Indagini campionarie e sondaggi demoscopici
6 CFU – 52 ore

Ornella Giambalvo
ornella.giambalvo@unipa.it

INDAGINI CAMPIONARIE E SONDAGGI DEMOSCOPICI

Cosa è un campione?

Avete mai lavorato con campioni?

Che vuol dire campione probabilistico?

Cosa è una distribuzione di campionamento?

È preferibile un campione di grandi dimensioni o basta un piccolo campione?

È preferibile un censimento al campione?

Quanti tipi di campioni conoscete?

Argomenti da trattare

- Generalità sul campionamento
- Il campionamento da popolazioni finite
  - Implicazioni metodologiche e aspetti peculiari
- Piani di campionamento
  - Casuale semplice
  - Casuale stratificato
    - Stima della media, del totale e della proporzione
    - Dimensione campionaria
    - Confronti fra piani di campionamento

Premessa

Prendere decisioni è un'esperienza quotidiana familiare a tutti e diventa tanto più facile quante più informazioni si possiedono

Le indagini campionarie hanno proprio lo scopo di ***ottenere informazioni sul fenomeno*** *quindi sulla variabile oggetto di studio*

L'obiettivo è presentare come si perviene alla costruzione dell'informazione su un collettivo statistico (popolazione) sulla base dell'osservazione di una parte di esso (il campione), sottolineando come evitare di fornire risultati "solo" desiderati e non "ottenuti" dall'elaborazione dei dati disponibili.

Premessa

L’esigenza di conoscere in tempo reale i fenomeni sociali ha reso **sempre più diffuso l’impiego dell’indagine campionaria** a fini di ricerca.

Parallelamente si è sviluppata dal punto di vista statistico-matematico la teoria del campionamento.

