INTRODUZIONE ELEMENTARE AL CAMPIONAMENTO STATISTICO DA POPOLAZIONI FINITE

A. GIOMMI

A. PETRUCCI

Dipartimento di Statistica "G. Parenti"

Università degli Studi di Firenze

1 INTRODUZIONE

L'indagine è lo strumento statistico mediante il quale si acquisiscono informazioni su uno o più fenomeni attinenti ad una popolazione.
L'informazione può essere acquisita osservando tutte le unità componenti la popolazione o soltanto parte di esse. Nel primo caso, l'indagine è detta completa, nel secondo, parziale o campionaria.
L'indagine completa è teoricamente semplice ma all'atto pratico presenta molti lati negativi. Se la popolazione che si desidera studiare è molto numerosa, le risorse economiche e personali necessarie al suo corretto svolgimento possono essere superiori a quelle disponibili. Anche i tempi di esecuzione possono spesso superare limiti accettabili o comunque limitarne notevolmente la frequenza. Si pensi ai censimenti che per la spesa e la mole di lavoro che comportano - sia in fase organizzativa che di esecuzione - non potrebbero avere cadenza più stretta di quella decennale.
Inoltre, le indagini complete non possono essere svolte:

(i) su popolazioni non finite (come ad esempio può essere concettualmente considerata ogni popolazione che origini da un processo produttivo di tipo industriale);
(ii) su popolazioni per le quali l'osservazione del fenomeno di studio comporti la distruzione dell'unità che si osserva (come ad esempio la durata di accensione di una lampada o la resistenza alla rottura di una barra metallica, ecc.).

Per contro, l'indagine campionaria offre all'atto pratico una serie di vantaggi. In primo luogo, non vi sono limitazioni legate alla dimensione della popolazione o alla natura delle unità componenti. In secondo luogo, la possibilità di limitare la rilevazione ad un insieme di unità di dimensione ben inferiore a quella della popolazione consente di:

(i) contenere i costi dell'indagine entro limiti accettabili;
(ii) svolgere l'indagine in tempi relativamente brevi;
(iii) raccogliere per ogni unità inclusa nell'indagine un maggior numero di informazioni;
(iv) raccogliere le informazioni con maggior accuratezza grazie all'utiliz-zazione di
(v) personale qualificato *e/o* di tecniche specialistiche.

Sul piano teorico tuttavia l'indagine campionaria presenta due notevoli problemi: il primo, legato al modo in cui deve essere scelto il campione; il secondo, relativo ai procedimenti da adottare per estendere l'evidenza campionaria alla popolazione. Lo studio di questi problemi, che come si vedrà sono strettamente collegati, costituisce l'oggetto della teoria del campionamento statistico.
Il presente scritto contiene gli elementi introduttivi di tale teoria. Il suo scopo è quello di illustrare in estrema sintesi i principali aspetti teorici e tecnici che stanno alla base di alcuni metodi campionari di larga diffusione.

2 POPOLAZIONE E CAMPIONE

Qualunque indagine nasce da esigenze conoscitive. Se a seguito di tali esigenze si stabilisce di effettuare una ricerca, in primo luogo, si dovranno definire con precisione i suoi obiettivi. Possiamo considerare questo momento come la prima fase dell'indagine. Tra gli obiettivi da definire vi è la popolazione oggetto di studio o "popolazione obiettivo". In generale, per popolazione si intende un insieme finito o infinito di unità che non interessano prese singolarmente ma per il contributo che danno alle proprietà statistiche dell'insieme di appartenenza. In seguito, faremo riferimento esclusivamente a popolazioni di dimensione finita ed indicheremo con N il numero complessivo di unità componenti la popolazione.

Definire la popolazione obiettivo significa individuare con esattezza la natura dei suoi elementi componenti, cioè delle unità oggetto di studio, e la sua estensione spaziale e temporale.

In questa stessa fase vengono definiti in dettaglio gli aspetti o, meglio, le caratteristiche della popolazione che si intendono studiare e, conseguentemente, si stabiliscono le modalità di rilevazione delle stesse. Nella maggior parte delle indagini in campo sociale le modalità di rilevazione sono rappresentate da un insieme di domande raccolte in una scheda di rilevazione o questionario. Il questionario può essere redatto su carta o implementato su supporto informatico.

Si definisce campione un qualsiasi sottoinsieme di n unità ($n \leq N$) della popolazione. L'indagine completa, della quale non tratteremo in queste note, può comunque essere vista come un caso particolare di quella campionaria nel caso in cui $n = N$.

Vi sono numerosi metodi per selezionare un campione e diverse possibilità di classificarli. Una distinzione di importanza fondamentale è quella tra campioni probabilistici (o casuali) e non probabilistici.

Si parla di campione probabilistico quando:

(i) è possibile definire l'insieme di tutti i possibili campioni che possono essere formati seguendo una determinata procedura di estrazione (detta schema di selezione);

(ii) è possibile associare a ciascuno di essi una probabilità di selezione nota;

(iii) la procedura di selezione permette di attribuire a ogni unità componente la popolazione una probabilità strettamente positiva di essere estratta;

(iv) la procedura consente di selezionare un campione con probabilità esattamente corrispondente a quella che gli era stata associata a priori.

Sono non probabilistici i campioni che non hanno i requisiti suddetti.

Se la probabilità di estrazione è costante per ogni unità della popolazione o di sotto-popolazioni in cui viene suddivisa la popolazione, si parla di campionamento equi-probabilistico.

Nella pratica, la selezione di un campione probabilistico viene effettuata utilizzando routine di programmi informatici. In passato l'estrazione era effettuata con l'ausilio delle tavole di numeri casuali che surrogavano, per grandi popolazioni, i meccanismi per selezione casuale propri dei giochi di sorte come l'urna per il lotto, la sacca dei numeri per la tombola ecc..

