

Il trattamento delle mancate risposte nelle indagini statistiche: l'approccio dell'imputazione

Orietta Luzi

20 Novembre 2009 - Università degli studi di Napoli Parthenope

Outline

- Origine dei valori mancanti
- Tipi di valori mancanti: patterns e meccanismi
- Metodi per il trattamento dei dati mancanti
- Imputazione

Mancata Risposta

Uno dei motivi principali dell'incompletezza dell'informazione è **la mancata risposta**.

- **Mancata risposta totale (MRT):**
nessuna informazione richiesta dall'indagine è disponibile per una data unità
- **Mancata risposta parziale (MRP):**
non sono fornite le risposte ad alcuni quesiti
- **Casi intermedi**
intere sezioni del questionario mancanti, intero questionario mancante ma presenza di informazioni storiche da altre indagini, ...

Mancata Risposta Totale

Cause:

- unità non rilevata
- rifiuto
- ...

generalmente (ma non sempre) viene trattata aggiustando i pesi campionari dei rispondenti

Mancata Risposta Parziale

Cause:

- risposta non conosciuta
- quesito ambiguo (difetto del questionario)
- "attrito" lungo il questionario
- ...

generalmente (ma non sempre) viene trattata imputando i valori mancanti

Valori mancanti derivanti dalla fase di editing

I valori mancanti possono derivare dalla cancellazione di valori ritenuti erronei nella fase di *editing*. Questo succede specialmente nella localizzazione di errori casuali, quando cioè non si conosce il meccanismo che ha generato l'errore e quindi il valore "vero".

Ad esempio, se un vincolo di quadratura tra le diverse voci di spesa di un'impresa è violato e la procedura di localizzazione degli errori localizza come errata una particolare spesa, il valore di quest'ultima è cancellato.

Integrazione di fonti diverse

Un caso particolare di informazione incompleta si ha quando si integrano i dati di due indagini (fonti) diverse **A** e **B** e non tutte le variabili sono rilevate in entrambe le indagini.

X : variabili rilevate solo in **A**

Y : variabili rilevate solo in **B**

Z : variabili rilevate in **A e B**

Siamo interessati alla distribuzione congiunta di X, Y, Z

Caratteristiche della mancata risposta

I due aspetti principali della mancata risposta sono:

- **Il meccanismo**

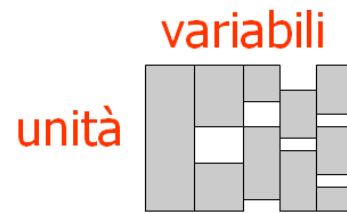
Elemento probabilistico che *genera* le mancate risposte. La comprensione del meccanismo di mancata risposta è essenziale al fine di avere inferenze valide.

- **Il pattern**

Struttura della mancata risposta sulle variabili oggetto di indagine. Può dipendere dal meccanismo.

