Laboratorio

di

Statistica - 2

IL CAMPIONAMENTO

Qualche cenno storico

Un primo tentativo di formalizzare la teoria del campionamento è dovuto a Anders Kiaer, capo dell'ufficio centrale di statistica norvegese, che nel 1895 propone alla Società Internazionale di Statistica un *metodo rappresentativo* di selezione dei campioni al fine di ottenere dei campioni rappresentativi ed utili per l'inferenza sull'intera popolazione.

Il problema centrale del metodo campionario è dimostrare la *rappresentatività* del campione rispetto alla popolazione.

Kiaer utilizza un concetto intuitivo di rappresentatività: il campione deve essere una *miniatura* della popolazione, ovvero il più possibile simile alla popolazione per tutte quelle variabili che sono note dal Censimento.

Questa scelta delle unità da estrarre ha tuttavia dei limiti:

- ✓ richiede che siano disponibili informazioni "certe" sulla popolazione (Censimenti);
- ✓ non esiste un criterio oggettivo che dimostri la rappresentatività del campione; se il campione è rappresentativo rispetto ad un insieme di caratteristiche rispetto a cui possiamo verificare la sua somiglianza con la popolazione non abbiamo elementi per dire che lo sarà anche su altre.

Nei primi decenni del 1900 esperienze parallele vengono condotte negli *Stati Uniti*, soprattutto nell'ambito dei sondaggi di opinione (es.: previsione dei risultati delle elezioni presidenziali). Si arriva così a definire il *campionamento per quote*.

Il campionamento per quote non è molto dissimile dal metodo di Kiaer: si divide la popolazione in sottoinsiemi disgiunti (e tali che uniti ci diano l'intera popolazione) in base ad alcune caratteristiche demografiche e sociali: ad esempio classe d'età, gruppo razziale, regione di residenza.

Le quote di queste sottopopolazioni sul totale della popolazione vengono assunte note sulla base dei risultati del Censimento più recente.

Un *campione estratto* dalla popolazione dovrà rispettare le *quote* dei vari sottoinsiemi nella popolazione.

Nel rispetto delle quote, il campionamento è piuttosto libero. Il campionamento per quote permette di ottenere alcuni risultati brillanti: ad esempio si afferma l'idea che la ***dimensione del campione è molto meno importante della sua rappresentatività***.

Un esempio molto noto riguarda la previsione delle elezioni presidenziali del 1936 negli USA.

Una rivista, il *Literary Digest* promuove una grande indagine postale per prevedere l'esito delle elezioni inviando centinaia di migliaia di questionari sia ai suoi abbonati, sia ad altri destinatari selezionati dagli elenchi telefonici e da quelli di iscritti a vari club e associazioni sportive.

L'esito del sondaggio è che la vittoria dovrebbe essere per Landon, attribuendogli il 57% dei voti. Molti non rispondono.

In parallelo l'istituto *Gallupp*, che allora è una piccola impresa, sulla base di un campione molto più piccolo (qualche migliaio) prevede invece la vittoria di Roosvelt.

L’istituto *Gallup* usa una tecnica di campionamento detta “per quote” ossia il campione deve rispettare le quote di popolazione di alcune caratteristiche (es. sesso, gruppo etnico, età, regione di residenza).

Alle elezioni Landon si ferma al 38.5% e Roosvelt trionfa.

Il campione del *Literary* era enorme, ma tragicamente non rappresentativo. I suoi lettori erano benestanti, di cultura medio-alta, tendenzialmente più repubblicani che non democratici. Anche avere un telefono o essere soci di un club non era socialmente e politicamente neutro nell’America degli anni ’30.

Errore copertura: le liste della popolazione utilizzate non erano complete.

Errore non risposta: caratteristiche rispondenti ≠ caratteristiche non rispondenti.

Tuttavia manca ancora un *criterio oggettivo* per la rappresentatività del campione e una teoria matematica rigorosa capace di “misurare” l’approssimazione implicita nella stima campionaria.

Il campione Gallup del 1936 utilizza un’intuizione felice delle variabili rispetto a cui era necessario bilanciare il campione (gruppo etnico, regione di residenza, sesso), ma questo meccanismo non necessariamente funziona sempre.

