Analisi dei Dati

Domenico De Stefano

a.a. 2022/2023

Indice

Definizione target population

Le attività relative ad un'indagine statistica partono dalla definizione della target population
Quando definiamo la target population dobbiamo individuare con precisione l'insieme di unità statistiche alle quali si intende estendere i risultati dell'indagine.

- In questa fase occorre specificare le condizioni di eleggibilità, ovvero le caratteristiche che consentono di risolvere l'unità statistica ossia di determinarne la sua inclusione (o esclusione) dalla popolazione

Operativamente ci si concentrerà sulla frame population, ossia sulla lista disponibile di unità statitiche da includere nel campione
(non è escluso infatti che la target population, compatibilmente con l'obiettivo d'indagine, possa essere aggiustata sulla base della lista disponibile)

Definizione della frame population e disegno di campionamento

La realizzazione concreta della target population è data dall'insieme di operazioni che consistono in:

- costruzione della lista di selezione relativa alla target population e contenente, per ciascuna unità della popolazione, le informazioni identificative e necessarie per il contatto, eventuali variabili ausiliarie utili per la definizione del campione (variabili di stratificazione, variabili identificative degli eventuali stadi di selezione, ecc.)
- progettazione del disegno di campionamento che, sulla base degli obiettivi di ricerca e dei vincoli operativi e di costo, consenta di ottenere stime affidabili.

Progettazione della lista di campionamento

Le caratteristiche della lista di campionamento sono rilevanti per la corretta definizione del disegno di campionamento.

- E' necessario che la lista risponda a criteri di qualità in termini di aggiornamento, copertura e accuratezza delle informazioni in essa riportate.
- teoricamente la lista di selezione ideale dovrebbe possedere i seguenti requisiti:
  - essere costituita dalle sole unità appartenenti alla popolazione di interesse al momento di riferimento dell'indagine;
  - includere ogni unità della popolazione una sola volta;
  - contenere dati aggiornati e corretti relativamente alle informazioni identificative (nome e indirizzo) e alle eventuali informazioni descrittive (altri dati strutturali importanti) delle unità.

Progettazione della lista di campionamento (2)

- Le possibili situazioni di allontanamento dalla lista ideale sono:
  - sottocopertura, nel caso in cui alcuni elementi della target population non sono contenuti nella lista e non devono, pertanto, essere inclusi nel campione;
  - sovracopertura, quando alcuni elementi della lista sono inesistenti e/o non appartengono alla target population;
  - duplicazione di alcune unità, se alcuni elementi della popolazione sono presenti più volte nella lista;
  - grappoli di unità, quando alcuni elementi della lista contengono più elementi della target population

Indice

Dalla definizione della popolazione al campionamento

Concetti chiave

I concetti chiave del campionamento sono i seguenti:

- con popolazione **P** (la target population) indichiamo un insieme di N unità statistiche U_i, con $i = 1, 2, \ldots, N$
- Oggetto del campionamento da una popolazione finita è la selezione di un sottoinsieme $\mathbf{S} \subset \mathbf{P}$, detto campione, la cui ampiezza n (detta numerosità campionaria o sample size) è molto minore di N.
- Scopo del campionamento è di esaminare le unità statistiche di **S** per studiare una (o più) variabile X la quale nella popolazione $\mathbf{P} = (U_1, U_2, \ldots, U_N)$ assume valori $X_1, X_2, \ldots, X_N$ in corrispondenza di ciascuna unità statistica $U_1, U_2, \ldots, U_N$

Concetti chiave (2)

- Il rapporto tra numerosità campionaria e numerosità della popolazione $\frac{n}{N}$ è detto frazione di campionamento (o tasso di sondaggio)
- Il campione **S** è un sottoinsieme di unità statistiche di **P** per il quale sono note le etichette $\{i_1, i_2, \ldots, i_n\}$ ossia codici numerici (o alfanumerici) che consento di identificare univocamente ciascuna unità statistica "campionata" le cui informazioni di contatto sono presenti nella lista di campionamento (sampling frame o fram population)
- Definiamo spazio campionario $\mathbf{\Omega_n(S)}$ l'*insieme di tutti i possibili campioni di numerosità n* derivabili mediante un prescelto disegno di campionamento. La numerosità dello spazio campionario (il numero di campioni possibili) è $\binom{N}{n}$ se non ammettiamo ripetizioni (vedi dopo) per cui la probabilità di estrarre un certo campione è $\binom{N}{n}^{-1}$
- Lo studio della popolazione **P** avviene riassumendo gli aspetti più importanti delle variabili X, Y, ecc. mediante parametri, ad es.:

Concetti chiave (3)

- media: $\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i$
- varianza $\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)^2$
- covarianza $\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)$ e correlazione $r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

Concetti chiave (4)

- Disegno di campionamento: è la definizione di una procedura di selezione di n unità statistiche per formare il campione **S**
- Schema di campionamento: è la procedura operativa (il modo concreto) con cui si perviene alla effettiva estrazione delle n unità statistiche di **S**

Modalità di campionamento

Con riferimento alle modalità di campionamento si distingue tra:

- Campionamento con ripetizione (o campionamento Bernoulliano)
  - la modalità più semplice e con le proprietà statistiche più "desiderabili"
  - le unità statistiche nella popolazione **P** una volta estratte vengono reintrodotte nella popolazione e possono essere riestratte
  - si immagini un'urna in cui le palline (che rappresentano le unità statistiche): Una pallina, dopo essere stata estratta, entra a far parte del campione **e** viene re-immessa nuovamente nell'urna
- Campionamento senza ripetizione (o campionamento in blocco):
  - Le unità statistiche nella popolazione **P** una volta estratte **non** vengono reintrodotte nella popolazione e possono essere riestratte
  - Una pallina, dopo essere stata estratta, entra a far parte del campione **e non** viene re-immessa nell'urna

Probabilità di inclusione e probabilità di selezione

In relazione alle unità statistiche si definiscono:

- Probabilità di inclusione
  - La probabilità di inclusione (del primo ordine) indica la probabilità che una generica unità statistica U_i, appartenente ad una popolazione **P** di numerosità sia inclusa nel campione estratto sulla base di uno specifico disegno di campionamento
- Probabilità di selezione (o estrazione)
  - La probabilità di selezione indica la probabilità che una generica unità statistica U_i entri nel campione alla j-esima estrazione

Tipi di campionamento

- Campionamento probabilistico
  - Campionamento casuale semplice (CCS)
    - con ripetizione (CCSCR)
    - senza ripetizione (CCSSR)
  - Campionamento casuale sistematico (STM)
  - Campionamento casuale stratificato (STR)
  - Campionamento casuale a (due) stadi (STA)
  - Campionamento casuale a grappoli (GRA)
- Campionamento non probabilistico
  - Campionamento per quote
  - Campionamento a scelta ragionata
  - Campionamento a valanga

Indice

Campionamento casuale semplice con ripetizione (CCSCR)

- Il CCSCR è il disegno più semplice ed equivale all'estrazione di un campione **S** di n palline da un'urna **P** che contiene le N palline U_i, $i = 1, \ldots, N$ tutte identiche per forma, peso, dimensione, ecc. tranne che per le modalità di una variabile X "impressa" su di esse (che è la caratteristica che vogliamo studiare)
- Pertanto, la probabilità di estrarre una unità statistica è costante... in particolare si dimostra che:
  - La probabilità di inclusione è uguale alla frazione di campionamento $\frac{n}{N}$ per ogni unità
  - La probabilità di selezione è $\frac{1}{N}$ per ogni unità
- Il Campionamento casuale semplice (sia con, che senza ripetizione) presenta numerosi vantaggi concettuali e formali (è sicuramente rappresentativo della popolazione) ma ha l'inconveniente di richiedere una lista completa ed aggiornata delle unità statistiche per cui diventa ingestibile in caso di unità mancanti, estranee, o erroneamente riportate

CCSCR: come funziona

- La selezione di un campione di questo tipo si effettua molto semplicemente. Immaginiamo di avere la lista degli studenti iscritti alla corso di laurea in Scienze Politiche e dell'Amministrazione (ipoteticamente, $N = 3000$)
- Supponiamo di volere un campione di $n = 300$ soggetti
- In teoria si tratta di assegnare a ciascuno studente dei numeri da 1 a 3000, inserire in un'urna i numeri da 1 a 3000 ed estrarne 300. I possessori dei 300 numeri estratti entreranno a far parte del campione.
  - Se la numerazione dei soggetti è un'operazione sempre necessaria, il ricorso all'urna non lo è. Esistono programmi di computer che consentono la generazione di numeri casuali (o, meglio, pseudocasuali) analoghi a quelli che si estrarrebbero da un'urna.
  - In alternativa, si ricorre ad apposite tavole dei numeri casuali, prodotte da programmi come quelli citati, riportate in genere nei (vecchi) manuali di statistica