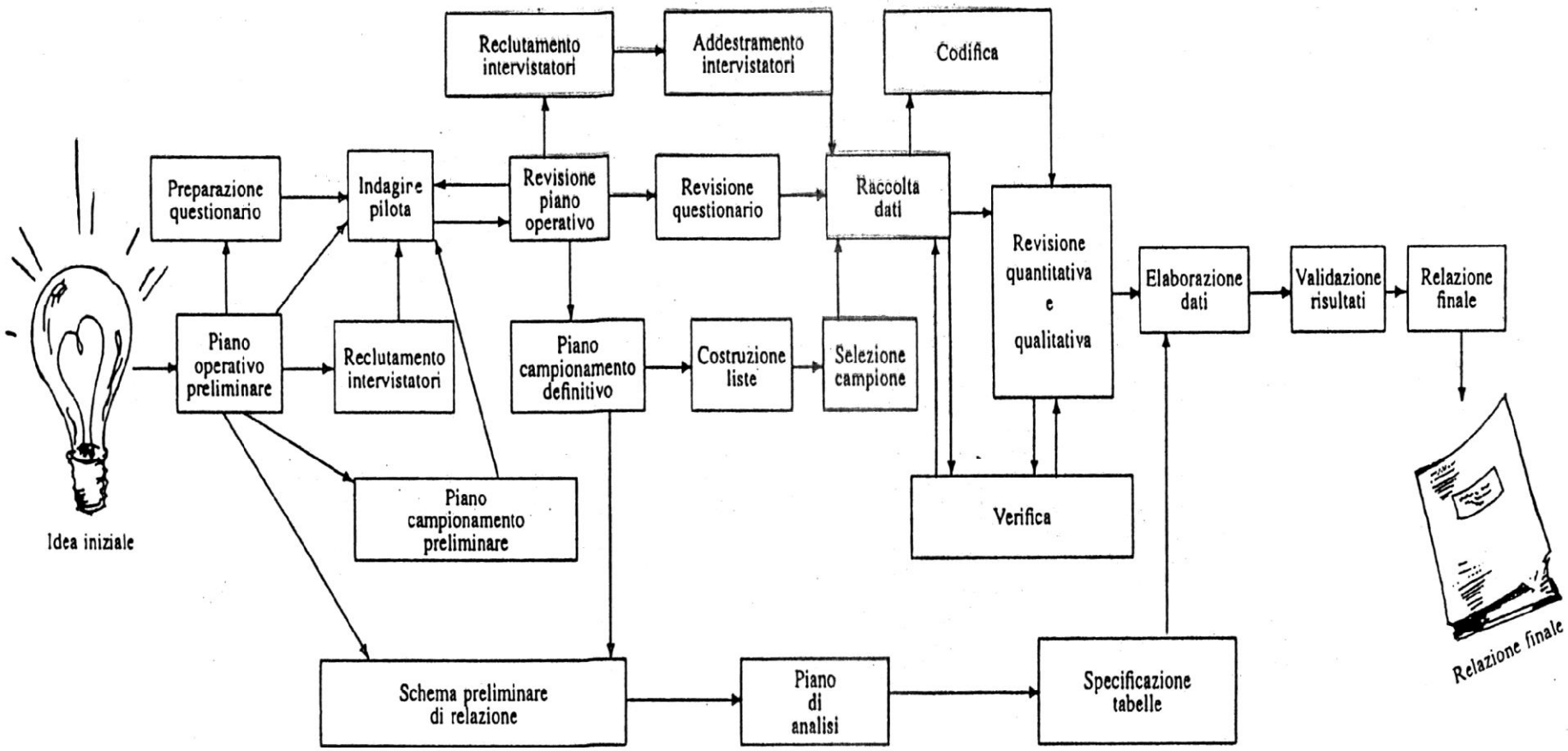


Fig. 1. Fasi dell'indagine campionaria (figura tratta da Ferber *et al.*, 1990).

Generalità:
Cenni storici sul campionamento

- Halley (1656-1742)

costruì le tavole di mortalità soltanto con i dati della città di Breslau (*prima indagine parziale*)

- Vauban (1633-1707)

stimò l'estensione delle terre coltivate in Francia sulla base di un campione di aree in un certo numero di province

- Laplace (1802)

stimò la popolazione della Francia tramite un campione di statistiche

- Marx (1880)

inviò un questionario a 2500 lavoratori francesi per studiare il fenomeno dello sfruttamento

- solo nel 1895, Kiaer *(direttore dell'ufficio centrale di statistica norvegese)*

discusse, per la prima volta, *la validità delle indagini campionarie* formulando il

Criterio della rappresentatività:

"il campione deve costituire una miniatura un'immagine su scala ridotta della realtà"

Generalità:
Cenni storici sul campionamento

- Il criterio della rappresentatività venne accolto con scetticismo
  - anche se Kiaer lo ripropose nel 1899 e nel 1901 eseguì un confronto con i dati censuari.
- Nel 1926 Bowley propose
  - la prima formulazione di campione
  - una teoria del campionamento

metodo per valutare la precisione delle stime ottenute da campioni con "n" grande, estratti da ampie popolazioni

Generalità:
Cenni storici sul campionamento

- Nel 1934 Neyman

fornì un insieme di regole per l'estrazione di campioni, che formarono i

fondamenti del campionamento probabilistico

- È possibile estrarre campioni casuali di unità misurabili;
- Il numero di unità deve essere elevato;
- È possibile usare le informazioni oggettive della popolazione senza turbare la casualità del campione;
- Vi sono situazioni in cui il campione a scelta ragionata può essere usato con successo (ma non è la norma).

Generalità:
Cenni storici sui sondaggi demoscopici

U.S.A. nel 1930

- Crossley, Gallup e Roper
  applicando le tecniche note e usate:
  - *per le ricerche di mercato*, in particolare il campionamento per quote
  - *per i sondaggi elettorali*
- nel 1936 primo clamoroso successo per le Elezioni presidenziali in U.S.A.

il Literary Digest realizzò un'enorme indagine che predisse la vittoria di Landon su Roosvelt

Contemporaneamente furono effettuati tre sondaggi d'opinione che predissero tutti l'esito opposto

Risultato: Vinse Roosvelt come previsto dai tre sondaggi

Generalità:
Cenni storici sui sondaggi demoscopici

Perché l'indagine del Literary Digest fallì?

a) le persone da intervistare erano state scelte da una lista di nominativi facendo riferimento

- all'elenco telefonico
- ai registri dei possessori di automobili

Veniva quindi esclusa gran parte della popolazione

l'automobile e il telefono in quegli anni non erano diffusi

b) il questionario fu rispedito solo da coloro che erano effettivamente interessati alla vittoria di uno dei due candidati

n grande non è una garanzia di successo !!

