato del valore della variabile casuale dalla media μ per la sua probabilità, ossia

$$Var=\sigma^2=\sum_{i=1}^{n} (x_i - \mu)^2 f(x_i)$$

Spazi discreti e infiniti (infinito numerabile)

$$\mu=\sum_{i=1}^{\infty} x_i f(x_i)$$

$$Var=\sigma^2=\sum_{i=1}^{\infty} (x_i - \mu)^2 f(x_i)$$

con la limitazione della convergenza della serie.

Spazi continui

Bisogna ricordare che la probabilità è riferita ad intervalli dell'asse reale ed è l'integrale della funzione con estremi corrispondenti all'intervallo.

$$P(a\text{-}b) = \int f(x)dx \qquad \text{con} \quad a \leq x \leq b$$

$$\mu = \int_R x\, f(x)\, dx$$

$$Var = \sigma^2 = \int_R (x - \mu)^2 f(x)\, dx$$

DISTRIBUZIONE BINOMIALE

Nella sua *Ars Conjectandi*, lo svizzero Jakob Bernoulli (1654-1705) formulò una legge matematica che costituisce la base teorica della *distribuzione binomiale* e che oggi è considerata uno dei fondamenti del calcolo della probabilità.

Consideriamo n prove ripetute e indipendenti di un esperimento con due esiti; definiamo uno dei due esiti *successo*, e definiamo l'altro *insuccesso*. Sia p la probabilità del successo, cosicché $q = 1-p$ è la probabilità dell' insuccesso. Se ci interessa solo il numero dei successi e non l'ordine in cui essi si presentano, allora vale il seguente teorema:

- La probabilità di ottenere esattamente k successi in n prove ripetute viene indicata e calcolata mediante l'espressione:

$$b(k;n,p) = \binom{n}{k} p^k\, q^{n-k}$$

$\binom{n}{k}$ è il coefficiente binomiale. Si noti che la probabilità che non si verifichi alcun successo è q^n, e quindi la probabilità che si verifichi almeno un successo è $1-q^n$.

Se consideriamo n e p costanti, allora $b(n,p,k) = b(k) = P(k)$ è una funzione di probabilità solo di k.

k	0	1	2	...	N
$P(k)$	q^n	$\binom{n}{1}q^{n-1}p$	$\binom{n}{1}q^{n-2}p^2$	...	p^n

$b(k,n,k)$ è detta *distribuzione* binomiale poiché per $k=0,1,2,...,n$ essa corrisponde ai termini successivi dello sviluppo del binomio:

$$(q+p)^n = q^n + \binom{n}{1}q^{n-1}p + \binom{n}{2}q^{n-2}p^2 + ... + p^n$$

Questa distribuzione è detta anche distribuzione di Bernoulli, e le prove indipendenti con due esiti sono dette prove di Bernoulli.

Le proprietà della distribuzione binomiale sono:

Valor medio $\mu = np$

Varianza $\sigma^2 = npq$

Deviazione standard $\sigma = \sqrt{npq}$

Riassumendo: uno schema di Bernoulli possiede, in sostanza, le seguenti caratteristiche:

a) ogni prova è un esperimento casuale che può avere soltanto due esiti possibili, con probabilità effettive p e $q=1-p$;

b) ogni prova effettuata è indipendente da ogni altra prova

c) per ogni prova la probabilità p del successo è costante.

In generale, se i lanci sono un numero qualsiasi n e se i successi (per esempio che esca testa) sono un qualsiasi numero k, se indichiamo con p la probabilità

del successo (evento) e con q= 1-p la probabilità dell'insuccesso (evento opposto), la probabilità di avere k successi sarà uguale a :

$$b(n,p,k) = P(k) = \binom{n}{k} \times p^k \times q^{n-k}$$

Questa formula vale non solo per il gioco "testa o croce", ma anche per tutte le variabili aleatorie con due sole possibilità. Ad esempio: qual è la probabilità di avere due figli maschi in una famiglia di 3 figli?(n=3, p=1/2,k=2). Oppure: qual è la probabilità in una famiglia di 6 figli in cui tutti e due i genitori sono portatori dei tratto talassemico di avere 4 figli malati? In questo caso n=6; p=1/4; k=4.

Gli eventi possibili sono n+1 e corrispondono ai seguenti valori di k: 0,1,2,3.....n.

- "Qual è la probabilità di avere su 5 lanci 2 volte testa e 3 volte croce a condizione che si abbia nei primi due lanci testa e negli altri 3 croce?" La combinazione è una sola e quindi $P(TT) = 1 \cdot (1/2)^2 \cdot (1/2)^3$. Questo equivale a $P(TTCCC) = 1/2 \cdot 1/2 \cdot 1/2 \cdot 1/2 \cdot 1/2$ per il principio della probabilità composta poiché sono eventi indipendenti.
- "Qual è la probabilità di avere su 5 lanci 2 volte testa e 3 volte croce?" Poiché i due eventi "successo (testa)" possono presentarsi in diversi modi (nel primo e nel secondo lancio, nel primo e nel quarto lancio, nel secondo e terzo lancio...e così via), dobbiamo moltiplicare la probabilità dell'evento elementare per un coefficiente che esprime il numero delle combinazioni di n oggetti a "k" a "k" e che nel caso dei nostro esempio è uguale al numero delle combinazioni di 5 oggetti a 2 a 2:

  $$P(TT) = \binom{5}{2} \cdot (1/2)^2 \cdot (1/2)^3$$

DISTRIBUZIONE NORMALE

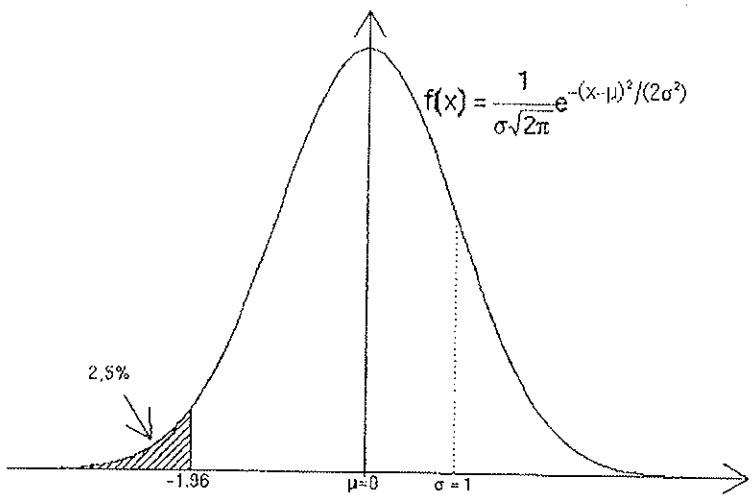
Deve il nome a Karl Friederick Gauss, che la propose per la descrizione delle deviazioni delle misure astronomiche rispetto al loro andamento medio. Egli ipotizzò, infatti, che tali deviazioni fossero dovute ad *errori casuali di misura* e, in base ad argomenti abbastanza generali, derivò una funzione densità di probabilità per gli errori casuali per cui viene comunemente detto che gli errori casuali "seguono normalmente" tale distribuzione e la distribuzione stessa è stata chiamata *distribuzione normale (detta anche curva di Gauss o curva degli errori).* Essa è una distribuzione continua con due parametri, indicata con $N(\mu,\sigma^2)$ ed è una distribuzione statistica teorica che ben riproduce alcune delle caratteristiche di una popolazione (es. altezza). In particolare si dice che un carattere (o una variabile) si distribuisce secondo una (distribuzione) normale quando nella popolazione la maggior parte degli individui presentano valori centrali del carattere, mentre i rimanenti individui presentano i valori estremi a destra e a sinistra dei centrali. È una distribuzione simmetrica, con entrambe le code che tendono ad infinito con la caratteristica forma a campana; ha uguali media, moda e mediana; la sua forma è determinata completamente dalla media μ e dalla varianza σ^2 (la figura mostra alcune curve normali che differiscono per i valori di media e varianza).

La curva normale ha la seguente funzione di probabilità:

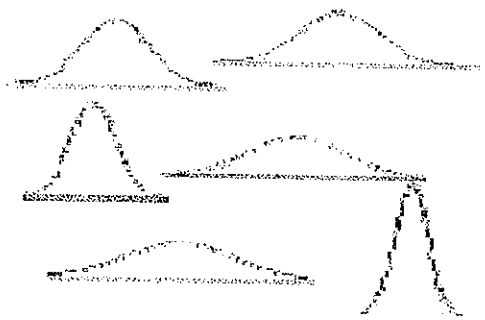
$$f(x)=\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

con $-\infty < x < \infty$

dove μ e σ^2. sono appunto la media e la varianza.

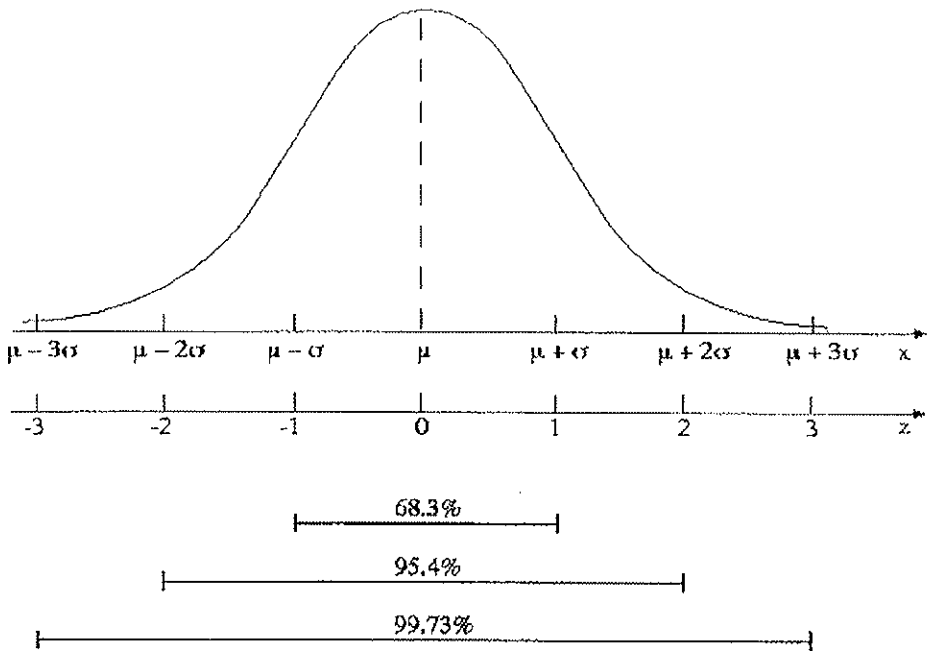


Alcuni esempi di gaussiane:



Gli eventi sono segmenti della x e l'area sovrastante il segmento rappresenta la probabilità dell'evento stesso. L'area totale è uguale a 1.

PROPRIETA' DELLA CURVA NORMALE



$$68{,}3\% = P\{\mu - 1\sigma \prec X \prec \mu + 1\sigma\}$$

$$95{,}0\% = P\{\mu - 1.96\,\sigma \prec X \prec \mu + 1.96\sigma\}$$

$$99{,}0\% = P\{\mu - 2.58\,\sigma \prec X \prec \mu + 2.58\sigma\}$$

I numeri in neretto vengono detti valori critici di z e vengono indicati con z_c.

Per semplicità prenderemo

$$68\% = P\{\mu - 1\sigma \prec X \prec \mu + 1\sigma\}$$

$$95\% = P\{\mu - 2\sigma \prec X \prec \mu + 2\sigma\}$$

$$99\% = P\{\mu - 3\sigma \prec X \prec \mu + 3\sigma\}$$

Capitolo 5

STATISTICA DESCRITTIVA

ELEMENTI DI BASE

Il problema centrale della statistica è quello di affrontare grandi quantità di informazioni relative agli oggetti della propria indagine, teoricamente disponibili ma di fatto difficilmente gestibili. Tutte le informazioni, perchè contribuiscano effettivamente ad accrescere la conoscenza di un fenomeno hanno bisogno di essere rilevate accuratamente, devono essere selezionate, organizzate e sintetizzate.

I dati (informazioni): costituiscono il materiale di base della statistica. Essi vengono sempre ricondotti a numeri.
Unità statistica: è il più piccolo componente di un insieme di "soggetti" che si vuole esaminare.
Variabile: è una caratteristica in un insieme di "soggetti", che, in generale, assume valori differenti tra i vari "soggetti". Una variabile può essere: numerica (quantitativa) o nominale (qualitativa).
Campione: é un gruppo di soggetti estratto da una popolazione. Esso deve essere rappresentativo della popolazione.
Popolazione é un insieme, finito o infinito, di soggetti di natura qualsiasi che noi siamo interessati a studiare.

La statistica descrittiva è la branca della Statistica che studia i criteri di rilevazione, di classificazione e di sintesi delle informazioni relative a una popolazione oggetto di studio. Quando si effettua una statistica descrittiva ci si può riferire:

- ad un solo fenomeno e si parlerà allora di statistica descrittiva univariata
- a due fenomeni e si parlerà allora di statistica descrittiva bivariata
- a tre o più fenomeni e si parlerà allora di statistica descrittiva multivariata

Se si effettua una statistica descrittiva univariata essa sarà:

- qualitativa, se la rilevazione del fenomeno viene effettuata su scala di modalità di tipo nominale o ordinale;
- quantitativa, se la rilevazione del fenomeno viene effettuata su scala di modalità di tipo intervallare o di rapporto.

Se si effettua una statistica descrittiva bivariata o multivariata essa sarà:

- qualitativa quando i fenomeni oggetto di studio sono entrambi di tipo qualitativo;
- quantitativa quando i fenomeni oggetto di studio sono entrambi di tipo quantitativo;

e così via.

Con variabili qualitative si costruiscono:

1. *Tabelle di frequenza* mediante calcolo di:
    - Frequenze assolute (numero di casi per ogni categoria)
    - Frequenze relative (rapporto della frequenza assoluta sul totale)
    - Frequenze cumulate (somma delle frequenze di ogni categoria alle modalità precedenti)
    - Frequenze 'valide' [frequenze (relative o cumulate) calcolate sul totale senza eventuali valori mancanti]

2. *Tabelle di contingenza* (Cross-tabulations) mediante calcolo di:
    - Frequenze assolute
    - Frequenze relative al totale di riga

- Frequenze relative al totale di colonna
- Frequenze relative al totale generale

Se il campione è abbastanza grande, la legge dei grandi numeri dice che possiamo considerare la frequenza relativa dell'evento uguale alla sua probabilità statistica.

Con variabili quantitative si calcolano:

1. *Frequenze*
2. *Indici della tendenza centrale:*
   - Media aritmetica
   - Moda
   - Mediana
3. *Indici della dispersione:*
   - Scarto semplice o range
   - Scarto quadratico medio o Deviazione Standard
   - Varianza
   - Percentili

INDICI DELLA STATISTICA DESCRITTIVA

INDICI DELLA TENDENZA CENTRALE

MEDIA ARITMETICA

Se si considera una serie di n termini $x_1, x_2, ...x_n$, la media aritmetica $\bar{x}$, è data dalla somma degli n termini diviso la loro numerosità.

In simboli:

$$\bar{x} = (x_1 + x_2 + ...x_n)/n$$

Essa viene indicata con μ se calcolata su una popolazione e con $\bar{x}$ se calcolata su un campione.

Se i dati sono raggruppati in classi, le x_i sono i valori centrali delle classi.

MODA

La moda m_0 rappresenta il valore (o la classe) avente massima frequenza. In una distribuzione possono esservi più mode.

MEDIANA

La mediana M è il valore centrale dei valori assunti da una variabile dopo che tali valori sono stati ordinati. Per esempio, se abbiamo 99 individui ordinati secondo la statura, la mediana sarà il valore dell'altezza dell'individuo che si trova al 50° posto.

