2. Ricerche per campione

1. Indagini complete e indagini campionarie

Alla conoscenza statistica di un fenomeno – di tipo socio-economico o fisico – si può pervenire sia mediante un'indagine **completa**[1] (o **censuaria**) sulle sue manifestazioni, sia tramite un'indagine **parziale**, che permetta di risalire – con sufficiente approssimazione – alle caratteristiche del fenomeno nella sua globalità, cioè, alle sue possibili modalità, comprese quelle relative alle unità escluse dal **campione**[2].

Quando la dimensione della **popolazione**[3] **obiettivo** (*target population*) è molto elevata, o infinitamente grande[4], non è praticabile

---

[1] La rilevazione **completa** di un carattere (professione, titolo di studio, possesso di un bene, ecc.) è svolta su tutte le unità o elementi costituenti una **popolazione** o **collettivo**, in un determinato intervallo spazio-temporale. Esempi di rilevazioni complete sono quelle censuarie, svolte in Italia ogni dieci anni dall'Istat.

[2] Nel linguaggio comune con il termine **campione** si intende la parte di un tutto, ovvero, un sotto-insieme di una popolazione di elementi deputato a **rappresentare** la popolazione medesima.

Il **campione statistico**, pur non discostandosi da questa nozione intuitiva, può non rappresentare un sotto-insieme della popolazione; ciò accade quando l'operazione di estrazione delle unità campionarie è **con ripetizione** (parr. 1.1; 5.1; 10.1), cioè quando uno stesso elemento può essere estratto dalla popolazione e incluso nel campione più volte: un siffatto campione non costituisce in senso proprio un sotto-insieme della popolazione.

[3] È detta **popolazione** un collettivo composto da unità aventi uno o più caratteri in comune; ad esempio, la popolazione dei potenziali acquirenti di un'automobile è costituita dai soggetti di età superiore a 18 anni e forniti di patente di guida.

[4] È infinitamente grande, ad esempio, una popolazione teorica alla quale si associano variabili **continue**; le variabili osservate sulle popolazioni concrete (*finite*) sono, invece, **discrete** o rilevate come tali.

un'indagine completa (rilevazione di tutte le unità statistiche), per motivi pratici e/o per l'insostenibilità dei costi.

Quando la popolazione allo studio è di dimensione elevata e si presenta la necessità di ricavare informazioni sulla stessa, senza dover rilevare tutti i suoi elementi, si fa ricorso all'*indagine campionaria*, cioè ad un tipo di indagine parziale svolta con il **metodo del campione**.

Mentre l'indagine **parziale** comprende semplicemente una parte o sottoinsieme (*sample*) delle unità della popolazione, quella **campionaria** considera sì una parte, ma scelta con procedure atte a garantire la **rappresentatività** della popolazione di provenienza (**campione a scelta ragionata** o **per quote**) e/o la **casualità** di estrazione delle unità campionarie (**campione probabilistico**).

È importante sottolineare la profonda differenza, anzi, la contrapposizione, esistente tra il campione *a scelta ragionata*, che mira ad essere *rappresentativo*, e il campione *probabilistico*, che vuole essere solo uno dei possibili campioni (Herzel *et alii*, 1994, p. 148). Il secondo tipo di campione ha il vantaggio di assicurare l'estrazione degli elementi campionari con criteri **oggettivi** e consente di costruire un *modello matematico* basato su alcuni risultati del **calcolo delle probabilità**.

Il **campione rappresentativo** assicura una buona qualità delle informazioni acquisite, dando un'immagine fedele della popolazione da cui esso proviene, con riguardo al carattere indagato; la scelta delle unità campionarie è lasciata all'arbitrio del rilevatore[5]; la dimensione del campione è fissata secondo criteri di convenienza economica e di praticità.

La differenza tra gli anzidetti due tipi di indagine è la seguente: se da una popolazione di 100.000 consumatori di un prodotto si estrae un sottogruppo – comunque scelto – di 1.000 soggetti, sui quali si effettua una rilevazione, si ha un'indagine **parziale**; se i 1.000 consumatori sono selezionati in modo da includere uomini e donne, giovani e anziani, ecc., in proporzioni tali che il campione rispecchi la composizione della popolazione, si ha un campione **rappresentativo**; se, invece, le unità campionarie sono scelte con criterio di casualità dalla popolazione, si ha un campione **probabilistico**.

[5] Gli elementi del campione sono estratti con criteri **soggettivi**, ciò costituisce una fonte di distorsione del campione.

L'indagine campionaria si sostituisce a quella **censuaria** (*census*) nei seguenti casi:

a) la rilevazione di tutte le unità della popolazione da analizzare comporta costi elevati;
b) i risultati della ricerca devono essere disponibili in tempi brevi;
c) la rilevazione, misurazione e controllo delle informazioni comporta la distruzione delle unità da esaminare (ad esempio, prova di degustazione di un vino o assaggio di un piatto pronto, prova di resistenza di una lampadina o di durata di una batteria);
d) le prove di gradimento di un prodotto non possono esser svolte sull'intero mercato, per motivi di riservatezza verso la concorrenza, o per l'incompleta definizione delle modalità di alcuni fattori di marketing;
e) la popolazione è di dimensione infinita[6].

Nei casi: *c*), *d*), *e*), l'indagine per campione è l'unica via praticabile per la rilevazione delle informazioni di interesse.

È da osservare, tuttavia, che, spesso, indagini campionarie e censimenti si **integrano** tra di loro: per lo svolgimento delle prime ci si avvale dei risultati dei secondi; d'altro canto, nello studio dei risultati di un censimento si fa ricorso ad indagini campionarie integrative post-censuarie, volte a valutare gli errori di risposta e la **qualità dei dati** in generale (v. cap. 5).

[6] Nelle ricerche di mercato sono considerate **finite** le popolazioni costituite da un insieme di unità le quali, pur se numerose, sono enumerabili ed **identificabili** (cioè "etichettabili" e distinguibili tra di loro).

Non sempre la popolazione di interesse è identificabile; ad esempio, l'insieme dei potenziali acquirenti di una stampante *laser* a colori non può considerarsi una popolazione, non potendosi identificare tutte le unità statistiche appartenenti all'insieme considerato (potenziali acquirenti del prodotto).

4. Campioni non probabilistici

I campioni **non probabilistici**, detti anche *campioni senza una legge probabilistica definita a priori* o campioni **non casuali** (per i quali la scelta delle unità campionarie prescinde dai criteri di casualità), sono campioni **rappresentativi**, che raffigurano la popolazione studiata rispetto a determinati caratteri.

Questi campioni, diversamente da quelli probabilistici, non consentono di individuare un intervallo – intorno al valore di stima del parametro del carattere analizzato – entro il quale collocare, con ragionevole fiducia (**livello di confidenza**), il valore del parametro incognito.

La ragione per la quale i campioni non probabilistici sono preferiti, a volte, a quelli ad estrazione casuale è da ricercare nella necessità di operare su campioni rappresentativi o di limitare i *costi*[14] (soprattutto di rilevazione) ed i *tempi* della ricerca, o nella circostanza che, talvolta, possano essere ritenuti sufficienti dei risultati di massima (come ad esempio, in indagini **preliminari**), o nel fatto che la popolazione osservata è omogenea rispetto al carattere studiato o, infine, nella circostanza che la popolazione di interesse si compone di elementi non individuabili facilmente (ad esempio, il collettivo dei potenziali acquirenti di un elettrodomestico che prepara il caffè).

[14] Le indagini svolte su campioni *casuali* risultano più costose in quanto richiedono: *a*) di approntare l'elenco delle unità appartenenti alla popolazione; *b*) di estrarre con criterio di casualità statistica le unità del campione; *c*) di raggiungere effettivamente le unità campionarie individuate nominativamente (a volte si devono contattare unità remote e difficilmente accessibili).

La tecnica del campionamento non probabilistico (che storicamente precede quella del campione probabilistico[15]) è diffusamente adottata in ricerche di natura **qualitativa** (cap. 8), ma viene applicata anche in ricerche **quantitative**.

Il campionamento non probabilistico, pur se discutibile sul piano della correttezza metodologica, può dare buoni risultati, qualora si adottino opportuni accorgimenti.

La selezione delle unità campionarie viene effettuata generalmente con criteri **soggettivi** e, in alcuni casi, in parte oggettivi.

Si fa distinzione, quindi, tra:

a) campioni **assolutamente non probabilistici** (*a scelta ragionata, di comodo, a valanga*);

b) campioni **quasi probabilistici** (*campione per quote*).

Tra i campioni non probabilistici, nelle ricerche di mercato sono maggiormente usati quelli a scelta *ragionata* e per *quote*. Entrambi questi metodi si basano su un modello relazionale inerente ai caratteri che guidano il campionamento e quelli incogniti allo studio.

Come già rilevato, i campioni non probabilistici non consentono di calcolare il grado di precisione dei risultati (cioè degli **stimatori**), tuttavia, i sostenitori di questi campioni ritengono che, in alcune circostanze, la qualità di un'informazione non va valutata in termini assoluti, ma nell'ambito dello specifico contesto decisionale, prescindendo, quindi, dalla possibilità di poter estendere i risultati alla popolazione da cui è tratto il campione.

4.1. *Campione a scelta ragionata*

In questo tipo di campionamento (*assolutamente non probabilistico*) la numerosità del campione – in genere non elevata – è fissata sulla base di criteri di convenienza; l'organizzazione dell'indagine è snella; i tempi ed i costi sono contenuti.

La scelta – soggettiva – delle unità campionarie è effettuata: *a)* sulla base delle conoscenze e del **giudizio** del ricercatore sul fenomeno allo

[15] Il campionamento non probabilistico nasce negli anni '20 del secolo scorso, quello probabilistico intorno agli anni '40.

studio (un esperto, con criteri più o meno personali, compone un campione **rappresentativo** della popolazione[16]); *b)* su particolari – importanti – comunità di soggetti, quali, ad esempio, **opinion leaders** o testimoni privilegiati.

I criteri di scelta delle unità possono basarsi, oltre che su valutazioni personali del ricercatore (*criteri soggettivi*), anche su *criteri oggettivi*, cioè, sulla somiglianza delle unità campionarie (in genere territoriali) alla popolazione di provenienza (campioni di **unità tipiche**[17]), con riferimento ai caratteri strutturali di quest'ultima. (Nel **campionamento bilanciato** le unità campionarie vengono scelte in modo tale che la media campionaria di uno o più caratteri noti sia uguale o molto vicina alla media della popolazione di provenienza del campione). Si perviene, quindi, a campioni costituiti da **elementi rappresentativi** ed a campioni estratti da **aree barometro**.

I primi vengono utilizzati in *ricerche preliminari*, per scegliere, ad esempio – nell'ambito di una data area geografica – zone che presentano caratteristiche *tipiche*. Con questa metodica vengono selezionate – sulla base dell'analisi delle distribuzioni di opportuni caratteri della popolazione – le zone o località (ad esempio, comuni) che, rispetto ai ca-

[16] L'esperto sceglie le caratteristiche della popolazione da riprodurre nel campione, affinché questo risulti rappresentativo. Nel campione vengono incluse, quindi, solo le unità che presentano determinati caratteri della popolazione, rilevanti per il fenomeno allo studio (comportamento di consumo, atteggiamenti verso un prodotto, ecc.).

Se, ad esempio, l'obiettivo della ricerca è quello di accertare il gradimento di un nuovo "piatto pronto" da lanciare sul mercato, pur se le potenziali acquirenti dello stesso sono donne di casa, è probabile che il successo del prodotto sia decretato dal consenso di donne-lavoratrici; è probabile, inoltre, che le potenziali consumatrici vivano e lavorino soprattutto in città. In tal caso, per giungere ad una valutazione di massima di gradimento del prodotto, si può formare un campione ragionato di donne, che vivono in città e che svolgono un'attività professionale. Successivamente, sulle donne del campione così formato, si effettueranno discussioni di gruppo e/o colloqui individuali non strutturati (cap. 8), prima e dopo l'assaggio del prodotto (**quasi-esperimento**: disegno **prima-dopo** senza gruppo di controllo; A. De Luca, *Programmazione ed analisi degli esperimenti nel marketing - Applicazione dei metodi statistici*, FrancoAngeli, Milano, 2004, p. 50).

[17] Unità che rispecchiano le principali caratteristiche della popolazione. Dette unità sono estratte da gruppi omogenei rispetto a caratteri predeterminati, ipotizzando che le variabili rilevate su un'unità statistica siano interdipendenti (se un'unità è vicina alla media della popolazione rispetto a determinati caratteri noti, si presume che la stessa sia vicina a tale media anche per i caratteri non noti).

ratteri allo studio, presentano valori prossimi a quelli dell'intera area geografica.

I campioni scelti da aree barometro rispecchiano il comportamento *medio* dell'intera popolazione; pertanto, se un prodotto è accolto con favore dai consumatori appartenenti ad una di queste aree, si potrà supporre che il medesimo avrà la stessa accoglienza sull'intera area di interesse della ricerca.

In questo tipo di campionamento, quindi, le valutazioni e le conoscenze del ricercatore si sostituiscono al caso. Ma, sostituendosi al criterio dell'estrazione casuale la discrezionalità del rilevatore, si presenta il rischio che questi possa introdurre – pur se involontariamente – delle **tendenziosità**. Nonostante ciò, i campioni ragionati forniscono informazioni di ottima qualità, se la selezione delle unità campionarie è effettuata da un esperto, che vanta una profonda conoscenza del fenomeno allo studio.

I campioni in argomento sono preferibili a quelli probabilistici nei casi seguenti:

a) in **vaste** indagini, su scala nazionale, che interessano poche unità territoriali[18], o nelle ricerche inerenti un *test* su un prodotto o su una campagna pubblicitaria (l'area di prova viene scelta con oculatezza e non sulla base di criteri di casualità [19]);

b) in ricerche su **prodotti industriali**, quando la particolare struttura produttiva e distributiva richiede l'inclusione di aree geografiche *tipiche* per il settore, le quali consentono di ottenere un' informazione completa sul fenomeno allo studio.

[18] Quando la dimensione del campione n è molto piccola ($n < 10$) un campione *ragionato* può risultare più affidabile e maggiormente rappresentativo di un campione probabilistico. Ad esempio, se si è accertato che due città ne rappresentano 20 (con riferimento all'ampiezza demografica, ai comportamenti dei consumatori, alle caratteristiche dei punti di vendita, ecc.), conviene scegliere queste due città per formare un campione rappresentativo delle 20. L'estrazione **casuale** di due città dalle 20 darebbe, molto probabilmente, un campione **non rappresentativo**.

[19] Se si intende svolgere un'indagine regionale – onde pervenire alla stima di un parametro a livello nazionale – non è opportuno considerare alla pari tutte le regioni italiane (estraendo poi casualmente alcune di esse per comporre il campione); conviene, invece, individuare e scegliere le regioni **tipiche** rispetto al fenomeno allo studio.

4.2. *Campione per quote*

Il **campione per quote** (*quota sampling*) è il più usato tra i campioni non probabilistici (soprattutto nelle ricerche di mercato e nei sondaggi di opinione), pur non rispettando esso i criteri della stretta casualità delle unità campionarie (la quale, a volte, non è applicabile o è di costosa realizzazione). Tuttavia, la procedura consente di non rinunciare del tutto al principio della casualità (campione **quasi probabilistico**).

Con questo tipo di campione – fatto importante – non è necessario disporre di una **lista** (anagrafica, elettorale, ecc.)[20] nominativa delle unità componenti la popolazione studiata.

La popolazione di individui viene suddivisa in **classi** o sottogruppi omogenei, secondo caratteri strutturali discriminanti socio-demografici o professionali o economici (età, grado di istruzione, zona di residenza, sesso, numero di componenti il nucleo familiare, condizione lavorativa, tipo di abitazione, possesso di determinati prodotti, ecc.).

La dimensione totale del campione, fissata *a priori*, viene ripartita tra le anzidette classi in base al loro peso percentuale – rilevato da idonee fonti statistiche – in modo tale che il campione rispecchi la struttura della popolazione.

Agli intervistatori si assegnano le **quote**, cioè il numero di interviste da effettuare in ogni classe[21], imponendo il **vincolo** di formare un campione che presenti la stessa **composizione** della popolazione.

Normalmente ciascuna quota assegnata è indipendente dalle altre (**quote marginali**); solo in casi particolari si dispone della documentazione statistica *multidimensionale* (a due o più variabili), utile per assegnare quote associate (basate, cioè, su più caratteri congiuntamente considerati). È da rilevare che solo il rispetto delle **quote congiunte** assicura un campione effettivamente rappresentativo della popolazione (v. Quadro 1).

[20] La lista, detta **base campionaria** (*frame*), è, come noto, l'elenco delle unità che compongono la popolazione (ad esempio, l'elenco degli abbonati al telefono, quello delle aziende appartenenti ad un'associazione di categoria e simili), da cui estrarre casualmente gli elementi campionari.

[21] Ogni quota introduce un vincolo, il quale garantisce un'equilibrata formazione del campione, che risulta così rappresentativo.

Quadro 1 – Un esempio di campione per quote non rappresentativo (continua)

Una società petrolifera sia interessata a valutare la proporzione di donne che effettua personalmente la manutenzione della propria vettura, su una popolazione obiettivo, classificata secondo due attributi: **età** anagrafica (donne con età < 35 anni e donne di età ≥ 35 anni) e **categoria occupazionale** (casalinghe e professioniste).

La composizione della popolazione, secondo i due caratteri strutturali indicati, è riportata in Tab. 1.

Tab. 1 – Composizione di una popolazione di donne

Età / Categoria occupazionale	< 35 anni	≥ 35 anni	*Totale*	*Composizione della popolazione*
professionista	300 [30%]	200 [20%]	500	50%
casalinga	200 [20%]	300 [30%]	500	50%
Totale	500	500	1.000	100%
Composizione della popolazione	50%	50%	100%	

Per raggiungere gli obiettivi della ricerca si sia deciso di fare ricorso al campionamento per quote. Fissata la dimensione campionaria complessiva $n = 100$, il campione di intervistate è stato scelto come indicato in Tab. 2.

Tab. 2 – Campione per quote, secondo due caratteri, di una popolazione di donne

Età / Categoria occupazionale	< 35 anni	≥ 35 anni	*Quote marginali*	*Composizione della popolazione*
professionista	50	0	50	50%
casalinga	0	50	50	50%
Quote marginali	50	50	100	100%
Composizione della popolazione	50%	50%	100%	

Dalla Tab. 2 si desume che il campione, pur rispettando per i due caratteri – come impone la procedura – le quote marginali (50% e 50%), non è rappresentativo, infatti esso non riproduce tutte le quote associate (30%, 20%, 20%, 30%), risultando assenti quelle delle casalinghe con età < 35 anni e le donne professioniste di età ≥ 35 anni.

Sulla base delle quote loro assegnate, gli intervistatori scelgono **discrezionalmente** le unità da contattare, senza far riferimento ad un elenco nominativo delle stesse.

Una siffatta procedura facilita la rilevazione campionaria, cadendo la necessità dell'*identificazione nominativa* degli intervistandi, ma non garantisce la equiprobabilità di estrazione delle unità della popolazione, le quali non detengono la medesima probabilità di essere estratte (ad esempio, gli inquilini dei piani superiori dei palazzi residenziali, come

pure gli abitanti delle zone periferiche o di località remote o non ben collegate alle zone di residenza degli intervistatori, hanno scarse probabilità di essere contattati).

Questo tipo di campione presenta una certa analogia con quello – probabilistico – stratificato (cap. 3); la quota può essere considerata uno **strato**, con la differenza fondamentale che nel secondo tipo di campione l'estrazione delle unità campionarie è effettuata con criterio di stretta **casualità**, mentre nel campione per quote tale scelta è lasciata all'**arbitrio** del rilevatore, il quale può privilegiare interviste che richiedono meno sforzo.

Pertanto, l'intervistatore[22], da un lato è libero di scegliere gli intervistandi all'interno di gruppi prestabiliti, dall'altro è vincolato dalla condizione di intervistare gruppi di soggetti, numericamente definiti (quote), che presentano determinate modalità dei caratteri strutturali.

L'intervistatore è, quindi, libero di scegliere tra due soggetti, che siano entrambi uomini, di età compresa tra 25 e 30 anni, entrambi impiegati, ma non potrà sostituire un impiegato di età 25-30 anni con un'impiegata della stessa età.

Il campione per quote costituisce un artificio volto a limitare l'arbitrio del rilevatore nella scelta degli elementi campionari e, quindi, la **tendenziosità** del campione[23].

Per ridurre l'arbitrarietà del rilevatore nella scelta delle unità campionarie alcuni istituti di ricerca fissano – con scelta casuale – il punto di partenza di ciascun grappolo di interviste da effettuare, l'intervistatore deve, perciò, rispettare un itinerario specificato (**campionamento per itinerari**)[24]; a volte egli è tenuto anche ad osservare determinati orari nello svolgimento delle interviste.

I principali **limiti** che può presentare il campione per quote sono i seguenti:

[22] Per questo tipo di campionamento si ricorre normalmente all'intervista **diretta** (detta anche **personale** o **faccia-a-faccia**).

[23] Le sottoquote frazionano l'universo in gruppi omogenei tali che il campione, pur se non estratto dagli stessi con criteri di casualità, possa fornire una buona stima del carattere analizzato.

