Analisi dei dati

Domenico De Stefano

a.a. 2018/2019

Indice

Le dimensioni principali di un'indagine

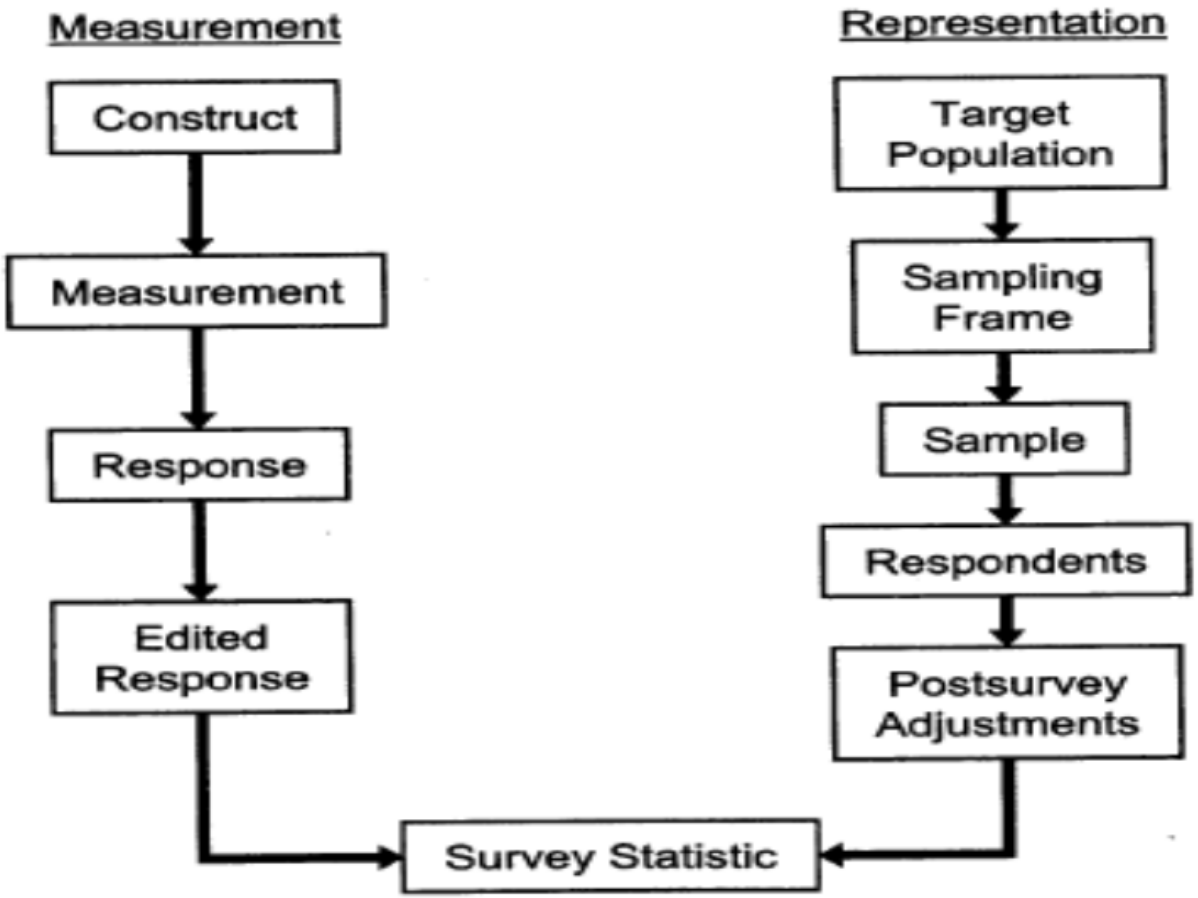
- Misurazione
  - quali dati devono essere collezionati sulle singole unità statistiche nel campione
  - Di cosa si interessa l'indagine?
- Rilevazione (rappresentazione)
  - a quale popolazione si riferisce l'indagine
  - Su chi è condotta l'indagine?