Generalità:
Cenni storici sui sondaggi demoscopici

Campioni di dimensioni più modeste, ma scelti con cura, con il criterio della rappresentatività, sono spesso più affidabili

i sondaggi avevano, invece, riprodotto la composizione elettorale americana usando la tecnica del campionamento per quote

Altri sondaggi effettuati

- nel 1937 in Gran Bretagna
- nel 1938 in Francia
- nel 1940 in Australia
- nel 1941 in Canada
- *nel dopoguerra in Italia*

tutti prevalentemente con lo scopo di fornire previsioni pre-elettorali

Dal 1970 in seguito ad un insuccesso nella previsione di vittoria Laburista in Gran Bretagna

i sondaggi iniziarono a perdere popolarità

Definizioni

L'indagine campionaria

Scelta di una *parte* di un insieme finito di unità per inferire sull'**intero** insieme (popolazione) sulla base della parte scelta

I sondaggi demoscopici

Inchieste con l'obiettivo di dar voce ai pensieri, alle opinioni e ai desideri della **gente**

"… polling is merely an instrument for ganging public opinion. When a president, or any other leader pays attention to poll results he is, in effect, paying attention to the views of the people … " GALLUP, 1972

Indagine campionaria o sondaggio demoscopico?

Il sondaggio demoscopico si differenzia dall'indagine campionaria solo per alcuni connotati tipici:

- *La popolazione obiettivo* è costituita quasi sempre da adulti di uno Stato o di un'area geografica
- *La dimensione campionaria* è sempre fra 1.000 e 2.000 unità
- *L'ambito territoriale* è ristretto, perché la risposta deve pervenire entro 2-3 giorni
- *Le stime* fornite riguardano la percentuale di casi che presentano una certa caratteristica
- *Il questionario* è generalmente composto da un numero non elevato di domande che *tendono all'accertamento delle opinioni e degli atteggiamenti delle persone nei confronti di questioni pubbliche o sociali*
- *I risultati* sono rivolti ad un vasto uditorio

il campione può essere estratto senza seguire regole formali ma usando

- il criterio della prevalenza (di una unità statistica su un'altra)
- l'osservazione delle unità
- l'esperienza nella scelta dell'unità

La qualità di un sondaggio demoscopico

Per definire **“buono”** un sondaggio occorre fornire indicazioni su:

- Tipo e struttura del campione
- Modalità di acquisizione delle informazioni
- Incidenza delle non risposte e dei mancati contatti
- Entità dell’errore di campionamento (solo se l’indagine è probabilistica)

La qualità di un sondaggio d’opinione dipende:

- *dall’attendibilità delle stime*
  - errori da incomplete osservazioni
  - errori di misurazione
- *dall’attendibilità complessiva dell’indagine*
  - decresce all’aumentare del numero e del tipo di sondaggi effettuati
  - dipende da un inevitabile calo di cooperazione da parte della popolazione e dal calo negli standard tecnici adottati per la realizzazione

L’indagine campionaria o censuaria?

- **L’indagine censuaria**
  - rileva informazioni da tutta la popolazione oggetto di studio
  - presenta elevati tassi di risposta
- **L’indagine campionaria (parziale)**
  - rileva informazioni da un sottoinsieme della popolazione cui si riferisce l’indagine
  - consente di ottenere informazioni in tempi brevi (quindi le informazioni saranno meno obsolete)
  - presenta costi medio-bassi
  - studia fenomeni che prevedono la distruzione delle unità (es. la durata delle lampadine)
  - permette di ottenere rilevazioni più accurate e approfondite

L’indagine campionaria o censuaria?

perché l’alternativa ?

Possono completarsi a vicenda:

a) nella progettazione di un’indagine campionaria (parziale) ci si serve di dati censuari;

b) i risultati di un’indagine censuaria possono essere completati e/o verificati da indagini campionarie (parziali) per:

- **valutare gli errori**
- **accertare la qualità del dato**

L’indagine campionaria o censuaria?