[24] Gli Istituti di ricerca, dopo aver estratto casualmente i nominativi degli intervistandi (ricavati da liste elettorali o anagrafiche, da registri automobilistici, da base di dati aziendali, ecc.), suddividono gli stessi per zone omogenee, che assegnano agli intervistatori.

- possibile **distorsione** del campione dovuta all'esclusione degli intervistandi trovati assenti dal loro domicilio;
- **autoselezione** degli intervistati: gli intervistatori possono tendere a contattare i soggetti più disponili;
- possibile *sottostima* della variabilità[25] del fenomeno allo studio, se il rilevatore tende ad avvicinare – pur rispettando le quote – soggetti con caratteristiche simili;
- possibile scarso impegno dei rilevatori, provocato dalla mancanza di controllo – da parte dell'Istituto di ricerca – delle ragioni di rifiuti all'intervista.

In sintesi, i vantaggi che offre il campione per quote – rispetto ad altri tipi di campioni – sono costituiti da *minori costi* e da tempi di rilevazione *più brevi*.

Un caso particolare è costituito dal campione formato da **popolazioni mobili** o **di flusso** (indagini sui visitatori di una fiera campionaria, sugli automobilisti in transito da un casello autostradale, sugli spettatori all'uscita da uno spettacolo di massa, ecc.). In questi casi, non conoscendo la dimensione della popolazione oggetto dell'indagine, non è possibile determinare le quote; si impone, pertanto, la condizione di intervistare un soggetto ogni k, fissando il **modulo** k in modo empirico, sulla base della distribuzione dell'affluenza del pubblico nell'intervallo temporale di rilevazione.

Questa procedura di scelta delle unità campionarie è simile a quella del campione **sistematico** (par. 14), ma, contrariamente a quest'ultimo, nel campione per quote non è nota la dimensione della popolazione allo studio e la **frazione di campionamento** (rapporto tra la dimensione del campione e quella della popolazione) è fissata in modo empirico.

4.3. *Campione di comodo*

I campioni di *comodo* o di *convenienza* sono usati allo scopo di acquisire informazioni in modo veloce ed **economico**. Per essi non si fa

[25] La rappresentatività che si consegue con questo tipo di campione è solo **presunta**, siccome si sottopongono all'intervista le persone più disponibili o più facilmente reperibili e, quindi, tra di loro più simili per comportamenti di acquisto ed atteggiamenti.

ricorso a liste della popolazione; gli intervistatori scelgono arbitrariamente le unità da avvicinare (in genere volontari[26]), tra quante si trovano in una determinata situazione (ad esempio, in un dato negozio o centro commerciale). Con questo tipo di campione, di rapida rilevazione, si privilegia il **numero** di interviste più che la loro qualità.

Con riferimento al campionamento svolto presso esercizi commerciali, è da rilevare che si possono presentare tre fonti di distorsione, inerenti[27]:

- al punto di vendita: **autoselezione** degli intervistati (residenti nella zona di ubicazione del negozio e frequentatori più assidui del punto vendita), che possono non essere rappresentativi della popolazione di utenti della catena di distribuzione cui appartiene il punto di vendita;
- alla **postazione** dell'intervistatore (che può essere situata all'ingresso, all'interno o all'uscita del negozio), la quale influenza l'accettabilità del contatto da parte della clientela e la qualità dell'intervista;
- al **momento** dell'intervista: il giorno e l'ora in cui è svolta la rilevazione può influenzare il clima dell'intervista.

I campioni di comodo – come tutti i campioni non probabilistici – danno risultati di massima (fine a sé stessi) e **non sono generalizzabili** ad una popolazione più ampia, né consentono di valutare l'attendibilità delle stime campionarie[28].

4.4. *Campione a valanga*

I campioni a valanga (*snowball sampling*[29]) o *putativi* si utilizzano

[26] Gli intervistati appartengono, in genere, a comunità (università, centri religiosi o di volontariato o enti di beneficienza). Le interviste sono condotte a costo zero.

[27] G. Guido, *Aspetti metodologici e operativi del processo di ricerca di marketing*, Cedam, Padova, 1999.

[28] Nel campione in argomento rientra quello inerente al **televoto** (raccolta per telefono delle opinioni degli ascoltatori/telespettatori sulle trasmissioni radiofoniche o televisive). Essendo la partecipazione al sondaggio spontanea, i dati rilevati non sono estensibili all'intera popolazione di radio-telespettatori, né, a maggior ragione, alla popolazione degli italiani (essendo il campione – formato tramite un processo di **autoselezione** degli intervistati – di natura non probabilistica).

[29] L.A. Goodman (1961), "Snowball sampling", in *Annals of Mathematical Statistics*, 32, pp. 148-170.

in indagini su **popolazioni rare**[30] (cap. 6). La procedura di formazione del campione è la seguente: ad un gruppo iniziale di intervistati si chiede, dopo aver concluso l'intervista, i nominativi di altri soggetti appartenenti alla popolazione di interesse[31]; a questi ultimi si chiedono – progressivamente – liste (nomi e indirizzi) di altri nominativi, così via..., fino a formare il campione.

Esempio

Per svolgere un'indagine sulla comunità scientifica interessata alle applicazioni di alta tecnologia (ad esempio, quella laser) e individuare il *target group*, si può attuare la procedura seguente: si contattano organizzazioni specialistiche (uffici di ricerca nazionale, università, laboratori, aziende industriali) e ai soggetti intervistati si chiedono, alla fine dell'intervista, i nominativi di altri soggetti interessati alla tecnologia in questione, che entrano a far parte del campione; si procede in tal modo – sequenzialmente – fino a comporre il necessario campione.

Questo campionamento è adottato:

- in indagini *esplorative*;
- per l'analisi di *singoli casi*, preliminari ad una ricerca più vasta;
- in sondaggi su piccole comunità (gruppi etnici, clandestini) disperse territorialmente.

La procedura ingenera elevati rischi di **distorsione** delle stime.

5. Campioni probabilistici

I campioni probabilistici presuppongono che ciascuna unità della popolazione abbia una **probabilità nota** e **diversa da zero** di essere estratta e inclusa nel campione. Questi campioni, ottenuti sulla base di

[30] La letteratura corrente propone vari metodi campionari per le indagini su popolazioni rare o **elusive**, o per i casi in cui non è nota la dimensione e/o la localizzazione della popolazione (finita), o non sia possibile approntare una lista completa delle unità che compongono la popolazione. Tra tali approcci è da segnalare la teoria **dual/multiple frame** di campionamento per centri (C. Colleoni, "I campionamento da popolazioni *difficult-to-sample*: stato dell'arte e nuove prospettive", Università degli Studi di Milano, 2005).

[31] Ciò in quanto gli appartenenti a popolazioni rare (ad esempio, genitori di gemelli, esploratori di abissi marini, collezionisti di particolari oggetti d'arte, ecc.) conoscono altri soggetti che condividono la loro condizione.

modelli combinatori e probabilistici, vengono formati con *scelta casuale* delle unità campionarie.

Per **scelta casuale** si intende una procedura di selezione equivalente all'estrazione di palline numerate – di forma e peso uguali – da un'urna. Estrazione a caso non significa, perciò, scelta senza criterio, o *a casaccio*, delle unità del campione[32], al contrario, tale scelta è effettuata con metodo rigoroso, che garantisce l'imprevedibilità (aleatorietà) di un risultato tra tanti possibili[33].

Facendo riferimento, per fissare le idee, al campione **casuale semplice** (estrazione di palline da un'urna), si può convenire che nell'insieme di tutti i campioni possibili di una data numerosità (**universo dei campioni**) sono compresi campioni **non conformi** alla popolazione (*campioni difformi*)[34] di provenienza; tali campioni danno una stima distorta dei parametri[35] dell'universo ed hanno normalmente basse probabilità di essere estratti.

Un campione scelto casualmente è uno dei campioni possibili e il suo grado di **rappresentatività** (*conformità* alla popolazione di provenienza) non è determinabile. Tuttavia la teoria statistica assicura che i campioni probabilistici, appartenenti all'insieme dei possibili campioni, sono **rappresentativi** (essendo casuali e, quindi, non distorti per cause sistematiche), anche se **non tutti** i campioni riproducono con precisione un parametro della popolazione.

L'importanza dei campioni casuali è dovuta al fatto che sono note alcune relazioni che legano i *valori caratteristici* o **stimatori** dell'u-

[32] È da osservare, tuttavia, che se le unità campionarie sono scelte senza l'applicazione di criteri di casualità, ma in numero così ampio da far cogliere la variabilità del fenomeno indagato, il relativo campione può essere ritenuto **equivalente ad un campione casuale**.

[33] Se si lasciasse la scelta delle unità campionarie all'arbitrio del rilevatore il campione risulterebbe probabilmente **distorto**, in quanto le unità della popolazione non avrebbero tutte la stessa probabilità di essere selezionate, infatti, l'intervistatore tenderebbe ad avvicinare, in base a criteri soggettivi – anche inconsci –, determinate unità piuttosto che altre.

[34] Nell'universo campionario la presenza di campioni difformi rispetto ad un carattere (o parametro) della popolazione è dovuta – escludendo cause **sistematiche** di distorsione – a cause **accidentali**.

[35] Tali parametri possono essere: l'ammontare complessivo (*totale*) di un carattere, la sua media, varianza, ecc..

niverso dei campioni (media, varianza, momenti di ordine qualunque) ai **parametri** della popolazione.

I risultati di un'indagine per campione sono affetti – inevitabilmente – da un margine di errore: per interpretare correttamente tali risultati occorrono informazioni aggiuntive. Dette informazioni sono costituite dagli **intervalli di confidenza** (intervalli che contengono, verosimilmente, il valore del parametro allo studio con una probabilità elevata, ad esempio, del 95%[36]).

Tramite questi intervalli è possibile valutare la probabilità che la stima di un parametro della popolazione – ottenuta sulla base di un *solo campione* – sia affetta da un errore casuale di ampiezza predeterminata (**errore ammesso**). Ciò in virtù delle succitate relazioni, che collegano i principali parametri dell'universo campionario alla popolazione di riferimento.

5.1. *Significato dei principali termini di statistica inferenziale*

Di seguito si riporta sinteticamente il significato dei principali termini utilizzati nella presente trattazione.

- **Unità statistica**: elemento che presenta caratteristiche tali da farlo rientrare nel campo di osservazione di un'indagine. L'unità statistica può essere il consumatore, la famiglia, l'azienda, un distretto territoriale e simili.
- **Popolazione** o **universo**: insieme (finito o infinito) di unità che soddisfano una comune definizione (ovvero, che hanno in comune un carattere) e che costituiscono il *campo di osservazione*[37].
- **Popolazione finita**: insieme di unità statistiche di limitate dimensioni (ad esempio, numero di vetture circolanti in Italia nell'anno 2005).
- **Popolazione infinita**: insieme di unità statistiche molto elevato o infinito (astrazione concettuale).

[36] Siccome la probabilità *a priori* di incorrere in un campione al quale è associato un intervallo di stima del parametro valido è – nel caso in esempio – elevato (95%), si ha **fiducia** in tale intervallo.

[37] Ad esempio, in un'indagine sui consumatori di un prodotto appartenenti ad un'area commerciale la popolazione è costituita dai residenti nell'area; in un'indagine volta a stimare l'età media dei lettori di una certa rivista, la popolazione di riferimento è data dai soggetti (uomini e donne) definiti – sulla base di opportuni criteri – "lettori" della rivista.

- **Carattere continuo**: la variabile (X) di interesse presenta carattere di continuità (ad esempio: costo, reddito, lunghezza, ecc.) e può assumere tutti gli infiniti valori compresi tra due estremi di un intervallo (ad esempio, costi rilevati su n unità: $x_1 = 15{,}56$; $x_2 = 35{,}56$, ..., $x_n = 54{,}78$; v. capp. 6 - 7).
- **Carattere discreto** o **enumerabile**: il carattere (X) può assumere solo valori interi ($x_1 = 0$, $x_2 = 1$, $x_3 = 2$, ..., $x_n = n$; v. cap. 6).

  Alcuni caratteri presentano due sole modalità (ad esempio: buono, difettoso; uomo, donna; consumatore non consumatore, ecc.) e sono detti **dicotomici**; la popolazione di pertinenza è denominata *binomiale*.
- **Campione**: sottogruppo, parte, sottoinsieme, tratto dalla popolazione per studiare alcune caratteristiche della stessa.
- **Campione probabilistico** o **casuale**: sotto-insieme estratto con criteri di casualità da una popolazione, per la quale ciascuna unità ha probabilità nota *a priori* di essere scelta.
- **Campioni equiprobabili**: campioni per i quali ciascuna unità della popolazione di provenienza ha la *medesima* probabilità – nota e diversa da 0 – di essere estratta (ad esempio, campione casuale semplice).
- **Campioni con probabilità diseguali**: campioni per i quali le unità della popolazione hanno probabilità nota di essere estratte, ma tale probabilità non è uguale per tutte (ad esempio, campione stratificato con assegnazione *proporzionale* o *inversamente proporzionale* o con attribuzione *ottimale* di Neyman (cap. 3)).
- **Campione non probabilistico**: campione per il quale non è possibile attribuire una probabilità di estrazione alle unità della popolazione di provenienza del campione[38].
- **Piano di campionamento**: è costituito: *a)* da una *regola* che specifica la procedura di estrazione del campione dalla popolazione; *b)* da *formule* di stima dei parametri della popolazione, basate sui dati campionari.
- **Estrazione campionaria con reinserimento** (o bernoulliana): le unità sono estratte una di seguito all'altra e ciascuna di esse, dopo l'estrazione (ed osservazione), viene rimessa nell'urna, prima della successiva estrazione[39].
- **Estrazione campionaria senza reinserimento** (o **esaustiva**), tipica nelle ricerche di mercato, può essere effettuata in due modi:

[38] La selezione degli elementi campionari viene effettuata sulla base di criteri soggettivi; il criterio della casualità è applicato in senso molto lato (campioni **quasi probabilistici**), come nel campionamento *per quote*, o non è applicato affatto, come nel campione *a scelta ragionata*. I campioni non probabilistici (empirici), semplici da utilizzare, si diffusero verso il 1920 e furono largamente adottati fino al 1940. Successivamente essi furono sostituiti dai campioni probabilistici (statistici). Oggi le ricerche di mercato si avvalgono, principalmente, di campioni probabilistici (stratificato, a stadi, a grappoli) o semi-probabilistici (campione per quote).

[39] Ad ogni estrazione la composizione della popolazione non viene alterata (**popolazione costante**), ciascuna unità può essere prelevata più volte con la medesima probabilità.

a) con "schema successivo": prelevando una pallina di seguito all'altra, dall'urna (reale o simbolica), senza reinserire l'elemento estratto di volta in volta;
b) con "schema in blocco", estraendo simultaneamente un gruppo di unità.

In entrambi i casi ciascuna unità può essere estratta una sola volta e la composizione dell'urna viene alterata ad ogni estrazione. Pertanto, dopo ogni prelievo, la probabilità di scelta delle restanti unità è influenzata dalle precedenti estrazioni (unità selezionate *dipendenti* tra di loro).

- **Probabilità di un evento**. La realizzazione di un'indagine campionaria, e la valutazione dei relativi risultati, si basano sul concetto di **caso** (v. *ultra*) e di **probabilità**.

  In letteratura si riportano varie definizioni di probabilità, ma dal punto di vista interpretativo e concettuale si hanno – sostanzialmente – due posizioni, tra loro contrapposte, che danno luogo a due impostazioni: quella *oggettivista* e quella *soggettivista*[40].

  La definizione **oggettivista** definisce la probabilità muovendo dalla considerazione che se non è possibile prevedere il risultato di una singola prova (potendo questo essere costituito da uno qualsiasi dell'insieme dei possibili risultati), si può, però, prevedere quello ottenibile su un gran numero di prove[41]. Ad esempio, una compagnia di assicurazioni non è in grado di predire se un dato appartamento subirà o meno un furto, ma essa può calcolare – sulla base del numero di furti rilevati, nel tempo, su appartamenti simili (basi tecniche) –, con un certo grado di precisione, la probabilità che l'immobile in questione sia oggetto di un furto.

  Secondo l'impostazione **soggettivista** invece, un "evento" è ritenuto possibile in termini di giudizio di un valutatore. Il valore numerico esprimente il grado di fiducia che un soggetto ripone – sulla base di informazioni raccolte[42] – nell'accadimento di un evento è detto "probabilità" dell'evento[43].

---

[40] Esiste anche l'"impostazione assiomatica" del calcolo delle probabilità, che assume come primitivi i concetti di **prova**, **evento** e **probabilità** (la *prova* genera l'*evento*, il quale si può presentare con una certa *probabilità*). Tali concetti vengono collegati tra di loro da alcuni postulati e teoremi per l'elaborazione della teoria (senza la necessità, peraltro, di definire esplicitamente il concetto di probabilità; G. Pompilj, *Le variabili casuali – Assiomatizzazione del calcolo delle probabilità*, vol. I, Eredi Veschi, Roma, 1967; B. De Finetti, *Teoria della probabilità*, Einaudi, Torino, 1970; G. Dall'Aglio, *Appunti sulle variabili casuali*, Università degli Studi "La Sapienza" di Roma, Facoltà di Scienze Statistiche ed Attuariali, Roma, 1964).

[41] «La osservazione che eventi imprevedibili singolarmente risultino invece regolari nel complesso è, probabilmente, vecchia quanto il mondo»; V. Castellano (*Istituzioni di statistica*, Edizioni Ilardi, Roma, 1965, p. 475).

[42] L'impostazione soggettivista della probabilità (B. De Finetti, L. J. Savane) risulta interessante per le ricerche di mercato, essa potrebbe essere utilizzata, ad esempio, nelle valutazioni economiche di esperti o per stimare la probabilità di futuri acquisti dei consumatori.

[43] Attraverso questo approccio il soggettivista – sulla base delle proprie convinzioni – riesce ad attribuire un valore di probabilità a qualsiasi fenomeno, ad esempio, che

8. Proprietà dell'universo campionario

È detto **universo dei campioni** l'insieme dei campioni possibili di *n* unità che si possono estrarre da un collettivo o popolazione attraverso una data operazione di scelta[56].

Con riferimento ai campioni probabilistici – estratti, cioè, secondo criteri di **casualità** e sulla base di una *precisa probabilità* di selezione, *nota* per tutti gli elementi della popolazione considerata – è possibile identificare un campione, nell'universo dei possibili campioni estraibili casualmente (Quadro 4), che consente un raccordo del risultato campionario con la popolazione di provenienza.

È noto, infatti, che nell'universo campionario alcune statistiche (media, varianza) del carattere allo studio sono collegate ai corrispondenti valori incogniti dei parametri da stimare.

[56] La *teoria dei campioni* si fonda sulla costruzione teorica dell'**universo dei campioni** (insieme dei possibili campioni, ognuno considerato con la probabilità che gli compete), sulla base della quale le proprietà dei diversi piani di campionamento *non* sono riferite ad un *singolo* campione [estratto], ma all'insieme di tutti i piani corrispondenti al tipo di campionamento considerato.

Quadro 4 - Universo dei campioni (continua)

L'insieme dei possibili campioni estraibili da una popolazione, con determinate procedure di scelta casuale, forma l'**universo dei campioni**. Il numero di tali campioni è determinato dal numero dei diversi modi – schematizzabili tramite il **calcolo combinatorio** – nei quali le unità elementari si possono "combinare" nel comporre il campione.

Data una popolazione di N unità (indicate con le lettere: A, B, C, D,...), l'estrazione campionaria di $n = 2$ elementi può dar luogo ai seguenti campioni:

$$AB,\ AC,\ BA,\ BC,\ \ldots$$

Il primo ed il terzo di tali campioni differiscono solo per l'ordine di estrazione delle unità. Questi due campioni (che presentano le stesse unità, ma con ordine differente) possono essere ritenuti: *a)* **coincidenti** se non interessa l'ordine di uscita delle unità elementari; *b)* **distinti**, se si tiene conto di tale ordine.

Se la popolazione è **infinita** si può estrarre un numero pure infinito di campioni costituiti da n unità, sia nell'ipotesi di **reinserimento**[57], sia in quella di **non reinserimento**, sia che i campioni estratti – con stessi elementi – siano considerati distinti o coincidenti.

Se, invece, la popolazione è **finita**, di dimensione N, e da questa si estraggono campioni di dimensione n, occorre fare distinzione tra lo schema di estrazione *con reinserimento* (il campionamento non esaurisce mai la popolazione, come se questa fosse di dimensione infinita) e quello di estrazione *senza reinserimento* o *in blocco*.

Questo secondo tipo di schema ha maggiore interesse pratico e rispecchia le procedure di campionamento adottate nelle ricerche di mercato, ma è più complesso da studiare; infatti, mentre nello schema con reinserimento le singole estrazioni sono tra loro *indipendenti*, nel secondo caso – venendo meno, ad ogni estrazione, la ricomposizione dell'urna – i risultati di estrazioni successive risultano tra loro *dipendenti*.

Campioni estraibili da una popolazione finita (elementi di calcolo Combinatorio)

Se la popolazione è finita, di dimensione N, e da questa si selezionano campioni di numerosità n, occorre, innanzitutto, fare distinzione tra lo schema di estrazione con reinserimento e quella senza reinserimento.

Nel secondo caso, dopo ogni estrazione, la probabilità che le restanti unità vengano selezionate si modifica e risulta influenzata dal risultato delle precedenti estrazioni.