Entrambe le dimensioni richiedono attenzione in fase di pianificazione e realizzazione dell'indagine

Approcci alla progettazione/realizzazione di un'indagine

1. Approccio da disegno: dai concetti astratti alle azioni concrete (operative)
2. Approccio alla qualità: possibili fonti di errore che possono influenzare (distorcere) i risultati

Progettazione d'indagine: approccio da disegno



Approccio da disegno / Misurazione

- Costrutti (Constructs): sono gli "elementi/caratteristiche" che si cerca di "misurare" attraverso l'indagine
  - ad esempio nell'indagine multiscopo "sicurezza dei cittadini" un elemento che si vuole misurare è la percezione della sicurezza nella zona in cui si vive
  - ... altro esempio, l'altezzo o il peso
  - a questo livello la formulazione è astratta e suscettibile di più interpretazioni
- Misurazioni (Measurements): sono "elementi" dell'indagine più concreti dei costrutti. Le misurazioni sono le realizzazioni dei costrutti
  - ad esempio nell'indagine multiscopo "sicurezza dei cittadini" la sicurezza nella zona in cui si vive è misurata attraverso una serie di domande nel questionario, tra cui:

    Q: "quanto si sente sicuro camminando per strada quando è buio ed è solo nella zona in cui vive?"

    - Molto sicuro
    - Abbastanza sicuro
    - Poco sicuro
    - Per niente sicuro
  - ... per il peso e l'altezza ad es. Kg o Grammi; cm e metri

Approccio da disegno / Misurazione (2)

- Risposta (Response): il dato prodotto dall'unità statistica a seguito della misurazione
  - La natura delle risposte è determinata dal modo in cui si definisce la misurazione (es: natura qualitativa, quantitativa)
  - Quando la misurazione è effettuata mediante domande in un questionario l'unità statistica ha vari modi di produrre la risposta
    - Può ricorrere al ricordo
    - Può ricorrere ad un documento (es: domanda su personale impiegato presso un'azienda)
    - Può ricorrere ad informazioni fornite da un'altra persona (es: chiedere al partner)
  - Nel caso delle domande ad un questionario l'unità statistica può scegliere tra una serie di opzioni di risposta oppure produrre una risposta libera senza vincoli

Approccio da disegno / Misurazione (3)

- Risposta processata (Edited Response): il dato processato dal sistema di rilevazione o dal ricercatore posteriore alla raccolta dati
  - Risposta processata dal sistema: alcuni strumenti di rilevazione dati di tipo computer-assisted effettuano contestualmente una verifica di coerenza delle risposte fornite alla misurazione
    - Ad es: nel caso si chieda l'anno di nascita specifici range di valori possono essere controllati dal sistema (anno < 1890) e si può richiedere una verifica (**range check**)
    - Ad es: se una us dichiara di avere 14 anni e in un'altra domanda dichiara di avere 5 figli lo strumento di rilevazione può richiedere una verifica della risposta (**consistency check**)
  - Risposta processata dal ricercatore: dopo la fase di rilevazione dati il ricercatore può controllare l'intera distribuzione delle risposte e individuare eventuali valori anomali (**outliers**) da trattare in modo opportuno

Approccio da disegno / Rappresentazione

- Universo o popolazione di riferimento/obiettivo (Target population)
  - La popolazione oggetto di interesse per l'indagine
  - questo è il livello più astratto di definizione della popolazione di interesse
  - si definisce in base agli obeittivi dell'indagine
  - ad es. nell'indagine multiscopo "sicurezza dei cittadini" la target population è l'insieme dei residenti in Italia con età superiore a 14
  - il concetto di tale popolazione è astratto in quanto non viene menzionato, ad esempio, il riferimento temporale ($>$ 14 anni al 1° Gennaio 2016)

Approccio da disegno / Rappresentazione (2)