Vantaggi dell’indagine campionaria

Si ricorre certamente all'indagine campionaria quando:

- la rilevazione completa è impossibile
- la determinazione della modalità posseduta dalle unità in esame ne comporta la distruzione

ATTENZIONE: le informazioni ottenute da un'indagine si ritengono valide solo se sono al di sopra di una certa soglia di precisione

Vantaggi dell’indagine censuaria

- fornisce dati per un qualunque sottinsieme o dominio della popolazione, indipendentemente dalla sua ampiezza
- permette così, in un secondo momento, di verificare la rappresentatività di un suo eventuale campione

Il piano di campionamento

"La prima cosa da fare - si diceva - è di riacquistare la mia vera statura. Poi devo ritrovare la via che porta a quel meraviglioso giardino. Questo è il piano migliore. Senza dubbio questo era un piano eccellente, semplice e davvero ben congegnato. C'era solo una difficoltà: Che Alice non aveva la più piccola idea di come realizzarlo." **[Alice nel Paese delle meraviglie - Lewis Carroll]**

- **Qualsiasi indagine campionaria richiede un piano di lavoro dove vengono definiti e precisati gli aspetti fondamentali dell'indagine sintetizzati nei seguenti passi:**

Le fasi per un piano di campionamento

- **Formulazione degli obiettivi dell'indagine**

Si tratta di precisare gli scopi dello studio, le sue finalità conoscitive e le variabili di indagine (es. età, genere…). Rientra nella definizione degli obiettivi l'indicazione del grado di precisione voluto dal committente per il risultato dell'indagine. Si tratta di un problema delicato perché, dal livello di precisione prescelto, discende anche il costo dell'indagine.

- **Periodo di svolgimento**

È il tempo impiegato per la realizzazione dell'indagine, compreso anche il tempo per la divulgazione dei risultati.

Le fasi per un piano di campionamento

- **Periodo di riferimento**

È il lasso di tempo cui vanno riferite le informazioni riguardanti le unità indagate. Può essere conveniente scegliere periodi di riferimento diversi per le stesse caratteristiche studiate.

- **Determinazione della lista**

È l'elenco degli elementi che costituiscono la popolazione e rappresenta la base per la scelta delle unità da inserire nel campione. La disponibilità di una buona lista è un requisito essenziale per la riuscita di un'indagine campionaria.

Le fasi per un piano di campionamento

- **Scelta del piano di campionamento**

Per piano di campionamento si intende il procedimento con cui viene formato il campione. Nelle indagini su larga scala il piano di campionamento ha generalmente una struttura complessa; è quasi sempre a più stadi, nel senso che le unità finali non vengono scelte direttamente, ma attraverso tappe successive (Es. indagini sulle forze di lavoro redatte dall'Istat).

Può essere probabilistico o non probabilistico

- ***Processo selezione***
- ***Processo stima***

(solo nel caso di campionamento probabilistico)

Le fasi per un piano di campionamento

- **Metodo di raccolta dei dati**

Si tratta di stabilire come acquisire le informazioni desiderate attraverso un processo di misurazione dei dati. Lo strumento più utilizzato per la raccolta dei dati è il questionario che può essere somministrato in vari modi (intervista diretta, postale, telefonica, on line…). In alternativa si potrebbe affidare ad intervistatori perfettamente addestrati il compito di formulare le domande e di interpretare le risposte.

Le fasi per un piano di campionamento

- **Lavoro sul campo**

Si tratta della fase esecutiva della ricerca in cui vengono raccolti i dati. Riguarda le modalità organizzative con cui la raccolta dei dati viene espletata. Queste comprendono l'addestramento dei rilevatori, i quali devono avere non soltanto la competenza nei settori di indagine, ma anche la capacità di acquisire informazioni da diverse fonti.

Le fasi per un piano di campionamento

- **Elaborazione ed analisi dei dati**

È il processo mediante il quale i dati raccolti vengono "trattati" con opportuni metodi statistici. Ci si avvale, in genere, dell'utilizzo di software specializzati che ci permettono di organizzare, elaborare e analizzare i dati raccolti, producendo tabelle e grafici che costituiscono la base per l'analisi dei risultati dell'indagine.