CCSCR: come funziona (2)

- nel caso specifico la frazione di campionamento è $\frac{300}{3000} = 1/10 = 0.1$
- la probabilità di selezione di ciascuno studente è $\frac{1}{3000} = 3.333e-4 = 0.0003$
- dimostriamo che la probabilità di inclusione è uguale alla frazione di campionamento.
  ragioniamo cosí: ogni unità ha probabilità $1/3000$ di essere estratto ad ogni estrazione. Le estrazioni sono 300. dunque la probabilità di tale inclusione è $300 \times \frac{1}{3000} = 1/10$

Campionamento casuale semplice senza ripetizione (CCSSR)

- È il campionamento più usato nelle scienze sociali (non intervisteremo mai due volte la medesima persona nella stessa indagine)
- L'unica differenza con il CCSCR è che qui l'estrazione deve avvenire senza reimmettere la "pallina" estratta nell'urna
- dal punto di vista operativo non cambia nulla. Se si usano programmi per la generazione di numeri casuali ovviamente si opererà impedendo la ripetizione del numero casuale già estratto tra gli n
- infatti si parla anche di campionamento a blocco, perché è come se si estraessero interi blocchi di n unità statistiche dalla popolazione

Campionamento casuale semplice senza ripetizione (CCSSR) (2)

Lista di campionamento nel CCS

L'unico problema nel CCS su cui bisogna porre attenzione è quello della lista di campionamento (la frame population). Essa deve contenere (possibilmente) tutti i membri della popolazione (la target population) e soltanto loro. Inoltre ogni unità statistica deve figurare una sola volta, altrimenti la probabilità di selezione sarebbe variabile da caso a caso.

Campionamento casuale semplice senza ripetizione (CCSSR) (3)

Applicazione reale del CCS

CCS con o senza ripetizione non è il disegno più usato per la difficoltà nel reperire le liste di campionamento soprattutto nei casi in cui la popolazione è distribuita in un territorio esteso: ad es., l'anagrafe del comune di Trieste ha la lista di tutti i cittadini, ma nessuno ha la lista dei cittadini della provincia di Trieste. Per avere quest'ultima occorrerebbe assemblare le liste dei cittadini di tutti i comuni della provincia e ciò ha dei costi non indifferenti (oltre ad ovvi problemi di accessibilità delle liste legati alla privacy). In casi come questi si preferisce il successivo tipo di campionamento...

Campionamento casuale sistematico (STM)

- Il campionamento sistematico si considera equivalente a quello casuale semplice e si può usare tutte le volte che si usa quest'ultimo. Differisce dal casuale semplice solo per la tecnica di estrazione
- Di solito si usa quando la popolazione consiste di unità statistiche organizzate in "elenchi" predisposti secondo un ordine logico, cronologico o di altro tipo indipendente rispetto alle variabili X che si intendono studiare.
- È il caso dei dati anagrafici, amministrativi, fiscali, commerciali. Per es. i residenti di un comune, coloro che hanno presentato la dichiarazione dei redditi, gli utenti delle concessionarie dei serivizi pubblici (luce, acqua, gas), i possessori di patenti, gli abbonati inclusi negli elenchi telefonici

Campionamento casuale sistematico (STM) (2)

- Affinchè il campionamento STM conservi la natura di "campionamento statistico" (probabilistico) ci si deve assicurare che l'ordinamento delle unità nelle liste **non dipenda** da alcuna delle variabili X che si vogliono studiare (ad es., negli elenchi telefonici l'ordinamento è alfabetico per cui si può supporre indipendenza tra lettera del proprio cognome ed una caratteristica da studiare)
- L'importanza dell'indipendenza tra ordinamento e caratteristiche oggetto di studio è data dal fatto che di solito il campionamento STM si basa sull'ordine delle unità statistiche nelle liste

Campionamento STM: come funziona

- Si scorre la lista di campionamento (nel suo ordine "naturale", ossia così come ci viene fornita) e si seleziona un'unità statistica ogni k
- dove k è un numero intero che si chiama passo di campionamento (o intervallo di campionamento)
- Il valore di k è pari a N/n, dove come al solito N è l'ampiezza della popolazione e n l'ampiezza desiderata del campione.
- Nell'esempio degli studenti di Scienze politiche e dell'amministrazione $n = 300$ e $N = 3000$, perciò $k = 3000/300 = 10$.
- Si seleziona pertanto uno studente ogni 10, ad esempio il primo, l'undicesimo, il ventunesimo, ecc.

