

Statistica medica

IMPORTANZA DELLA (BIO) STATISTICA

- Importanza di quantificare
- Importanza di distinguere il rischio assoluto dal rischio relativo

CONCETTI FONDAMENTALI

Variabili:

- Continue → esprimono tutti i valori da zero a infinito
 - Distribuzione di frequenza
 - Normale (Gaussiana, di Bell, parametrica) → media, moda e mediana coincidono (o quasi), e sono tutti ottimi indicatori della tendenza centrale
 - Skewed (non parametrica) → la media è molto diversa da moda e mediana e non identifica la tendenza centrale, a causa della lunga coda di pochi soggetti con valori molto alti
 - Misure di tendenza centrale
 - Media: somma dei dati / numero dei dati
 - Mediana: valore al di sotto del quale cade la metà dei dati
 - Moda: valore che ricorre con maggiore frequenza
 - Misure di dispersione
 - Deviazione standard → valore che, aggiunto o sottratto alla media, comprende il 68% (2/3) delle osservazioni, misura della distanza media dei dati dalla loro media, è il più importante indice di variabilità, ed è da utilizzare sempre quando una distribuzione è normale
 - Range
 - Range interquartile → valore che, aggiunto o sottratto alla media, comprende il 50% delle osservazioni, misura della distanza media dei dati dalla loro mediana, è da utilizzare quando una distribuzione non è normale
- Categoriche → descrizione di tutto ciò che esiste in ambito biomedico, indicate da numero e percentuale
 - Misure di frequenza (incidenza, prevalenza)

Incidenza → numero di nuovi casi di malattia verificatisi durante un dato intervallo di tempo / numero di individui a rischio di ammalarsi (le persone già ammalate si escludono) nella popolazione all'inizio del periodo di tempo considerato

Prevalenza → numero dei casi di malattia presenti nella popolazione in un dato periodo / numero totale di individui nella popolazione in quel periodo

La prevalenza e l'incidenza sono rapporti, che variano da 0 ad 1 (o 0% a 100%)

La prevalenza è una stima della probabilità di essere malati (se si fa parte di una determinata popolazione in un determinato periodo di tempo)

L'incidenza è una stima della probabilità di ammalarsi (se si fa parte di una determinata popolazione in un determinato periodo di tempo)

La prevalenza e l'incidenza sono approssimativamente correlate con la seguente formula: $\text{Prevalenza} = \text{Incidenza} \times \text{durata della malattia}$
 - Misure di associazione (rischio relativo)

Stimano la forza dell'associazione tra un determinato fattore di rischio o fattore protettivo (es. fumo o farmaco) e una determinata malattia (es. tumore).

La misura "principe" è il Rischio Relativo (RR), che è calcolabile in varie forme: Risk Ratio (RR), Odds Ratio (OR), Hazard Ratio (HR), Incidence Rate Ratio (IRR), Standardized Incidence Ratio (SIR), etc. Il più comune, e frequente, è il Risk Ratio, che noi chiamiamo comunemente Rischio Relativo (RR), ed è: $\text{RR} = \frac{\text{incidenza negli esposti}}{\text{incidenza cumulativa nei non esposti}}$

 - $\text{RR} = 1$ → l'esposizione non è associata alla malattia, non influenza lo sviluppo della malattia
 - $\text{RR} > 1$ → l'esposizione è associata ad un maggior rischio di malattia - fattore di rischio
 - $\text{RR} < 1$ → l'esposizione è associata ad un minor rischio di malattia - fattore protettivo

