INDICE

PREFAZIONE

La validità di un'indagine empirica si basa essenzialmente sulla bontà della teoria che la sostiene, ovvero sulle modalità di individuazione e messa a punto dei problemi, di scelta ed uso delle tecniche più idonee. Ben prima di impegnarsi nella raccolta dei dati, occorre quindi interrogarsi attentamente sugli scopi che si vogliono raggiungere e sugli itinerari più appropriati. Su questi punti non sempre si riflette abbastanza, un po' perché spesso si ha la necessità di acquisire quanto prima dei risultati, quali che siano, un po' perché non si è sufficientemente sensibilizzati in proposito.

Però, chi si occupa di metodologia – in qualità di studioso, di professionista impegnato nelle scienze sociali o di discente – non può esimersi dal realizzare una costante interazione fra teoria e tecniche.

Per decidere quanto ampio debba essere, ad esempio, un campione di consumatori di un certo bene o di persone di cui si vogliano conoscere gli atteggiamenti, occorre prima rispondere ad alcuni quesiti preliminari sulla configurazione della popolazione, sui mezzi e sugli strumenti disponibili per la ricerca empirica, sull'arco temporale della realizzazione, sull'affidabilità ammissibile dei dati. È proprio per sottolineare questa necessità, tante volte rilevata e approfondita durante vari corsi universitari di metodologia e tecniche d'indagine nelle scienze sociali, così come nella cura delle tesi di laurea, che si è ritenuto utile dar vita a questo volume. Esso propone, infatti, una ricca rassegna di problematiche utili a fini sia didattici sia applicativi: da diverse angolazioni, collegate ed interagenti l'una con l'altra, il lettore può trarre giudizi, spunti, curiosità e suggerimenti operativi necessari per «cam-

pionare» con sicurezza, cioè per poter formulare valutazioni di portata generale muovendo da un numero ristretto di unità. L'ambito di applicazione è quello delle scienze dell'uomo, sui versanti dell'indagine psicosociale, economica, politico-demografica.

Un vivo ringraziamento va a tutti gli studenti e agli operatori che a vario titolo hanno fornito stimolazioni e contributi. Fra essi, in particolare, Cecilia Cerra, Franca Bovo e Alessandra Anguillari hanno concorso attivamente alla messa a punto della rassegna bibliografica e alla compiuta articolazione dei problemi con cui frequentemente si misura chi si occupa di ricerca.

Nicola A. De Carlo
Egidio Robusto

Università di Padova
II Università di Napoli
Gennaio 1996

1.

INTRODUZIONE ALLE INDAGINI CAMPIONARIE

1.1. Utilità delle tecniche di campionamento

L'informazione statistica è oggi alla base di molti tipi di decisione, sia in campo sociopolitico sia economico.

Nei paesi ad economia avanzata esistono istituti privati ed enti pubblici – come ad esempio l'ISTAT, Istituto Centrale di Statistica, per quanto riguarda l'Italia – che forniscono periodicamente dati relativi al commercio, l'industria, l'economia, le condizioni di vita dei cittadini, la situazione generale del paese.

> La conoscenza statistica di un fenomeno può essere acquisita sia mediante una rilevazione completa delle sue manifestazioni, sia attraverso una rilevazione parziale che consenta di risalire con sufficiente approssimazione alle caratteristiche complessive del fenomeno, comprese quelle relative ai casi non considerati nel campione. In generale, si utilizzano rilevazioni parziali sia quando l'esame di tutte le unità comporta costi eccessivi, organizzazione troppo complessa, impossibilità di ottenere in tempi brevi i dati elaborati, sia nei casi che implicano la distruzione delle unità osservate, come avviene durante alcuni controlli di qualità (Marbach, 1992, p. 79).

Secondo Marbach, i censimenti demografici del 1990 sono costati 220 miliardi di lire in Francia, 300 nel Regno Unito, 3000 negli Stati Uniti e, per i censimenti del 1991 sulla popolazione e le attività produttive, in Italia sono stati spesi 500 miliardi di lire.

Il National Research Council ha stimato nel 1981 che negli Stati Uniti venivano condotte 100 milioni di interviste l'anno. Se pensiamo che il costo medio di un'intervista si aggira tra i 20 e i

40 dollari, negli Stati Uniti si spendono circa dai 2 ai 4 miliardi di dollari l'anno per le interviste, che sono solo una componente dell'intera operazione di un'indagine statistica (Särndal, Swensson, Wretman, 1992, p. 4).

> Il costo complessivo di un campione è inferiore a quello di una rilevazione complessiva (censimento) anche se il costo per unità elementare intervistata può risultare superiore in conseguenza della necessità di impiegare intervistatori esperti, alle spese amministrative di progettazione del campione e di generale impostazione del lavoro (Chisnall, 1990, pp. 77-78).

Va rilevato che per i censimenti sono necessari tempi lunghi sia nella fase di raccolta sia in quella di elaborazione dei dati; può quindi accadere che i risultati ottenuti da un'indagine durata un ampio arco di tempo non siano più attendibili data la grande velocità con cui mutano le caratteristiche della società contemporanea.

> Le indagini campionarie sono preferibili a quelle complete nei casi elencati di seguito:
> a) quando le rilevazioni di tutte le unità della popolazione da analizzare comportano costi elevati;
> b) quando i risultati della ricerca devono essere forniti in tempi brevi;
> c) quando la rilevazione, misurazione e controllo delle informazioni comporta la distruzione delle unità esaminate, come nel controllo statistico di qualità della produzione (valutazione della vita media di una lampadina elettrica o dell'affidabilità di un dato componente, giudizio su di un prodotto alimentare);
> d) nelle prove di mercato di un prodotto, quando esse non si possono svolgere sull'intero mercato, per motivi di riservatezza verso la concorrenza o per incompleta definizione di alcune variabili di marketing;
> e) quando la popolazione è di dimensione infinita o è un'entità astratta (ad esempio la popolazione dei prezzi di un certo prodotto nel tempo).
>
> Nei casi c), d), e), l'indagine campionaria diventa l'unica via praticabile per la rilevazione delle informazioni che interessano (De Luca, 1990, pp. 4-5).