Campionamento STM: come funziona (2)

- Non si deve partire per forza dal primo (altrimenti tutti i campioni sistematici estratti da questa lista sarebbero uguali tra loro). Per ottenere un campione esattamente di 300 casi è sufficiente partire da uno qualunque degli studenti compresi tra il primo e il decimo.
- Di solito si estrae casualmente un numero compreso tra 1 e k (10) e si inizia dal soggetto corrispondente al numero estratto.
- **NB**: È evidente che l'estrazione di questo tipo di campionamento è indipendente dal supporto su cui si trova la lista (vanno bene anche schedari con una scheda per unità) e non è strettamente necessario numerare le unità

Campionamento STM: come funziona (3)

Lista di campionamento nel STM

Rispetto al CCS nel STM c'è un problema aggiuntivo. Oltre alla indipendenza dell'ordinamento con le variabili X da studiare, la lista non deve contenere delle ricorrenze che abbiano lo stesso passo del campionamento. Ad es., se $k = 10$ e la lista comprende militari elencati per squadra, prima il sergente, poi invariabilmente 10 militari semplici, è chiaro che si selezionano soltanto i sergenti, oppure soltanto militari semplici, secondo il punto di partenza dell'estrazione. Di solito, l'elenco alfabetico esclude periodicità di questo tipo.

Campionamento STM: come funziona (4)

Applicazione del STM senza lista di campionamento

La particolarità di questo tipo di campionamento è che può essere usato anche senza una preventiva lista di campionamento. Viene usato ad es. negli exit-polls, i sondaggi effettuati all'uscita dal seggio elettorale. Si intervista un elettore ogni k tra quelli che escono dal seggio tra l'apertura e la chiusura del seggio. La stessa cosa si può fare per campionare i clienti di un supermercato, magari estraendo casualmente anche i giorni delle interviste (non andando solo il sabato ad esempio)

Esempio STM senza lista di campionamento

- Il campionamento sistematico consente di ottenere campioni casuali anche nella situazione in cui manchi la lista della popolazione e N sia ignoto ma stimabile in base a dati precedenti:
  - per es., sapendo che un supermercato ha circa 1800 clienti al giorno (giornata di 10 ore continuate: 3 ogni minuto), volendo ottenere un campione rappresentativo di 200 clienti (k=1800/200=9) se ne intervista uno ogni 9, o, in alternativa, uno ogni 3 minuti.
  - Deve essere possibilmente evitata ogni selezione diversa da quella predeterminata dal passo di campionamento (ad esempio quella in base alle carattersitiche delle persone).

Considerazioni finali

- I tipi di campionamento considerati finora sono puramente casuali
- eccetto la considerazione fatta per il campionamento sistematico in assenza di lista di campionamento non viene sfruttata nessuna informazione nota a priori sulla composizione della popolazione
- Il CCS (con e senza ripetizione) e il STM (con lista di campionamento) sono inoltre inapplicabili nel caso di indagini su vasta scala in quanto comportano un piano di rilevazione costoso e di difficile realizzazione dal punto di vista organizzativo, necessitando inoltre della lista completa della popolazione

Campionamento casuale stratificato (STR)

- Il campionamento si dice stratificato tutte le volte che la popolazione $\mathbf{P}$ di numerosità N può essere suddivisa in L strati, $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_L$, ciascuno di numerosità $N_1, N_2, \ldots, N_L$, dove $\sum_{i=1}^{L} N_i = N$
- quindi, mediante estrazione senza ripetizione, da ciascuno strato si estraggono casualmente i campioni $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_L$ di numerosità $n_1, n_2, \ldots, n_L$, dove $\sum_{i=1}^{L} n_i = n$
- Il campionamento da ciascuno strato $\mathbf{P}_i$ è equivalente ad un CSSR. Per cui **all'interno dello strato** sono costanti le probabilità di inclusione e di selezione
- il disegno STR è **ottimale** tutte le volte che, per ragioni "strutturali" la popolazione è divisa in strati tali che in "media" il fenomeno di interesse è molto diverso **tra gli strati** e la "varianza" in ciascuno strato è più piccola di quella della popolazione (omogeneità negli strati)

Campionamento casuale stratificato: quando si usa

La stratificazione si usa quando si vuole:

- evidenziare insiemi di unità significative per la ricerca;
- separare sottopopolazioni con caratteristiche speciali;
- utilizzare informazioni note, mantenendo tuttavia la casualità dell'estrazione;
- inoltre se le sottopopolazioni (gli strati) sono omogenei rispetto alla variabile da studiare si ottengono stime più "efficienti" di quelle che si possono avere con un CCS o un STM.

