Lezione 9
Inferenza statistica e campionamento

PROF. ROBERTO COSTA

SCIENZE DELL'EDUCAZIONE - STATISTICA SOCIALE (305SF)

Dove eravamo rimasti

Nella scorsa lezione abbiamo concluso la parte di statistica descrittiva, completando il tema delle relazioni tra due variabili.

Abbiamo visto come possiamo misurare la relazione tra due variabili qualitative, o in presenza dei soli macrodati (tabella a doppia entrata).

Abbiamo visto tre diversi indici: l'indice di associazione χ^2 di Pearson, l'indice Φ^2 sempre di Pearson e infine l'indice V di Cramer.

Abbiamo poi visto come possiamo calcolare la retta di regressione stimando i due parametri

α e β con il metodo dei minimi quadrati.

Ci sono dei dubbi?

Inferenza statistica

Fino alla scorsa lezione abbiamo visto come rispondere ad alcuni interrogativi di ricerca tipici delle scienze sociali.

Abbiamo imparato a utilizzare diversi strumenti di analisi per sintetizzare i dati disponibili: dalle distribuzioni di frequenza alle misure di centralità e di disuguaglianza.

Abbiamo poi imparato a utilizzare gli strumenti dell'analisi bivariata.

Abbiamo anche imparato che i risultati ottenuti si caratterizzano per un margine di incertezza, che può avere diverse origini: dagli errori materiali nella compilazione del questionario, all'utilizzo di variabili non perfettamente corrispondenti col fenomeno che vogliamo analizzare, ecc.

Stime campionarie ed errori

Se, invece di studiare un fenomeno sull'intera popolazione di riferimento, ci limitiamo ad osservarne solo una porzione di essa, introduciamo una tipologia di errore detta **errore di campionamento**.

In estrema sintesi, osservando solo una parte della popolazione e non la sua totalità non potremo mai essere certi che i risultati che otteniamo corrispondano perfettamente alla realtà.

In statistica si parla di **inferenza**. In particolare si dice inferire (che significa arguire, desumere, derivare) dai dati di un campione delle informazioni sulla popolazione da cui è stato estratto.

La statistica inferenziale

Definiamo **statistica inferenziale** l'insieme delle teorie e delle tecniche che ci consentono di estendere i risultati di un'indagine di tipo campionario all'intera popolazione di riferimento.

Siamo consapevoli che questo comporta un certo grado di incertezza.

Quantificare il grado di incertezza, significa associare ai risultati dell'indagine campionaria un livello di probabilità, ovvero calcolare la probabilità che quei risultati siano validi per la popolazione.

La statistica inferenziale

L’incertezza deriva da:

- Da ogni popolazione di riferimento sufficientemente grande possiamo estrarre un numero pressoché infinito di campioni di una certa ampiezza.
- Ognuno dei possibili campioni rappresenta la popolazione di riferimento in modo imperfetto. Nella maggior parte dei casi l’imperfezione è piuttosto contenuta in altri potrebbe restituire un risultato molto difforme dalla popolazione di riferimento.
- In ogni studio viene estratto e osservato uno solo dei possibili campioni.
- Le caratteristiche della popolazione di riferimento spesso non sono note e pertanto non siamo in grado di stabilire in che misura il campione estratto è rappresentativo della popolazione di riferimento.

La popolazione di riferimento

Facciamo un esempio: vogliamo valutare il livello di soddisfazione sulle strutture messe a disposizione dalle università (aule, luoghi di studio, laboratori, luoghi di ristoro, ecc.) degli studenti iscritti agli atenei del Friuli Venezia Giulia nell'anno accademico 2023/2024.

Abbiamo definito correttamente la popolazione di riferimento?

Studenti che si sono iscritti agli atenei presenti in Friuli Venezia Giulia nell'anno accademico 2023/2024.

La popolazione di riferimento

Quanti sono gli studenti che si sono iscritti agli atenei presenti in Friuli Venezia Giulia nell'anno accademico 2023/2024?