- Se i valori dei parametri sono in numero dispari la mediana sarà esattamente il valore centrale
- Se i valori sono in numero pari, si prendono i due valori centrali e la mediana sarà la media aritmetica di essi.

Nel caso di una distribuzione simmetrica media, mediana e moda coincidono.

INDICI DELLA DISPERSIONE

SCARTO SEMPLICE O RANGE (R)

È l'indice di dispersione più semplice da calcolare ed è dato dalla *differenza fra il maggiore e il minore dei valori rilevati.* Talvolta, il campo di variazione si esprime indicando gli estremi dell'intervallo invece della differenza fra il maggiore e il minore dei valori rilevati. Il campo di variazione è un indice molto semplice da calcolare ma di scarsa importanza perché tiene conto in un insieme di dati solo dei valori estremi e non degli altri.

DEVIAZIONE STANDARD

Data una popolazione, consideriamo l'insieme di tutti i valori di una certa variabile e gli scarti di tali valori dalla media aritmetica, ossia consideriamo le differenze $x_i - \mu$. Queste differenze vanno elevate al quadrato per ovviare al fatto che differenze positive possano elidersi con differenze negative. Poiché in una distribuzione di valori possono essere presenti valori che si discostano moltissimo o per niente dalla media aritmetica, si calcola un valore medio degli scarti . Questo valore è la VARIANZA. La radice quadrata del valore medio della somma

dei quadrati degli scarti dalla media aritmetica si chiama DEVIAZIONE STANDARD.

- Se si sta esaminando una popolazione o un grande campione la deviazione standard si indica con σ

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

- Se si sta esaminando un piccolo campione la deviazione standard si indica con s:

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

Tale indice è tanto più piccolo quanto più i dati sono prossimi al valore medio ed è uguale a zero se e solo se i dati sono tutti eguali fra loro; è un indice i molto sensibile per misurare la distanza di dati che si scostano molto dal valore medio.

PERCENTILE

Il concetto di percentile generalizza quello di mediana *(la mediana è il dato che separa il primo 50% dei dati (ordinati) dai rimanenti dati).*

Se i dati sono suddivisi in quattro parti uguali si hanno i quartili Q1, Q2, Q3. Essi vengono definiti come quei valori che, in una seriazione ordinata, separano il 25%, 50%, 75% delle osservazioni.

I decili $D_1, D_2, \ldots D_9$ ed i centili $C_1, C_2 \ldots C_{99}$ si ottengono dividendo la seriazione rispettivamente in 10 ed in 100 parti.

RAPPRESENTAZIONI GRAFICHE

In un esperimento la quantità che controlliamo o che deliberatamente facciamo variare costituisce la variabile indipendente e viene posta sull'asse delle ascisse (asse orizzontale o *asse x*). La quantità che varia in corrispondenza delle

variazioni della variabile indipendente è detta variabile dipendente e viene rappresentata sull'asse delle ordinate (asse verticale o *asse y*).

Quando si analizzano dei dati ottenuti da un determinato esperimento, una rappresentazione grafica ci aiuta meglio di una tabulazione ad estrarre informazioni su come i dati sono distribuiti e sulle loro eventuali correlazioni.
Questo fatto è ancora più evidente nel momento in cui l'insieme dei dati da trattare è costituito da un numero elevato di valori. Le rappresentazioni grafiche, quindi, costituiscono un eccellente mezzo per riassumere molte caratteristiche di un esperimento.

Un grafico può dare indicazioni su:

- l'intervallo (valore minimo e massimo) di una/più misure;
- l'esistenza o meno di relazioni tra i dati. Per esempio i punti in un grafico potrebbero giacere tutti su di una linea retta, una curva, oppure riempire il grafico in modo completamente casuale;
- i punti che si discostano in modo sensibile dall'andamento della maggior parte dei dati.

Essendo le rappresentazioni grafiche dei dati un mezzo di comunicazione visiva, permettono al lettore di cogliere immediatamente l'andamento di una variabile. Perciò devono fornire al lettore un'informazione sintetica e facile da interpretare e devono essere di supporto durante la lettura dei risultati. È sempre bene specificare su entrambi gli assi la natura delle variabili, l'unità di misura ed il loro orientamento. Per la comprensione di un grafico è anche utile apporre un titolo che brevemente riassuma gli elementi posti su esso.

Una rappresentazione grafica diventa indispensabile nello studio di fenomeni di elevate dimensioni, infatti, una lunga serie di dati non è sempre idonea alla comprensione.

SCELTA DI UNA RAPPRESENTAZIONE GRAFICA

La scelta di una rappresentazione grafica dipende dal fenomeno che stiamo studiando in quanto un grafico deve essere adeguato al tipo di variabile che stiamo esaminando. Ne trattiamo solo alcune.

DIAGRAMMA CIRCOLARE

Il diagramma circolare viene usato per variabili qualitative nominali.

Viene rappresentato generalmente con un cerchio suddiviso in settori proporzionali alle frequenze di ogni modalità della variabile in studio. Il cerchio rappresenta la totalità delle frequenze assolute.

È molto usato nella statistica descrittiva di fenomeni di natura economica, demografica e sanitaria e mette bene in evidenza la ripartizione dell'insieme.

DIAGRAMMA A BARRE

Il diagramma a barre viene usato per variabili qualitative ordinali o quantitative discrete.

- Se stiamo esaminando una variabile qualitativa ordinale, sull'asse x vengono riportate le varie modalità della variabile, preferibilmente a uguale distanza fra foro.
- Se stiamo esaminando una variabile quantitativa discreta, sull'asse x vengono riportati i singoli valori della variabile rispettando le loro distanze in relazione all'unità grafica scelta.

ISTOGRAMMA DI FREQUENZA

L'istogramma di frequenza viene usato per variabili quantitative continue

DIAGRAMMI CARTESIANI (POLIGONO DI FREQUENZA ED OGIVA)

- Il poligono di frequenza è il grafico lineare dell'istogramma di frequenza e si usa se siamo interessati alle frequenze assolute o relative di una variabile quantitativa continua. Si disegna unendo i valori medi di ogni classe presi sulla base superiore di ogni rettangolo di un istogramma di frequenze assolute o relative.
- L'ogiva è la rappresentazione grafica delle frequenze cumulate assolute o relative di una variabile quantitativa continua e si costruisce unendo i limiti superiori di ogni rettangolo di un istogramma di frequenza cumulativa o cumulativa relativa.

DIFFERENZA TRA SPAZI NUMERICI DISCRETI E SPAZI NUMERICI CONTINUI

- Nel caso di spazi di probabilità discreti e finiti: gli eventi x_i sono in numero finito. (È chiaro che se gli eventi non sono numerici, non ha alcun senso parlare di media).
- Nel caso di spazi di probabilità discreti e infiniti (infinità numerabile), le x_i sono una serie di valori per cui bisogna verificare la loro convergenza altrimenti la media non corrisponde ad un valore finito.
- Nel caso di spazi di probabilità continui la sommatoria che compare nella formula di μ e σ si indica con il simbolo di integrale, ma la struttura è fondamentalmente la stessa.

  Esempi di spazi continui sono quelli rappresentati dalla distribuzione di parametri fisiologici quali peso, azotemia, pressione arteriosa per i quali possiamo immaginare che il parametro possa assumere tutti i valori compresi tra il minimo ed il massimo valore
- Nel caso di spazi di probabilità discreti e finiti o infiniti viene considerata la probabilità relativa ai singoli eventi numerici.
- Nel caso di spazi di probabilità continui la probabilità viene riferita ad intervalli di eventi numerici. In altre parole gli eventi sono intervalli di numeri. In questo caso la probabilità è data dall'area al di sopra dell'intervallo numerico ossia è data dall'integrale della nostra funzione calcolato nell'intervallo [a,b], dove a e b costituiscono gli estremi del nostro intervallo.
- La rappresentazione grafica di una distribuzione di probabilità discreta è un insieme di bastoncini (o rettangoli), la cui altezza (di solito sull'asse y) rappresenta la probabilità del singolo evento numerico e la base (di solito sull'asse x) ogni valore della variabile discreta. *Non può essere mai descritta da una linea continua.*
- La rappresentazione grafica di una distribuzione di probabilità continua può essere descritta da una linea continua.

FREQUENZA

In statistica si definiscono 2 tipi di frequenze:

- *FREQUENZA ASSOLUTA* (p): è il numero di volte che si verifica un evento a prescindere dal numero totale delle prove.
- *FREQUENZA RELATIVA* ($f = p / n$): è il rapporto tra il numero di volte che si verifica un evento ed il numero totale delle prove.

La legge empirica del caso, o legge dei grandi numeri, ci permette di confondere la frequenza f con la probabilità dell'evento, purché il numero di prove sia molto grande.

- *FREQUENZA ASSOLUTA CUMULATIVA* (P): frequenza assoluta di una classe più le frequenze assolute delle classi precedenti.
- *FREQUENZA RELATIVA CUMULATIVA* (F): frequenza relativa di una classe più le frequenze relative delle classi precedenti.

- Se stiamo studiando una variabile qualitativa (categorica) possiamo solo contare i casi di ciascuna categoria. Il numero di casi di una categoria si chiama frequenza assoluta di quella categoria.
- Se stiamo studiando una variabile quantitativa continua può essere conveniente raggruppare i valori della variabile in classi

DISTRIBUZIONE DI FREQUENZA

Regole generali per formare una distribuzione di frequenza nel caso di dati quantitativi:

a) Si ordinano i dati dal valore più piccolo al valore più grande (seriazione) assegnando a ciascun dato il suo rango.

b) Si trova il campo di variazione dell' intero insieme di osservazioni (*range*).

c) Si divide il campo di variazione in un numero conveniente di classi della stessa ampiezza (usualmente si prende un numero di classi compreso fra 5 e 20 e possibilmente un multiplo di 5). Le classi possono anche essere scelte in modo che i valori centrali coincidano con i dati realmente osservati.

d) Si conta il numero di osservazioni che cadono all'interno di ciascuna classe e si hanno, così, le frequenze assolute di ciascuna classe.

Nel caso di dati qualitativi, in generale, non si raggruppano i dati in classi.

Riassumendo:

- La distribuzione di frequenza si ottiene dividendo l'intervallo di variazione dei dati in una serie di classi dopo aver scelto il tipo di classe.

- Il tipo di classe maggiormente usato è il tipo chiuso-aperto. Sull'asse delle ascisse vanno riportate le classi mentre sull'asse delle ordinate vanno riportate le frequenze di ciascuna classe. Se si fanno classi chiuse-aperte, l'ultimo valore di ogni classe viene contato nella classe successiva (vedere esempio a pag. 81).

- La rappresentazione grafica di una distribuzione di frequenza si chiama "istogramma"; esso indica in che modo i valori di una variabile si sono distribuiti nel campo di variazione.

- Le frequenze cumulate si formano sommando alle frequenze di una certa classe tutte le frequenze delle classi precedenti.

- La curva cumulativa, disegnata dalla distribuzione di frequenze cumulate, ci dice visivamente la percentuale di campioni che hanno superato una certa classe.

Capitolo 6

INFERENZA STATISTICA

Mentre la statistica descrittiva si occupa di descrivere la massa dei dati sperimentali con pochi indici o grafici, la statistica inferenziale utilizza i dati, opportunamente sintetizzati dalla statistica descrittiva, per fare previsioni di tipo probabilistico su situazioni future o comunque incerte. Dall'esame, ad esempio, di un piccolo campione estratto da una grande popolazione mediante l'inferenza statistica si cerca di valutare per esempio la frazione della popolazione che possiede una certa caratteristica, un certo reddito o voterà per un certo candidato. Possiamo dire che l'inferenza statistica è un procedimento induttivo, che avvalendosi del calcolo delle probabilità, consente di estendere all'intera popolazione le informazioni fornite da un campione.
I due obiettivi dell'inferenza statistica sono: la Stima dei parametri ed il Test di ipotesi.

STIMA DI PARAMETRI

- Una popolazione è un insieme, finito o infinito, di individui di natura qualsiasi che noi siamo interessati a studiare.
- Un campione è un gruppo di individui estratto da una popolazione che noi usiamo per esaminare qualche problema riguardante la popolazione. Il campione deve essere rappresentativo della popolazione.

Se vogliamo valutare in una popolazione un parametro (per es. la media μ della pressione sanguigna), estraiamo un campione casuale da questa popolazione e stimiamo tale parametro.
Se estraiamo da questa popolazione 20 campioni tutti della stessa dimensione, i valori medi della pressione sanguigna nei vari campioni saranno differenti fra loro e saranno in generale anche differenti dal valore medio effettivo della pressione sanguigna della popolazione.

Le 20 medie campionarie formeranno una distribuzione.

La distribuzione di tutte le possibili medie campionarie è chiamata distribuzione campionaria delle medie.

Se i campioni hanno numerosità grande (n>30) la distribuzione campionaria delle medie tende ad assumere le caratteristiche di una distribuzione normale.

La distribuzione campionaria delle medie ha una propria media ed una deviazione standard.

La media della distribuzione campionaria delle medie approssima la media μ della popolazione: $\mu_{\bar{x}} \approx \mu$.

La deviazione standard della distribuzione delle medie campionarie è direttamente proporzionale alla deviazione standard della popolazione $\sigma_{\bar{x}} \approx \sigma$ ed inversamente proporzionale alla radice quadrata della numerosità dei campioni, ossia la deviazione standard della distribuzione delle medie campionarie è chiamato ERRORE STANDARD ($ES = \frac{\sigma}{\sqrt{n}}$ con n = dimensione del campione).

Se il campione è grande, la deviazione standard del campione s sarà una stima buona e corretta di σ.

L'errore standard è un indice della distanza del valore stima (media di un campione) dal valore vero (media della popolazione).

Data la distribuzione delle medie campionarie, il valore stima $\bar{x}$ cadrà:

- con il 68% di probabilità entro $\mu \pm 1$ E.S.
- con il 95% di probabilità entro $\mu \pm 2$ E.S.
- con il 99% di probabilità entro $\mu \pm 3$ E.S.

INTERVALLO DI CONFIDENZA PER UNA MEDIA

La media $\bar{x}$ di un campione estratto da una popolazione è chiamata stima puntiforme della media della popolazione. È difficile che accada che la stima puntiforme sia esattamente uguale alla media della popolazione, è invece verosimile che si avvicini di molto ad essa. Per misurare tale scostamento facciamo uso dell'errore standard.

Troviamo gli estremi di un intervallo (intervallo di stima) intorno alla stima puntiforme e diciamo con quale probabilità la media della popolazione cade entro questo intervallo.

Conoscendo il valore μ di una popolazione, sappiamo che circa il 95% delle medie campionarie cadono entro 2 errori standard dalla media della popolazione

Conoscendo la media $\bar{x}$ di un campione di dimensione n possiamo invertire il ragionamento ed asserire che l'intervallo di confidenza per μ (per es. al 95%) è:

$$\bar{x} - 2 \cdot ES < \mu < \bar{x} + 2 \cdot ES \quad \text{dove} \quad ES = \frac{\sigma}{\sqrt{n}}$$

La media del nostro campione più o meno 2 volte l'errore standard $(\bar{x} \pm 2 \cdot ES)$ costituiscono i limiti di confidenza al 95% per μ.

L' insieme dei valori tra $(\bar{x} - 2 \cdot ES)$ e $(\bar{x} + 2 \cdot ES)$ forma l'intervallo di confidenza al 95% intorno a μ.

INTERVALLO DI CONFIDENZA PER LA DIFFERENZA DI DUE MEDIE

Siano due popolazioni P_1 e P_2 distribuite normalmente con medie μ_1 e μ_2, e con varianze σ_1^2 e σ_2^2.

I campioni estratti dalle due popolazioni avranno ciascuno una propria distribuzione con una media ed una deviazione standard.

