Statistica sociale – 5

Prof. Antonio Mussino

a. a. 2022-2023

SAPIENZA
UNIVERSITÀ DI ROMA

Random sampling - 1

- Tutti questi ragionamenti sono possibili se, e solo se, il campione è stato individuato con una procedura rigorosamente probabilistica, ovvero si tratti di un **campione casuale** *(**random sampling**)*.

- Un campione è tale quando ***ciascuna unità della popolazione ha una chance nota (e non nulla) di essere scelto per l'inserimento nel campione***.

Random sampling - 2

- La procedura di campionamento probabilistico può essere simulata con l'estrazione di palline numerate da un urna (gioco della tombola):
- tutti i numeri da 1 a 90 hanno la **stessa chance** di essere estratti, se chi effettua l'estrazione è "**onesto**"!
- Ma nelle indagini nel campo sociale la popolazione è quasi sempre molto ampia, quindi la semplice estrazione da un'urna di una sequenza di palline risulta inapplicabile.

Random sampling - 3

- Una soluzione è quella di costruire una lista delle unità della popolazione, assegnando un numero a ciascuna unità e utilizzare una **tavola dei numeri aleatori** (**pseudo-aleatori**), ovvero una routine che generi numeri aleatori:
- questi numeri individueranno le unità da inserire nel campione.

Tavola numeri aleatori

91	54	2	11	17	58	4	92	47	89	73	55	1	58	64
57	19	78	52	3	26	9	88	35	89	22	80	3	29	88
25	89	21	71	30	55	5	74	96	14	43	56	99	0	71
64	76	53	27	59	0	82	33	4	33	48	95	80	81	71
38	70	36	91	95	50	14	20	85	9	89	26	96	64	36
94	82	58	28	13	44	30	48	40	57	42	19	11	80	75
87	33	65	27	10	38	63	59	58	7	48	73	59	76	87
78	93	85	26	6	72	42	39	99	16	85	46	51	94	64
15	79	32	60	7	72	89	84	89	14	84	41	27	15	11
92	47	52	61	68	99	6	62	40	28	46	76	84	74	89
67	93	24	50	38	61	23	92	41	56	21	90	46	57	73
10	49	53	66	0	81	92	9	62	59	80	2	21	66	42
12	16	20	60	98	85	34	67	43	9	67	64	66	38	3
62	10	34	51	90	43	33	85	33	27	8	11	56	72	74
94	73	76	92	78	23	62	29	56	94	92	96	95	88	51
54	53	69	53	93	91	84	52	51	41	56	20	3	30	85
67	10	71	12	99	63	40	3	46	19	92	54	74	24	14
90	7	13	58	79	94	50	24	10	25	46	92	9	65	4
71	20	44	94	57	9	35	98	94	28	82	93	28	89	58
30	42	71	11	60	88	76	88	4	24	95	95	55	19	75
59	29	59	0	81	23	30	40	29	54	57	93	48	64	6
91	24	53	17	17	3	84	30	24	18	92	73	46	62	36
63	97	26	3	59	21	27	1	51	77	49	18	65	99	41
37	39	82	29	81	27	97	11	12	36	34	34	21	99	61
81	68	27	41	21	80	67	66	45	29	13	43	0	88	61

Campionamento con reintroduzione

- Teoricamente vi è la possibilità di estrarre una pallina (un'unità), reinserirla nell'urna e procedere con una nuova estrazione: questa procedura viene definita campionamento ***con reintroduzione***.
- È una situazione ***teorica***, in quanto è una procedura alla base delle teorie statistiche sul campionamento, ma non è quasi mai usata nella realtà.
- Si pensi all'individuazione di un campione per un sondaggio pre-elettorale: l'estrazione di un elettore più volte vorrebbe dire considerare più volte il suo voto, che nella realtà è unico!

Campionamento in blocco - 1

- Se, invece, la pallina estratta viene messa da parte e non partecipa alle estrazioni successive, ovvero non vi è la reintroduzione, si parla di ***campionamento in blocco***.
- Oltre all'evidente maggiore concretezza di questa scelta va rilevato che la variabilità dei risultati è qui più bassa, in quanto gli eventi estremi possono capitare una sola volta.
- Stime -> come cambia l'intervallo di confidenza?

