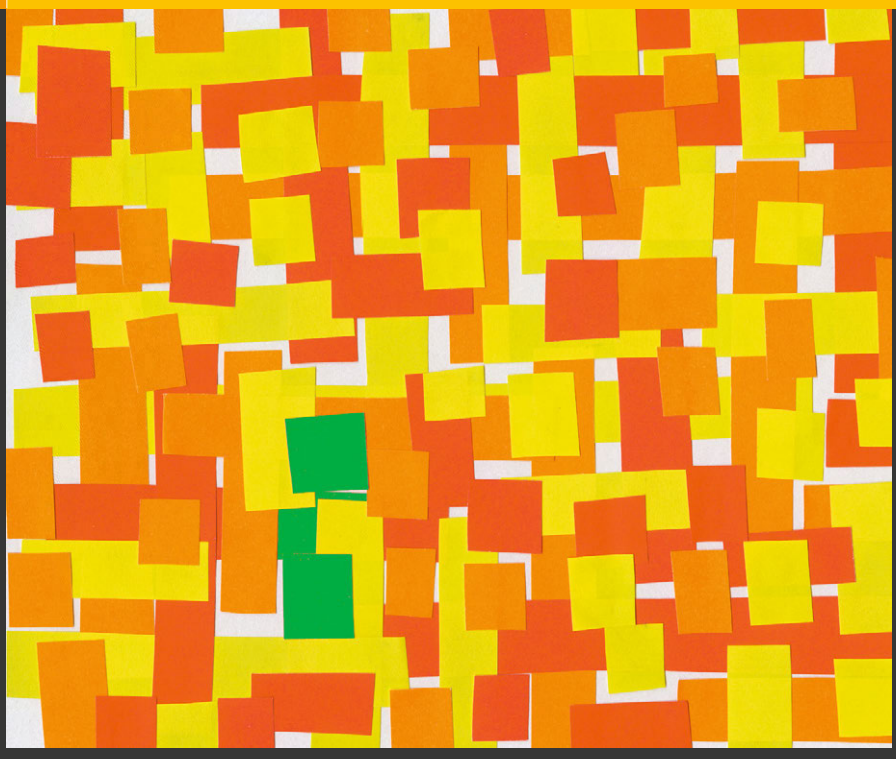
Giovanna Nicolini, Donata Marasini,
Giorgio Eduardo Montanari, Monica Pratesi,
Maria Giovanna Ranalli, Emilia Rocco

Metodi di stima in presenza di errori non campionari



UNITEXT

Springer

Metodi di stima in presenza di errori non campionari

Giovanna Nicolini · Donata Marasini ·
Giorgio Eduardo Montanari · Monica Pratesi ·
Maria Giovanna Ranalli · Emilia Rocco

Metodi di stima in presenza di errori non campionari

Springer

Giovanna Nicolini
Dipartimento di Economia,
Management e Metodi Quantitativi
Università degli Studi di Milano

Donata Marasini
Dipartimento di Economia,
Metodi Quantitativi e Strategie di Impresa
Università degli Studi di Milano-Bicocca

Giorgio Eduardo Montanari
Dipartimento di Economia,
Finanza e Statistica
Università degli Studi di Perugia

Monica Pratesi
Dipartimento di Statistica e Matematica
applicata all'Economia
Università degli Studi di Pisa

Maria Giovanna Ranalli
Dipartimento di Economia,
Finanza e Statistica
Università degli Studi di Perugia

Emilia Rocco
Dipartimento di Statistica
"G. Parenti"
Università degli Studi di Firenze

UNITEXT – Collana di Statistica e Probabilità Applicata
ISSN 2240-2640 ISSN 2240-2659 (elettronico)
ISBN 978-88-470-2795-4 ISBN 978-88-470-2796-1 (eBook)
DOI 10.1007/978-88-470-2796-1
Springer Milan Heidelberg New York Dordrecht London

9 8 7 6 5 4 3 2 1

Layout copertina: Beatrice Ч, Milano
Impaginazione: PTP-Berlin, Protago TEX-Production GmbH, Germany (www.ptp-berlin.eu)

Springer-Verlag Italia S.r.l., Via Decembrio 28, I-20137 Milano
Springer-Verlag fa parte di Springer Science+Business Media (www.springer.com)

Prefazione

Questo volume nasce dalla collaborazione di sei ricercatori, tutti di estrazione accademica e con un comune interesse nella metodologia delle indagini campionarie. Gli autori si sono trovati ad organizzare nel 2007 un corso dal titolo "Teoria e pratica delle indagini campionarie: approccio probabilistico ed errori non campionari" che si è svolto all'Università di Milano Bicocca nell'ambito dei corsi di alta formazione patrocinati dalla Società Italiana di Statistica. La teoria dei campioni rappresenta una parte della Statistica molto articolata e ogni autore ha apportato al corso un suo specifico contributo che, successivamente, si è trasformato in un argomento trattato nel presente libro.

La finalità di questo volume è quella di dare ai ricercatori gli strumenti necessari per correggere gli errori non campionari che inevitabilmente sorgono nelle diverse tipologie di indagine, mettendo in relazione tali errori anche con i più moderni metodi di condurre le indagini. Per come è stato impostato e per il rigore scientifico impiegato – che colloca nelle Dimostrazioni le parti metodologiche più sofisticate – il target di riferimento può essere sia accademico (corsi in lauree specialistiche, dottorati, master) sia non accademico (istituti di ricerca in generale). Il volume riporta argomenti che si trovano in letteratura, tuttavia il suo valore aggiunto si ritiene sia quello di averli unificati e di averli inseriti in un contesto moderno di indagine.

