Sommario: 1.1. Introduzione. – 1.2. Popolazione finita, indagine campionaria. – 1.2.1. Tipologie delle popolazioni finite. – 1.2.2. L'evoluzione dell'indagine campionaria. – 1.2.3. Tipologie e fasi dell'indagine campionaria. – 1.3. Errore statistico, qualità dei dati, “*error profile*”. – 1.4. Analisi della popolazione.

1.1. Introduzione

Per analizzare e conoscere un qualsiasi fenomeno che si manifesta in una popolazione¹ è necessario svolgere un'indagine statistica.

Tale indagine, qualunque sia la caratteristica progettuale che la contraddistingue, può essere realizzata o in forma esaustiva, ottenuta cioè attraverso la rilevazione completa delle manifestazioni del fenomeno, o svolta in forma parziale, laddove se ne consideri solo un sottoinsieme.

Ciascun fenomeno considerato prende il nome di variabile che può essere *unidimensionale*, quando per ciascuna unità di cui si compone la popolazione si analizza un solo carattere, *multidimensionale*, quando per la medesima unità vengono considerati più caratteri. Tale variabile può dunque essere rilevata sull'intera popolazione, vale a dire su tutte le unità elementari che la costituiscono, realizzando una rilevazione totale o *censuaria* ovvero su una parte delle unità che la compongono, attuando quindi una rilevazione parziale o *campionaria*.

Ovviamente, la scelta risulta del tutto vincolante nei casi in cui la rilevazione di tipo censuario sia di fatto impossibile (si pensi alle oltre 55 misure antropometriche rilevabili su tutta la popolazione italiana) o quando la rilevazione comporti la distruzione delle unità osservate (analisi di alcune caratteristiche di un prodotto industriale). Ma ancor più selettivamente l'indagine campionaria viene

¹ Nel prosieguo con il termine *popolazione* si intende una *popolazione statistica*.

scelta per motivi specifici quali la rapidità di realizzazione, il grado di approfondimento dell'informazione, il costo contenuto. Come anche l'indagine censuaria è necessaria quando si debbano effettuare comparazioni e analisi a livello di microaree, e risulta indispensabile quando deve costituire la base per i possibili confronti e raccordi nelle indagini campionarie ad essa connesse.

La struttura di un'indagine campionaria, che nella terminologia anglosassone prende il nome di *survey*, è caratterizzata da un insieme di fasi che devono essere considerate tutte egualmente importanti ai fini di un positivo giudizio di qualità.

La complessità di un'indagine varia notevolmente in funzione della dimensione della popolazione da studiare e della diversa facilità con cui le unità di cui si compone possono essere raggiunte ed intervistate. Svolgere un'indagine sugli abitanti di un piccolo comune o sulle strutture di una Azienda Sanitaria Locale (ASL) o infine sui membri di un ordine professionale, potrebbe essere considerato abbastanza semplice; viceversa, svolgere un'indagine a livello nazionale su più variabili riferite alla popolazione di un grande paese, o su più sottopopolazioni (chiamate campi o domini di studio), è un'operazione molto complessa.

In questo capitolo verranno presentate le diverse tipologie di indagini campionarie e le fasi in cui si articolano, inoltre verranno introdotte le diverse tipologie di popolazioni finite.

1.2. Popolazione finita, indagine campionaria

L'indagine campionaria (*sample survey*) si caratterizza rispetto a quella censuaria o totale per almeno due aspetti fondamentali:

- a) la scelta di una specifica procedura di selezione delle unità della popolazione $\mathcal{U}$;
- b) l'induzione, in base all'analisi delle osservazioni campionarie, dal campione all'intera popolazione.

Intanto va precisato, con riferimento al punto a), che quando si parla di popolazione in questo contesto si intende una *popolazione finita* cioè costituita da un insieme finito e noto di unità elementari che, sebbene numerose, siano comunque enumerabili e quindi etichettabili, tali cioè da rendere identificabile l'unità i -esima rispetto alla j -esima.

Per quanto detto, la popolazione finita sarà nel seguito indicata con $\mathcal{U} = \{1, 2, \dots, N\}$ dove i numeri naturali $1, 2, \dots, N$ indicano le unità che la compongono; $\mathbf{U}$ connota una caratteristica qualitativa o quantitativa unidimensionale analizzata e $Y_1, Y_2, \dots, Y_N$ i valori o più in generale le modalità, che la caratteristica medesima assume rispettivamente nelle unità $1, 2, \dots, N$.