Non bisogna avere l'illusione che la rappresentatività dei dati raccolti attraverso rilevazioni complete sia sempre superiore a quella riscontrabile utilizzando metodologie campionarie.

In tal senso Marbach (1992, p. 81), citando Giusti, afferma che:

> L'errore di osservazione nelle rilevazioni complete è altrettanto importante o forse più importante che nelle rilevazioni parziali; e mi riferisco all'idea, erronea, ma ancora molto diffusa tra gli utilizzatori, che i risultati di un censimento siano assolutamente esatti, o comunque più esatti di quelli forniti da indagini per campione. La circostanza che i risultati di un campionamento siano affetti da un errore probabilistico dovuto alla natura parziale della rilevazione non può e non deve far concludere che i risultati censuari siano di qualità superiore; questi ultimi, in realtà, sono ottenuti sotto il dominio delle stesse cause di errore e possono contenere deformazioni sistematiche almeno dello stesso ordine di grandezza. Anzi, nelle indagini per campione molte fonti di errore possono essere meglio tenute sotto controllo ... Può accadere, quindi, che un ampliamento della numerosità del campione accresca l'entità delle deformazioni sistematiche; ciò induce a sospettare che il massimo errore si registri proprio nei censimenti e che quindi un'indagine campionaria sia atta a fornire risultati anche più precisi dei censimenti stessi, specialmente allorché l'errore probabilistico possa ritenersi di scarsa rilevanza.

È opportuno evidenziare che esistono due tipi di errori di campionamento: un errore casuale (determinato dalle fluttuazioni casuali del campione) che può essere stimato ed eventualmente ridotto aumentando la numerosità del campione stesso, e un errore sistematico (o costante) più difficile da cogliere perché direttamente connesso alle problematiche relative al piano di campionamento e quindi al metodo di rilevamento utilizzato, alle tecniche impiegate (De Carlo, 1983, p. 38).

Se, ad esempio, ci rechiamo in un'aula universitaria alla fine di una lezione e selezioniamo alcuni studenti ritenendo che essi costituiscano un campione rappresentativo di tutti gli iscritti al corso di laurea in questione, molto probabilmente commettiamo un errore di tipo sistematico: non solo alcuni allievi potrebbero essere assenti in quel giorno, ma molti altri potrebbero, per svariati motivi, non frequentare assiduamente le lezioni. Il nostro campione risulterebbe quindi indicativo solo degli studenti costantemente frequentanti.

> Distorsioni sistematiche possono infatti verificarsi, in indagini totali come in quelle parziali, in rapporto a:

a) quesiti presentati, anche involontariamente, in forma orientata, tale da provocare un proprio addensamento di posizioni e atteggiamenti, su una particolare modalità di risposta;
b) comportamento non neutrale, tecnicamente improprio, del rilevatore, il quale può contribuire in svariati modi a deformare le informazioni se compila il questionario con scarso scrupolo ed interpreta scorrettamente alcuni quesiti;
c) comportamento degli intervistati, soprattutto di fronte a quesiti che possono avere una sia pur lontana eco fiscale oppure riguardare fenomeni che implicano comportamenti percepiti come lesivi del prestigio personale (si pensi a quesiti sulla pulizia personale) o addirittura socialmente riprovevoli (atteggiamenti nei confronti dell'eutanasia, degli stupefacenti, e simili).

Inoltre ogni forma di rilevazione statistica è soggetta ad errori nelle diverse fasi di elaborazione delle informazioni: ciò può avvenire durante la codifica, la preparazione dell'input per il computer e l'elaborazione elettronica (Marbach, 1992, p. 80).

Nelle indagini campionarie si possono ottenere risultati più validi curando molto attentamente l'indagine «sul campo»: avvalendosi di intervistatori meglio addestrati, controllando efficacemente il loro operato, affidando a ciascuno un numero adeguato di interviste (Pedon, 1995).

È possibile poi valutare l'accuratezza e la precisione dei dati raccolti mediante indagini complete (Marbach, 1992, pp. 81-82).

> Nei censimenti statunitensi del 1940 i giovani bianchi di 21-35 anni di età furono sottostimati del 4.5% e di ben 18% quelli di colore. Nel 1970 la popolazione complessiva fu sottostimata del 2.2%, in particolare nella misura dell'1.5% quella bianca e del 7.6% quella di colore. In occasione del censimento 1980 si tentò di prendere contatto anche con i circa 3 milioni di stranieri sprovvisti dei prescritti documenti di residenza, ma tuttora si ritiene che nonostante ciò la popolazione complessiva sia affetta da una sottostima massima del 2% che può giungere fino al 7.2% per quella di colore... In Italia, l'ISTAT ha effettuato accurati ed estesi controlli sui dati dei censimenti nazionali del 1981. Il grado di *copertura* del censimento della popolazione è risultato pari al 96%, ma soltanto dell'83% per le abitazioni non occupate. Su base di un ampio campione (8085 persone) realizzato dopo tale censimento è stato possibile accertare che nel 15.5% dei casi la variabile *istruzione* è stata contraddistinta da modalità nelle quali una sorta di *effetto prestigio*, presente al momento dell'autocompilazione del modello censuario, potrebbe aver inciso in senso distorcente. Anche la *copertura* ottenuta dal censimento delle attività produttive del 1981 non è stata totale: secondo stime cautelative, infatti, il 13.3% delle *posizioni* sarebbe sfuggito alla rilevazione censuaria.