I primi due capitoli, pur non essendo dedicati agli errori non campionari, costituiscono un utile link con i metodi più tradizionali di campionamento e stima e unificano simboli e terminologie. Il Cap. 3 presenta una trattazione sistematica dell'utilizzo delle variabili ausiliarie nell'inferenza descrittiva su popolazioni finite; questo argomento non è facilmente reperibile in letteratura e costituisce la base necessaria per lo sviluppo dei capitoli successivi. Inoltre in questo capitolo si fa un accenno all'approccio basato sul modello, anch'esso indispensabile, per comprendere appieno i metodi proposti, e si analizzano gli effetti degli errori di misura, in modo da evidenziare la necessità della loro prevenzione, rinviando però ai testi specializzati i metodi per il loro trattamento.

I successivi tre capitoli sono dedicati al trattamento degli errori di copertura e di mancata osservazione e dei loro effetti sui risultati di una indagine statistica. Il Cap. 4 si occupa in particolare dei problemi connessi alla disponibilità di liste delle unità da campionare e alla loro capacità di rappresentare correttamente la popolazione indagata. Si analizzano le circostanze che portano ad errori e le loro ripercussioni sulla qualità delle stime ottenute e vengono suggeriti i metodi più utilizzati per prevenire e/o correggere gli errori di copertura, anche nei casi, sempre più frequenti, delle indagini via web. I Capp. 5 e 6 trattano il caso della mancata osservazione delle unità designate a far parte del campione, cioè la non riposta. Il Cap. 5 analizza gli effetti e presenta i metodi di correzione della non risposta totale basati principalmente sulla riponderazione dei dati e, tra questi, particolare attenzione è riservata ai metodi di calibrazione. Il Cap. 6 si occupa invece della non risposta parziale e delle tecniche di imputazione correntemente utilizzate per farvi fronte, senza tralasciare argomenti più avanzati frutto della ricerca più recente sul tema.

Al fine di agevolare la comprensione di quanto illustrato, in ogni capitolo sono stati inseriti diversi esempi, tratti a volte da indagini reali, per declinare nel caso considerato le metodologie descritte.

Tutti i capitoli sono stati ampiamente discussi dagli autori che hanno condiviso l'impostazione metodologica, tuttavia i Capp. 1 e 2 sono stati scritti da Donata Marasini e Giovanna Nicolini, il Cap. 3 da Giorgio E. Montanari, Monica Pratesi e Maria Giovanna Ranalli, il Cap. 4 da Giovanna Nicolini e Monica Pratesi, il Cap. 5 da Giorgio E. Montanari e Maria Giovanna Ranalli e, infine, il Cap. 6 da Emilia Rocco.

Milano, settembre 2012

Giovanna Nicolini
Donata Marasini
Giorgio Eduardo Montanari
Monica Pratesi
Maria Giovanna Ranalli
Emilia Rocco

Indice

Simboli

Popolazione	
U	Popolazione obiettivo (o target)
N	Numero di unità statistiche o elementari che formano la popolazione finita
F	Lista della popolazione (o popolazione frame)
M	Numero delle unità di rilevazione che formano la lista della popolazione
N_h	Numero di unità statistiche o elementari nella popolazione che formano lo strato h-esimo
$h = 1, \ldots, H$	Indice di strato
M_i	Numero di unità statistiche o elementari nella popolazione che formano il grappolo i-esimo
$i = 1, \ldots, M$	Indice di grappolo
N_l	Numero di unità statistiche o elementari nella popolazione che formano il post-strato l-esimo
$l = 1, \ldots, L$	Indice di post-strato
N_D	Numero di unità statistiche o elementari nella popolazione che formano il dominio D-esimo
N_F	Numero delle unità statistiche della popolazione target contenute nel frame
N_{F^*}	Numero delle unità statistiche non appartenenti alla target ma presenti nel frame
N_M	Numero delle unità statistiche nella popolazione target mancanti nel frame
W_M	Tasso di sotto-copertura
W_{F^*}	Tasso di sovra-copertura
N_R	Numero di unità statistiche o elementari nella sotto-popolazione dei rispondenti
N_{NR}	Numero di unità statistiche o elementari nella popolazione che formano la sottopopolazione dei non rispondenti

Variabili	
y	Variabile oggetto di studio
y_j	Valore della variabile y riferito alla j-esima unità della popolazione
x	Variabile ausiliaria
x_j	Valore della variabile x riferito alla j-esima unità della popolazione
$I_j(s)$	Variabile indicatrice di appartenenza ad s
$\boldsymbol{x} = (x_1, x_2, \ldots, x_P)$	Vettore di variabili ausiliarie
$\mathbf{x}_j$	Vettore dei valori di $\boldsymbol{x}$ nella j-esima unità
$\mathbf{y} = (y_1, y_2, \ldots, y_N)$	Vettore degli N valori di y nella popolazione
$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$	Vettore delle N variabili casuali Y_j associate alle unità della popolazione
ϕ	Variabile molteplicità
ϕ_j	Valore della molteplicità dell'unità statistica j-esima
N_ϕ	Frequenza assoluta della molteplicità con valore ϕ
Y_ϕ	Totale della variabile y sulle unità che hanno una molteplicità pari a ϕ
R_j	Variabile indicatrice di appartenenza all'insieme dei rispondenti r
$\mathbf{y}_\bullet = \{y_{\bullet j} : j \in r\}$	Insieme completato dei dati mediante un qualsiasi metodo di imputazione