- $\sqrt{\{(N-n)/(N-1)\}}$

Campionamento in blocco - 2

- In questo caso è ininfluente l'ordine con il quale si sceglie l'unità (la pallina).
- Nel caso della tavola di numeri aleatori, poiché una sequenza di cifre che individuano il numero si può ripetere con l'algoritmo scelto, si dovrà consentire l'eliminazione di una sequenza già scelta.

- Nel caso di popolazioni molto grandi (infinite) le due strategie di fatto coincidono.

La stratificazione - 1

- Nel caso citato del campione di 900 unità, usato per l'indagine sulla partecipazione sportiva, potrebbe verificarsi un evento *rarissimo*, ma non impossibile in cui **tutti i prescelti siano maschi**

  (la partecipazione ad attività sportive è diversa per genere!).
- Per evitare questa possibilità si suddivide la popolazione per genere, come se avessimo due urne (o due liste), e poi si campiona separatamente.

La stratificazione - 2

- Questa procedura si chiama ***stratificazione***: oltre che per genere sono comuni la stratificazione territoriale, quella per fasce d'età e per altre caratteristiche importanti per la ricerca che si sta compiendo.
- In questo modo si hanno tanti campioni casuali quanti sono gli strati, che a loro volta sono tanti quante sono le possibili associazioni fra le modalità delle variabili rispetto alle quali si stratifica.

La stratificazione - 3

- Ad esempio, stratificando per genere (maschi e femmine), fasce d'età (adolescenti, giovani, adulti, anziani) e ripartizione territoriale (Nord Est, Nord Ovest, Centro, Sud e Isole), gli strati sono 32 e stiamo parlando di ***stratificazione multipla***.
- In realtà il campione complessivo non è più trattabile come un campione casuale semplice.
- Le strategia di scelta e tutte le implicazioni di queste scelte sulle stime sono materia dei corsi di **Teoria dei campioni**.

La stratificazione - 4

- La stratificazione permette di migliorare la precisione delle stime.
- Si è già visto come calcolare la dimensione ottimale (con un'analisi costi/benefici) del campione "n", partendo da una popolazione che abbia una dimensione pari a "N".
- Ora il problema è come allocare negli strati le unità campionarie?

La stratificazione - 5

- La soluzione più semplice sembra essere quella di utilizzare la stessa frazione campionaria (n/N) in ogni strato,
  - N_1, N_2, N_3, N_4 (con $N_1+N_2+N_3+N_4=N$),
  - n_1, n_2, n_3, n_4
  - con $n_1/N_1= n_2/N_2= n_3/N_3= n_4/N_4= n/N$.

- Questa strategia è la più comoda e la più efficace, ma le frazioni campionarie possono anche essere diverse fra gli strati.

La stratificazione - 6

- La variabilità complessiva può scomporsi in
  - variabilità dovuta alle caratteristiche per le quali si stratifica (ad esempio gli uomini praticano sport più delle donne, i giovani più degli anziani, i residenti nel Nord più dei meridionali) e
  - variabilità interna agli strati.

- Nel campionamento stratificato la prima non entra nel calcolo della stima, perché tenuta sotto controllo.

La stratificazione - 7

- Ovviamente non c'è nessun guadagno nello stratificare se il carattere rispetto al quale si stratifica non incide sulla variabile di cui dobbiamo effettuare la stima,
- ad esempio se si vuole stimare il livello di partecipazione sportiva degli studenti universitari (a parte il caso specifico di Scienze motorie!).

La stratificazione - 8

- Ovviamente in ogni strato ci sarà una stima per il valore cercato e quindi la stratificazione è utile anche per studiare i diversi comportamenti all'interno degli strati.

- La stima complessiva su tutta la popolazione, può essere calcolata facilmente con una media delle stime degli strati, ponderate con le frazioni campionarie.

La stratificazione - 9

- Vi sono casi in cui è preferibile che la frazione campionaria sia diversa da strato a strato:
- ad esempio quando il carattere studiato presenta una variabilità maggiore in alcuni strati e quindi si ha la necessità di una numerosità maggiore per stima più accurata;
- vi sono situazioni, generalmente legate a problemi territoriali, in cui è più costoso contattare le unità di alcuni strati, allora si potrebbe prendere una frazione più piccola negli strati più costosi;
- e così via.

La stratificazione - 10

- La migliore stratificazione per studiare una variabile potrebbe non esserla per altre;
- Nelle ricerche multiscopo la soluzione è quella di stratificare per le variabili cosiddette ***strutturali*** (genere, età, ripartizione geografica, dimensione demografica del luogo di residenza e così via).