- Popolazione statistica oggetto di studio (Frame o Survey population)
  - Pop. effettiva che viene indagata, anche in relazione alla disponibilità/caratteristiche del sampling frame (lista di campionamento) usati per identificare le u.s. della target population
  - popolazione che ha la possibilità di entrare a far parte del campione
  - Nel caso più semplice (ideale) il sampling frame è la lista di tutte le unità statistiche nella target population
  - in altri casi il sampling frame è una lista di unità statistiche non perfettamente collegate alle unità statistiche nella target population (ad es: numeri telefonici, email, ecc.)

se la target population e la frame population non sono sovrapposte bisogna tener presente la possibilità di errori di copertura

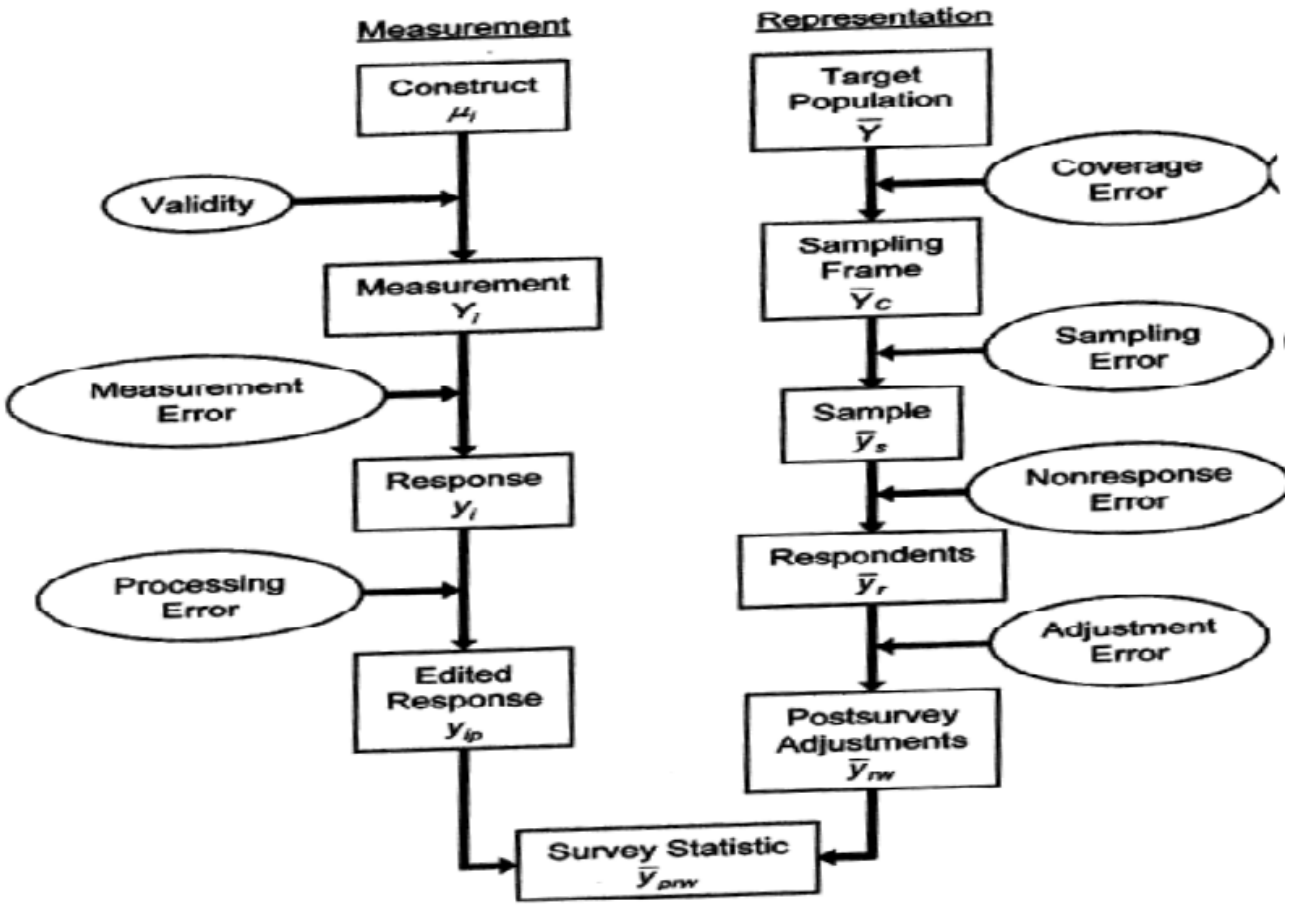
Approccio da disegno / Rappresentazione (3)