Lo STR è il disegno di campionamento piu usato nella pratica

Campionamento STR: esempio

Supponiamo di avere una popolazione di $N = 9$ soggetti distribuita in questo modo secondo età e reddito (in euro):

n. soggetto	1	2	3	4	5	6	7	8	9
età	30 anni	30 anni	30 anni	40 anni	40 anni	40 anni	50 anni	50 anni	50 anni
reddito	2000	2100	2200	3000	3100	3300	4000	4100	4200

Le medie e le deviazioni standard relative alla popolazione complessiva e ai tre gruppi di età sarebbero le seguenti:

	30 anni	40 anni	50 anni	Totale
Media	2100	3100	4100	3100
Dev.standard	81,65	81,65	81,65	820,57

Campionamento STR: esempio

- Come si vede, la variabilità nella popolazione è assai superiore a quella nei singoli strati di età
- Nella popolazione ci sono due tipi di variabilità: quella interna ai singoli strati di età e quella esterna tra gli strati
- In altri termini, i trentenni hanno stipendi diversi tra loro, così come i quarantenni e i cinquantenni (variabilità interna agli strati)
- D'altra parte i trentenni hanno stipendi molto differenti rispetto ai quarantenni e ai cinquantenni (variabilità esterna, tra gli strati)
- Se si guarda alla distribuzione dei valori nella popolazione, si vede anche a colpo d'occhio che la variabilità interna è assai inferiore a quella esterna: le differenze tra i trentenni sono molto inferiori alle differenze tra questi e i quarantenni/cinquantenni.

Campionamento STR: esempio

Cosa accade quando campioniamo da questa popolazione ($n = 3$).
Possiamo procedere in due modi:

1. **CCSSR o STM**: Selezioniamo tre casi dai nove complessivi con campionamento casuale semplice o sistematico.
2. **STR**: Selezioniamo un caso su tre entro ciascuno strato di età (un trentenne tra i trentenni, un quarantenne tra i quarantenni, ecc.) con un separato campionamento casuale semplice (o sistematico, ma qui non utilizzabile dato che estraiamo un solo caso)

Si dimostra che in casi come questo l'errore campionario è maggiore se usiamo un disegno del tipo CCS o STM rispetto al campionamento STR (vedremo in seguito)

Campionamento STR: come funziona

- Si tratta di sfruttare le informazioni disponibili sulla popolazione. Se disponiamo nella lista di campionamento delle informazioni circa una variabile correlata a quelle oggetto di studio (es.: l'età correlata al reddito) possiamo suddividere la popolazione in strati secondo i valori di questa variabile.
- In altri termini, dividiamo la lista di campionamento in liste separate per ciascuno strato.
- Effettueremo campioni casuali semplici (senza ripetizione) o sistematici separati per ciascuna di queste liste

Campionamento STR: come funziona (2)

- Esempio: studenti di Scienze politiche e dell'amministrazione. Nella lista di campionamento fornita dalla segreteria didattica, ad es., compare implicitamente o esplicitamente il sesso dello studente.
- Supponiamo di avere 2000 femmine e 1000 maschi
- Separiamo le due liste poi stabiliamo quante femmine dobbiamo estrarre dalla lista delle femmine e quanti maschi da quella dei maschi
- Il campione complessivo deve essere formato da 300 studenti, cioè da 1/10 della popolazione (frazione di campionamento).
- Estraiamo pertanto 1/10 delle femmine (200) e 1/10 dei maschi (100). In questo modo il nostro campione è stratificato e proporzionale (allocazione proporzionale negli strati): nel senso che femmine e maschi vi compaiono nelle identiche **proporzioni o frequenze relative in cui compaiono nella popolazione** (femmine 200:300 nel campione, 2000:3000 nella popolazione).