3 DISEGNO DI CAMPIONAMENTO E DISEGNO DI INDAGINE

Un'indagine campionaria può avere molteplici obiettivi conoscitivi. Nella maggior parte delle indagini in campo sociale, l'obiettivo principale è rappresentato dalla "stima" di grandezze caratteristiche della popolazione dette “parametri”.
Sul termine stima torniamo successivamente. Per il momento possiamo osservare che la selezione del campione e la stima dei parametri della popolazione rappresentano senz'altro i due momenti di maggiore interesse teorico dell'indagine campionaria. Questa a sua volta può essere vista come un insieme di fasi interrelate che nel loro complesso vengono identificate con il termine disegno di indagine (dall'inglese *survey design*) o piano di indagine (*survey plan*). Le fasi relative alla selezione del campione e alla stima dei parametri della popolazione costituiscono il cosi detto piano o disegno di campionamento (*sampling design*).
Il disegno di indagine, di contenuto più ampio, comprende oltre agli aspetti appena elencati:
- la definizione della popolazione oggetto di indagine;
- la scelta dei caratteri (variabili) da studiare, del modo di definirli e di osservarli;
- la scelta e la definizione dei livelli, o domini, spaziali e temporali di indagine;
- la definizione dei metodi di raccolta, di codifica e di elaborazione dei dati;
- l'individuazione dei costi e dei livelli di precisione e accuratezza desiderati
- la scelta delle analisi statistiche da affiancare ai metodi di stima;
- la metodologia di calcolo degli errori campionari;
- i metodi di controllo rilevazione e correzione degli errori non campionari
- la presentazione di dati statistici e dei risultati.

Occorre tenere presente che l'articolazione in fasi non implica il loro succedersi secondo l'ordine precedentemente dato. L'elenco ha uno scopo essenzialmente didattico. In pratica, parte di queste fasi procedono in simultanea e possono essere ripercorse a più riprese. Esse, inoltre, interagiscono tra loro in modo diverso da un indagine all'altra e ciò avviene in particolare tra il disegno di campionamento e le restanti operazioni.

4 STIMA

Scopo principale dell'indagine campionaria è la stima di una o più costanti caratteristiche (parametri) della popolazione. La stima è il procedimento statistico mediante il quale un valore ricavato come funzione (cioè elaborazione) delle osservazioni campionarie viene assunto a rappresentare il valore incognito della corrispondente funzione nella popolazione.
I parametri di maggior interesse sono rappresentati da medie, totali e differenze o rapporti tra queste grandezze, per i caratteri (o variabili) quantitativi e da proporzioni o percentuali, per i caratteri qualitativi dicotomici. Per i caratteri che non ha senso o non è utile esprimere in forma dicotomica, sono oggetto di stima le distribuzioni di frequenza, assolute e/o relative, nella popolazione.
Le stime campionarie devono possedere delle proprietà. Poiché la stima è effettuata su un campione, cioè su un sottoinsieme della popolazione, essa non coinciderà normalmente con il valore che si desidera stimare. La più ovvia proprietà ed anche quella che le riassume

tutte è che la stima sia più prossima possibile al parametro incognito della popolazione che si desidera stimare.
Proviamo ad esprimere questa proprietà in modo diverso. La differenza tra stima e vero valore (che purtroppo non è dato conoscere) viene denominata, nella teoria, errore di campionamento. Dunque la proprietà prima citata può essere vista anche come possibilità di ridurre ai minimi termini l'errore di campionamento.
E' possibile ridurre o addirittura annullare l'errore di campionamento e, se si, in che modo? E'intuitivo che la dimensione del campione ha un ruolo fondamentale nella riduzione dell'errore di campionamento. In effetti l'errore si riduce all'aumentare della dimensione del campione. L'errore di campionamento è assente nei censimenti, dal momento che nell'indagine censuaria si rilevano (almeno in teoria) tutte le unità della popolazione, ma, come abbiamo ricordato nell'introduzione, i censimenti proprio per le ingenti risorse che richiedono non possono essere effettuati che a cadenza decennale. La dimensione del campione è chiaramente legata alle risorse disponibili e per questo non si è liberi di variarla se non entro i limiti imposti dalle risorse stesse.
Per chiarire ulteriormente con un esempio il ruolo rilevante della dimensione campionaria è possibile pensare all'indagine sulle forze di lavoro che il comune di Firenze effettua a proprio carico in parallelo alla rilevazione dell'ISTAT. Le 464 famiglie intervistate trimestralmente dall’ISTAT corrispondono a circa 900 individui, troppo pochi per stimare con ragionevole precisione, cioè con un ridotto errore campionario, le principali grandezze di interesse a livello comunale. Già dal 1995 il comune aveva portato a 1200 il numero di famiglie intervistate trimestralmente selezionando ed intervistando a proprie spese un campione aggiuntivo di 736 famiglie. Un campione di 1200 famiglie (per circa 2500 individui) ha un errore di stima indubbiamente inferiore di un campione di 464.
E' logico chiedersi perché non 2000 o 2500 famiglie? La risposta è ovvia: le risorse disponibili non avrebbero coperto i costi che si sarebbero sostenuti per aumentare ulteriormente il numero delle interviste.
Viene allora naturale chiederci: la dimensione campionaria vincolata com'è alle risorse disponibili rappresenta la sola possibilità di riduzione dell'errore di campionamento? La risposta è no; a parità di dimensione campionaria e quindi a parità di costi legati alle interviste, vi sono tecniche di campionamento e di stima che consentono, sotto certe condizioni, una maggiore riduzione dell'errore campionario rispetto ad altre. Un esempio di questa ultima affermazione è rappresentato dalla nuova impostazione che è stata data all'indagine del Comune di Firenze sulle forze di lavoro prima citata. A partire dal 2002, il Comune ha deciso di estrarre non un campione di famiglie, come nel piano adottato dall'ISTAT, bensì un campione di singoli individui. Questo accorgimento, unitamente ad altre caratteristiche del piano di campionamento delle quali parleremo successivamente, consente di ottenere, con un campione di 1200 individui, delle stime il cui grado di precisione è analogo a quello che si otterrebbe con 1200 famiglie cioè con un numero quasi doppio di individui utili per l'indagine (si ricorda che non vengono intervistati i giovani minori di 15 anni dato che per legge non possono lavorare).
Nei successivi paragrafi descriveremo in termini informali e sintetici i seguenti piani di campionamento, di uso frequente nella pratica, cercando di evidenziare le caratteristiche che li rendono più o meno adeguati nelle diverse situazioni operative:
- campionamento casuale semplice;

- campionamento casuale stratificato;
- campionamento sistematico;
- campionamento a grappoli;

