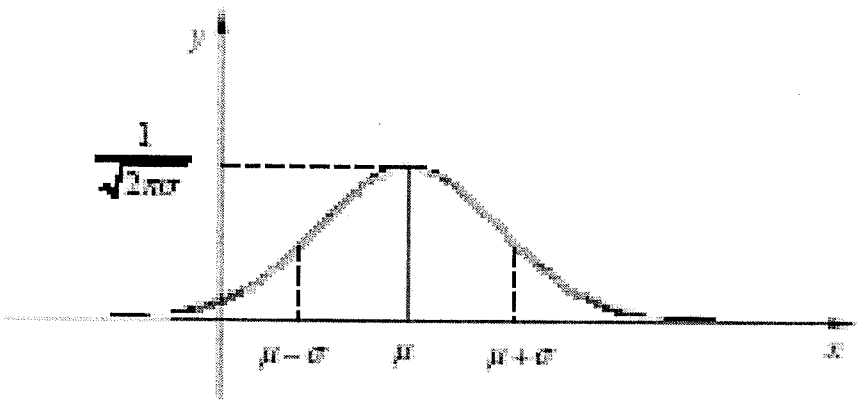
Gravito Fiorelli

APPUNTI ED ESERCIZI DI PROBABILITÀ E STATISTICA

Fulvia Gloria



Anno Accademico 2005-2006

APPUNTI

02/03/10

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:

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

Nella maggior parte delle applicazioni della teoria degli insiemi, tutti gli insiemi che vengono considerati sono sottoinsiemi di un certo insieme universo. Al contrario, l'insieme vuoto, indicato con Ø, è un insieme privo di elementi.

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:

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

$$A \cap B = \varnothing$$

Differenza

Dati due insiemi *A* e *B*, si dice differenza di *A* e *B* l'insieme che contiene gli elementi che stanno in *A* e non in *B*, cioè:

$$A \setminus B = \{x:\ x \in A \ \ e \ \ x \notin B\}$$

Esempio:

$$A = \{1, 3, 5, 6\};\quad B = \{1, 3, 4\}$$

La differenza è l' insieme:

$$A \setminus B = \{5, 6\}$$

Notiamo che (al contrario di unione ed intersezione) la differenza non é un'operazione commutativa, infatti dall'esempio precedente:

$$B \setminus A = \{4\}$$

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:

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

$(3,1), (3,2), (3,3), (1,1), (1,2), (1,3)\}$

$A \times B$ è commutativo, cioè

$(2,1), (3,5), (3,6), (3,3), (3,1)\}$

$(3,2), (3,3), (1,1), (1,2), (1,3)\}$

piccoli grafici schematici noti come

Venn.

04/03/10

matematica ancillare al Calcolo delle

modi con i quali possono essere

menti (sottoinsiemi) di uno o più

sieme o in un sottoinsieme è detto

attro elementi A, B, C, D

mi di ordine 2 (ossia

uattro elementi

rtenza si

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 $k<n$ 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

Si ha:

- $D_{n,1} = n$ 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:

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

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

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

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*, quindi il calcolo del numero delle permutazioni è uguale al calcolo del numero delle disposizioni semplici di *n* elementi di classe *n*, in pratica:

$$P_n = D_{n,n} = n(n-1)(n-2)...[n-(n-2)][n-(n-1)] = n(n-1)(n-2)...2 \cdot 1$$

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

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

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.

La condizione k≤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.

Combinazioni

Partiamo da un esempio : "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: $C_{n,k} = \frac{D_{n,k}}{P_k} = \frac{n(n-1)(n-2)...[n-(k-1)]}{k!}$

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.

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*, quindi il calcolo del numero delle permutazioni è uguale al calcolo del numero delle disposizioni semplici di *n* elementi di classe *n*, in pratica:

$$P_n = D_{n,n} = n(n-1)(n-2)...[n-(n-2)][n-(n-1)] = n(n-1)(n-2)...2 \cdot 1$$

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

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

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.

La condizione k≤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.

Combinazioni

Partiamo da un esempio : "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: $C_{n,k} = \frac{D_{n,k}}{P_k} = \frac{n(n-1)(n-2)...[n-(k-1)]}{k!}$

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.

X 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.

CALCOLO DELLE PROBABILITÀ 11/03/10

X ELEMENTI DI BASE

- Lo spazio campionario è l'insieme S di tutti i possibili esiti di un esperimento.
- Un particolare esito (cioè un elemento di S) è chiamato punto campionario o campione.
- Un evento A è un insieme di esiti (A è un sottoinsieme di S).
- Se A è formato da un solo elemento si parlerà di evento elementare.
- L'insieme vuoto ∅ ed S sono anch'essi eventi; ∅ è chiamato "evento impossibile" ed S "evento certo".

Dal momento che gli eventi sono insiemi, ogni affermazione concernente gli eventi può essere traslata nel linguaggio della teoria degli insiemi e viceversa.

- Se gli esperimenti sono semplici, lo spazio campionario è esattamente l'insieme di tutti i possibili esiti. 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).

- Se gli esperimenti sono composti, lo spazio campionario è un insieme che comprende tutti i possibili esiti ed altri elementi che dipendono dall'esperimento.

- Con il CALCOLO DELLE PROBABILITÀ si studiano eventi casuali, probabili, cioè eventi che possono o non possono verificarsi e che dipendono unicamente dal caso. Tale studio permette di assegnare agli eventi casuali o aleatori un valore numerico al fine di poter confrontare oggettivamente tali eventi e decidere quali tra essi ha maggiore probabilità di verificarsi. Ci sono due importanti procedure mediante le quali è possibile stimare la probabilità di un evento.

- **Approccio classico o a priori**

In n prove, la probabilità di un evento è il rapporto tra il numero dei casi favorevoli ed il numero dei casi possibili, tutti ugualmente probabili:

$$P(E) = n_f / n_p$$

Questa definizione classica è essenzialmente circolare poiché l'idea di "ugualmente probabili" è la stessa cosa di "con uguale probabilità"non ancora definita.

-Nel caso finito ed infinito numerabile, se S= $\{a_i\}$ e indichiamo con p_i le probabilità associate ad a_i si ha:

- $p_i \geq 0$

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

-Nel caso infinito non numerabile la sommatoria diventa un integrale.

- **Approccio frequentistico o a posteriori**

In n prove, la probabilità di un evento è il limite della frequenza relativa dei successi,

$$P(E) = \lim_{n \to \infty} n_f / n_p$$

L'approccio frequentistico si basa sulla sperimentazione e poiché n deve essere molto grande poggia sulla Legge dei grandi numeri

Legge dei grandi numeri

Le frequenze relative del successo in n prove indipendenti, convergono stocasticamente verso la probabilità p del successo.

Cioè se, per esempio, lanciamo spesso un dado e se l'uscita del 6 è il successo, le frequenze relative del successo (ovvero: quante volte è uscito il sei, diviso il numero dei lanci effettuati) vanno stabilizzandosi e rappresentano bene la probabilità del successo.

Quindi in una successione di prove fatte nelle stesse condizioni, la frequenza di un evento si avvicina alla probabilità dell'evento stesso e tale approssimazione tende a migliorare con l'aumentare delle prove.

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

Sia l'approccio classico sia quello frequentistico vanno incontro a serie difficoltà: il primo a ragione dell'espressione "equiprobabili" ed il secondo per aver presupposto "n molto grande" per cui i matematici preferiscono l'approccio assiomatico alla probabilità che fa uso degli insiemi.

- **Approccio assiomatico**

ASSIOMI DELLA PROBABILITÀ

Sia S uno spazio campionario, E la classe degli eventi e P una funzione a valori reali definita su E. P è detta funzione di probabilità e $P(A)$ è detta probabilità dell'evento A se sono soddisfatti i seguenti assiomi:

1. Per ogni evento A $\quad 0 \le P(A) \le 1$

2. $P(S) = 1$

3. Se A e B sono eventi mutuamente escludentesi, allora:
$$P(A \cup B) = P(A) + P(B)$$

4. Se $A_1, A_2 ... A_n ....$è una sequenza di eventi mutuamente escludentesi, allora:
$$P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$$

ALCUNI TEOREMI SULLA PROBABILITÀ

1. Se $A_1 \subset A_2$ allora $P(A_1) \le P(A_2)$ $\quad e \quad P(A_2 - A_1) = P(A_1) - P(A_2)$
2. Se $\varnothing$ è l'insieme vuoto, allora $P(\varnothing) = 0$
3. Se A^c è l'evento complementare ad A allora $P(A^c) = 1 - P(A)$

Gli eventi con intersezione vuota sono detti disgiunti o incompatibili o escludentesi

Se due eventi E_1 e E_2 sono incompatibili, la probabilità dell'evento unione (detto evento totale) è uguale alla somma delle probabilità dei singoli eventi:
$$(E_1 \cap E_2) = \varnothing \Leftrightarrow P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

ne segue il "PRINCIPIO DELLA PROBABILITÀ TOTALE":

Se un evento E può manifestarsi secondo due o più modalità $E_1, E_2 ... E_n$ tra loro mutuamente escludentesi, ne consegue che $P(E) = P(E_1) + P(E_2) + ... + P(E_n)$ ossia $P(E)$ è uguale alla somma delle probabilità relative alle singole modalità E_1, E_2,.... E_n secondo le quali l'evento E può manifestarsi:

$$P(E) = P(E_1) + P(E_2) + ... + P(E_n) = \sum_n P(E_n)$$

INDIPENDENZA DI DUE EVENTI

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

$$A \text{ e } B \text{ sono indipendenti} \Leftrightarrow P(A \cap B) = P(A) \times P(B)$$

ne segue il "PRINCIPIO DELLA PROBABILITÀ COMPOSTA":

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

$$P(E) = P(E_1) \times P(E_2) \times ... \times P(E_n)$$

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.