Campionamento STR: allocazione non proporzionale (uniforme)

- Se in una popolazione uno strato è di dimensioni ridotte seguire il criterio della proporzionalità potrebbe portare alla formazione di uno strato campionario di ampiezza troppo ridotta per stime affidabili al suo interno
- **NB**: ricordatevi che più è bassa la numerosità su cui si calcolano le statistiche (media, percentuali, ecc.) più aumenta l'incertezza della stima!
  - **Esempio**: se in una popolazione di 10000 soggetti stratificata per confessione religiosa, i musulmani sono 300 (3%) e il campione da estrarre è $n = 500$, in esso i musulmani dovrebbero essere 15 ($500 \times 3/100$)
  - La distribuzione di qualunque variabile entro lo strato dei musulmani sarebbe calcolata su un totale di 15 e ci potrebbero essere problemi come ad es. sottorappresentazione delle donne

Campionamento STR: allocazione non proporzionale (uniforme) (2)

- In questi casi come questi si ricorre ad un campionamento stratificato non proporzionale con numero uguale di unità per ogni strato: ad es., se le confessioni religiose sono 4 si selezioneranno 125 casi da ciascuno strato cioè...
  - $n_i = n/L$ per ogni strato i-mo (dove: $L =$ numero di strati)
- ovviamente in questo caso si risolve il problema della numerosità negli strati, ma rende il campione non rappresentativo (non è più una "copia in miniatura della popolazione")
- Nell'esempio i musulmani sarebbero sovrarappresentati, mentre gli altri strati sarebbero di conseguenza sottorappresentati.
- Per ovviare al problema della rappresentatività, si procede ad aggiustamenti post-rilevazione in particolare all'uso di **pesi** per ristabilire l'equilibrio tra gli strati

Campionamento STR: allocazione non proporzionale (esempio sovrastima)

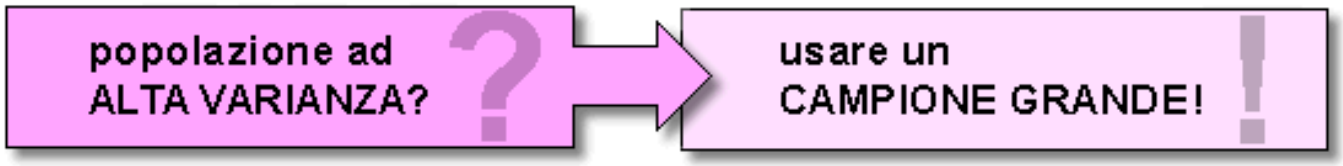
- Un problema del campionamento stratificato non proporzionale è ad esempio la sovrastima delle statistiche tratte dagli strati che nella popolazione hanno il minor peso
- ad esempio, se lo scopo d'indagine è quello di stimare il reddito medio della popolazione
- immaginiamo di conoscere la distribuzione degli strati formati dalle differenti professioni: operai, impiegati, lavoratori autonomi e liberi professionisti sono rispettivamente il 35, 45, 15 e 5% della popolazione,
- decidiamo di estrarre un campione di n=1000
- se riteniamo che i liberi professionisti vengano sottorappresentati con allocazione proporzionale allora decidiamo di selezionare in maniera non proporzionale e prendiamo 250 casi per ciascuno strato

Campionamento STR: allocazione non proporzionale (esempio sovrastima) (2)

- in tal modo il reddito dei liberi professionisti concorrerebbe al calcolo del reddito medio in misura 5 volte maggiore di quanto avviene nella popolazione
- La soluzione di questo problema consiste (dopo l'indagine) nel moltiplicare le stime ottenute in ogni strato per i rispettivi "coefficienti di espansione"
- $k_1 = f_1 n/n_1, \ldots, k_L = f_L n/n_L$ (dove f_i è la frequenza relativa dello strato nella popolazione)
- Nell'esempio, i coefficienti di espansione sarebbero: $k_1 = 350/250 = 1.4$ per gli operai, $k_2 = 450/250 = 1.8$ per gli impiegati, $k_3 = 150/250 = 0.6$ per i lavoratori autonomi, $k_4 = 50/250 = 0.2$ per i liberi professionisti.

Campionamento STR: allocazione ottimale

- L'allocazione non proporzionale descritta in precedenza non è l'unica
- Dall'inferenza statistica sappiamo che l'ampiezza ottimale di un campione è **direttamente proporzionale** alla variabilità della popolazione (ricordate la formula dell'intervallo di confidenza per la media?)
  - $n = (\frac{z_{\alpha/2}\sigma}{\delta})^2$



perché?

- al contrario tanto più omogenea la popolazione (bassa varianza) tanto più piccolo potrà essere il campione

Campionamento STR: allocazione ottimale (2)

Nel caso del campionamento stratificato... Se

1. possiamo dividere la popolazione in strati,
2. di tali strati conosciamo la varianza o la deviazione standard (cioè la variabilità) e
3. gli strati hanno variabilità diverse, possiamo effettuare un campionamento in cui selezioniamo da ogni strato un numero di casi proporzionale alla variabilità dello strato: più casi dagli strati più eterogenei

Questo tipo di campionamento è noto come campionamento stratificato ottimale (o con allocazione ottimale).
Questo tipo di campionamento è ancora più efficiente di quello stratificato proporzionale (meno casi per ottenere la stessa precisione) e ancor di più del precedente non proporzionale.

