Tecniche di indagine statistica

Lezione 25

Alcune considerazioni su non risposta (NR)

- Totale (unit non response)

  mancanza di informazione su un rispondente
  - rifiuto
  - assenza
  - impossibilità a rispondere
  - perdita questionario

- Parziale (item non response)

  mancanza di informazione su uno o più quesiti
  - rifiuto
  - incapacità o non volontà a rispondere
  - inconsistenze
  - errata registrazione

effetto $\Longrightarrow$ variabilità (campione finale più piccolo)
e distorsione (se non risposta selettiva)

Motivi non riposta totale e def.ne tasso di *risposta*

Composizione campione (in generale)

$$n = n_{NC} + n_{OC} + n_{RF} + n_{NA} + n_R$$

n_{NC} = Non Contattati
n_{OC} = Non Eligibili tra i contattati
n_{RF} = Rifiuti
n_{NA} = Impossibilitati a rispondere
n_R = Rispondenti

Tasso di risposta $TR = n_R \,/\, n_E$ E = Eligibili

se n_{NC} tutti elegibili $TR = n_R \,/\, n_{NC} + n_{RF} + n_{NA} + n_R$

se proporzione E tra NC = proporzione E tra C

$$TR = \frac{n_R}{n_{NC}\,[(n_{RF} + n_{NA} + n_R)/(n_{OC} + n_{RF} + n_{NA} + n_R)] + n_{RF} + n_{NA} + n_R}$$

Ipotesi su eligibilità non contatti

In indagini SELF (web)
(in generale) $TR = n_R \,/\, (n_R + n_{NR})$ $(n_{NR}=Non\ rispondenti)$

Non risposta e trattamento

Da considerare:

Entità: proporzione di casi mancanti (in totale e per singola variabile)
Distorsione: in generale, più frequente in gruppi particolari

Come trattarla? Decisioni necessarie per analisi dei dati

Tenere solo i casi completi può far perdere troppa informazione

Se si può ipotizzare che NR sia:

1. completamente casuale (*missing completely at random,* MCAR)
   probabilità NR non è collegata né al valore mancante sulla variabile, né al valore di ogni altra variabile osservata
2. casuale (*missing at random,* MAR)
   probabilità NR non è collegata al valore mancante sulla variabile ma dipende da altre variabili (nr su gruppi di unità che potrebbero essere identificati da valori su altre variabili)

alcuni 'aggiustamenti' possono ridurre la distorsione e ripristinare la numerosità campionaria

- unit non response: aggiustamenti con pesi che assegnano peso maggiore a categorie (*uso var. ausiliarie*) sotto rappresentate
- item non response: metodi di imputazione: valori mancanti replicati da valori *"di sintesi"*

Non risposta: modello generale per imputazione singola

$$\hat{y}_i = b_0 + \sum_{j=1}^{p} b_j X_{ij} + e_i$$

$\hat{y}_i$ = valore imputato per l'unità i (il cui valore è mancante)