Il limite dei campioni ragionati o per quote consiste in prima approssimazione nel *rischio di non essere rappresentativi* per effetto dell’infelice scelta delle variabili di definizione delle quote o in generale per la scarsa conoscenza della popolazione. Questo problema ne porta inoltre con sé un altro: la *soggettività*.

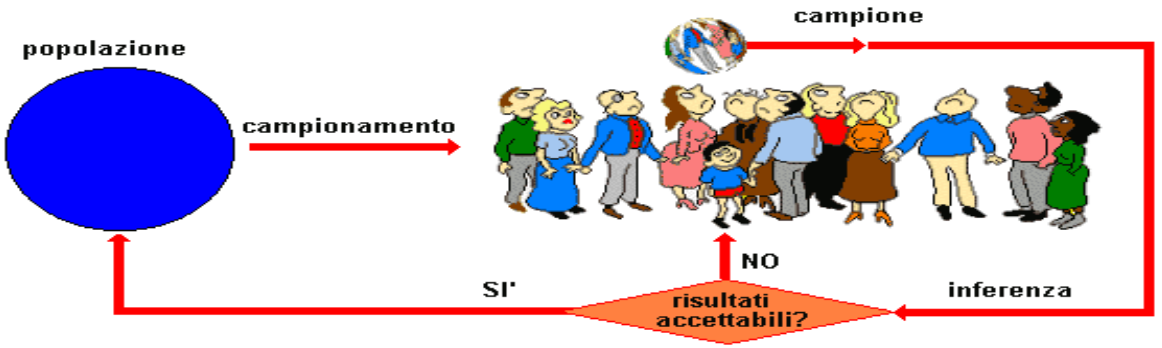
Se il meccanismo di selezione contiene alcune scelte soggettive, una volta che i risultati sono pubblicati, essi saranno sempre esposti agli attacchi di chi non li gradisce in nome della possibilità, mai eliminabile, di non rappresentatività.

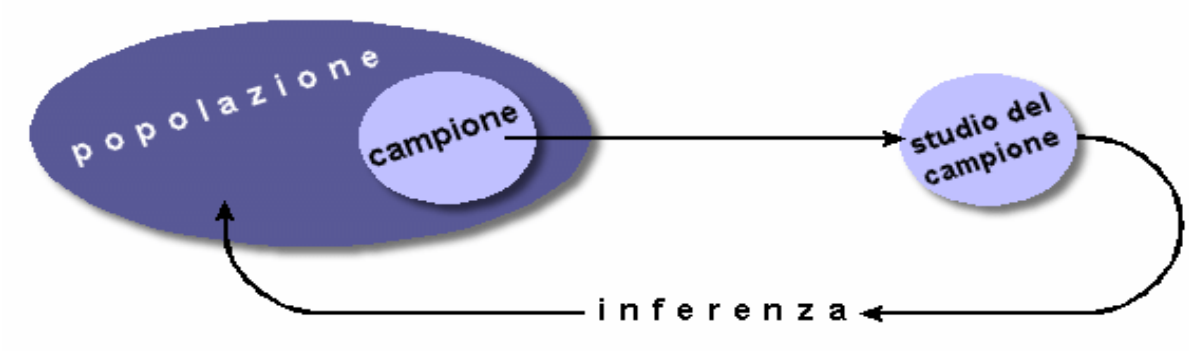
Il campionamento ragionato e per quote, per quanto condotti con onestà e attenzione limitano quindi l'intersoggettività dei risultati, un'esigenza che è invece fondamentale in statistica.

Ciò che serve è quindi *un metodo di campionamento che elimini in modo radicale la possibilità di un condizionamento soggettivo nella selezione del campione*. La soluzione a questo problema della rappresentatività è rappresentata dal *campionamento casuale*.

L’idea base del campionamento casuale è che se trattiamo le unità della popolazione come palline in un’urna e attribuiamo a ciascuna la stessa probabilità di essere estratta, il campione è rappresentativo in quanto il meccanismo di selezione non contiene nessun elemento soggettivo capace di distorcere in un senso o in un altro i risultati.

In un'indagine campionaria l'obiettivo è quello di generalizzare quanto osservato su un campione, attraverso procedure inferenziali appropriate, all'intera popolazione o a suoi sottoinsiemi.