5 CAMPIONAMENTO CASUALE SEMPLICE

Il campionamento casuale semplice rappresenta il naturale punto di partenza per lo studio di tutti gli altri piani di campionamento probabilistici. Il campionamento casuale semplice può essere definito come segue:
Si consideri una popolazione di dimensione N dalla quale si desideri estrarre un campione di n unità. Il campionamento casuale semplice è il piano che attribuisce la stessa probabilità di selezione a ciascun possibile insieme di n unità distinte della popolazione.
Utilizziamo un semplice esempio didattico per chiarire i contenuti della definizione. Supponiamo che la popolazione sia formata da $N = 4$ unità:

U_1, U_2, U_3, U_4

e che si desideri estrarre un campione casuale di $n = 2$ unità. E' facile verificare che le possibili coppie di unità distinte sono le seguenti:

$(U_1\ U_2)$ $(U_1\ U_3)$ $(U_1\ U_4)$
$(U_2\ U_3)$ $(U_2\ U_4)$ $(U_3\ U_4)$.

Affinché ciascuna di queste coppie, cioè ciascuno di questi sei campioni sia un campione casuale semplice è sufficiente che la sua probabilità di estrazione sia pari a 1/6.
In altre parole se estraiamo una delle sei coppie di unità con probabilità costante pari a 1/6 formiamo un campione casuale semplice di $n = 2$ unità.
Si noti che, operativamente, per formare uno dei sei campioni, possiamo semplicemente selezionare una prima unità dalle quattro presenti nella popolazione, assegnando a ciascuna probabilità di estrazione pari a 1/4 e, successivamente, una tra le tre rimanenti, assegnando a ciascuna probabilità di estrazione pari a 1/3.
Facciamo adesso un passo avanti per vedere in che modo è possibile procedere ad un'operazione di stima mediante questo piano di campionamento. Supponiamo che le quattro unità introdotte siano persone e ci interessi stimare il numero di ore lavorate alla settimana. Indichiamo il numero di ore di lavoro settimanali con Y e la media incognita che desideriamo stimare con $\overline{Y}$. La media ha in generale la seguente espressione:

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad ; \quad (i = 1,..., N) \tag{1}$$

Y_i è valore del carattere associato ad una generica unità della popolazione e i varia da 1 a N. Nel nostro semplice esempio, i assume solo i valori 1, 2, 3 e 4 e, quindi, la media (1) Può essere scritta:

$$\bar{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$$

Supponiamo infine che nella popolazione i valori del carattere Y siano i seguenti:

$Y_1 = 20$, $Y_2 = 40$, $Y_3 = 36$, $Y_4 = 48$,

E' quindi ovvio che in corrispondenza dei possibili 6 campioni selezionabili siano osservabili le seguenti coppie di valori:

($Y_1 = 20$, $Y_2 = 40$) ($Y_1 = 20$, $Y_3 = 36$) ($Y_1 = 20$, $Y_4 = 48$)
($Y_2 = 40$, $Y_3 = 36$) ($Y_2 = 40$, $Y_4 = 48$) ($Y_3 = 36$, $Y_4 = 48$)

In pratica si osserva un unico campione dei 6 possibili e da quello occorre stimare la media incognita della popolazione.
Lo stimatore della media della popolazione che si utilizza è la media calcolata sullo stesso campione.
Le possibili medie campionarie osserva bili sono:

30, 28, 34, 38, 44, 42

e si può osservare che nessuna di queste corrisponde alla media della popolazione che è pari a (20 + 40 + 36 + 48)/4 = 36. Se il campione estratto è quello con madia 44 la differenza con il vero valore è piuttosto consistente, mentre se il campione estratto è quello con media 34 o quello con media 38 il valore è più vicino a quello, incognito, della popolazione. Va da sé che l'esempio è puramente didattico e lontano anche dalla più semplice delle situazioni reali, ma è comunque valido per evidenziare due aspetti molto importanti della teoria del campionamento:

(i) il singolo campione può produrre una stima anche abbastanza diversa dal vero valore da stimare;
(ii) non ci sono proprietà riferibili ad un singolo campione; ma soltanto all'insieme dei possibili campioni che si possono selezionare.

Il primo punto è ovvio. Riguardo al secondo punto è possibile utilizzare l'esempio per verificare, sia pure numericamente, una proprietà della stima utilizzata (la media campionaria) e il legame tra dimensione campionaria e precisione della stima.
Si deve in primo luogo osservare che la media incognita della popolazione è uguale alla media calcolata sulle medie dei possibili campioni:

$(30 + 28 + 34 + 38 + 44 + 42)/6 = 36 = \bar{Y}$

Questa è una proprietà generale della media campionaria, che per questo motivo è definita come "stimatore corretto" o "non distorto" della media della popolazione.
Disporre di uno stimatore corretto può non essere sufficiente; la correttezza dello stimatore non ci garantisce, e lo abbiamo visto nell'esempio, che la stima del campione osservato sia

prossima al valore da stimare. Sarebbe importante che non ci fossero possibili campioni che producono stime distanti dal vero valore da stimare. Abbiamo inoltre osservato in precedenza che stime più precise possono essere ottenute aumentando la dimensione campionaria. Possiamo verificare questa affermazione attraverso l'esempio numerico. Prima tuttavia è necessario introdurre un indice che misuri la precisione della stima. E' cioè necessario sintetizzare in un unico valore l'entità media delle differenze tra stima e vero valore.
L'indice che si utilizza per valutare la precisione dello stimatore (nel nostro caso la media campionaria) è la varianza, che si ottiene come media degli scarti quadratici delle possibili stime dal vero valore da stimare. Nel nostro esempio, indicando con $V(\bar{Y})$ la varianza dello stimatore media campionaria:

$$V(\bar{Y}) = [(30-36)^2 + (28-36)^2 + (34-36)^2 + (38-36)^2 + (44-36)^2 + (42-36)^2]/6$$
$$= 34{,}67$$

otteniamo un valore che esprime una misura di precisione dello stimatore. Un ulteriore indice anche più utile all'atto pratico è la radice quadrata della varianza: $ES(\bar{Y}) = \sqrt{V(\bar{Y})}$ che prende il nome di errore medio di stima (o errore standard della stima), ed è pari nel nostro esempio a 5,89.
Questi valori sono poco indicativi presi singolarmente. Ma sono interessanti in termini comparativi.
Tornado alla nostra popolazione di quattro unità, supponiamo di poter estendere a tre unità la dimensione campionaria. I possibili campioni casuali sono ora soltanto quattro.