Due eventi A e B disgiunti non sono tra loro indipendenti, infatti:

$$P(A \cap B) = 0 \ \textit{mentre} \ P(A) \neq 0 \ \ e \ \ P(B) \neq 0 \ \ \textit{quindi} \ \ P(A) \times P(B) \succ 0$$

Probabilità soggettiva

Quando non è noto o il numero dei casi favorevoli o dei casi possibili o entrambi non si può che assegnare alla probabilità dell'evento un valore secondo una nostra personale stima basata sulle conoscenze che abbiamo circa la natura dei fenomeni. Dobbiamo però rispettare sempre le due regole fondamentali:

- la probabilità di un evento deve essere un valore compreso tra 0 ed 1;
- la somma delle probabilità di tutti i possibili eventi deve essere uguale ad 1.

L'uso della stima personale, se da un lato elimina molte difficoltà di calcolo dall'altro non offre sufficienti garanzie che le nostre stime siano fondate. Molto dipende dalla quantità di informazioni che abbiamo sul fenomeno aleatorio, ma anche dal modo in cui riusciamo a valutarle. Ad esempio, molti ritengono erroneamente che un numero "ritardatario" abbia maggiori probabilità di uscire rispetto ad altri numeri che sono già usciti al gioco del lotto.

VARIABILE CASUALE E DISTRIBUZIONE DI PROBABILITÀ

Un concetto molto importante del calcolo delle probabilità è il concetto di variabile aleatoria o casuale. Alcune di queste variabili costituiscono dei modelli teorici per rappresentare fenomeni aleatori reali. In molti fenomeni aleatori il risultato di un esperimento è una grandezza che assume valori variabili in modo casuale. Alcuni esempi di variabili aleatorie : il numero di autovetture che giungono in un giorno ad una stazione di distribuzione di carburante in autostrada; nel gioco del Lotto, il primo numero estratto nella ruota di Roma in una determinata settimana; ....

Non sapendo, in una data prova, quale valore assumerà una variabile aleatoria si può associare ad essa la relativa probabilità. Per alcune variabili aleatorie i valori sono in numero finito, come nell'esempio del gioco del Lotto, per altre sono un'infinità numerabile come nell'esempio delle autovetture. L'analisi delle variabili casuali è un'analisi più approfondita dei fenomeni aleatori rispetto al solo calcolo delle probabilità.

Definizione di variabile casuale

Richiamiamo il concetto di funzione:Siano S e T due insiemi arbitrari. Supponiamo che a ciascun $s \in S$ venga associato un unico elemento di $t \in T$; l'insieme f di tutte queste associazioni è chiamato "funzione" (o mappa) da S in T e si scrive f:S→T. Ogni t, immagine di s, verrà indicato con f(s)., e ogni elemento s, preimmagine di t, verrà indicato $f^{-1}(t)$.

Una variabile casuale X su uno spazio campionario S è una funzione a valori reali di S in R tale che la preimmagine di ogni intervallo di R sia un evento di S. Tale funzione di solito si indica con X,Y... ed i valori immagine con $x_i, y_i \ldots.$,

Distribuzione di probabilità di una variabile casuale

Sia X una variabile casuale su uno spazio campionario S finito, cioè X(S)= $\{x_1, x_2, \ldots x_n\}$. Se ad ogni x_i si associa la sua probabilità $f(x_i)$ mediante la funzione f , tale funzione è chiamata funzione o distribuzione di probabilità di X

La distribuzione f soddisfa le seguenti condizioni:

- $f(x_i) \geq 0$
- $\sum_{i=1}^{n} f(x_i) = 1$ (nel caso discreto)
- $f(x_i) \geq 0 \quad e \quad \int_{-\infty}^{+\infty} f(x)\,dx = 1$ (nel caso continuo)

Esempio:

Supponiamo di lanciare una moneta non truccata due volte, lo spazio campionario sarà:

$$S = \{TT, TC, CT, CC\}$$

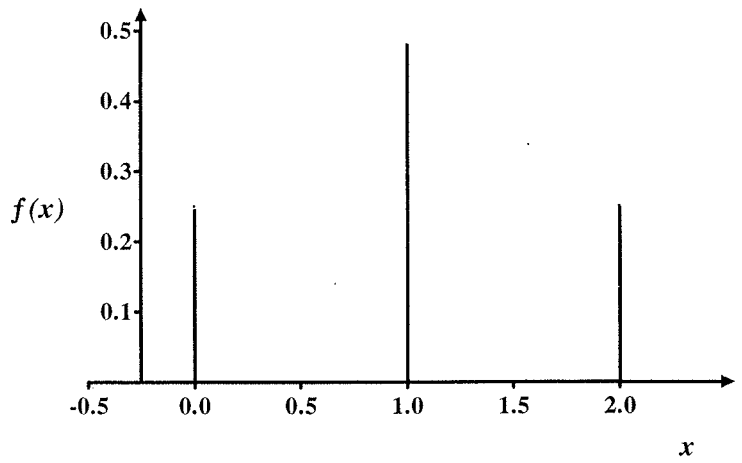
Sia X una variabile casuale che associa ad ogni elemento di S il numero delle volte in cui si verifica testa Ne consegue che ad ogni punto campione sarà associato un numero reale mediante la X:

Punto campione	TT	TC	CT	CC	
X	2	1	1	0	*(variabile casuale)*

Se associamo ad ogni valore numerico x_i la relativa probabilità $f(x_i)$:

x_i	2	1	0
$f(x_i)$	1/4	1/2	1/4

l'insieme delle $f(x_i)$ formerà una distribuzione di probabilità con grafico:



Introdurremo due distribuzioni di probabilità, distribuzione binomiale o variabile casuale binomiale e distribuzione normale o variabile casuale normale, appartenenti ad una famiglia di distribuzioni parametriche che hanno un ruolo di particolare importanza in statistica. In alcuni casi queste distribuzioni sono rilevanti perché si presentano come limite di altre, in altri casi sono importanti perché possono rappresentare un'ampia varietà di fenomeni aleatori.

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.

La funzione Binomiale o Bernoulliana è una funzione discreta che viene usata in particolar modo per descrivere una variabile aleatoria con due valori ai quali sono associate le probabilità p e 1-p.

Definizione

Consideriamo n prove ripetute e indipendenti (n sia finito o infinito numerabile) 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,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 $1-q^n$.

Lo spazio campionario delle n prove ripetute consta di tutte le n-ple ordinate le cui componenti sono o s (successo) o i (insuccesso). Se l'evento A consta di tutte le n-ple ordinate con k successi ed n-k insuccessi.allora il numero delle n-ple ordinate dell'evento A sono $\binom{n}{k}$. Poiché la probabilità di ciascun punto di A è $\mathbf{p^k q^{n-k}}$, si ha:

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

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

$b(k,n,k)$ calcolata per $k=0,1,2,...,n$ è detta *distribuzione* binomiale in quanto 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 $\sigma^2 = npq$

Scarto quadratico medio $\sigma = \sqrt{npq}$

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 rispettive probabilità p e q=1-p;

b) ogni prova effettuata è indipendente da ogni altra prova

c) per ogni prova la probabilità p del successo (e quindi quella dell'insuccesso) è costante.

Per esempio, se giocando a testa o croce ammettiamo che l'evento testa (T) sia equiprobabile all'evento croce (C) e che le loro probabilità valgano entrambe 1/2 (50%), la probabilità di avere 1 testa (successo) in due lanci sarà:

$$P(T) = \binom{2}{1}(1/2)^1(1/2)^1 = 1/2 \cdot 1/2 = 1/4$$

Se ,invece, i due eventi non sono equiprobabili:

-se i lanci sono due, e ci chiediamo la probabilità di 1 successo

$$b(2, p, 1) = \binom{2}{1}(p)^1(q)^1$$

- in generale, se i lanci sono un numero qualsiasi n e se i successi sono un qualsiasi numero k, se indichiamo con p la probabilità del successo e con q= 1-p la probabilità dell'insuccesso , 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? 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 $P(TTCCC) = 1/2 \cdot 1/2 \cdot 1/2 \cdot 1/2 \cdot 1/2$ che si ha applicando 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 testa si possono presentare in diversi modi (nel primo e nel secondo lancio, nel primo e nel quarto lancio, nel secondo e terzo lancio.....), dobbiamo moltiplicare la probabilità dell'evento elementare per un coefficiente che esprime il numero delle combinazioni di n oggetti a "k" a "k" .Nel caso dei nostro esempio tale coefficiente 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 di probabilità continua con due parametri, μ e σ^2 ,indicata con $N(\mu,\sigma^2)$. E' una distribuzione statistica teorica che ben riproduce molte delle caratteristiche di una popolazione (es. peso, altezza, etc.). 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).