Il **Calcolo Combinatorio** consente di "contare" i raggruppamenti di diverso tipo

[57] Nello schema di estrazione del campione **con reinserimento** le unità che vengono man mano estratte sono rimesse nella popolazione (urna virtuale) e possono, quindi, essere prelevate successivamente.

Nello schema **senza reinserimento** – adottato nelle ricerche di mercato e nei sondaggi di opinione – le unità campionarie non sono reinserite nella popolazione e, pertanto, non possono essere rilevate più volte.

Quadro 4 - Universo dei campioni (segue)

che si possono formare con determinati oggetti.

1. Estrazione con reinserimento

Secondo questo schema – il più generale – i campioni estraibili da una popolazione possono essere formati:

- da unità ripetute (AAA, BBB, ...);
- da unità differenti (ABD, ABC, ...);
- dalle stesse unità ma con ordine differente (ABD, BDA, ...).

Si dimostra che con questo schema il numero dei possibili campioni è pari a:

$$n_{camp} = N^n.$$

L'insieme di questi campioni costituisce l'**universo bernoulliano**: il più ampio.

Esempio: data una popolazione di $N = 5$ unità (A, B, C, D, E)[58], il numero dei diversi campioni di $n = 2$ elementi (campioni binari) che si possono estrarre con reinserimento è dato da:

$$n_{camp} = N^n = 5^2 = 25;$$

infatti, i campioni possibili sono:

AA	AB	AC	AD	AE
BA	BB	BC	BD	BE
CA	CB	CC	CD	CE
DA	DB	DC	DD	DE
EA	EB	EC	ED	EE

I campioni con unità ripetute sono collocati sulla diagonale principale, quelli composti dalle stesse unità, ma di ordine inverso, sono disposti simmetricamente nel triangolo superiore ed inferiore della matrice.

Appena N e n diventano un po' elevati, il numero dei possibili campioni diviene molto grande. Ad esempio, da una popolazione di $N = 10$ unità si possono estrarre 10.000.000 campioni possibili composti da 7 unità, infatti:

[58] In questo esempio, come negli altri che seguiranno, vengono considerate popolazioni di poche unità, al fine di descrivere le proprietà dei campioni e per verificare con più facilità le formule di calcolo del numero di campioni possibili; l'estensione di dette proprietà a campioni estratti da **grandi popolazioni** (che si incontrano nella pratica) risulterà immediata.

$$n_{camp} = N^n = 10^7 = 10.000.000.$$

2. Estrazione senza reinserimento

È quello maggiormente utilizzato nelle ricerche di mercato e nei sondaggi di opinione, poiché le unità campionarie estratte non vengono reinserite nella popolazione (per evitare distorsioni nei risultati) e non possono, quindi, essere successivamente prelevate.

L'ipotesi di non reinserimento sussiste anche nell'estrazione **in blocco**, con la quale le unità campionarie sono prelevate simultaneamente.

Nell'estrazione *senza reinserimento* si distinguono due procedure – alternative – di estrazione, a seconda che venga tenuto presente l'ordine di selezione delle unità (**disposizioni**) o che vi si prescinda (**combinazioni**).

a) Disposizioni

Se si tiene conto dell'ordine di uscita delle unità campionarie, il numero dei possibili campioni è dato, secondo il calcolo combinatorio, da:

$$n_{camp} = N \cdot (N - 1) \cdot (N - 2) \cdot \ldots \cdot (N - n + 1),$$

cioè: «*il numero delle disposizioni semplici di N oggetti, presi ad n ad n* (con $n < N$), è dato dal prodotto dei primi n numeri interi decrescenti a partire da n»[59].

Detto numero è equivalente a quello che dà la seguente formula:

$$D_{N,n} = \frac{N!}{(N-n)!},$$

dove $N!$ si legge «*enne* fattoriale»; esso indica il prodotto dei primi N numeri interi; $(N - n)!$ ha un significato analogo.

Esercizio 1

Da una popolazione di $N = 5$ unità (A, B, C, D, E) i campioni possibili che si possono estrarre – senza reinserimento – di $n = 2$ unità, tenendo conto dell'ordine di uscita delle stesse, è pari a:

$$n_{camp} = N \cdot (N - 1) \cdot \ldots \cdot (N - n + 1) = 5 \cdot 4 = 20;$$

oppure, in base alla seconda formula sopra riportata:

[59] Nel calcolo combinatorio di definiscono «disposizioni semplici di N oggetti di classe n» tutti i possibili raggruppamenti che si possono formare con gli N oggetti, presi ad n ad n, in modo che differiscano tra di loro per un oggetto o, almeno, per l'ordine.

$$D_{5,2} = \frac{5!}{(5-2)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 20.$$

Dalla popolazione considerata si possono estrarre, quindi, 20 campioni differenti, infatti le disposizioni di classe 2 sono le seguenti:

AB	*BA*	*BC*	*CB*	*CD*	*DC*	*DE*	*ED*
AC	*CA*	*BD*	*DB*	*CE*	*EC*		
AD	*DA*	*BE*	*EB*				
AE	*EA*.						

b) Permutazioni

Si definiscono «*permutazioni semplici di N oggetti*» le disposizioni di classe N, per le quali ciascun raggruppamento contiene tutti gli oggetti e differisce dagli altri gruppi solo per l'ordine degli elementi.

Il numero delle permutazioni di N oggetti è dato da:

$$P_N = N!$$

cioè: «*il numero delle permutazioni semplici di N oggetti è uguale al prodotto dei primi N numeri interi*».

Esercizio 2

Da una popolazione di $N = 4$ unità (A, B, C, D) il numero di campioni possibili di $n = 4$ unità che si possono estrarre è dato da:

$$P_4 = 4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24.$$

Tali permutazioni sono le seguenti:

ABCD	*BACD*	*CABD*	*DABC*
ABDC	*BADC*	*CADB*	*DACB*
ACBD	*BCAD*	*CBAD*	*DBAC*
ACDB	*BCDA*	*CBDA*	*DBCA*
ADBC	*BDAC*	*CDAB*	*DCAB*
ADCB	*BDCA*	*CDBA*	*DCBA*.

c) Combinazioni

Se si prescinde dall'ordine di uscita delle unità, il numero dei possibili campioni è dato da:

Quadro 4 - Universo dei campioni (segue)

$$n_{camp} = \frac{N \cdot (N-1) \cdot (N-2) \cdot \ldots \cdot (N-n+1)}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot n} = \binom{N}{n}^{60},$$

cioè: «*il numero delle combinazioni semplici di N oggetti presi ad n ad n è pari al rapporto fra il prodotto dei primi n numeri interi decrescenti a partire da N ed il fattoriale di n*»[61].

Tale numero è equivalente a quello che dà la seguente formula:

$$D_{N,n} = \frac{N!}{n!(N-n)!}$$

Esercizio 3

Dalla popolazione dell'esercizio 1, di $N = 5$ elementi si possono estrarre campioni di $n = 2$ unità senza reinserimento (considerando, quindi, coincidenti i campioni che figurano con le stesse unità e con ordine differente), in numero pari a:

$$C_{N,n} = n_{camp} = \binom{N}{n} = \frac{5!}{2!(5-2)!} = 10.$$

Le combinazioni di 5 oggetti, di classe 2, sono, infatti:

AB BC CD DE
AC BD CE
AD BE
AE.

Il numero di possibili campioni, pari a 10, è inferiore al numero di raggruppamenti ottenuti con le disposizioni (uguale a 20).

Data una **popolazione finita**[62] di cinque elementi (famiglie) che presentano le modalità quantitative (reddito mensile) riportate in Tab. 4, l'universo dei campioni di due unità, estratte senza ripetizione, è costituito dalle **combinazioni** *di cinque elementi presi a due a due*[63], indicate con il simbolo $\binom{5}{2}$, riportate in Tab. 5.

[60] L'espressione: $\binom{N}{n}$ è detta **coefficiente binomiale**, esso indica il rapporto tra il prodotto di n numeri interi consecutivi e decrescenti, partendo da N, ed il prodotto di n numeri interi e crescenti muovendo da 1.

[61] Si denominano «*combinazioni semplici di N oggetti di classe n*» tutti i raggruppamenti che si possono formare con gli N oggetti, presi ad n ad n, che differiscono tra di loro almeno per un oggetto.

[62] Per *popolazione finita* si intende un insieme finito e noto di unità statistiche che, pur se numerose, sono *enumerabili* e *identificabili* (cioè, etichettabili).

Tab. 4 - Popolazione di famiglie

Famiglie	Reddito mensile (migliaia di euro)
A	1,50
B	3,15
C	1,20
D	1,35
E	1,05

Si noti che la *media* dello **stimatore** della variabile X, cioè $E(\bar{x})$ (con $E(\cdot)$ si intende il valore della media (*expectation*) della variabile indicata nell'argomento), ottenuta dalle medie ($\bar{x}$) dei campioni, è uguale alla *media* μ della **popolazione**[64]. Quest'ultima, infatti, è:

$$\mu = (1{,}50 + 3{,}15 + 1{,}20 + 1{,}35 + 1{,}05)/5 = 1{,}65,$$

e quella della v.c $\bar{x}$ è data da:

$$E(\bar{x}) = (2{,}325 + 1{,}350 + 1{,}425 + 1{,}275 + 2{,}175 + 2{,}250 + 2{,}100 + 1{,}275 + 1{,}125 + 1{,}200)/10 = 1{,}65^{65}.$$

[63] Le **combinazioni** di $N = 5$ unità a 2 a 2 sono di numero pari a: $\binom{5}{2} = \frac{5 \times 4}{2 \times 1} = 10$ (Quadro 4). Si fa ricorso al numero di *combinazioni* – per determinare l'universo dei possibili campioni – quando non interessa l'*ordine* di estrazione delle diverse unità campionarie; se tale ordine *è rilevante* si fa riferimento alle **disposizioni**, il cui numero, nell'esempio, è pari a: $5!/3! = 5 \cdot 4 = 20$.

[64] In questo caso si dice che la media di ogni possibile campione è **stima corretta** (*unbased estimate*) della media della popolazione.

[65] Si osservi che quanto sopra (cioè: $E(\bar{x}) = \mu$) vale anche per l'universo campionario delle **disposizioni** binarie **con ripetizione**, composto da N^n campioni possibili.

Tab. 5 - Campioni di $n = 2$ elementi (famiglie) estratti da un collettivo di $N = 5$

Combinazioni di elementi (famiglie)	Reddito mensile totale	Reddito mensile medio $\bar{x}$
AB	1,50 + 3,15 = 4,65	2,325
AC	1,50 + 1,20 = 2,70	1,350
AD	1,50 + 1,35 = 2,85	1,425
AE	1,50 + 1,05 = 2,55	1,275
BC	3,15 + 1,20 = 4,35	2,175
BD	3,15 + 1,35 = 4,50	2,250
BE	3,15 + 1,05 = 4,20	2,100
CD	1,20 + 1,35 = 2,55	1,275
CE	1,20 + 1,05 = 2,25	1,125
DE	1,35 + 1,05 = 2,40	1,200
	Totale	16,500

Il precedente risultato è generalizzabile. Nello schema di estrazione **senza ripetizione** la **media campionaria** è:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{N} D_i X_i}{n},$$

dove D_i è una **variabile indicatrice** che assume solo i valori 0 o 1, a seconda che nel campione sia assente o presente, rispettivamente, l'unità i-ma della popolazione da cui proviene il campione.

Pertanto si ha:

$$E(\bar{x}) = \frac{\sum_{i=1}^{N} [E(D_i) X_i]}{n} = \frac{\sum_{i=1}^{N} X_i}{N} = \mu \text{ [66]}, \qquad [1]$$

tenendo presente che nella precedente espressione è: $E(D_i) = \frac{n}{N}$.

[66] Essendo la media delle medie di tutti i possibili campioni ($E(\bar{x})$) uguale a μ (media della popolazione), la media $\bar{x}$ calcolata su un singolo campione è una stima corretta (**stimatore corretto**), non affetta da errore sistematico, di μ.

Indicando con: $Var(\bar{x})$ la varianza delle medie campionarie; con σ^2 la varianza della popolazione; con n e N, rispettivamente, la dimensione del campione e della popolazione, si dimostra che è:

$$Var(\bar{x}) = \frac{\sigma^2}{n}, \qquad [2]$$

per campioni **con ripetizione**, ovvero, nel caso di popolazione **infinita** o molto grande.

È, invece:

$$Var(\bar{x}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} \text{ [67]}, \qquad [3]$$

per campioni **senza ripetizione** ovvero, nel caso di popolazione **finita**.

Nella presente trattazione si fa riferimento a **popolazioni finite**.

Quadro 5 - Calcolo della varianza della distribuzione delle medie campionarie con la formula indiretta e diretta, per differenti valori di n (continua)

Formula indiretta

La [3], riferita al caso di campioni senza ripetizione, può essere accertata con i dati relativi al precedente esempio, inerenti ai redditi mensili del collettivo di cinque famiglie.

Per $n = 2$, essendo (Tab. 4):

$$\sigma^2 = \frac{(1{,}50-1{,}65)^2 + (3{,}15-1{,}65)^2 + \ldots + (1{,}05-1{,}65)^2}{5} = 0{,}585,$$

si ha:

$$Var(\bar{x}) = \frac{0{,}585}{2} \cdot \frac{5-2}{5-1} = 0{,}219.$$

[67] Per la dimostrazione v. Giardina (1990, pp. 228-231).

Quadro 5 - Calcolo della varianza della distribuzione delle medie campionarie con la formula indiretta e diretta, per differenti valori di *n* (segue)

Formula diretta

Se si calcola direttamente la varianza $Var(\bar{x})$ dei valori dei redditi medi si ottiene:

$$Var(\bar{x}) = \frac{(2,325-1,650)^2 + (1,350-1,650)^2 + \ldots + (1,200-1,650)^2}{10} = 0,219. \quad c.v.d.$$

Per $n = 3$ l'universo dei campioni di tre unità, estratte senza ripetizione, è costituito dalle **combinazioni** *di cinque elementi presi a tre a tre*; queste sono riportate in Tab. 6, unitamente ai corrispondenti valori medi.

Tab. 6 - Campioni di $n = 3$ elementi (famiglie) estratti da un collettivo di $N = 5$

Combinazioni di elementi (famiglie)	Reddito mensile totale	Reddito mensile medio $\bar{x}$
ABC	1,50 + 3,15 + 1,20 = 5,85	1,95
ABD	1,50 + 3,15 + 1,35 = 6,00	2,00
ABE	1,50 + 3,15 + 1,05 = 5,70	1,90
ACD	1,50 + 1,20 + 1,35 = 4,05	1,35
ACE	1,50 + 1,20 + 1,05 = 3,75	1,25
ADE	1,50 + 1,35 + 1,05 = 3,90	1,30
BCD	3,15 + 1,20 + 1,35 = 5,70	1,90
BCE	3,15 + 1,20 + 1,05 = 5,40	1,80
BDE	3,15 + 1,35 + 1,05 = 5,55	1,85
CDE	1,20 + 1,35 + 1,05 = 3,60	1,20
	Totale	16,50

La media delle medie dei campioni è: $E(\bar{x}) = 16,50/10 = 1,65$. Anche in questo caso il valore di $E(\bar{x})$ coincide con $\mu = 1,65$.

Si ha, inoltre:

$$Var(\bar{x}) = \frac{0,585}{3} \cdot \frac{5-3}{5-1} = 0,0975.$$

Il calcolo diretto (effettuato sui valori medi riportati in Tab. 6) dello **scarto quadratico medio** (s.q.m.) dei redditi medi dà 0,0975.

Per $n = 4$ l'universo dei campioni di quattro unità, estratte senza ripetizione, è costituito dalle **combinazioni** *di cinque elementi presi a quattro a quattro* riportate in Tab. 7, unitamente ai valori medi.

Quadro 5 - Calcolo della varianza della distribuzione delle medie campionarie con la formula indiretta e diretta, per differenti valori di *n* (segue)

Tab. 7 - Campioni di $n = 4$ elementi (famiglie) estratti da un collettivodi $N = 5$

Combinazioni di elementi (famiglie)	Reddito mensile totale	Reddito mensile medio $\bar{x}$
ABCD	1,500 + 3,150+ 1,200 + 1,350 = 7,200	1,8000
ABCE	1,500 + 3,150+ 1,200 + 1,050 = 6,900	1,7250
ABDE	1,500 + 3,150+ 1,350 + 1,050 = 7,050	1,7625
ACDE	1,500 + 1,200+ 1,350 + 1,050 = 5,100	1,2750
BCDE	3,150 + 1,200+ 1,350 + 1,050 = 6,750	1,6875
	Totale	8,2500

La media delle medie dei campioni è: $E(\bar{x}) = 8,250/5 = 1,65$. Anche in questo caso il valore di $E(\bar{x})$ coincide con $\mu = 1,65$.

Si ha, inoltre:

$$Var(\bar{x}) = \frac{0,585}{4} \cdot \frac{5-4}{5-1} = 0,0366.$$

Il calcolo diretto (effettuato sui valori medi riportati in Tab. 7) dello s.q.m. dei redditi medi dà nuovamente 0,0366.

Raggruppando le medie – calcolate sui campioni di differenti dimensioni – in classi di reddito, si giunge alla **distribuzione delle medie campionarie** riportata in Tab. 8.

Da detta tabella si evince che all'aumentare di n (da 2 a 4) la distribuzione di frequenza delle medie $\bar{x}$ presenta un addensamento di valori attorno alla media della popolazione (media delle medie) – classe incorniciata – nella quale cade $E(\bar{x}) = \mu = 1,65$. Detto addensamento evidenzia che all'aumentare di n da 2 a 4 diminuisce la dispersione delle medie (cioè il valore dello s.q.m)[68].

Si deduce, quindi – per via empirica – che al crescere di n la distribuzione delle medie campionarie tende – asintoticamente – alla normale.

[68] La tabella sottostante evidenzia la riduzione dello s.q.m. al crescere di *n*.

Dimensione del campione	s.q.m. delle medie campionarie
2	0,4684
3	0,3123
4	0,1912

Quadro 5 - Calcolo della varianza della distribuzione delle medie campionarie con la formula indiretta e diretta, per differenti valori di *n* (segue)

Tab. 8 - Frequenza di campioni di *n* unità per classi di reddito medio						
Classi di reddito medio	*Frequenza di campioni di n unità ricadenti nelle classi di reddito medio indicate*					
	$n=2$		$n=3$		$n=4$	
1.000-1.199	x	1	-		-	
1.200-1.349	xxx	2	xxx	3	x	1
1.350-1.499	xx	3	x	1	-	
1.500-1.649	-		-		-	
1.650-1.799	-		-		xxx	3
1.800-1.949	-		xxxx	4	x	1
1.950-2.099	-		xx	2	-	
2.100-2.249	xx	2	-		-	
2.250-2.399	xx	2	-		-	
2.400-2.549	-		-		-	
2.550-2.699	-		-		-	
2.700-2.849	-		-		-	
2.850-2.999	-		-		-	
3.000-3.149	-		-		-	
3.150-3.299	-		-		-	
N. totale di campioni		10		10		5
media dei redditi medi		1,65		1,65		1,65
s.q.m.		0,4684		0,3123		0,1912

È noto che se X e Y sono variabili casuali **normali**, allora qualunque combinazione lineare z:

$$z = aX + bY, \qquad \text{per: } a, b \neq 0,$$

è una variabile casuale (v. c.) *normale*.

Ciò consente di dimostrare che la distribuzione campionaria delle medie è **normale**, se la popolazione è normale.

Nel caso in cui la popolazione dalla quale è estratto il campione non è normale si può sostenere ancora la normalità **asintotica** della distribuzione delle medie campionarie, in virtù del **teorema del limite centrale**, detto anche di *convergenza stocastica*.

8.1. *Teorema del limite centrale*

Questo teorema afferma che, indipendentemente dalla circostanza che la v. c. X in una popolazione *infinita* sia distribuita normalmente o no, se la *media* μ e la *varianza* σ^2 (parametri) di tale popolazione sono *finite*, iterando l'operazione di campionamento, la distribuzione delle medie campionarie **approssima** – all'aumentare della dimensione campionaria n (in pratica, se è $n > 30$, caso di **grandi campioni**, ai quali si fa riferimento nella presente trattazione[69]) – una distribuzione **normale** con media $E(\bar{x})$ (media dei valori medi campionari $\bar{x}$) pari a μ e varianza $Var(\bar{x})$ data da: $Var(\bar{x}) = \frac{\sigma^2}{n}$.

I risultati ora richiamati rappresentano l'impianto della **teoria dei campioni**. Infatti, considerando l'universo dei possibili campioni di dimensione n, il valore medio delle medie campionarie è uguale al valore medio incognito da stimare sulla popolazione e la distribuzione di dette medie campionarie approssima le legge normale.

Grazie a quest'ultimo risultato è possibile conoscere l'aliquota dei campioni dell'universo campionario che presenta un dato scarto dal valore μ. Si può, allora, fissare un elevato valore di questa aliquota, ad esempio, pari al 95% o al 99% e, se vale l'ipotesi di **normalità**, si potrà ritenere che la media aritmetica del 95% o del 99% dei campioni si discosterà da quella incognita della popolazione non oltre, rispettivamente, il valore di $1{,}96 \cdot \sigma/\sqrt{n}$ (per una percentuale del 95%) o il valore $2{,}56 \cdot \sigma/\sqrt{n}$ (per il 99%). In altre parole, nel primo caso, 95 medie su 100 non si discosteranno – per eccesso o per difetto – più dell'anzidetto valore dalla media esatta cercata (μ). I coefficienti 1,96 e 2,56 sono ricavati dalla tavola della distribuzione normale (Tab. 9).