Parametri	
T	Parametro della popolazione
Y	Totale nella popolazione della variabile y
$\overline{Y}$	Valore medio nella popolazione di y
A	Frequenza assoluta nella popolazione
P	Frequenza relativa o proporzione nella popolazione
C_y	Coefficiente di variazione nella popolazione di y
Q_q	Quantile di ordine q
$\Phi(y)$	Funzione di ripartizione nella popolazione della variabile y
X	Totale della variabile ausiliaria x
$\overline{X}$	Valore medio nella popolazione di x
$\mathbf{X} = (X_1, X_2, \ldots, X_P)$	Vettore dei totali nella popolazione di x
S_y^2	Varianza nella popolazione di y
S_x^2	Varianza nella popolazione di x
S_{yx}	Covarianza tra le variabili y e x
β_{yx}	Coefficiente di regressione
ρ_{yx}	Coefficiente di correlazione lineare
R	Rapporto tra totali o tra medie
θ	Vettore dei parametri di superpopolazione
$f(\cdot)$	Funzione di densità di una variabile casuale
$\varepsilon \quad \sigma^2$	Componente casuale del modello di superpopolazione e sua varianza
$\mu_j \quad \sigma_j^2 \quad \sigma_{jj'}$	Rispettivamente valore atteso e varianza di Y_j e covarianza tra Y_j e $Y_{j'}$ secondo il modello di superpopolazione

Campione	
s	Generico campione
n	Dimensione fissa di s
$n(s)$	Dimensione variabile di s
$f = n/N$	Frazione di sondaggio
S_0	Spazio dei campioni ordinati, con e senza reinserimento, di ampiezza variabile
S	Spazio dei campioni non ordinati e senza reinserimento di ampiezza variabile
$p(s)$	Piano di campionamento (pdc)
p_j	Probabilità di selezione della j-esima unità della popolazione
π_j	Probabilità di inclusione del primo ordine dell'unità j-esima
$\pi_{j,j'}$	Probabilità di inclusione del secondo ordine delle unità j e j'
$F_j(s)$	Frequenza attesa di inclusione
ψ_j	Numero medio di volte in cui l'etichetta j è compresa in s
$a_j = 1/\pi_j$	Peso base di riporto all'universo
$a_{j,j'} = 1/\pi_{j,j'}$	Reciproco della probabilità di inclusione del secondo ordine
π_k	Probabilità di inclusione del primo ordine del network k
π_{kl}	Probabilità di inclusione del secondo ordine dei network k e l
r	Campione dei rispondenti
n_r	Dimensione del campione dei rispondenti
$q(r\|s)$	Processo di selezione dei rispondenti dal campione s
ϑ_j	Probabilità di risposta dell'unità j-esima
r_c	Insieme delle unità completamente osservate
r_y	Insieme delle unità che hanno fornito un valore per la variabile di interesse y
n_{r_y}	Dimensione insieme delle unità che hanno fornito un valore per la variabile di interesse y

Piani di Campionamento	
srs	pdc casuale semplice senza reinserimento
srswr	pdc casuale semplice con reinserimento
sm	pdc sistematico
Po	pdc di Poisson
pps	pdc con reinserimento proporzionale alla misura di ampiezza
πps	pdc senza reinserimento proporzionale alla misura di ampiezza
st	pdc stratificato
gr	pdc a grappoli
ds	pdc a due stadi

Stimatori	
$\hat{T}$	Stimatore generico del parametro T
$\hat{Y}_\pi$,	Stimatore o stima di Horvitz-Thompson (HT) del totale Y
$\hat{X}_\pi$	Stimatore o stima di HT del totale X
$\hat{R}_\pi$	Stimatore o stima di HT del rapporto R
$\hat{\overline{Y}}_\pi, \hat{\overline{X}}_\pi$	Stimatore o stima di HT di $\overline{Y}$ e di $\overline{X}$
$\hat{Y}, \hat{X}, \hat{A}$	Stimatore di HT di Y, di X e di A per un *srs*
$\overline{y}, \overline{x}$	Media campionaria di y e di x.
$\hat{Y}_{HH}$	Stimatore o stima di Hansen-Hurwitz
$\hat{Y}_{PS}$, $\hat{Y}_{PSS}$	Stimatore post-stratificato in presenza di mancate risposte
$\hat{Y}_{\pi st}$	Stimatore o stima di HT del totale Y per un *st*
$\hat{Y}_{\pi gr}$	Stimatore o stima di HT del totale Y per un *gr*
$\hat{Y}_{\pi ds}$	Stimatore o stima di HT del totale Y per un *ds*
$\hat{Y}_{2f}$	Stimatore nel campionamento a due fasi
$\hat{Q}_q$	Stimatore o stima di un quantile
$\hat{Y}_q$,	Stimatore o stima per quoziente del totale Y
$\hat{Y}_{ps}$	Stimatore o stima post-stratificata del totale Y
$\hat{Y}_d$	Stimatore o stima per differenza del totale Y
$\hat{Y}_{reg}$	Stimatore o stima per regressione generalizzata del totale Y
$\hat{Y}_{cal}$	Stimatore o stima di calibrazione del totale Y
$\hat{Y}_\xi$	Stimatore basato sul modello del totale Y
$\hat{Y}_{Ha}$	Stimatore di Hartley
$\hat{Y}_M$	Stimatore del totale in presenza di molteplicità
$\hat{Y}_H$	Stimatore di Hajek
$\hat{T}_\bullet$	Stimatore del parametro T applicato all'insieme completato dei dati