- Campione (sample): campione estratto dalla frame population
  - Esso è spesso una piccolissima frazione della frame population (e quindi anche della target population)
- Rispondenti (respondents): le unità statistiche campionate che effettivamente hanno fornito risposte alle misurazioni (il loro complementare sono i **non rispondenti**)
  - Il tentativo di "misurazione" può non andare a buon fine e quindi ci sarà una frazione di non rispondenti tra le unità campionate
  - se un'unità statistica non risponde completamente alle misurazioni si parla di unit nonresponse
  - a volte le definizioni di rispondente/non rispondente non sono univoche. Ad es: potrebbe essere definito non rispondente anche l'unità statistica che ha riposto solo ad alcune misurazioni
  - occorre operare delle scelte post-rilevazione per decidere se escludere dalla matrice dati le unità che non hanno fornito risposte a tutte le misurazioni oppure lasciare l'unità statistica ma classificare come **item nonresponse** o **missing data** le risposte mancanti

Approccio da disegno / Rappresentazione (4)

- Aggiustamenti post-rilevazione (postsurvey adjustments): aggiustamento dei dati sulla base delle caratteristiche dei rispondenti
  - dopo la fase di rilevazione e aver costruito la matrice dati per i rispondenti è possibile avere dei problemi di sovra- o sotto-rappresentazione di particolari sottogruppi della target population dovuti a:
    - Non corrispondenza della frame population con la target population (problemi di copertura)
    - Non rispondenti
  - ad es: in un'indagine via internet potremmo sottorappresentare il sottogruppo degli over 50 (se presenti nella target population)
  - a tal fine è possibile utilizzare dei pesi per ristabilire l'equilibrio tra i sottogruppi
  - oppure individuare dei metodi di imputazione per i missing data (procedure di stima per dati mancanti)

Progettazione d'indagine: approccio in termini di qualità



Approccio in termini di qualità / Misurazione

- Ogni fase della misurazione contiene elementi che possono inficiare la qualità del dato finale
- Lo scopo di chi pianifica l'indagine è quello di massimizzare la qualità e minimizzare gli errori che si possono commettere in ciascuna fase dell'indagine
- È importante notare che ciascun termine di errore riguarda la singola risposta (relativa al singolo item/domanda del questionario) non tutta l'indagine