$(U_1\ U_2\ U_3)$ $(U_1\ U_2\ U_4)$ $(U_2\ U_3\ U_4)$

con i corrispondenti valori di Y:

(20 40 36) (20 40 48) (20 36 48) (40 36 48)

e le corrispondenti medie:

32 36 34,67 41,33

La media delle quattro medie è ancora pari a 36, a conferma della correttezza dello stimatore. La varianza invece: $V(\bar{Y}) = 11{,}54$ e l'errore medio di stima $ES(\bar{Y}) = 3{,}4$ sono ben inferiori ai precedenti, a dimostrazione della relazione tra precisione dello stimatore e dimensione del campione.
E' interessante osservare che la varianza dello stimatore o il suo errore medio di stima, che abbiamo calcolato utilizzando l'insieme dei possibili campioni casuali estraibili dalla nostra popolazione, possono essere calcolati anche in modo diverso. Gli stessi valori di $\sqrt{V(\bar{Y})}$ e

di $ES(\bar{Y})$ possono essere ricavati calcolando la varianza dei valori della popolazione e dividendo il risultato per la dimensione del campione.
Nella teoria del campionamento la varianza dei valori della popolazione (detta varianza elementare) si indica normalmente con il simbolo S^2 ed ha la seguente espressione:

$$S^2 = \frac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}{N-1}, \qquad (2)$$

e la varianza dello stimatore, che indichiamo con $\bar{y}$:

$$V(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right). \qquad (3)$$

In questa espressione, in realtà, oltre a dividere la varianza della popolazione per *n*, si è anche moltiplicato per il fattore (1- *n*/*N*) che viene denominato fattore di correzione per popolazione finita (mentre *n*/*N* è detta frazione di campionamento) ed è un termine che si applica quando il campione vene estratto senza reimmissione delle unità nella popolazione. Nella pratica l'estrazione del campione è sempre effettuata senza reimmissione, ma il fattore (1- *n*/*N*) può essere omesso tutte le volte che *N* è molto grande rispetto a *n* e, di conseguenza, è uguale approssimativamente a 1.
Applicando ai dati della nostra popolazione le formule (2) e (3), abbiamo per campioni di dimensione *n* = 2:

$$S^2 = [(20-36)2 + (40-36)2 + (36-36)2 + (48-36)2]/3 = 138{,}67$$

$$V(\bar{y}) = \frac{138{,}67}{2}\left(1 - \frac{2}{4}\right) = 34{,}67$$

e per campioni di dimensione *n* = 3:

$$V(\bar{y}) = \frac{138{,}67}{3}\left(1 - \frac{3}{4}\right) = 11{,}54\,,$$

come avevamo ricavato in precedenza.
Prima di concludere questa parte dedicata al campionamento casuale semplice, occorre avvertire che sia $\sqrt{V(\bar{y})}$ che $ES(\bar{y})$ non possono essere calcolati nella pratica perché dipendono dai valori Y_i incogniti della popolazione. Vi è tuttavia la possibilità di stimare sia la varianza che l'errore medio di stima dal campione. Le procedure di stima della varianza sono in qualche caso complesse e pertanto non riteniamo opportuno discuterne in queste note. Nel nostro esempio, tuttavia, e in generale per il campionamento casuale semplice, la stima della varianza è invece semplice. È sufficiente applicare l'espressione (3)

ai dati del campione estratto, cioè sostituendo la formula di S^2, nella (2), che ovviamente non è calcolabile, con l'analoga espressione calcolata su dati campionari. Indichiamo questa ultima con s^2:

$$s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1} ,$$

con la somma estesa ai valori y_i campionari dei quali $\bar{y}$ è la media. Pertanto la stima campionaria della varianza dello stimatore della media, $\hat{V}(\bar{y})$, è data dalla seguente espressione:

$$\hat{V}(\hat{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right).$$

E' senz'altro opportuno esaminare, sia pure molto sinteticamente, anche il caso di stima di una proporzione mediante un campione casuale semplice. La proporzione di unità che nella popolazione detengono un particolare attributo è frequentemente uno dei parametri di maggiore interesse nelle indagini. I risultati teorici per le proporzioni derivano direttamente da quelli appena illustrati per le medie, considerando la proporzione come una media calcolata su un carattere che possa assumere solo due valori: 1, ad indicare il possesso di un certo attributo; 0, ad indicarne la mancanza. Pertanto, la proporzione P di unità che hanno un certo attributo nella popolazione è equivalente alla media $\bar{Y}$ del carattere stesso e la corrispondente proporzione campionaria p alla media campionaria $\bar{y}$ o, dato che il campione è casuale semplice, $\bar{y}$.

La formula della varianza S^2 dato che Y_i può assumere solo valore 0 o 1, può essere scritta in una forma alternativa rispetto alla (2.5) e lo stesso vale per la sua stima campionaria. In particolare, $S^2 = NPQ/(N-1)$, con $Q = (1-P)$ e $s^2 = npq/(n-1)$, con $q = (1-p)$.
Pertanto le espressioni della varianza dello stimatore e della sua stima campionaria, saranno le seguenti:

$$V(p) = \left(1 - \frac{n}{N}\right)\frac{NPQ}{(N-1)n} = \frac{PQ}{n}\frac{N-n}{N-1}$$

e

$$\hat{V}(p) = \left(1 - \frac{n}{N}\right)\frac{pq}{(n-1)}$$

6 CAMPIONAMENTO CASUALE STRATIFICATO

Prima di intraprendere un'indagine sono spesso disponibili, per tutte le unità della popolazione, alcune informazioni che possono essere utilizzate nel piano di campionamento per migliorare la qualità dell'indagine stessa. Per esempio, se dobbiamo formare un campione di comuni sono normalmente noti, prima dell'estrazione, i dati relativi alla loro localizzazione, alla loro dimensione demografica, alla loro attività economica prevalente, ecc..

La stratificazione è una tecnica che consente di utilizzare questo tipo di informazioni, dette ausiliarie o supplementari, per ottenere alcuni vantaggi tra i quali il più importante è una maggiore precisione degli stimatori.