La stratificazione - 11

- Un'ultima cautela: la numerosità del campione calcolato precedentemente consentiva una stima con un errore massimo di più o meno 1%.
- Se il campione fosse stratificato, ad esempio per genere e ripartizione di residenza, e la numerosità dello strato "maschi residenti al Sud e Isole" fosse di 784 unità, una stima per la percentuale di sportivi in questo strato sarebbe ben più ampia (+ o - 3%) di quella totale (+ o - 1%).
- ***Attenzione, quindi alle inferenze che si effettuano!***

Il *cluster sampling* - 1

- Quando andiamo a stratificare la popolazione, individuiamo insiemi di unità omogenee per uno (o più) carattere (i);
- in ogni strato poi si campionano le unità (solo una parte).
- Potremmo invece prendere **tutte!**
- Questa è la logica del ***campionamento a grappoli*** ed è efficace se le unità sono fortemente eterogenee all'interno degli strati, mentre questi sono omogenei fra di loro.

Il *cluster sampling* - 2

- Campionamento di studenti in una scuola media per un'indagine sulla pratica sportiva:
  - gli studenti sono suddivisi in strati amministrativi già precostituiti (le sezioni A,B,C e D e le classi I, II e III, con circa 25 studenti in ogni classe).

- Noi potremmo utilizzare questi strati per effettuare il campionamento, prendendo ad esempio 5 studenti in ogni classe in tutte le sezioni, per avere un campione di 60 studenti.

Il *cluster sampling* - 3

- Questa procedura è costosa in termini di tempo, perché dobbiamo andare in ogni classe e convincere il professore a lasciare liberi i 5 studenti.
- È molto più semplice scegliere casualmente 3 classi e, con un impegno temporale minore, avere 75 questionari!
- Questo risparmio non va a scapito della accuratezza della stima, perché è prevedibile che il comportamento sportivo non sia differente nelle varie classi e sezioni, ovvero non dipenda dalla appartenenza a questi strati amministrativi.

Il *cluster sampling* - 4

- Quindi il campionamento casuale avviene fra classi e non fra le unità campionarie finali.
- Stratificare le unità "classe" per l'anno?

- Un altro esempio potrebbe essere quello delle sezioni di censimento: queste potrebbero essere le unità di campionamento, mentre all'interno di ciascuna sezione tutti i cittadini sarebbero coinvolti.

Il *cluster sampling* - 5

- Variabilità del fenomeno studiato negli strati:

DT = DS + DI

- **DT** (variabilità totale)
- **DS** (variabilità *inter* strati)
- **DI** (variabilità *intra* strati

- se DI è più importante sceglieremo un campionamento a grappoli, altrimenti sarà preferibile un campionamento casuale all'interno di ogni strato.

Campionamento a più stadi – 0

- In effetti il disegno campionario per una buona indagine sociale è sempre più complesso del caso del campionamento casuale semplice:

- ***campionamento a più stadi***, in cui possono interagire stratificazione, cluster sampling e altro,

- mantenendo però il principio della scelta casuale in ognuno degli stadi!

Campionamento a più stadi - 1

- In questa strategia si considera la popolazione composta da ***unità primarie*** di campionamento, in ognuna delle quali è presente un insieme di ***unità secondarie*** e così via se il campione è a più stadi.
- Una scelta campionaria è effettuata a ogni stadio (se a grappolo è preso il 100% delle unità).
- La necessità di utilizzare più stadi è statistica, ma anche organizzativa -> Indagine multiscopo.

Campionamento a più stadi - 2

- Indagine sugli studenti delle scuole medie di un comune, possibili strategie:

- tramite le liste (in ordine alfabetico) fornite da tutte le scuole costruisco un'unica lista complessiva e, con l’ausilio di una tavola di numeri aleatori, estraggo 2000 nomi di studenti: si tratta di un campione casuale semplice (a uno stadio);

Campionamento a più stadi - 3

- costruisco una lista delle scuole e estraggo un campione di 4 scuole; in ognuna delle 4 scuole, basandomi sulla lista degli studenti, estraggo 500 nomi: si tratta di un campione casuale a due stadi, con 2000 interviste;