- **Preparazione della relazione finale**

La relazione deve contenere specifici paragrafi contenenti un sintetico e completo quadro generale del lavoro svolto.

"STATISTICAL THINKING WILL ONE DAY BE AS NECESSARY FOR EFFICIENT CITIZENSHIP AS THE ABILITY TO READ AND WRITE" H. G. Wells

Il campionamento probabilistico o … non probabilistico?

Definizioni:

- *il campionamento probabilistico* è quello in cui, per ogni elemento della popolazione è nota la probabilità di far parte del campione

- *il campionamento non probabilistico*, è quello in cui, invece, non è nota la probabilità che le unità della popolazione hanno di far parte del campione.

Il campionamento non probabilistico

- I metodi di formazione del campione, detti non probabilistici, prescindono dai criteri di casualità nella scelta delle unità campionarie.

Si parla anche di campionamento a scelta arbitraria, quando le unità campionarie vengono selezionate in modo non casuale sulla base di informazioni preliminari riguardanti la popolazione indagata. Questa tecnica è appropriata soprattutto per piccoli campioni.

Assunto di base

le caratteristiche studiate nella popolazione sono distribuite

uniformemente e *casualmente*

Tipi di campionamenti non probabilistici

- **Campioni a casaccio o fortuiti**
- **Campioni di soggetti volontari** disposti a sottoporsi a particolari "trattamenti" (es. test sperimentali su nuovi farmaci…)
- **Campioni di esperti** (protagonisti)
  - Purpositive samples
- **Testimoni privilegiati** (conoscitori)
- **Campione a "valanga"**

  Impiegato nelle indagini su popolazioni rare, consiste nello scegliere un gruppo iniziale di persone, dalle quali poi ottenere nomi e indirizzi di altre unità appartenenti alla stessa popolazione.
- **Campioni cattura – ricattura** (popolazioni mobili)

  Usato su popolazioni di animali. Il totale della popolazione è stimato dalla proporzione di individui "ri-catturati"

Tipi di campionamenti non probabilistici

- **Campioni per quote** (utilizzati nelle indagini di mercato, nei sondaggi di opinione e per la misura di atteggiamenti)

  Si suddivide la popolazione in classi o sotto gruppi omogenei e dai dati censuari, o da altre fonti, si ricava il peso percentuale di ogni classe; il totale delle unità nel campione viene, poi, suddiviso tra le classi in modo da rispecchiare le proporzioni esistenti nella popolazione. Si perviene dunque alla definizione delle quote, cioè del numero delle osservazioni da effettuare in ciascuna classe. L'elemento caratteristico del campionamento per quote è che la scelta delle unità da inserire nel campione è demandata all'intervistatore stesso nell'ambito delle quote assegnate.

Il campionamento non probabilistico

VANTAGGI:

bassi costi

velocità di esecuzione (riduzione dei tempi)

SVANTAGGI:

non è possibile misurare il livello di “precisione” dei risultati

ha una pesante dipendenza dalle assunzioni (più o meno valide) sulle distribuzioni delle variabili di indagine della popolazione. In questo caso non è possibile fare inferenza secondo la concezione “classica”.