Gli elementi essenziali sono:

1. La popolazione finita

In statistica con “popolazione” non si fa riferimento necessariamente ad un insieme di individui, ma più genericamente ad un insieme di elementi che possono essere individui ma anche famiglie, imprese, porzioni di territorio, istanti temporali successivi *etc*… Una popolazione viene detta finita se è composta da un numero finito di elementi. Inoltre gli elementi di una popolazione finita devono essere distinguibili gli uni dagli altri. Per ciò è essenziale che la popolazione sia definita in modo preciso nel tempo e nello spazio.

2. Un campione è un sottoinsieme di una popolazione

Nelle indagini campionarie procediamo ad osservare alcune caratteristiche relative agli elementi della popolazione appartenenti al campione. Il fatto di concentrare lo sforzo di osservazione su una parte e non sul tutto è ciò che caratterizza le indagini campionarie rispetto ai Censimenti.

3. Le caratteristiche osservate sulle unità campionate vengono dette caratteri.

I caratteri rilevabili sono i più vari (su una popolazione di individui: reddito, titolo di studio, opinione su una determinata legge, grado di soddisfazione nel proprio lavoro, numero di visite mediche nell'ultimo anno, ecc…).

4. L’obiettivo è di fare inferenza su alcune caratteristiche della popolazione nel suo complesso.

L’obiettivo non è il campione estratto, ma la popolazione dalla quale esso è stato estratto.

Gli elementi che costituiscono la popolazione vanno definiti in modo non ambiguo in modo tale che sia chiaro a cosa si riferiscono le informazioni prodotte attraverso l’indagine. Occorre cioè definire l’unità statistica dell’indagine.

Dopo aver estratto le unità statistiche, queste vengono contattate per valutare la loro la disponibilità a partecipare all’indagine e successivamente vengono raccolti i dati.

Esistono vari mezzi in un'indagine per raccogliere le risposte delle unità campionate alle domande che formano l'oggetto dell'indagine.

Il contatto può essere effettuato:

- con contatto personale (intervista personale)
- per telefono (intervista telefonica),
- per lettera (intervista postale o via mail),
- per e-mail (questionario inviato via web).

Intervista personale	
Vantaggi	Svantaggi
• Flessibilità (l'intervistatore può spiegare le domande, aiutare il rispondente ecc…) • E' possibile somministrare questionari anche lunghi • Produce alti tassi di risposta alle singole domande • L'intervistatore è in grado di valutare e controllare i condizionamenti e di raccogliere informazioni sul contesto in cui avviene l'intervista	• E' molto costosa (incidono soprattutto i costi di viaggio) L'intervista è una forma di conversazione sociale per cui • errori di risposta dovuti all'effetto della "desiderabilità sociale" o dell'"acquiescenza" • L'intervistatore può influenzare le risposte • L'intervistatore può falsificare in tutto o in parte l'intervista

Intervista telefonica	
Vantaggi	Svantaggi
• Costi più contenuti • Possibilità di una maggiore standardizzazione e controllo del comportamento degli intervistatori (organizzazione di call centers) • L'effetto intervistatore e desiderabilità sociale sono solitamente più contenuti (minor interazione tra intervistatore e rispondente) • Campioni geograficamente dispersi non causano problemi	• Minore flessibilità (comunicazione solo verbale) • Necessità di semplificare le domande e ridurre la lunghezza del questionario • Minore libertà nella predisposizione delle risposte (es.: numero di alternative contenuto) • Minore possibilità di controllare le condizioni al contorno dell'intervista • Mancata copertura delle unità senza telefono

Intervista postale	
Vantaggi	Svantaggi
• Costi molto bassi • Nessuna interazione tra intervistatore e intervistato (e quindi niente desiderabilità sociale o effetto intervistatore) • Possibilità di affrontare temi sensibili • L'intervistato ha tempo per rispondere (ovvero si può prendere tutto il tempo che ritiene necessario per rispondere)	• Nessun controllo sul processo di risposta (chi risponde? Il rispondente risponde da solo?) • Nessuna flessibilità (il rispondente non può chiedere nessuna informazione) • Il rispondente deve essere alfabetizzato e in grado di leggere testi di una certa complessità • Bassi tassi di risposta sia in generale, sia relativi a singole domande