Bisogna anche sottolineare le crescenti difficoltà che si incontrano nelle rilevazioni censuarie: l'ormai diffusa mobilità della popolazione e l'incidenza dei cittadini stranieri non registrati; il costante aumento della diffidenza nei confronti della pubblica amministrazione; il sempre più forte inserimento femminile nel mondo del lavoro, che rende più difficoltosa la consegna ed il ritiro dei questionari; l'accresciuta percentuale delle famiglie «monocomponenti», assai poco reperibili.

Si può quindi affermare che le ricerche per campione non costituiscono in generale un *minus* rispetto alle rilevazioni complete.

> Le indagini campionarie su larga scala e i censimenti sono raramente in reale alternativa; anzi si completano a vicenda: nella progettazione delle prime ci si giova in larga misura dei risultati censuari; d'altra parte, i risultati di un censimento possono essere completati da specifiche indagini campionarie: si pensi alle indagini post-censuarie dirette alla valutazione degli errori di risposta e più in generale all'accertamento della qualità dei dati (Cicchitelli, Herzel, Montanari, 1992, p. 24).

1.2. Definizioni e concetti statistici essenziali

È opportuno precisare subito alcuni dei termini più utilizzati. Con *popolazione* (o *universo*) ci si riferisce ad un

> qualsiasi insieme di elementi simili tra loro per una o più caratteristiche, che rappresentano l'oggetto di studio di una particolare indagine (Chisnall, 1990, p. 73).

Una popolazione si definisce *finita* quando è costituita da un numero finito di unità che può anche essere assai grande (ad esempio il numero dei nati vivi in una certa località in un deter-

minato periodo, il numero di motoveicoli circolanti in una data città); essa è *infinita*, invece, quando risulta costituita da un numero infinito di unità, come ad esempio per tutti i lanci che è possibile effettuare con una moneta (De Carlo, 1983, p. 30).

Le caratteristiche di una popolazione vengono chiamate *caratteri* o *variabili*, e possono essere qualitative o quantitative, a seconda che possano essere descritte con espressioni verbali o con numeri, pur se, a rigore, tali distinzioni hanno valore relativo.

> I caratteri quantitativi si distinguono, a loro volta, in *continui* e *discreti*: si definiscono continui quei caratteri, come l'età, la statura, il peso delle persone, che possono assumere tutti i valori all'interno di un certo intervallo; si dicono discreti quelli che, viceversa, possono assumere un numero limitato di valori, come la dimensione della famiglia, il numero di vani di una abitazione, e così via (Cicchitelli, Herzel, Montanari, 1992, p. 34).

> Il *campione* è quella parte limitata di popolazione che viene presa in esame (De Carlo, 1983, p. 29).

> Nel linguaggio corrente la parola campione significa parte di un tutto, sottoinsieme di una totalità di elementi che viene assunto a *rappresentare* la totalità stessa (Cicchitelli, Herzel, Montanari, 1992, p. 45).

> Perché si possano estendere alla popolazione i dati tratti dal campione occorre che esso sia *rappresentativo*, cioè che la distribuzione delle osservazioni tratte dal campione corrisponda alla distribuzione propria della popolazione (De Carlo, 1983, p. 29).

Quindi, possiamo definire un campione come quella parte limitata di una popolazione che rappresenti però la popolazione stessa; il *campionamento* è il procedimento di individuazione del campione, cioè di un insieme di osservazioni rappresentativo della popolazione.

Per *parametro* si intende la misura di una caratteristica della popolazione e per *statistica* la misura corrispondente nel campione.

Tutto ciò che riguarda lo studio delle relazioni esistenti fra popolazione e campione è oggetto della *teoria dei campioni*.

1.2.1. *Stima dei parametri*

L'obiettivo di un'indagine campionaria consiste nella stima, attraverso i dati del campione, dei parametri incogniti della popolazione oggetto di studio. Vengono assegnati così dei valori che si approssimano ai parametri della popolazione, utilizzando le informazioni tratte dal campione osservato (o da più campioni). Le tecniche di stima si basano su indici, detti *stimatori*, ottenuti dai dati campionari.

Sono stimatori, ad esempio, la media e la varianza del campione che vengono utilizzate per stimare la media e la varianza della popolazione.

Riportiamo di seguito le caratteristiche più comuni che può assumere uno stimatore.

Uno stimatore si definisce *corretto* se il suo valore medio (al variare del campione) corrisponde al valore del parametro della popolazione di riferimento.

> Lo stimatore gode della proprietà enunciata se il valore atteso della media campionaria coincide con la media della popolazione; in altri termini, lo stimatore in parola *riproduce* in media il valore del parametro da stimare (Cicchitelli, Herzel, Montanari, 1992, p. 55).

Se invece vi è una differenza tra il valore stimato e quello del parametro corrispondente tale stimatore viene definito *distorto*.

