L’indagine campionaria: l’affidabilità statistica dei dati raccolti

Quando leggiamo la nota metodologica di una determinata indagine troviamo un frase del tipo: “errore massimo atteso compreso nel range ±3%; intervallo di confidenza: 95%.

Cosa significa questa frase?

Le ricerche nascono solitamente dalla necessità di fornire una stima del valore effettivo di una o più variabili di una popolazione in un determinato periodo. Ma la stima che si ricava da una ricerca non corrisponde esattamente al valore effettivo a livello dell’universo di riferimento; il fatto di condurre l’indagine su un campione e non sull’intera popolazione mostra una “*stima*” del parametro che si vuole conoscere e non il suo valore esatto che potrebbe, invece, essere ottenuto solo con un’indagine condotta su tutto l’universo di riferimento. Se in una indagine campionaria volta a conoscere la quota di laureati presenti in determinata regione ottenessimo un valore del 25% questa dovrebbe essere interpretata in questo modo: il valore effettivo del parametro studiato (la quota percentuale dei laureati di una determinata area) corrisponde con una probabilità del 95% al 25% ±3 della popolazione (il valore effettivo si colloca cioè, dando per scontata l’assenza di altri tipi di errore non derivanti dal campionamento, nell’intervallo 22-28%).

Il “*margine d’errore*” rappresenta, dunque, la sintesi dell’errore di campionamento che quantifica il margine di incertezza che dobbiamo attenderci dai risultati di una ricerca. Il margine d’errore si basa sulle leggi della probabilità. Se il nostro campione è probabilistico il margine d’errore che siamo disposti ad accettare può essere quindi stabilito in anticipo.

Un importante fattore nella determinazione del margine d’errore è la dimensione del campione. I campioni più ampi producono stime più vicine al valore effettivo della popolazione (e dunque margini d’errore inferiori rispetto a campioni numericamente meno consistenti). Facciamo un altro esempio; se in un sondaggio pre-elettorale emerge che il 55% degli intervistati è favorevole al candidato X, e noi avevamo accettato un margine d’errore del ±10%, con intervallo di confidenza del 95%, la quota effettiva della popolazione che è favore del candidato X si trova all’interno dell’intervallo che va da (55%-10%) a (55%+10%) ossia dal 45% al 65%; un margine di errore massimo del 10% implica, a causa del campione modesto, una qualche incertezza sulla reale quota della popolazione favorevole al candidato X. Accettando, invece, un margine d’errore del ±3%, aumentando cioè la numerosità campionaria da 100 a 1.000 persone e supponendo sempre che il 55% risulti a favore del candidato X, potremmo affermare che il livello effettivo della popolazione favorevole al candidato X si trova, con una probabilità del 95%, all’interno dell’intervallo che va da (55%-3%) a (55%+3%) ossia dal 52% al 58%, che rappresenta una stima certamente più precisa di quella precedente.

Da cosa dipende il margine d’errore

I tre fattori che incidono sul margine d’errore sono l’ampiezza del campione, il tipo di campionamento effettuato e la dimensione della popolazione.

Dimensione del campione

La dimensione del campione è un elemento che incide in modo cruciale sul margine d’errore. Un campione di 100 casi produrrà un margine d’errore non superiore al 10%, un campione di 500 produrrà una margine d’errore non superiore al 4.5%, e un campione di 1.000 produrrà un margine di non più del 3%. Questo dimostra che i tassi percentuali di errore diminuiscono all’aumentare della dimensione campionaria ma non dipendono dall’ampiezza della popolazione se abbiamo effettuato un campionamento casuale semplice (ad esempio, con un campione di 2.500 casi avremo stime con un margine d’errore del ±2% e con 400 casi avremo stime con un margine d’errore del ±5% indipendentemente dall’ampiezza della popolazione).

Probabilmente sorprenderà il fatto che l’ammontare della popolazione poco incida poco sul margine d’errore. In effetti, un campione di 100 soggetti di una popolazione di 10.000 avrà almeno lo stesso margine di errore di un campione di 100 persone in una popolazione di 10 milioni di persone.

Tipo di campionamento: campioni probabilistici

Campionamento casuale semplice

È un tipo di campionamento col quale ogni unità della popolazione ha la stessa probabilità di essere scelta per entrare a far parte del campione. Questo presuppone la disponibilità di una lista dell'intera popolazione dalla quale estrarre i casi (ad ogni unità della lista corrisponderà a un numero random).

Campionamento sistematico

Anche il campionamento sistematico genera un campione casuale semplice; la differenza con quest'ultimo consiste nel modo in cui vengono estratti i casi. Le unità non vengono più selezionate tramite sorteggio ma scegliendone "sistematicamente" una ogni tot unità o, meglio, ogni dato intervallo. Anche in questo caso deve essere nota la lista della popolazione dopodiché possiamo determinare l'intervallo all'interno del saranno scelte le unita mediante una semplice formula:

$$k = N/n$$

dove K indica l'unità della popolazione da selezionare, N l'ampiezza della popolazione e n la numerosità del campione desiderata. Il valore che si ottiene da questa formula rappresenta l'ambito dal quale devono essere scelte le singole unità; ad esempio, con N = 4522 e n = 400 ottengo k = 11.3; questo vuol dire che per formare il campione sceglierò un'unità ogni 11 iniziando con un numero estratto a caso fra 1 e 11. Se il numero estratto è 8 allora le unità scelte saranno {la n° 8, la n° 16, la n° 24 ecc. in questo modo si selezionano 400 casi}; se alla fine avessimo ottenuto un numero superiore a 400, ad esempio 412 unità, avremmo dovuto scartare le ultime 12 unità (n = 400).

Campionamento stratificato

Con il campionamento stratificato si suddivide la popolazione in gruppi (strati) quanto più omogenei possibile rispetto alla variabile che si vuole indagare utilizzando una variabile ad essa collegata; da ogni gruppo (strato) si estrae poi un campione casuale semplice; il campione globale si ottiene unendo i diversi sottocampioni. Se, ad esempio, vogliamo stimare il reddito medio di una popolazione possiamo stratificare la popolazione in base alla variabile "professione" che è fortemente correlata al reddito e, se gli strati considerati sono 4 (professionista/dirigente, commerciante/artigiano, lavoratore dipendente pubblico e lavoratore dipendente privato), estrarre 4 sottocampioni.

Il vantaggio che offre il campionamento stratificato è quello di suddividere una popolazione eterogenea (che richiederebbe un campione molto ampio) in strati relativamente omogenei che richiedono campioni più piccoli la cui somma è inferiore all'ampiezza del campione che avremmo dovuto estrarre dalla popolazione eterogenea.

Naturalmente questo campionamento è possibile solo quanto la variabile da indagare presenta delle zone di maggiore omogeneità.