Anche senza conoscere il dato esatto, possiamo fare qualche prima ricerca in rete e scoprire che potrebbero essere poco meno di 30.000.

In un mondo ideale dove tempi e costi sono irrilevanti, potremmo immaginare di intervistarli tutti.

I dati ottenuti (al netto di alcuni errori di tipo non campionario) sarebbero di fatto riferibili con certezza all'intera popolazione di riferimento.

https://ustat.mur.gov.it/dati/didattica/friuli-venezia-giulia/atenei#tabstudenti

Studenti per tipologia di Corso di Laurea a.a. 2021/22

Corsi di Laurea	Iscritti
Laurea	18.810
Laurea Magistrale	5.606
LM a Cliclo Unico	5.202
Vecchio Ordinamento	284
Totale	29.902

La popolazione di riferimento

Nella pratica se volessimo intervistare tutti dovremmo sostenere ingenti spese e i tempi per realizzare tutte le interviste e per le successive fasi di controllo ed elaborazione dei dati sarebbero molto lunghi.

Nella ricerca sociale spesso ci dobbiamo confrontare con popolazioni piuttosto ampie e, al contempo, con risorse economiche e tempi per produrre i risultati contenuti, rendendo così irrealizzabile l'ipotesi di indagare su tutti i componenti della popolazione di riferimento.

Il ricercatore di conseguenza prenderà in considerazione solo un sottoinsieme della popolazione definito **campione**.

Il complesso delle procedure adottate per estrarre il campione dalla popolazione di riferimento è detto **campionamento**.

Stime e valore «vero»

Dalla nostra popolazione di riferimento di 30.000 studenti degli atenei del Friuli Venezia Giulia, decidiamo di estrarre casualmente 800 individui.
Quanti sono i diversi campioni di 800 individui che potrei estrarre da una popolazione di 30.000 unità?

Se chiamiamo:

n la numerosità campionaria e

N la numerosità della popolazione

Possiamo estrarre un numero di campioni diversi pari a:

$$\frac{N!}{(n!((N-n)!)}$$

Stime e valore «vero»

$$\frac{N!}{(n!((N-n)!)}$$

Cosa significa?

Il simbolo N! Significa Fattoriale di N, ovvero N * (N-1) * (N-2)* … *2 * 1

Nel nostro esempio equivale a:

$$\frac{30000!}{(800!((30000-800)!)} = \frac{30000!}{800!*29200!}$$

Stime e valore «vero»

Vogliamo fare un esempio più "facile"?

Supponiamo di voler estrarre un campione di 5 persone tra i 20 studenti presenti ad una lezione di statistica ufficiale. Quante sono le possibili combinazioni?

$$\frac{N!}{(n!((N-n)!)}$$

Nel nostro esempio equivale a:

$$\frac{20!}{(5!((20-5)!)} = \frac{20!}{5!*15!} = 2.432.902.008.176.640.000/120*1.307.674.368.000$$

Stime e valore «vero»

Questo vi fa capire come a fronte di un solo valore «vero» possiamo ottenere, a seconda del campione estratto, delle stime che differiscono tra di loro anche in modo importante.

Nella maggior parte dei casi la stima che otteniamo è prossima al valore «vero», in altre si discosterà anche in modo rilevante.

La variabilità delle stime campionarie non è mai erratica, ma tende ad assumere una forma precisa.

Variabilità delle stime

Il grado di precisione delle stime campionarie dipende da diversi fattori:

- ampiezza del campione selezionato

- la variabilità del fenomeno osservato

Data la stima campionaria di un parametro di interesse, vogliamo definire l'intervallo di valori entro il quale molto probabilmente si colloca il valore vero della popolazione di riferimento.

Questo intervallo viene chiamato **intervallo di confidenza** e la misura che lo definisce si chiama **errore standard della stima**.

Variabilità delle stime

Tornando al nostro esempio:

Alla fine di un'indagine su tutta la popolazione otteniamo un livello medio di soddisfazione dei servizi forniti dagli atenei della regione pari a 8,5.