- Metodologia di calcolo in Excel e la presentazione dei risultati
Secondo stime consolidate, l'uso dei contraccettivi orali aumenta il rischio di trombo-embolismo venoso profondo (trombosi) di circa 3 volte.
Il rischio assoluto medio annuale di trombosi, per una donna "normale" giovane, che non usa contraccettivi orali, è di circa 20×100.000 .
In 10 anni, il rischio assoluto di trombosi è quindi $20 \times 10 = 200 \times 100.000$. Di conseguenza, assumendo contraccettivi orali, il nostro rischio nei 10 anni arriva a $200 \times 3 = 600 \times 100.000$. Ciò significa che: su 100.000 donne che non usano contraccettivi, 200 avranno una trombosi; su 100.000 donne che usano contraccettivi, 600 avranno una trombosi. In un'ottica di popolazione, questo rischio è rilevante, perché in Italia significherebbe, su 6.000.000 di donne di 20-40 anni di età, $40 \times 60 = 2400$ trombosi in più all'anno. In un'ottica individuale, su 100.000 donne che usano contraccettivi, 99.400 non avranno mai una trombosi in quegli anni. Inoltre, delle 600 che hanno avuto una trombosi, 200 la avrebbero avuto comunque. Di conseguenza, se non usano contraccettivi, la probabilità di non avere una trombosi passa da 99,80% a 99,40%. In altri termini, usando contraccettivi orali la nostra probabilità di avere una trombosi aumenta dello 0,4%: l'uso dei contraccettivi danneggerà 4 su 1.000, mentre non avrà effetto sulle rimanenti 996. Secondo alcuni studi, l'uso del cellulare aumenta il rischio di gliomi di circa il 20%. Il rischio assoluto medio annuale di gliomi è di circa 5×100.000 . In 60 anni di vita, il rischio assoluto è di $60 \times 5 = 300 \times 100.000$. Di conseguenza, facendo uso del cellulare, il nostro rischio nella vita arriva a $300 + (300 \times 0,2) = 360 \times 100.000$. Ciò significa che: nella vita, su 100.000 persone che non usano il cellulare, 300 avranno un glioma; su 100.000 persone che usano il cellulare, 360 avranno un glioma. In un'ottica di popolazione, questo rischio è rilevante, perché in Italia significherebbe, su 60.000.000 di abitanti, 600 gliomi in più all'anno. In un'ottica individuale, tuttavia, su 100.000 persone che fanno uso del cellulare, 99.640 non avranno mai un glioma nella vita. Inoltre, dei 360 che hanno avuto un glioma, 300 lo avrebbero avuto comunque. Di conseguenza, se non si usa il cellulare, la probabilità di non avere un glioma nella vita passa da 99,70% a 99,64%. In altri termini, usando il cellulare la nostra probabilità di avere un glioma aumenta dello 0,06%: l'uso del cellulare danneggerà 6 su 10.000, mentre non avrà effetto sui rimanenti 9.994.

L'INFERENZA STATISTICA

Poiché tutti gli studi (o quasi) sono svolti su campioni, e non su tutta la popolazione, inevitabilmente tutti i risultati hanno un certo margine di errore, esattamente come quando si fa un sondaggio.

Siamo abituati a vedere "la forchetta" di valori quando ci sono le elezioni, che funziona più o meno così: voti presi sul campione scrutinato: 51% (47% - 55%).

Quei due valori inseriti tra parentesi indicano il margine di incertezza, perché non avendo ancora concluso lo scrutinio c'è un margine di errore, ed indicano i due valori entro i quali, quasi sicuramente, ci sarà quello vero e "finale", che otterremo solo al termine dello scrutinio. Questi due valori sono stati calcolati tramite complessi test statistici, attraverso un processo chiamato "inferenza statistica".

L'inferenza statistica consiste nel trasferire i risultati del campione alla popolazione generale, ovvero a calcolare il margine di errore del risultato trovato nel campione.

Si attua tipicamente calcolando l'intervallo di confidenza, ovvero l'intervallo di valori entro i quali starà (quasi) sicuramente il vero valore della popolazione generale.

INTERVALLO DI CONFIDENZA

Esempio:

- In uno studio su 200 studenti, a 40 di loro viene trovata una candidosi
- L'incidenza nel campione è pari al 20% (40 su 200)
- Ma siamo proprio sicuri che il 20% di tutti gli studenti ha un'infezione da Candida?
- Chiaramente, esiste un margine di errore, che va calcolato, e questo processo è l'inferenza statistica
- Il valore esatto non lo sappiamo, ma tramite semplici test statistici possiamo calcolare che nella popolazione generale di studenti l'incidenza di candidosi sarà compresa tra 14.7% e 26.2%

Nel nostro campione, la malattia è stata trovata nel 20% di studenti, ma questi erano "solo" 200, e non sappiamo il vero valore nell'intera popolazione generale di studenti.

Abbiamo quindi dovuto fare un'inferenza statistica (si fa sempre), e calcolare il margine di errore.

I due valori che abbiamo trovato, ovvero 14.7% e 26.2%, significano che, quasi sicuramente, il vero valore di incidenza di malattia nella popolazione complessiva di studenti sarà compreso tra 14.7% e 26.2%. Non è un valore preciso, ma almeno sappiamo l'intervallo entro il quale è compreso. Quel "quasi sicuramente" di cui sopra, è riferito al fatto che, nei calcoli che facciamo, impostiamo il nostro grado di sicurezza al 95%, e non al 100%. Di conseguenza, quell'intervallo di valori viene chiamato Intervallo di confidenza al 95%, e si indica con IC 95%.

