GTND Giornale di Tecniche Nefrologiche e Dialitiche

Epidemiologia e Statistica

Metodi di campionamento negli studi epidemiologici

Giornale di Tecniche Nefrologiche e Dialitiche
2019, Vol. 31(3) 171–174

DOI: 10.1177/0394936219869152
journals.sagepub.com/home/gtn
SAGE

Francesco Franco e Anteo Di Napoli

Abstract

Sampling methods in epidemiological studies

Sampling allows researchers to obtain information about a population through data obtained from a subset of the population, with a saving in terms of costs and workload compared to a study based on the entire population. Sampling allows the collecting of high quality information, provided that the sample size is large enough to detect a true association between exposure and outcome. There are two types of sampling methods: probability and non-probability sampling. In probability sampling the subset of the population is extracted randomly from all eligible individuals; this method, as all subjects have a chance of being chosen, allows researchers to generalize the findings of their study. In non-probability sampling, some individuals have no chance of being selected, because researchers do not extract the sample from all eligible subjects of a population; the sample is probably non-representative, the effect of sampling error cannot be estimated, so that the study produces non-generalizable results. Examples of probability sampling methods are: simple random sampling, systematic sampling, stratified sampling, and clustered sampling. Examples of non-probability sampling methods are: convenience sampling, judgement sampling.

Keywords

Simple random sample, systematic random sample, stratified random sample

Introduzione

Al fine di valutare le caratteristiche di una popolazione di interesse (source population), degli studi epidemiologici effettuano stime basate su campioni estratti da tale popolazione.

L'insieme dei metodi statistici utilizzati per produrre queste stime prende il nome di "statistica inferenziale", nata all'inizio del secolo scorso grazie al lavoro di eminenti statistici quali Pearson, Fisher, Gosset, Neyman, Wald e Tukey.[1]

Effettuare misurazioni su un'intera popolazione è un'operazione spesso non praticabile e meno conveniente rispetto alle misurazioni compiute su un suo sottoinsieme, sia in termini di tempo che di costi. Il campionamento deve essere però ben pianificato, in quanto le inferenze basate su dati campionari sono soggette a errori. Gli elementi che più influiscono sui risultati ottenuti sono: la variabilità dell'esito rilevato, l'errore di misurazione e la variabilità della stima da campione a campione.

La finalità del campionamento è infatti l'estrazione di un campione rappresentativo delle caratteristiche della popolazione di interesse. La metodologia di estrazione più corretta è la selezione casuale o randomizzata dei soggetti, procedura che dovrebbe assicurare a ciascun soggetto la stessa probabilità di essere estratto.[2] Tuttavia, anche questo metodo può non sempre garantire la reale rappresentatività della popolazione, ad esempio se non si ha una buona copertura della lista da cui si vuole estrarre il campione o nei casi di non reperibilità di individui campionati.

La scelta della dimensione campionaria è un elemento cruciale nella pianificazione di uno studio epidemiologico e costituisce uno dei motivi di fallimento della ricerca dell'effetto di un fattore di rischio per un esito.

Difatti, i motivi per cui uno studio potrebbe non trovare alcun effetto del fattore sotto indagine sono:[3]

- il fattore di esposizione non ha realmente alcun effetto sull'esito

Comitato Editoriale GTND, Roma, Italy

Correspondenza:
Francesco Franco, Comitato Editoriale GTND, Roma, Italy.
E-mail: franco_francesco@fastwebnet.it

- il disegno dello studio è inadeguato
- sfortuna
- la dimensione del campione è troppo piccola (bassa potenza)

Un campione di piccole dimensioni può portare a una stima poco precisa e affidabile del parametro di interesse. Di converso, un campione di grandi dimensioni può rendere difficile la conduzione dello studio in termine di tempo e di risorse.