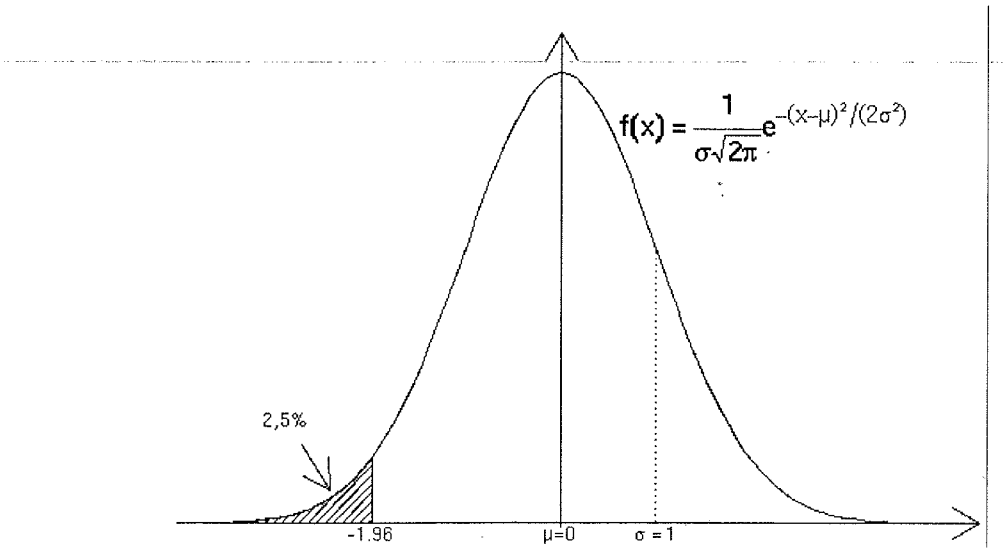
La curva normale ha la seguente funzione di probabilità:

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

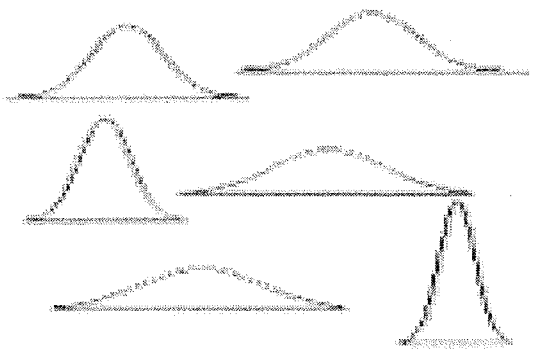
con $-\infty < x < \infty$

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

ed il suo grafico è:

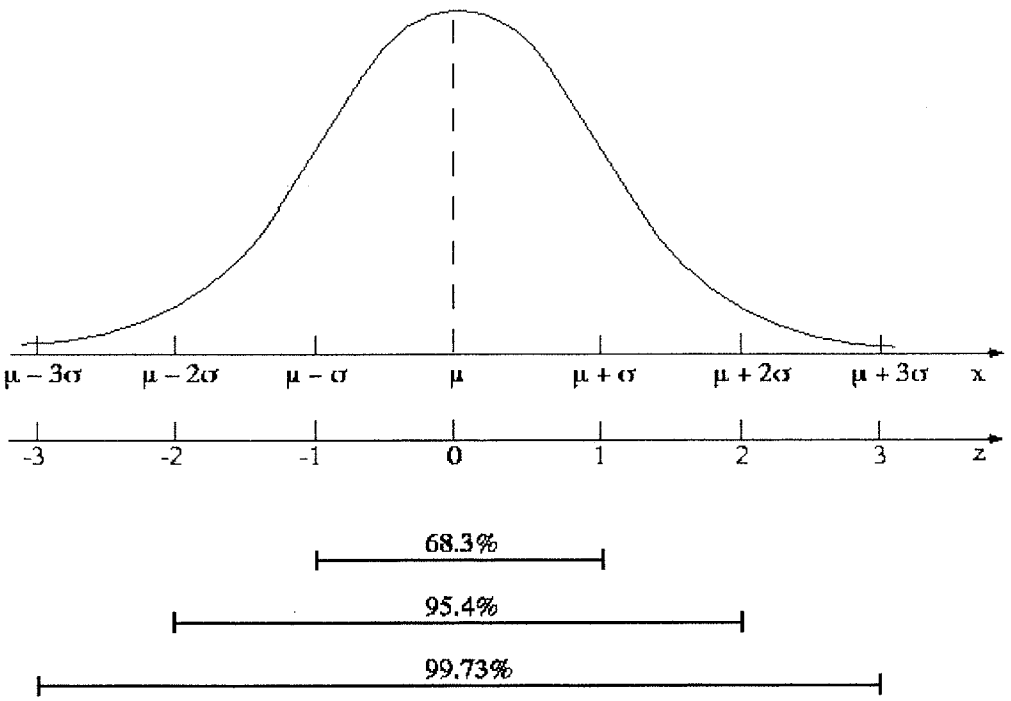


Alcuni esempi di gaussiane:



L'area sottesa alla curva rappresenta la probabilità di un evento. L'area totale è uguale a 1.

PROPRIETA' DELLA DISTRIBUZIONE NORMALE



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

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

$$95{,}5\% = P\{\mu - \mathbf{2}\,\sigma \prec X \prec \mu + \mathbf{2}\sigma\}$$

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

$$99{,}7\% = P\{\mu - \mathbf{3}\,\sigma \prec X \prec \mu + \mathbf{3}\sigma\}$$

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

Per semplicità esprimeremo:

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

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

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

Essendo f(x) una funzione simmetrica, è sufficiente conoscere la funzione di ripartizione dei valori positivi per conoscere pure quella dei valori negativi.

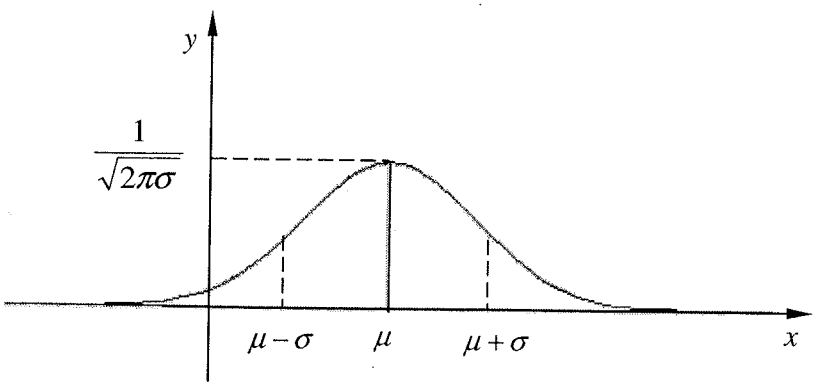
Per fortuna esiste la possibilità di operare in modo estremamente più semplice. A tale scopo occorre definire una particolare distribuzione normale detta distribuzione normale standardizzata mediante la relazione $z=(x-\mu)/\sigma$

L'equazione della funzione di probabilità della distribuzione normale standardizzata è:

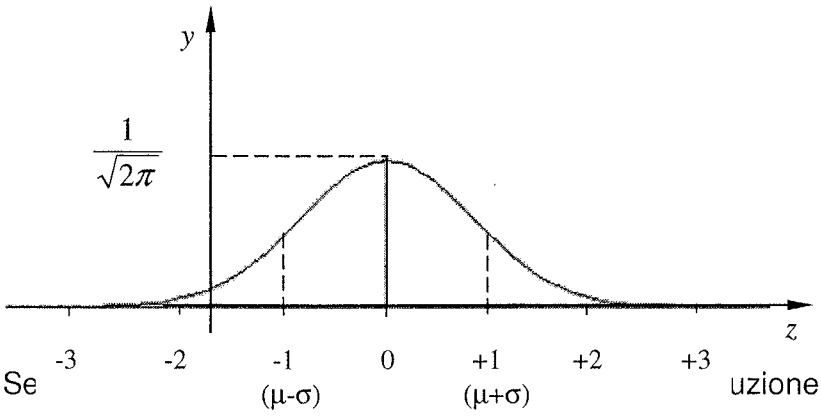
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Da una distribuzione normale $N(\mu,\sigma^2)$ si può sempre passare ad una distribuzione N(0,1) semplicemente trasformando la x in z.

$N(\mu,\sigma^2)$:



N(0,1):



Se uzione normale,

Se vogliamo calcolare la probabilità che la variabile casuale X cada tra a e b, prima cambiamo a e b in unità standard a' e b' e poi calcoliamo la probabilità che la variabile casuale standardizzata X* corrispondente a X cada tra a' e b'.

Per il calcolo delle probabilità relative alle varie aree vengono usati valori tabulati.

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.
- L'unità statistica: è il più piccolo componente di un insieme di "soggetti" che si vuole esaminare.

- La variabile è una caratteristica di un insieme di "soggetti", che, in generale, assume valori differenti tra i vari "soggetti". Una variabile può essere: numerica (quantitativa) o nominale (qualitativa).
- Il campione: é un gruppo di soggetti estratto da una popolazione. Esso deve essere rappresentativo della popolazione.
- La 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 parametro e si parlerà allora di statistica descrittiva univariata
- a due parametri e si parlerà allora di statistica descrittiva bivariata
- a tre o più parametri e si parlerà allora di statistica descrittiva multivariata

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

- qualitativa, se i valori del parametro o dei parametri in studio vengono effettuati su scala di modalità di tipo nominale o ordinale;
- quantitativa, se i valori del parametro o dei parametri in studio vengono effettuati su scala di modalità di tipo intervallare o di rapporto.

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. *Tavole 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 ci 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 la massima frequenza.

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 sono in numero dispari, la mediana sarà esattamente il valore centrale della distribuzione dei valori
- Se i valori sono in numero pari, si prendono i due valori centrali della distribuzione 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 O SCARTO QUADRATICO MEDIO

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. La radice quadrata del valore medio della somma dei quadrati degli scarti dalla media aritmetica si chiama "deviazione standard" o "scarto quadratico medio" (la deviazione standard si chiama anche "scarto quadratico medio" perché è la media quadratica, semplice o ponderata, degli scarti dei valori dalla media aritmetica).