Se i campioni sono di numerosità grande, s di ogni campione è un buon stimatore di σ della popolazione da cui è stato estratto.

La distribuzione della differenza tra le medie campionarie avrà una media ed una deviazione standard.

Risulta:

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$$

Per popolazioni finite e di numerosità ν_1 e ν_2 si ha:

$$\sigma_{\bar{x}_1-\bar{x}_2} = ES_{\bar{x}_1-\bar{x}_2} = \sqrt{(ES_1)^2+(ES_2)^2}$$

Se $\bar{x}_1$ ed $\bar{x}_2$ le medie di due campioni di grandi dimensioni n_1 ed n_2 e varianze s_1^2 ed s_2^2 estratti dalle due popolazioni allora i limiti fiduciali con una affidabilità del 95% per la differenza delle medie di due popolazioni saranno:

$$(\bar{x}_1-\bar{x}_2)-2\cdot ES_{\bar{x}_1-\bar{x}_2} \prec \mu_1-\mu_2 \prec (\bar{x}_1-\bar{x}_2)+2\cdot ES_{\bar{x}_1-\bar{x}_2}$$

dove $ES_{\bar{x}_1-\bar{x}_2} = \sqrt{s_1^2/n_1+s_2^2/n_2}$

In generale:

$$(\bar{x}_1-\bar{x}_2)-z_c\cdot ES_{\bar{x}_1-\bar{x}_2} \prec \mu_1-\mu_2 \prec (\bar{x}_1-\bar{x}_2)+z_c\cdot ES_{\bar{x}_1-\bar{x}_2}$$

Dove z varia a seconda del grado di affidabilità che vogliamo avere:

quindi i limiti fiduciali per la differenza delle medie di due popolazioni sono:

$$\left(\overline{x}_1 - \overline{x}_2\right) \pm z_c \cdot ES_{\overline{x}_1 - \overline{x}_2}$$

TEST DI IPOTESI

Il test di ipotesi si utilizza per verificare la bontà di un'ipotesi.

Per ipotesi si intende un'affermazione su fenomeni reali che vogliamo confermare o smentire in seguito ad osservazione.

Un test statistico può essere:

- parametrico *(Si intendono "Parametrici" quei Test di significatività che si basano sui Parametri fondamentali della statistica (Media e Deviazione Standard), quindi applicabili a variabili quantitative con Curva di distribuzione approssimabile alla distribuzione Normale.)*
- non parametrico

Un'ipotesi statistica è un'affermazione sulla distribuzione di probabilità di una variabile casuale. Nel test statistico viene verificata in termini probabilistici la validità di un'ipotesi statistica, detta ipotesi nulla, indicata di solito con H_0.

Un test di ipotesi conduce ad una decisione statistica : la conclusione potrà essere di rifiutare l'ipotesi nulla in favore di quella alternativa, o di non poter rifiutare l'ipotesi nulla. Nel caso l'ipotesi nulla venga rifiutata si accetterà l'ipotesi alternativa, indicata con H_1.

LIVELLO DI SIGNIFICATIVITÀ

Prima di decidere quali sono i valori che cadono nella regione di rifiuto e quelli che cadono nella regione di non rifiuto bisogna fissare il livello di significatività.
Il livello di significatività α è la probabilità di rifiutare un'ipotesi nulla vera. Poiché rifiutare un'ipotesi nulla vera potrebbe rappresentare un errore, è buona cosa prendere α piccolo. I valori più frequentemente usati per α sono 0.05 e 0.01.

ERRORE

La decisione che prendiamo può essere corretta o errata. Esistono due tipi di errore, a seconda di quale delle due ipotesi è vera:

1. Errore di prima specie consiste nel rifiutare l'ipotesi nulla quando è vera
2. Errore di seconda specie consiste nel non rifiutare l'ipotesi nulla quando è

Di solito esiste un compromesso tra le probabilità di errori di prima e seconda specie. Se riduciamo la probabilità di un errore di prima specie, incrementiamo necessariamente la probabilità di errore di seconda specie.

POTENZA DI UN TEST

Se H_1 è vera, la probabilità di rifiutare H_0 (e prendere quindi una decisione corretta), è detta potenza del test.

TEST DEL CHI-QUADRATO

Il test del Chi-quadrato di Pearson è un test non parametrico applicato a grandi campioni quando si è in presenza di variabili nominali e si vuole verificare se il campione è stato estratto da una popolazione con una predeterminata distribuzione o che due o più campioni derivino dalla stessa popolazione.

La statistica test del Chi-quadrato è

$$\chi^2 = \sum_{i=1}^{n} (o_i - a_i)^2 / a_i$$

Dove o sono i casi osservati ed a sono i casi attesi

- Se vengono utilizzati i dati di un solo campione e si vuole testare l'ipotesi nulla che il campione è stato estratto da una popolazione di cui è nota la distribuzione, il test del Chi-quadrato si chiama "test della bontà dell'adattamento".

- Se si vuole testare l'ipotesi che due campioni sono indipendenti e derivano dalla stessa popolazione, di cui non è richiesto conoscere la distribuzione il test del Chi-quadrato si chiama "test di indipendenza di due campioni". I dati vengono organizzati in una tabella.

Il test del Chi quadrato si basa sulla distribuzione del Chi-quadrato. Essa è una distribuzione statistica determinata completamente dai gradi di libertà. È asimmetrica e definita sull'asse positivo.

Per 1 grado di libertà

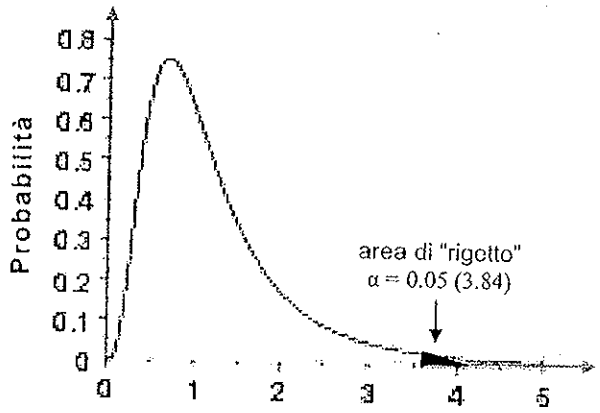


TABELLA DI CONTINGENZA

È rappresentata da una tabella m x n dove m è il numero delle righe ed n il numero delle colonne. Considerando le colonne come una classe di eventi (A) e le righe come un'altra classe di eventi (B), si pone il problema di determinare se gli eventi di A sono indipendenti dagli eventi di B.

L'applicazione in ambito medico di una tabella di contingenza ci viene offerto da un esperimento di questo tipo: immaginiamo di trattare con un farmaco un gruppo di individui malati di una certa malattia (gruppo A) e di trattare con placebo un gruppo paragonabile di soggetti malati della stessa malattia (gruppo B). Dopo un congruo periodo di trattamento andiamo a verificare la situazione clinica nei due

gruppi definendo con criteri opportuni l'evento miglioramento e l'evento stazionarietà.

Dal punto di vista clinico c'è interesse a vedere se il farmaco ha efficacia, ossia a verificare se la proporzione dei soggetti migliorati è maggiore nel gruppo A che nel gruppo B.

Dal punto di vista probabilistico c'è interesse a vedere se l'evento miglioramento è indipendente dall'evento trattamento ossia se il farmaco ha lo stesso effetto del placebo. Se l'analisi statistica ci porta al rigetto di questa ipotesi, concludiamo che il farmaco ha un effetto diverso dal placebo.

PROCEDIMENTO PER IL CALCOLO DEL CHI QUADRATO

- Si costruiscono le due tabelle delle frequenze osservate e delle frequenze attese. (S = situazione stazionaria, M = migliorati, A = individui trattati con il farmaco A, B = individui trattati con placebo, r = marginale di riga, c = marginale di colonna, t = totale generale).

La tabella delle frequenze attese viene costruita con la seguente regola:

Il valore atteso, per ciascun evento congiunto, sotto l'ipotesi di indipendenza, si ottiene facendo il prodotto dei marginale di riga (r) per il marginale di colonna (c) e dividendo per il totale (t).

Esempio 1:

OSSERVATI

	S	M	Totale
A	50	50	100 (r1)
B	70	30	100 (r2)
Totale	120 (c1)	80 (c2)	200 (t)

ATTESI

	S	M	Totale
A	60	40	100 (r1)
B	60	40	100 (r2)
Totale	120 (c1)	80 (c2)	200 (t)

Una volta che per ogni coppia (AS,AM,BS,BM) sono state calcolate le frequenze attese, ad esse verrà applicata l'uguaglianza:

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - a_i)^2}{a_i}$$

che si legge "sommatoria rispetto ad i della differenza valore osservato meno valore atteso, elevata al quadrato e divisa per il valore atteso". I gradi di libertà sono uguali a : (n° righe -1)(n° colonne -1).

Nel nostro esempio si ha:

$$\chi^2 = \frac{(50-60)^2}{60} + \frac{(50-40)^2}{40} + \frac{(70-60)^2}{60} + \frac{(30-40)^2}{40} = 8.33$$

È compito della statistica valutare se eventuali differenze tra valori osservati e valori attesi siano da imputarsi a fluttuazioni campionarie oppure se tali differenze debbano essere considerate eccezionali sotto l'ipotesi di indipendenza. In quest'ultimo caso la statistica ci consiglia di rigettare l'ipotesi di partenza.

Il modello logico che si segue in questo esempio è il seguente: si assume che i due eventi farmaco e guarigione siano indipendenti tra di loro, ossia che il farmaco abbia lo stesso effetto del placebo (nessun effetto o ipotesi zero).

In una tabella 2 x 2 e quindi con 1 grado di libertà,tenendo presente che un livello di significatività α=0.05 corrisponde ad un valore di χ^2=3.84, se il Chi-quadrato ha un valore, per esempio, pari a 4, possiamo dedurre che se la nostra

H_0 è vera ci troviamo di fronte ad un evento raro (meno di 5/100) e quindi il buon senso ci consiglia di rigettare Ho. Infatti, se assumiamo tale comportamento ci sbaglieremo solo 5 volte su 100.

Se il valore del Chi-quadrato è maggiore o uguale a 6.6, poiché un valore così fatto si osserva solo una volta su 100, a maggior ragione rigettiamo Ho perchè ci troviamo di fronte ad un evento ancora più raro. Se assumiamo un tale comportamento rigettando Ho tutte le volte che ci troviamo di fronte a differenze che danno un valore di chi-quadrato maggiore o uguale a 6.6, ci sbaglieremo solo una volta su 100.

Nel nostro esempio, essendo χ^2= 8.33,rigettiamo l'ipotesi che il miglioramento è indipendente dal trattamento in quanto un valore simile di χ^2, a parità di condizioni,si ottiene meno di una volta su 100.

Poiché il Chi-quadrato è una funzione che viene usata correntemente per variabili nominali, l'approssimazione è ottima per campioni molto grandi ma meno buona per campioni piccoli.

Se il sistema ha 1 grado di libertà, il valore minimo per il valore atteso in ogni casella, non deve essere inferiore a 5. Nei sistemi con più di 1 grado di libertà, è tollerabile che qualche casella abbia un atteso inferiore a 5 ma non inferiore ad 1.

Esempio 2:

Consideriamo un'indagine epidemiologica nella quale si esaminano 90 individui, classificandoli E = esposto ad un certo fattore di rischio e M = malato di un certa malattia. Dall'indagine si ricavano i valori osservati con i quali si costruisce la seguente tabella di contingenza :

	M	**non M**	**totale**
E	30	32	62
non E	6	22	28
totale	36	54	90

Siamo interessati a vedere se la malattia dipende dall'esposizione al fattore di rischio con un livello di fiducia del 95%.

Si costruisce la tabella delle frequenze attese

	M	non M	totale
E	24.8	37.2	62
non E	11.2	16.8	28
totale	36	54	90

Si trova un χ^2=5.841. Il risultato indica una dipendenza della malattia dalla esposizione al fattore di rischio.

VERIFICA DI IPOTESI SULLE MEDIE

VERIFICA DELL' IPOTESI SULLA DIFFERENZA DI DUE MEDIE

-Siano dati due campioni con medie $\bar{x}_1$ ed $\bar{x}_2$, varianze s_1^2 ed s_2^2 e dimensioni n_1 ed n_2 ,ambedue ≥30, estratti indipendentemente da due popolazioni P_1 e P_2 con medie μ_1 e μ_2 e distribuite normalmente.

L'ipotesi zero sarà:

$H_0: \ \mu_1 - \mu_2 = 0$ con μ_1 e μ_2 medie di P_1 e P_2

Campionamento da popolazioni con varianze non note:

Se si suppone che le varianze delle popolazioni siano diverse:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Se i campioni hanno numerosità n_1 ed n_2 ambedue <30, e le varianze delle popolazioni non sono note:

Se si suppone che le varianze siano diverse :

$$t = \frac{\left[\left(\bar{x}_1 - \bar{x}_2\right) - 0\right]}{ES_{\bar{x}_1 - \bar{x}_2}} \qquad \text{dove} \quad ES_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

TEST DEL t di STUDENT

Il t di Student è un test parametrico che permette di testare l'ipotesi nulla:

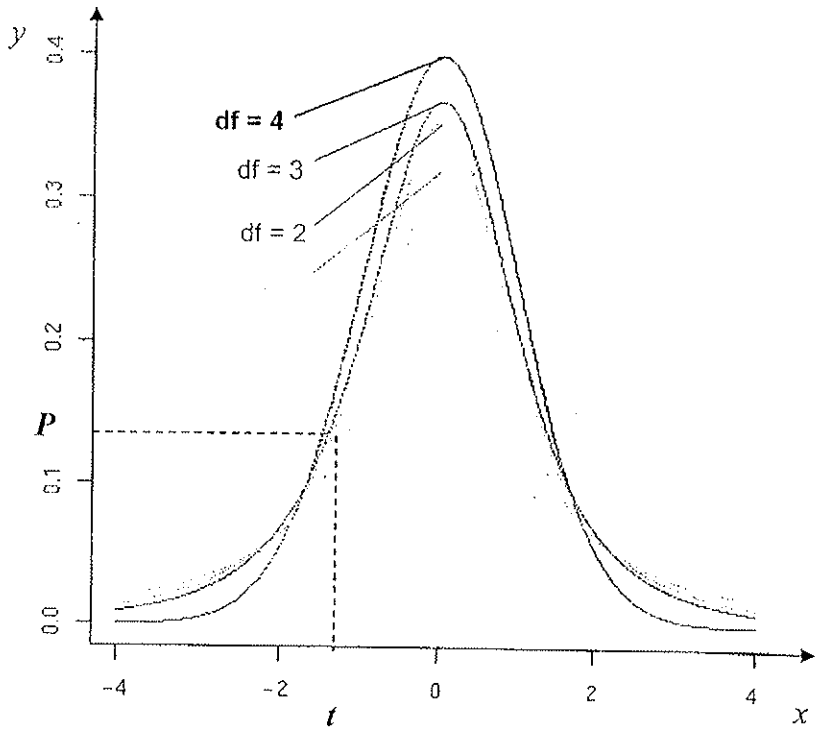
- che la media di una certa popolazione sia uguale ad un certo valore,
- che le medie di due gruppi diversi siano uguali.

Viene usato in statistica con campioni di piccole dimensioni (minore o uguale a 30 elementi). Se il campione è più numeroso la distribuzione normale e quella di Student differiscono di poco, pertanto è indifferente usare una o l'altra.

DISTRIBUZIONE t DI STUDENT

Questa è una distribuzione di probabilità teorica con andamento simile alla distribuzione normale, rispetto alla quale presenta delle code più alte. E' determinata da un solo parametro detto gradi di libertà e tende alla curva normale aumentando i gradi di libertà.