Un esempio può chiarire quanto detto: se si pensa alle talpe che sono presenti su un determinato territorio si fa riferimento a una popolazione finita; ma le talpe non essendo identificabili, e non essendo perciò possibile associare a ciascuna di esse un'etichetta, non possono costituire una popolazione finita nei termini predetti. Al contrario, le matricole di una certa università italiana di un certo anno accademico costituiscono una popolazione finita in quanto ciascuna di esse è etichettabile e quindi identificabile. Se inoltre, per ciascuna delle matricole, si volesse analizzare il sesso e la statura si avrebbe a che fare con lo studio di un carattere rispettivamente qualitativo discreto e quantitativo continuo.

Per quanto detto la popolazione delle matricole può essere studiata o in forma censuaria, cioè contattandole tutte o ricorrendo a una parte di esse, cioè ad un campione. Nel primo caso si generano dati utilizzabili per qualunque sottoinsieme o dominio della popolazione, ma sicuramente si va incontro a costi e a tempi di realizzazione più elevati; nel secondo caso si ottengono informazioni in tempi più brevi, con costi più contenuti ed inoltre si può disporre di informazioni più accurate ed estese ottenute attraverso l'impiego di questionari anche complessi che possono essere somministrati da personale altamente qualificato e addestrato.

1.2.1. Tipologie delle popolazioni finite

La scelta di un campione da una popolazione finita non è cosa semplice, anzi costituisce uno dei punti più critici di un'indagine campionaria in quanto tale scelta è soggetta a numerose imperfezioni, difficili da eliminare completamente.

A tale scopo e con riferimento a quanto riportato da Kish (1987) e ripreso da Groves (1989) appare utile delineare le quattro più note tipologie di popolazione che entrano in gioco nell'operazione di scelta di un campione da una popolazione finita. Le popolazioni sono:

1. *la popolazione sotto inferenza (PI) (population of inference)*. È l'insieme degli individui che sono oggetto di studio, in un intervallo di tempo ben definito. Ad esempio, può interessare la popolazione presente sul territorio nazionale nel primo semestre di un certo anno, ai fini di un'indagine sui consumi alimentari; oppure il numero di ricoverati negli ospedali di una regione che appartengono alla popolazione residente nella stessa regione ad una certa data; oppure, infine, il numero di studenti universitari fuori corso sul totale degli studenti universitari iscritti alle università di una città o di una nazione ad una certa data;
2. *la popolazione obiettivo (PO) (target population)*. È l'insieme degli individui di ampiezza finita, scopo dello studio. Della popolazione obiettivo possono non fare parte alcuni gruppi di unità appartenenti alla PI come, ad esempio, i ri-

coverati negli ospedali, i detenuti negli istituti di pena o anche gli abitanti di alcune zone del paese o dei comuni al di sopra di una certa altitudine. La distinzione operativa tra la popolazione obiettivo e la popolazione sotto inferenza viene effettuata generalmente dal ricercatore, sia implicitamente, sottacendo tale differenza, sia esplicitamente, evidenziando i gruppi non considerati;

3. *la popolazione base per il campionamento (PC) (frame population)*. È la lista di unità utilizzata per estrarre il campione; può essere una lista di unità elementari, ad esempio individui, o di aggregati di unità elementari, ad esempio famiglie, sezioni elettorali, sezioni di censimento, ospedali, scuole, rioni, quartieri ecc. e può includere informazioni ausiliarie di ogni tipo. In un'indagine campionaria, le unità della PC, sono quelle su cui viene realizzata la procedura di campionamento;
4. *la popolazione effettivamente indagata (PE) (survey population)*. È l'insieme degli individui che, qualora selezionati, dovrebbero essere oggetto di intervista in quanto dispongono delle necessarie capacità fisiche e mentali per partecipare all'indagine. Ovviamente, la propensione a rispondere alle interviste varia nei diversi individui e tale propensione è indipendente dalla scelta campionaria.