Operativamente si osserverà, però, un solo campione e ci si chiederà se esso appartiene alla categoria dei campioni "buoni" (**conformi**), che hanno cioè una media non molto differente da quella della popolazione, o a quella dei campioni "non buoni" (**non conformi**), aventi una media

[69] Nel caso di piccoli campioni ($n < 30$), provenienti da una popolazione distribuita normalmente, in luogo della variabile normale standardizzata z (v. Quadro 6) si utilizza la variabile t di **Student** (Quadro 10).

sottostimata o sovrastimata per più di $1{,}96 \cdot \sigma / \sqrt{n}$ volte la media esatta μ. Siccome i campioni "buoni" sono 95 (contro i "non buoni", che sono 5), si confiderà che il campione estratto dia un risultato rientrante nella prima categoria. Il rischio che questa ipotesi sia errata è, perciò, del 5%.

9. Campionamento casuale semplice

Nel campionamento casuale semplice le unità della popolazione hanno tutte la **medesima probabilità** di essere incluse nel campione.

Come già rilevato, lo schema di estrazione degli elementi campionari può essere di due tipi, estrazione con o senza ripetizione.

Nello schema **con ripetizione** le palline estratte dall'urna sono, man mano, reimmesse nella stessa e la probabilità di estrazione di un elemento resta costante ad ogni estrazione (*estrazioni indipendenti* tra di loro). Pertanto, con questo schema – non utilizzato nelle ricerche di mercato – una stessa unità può essere estratta più volte.

Il campionamento casuale semplice **senza ripetizione** rappresenta lo schema di base e dà una formazione del campione semplice ed intuitiva. Le unità campionarie vengono estratte una ad una, escludendo dalla popolazione quelle di volta in volta prelevate; in tal modo la probabilità di estrazione di un'unità varia ad ogni estrazione (*eventi condizionati*).

Questo schema costituisce il termine di confronto per la misura dell'efficienza delle altre strategie di campionamento (cap. 3).

Il modo più semplice per realizzare questa procedura di estrazione è il seguente: alle N unità della popolazione – numerate da 1 a N – si fanno corrispondere N palline; inserite queste nell'urna e mescolatele, se ne estraggono n, una dopo l'altra, senza reimmissione. Il campione sarà composto dalle n unità corrispondenti ai numeri estratti.

Quando la dimensione della popolazione è elevata il procedimento di estrazione descritto viene simulato attraverso una **tavola dei numeri casuali** (par. 7.1) o l'uso di *routines* di calcolo (disponibili in EXCEL, SAS, SPSS, STATISTICA, ecc.), idonee a produrre successioni di numeri **pseudocasuali**, compresi nell'intervallo $[1, N]$.

9.1. *Dimensione campionaria e stimatori*

Uno dei problemi più rilevanti in un'indagine statistica campionaria è quello della determinazione della dimensione del campione (probabilistico). Pur essendo intuitivo che al crescere della grandezza del campione cresce l'**attendibilità** delle stime campionarie, la scelta della

numerosità campionaria più opportuna deve essere effettuata sulla base di regole derivanti dalla **teoria dei campioni**.

Si tratta di scegliere la dimensione minima del campione che riduca i costi della ricerca e che garantisca, nel contempo, la desiderata precisione delle stime.

Per **precisione** delle stime si intende il margine di **errore ammesso**, con un prefissato **livello di confidenza** (probabilità che la stima campionaria cada all'interno dell'intervallo definito da detto errore).

La teoria dei campioni fornisce – per i vari tipi di campionamento – le formule di calcolo della numerosità campionaria, sulla base del margine di errore tollerato e del livello di confidenza, nota la dimensione della popolazione su cui svolgere l'indagine, nel caso questa sia finita.

A questo proposito, nella realtà operativa si possono presentare due casi contrapposti.

a) La **dimensione** campionaria n – scelto lo schema di campionamento – viene determinata sulla base della *formula* del piano prescelto, la quale considera: la grandezza della popolazione, la varianza del carattere allo studio, l'errore ammesso ed il livello di confidenza della stima. Livelli di confidenza tipici sono quelli del 95% o del 99%, che si ritiene diano una pratica certezza di inclusione del parametro incognito (Tabb. 10 - 11).

b) La dimensione del campione – per un prefissato tipo di schema campionario – è determinata (ad esempio, dai limiti dello stanziamento complessivo della ricerca[78]); in tale circostanza, utilizzando la formula del calcolo di n, di cui al caso *a)*, si determina [a posteriori] l'**errore massimo** commesso per determinati livelli di confidenza[79] (Tab. 12).

Nel seguito (par. 10) verranno fornite le formule della dimensione campionaria per il caso di campionamento casuale **per attributi**, nel quale il parametro oggetto di interesse è costituito dalla **proporzione** o **frequenza relativa** di unità che nella popolazione possiedono un determinato attributo (ad esempio, proporzione di soggetti consumatori di un dato prodotto).

[78] Attualmente il costo base unitario di un sondaggio è di circa 36 euro per intervista (comprendente tutte le fasi della ricerca: dalla formulazione del questionario fino al rapporto sui risultati).

[79] In una ricerca per campione è possibile scegliere tra diverse combinazioni – alternative – delle due grandezze: *errore massimo*, *livello di confidenza*.

Successivamente (par. 11) verrà sinteticamente trattato il campionamento casuale **per variabili**. Il parametro di interesse è rappresentato in questo caso dal *valore medio* che un carattere quantitativo assume nella popolazione (ad esempio, il reddito medio delle famiglie di un collettivo).

Nelle ricerche di mercato gli stimatori più utilizzati sono la *proporzione* (o *frequenza relativa*), la *media* e la *varianza* campionaria.

La scelta dello specifico stimatore[80] viene effettuata sulla base delle **proprietà** godute dallo stesso e della natura del carattere statistico (qualitativo o quantitativo[81]) rilevato.

Indicando con θ un **valore caratteristico** (denominato **parametro**) della popolazione, che deve essere stimato, uno **stimatore** $\hat{\theta}$ (del parametro θ) è una funzione delle osservazioni – disponibili sul carattere allo studio – che permette di giungere ad una stima di θ. Il valore che la funzione campionaria assume in un campione è detto **stima**. Su un dato carattere si possono determinare stimatori diversi.

Uno stimatore può essere:

a) **corretto** o non distorto (*unbiased*) se il suo valore medio $E(\hat{\theta})$ è uguale a θ ($E(\cdot)$ indica, come noto, l'operatore matematico **valore atteso**). Uno stimatore è *distorto* se sussiste una scarto (Δ) tra i due valori, se si ha, cioè: $\Delta = E(\hat{\theta}) - \theta \neq 0$[82];

b) **consistente** se la sua precisione cresce al crescere della dimensione campionaria n, fino a convergere al valore del parametro per $n = N$, in tal caso è:

$$\lim_{n \to N} E(\hat{\theta}_n) = \theta$$

[80] Che può essere un indice di posizione (media analitica, moda, mediana) o di variabilità.

[81] Ad esempio, dal collettivo delle imprese italiane si può estrarre un campione di n imprese e rilevare un carattere **qualitativo** (il settore industriale di appartenenza delle unità campionarie, la loro collocazione territoriale o amministrativa, ecc.) o un carattere **quantitativo** (il numero dei dipendenti delle aziende estratte, l'ammontare del fatturato delle stesse, ecc.).

[82] Δ è ipotizzato *costante* su tutti i possibili campioni.

e la varianza dello stimatore è nulla;

c) **più efficiente**[83] di un altro stimatore se ha un più basso *errore quadratico medio* (*Mean Squared Error* – MSE). Quest'ultimo è costituito dalla media quadratica degli scarti tra i valori dello stimatore ($\hat{\theta}$) e il parametro (θ):

$$\text{MSE} = E(\hat{\theta} - \theta)^2.$$

Il MSE dà una misura inversa della **precisione** dello stimatore. Lo stimatore che presenta il MSE minore è *più efficiente*. Se lo stimatore è corretto il suo valore medio è uguale al parametro della popolazione e il MSE coincide con la varianza dello stimatore $E(\hat{\theta} - \theta)^2$; la precisione dello stimatore cresce al decrescere della sua varianza.

Se uno stimatore gode della proprietà *a*) (è scevro, quindi, da errori sistematici) le stime di un parametro ottenute su un campione probabilistico presentano differenze *casuali* dal parametro incognito (tali differenze si compensano in media tra di loro). Se uno stimatore gode anche della proprietà *b*), al crescere della dimensione campionaria *n* il valore dello stimatore tende a concentrarsi intorno al valore del parametro incognito e diviene sempre più improbabile avere differenze elevate tra il valore dello stimatore e quello del parametro[84].

Esistono vari metodi per determinare stimatori che garantiscono le proprietà sopra elencate: dei **minimi quadrati** (cap. 1, Quadro 5) dei **momenti**, della **massima verosimiglianza**. Per la trattazione di questi metodi si rinvia ai testi di statistica metodologica[85].

[83] L'efficienza di uno stimatore $\hat{\theta}'$ è data dal suo grado di precisione in rapporto ad un altro stimatore di riferimento ($\hat{\theta}$): **Efficienza** $(\hat{\theta}' \mid \hat{\theta}) = Var(\hat{\theta}') / Var(\hat{\theta})$; se il rapporto è minore dell'unità è maggiore l'efficienza dello stimatore $\hat{\theta}'$ rispetto a $\hat{\theta}$, a parità di numerosità campionaria, e viceversa.

[84] Uno stimatore – oltre alle tre precedenti proprietà può averne una quarta – inerente la *sufficienza*. Uno stimatore è **sufficiente** se utilizza tutte le informazioni disponibili. Ad esempio, la media aritmetica è sufficiente, ed è preferibile alla semisomma del valore minimo e massimo, che tiene conto di due sole informazioni.

[85] Ad esempio, A. Rizzi (1992); B. Giardina (1990).

La stima di un parametro incognito può essere di tipo puntuale o intervallare.

La prima è costituita da un valore unico (**stima puntuale**), la seconda consta di un intervallo (**stima intervallare**) di valori, entro i quali si suppone che cada il vero parametro incognito della popolazione, con un prefissato livello di confidenza.

La stima puntuale di un parametro ha un interesse limitato; è preferibile, perciò (oltre che più prudente), la stima intervallare.

10. Campionamento casuale semplice per attributi e determinazione della dimensione campionaria

Nella teoria dei campioni si fa distinzione tra la stima della **proporzione** (caso di attributo) e la stima del **valore medio** (caso di carattere quantitativo).

Nella realtà applicativa i due casi, in genere, coesistono in una medesima indagine, pertanto, nell'ambito di una ricerca di mercato – avuto riguardo ai principali obiettivi perseguiti – si determina cautelativamente la dimensione campionaria più elevata tra diverse alternative.

10.1. *Stima di una proporzione*

Nel caso di campionamento **per attributi** i parametri campionari di base sono formulati come di seguito descritto.

Si fa nuovamente riferimento, per semplicità espositiva, allo schema dell'urna. Questa contenga N palline, k delle quali contraddistinte con 1 e le restanti $(N - k)$ con 0.

La probabilità di estrarre dall'urna una pallina contrassegnata con 1 è P, mentre è $Q = (1 - P)$, ovvero $(1 - k/N)$, la probabilità di estrarre una pallina numerata 0. Il valore medio e la varianza del risultato sono, rispettivamente:

$$E(X) = 1\,P + 0\,P = P.$$

$$Var(X) = (1 - P)^2\,Q = PQ.$$

Le n successive estrazioni possono essere effettuate secondo i due seguenti schemi:

1) la pallina estratta è reimmessa – successivamente – di volta in volta nell'urna; quest'ultima mantiene fissa la sua composizione iniziale ad ogni estrazione (schema con **ripetizione** o **bernoulliana**). Il valore *medio* del risultato delle estrazioni è:

$$E(p) = \frac{\sum E(X)}{n} = \frac{nP}{n} = P.$$

Data l'indipendenza delle singole estrazioni, la *varianza* dello stimatore p si calcola ponderando la somma delle varianze di ognuna con il quadrato $1/n^2$:

$$Var(p) = \frac{1}{n^2}\sum PQ = \frac{nPQ}{n^2} = \frac{PQ}{n};$$

2) la pallina estratta non viene successivamente reimmessa nell'urna (estrazione **senza ripetizione**). Si dimostra allora che il valore medio degli n risultati è sempre P, mentre la varianza è:

$$Var(p) = \frac{N-n}{N}\frac{N}{N-1}\frac{PQ}{n} \cong (1-f)\frac{PQ}{n},$$

dove $(1 - f)$ è il **fattore correttivo** per popolazioni finite, con $f = n/N$, **frazione di sondaggio**. Questa varianza è sempre inferiore a quella del precedente caso **1)**, a parità di n.

Nel caso di estrazione *senza ripetizione* è noto che, per n sufficientemente **elevato** (in genere *non inferiore* a 30) e per P non *molto prossimo* a 0 o a 1, si può sostenere che nel 95% dei casi una determinazione campionaria p_c non risulterà esterna all'intervallo:

$$p \mp 2\sqrt{Var(p)}.$$

Si potrà, quindi, dire che nell'universo dei campioni di dimensione n, estratti senza ripetizione dalla popolazione allo studio, il valore medio delle determinazioni p_c coincide con il valore P della popolazione e che è, inoltre, possibile individuare un intervallo intorno a P entro il quale cadrà una determinazione p_c, con un prefissato livello (*Prob*) di probabilità (**livello di fiducia**).

Normalmente, però, non interessa determinare l'anzidetto intervallo: interessa invece risalire – **inferenza induttiva inversa**[86] – dal valore campionario p_c ad una valutazione del valore incognito P della popolazione, con un intervallo di **confidenza**, espresso in termini di p_c e di *Prob*.

Si tratta di attribuire la stessa probabilità *Prob* della precedente *relazione diretta*, agli intervalli *variabili* di confidenza (*relazione inversa*) del tipo:

$$p_c \mp z_{\alpha/2}\sqrt{Var(p)},$$

dove: la costante $z_{\alpha/2}$ è fissata in funzione di *Prob*, con l'ausilio della tavola della curva normale (Tab. 9); $Var(p)$ è una stima ottenuta sul campione estratto.

Per i **grandi campioni**, cioè di numerosità $n \geq 30$, sotto l'ipotesi *cruciale* di *normalità* della frequenza relativa o proporzione da stimare, su una popolazione *finita*, si può affermare che, con probabilità *Prob*, il parametro incognito P cade entro l'intervallo:

$$p_c \mp z_{\alpha/2}\sqrt{(1-f)\frac{p_c q_c}{n-1}}$$ [87],

dove è $q_c = 1 - p_c$.

[86] In quanto dal particolare (*campione*) si intende risalire al generale (*popolazione*).

[87] Per questo problema Cochran ha proposto come regola pratica, per accertare la **normalità** di una distribuzione, la seguente: $n\, p_c\, q_c > 9$. Pertanto se è $n = 42$ e $p = 0{,}28$, si ha: $npq = 8{,}47$ e non si è certi della richiesta normalità della distribuzione allo studio. Necessita allora considerare una numerosità minima pari a 45, che soddisfa alla precedente relazione empirica, infatti il risultato 9,07 supera il limite fissato.

10.2. *Determinazione della numerosità campionaria*

Per stimare una *proporzione*[88] o *frequenza relativa* (campionamento per attributi) tramite un campione casuale semplice, si indichi, innanzi tutto, con:

N la dimensione della popolazione;
P la **proporzione incognita**;
p la stima della **proporzione campionaria**;
Δ il margine di **errore ammesso** (la differenza – in più o in meno – della stima (p) rispetto al vero valore incognito del parametro (P) non deve superare Δ);
$(1-\alpha)$ il **livello di fiducia** della stima.

Nelle condizioni considerate p è uno stimatore corretto di P e, se la dimensione campionaria è sufficientemente elevata, per il teorema del limite centrale[89] p si distribuisce approssimativamente in modo normale, con media:

$$E(p) = P$$

e varianza:

$$Var(p) = \frac{P(1-P)}{n}\frac{N-n}{N-1}.$$

Si introduca ora la seguente variabile z, con media 0 e varianza 1:

$$z = \frac{p-P}{\sqrt{Var(p)}},$$

la quale si distribuisce approssimativamente come una normale standardizzata.

Fissato il *margine di errore* o di precisione Δ e il *livello di fiducia* $(1-\alpha)$, si ha:

[88] In letteratura è invalso l'uso di denominare p "percentuale", anche se trattasi propriamente di proporzione; la percentuale è, infatti, $p \cdot 100$.
[89] Per la dimostrazione del teorema si rinvia a B. Giardina, 1990, p. 200 e ss.

$$\text{Prob}\{|p-P| \le \Delta\} = \text{Prob}\left\{|z| \le \frac{\Delta}{\sqrt{Var(p)}}\right\} = 1-\alpha,$$

dove *Prob* indica, al solito, il valore di probabilità.

Dalla tavola della curva normale standardizzata si rileva che la precedente relazione si verifica per: $\frac{\Delta}{\sqrt{Var(p)}} = z_{\alpha/2}$[90],

da cui si ricava la:

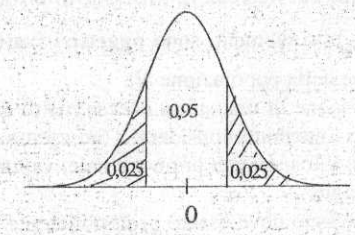
$$Var(p) = \frac{P(1-P)}{n}\frac{N-n}{N-1} = \frac{\Delta^2}{z_{\alpha/2}^2}. \qquad [6]$$

La [6] è l'espressione di un'equazione nell'incognita n, numerosità del campione. È facile, a questo punto, ricavare n in funzione dei restanti termini noti del problema:

$$n = \frac{z_{\alpha/2}^2 P(1-P)N}{(N-1)\Delta^2 + z_{\alpha/2}^2 P(1-P)}. \qquad [7]$$

L'errore massimo ammesso Δ può essere posto, ad esempio, pari a 1,96 lo s.q.m. della proporzione stimata p_c. In questo caso i limiti

[90] Il **percentile** $z_{\alpha/2}$, corrispondente al grado di fiducia $(1-\alpha)$, è il valore che nella tavola della curva normale standardizzata (Tab. 9) esclude dal valore dell'area totale (pari a 1) una frazione uguale a α: metà ($\alpha/2$) sulla destra e metà ($\alpha/2$) sulla sinistra della distribuzione, centrata sul valore medio (0). Ad esempio, per $(1-\alpha) = 0{,}95$, cioè, per $\alpha/2 = 0{,}025$ (come rappresentato nella sottostante figura) da Tab. 9 si deduce un valore $z_{0{,}025} = 1{,}96$.



Per $(1-\alpha) = 0{,}99$ è $z_{\alpha 2} = z_{0{,}005} = 2{,}58$.

dell'intervallo di stima (intervallo di **confidenza**) che ne derivano includeranno la proporzione incognita P nel 95% dei casi.

Se si sceglie, ad esempio, un livello di fiducia $(1- \alpha) = 95\%$, la [7] dà:

$$n = \frac{1{,}96^2 P(1-P)N}{(N-1)\Delta^2 + 1{,}96^2 P(1-P)}$$ [91].

Nella [7], noti o fissati i valori di N, Δ, $z_{\alpha/2}$, per ricavare n è necessario disporre di un valore di P[92], il quale è proprio il valore incognito cercato, obiettivo dell'indagine; si entra quindi in un *circolo chiuso* dal quale si può uscire in due modi alternativi:

a) tramite il ricorso ad un **sondaggio pilota**, cioè ad un campione ragionato di dimensione ridotta (che può dare, però, risultati distorti e scarsamente affidabili[93]), allo scopo di ottenere una stima preventiva di P (v. Tab. 11);

b) si fissa, cautelativamente, $P = 0{,}5$, determinando una stima **per eccesso** della numerosità n (v. Tab. 10). Infatti, considerando che la [7] è equivalente a:

[91] La dimensione del campione è funzione *1*) del tipo di *parametro* che si intende stimare o dell'*ipotesi* che si intende verificare, *2*) della *dimensione della popolazione* (finita o infinitamente grande), *3*) della *variabilità* del carattere sulla popolazione, *4*) dell'*errore* che si è disposti ad accettare, *5*) della *probabilità dell'errore* (livello di fiducia prescelto).

È da osservare che nella determinazione della grandezza campionaria, alcuni elementi sono **scelti** dal ricercatore (margine di errore Δ, livello di confidenza $(1 - \alpha)$ e, quindi, $z_{\alpha/2}$), altri elementi sono **oggettivi** (variabilità del carattere indagato $P(1 - P)$ e dimensione della popolazione N).

[92] A volte si dispone di una stima *preventiva* di P, cioè della proporzione delle unità che possiedono l'attributo considerato (ad esempio, proporzione presunta di aderenti ad un certo tipo di campagna promozionale, valutata sulla base dei risultati di passate analoghe campagne).