Questionario sul web	
Vantaggi	Svantaggi
• Costi e tempi di invio e raccolta virtualmente nulli • Nessuna interazione tra intervistatore e intervistato • L'intervistato ha tempo per rispondere • Spiegazioni alle domande in ipertesto • Grande libertà di disegno del questionario, utilizzo di strumenti grafici, multimediali • Possibilità di controllare la qualità e coerenza dei dati durante la compilazione. • Possibilità di elaborazioni in tempo reale e fornitura di feed-back al rispondente. • Controllo dei tempi e modalità di risposta	• Costi di preparazione tecnica del questionario non irrilevanti • Il rispondente deve essere alfabetizzato dal punto di vista informatico e in grado di navigare nel questionario. • Il rispondente deve essere un utente internet. • Nessun controllo sul processo di risposta (chi risponde? Il rispondente risponde da solo?) • Limiti alla flessibilità (disegno dell'ipertesto). • Tendenza a risposte poco motivate e di bassa qualità (frettolose) • Bassi tassi di completamento e risposta. "Drop out" alla prima difficoltà.

All’interno delle modalità di raccolta dati è in corso una profonda rivoluzione generata dall’avvento della “information technology” che ha cambiato il modo in cui interviste personali e telefoniche vengono condotte.

Modalità Computer Assisted:

- CAPI, Computer Aided Personal Interviewing
- CATI, Computer Aided Telephone Interviewing
- CASI Computer Aided Self Interviewing

La Computer Assistance consiste essenzialmente nella possibilità da parte dell’intervistatore di archiviare direttamente i dati in un computer man mano che il rispondente li fornisce.

I vantaggi principali della Computer Assistance sono principalmente:

- riduzione degli errori di editing e controllo in tempo reale della coerenza interna dei dati forniti dal rispondente;
- eliminazione della fase di trasferimento dei dati da supporto cartaceo a magnetico con conseguente riduzione dei tempi di elaborazione dati;
- ampliamento (soprattutto per le interviste personali) delle possibilità di presentazione delle domande.

I metodi di campionamento

E’ possibile distinguere tra metodi di campionamento:

- **probabilistici** – è possibile definire a priori la probabilità che ogni unità ha di essere estratta,
- **non probabilistici** – non è possibile definire a priori tale probabilità.

Campionamenti probabilistici

Campionamento casuale semplice

Tutti gli elementi della popolazione vengono presi in considerazione ed hanno tutti *uguale probabilità* di essere selezionati; ognuno di essi cioè può "casualmente" costituire una delle unità del campione.

Il campionamento avviene estraendo unità per unità gli **N** elementi della popolazione fino ad ottenere le **n** unità del campione.

Per procedere in questo modo si deve disporre di un elenco – ***LISTA*** – numerato da 1 a N, degli elementi della popolazione tra i quali vengono presi quelli i cui numeri corrispondono ad una successione di **n** numeri estratti casualmente.

Campionamento casuale

Vantaggi

- E’ il tipo di campionamento più semplice
- Minima conoscenza a priori delle caratteristiche della popolazione
- Garantisce una scelta obiettiva delle unità da rilevare
- Conveniente quando la popolazione non è molto grande

Svantaggi

- Costi di rilevazione più elevati
- Stime dei parametri della popolazione meno precise
- Non vengono utilizzate tutte le informazioni che si posseggono sulla popolazione

Campionamento sistematico

E’ equivalente, dal punto di vista del risultato, al campionamento casuale semplice.

Le unità non vengono estratte mediante sorteggio, ma si scorre la lista dei soggetti selezionandone sistematicamente uno ogni dato intervallo; ad es. una unità ogni *k*, *k* = *intervallo di campionamento*.

E’ utile anche quando manca la lista della popolazione ed N è sconosciuto (ad es. quando si intervista una persona ogni *tot* all’uscita da un seggio elettorale, da un supermercato, …).

Tutte le unità *devono* avere la stessa probabilità di essere incluse nel campione.

Campionamento casuale stratificato

Permette di raggiungere una più grande accuratezza ad uno stesso costo o, analogamente, con un minor costo la medesima accuratezza.

In una prima fase gli N elementi della popolazione vengono suddivisi in ***k*** gruppi o ***strati*** il più possibile omogenei fra di loro rispetto ad una opportuna caratteristica.

Ciò significa che ogni strato è formato in modo tale che non ci sia sovrapposizione, cioè ogni elemento compare in un solo strato. Quindi il campione viene formato estraendo da ogni strato unità in modo indipendente.