Dall'analisi delle caratteristiche delle quattro popolazioni considerate (PI, PO, PC, PE) ne consegue, per la stessa indagine, una diversa ampiezza nelle numerosità delle medesime; tali differenze sono generate proprio dai diversi tipi di imperfezione che compaiono sia prima di dare inizio all'indagine, sia durante l'esecuzione della stessa. Infatti, le principali imperfezioni riguardano:

- *la non copertura (noncoverage)* che si presenta quando la lista, utilizzata per l'identificazione delle unità che compongono una popolazione finita, risulta incompleta; ciò genera il conseguente *errore di copertura (noncoverage error)*; questa prima imperfezione è la causa della differenza in termini di numerosità tra la PO – generalmente più numerosa – e la PC;
- *le mancate risposte totali (total nonresponse)* che sono generate principalmente da unità che, per diversi motivi, non rispondono all'intervista (*unit nonresponse*); ciò genera il conseguente *errore da mancate risposte totali (unit nonresponse error)*; questa seconda imperfezione è alla base della differenza tra la PC – in questo caso più numerosa – e la PE.

È molto importante tenere distinte le quattro popolazioni e le imperfezioni connesse a causa degli effetti che tali differenze generano sulle statistiche. Naturalmente, il fenomeno della non copertura è più grave in quanto è più difficile da stimare, mentre le mancate risposte totali possono almeno essere misura-

te, anche se i motivi per cui alcuni rifiutano l'intervista risultano molto differenti e, in alcuni casi, difficili da spiegare.

1.2.2. L'evoluzione dell'indagine campionaria

I punti salienti del processo attraverso il quale si è potuta affermare la validità dell'indagine campionaria partono dalle sessioni scientifiche dell'International Statistical Institute (ISI) del 1895, 1899 e 1901 nelle quali il norvegese Kiaer propose il metodo del campione rappresentativo riuscendo a dimostrare empiricamente che, tramite un campione stratificato, si potevano ottenere buone stime della media e del totale dei caratteri presenti in popolazioni finite. Nella sessione dell'ISI del 1903, tenuta a Berlino, il metodo venne accolto anche se con alcune condizioni.

Nel 1926 apparvero sul Bollettino dell'ISI due importanti lavori sull'argomento: uno, a opera di Bowley nel quale, bilanciando opportunamente tecnica campionaria e inferenza statistica, venne presentata un'organica teoria del metodo campionario e dimostrata l'accuratezza di stime per grandi campioni provenienti da popolazioni finite; l'altro, scritto da Jensen, nel quale venne chiarito quanto fosse consistente, in termini di risultati, la differenza tra campionamento casuale e campionamento a scelta ragionata.

Ben più significativo è il contributo alla teoria dei campioni offerto da Neyman (1934) che introdusse i principi base su cui fondare il campionamento probabilistico; in particolare, egli definì un intervallo di confidenza per la media e il totale di un carattere basandosi esclusivamente sulla distribuzione di tutti i campioni estraibili, introdusse il concetto di efficienza associato all'intervallo di confidenza e pose in evidenza i vantaggi del metodo del campionamento casuale stratificato delegittimando con ciò definitivamente il campionamento a scelta ragionata che aveva mantenuto, fino ad allora, un ruolo importante.

Dopo il fondamentale lavoro di Neyman numerosi altri autori contribuirono successivamente a dare alla teoria dei campioni un assetto sempre più sistematico. In termini cronologici, si ricordano i lavori di Hansen e Hurwitz (1943) che per primi proposero l'associazione alla variabile sotto analisi di un'altra variabile presente sulle unità della popolazione, la cui dimensione potesse costituire il supporto per la selezione delle unità campionarie con probabilità proporzionali alla dimensione medesima. Horvitz e Thompson (1952) svilupparono il tema del campionamento senza rimessa con conseguenti probabilità variabili di selezione delle unità della popolazione. Nel 1953 Cochran e nel 1950 Sukhatme fornirono una sistematica della teoria dei campioni che ancora oggi, con le aggiornate nuove edizioni, rappresentano essenziali punti di riferimento.