[93] Per questa ragione deve essere preferibilmente $n > 30$, per una sufficiente approssimazione delle medie campionarie alla distribuzione normale.

$$n = \frac{z^2_{\alpha/2} N}{\frac{(N-1)\Delta^2}{P(1-P)} + z^2_{\alpha/2}}$$ [8],

sostituendo nella precedente espressione al prodotto $P(1 - P)$ – che appare al denominatore e dal quale dipende il valore della varianza dello stimatore p (v. la [6]) – il suo valore massimo, cioè $max\ P(1 - P) = 0{,}25$[94], si giunge ad una **stima prudenziale** di n (la massima compatibile sotto le condizioni date).

Per $P = 0{,}5$ la [7] diviene (tenendo presente che è: $max\ P(1 - P) = 1/4$):

$$n = \frac{z^2_{\alpha/2} N}{4(N-1) + z^2_{\alpha/2}}$$ [9].

Quadro 7 - Applicazione: determinazione della dimensione campionaria per la stima di una proporzione su una popolazione finita (continua)

Si voglia determinare la numerosità di un campione di soggetti da intervistare, in una città di 90.000 abitanti, per stimare la proporzione P di possessori di carta di credito sull'intera popolazione. Si desidera che la stima p, al livello di confidenza del 95%, non si discosti da P più del 3% in valore assoluto (errore ammesso).

Ponendo $P = k/N$ la proporzione incognita di possessori di carta di credito nella popolazione, per n sufficientemente grande e $\mathbf{P} = \mathbf{Q}$, la frequenza relativa campionaria si distribuisce approssimativamente come una **normale**.

Per il caso di **popolazione finita** (come quello in argomento) è (v. la [6]) :

$$\sqrt{Var(p)} = \sqrt{\frac{PQ}{n}\frac{N-n}{N-1}}.$$

L'errore **massimo** (errore ammesso) che si è disposti ad accettare, al livello di confidenza del 95%, è, in termini di s.q.m. dello stimatore:

$$\Delta = z_\alpha \cdot \sqrt{Var(p)} = 1{,}96 \cdot \sqrt{0{,}25/n} = 0{,}03,$$

[94] P varia tra 0 e 1, come pure $(1 - P)$; il **massimo** della funzione di $P(1 - P)$ è pari a 0,25, che viene raggiunto per $P = 1 - P = 0{,}5$.

Quadro 7 - Applicazione: determinazione della dimensione campionaria per la stima di una proporzione su una popolazione finita (segue)

avendo posto $P = Q = 0,5$ (per massimizzare il prodotto PQ).

Il valore **minimo** di n che soddisfa le condizioni imposte è, per la [9]:

$$n = \frac{1,96^2 \times 90.000}{4 \times 89.999 \times 0,03^2 + 1,96^2} = 1.054,62 \cong 1.055 .$$

Nelle Tabb. 10, 11 sono riportate le dimensioni campionarie ricavate, rispettivamente, con la [9] e [8], per varie combinazioni: *a)* della dimensione della *popolazione* di riferimento (da $N = 500$ a $N = 1.000.000$)); *b)* del *margine di errore* Δ, indicato in percentuale[95] della stima ($\mp 1\%, \mp 2\%, ..., \mp 10\%$); *c)* del *livello di confidenza* ($1 - \alpha$) (pari a 95% e 99%).

La Tab. 10 riporta le grandezze campionarie, per le citate combinazioni, nel caso di $P = 0,50$, cioè di **varianza massima** dello stimatore p (campione cautelativo). In Tab. 11 figurano, invece, le dimensioni campionarie necessarie nel caso si disponga di una **stima preventiva** o preliminare di P, pari a 0,20 nel caso esemplificato. L'estensione di questa tavola per differenti valori di stima preventiva P è facilmente realizzabile dal ricercatore, tramite la [8].

Dalla lettura delle Tabb. 10, 11, si deducono le seguenti considerazioni:

- la dimensione campionaria varia al variare della grandezza della popolazione – in modo piuttosto *debole* – a partire da un margine di errore del 2% e da un valore di $N \geq 2.000$;
- la variabilità in argomento, misurata con il *campo di variazione* o *range*[96] si riduce decisamente all'aumentare del margine di errore, per collettivi di almeno 2.000 unità. Ad esempio, con riferimento alla sotto-tabella di sinistra (relativa al livello di confidenza del 95%) di Tab. 10, la dimensione campionaria passa da 1.091 (in corrispondenza di $N = 2.000$) a 2.395 (per $N = 1.000.000$), per un margine di errore $\Delta = 2\%$, con un aumento, perciò, del 119%. Per un errore del 3% la dimensione campionaria limite supera quella minima del 53%; per il 4% di errore il campione più elevato supera quello corrispondente a $N = 2.000$ solo del 30%; si scende, quindi, al 19% e al 5% di incremento, passando dal campione minimo a quello massimo, per il 5% e 10% di errore, rispettivamente;
- per un dato *margine di errore* della stima e un dato livello di confidenza, all'aumentare della dimensione della popolazione la grandezza campionaria aumenta in misura **meno che proporzionale**;
- per una data *dimensione della popolazione*, la riduzione del margine di errore richiede incrementi **più che proporzionali** della dimensione del campione;
- fissato un *margine di errore*, se la grandezza del collettivo di riferimento è maggiore di 5.000 unità, il passaggio dal livello di confiden-

[95] Va tenuto presente che nelle formule [8], [9] – che danno le dimensioni campionarie riportate nelle Tabb. 10, 11 – il margine di errore Δ da intendere in *proporzione*, anche se nelle tavole tale errore è indicato come *percentuale* (1%, 2%, ..., 10%, in luogo dei valori di proporzione 0,01; 0,02; ...; 0,10 usati), come è consuetudine in letteratura.

[96] Differenza tra il valore massimo e il valore minimo delle modalità di una distribuzione.

Tab. 10 - Numerosità *massima* di un campione estratto da una popolazione finita per due livelli di confidenza e per assegnati valori di errore ammesso: proporzione $P = 0,50$

Popolazione (N)	Margine di errore						Margine di errore					
	1%	2%	3%	4%	5%	10%	1%	2%	3%	4%	5%	10%
500	475	414	341	273	217	81	485	446	394	338	286	125
1.000	906	706	516	375	278	88	943	806	649	510	400	143
2.000	1.655	1.091	696	462	322	92	1.786	1.351	961	684	500	154
3.000	2.286	1.334	787	500	341	93	2.542	1.743	1.144	773	545	158
4.000	2.824	1.501	843	522	351	94	3.225	2.040	1.265	826	571	160
5.000	3.288	1.622	880	536	357	94	3.845	2.271	1.350	861	588	161
7.000	4.049	1.788	926	553	364	95	4.928	2.610	1.463	906	608	163
10.000	4.899	1.936	964	566	370	95	6.247	2.938	1.561	942	624	164
15.000	5.855	2.070	996	577	375	95	7.889	3.257	1.646	973	637	165
25.000	6.939	2.191	1.023	586	378	96	9.991	3.567	1.722	999	648	165
50.000	8.057	2.291	1.045	593	381	96	12.486	3.841	1.783	1.019	657	166
100.000	8.763	2.345	1.056	597	383	96	14.267	3.994	1.815	1.029	661	166
200.000	9.164	2.373	1.061	598	383	96	15.363	4.075	1.832	1.035	663	166
500.000	9.423	2.390	1.065	600	384	96	16.105	4.126	1.842	1.038	665	166
1.000.000	9.513	2.395	1.066	600	384	96	16.369	4.143	1.846	1.039	665	166
	Livello di confidenza del 95%						Livello di confidenza del 99%					

za del 95% al livello del 99% comporta incrementi apprezzabili della dimensione campionaria;

- la dimensione campionaria è direttamente **proporzionale alla variabilità** dell'attributo allo studio, a parità di N, di Δ e del livello di confidenza (1- α).

Tab. 11 - Numerosità di un campione estratto da una popolazione finita per due livelli di confidenza e per assegnati valori di errore ammesso: stima *preventiva* di $P = 0{,}20$

Popolazione (N)	Margine di errore						Margine di errore					
	1%	2%	3%	4%	5%	10%	1%	2%	3%	4%	5%	10%
500	462	377	289	217	165	55	478	421	352	286	230	88
1.000	860	606	406	278	198	58	914	727	542	400	299	96
2.000	1.509	869	509	322	219	60	1.684	1.142	744	500	351	101
3.000	2.016	1.016	556	341	227	60	2.341	1.411	849	545	373	103
4.000	2.423	1.110	583	351	232	61	2.908	1.599	913	571	385	104
5.000	2.757	1.176	601	357	234	61	3.403	1.738	957	588	393	104
7.000	3.273	1.260	622	364	238	61	4.224	1.929	1.012	608	402	105
10.000	3.807	1.332	639	370	240	61	5.158	2.103	1.058	624	409	105
15.000	4.360	1.394	653	375	242	61	6.228	2.261	1.097	637	414	106
25.000	4.934	1.448	665	378	243	61	7.469	2.406	1.130	648	419	106
50.000	5.474	1.491	674	381	245	61	8.780	2.528	1.156	657	422	106
100.000	5.791	1.513	678	383	245	61	9.625	2.594	1.170	661	424	106
200.000	5.963	1.525	681	383	246	61	10.112	2.628	1.176	663	425	106
500.000	6.072	1.532	682	384	246	61	10.428	2.648	1.181	665	426	106
1.000.000	6.109	1.534	682	384	246	61	10.538	2.655	1.182	665	426	106
	Livello di confidenza del 95%						**Livello di confidenza del 99%**					

11. Campionamento casuale semplice per variabili

Intervallo di confidenza

Nelle ricerche con campione casuale si è interessati a determinare gli estremi di un intervallo (di confidenza) entro il quale si ritiene compreso, con un prefissato **livello di confidenza**, il valore del parametro cercato.

In questo contesto, come già rilevato, si presenta il cosiddetto *problema inverso*. Nota, cioè, la media $\bar{x}$ e la varianza s^2 del carattere considerato, su un solo campione, si desidera risalire alla stima della media incognita del collettivo di provenienza dello stesso.

Con l'inversione in argomento, per un **livello di confidenza** $(1 - \alpha) = 0{,}95$ (α è denominato **livello di significatività** ed indica il *rischio di errore prefissato* che si è disposti ad accettare) si ha, nel caso di popolazione *infinita*:

$$\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}.$$

A questo punto bisogna porre attenzione: con l'inversione considerata *non si può dire* che nel 95% dei casi μ cade nell'intervallo delimitato dai punti:

$$\bar{x} - 1{,}96\cdot\frac{\sigma}{n} \quad \text{e} \quad \bar{x} + 1{,}96\cdot\frac{\sigma}{n},$$

si può, invece, dire che: *su una lunga serie di campioni di n elementi, tratti da una popolazione distribuita normalmente*, con media μ **incognita** e σ **nota**, si può determinare – corrispondentemente – una serie di intervalli:

$$\bar{x} \mp 1{,}96\cdot\frac{\sigma}{\sqrt{n}}, \qquad [10]$$

e *che il 95% di tali intervalli*, di **confidenza**[102] (volendo con questo termine esprimere l'affidabilità del metodo), include la media incognita della popolazione[103].

Nel caso di popolazione finita dalla [10], considerando il **fattore correttivo** per popolazioni **finite**, si ha l'intervallo di confidenza:

$$\bar{x} \mp 1{,}96\cdot\frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}.$$

[102] La teoria degli **intervalli di confidenza** permette di valutare la probabilità (o il rischio) di commettere errori casuali di ampiezza prefissata, nella stima del carattere allo studio. È, infatti, possibile determinare una dimensione *n* tale che, con una prefissata probabilità *Prob*, la media campionaria cada all'interno di un dato intervallo $(\mu - \Delta, \mu + \Delta)$, dove Δ indica l'**errore ammesso** per la stima cercata. Facendo il ragionamento inverso: sarà sempre possibile determinare un valore di *n* tale che una data quota (*Prob*) degli intervalli di confidenza $(\bar{x} - \Delta, \bar{x} + \Delta)$ contenga il parametro ignoto μ. Ipotizzando la normalità della distribuzione campionaria delle medie, Δ viene espresso in termini di s.q.m. di tale distribuzione; si pone, perciò: $\Delta = z\cdot\sqrt{Var(\bar{x})}$, dove *z* è definito in relazione al valore di probabilità *Prob*.

[103] *Si può dire* che la media campionaria $\bar{x}$ ("variabile") dei diversi possibili campioni nel 95% dei casi cade tra i limiti **fissi**: $\mu \mp 1{,}96\cdot\frac{\sigma}{\sqrt{n}}$; *non* si può dire che μ "fisso" cade nel 95% dei casi entro i limiti **variabili**: $\bar{x} \mp 1{,}96\cdot\frac{\sigma}{\sqrt{n}}$: si può **correttamente** dire che il valore "fisso" di μ cade nel 95% degli intervalli "variabili" (De Luca, 1990, p. 91-92).

Determinazione della numerosità campionaria: caso di popolazione finita

Nel **campionamento per variabili** il parametro cercato è normalmente il valore **medio** del carattere allo studio.

Con riferimento alla dimensione campionaria, da determinare in funzione degli elementi sopra indicati, la procedura è analoga a quella seguita per il caso del campionamento per attributi.

Si consideri nuovamente lo schema del campionamento casuale semplice. Si indichi con μ la media incognita della popolazione di riferimento finita, con σ^2 la varianza incognita e con $\bar{x}$ lo stimatore **media campionaria**.

Ora:

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

è stimatore corretto di μ e, se la dimensione del campione è sufficientemente elevata, il teorema del limite centrale assicura che la distribuzione di $\bar{x}$ è *approssimativamente normale*, con media:

$$E(\bar{x}) = \mu$$

e varianza:

$$Var(\bar{x}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$ [104].

La statistica campionaria:

$$z = \frac{\bar{x} - \mu}{\sqrt{Var(\bar{x})}}$$

si distribuisce approssimativamente secondo una variabile normale standardizzata. Fissato il *margine di errore* Δ e il *livello di confidenza* $(1-\alpha)$, si ha:

[104] Nel caso di popolazione di grandi dimensioni (estrazione con ripetizione) la varianza dello stimatore è: $Var(\bar{x}) = \sigma^2 / n$.

$$\text{Prob}\{|\bar{x} - \mu| \leq \Delta\} = \text{Prob}\left\{|z| \leq \frac{\Delta}{\sqrt{Var(\bar{x})}}\right\} = 1-\alpha.$$

Dalla tavola della curva normale standardizzata si desume che la precedente relazione si verifica per: $\frac{\Delta}{\sqrt{Var(\bar{x})}} = z_{\alpha/2}$,

da cui si ricava la:

$$Var(\bar{x}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{\Delta^2}{z^2_{\alpha/2}}.$$ [11]

La stima **corretta** (s_c^2) di σ^2 è[105]:

$$s_c^2 = s^2 \frac{n}{n-1}\frac{N-1}{N}.$$

Pertanto una stima **corretta** della varianza dello stimatore $\bar{x}$ è la seguente:

[105] Per una popolazione infinitamente grande la quantità: $s_c^2 = s^2 \cdot \frac{n}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ è uno stimatore **corretto** della varianza della popolazione (σ^2); s^2 è la **varianza campionaria**. Pertanto, la media dei valori di s_c^2 di tutti i possibili campioni di n elementi, estraibili dalla popolazione di riferimento, è uguale a σ^2, varianza della popolazione.

Il **fattore di correzione** $n/(n-1)$, da applicare per ottenere una stima corretta di σ^2, tende a 1 all'aumentare di n. Per n sufficientemente grande detto fattore si può trascurare (dividendo al solito per n la somma dei quadrati della scarti di cui alla precedente formula riportata in questa nota).

$$Var(\bar{x}) = \frac{s_c^2}{n}\frac{N-n}{N-1} = \frac{s^2}{n-1}\cdot(1-f), \qquad [12]$$

nella quale $f = n/N$ è la *frazione di sondaggio* e $(1 - f)$ è il **fattore correttivo** per popolazioni finite.

Queste espressioni sono valide per campioni estratti da popolazioni nelle quali il fenomeno (*variabile*) allo studio è distribuito normalmente, con media μ e varianza σ^2. Quando la popolazione – molto numerosa – **non si distribuisce in modo normale**, le espressioni precedenti sono ancora valide ma si deve tenere presente che la distribuzione delle medie è solo *approssimativamente normale*, purché la dimensione campionaria sia sufficientemente elevata.

♦ *Dimensione del campione per la stima della media*

La [11] è l'espressione di un'equazione nell'incognita n (numerosità del campione), che dà la soluzione seguente[106]:

[106] È da tener presente che nella determinazione della numerosità campionaria ciascuna domanda del questionario deve essere considerata separatamente. È da presumere infatti che, a fronte di differenti domande, si otterranno percentuali di risposte affermative differenti, pertanto l'errore *standard* varierà tra domanda e domanda.

Si desume allora che la dimensione del campione dipende anche dalla complessità del questionario: ciascun quesito dovrà essere analizzato, perciò, in relazione alle domande che lo precedono (domande «filtro»). Per queste ragioni, dopo avere eliminate le combinazioni di domande più semplici, la dimensione definitiva del campione sarà calcolata sulla base della combinazione più complessa (tale dimensione risulterà adeguata anche per la combinazione più semplice).

Per approfondimenti di natura empirica su questi aspetti si veda A. H. R. Delens (1954, pp. 94-95).

Per determinare la numerosità campionaria in un'indagine con questionario occorre, pertanto:

a) stabilire il grado di esattezza richiesto per ogni variabile sottostante una domanda;

b) determinare per ciascuna domanda – tramite le opportune formule – la dimensione del campione;

c) dimensionare la numerosità campionaria sulla domanda che richiede la numerosità più elevata del campione, verificando che essa sia adeguata per tutte le restanti domande del questionario; in caso negativo sarà necessario aumentare la dimensione campionaria in relazione ai quesiti che richiedono un livello di confidenza più elevato.

$$n = \frac{z_{\alpha/2}^2 \cdot \sigma^2 N}{(N-1)\cdot\Delta^2 + z_{\alpha/2}^2 \cdot \sigma^2}. \qquad [13]$$

Quadro 8 - Applicazione: determinazione dell'intervallo di confidenza di una media su una popolazione finita (continua)

Un'impresa fornitrice di un prodotto industriale ha svolto un'indagine con questionario su un campione di $n = 30$ clienti, scelti casualmente da una popolazione di $N = 100$ unità, per stimare il volume della domanda di prodotto della popolazione, sulla base delle necessità espresse dagli intervistati, per il semestre successivo. I valori di previsione dichiarati dal campione di clienti sono riportati in Tab. 13.

Tab. 13 - Distribuzione di frequenza dei valori di domanda di un prodotto

Domanda del prodotto (in unità)	Numero di clienti (frequenze)
100	3
200	5
300	7
400	3
500	12
Totale	30

La distribuzione della v.c. domanda non risulta normale, ma è noto che, per grandi campioni ($n \geq 30$), la distribuzione delle medie campionarie $\bar{x}$ si approssima alla normale (con media $E(\bar{x}) = \mu$ (media della popolazione) e s.q.m. pari a (v. la [11]):

$$\sqrt{Var(\bar{x})} = \frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}.$$

La media campionaria è:

$$\bar{x} = \frac{100\cdot 3 + 200\cdot 5 + 400\cdot 7 + 400\cdot 3 + 500\cdot 12}{30} = 353{,}33.$$

Lo s.q.m. o deviazione standard sul campione è $s = 140{,}79$.

La stima corretta di σ (s.q.m. della variabile nella popolazione) è l'**errore standard** dello stimatore, quindi (per la [12]):

Quadro 8 - Applicazione: determinazione dell'intervallo di confidenza di una media su una popolazione finita (segue)

$$\sqrt{Var(\bar{x})}^{107} = \frac{s}{\sqrt{n-1}}\sqrt{1-\frac{n}{N}} = \frac{140{,}79}{30-1}\sqrt{1-\frac{30}{100}} = 21{,}87.^{108}$$

Si può sostenere, quindi, che con un livello di confidenza del 95% la media della popolazione cadrà nell'intervallo (*di confidenza*):

$$\bar{x} \mp 1{,}96 \cdot \sqrt{Var(\bar{x})} = 353{,}3 \mp 1{,}96 \cdot 21{,}87 = \begin{cases} 310{,}1 \\ 395{,}9 \end{cases}$$

Si presumerà, quindi, che il numero medio di unità di prodotto richiesto da ciascun cliente possa variare da un minimo di 310,1 ad un massimo di 395,9 unità, con un **livello di confidenza** del 95%.

♦ *Dimensione del campione per la stima dell'ammontare totale del carattere*

Se il parametro di interesse è l'**ammontare totale** ($\hat{x}$) di una variabile, si procede in modo analogo a quello descritto in precedenza. Lo stimatore del totale è:

$$\hat{x} = N\,\bar{x},$$

la **varianza** dello stimatore del **totale** è:

$$Var(\hat{x}) = N^2 \frac{\sigma^2}{n}\frac{N-n}{N-1}.$$

Il valore soluzione di n risulta, pertanto:

$$n = \frac{z_{\alpha/2}^2 \sigma^2 N}{\frac{(N-1)\Delta^2}{N^2} + z_{\alpha/2}^2 \sigma^2}.$$

Per la **stima del totale** di un carattere nel caso di campionamento *stratificato*, *a grappoli*, *sistematico* ed *a due stadi*, si rinvia alla letteratura specializzata (ad esempio, Frosini *et alii*, 1994; pp.107-137).

