

Il calcolo statistico e le leggi della probabilità



ERRORI DEI TEST

- Quando si conduce un test, 2 tipi di errore sono possibili rispetto all'ipotesi nulla che abbiamo formulato:

Ipotesi	Ipotesi 0 vera	Ipotesi 0 falsa
Rigetto l'ipotesi 0	Errore di 1° tipo: α	OK
Accetto l'ipotesi 0	OK	Errore di 2° tipo: β

ERRORI DEI TEST

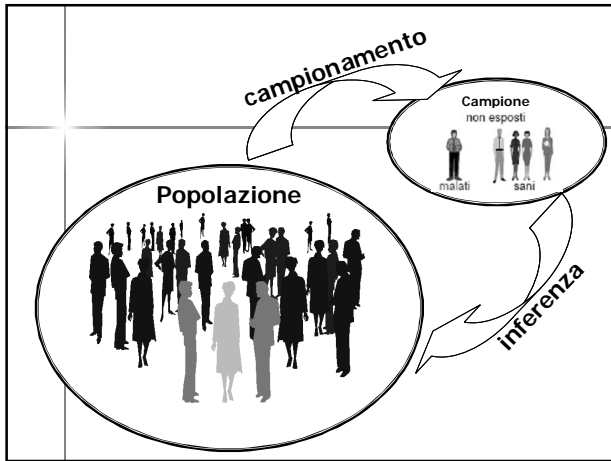
- Errore di primo tipo, con probabilità α di commetterlo:
 - Rigettare l'ipotesi nulla mentre invece essa è vera nella popolazione (falso positivo).
- Errore di secondo tipo, con probabilità β di commetterlo:
 - Accettare l'ipotesi nulla mentre invece essa è falsa nella popolazione (falso negativo).

Un esempio: il PAP test		
<u>ca cervice</u>	presente	assente
Pap Test +		
Pap Test -		

ERRORI DEI TEST
<ul style="list-style-type: none"> ■ In un campione, è possibile stabilire la probabilità α di commettere un errore del 1° tipo, ad una soglia predefinita (di solito 1% o 5%). ■ Allora, nell'eseguire il test, si ha una probabilità dell'1% o del 5% di fare un errore del 1° tipo, cioè rifiutare l'ipotesi nulla (essendo invece vera nella popolazione). ■ Non è possibile in generale fissare a priori la probabilità β. ■ Si può calcolare la 'potenza' del test: $1-\beta$, cioè la probabilità di rifiutare l'ipotesi nulla essendo in effetti falsa nella popolazione.


GENERALIZZAZIONE DEI DATI DI UN CAMPIONE
<ul style="list-style-type: none"> ■ Dopo aver dimostrato con un certo grado di probabilità che un campione è coerente con la popolazione da cui deriva (cioè RAPPRESENTATIVO), i risultati ottenuti dalle misure effettuate sul campione possono essere estesi e riferiti alla popolazione nel suo insieme.

GENERALIZZAZIONE DEI DATI DI UN CAMPIONE
<ul style="list-style-type: none"> ■ Le caratteristiche di interesse della popolazione, ad esempio la media e la deviazione standard per un fenomeno oggetto di studio, costituiscono i 'parametri' della popolazione, i cui valori sarebbero noti solo analizzandone tutte le unità. ■ Gli indici calcolati sul campione rappresentano una 'stima' di tali parametri ignoti.



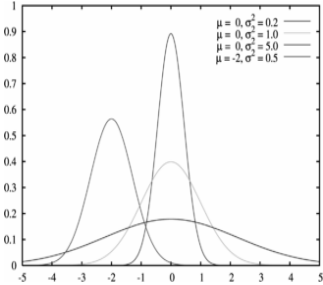
Inferenza statistica

- Tramite l'inferenza statistica si determinano i parametri della popolazione, utilizzando i risultati che abbiamo osservato sul campione.
- Come al solito, non abbiamo la certezza ma un certo grado di **Probabilità**.
- Come sperimentatori possiamo decidere se accettare o meno quel grado di probabilità.
- Si tratta di **'stimare'** dal campione il valore di parametri ignoti per la popolazione.



STIMA

- Significa effettuare un calcolo di valori con una possibilità di errore nota e accettata.
- Si basa sul calcolo delle probabilità e sulla distribuzione gaussiana.