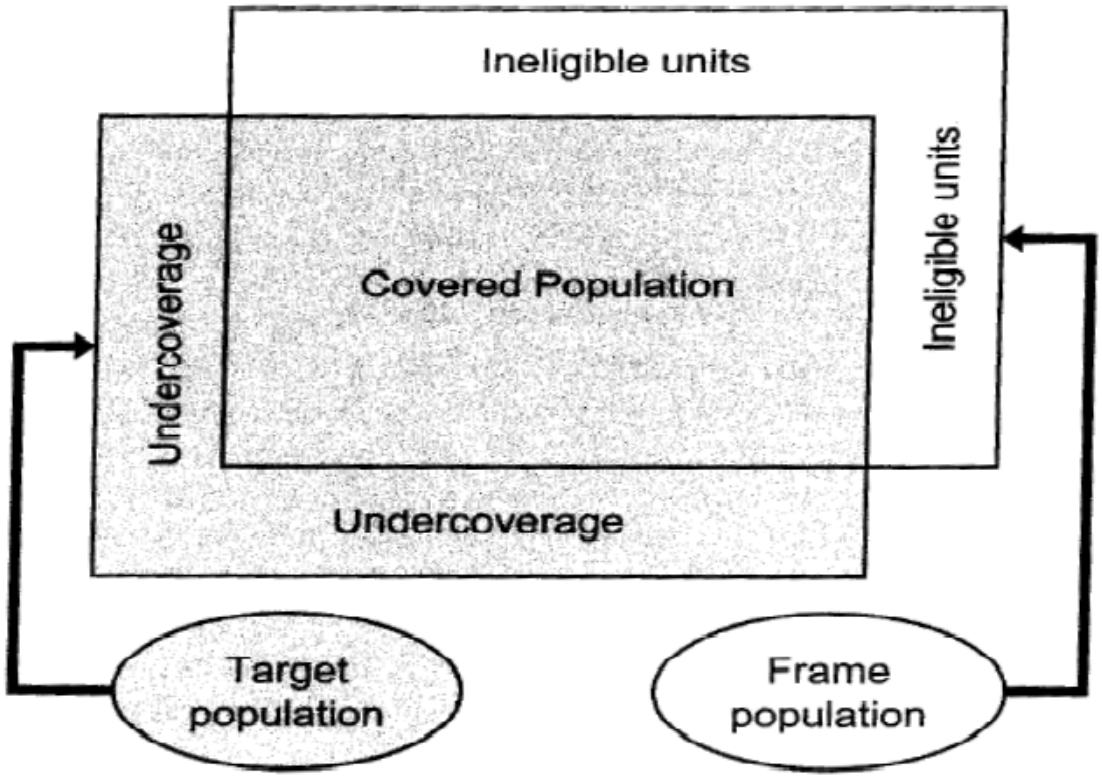
Approccio in termini di qualità / Rappresentazione

- Relativamente agli errori nella prospettiva della rappresentazione passiamo dalla singola risposta alle statistiche vere e proprie (es. media, mediana, ecc.)
- Nello schema semplificato tratteremo della media $\bar{Y}$ (ma in realtà il discorso vale per qualunque altro indice)

Approccio in termini di qualità / Rappresentazione (2)

- Quando la frame population non coincide con la target population si ha il cosiddetto **errore di copertura** (es. sample frame costituito da numeri telefonici)
- Quando alcune unità statistiche della target population non sono incluse nella frame population si parla di sottocopertura o non copertura (es. non possessori di apparecchi telefonici)
- Quando invece nella frame population ci sono unità statistiche non incluse nella target population si parla di ineleggibilità di tali unità ad entrare nel campione
- la copertura è la sovrapposizione delle unità statistiche nella target population e nella frame population. Si parla di eleggibilità di tali unità ad entrare nel campione

Approccio in termini di qualità / Rappresentazione



Approccio in termini di qualità / Errore di copertura

Indicando con:

- $\bar{Y}$ = media nell'intera target population (incognita) NB: in inferenza la indichiamo con la lettera greca μ
- $\bar{Y}_C$ = media della frame population
- $\bar{Y}_U$ = media della target population non inclusa nella frame population
- N = numero di unità statistiche target population
- C = numero di unità statistiche nella frame population eleggibili ad entrare nel campione (copertura)
- U = numero di unità statistiche nella target population non presenti nella frame population (sottocopertura o non copertura)

L'errore di copertura sarà:

$$\bar{Y}_C - \bar{Y} = \frac{U}{C}(\bar{Y}_C - \bar{Y}_U)$$

ossia l'errore nella stima della media della target population dovuta alla sottocopertura è il prodotto del tasso di non copertura ($\frac{U}{C}$) e la differenza tra la media delle unità statistiche sia nella target che nella frame population e quella delle unità statistiche nella target ma non nella frame population