[107] L'errore standard dello stimatore ($\sqrt{Var(\bar{x})}$) della distribuzione campionaria delle medie tende a 0 al tendere di n ad N, nel caso di popolazione finita (estrazione senza ripetizione), o al tendere di n a ∞, nel caso di popolazione infinita (estrazioni bernoulliana con ripetizione).

[108] Si osservi che essendo nell'esempio la dimensione campionaria piuttosto elevata (pari a circa 1/3 di quella della popolazione), lo s.q.m. $\sqrt{Var(\bar{x})}$ dello stimatore $\bar{x}$ è piccolo e gli intervalli di confidenza risultano ridotti, come pure ridotto è – ovviamente – il rischio di errore.

14. Campionamento sistematico

Il campionamento **sistematico** rappresenta una variante semplificata del campionamento **casuale semplice** (cui finora si è fatto implicitamente riferimento), largamente adottato nella pratica.

È da tener presente che l'estrazione di un campione puramente *casuale* da una popolazione comporta una serie di operazioni piuttosto complesse e costose. Si deve disporre, innanzitutto, di una lista nominativa delle unità della popolazione, si deve, poi, associare un numero o un'etichetta a ciascuna unità (che deve essere ben individuata e deve avere probabilità *nota* di essere estratta ed inclusa nel campione); infine l'estrazione degli elementi campionari deve essere effettuata con procedimento equivalente alla scelta casuale delle palline da un'urna.

Il campione sistematico richiede, invece, un procedimento di scelta delle unità campionarie **più semplice**, con una consistente riduzione dei costi di rilevazione.

Per estrarre un campione di n unità da una popolazione di N elementi, la procedura da seguire è la seguente (v. Fig. 11):

1) si ordinano e si numerano, da 1 a N, le N unità della popolazione;

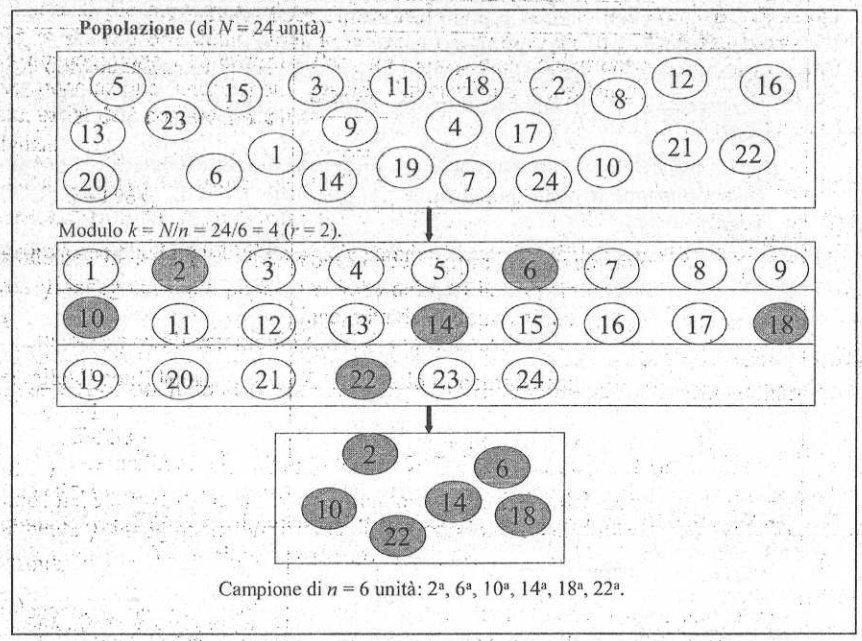
2) calcolato il **tasso di campionamento** k, dato dal rapporto tra la dimensione della popolazione e quella del campione ($k = N/n$), si preleva con criterio di casualità – facendo ricorso allo schema dell'urna o ad una tavola di numeri casuali – un numero intero compreso tra 1 e k, ovvero, la prima unità campionaria fra le prime k;

3) le restanti unità campionarie vengono scelte prelevando un elemento ogni k. La selezione degli elementi del campione viene effettuata, pertanto, secondo un **passo costante** (modulo di estrazione k), dovendo estrarre dalla popolazione 1 unità ogni k.

La scelta sistematica si configura, quindi, come un'estrazione casuale della **sola prima unità** campionaria r (che deve essere minore o uguale a k), gli altri elementi sono scelti a partire da questa – sequenzialmente – secondo una regola di estrazione (stabilita *a priori*), in base alla quale gli elementi del campione si susseguono in progressione aritmetica entro l'intervallo $[1, N]$, cioè, secondo l'ordine:

$$(r + k), (r + 2k), (r + 3k), \ldots, [r + (n - 1)k].$$

Fig. 11 – Rappresentazione schematica dell'estrazione di un campione casuale sistematico



Ad esempio, se la popolazione è di $N = 1.000$ unità e vincoli di bilancio impongono un campione di $n = 20$ elementi, la frazione sondata è: $f = 20/1.000 = 1/50$. È possibile, quindi, formare 50 campioni sistematici differenti (numero corrispondente al reciproco della frazione f sondata), secondo la regola di tenere equidistanziate le unità campionate, come si può vedere in Tab. 18[115].

[115] Il reciproco della frazione sondata ($1/f = k$) dà il numero **possibile di campioni** sistematici di una data dimensione. È da tener presente che se N non è **multiplo** intero di k, alcuni dei possibili campioni sistematici estratti dalla popolazione (finita) possono avere una dimensione (n - 1) e possono risultari **distorti**, in quanto, mentre per $N = nk$ la media $\bar{x}$ di un carattere rilevato su un campione sistematico è una **stima corretta** della media μ della popolazione, per $N \neq nk$, ciò non è più vero. La distorsione, cui in tal caso si incorre non è in genere rilevante (per $n > 50$ essa viene trascurata per semplicità). Tuttavia, si può evitare questo inconveniente considerando la lista della popolazione – le cui unità sono state preliminarmente ordinate – come **ciclicamente continua**, facendo così seguire all'ultimo elemento del collettivo il primo e i successivi elementi.

Tab. 18 - Campioni sistematici estraibili da unapopolazione

Campioni estratti	Unità considerate
1)	1, 51, 101, 151, ... 951
2)	2, 52, 102, 152, ... 952
⋮	⋮
50)	50, 100, 150, 200, 1.000

Una volta scelta la prima unità tra i numeri interi compresi tra 1 e 50, il campione è automaticamente determinato.

L'**efficienza** del campione sistematico (cap. 3) può risultare *uguale*, *superiore* o *inferiore* a quella del campione puramente casuale, in relazione al criterio con cui è ordinata la lista della popolazione da cui si estrae il campione.

Se tale lista è formata con ordine puramente **casuale** (caso piuttosto infrequente) il campione sistematico è **equivalente** a quello casuale. Se in ciascuno degli *N/n* possibili campioni sistematici di *n* unità il carattere studiato presenta un'elevata variabilità, questo tipo di campionamento dà risultati più attendibili rispetto al campione puramente casuale; ciò accade, ad esempio, quando le unità della popolazione sono ordinate **monotònicamente** (in modo crescente o decrescente) rispetto alla variabile di interesse.

Ad esempio, nel caso di un'indagine sul comportamento di acquisto di un collettivo di aziende, se per l'estrazione campionaria si fa ricorso ad una lista delle imprese ordinata per entità di fatturato conseguito, il campione formato dà risultati **più attendibili** di quello casuale semplice.

Se, ordinate le unità della popolazione il carattere analizzato presenta un andamento **ciclico**, cioè, le modalità del fenomeno osservato sulle unità rilevate si susseguono secondo una periodicità costante, può presentarsi l'inconveniente che il tasso di campionamento coincida con l'ampiezza del periodo del ciclo o con un suo multiplo. In tal caso il campione sistematico dà risultati distorti, poiché viene sottostimata la variabilità del carattere analizzato.

In questa evenienza il campione sistematico è *meno* **accurato** di quello casuale semplice e dà risultati non più precisi di quelli ottenuti estraendo in modo casuale solo una delle unità campionarie[116].

Ad esempio, se in una certa via – per puro effetto del caso – è collocato un bar ogni 10 negozi e si effettua un campionamento sistematico, con un valore di $k = 10$, c'è il rischio che il campione risulti composto solo da bar, con gravi distorsioni dei risultati campionari, se la popolazione obiettivo (*target population*) include anche private abitazioni.

Altri esempi di situazioni in cui si perviene a stime distorte sono i seguenti: rilevazioni sull'intensità del traffico ogni 24 nell'ora di punta; osservazione del numero di unità vendute in un negozio ogni 7 giorni (ad esempio, il sabato).

Bibliografia essenziale

Centre de Formation aux Applications Industrielles de la Statistique (1959), *Tables statistiques*, Institut de Statistique de l'Université de Paris, vol. II, n. 4.

Cicchitelli G., A. Herzel, G. E. Montanari (1997), *Il campionamento statistico*, Il Mulino, Bologna.

Colonel R.T. (2003), Il campionamento, in *Introduzione alle ricerche di marketing*, McGraw-Hill, Milano.

Delens A.H.R. (1954), *L'analisi del mercato*, Boringhieri, Torino.

De Luca A. (1990), *Metodi statistici per le ricerche di mercato*, Utet Libreria, Torino.

De Luca A. (1995), "Ricerche quantitative - il Campionamento", in Valdani E. (a cura di), *Enciclopedia dell'impresa - Marketing*", Utet Università, Torino.

Frosini B.V., M. Montinaro, G. Nicolini (1999), *Il campionamento da popolazioni finite*, Utet Università, Torino.

Giardina B. (1990), *Statistica per aziende e ricercatori*, FrancoAngeli, Milano, 8ª ed.

Herzel A., in D. Costantini *et alii.* (1994), *Metodi statistici per le scienze economiche e sociali*, Monduzzi Editore, Bologna.

Marbach G. (2000), *Le ricerche di mercato*, Utet, Torino.

Muttarini L. (1974), *Metodi statistici applicati alle ricerche economiche e sociali*, Giuffrè, Milano.

Rizzi A. (1992), *Inferenza statistica*, Utet Libreria, Torino.

Vajani L. (1969), *Metodi statistici nelle ricerche di mercato*, Etas Kompass, Milano.

Vianelli S. (1977), *Documentazione statistica*, Calderini, Bologna.

[116] In pratica difficilmente si incontrano popolazioni con un carattere il cui andamento è perfettamente periodico, pur tuttavia, un fenomeno può essere approssimativamente tale. Pertanto, quando si sospetta nella popolazione l'esistenza di variazioni cicliche è bene evitare il ricorso al campione in questione.

3. Piani di campionamento complessi

1. Introduzione

Nelle ricerche di mercato e nei sondaggi di opinione si fa ricorso, spesso, a piani di campionamento alternativi al piano *casuale semplice* o *sistematico*. Trattasi di piani complessi, utilizzati allo scopo di **contenere la dimensione** campionaria (con conseguente riduzione dei costi e dei tempi della ricerca) o per giungere a stime più accurate.

Nei piani complessi i campioni dell'universo campionario sono "più simili" alla popolazione, rispetto a quelli relativi al campionamento casuale semplice (Herzel *at alii*, 1994; p. 150).

Tali piani campionari possono essere ottenuti in uno dei modi seguenti:

a) dopo aver *stratificato* la popolazione obiettivo, ovvero, dopo aver suddiviso il collettivo in sotto-insiemi omogenei al loro interno (**strati**), si estrae da ciascuno di questi un subcampione di unità;

b) dalla popolazione obiettivo, ripartita in sotto-insiemi eterogenei al loro interno e simili tra di loro, si prelevano – al primo stadio – alcuni sotto-insiemi, da questi si selezionano poi (secondo stadio) ulteriori sotto-insiemi, così via ..., pervenendo al campione [finale] a **due** o a **più stadi**.

Talvolta dalle unità statistiche della popolazione, raggruppate in sotto-insiemi (denominati **grappoli** o *clusters*, in lingua inglese), si scelgono casualmente solo alcuni di essi, osservando poi tutti gli elementi che li compongono (rilevazione completa).

È da osservare che gli *strati* sono costruiti in funzione del disegno campionario, i *grappoli*, invece, sono raggruppamenti **precostituiti** di unità (ad esempio comuni, distretti commerciali, nuclei familiari, ecc.), che rispondono a requisiti di natura economico-territoriale o sociale.

Nel campionamento casuale semplice le probabilità di scelta dei singoli elementi del collettivo sono **uguali** tra di loro, in quello stratificato e a grappoli dette probabilità variano, invece, in ragione dello specifico strato o grappolo di appartenenza (campionamento con **probabilità disuguali**).

Quando nel campionamento a grappoli – piuttosto che rilevare tutte le unità di ogni grappolo – si osserva un campione di elementi si perviene al campione **a due stadi**, idoneo per indagini di vaste dimensioni. Con criterio analogo si ottengono campioni **a tre** o a **più stadi**.

La scelta del piano di campionamento e degli stimatori del carattere analizzato si effettua sulla base di informazioni disponibili sulla popolazione[1].

Nella realtà operativa i piani campionari complessi richiamati possono essere combinati tra di loro, in differenti modi, e generare un'ampia varietà di procedure campionarie (pertanto, negli stadi successivi possono essere combinati campioni casuali semplici con campioni stratificati, o sistematici, ecc.).

È possibile individuare due elementi *comuni* a tutti i tipi di piani:

a) nell'universo campionario i campioni **conformi** alla popolazione (che rispecchiano il profilo di quest'ultima) sono più *numerosi* dei non conformi;

b) al crescere della *dimensione* dei campioni cresce l'**attendibilità** delle stime campionarie.

Il primo elemento spiega il diffuso ricorso al metodo campionario, il secondo conferma l'importanza della determinazione della grandezza campionaria.

È da rilevare che, qualora lo stanziamento della ricerca non permetta di raggiungere il desiderato livello di precisione dei risultati (che richiede una certa dimensione campionaria, pur se minima), si deve ri-

[1] Quando tali informazioni non sono reperibili si fa ricorso, talvolta, ad un primo **campione informativo**, che viene utilizzato, successivamente, nel campione definitivo (*campione ripetuto*).

nunciare al campionamento e far ricorso ad altri metodi di osservazione. Ciò accade, tipicamente, quando il carattere di interesse presenta nella popolazione una distribuzione estremamente variabile o quando quest'ultima è di dimensioni molto ridotte (popolazioni **rare**[2]).

2. Campionamento stratificato

Nel cap. 2 si è visto che per formare un campione puramente casuale occorre disporre di una lista nominativa delle unità costituenti la popolazione obiettivo.

Nella realtà applicativa non è, in genere, possibile utilizzare un piano con scelta casuale delle unità da un'*unica* lista generale della popolazione; ad esempio, nelle indagini sui consumatori/clienti non è possibile attingere da un'anagrafe nazionale centralizzata. In ogni caso un'estrazione secondo il piano casuale semplice potrebbe comportare costi insostenibili, potendo risultare il campione molto disperso territorialmente. Per evitare questo problema si fa ricorso a varie strategie campionarie. Una di queste fa riferimento ad **unità ausiliarie**, non direttamente osservate, le quali contengono le unità da rilevare (campionamento **a grappoli**).

A volte, oltre all'anzidetta lista anagrafica del collettivo, si utilizzano altre informazioni per incrementare la precisione dei risultati, senza aumentare la dimensione del campione. È questa la strategia del **campionamento stratificato**, nel quale, sulla base di informazioni supplementari (*variabili ausiliarie*) disponibili, la popolazione viene suddivisa in sotto-insiemi o strati **omogenei**, da ciascuno dei quali vengono estratte le unità campionarie con criterio di casualità (Fig. 1).

Base della stratificazione può essere la stesso fenomeno di interesse (*variabile* o *attributo*) della ricerca, oppure una o più variabili correlate (quantitative) o connesse (qualitative) al carattere allo studio[3].

[2] Ad esempio, popolazione di collezionisti di monete antiche, di filatelici e simili.

[3] È preferibile usare come *base di stratificazione* lo stesso carattere oggetto di stima, anche se la classificazione cui si perviene risulta approssimativa.
Si è riscontrato che nelle indagini *territoriali* (*ecologiche*) la **stratificazione territoriale** delle unità statistiche consente di aumentare l'efficienza degli stimatori con riguardo al maggior numero di variabili allo studio.

Se, rispetto al carattere di interesse, gli strati sono effettivamente **omogenei** al loro *interno* (*within*) ed **eterogenei** *tra di loro* (*between*), la variabilità del medesimo risulta ridotta nei diversi sotto-insiemi; in tal modo si giunge a stime campionarie del carattere più precise all'interno degli strati e, quindi, ad una stima più attendibile sul campione complessivo.

Il campionamento stratificato è utilizzato frequentemente nelle ricerche di mercato per le seguenti due ragioni: *a)* spesso le indagini interessano unità territoriali intrinsecamente stratificate, cioè **aree amministrative** (regioni, province, comuni), **aree commerciali** (distretti) e **piccole aree** (studiate nel *geomarketing*); *b)* la procedura riduce in modo consistente la varianza degli stimatori.

È intuitivo che se si suddivide la popolazione in strati omogenei (sotto-insiemi), tali che il carattere allo studio assuma valori molto simili su tutte le unità appartenenti ai diversi sotto-insiemi, è sufficienti prelevare poche unità campionarie da ciascuno di questi per avere una stima quasi esatta del valore medio o del totale del carattere. Al limite, se tutte le unità di uno strato fossero uguali, sarebbe sufficiente estrarre dallo stesso *una sola* unità, per giungere ad una stima campionaria perfetta.

Ripartizione del campione tra gli strati

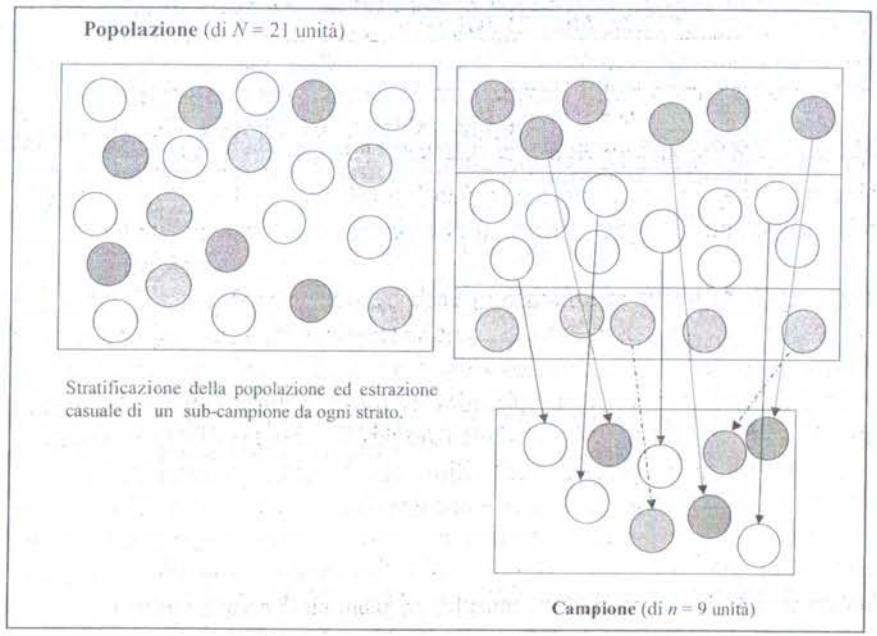
Nel campionamento stratificato interessano almeno tre elementi per ogni strato della popolazione: *1)* il numero delle unità che compongono lo strato (N_h), *2)* la media o proporzione del carattere allo studio, *3)* la varianza di tale carattere.

Si indichi con N la grandezza della popolazione obiettivo, con H il numero degli strati, con N_h ($h = 1, 2, \ldots, H$) la dimensione del generico strato h, con $w_h = N_h/N$ la frequenza relativa delle unità appartenenti allo strato h-mo. Il campione complessivo, di ampiezza n, è dato dall'unione dei campioni di grandezza n_h, estratti da ciascuno strato. La **frazione di campionamento** in ogni strato è uguale a $f_h = n_h/N_h$.

Determinata, come si vedrà in seguito, la dimensione n del campione, sorge il problema della sua ripartizione (*allocation problem*) tra gli H strati. Tale ripartizione può essere effettuata secondo vari criteri, alcuni semplici, altri più sofisticati e volti ad ottenere una maggiore precisione delle stime.

Di seguito si richiamano i principali criteri di suddivisione del campione complessivo (n) in subcampioni.

Fig. 1 – Rappresentazione schematica dell'estrazione di un campione stratificato



Per economia e semplicità espositiva, nelle applicazioni esemplificative che seguiranno, per la dimensione campionaria e per i principali stimatori si riporteranno solo le formule finali, rinviando, per una trattazione approfondita dell'argomento (dimostrazioni algebriche delle formule relative alla numerosità del campione, agli stimatori della media, della proporzione e dei totali) alle opere specialistiche (ad esempio, Herzel *et alii*, 1994; pp. 181-185).

1) Ripartizione uniforme del campione

È il criterio più semplice per ripartire le n unità campionarie tra i vari strati: a ciascuno strato viene assegnato lo stesso numero di elementi campionari (*frazione di campionamento* **costante**), dato da: $n_h = n/H$.