Distribuzioni t



TEST PER DUE CAMPIONI INDIPENDENTI

Si usa per confrontare due gruppi diversi rispetto ad una singola variabile. Dati due campioni di numerosità n_1 ed n_2 con medie $\bar{x}_1$ ed $\bar{x}_2$ e con deviazioni standard s_1 ed s_2 (buoni stimatori delle deviazioni standard delle popolazioni) estratti da due popolazioni distribuite normalmente con medie μ_1 e μ_2 e con varianze non note la statistica test si basa sulla differenza delle medie campionarie $(\bar{x}_1 - \bar{x}_2)$. Se le varianze delle popolazioni non sono note e si suppone che non siano uguali, esse devono essere determinate utilizzando le varianze campionarie:

$H_0: \ \mu_1 = \mu_2$ e la statistica test è:

$$t = \frac{\left[\left(\bar{x}_1 - \bar{x}_2\right) - 0\right]}{ES_{\bar{x}_1 - \bar{x}_2}}$$

dove $ES_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_1^2/n_1 + s_2^2/n_2}$ con $n_1 + n_2 - 2$ = gradi di libertà

TEST PER DATI APPAIATI

Si usa per confrontare tra loro due rilevazioni diverse fatte sullo stesso soggetto (per es. valutare se la pressione diastolica varia tra mattino e sera).
Dato un campione di n soggetti consideriamo n_1 valori di una certa variabile in una condizione ed n_2 valori della stessa variabile in un'altra condizione, ne facciamo le differenze tra questi valori e ne calcoliamo il valor medio $\bar{x}_d$ e l'errore standard: ES_d

$H_0: \ \bar{x}_d = 0$ e la statistica test è:

$t = \bar{x}_d / ES_d$ con n-1 gradi di libertà

ANALISI DELLA VARIANZA- Test F

L'analisi della varianza permette di analizzare due o più gruppi contemporaneamente, evidenziando eventuali differenze nella globalità di essi. Il Test F si basa sul rapporto tra due modi differenti di stimare la varianza: confronta la variabilità *interna* a questi gruppi con la variabilità *tra* i gruppi.

L'ipotesi nulla H_0 solitamente prevede che i dati di tutti i gruppi abbiano la stessa distribuzione e che le differenze osservate tra i gruppi siano dovuti solo al caso.

La statistica test è: $F=s^2$ tra/s^2 entro;

- gradi di libertà del numeratore=n° gruppi -1;

- gradi di libertà del denominatore = n° gruppi x (numerosità dei gruppi -1)

se non ci fosse differenza tra i gruppi questo rapporto dovrebbe essere approssimativamente 1.

L'analisi della varianza viene applicata a variabili di tipo nominale. Ne caso di variabili di tipo ordinale o continuo è meglio usare tecniche alternative (per es. regressione lineare).

Bisogna distinguere se la variabilità è dovuta ad:

- una sola causa: (per es. il gradimento di un cibo dipende dal colore del medesimo) – Analisi della varianza (ANOVA) ad una via-

- più di una causa (per es. il successo scolastico dipende sia dal genere (maschi, femmine) che dallo sport praticato (calcio, tennis, box,...) o l'interazione tra più cause: per es. la velocità di guarigione dipende da due farmaci, i quali si annullano (o rinforzano) a vicenda -Analisi della varianza (ANOVA)a più vie-

L'Analisi della Varianza fatta su 2 gruppi è analoga al Test t-Student per campioni Indipendenti.

CORRELAZIONE

Per correlazione si intende una relazione tra due variabili tale che a ciascun valore della prima variabile corrisponda con una certa regolarità un valore della seconda. Non si tratta necessariamente di un rapporto di causa ed effetto ma semplicemente della tendenza di una variabile a variare in funzione di un'altra.

La correlazione si dice:

- diretta o positiva quando variando una variabile in un senso anche l'altra varia nello stesso senso (alle stature alte dei padri corrispondono stature alte dei figli);
- indiretta o inversa quando variando una variabile in un senso l'altra varia in senso inverso (a una maggiore produzione di grano corrisponde un prezzo minore).
- semplice quando i fenomeni posti in relazione sono due (per esempio, numero dei matrimoni e il numero delle nascite);
- doppia quando i fenomeni sono tre (per esempio, circolazione monetaria, prezzi e risparmio); tripla, quadrupla, etc.

Il grado di correlazione fra due variabili viene espresso mediante il coefficiente di correlazione.

Il coefficiente di correlazione tra 2 variabili statistiche x e y indica quanto le due variabili sono collegate tra di loro. Un valore di 0 indica che non c'é nessun collegamento, +1 indica che i punti (x,y) sono disposti su una retta con valori alti di x corrispondenti a valori alti di y. Invece -1 corrisponde a una retta con valori alti di x corrispondenti a valori bassi di y.

Capitolo 7

FASI DI UNA RICERCA

Le fasi di una ricerca sono:

- raccolta delle informazioni
- analisi delle informazioni
- interpretazione dei risultati

L'analisi delle informazioni, di natura numerica, opportunamente selezionate, viene fatta utilizzando il metodo statistico.

La Statistica è una scienza che si occupa dello studio di fenomeni per lo più complessi e di qualsiasi natura, situati in un universo variabile. Infatti per lo studio di fenomeni complessi e casuali bisogna determinare delle leggi statistiche che permettono di spiegare tali fenomeni.

Le informazioni devono essere valide, senza errori, e rappresentative del fenomeno che si vuole studiare. L'analisi della validità porta automaticamente allo studio di tutti i possibili errori che possono apparire in qualunque fase del processo statistico, dalla raccolta dei dati fino all'interpretazione dei risultati.

TIPI DI VARIABILE

Variabile qualitativa (nominale)

Una variabile si dice qualitativa quando una caratteristica non viene misurata per mezzo di numeri ma per mezzo di categorie. Prendiamo, per esempio, il sesso. Il sesso viene classificato in due modalità: maschile e femminile. Inoltre le due modalità sono disgiunte, ossia tra una modalità e l'altra non ci sono gradazioni e

l'unica operazione matematica possibile è "uguale" o "diverso". Prendiamo ora il colore degli occhi. Anche il colore degli occhi viene classificato in modalità, ma il colore varia insensibilmente da un colore chiaro a un colore scuro.
Il sesso è una variabile qualitativa dicotomica e può essere espressa solo con attributi non ordinabili. Il colore degli occhi è una variabile qualitativa ordinabile. Fra una modalità e l'altra siamo sempre in grado di individuarne una terza, intermedia. Una variabile qualitativa può essere quindi classificata come:
- *dicotomica* (o binaria), se assume solo due stati (sì/no, presenza/assenza, vivo/morto, etc.)
- *ordinale*, se assume più di due stati che vengono ordinati secondo una certa gradualità (fumo di sigarette: non fumatore ,piccolo fumatore, medio fumatore, grande fumatore, etc.).
- *politomica* (o nominale), se assume più di due stati non ordinabili e mutuamente esclusivi (gruppo sanguigno ABO: A, B, AB, O).

Notiamo che, per essere utile, questa classificazione non dovrà essere né troppo grossolana, né troppo fine: se intendiamo studiare l'effetto dell'obesità sulla salute, dividere la popolazione soltanto in grassi e magri potrebbe non far notare effetti più sottili, d'altra parte, se la classificassimo in modo così fine che ogni individuo vada a finire in una classe diversa, la classificazione sarebbe priva di ogni interesse. Come trovare il "giusto mezzo" dipenderà da caso a caso.

Variabile quantitativa (numerica)

Una variabile si dice quantitativa quando è misurata tramite un valore numerico. Sono variabili quantitative: l'età, il peso, l'altezza, il numero dei componenti della famiglia.
Si distingue in:
- *continua* quando può assumere tutti i valori di un intervallo
- *discreta* quando non può assumere tutti i valori di un intervallo

La distinzione tra variabili continue e discrete è molto più complessa di quanto non sembri a prima vista. Vi sono variabili discrete che possono assumere un infinito numero di valori (almeno teoricamente) e pertanto vengono trattate comunemente come delle variabili continue Oppure vi sono variabili continue che vengono trattate come variabili discrete. Prendiamo, per esempio, la statura.

Essa è continua ma può essere trattata come una variabile discreta. Dipende dal mezzo di misura.

TIPI DI CLASSE

Se dobbiamo effettuare uno studio statistico su una variabile quantitativa continua é utile raggruppare i valori della variabile in classi.
In generale una classe può essere:
- Chiusa a sinistra e chiusa a destra (classe chiusa-chiusa): [ ], se comprende gli estremi della classe.
- Chiusa a sinistra e aperta a destra (classe chiusa-aperta): [ ), se non comprende l'estremo destro della classe.
- Aperta a sinistra e chiusa a destra (classe aperta-chiusa): ( ], se non comprende l'estremo sinistro della classe.
- Aperta a sinistra e aperta a destra (classe aperta-aperta): ( ), se non comprende gli estremi della classe.

LIMITI DI UNA CLASSE

I valori numerici che delimitano una classe sono detti limiti della classe. Il valore più piccolo è detto "limite inferiore della classe" ed il valore più grande è detto "limite superiore della classe".
La differenza tra il limite superiore ed il limite inferiore di una classe si chiama "ampiezza della classe".
Può accadere che nella formazione delle classi il limite inferiore della prima classe ed il limite superiore dell'ultima classe non siano valori realmente osservati; essi vengono ugualmente tabulati per avere classi di ampiezza tutte uguali.

VALORE CENTRALE DI UNA CLASSE

Il valore centrale di una classe o rappresentante della classe è ottenuto sommando i limiti inferiore e superiore di una classe e dividendo il risultato per 2.

SCALE DI MISURA

Quando si raccolgono i dati, il processo di misurazione è costituito dall'assegnare un valore al fenomeno osservato. Ci è molto familiare associare una misura a caratteristiche di tipo quantitativo (altezza,peso, etc.) ma ci è più difficile associarla a caratteristiche di tipo qualitativo (intelligenza, salute, colore degli occhi, etc.). Pertanto, il processo dì misurazione non è una semplice attribuzione numerica ma è una procedura di classificazione che permette di attribuire un oggetto ad una classe mediante un insieme di regole. Tutte le variabili sono, così, raccolte in livelli o scale di misura. S.S. Stevens (1946) ha identificato quattro livelli di misura: nominale, ordinale, ad intervalli e di rapporto.

1) Livello di misura nominale. è il livello di misurazione più basso. In questo livello non si fa nessuna assunzione sui valori che vengono assegnati ad una variabile ma le osservazioni vengono solo classificate in categorie tra loro escludentesi .In questo livello si possono fare solo confronti di tipo "uguale" o "diverso", "si" o "no".

2) Livello di misura ordinale si ha quando è possibile ordinare le categorie di una variabile seguendo un certo criterio. Alle categorie di una variabile non si applica solo la relazione di uguale/diverso ma anche la relazione di maggiore/minore.

3) Livello di misura ad intervalli si ha quando non solo è possibile ordinare le classi di misura di una variabile ma è anche possibile misurare la distanza fra due classi secondo una unità di misura. Tale distanza deve essere sempre costante. La scala intervallare misura i fenomeni per i quali lo zero non corrisponde all'assenza del carattere ma è fissato arbitrariamente; questo avviene, ad esempio, per la temperatura. In questo caso specifico lo zero corrisponde alla

temperatura alla quale l'acqua gela e il rapporto fra due misurazioni non avrebbe senso mentre si può apprezzare la loro differenza.

4) Livello di misura di rapporto è il più alto livello di misurazione. Tale livello di misura è determinato non solo dall'uguaglianza delle distanze ma anche dall'uguaglianza dei rapporti tra due classi. Tale livello ha la fondamentale proprietà che lo zero è una misura. Tale scala ci consente di apprezzare il rapporto esistente fra due misurazioni (potremo dire, ad esempio, che una famiglia ha il doppio dei componenti di un'altra o che una persona è alta una volta e mezza un'altra).
Le quattro scale di misura considerate esprimono una gerarchia che condiziona le possibilità nell'elaborazione statistica. Esse risultano via via più ampie passando dalla scala nominale a quella di rapporto

TIPI DI STUDIO

A) **Studio Sperimentale**: Sperimentazione clinica, Sperimentazione sul campo (per es. laboratori, studenti etc.), Sperimentazione su comunità (per es. popolazione).

Sperimentazione clinica controllata

La sperimentazione clinica controllata (*randomized controlled trial*, RCT) è uno studio sperimentale che permette di valutare l'efficacia di uno specifico trattamento in una determinata popolazione ((con il termine trattamento si intendono convenzionalmente non solo le terapie, ma tutti gli interventi (diagnostici, di *screening*, di educazione sanitaria) o anche l'assenza di intervento)).
Viene effettuata su soggetti affetti da una determinata malattia.
Lo studio è "*sperimentale*" (*trial*) perchè le modalità di assegnazione dei soggetti alla popolazione da studiare vengono stabilite dallo sperimentatore. Una volta reclutata la popolazione, sulla base di tutte le variabili considerate dal ricercatore per il loro significato prognostico (natura e gravità della malattia, età, parità etc.),

si verifica l'effetto di un trattamento (ad esempio, la somministrazione di un farmaco) confrontandolo con l'effetto di un altro diverso trattamento (ad esempio, un altro farmaco, nessun farmaco o un placebo).

È "*controllato*" (*controlled*) perchè i soggetti coinvolti nello studio vengono suddivisi in due gruppi: un gruppo che riceve il trattamento, ed un gruppo che ne riceve uno diverso o nessun trattamento.

È "*randomizzato*" (*randomized*) perchè l'assegnazione del trattamento ai soggetti deve avvenire con un metodo casuale (random) (la randomizzazione aumenta la probabilità che altre variabili, non considerate nel disegno dello studio, si distribuiscano in maniera uniforme nel gruppo sperimentale e in quello di controllo. In questo modo, le differenze eventualmente osservate tra i due gruppi possono essere attribuite al trattamento).

È *prospettico* perchè la sperimentazione viene condotta parallelamente nei due gruppi e i risultati ottenuti vengono analizzati alla fine dello studio.

Tipi di randomizzazione

- *Assegnazione alternata*: si alterna farmaco-placebo
- *Randomizzazione semplice*: uso della tavola dei numeri casuali
- *Randomizzazione vincolata* o per *cluster*: tutti i gruppi sono costituiti dallo stesso numero di soggetti.

Quando possibile, né lo sperimentatore né i soggetti coinvolti sono a conoscenza del trattamento assegnato (cioè entrambi sono *in cieco*, da cui il termine "doppio cieco") per ridurre la probabilità che ne vengano influenzati (i pazienti potrebbero comportarsi in maniera diversa a seconda del gruppo al quale appartengono e gli operatori sanitari potrebbero valutare diversamente le loro condizioni).

N.B.: È importante che l'analisi dei dati venga effettuata su tutti i soggetti inizialmente reclutati e che nessuno sia escluso dallo studio. È infatti possibile che alcuni pazienti ammessi allo studio e assegnati ad uno dei trattamenti manifestino sintomi o condizioni tali da ritenere necessari la sospensione o il cambio del trattamento (aggravamento della malattia, intollerabilità o tossicità del farmaco etc.). Anche le informazioni riguardanti chi non ha seguito il protocollo, o chi si è ritirato dallo studio, devono essere comprese nell'analisi finale dei dati.

B) **Studio Pianificato**: Studio caso-controllo, Studio per coorte, Studio trasversale.

- Lo studio "*caso-controllo*" è uno studio retrospettivo orientato alla malattia. È inteso a determinare la frequenza con cui un sospetto fattore di rischio per una certa malattia compare in una popolazione suddivisibile in due gruppi a seconda della presenza (caso) o meno (controllo) della malattia.