Operatori	
$\mathrm{B}(\cdot)$	Distorsione rispetto al piano di campionamento
$\mathrm{E}(\cdot)$	Valore atteso rispetto al piano di campionamento
$\mathrm{V}(\cdot)$	Varianza rispetto al piano di campionamento
$\mathrm{C}(\cdot,\cdot)$	Covarianza rispetto al piano di campionamento
$\mathrm{v}(\cdot)$	Stima della varianza
$\mathrm{c}(\cdot,\cdot)$	Stima della covarianza
Deff	Effetto del disegno
Eff	Efficienza relativa
MSE	Errore quadratico medio
ξ	Modello di superpopolazione
$\mathrm{E}_\xi(\cdot)$	Valore atteso rispetto al modello di superpopolazione
$\mathrm{V}_\xi(\cdot)$	Varianza rispetto al modello di superpopolazione
$\mathrm{V}_{p\xi}(\cdot)$	Varianza congiunta rispetto al piano di campionamento ed al modello di superpopolazione
$\mathrm{C}_\xi(\cdot,\cdot)$	Covarianza rispetto al modello di superpopolazione
$\mathrm{E}_m(\cdot)$	Valore atteso rispetto al modello di misura
$\mathrm{V}_m(\cdot)$	Varianza rispetto al modello di misura

$\mathrm{V}_{pm}(\cdot)$	Varianza congiunta rispetto al piano di campionamento e al modello di misura
$\mathrm{E}_{q}(\cdot\|s)$	Valore atteso rispetto al processo di risposta, condizionato al campione estratto
$\mathrm{V}_{q}(\cdot\|s)$	Varianza rispetto al processo di risposta, condizionato al campione estratto
$\mathrm{V}_{pq}(\cdot)$	Varianza congiunta rispetto al piano di campionamento ed al processo di risposta
$\mathrm{v_q}(\cdot\|s)$	Stima della varianza rispetto al processo di risposta, condizionato al campione estratto
E_{MI}	Valore atteso rispetto a un modello di imputazione
V_{MI}	Varianza rispetto a un modello di imputazione
$\mathrm{v_{JK}}(\cdot)$	Stima Jackknife della varianza
$\mathrm{v_{bo}}(\cdot)$	Stima bootstrap della varianza

1

Introduzione al campionamento da popolazioni finite

1.1 Introduzione

Lo scopo introduttivo di questo capitolo è quello di richiamare i concetti fondamentali del campionamento da popolazioni finite. Dopo alcune definizioni preliminari e una breve disquisizione sulla terminologia in uso, vengono ricordate le fasi di una indagine statistica e individuati gli errori che in ciascuna di esse possono aver origine. Come si vedrà nei successivi capitoli di questo volume, dedicati proprio alle diverse tipologie di errore, l'impiego delle variabili ausiliarie per la individuazione e la correzione degli errori è fondamentale. Tuttavia, come è noto, ad esse si può ricorrere anche per la costruzione del piano di campionamento. Questo non implica la correzione di alcun errore, ma una maggiore precisione ed aderenza alla realtà che si vuole indagare e il ricorso alle variabili ausiliarie in questo caso può essere inteso come un metodo preventivo degli errori. Non sempre è possibile disporre delle variabili ausiliarie per i diversi impieghi, può allora essere necessario costruire ad hoc un data-set di tali variabili, per esempio con il campionamento a due fasi, che viene brevemente richiamato. Infine, vengono ricordati alcuni metodi di indagine che non seguono i canoni tradizionali del campionamento da popolazioni finite – sono i così detti *campionamenti non probabilistici* – alcuni molto in uso nelle indagini in ambito sociale. Anche se per essi non è possibile conoscere l'errore campionario, tuttavia, in questi casi, si può ricorrere a metodi inferenziali diversi da quello tradizionale usato nel campionamento da popolazioni finite, noto come *design-based*.

1.2 Indagine campionaria

Si supponga di voler studiare una variabile y su una popolazione finita formata da N unità statistiche, con lo scopo di conoscerne il totale, il suo valor medio o una qualsiasi altra funzione dei dati che ne definisce un *parametro* T. Se si effettua un'*indagine censuaria* tutte le unità della popolazione vengono

G. Nicolini et al., *Metodi di stima in presenza di errori non campionari*,
UNITEXT – Collana di Statistica e Probabilità Applicata,
DOI 10.1007/978-88-470-2796-1_1,

osservate e, nell'ipotesi che non si verifichi alcun tipo di errore nella rilevazione dei dati, si determina il reale valore del parametro di interesse T. Se invece si effettua una *indagine campionaria* solo una parte delle unità della popolazione viene osservata; ne consegue che, anche in assenza di errori di rilevazione, il valore reale del parametro è ignoto, tuttavia può essere stimato attraverso i valori delle unità che costituiscono il campione.

La non conoscenza del reale valore del parametro comporta la necessità di una sua stima che deve essere la "migliore"; questa esigenza pone il problema dell'attendibilità della stima e della scelta delle tecniche di campionamento. La presenza di anomalie nella costruzione del piano di campionamento e/o della stima allontanano quest'ultima dal valore reale del parametro. Pertanto, in un'indagine concepita secondo la tradizione classica (basata sul disegno), la conoscenza e la identificabilità delle unità della popolazione di riferimento sono fondamentali, come fondamentali sono alcuni presupposti logici presentati nel seguito.