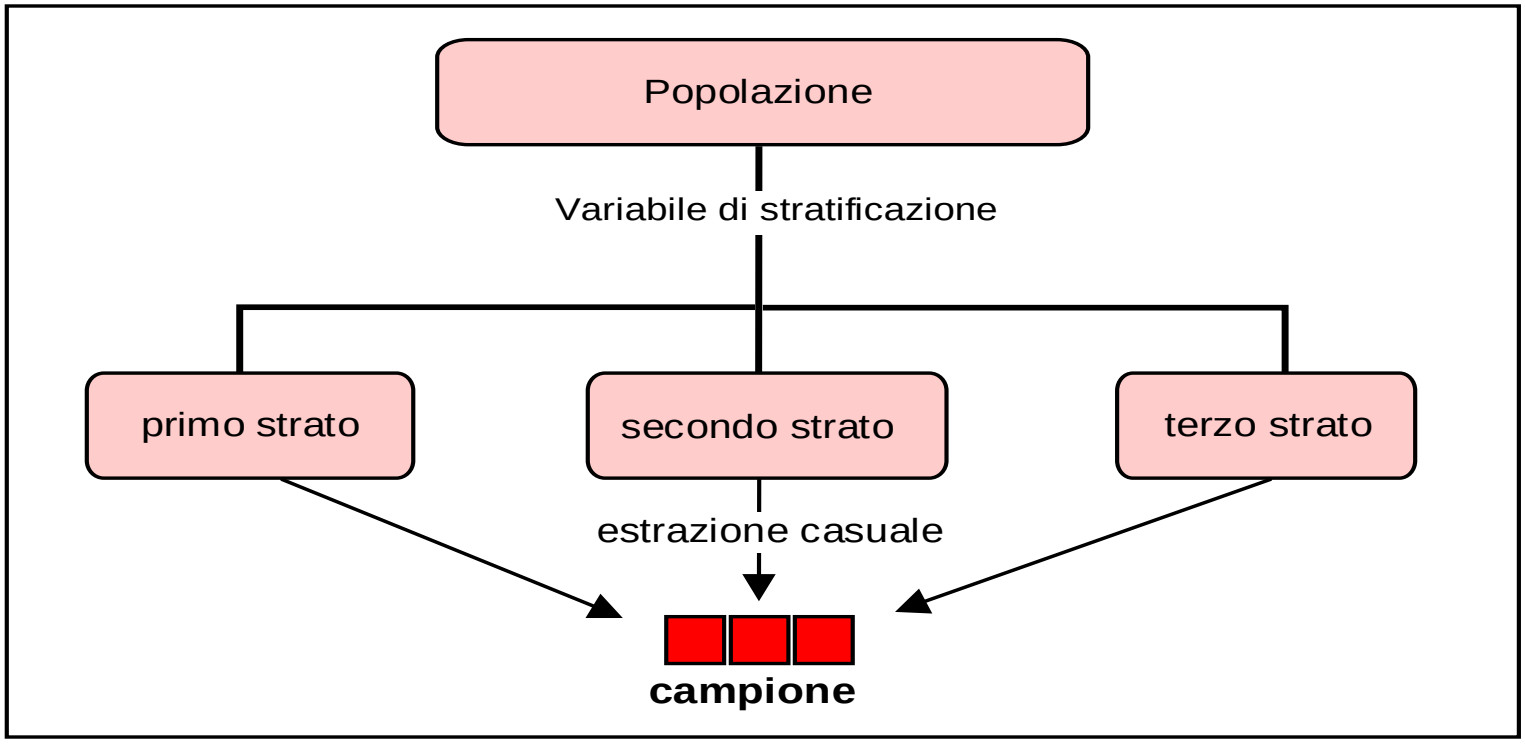
I motivi che rendono frequente il ricorso a questa tecnica campionaria possono essere:

1. Se si desidera un certo grado di precisione per certe suddivisioni della popolazione è consigliabile trattare ogni "suddivisione" come una popolazione a se stante.

2. L'analisi campionaria può riguardare diversi tipi di popolazioni che potrebbero essere difficilmente accorpati in una unica popolazione ed inoltre lo svolgimento dell'indagine può essere effettuato in tempi diversi da differenti operatori.

3. E’ possibile dividere una popolazione altamente eterogenea in sottopopolazioni ognuna delle quali al suo interno sia omogenea. Ciò, oltre a consentire analisi di

popolazioni i cui elementi sono molto diversi tra loro, permette anche di ridurre l'ampiezza del campione nei singoli strati composti da elementi omogenei tra loro, così che presentano una piccola variabilità interna.