Negli studi che utilizzano un campione di osservazioni le inferenze basate su tali dati sono soggette ad errori. Infatti, nel testare le ipotesi di uno studio analitico si può andare incontro a due tipi di errore:

- errore di tipo I (α): si rifiuta l'ipotesi nulla, concludendo che esiste un'associazione, in realtà inesistente, tra esposizione ed esito, ovvero che gli esiti nei gruppi confrontati sono diversi, quando invece non lo sono.
- errore di tipo II (β): si accetta l'ipotesi nulla, concludendo che non esiste alcuna associazione tra esposizione ed esito, in realtà presente, ovvero che gli esiti nei gruppi confrontati non sono diversi (cioè non esiste alcuna associazione tra l'esposizione e l'esito), quando invece lo sono.

Ma quanti soggetti sono realmente necessari per avere un buono studio?

La risposta a questa domanda è piuttosto articolata e si basa sostanzialmente su quattro elementi:

1) l'entità dell'effetto di interesse da rilevare negli studi comparativi o il grado di errore ammesso nella stima (è intuitivamente ovvio che se si desidera rilevare una piccola dimensione dell'effetto è necessario un campione più grande);
2) la variazione (cioè l'errore standard) dell'esito dello studio; con una variazione maggiore, è richiesta una dimensione maggiore del campione;
3) il livello di confidenza: un livello di confidenza più elevato nel rilevare un effetto desiderato richiede una maggiore dimensione del campione; il livello di confidenza è di solito fissato al 95%; il p-value (p) esprime la probabilità che una differenza grande (o maggiore) rispetto a quella osservata tra gruppi a confronto possa essere dovuta al caso, se l'ipotesi nulla è vera; il p-value rappresenta pertanto la probabilità di commettere un errore di tipo I (α); per valori di p-value (p) inferiori di 0,05, siamo "ragionevolmente" sicuri che l'effetto rilevato non sia dovuto al caso;
4) potenza dello studio; per potenza statistica si intende la probabilità di rifiutare l'ipotesi nulla quando questa è falsa (potenza=1-β).

Il presente articolo si propone di introdurre i concetti di campionamento probabilistico e non probabilistico e fare una sommaria rassegna di alcuni di essi.

Campionamenti probabilistici

I campioni probabilistici sono ottenuti selezionando gli individui sulla base delle loro probabilità note di essere inclusi nel campione; tale probabilità non deve essere mai pari a zero. Esempi di campionamento probabilistico sono: campionamento casuale semplice, campionamento sistematico, campionamento stratificato, campionamento a cluster.

Campionamento casuale semplice

Il campionamento casuale semplice si basa sul principio della uguale probabilità di estrazione per ogni unità della popolazione da campionare. La procedura consiste nel numerare tutte le unità della popolazione, dalle quali estrarre in maniera "casuale" le unità da inserire nel campione. I suoi vantaggi sono che la procedura è semplice, l'errore di campionamento è facilmente misurabile e si riduce il rischio di bias di selezione. È il metodo più lineare tra quelli di campionamento probabilistico.

Gli svantaggi sono soprattutto dovuti alla necessità di avere una lista completa delle unità; inoltre, può richiedere una numerosità maggiore di altre strategie di campionamento, specialmente nel caso in cui qualche caratteristica di interesse non sia particolarmente frequente nella popolazione. I costi usualmente aumentano all'aumentare della numerosità campionaria, anche perché si richiedono tempi più lunghi per l'arruolamento. Va tenuto presente che se all'aumentare della dimensione campionaria si ottengono risultati più accurati, l'accuratezza ha incrementi sempre più piccoli a fronte dell'aumento della dimensione campionaria. Inoltre, può essere difficile definire uno schema di campionamento completo, per la difficoltà di contattare le singole unità del campione. In appendice è riportato un esempio di calcolo della dimensione campionaria.