> Sappiamo che in generale possiamo servirci delle statistiche per effettuare stime di parametri e sono utili alcune precisazioni: posto che la media di una distribuzione campionaria sia uguale al parametro corrispondente, definiamo quella statistica *stimatore corretto* del parametro (in caso contrario tale statistica è detta *stimatore distorto* e la misura della distorsione è data dalla differenza fra il valore stimato e quello del parametro corrispondente); *l'accuratezza* di una stima derivata da un campione dipende dal fatto che vi sia, oppure no, una relazione di uguaglianza fra il suo valore e quello del parametro corrispondente: la sua eventuale inaccuratezza è così misurata dalla differenza fra i due valori; è possibile stimare dal campione la probabile accuratezza di una stima, che può essere indicata come *precisione* dello stimatore, e che è convenientemente misurata dalla deviazione standard della di-

stribuzione campionaria, cioè dall'errore standard; osserviamo che lo stimatore, per un certo disegno di campionamento, è il metodo di stimare il parametro servendosi dei dati del campione e che la stima è il valore ottenuto dal campione stesso; la correttezza della stima dipende sia dal metodo di stima sia dal metodo di campionamento; qualora le distribuzioni campionarie di più statistiche avessero la stessa media, la statistica che ha varianza minore viene definita lo *stimatore efficiente* della media, mentre le altre statistiche sono dette *stimatori inefficienti* (nelle applicazioni vengono usate assai sovente stime inefficienti).
La stima della varianza della popolazione si effettua servendosi della formula

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

dove s^2 rappresenta la stima della varianza della popolazione, ed n il numero di unità del campione; osserviamo che se n è sufficientemente grande, non abbiamo differenze rilevanti sostituendo n a $n-1$, e otteniamo

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

La stima dell'errore standard è espressa dalla formula

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

dove $s_{\bar{x}}$ costituisce la stima dell'errore standard della distribuzione campionaria.
Nelle applicazioni ci si serve per lo più di un unico campione per effettuare la stima del parametro che si vuole considerare (pur se è certo possibile valutarne più d'una da campioni diversi).
Se estraiamo vari campioni in un campionamento casuale e stimiamo da essi la media della popolazione, osserviamo che le varie stime in certa misura differiscono tra loro. Ciò è dovuto alle fluttuazioni casuali di ogni campione ed è necessario determinare quanto la stima possa essere influenzata dal campione estratto, ovvero di quanto possano differire tra loro i diversi campioni; la cosa è possibile considerando la dispersione della distribuzione campionaria che, come sappiamo, viene per lo più indicata mediante la deviazione standard, in termini di errore standard (*errore casuale*).

L'errore standard può essere valutato servendosi dei dati del singolo campione, usando la formula, che vale per popolazioni finite,

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}}$$

dove $\sigma_{\bar{x}}$ rappresenta l'errore standard e σ^2 la varianza della popolazione, n il numero dei dati del campione, N il numero dei dati della popolazione.
Se non è nota la varianza della popolazione ci serviamo della sua stima, e così da

$$s_{\bar{x}}^2 = \frac{s^2}{n}$$

sostituendo risulta

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}\cdot\frac{N-n}{N-1}}$$

in cui $s_{\bar{x}}$ è la stima dell'errore standard.

Osserviamo che il fattore $\frac{N-n}{N-1}$ (che può venire indicato come *fattore di correzione della popolazione finita*) si approssima all'unità per valori di N molto grandi relativamente a n, cioè quando la popolazione è assai ampia rispetto al campione. In tal caso questo fattore può venire non considerato: nelle applicazioni si tende a ometterlo quando la numerosità del campione non supera il cinque, il dieci per cento all'incirca, della numerosità della popolazione.
All'aumentare di n diminuisce il valore del fattore: inoltre, quando $N = n$, ovvero quando la popolazione coincide col campione, l'errore standard si annulla.
A questo punto è utile rivedere alcune delle questioni fin qui esposte, per apprezzare meglio la applicazioni.
Osserviamo come ciascuna stima del parametro possa essere tanto più precisa quanto minore è la dispersione delle stime tratte dai campioni, cioè quanto minore è la stima della dispersione della distribuzione campionaria.
La deviazione standard della distribuzione campionaria, cioè il suo errore standard, costituisce una misura idonea a valutare tale dispersione.
Già conosciamo la formula per calcolare l'errore standard della media utilizzando i dati del singolo campione, $\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}}$; omettendo di considerare ancora il fattore $\frac{N-n}{N-1}$, osserviamo che al crescere di n

il valore di $\frac{\sigma^2}{n}$ tende a diminuire e di conseguenza si riduce l'errore standard (è anche intuitivo che il più ampio fra vari campioni consente di trarre osservazioni più precise).
Nelle applicazioni è difficile che sia conosciuta la deviazione standard della popolazione, cioè il valore di σ; in tal caso sostituiamo σ con s ed effettuiamo una stima dell'errore standard.

Così, qualora il fattore $\frac{N-n}{N-1}$ venga tralasciato, troviamo ancora che $s_{\bar{x}} = \frac{s}{\sqrt{n}}$, dove $s_{\bar{x}}$ costituisce la stima dell'errore standard della distribuzione campionaria della media (De Carlo, 1983, pp. 47-52).

Uno stimatore viene definito *consistente* se al crescere della numerosità campionaria aumenta anche la probabilità che il valore della stima sia uguale al valore del parametro della popolazione. Per un n molto grande è quasi certo che la stima coincida con il parametro.

Uno stimatore si dice *efficiente* se la sua varianza, a parità di altre condizioni, risulta minore della varianza ottenibile con altri diversi stimatori.

Uno stimatore è *sufficiente* se contiene o sintetizza tutte le informazioni rilevanti contenute in un campione ai fini della stima di un parametro.

Uno stimatore per essere *naturale* deve avere lo stesso significato del parametro incognito.