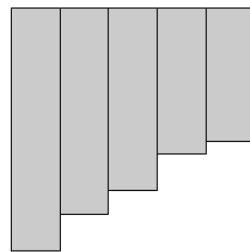
Patterns di mancata risposta

- Pattern generale

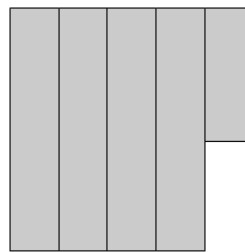


- Pattern speciali

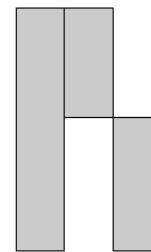
monotono



univariato



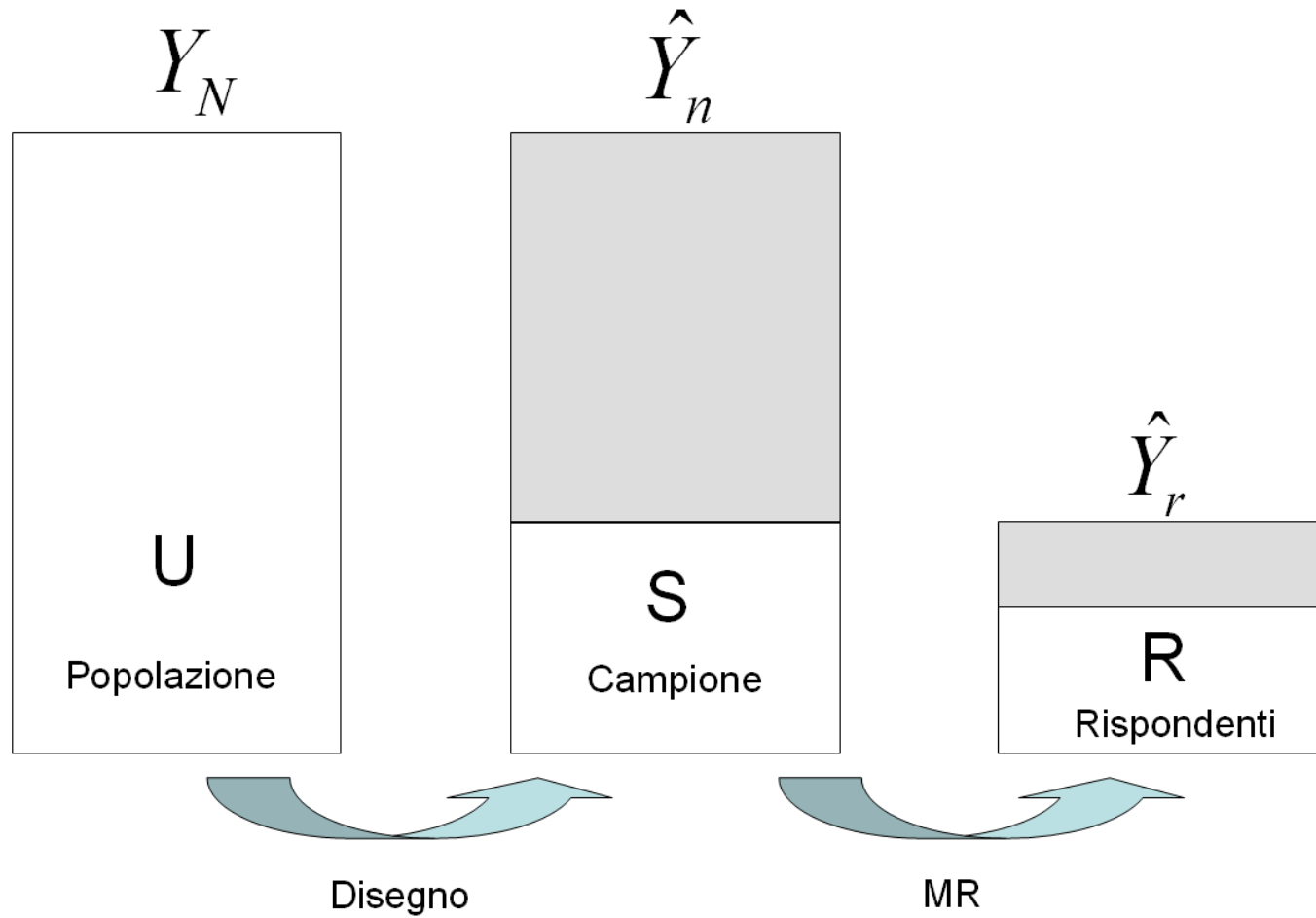
file matching



Quasi Randomization

- Un approccio all'inferenza in presenza di dati incompleti consiste nel considerare la mancata risposta come una "seconda fase" del piano di campionamento. Tuttavia il meccanismo di mancata risposta non è controllato dal ricercatore.
- La validità delle inferenze basate sui rispondenti dipende dalla validità delle assunzioni fatte sul meccanismo di mancata risposta.

Meccanismi aleatori:



Esempio SRS

Y = variabile target in una popolazione di N unità;

\widehat{Y}_n media campionaria (stimatore della media nella popolazione) su un campione casuale semplice di n unità:

$$\widehat{Y} = \frac{\sum_{i=1}^n y_i}{n}$$

Se ci sono solo r rispondenti lo stimatore diventa \widehat{Y}_r calcolato su r unità anzichè n .

Lo stimatore \widehat{Y}_r è corretto?

E' possibile stimarne la precisione?

Dipende dal *meccanismo di mancata risposta*.

Esempio SRS (Cntd)

Se i rispondenti possono essere considerati un campione casuale semplice delle unità incluse nel campione lo stimatore \hat{Y}_r è corretto. E' come se fosse uno stimatore 'espansione' basato su r unità invece di n . Ci sarebbe solo una perdita di precisione dello stimatore. In effetti la varianza aumenterebbe di un fattore $\sim n/r$.