Campionamento sistematico

Nei casi in cui la randomizzazione non è effettuabile si ricorre al campionamento casuale sistematico che è il metodo che più si avvicina ai criteri della randomizzazione.[2] Gli elementi del campione sono scelti in modo tale che la distanza fra le unità campionate, precedentemente ordinate secondo un certo criterio, sia costante e di ampiezza prefissata dipendente dalla grandezza del campione. È assicurata la uguale probabilità di estrazione per ogni unità. La procedura consiste nel calcolare un intervallo di campionamento (k=N/n; dove N è la numerosità della Popolazione di riferimento e n è la numerosità del

campione); il primo elemento è estratto in modo random ed i successivi elementi sono estratti ogni k unità a partire dal primo. I vantaggi sono che assicura la rappresentatività della popolazione ed è abbastanza facile da implementare, tanto che può risultare più conveniente del campionamento casuale semplice. Gli svantaggi consistono nel fatto che può essere distorto se la lista contiene cicli, vale a dire se ciclicamente alcune caratteristiche sono associate a determinate unità. In tal caso si ha un rischio molto elevato di bias di selezione.

Campionamento stratificato

Nel campionamento stratificato la popolazione da campionare è suddivisa in sottogruppi omogenei, detti strati. La procedura consiste nell'estrarre campioni randomizzati separatamente in ogni strato di entità proporzionale alla percentuale di quel sottogruppo nella popolazione totale e poi combinare i risultati del campionamento da tutti gli strati.[2] Un esempio di campionamento stratificato è la valutazione del tasso di copertura delle vaccinazioni in Italia: si estrae un campione in ogni regione che rappresenta lo strato, le stime sono calcolate per ogni strato, infine ogni strato viene "pesato" per ottenere la stima nazionale complessiva. I vantaggi sono che il campionamento è più preciso se l'esito è associato alla variabile di stratificazione; infatti, tutti i sottogruppi sono rappresentati, permettendo conclusioni separate su ognuno di loro. Gli svantaggi sono che l'errore di campionamento è difficile da misurare e ci può essere una mancanza di precisione se in alcuni strati i numeri sono piccoli.

Campionamento a grappolo (cluster)

In assenza di liste di unità elementari da cui campionare o quando risulta troppo oneroso costruirle in termini di tempo e costi ma si dispone soltanto dei loro raggruppamenti, è possibile procedere con un piano di campionamento detto a cluster.[2] Nel campionamento a cluster invece di selezionare un campione casuale di soggetti, viene selezionato un campione casuale di gruppi di individui (aree geografiche, villaggi, scuole) chiamati "cluster". Il campionamento a cluster è indicato ed utilizzabile solo quando la variabilità è minima tra i cluster e massima al loro interno. I vantaggi consistono nel fatto che è più semplice soprattutto se la lista completa delle unità non è disponibile; inoltre sono richieste meno risorse in termini di spazio e di tempo necessario al campionamento. Gli svantaggi sono che può essere impreciso se i cluster sono omogenei e di conseguenza la variazione campionaria del cluster risulterà più grande della variabilità della popolazione (grande "design effect"); inoltre, l'errore di campionamento è difficile da misurare. Il campionamento a cluster differisce dal campionamento stratificato: infatti il campionamento stratificato cerca di dividere il campione in gruppi eterogenei, così che la varianza entro strato è bassa e quella tra gli strati è alta; il campionamento a cluster idealmente cerca di avere ogni cluster con una variabilità pari a quella della popolazione (ogni cluster come una "mini" popolazione).

Campionamenti non probabilistici

I campioni non probabilistici sono ottenuti selezionando gli individui senza tener conto della loro probabilità di essere inclusi nel campione; pur essendo più rapidi da ottenere e a costi più bassi, hanno lo svantaggio di non essere accurati, a causa della presenza di un bias di selezione nel campionamento; è noto, infatti, come in alcune situazioni gli individui si auto selezionano e questo non consente di generalizzare i risultati. Campioni della popolazione estratti in modo non casuale sono detti campioni di convenienza, che possono anche avere un loro valore in determinate situazioni, ma non consentono di inferire i risultati generalizzandoli, come avverrebbe basandosi su un campione ottenuto con metodi random. Esempi di campionamento non probabilistico sono: campionamento di convenienza, campionamento con esperti.