- suddivido la città in 4 circoscrizioni amministrative e estraggo un campione di 2 scuole in ogni circoscrizione; in ogni scuola estraggo 250 nomi dalla lista unica: si tratta di un campione casuale a due stadi, stratificato al primo, con 2000 interviste;

Campionamento a più stadi - 4

- suddivido la città in 4 circoscrizioni amministrative e estraggo un campione di 2 scuole in ogni circoscrizione; in ogni scuola considero una stratificazione per sezioni(8) e classi(3); estraggo 10 o 11 nomi in ogni classe: si tratta di un campione casuale a due stadi, stratificato in entrambi gli stadi, con circa 2000 interviste;

Campionamento a più stadi - 5

- suddivido la città in 4 circoscrizioni amministrative e estraggo un campione di 2 scuole in ogni circoscrizione; in ogni scuola considero una stratificazione per sezioni(8) e classi(3); estraggo 9 classi, 3 prime, 3 seconde e 3 terze e intervisto tutti gli alunni delle classi scelte: si tratta di un campione casuale a due stadi, stratificato nel primo stadio e a grappolo nel secondo, con un numero di interviste che dipende dalla numerosità delle classi, ma comunque di circa 2000.

Campionamento a più stadi - 6

- Se è utile (o obbligato) avere lo stesso numero di unità in ogni strato, anche se questi hanno ampiezza diversa
- i risultati andranno ***riponderati*** per ricostruire la composizione complessiva della popolazione.

Campionamento per aree - 1

- Nel ***campionamento per aree*** il territorio che deve essere coperto dall'indagine è suddiviso in un numero di piccole aree (anche dette ***blocchi***), che corrispondono alle unità campionarie del primo stadio, ovvero è effettuato un campione probabilistico di esse.
- All'interno di ognuna di esse si effettua una rilevazione totale, oppure si effettua un ulteriore campionamento.

Campionamento per aree - 2

- È un campionamento a più stadi nel quale le ***mappe***, piuttosto che liste o registri, rappresentano lo schema di campionamento.
- Le sezioni di censimento, o quelle elettorali, possono essere le aree alle quali si fa riferimento -> stratificazione geografica;
- in genere si tratta di un campionamento a più stadi e nell'ultimo blocco tutte le unità dovrebbero essere coinvolte.

Campionamento multifase - 1

- Nel ***campionamento multifase*** si parte dalle unità scelte per essere inserite in un campione e se ne campiona una parte: in genere le fasi sono solo due e si parla di ***doppio campionamento***.
- Al campione base si sottopone un questionario semplificato, sul campione selezionato si approfondisce l'analisi con un secondo questionario più dettagliato o con una scheda di rilevazione, un diario e così via.

Campionamento multifase - 2

- L'obiettivo è quello di ridurre i costi e la complessità della rilevazione.
- Le informazioni raccolte sul campione ampio sono utilizzate per stimare al meglio i risultati del gruppo più ristretto.
- Il risultato si migliora con ponderazioni a posteriori e con tecniche di stima basate su modelli di regressione.
- Es. Uso del tempo, Censimento, Multiscopo…..

Bias e vincoli esterni - 1

- Errori sistematici nel campionamento:

- mancanza di casualità nella scelta del campione: se questo non è probabilistico, nelle sue varie fasi, le procedure di stima dell'inferenza statistica non possono essere applicate;

- assenza di una lista completa che copra tutte le unità della popolazione, o almeno le unità amministrative rispetto alle quali si stratifica;

Bias e vincoli esterni - 2

- impossibilità di prevedere prima del campionamento che una o più parti del campione siano irreperibili o rifiutino di collaborare.

- C'è poi un elemento esterno alla discussione metodologica fin qui effettuata: il **costo** delle interviste.

- Questo elemento gioca un ruolo cruciale, costringendo spesso il ricercatore a rinunciare a una maggiore accuratezza delle stime, o forzando la scelta della strategia campionaria nelle sue fasi.

I cittadini e il tempo libero. **Strategia di campionamento - 1**

- L´indagine, svolta sui medesimi temi nel 1995*, 2000, 2006 e 2015, fa parte del filone delle Indagini Multiscopo.
- Popolazione di interesse l'insieme delle famiglie residenti in Italia (unità campionarie), con tutti i membri che le compongono, escludendo pertanto le convivenze (Istat, 2015).