Campionamento STR: allocazione ottimale / Esempio

Prendiamo di nuovo il caso delle confessioni religiose...

- Dobbiamo estrarre un campione di 500 unità
- Innanzitutto si calcola la proporzione (p) di casi che appartengono a ciascuno strato nella popolazione
- I musulmani sono 300 su 10000 (N), quindi la proporzione è di 0,03 (3%)
- Supponiamo che la deviazione standard (sd) della variabile oggetto di studio sia pari a 83,3 nello strato dei musulmani
- Moltiplichiamo la proporzione delle unità statistiche appartenenti allo strato per la deviazione standard di questo stesso strato ($sd \times p = 83,3 \times 0,03 = 2,49$), allo stesso modo procediamo per i restanti strati
- Alla fine sommiamo tra loro tutti i prodotto ottenuti. Supponiamo di avere ottenuto come totale 15

Campionamento STR: allocazione ottimale / Esempio (2)

- Calcoliamo la proporzione di ciascun prodotto rispetto a quest'ultimo totale: per i musulmani 2,49/15=0,166.
- Applichiamo questa proporzione all'ampiezza dei campione e per ogni strato otteniamo il numero delle unità da selezionare: per i musulmani avremo $0,166 \times 500 = 83$
- cioè dovremo selezionare 83 unità da questo strato

NB: Ovviamente anche questo campionamento richiederà degli aggiustamenti post-rilevazione per correggere il sovra- o sotto-dimensionamento dei singoli strati

Campionamento a grappoli (GRA)

- Il campionamento a grappoli assomiglia vagamente al campionamento stratificato ma possiede proprietà statistiche molto differenti
- Prima tra tutte à la sostanziale **perdita di precisione** anche rispetto al CCS (quello stratificato è invece più efficiente del CCS)
- Si usa questo metodo quando vi sono vantaggi compensativi nel costo delle operazioni che sovrastano la perdita accennata
- L'**unità statistica da campionare** è ora un gruppo o grappolo (o cluster) di unità della popolazione
  - un grappolo è un **raggruppamento naturale** della popolazione, essenzialmente legato alla contiguità spaziale o istituzionale
  - In sostanza per contenere i costi si sfrutta l'esistenza di tali raggruppamenti naturali della popolazione

Campionamento a grappoli (GRA) (2)

- Nel campionamento a grappolo quindi si seleziona un campione casuale di G grappoli e **tutte le unità elementari** ad esso appartenenti sono oggetto di rilevazione

La differenza con il campionamento stratificato e quello a grappoli è che nello stratificato **si prendono tutti gli strati** mentre qui si selezionano casualmente **solo alcuni grappoli**

Campionamento a (2 o più) stadi (STA)

- Il Campionamento a due o più stadi è un disegno di campionamento complesso dove l'estrazione di una unità avviene **mediante scelte successive**
- la popolazione viene suddivisa in grappoli o strati (es. divisa per comune di residenza), solo alcuni dei quali vengono estratti a caso
- I grappoli (o gli strati) sono detti unità primarie o complesse. Le unità elementari al loro interno sono dette unità secondarie.

Campionamento a due stadi

1° stadio si scelgono casualmente un certo numero di grappoli (o strati)

2° stadio dentro ogni grappolo si scelgono casualmente un certo numero di unità elementari (secondo un ulteriore disegno di campionamento)

Campionamento a (2 o più) stadi (STA) (2)

Si ricorre a questo tipo di campionamento per una o entrambe le seguenti ragioni:

1. quando manca la lista della popolazione
2. quando la popolazione è distribuita su un territorio ampio e quindi l'indagine comporterebbe consistenti costi (es. quelli di trasferimento per gli intervistatori)

Campionamento STA: esempio

Immaginiamo di voler condurre un sondaggio tra gli elettori ricorrendo ad un campione di 2000 unità.
Ovviamente nella pratica non sarebbe una buona idea effettuare un CCS dalla lista degli italiani iscritti nelle liste elettorali, che comunque non esiste (troppo costoso raggiungerli ad esempio ma anche poco efficiente).
Possiamo però procedere in questo modo seguendo un campionamento a più stadi:

- estrarre casualmente 25 province dalle 110 totali.
- Costruire la lista dei comuni di ciascuna delle venti province estratte.
- Da ciascuna di queste liste estrarre casualmente 5 comuni, in ciascun comune estrarre 4 seggi elettorali (i seggi come sapete sono numerati).