POTENZA STATISTICA

L'intervallo di confidenza va calcolato sempre ed è una misura fondamentale, che ci permette di capire se le nostre stime sono precise o meno (se l'intervallo è stretto, sono ovviamente più precise, se è largo, meno).

Se l'intervallo di confidenza è troppo largo (es. tra il 5% ed il 35%), il nostro campione era troppo piccolo, ed il margine di errore è risultato troppo ampio, per cui lo studio non è servito. In termini tecnici, lo studio non ha potuto darci le indicazioni che speravamo, ed ha quindi avuto una "potenza statistica" insufficiente.

Per essere utili, gli studi devono avere una potenza statistica sufficiente, per cui il campione deve essere abbastanza grande per il nostro specifico quesito.

La dimensione del campione è quindi un parametro fondamentale, che si deve calcolare prima di fare ogni studio.

CONCETTO DI PROBABILITÀ

Fondamentalmente, di solito si riferisce alla probabilità che una differenza tra due valori sia dovuta al caso.

Per quanto ci riguarda, il punto fondamentale è sapere che la probabilità, solitamente indicata solo con "p" è un valore che si trova svolgendo l'inferenza statistica, e si riferisce alla probabilità che un risultato sia "vero".

Se, ad esempio, si confrontano i valori di incidenza di tumore al polmone nei tassisti che vivono in città inquinate con quelli dei tassisti che vivono in città non inquinate, la p ci permette di sapere se questa differenza sia "significativa", cioè "reale" e non sia dovuta al caso.

Se, dopo aver calcolato la p con gli stessi test statistici che si usano per il calcolo dell'intervallo di confidenza, troviamo una $p < 0.05$ (o 5%), significa che la probabilità che i valori dei due gruppi siano uguali (ovvero che la differenza sia casuale) è inferiore al 5%, e quindi possiamo concludere che i due valori sono "veramente" diversi.

Se invece, troviamo una $p > 0.05$, significa che la probabilità che i valori dei due gruppi siano uguali (ovvero che la differenza sia casuale) è maggiore del 5%, e quindi non possiamo concludere che i due valori sono "veramente" diversi.

Quindi:

- $p < 0.05$ = risultato "statisticamente significativo" - differenze "vere"
- $p > 0.05$ = risultato "non statisticamente significativo" - no differenza

Da notare che il 5% è un valore di incertezza preso convenzionalmente ed accettato da tutta la comunità scientifica in ambito biomedico. Anche l'intervallo di confidenza viene stabilito al 95% per lo stesso motivo (è il concetto inverso: 100%-5%).

STANDARDIZZAZIONE DEI TASSI

Solitamente, quando si calcola il tasso di una patologia (o di mortalità) in una popolazione, questo valore non viene fornito come tasso grezzo, ossia come il numero di casi per 100.000 abitanti in un anno, bensì come tasso standardizzato, normalmente per età.

Poiché, infatti, l'incidenza di morte o di patologia cambia a seconda dell'età, è utile annullare l'effetto della distribuzione per età delle popolazioni, per poter effettuare confronti precisi.

Se, ad es., confrontiamo i tassi annuali di mortalità in Italia e nello Zambia, non possiamo non tenere conto del fatto che la popolazione italiana nel complesso è molto più anziana, per cui sarebbe normale trovare dei tassi grezzi di mortalità molto più bassi nello Zambia, ma non è dovuto al fatto che davvero si muore di meno, bensì al fatto che sono tutti molto giovani.

Appare ovvio, quindi, come sia necessario standardizzare il tasso, ovvero annullare l'effetto dell'età con un particolare procedimento statistico.

La standardizzazione viene solitamente fatta per età, ma si può anche standardizzare per sesso e per tante altre caratteristiche, con o senza l'inclusione dell'età.

Ribadiamo, quindi, che un tasso di mortalità standardizzato per età elimina l'influenza della distribuzione di per età nelle popolazioni a confronto, dicendoci quanti decessi avverrebbero in ogni popolazione se queste avessero la stessa composizione per fasce di età.

I tassi standardizzati per età possono essere anche molto differenti rispetto ai tassi grezzi; è ovvio che occorre considerare soprattutto i tassi standardizzati nel confronto tra diverse popolazioni.