La progettazione di un piano di campionamento non sarebbe possibile senza aver proceduto alla scelta preliminare di uno stimatore. Ma, per effettuare tale scelta, occorre poter fare riferimento a dei criteri, che consentano di rendere praticamente possibile la scelta stessa.
Un criterio di scelta tra due o più stimatori disponibili, proposto da alcuni studiosi (sostenitori di un punto di vista comportamentista o pragmatista), consiste nel tener conto delle loro proprietà, scegliendo lo stimatore che goda di un maggior numero di possibili proprietà o delle proprietà ritenute più desiderabili: ciò presuppone, naturalmente, di poter indicare le proprietà degli stimatori secondo una graduatoria di importanza; ma questo punto è ben lungi dall'essere risolto in maniera soddisfacente.
Cominciamo col considerare la consistenza di uno stimatore. Questa è una proprietà presentata coma essenziale da R.A. Fisher, che ha, appunto, proposto il termine *consistent*; col significato di *stretta coerenza*. In effetti, uno stimatore non consistente non ci dà alcuna fiducia di avvicinarsi al vero valore del parametro da stimare, neanche aumentando l'ampiezza del campione. Ma, se l'errore sistematico asintotico di uno stimatore è noto, od è praticamente trascurabile, è evidente che la consistenza può passare in secondo ordine, specie se uno stimatore è caratterizzato da una minore varianza rispetto ad un altro stimatore sia pur consistente.
Per quanto riguarda la correttezza, si può notare che gli statistici accordano generalmente la loro preferenza agli stimatori corretti. Ma questa proprietà, da sola, può essere di scarso interesse pratico, specie se uno stimatore non è consistente. Inoltre, nel valutare l'importanza della correttezza conviene tener conto anche della precisione di uno stimatore. Potrebbe, infatti, esistere uno stimatore non corretto, o solo asintoticamente corretto, molto più preciso (pur tenendo conto del suo errore sistematico) di uno stimatore corretto.
Oltre alla consistenza ed alla correttezza (eventualmente asintotica) uno stimatore viene giudicato in base alla sua varianza. In particolare, tra tutti gli stimatori consistenti e corretti si sceglie quello con la più piccola varianza.
Un criterio per la scelta di uno stimatore molto più restrittivo del precedente è noto come *criterio BLUE*, così denominato dalle iniziali dell'inglese Best Linear Unbiased Estimator, che letteralmente vuol dire: «*il migliore stimatore lineare corretto*».
Il criterio BLUE è un criterio molto famoso tra gli statistici, la cui origine storica può addirittura essere fatta risalire a C.F. Gauss. Si dice che T è uno stimatore BLUE di θ se è definito da una funzione lineare delle osservazioni campionarie, è corretto ed ha varianza minima nella classe degli stimatori lineari corretti di θ.
L'interesse verso gli stimatori BLUE è dovuto in primo luogo al fatto di essere corretti e con una varianza minima, che, come si è suaccennato, sono sempre state considerate delle proprietà desiderabili di uno stimatore, fin dal periodo classico. D'altra parte, gli stimatori lineari hanno una distribuzione asintotica normale in virtù del teorema limite centrale, ed anche questa è una caratteristica da non sottovalutare, specie quando non è nota la distribuzione di probabilità da cui è definita una popolazione.
Un criterio a cui si fa talvolta ricorso nel campionamento statistico, che merita di essere qui ricordato, è noto come criterio *analogico*. Il criterio analogico consiste nello scegliere uno stimatore in base allo stesso tipo di funzione usata nella popolazione per determinare un parametro. Così, la media delle osservazioni di un campione è uno stimatore analogico della media della popolazione.

Sotto certe condizioni si dimostra che uno stimatore analogico è uno stimatore corretto e di minima varianza del corrispondente parametro della popolazione. In ogni caso, il criterio analogico può servire a ricavare uno stimatore di cui è possibile successivamente esaminare le proprietà.
Un criterio per la scelta di uno stimatore, dovuto ad R.A. Fisher, a cui si può far ricorso quando sia nota la distribuzione di probabilità da cui è definita la popolazione, e quello della *massima verosimiglianza*. Al riguardo, ci limitiamo ad osservare che, sotto delle condizioni molto generali, uno stimatore di massima verosimiglianza è consistente, asintoticamente corretto ed efficiente in senso Fisheriano (De Cristofaro, 1979, pp. 101-103).

1.2.2. *Verifica di ipotesi*

Con verifica di ipotesi si intende quel processo in base al quale si analizza, in senso statistico, un'affermazione su di una popolazione; cioè se tale affermazione debba ritenersi vera o falsa, considerando i dati campionari.

Possiamo assumere come ipotesi che la differenza fra le medie dei campioni non sia tale da suggerire che essi siano stati estratti da popolazioni diverse, ma che tale differenza sia dovuta al caso.
Questa ipotesi è detta *ipotesi nulla*, e viene indicata con H_0.
L'altra ipotesi, quella cioè che non si verifichi H_0, e che la differenza fra le medie sia tale da far supporre che i campioni siano stati estratti da popolazioni diverse, è detta *ipotesi alternativa* e viene indicata con H_1.
Perché si possa stabilire se la differenza fra le medie dei campioni sia rilevante oppure no, e quindi tale da farci rifiutare H_0 o H_1, occorre stabilire se la loro diversità sia *significativa*, e ciò è possibile mediante i *test di ipotesi* (o *test di significatività*).
Tali test consentono di prendere decisioni riguardo alla possibilità di compiere errori: viene definito *errore di 1° tipo* quello che si commette riguardo ad un'ipotesi che dovrebbe essere accettata, *errore di 2° tipo* quello in cui si incorre ritenendo valida un'ipotesi che dovrebbe essere respinta.
La probabilità massima di incorrere in un errore di 1° tipo è definita *livello di significatività* del test, viene solitamente indicata col simbolo α, e viene scelta sulla base delle esigenze della ricerca (di solito vengono usati i livelli di significatività $\alpha = .05$ e $\alpha = .01$).