PROPRIETA' DELLA DISTRIBUZIONE GAUSSIANA

Circa il 95% dei valori in una distribuzione gaussiana (normale) cade nell'intervallo definito da due valori:

- Media - 2 volte la deviazione standard
- Media + 2 volte la deviazione standard

DISTRIBUZIONE GAUSSIANA IN UNA POPOLAZIONE

- Consideriamo una variabile che segue la distribuzione gaussiana, ad esempio l'altezza di una popolazione.
- Sorteggiando (quindi a caso) un individuo della popolazione si può ragionevolmente affermare che con circa il 95% di probabilità la sua altezza cade nell'intervallo compreso tra i valori:
 - Media + 2 volte la deviazione standard
 - Media - 2 volte la deviazione standard
- Media e deviazione standard sono i nostri parametri

DISTRIBUZIONE GAUSSIANA IN UNA POPOLAZIONE

- Se i parametri della variabile nella popolazione fossero noti, sarebbe immediato ottenere i valori cercati. Ad esempio, se la media fosse 175 cm e la deviazione standard 10, avremmo:
- Media - 2 volte la deviazione standard: $175 - 2 \cdot 10 = 175 - 20 = 155$
 - Media + 2 volte la deviazione standard: $175 + 2 \cdot 10 = 175 + 20 = 195$
- In questo caso, estraendo a caso un soggetto da quella popolazione avremmo il 95% di probabilità che la sua altezza sia compresa tra 155 e 195 cm.

STIMA DI UNA MEDIA

- Di solito, però, i parametri sono ignoti (a meno di un censimento)...
- In questi casi, dobbiamo estrarre un campione di individui e ne calcoliamo l'altezza media...
- A che livello di probabilità la media del nostro campione rappresenta la media vera e sconosciuta della popolazione?

TEOREMA DEL LIMITE CENTRALE

- Data una variabile, qualsiasi sia la sua distribuzione, la media di tutti i suoi campioni di ampiezza n ha una distribuzione normale.
- In termini rozzi, il teorema del limite centrale afferma che la distribuzione della somma di un numero elevato di variabili casuali indipendenti e identicamente distribuite tende distribuirsi normalmente, indipendentemente dalla distribuzione delle singole variabili.

TEOREMA DEL LIMITE CENTRALE

■ Se una popolazione è distribuita con media uguale a μ e deviazione standard σ , e si estraggono a caso campioni con numerosità n , allora la media campionaria \bar{x} di questi campioni tenderà a distribuirsi approssimativamente secondo una curva normale, la cui media sarà pari a μ , e deviazione standard pari a :

$$\sqrt{V(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

■ Grazie a questo teorema è possibile sempre ottenere un intervallo di confidenza per la stima della Media n di una popolazione.

TEOREMA DEL LIMITE CENTRALE

■ Pertanto, se la dimensione del campione n è "grande," allora la distribuzione delle somme parziali Y_n (o, equivalentemente, della media campionaria M_n) è approssimativamente normale.

■ Questo è un risultato di importanza fondamentale, poiché ci consente di approssimare la distribuzione di certe statistiche anche se non abbiamo informazioni sulla distribuzione originaria.

INTERVALLI DI CONFIDENZA

- Confidenza: livello di fiducia probabilistico che si attribuisce alla stima campionaria.
- Gli **intervalli di confidenza (IC)**: definiscono entro quali valori sia localizzata la media vera (e ignota), in base alle stime campionarie, per N sufficientemente grande.

$$IC_{95\%} = m \pm 1.96s_m$$

$$IC_{99\%} = m \pm 2.58s_m$$

INTERVALLI DI CONFIDENZA

- Per meno di 30 unità, lo scostamento della media campionaria dalla gaussiana è notevole: ad esempio, per comprendere il 95% dell'area intorno alla media con un campione di 10 misure, occorre considerare 2.26 volte l'errore standard anziché 1.96.
- I valori di correzione si trovano su apposite tabelle.

CONSIDERAZIONI SULL'INFERENZA

- È utile utilizzarla solo in casi semplici (ad esempio la stima dell'intervallo della media di una popolazione).
- È fondamentale tenere presente che se si prende un campione a caso, questo non è sempre attendibile (basta ricordare la probabilità di errore).
- La stima comporta sempre un rischio di essere errata.



MODELLO GENERALE DI ASSOCIAZIONE CASUALE O CAUSALE

- Una associazione statistica può indicare l'esistenza di una correlazione tra variabili.
- L'associazione, però, non dimostra l'esistenza di un rapporto causa-effetto.

ASSOCIAZIONE E CAUSALITA': TIPI DI ASSOCIAZIONE

- Molti studi, ad esempio quelli su malattie e farmaci, sono finalizzati a stabilire l'esistenza di un'associazione statistica tra un presunto determinante ed un evento (ad esempio guarigione dopo trattamento con un farmaco).
- L'accertamento dell'associazione è soltanto il **primo passo**, che richiede poi l'interpretazione del **significato** dell'associazione.
- Associazione non è sinonimo di causalità.

TIPI DI ASSOCIAZIONE

- **CAUSALE**: il fattore è realmente causa dell'avvenimento.
- **SPURIA**: falsa associazione (dovuta ad esempio ad errori metodologici).
- **NON CAUSALE**: è l'avvenimento che causa il fattore, oppure c'è un terzo elemento che determina gli altri due.

ASSOCIAZIONE SPURIA

- Un'associazione **spuria** è quella dovuta alla presenza di **errori sistematici**.
- L'errore sistematico è un errore che prevedibilmente causa lo stesso tipo di errore ad ogni osservazione ed è dovuto a vizi di impostazione o di esecuzione di uno studio (campionamento, metodo, disegno dell'esperimento...).