Si parla di:

- *campionamento stratificato proporzionale* – se si conosce la proporzione di unità per ogni strato, per ottenere un campione più rappresentativo, si estrae da ogni strato una certa quantità di unità in proporzione alla numerosità dello strato;

- *campionamento stratificato "ottimale"* – quando la variabilità interna di ogni strato è elevata, per migliorare l'efficienza del campionamento, si estrae da ogni strato una frazione differente di unità. Si ricerca l'ottima ripartizione.

Campionamento stratificato

Vantaggi

- Aumenta la precisione delle stime senza accrescere la dimensione totale del campione
- Utile quando la distribuzione statistica della variabile da rilevare è fortemente asimmetrica

Svantaggi

- Costruzione degli strati può risultare alquanto costosa
- Se la stratificazione è errata si possono ottenere risultati fuorvianti

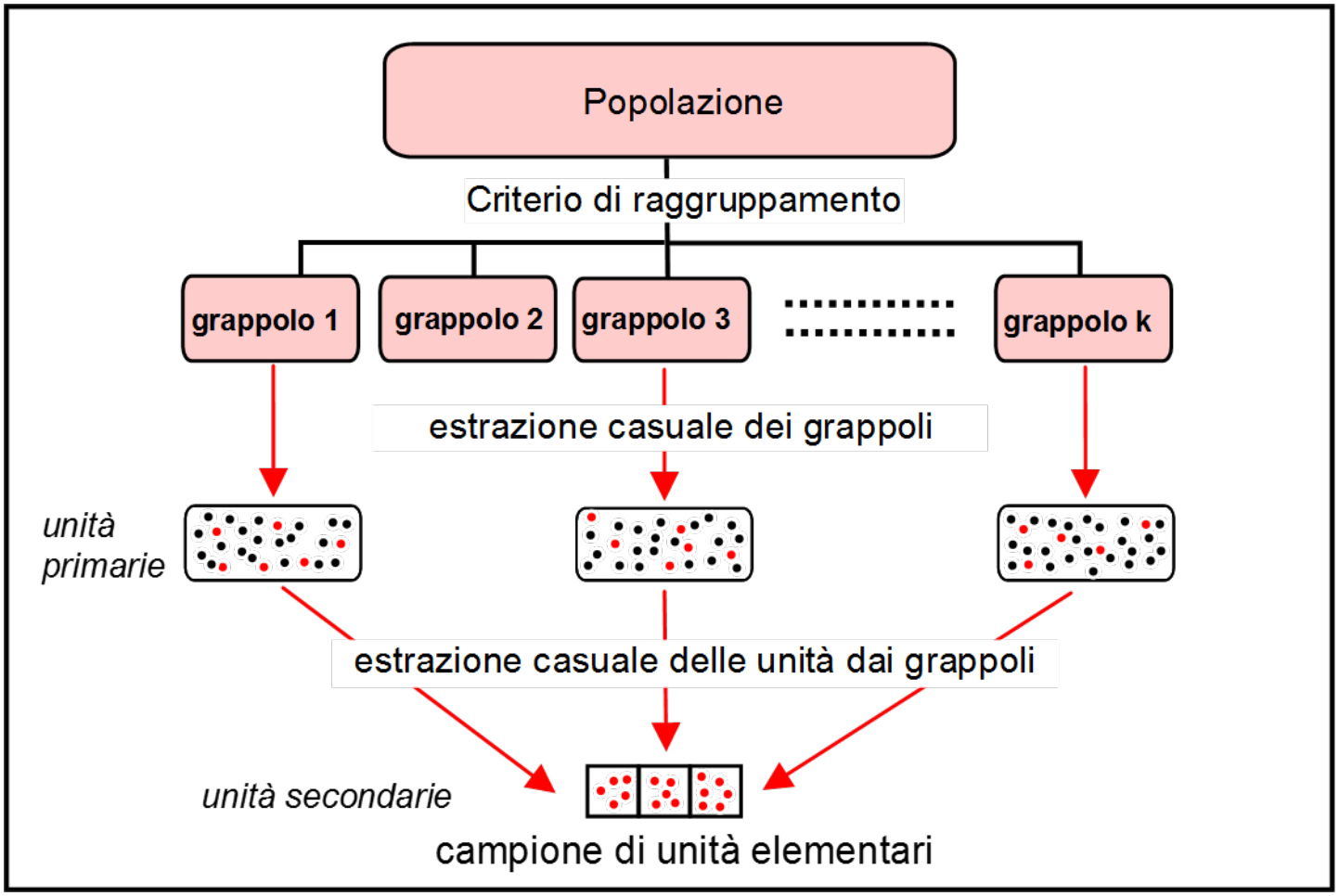
Campionamento casuale a grappoli

La popolazione viene suddivisa in un gran numero di sottoinsiemi detti *grappoli* (clusters), si effettua un campionamento casuale tra i grappoli e vengono considerate tutte le unità del grappolo.

Ad esempio si selezionano 20 quartieri di una città avente 100 quartieri ed si inseriscono nel campione tutti gli abitanti dei 20 quartieri.

Il metodo non prevede quindi il campionamento diretto degli elementi, ma vengono campionati *grappoli di elementi*.

Viene spesso fatto ricorso ad un tale campionamento per ridurre il costo della raccolta dei dati.


Popolazione
Criterio di raggruppamento
grappolo 1
grappolo 2
grappolo 3
grappolo k
estrazione casuale dei grappoli
unità primarie
estrazione casuale delle unità dai grappoli
unità secondarie
campione di unità elementari

Campionamento casuale a più stadi

E' una tecnica di campionamento che risulta molto vantaggiosa quando la popolazione da studiare è molto numerosa e gli elementi possono essere raggruppati in diversi sottoinsiemi.

Dopo aver suddiviso la popolazione di partenza in successive sottoclassi o stadi (es. province, comuni, scuole e così via), si estrae un campione di unità di primo stadio (province) e nell'ambito delle unità ottenute si procede alla scelta dei campione di secondo stadio (comuni) e cosi via. Il campione è costituito dalle unità estratte dall'ultimo stadio.

Questa tecnica viene spesso utilizzata, ad esempio, nei sondaggi di opinione, quando si procede al successivo campionamento delle città, quindi dei quartieri ed in ultimo dei soggetti da intervistare.