Non può essere comunque dimenticato, in questa breve rassegna storica, il

lavoro di Godambe che nel 1955 propose una significativa classe di stimatori lineari e nel 1960 e 1966 dimostrò, dapprima, l'inesistenza di uno stimatore lineare corretto per la media di un carattere con varianza uniformemente minima e, successivamente, cercò di dimostrare, anche nel campo dell'inferenza da popolazioni finite, quanto già codificato per l'inferenza da popolazioni illimitate, anche se i risultati non furono sempre completamente confrontabili, specie per i concetti di *sufficienza*, *ammissibilità* e, in particolare per quello di *verosimiglianza*, dei quali si fa cenno nel seguito.

La notevole considerazione per questi problemi da parte degli statistici teorici e operativi è stata testimoniata, a livello internazionale, dall'ampio spazio riservato al campionamento nelle sessioni dell'ISI tenute a Nuova Delhi nel 1977, a Buenos Aires nel 1981, ad Amsterdam del 1983; in quest'ultima sessione si è cercato addirittura di definire l'entità del reciproco contributo scientifico e applicativo alla soluzione dei problemi connessi con la teoria e la pratica del campionamento statistico.

Questa rassegna, ovviamente non esaustiva, dei numerosi contributi alla teoria del campionamento non poteva terminare senza menzionare i più recenti lavori di Brewer e Hanif (1983) e di Chaudhuri e Vos (1988) nei quali vengono presentati quasi tutti i metodi di estrazione senza ripetizione di campioni di ampiezza n dalla popolazione, con probabilità proporzionali alla dimensione di una variabile ausiliaria, alcuni dei quali sono oggetto di successiva specifica trattazione (cfr. par. 3.4).

1.2.3. Tipologie e fasi dell'indagine campionaria

Si possono considerare diverse tipologie di indagini campionarie in relazione agli obiettivi, alla diffusione territoriale della popolazione coinvolta e alle variabili sotto analisi. In molti casi si parla di indagini *ad hoc*, cioè di indagini molto diffuse che vengono programmate e realizzate per raggiungere un determinato obiettivo conoscitivo e si concludono con il soddisfacimento del medesimo. Le indagini ad hoc possono avere molteplici e specifici obiettivi e in molti casi si rivolgono a popolazioni di limitata entità; esse sono programmate e realizzate con notevole frequenza da istituti di ricerca privati. Tali istituti sono prevalentemente presenti nel nord-Italia, data la forte concentrazione industriale, e al centro-Italia per la notevole concentrazione degli enti pubblici.

Si parla invece di indagini *longitudinali* quando l'indagine viene ripetuta nel tempo, a cadenze stabilite. Se l'obiettivo è quello di osservare un fenomeno in tempi diversi, ad esempio si vuole misurare il consumo di certi beni o l'efficienza e l'efficacia di certi servizi, si effettuerà la stessa indagine su differenti campioni di consumatori. Se invece si vuole mettere in evidenza la tendenza del fe-

nomeno sarà opportuno considerare lo stesso campione di unità in tutte le occasioni di rilevazione; in questo contesto il campione viene chiamato *panel*. Tuttavia, per la difficoltà di disporre per molte occasioni dello stesso insieme di unità, solitamente si fa riferimento non a un panel di consumatori fissi ma a *campioni ruotati*, vale a dire a campioni composti in parte da unità che hanno partecipato alle indagini precedenti ed in parte da nuove unità².

Un'altra tipologia di indagini riguarda i *sondaggi*. La principale caratteristica di questi si basa sulla selezione di poche variabili – si utilizzano quindi questionari molto ridotti – riferite ad argomenti di attualità. Di solito i sondaggi coinvolgono campioni scelti da tutta la popolazione nazionale.

Infine, è doveroso menzionare le indagini campionarie condotte da istituzioni pubbliche quali, ad esempio l'*Indagine sui bilanci delle famiglie italiane* (IBF) svolta a cadenza biennale dalla Banca d'Italia; la *Rilevazione sulle forze di lavoro* (IFL), indagine trimestrale condotta dall'ISTAT e, sempre da quest'ultimo Istituto, l'*Indagine sui consumi delle famiglie* (ICF) e l'*Indagine multiscopo sulle famiglie* (IMF). Come si può notare sono indagini su diversi argomenti di interesse nazionale, da cui è possibile ottenere stime dei parametri dei vari fenomeni considerati anche a livello regionale e provinciale.