Campionamento STA: esempio (2)

- Dalle liste elettorali di ciascuno dei 4 seggi, disponibili nell'ufficio elettorale di ciascun comune, estrarre 4 elettori. L'ampiezza del campione sarà quindi pari a $n = 25 \times 5 \times 4 \times 4 = 2000$ come desiderato
- Si possono usare dunque 25 intervistatori, ognuno dei quali copre una provincia (cioè i 5 comuni estratti) si incarica di reperire le liste dei seggi e degli elettori dei seggi estratti e infine effettua 80 interviste.
- Con un CCS avremmo avuto bisogno di più intervistatori o di far spostare molto gli intervistatori: non si sarebbe certo potuto assegnare a ciascuno 80 intervistati abitanti in soli 5 comuni di una stessa provincia!

A quanti stadi è il campionamento appena descritto?

Campionamento STA: precisazioni

Perché è a stadi?

Non si tratta di un unico campionamento **ma di più campionamenti a cascata**

- C'è un primo stadio in cui le unità da selezionare sono le province: dalla lista di queste si estraggono 25 unità con campionamento casuale semplice, sistematico o stratificato
- C'è un secondo stadio, quello dei comuni, anche qui può essere usata una delle tecniche di campionamento che abbiamo visto in precedenza.
- C'è un terzo stadio, quello dei seggi elettorali.
- C'è infine un quarto e ultimo stadio in cui si selezionano i soggetti da intervistare (le unità elementari), sempre con una delle tecniche precedenti.

Campionamento STA: precisazioni (2)

- Si noti che **ad ogni stadio** le singole unità complesse da estrarre (province, comuni, ecc.) in termini della variabile oggetto d'indagine dovrebbero essere *simili tra loro* e mantenere il **massimo di eterogeneità al loro interno**
- Nel linguaggio usato in precedenza, la variabilità esterna dovrebbe essere pressoché nulla, dovrebbe essere elevata la variabilità interna (**diversamente da quanto richiesto nel campionamento stratificato!**).
- Se così non fosse, escludendo una provincia in cui gli elettori sono molto diversi da quelli delle altre, escluderemmo in via definitiva dal campione questo tipo di elettori. Lo stesso rischio possiamo correrlo ad ogni stadio successivo.

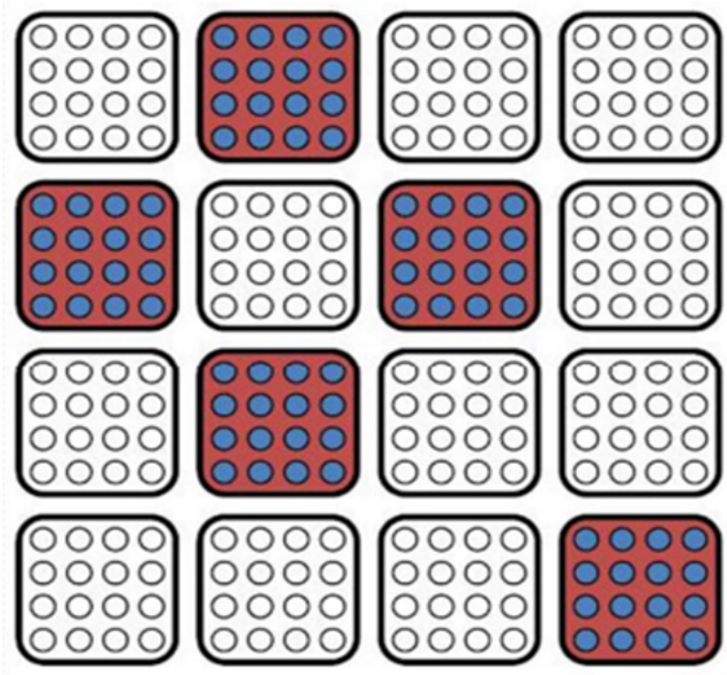
Campionamento STA: precisazioni (3)

- Per questo motivo il campionamento a più stadi è a volte meno efficiente degli altri. Spesso inoltre questo tipo di campione richiede aggiustamenti post-rilevazione (ponderazione) assai complessi

Campionamento a Grappoli vs a (due) Stadi

Campionamento a grappoli

G = 5



Campionamento a 2 stadi

I stadio: G = 5; II stadio: $n_i = 3$