Il campionamento viene definito a **k** stadi in riferimento al numero dei campionamenti successivi (nel caso città – quartieri – persone si parla di campionamento a 3 stadi).

Un altro esempio di campionamento a più stadi:

I stadio: campioniamo distretti scolastici dalla lista dei distretti;

II stadio: campioniamo scuole elementari dalla lista delle scuole del distretto;

III stadio: campioniamo classi dalla lista delle classi delle scuole;

IV stadio: campioniamo alunni dalle classi.

Il campionamento a più stadi ha diversi vantaggi:

- semplifica il problema di ottenere liste adeguate nelle indagini di media e grande scala. Permette di non dover disporre di una lista di tutte le unità della popolazione, ma solo di quelle che vengono campionate ad ogni stadio;
- riduce i costi di rilevazione, pur preservando una buona rappresentatività geografica dei campioni.

Campionamento per aree

E' molto utile quando manca la lista della popolazione.

Il territorio viene suddiviso in aree via via più piccole fino ad individuare le unità da intervistare.

Utile anche quando le liste risultano essere incomplete; con questa tecnica si rilevano anche i residenti temporanei, i clandestini, … .

Campionamenti non probabilistici

In molte indagini risulta impossibile o comunque impraticabile ricorrere al campionamento casuale in quanto non risultano valide alcune implicazioni procedurali statistiche.

Di fatto si è costretti a campionare una parte della popolazione che risulta quella realmente accessibile. Ad esempio egli esperimenti di laboratorio su animali il ricercatore frequentemente è costretto ad "usare" quegli animali di cui dispone; altrimenti, se dovesse selezionare casualmente le cavie, difficilmente riuscirebbe a fare della ricerca. Naturalmente anche in queste situazioni è lecito supporre che il campione sia equivalente ad un campione casuale, non essendoci grossi motivi per ritenere che gli animali di cui dispone non siano rappresentativi.

In molti progetti di ricerca in campo sanitario, si è costretti a ricorrere a dei campioni ad *hoc*, i cui elementi sono dei "volontari" o comunque soggetti "disponibili" a sottoporsi alla ricerca. Ciò accade quando, ad esempio, si indaga su situazioni delicate, oppure quando questionari vengono inviati per posta e la percentuale dei non rispondenti potrebbe comunque alterare il campione (in quanto coloro che non accettano di rispondere hanno delle caratteristiche diverse dai rispondenti).