Il campionamento probabilistico

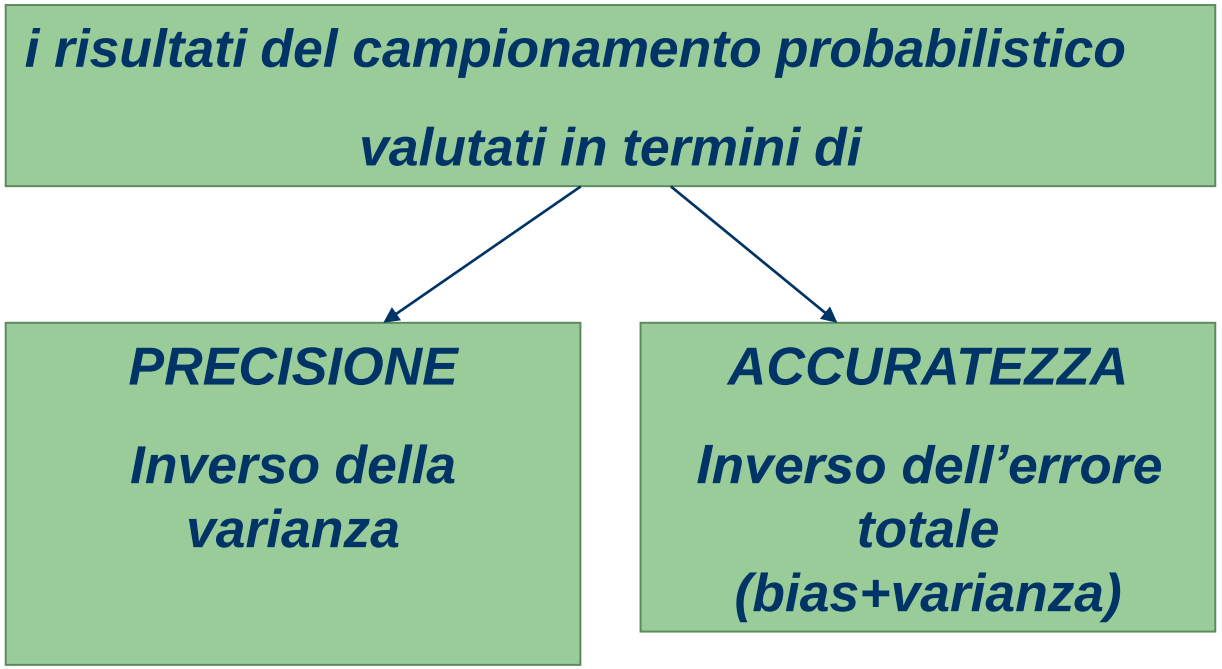
È quel campione per cui le

costanti caratteristiche della popolazione *sono oggetto naturale dell'inferenza*

- Deve essere **orientato** verso l'obiettivo
  - Nel processo di selezione e stima orientato agli obiettivi della ricerca

- Deve essere **misurabile**
  - Nota la probabilità di ogni elemento della popolazione si possono calcolare
    - Errori standard
    - Funzioni di verosimiglianza (?)
    - Distribuzioni di campionamento (?)

Il campionamento probabilistico



Il campionamento probabilistico

Prima di procedere con il campionamento probabilistico

- Occorre un'analisi economica
  - Raggiungimento degli obiettivi al minimo costo (denaro-tempo-fatica)

- Occorre un'analisi di fattibilità
  - Cura nel tradurre un modello teorico di selezione in un insieme di istruzioni utili per gli esecutori
    - Chiare
    - Semplici
    - Pratiche
    - Complete

La selezione delle unità statistiche nel campionamento probabilistico

- Il processo di selezione è
  - guidato da un corpo di *principi e procedure* per l'inferenza statistica
  - può coinvolgere parecchie *fonti di randomizzazione*
    - *Nel metodo usato per la scelta delle unità*
    - *Nel metodo usato per ottenere una misura per l'unità scelta*
    - *Nella conoscenza di un processo che genera la **vera** misura per una unità*

La selezione delle unità statistiche nel campionamento probabilistico

- *Metodo usato per la scelta delle unità*

Approccio da una popolazione fissata

- Il meccanismo di randomizzazione determina quale sottoinsieme osservare
- Ogni elemento della popolazione è associato ad un numero ‘reale’ ignoto
- Fissa il valore della variabile in studio (approccio classico)

La selezione delle unità statistiche nel campionamento probabilistico

- *Metodo usato per ottenere una misura per l'unità scelta*

Modelli per errori di misura

- *Teoria degli errori non campionari*
- *La misura è "imperfetta" e gli errori possono essere modellati statisticamente*

La selezione delle unità statistiche nel campionamento probabilistico

- *Conoscenza di un processo che genera la **vera** misura per una unità*

Approccio da "superpopolazione"

- Ogni unità della popolazione è associata ad una v.c. di cui si specifica la struttura stocastica
- Il valore osservato associato a quello della popolazione è una determinazione della variabile casuale (approccio predittivo)
- Ciascuna popolazione finita è considerata un campione casuale estratto da una popolazione più grande