X_{ij} = variabile ausiliaria j relativa all'unità i

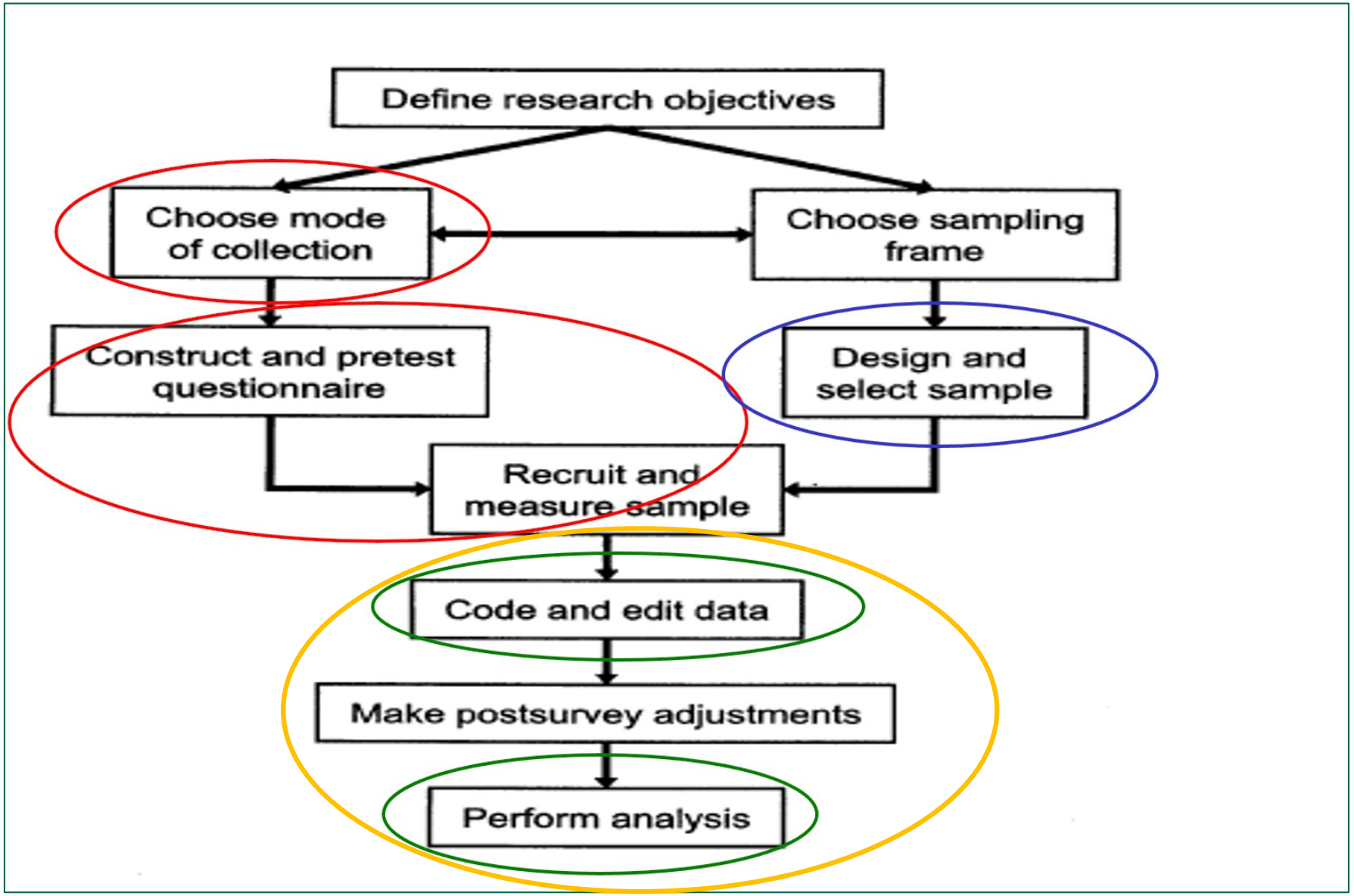
b_j = coefficienti di regressione ($j = 0, \ldots, p$) (stimati sui dati dei rispondenti)

e_i = termine di errore (determinato dalla specifica tecnica di imputazione)

con $b_0 = \bar{y}$, $b_j = 0$, $e_i = 0$: imputazione di $\bar{y}$ calcolata sui rispondenti

Se variabili X_j sono variabili indicatrici (*dummy)* per appartenenza ad un gruppo ($X_{ij} = 1$ se i è nel gruppo j, per es.: laureati), si imputano le medie di gruppo $\bar{y}_j$

Ciclo di vita indagine come processo (produttivo) – Fasi indagine



(Groves et al., 2009, p.47)

Analisi dei dati /risultati

By the time you get to the analysis of your data, most of the really difficult work has been done.

It's much more difficult to: define the research problem; develop and implement a sampling plan; conceptualize, operationalize and test your measures; and develop a design structure.

If you have done this work well, the analysis of the data is usually a fairly straightforward affair. (https://conjointly.com/kb/, William M.K. Trochim, Cornell University)

Data analysis: in genere 3 steps principali

1. Cleaning/cleansing and organizing the data for analysis (Data Preparation)

Operazioni anche dette: data *wrangling (munging)* che può seguire il data *harvesting* (più tipiche per dati da siti web, social media, survey online e/o dati d'archivio)

2. Describing the data (Comprensione e descrizione dei dati)

3. Testing Hypotheses and Models (Inferenza statistica e previsione)

Analisi dei dati - risultati

1. codifica
   - piano di codifica
     - variabili/informazioni specifiche rilevate (o calcolate e rese disponibili nel file dati)
2. descrizione (*editing*) per individuare dati mancanti, outlier, inconsistenze, …
   - distribuzioni di frequenza
   - indici sintetici e misure di dispersione per variabili quantitative
   - rappresentazioni grafiche
   - incroci di vario tipo

trasformazioni di variabili
(ricodifiche, accorpamento di modalità, variabili composte), standardizzazione, …