- Lo studio "*per coorte*" è uno studio prospettico orientato all'esposizione ed uno studio longitudinale. È inteso a determinare la frequenza con cui una determinata malattia si presenterà in una popolazione dopo un certo tempo T dall'inizio dell'indagine. La popolazione viene suddivisa in genere in due gruppi a seconda della presenza (persone esposte a rischio e sane) o meno (persone non esposte a rischio e sane) del fattore eziologico (fattore di rischio).

L'obiettivo è quindi principalmente rivolto a misurare l'evoluzione nel tempo delle caratteristiche di interesse (fattore di rischio) mediante l'espediente di ricontattare le unità (persone esposte e non esposte) per analizzarne i cambiamenti.

- Lo studio "trasversale" è uno studio di prevalenza che stima le caratteristiche di una popolazione in un particolare momento o periodo di tempo. Le unità statistiche vengono formate raccogliendo informazioni di interesse riferite a quel particolare momento o periodo di tempo. Ha essenzialmente lo scopo di descrivere la popolazione e può essere eseguito preliminarmente ad uno studio per coorte.

CAMPIONAMENTO

Il campione si usa nelle indagini pianificate, soprattutto trasversali quando si vuole conoscere uno o più parametri di una popolazione, finita, ma troppo numerosa, ed il costo di studio è troppo oneroso sia per risorse umane che finanziarie.

I tipi di campionamento più noti sono:

- casuale semplice, se tutti gli elementi della popolazione hanno la stessa probabilità di entrare a far parte del campione
- randomizzato, se per estrarre il campione si utilizza la tavola dei numeri casuali
- stratificato, se il campionamento casuale si ottiene entro uno schema di stratificazione
- a grappolo, se la scelta casuale, randomizzata o sistematica viene effettuata dopo che la popolazione viene suddivisa in un elevato numero di gruppi (grappoli o *cluster*) composti da elementi il più eterogenei possibile. Si procede poi all'estrazione casuale di un certo numero di grappoli, le cui unità vengono tutte incluse nel campione.

ESERCIZI

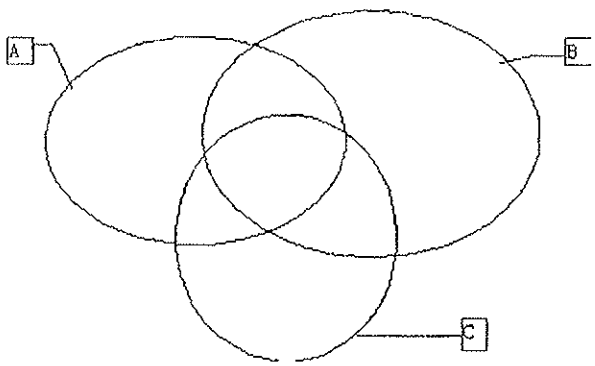
1. INSIEMISTICA

Appartenenza e non appartenenza

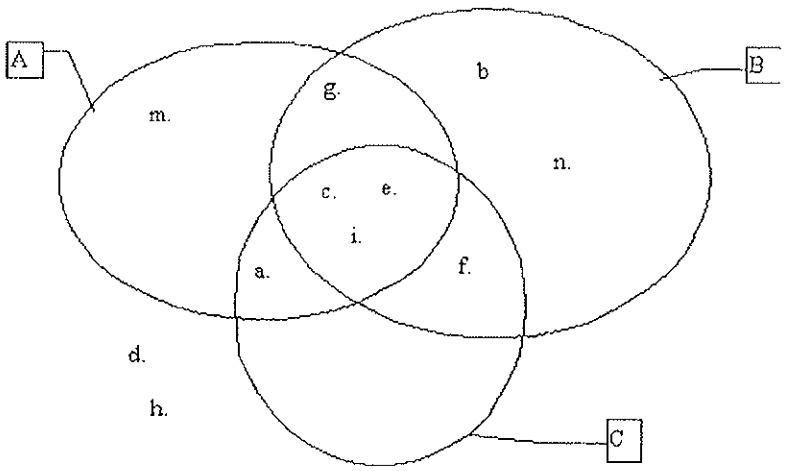
Dati i seguenti elementi:

$a \in A$	$d \notin A$	$g \in A$	$m \in A$
$a \in C$	$d \notin B$	$g \in B$	$m \notin B$
$a \notin B$	$d \notin C$	$g \notin C$	$m \notin C$
$b \in B$	$e \in B$	$h \notin A$	$n \in B$
$b \notin A$	$e \in C$	$h \notin B$	$n \notin A$
$b \notin C$	$e \in A$	$h \notin C$	$n \notin C$
$c \in B$	$f \in B$	$i \in A$	
$c \in A$	$f \in C$	$i \in B$	
$c \in C$	$f \notin A$	$i \in C$	

mettili al loro posto all'interno del diagramma di Eulero-Venn.



Soluzione:



Unione e intersezione

Dati i seguenti insiemi:

A={1,2,3,4}; B={ 3,4,5,6}; C={ 3,4,7,8}; D={7,8,9}

Rappresentare le seguenti situazioni:

$A \cap B$ $A \cap C$ $B \cap C$ $B \cap D$ $A \cap B \cap C$
$A \cup B$ $A \cup C$ $B \cup D$

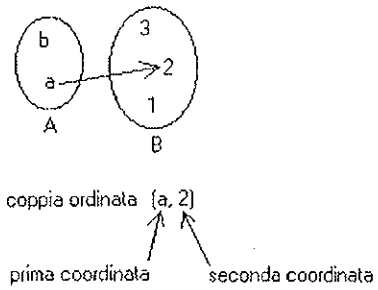
Soluzione:

$A \cap B=(3,4)$ $A \cap C=(3,4)$ $B \cap C=(3,4)$ $B \cap D=(\varnothing)$

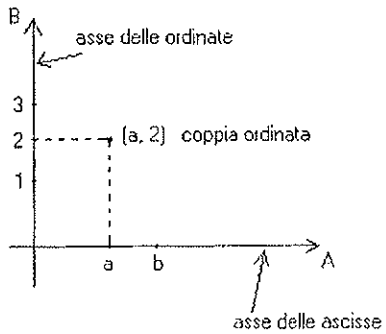
$A \cup B=(1,2,3,4,5,6)$ $A \cup C=(1,2,3,4,7,8)$ $B \cup C=(3,4,5.6,7,8)$ $A \cap B \cap C=(3,4)$

COPPIA ORDINATA

Consideriamo l'insieme A = {a,b} e l'insieme B = {1,2,3}. Se prendiamo un elemento di A, per esempio, "a", ed un elemento di B, per esempio "2", possiamo costruire la coppia ordinata (a,2) dove è essenziale l'ordine con cui si scelgono gli elementi dai due insiemi. Il primo elemento della coppia ordinata, quello scritto a sinistra, si chiama prima coordinata mentre il secondo, quello scritto a destra, si chiama seconda coordinata.



Nei diagrammi di Venn una coppia ordinata viene rappresentata da una freccia che parte dalla prima coordinata della coppia ordinata e punta alla seconda coordinata della medesima. Vi è un altro modo molto proficuo di rappresentare le coppie ordinate utilizzando gli assi cartesiani.



Sugli assi cartesiani una coppia ordinata viene rappresentata con un punto come illustrato in figura. Utilizzando gli assi cartesiani occorre sottolineare che l'insieme da cui si prendono le prime coordinate va posto sull'asse delle ascisse (l'asse orizzontale) mentre l'altro insieme, da cui si prendono le seconde coordinate, va posto sull'asse delle ordinate (l'asse verticale).

PRODOTTO CARTESIANO

L'insieme di tutte le coppie ordinate che si possono formare prendendo le prime coordinate dall'insieme A e le seconde coordinate dall'insieme B si chiama prodotto cartesiano di A per B e si indica con: A x B Il prodotto cartesiano A x B è definito allora da

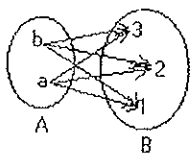
$$A \times B = \{(x, y); x \in A, y \in B\}$$

che si legge "il prodotto cartesiano dell'insieme A per l'insieme B è l'insieme di tutte le coppie ordinate che si ottengono prendendo la prima coordinata in A e la seconda coordinata in B".

Considerando gli insiemi A e B, sopra definiti, si ha allora :

A x B = {(a,1), (a,2), (a,3), (b,1), (b,2), (b,3)}.

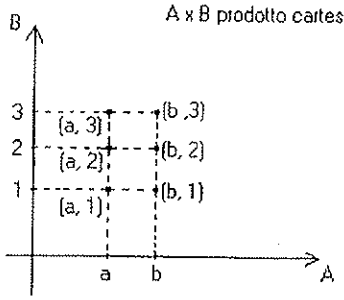
Graficamente, usando i diagrammi di Venn, il prodotto cartesiano A x B sarà visualizzato come:



A x B prodotto cartesiano di A per B

ovvero prendendo tutte le possibili frecce dagli elementi di A agli elementi di B .Usando invece gli assi cartesiani si ottiene il seguente grafico:

A x B prodotto cartesiano di A per B



dove si vede bene che le coppie ordinate del prodotto cartesiano sono indicate da tutti i possibili punti che si possono ottenere considerando gli elementi dei due insiemi.

2. CALCOLO COMBINATORIO

DISPOSIZIONI

1. Calcolare le disposizioni di 5 elementi a 3 a 3

$D_{5,3}$=5x4x3=(5x4x3x2x1)/(2 x1)=60

2. Un urna contiene 7 palline differenti, trovare il numero possibile di teme ordinate che si possono ottenere senza reimbussolamento

$D_{7,3}$=7x6x5=(7x6x5x4x3x2 x1)/(4x3x2x1)=

3. In quanti modi si possono prelevare da 8 libri gruppi ordinati di 5 libri?

$D_{8,5}$=8x7x6x5x4= (8x6x5x4x3x2x1)13x2x1

4. Quanti numeri di 3 cifre distinte possono essere formati a partire dai numeri 1,2,3,4,5,6? (ogni numero va preso una sola volta).

$D_{6,3}$= 6 x 5x4 = (6x 5x4x3x2 x 1)1(3 x2 x 1) = 120

5. Quanti numeri di 3 cifre distinte possono essere formati a partire dai numeri 0,1,2,3,4,5? (ogni numero va preso una sola volta).

La prima cifra può essere scelta fra 5 cifre (tutte meno lo 0);

la seconda cifra fra 5 cifre (tutte tranne quella già scelta)

la terza cifra tra 4 cifre (tutte tranne le due già estratte) 5 x 5 x 4=100

6. In quanti modi possono essere scelti tra 9 pazienti 4 pazienti da sistemare in modo ordinato in una stanza d'ospedale?

$D_{9,4}$=9x8x7x6=

PERMUTAZIONI

1. Calcolare le disposizioni di 3 oggetti a 3 a 3

$D_{3,3} = P_{3,3} = 3! = 3 \times 2 \times 1 = 6$

2. In quanti si possono sistemati 4 libri in uno scaffale?

$P_{4,4} = 4! = 4 \times 3 \times 2 \times 1 = 24$

3. Quanti anagrammi si possono formare con la parola ROMA?

$P_{4,4} = 4! = 4x3x2x1 = 24$

4. Si vogliono disporre 5 uomini e 4 donne su una panca. Quanti ordinamenti sono possibili?

- Senza separare gli uomini dalle donne: $P_{9,9} = 9!$
- Ponendo le donne a destra e gli uomini a sinistra: $P_{5,5} \times P_{4,4} = 5!\ 4!$
- Ponendo le donne da un lato e gli uomini dall'altro :

$2(P_{5,5} \times P_{4,4}) = 2 \times 5! \times 4!$

COMBINAZIONI

1. Calcolare le combinazioni di 5 elementi a 3 a 3

$C_{5,3} = n!/k!(n-k)! = D_{n,k}/ k! = (5 \times 4 \times 3) / (3 \times 2 \times 1)$

2. In quanti modi un insegnante può scegliere 3 alunni da interrogare in una classe di 24 alunni?

$C_{24,3} = (24 \times 23 \times 22)/(3 \times 2 \times 1) = 2024$

3. In quanti modi si possono scegliere 3 libri di matematica tra 5 libri di matematica e 2 libri di medicina tra 4 libri di medicina?

$C_{5,3} \times C_{4,2} = [(5 x4 x3)/ (3 x2 x1)] x [(4 x3)/ (2x1)]$

PROPRIETÀ DEI COEFFICIENTI BINOMIALI

Varie sono le proprietà dei coefficienti binomiali $\binom{n}{k}$ utilizzati nel calcolo delle probabilità e nell'analisi matematica e qui vogliamo indicarne solo alcune.

La formula dei coefficienti binomiali si può esprimere per mezzo dei *fattoriali*:

$$\binom{n}{k} = D_{n,k} = \frac{n!}{n!(n-k)!}$$

1. Proprietà simmetrica dei coefficienti binomiali. Si può dimostrare la seguente eguaglianza:

$$\binom{n}{k} = \binom{n}{n-k}$$

Infatti, applicando la formula precedente si ha:

$$\binom{n}{n-k} = \frac{n!}{(n-k)![n-(n-k)]!} = \frac{n!}{(n-k)!k!} = \binom{n}{k}$$

2. Precedentemente si sono considerate le combinazioni di classe k e si è posto $k \leq n$. Per $k=n$ si ha immediatamente:

$$\binom{n}{n} = \frac{n!}{n!0!} = 1$$

poiché, per convenzione, 0! vale 1.

Per k=0 si ha $\binom{n}{0}\frac{n!}{0!(n-0)!} = \frac{n!}{n!} = 1$

3. Proprietà di Stiefel

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$$

Questa proprietà permette di costruire una tabella di coefficienti binomiali, nota come *triangolo di Tartaglia*.

TRIANGOLO DI TARTAGLIA

	K						
n	0	1	2	3	4	5	6....
0	1						
1	1	1					
2	1	2	1				
3	1	3	3	1			
4	1	4	6	4	1		
5	1	5	10	10	5	1	
6	1	6	15	20	15	5	1

3. ELEMENTI DI CALCOLO DELLE PROBABILITÀ

-TEORIA INGENUA-

Probabilità semplice

1. La probabilità che esca testa dal lancio di una moneta è 1/2.
2. La probabilità che esca 2 dal lancio di un dado è 1/6.
3. La probabilità di avere un figlio maschio è 1/2.
4. Da un'urna contenente 3 palline rosse (R) e 2 palline bianche (B), la probabilità di estrarre una pallina rossa è P(R)=3/5.
5. Data un'urna contenente 20 palline numerate da 1 a 20:
    a) la probabilità di estrarre un numero dispari è 10/20=1/2.
    b) la probabilità di estrarre un numero divisibile per 3 é 6/20=3/10
    c) la probabilità di estrarre un numero divisibile per 5 è 4/20=1/5

Spazio di probabilità

Lo spazio di probabilità dell'evento "lancio dì due monete" è costituito da 4 eventi S(E): {TT,TC,CT,CC }ai quali sono associate le probabilità P(E):{ 1/4,1/4,1/4,1/4 }.

Principio della probabilità composta

Supponiamo che gli eventi che costituiscono l'evento composto siano tra loro indipendenti.

1. Dal lancio di due monete P(TT)=1/2 x 1/2=1/4, P(TC)=1/4, P(CT) = 1/4, P(CC)=1/4.

2. Rifacendoci all'esempio n°4 della probabilità semplice, P(RB)=3/5 x 2/5 se si rimette la prima pallina estratta nell'urna mentre P(RB)=3/5 x 2/4 se non si rimette la prima pallina estratta nell'urna.

3. La probabilità di avere in una famiglia di tre figli la sequenza MMF è P(MMF)=1/2 x 1/2 x 1/2.

4. La probabilità di avere dal lancio di due dadi due volte la faccia 3 è P(3,3)=1/6 x 1/6.

Principio della probabilità totale

Supponiamo che gli eventi che rappresentano le modalità secondo cui l'evento di interesse può manifestarsi siano tra di loro disgiunti

1. Rifacendoci all'esempio n° 1 della probabilità composta, la probabilità che dal lancio contemporaneo o in successione di due monete si abbia prima testa e poi croce oppure prima croce e poi testa è P(TC o CT)=1/4 + 1/4=1/2..