Ricordiamo ancora che l'ipotesi nulla e quella alternativa non possono essere vere entrambe, ma il verificarsi dell'una esclude l'altra: la decisione per la prima o per la seconda ha un valore probabilistico, cioè tiene in conto la probabilità di sbagliare.
Se, ad esempio, poniamo in ipotesi che una moneta non sia truccata, la probabilità H_0 che in un lancio risulti testa è $p = .5$; l'ipotesi alternativa H_1, che invece la moneta sia truccata, risulta $p \neq .5$: qualora in 6 lanci consecutivi ottenessimo testa per sei volte, la probabilità che ciò si verifichi casualmente è $p = (.5)^6$, cioè $p = .0156$, il che significa $p < .02$; se il livello di significatività prescelto è $\alpha = .05$ possiamo rifiutare l'ipotesi nulla e accettare che la moneta sia truccata, considerando che esiste una probabilità inferiore a .05 di incorrere in errore prendendo tale partito.
In questo esempio abbiamo considerato la possibilità di commettere un errore di 1° tipo: nelle applicazioni occorre rendere minima la probabilità di compiere errori sia del 1° sia del 2° tipo. Non è cosa facile perché assumendo valori di α assai piccoli, in modo da ridurre la possibilità dell'errore di 1° tipo (in pratica accettando quasi sempre H_0), ci esponiamo maggiormente all'errore di 2° tipo (accettare H_0 anche quando essa è falsa). In altri termini, limitare la probabilità di incorrere nell'errore di 1° tipo fa aumentare il rischio di commettere l'errore di 2° tipo: occorre decidere, considerando l'argomento e lo scopo della singola ricerca, se convenga minimizzare l'uno o l'altro tipo di errore...
È utile osservare che al crescere di n (aumentando dunque l'ampiezza del campione) diminuisce la probabilità di incorrere in entrambi i tipi di errore (De Carlo, 1983, pp. 61-63).

Facendo riferimento ad uno schema riportato in Levy, Lemeshow (1991, p. 14), richiamiamo e precisiamo alcuni indici di base.

Popolazione

La popolazione di una caratteristica X viene generalmente indicata con X ed è la somma dei valori della caratteristica per tutti gli elementi della popolazione. La popolazione è data da

$$X = \sum_{i=1}^{N} X_i$$

Media della popolazione

La media della popolazione rispetto alla caratteristica X è data da

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Proporzione della popolazione

Quando nella popolazione sono presenti attributi dicotomici e si vuole stimare la proporzione di un attributo *X* al suo interno, tale proporzione è data da

$$P_x = \frac{X}{N}$$

Varianza della popolazione

Indice di variabilità dei dati sulla base degli scarti dalla media. Poiché la somma degli scarti dalla media risulta nulla si utilizzano gli scarti al quadrato:

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N}$$

Deviazione standard della popolazione

L'indice di variabilità più usato risulta essere lo scarto quadratico medio, che corrisponde alla radice quadrata della varianza:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N}}$$

Varianza su attributi dicotomici

$$\sigma_x^2 = P_x\left(1 - P_x\right)$$

Coefficiente di variazione

Il coefficiente di variazione di una distribuzione viene indicato con V_x e corrisponde al rapporto tra la deviazione standard e la media della distribuzione:

$$V_x = \frac{\sigma_x}{\overline{X}}$$

Passiamo ora alle statistiche campionarie più comuni (Levy, Lemeshow, 1991, p. 21).

Totale campionario

Il totale campionario è generalmente indicato con *x* ed è la somma dei valori di una caratteristica per tutti gli elementi del campione:

$$x = \sum_{i=1}^{n} x_i$$

Media campionaria

La media campionaria di una caratteristica *X* è data da

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Proporzione campionaria

La proporzione campionaria su attributi dicotomici è data da

$$p_x = \frac{x}{n}$$

Varianza campionaria

Indice di variabilità dei dati campionari:

$$s_x^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n-1}$$

Deviazione standard campionaria

Lo scarto quadratico medio campionario è dato da

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n-1}}$$

Varianza su attributi dicotomici

$$s_x^2 = \frac{np_x\left(1 - p_x\right)}{n-1}$$

Se n è sufficientemente grande, non abbiamo differenze rilevanti sostituendo n a $n-1$, e otteniamo

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Di conseguenza, anche la varianza su attributi dicotomici risulta essere:

$$s_x^2 = p_x(1 - p_x)$$

oppure

$$s_x^2 = p_x q_x$$

se indichiamo con p la percentuale del campione in possesso dell'attributo ricercato e con $q = (1-p)$ la percentuale del campione che non possiede l'attributo ricercato.

La somma di p e q dà sempre 1.

1.2.3. *Distribuzione campionaria*

Consideriamo per esempio una popolazione costituita da 80 palline ($N = 80$), il cui valore vada da 1 a 80, e che ciascuna pallina rechi impresso su di sé il proprio valore, diverso da quello di tutte le altre. A ciascun numero potremo associare un elemento della popolazione. Da questa popolazione possiamo ottenere un campione casuale di 10 palline ($n = 10$), ciascuna contraddistinta dal suo valore, inserendo tutte le palline in un'urna ed estraendo le 10 palline che ci occorrono.

Se dopo ogni estrazione non reinseriamo la pallina estratta nell'urna, realizziamo un *campionamento senza ripetizione*, e otteniamo un numero finito di campioni di ampiezza n. Se invece dopo ogni estrazione rimettiamo la pallina nell'urna, effettuiamo un *campionamento con ripetizione* e possiamo considerare infinita la popolazione. Osserviamo che quando la popolazione finita è molto grande, per molti fini applicativi si può considerare il campione come se fosse estratto da una popolazione infinita.

Di seguito riportiamo la tabella 1.1. così ottenuta, con la media campionaria per ciascuna estrazione.

Riprendiamo l'esempio precedente e proviamo ad estrarre 8 campioni di 10 palline ($n = 10$), dalle 80 palline i cui valori vanno da 1 a 80. Le palline si trovano nell'urna e dopo ogni estrazione ciascuna pallina viene reinserita in essa.