Risulta, pertanto: $n_1 = n_2 = \ldots = n_H$ (con $\sum_{h=1}^{H} n_h = n$).

Con questo criterio di ripartizione il campione stratificato risulta *più preciso* di un campione casuale semplice, di pari numerosità, se le medie degli strati *non* sono tutte uguali tra di loro.

2) Ripartizione del campione proporzionalmente alla dimensione degli strati

I subcampioni vengono assegnati ai vari strati sulla base della dimensione (N_h) di questi ultimi. Il campione stratificato proporzionale riproduce la composizione della popolazione obiettivo[4]. In ciascuno sotto-insieme la probabilità di scelta di un elemento è uguale alla *frazione di campionamento* dello strato. Tale frazione è **uguale** per tutti gli strati, infatti si ha:

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f \qquad (h = 1, 2, \ldots, H).$$

Il subcampione assegnato allo strato h ha dimensione $n_h = w_h \cdot n$. Il campione così ottenuto è **autoponderante**, nel senso che per ottenere gli stimatori della popolazione non occorre ponderare i dati a posteriori (Quadro 1). Infatti, la media aritmetica *semplice* calcolata sui valori osservati sul campione complessivo coincide con la media aritmetica *ponderata* delle medie degli strati (Herzel *et alii*, 1994; p. 183).

Quadro 1 - Alcuni aspetti statistici del campione stratificato proporzionale (continua)

Come si è in precedenza rilevato, l'errore campionario dipende, principalmente: *a)* dalla variabilità del carattere allo studio; *b)* dalla numerosità del campione.

I costi di campionamento limitano, in genere, la dimensione del campione.

È da osservare, però, che, se non si intende ridurre l'errore di stima tramite un incremento della dimensione campionaria, si può ridurre la varianza degli stimatori: è questa la **strategia** del campione stratificato.

Di essa si dà di seguito un'applicazione esemplificativa.

In Tab. 1 sono richiamati i valori dei redditi mensili di una popolazione di famiglie, riportati nel cap. 2, par. 8.

[4] Sotto questo profilo il campione stratificato risulta *più rappresentativo* di quello casuale semplice.

Tab. 1 - Reddito mensile di un collettivo di famiglie

Famiglie	Reddito mensile (migliaia di euro)
A	1,50
B	3,15
C	1,20
D	1,35
E	1,05

Nella popolazione sono individuabili due strati (Tab. 2):
a) il primo costituito dalle famiglie con reddito mensile < 3.000 euro;
b) il secondo composto dalle famiglie il cui reddito è ≥ 3.000 euro.

Tab. 2 - Collettivo di famiglie stratificato secondo il reddito mensile

Strato	Famiglie	Peso dello strato (w_h)
I (reddito < 3.000 euro)	*A, C, D, E*	4/5
II (reddito ≥ 3.000 euro)	*B*	1/5

I possibili campioni di due elementi, tratti da entrambi gli strati, sono 4. Questi campioni, unitamente al corrispondente stimatore "reddito mensile medio", sono riportati in Tab. 3.

Tab. 3 - Possibili campioni di famiglie e corrispondenti valori medi di reddito

Famiglie	Reddito mensile ponderato
AB	1,50 · 4/5+ 3,150 · 1/5 = 1,83
CR	1,20 · 4/5+ 3,150 · 1/5 = 1,59
DB	1,35 · 4/5+ 3,150 · 1/5 = 1,71
FB	1,05 · 4/5+ 3,150 · 1/5 = 1,47
Totale	6,60

La ponderazione è effettuata sulla base dei pesi associati agli strati, pertanto, l'elemento del campione tratto dal primo strato è ponderato per 4/5, quello prelevato dal secondo strato è ponderato per 1/5.

La media dei valori medi campionari è $E(\bar{x}) = 1{,}65$, valore uguale alla media della popolazione (v. cap. 2, par. 8); la media di ogni possibile campione è, perciò, stima **corretta** della media della popolazione.

Lo s.q.m. dei redditi medi risulta pari a 134,20. Nell'esempio riportato nel cap. 2, par. 8, lo s.q.m. dei redditi medi sui possibili campioni di due elementi, estratti con il metodo del campionamento **casuale semplice**, è risultato pari a 468,40. Lo s.q.m.

del reddito medio calcolato su tutti i possibili campioni del *piano stratificato* risulta, quindi, sensibilmente *inferiore* a quello del piano *casuale semplice*.

Il campionamento stratificato risulta **più efficiente**[5] di quello casuale semplice, in quanto il primo consente di individuare in anticipo nella popolazione le unità con modalità eccezionali del carattere allo studio; con il piano casuale semplice, invece, la famiglia *B* – nel caso in esempio – detiene la stessa probabilità delle altre unità: *A, C, D, E*: ciò è inefficiente.

Applicazione esemplificativa: ripartizione del campione stratificato in un'indagine sul servizio mensa aziendale

In un'azienda metalmeccanica è stato effettuata un'indagine campionaria sui dipendenti, volta a rilevare il gradimento di vari menù per il servizio mensa.

Si è deciso di far ricorso al campionamento stratificato, ritenendo che il gradimento verso differenti combinazioni di piatti potesse variare tra i sotto-insiemi di lavoratori (strati) individuati dai seguenti caratteri: 1) qualifica aziendale (impiegati, operai); categoria aziendale (impiegati: 5ª, 4ª, 3ª; operai: qualificati, generici); sesso (*M, F*); età (≤ 30 anni, > 30 anni).

Vincoli di bilancio avevano imposto una dimensione campionaria pari a $n = 300$.

La ripartizione del campione in subcampioni è stata effettuata con il criterio della proporzionalità rispetto al peso degli strati.

In Tab. 4 è riportato lo schema di ripartizione del campione.

I subcampioni di alcuni strati (ad esempio, lo strato formato da: impiegati, femmine > 30 anni), pur se esigui, sono stati ugualmente considerati, onde formare un campione rappresentativo della popolazione; in fase di stima delle proporzioni dei caratteri di interesse tali strati sono stati aggregati ad una delle classi contingue.

3) Ripartizione di Neyman e ripartizione ottimale del campione

Il campione complessivo viene suddiviso nei vari strati della popolazione in rapporto alla dimensione e **variabilità** del carattere di stratificazione: disponendo delle necessarie informazioni *a priori*, si attribuiscono sub-campioni di più elevata dimensione agli strati con maggiore variabilità.

La frazione di campionamento f_h varia, perciò, tra gli strati ed è **proporzionale** alla variabilità degli stessi.

[5] Per la misura dell'efficienza relativa dei disegni di campionamento v. par. 8.

Tab. 4 - Suddivisione di un campione di 300 unità tra vari strati (i valori tra parentesi indicano il numero di lavoratori)

Strati				Dimensione campionaria
Impiegati (2,201)	5ª (447)	*M*	≤30	6
			> 30	27
	4ª (762)	*M*	≤30	33
			> 30	24
	3ª (992)	*M*	≤30	48
			> 30	6
		F	≤30	20
			> 30	2
Operai (1.730)	Qualificati (1.045)	*M*	≤30	52
			> 30	29
	Generici (685)	*M*	≤30	40
			> 30	13
Totale				300

Se si intende rendere *minima* la varianza **complessiva** del campione stratificato, con il vincolo che sia $\sum_{h=1}^{H} n_h = n$, si perviene al **campione di Neyman** (1934), che soddisfa la seguente condizione:

$$n_h \propto w_h S_h, \qquad [1]$$

dove va tenuto presente che la varianza del carattere nel generico strato è: $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (x_{hi} - \bar{x}_h)^2$ (in cui x_{hi} è il valore che il carattere quantitativo di stratificazione assume sull'unità *i*-ma, $i = 1, 2, \ldots, N_h$, dello strato h, $h = 1, 2, \ldots, H$; $\bar{x}_h$ è la media aritmetica del carattere nello strato *h*-mo); w_h indica il peso dello strato (N_h/N); $\propto$ indica l'operazione di proporzionamento.

Dalla [1] discende che se i valori di S_h sono uguali tra di loro il campione di Neyman si identifica con quello proporzionale.

Per le formule di calcolo della dimensione campionaria *n*, nel caso di campionamento stratificato di Neyman, si rinvia alla letteratura specializzata (ad esempio, A. De Luca, 1990; pp. 159-160).

Nella realtà applicativa questo tipo di campione è raramente utilizzato, richiedendo esso la conoscenza del valore caratteristico S_h.

È da osservare che nelle ricerche **multiscopo**, aventi come obiettivo la stima di più caratteri, per ciascuno di questi il campione di Neyman può avere una differente struttura (per quanto detto con riguardo alla dimensione campionaria rapportata alla variabilità del carattere considerato); in tal caso occorre adottare una soluzione di compromesso, che si individua normalmente nel campione proporzionale.

La [1] vale anche nel caso di **attributi**, ponendo:

$$S_h^2 = \frac{N_h}{N_h - 1} P_h (1 - P_h) .$$

Considerando, inoltre, il **costo di rilevazione**: supponendo che questo *vari* tra gli strati e che sia pari a c_h il costo per il generico strato h, se si rende minima la varianza complessiva del carattere allo studio per un prefissato costo, oppure, se si minimizza il costo per una prefissata varianza, si ottiene il **campione ottimo**. Nel caso di stima di **medie**[6] la ripartizione ottimale del campione è la seguente,:

$$n_h \propto w_h S_h / \sqrt{c_h} . \qquad [2]$$

I valori n_h conformi alla [2] rendono minimo il valore della varianza del carattere allo studio, per un prefissato valore di costo e, viceversa, minimizzano il costo per un dato valore della varianza.

Il campione stratificato ottimo è idoneo nei casi in cui sia la varianza del carattere di interesse, sia il costo di un'intervista, differiscono tra gli strati.

Con il campionamento in parola, alcuni strati sono *sovrarappresentati*, altri risultano *sottorappresentati*. Gli elementi del collettivo non hanno uguale probabilità di essere estratti, pertanto per la stima di valori medi o di totali o di proporzioni, si fa ricorso a **schemi di pondera-**

[6] Jerzy Neyman, "On the two differente aspects of the representative method. The method of stratified sampling and the method of purposive selection", *Journal of the Royal Statistical Society*, 1934, pp. 97, 558-606.

zione, con pesi costituiti dalle probabilità – note – di inclusione delle unità degli strati nel campione.

4) **Ripartizione del campione non proporzionale e non ottimale**

A questo tipo di ripartizione di n si fa ricorso quando si desiderano stime attendibili non solo sull'intera popolazione obiettivo, ma anche sui singoli strati. Ad esempio, nell'ambito di una ricerca su scala nazionale, con popolazione stratificata su base provinciale, si desiderino stime affidabili a livello provinciale, oltre che nazionale. A tale scopo, fissata una numerosità minima per ogni strato, se questa non è assicurata direttamente in alcuni sotto-insiemi – applicando il criterio della proporzionalità – si **sovracampiona** (*oversampling*) dagli altri strati.

5) **Ripartizione del campione in modo inversamente proporzionale**

Nelle ricerche di marketing, al fine di approfondire la conoscenza di **piccoli strati** o segmenti di mercato (*nicchie*), si possono utilizzare campioni di grandezza inversamente proporzionale al numero degli elementi dei sotto-insiemi di riferimento (Quadro 2).

In generale, la stratificazione viene eseguita all'interno del piano di campionamento, sulla base di uno o pochi caratteri[7]. Nel caso in cui per ciascun elemento della popolazione si disponga di più di tre caratteri si può fare ricorso all'**analisi dei gruppi** o *cluster analysis* (De Luca, 2005[5]), cioè a tecniche di formazione di strati il più possibile omogenei, tali che ciascuna unità appartenga ad un solo gruppo (i gruppi – se composti da unità territoriali – possono presentare, eventualmente, il **vincolo di contiguità**)[8].

[7] La *scelta* dei criteri di stratificazione è di importanza *cruciale* nel disegno campionario. Ottimale sarebbe scegliere come base di stratificazione la variabile oggetto di stima; siccome ciò risulta spesso irrealizzabile, si fa ricorso a 2-3 variabili *ausiliarie*, correlate con quella di interesse. È necessario, inoltre, che i caratteri di stratificazione consentano di individuare facilmente le unità appartenenti ai vari strati.

Nella pratica la stratificazione viene effettuata spesso sulla base di caratteri enumerabili (numero di abitanti di un comune, numero di addetti nelle unità produttive, ecc.) o di natura *geografica* (circoscrizioni territoriali, regioni, province, comuni, distretti commerciali, ecc.), che presentano il vantaggio di essere stabili nel tempo.

[8] Nella stratificazione su base territoriale può risultare opportuno fare ricorso alla *cluster analysis* con il **vincolo di contiguità** (A. Daggiano, *Cluster analysis con vincolo di contiguità per la segmentazione territoriale e il calcolo del potenziale di mercato*,

Le tecniche di analisi dei gruppi maggiormente utilizzate sono implementate nei *package* statistici più noti (SPSS, SAS, STATISTICA, ecc.).

La stratificazione consente di:

a) **aumentare la precisione** delle stime campionarie a parità di n, oppure, di **ridurre** n a parità di precisione richiesta;

b) analizzare con **diversa precisione** i singoli strati (indipendenti), aumentando – quando necessario – la dimensione di quelli particolarmente interessanti per la ricerca.

Quadro 2 - Un caso di utilizzazione del campionamento proporzionale e inversamente proporzionale (continua)

Un'impresa locale, produttrice di un superalcolico, intende svolgere un'indagine conoscitiva sui consumatori del suo prodotto. La popolazione obiettivo sia composta da $N = 600$ soggetti, residenti in un certo distretto territoriale. Si sia deciso di svolgere l'indagine sul 10% della popolazione.

Caso A)

Se l'impresa vuole svolgere un'indagine di *customer satisfaction* sulla sua clientela può fare ricorso al campionamento *stratificato proporzionale*, considerando come base di stratificazione il comportamento di acquisto (di fedeltà o non).

Campionamento stratificato proporzionale

Il peso degli strati è $w_h = N_h/N$ (con $h = 1, 2$), la dimensione dei subcampioni è data da: $n_h = w_h \cdot n$ (con $h = 1, 2$); in Tab. 5 si riporta la struttura del campione.

Tab. 5 - Campione stratificato proporzionale

	Strati	Dimensione dello strato (N_h)	Dimensione dei subcampioni (n_h)	
	consumatori fedeli	400	40	
	consumatori infedeli	200	20	
	Totale	600	60	

Caso B)

Se l'impresa è interessata, invece, principalmente a rilevare il parere dei forti consumatori, il campione – stratificato in base alle modalità forte-debole bevitore – può essere di **tipo inversamente proporzionale**.

Elaborato finale di laurea triennale in Scienze Statistiche ed Economiche, relatore A. De Luca, Università Cattolica del Sacro Cuore di Milano, A.A. 2003-04).

Quadro 2 - Un caso di utilizzazione del campionamento proporzionale e inversamente proporzionale (segue)

Campionamento stratificato inversamente proporzionale

Ponendo: $w_h = N/N_h$ (con $h = 1, 2$); $w = w_1 + w_2$, si ha:

$$n_1 = \frac{w_1}{w} \cdot n \; ; \quad n_2 = \frac{w_2}{w} \cdot n \; .$$

Tab. 6 - Campione stratificato inversamente proporzionale

Strati	**Dimensione dello strato** (N_h)	**Dimensione dei subcampioni** (n_h)
forti bevitori	200	40
deboli bevitori	400	20
Totale	600	60

La dimensione dei due subcampioni si calcola, quindi, nel modo seguente:

$$w_1 + w_2 = \frac{600}{200} + \frac{600}{400} = 3 + 1{,}5 \; ,$$

$$w = w_1 + w_2 = 4{,}5 \; .$$

Si ha quindi:

per i *forti* bevitori: $3/4{,}5 = 0{,}667$; $n_1 = 0{,}667 \cdot 60 \cong 40$ (numerosità del campione);

per i *deboli* bevitori: $1{,}5/4{,}5 = 0{,}333$; $n_2 = 0{,}333 \cdot 60 \cong 20$ (numerosità del campione).

Qualunque sia il criterio di stratificazione della popolazione e di ripartizione del campione, il subcampione relativo a ogni strato è scelto in modo *indipendente* dagli altri.

3. Campionamento per area

Non rientra nell'economia espositiva di quest'opera la trattazione completa dei vari piani di campionamento. È intendimento del lavoro trattare principalmente il campionamento casuale semplice e quello stratificato. Si ritiene opportuno menzionare, tuttavia, per completezza espositiva, altri tipi di campionamento complessi (oltre al campionamento stratificato già trattato), che possono essere adottati nelle ricerche di mercato e nei sondaggi di opinione.

Il campionamento di aree[14], metodo pratico ed economico per giungere alle unità campionarie, fornisce stime precise.

[14] Trae origine da applicazioni svolte in **agricoltura**. Per valutare la produzione e le caratteristiche del suolo di una data zona, questa viene preliminarmente suddivisa in

La procedura consiste nel suddividere una macro area territoriale in aree omogenee e comparabili (ad esempio, nelle indagini **ecologiche** si considerano aree omogenee *urbane* e *rurali*), sia per *dimensione* che per *densità* della popolazione; si estraggono, quindi, casualmente – con procedura di *stratificazione* – **alcune** di queste aree e, all'interno delle stesse, vengono prelevati dei subcampioni (unità finali di rilevazione; v. Fig. 2), oppure tutti i loro elementi (**campionamento a grappoli**; par. 4).

Il **campione areale** è estratto dalla lista di porzioni del territorio complessivo.

A volte l'indagine riguarda un'area limitata; in tal caso, dopo aver suddivisa quest'ultima in *lotti* (o *spicchi*) simili, se ne seleziona un certo numero (ad esempio, blocchi di *isolati* di una città) e, tramite l'ausilio di carte topografiche dettagliate, si individua l'unità campionaria (famiglia o individuo). La scelta degli elementi del **campione** all'interno dell'area è guidata da rigidi criteri (stabiliti *a priori)* che ne assicurano la casualità.

Per accrescere la correttezza delle stime campionarie conviene – ove possibile – ridurre la dimensione delle aree, aumentando, così, il loro numero. Risultando – in tal modo – le unità primarie più piccole e più numerose, è più probabile che l'errore campionario su una di esse possa essere compensato dall'errore (in senso opposto) associato ad un'altra unità campionaria.

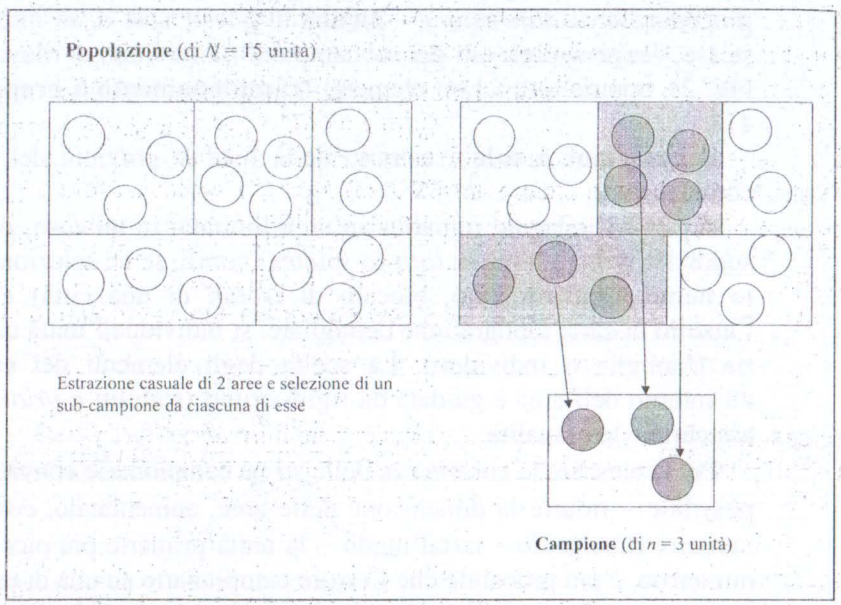
Ad esempio, se si decidesse di campionare 3 unità primarie (aree) di grandi dimensioni su un totale di 40 aree, un errore rilevante su un'unità campionaria peserebbe in maniera apprezzabile sulla correttezza della stima campionaria. Se, al contrario, su un totale di 500 unità primarie se ne scegliessero 100, un errore consistente su un'unità campionaria potrebbe avere presumibilmente un peso trascurabile sulla stima finale.

Aumentando, però, il numero delle unità primarie, diminuisce normalmente quello degli elementi da estrarre da ciascuna di esse. Ciò

lotti omogenei, da ciascuno dei quali si estrae un'unità campionaria con criterio di casualità; valutata la produzione dei lotti del campione, si stima quella dell'intera zona. Affinché il campione estratto sia rappresentativo, i lotti devono avere le stesse *dimensioni* e *caratteristiche*.

comporta un incremento dei costi, a causa dei più lunghi percorsi da

Fig. 2 – Rappresentazione schematica dell'estrazione di un campione per area



coprire per contattare le unità del campione[15].

In Italia questo tipo di campionamento non ha suscitato – fino ad epoca recente – grande interesse, anche perché non si disponeva di liste affidabili e di una mappa tipologica del territorio, strutturata per **microaree**. Attualmente sono disponibili mappe territoriali dettagliate secondo subaree molto piccole[16], che schiudono nuove prospettive al

[15] Il lavoro preparatorio di disegno delle aree – necessario per l'approntamento delle carte topografiche e degli itinerari – è, di norma, complesso e costoso, ma può essere sfruttato in indagini successive, ripartendo su queste il costo iniziale.