- * Nel 1995 *Tempo libero e cultura*

I cittadini e il tempo libero. Strategia di campionamento - 2

- I parametri della popolazione da stimare vanno riferiti ai seguenti *domini di studio*:
  - intero territorio nazionale;
  - cinque ripartizioni geografiche (Nord ovest, Nord est, Centro, Sud, Isole)
  - diciannove regioni più le province autonome di Trento e Bolzano;
  - sei tipologie comunali (A1 comuni centro di un´area metropolitana; A2 altri comuni della area metropolitana; B1 comuni fino a 2.000 abitanti; B2 comuni da 2001 a 10000 abitanti; B3 comuni da 10001 a 50000 abitanti; B4 comuni con oltre 50000 abitanti).

I cittadini e il tempo libero. Strategia di campionamento - 3

- Il disegno di campionamento è a due stadi, con doppio livello di stratificazione dei comuni: in base al territorio e alla dimensione demografica.
- La lista delle unità campionarie per il primo stadio è costituita dall'archivio dei ***comuni*** italiani; quella per il secondo stadio dagli archivi anagrafici dei comuni selezionati al primo stadio, ovvero dai fogli di ***famiglia***.

I cittadini e il tempo libero. Strategia di campionamento - 4

- Una volta costruiti gli strati, all'interno di ciascuno di essi si adottano due diversi schemi di campionamento:
- nel caso si sia in presenza di comuni di dimensione demografica "maggiore", relativamente al proprio strato, questi sono classificati come ***auto rappresentativi*** (**AR**);
- nel caso contrario, ovvero per tutti gli altri comuni, questi sono classificati come ***non auto rappresentativi*** (**NAR**).

I cittadini e il tempo libero. **Strategia di campionamento - 5**

- I comuni **AR** costituiscono uno strato a se stante e vengono presi tutti (campionamento a grappolo): le famiglie sono estratte in modo **sistematico** dalle anagrafi dei comuni stessi.
- Nello ambito dei comuni NAR, invece, il disegno individua sempre due stadi: nel primo vengono estratti i comuni allo interno di ciascuno strato e nel secondo nei comuni scelti vengono estratte in modo **sistematico** le famiglie.

Il *campionamento sistematico* - 1

- Il ***campionamento sistematico*** non è teoricamente considerato un campionamento casuale.
- Se si possiede una lista delle unità della popolazione "N", si possono individuare tecnicamente "n" strati, ognuno composto da k unità (k=N/n).
- Si sceglie casualmente un numero compreso fra 1 e k (estrazione, tavola dei numeri aleatori): ad esempio "j".
- Le unità da inserire nel campione saranno allora j, j+k, j+2k, j+3k e così via.

Il *campionamento sistematico* - 2

- È evidente che, se la lista non contiene alcuna sistematicità, siamo in presenza di un campione casuale; ben diversi sono i casi in cui ci sia una gerarchia, o una ciclicità nella lista stessa.
- L'ordine alfabetico non può essere considerato tra questi casi, pertanto la scelta della Istat di questa procedura è corretta.

I cittadini e il tempo libero. Strategia di campionamento - 6

- Il calcolo della dimensione campionaria è legato alla necessità di accurati livelli di stima anche a livello disaggregato:
- ma l'indagine è, come dice il nome, con obiettivi multipli, quindi non è realistico pensare a una soluzione univoca.
- <<determinare la numerosità nazionale e poi ripartirla fra le regioni in modo proporzionale alla loro dimensione demografica>> (Istat, op.cit.).

I cittadini e il tempo libero. Strategia di campionamento - 7

- Poiché, invece, vi è bisogno di stime a livello regionale, una soluzione potrebbe essere quella di <<selezionare un campione uguale in tutte le regioni>>, che però sarebbe <<poco efficiente per le stime a livello nazionale>>.

- In aggiunta vi è, come in tutte le indagini sul sociale, un problema di *budget* per l'indagine, in termini di costo e organizzativi.

I cittadini e il tempo libero. Strategia di campionamento - 8

- In sintesi la dimensione totale in termini di famiglie a livello nazionale è stata di circa **24.000** unità, con una rilevazione da effettuarsi in circa **900** comuni; il numero minimo di famiglie intervistate in ciascun comune è pari a **23**.

  <<L'allocazione del campione di famiglie e di comuni fra le varie regioni è stata quindi calcolata adottando un criterio di compromesso tale da garantire sia l'affidabilità delle stime a livello nazionale che quella delle stime a livello di ciascuno dei domini territoriali ...>>.