2. Rifacendoci all'esempio n° 4 della probabilità semplice, se si rimette la prima pallina estratta nell'urna, la probabilità di estrarre prima una pallina rossa e poi una pallina bianca oppure prima una pallina bianca e poi una pallina rossa è P(RB o BR)=6/25 + 6/25.

3. La probabilità di estrarre da un mazzo di 40 carte un asso o una figura è P(A o F) =4/40 + 12/40

Principio di indipendenza

1. Consideriamo lo spazio degli eventi del lancio di tre monete S=(TTT,TTC,TCT,TCC,CTT,CTC,CCT,CCC) ed indichiamo con A l'evento "una testa o due teste" ossia A = (TTC, TCT, TCC, CTT, CTC,CCT) e con B l'evento "almeno due teste" ossia B = (TTT, TTC,TCT,CTT) ne segue che l'evento ($A \cap B$) è formato dagli eventi (TTC,TCT,CTT) . Diciamo che gli eventi A e B sono fra loro indipendenti, infatti:

P(A)=6/8, P(B)=4/8, P(A∩B)=3/8 e quindi P(A∩B)=P(A) x P(B)

2. Rifacendoci all'esempio n° 5 della probabilità semplice la probabilità di estrarre un numero dispari e divisibile per 3 è pari a 3/20. Poiché P (Numero dispari) x P(Numero divisibile per 3)=1/2 x 3/10=3/20, si deduce che i due eventi "Numero dispari" e "Numero divisibile per 3" sono indipendenti.

FUNZIONE BINOMIALE

ESERCIZI

- Qual è la probabilità che lanciando 5 volte una moneta esca 0,1,2,3,4,5 volte testa? (Vedi grafico).

P (0 teste)= $b(5,1/2,0)=C_{5,0}\,(1/2)^0\,(1/2)^5=1/32$

P(1 testa)= $b(5,1/2,1)=C_{5,1}\,(1/2)^1\,(1\,/2)^4$,= 5/32

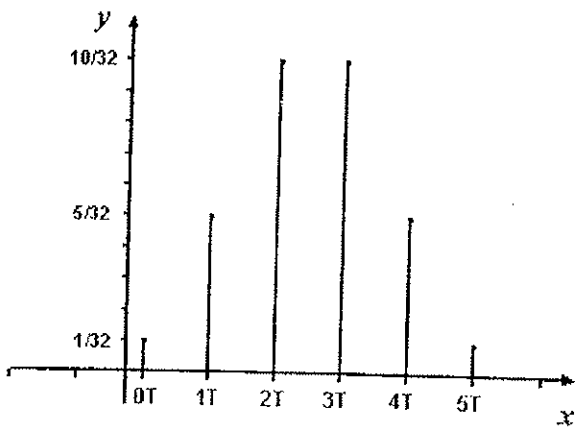
P(2 teste)=$b(5,1/2,2)=C_{5,2}\,(1/2)^2$x.$(1/2)^3=10/32$

P(3 teste)=$b(5,1/2,3)=C_{5,3}(1/2)^3(1/2)^2=10/32$

P(4 teste)= $b(5,1/2,4)=C_{5,4}(1/2)^4(1/2)^1=5/32$

P (5 teste)= $b(5,1/2,5)=C_{5,5}\,(1/2)^5(1/2)^0$ =1/32

Rappresentazione grafica



- Qual è la probabilità di avere 2 figli maschi in una famiglia di 6 figli?

$B(6,1/2,2)=C_{6,2}\,(1/2)^2(1/2)^4=15/64$

- Qual è la probabilità in una famiglia di 6 figli in cui tutti e due i genitori sono portatori dei tratto talassemico di avere 4 figli malati?
Ricordiamo che nel caso di una malattia autosomica, recessiva p=1/4 e q=3/4.
$B(6,1/4,4)=C_{6,4}(1/4)^4 \times (3/4)^2=15x1/256\ x9/16=135/4096=0.03296$

- Qual è la probabilità che in una famiglia di 4 figli di avere esattamente 2 figli maschi?
$B(4,1/2,2)=C_{4,2}\,(1/2)^4$

- Qual è la probabilità che nella stessa famiglia gli ultimi 2 figli siano maschi?
$P(FFMM)= (1/2)^4$

- Qual è la probabilità in una famiglia di 5 figli in cui tutti e due i genitori siano portatori dei tratto talassemico di avere nessun figlio malato?
B(5,1/4,0)=
- Qual è la probabilità nella stessa famiglia di avere 2 figli talassemici e gli altri sani?
B(5,1/4,2)=

- Qual è la probabilità, nella stessa famiglia, di avere i primi due figli talassemici e gli altri 3 no?
$P(TTSSS)= (1/4)^2 \times (3/4)^3$

FUNZIONE NORMALE (tab. 1)

<u>COME SI USA LA TAVOLA DELLA CURVA NORMALE STANDARDIZZATA (CURVA DI t)</u>

Nella tavola della curva normale standardizzata, la prima colonna contrassegnata da t riporta i valori di t con una sola cifra decimale, la prima riga riporta la seconda cifra decimate di t. Tutti gli altri valori sono valori di area.

Per trovare il valore di area che ci interessa si procede verso il basso nella prima colonna fino al valore di t desiderato prendendolo con una sola cifra decimale. Quindi si procede verso destra e ci si ferma al valore di area che si trova all'incrocio con la perpendicolare abbassata dalla seconda cifra decimale .

1. Trovare l'area compresa tra t=0 e t=1,5

Nella tavola si procede verso il basso nella colonna segnata da t fino a raggiungere il valore 1.5 poi si procede verso destra fino alla colonna segnata dallo 0. Infatti 1.5=1.50.

Il valore 0.432 è l'area richiesta e rappresenta la probabilità che t sia compresa tra 0 e 1.5.

$P (0 \leq t \leq 1.5) = 0.4032$

2. Trovare l'area compresa tra t=-0.65 e t=0

Nella tavola si procede verso il basso nella colonna segnata da t fino a raggiungere il valore t=0.6 poi si procede verso destra fino alla colonna segnata da 5. Il valore 0.2422 è l'area richiesta.

$P (-0.65 \leq t \leq 0) = 0.2422$

3. Trovare l'area compresa tra t=-0.65 e t=1.5

$P(-0.65 \leq t \leq 1.5) = P(-0.65 \leq t \leq 0) + P(0 \leq t \leq 1.5) = 0.4032 + 0.2422 = 0.6454$

4. Trovare l'area per $t \geq 2.54$

$P(t \geq 2.54) = P(0 \leq z \leq \infty) - P(0 \leq t \leq 2.54) = 0.5000 - 0.4945 = 0.0055$

5. Trovare l'area per $t \leq -1.42$

$P(t:: \leq -1.42) = P(-\infty \leq t \leq 0) - P(-1.42 \leq t \leq 0) = 0.5000 - 0.4222 = 0.0778$

6. Il valore medio dell'altezza dei giovani di leva è μ=170cm con una deviazione standard σ=10cm. Assumendo che l'altezza sia distribuita normalmente, qual è la proporzione di giovani con altezza tra 178cm e 180cm ossia qual è la probabilità di trovare un giovane con un'altezza compresa tra 178crn e 180cm?

Per prima cosa si esprimono i valori 178cm e 180 cm in unità standard.

178cm corrispondono a:

(178 -170)/10 = 0.80 unità standard

180cm corrispondono a:

(180 -170)/10 = 2.00 unità standard

$P(178 \leq \bar{x} \leq 180) = P(0.80 \leq t \leq 2.00) = 0.4772 - 0.2881 = 0.1891$

La proporzione di giovani con altezza tra 178cm e 180cm è 18.91%

5. RAPPRESENTAZIONE GRAFICA

Dati categorici

Vogliamo rappresentare graficamente il sesso di 400 neonati nella Clinica Ostetrica della Università di Roma La Sapienza.

Il campione era formato di 215 maschi (45.8%) e 182 femmine (54.2%). Di 3 bambini non si aveva il dato sul sesso.

Al dato "maschio" è stato associato il valore 1

Al dato "femmina" è stato associato il valore 0

Messi i dati numerici al computer abbiamo ottenuto i seguenti diagrammi:

Diagramma a barre:

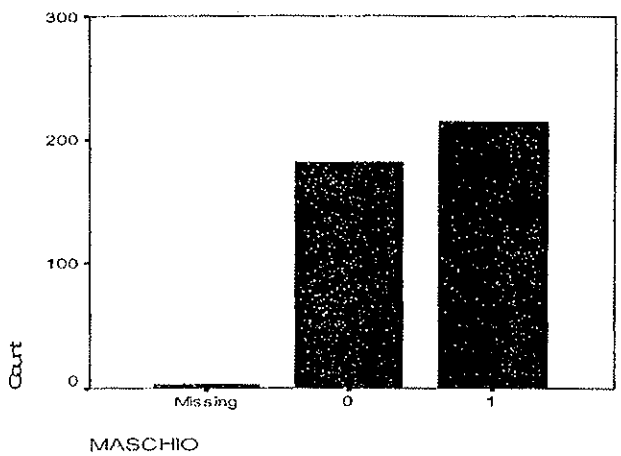
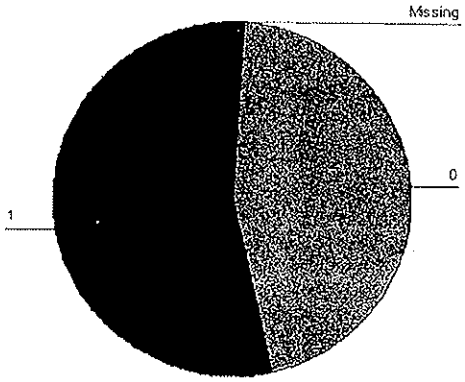


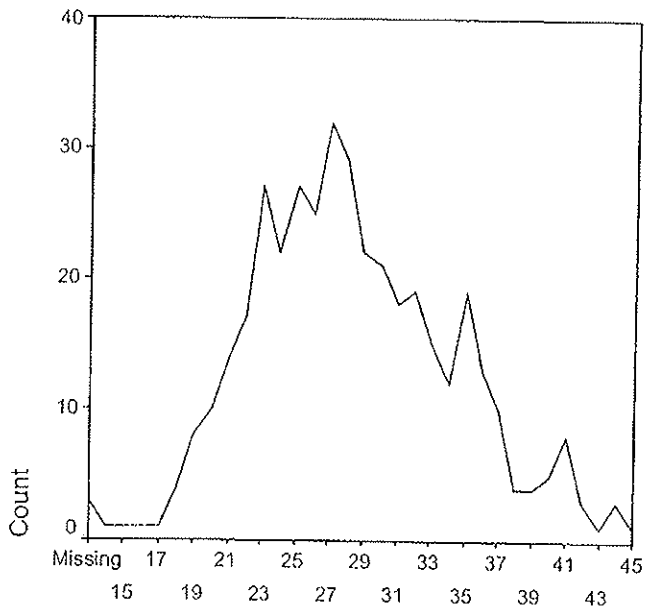
Diagramma circolare:



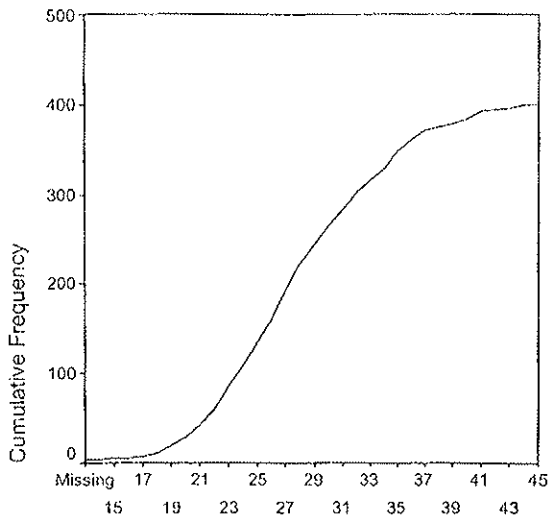
Dati numerici

Se l'età della madre di questi bambini è trattata come una variabile continua, introdotte le età materne al computer abbiamo ottenuto i seguenti diagrammi:

il polígono di frequenze:



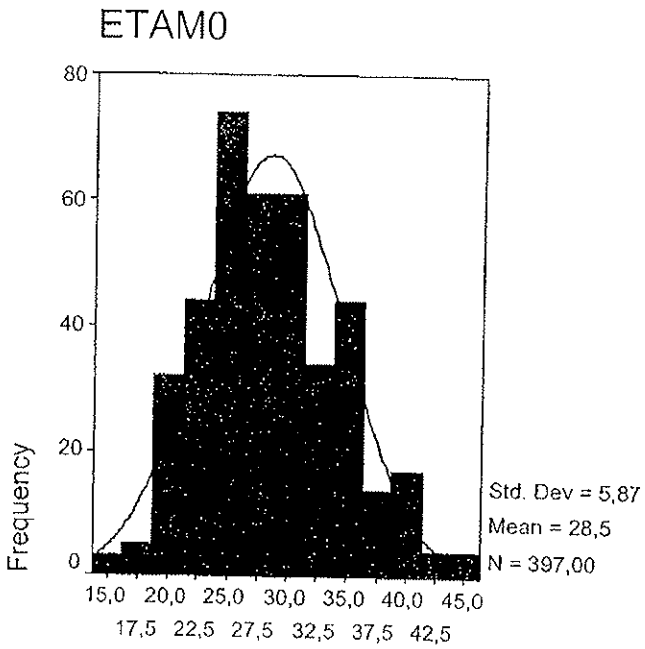
e l'ogiva



Se l'età della madre viene raggruppata in classi:

Classi	valore centrale	frequenza assoluta
[14-16]	15.0	3
[17-18]	17.5	5
[19-21]	20.0	32
[22-23]	22.5	44
[24-26]	25.0	74
[27-28]	27.5	61
[29-31]	30.0	61
[32-33]	32.5	34
[34-36]	35.0	44
[37-38]	37.5	14
[39-41]	40.0	17
[42-43]	42.5	4
[44-46]	45.0	4

si ottiene *l'istogramma di frequenze*



Di tre madri non si aveva il dato.

6. STIMA DI PARAMETRI

ERRORE STANDARD E DEVIAZIONE STANDARD

L'errore standard è la deviazione standard della media campionaria (l'unità statistica è il campione), tuttavia la convenzione è di usare il termine "errore standard" quando vogliamo misurare la precisione di una stima e di usare il termine "deviazione standard" quando vogliamo valutare la variabilità di un parametro in una popolazione (l'unità statistica è l'individuo).

INTERVALLO DI CONFIDENZA (O LIMITI FIDUCIALI) PER UNA MEDIA

In un campione di 100 giovani di 20 anni si è trovato un peso medio di 67 kg ed una deviazione standard di 8kg, Supponendo che il peso dei giovani di 20 anni sia distribuito normalmente, trovare l'intervallo di confidenza al 95% e al 99% dei peso medio dei giovani di 20 anni.

(95%) $67 - 2(0.8) < \mu < 67 + 2(0.8)$ $65.4 < \mu < 68.6$

(99%) $67 - 3(0.8) < \mu < 67 + 3(0.8)$ $64.6 < \mu < 69.4$

INTERVALLO DI CONFIDENZA (O LIMITI FIDUCIALI) PER LA DIFFERENZA TRA DUE MEDIE

Ricordiamoci sempre che quando i campioni sono grandi e abbastanza numerosi possiamo assumere che le medie campionarie sono distribuite normalmente.