A questo punto potremo dire che: «il livello di soddisfazione medio è pari a 8,5»

Invece, se al termine dell'indagine campionaria stimiamo che il livello medio di soddisfazione dei servizi forniti dagli atenei della regione sia pari a 8,3 dovremo dire qualcosa del tipo: «c'è il 95% di probabilità che l'intervallo 8,3 ± x contenga il livello di soddisfazione medio»

L’intervallo di confidenza

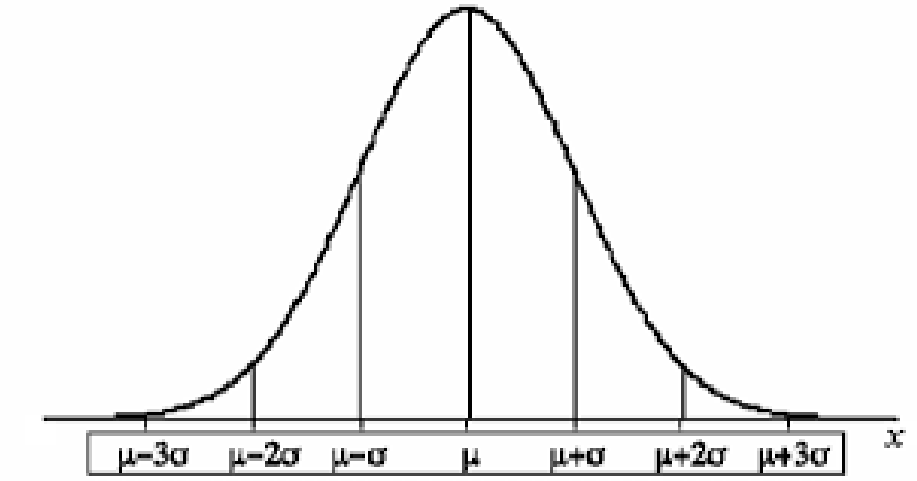
La stima è quindi caratterizzata da un certo livello di fiducia e consiste in un intervallo.

Come calcolo questo intervallo?

Supponiamo di voler stimare un parametro β

Indichiamo la stima ottenuta con **$\hat{\beta}$** e con **$\sigma(\hat{\beta})$** l’errore standard, ovvero la stima della deviazione standard, **z** è un coefficiente che dipende dal livello di fiducia che vogliamo avere per la nostra stima.

$\hat{\beta} \pm z\sigma(\hat{\beta})$



L’intervallo di confidenza

Il problema non è solo calcolare la stima del parametro tramite un campione, ma anche calcolare l’errore di campionamento.

Per calcolarlo dovremmo conoscere alcune informazioni sulla popolazione, che però non abbiamo.

Se il campione è stato estratto in modo rigorosamente casuale (campione probabilistico) la statistica ci permette di calcolare tale errore.

Abbiamo visto che volendo calcolare un parametro β, parto da una sua stima a cui assommo l’errore di campionamento e: $\beta = \hat{\beta} \pm e$

Abbiamo visto prima che $e = z\sigma(\hat{\beta})$

L’errore di campionamento

Supponiamo di voler stimare la media di un fenomeno. L’errore sarà e = zσ($\hat{m}$)

Dove σ($\hat{m}$) è l’errore standard della media campionaria. Come lo calcolo?

e = zσ($\hat{m}$) = z$\frac{s}{\sqrt{n}}\sqrt{1-f}$

Dove:

z è il coefficiente dipendente dal livello di fiducia della stima, nel caso del 95% = 1,96

s è la deviazione standard campionaria della variabile studiata

n è l’ampiezza del campione

1 – f è un fattore di correzione per popolazioni finite, dove f è la frazione di campionamento n/N (numerosità campionaria/ampiezza della popolazione).

L’errore di campionamento

Abbiamo detto che 1 – f è un fattore di correzione per popolazioni finite, dove f è la frazione di campionamento n/N (numerosità campionaria/ampiezza della popolazione).