1.2.1 Definizioni preliminari

Il campionamento da popolazione finita si basa su presupposti logici che sono ormai noti in letteratura (Hansen et al. 1953; Cochran 1977; Kish 1965, 1987; Särndal et al. 1992; Frosini et al. 2011); tuttavia, qui di seguito, se ne vogliono richiamare alcuni ai quali si farà riferimento nel corso del presente volume:

- *Popolazione obiettivo* o *popolazione target.* È la popolazione formata da un numero finito N di *unità statistiche* identificabili, su cui si analizzano la variabile o le variabili oggetto dell'indagine. Se ad ogni unità si associa un numero intero da 1 a N, tale popolazione viene indicata con $U = \{1, \cdots, j, \cdots, N\}$.
- *Lista della popolazione* o *popolazione frame.* È l'elenco delle M *unità di rilevazione* e viene indicato con $F = \{F_1, \cdots, F_i, \cdots, F_M\}$. L'unità di rilevazione può coincidere con l'unità statistica – in tal caso $N = M$ – ovvero è un grappolo di unità statistiche e pertanto ogni unità di rilevazione F_i è una subpopolazione composta da M_i unità statistiche, il cui totale è $\sum_{i=1}^{M} M_i = N$. Ad esempio, nel primo caso la popolazione target è formata dagli individui residenti in un comune e la popolazione frame ne è l'elenco anagrafico; nel secondo caso la popolazione target è ancora formata dagli individui residenti e quella frame è l'elenco dalle famiglie residenti in quel comune. Rientra nel secondo caso il così detto *frame areale*, le cui unità di rilevazione sono le *areole* che rappresentano parti ben definite sul territorio di riferimento quali, ad esempio, i quartieri di una città, le sezioni di censimento o anche parti di territorio definite ad hoc. Ogni areola contiene un numero differente di unità statistiche che possono essere, a seconda dei casi, famiglie, individui, aziende agricole, unità commerciali, ecc. (per una analisi più approfondita delle relazioni tra popolazioni frame e target si veda il Cap. 4).

- *Parametri della popolazione.* Indicato con $y_j (j = 1, \ldots, N)$ il valore della variabile in esame y riferito alla j-esima unità della popolazione, si definisce parametro una qualsiasi funzione degli N valori: $T = f(y_1, \cdots, y_N)$. Numerosi sono i parametri di una variabile, nel seguito vengono riportati quelli descrittivi più comunemente usati. Nel caso di una variabile quantitativa, il parametro di maggior interesse è il *totale*

$$Y = \sum_{j=1}^{N} y_j, \tag{1.1}$$

o anche il *valor medio* che è dato dal rapporto tra totale e dimensione della popolazione:

$$\overline{Y} = \frac{Y}{N}. \tag{1.2}$$

Se invece la variabile è qualitativa e dicotoma, i due parametri più utilizzati sono ancora il totale e il valor medio che, tuttavia, in questo contesto, equivalgono rispettivamente ad una *frequenza assoluta* A e ad una frequenza relativa o *proporzione* P. Infatti, per un fenomeno qualitativo, indicata con C la modalità di interesse e con $\overline{C}$ la modalità associata a tutto ciò che ne è complementare, se l'unità j possiede C si conviene di porre $y_j = 1$, mentre se possiede $\overline{C}$, si pone $y_j = 0$. Ne consegue che il numero di unità della popolazione con la modalità C è pari a

$$A = \sum_{j=1}^{N} y_j = Y \tag{1.3}$$

e rappresenta una frequenza assoluta. Dividendo la (1.3) per la dimensione della popolazione si ottiene una frequenza relativa, cioè la proporzione delle unità della popolazione con la caratteristica di interesse

$$P = \frac{A}{N}. \tag{1.4}$$

Parametri che identificano posizioni particolari sulla distribuzione della variabile sono i *quantili*. Indicata con $\Phi(y)$ la funzione di ripartizione della variabile y, sempre che la medesima possa essere ordinata, il *quantile di ordine* q $(0 < q < 1)$, in termini operativi, si può definire nel modo seguente:

$$Q_q = \inf\{y : \Phi(y) \geq q\}. \tag{1.5}$$

Un particolare quantile molto utilizzato per la sintesi di distribuzioni asimmetriche è la *mediana* che si ha per $q = 0,5$. Un altro parametro di interesse per una variabile quantitativa è la *varianza*, che esprime la variabilità della

variabile y nella popolazione, data in questo contesto da (Cochran 1997)[1]

$$S_y^2 = \frac{1}{N-1}\sum_{j=1}^{N}(y_j - \overline{Y})^2 \tag{1.6}$$

e la sua radice quadrata S_y, nota come scarto quadratico medio. Un parametro molto usato per confrontare la variabilità è il *coefficiente di variazione* dato dal rapporto

$$C_y = \frac{S_y}{\overline{Y}}. \tag{1.7}$$

Oltre ai parametri funzioni di una sola variabile potrebbero interessare anche quelli che mettono in evidenza la relazione tra la variabile y e una variabile x presente sulle unità della popolazione, come il *rapporto* tra totali o tra medie

$$R = \frac{Y}{X} = \frac{\overline{Y}}{\overline{X}}, \tag{1.8}$$

con $\sum_{j=1}^{N} x_j = X$ totale della variabile x ; il *coefficiente di regressione*

$$\beta_{yx} = \frac{S_{yx}}{S_x^2}, \tag{1.9}$$

dove S_x^2è la varianza della variabile x e S_{yx} è la covarianza tra le due variabili

$$S_{yx} = \frac{1}{N-1}\sum_{j=1}^{N}(y_j - \overline{Y})(x_j - \overline{X}) \tag{1.10}$$