Campionamento di convenienza (convenience sampling)

Il campione è identificato principalmente per convenienza, opportunisticamente o incidentalmente.[4] È una tecnica non probabilistica. Le unità sono selezionate senza conoscere la loro probabilità di essere estratte. Alcuni esempi di questa tipologia di campionamento sono rappresentati dalle interviste via e-mail, per strada o dai volontari per una ricerca. Il vantaggio è che questo tipo di campionamento è probabilmente relativamente semplice, veloce e pertanto economico. Lo svantaggio però sta nel fatto che è impossibile valutare quanto sia rappresentativo della popolazione studiata.

Campionamento con esperti (judgment sampling)

È anche noto come campionamento soggettivo, basandosi su una metodologia che prevede la selezione, da parte di esperti della materia oggetto dello studio, di persone o gruppi di persone che ritengono rappresentative della popolazione o comunque delle caratteristiche da studiare. La metodologia è vantaggiosa in termini di tempo e di costi ed è particolarmente utile nella ricerca qualitativa. Gli svantaggi derivano dal fortissimo rischio di "volunteer bias", dal potenziale errore di giudizio del ricercatore che non di rado rende i risultati non rappresentativi di quelli che si osserverebbero nella popolazione di riferimento.

Dichiarazione di assenza di conflitto di interessi

Gli Autori dichiarano di non aver conflitti di interessi.

Finanziamenti

Gli Autori dichiarano di non aver ricevuto finanziamenti specifici da qualsiasi ente nel settore pubblico, privato o senza fini di lucro.

Bibliografia

1. Levine DM, Krehbiel TC and Berenson ML. *Statistica*. Milano: Apogeo. 2002; 1:4–6.
2. Attena M. *Epidemiologia e valutazione degli interventi sanitari*. Padova: Piccin. 2004; 3:43–7.
3. Dohoo I, Martin W and Stryhn H. *Methods in Epidemiologic Research*. 1st edition Charlottetown: VER Inc. 2002; 2:35–59.
4. Tyrer S and Heyman B. Sampling in epidemiological research: issues, hazards and pitfalls. *The British Journal of Psychiatry Bulletin* 2016; 40(2):57–60.

Appendice

Esempio di calcolo della dimensione campionaria nel caso di un campionamento casuale semplice

Si voglia valutare la proporzione di soggetti con una data patologia (prevalenza) in una data popolazione. In base ad indagini effettuate nel passato nella stessa popolazione si prevede che la prevalenza sia pari a 0,1 (10%). Si ritiene accettabile un errore del 5% (precisione). Quanti soggetti della popolazione dovranno essere selezionati se la stima derivante dal campione deve cadere entro 5 punti percentuali rispetto alla vera prevalenza, con una confidenza del 95%?

$$n = \frac{1,96^2 * prev_{att} * \left(1 - prev_{att}\right)}{D^2}$$

sostituendo i valori nella formula:

$$n = \frac{1,96^2 * 0,1 * \left(1 - 0,1\right)}{0,05^2} = 138$$

Cosa succede se a parità di precisione (5%) e confidenza del 95% la prevalenza attesa fosse sconosciuta e assunta, quindi, pari al 50%? Quanti soggetti della popolazione dovranno essere selezionati?

$$n = \frac{1,96^2 * 0,5 * \left(1 - 0,5\right)}{0,05^2} = 384$$

Cambiando scenario, con un errore del 5% e una confidenza del 99%, quanti soggetti della popolazione dovranno essere selezionati se la prevalenza attesa fosse del 10%?

$$n = \frac{2,58^2 * 0,1 * \left(1 - 0,1\right)}{0,05^2} = 240$$

Dagli esempi sopra descritti si evince che il calcolo della dimensione campionaria dipende dalla varianza nella popolazione del fenomeno in esame, che nel nostro caso è uguale a: $prev_{att} * \left(1 - prev_{att}\right)$, dall'ampiezza dell'intervallo di confidenza desiderata e dal margine di errore ammesso.