Qualunque sia l'indagine – campionaria o censuaria – la sua organizzazione si articola in una serie di procedure che devono essere accuratamente svolte. Ad esempio, secondo l'ISTAT l'indagine si deve articolare nelle seguenti fasi:

1. *progettazione*, che consiste nel definire gli obiettivi e le unità che si intendono raggiungere con l'indagine, i costi ed i tempi di esecuzione; se censuaria o campionaria e, se campionaria, quale piano di campionamento è ritenuto più idoneo; infine, quali caratteristiche devono connotare il questionario;
2. *rilevazione*, che definisce l'insieme delle operazioni necessarie alla raccolta delle informazioni, presso le unità oggetto di analisi, tramite la selezione e l'impiego di intervistatori e supervisori che, utilizzando il questionario predisposto, rispettivamente effettuano l'intervista e ne controllano la validità;
3. *registrazione*, mediante la quale, i dati rilevati tramite il questionario vengono trasferiti su supporto informatico, assumendo la caratteristica di dati elaborabili;
4. *revisione e codifica*, che consiste nella verifica e nella correzione di dati grezzi oltre che nella messa a punto delle procedure informatiche finalizzate a tale scopo;

² Benché il sistema di rotazione delle unità in un campione ruotato renda l'indagine piuttosto complessa e onerosa, tuttavia negli ultimi anni l'indagine longitudinale è stata sempre più frequente.

5. *elaborazione*, che si esplica attraverso la definizione di tabelle semplici e multiple e di indicatori statistici finalizzati alla costruzione di files di analisi;
6. *validazione*, che consiste nell'individuazione della coerenza interna o, eventualmente, esterna qualora risulti necessario il ricorso a fonti esterne;
7. *diffusione*, mediante la quale, i risultati ottenuti vengono resi disponibili per gli utilizzatori finali, corredati spesso da alcune dichiarazioni circa le condizioni generali in cui si è svolta l'indagine.

Si fa notare che, in ciascuna di queste fasi, vengono generate alcune tipologie di errore connesse tipicamente con la fase medesima; nel complesso si considerano 21 tipologie di errore riportate in dettaglio nel capitolo sesto dedicato agli errori non campionari.

1.3. Errore statistico, qualità dei dati, "error profile"

Nella realizzazione di un'indagine statistica sia essa esaustiva o censuaria ovvero parziale o campionaria, è indispensabile associare alla medesima l'*errore statistico* inteso come discrepanza tra *valore vero* e *valore disponibile* dall'indagine statistica (valore osservato o rilevato).

Ovviamente, quest'ultima definizione di errore statistico assume diverso contenuto quando si associa alla singola *unità statistica* oppure a una *statistica di sintesi* dei dati rilevati mediante indagine campionaria. Nel primo caso, la discrepanza predetta fa riferimento al complesso degli *errori non campionari* cioè agli errori statistici ottenuti come somma di tutti gli errori che possono essere commessi in una qualsiasi fase del processo di indagine comprendendo le possibili interazioni ed escludendo gli errori campionari. Nel secondo caso, agli errori non campionari si aggiungono quelli campionari per cui la statistica campionaria differisce generalmente dal valore osservabile in un'indagine esaustiva anche quando si ipotizzi, teoricamente, che i valori di ogni unità vengono rilevati con esattezza.

In generale, l'errore non campionario essendo presente in tutti i tipi di indagine e connotandosi in forma e dimensione differente per ciascuna di esse, può essere opportunamente considerato come una guida all'informazione sulla *qualità dei dati* prodotti; qualità intesa in termini di rispondenza alla realtà fattuale che tali dati permettono di dimensionare.

Nella pratica, ogni indagine deve essere vissuta come un qualsiasi processo produttivo e in tale ottica per qualità dei dati si intende la loro capacità di soddisfare le proprietà garantite dal produttore dell'indagine medesima. Più in dettaglio, tali proprietà riguardano l'*attendibilità*, intesa in termini di grado di accu-

ratezza delle stime e l'*adeguatezza* riferita più precisamente alla definizione degli obiettivi e alla diffusione dei risultati dell'indagine.

È quindi molto utile che, per ogni indagine, vi sia un progetto che definisca non solo i modelli di adeguatezza e la struttura delle fasi che la caratterizzano, ma anche l'*error profile*, cioè il profilo dell'errore dell'indagine, ovvero la descrizione completa e ordinata delle operazioni e delle potenziali fonti di errore nonché dell'effetto dell'errore di ciascuna operazione sull'errore complessivo.