e il *coefficiente di correlazione lineare*

$$\rho_{yx} = \frac{S_{yx}}{S_y S_x}. \tag{1.11}$$

Vale la pena notare che i parametri descritti possono essere interpretati come funzioni di uno o più totali relativi alla sola variabile y o alla coppia (y, x). Infatti, è immediato osservare che i parametri (1.2) e (1.4) sono funzioni del solo totale Y, mentre è facile verificare che il parametro (1.6) è funzione dei totali $\sum_{j=1}^{N} y_j^2$ e Y potendosi scrivere

$$S_y^2 = \frac{1}{N-1}\sum_{j=1}^{N} y_j^2 - \frac{1}{N(N-1)}Y^2. \tag{1.12}$$

I parametri (1.8), (1.9), (1.10) e (1.11) chiamano invece in causa totali relativi anche alla variabile x e alla coppia (y, x).

[1] Conviene notare che nel presente contesto i parametri (1.6) e (1.10) sono rapportati alla quantità $N-1$, anziché a N come è naturale, esclusivamente per motivi di semplicità quando si passa alla stima dei medesimi.

1.2.2 Campione casuale, probabilistico, rappresentativo

Campione *casuale*, *probabilistico*, *rappresentativo*: sono tre concetti fondamentali nel campionamento da popolazioni finite che non necessariamente devono coesistere, avendo ciascuno la sua specificità.

La casualità riguarda il modo in cui sono selezionate le unità della popolazione destinate a formare il campione; pertanto la selezione deve avvenire seguendo una qualsiasi procedura riconducibile ad un esperimento aleatorio. Se la dimensione N della popolazione è nota, l'estrazione casuale è assimilabile all'estrazione di palline da un'urna che contiene appunto N palline; se invece la dimensione della popolazione non è nota, come ad esempio possono essere i clienti di un grosso emporio di abbigliamento, è sempre possibile pensare ad un meccanismo di casualità che selezioni alcune unità; ad esempio si può pensare di intervistare un cliente che entra nell'emporio ogni m ingressi.

Un campione è probabilistico quando la probabilità che ciascuna unità della popolazione ha di farne parte, in base al meccanismo di selezione, è positiva e calcolabile, secondo la definizione fornita nel prossimo capitolo. In questo senso un campione probabilistico è anche casuale.

La rappresentatività di un campione, invece, è un concetto più articolato al quale si possono ricondurre più definizioni. Kruskall e Mosteller (1979) hanno individuato sei definizioni di rappresentatività presenti nella letteratura statistica, alle quali ne sono state aggiunte dagli autori altre tre. Per definire la rappresentatività occorre individuare il contesto di riferimento. Una accezione molto diffusa di rappresentatività del campione è che esso sia una "miniatura" della popolazione stessa. Ovviamente è utopistico pensare di raggiungere questo obiettivo per tutte le variabili, più realisticamente ciò avviene solo per alcune variabili note (ad esempio il genere e l'età degli individui). Naturalmente l'aspettativa è quella di realizzare in tutto o in parte la rappresentatività di un campione anche rispetto alle variabili di indagine. Tale aspettativa dipende però dal legame tra le variabili rispetto alle quali il campione è rappresentativo e quelle oggetto di indagine e non sempre viene realizzata.

Da quanto detto discende che un campione casuale può essere probabilistico e non necessariamente rappresentativo. Viceversa l'essere probabilistico per un campione implica la casualità ma non la rappresentatività, mentre un campione rappresentativo può essere o non essere casuale e/o probabilistico. Il ricercatore sa se il campione è casuale e probabilistico, mentre non sa se è rappresentativo, se non facendo riferimento a esperienze passate o a particolari verifiche a posteriori, ad esempio con l'analisi delle sub-popolazioni, come si avrà occasione di esaminare nel Cap. 3. L'assenza dei primi due attributi implica decisioni sulla scelta dell'approccio inferenziale, come emergerà nel Cap. 3, mentre l'assenza di rappresentatività[2] potrebbe indicare la presenza

[2] Negli ultimi anni è stato sviluppato in letteratura un nuovo concetto di rappresentatività legato al contesto di qualità dell'indagine. È opinione diffusa che non è il tasso di non risposta a rendere le stime distorte quanto le differenze tra rispondenti e non nei confronti della variabile di interesse (J. Bethlehem et al.

degli errori non campionari, che devono quindi essere individuati e corretti con le metodologie proposte nei Capp. 4, 5 e 6.

1.2.3 Fasi di un'indagine campionaria

L'indagine statistica, sia essa censuaria o campionaria, è una procedura complessa che si articola in una successione di fasi di seguito elencate:

1. *Formalizzazione degli obiettivi, individuazione della popolazione obiettivo, definizione spazio-temporale.* Formalizzare gli obiettivi significa individuare le variabili che verranno osservate nella popolazione di riferimento. Il numero di variabili oggetto della ricerca dipende dalla complessità degli obiettivi: un'indagine sulla lettura dei quotidiani è molto più semplice dell'indagine sul tempo libero, che coinvolge un più ampio numero di variabili rispetto alla prima. La definizione già fornita di popolazione target deve essere circoscritta nel tempo e nello spazio. Ad esempio, nell'indagine sulla lettura dei quotidiani la popolazione obiettivo potrebbe essere costituita dai residenti con più di 15 anni (i residenti con meno di 15 anni potrebbero essere scarsamente interessati alla lettura dei quotidiani) in una definita area territoriale (comune, provincia, ecc.) in un fissato arco temporale (settimana, mese, anno, ecc.).
2. *Costruzione del frame.* Il frame deve essere costruito con cura in modo che contenga tutte le unità della popolazione target. Ad ogni unità del frame corrisponde una etichetta che identifica l'unità di rilevazione, il cui ordinamento può essere casuale o non casuale. Ad esempio, l'ufficio del personale di una azienda può costruire un frame dei suoi dipendenti utilizzando l'ordine alfabetico del cognome (in questo caso l'ordinamento delle unità della lista può ritenersi casuale rispetto ad una usuale variabile d'indagine) ovvero può elencare i dipendenti in funzione della data di assunzione (in questo caso l'ordinamento non è casuale, ma in funzione del tempo, e può essere un vantaggio se la variabile oggetto d'indagine è sensibile a tale ordinamento come, ad esempio, l'avanzamento di carriera o gli scatti di stipendio). La costruzione del frame implica anche l'inserimento, per ogni unità, di tutte le informazioni ausiliarie che possono essere utili per la estrazione del campione o la costruzione dello stimatore; ad esempio, per quanto riguarda l'elenco dei dipendenti tali informazioni potrebbero essere: il genere, il titolo di studio, il livello attribuito al momento dell'assunzione, il livello attuale, lo stipendio, ecc.

2008; 7mo Programma Quadro dell'UE e i Work Package 1, 2, 3, 4, 5, 6 a cura di N. Shlomo et al. 2009). Pertanto, il campione dei rispondenti potrebbe essere ancora rappresentativo del campione programmato e le stime calcolate su tale campione potrebbero essere ancora realistiche. Per sostenere questa tesi occorre quindi individuare degli indicatori di rappresentatività del campione osservato, che sono stati chiamati R-indicators.

3. *Scelta del piano di campionamento e dello schema di estrazione.* Piano di campionamento e schema di estrazione vengono descritti nel Par. 2.2. Tuttavia, lo schema, proprio perché indica il modo con cui le unità vengono selezionate, si ritiene fondamentale per definire il campione probabilistico o meno. Ad esempio, nel piano di campionamento stratificato la popolazione viene suddivisa, come è noto, in sottoinsiemi (strati) omogenei e disgiunti, per ciascuno viene costruito il frame e individuato uno schema di estrazione, in base al quale in ogni strato viene selezionato un campione. Se il metodo di scelta prevede un esperimento casuale con probabilità di selezione costante per ogni unità in ognuno degli strati, si ha un campione probabilistico e l'unione di questi campioni definisce il *piano di campionamento casuale stratificato* (si veda Par. 2.5.7). Al contrario se non è previsto alcun esperimento casuale per la selezione, che viene demandata all'intervistatore, il campione di strato è non casuale e non probabilistico e l'unione di tali campioni porta al *campionamento per quote* (si veda Par. 1.4).
4. *Rilevazione dei dati.* I metodi di rilevazione o raccolta dei dati sono diversi e vengono scelti in relazione ai differenti contesti in cui l'indagine si svolge. Ad esempio, nelle indagini sugli individui e le famiglie, la raccolta dati richiede spesso l'intervento di un intervistatore. In questo ruolo, il rilevatore ha il compito di contattare il possibile rispondente e di somministrare il questionario. Le modalità di rilevazione possono essere diverse in relazione alle risorse disponibili per l'indagine, ai tempi richiesti per la sua realizzazione e alla natura della popolazione obiettivo da studiare. La rilevazione può essere effettuata con interviste faccia a faccia compilando un questionario cartaceo (modalità PAPI – *Paper and Pencil Interviewing*) oppure un questionario elettronico (modalità CAPI – *Computer Assisted Personal Interviewing*), con interviste telefoniche (modalità CATI – *Computer Assisted Telephone Interviewing*) oppure via Internet, proponendo al soggetto l'auto-compilazione del questionario via Web (modalità CAWI – *Computer Assisted Web Interviewing*). In taluni casi l'auto compilazione può essere proposta anche inviando il questionario cartaceo via posta ordinaria.
5. *Analisi dei dati.* I dati campionari vengono utilizzati in una logica inferenziale per stimare i parametri delle variabili di interesse nella popolazione target. In un'indagine ideale, cioè senza alcun tipo di errore, viene utilizzato prevalentemente l'approccio inferenziale *design-based* (Par. 3.1). Tuttavia l'indagine senza alcun tipo di errore difficilmente si realizza; gli argomenti dei capitoli che seguono sono rivolti proprio ai metodi inferenziali per le indagini in presenza di errori. Comunque, una regola che deve essere sempre rispettata nell'analisi delle variabili oggetto di studio è quella di considerare, congiuntamente alle stime, l'errore quadratico medio, l'errore standard, nonché, quando possibile, l'intervallo di confidenza.
6. *Diffusione dei risultati.* Una volta raccolti i dati ed effettuate le analisi, inizia la fase di pubblicizzazione e diffusione dei risultati, che avviene principalmente tramite la diffusione del cosiddetto rapporto di indagine. Esso contiene la descrizione delle fasi d'indagine e l'illustrazione dei principali

risultati, con attenzione alla rappresentatività del campione e all'attendibilità delle stime. In tale fase, il soggetto responsabile dell'indagine svolge un'attività di promozione, diffusione e trasferimento dei risultati descritti nel rapporto con un programma di iniziative finalizzato a raggiungere tutti gli interessati.

Si fa notare che nell'indagine *censuaria* non è prevista la fase 3; mentre in quella c*ampionaria* le fasi 5 e 6 sono presenti con una logica diversa rispetto a quella dell'indagine globale.