- Se si sta esaminando una popolazione la deviazione standard si indica con σ

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

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

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

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 molto sensibile per misurare la distanza di dati che si scostano molto dal valore medio.

VARIANZA

Il quadrato dello scarto quadratico medio, σ^2 o s^2, è detto varianza.

La varianza è eguale alla media della somma degli scarti dalla media aritmetica al quadrato.

Se i dati sono raggruppati in classi, le x_i sono i valori centrali di ogni classe anche se questo comporta un errore di approssimazione dovuto proprio al raggruppamento.

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 $D_1,D_2,...D_9$ ed i centili $C_1,C_2...C_{99}$ si ottengono dividendo la seriazione rispettivamente in 10 ed 1n 100 parti.

Si hanno i quartili se p=1/4;2/4;3/4

i decili se p=1/10;2/10;...9/10

i centili se p=1/100;2/100;3/100;...99/100

In generale:

- se *p* è un numero tra 0 e 100, il percentile *di ordine p* (o *p °percentile*, se *p* è intero) è il dato che delimita il *p%* dei dati (ordinati) dai rimanenti dati.

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 numerabili 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 finiti o infiniti numerabili 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 DI UNA VARIBILE QUANTITATIVA

Regole generali:

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.

I 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. 82).

- 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.

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

Ricordiamo:
- 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 una statistica (per es. la media μ della pressione sanguigna), estraiamo un campione casuale da questa popolazione e stimiamo tale parametro.
Se estraiamo da questa popolazione per esempio 20 campioni tutti della stessa dimensione che indichiamo n, 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 delle medie campionarie

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

Tale distribuzione ha una propria media ed una deviazione standard.

La media della distribuzione delle medie campionarie approssima la media μ della popolazione: $\mu_{\bar{x}} \approx \mu$ mentre la deviazione standard è direttamente proporzionale alla deviazione standard della popolazione ed inversamente proporzionale alla radice quadrata della numerosità dei campioni:

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} \quad .$$

La deviazione standard della distribuzione delle medie campionarie è chiamato ERRORE STANDARD ($ES = \frac{\sigma}{\sqrt{n}}$ con n = dimensione della popolazione).

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 (per esempio media di un campione) dal valore vero (per esempio 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 μ cade (per es. al 95%) nl"intervallo:

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

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.

I valori $(\bar{x} - 2 \cdot ES)$ e $(\bar{x} + 2 \cdot ES)$ sono gli estremi di un intervallo di confidenza al 95% intorno a μ. ed $(\bar{x} \pm 2 \cdot ES)$ sono detti anche "limiti fiduciali per 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.

Ciascun campione estratto da ogni popolazione avrà una distribuzione con una media ed una deviazione standard.

Per ogni popolazione, la distribuzione della differenza tra le medie campionarie avrà una media ed una deviazione standard.

Risulta:

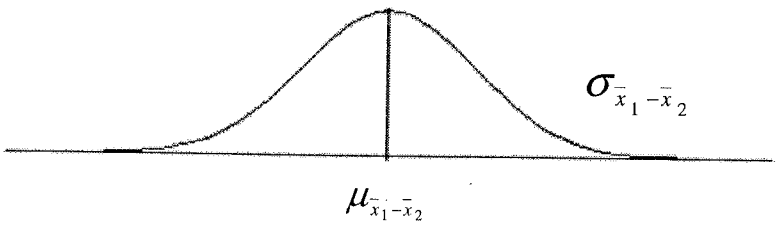
$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} \neq \sigma_{\bar{x}_1} - \sigma_{\bar{x}_2}$$

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

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_1^2/\nu_1 + \sigma_2^2/\nu_2} = ES_1 - ES_2 = ES_{\bar{x}_1 - \bar{x}_2}$$

Se le popolazioni di partenza sono distribuite normalmente anche la distribuzione della differenza tre le medie campionarie sarà una distribuzione normale con media $\mu_{\bar{x}_1 - \bar{x}_2}$ e deviazione standard $\sigma_{\bar{x}_1 - \bar{x}_2}$



Ricordiamo che se il campione è grande l's del campione stima bene ili σ della popolazione

Siano $\bar{x}_1$ 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 P_1 e P_2 con medie μ_1 e μ_2 , e deviazioni standard σ_1 e σ_2 ,possiamo dire che $\mu_1 - \mu_2$ è compreso nell'intervallo di confidenza:

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

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

In generale diciamo che:

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

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

$$(\bar{x}_1 - \bar{x}_2) \pm z_c \cdot ES_{\bar{x}_1 - \bar{x}_2}$$

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.

Il test di ipotesi si può applicare in campo:

- deterministico
- aleatorio

- Nel primo caso si è in grado di pervenire a conclusioni certe.

- Nel secondo caso il risultato può essere modificato dal caso. Si supponga di avere una moneta con due facce contrassegnate con testa e croce. Se la moneta è buona, dal lancio di n volte di una moneta ci si aspetta n/2 volte testa ed n/2 volte croce (è la nostra ipotesi). Tuttavia, anche nel caso in cui la moneta sia buona, lanciandola 20 volte non si può escludere di osservare 20 teste. Il numero di teste può variare da 0 a 20. In tale contesto una qualsiasi decisione in merito all'ipotesi da verificare comporta un rischio di errore. Ad esempio rigettare l'ipotesi avendo osservato 20 teste comporta il rischio di prendere una decisione errata in quanto è possibile ottenere 20 teste da una moneta buona. Per verificare tale ipotesi (moneta buona) si cerca di individuare una conseguenza della bontà della moneta. L'esperimento della moneta porta alla costruzione di una variabile casuale binomiale b (20,1/2,k).

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

Possiamo dire che un'ipotesi statistica è un'affermazione sulla distribuzione di probabilità di una o più variabili casuali. Nel test statistico viene verificata in termini probabilistici la validità di un'ipotesi statistica, detta appunto ipotesi nulla, indicata di solito con H_0.

STEP PER LA VERIFICA DI UN'IPOTESI STATISTICA

1. Esame della natura dei dati
2. Assunzioni sulle statistiche (media, varianza, deviazione standard, etc.) delle popolazioni e del/dei campioni
3. Ipotesi nulla ed ipotesi alternativa
4. Formulazione della statistica test dai dati del campione (una formula generale per la statistica test:

statistica test = (statistica di interesse – parametro ipotizzato)/errore standard della statistica di interesse

5. Decisione statistica

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 falsa.

		Ipotesi nulla	
		Vera	Falsa
Possibile Scelta	*Non rifiutare* H_o	Non rifiuto H_o Decisione corretta	Non rifiuto H_o (β) Errore di 2° specie
	Rifiutare H_o	Rifiuto H_o (α) Errore di 1° specie	Rifiuto H_o (1-β) Decisione corretta

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. Di solito si costruisce la regione di rifiuto in modo che il livello di significatività sia un valore prefissato e piccolo.

POTENZA DI UN TEST

Se H_1 è vera, la probabilità (1-β) di rifiutare H_0 (e prendere quindi una decisione corretta), è detta potenza del test.

GRADI DI LIBERTÀ

I gradi di libertà di una statistica, in genere, esprimono il numero di dati effettivamente disponibili per valutare la quantità d'informazione contenuta nella statistica. Infatti, quando un dato non è indipendente, l'informazione che esso fornisce è già contenuta implicitamente negli altri. È possibile quindi calcolare le statistiche utilizzando soltanto il numero di osservazioni indipendenti, consentendo in questo modo di ottenere una maggiore precisione nei risultati.

TEST CHI-QUADRATO

Il test 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

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

- 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" ed 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

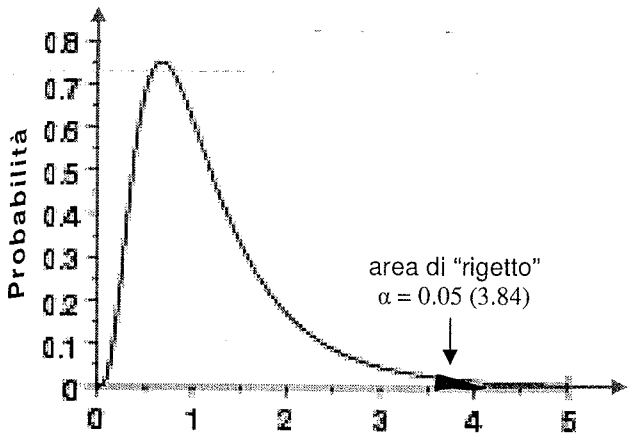


TABELLA DI CONTINGENZA

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 e le righe come un'altra classe di eventi, si vuole determinare se l'evento che compone le righe è indipendente dall'evento che compone le colonne.

Applicazione in ambito medico

Esempio 1

Vogliamo verificare l'effetto di un farmaco su 200 individui divisi in malati di una certa malattia e sani

Prendiamo 80 soggetti malati e ne trattiamo un gruppo con il farmaco (gruppo A) ed un altro gruppo con il placebo (gruppo B), poi prendiamo 120 soggetti sani e trattiamo un gruppo con il farmaco (gruppo A) ed un gruppo con il placebo (gruppo B). Dopo un congruo periodo di trattamento vogliamo 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

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)

- Costruiamo 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).

- Una volta che per ogni coppia (AS,AM,BS,BM) sono state calcolate le frequenze attese, ad esse applicheremo 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:

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

- .La statistica ci aiuterà a 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 consigliandoci di rigettare l'ipotesi di partenza.