[16] Il tema della stima su piccole aree (*Small Area Estimation* - SAE) interessa oggi nel nostro Paese – anche se con inspiegabile ritardo – sia il mondo accademico che quello aziendale, come testimonia lo sviluppo del **marketing del territorio** o **Geomarketing** (v. J. N. K. Rao, *Small area estimation*, Wiley, New York, 2003; L. Moretti, *Segmentazione, geomarketing e stima su piccole aree*, Elaborato finale di laurea triennale in Scienze Statistiche ed Economiche, relatore A. De Luca, Università Cattolica del Sacro Cuore di Milano, A.A. 2003-04).

campionamento areale. Infatti a partire dal Censimento del 1991, sono fornite in Italia liste di unità elementari, calibrate sulla **sezione di censimento** o sull'**isolato** (caratterizzato tipologicamente secondo *stili di vita*).

Il ricorso al campionamento areale è idoneo per ricerche su popolazioni obiettivo delle quali non si dispone di una lista; ad esempio, in indagini su popolazioni *clandestine* (di evasori fiscali, di immigrati senza permesso di soggiorno e simili).

Per l'analisi di dette popolazioni in letteratura si suggerisce il campionamento **a valanga** (cap. 2), ma questa procedura presenta rischi di distorsione nelle stime e risulta meno corretta – sul piano metodologico – del campionamento di aree.

4. Campionamento a grappoli

A volte, le unità elementari di un collettivo non vengono selezionate una ad una (come accade nel campionamento casuale semplice), ma per gruppi. Ciò avviene quando le unità statistiche della popolazione sono raggruppate in insiemi o aggregati – detti **grappoli** – i cui elementi sono spazialmente contigui (e solitamente di dimensioni differenziate); in questi casi si rivela idoneo il **campionamento a grappoli** (*cluster sampling*); v. Fig. 3.

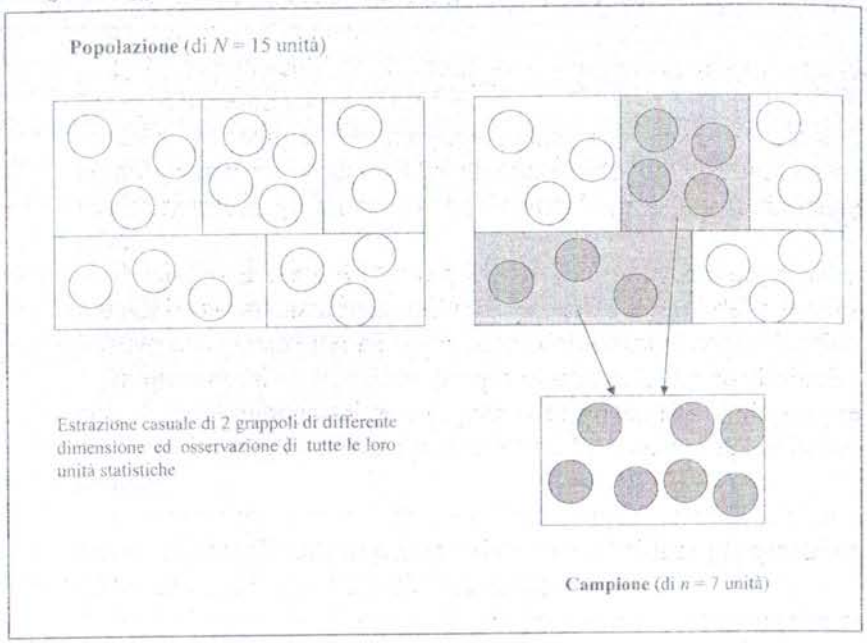
Questo tipo di campionamento è idoneo quando:

a) **non esiste la lista** nominativa delle singole unità della popolazione da cui estrarre il campione e la formazione della stessa per la sola rilevazione in atto risulta molto onerosa;

b) si ha interesse a **ridurre i costi** di rilevazione (l'osservazione di unità vicine è più economia e più facile – dal punto di vista operativo – di quella su unità sparse e distanti tra di loro).

Si va facendo oggi strada anche il cosiddetto "**marketing urbano**" per lo *sviluppo del territorio*, dal punto di vista commerciale, promosso con il supporto di operatori locali (commercianti, artigiani ed operatori), avente l'obiettivo di rivitalizzare e rilanciare l'area in cui sono insediati i punti vendita al dettaglio. Esempi in tal senso si registrano in alcune città del Regno Unito (Londra, Carline, Bolton) e della Germania (ad esempio, Berlino ed Amburgo). In Italia sono state promosse iniziative di marketing urbano in Emilia Romagna e, di recente, nel sud d'Italia (Puglia), ad opera di M. Quarta (*Project Consulting*; www.assovia.it).

Fig. 3 – Rappresentazione schematica dell'estrazione di un campione a grappoli



Richiede minori costi e minor tempo la rilevazione svolta presso 30 abitazioni collocate in un blocco di palazzi (grappoli) di un comune, che l'osservazione di un campione – di pari numerosità di abitazioni – estratto in modo casuale dall'intero territorio comunale.

Contrariamente agli strati del campionamento stratificato, i grappoli sono **eterogenei** al loro interno e **omogenei** tra di loro.

La varianza dello stimatore ottenuto con questo tipo di campionamento è normalmente *maggiore* di quella relativa al campionamento casuale semplice (il *Deff* risulta maggiore di 1; v. par. 8), d'altra parte, rispetto a quest'ultimo, i costi di rilevazione sono inferiori. Pertanto la **precisione** di un campione a grappoli – rapportata ai costi – può essere, in definitiva, **maggiore** di quella ottenuta con il campione *casuale semplice*.

Contrariamente al campionamento stratificato, nel quale il campione complessivo include tutti gli strati, nel campionamento in argomento **solo una parte dei grappoli** è scelta per formare il campione; tale scelta è effettuata con la procedura del campionamento casuale semplice, con o senza ripetizione.

Sui grappoli scelti si rilevano, poi – normalmente – tutti gli elementi (campionamento ad **uno stadio**).

Uno **svantaggio** del campionamento a grappoli è costituito dalla complessità dei metodi di stima dei parametri[17] e, in alcuni casi, dal fatto che non tutti gli stimatori soddisfano le proprietà statistiche desiderabili.

Il campionamento a **due** o **a più stadi** è utilizzato principalmente in indagini su scala nazionale (ad esempio sui consumi delle famiglie; sullo stato occupazionale; sulla *audience* dei mezzi di comunicazione).

In questi casi si estrae un gruppo di comuni, che costituiscono grappoli di unità elementari, e da ognuno di questi si sceglie un prefissato numero di famiglie, senza utilizzare una lista anagrafica delle unità statistiche.

La procedura consente di contattare soggetti non dispersi territorialmente, ma localizzati in insiemi ristretti (ciò riduce gli spostamenti dei rilevatori e, quindi, i costi della ricerca).

È da rilevare, però, che se i grappoli non sono sufficientemente omogenei tra di loro, i vantaggi del disegno campionario in parola potrebbero ridursi o annullarsi, a causa dell'elevata variabilità degli stimatori ottenuti.

5. Campionamento a stadi

Oltre ai campionamenti **a uno stadio** (per i quali le unità scelte sono anche le *unità di rilevazione*), nelle applicazioni concrete vengono spesso adottati piani **a due** o **a più stadi**. Con questi piani si fa riferimento a unità statistiche **ausiliarie**, le quali non formano oggetto di rilevazione diretta, ma contengono le unità da rilevare.

Dopo aver estratto casualmente alcune grandi unità (*grappoli*), dette di *primo ordine*, non si rilevano tutte le unità elementari in esse contenute, ma un loro campione (di *secondo ordine*), le quali possono essere:

[17] Per le formule degli stimatori v. A. Herzel (1994, pp. 186-193).

a) le *unità finali* elementari (campionamento **a due stadi**);
b) altre *unità più ampie* rispetto alle unità finali (campionamento a **tre stadi**); così via, facendo crescere il numero degli stadi finché è necessario.

Nello scegliere tra il campionamento ad uno o a più stadi si deve considerare che ad ogni stadio i *costi si riducono* e che *aumenta la varianza* dello stimatore.

Questa procedura è consigliata quando le unità di ordine superiore hanno una composizione **omogenea** tra di loro, tanto da rendere superflua una rilevazione di tutte le grandi unità o grappoli (Fig. 4*a*).

Pertanto, invece di campionare direttamente i consumatori di un dato prodotto, si possono considerare sequenzialmente le unità ausiliarie: regioni, province, comuni, ecc. Nel primo stadio si estrarrà un campione di regioni, all'interno di ciascuna regione si sceglieranno casualmente alcune province, all'interno di ciascuna provincia estratta si campioneranno alcuni comuni; infine, dai comuni si estrarranno casualmente i consumatori o le famiglie da intervistare (Fig. 4*b*).

Il campionamento a stadi comporta costi inferiori a quello stratificato, ma allo stadio finale – a parità di elementi di un campione casuale semplice – fornisce, nel complesso, risultati che possono essere più precisi[18].

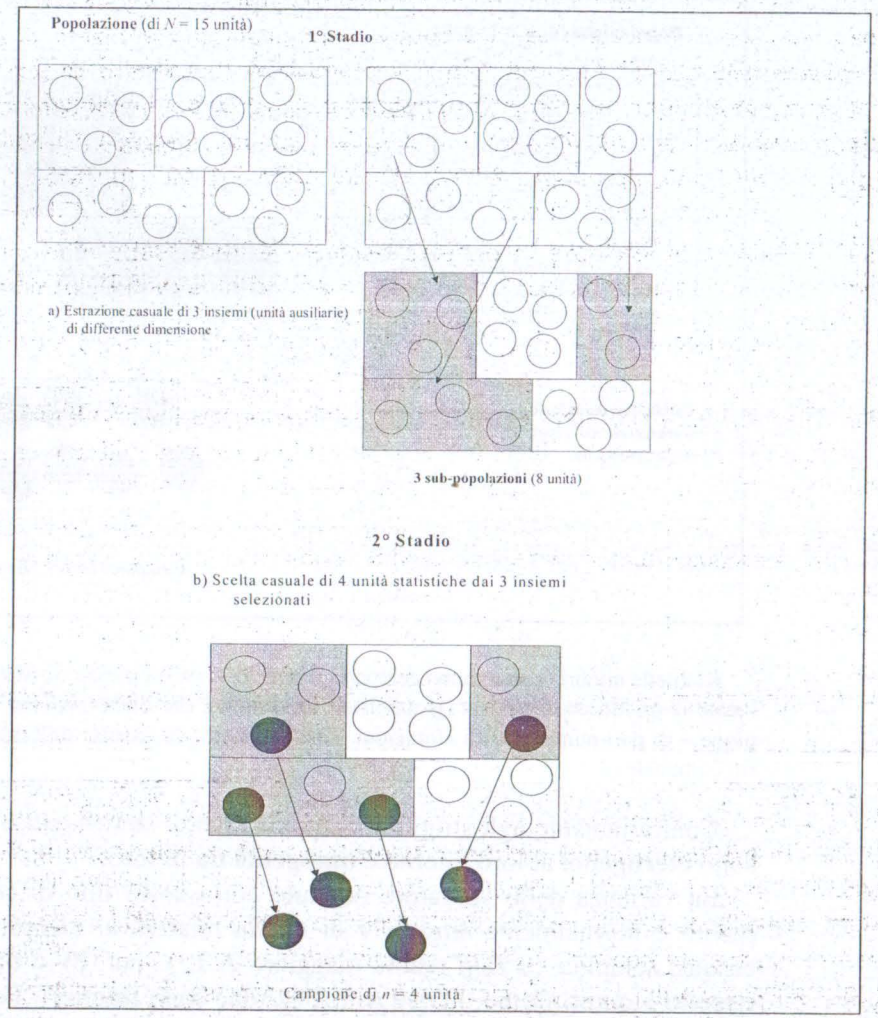
Esempio di campionamento a due stadi per la stima di una media

Il responsabile di una società produttrice di prodotti per l'igiene personale era interessato a conoscere – tramite un'indagine campionaria – il numero di marche in concorrenza trattate dai suoi rivenditori, distribuiti in 300 comuni, capoluoghi esclusi.

Per costruire il campione si procedette ad un'aggregazione dei comuni (tra loro contigui) in 10 grappoli, in modo tale che in ciascuno di

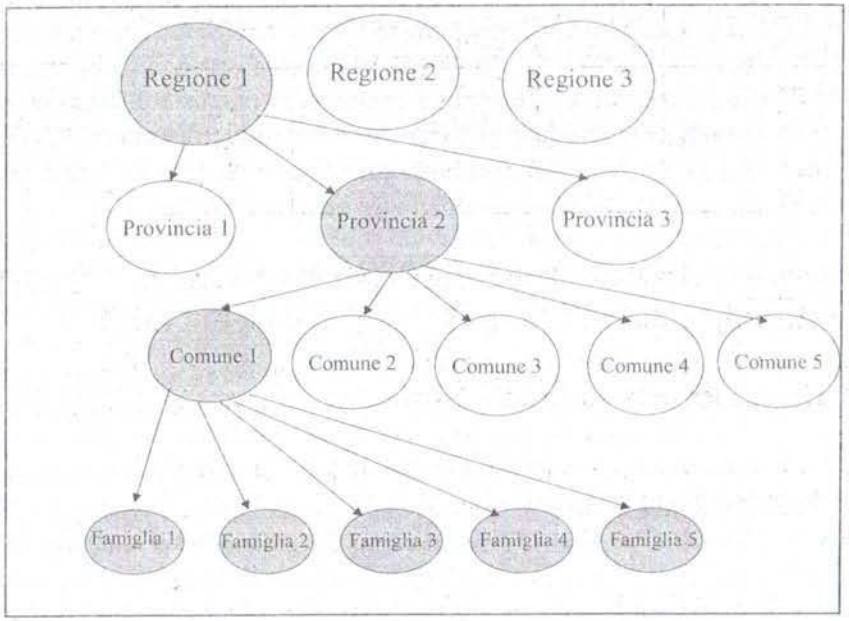
[18] A causa dei suoi più **bassi costi**, questo tipo di campionamento consente di usare – rispetto a quello casuale semplice – campioni con un maggior numero di elementi allo stadio finale, ciò compensa gli svantaggi derivanti dalla minore precisione dei risultati cui si perviene con la procedura.

Fig. 4a – Rappresentazione schematica dell'estrazione di un campione a stadi



essi rientrasse lo stesso numero di rivenditori, pari a 40. In tal modo si fissarono:
$N_1 = 10$ unità primarie (grappoli);
$N_2 = 40$ unità secondarie (famiglie);
$N = N_1 \cdot N_2 = 400$ (N dimensione della popolazione).

Fig. 4b – Estrazione di un campione a stadi



Furono scelti, con criterio di casualità, 5 grappoli (gruppi) di comuni, limitando a questi l'indagine. Su ciascuno di detti grappoli si eseguì un secondo campionamento, estraendo – sempre casualmente – 4 rivenditori. Si ebbe pertanto:

$n_1 = 5$

$n_2 = 4$;

$n = 4 \cdot 5 = 20$ (numero di rivenditori campionati).

I risultati dell'indagine sono riportati in Tab. 8.

La stima della media generale risultò:

$$E(\bar{x}) = \frac{\sum_{h=1}^{H} \bar{x}_h}{n_1} = \frac{\sum_{h=1}^{5} \bar{x}_h}{5} = \frac{18{,}50}{5} = 3{,}7 .$$

Tab. 8 - Valori medi di numero di marche trattate da un campione di rivenditori in 5 grappoli territoriali

Grappolo	Numero di marche trattate dai quattro rivenditori del grappolo	Numero medio di marche ($\bar{x}$)
1	4 3 4 2	3,25
2	3 5 6 2	4,00
3	4 3 2 6	3,75
4	5 5 4 3	4,25
5	2 3 4 4	3,25

L'*errore di campionamento*, calcolato utilizzando – ovviamente – la formula relativa al piano in parola[19], risultò: $\sqrt{Var(\bar{x})} = \sqrt{0{,}101} = 0{,}318$: valore piuttosto elevato (a conferma della scarsa precisione del campionamento in argomento), rispetto a quello della media $E(\bar{x})$.

È da rilevare che nel campionamento a grappoli, ad uno o a più stadi (**stratificazione dei grappoli**), la stratificazione viene applicata in modo naturale. Le indagine con campionamento a grappoli prevedono quasi sempre la stratificazione dei grappoli stessi (visti come unità primarie nel campionamento a più stadi, v. Cicchitelli *et alii*, 1997; pp. 213-215).

8. Effetto del disegno di campionamento (*Deff*)

Rispetto ai piani campionari semplici, per i piani di campionamento **complessi** è più difficile determinare direttamente la numerosità campionaria, fissato il valore dell'errore ammesso (precisione).

Per calcolare detta numerosità viene utilizzato il **design effect** (*Deff*), introdotto da Kish nel 1965[22].

Trattasi di un quoziente di misurazione dell'**efficienza relativa**, nel cui numeratore è posta la varianza dello stimatore associato al campione *complesso*, mentre nel denominatore è riportata la varianza dello stimatore inerente al piano casuale *semplice*, di pari numerosità.

Indicando con $\hat{\theta}$ lo stimatore corretto del parametro θ in un piano *complesso*, con $Var(\hat{\theta})$ la *varianza* del primo e con $Var_O(\hat{\theta})$ la varianza dello stimatore associato al piano *casuale semplice*, l'effetto del disegno di campionamento (*Deff*) è dato, quindi, dal rapporto:

$$Deff = Var(\hat{\theta}) / Var_O(\hat{\theta}) \,. \qquad [9]$$

Il quoziente *Deff* consente di determinare la **dimensione campionaria** del piano complesso: a questo scopo è sufficiente, infatti, moltiplicare la numerosità del campione casuale semplice (par. 10, 11 del cap. 2) per il *Deff*.

Il *Deff* permette anche di valutare l'efficienza relativa dei piani complessi. Infatti nella [9] un risultato **minore** (*maggiore*) di **1** segnala che l'efficienza dello stimatore ottenuto con il campionamento complesso è **maggiore** (*minore*) di quella del piano casuale semplice (di uguale numerosità). Con riguardo alla stima di una **proporzione**, l'efficienza del campionamento considerato si può indicare nel modo seguente:

$$Deff = nVar(\hat{p}) / [(1-f)P(1-P)] \,,$$

dove: $f = n/N$ è la frazione di campionamento; N e n indicano, rispettivamente, la dimensione della popolazione e quella del campione; P e $\hat{p}$ rappresentano la proporzione (o frequenza relativa) del carattere di interesse, rispettivamente, nella popolazione e nel campione;

[22] L. Kish (1965).

$Var(\hat{p})$ indica la varianza nel disegno campionario utilizzato. In termini generali si può dire che il *Deff* nei campioni **stratificati è minore di 1**, mentre in quelli **a grappoli** è **maggiore di 1**.

Quadro 5 - Efficienza di differenti strategie di campionamento

• Nel campionamento **sistematico** (nel quale le unità sono ordinabili secondo una variabile correlata fortemente a quella allo studio) l'effetto del disegno di campionamento è:

$$Deff = (k + 1)/(N + 1) < 1,$$

dove $k = N/n$ è il passo di campionamento. Si può notare che l'efficienza di questo disegno aumenta con la dimensione del campione.

• Nel campionamento **a grappoli** l'effetto del disegno è dato da:

$$Deff = 1 + (n - 1)\,\rho\,,$$

dove ρ indica il coefficiente di correlazione nei grappoli (*intragrappoli*), tutti di pari ampiezza (Cicchitelli et *alii.*, 1997; pp. 188-190).

• Nel campionamento **stratificato** l'efficienza del disegno *aumenta* al *crescere* del *numero* degli strati. Misurando la relazione tra la variabile di interesse e le variabili di stratificazione con il coefficiente di correlazione multipla R^2, si dimostra che (Cochran, 1977; p. 133) l'effetto della stratificazione con L strati, misurato con l'effetto del disegno, è tale da ottenere:

$$Deff \geq 1 - R^2\,\frac{L^2 - 1}{L^2}. \qquad [10]$$

Dalla [10] si evince che all'aumentare di L il *Deff* tende *rapidamente* ad $(1 - R^2)$, che ne costituisce il valore minimo. Ad esempio, per $L = 3$ dalla [10] si ha $(1 - 8/9\ R^2)$, ma per $L = 4$ il *Deff* è già $(1 - 15/16\ R^2)$, che dà un valore vicino al minimo.

Se ad esempio tra il carattere di interesse e le variabili di stratificazione il coefficiente di correlazione lineare è $R = 0{,}85$, si ottiene:

$$Deff = 1 - 0{,}7225 \cdot 15/16 = 0{,}3227,$$

ciò significa che, a parità di ampiezza, il campionamento casuale semplice ha efficienza pari solo al 32% del campionamento a *quattro* strati.

Dalla [10] si deduce che conviene effettuare la stratificazione utilizzando **più di una variabile**, invece che più *modalità* di una medesima variabile; inoltre, le basi di stratificazione dovrebbero risultare *incorrelate* o poco correlate tra di loro (solo così, infatti, il valore di R^2 può aumentare con il minore numero di variabili).