1.2.4 Gli errori di un'indagine campionaria

Nelle fasi dell'indagine possono verificarsi delle imperfezioni, alcune con conseguenze molto gravi, altre meno, che in entrambi i casi richiedono interventi correttivi. In generale le imperfezioni determinano delle divergenze, chiamate comunemente *errori*, tra quello che "si sarebbe teoricamente dovuto osservare" e quello che "realmente è stato osservato". Diverse sono le tipologie di errore (Lessler e Kalsbeek 1992), tuttavia ciò che mina la qualità della ricerca non è il tipo di errore ma l'entità del medesimo. L'inferenza statistica insegna che gli errori di piccole dimensioni sono accettabili, non lo sono se di elevate dimensioni. La linea di demarcazione dell'accettabilità non può essere individuata in modo oggettivo, in quanto varia in funzione di diversi fattori come l'obiettivo dell'indagine, gli strumenti di rilevazione utilizzati, il livello di precisione richiesto, i tempi e i costi imposti; ed è per questo che tale linea spesso viene suggerita dall'esperienza e dalla sensibilità del ricercatore.

In genere, nelle indagini si è consapevoli della presenza dell'errore e della sua origine e la letteratura propone diversi metodi di correzione. Nel seguito sono brevemente elencate (mantenendo la stessa numerazione delle fasi) le anomalie che si possono verificare nelle fasi dell'indagine con le loro conseguenze, rinviando ai capitoli successivi i metodi di correzione:

1. La popolazione obiettivo non è specificata, obiettivi non chiari, variabili non coerenti con gli obiettivi: sono errori molto gravi con pesanti conseguenze perché potrebbero portare ad invalidare la ricerca. Non esistono metodi generali di correzione.
2. La non corrispondenza tra la popolazione frame e la popolazione target genera gli *errori di copertura*, che saranno trattati nel Cap. 4.
3. La scelta di procedere alla selezione di un campione implica a priori la rinuncia a conoscere il reale valore del parametro T e ad accettarne una stima $\hat{T}$. Conseguenza di tale scelta è la presenza dell'*errore campionario*, definito come la differenza tra la stima e il reale valore. Per ridurre questo errore si può incrementare la dimensione del campione e/o scegliere un piano di campionamento più efficiente di un altro (si veda Par. 2.7). Nell'ambito dell'inferenza *design-based* è possibile conoscere la dimensione dell'errore campionario solo se il campione è probabilistico.

4. In fase di rilevazione sono frequenti gli *errori di non risposta* che consistono nella mancata partecipazione all'indagine o nel non rispondere ad alcune domande del questionario. Nel primo caso la non risposta è *globale* o *totale*, nel secondo è *parziale*; tali errori saranno ampiamente trattati rispettivamente nei Capp. 5 e 6. La non risposta dipende da molti fattori tra cui un senso di diffidenza nei confronti delle indagini, l'obiettivo della ricerca che può coinvolgere variabili sensibili, il metodo di rilevazione e la comprensione del questionario. In questa fase si possono osservare anche gli *errori di misurazione*, (differenze tra il valore reale della variabile y e il valore rilevato); di questi alcuni sono imputabili all'intervistatore, altri all'intervistato e altri ancora all'imputazione o alla codifica dei dati.
5. È questa la fase più compromessa dalla presenza degli errori di cui sopra. Infatti, l'analisi dei dati di un'indagine campionaria riguarda le stime che vengono calcolate per i parametri descrittivi delle variabili di interesse, per quelli che ne esprimono le relazioni o per l'analisi delle variabili stesse. È evidente che la presenza degli errori provoca allontanamenti incontrollati delle stime dai rispettivi parametri.
6. L'errore che si commette in questa fase consiste nella non esatta descrizione dell'indagine e dei suoi risultati.

Gli errori descritti possono essere classificati in modi diversi; ad esempio:

- Se si vuole evidenziare l'errore della stima, si distingue l'*errore campionario* dagli *errori non campionari*. Il primo è presente solo nelle indagini campionarie e, come già detto, può essere tenuto sotto controllo in vari modi; i secondi sono gli errori di copertura, di non risposta e di misura che sono presenti nelle indagini campionarie ma anche in quelle censuarie; inoltre, in queste ultime la loro presenza è più rilevante in quanto tali errori aumentano con l'aumentare del numero delle osservazioni. Nei capitoli seguenti saranno proposti i metodi per correggere gli errori di copertura e di non risposta (globale e parziale), mentre non saranno considerati i metodi per correggere gli errori di misura, in quanto nella pratica delle indagini campionarie sono scarsamente utilizzati. Tuttavia, poiché questi errori possono distorcere anche sensibilmente le stime, alla fine del Cap. 3 vengono analizzati gli effetti che producono, per i quali si suggerisce la prevenzione come unico approccio valido di contrasto.
- Se si vogliono distinguere gli errori in base al fatto di avere o meno osservato le unità statistiche oggetto dell'indagine, si distinguono gli errori da *non osservazione* da quelli da *osservazione* (Groves 1989). I primi nascono perché non sono state osservate tutte le unità della popolazione, sono tali l'errore campionario, quello di sotto-copertura e di non risposta; i secondi sono tutti gli errori di misurazione che possono essere generati da chi, a vario titolo, si trova coinvolto nell'indagine, come ad esempio gli intervistatori e gli intervistati, o gli errori generati dal metodo di raccolta dei dati o dalla costruzione del database, ecc.