Il modello logico che si segue 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 con 1 grado di libertà, 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 Ho è vera ci troviamo dì 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, χ^2= 8.33 e quindi rigettiamo l'ipotesi che il miglioramento è indipendente dal trattamento sbagliando meno di 1 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 con 1 grado di libertà. Il risultato indica una dipendenza della malattia dalla esposizione al fattore di rischio.

VERIFICA DI IPOTESI SULLE MEDIE

VERIFICA DELL'IPOTESI SU UNA MEDIA

A) Sia dato un campione grande (n≥30) con media $\bar{x}$, varianza s^2, dimensione n ed estratto da una popolazione P con media μ e distribuita normalmente. Sia fatta l'ipotesi zero:

$$H_0: \ \mu = \mu_0 \qquad \text{con } \mu \text{ media della popolazione}$$

Distinguiamo:

- Campionamento da una popolazione con varianza nota, la statistica test è:

$z = \dfrac{(\bar{x} - \mu_0)}{\sigma/\sqrt{n}}$ e si usa la tabella normale per verificare l'ipotesi,

Se l'ipotesi viene verificata usando i limiti fiduciali

$$-1.96 \leq \frac{(\bar{x} - \mu_0)}{\sigma/\sqrt{n}} \leq 1.96 \quad \text{(al 95\%)}.$$

- Campionamento da una popolazione con varianza non nota, la statistica test è:

$$z = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$$

Se l'ipotesi viene verificata usando i limiti fiduciali

$$-z_c \leq \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}} \leq z_c$$

B) Sia dato un campione piccolo (n<30) con media $\bar{x}$, varianza s^2, dimensione n, estratto da una popolazione P con media μ e distribuita normalmente:

$H_0: \ \mu = \mu_0$ con μ media della popolazione, la statistica test è:

$$t = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n-1}}$$ gradi di libertà = n – 1 e si usa la tabella del t per verificare l'ipotesi

Se l'ipotesi viene verificata usando i limiti fiduciali

$$-t_c \leq \frac{(\bar{x} - \mu_0)}{s/\sqrt{n-1}} \leq t_c$$

VERIFICA DELL' IPOTESI SULLA DIFFERENZA DI DUE MEDIE

A.) 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

Distinguiamo

- Campionamento da popolazioni con varianze note:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$ (poiché i campioni sono grandi s è un buon stimatore di σ)

- Campionamento da popolazioni con varianze non note:

a) se si suppone che le varianze delle popolazioni siano uguali la statistica test è:

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

con $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ stima congiunta

b) se si suppone che le varianze delle popolazioni siano diverse la statistica test è:

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

In entrambi i casi si usa la tabella dello z per verificare l'ipotesi

B) 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, distribuite normalmente e con le varianze non note:

a) se si suppone che le varianze delle due popolazioni siano uguali ,la statistica test è:

$t = \frac{[(\bar{x}_1 - \bar{x}_2) - 0]}{ES_{\bar{x}_1 - \bar{x}_2}}$ dove $ES_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2/n_1 + s_p^2/n_2}$ ed

$s_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ con $n_1 + n_2 - 2$ = gradi di libertà

b) se si suppone che le varianze delle due popolazioni siano diverse, la statistica test è :

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

In entrambi i casi si usa la tabella del t per verificare l'ipotesi.

TEST t di STUDENT

W.S. Gossett il cui pseudonimo era Student mostrò che se la popolazione è distribuita normalmente, la distribuzione delle medie dei campioni di dimensioni p sono distribuite secondo la distribuzione di Student di ordine p-1 che è il numero dei gradi di libertà.

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.

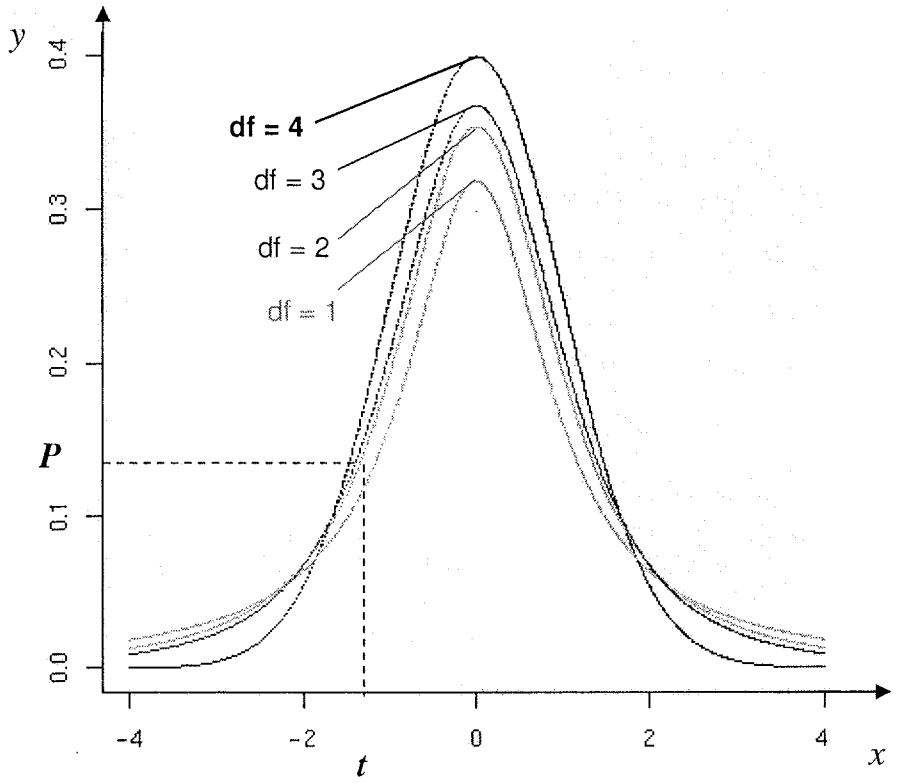
Viene usato in statistica con campioni di piccole dimensioni (minore o uguale a 30 elementi). Se il campione è più numeroso la distribuzione normale e quella di Student differiscono di poco, pertanto è indifferente usare una o l'altra.

Si parla di t test per dati appaiati quando un campione rispetto ad una caratteristica viene esaminato dopo un certo periodo di tempo.

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. La media è sempre uguale a 0 e la varianza è generalmente poco maggiore di 1; è completamente determinata da un solo parametro detto gradi di libertà. Essa tende alla curva normale aumentando i gradi di libertà.

Distribuzioni t



TEST PER UN SOLO CAMPIONE

-Dato un campione di n ($n \leq 30$) valori indipendenti di una variabile,avente media $\bar{x}$ e deviazione standard s estratto da una popolazione distribuita normalmente con varianza non nota, si fa l'ipotesi che la media della popolazione sia uguale ad un certo valore. Si testa questa ipotesi rispetto alla media campionaria:

H_0 : $\mu = \mu_0$ e la statistica test:

$$t = \frac{(\bar{x} - \mu)}{s/\sqrt{n-1}}$$

gradi di libertà =n - 1

TEST PER DUE CAMPIONI INDIPENDENTI

Tale test si usa per confrontare due gruppi diversi rispetto ad una singola variabile.

-Dati due campioni di grandezza n_1 ed n_2 (entrambi i valori ≤30) con medie $\bar{x}_1$ ed $\bar{x}_2$ e con deviazioni standard s_1 ed s_2 estratti da due popolazioni distribuite normalmente con medie μ_1 e μ_2 e con varianze non note (ma può essere giustificata l'uguaglianza delle varianze) si fa l'ipotesi che entrambi i campioni siano stati estratti dalla stessa popolazione.

$$H_0 : \mu_1 = \mu_2$$

La statistica test si basa sulla differenza delle medie campionarie $(\bar{x}_1 - \bar{x}_2)$. Per il teorema del limite centrale si può dimostrare che $(\bar{x}_1 - \bar{x}_2)$ si distribuisce asintoticamente secondo una normale standardizzata. Poichè le varianze delle popolazioni non sono note esse devono essere determinate utilizzando le varianze campionarie:

a)se si suppone che le varianze delle popolazioni siano uguali, la statistica test è:

$$t = \frac{[(\bar{x}_1 - \bar{x}_2) - 0]}{ES_{\bar{x}_1 - \bar{x}_2}} \qquad ES_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p{}^2 / n_1 + s_p{}^2 / n_2}$$

e $$s_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$ con $n_1 + n_2 - 2$ = gradi di libertà

b) se si suppone che le varianze non siano uguali la statistica test è

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

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. Nulla impedisce di usare queste tecniche anche in presenza di variabili di tipo ordinale o continuo, ma in tal caso sono meno efficienti delle tecniche alternative (per es. regressione lineare).

Bisogna distinguere se la variabilità è dovuta a:

- 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.

Definizioni:
- Il fenomeno statistico la cui variabilità si vuole spiegare in base ad una o più variabili che formano i gruppi viene definita come variabile risposta che deve essere necessariamente rappresentata da una variabile quantitativa continua.

- Le variabili statistiche che stabiliscono una ripartizione fenomeno statistico in classi o strati vengono chiamate fattori (o trattamenti) che devono essere necessariamente variabili categoriche.

- Le modalità, cioè i valori, che i fattori possono assumere si definiscono livelli del fattore.

Un fattore può essere classificato in due categorie:

- fisso, se il livello del fattore è controllato dallo sperimentatore
- casuale, se il livello del fattore è determinato da un campionamento di una popolazione

Le procedure statistiche per l'analisi della varianza (ANOVA) si suddividono in quelle che sottintendono un modello:

- bilanciato, cioè il numero di osservazione per ogni livello (ANOVA ad una via) o per ogni combinazione di livelli (ANOVA a due o più vie) deve essere il medesimo; oppure
- non bilanciato, in caso contrario rispetto a sopra

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.