ASSOCIAZIONE SPURIA

- **Gli errori sistematici possono** conseguire a svariati fattori (es. errato campionamento, imprecisione nella diagnosi, ecc.) e, in definitiva, conducono ad una stima errata (per difetto o per eccesso) della forza dell'associazione.

ASSOCIAZIONE NON CAUSALE

- Un'associazione **non causale** può essere la conseguenza di una *confusione* fra causa ed effetto.
- Un'associazione non causale può essere la conseguenza di:
 - confusione fra causa ed effetto
 - altro fattore *x*, vero responsabile dell'associazione

CAUSA/EFFETTO

- Le regole per dichiarare l'esistenza di una relazione causa/effetto variano in rapporto ai settori di studio.
- Risulta più facile dichiararne l'esistenza se si possono disegnare esperimenti nei quali un solo componente può essere isolato.

I CRITERI DI CAUSALITA'

- I metodi statistici non possono costituire da soli la prova che un'associazione tra due fenomeni sia basata su una relazione causa/effetto.
- Bisogna verificare la rispondenza a precisi criteri di causalità.

I CRITERI DI CAUSALITA'

I cinque criteri di causalità derivano dalla proposta dello statistico inglese **Bradford Hill** (1965) e del Comitato Consultivo per la Salute Pubblica degli U.S.A. (1966).

- I criteri sono unanimemente accettati nella comunità Scientifica.

I CRITERI DI CAUSALITA'

- **Consistenza**
- **Forza**
- **Specificità**
- **Temporalità**
- **Coerenza**

CONSISTENZA

- La consistenza di un'associazione richiede che studi diversi, eseguiti in tempi diversi ed in diverse condizioni sperimentali, evidenzino la stessa associazione.

FORZA

- La forza di una associazione è un concetto più complesso e misura la correlazione tra un presunto determinante di malattia e la malattia stessa e può essere quantificata attraverso il rischio relativo.

$$RR = \frac{I_{exp+}}{I_{exp-}}$$

FORZA

$$RR = \frac{I_{exp+}}{I_{exp-}}$$

- Ovviamente, più alto è il rischio relativo, più è probabile che l'associazione sia causale
- Inoltre, la forza dell'associazione aumenta se si riesce ad individuare l'esistenza di un effetto dose-effetto (cioè, più intensamente o a lungo agisce la presunta causa, più aumenta il rischio relativo).

SPECIFICITA'

- La specificità misura la costanza con cui una specifica variabile produce un determinato effetto.
- Più la risposta è costante, più è probabile che la variabile in esame sia una causa effettiva.

SPECIFICITA'

- Questo criterio è applicabile soprattutto alle malattie infettive, nelle quali - generalmente - l'esposizione ad un particolare agente di malattia provoca la comparsa di quella stessa malattia.
- È invece difficilmente applicabile a molte malattie cronic-degenerative ove un singolo determinante (es. fumo di sigaretta) può provocare effetti molto diversi (bronchiti, tumori, malattia cardiovascolari, ecc.).

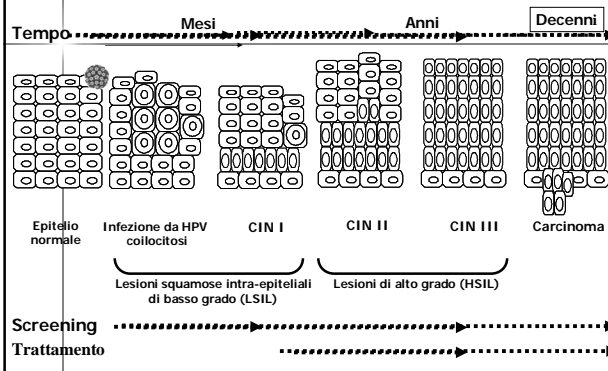
TEMPORALITA'

- La temporalità dell'associazione è basata sul principio che ogni causa deve precedere il relativo effetto.

TEMPORALITA'

- Bisogna però considerare che, soprattutto per le malattie croniche, la successione temporale degli eventi può essere difficile da stabilire, anche perché l'inizio dell'effetto non sempre è facilmente evidenziabile.
- Ad esempio: quando inizia un tumore?
In alcune condizioni è possibile addirittura incorrere nell'errore di assumere che una variabile abbia preceduto temporalmente un'altra variabile quando invece si è verificato l'opposto.

Progressione dell'infezione HPV verso il cancro della cervice uterina



COERENZA

- La coerenza può venire definita anche plausibilità biologica (fisiologica e patologica).
- Essa richiede che la presunta causa sia verosimilmente inquadrabile nel contesto delle conoscenze sull'argomento e sulla patogenesi.
- Può essere stabilita, per esempio, su modelli animali o su cellule viventi o su colture d'organo.