Più precisamente B.A. Bailar (1983) sostiene che "Ideally an *error profile* would not only list the survey operations and the potential sources of error, but also the impact of the error in each operation on the total survey error" e che "This ideal is rarely, if ever, possible because the measurement of nonsampling errors is rarely undertaken" e che "Another reason it is as yet impossible to get a measurement of total survey error is that we do not yet have a mathematical model that adequately reflects the interaction of the errors from different sources".

Naturalmente gli effetti positivi della messa a punto di un profilo dell'errore in un'indagine sono notevoli sia per chi la realizza sia per chi la utilizza; infatti:

- permette il dimensionamento di un errore complessivo oltre che di quello campionario e non campionario;
- definisce la documentazione che caratterizza la procedura di indagine a fini divulgativi;
- sensibilizza gli utilizzatori sugli effetti dei diversi tipi di errore; stimola la messa a punto di programmi per il controllo della qualità dei dati.

L'importanza dei predetti effetti positivi lascerebbe pensare a una notevole diffusione dei profili dell'errore; ma non è così. Sono molto rari i casi di indagine, siano esse di *mercato* o di *opinione*, con *finalità statistiche* o con *finalità amministrative* che siano dotate di un profilo dell'errore inteso nel senso precedentemente delineato. Ma è opportuno notare che esistono alcuni seri motivi che limitano la messa a punto di tali supporti, quali ad esempio:

- la complessità degli staff di ricerca spesso oberati da specifiche necessità connesse con l'indagine stessa;
- la considerazione che un rapporto di ricerca che riporti gli errori dell'indagine riduce la credibilità dei dati e delle statistiche prodotte;
- l'ammissione dell'esistenza di errori che di fatto induce implicitamente ad ammettere un lavoro scadente.

A questi motivi, pur validi, si può tuttavia obiettare affermando che la descrizione delle possibili fonti di errore nonché della dimensione e della caratteristica degli stessi tende ad aumentare la credibilità e la significatività dell'indagine.

Quanto, infine, alla deviante, secondo alcuni, denominazione di *error profile* essa potrebbe essere, più positivamente trasformata in *quality profile*.

Ma la vera causa della scarsa diffusione nelle indagini dei corrispondenti profili dell'errore sta nella difficoltà a definire l'elemento caratterizzante di essi cioè la formulazione di *modelli matematici* per la quantificazioni degli errori parziali e dell'errore complessivo dell'indagine medesima (Montinaro, 1988; 1991). Si ritiene, comunque, che l'esigenza di *error* o *quality profile* vada sviluppandosi parallelamente col crescere della competitività fra produttori di indagine e con la maggiore domanda di qualità da parte degli utilizzatori.

Per alcune indagini effettuate dall'Istat in Italia sono stati approntati degli *error profile* ma da ricercatori diversi da quelli che hanno progettato e condotto le indagini medesime (Montinaro, 1988; 1991).

1.4. Analisi della popolazione

Per conoscere le caratteristiche di una o più variabili nella popolazione finita $\mathcal{U}$ vengono generalmente impiegati alcuni noti indici statistici che permettono di disporre di un quadro di sintesi sufficientemente adeguato a descrivere sinteticamente la popolazione medesima. Nel seguito vengono elencati gli indici più utilizzati riferiti alle variabili $\mathcal{Y}$, $\mathcal{X}$, ... presenti in $\mathcal{U}$ ³.

- La *media aritmetica* della variabile $\mathcal{Y}$, indicata con $\bar{Y}$, è fornita dalla quantità:

$$\bar{Y} = \mu = \frac{1}{N} \sum_{i=1}^N Y_i \quad (1.1)$$

- Il *totale*, indicato con Y , è dato dalla quantità:

$$Y = N\bar{Y} = \sum_{i=1}^N Y_i \quad (1.2)$$

- La *varianza*, indicata con σ^2 , è pari a:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad (1.3)$$

che si presenta come una media di scarti al quadrato e gode della nota proprietà di minimo della media aritmetica.

³ Tali indici sono tipicamente riferiti a caratteri quantitativi; per l'impiego di analoghi indici per caratteri qualitativi si veda Frosini (2009); Montinaro, Nicolini (2007).