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 prospettivo 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

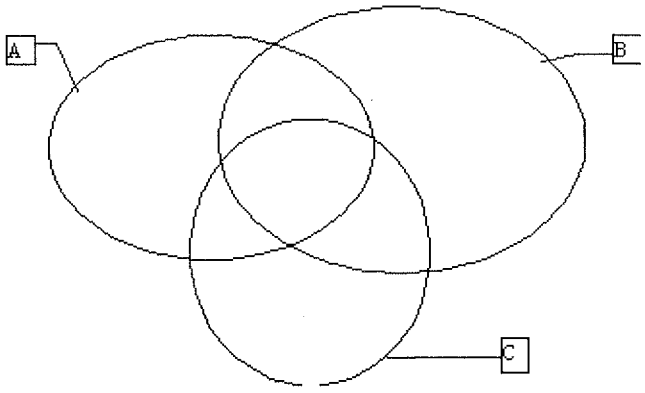
INSIEMI

Appartenenza e non appartenenza

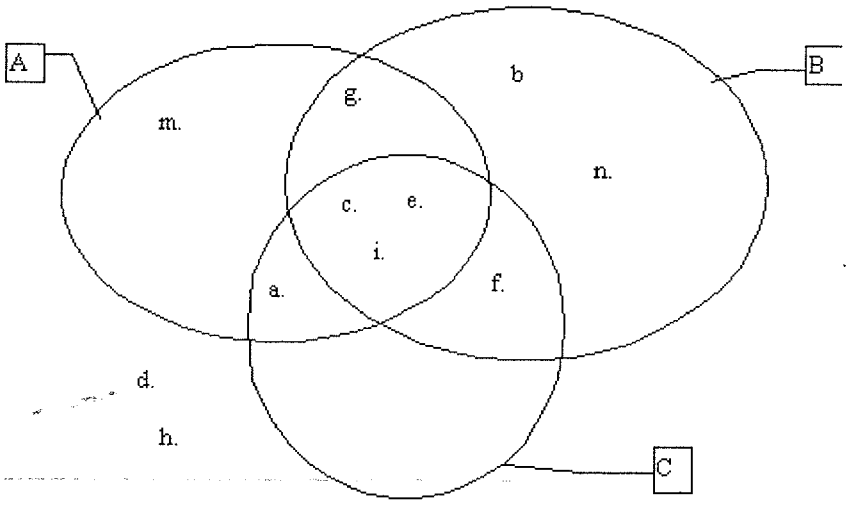
Dati i seguenti elementi:

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

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:

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

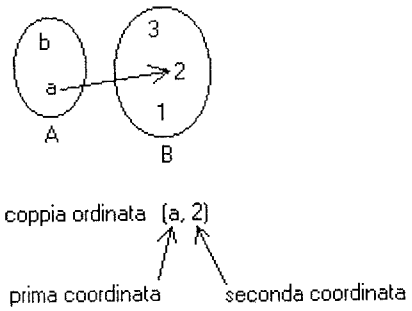
Soluzione:

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

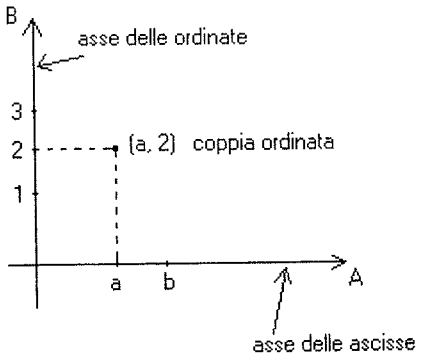
$A \cup B=(1,2,3,4,5,6)$ $A \cup C=(1,2,3,4,7,8)$ $B \cup C=(3,4,5.6,7,8)$ $A \cap B \cap C=(3,4)$

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

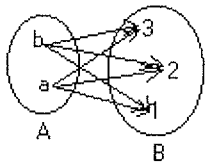
$$A \times B = \{(x,y), x \in A, y \in B\}$$

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)}.

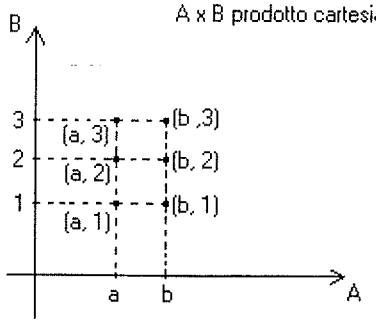
Graficamente, usando i diagrammi di Venn, il prodotto cartesiano A x B sarà visualizzato come:



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.

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 x 2 x 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 :

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

<u>COMBINAZIONI</u>

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?

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

4. Una scatola contiene 5 palline rosse (R), 4 palline bianche (B) e 3 palline nere (N). Se si estraggono a caso tre palline:

- qual è la probabilità che siano tutte e tre rosse?

P(3 R)= n° di scelte di 3 palline rosse su 5 / n° di scelte di 3 palline su 12= $C_{5,3}/ C_{12,3}$=

- qual è la probabilità che 2 palline siano rosse e 1 sia bianca?

P(2R1 B) = (n° di scelte di 2 palline rosse su 5)x(N° di scelte di 1 pallina bianca su 4) / n° di scelte di 3 palline su 12 $= (C_{5,2} \times C_{4,1}) / C_{12,3} =$

- qual è la probabilità di estrarre una pallina di ogni colore?

$(C_{5,1} x C_{4,1} x C_{3,1}) / C_{12,3} =$

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*:

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

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:

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

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.

Applicando la proprietà simmetrica dei coefficienti binomiali si può attribuire, per convenzione, un valore al coefficiente binomiale $\binom{n}{k}$ anche per $k = 0$; si ha infatti:

$$\binom{n}{0} = \binom{n}{n-0} = \binom{n}{n} = 1$$

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

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 dì 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:

$P(A)=6/8$, $P(B)=4/8$, $P(A \cap B)=3/8$ e quindi $P(A \cap B)=P(A) \times P(B)$

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

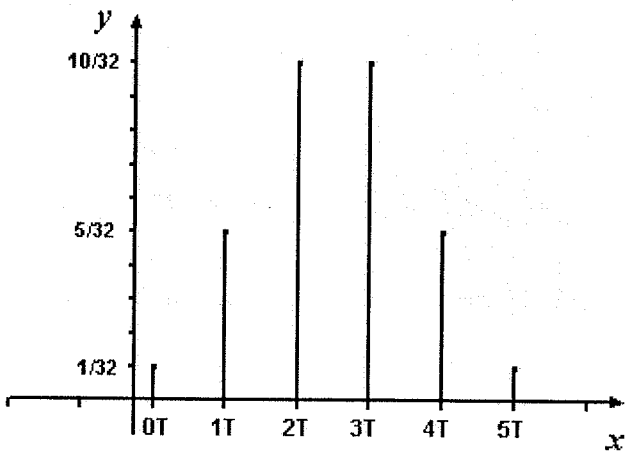
P(2 teste)=$b(5,1/2,2)=C_{5,2}\,(1/2)^2 x.(1/2)^3=10/32$

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

COME SI USA LA TAVOLA DELLA CURVA NORMALE STANDARDIZZATA (CURVA DELLO z)

Nella tavola della curva normale standardizzata, la prima colonna contrassegnata da z riporta i valori di z con una sola cifra decimale, la prima riga riporta la seconda cifra decimate di z. 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 z 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 z=0 e z=1,5

Nella tavola si procede verso il basso nella colonna segnata da z 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 z sia compresa tra 0 e 1.5.

$P (0 \leq z \leq :1.5) = 0.4032$

2. Trovare l'area compresa tra z=-0.65 e z=0

Nella tavola si procede verso il basso nella colonna segnata da z fino a raggiungere il valore z=0.6 poi si procede verso destra fino alla colonna segnata da 5. Il valore 0.2422 è l'area richiesta.

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

3. Trovare l'area compresa tra z=-0.65 e z=1.5

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

4. Trovare l'area per $z \geq 2.54$

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

5. Trovare l'area per $z \leq -1.42$

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

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
$P(178 \leq \bar{x} \leq 180)=P(0.80: \leq z \leq 2.00) = 0.4772-0.2881= 0.1891$
La proporzione di giovani con altezza tra 178cm e 180cm è 18.91%

MISURE DESCRITTIVE

MEDIA- MODA- MEDIANA- RANGE- VARIANZA- DEVIAZIONE STANDARD

1. Dei seguenti numeri: 2, 3, 4, 5, 1, 1 calcolare le misure descrittive:

-media aritmetica : $\bar{x} = \sum_{i=1}^{n} x_i / n$ =16/6=2.66

-moda: m_o =1

-mediana:M=4.5

-range: R=3

-varianza : $s^2 = \frac{(-0.66)^2 + (0.34)^2 + (1.34)^2 + (2.34)^2 + 2(-1.66)^2}{6-1}$

- deviazione standard: $s = \sqrt{13.3336} = 3.651$

2. Della distribuzione A di pag.80 calcolare le misure descrittive:

-media aritmetica: $\mu = \sum_{i=1}^{n} x_i f_i = 4.055$

-moda: m_o=3.6

-mediana: M=4.00

-range : R=3.8;

-varianza : $\sigma^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2 / n = 0.488$

- deviazione standard : $\sigma = 0.698$

RAPPRESENTAZIONE GRAFICA

A) 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:

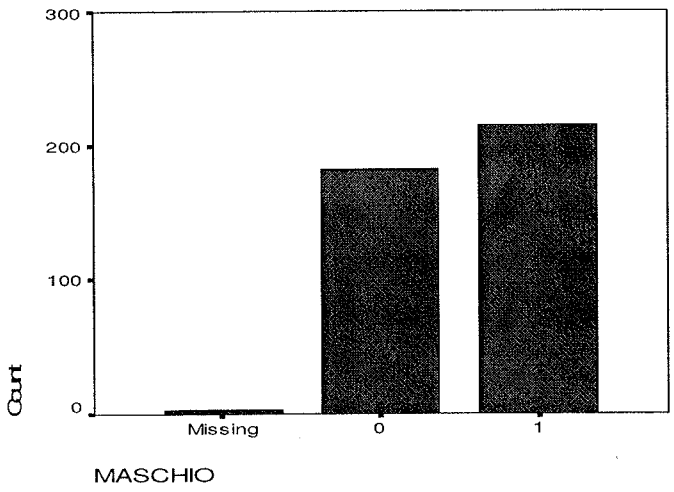
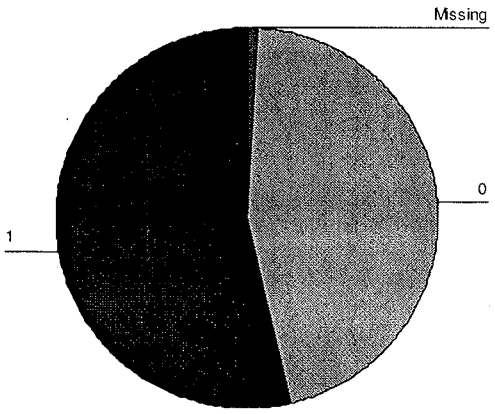


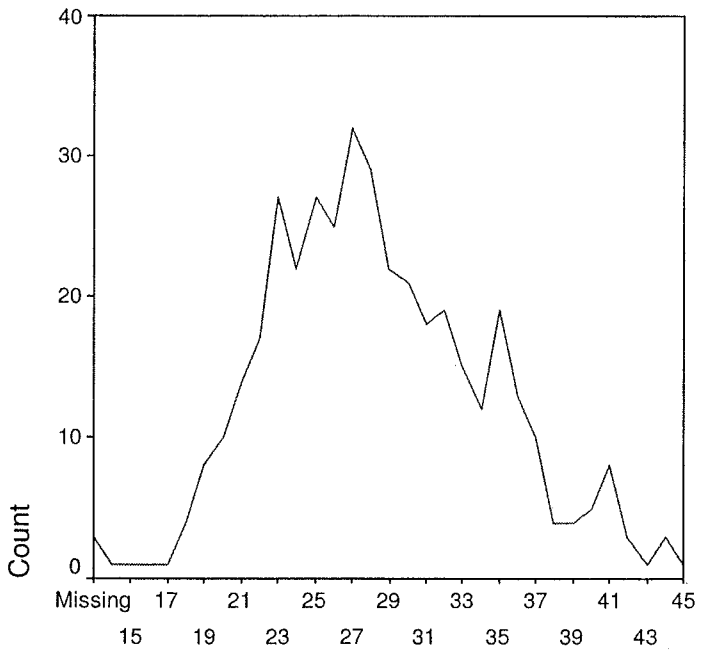
Diagramma circolare:



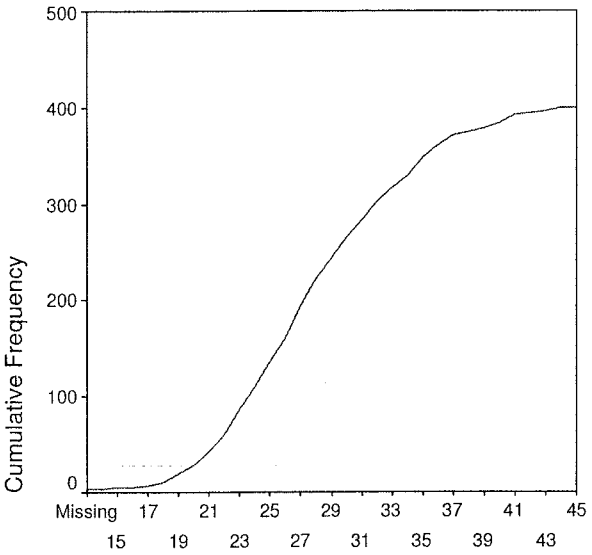
B)Dati numerici

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 poligono di frequenze:



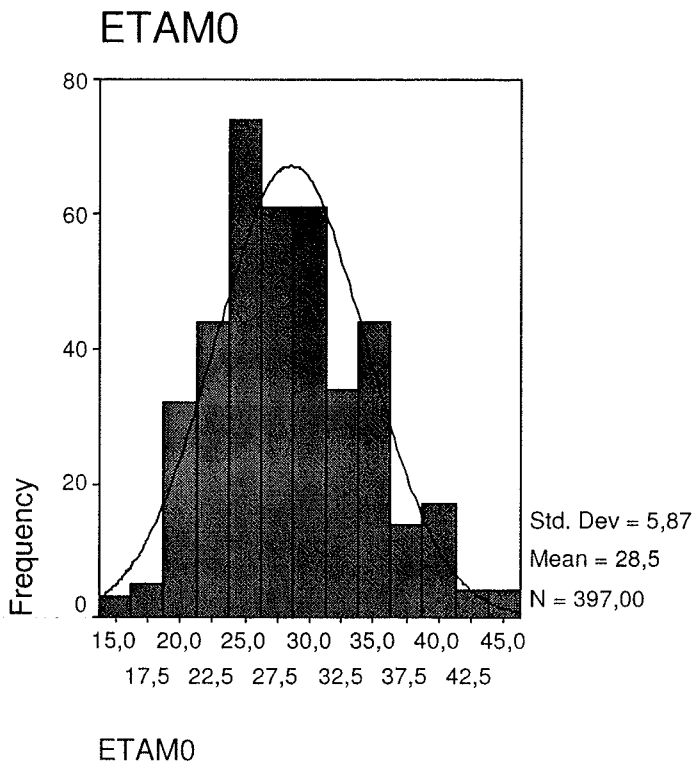
- e l'ogiva



C) 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.

FREQUENZA

A) Sia data la distribuzione dei livelli di glucosio nel sangue (mmol/l)di un gruppo di studenti del 1° anno di medicina:

4.7	3.6	3.8	2.2	4.7	4.1	3.6	4.0	4.4	5.1
4.2	4.1	4.4	5.0	3.7	3.6	2.9	3.7	4.7	3.4
3.9	4.8	3.3	3.3	3.6	4.6	3.4	4.5	3.3	4.0
3.4	4.0	3.8	4.1	3.8	4.4	4.9	4.9	4.3	6.0

La distribuzione di frequenze:

Classi	v. centrale di classe	frequenza assoluta	frequenza relativa %	frequenza assoluta cumulata	frequenza relativa cumulata
	x_i	p_i	f_i	P_i	F_i
[1.75-2.25)*	2.00	1	2.5	1	2.5
[2.25-2.75)	2.50	0	0	1	2.5
[2.75-3.25)	3.00	1	2.5	2	5.0
[3.25-3.75)	3.50	12	30	14	35.0
[3.75-4.25)	4.00	11	27.5	25	2.5
[4.25-4.75)	4.50	9	22.5	34	85.0
[4.75-5.25)	5.00	5	12.5	39	97.5
[5.25-5.75)	5.50	0	0	39	97.5
[5.75-6.25)	6.00	1	2.5	40	100.0

*la classe [1.75-2.25) può essere anche indicata $\geq 1.75 < 2.25$ e così via per le altre classi.

STIMA DI PARAMETRI

DISTRIBUZIONE DELLE MEDIE CAMPIONARIE

Data una popolazione di 35 numeri casuali

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

μ=4.86 σ=2.84

Estraiamo campioni di numerosità n= 3 saltando per esempio due numeri e prendendo il terzo.

	9 7 5	4 1 2	3 0 5	2 9 8	6 3 2
Media ($\bar{x}$)	7	2.33	2.67	6.33	3.67

La media della distribuzione delle medie campionarie sarà: $\mu_{\bar{x}} = 4.40 \approx \mu$

La deviazione standard della distribuzione delle medie campionarie sarà:

$$\sigma_{\bar{x}} = \frac{\sqrt{(7-4.4)^2 + (2.33-4.4)^2 + (2.667-4.4)^2 + (6.33-4.4)^2 + (3.667-4.4)^2}}{5} = 1.91$$

$\sigma_{\bar{x}} \approx \sigma/\sqrt{n} = 2.84/\sqrt{3} = 1.64$ (Errore standard) $\neq \sigma$

La media della distribuzione delle medie è molto vicina alla media della popolazione

La deviazione standard della distribuzione delle medie è molto vicina alla deviazione standard della popolazione fratta la radice quadrata della numerosità del campione e viene chiamata errore standard .

Questi valori si avvicinano sempre di più ai parametri della popolazione aumentando la numerosità dei campione

In molte situazioni pratiche non conosciamo il vero valore della varianza della popolazione σ^2 ma solo la sua stima s^2 (dove con s^2 si intende la varianza del campione).