Se la popolazione è infinita, o comunque quando il campione è inferiore al 5% della popolazione, il fattore di correzione si approssima a 1 e si può trascurare.

L’errore di campionamento

Riassumendo:

L’errore è tanto più grande:

- Quanto è più elevato il livello di fiducia che vogliamo avere. Se 95% z = 1,96, se 99% z = 2,58, ecc. e = $\mathbf{z}\frac{s}{\sqrt{n}}\sqrt{1-f}$
- Quanto più è elevata la variabilità del fenomeno oggetto di studio -> e = $z\frac{\boldsymbol{s}}{\sqrt{n}}\sqrt{1-f}$
- Quanto più piccola è la numerosità campionaria -> e = $z\frac{s}{\sqrt{\boldsymbol{n}}}\sqrt{1-f}$

L’errore di campionamento

Nel caso di variabili qualitative, la misura di sintesi più comune è una proporzione. Ad esempio quanto hanno votato un determinato partito, quanti hanno fatto ricorso ad un determinato servizio sociale, ecc.).

In questo caso l’errore campionario si calcola tramite la seguente formula:

$$e = z\sqrt{\frac{pq}{n-1}}\sqrt{1-f}$$

z, n e f hanno lo stesso significato di prima

p è la proporzione nel campione per la categoria che abbiamo rilevato

q = 1-p

Ampiezza del campione

Definire l'ampiezza del campione è uno dei primi passi che il ricercatore deve fare.

La scelta solitamente dipende dall'errore tollerabile e dai tempi e risorse disponibili.

Ritorniamo alla formula per una proporzione:

$$e = z\sqrt{\frac{pq}{n-1}}\sqrt{1-f}$$

Partiamo dal presupposto che la popolazione sia sufficientemente grande e non consideriamo il fattore di correzione, inoltre consideriamo n $\cong n-1$.