Fra le varie distribuzioni possibili, per ciascun campione otteniamo le distribuzioni che descriviamo di seguito.

Tabella 1.1. – Valori per ciascun campione nelle 10 estrazioni (con ripetizione)

Campione											Medie campionarie $\bar{X}$
1	61	31	54	12	6	49	34	39	10	63	35.9
2	80	33	13	54	26	52	3	13	23	76	37.3
3	51	57	66	56	70	37	70	10	65	41	52.3
4	8	49	36	44	41	69	7	34	27	28	34.3
5	77	41	38	18	10	27	32	44	30	7	32.4
6	10	53	16	55	55	15	78	54	69	75	48.0
7	35	48	12	43	44	8	12	44	56	2	30.4
8	57	78	74	79	41	22	5	20	39	14	42.9
											$\bar{X}_{\bar{X}} = 39.2$

In tabella sono riportati i valori delle palline estratte per ciascuno degli otto campioni; ne è stato calcolato il valore medio usando la formula

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

in cui n e $\bar{X}$ indicano rispettivamente, il numero di unità del campione e il suo valore medio.

Così, sostituendo, da $\bar{X} = \frac{i}{10}\sum_{i=1}^{10} X_i$, si hanno per ciascun campione i valori $\bar{X}_1 = 35.9$, $\bar{X}_2 = 37.3$, $\bar{X}_3 = 52.3$, $\bar{X}_4 = 34.3$, $\bar{X}_5 = 32.4$, $\bar{X}_6 = 48.0$, $\bar{X}_7 = 30.4$, $\bar{X}_8 = 42.9$.

Da tale distribuzione di medie si ricava $\bar{X}_{\bar{X}} = \frac{1}{8}\sum_{i=1}^{8} \bar{X}_i = 39.2$, dove $\bar{X}_{\bar{X}}$ rappresenta la media degli otto campioni. Se indichiamo con μ la media della popolazione e con N il numero dei suoi elementi, usando la formula consueta

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

otteniamo

$$\mu = \frac{1}{80}\sum_{i=1}^{80} X_i = 40.5$$

È utile sottolineare che, nell'esempio fin qui condotto, la media delle medie dei campioni ($\overline{X}_{\overline{X}} = 39.2$) è più vicina alla media della popolazione ($\mu = 40.5$) di quanto lo siano le medie dei singoli campioni ($\overline{X}_1 = 35.9$, $\overline{X}_2 = 37.3$, $\overline{X}_3 = 52.3$, $\overline{X}_4 = 34.3$, $\overline{X}_5 = 32.4$, $\overline{X}_6 = 48.0$, $\overline{X}_7 = 30.4$, $\overline{X}_8 = 42.9$).

Se disponessimo di un numero infinito di palline, aventi ciascuna un valore diverso da quello dell'altra e se continuassimo all'infinito la nostra estrazione di campioni casuali di ampiezza $n = 10$, la media delle medie dei campioni andrebbe a coincidere con m, media della popolazione. In questo modo otterremmo la *distribuzione campionaria della media*, ovvero la distribuzione delle medie tratte da tutti i diversi campioni.
Si può dimostrare che per la distribuzione campionaria della media, in una popolazione infinita, valgono le equazioni

$$\mu_{\bar{x}} = \mu \quad \text{e} \quad \sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

dove $\mu_{\bar{x}}$ e $\sigma^2_{\bar{x}}$ sono rispettivamente media e varianza della distribuzione campionaria della media, mentre μ e σ^2 corrispondono alla media e alla varianza della popolazione (*teorema di convergenza stocastica*, detto anche *teorema del limite centrale*).
Tale distribuzione campionaria ha un valore unicamente teorico: non è possibile nella pratica ottenere un numero infinito di campioni; essa ha un andamento che tende alla normalità se viene tratta da una popolazione che si distribuisce normalmente, oppure, senza riguardo al tipo di distribuzione della popolazione, per valori di *n* sufficientemente grandi...

Posto $n = 1$ da $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$ deriva $\sigma^2_{\bar{x}} = \sigma^2$, il che significa che la dispersione della distribuzione campionaria della media coincide con la dispersione dei dati della popolazione, infatti la media di ogni campione coincide con ogni singolo dato della distribuzione della popolazione.
Per $n > 1$ osserviamo che in ogni campione vengono estratti valori che raramente coincidono con i valori estremi della distribuzione della popolazione: se ne calcoliamo la media, i valori alti estratti si compensano con quelli bassi... La dispersione della distribuzione campionaria della media è minore della dispersione dei dati della popolazione perché essa viene calcolata su valori che sono essi stessi medi. Così, in altri termini, la dispersione della distribuzione campionaria tende a raccogliersi intorno alla media della popolazione.
Osserviamo che anche in formula, da

$$\sigma^2_{\bar{x}} \cdot n = \sigma^2,$$

considerando che il prodotto $\sigma^2_{\bar{x}} \cdot n$ è costante, quanto più grande è *n* tanto più diminuisce la dispersione della distribuzione campionaria della media intorno al parametro (De Carlo, 1983, pp. 39-45).