Se la numerosità n del campione è grande,la distribuzione delle medie campionarie tende ad una distribuzione normale e si può assumere s^2 come una buona stima di σ^2

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 PER UNA MEDIA

In un campione di 100 giovani di 20 anni si è trovato un peso medio dì 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 dì 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 PER LA DIFFERENZA DI 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°	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

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

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

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

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

TEST DI IPOTESI

CHI-QUADRATO--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 dei 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 del 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

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 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 s_1^2 = 17.741

Campione B:

Media $\bar{x}_2$= 55.23 s_2= 4.133 n_2= 16 s_2^2 =17.082

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

$$t = \frac{[(\bar{x}_1 - \bar{x}_2) - 0]}{ES_{\bar{x}_1 - \bar{x}_2}}$$ con $n_1 + n_2 - 2$ = gradi di libertà

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

Sostituendo i valori dell'esercizio, si ha:

$$t = \frac{[(53.50 - 55.23) - 0]}{\sqrt{17.741/15 + 17.082/16}} = -1.153 \text{ con 29 gradi di libertà, } p<0.25.$$

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 poiítomíca
Gruppo Rh	categorica dicotomica (Rh+1Rh-)
Sesso	categorica dicotomica
Ordine di genitura	numerica discreta
Età gestazionale(ín settimane)	numerica discreta
Fototerapia	categorica dicotomica (si/no)
Deficit di G-6-PD	categorica dicotomica

Dati riguardanti la madre

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

SCALE DI MISURA

Misura nominale (categorica): appartengono a questa scala di misura:

- le forme dicotomiche: maschio-femmína, sano-maiato, vivo-morto, celibe-sposato, fumatore-non fumatore.
- le forme politomiche: lo stato di fumatore, per esempio, può essere specificato in categorie esaustive e mutuamente escludentesi (non fumatore, fumatore lieve, grande fumatore). Ogni categoria può essere ulteriormente specificata.

Ad ogni categoria verrà associato un valore numerico semplicemente per poterle confrontare tra foro.

Misura ordinale: in questa scala di misura le forme politomiche vengono ordinate secondo qualche criterio.

Nell'esempio precedente di stato di fumatore, le tre categorie: non fumatore, fumatore lieve e grande fumatore, se stiamo studiando una patologia bronchiale, ci danno una significativa notizia su un fattore di rischio.

Misura ad intervalli: in questa scala di misura dobbiamo fissare a priori un'unità di misura arbitraria ed un punto zero che non significa assenza totale della variabile che si sta esaminando ma serve semplicemente per un confronto.

Esempi sono l'altezza di un individuo in cm, il peso di un individuo in g, la FEV (capacità vitale espiratoria) in litri, etc.

Misura di rapporto: in questa scala di misura lo zero significa mancanza totale della variabile.

Si prendono in considerazione gli stessi esempi della misura ad intervalli, facendo le ulteriori assunzioni che il rapporto tra due valori di una variabile si mantiene costante e che lo zero é una misura ben precisa.

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 GENICHE ED 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.

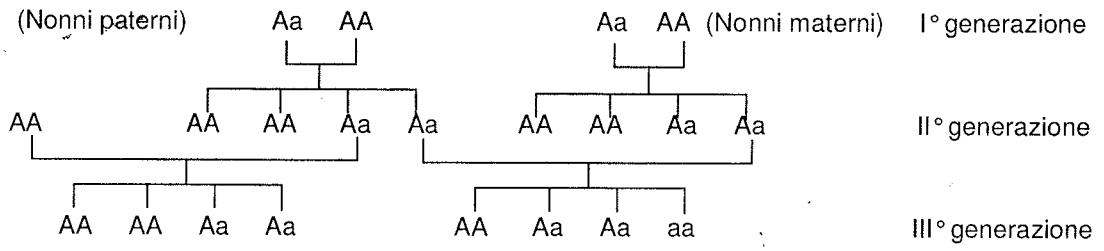
Nel *locus* MN si hanno 2 alleli, L^M e L^N codominanti per cui i genotipi possibili sono: L^M/L^M, L^M/L^N, L^N/L^N ed i fenotipi sono M, MN, N. Supponiamo che le frequenze assolute dei tre fenotipi (genotipi) siano rispettivamente 40, 50 e 10. Si avrà:

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

(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.

La procedura per valutare il rischio di avere un figlio omozigote recessivo 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 il cugino primo germano 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 lo zio di un soggetto malato si accoppia con un soggetto preso a caso nella popolazione?
4. 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?
5. Qual è la probabilità di avere un figlio malato se la sorella di un soggetto malato si accoppia con lo zio (fratello di sua madre) di un malato?
6. Qual è la probabilità di avere un figlio malato se la sorella di un soggetto malato si accoppia con un cugino primo di un malato?
7. 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. 1/4 x2/100 x1/4=1/800
3. 1/2 x 2/100 x 1/4=1/400
4. 2/3 x 2/100 x 1/4=1/300

5. 2/3 x 1/2 x1/4=1/12

6. 2/3x1/4 x1/4=1/24

7. 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 x 0.99 ~ 0.98, f(Aa)= 2pq= 2 x 0.99 x 0.01 ~ 2q ossia ~ 0.02, f(aa)= q^2=0.01 x 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

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

Esempio:

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

M=40, MN=40 e N=20

Soluzione:

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

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
8	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

Area tra 0 e z

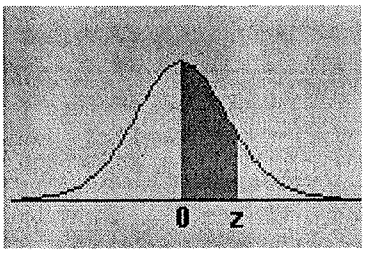


Tabella di z

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

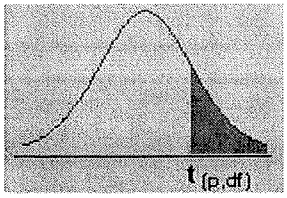


Tabella del t di Student

df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

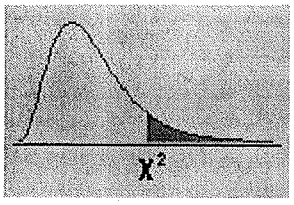


Tabella del Chi Quadrato

df\area	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.00004	0.00016	0.00098	0.0039	0.01579	0.10153	0.45494	1.32330	2.70554	3.84146	5.02389	6.63490	7.879
2	0.01003	0.02010	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.99146	7.37776	9.21034	10.59
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.81473	9.34840	11.3448	12.83
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.48773	11.1432	13.2767	14.86
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.67460	4.35146	6.62568	9.23636	11.0705	12.8325	15.0862	16.74
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.45460	5.34812	7.84080	10.6446	12.5915	14.4493	16.8118	18.54
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.0170	14.0671	16.0127	18.4753	20.27
8	1.34441	1.64650	2.17973	2.73264	3.48954	5.07064	7.34412	10.2188	13.3615	15.5073	17.5345	20.0902	21.95
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.89883	8.34283	11.3887	14.6836	16.9189	19.0227	21.6659	23.58
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182	12.5488	15.9871	18.3070	20.4831	23.2092	25.18
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.3410	13.7006	17.2750	19.6751	21.9200	24.7249	26.75
12	3.07382	3.57057	4.40379	5.22603	6.30380	8.43842	11.3403	14.8454	18.5493	21.0260	23.3366	26.2169	28.29
13	3.56503	4.10692	5.00875	5.89186	7.04150	9.29907	12.3397	15.9839	19.8119	22.3620	24.7356	27.6882	29.81
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.1653	13.3392	17.1169	21.0641	23.6847	26.1189	29.1412	31.31
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.0365	14.3388	18.2450	22.3071	24.9957	27.4883	30.5779	32.80
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.9122	15.3385	19.3688	23.5418	26.2962	28.8453	31.9999	34.26
17	5.69722	6.40776	7.56419	8.67176	10.0851	12.7919	16.3381	20.4886	24.7690	27.5871	30.1910	33.4086	35.71
18	6.26480	7.01491	8.23075	9.39046	10.8649	13.6752	17.3379	21.6048	25.9894	28.8693	31.5263	34.8053	37.15
19	6.84397	7.63273	8.90652	10.1170	11.6509	14.5620	18.3376	22.7178	27.2035	30.1435	32.8523	36.1908	38.58
20	7.43384	8.26040	9.59078	10.8508	12.4426	15.4517	19.3374	23.8276	28.4119	31.4104	34.1696	37.5662	39.99
21	8.03365	8.89720	10.2829	1.59131	13.2396	16.3443	20.3372	24.9347	29.6150	32.6705	35.4788	38.9321	41.40
22	8.64272	9.54249	10.9823	12.3380	14.0414	17.2396	21.3370	26.0392	30.8132	33.9244	36.7807	40.2893	42.79
23	9.26042	10.1957	11.6885	13.0905	14.8479	18.1373	22.3368	27.1413	32.0069	35.1724	38.0756	41.6384	44.18
24	9.88623	10.8563	12.4011	13.8484	15.6586	19.0372	23.3367	28.2411	33.1962	36.4150	39.3640	42.9798	45.55
25	10.5196	11.5239	13.1197	14.6114	16.4734	19.9393	24.3365	29.3388	34.3815	37.6524	40.6464	44.3141	46.92
26	11.1602	12.1981	13.8439	15.3791	17.2918	20.8434	25.3364	30.4345	35.5631	38.8851	41.9231	45.6416	48.28
27	11.8075	12.8785	14.5733	16.1514	18.1139	21.7494	26.3363	31.5284	36.7412	40.1132	43.1945	46.9629	49.64
28	12.4613	13.5647	15.3078	16.9278	18.9392	22.6571	27.3362	32.6204	37.9159	41.3371	44.4607	48.2782	50.99
29	13.1211	14.2564	16.0470	17.7083	19.7677	23.5665	28.3361	33.7109	39.0874	42.5569	45.7222	49.5878	52.33
30	13.7867	14.9534	16.7907	18.4926	20.5992	24.4776	29.3360	34.7997	40.2560	43.7729	46.9792	50.8921	53.67