Superpopolazione

- Il parametro non è fisso ma una variabile casuale multidimensionale

Il processo inferenziale classico

FISHER

Inferenza da popolazioni infinite

La struttura dei dati:

- *Campione costituito da n unità i.i.d.* $(x_1, x_2, x_3, \ldots x_n)$ di una v.c. con $f(x, \theta)$

Obiettivo

- *Stima del parametro* θ

Tipo di Inferenza

classica – descrittiva

- Indaga su singoli parametri di una variabile osservata sul campione
- Stima di un parametro (es. reddito medio di un'area geografica)

analitica

- Tenta di "spiegare" la distribuzione di una variabile nella popolazione per mezzo di relazioni con altre variabili
- Stima di una relazione (es. come varia il reddito al variare dell'età del CF)

Il processo inferenziale classico

Inferenza classica – descrittiva

o

Inferenza analitica?

Dipende dal tipo di modello statistico che si sceglie

Il processo inferenziale, ***fondamentale per garantire la casualità****, è piuttosto diverso:*

- Tipo di campionamento effettuato
- Modello interpretativo scelto

Il processo inferenziale per le IC

GODAMBLE

Inferenza da popolazioni finite

Caratteristiche

- *La popolazione è reale*
- *Problema dell'identificabilità delle unità della popolazione*
  - *IDENTIFICABILITA'*
    - *è possibile etichettare le unità della popolazione*

I concetti di parametro – campione - dati – stimatore
assumono un altro significato

Il processo inferenziale per le IC

Esempio 1:

$$\{1, 28\}\{2, 20\}\{3, 20\} \quad N = 3$$

- *Se estraggo campioni di dimensione n=2*

(28, 20) prima coppia (y_1, y_2)

(28, 20) seconda coppia (y_1, y_3)

Non è detto che diano luogo alla stessa "inferenza"

Il processo inferenziale per le IC

Se si ipotizza

n=2 estraggo tutti i possibili campioni senza re-immissione

$$\begin{array}{l} 1^{o}\ (1,2) \rightarrow 28,20 \quad p(s_1) = 0,32 \\ 2^{o}\ (1,3) \rightarrow 28,20 \quad p(s_2) = 0,40 \\ 3^{o}\ (2,3) \rightarrow 20,20 \quad p(s_3) = 0,28 \end{array} \qquad N = 3 \quad n = 2$$

Elementi distinti

Il processo inferenziale per le IC

Esempio 2:

I campioni hanno probabilità diversa di essere estratti perché ogni unità y_i ha probabilità diversa di far parte del campione

$$\sum_{i=1}^{3} p(s_i) = 1$$

si definisce lo stimatore t

$t = \frac{y_1 + y_2}{6 \times p_{s1}}$ se si verifica s_1 $t = \frac{y_1 + y_3}{6 \times p_{s2}}$ se si verifica s_2 $t = \frac{y_2 + y_3}{6 \times p_{s3}}$ se si verifica s_3

La distribuzione di campionamento ? È difficile da costruire

$E(t) = \mu$????????

Le tipologie di popolazioni finite

- *Popolazione su cui si fa inferenza*

PI = insieme di individui oggetto di studio in un intervallo di tempo definito

- *Popolazione obiettivo 'target population'*

PO = insieme di individui di ampiezza finita, scopo dello studio

- *Popolazione base per il campionamento 'frame population'*

PC = è la lista di unità utilizzate per estrarre il campione

su cui viene realizzata la procedura di selezione.

- *Popolazione effettivamente indagata 'survey population'*

PE = insieme di individui che, se selezionati, sono/diventano oggetto dell'intervista

Le tipologie di popolazioni finite

Esempio 3:

Campione per un'indagine sui consumi alimentari

Si considera la popolazione presente nel territorio nazionale nel semestre precedente

- *PI (inferenza)*

Escludiamo i detenuti e i ricoverati

- *PO (obiettivo)*

Dall'anagrafe prendiamo i nominativi

- *PC (lista/frame)*

E procediamo con le operazioni per etichettare la lista

- *PE (effettiva/sample)*

PI>PO PC>PE