La stratificazione consiste:

(i) nella suddivisione della popolazione in sottopopolazioni effettuata in base alle informazioni ausiliarie; tali sottopopolazioni sono dette strati.

(ii) nella selezione di campioni indipendenti da ciascuno strato, in modo che la somma delle dimensioni dei campioni di strato sia in totale pari a n, cioè la dimensione campionaria programmata rispetto alla popolazione obiettivo nel suo complesso.

I maggiori vantaggi della stratificazione discendono dal fatto che la dimensione dei campioni negli strati anziché essere determinata dalla casualità dell'estrazione - come avverrebbe nel campionamento casuale semplice - è sotto il controllo di chi effettua il campione.

Spesso i campioni sono formati applicando in tutti gli strati la stessa frazione di campionamento. Essi risultano in tal caso di dimensione proporzionale a quella dello strato di provenienza e la stratificazione stessa viene detta *proporzionale*. Con questo tipo di stratificazione si ha la garanzia di ottenere stime migliori (più precise) di quelle che proverrebbero da un campione casuale semplice della stessa dimensione complessiva. Altre forme di stratificazione non garantiscono questo obiettivo ma consentono di perseguire altri obiettivi che possono assumere notevole rilievo nell'indagine.

Si pensi alla possibilità di costruire strati ciascuno dei quali raccolga unità appartenenti ad una categoria, un gruppo, una sottopopolazione di particolare interesse nell'indagine, generalmente indicata col termine dominio di studio. Questo avviene, ad esempio, quando gli strati sono circoscrizioni territoriali per le quali è necessario disporre di risultati analoghi a quelli che si vogliono ottenere per la popolazione nel suo complesso. In questa situazione, sarà opportuno cercare di conferire a questi risultati (stime) lo stesso grado di precisione nei diversi strati. Ciò sarà spesso realizzabile selezionando campioni di strato che abbiano approssimativamente la stessa dimensione.

Per semplicità supporremo che negli strati le unità siano selezionate con un campionamento casuale semplice. I risultati che vengono riportati in questa sezione sono facilmente generalizzabili ad altri metodi di selezione negli strati.

Le notazioni già introdotte per il campione casuale semplice necessitano solo di piccole modifiche per tener conto della divisione della popolazione in strati. Denotiamo con N_h la dimensione dell'h-esimo strato e con H il numero di strati formati, pertanto: $h = 1,...,H$ e $\sum_h N_h = N$.

Analogamente denotiamo con n_h ($\sum_h n_h = n$) la dimensione del campione nel generico strato *h*. $\bar{Y}_h$ e $\bar{y}_h$ sono rispettivamente la vera media e la media campionaria nello strato *h*; S_h^2, la varianza elementare (per il carattere Y) nello strato *h*.
La frazione di campionamento nello strato *h* è indicata con $f_h = n_h/N_h$. Infine, è utile introdurre il nuovo simbolo: $W_h = N_h/N$ che rappresenta la proporzione della popolazione nello strato *h*; ovviamente, $\sum_h W_h = 1$.
Assunto un campionamento casuale semplice negli strati, i risultati visti nella precedente sezione possono essere estesi a ciascuno strato preso separatamente. Pertanto le medie campionarie $\bar{y}_h$ sono corrette per le rispettive medie $\bar{Y}_h$ e le relative varianze si ricavano immediatamente dall'espressione (3).
Il problema principale riguarda come combinare tra loro le stime di strato per ottenere uno stimatore della media generale $\bar{Y}$. Altri problemi relativi alla stima campionaria della varianza dello stimatore non vengono presi in esame in queste note. Il problema si risolve semplicemente ricorrendo allo stimatore analogico - cioè avente struttura analoga a quella del parametro da stimare - dato da:

$$\bar{y}_{str} = \sum_{h=1}^{H} W_h \bar{y}_h$$

E' immediato verificare che tale stimatore è corretto per $\bar{Y}$. Inoltre poiché i campioni sono selezionati indipendentemente da uno strato all'altro,

$$V(\bar{y}_{str}) = \sum_h W_h^2 V(\bar{y}_h).$$

Si può osservare come la varianza dello stimatore sia funzione di quella elementare, interna ai vari strati. La possibilità di ridurre la varianza dello stimatore è quindi legata a quella di ottenere strati che risultino (rispetto alla variabile di indagine) più omogenei della popolazione presa nel suo complesso. Questo obiettivo ha una probabilità di essere realizzato tanto maggiore quanto maggiore risulta l'associazione tra caratteri di stratificazione e carattere di indagine. Nella pratica i caratteri di indagine sono numerosi ed è estremamente improbabile che negli strati le unità risultino omogenee per ciascuno di essi.

7 STRATIFICAZIONE PROPORZIONALE E NON

Le espressioni appena viste si riferiscono ad un qualsiasi tipo di distribuzione delle unità campionarie negli strati. Se la stratificazione è proporzionale, cioè se: $f_h = f$ per ogni *h*, possono essere riscritte in termini più semplici.
In primo luogo, lo stimatore della media, che denotiamo $\bar{y}_{stp}$, può essere espresso come

una media semplice delle osservazioni campionarie:

$$\bar{y}_{stp} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i \in s_h} y_{hi}$$

dove y_{hi} è il valore dell'*i*-esima unità campionaria dello strato *h* e la seconda sommatoria è estesa a tutte le unità che appartengono al campione s_h dello strato *h*. Inoltre, anche l'espressione della varianza si semplifica come segue:

$$V\left(\bar{y}_{stp}\right) = \frac{1-f}{n}\sum_h W_h S_h^2 = \frac{1-f}{n} S_w^2,$$

in cui S_w^2 è la media ponderata delle varianze di strato.
La varianza viene stimata dalla:

$$v\left(\bar{y}_{stp}\right) = \frac{1-f}{n}\sum_h W_h s_h^2.$$

Si può osservare che la varianza dello stimatore della media nel campionamento stratificato proporzionale è simile a quella dello stesso stimatore nel campionamento casuale semplice. La differenza è rappresentata dal termine S_w^2 che sostituisce il termine S_h^2 presente nella varianza dello stimatore $\bar{y}_{ccs}$.
La stratificazione proporzionale è molto diffusa in quanto dà luogo a stimatori molto semplici e di precisione non inferiore a quella che si otterrebbe con un campione casuale semplice di identiche dimensioni. Altri tipi di stratificazione sono comunque frequentemente utilizzati.
Volendo massimizzare la precisione delle stime, tenuto conto delle risorse economiche disponibili, la frazione di campionamento negli strati dovrà essere stabilita in proporzione diretta alla variabilità (deviazione standard) elementare degli strati stessi (per la variabile di indagine) e in proporzione inversa alla radice quadrata del costo di rilevazione negli strati. Tale frazione di campionamento dà luogo alla ripartizione denominata ottimale ed è data da:

$$f_h \propto S_h \Big/ \sqrt{c_h}$$

dove c_h è il costo di osservazione di un'unità nello strato *h*. La formula esprime un risultato intuitivo: negli strati più eterogenei si deve applicare una frazione di campionamento maggiore di quella che si applica negli strati più omogenei, anche se è necessario tenere

conto delle eventuali differenze nel costo di rilevazione nei diversi strati. Un maggior costo di rilevazione implica una riduzione della frazione di campionamento. Se il costo di rilevazione non varia da strato a strato, la frazione di campionamento è semplicemente proporzionale al prodotto della dimensione dello strato per la deviazione standard:

$$f_h \propto S_h$$

All'atto pratico l'applicazione della ripartizione ottimale presuppone una qualche conoscenza sulla deviazione standard della variabile di studio nella popolazione. Un'approssimazione non troppo grossolana è spesso sufficiente per non vanificarne gli effetti. Diversamente da quanto avviene con la stratificazione proporzionale, infatti, con quella ottimale, l'uso di approssimazioni non adeguate per i termini S_h può portare ad una perdita in precisione rispetto al campionamento casuale semplice. Inoltre, dato che le variabili di indagine sono normalmente numerose, non è detto che la ripartizione ottimale per una o alcune lo sia anche per le altre.
Spesso il principale obiettivo che si persegue con la stratificazione è quello di ottenere stime di adeguata precisione per particolari sottopopolazioni, dette domini di studio, che vengono fatte coincidere con gli strati. Se un dominio e rappresentato da uno strato molto più piccolo rispetto agli altri è probabile che una stratificazione proporzionale non risulti adeguata a garantire al suo interno una sufficiente precisione degli stimatori. La soluzione consiste nell'applicare in quello strato una frazione di campionamento diversa (maggiore) dalle altre.
Una situazione analoga è quella nella quale si è interessati più a confrontare tra loro le stime dei vari strati che non a fonderle in un unico stimatore. Facendo riferimento per semplicità al caso di due soli strati, la distribuzione ottimale delle unità per la stima della differenza tra le medie di strato è:

$$\frac{n_1}{n_2} = \frac{S_1 / \sqrt{c_1}}{S_2 / \sqrt{c_2}}$$

Se, come spesso avviene, le varianze e i costi di rilevazione possono essere ipotizzati approssimativamente uguali nei due strati, la ripartizione ottimale per il confronto tra medie si riduce a:

$$n_1 \cong n_2$$

L'uguaglianza, sotto analoghe condizioni, vale in generale per un qualsiasi numero di strati.
Nel confronto tra strati la loro dimensione è dunque irrilevante, ma non lo è nel computo delle stime per l'intera popolazione. Ciò crea dei conflitti quando sia necessario perseguire ambedue gli obiettivi. Si pensi, ad esempio, a una popolazione suddivisa ancora in due soli strati nei quali la variabilità (per la variabile di indagine) sia approssimativamente uguale ma che abbiano dimensione notevolmente diversa: il primo strato contiene il 90% delle unità. La ripartizione ottimale, per la stima della media della popolazione, di un campione di 1000 unità, che corrisponde in questo caso anche a quella proporzionale, è di 900 unità

nel primo strato e 100 nel secondo; mentre quella ottimale per il confronto delle medie dei due strati è di 500 unità in ciascuno di essi. L'unico modo per uscire da questa situazione è quello di utilizzare una ripartizione intermedia che pur non essendo ottima per nessuno dei due obiettivi rappresenti un ragionevole compromesso delle inconciliabili esigenze dell'indagine.

8 SCELTA DEGLI STRATI

L'adozione di un campionamento stratificato è soggetta a due condizioni:
- deve essere nota la proporzione, W_h, di popolazione negli strati che si vogliono formare;
- ogni unità della popolazione deve essere attribuibile senza equivoci ad uno ed uno soltanto dei possibili strati.

Se si vogliono utilizzare stimatori corretti, un'ulteriore restrizione è rappresentata dalla necessità di selezionare almeno una unità da ogni strato. Se inoltre si vuole stimare correttamente da campione la varianza degli stimatori adottati è necessario selezionare da ogni strato almeno due unità.
Le ulteriori scelte dipendono dagli obiettivi della stratificazione e dalla quantità e qualità dell'informazione disponibile.
Quando il principale obiettivo della stratificazione è quello dell'efficienza degli stimatori, gli strati dovranno essere formati in modo da risultare più omogenei possibile al loro interno rispetto alla variabile di studio. Ciò richiede evidentemente la disponibilità di informazioni cioè di variabili che si assumono avere uno stretto legame con la variabile di indagine. Spesso tuttavia le variabili di indagine sono numerose ed è assai difficile che la stratificazione abbia efficacia per ciascuna di esse. Quando è necessario effettuare stime separate per categorie di unità, o domini, come li abbiamo denominati in precedenza, è senz'altro opportuno cercare di separarle in strati diversi. Se una (o alcune) di queste categorie ha dimensione molto piccola, sarebbe improbabile trovarle adeguatamente rappresentate nel campione senza il controllo campionario garantito dalla stratificazione.
Talvolta, inoltre, può essere utile isolare in strati distinti unità che per le loro caratteristiche o per il tipo di lista che le raccoglie devono o possono essere selezionate con metodologie diverse. Un esempio può essere utile per chiarire la problematica relativa alla scelta degli strati. Supponiamo di effettuare un'indagine sugli studenti universitari per stimare il tempo passato a guardare la televisione. Assumiamo di disporre per ciascuno studente, oltre all'informazione sull'anno di corso cui è iscritto, anche di quella sul sesso, sulla votazione finale conseguita uscendo dalla scuola media superiore (classificata come: alta media bassa) e sulla sua nazionalità (italiana o straniera).
Non esiste alcuna regola oggettiva su come utilizzare queste informazioni per formare gli strati. Se l'obiettivo è formare gruppi omogenei rispetto alla durata dell'ascolto televisivo, il ricercatore dovrà essere in grado di valutare soggettivamente la misura del legame di questa variabile con i fattori di stratificazione di cui dispone. Potrebbe ad esempio valutare irrilevante la relazione tra la variabile di indagine e la votazione al conseguimento della maturità; in tal caso dovrebbe evitare di utilizzare tale informazione nella stratificazione. Si deve ancora notare che gli strati risultanti utilizzando gli altri tre fattori: anno di corso (con 4 modalità), sesso (2 modalità) e nazionalità (2 modalità) non darebbero luogo