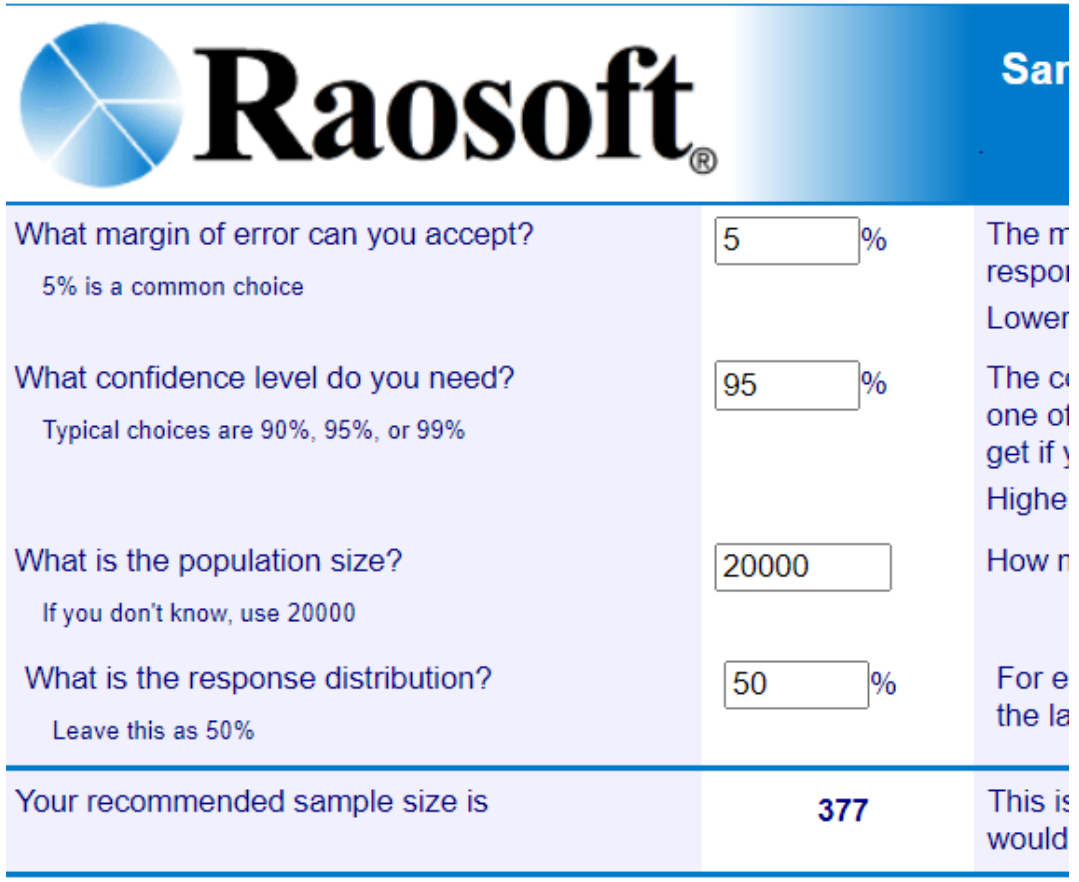
$$e = z\sqrt{\frac{pq}{n-1}}$$

$$n = \left(\frac{zs}{e}\right)^2 \text{ con s} = \text{pq n} = \frac{z^2pq}{e^2}$$

Ampiezza del campione

Esistono delle app online che consentono di calcolare la numerosità campionaria ideale.

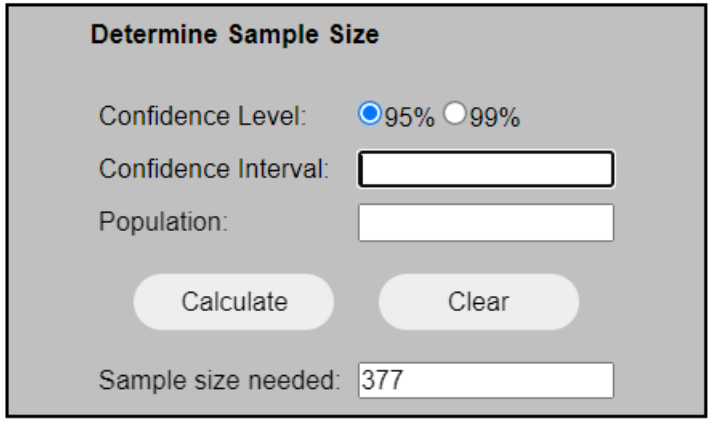
http://www.raosoft.com/samplesize.html

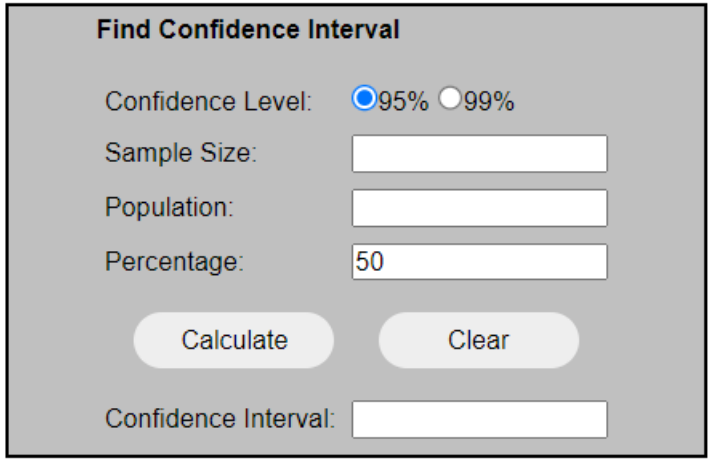


Ampiezza del campione

Esistono delle app online che consentono di calcolare la numerosità campionaria ideale.

https://www.surveysystem.com/sscalc.htm





Il campionamento

Perché ricorrere ad un campionamento?

- Riduzione dei costi di rilevazione
- Riduzione dei tempi di raccolta ed elaborazione dei dati
- Riduzione del numero di rilevatori e della loro formazione e gestione
- Vantaggi sull'accuratezza della rilevazione

E' una scelta obbligata nei casi in cui:

- La rilevazione implichi la distruzione (ad es. test sui prodotti)
- Nei casi di popolazione teorica (ad es. consumatori di un certo prodotto)

Il caso del Literary Digest

Siamo negli Stati Uniti nel 1936, alla vigilia delle elezioni presidenziali.