Esempio – Indagine Multiscopo sulle famiglie, Famiglie e soggetti sociali /1

Esempio – Indagine Multiscopo sulle famiglie, Famiglie e soggetti sociali /2

Relazioni di parentela con PR	n	%
PR	6377	69,30
coniuge PR	2307	25,07
convivente PR	37	0,40
genitore PR	214	2,33
genitore partner PR	105	1,14
figlio PR nato da ultima relazione	17	0,18
coniuge figlio PR	5	0,05
nipote (figlio di fratello/sorella) PR	2	0,02
fratello/sorella PR	75	0,82
fratello/sorella partener PR	16	0,17
coniuge fratello/sorella PR	2	0,02
convivente fratello/sorella PR	2	0,02
altro parente PR	38	0,41
amico PR	5	0,05
Tot	**9202**	**100**

Relazioni di parentela con PR	n	%
PR	6377	69,30
Partner di PR	2344	25,47
Figlio (di PR o da precedente/i matrimonio/i)	17	0,18
Altro (Fratello/sorella di PR/partner di PR; nipote, altre relazioni)	464	5,04
Tot	**9202**	**100**

Tipo famiglia	n	%
Single	2556	40,08
Coppia coniugata	2836	44,47
Coppia non coniugata	39	0,61
Altro	946	14,83
Tot	**6377**	**100**

Tipo famiglia	n	%
persona sola	2556	27,78
genitore con figli non celibi o nubili	81	0,88
insieme di parenti	181	1,97
persone non parenti	10	0,11
coppia coniugata senza figli, senza isolati	3882	42,19
coppia non coniugata senza figli, senza isolati	69	0,75
coppia coniugata con figli, senza isolati	1153	12,53
coppia non coniugata con figli, senza isolati	6	0,07
monogenitore maschio separato di fatto, senza isolati	6	0,07
monogenitore maschio divorziato, senza isolati	4	0,04
monogenitore maschio vedovo, senza isolati	72	0,78
monogenitore femmina nubile, senza isolati	9	0,10
monogenitore femmina separata di fatto, senza isolati	20	0,22
monogenitore femmina separata legalmente, senza isolati	6	0,07
monogenitore femmina divorziata, senza isolati	7	0,08
monogenitore femmina vedova, senza isolati	372	4,04
coppia coniugata senza figli, con isolati	263	2,86
coppia non coniugata senza figli, con isolati	10	0,11
coppia coniugata con figli, con isolati	235	2,55
coppia non coniugata con figli, con isolati	11	0,12
monogenitore maschio separato di fatto, con isolati	2	0,02
monogenitore maschio separato legalmente, con isolati	1	0,01
monogenitore maschio divorziato, con isolati	3	0,03
monogenitore maschio vedovo, con isolati	3	0,03
monogenitore femmina nubile, con isolati	5	0,05
monogenitore femmina separata di fatto, con isolati	15	0,16
monogenitore femmina separata legalmente, con isolati	11	0,12
monogenitore femmina divorziata, con isolati	6	0,07
monogenitore femmina vedova, con isolati	35	0,38
a due generazioni, senza isolati	147	1,60
a due generazioni, con isolati	2	0,02
di tipo fraterno, con isolati	13	0,14
binucleare di altro tipo, con isolati	3	0,03
tre o più nuclei, senza isolati	2	0,02
tre o più nuclei, con isolati	1	0,01
Tot	**9202**	**100**

Individui età 65+ (**dati non pesati**)

‘Bisogni’ informativi diversi

(https://conjointly.com/kb/, William M.K. Trochim, Cornell University)

- **Descrittivi**

  descrizione dell’esistente (variabile per variabile)

- **Relazionale**

  indagare le relazioni tra variabili

- **Causale**

  determinare se una (o più) variabile(i) causa un (o più) risultato(i)

  attenzione: associazione non significa relazione causale!

Usualmente sono *cumulati*

(analisi causale certamente la più complicata)

Analisi dei dati e scala di misura delle variabili

- Variabili qualitative (scala nominale o ordinale)
  - variabili categoriali

- Variabili quantitative (scala intervallo o rapporto)
  - a volte variabili ordinali trattate come quantitative se scala modalità adeguate
  - variabili discrete e continue

n.b.: metodi/tecniche diverse a seconda del tipo di dati