I pesi alla nascita di 200 bambini esaminati in un certo ospedale hanno mostrato la seguente distribuzione:

Sesso	n^O	media (gr)	d.s. (gr)
Maschi	97	3280	120
Femmine	103	3150	105

Assumendo che i pesi dei maschi e delle femmine siano distribuiti normalmente con medie μ_1 e μ_2 varianze uguali l'intervallo di confidenza al 95% per la differenza fra le medie delle due popolazioni è

$$(\bar{x}_1 - \bar{x}_2) - 2(ES_{\bar{x}_1 - \bar{x}_2}) \prec \mu_1 - \mu_2 \prec (\bar{x}_1 - \bar{x}_2) + 2(ES_{\bar{x}_1 - \bar{x}_2})$$

$$130 - 2\sqrt{\frac{120^2}{97} + \frac{105^2}{103}} \prec \mu_1 - \mu_2 \prec 130 + 2\sqrt{\frac{120^2}{97} + \frac{105^2}{103}}$$

$$130 - 2(15.98) \prec \mu_1 - \mu_2 \prec 130 + 2(15.98)$$

$$98.84 \prec \mu_1 - \mu_2 \prec 161.96$$

con un'affidabilità del 95%.

TEST DI IPOTESI

CHI-QUADRATO (tab. 2)

(COME SI USA LA TABELLA DEL CHI-QUADRATO)

Nel test che stiamo eseguendo dopo aver calcolato il chi-quadrato e individuato i gradi di libertà, si scorre nella tabella lungo la riga corrispondente al grado di libertà fino a trovare un valore di χ^2 maggiore del valore calcolato.

Ci portiamo sul valore di chi-quadrato immediatamente precedente e risaliamo verso la prima riga in cui sono riportate le probabilità. Il valore di probabilità corrispondente al valore di chi quadrato considerato ci indica la significatività dei test ossia la probabilità con la quale viene rigettata Ho (la probabilità di sbagliare rigettando H_o).

ESERCIZI

Tabella di contingenza:

Osservati

10	7	2	25
8	15	4	7

Attesi

10.15	12.41	3.38	18.05
7.85	9.59	2.61	13.95

Gradi di libertà: (2-1) x (4-1)=3

$$\chi^2 = \frac{\sum_{i=1}^{n}(o_i - a_i)^2}{a_i} = 12.85 \qquad p < 0.005$$

Osservati

78	4
85	20

Osservati

7	14	20	30
45	7	1	28
7	3	25	7

TEST DI IPOTESI MEDIANTE USO DELLA BINOMIALE

Supponiamo di avere costruito una tabella costituita da 1.000.000 di numeri casuali e di non avere la possibilità di contare quanti siano i pari e quanti i dispari -Ho: n° dei pari=n° dei dispari ossia la probabilità dei numeri pari = probabilità dei numeri dispari=1/2.

-Supponiamo di estrarre un campione di 10 numeri e di osservare 0 pari e 10 dispari. Il test del χ^2 ci aiuterà a valutare la differenza tra valori osservati e valori attesi assumendo l'ipotesi zero che gli attesi siano 5 pari e 5 dispari.

	pari	dispari
osservati	0	10
attesi	5	5

$$\chi^2 = 10 \quad gl = 1 \quad p<0.001$$

Se la nostra ipotesi zero è giusta un evento così fatto cioè 0 pari e 10 dispari si presenta con una frequenza inferiore al 1 % per cui il buon senso ci consiglia di rigettare l'ipotesi zero e di dedurre che nella nostra tabella il numero dei pari non è uguale al numero dei dispari.

Possiamo tuttavia utilizzare la distribuzione binomiale per arrivare allo stesso risultato.
Se "successo"= "presenza di un numero pari":

$$b(10,1/2,0)= 1\cdot p^0\cdot q^{10} = 1\ (1/2)^0(1/2)^{10} = 1\cdot 1\cdot(1/2)^{10} <0.001$$

t DI STUDENT (tab. 3)

1. Siano A e B due campioni indipendenti con n_1=15 ed n_2=16 pesi corporei. Essi sono stati estratti da due popolazioni distribuite normalmente con varianze non note che non si suppongono uguali. Si vuole testare se i campioni siano stati estratti dalla stessa popolazione.

A: 50.9; 52.7; 51.6; 59.2; 50.3; 51.1; 51.7; 53.8; 54.9; 55.7; 56.8; 58.2; 55.3; 42.4; 57.9.

B: 59.4; 51.3; 56.6; 58.8; 58.4; 51.0; 53.5; 51.4; 58.2; 56.7; 59.3; 57.4; 59.7; 55.8; 46.1; 50.0.

Campione A:

Media $\bar{x}_1 = 53.50$ $s_1 = 4.212$ $n_1 = 15$

Campione B:

Media $\bar{x}_2 = 55.23$ $s_2 = 4.133$ $n_2 = 16$

$H_0: \mu_1 = \mu_2$ e la statistica test è:

$$t = \frac{\left[\left(\bar{x}_1 - \bar{x}_2\right) - 0\right]}{ES_{\bar{x}_1 - \bar{x}_2}}$$

dove $ES_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_1^2/n_1 + s_2^2/n_2}$ con $n_1 + n_2 - 2$ = gradi di libertà

Sostituendo i valori dell'esercizio, si ha:

$$t = \frac{(53.5 - 55.23)}{\sqrt{1.18 + 1.07}} = \frac{-1.73}{1.5} = -1.15$$

t= -1.15 con 29 gradi di libertà, p>0.20.

Il test non è significativo e quindi i due campioni sono stati estratti dalla stessa popolazione.

2. Un campione di 10 donne viene sottoposto ad una dieta ipocalorica per 3 mesi. Indichiamo con A i pesi prima della dieta e con B i pesi dopo la dieta. Ci chiediamo se la dieta ha avuto effetto sul peso.

A: 55.60; 63.90; 60.00; 59.10; 55.30; 62.50; 63.70; 60.10; 72.90; 68.30.

B: 53.20; 60.40; 57.30; 51.60; 48 50; 52.30; 59.50; 58,12; 65.30; 62.40.

H_0: $\bar{x}_d = 0$

Si calcolano prima le differenze a due a due fra i valori, si calcola il valor medio della distribuzione delle differenze delle medie con l'errore standard.

A-B: 2.4 3.5...2.7 7.5 6.8 10.2 4.2 1.98 7.6 5.9

$\bar{x}_d = 5.278$ $ES_d = 0.8653$

t= 5.278/0.8653=6.0996 con 9 gradi di libertà $p<<0.001$

Il test è altamente significativo e quindi la dieta ha fatto effetto.

TIPI DI VARIABILI

Consideriamo le variabili contenute in un questionario preparato per un reparto di Pediatria per uno studio epidemiologico su bambini nati nei quattro anni 1993-1996 e vediamo che tipo di variabili sono.

Variabile	*Tipo di variabile*
<u>Dati riguardanti il bambino</u>	
mese di nascita	numerica discreta
anno di nascita	"
Peso alla nascita	numerica continua
Peso corretto per età gestazionale	ordinale (centili)
Peso della placenta	numerica continua
Gruppo ABO	categorica poiltomica
Gruppo Rh	categorica dicotomica (Rh+1Rh-)
Sesso	categorica dicotomica
Ordine di genitura	numerica discreta
Età gestazionale(in settimane)	numerica discreta
Fototerapia	categorica dicotomica (si/no)
Deficit di G-6-PD	categorica dicotomica
<u>Dati riguardanti la madre</u>	
Mese di nascita	numerica discreta
Anno di nascita	"
Età al momento dei parto	"
Cestosi	categorica dicotomica (si/no)
Gruppo ABO	categorica politomica

Gruppo Rh	categorica dicotomica (si/no)
Fumo	categorica dicotomica
Alcool	categorica dicotomica
Aborti	numerica discreta
Diabete	categorica dicotomica
Farmaci in gravidanza	categorica politomica

APPENDICE

ELEMENTI DI GENETICA

MEIOSI

La meiosi é un processo di divisione cellulare a cui vanno incontro le cellule della linea germinativa. Essa ha come evento finale la produzione di cellule aploidi (gameti).

Al momento della fecondazione, dall'unione di gameti di sesso opposto si forma una cellula diploide (zigote) che, attraverso un processo di sviluppo, darà origine all'individuo diploide che rappresenta il tipo della specie.

La meiosi é costituta da due divisioni cellulari successive: riduzionale ed equazionale.

PRIMA LEGGE DI MENDEL

Dall'incrocio di due linee pure "AA" e "aa"; ossia dall'incrocio di un individuo omozigote per uno dei due alleli al locus *A* con un altro individuo omozigote per l'altro allele al locus *A*, alla prima generazione, F1, si ottengono solo individui eterozigoti Aa.

Negli esperimenti di Mendel, poiché A é dominante su a, alla F1 si ottengono tutti individui con fenotipo A.

Sempre nel caso della dominanza, alla F2, ossia dall'incrocio Aa x Aa si ottengono 3/4 di soggetti con fenotipo "A" (genotipi AA e Aa) e 1/4 di soggetti con fenotipo "a" (genotipo aa).

Nel caso di alleli codominanti, invece, alla F2 si hanno 1/4 di soggetti con genotipo e fenotipo "AA", 2/4 di soggetti con genotipo e fenotipo "Aa" ed 1/4 di soggetti con genotipo e fenotipo "aa".

Alla meiosi ogni individuo con genotipo Aa produrrà gameti di tipo "A" con probabilità 1/2 e gameti di tipo "a" con probabilità 1/2.

L'incrocio di due individui con genotipo Aa darà (la probabilità relativa ad ogni genotipo è posta tra parentesi):

	A(1 /2)	a(1/2)	(gameti del 1° individuo)
A(1/2)	AA(1/4)	Aa(1/4)	
a(1/2)	aA(1 /4)	aa(1/4)	

(gameti del 2° individuo)

Infatti, se assumiamo che l'accoppiamento di un uovo con uno spermatozoo sia casuale, i due eventi sono tra di loro indipendenti e per il principio della probabilità composta si ha: P(AA)=P(A) x P(A)=1/4, P(Aa)=P(A) x P(a)=1/4, P(aA)=P(a) x P(A)=1/4, P(aa)=P(a) x P(a)=1/4.
Poiché non possiamo distinguere il genotipo Aa dal genotipo aA in quanto eterozigoti con fenotipo A, ci troviamo di fronte ad un evento che può manifestarsi secondo due modalità tra di loro escludentesi. Per il principio della probabilità totale, la probabilità dell'evento eterozigote è uguale alla somma delle probabilità delle due modalità Aa e aA, ossia P(eterozigote)=P(Aa) + P(aA).

FREQUENZE ALLELICHE

In un *locus* con due alleli "A" e "a" la frequenza dell'allele "A" è uguale al numero di *loci* occupati da "A"' sul numero totale di *loci* a disposizione per '"A" e "a" (la frequenza dell'allele "a" è il complemento a 1 della frequenza dell'allele "A"). È implicita l'assunzione che ogni individuo, essendo diploide, ha due *loci* per gene. Ossia rappresentiamo ciascun individuo con lo zigote da cui esso deriva.

Calcolo delle frequenze alleliche

Sia dato un campione di 100 individui e sia in esso esaminato il gruppo sanguigno MN, determinare la frequenza dei suoi alleli.
Nel *locus* MN si hanno 2 alleli, L^M e L^N codominanti per cui i genotipi possibili sono: L^M/L^M, L^M/L^N, L^N/L^N ed i fenotipi sono M, MN, N. Supponiamo che le

frequenze assolute dei tre fenotipi (genotipi) siano rispettivamente 40, 50 e 10. Si avrà:

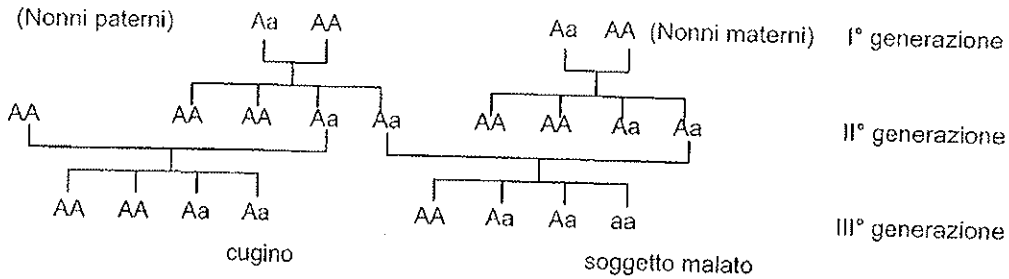
$f(L^M)=(80+50)/200=0,65$ $f(L^N)=(50+20)/200=0,35$

(*i loci* sono 200 perchè ogni individuo, essendo diploide, mette a disposizione 2 *loci*).

Nel caso di un allele dominante su un altro allele, il calcolo diretto delle frequenze alleliche non si può fare perchè è impossibile distinguere il genotipo omozigote dominante dal genotipo eterozigote.

Nel *locus* Rh ci sono 2 alleli D e d, con D dominante su d e tre genotipi: DD, Dd, dd. I primi due genotipi corrispondono al fenotipo Rh(+) ed il terzo genotipo corrisponde al fenotipo Rh(-).

VALUTAZIONE DELLA PROBABILITÀ DI ESSERE ETEROZIGOTE IN BASE AL GRADO DI PARENTELA CON UN SOGGETTO MALATO PER UNA MALATTIA AUTOSOMICA RECESSIVA.



Qual è la probabilità che un cugino primo germano di un individuo malato sia un portatore sano?

Per valutare la probabilità che il cugino di un soggetto malato sia eterozigote si risale l'albero genealogico dal soggetto eterozigote fino al progenitore comune. Entrambi i genitori dei soggetto malato devono essere eterozigoti. La probabilità che uno dei due nonni (per ciascuna coppia di nonni) sia eterozigote è 1/2, la probabilità che lo zio sia eterozigote è 1/2, la probabilità che il cugino sia eterozigote è 1/4.

VALUTAZIONE DEL RISCHIO DI AVERE UN FIGLIO OMOZIGOTE PER UNA MALATTIA AUTOSOMICA RECESSIVA .

La procedura può essere così riassunta:

1. Si definisce l'accoppiata genica che può dare un soggetto malato (coppia pericolosa ossia ambedue i genitori eterozigoti)
2. Si stabilisce per ciascun elemento della coppia la probabilità di essere eterozigote.
3. Si moltiplicano tra loro le due probabilità e si ottiene la probabilità di coppia pericolosa.
4. Poiché da una coppia pericolosa solo 1/4 dei figli sarà alatasi moltiplica la probabilità ottenuta nel punto 3. per 1/4.

ESEMPI:

Data una malattia autosomica recessiva di frequenza 1/10.000:

1. Qual è la probabilità di avere un figlio malato se un soggetto preso a caso nella popolazione si accoppia con un altro soggetto preso a caso nella popolazione?
2. Qual è la probabilità di avere un figlio malato se la sorella di un soggetto malato si accoppia con un soggetto preso a caso nella popolazione?
3. Qual è la probabilità di avere un figlio malato se la sorella di un soggetto malato si accoppia con il fratello di un soggetto malato?

Soluzioni:

Se indichiamo con "a" il gene responsabile della malattia e con q la sua frequenza:

1. $q = \sqrt{1/10000} = \sqrt{0.0001} = 0.01; \qquad p = 1 - q = 1 - 0.01 = 0.99;$

   $2pq = 2 \cdot 0.01 \cdot 0.99 \approx 2/100$

la probabilità che si accoppino due soggetti eterozigoti: P(Aa) x P(Aa)=2/100 x 2/100=4/10.000; la probabilità che un figlio sia malato è 1/4; la probabilità che si abbia un figlio malato dall'unione di due soggetti presi a caso nella popolazione è 4/10.000 x 1/4=1/10.000

2. 2/3 x 2/100 x 1/4=1/300
3. 2/3 x 2/3 x 1/4=1/9

La sorella sana di un soggetto malato ha una probabilità di essere una portatrice pari a 2/3. Si tratta di una probabilità condizionata all'essere fenotipicamente sana. Infatti, i 3/4 delle fratrie dove è presente un malato è costituito da soggetti sani e tra questi 2/3 sono Aa e 1/3 sono AA.