Nelle settimane precedenti più di una testata giornalistica vuole prevedere il vincitore tra il democratico Franklin D. Roosevelt, e il suo sfidante repubblicano, Alfred M. Landon, tramite dei sondaggi.

La nota rivista “Literary Digest”, che aveva correttamente previsto i risultati delle cinque elezioni presidenziali americane precedenti, decide di avviare il più ambizioso e costoso sondaggio di qualsiasi altro mai svolto in precedenza.

Ad agosto partono via posta 10 milioni di fac-simile di schede elettorali a nominativi estratti dai registri automobilistici e dagli elenchi telefonici. Entro il 31 ottobre il “Literary Digest” riceve ed elabora circa 2,4 milioni di voti.

Il repubblicano Landon ne esce vincitore con il 55% dei voti, contro il 41% di Roosevelt.

Il caso del Literary Digest

Pochi giorni dopo, l'esito delle elezioni smentisce completamente il pronostico del "Literary Digest": Roosevelt viene rieletto alla Casa Bianca con un ampio margine: ottiene infatti il 61% delle preferenze contro il 37% del candidato repubblicano.

Come si può sbagliare in modo così clamoroso, con un campione così grande?

Nel 1936 gli Stati Uniti erano ancora in preda alla Grande Depressione. Coloro che possedevano un'automobile e un telefono erano presumibilmente tra i più privilegiati nella società. Di conseguenza, l'elenco compilato dal "Digest" privilegiava gli elettori delle classi medie e alte che, con opinioni politiche più tendenti a destra, erano meno inclini a votare Roosevelt, e sottorappresentava la popolazione dei votanti del partito democratico.

Inoltre, non avevano tenuto conto del fenomeno dell'autoselezione, con circa 7,5 milioni di non rispondenti. Evidentemente le persone generalmente più ricche e istruite e che tendevano a votare repubblicano, erano anche più propense a rispondere al sondaggio.

Il caso del Literary Digest

Quello stesso anno, George Gallup, utilizzando un campione di poche migliaia di americani, predice correttamente la vittoria di Roosevelt.

Questo ci insegna che non è importante solo la dimensione del campione, ma ancor più la sua composizione.

Nel caso del Literary Digest ebbero un peso determinante:

- l'errore di copertura (le liste utilizzate per la rilevazione erano incomplete),
- l'errore di non risposta (fenomeno dell'autoselezione: chi ha risposto non era uguale a chi non ha partecipato alla rilevazione).

I disegni di campionamento

Distinguiamo due tipologie di disegni di campionamento:

- Probabilistici
- Non probabilistici

Un campione si dice probabilistico quando ogni unità della popolazione da cui viene estratto viene estratta con una probabilità nota e diversa da zero.

Ad esempio se voglio condurre un'indagine tra gli studenti ed estraggo un campione tra quelli che sono presenti all'università in un certo giorno non otterremo un campione probabilistico.

- chi non frequenta ha probabilità 0 di essere intervistata
- probabilmente sovrastimerò le matricole, che di solito frequentano le lezioni con maggiore assiduità rispetto ad esempio, agli studenti fuori corso, gli studenti-lavoratori, ecc.

I disegni di campionamento probabilistici

Ci sono diversi disegni di campionamento probabilistico:

- Campionamento casuale semplice
- Campionamento sistematico
- Campionamento stratificato
- Campionamento a stadi
- Campionamento a grappoli
- Campionamento per aree

Campionamento casuale semplice

Si parla di campionamento casuale semplice quando tutte le unità della popolazione di riferimento hanno la stessa probabilità di essere incluse nel campione.

Bisogna disporre della lista completa delle unità che compongono la popolazione di riferimento e poi procedere con una selezione casuale.