necessariamente ad un numero di strati pari al prodotto delle modalità di ciascuno di essi, cioè 16, se alcune combinazione dei criteri adottati si ritengono inefficaci a rendere più omogeneo l'ascolto televisivo. Così, si potrebbe pensare che la nazionalità discrimini la durata dell'ascolto solo per gli iscritti al primo anno che devono confrontarsi con i problemi dell'inserimento in un paese diverso da quello di origine. In tal caso si formerebbero soltanto 10 strati: 6 incrociando 3 anni di corso con le 2 modalità del sesso più 4 combinando al1 o anno di corso sesso e nazionalità.
In genere, non risulta conveniente formare un elevato numero di strati, sebbene sia intuitivo che il grado di omogeneità interna a ciascuno strato aumenta con la diminuzione di unità che contiene. l'aumento del numero degli strati per una data dimensione complessiva del campione, implica una riduzione della dimensione campionaria interna ad ogni strato e, conseguentemente, un aumento della variabilità delle stime.
Un altro problema che sorge con un numero elevato di strati, quando la stratificazione è proporzionale, è che alcune dimensioni campionarie non risultano uguali a numeri interi. l'arrotondamento all'intero più vicino non provoca effetti rilevanti sulla varianza quando tale numero è abbastanza grande. l'arrotondamento di piccoli numeri implica invece un allontanamento dalla proporzionalità e quindi anche un effetto sulla varianza delle stime. Un metodo molto diffuso per aggirare il problema è quello di sostituire la stratificazione esplicita con una stratificazione implicita: le unità della popolazione sono listate strato per strato; ma la selezione viene effettuata mediante un campionamento sistematico. In questo modo gli strati di dimensione maggiore avranno un numero di unità estratte - approssimativamente uguale a quello individuato da una stratificazione proporzionale mentre non è più garantita la presenza nel campione di unità provenienti da gli strati di piccole dimensioni. Per le caratteristiche della selezione sistematica (vedi paragrafo successivo) il disegno campionario sarà comunque equiprobabilistico.

9 CAMPIONAMENTO SISTEMATICO

Prima dell'avvento degli elaboratori e della loro rapida diffusione, l'estrazione di un campione casuale di grandi dimensioni poteva risultare estremamente laboriosa implicando, per ogni unità da estrarre, il ricorso alle tavole dei numeri casuali. Un metodo ideato per ridurre il lavoro sulle tavole e ancor oggi ancora molto utilizzato per la sua semplicità e la sua efficacia è rappresentato dal così detto campionamento sistematico che richiede di estrarre casualmente solo una (la prima) unità del campione. Il campione è infatti formato prendendo una unità ogni k presenti nella lista, a partire dalla prima estratta, con k pari al reciproco della frazione di campionamento.
Si supponga di dover estrarre un campione di 100 studenti da una lista di 1500 studenti. Il reciproco della frazione di campionamento, N/n, è uguale a 15. Per formare il campione è sufficiente selezionare un numero casuale compreso tra 1 e 15 (estremi inclusi) che individua la prima unità estratta e quindi procedere selezionando le altre unità con una progressione aritmetica di ragione 15 fino all'esaurimento della lista. Se, ad esempio, il primo numero estratto fosse 6, il campione risulterebbe formato dalle unità della lista contrassegnate dai numeri d'ordine:

6 6+15 6+2×15.....6+99×15

cioè,

6 21 36.....1491

Nell'esempio, volutamente molto semplice, la dimensione campionaria era tale da rendere il valore *k* intero. Nella pratica *k*, che prende il nome di ragione o intervallo di selezione, risulta spesso decimale. Se, ad esempio, la lista è composta da 1536 studenti, lo stesso campione di dimensione 100 dà luogo ad un valore di *k* pari a 15,36. In questa situazione è possibile arrotondare *k* all'intero inferiore o superiore a prezzo di un cambiamento nella dimensione campionaria. Con $k = 15$, infatti, si dovranno selezionare 102 o 103 unità per esaurire la lista, mentre con $k = 16$ non si potranno estrarre più di 96 unità. Se queste variazioni dimensionali sono accettabili il problema è risolto , altrimenti possono essere adottate diverse soluzioni alternative che tuttavia non prendiamo in esame in questa sede.
Nel campionamento sistematico, come in quello casuale semplice, ogni unità della popolazione ha la stessa probabilità di entrare a far parte del campione. La media della popolazione può quindi essere stimata ancora, come nel campione casuale semplice, per mezzo della media aritmetica semplice delle unità del campione:

$$\bar{y}_{sis} = \frac{1}{n}\sum_j y_j \quad ; \quad (j = 1,...,n)$$

Diversamente da quanto avviene nel campionamento casuale semplice, tuttavia, in quello sistematico non tutte le *n*-ple hanno la stessa probabilità di entrare a far parte del campione. Al contrario, fissato l'ordinamento della lista e stabilito di selezionare la prima unità tra le prime *k*, sono soltanto *k* le *n*-ple, cioè i possibili campioni, selezionabili. Ciascuna, ovviamente, con probabilità $1/k$ ($1/k = n/N$).
Ci si può comunque ricondurre ad una selezione del tutto equivalente a quella casuale semplice se si fa precedere l'estrazione da un'operazione che disponga le unità della lista in ordine casuale. Se questa operazione è parte integrante del processo di selezione allora il campionamento sistematico può essere assimilato in tutto e per tutto a quello casuale semplice.
L'operazione di ordinamento casuale della lista è però il più delle volte inopportuna poiché uno degli scopi della selezione sistematica può essere proprio quello di riuscire ad inserire nel campione unità con caratteristiche legate alla loro diversa posizione nella lista. Se, ad esempio, si tratta di unità ordinate in rapporto alla loro dimensione, dalla più piccola alla più grande, il campionamento sistematico assicurerà la presenza nel campione di unità di dimensioni piccole, medie e grandi in proporzione prossima a quella in cui sono presenti nella popolazione. Ciò potrebbe rispondere ad esigenze analoghe a quelle che ispirano la stratificazione. Infatti, è possibile pensare alle $k = N/n$ sottoliste nelle quali viene idealmente suddivisa la popolazione come a degli strati dai quali venga estratta una sola unità. Un'evidente analogia con il campionamento stratificato proporzionale rispetto al quale, tuttavia, verrebbe a mancare l'indipendenza tra le estrazioni nelle varie sottoliste.

Infatti, una volta determinata la posizione dell'unità da estrarre nella prima sottolista, sono automaticamente incluse nel campione le unità che hanno la stessa posizione nelle altre sottoliste.
Inoltre, senza che ciò contraddica le analogie precedentemente viste, il campione sistematico deve essere considerato come un caso particolare di campionamento a grappoli (vedi il successivo paragrafo), nel quale venga selezionato un solo grappolo. Il grappolo è un aggregato di unità elementari tra le quali esiste un qualche legame. Nel campione sistematico il legame è rappresentato dall'identica posizione che le unità estratte hanno all'interno delle sottoliste in cui viene suddivisa la lista della popolazione.

10 CAMPIONAMENTO A GRAPPOLI E A PIÙ STADI

In gran parte delle popolazioni oggetto di indagine le unità di studio sono raggruppate in sottopopolazioni di varia natura. la popolazione presente sul territorio italiano è la somma delle sottopopolazioni presenti sui territori regionali. All'interno di ciascuna regione, la popolazione è distribuita in province e, all'interno delle province, in comuni; nei comuni, infine, la popolazione è aggregata in famiglie. Gli studenti di un ateneo sono classificati in facoltà, quelli di una scuola, in classi, e così via dicendo.
Questi raggruppamenti di unità possono essere utilizzati come strati, come abbiamo illustrato nel paragrafo 6. Alternativamente, possono essere utilizzati come unità di selezione e in questo caso sono denominati grappoli. l'elenco dei grappoli forma la lista dalla quale viene estratto il campione. Se tutte le unità che appartengono ai grappoli estratti vengono incluse nel campione, il procedimento è detto campionamento a grappoli. Se nel campione vengono incluse solo alcune unità, selezionate da ciascuno dei grappoli estratti, il metodo è detto campionamento a due stadi o a più stadi.
Il numero di stadi dipende da quello dei livelli gerarchici di aggregazione delle unità che vengono individuati per effettuare la selezione. Un campione di italiani potrebbe essere estratto selezionando inizialmente alcune regioni, da ognuna di queste alcune province, da ciascuna provincia dei comuni, da questi delle famiglie e, infine, dalle famiglie, le persone che sono oggetto di studio.
Sebbene gli stessi aggregati di popolazione possano essere utilizzati come strati e come grappoli, gli scopi che si perseguono con la stratificazione sono profondamente diversi da quelli che si perseguono con la stadificazione. Gli strati devono o dovrebbero essere omogenei in quanto ognuno di essi è rappresentato nel campione. Al contrario, solo alcuni dei grappoli vengono selezionati, e questi devono rappresentare anche quelli esclusi dalla selezione. L'ideale sarebbe quindi che tutti i grappoli fossero più eterogenei possibile al loro interno e, conseguentemente, più simili possibile tra loro. Se, per fare un'ipotesi estrema, fossero tutti uguali, ciascuno sarebbe una copia ridotta della popolazione e sarebbe sufficiente selezionarne uno solo per avere la stessa informazione che si otterrebbe da un'indagine completa.
Purtroppo, come si intuisce da quanto fin qui osservato, i grappoli non vengono formati da chi estrae il campione, ma sono aggregazioni preesistenti nella popolazione, caratterizzate da una certa omogeneità interna che risulta generalmente tanto più marcata quanto minore è la loro dimensione. L'omogeneità che nella stratificazione è sinonimo di precisione degli

stimatori, nel campionamento a grappoli produce normalmente una perdita in precisione rispetto alla selezione casuale semplice.
La giustificazione del metodo sta negli aspetti pratici ed economici ad esso collegati. In primo luogo risulta spesso impossibile (economicamente o materialmente) formare una lista delle unità di studio, mentre può essere disponibile una lista di grappoli della popolazione. Inoltre, per una prestabilita dimensione campionaria, il campionamento a grappoli comporta costi generalmente molto inferiori a quelli del campionamento casuale semplice, in massima parte per la minore dispersione delle unità del campione.
Per chiarire con un esempio, si pensi alla formazione di un campione di individui a livello comunale. Se anziché selezionare singoli individui, si selezionano famiglie e si includono nel campione tutti i componenti delle famiglie stesse, il numero di spostamenti sul territorio per effettuare le interviste è ridotto in misura del numero medio di componenti delle famiglie, con conseguente riduzione dei costi. Ciò vale ovviamente per le indagini svolte mediante intervistatore mentre ha minore rilevanza, o non ne ha affatto, per quelle postali o telefoniche.
La riduzione dei costi si traduce all'atto pratico nella possibilità di selezionare campioni di dimensione assai superiore di quella che avrebbe avuto, per la stessa indagine, un campione casuale semplice. Le dimensioni dei campioni a grappoli o a più stadi sono normalmente tali da compensare la perdita in precisione indotta dal metodo di selezione.
La teoria della stima nel campionamento a grappoli e a più stadi è piuttosto complessa e non riteniamo opportuno inserirla in questa trattazione elementare. E' però importante osservare che alcune delle indagini di maggiore dimensione e rilevanza sono condotte con forme di campionamento complesso nelle quali entrano sia la stratificazione che la stadificazione .L'esempio più interessante è rappresentato proprio dalla maggiore indagine campionaria a livello nazionale: la rilevazione sulle forze di lavoro.