PRINCIPIO DI HARDY-WEINBERG

Esiste una corrispondenza tra frequenze alleliche e frequenze genotipiche: in un sistema con due alleli codominanti A^1 e A^2 di frequenza rispettivamente p e q, con p+q=1, la frequenza dei genotipi A^1A^1, A^1A^2, A^2A^2 è rispettivamente p^2, 2pq, q^2.

	A^1 (p)	A^2(q)
A^1 (p)	P.P	p.q
A^2 (q)	q.p	q.q

Si tratta di una generalizzazione della 1° legge di Mendel nel passaggio dalla F1 alla F2 in quanto nel caso dell'incrocio di Mendel l'esperimento era stato strutturato in modo da avere p=1/2 e q=1/2.

Nel caso di malattie autosomiche recessive la frequenza del gene letale, q, è piuttosto bassa, in generale q< 1%. Fanno eccezione il gene per la fibrosi cistica dei pancreas, che nella popolazione europea raggiunge il 2%, ed i geni per alcune malattie come la talassemia e l'emoglobinopatia S che in alcune popolazioni sottoposte nel passato ad endemia malarica possono raggiungere frequenze notevolmente più elevate.

In generale, se un gene letale ha la frequenza q=0.01, f(a)=0.01, f(A)= 1-0.01=0.99 (A è il gene dominante sano), $f(AA)= p^2 = 0.99 \times 0.99 \sim 0.98$, $f(Aa)= 2pq= 2 \times 0.99 \times 0.01 \sim 2q$ ossia ~ 0.02, $f(aa)= q^2=0.01 \times 0.01=0.0001$.

La frequenza dei portatori di un gene letale (eterozigoti) è circa due volte la frequenza del gene stesso infatti poiché f(A)=0.99 è circa 1, f(Aa) è circa 0.02.

TEST DI HARDY-WEINBERG

SISTEMA DI GRUPPO SANGUIGNO MN

ALLELI L^M, L^N (alleli codominanti)

GENOTIPI L^M/L^M, L^M/L^N, L^N/IL^N

FENOTIPI M MN N

$p=f(L^M), q=f(L^N)$; $p^2=f(L^M/L^M)=f(M)$; $q^2=f(L^N/IL^N)=f(N)$;

$2pq=f(L^M/L^N)=f(MN)$

Esempio:

Supponiamo di aver esaminato 100 soggetti e di aver ottenuto I seguenti risultati:

M=40, MN=40 e N=20

Calcolo delle frequenze alleliche:

$f(L^M)$=n° dei loci occupati da L^M fratto il n° totale dei loci a disposizione

$f(L^M)=[(40 \times 2) + 40]/200=0.6$

$f(L^N)$=n° dei loci occupati da L^N fratto il numero totale di loci a disposizione

$f(L^N)=[(20 \times 2)+40]/200=0.4$

Calcolo dei valori attesi secondo H.W.

Se la legge di Hardy-Weinberg è rispettata:

$p^2=0.6 \times 0.6 =0.36$; $2pq=2 \times 0.6 \times 0.4 = 0.48$; $q^2=0.4 \times 0.4= 0.16$

Queste sono probabilità quindi i valori attesi saranno:

M=36; N=16; MN=48

I gradi di libertà = 3-2=1(n° di genotipi - n° alleli} e applicando il test del chi-quadrato si ha: $\chi^2 = 3$.

Il test di equilibrio di Hardy-Weinberg è possibile solo nel caso di un sistema con due o più alleli codominanti tra loro.

Nel caso di sistemi con due alleli uno dominante sull'altro (es. Rh), il test non può essere eseguito perchè non è possibile il calcolo diretto delle frequenze alleliche.

TAPPE PER ESEGUIRE IL TEST DI HARDY-WEINBERG

1. Vengono calcolate le frequenze alleliche attese nel campione.
2. Tenendo presente il principio di Hardy-Weinberg si calcolano le frequenze genotipiche attese:

   $f(A^1/A^1)=p^2;\ f(A^1/A^2)=2pq;\ f(A^2/A^2)=q^2$
3. Si moltiplicano le frequenze genotipiche relative attese per il numero di individui studiati e si ottengono così le frequenze assolute attese.
4. Si valutano le differenze fra valori osservati e valori attesi mediante il test del Chi-Quadrato.

N.B.: Non bisogna confondere il test di Hardy-Weinberg con la tabella di contingenza che è un test di indipendenza.

Per il calcolo degli attesi nel test di H.W. si assume che valga il principio di Hardy-Weinberg per quanto concerne la relazione tra frequenze alleliche e genotipiche, mentre nella tabella di contingenza si assume che i parametri esaminati siano indipendenti fra loro. Si applica lo stesso test statistico a due situazioni differenti.

ESERCIZI

OSSERVATI

A^1A^1	A^1A^2	A^2A^2
25	7	31

ATTESI

A^1A^1	A^1A^2	A^2A^2
12.7	31.18	19.05

Gradi di libertà: (3 - 2) = 1

χ^2=36.5 p<<0.001

OSSERVATI

A^1A^1	A^1A^2	A^2A^2
	36	14

OSSERVATI

A^1A^1	A^2A^2	A^3A^3	A^1A^2	A^1A^3	A^2A
14	71	27	14	5	3

CONDIZIONI PER L'EQUILIBRIO DI HARDY-WEINBERG

Affinché l'equilibrio sia rispettato è necessario che siano verificati 5 condizioni:

1) Popolazione di effettivo (n° di individui) infinito
2) Panmixia, ossia che l'accoppiamento tra individui avvenga a caso e non in base a differenze o somiglianze genotipiche o fenotipiche
3) Non vi sia migrazione ossia immissione di soggetti da altre popolazioni con differenti frequenze alleliche
4) Non vi sia mutazione
5) Non vi sia selezione

Se la popolazione è piccola si hanno continuamente importanti fluttuazioni delle frequenze alleliche nel passaggio da una generazione all'altra. Questo porta alla perdita della variabilità genetica.

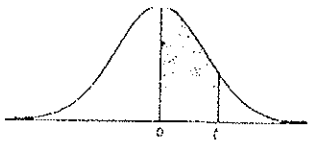
La panmixia nell'uomo prevede sia la panmixia in senso stretto cioè l'accoppiamento a caso tra individui, sia la pangamia ossia l'accoppiamento a caso tra gameti di sesso opposto.

L'accoppiamento tra individui imparentati (ad esempio cugini primi) porta ad un difetto di eterozigoti e ad un corrispondente aumento degli omozigoti nella progenie. Nell'ambito delle famiglie in cui sia presente una determinata malattia autosomica recessiva (ad esempio fenilchetonuria) la frequenza di matrimoni tra consanguinei è maggiore rispetto a quella osservata nella popolazione generale. Viceversa nell'ambito delle coppie costituite da soggetti imparentati la frequenza delle malattie autosomiche recessive è maggiore rispetto all'atteso.

La selezione naturale è chiaramente osservabile nel caso di geni letali che portano a morte prima dell'età riproduttiva o che impediscano l'attività riproduttiva. La selezione naturale è importante a livello riproduttivo: i portatori di geni associati a scarsa efficienza riproduttiva riescono a passare i propri geni alla generazione successiva con difficoltà maggiore rispetto ai portatori di geni associati ad alta efficienza riproduttiva.

La mancanza di rispetto delle condizioni 1,3,4,5 porta a variazione di frequenze geniche nel passaggio da una generazione all'altra, mentre il mancato rispetto della condizione 2 porta ad una differente associazione degli alleli allo stato diploide ma non a variazione delle frequenze da una generazione all'altra. Quindi affinchè sia rispettato l'equilibrio di H.W. Sono necessari 5 condizioni mentre affinchè sia rispettata la costanza delle frequenze alleliche nel passaggio da una generazione all'altra sono necessarie solo 4 condizioni.

Tab. I. AREE DELLA CURVA NORMALE STANDARD $\mu=0$ $\sigma=1$
La tabella dà le aree sotto la curva normale standard
in steps di 0.01



t	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Tab. 2. V VALORI CRITICI DELLA DISTRIBUZIONE DEL CHI-QUADRATO

υ=gradi di libertà α=probabilità (errore di I^a specie)

ν \ α	.995	.975	.9	.5	.1	.05	.025	.01	.005	.001	α / ν
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	10.828	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	13.816	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	16.266	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	18.467	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	20.515	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	22.458	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	24.322	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	26.124	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	27.877	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	29.588	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	31.264	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	32.910	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	34.528	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	36.123	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	37.697	15
16	5.142	6.908	9.312	15.338	23.542	26.296	28.845	32.000	34.267	39.252	16
17	5.697	7.564	10.085	16.338	24.769	27.587	30.191	33.409	35.718	40.790	17
18	6.265	8.231	10.865	17.338	25.989	28.869	31.526	34.805	37.156	42.312	18
19	6.844	8.907	11.651	18.338	27.204	30.144	32.852	36.191	38.582	43.820	19
20	7.434	9.591	12.443	19.337	28.412	31.410	34.170	37.566	39.997	45.315	20
21	8.034	10.283	13.240	20.337	29.615	32.670	35.479	38.932	41.401	46.797	21
22	8.643	10.982	14.042	21.337	30.813	33.924	36.781	40.289	42.796	48.268	22
23	9.260	11.688	14.848	22.337	32.007	35.172	38.076	41.638	44.181	49.728	23
24	9.886	12.401	15.659	23.337	33.196	36.415	39.364	42.980	45.558	51.179	24
25	10.520	13.120	16.473	24.337	34.382	37.652	40.646	44.314	46.928	52.620	25
26	11.160	13.844	17.292	25.336	35.563	38.885	41.923	45.642	48.290	54.052	26
27	11.808	14.573	18.114	26.336	36.741	40.113	43.194	46.963	49.645	55.476	27
28	12.461	15.308	18.939	27.336	37.916	41.337	44.461	48.278	50.993	56.892	28
29	13.121	16.047	19.768	28.336	39.088	42.557	45.722	49.588	52.336	58.301	29
30	13.787	16.791	20.599	29.336	40.256	43.773	46.979	50.892	53.672	59.703	30
31	14.458	17.539	21.434	30.336	41.422	44.985	48.232	52.191	55.003	61.098	31
32	15.134	18.291	22.271	31.336	42.585	46.194	49.480	53.486	56.329	62.487	32
33	15.815	19.047	23.110	32.336	43.745	47.400	50.725	54.776	57.649	63.870	33
34	16.501	19.806	23.952	33.336	44.903	48.602	51.966	56.061	58.964	65.247	34
35	17.192	20.569	24.797	34.336	46.059	49.802	53.203	57.342	60.275	66.619	35
36	17.887	21.336	25.643	35.336	47.212	50.998	54.437	58.619	61.582	67.985	36
37	18.586	22.106	26.492	36.335	48.363	52.192	55.668	59.892	62.884	69.346	37
38	19.289	22.878	27.343	37.335	49.513	53.384	56.896	61.162	64.182	70.703	38
39	19.996	23.654	28.196	38.335	50.660	54.572	58.120	62.428	65.476	72.055	39
40	20.707	24.433	29.051	39.335	51.805	55.758	59.342	63.691	66.766	73.402	40
41	21.421	25.215	29.907	40.335	52.949	56.942	60.561	64.950	68.053	74.745	41
42	22.138	25.999	30.765	41.335	54.090	58.124	61.777	66.206	69.336	76.084	42
43	22.859	26.785	31.625	42.335	55.230	59.304	62.990	67.459	70.616	77.419	43
44	23.584	27.575	32.487	43.335	56.369	60.481	64.202	68.710	71.893	78.750	44
45	24.311	28.366	33.350	44.335	57.505	61.656	65.410	69.957	73.166	80.077	45
46	25.042	29.160	34.215	45.335	58.641	62.830	66.617	71.201	74.437	81.400	46
47	25.775	29.956	35.081	46.335	59.774	64.001	67.821	72.443	75.704	82.720	47
48	26.511	30.755	35.949	47.335	60.907	65.171	69.023	73.683	76.969	84.037	48
49	27.249	31.555	36.818	48.335	62.038	66.339	70.222	74.919	78.231	85.351	49
50	27.991	32.357	37.689	49.335	63.167	67.505	71.420	76.154	79.490	86.661	50

Tab. 3. VALORI CRITICI DELLA DISTRIBUZIONE DEL t di STUDENT

ν=gradi di libertà α=probabilità (errore di I^a specie)

ν \ α	0.9	0.5	0.4	0.2	0.1	0.05	0.02	0.01	0.001	α / ν
1	.158	1.000	1.376	3.078	6.314	12.706	31.821	63.657	636.619	1
2	.142	.816	1.061	1.886	2.920	4.303	6.965	9.925	31.598	2
3	.137	.765	.978	1.638	2.353	3.182	4.541	5.841	12.924	3
4	.134	.741	.941	1.533	2.132	2.776	3.747	4.604	8.610	4
5	.132	.727	.920	1.476	2.015	2.571	3.365	4.032	6.869	5
6	.131	.718	906	1.440	1.943	2.447	3.143	3.707	5.959	6
7	.130	.711	.896	1.415	1.895	2.365	2.998	3.499	5.408	7
8	.130	.706	.889	1.397	1.860	2.306	2.896	3.355	5.041	8
9	.129	.703	.883	1.383	1.833	2.262	2.821	3.25[illegible]	4.781	9
10	.129	.700	.879	1.372	1.812	2.228	2.764	3.169	4.587	10
11	.129	.697	.876	1.363	1.796	2.201	2.718	3.106	4.437	11
12	.128	.695	.873	1.356	1.782	2.179	2.681	3.055	4.318	12
13	.128	.694	.870	1.350	1.771	2.160	2.650	3.012	4.221	13
14	.128	.692	.868	1.345	1.761	2.145	2.624	2.977	4.140	14
15	.128	.691	.866	1.341	1.753	2.131	2.602	2.947	4.073	15
16	.128	.690	.865	1.337	1.746	2.120	2.583	2.921	4.015	16
17	.128	.689	.863	1.333	1.740	2.110	2.567	2.898	3.965	17
18	.127	.688	.862	1.330	1.734	2.101	2.552	2.878	3.922	18
19	.127	.688	.861	1.328	1.729	2.093	2.539	2.861	3.883	19
20	.127	.687	.860	1.325	1.725	2.086	2.528	2.845	3.850	20
21	.127	.686	.859	1.323	1.721	2.080	2.518	2.831	3.819	21
22	.127	.686	.858	1.321	1.717	2.074	2.508	2.819	3.792	22
23	.127	.685	.858	1.319	1.714	2.069	2.500	2.807	3.767	23
24	.127	.685	.857	1.318	1.711	2.064	2.492	2.797	3.745	24
25	.127	.684	.856	1.316	1.708	2.060	2.485	2.787	3.725	25
26	.127	.684	.856	1.315	1.706	2.056	2.479	2.779	3.707	26
27	.127	.684	.855	1.314	1.703	2.052	2.473	2.771	3.690	27
28	.127	.683	.855	1.313	1.701	2.048	2.467	2.763	3.674	28
29	.127	.683	854	1.311	1.699	2.045	2.462	2.756	3.659	29
30	.127	.683	.854	1.310	1.697	2.042	2.457	2.750	3.646	30
40	.126	.681	.851	1.303	1.684	2.021	2.423	2.704	3.551	40
60	.126	.679	.848	1.296	1.671	2.000	2.390	2.660	3.460	60
120	.126	.677	.845	1.289	1.658	1.980	2.358	2.617	3.373	120
∞	.126	.674	.842	1.282	1.645	1.960	2.326	2.576	3.291	∞