Dieci buoni motivi per non credere (troppo) ai sondaggi – 5. Errore di campionamento

written by Paolo Natale | 8 Luglio 2022

È un tipo di errore insopprimibile in tutti i sondaggi basati su campioni probabilistici (dove cioè tutti hanno la medesima probabilità di essere estratti), per il semplice fatto che intervistando un numero limitato di individui (ad esempio, un migliaio di elettori) non avremo mai la certezza che i risultati siano effettivamente generalizzabili a tutta la popolazione (elettorale). Sinteticamente, può essere definito come una sorta di "fascia di indeterminatezza" (chiamata in termini statistici "intervallo di confidenza") entro la quale oscillano i risultati, ottenuti da qualsiasi indagine campionaria, se vogliamo da questi risalire a stime dell'universo. Conoscendo la numerosità campionaria di una rilevazione demoscopica, possiamo automaticamente calcolare la misura dell'errore di campionamento, utilizzando apposite tavole basate su specifici teoremi statistici. L'ampiezza dell'errore di campionamento è strettamente correlata con due elementi: la numerosità campionaria e l'omogeneità dei risultati ottenuti. Vediamo come.

Se da un sondaggio di 1000 casi rileviamo ad esempio che James Joyce è conosciuto dal 15% degli italiani intervistati, è possibile affermare (con una probabilità data di non sbagliare) che il livello di conoscenza dello scrittore è compreso tra il 13% e il 17% dell'universo di riferimento. Se intervistiamo 2000 persone, cioè il doppio, ottenendo supponiamo il medesimo risultato (il 15%), il margine di errore si restringe: possiamo in questo secondo caso affermare cioè che il livello di conoscenza di Joyce è compreso tra il 14% e il 16%. A parità di metodo di rilevazione, per migliorare la precisione delle stime campionarie occorre

dunque aumentare l’ampiezza del campione; al crescere del numero dei casi, diminuisce infatti parallelamente l’errore di campionamento.

Veniamo al secondo elemento che determina l’ampiezza dell’errore, cioè il livello di omogeneità delle risposte fornite dagli intervistati (misurabile attraverso il calcolo della “varianza” di una variabile). Se chiediamo ad un campione di 1000 individui la loro opinione su Papa Francesco ed il 99% ne dà un giudizio “ottimo”, l’errore di campionamento per questa risposta sarà uguale a 0,3%; se chiediamo al medesimo campione l’opinione sul Presidente Mattarella e soltanto il 50% ne dà un giudizio “ottimo”, l’errore di campionamento per questa risposta sarà invece più alto, pari al 3,1%. Perché? Perché più le risposte sono omogenee (cioè più è bassa la varianza), più siamo sicuri di non commettere un errore molto ampio nello stimare l’universo, vale a dire nello stimare le risposte di tutti coloro che non abbiamo effettivamente intervistato.

Per chiarire meglio, pensiamo a due amici (Federer e Nadal) che giocano a tennis da anni l’uno contro l’altro. Se nei 1000 incontri giocati tra loro ha vinto 990 volte Federer, avremo molte più probabilità di estrarre a caso, da quei 1000, un incontro dove lui ha vinto, di quante ne avremmo se nei 1000 incontri avessero vinto 500 volte a testa. L’incertezza del risultato (e la disomogeneità delle risposte, vale a dire un’elevata varianza) determina quindi un aumento dell’incertezza delle stime, vale a dire un maggior margine di “imprecisione” quando si vuole estendere la stima campionaria alla popolazione.

In questo caso il ricercatore, se già aveva utilizzato il miglior sistema di campionamento possibile, nulla può fare per diminuire l’errore campionario, non potendo sapere ovviamente in anticipo il livello di omogeneità delle risposte che otterrà. L’unico elemento cui potrà attingere, per stabilire la miglior numerosità campionaria, sarà l’analisi di

precedenti indagini sullo stesso tema. Sempre che ve ne siano.

Paolo Natale

**estratto del volume “Sondaggi”, in uscita nel prossimo autunno presso Laterza*