In altre situazioni è possibile introdurre la casualità nell'esperimento anziché nei soggetti; nel confronto di due trattamenti, ad esempio, i pazienti selezionati vengono casualmente attribuiti ad uno o all'altro trattamento; in tal caso le implicazioni statistiche (inferenza), saranno applicate ai trattamenti anziché ai soggetti.

Sotto le giuste condizioni, quindi, i campioni non probabilistici possono dare utili risultati. Essi però non sono trattabili con la teoria campionaria perché non si basano su alcun principio di casualità pertanto la loro validità è strettamente legata alla situazione cui si riferiscono e non si ha nessuna garanzia della loro validità in circostanze diverse.

Questi metodi "non corretti" sono meno impegnativi da applicare e, sebbene l'accuratezza sia discutibile, trovano frequentemente pratica applicazione. Sono i metodi molto usati da organizzazioni per le ricerche di mercato, per sondare gli orientamenti elettorali nella popolazione ecc.

Un esempio in ambito sanitario può essere il seguente. Si vuole valutare l'efficacia di un nuovo vaccino contro l'AIDS. Tutti i soggetti a rischio che volontariamente si presentano ai centri sieroprofilattici e che risultano sieronegativi vengono sottoposti alla vaccinazione. Avendo stimato in 1000 la numerosità campionaria richiesta, si procede al reclutamento di tutti i soggetti in campo nazionale fino al raggiungimento del numero previsto.

Campionamento accidentale

Si ha un campionamento accidentale (o di convenienza) quando il ricercatore sceglie come rispondenti alla sua indagine le prime persone che capitano, senza criteri definiti.

Ciò che si perde in accuratezza del campione, lo si risparmia in tempo e denaro.

Per questo motivo non è possibile applicare tecniche di inferenza statistica partendo da un campione così selezionato.

Campionamento per quote

E’ un tipo di campionamento non probabilistico che equivale al campionamento stratificato da cui si differenzia perché ogni strato è generalmente rappresentato nella stessa proporzione, proporzione che ha nella popolazione complessiva (che deve essere finita).

Dopo aver deciso quali strati possono essere rilevanti per l'indagine che si deve condurre, si stabilisce per ogni strato una quota proporzionata alla sua consistenza nella popolazione complessiva.

Occorre dividere la popolazione in strati il più possibile omogenei al loro interno e il più possibile eterogenei tra di loro.

La selezione degli individui negli strati viene lasciata agli intervistatori. Il totale di elementi campionati deve essere *n*.

In questo tipo di campionamento vengono occultati i problemi di "non risposta". I risultati possono essere distorti a causa della discrezionalità degli intervistatori.

Anche in questo caso non sono applicabili le tecniche della statistica inferenziale.

Campionamento a valanga

Questo tipo di campionamento è composto da più fasi: dopo aver intervistato alcune persone dotate delle caratteristiche richieste, queste persone servono per identificare altri soggetti che possono essere intervistati in una fase successiva e che a loro volta producono informazioni per identificare altri soggetti con le caratteristiche per essere inclusi nel campione, creando così un effetto a valanga.

Non sono applicabili le tecniche della statistica inferenziale.

Campionamento per dimensioni

In questa tecnica dopo aver specificato tutte le dimensioni (variabili) che ci interessa studiare all'interno della popolazione (che deve essere finita), si verifica che per ogni possibile combinazione delle diverse dimensioni ci sia almeno un caso. In tal modo anche con un campione piccolo si possono studiare le dimensioni suddette senza correre il rischio di avere combinazioni non rappresentate.

Non sono applicabili le tecniche della statistica inferenziale.

Campionamento a più stadi

Dopo aver suddiviso la popolazione (finita) in gruppi, si estrae un campione casuale di sottogruppi all'interno di ogni gruppo. Si ripete il processo fino a che non si giunge all'estrazione delle unità di analisi prescelte.

Non sono applicabili le tecniche della statistica inferenziale.

Campionamento a elementi rappresentativi

Si ha un campionamento a elementi rappresentativi quando si selezionano all'interno della popolazione gli elementi che il ricercatore ritiene rappresentativi per gli obiettivi della ricerca.

Non sono applicabili le tecniche della statistica inferenziale.