Non è molto utilizzato nelle ricerche sociali perché:

- Non include eventuali informazioni note a priori (ad es. il genere, l'età, ecc.)
- Nelle indagini su vasta scala il piano di rilevazione potrebbe rivelarsi costoso e complicato

Campionamento sistematico

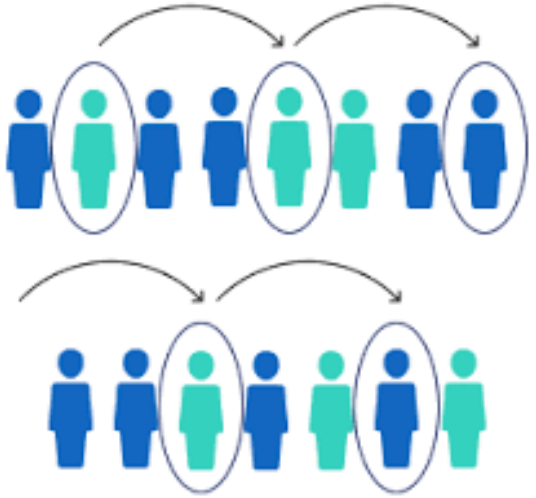
Le unità non vengono estratte tramite sorteggio, bensì individuando un **passo di campionamento** k. Si scorre la lista e si estrae un'unità ogni k.

Il passo di campionamento si ottiene dividendo la numerosità della popolazione N per quella campionaria n.

Passo di campionamento k = N/n

In caso di liste ordinate (ad es. per indirizzo di residenza, età, ecc.) il campionamento sistematico consente di garantire una adeguata copertura dei parametri utilizzati per l'ordinamento.

Viene usato ad es. negli exit polls, dove gli intervistatori contattano un elettore ogni k che escono dai seggi elettorali.



Campionamento stratificato

Abbiamo visto che l'errore di campionamento dipende anche dal grado di variabilità del fenomeno.

Se il fenomeno oggetto di studio contiene dei gruppi più omogenei al loro interno (ad es. conosciamo il genere, il titolo di studio) possiamo aumentare l'efficienza (maggiore precisione a parità di numerosità campionaria) con il campionamento stratificato.

Tre fasi:

- Suddividere la popolazione in strati (ovvero sottoinsiemi omogenei rispetto al fenomeno oggetto di studio).
- Estrarre con procedura casuale un campione da ogni strato.
- Unire i campioni ottenuti.

Campionamento stratificato

Ad esempio, se vogliamo studiare il reddito di una popolazione e disponiamo dell'informazione della posizione nella professione (operai, impiegati, dirigenti, liberi professionisti, ecc.), che sappiamo essere correlata con il reddito, dividiamo la popolazione in questi quattro strati e facciamo diverse estrazioni.

Se il campionamento riproduce la stessa composizione degli strati della popolazione, il campione si dice:

- **Proporzionale (o autoponderato)**

Se invece decidiamo di sovrarappresentare certi strati sottorapresentandone altri si chiama:

- **Non proporzionale**

In questo caso in fase di elaborazione dovremo attribuire dei pesi per ristabilire all'interno del campione la corretta proporzione degli strati nella popolazione.

Campionamento a stadi

Questa tecnica non produce un incremento dell'efficienza rispetto al campionamento casuale semplice, ma semplifica la procedura di estrazione e contiene i costi nella fase di rilevazione.

La popolazione viene suddivisa in più livelli organizzati in modo gerarchico. Supponiamo di voler condurre un'indagine tra i tesserati ad una federazione sportiva.

Devo costruire un primo stadio, che potrebbero essere le società (unità primarie)

Poi avrò un secondo stadio costituito dai tesserati (unità secondarie).

Faremo quindi due estrazioni successive:

Un campione di unità primarie (società sportive)

All'interno delle società selezionate, estraiamo un campione di tesserati.

Gli strati possono essere anche più di due: ad esempio, posso partire dalla provincia, poi seleziono alcune società in quelle province e poi seleziono i tesserati.