1.2.4. *Errore standard*

La deviazione standard della distribuzione campionaria della media è detta *errore standard*: se usiamo infatti la media di un campione per stimare la media della popolazione (il parametro) possiamo incorrere in errori dovuti alle specifiche caratteristiche (*le fluttuazioni casuali*) di quel campione.
Esprimendo in altri termini quanto abbiamo visto precedentemente, possiamo dire che al crescere di *n* (cioè aumentando l'ampiezza del campione) diminuisce la dispersione della distribuzione campionaria, e pertanto diminuisce l'errore standard. Ciò appare evidente considerando la formula tratta dalla formula della varianza della distribuzione campionaria della media,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

dove $\sigma_{\bar{x}}$ indica l'errore standard, σ la deviazione standard della popolazione, e *n* l'ampiezza del campione.
Osserviamo ancora che per qualsiasi valore di *n* la distribuzione campionaria della media assume forma normale se la popolazione dei dati si distribuisce normalmente.
Per *n* sufficientemente grande (ai fini applicativi l'ampiezza del campione in tal caso può essere indicata in un minimo di 30 osservazioni per ciascun campione, cioè ≥ 30) la distribuzione campionaria della media tende a disporsi secondo la curva normale qualunque sia l'andamento della popolazione da cui è tratta.
Così alla distribuzione campionaria della media, purché *n* sia sufficientemente grande, si possono applicare le proprietà della distribuzione normale.
Se ignoriamo la varianza della popolazione la formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ non è applicabile, ma si deve ricorrere a una stima tratta dal campione (De Carlo, 1983, pp. 45-46).

Se la popolazione di partenza è finita (estrazione senza reinserimento) la varianza risulta essere:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

dove il fattore $\frac{N-n}{N-1}$ è detto *fattore di correzione per popolazioni finite*. Quando la popolazione è molto grande rispetto al campione tale fattore tende all'unità e può non essere considerato.

Se la varianza della popolazione non è nota si può calcolare la deviazione standard della distribuzione campionaria della media sulla base della stima di tale varianza ottenendo una stima dell'errore standard:

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n} \cdot \frac{N-n}{N-1}}.$$

Se tralasciamo il fattore di correzione per popolazioni finite si può notare come al crescere di n diminuisca l'errore standard. Inoltre, se $N = n$, e cioè se la popolazione coincide con il campione l'errore standard diventa nullo.

Considerando la varianza calcolata su attributi dicotomici, un'altra formula per calcolare l'errore standard è:

$$\sigma_{\bar{x}} = \sqrt{\frac{p_x q_x}{n}}.$$

1.2.5. *Intervalli di fiducia*

Poiché la distribuzione campionaria della media si approssima alla curva normale valgono per essa le stesse proprietà di una distribuzione gaussiana.

Il 68.27% delle unità incluse nel campione sarà compreso nell'intervallo ± 1σ dalla media.

Il 95.45% delle unità incluse nel campione sarà compreso nell'intervallo ± 2σ dalla media.

Il 99.73% delle unità incluse nel campione sarà compreso nell'intervallo ± 3σ dalla media.

Tabella 1.2. – Valore dei limiti di confidenza

Livello di confidenza	99.73	99	95.45	95	68.27
Valore di Z	3.0	2.58	2.0	1.96	1.0

Nella tabella 1.2. (Chisnall, 1990, p. 133) vengono riportati alcuni dei valori delle aree sottese dalla curva normale standardizzata Z che ha media pari a zero e varianza unitaria.

$$Z = \frac{X - \bar{X}}{\sigma}$$

Quindi se la statistica S è la media campionaria $\bar{X}$, allora i *limiti di confidenza* (o di fiducia) al 95% e al 99% per le stime della media μ della popolazione sono dati da $\bar{X} \pm 1.96\sigma_{\bar{x}}$ e $\bar{X} \pm 2.58\sigma_{\bar{x}}$. I limiti di confidenza sono:

$\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}}$ popolazioni infinite;

$\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ popolazioni finite.

In genere σ è sconosciuto e allora per ottenere i limiti di confidenza si usa la stima ottenuta dal campione.

Proponiamo di seguito alcuni esempi relativi a diverse distribuzioni campionarie.

L'intervallo di confidenza per le proporzioni è dato da:

$$\hat{p} - z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} \le p \le \hat{p} + z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} \quad \text{per}$$

$$\mu_{\hat{p}} = p \quad \text{e} \quad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

Se S_1 e S_2 sono due statistiche aventi distribuzioni approssimativamente normali, i limiti di confidenza delle differenze tra i parametri dell'universo corrispondenti alle due statistiche sono:

$$S_1 - S_2 \pm z_c \sigma_{s_1 - s_2} = S_1 - S_2 \pm z_c \sqrt{\sigma_{s_1}^2 + \sigma_{s_2}^2};$$

mentre limiti di confidenza della somma dei parametri sono:

$$S_1 + S_2 \pm z_c \sigma_{s_1 + s_2} = S_1 + S_2 \pm z_c \sqrt{\sigma_{s_1}^2 + \sigma_{s_2}^2}$$

L'intervallo di confidenza della differenza tra le medie di due popolazioni infinite è:

$$\left(\overline{X}_1 - \overline{X}_2\right) - z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\overline{X}_1 - \overline{X}_2\right) + z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

I limiti di confidenza delle differenze delle proporzioni fra due campioni n_1 e n_2 sono:

$$p_1 - p_2 \pm z_c \sigma_{p_1 - p_2} = p_1 - p_2 \pm z_c \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\,.$$

1.2.6. *Design effect*

Il design effect viene definito come il rapporto tra la varianza che si ottiene con un campione complesso e quella che si ottiene con un campione casuale semplice della stessa numerosità (Kish, 1965).

$$\text{Deff} = \frac{\text{Varianza del campione complesso}}{\text{Varianza del campione casuale semplice con pari numerosità}}$$

$$\sqrt{\text{Deff}} = \frac{\text{Errore standard del campione complesso}}{\text{Errore standard del campione casuale semplice con pari numerosità}}$$

Il campione casuale semplice viene utilizzato come termine di confronto per valutare l'affidabilità degli altri metodi di campionamento casuale (Chisnall, 1990, p. 130).