Campionamento a grappoli

Questa tecnica è simile al campionamento a stadi e viene utilizzata quando la popolazione è composta da gruppi di unità contigue nello spazio.

Alcuni esempi: le famiglie, i reparti/uffici in un luogo di lavoro, ecc.

Questi gruppi vengono definiti grappoli.

In questo caso non vengono estratte le unità elementari, bensì i grappoli.

Le unità che compongono i grappoli vengono inserite tutte nel campione.

Ad esempio possiamo estrarre delle famiglie dalle liste anagrafiche di un comune e poi intervistare tutti i componenti.

Campionamento a grappoli

Questa tecnica ha diversi vantaggi operativi:

- Non bisogna avere gli elenchi di tutta la popolazione, ma solo quelle relative alle unità dove si procede con l'estrazione (nel nostro esempio le società).
- Concentriamo la fase di rilevazione nelle sole unità estratte.

Il campionamento a grappoli però determina una perdita di efficienza perché le unità, che appartengono alle medesime unità di ordine superiore, tendono ad assomigliarsi.

Campionamento per aree

In assenza delle liste della popolazione, posso usare questa tecnica, molto simile al campionamento a stadi.

- Suddividiamo il territorio in porzioni (ad es. le sezioni di censimento, isolai) che raggruppano un numero limitato di abitazioni.
- Concentriamo la fase di rilevazione nelle sole unità estratte.

Anche il campionamento per aree determina una perdita di efficienza perché le unità, che appartengono alle medesime unità di ordine superiore, tendono ad assomigliarsi.

Disegni di campionamento non probabilistici

Questa tecnica di campionamento prevede che la selezione delle unità da rilevare avvenga in base a un giudizio soggettivo piuttosto che a criteri probabilistici.

Quelli più utilizzati sono:

- Campionamento per quote.
- Campionamento a valanga.
- Campionamento a scelta ragionata

Campionamento per quote

Questa tecnica è molto diffusa nelle ricerche di mercato e nei sondaggi d'opinione.

Il primo passaggio consiste nel dividere la popolazione in un certo numero di strati definiti da alcune variabili di cui conosciamo la distribuzione.

Sulla base di queste informazioni si quantifica quante unità rientrano in ogni strato (il «peso») e, in base alla numerosità campionaria si definiscono le quote, ovvero quante unità vanno intervistate per ogni strato.

Il limite più grosso dipende dal fatto che:

- L'intervistatore tenderà a contattare, all'interno delle quote prefissate, gli individui più facilmente reperibili/disponibili, non badando alle ripetute sostituzioni di chi è più restio.
- Vengono sistematicamente sottorappresentate quelle unità più difficilmente reperibili.

Nei sondaggi d'opinione e ricerche di mercato è molto diffuso perché garantisce un grande risparmio economico e di tempi per la realizzazione delle interviste.

Campionamento a valanga

Questa tecnica è utilizzata in particolare nello studio di popolazioni di difficile reperibilità (in generale gruppi sociali che tendono a nascondersi come clandestini, evasori fiscali, appartenenti a sette religiose, lavoratori in nero, ecc.) o composte da elementi «rari» (piccole comunità in rete tra loro).

Partiamo da un piccolo gruppo di individui appartenenti a quella popolazione, a cui chiederemo altri contatti, in modo da aumentare di volta in volta la numerosità campionaria.

Ovviamente si corre il rischio di contattare gruppi di soggetti molto omogenei tra di loro.

Campionamento a scelta ragionata

Le unità campionarie non vengono individuate su base probabilistica, ma sulla base di alcune loro caratteristiche.

Questa tecnica si utilizza maggiormente nelle indagini qualitative e più, in generale, quando l'ampiezza del campione è molto limitate e si preferisce raccogliere le opinioni di tutti gli strati di una popolazione, inclusi anche quelli numericamente più contenuti.

Ad esempio potrei decidere di intervistare almeno 3 elettori di ogni partito presente alle ultime elezioni comunali, indipendentemente dai risultati ottenuti dai vari schieramenti.