In questo caso si dice che il meccanismo di mancata risposta è *completamente casuale (MCAR - Missing Completely at Random)*

Meccanismi di mancata risposta

Notazioni

$\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ matrice $n \times p$ dei dati

\mathbf{Y}_{obs} dati osservati

\mathbf{Y}_{mis} dati mancanti

\mathbf{M} matrice $n \times p$ degli indicatori di MR:

$M_{ij} = 0$ se la variabile Y_j è osservata sulla i -esima unità

$M_{ij} = 1$ altrimenti

$P(\mathbf{M}|\mathbf{Y})$ distribuzione della MR condizionata ai dati

$P(\mathbf{Y})$ distribuzione (densità) dei dati

Meccanismi di mancata risposta

- **MCAR** (Missing Completely At Random)

$$P(\mathbf{M}|\mathbf{Y}) = P(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{M})$$

- **MAR** (Missing At Random)

$$P(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{M}|\mathbf{Y}_{obs})$$

- **MNAR** (NMAR) (Missing Not At Random)

$P(\mathbf{M}|\mathbf{Y})$ non è semplificabile.

in parole ..

MCAR La probabilità che Y_{mis} sia mancante non dipende da Y (i rispondenti sono un sotto-campione casuale semplice del campione iniziale).

MAR La probabilità che Y_{mis} sia mancante dipende solo da Y_{obs} (non dipende dai valori mancanti)

NMAR La probabilità che Y_{mis} sia mancante dipende da Y_{mis} anche condizionatamente a Y_{obs}

Esempio meccanismi di MR

X = classe di addetti classi: 1-9; 10-49; 50-100; > 100

Y = fatturato

Supponiamo che X sia sempre nota, ma che la variabile Y abbia alcuni valori mancanti. Se la probabilità di mancata risposta $P(M)$ è indipendente da Y , allora il meccanismo è MCAR. Se, *in ogni classe di addetti*, la probabilità di MR è costante, ma classi diverse hanno probabilità diverse, allora il meccanismo è MAR. Infine, se anche entro ogni classe di addetti, la probabilità di non risposta dipende dal fatturato (Y), allora il meccanismo è NMAR.

Se vogliamo stimare la media (o il totale) del fatturato Y , nei primi due casi possiamo ottenere stimatori corretti, nel caso NMAR no.

Esempio meccanismi di MR (cntd)

n : numerosità campione

n_j : numerosità classe di addetti j ($j = 1, \dots, 4$)

r : numero rispondenti

r_j : numero rispondenti nella classe di addetti j

y_i : fatturato dell' i -esima impresa

Siano:

$$\widehat{Y}_n = \frac{1}{n} \sum_{i=1}^n y_i \quad \widehat{Y}_r = \frac{1}{r} \sum_{i=1}^r y_i \quad \widehat{Y}_{r_j} = \frac{1}{r_j} \sum_{i=1}^{r_j} y_i$$

la media campionaria sui dati completi, la media campionaria sui rispondenti, la media campionaria sui rispondenti nella classe j rispettivamente

Esempio meccanismi di MR (cntd)

Stimatori:

- **DATI COMPLETI:** \widehat{Y}_n
- **MCAR:** \widehat{Y}_r
- **MAR:** $\frac{1}{n} \sum n_j \widehat{Y}_{r_j}$ (preferibile anche nel caso MCAR)
- **NMAR:** ?

Aggiustamento per MR

Nel caso MAR abbiamo "riponderato" le unità in modo diverso nelle diverse celle (classe di addetti) per ridurre la distorsione. L'idea è simile a quella che si usa nel campionamento (randomization): Se la probabilità di inclusione dell'unità i nel campione è π_i , ogni l'unità rappresenta $w_i = \pi_i^{-1}$ unità nella popolazione.

w_i = peso campionario.

I pesi campionari sono noti perchè sono determinati dal disegno (inversi delle prob. di inclusione). La mancata risposta è un ulteriore processo di selezione in cui, però, le probabilità sono in generale sconosciute (quasi randomization)

Aggiustamento per MR (cntd)

Se conoscessimo le probabilità di risposta per ogni unità i campionata, p_{ri} , potremmo calcolare la probabilità di osservare l'unità i della popolazione come:

$$\tilde{\pi}_i = \Pr (i \text{ campionata e } i \text{ risponde }) = \Pr(i \text{ campionata}) \times \Pr(i \text{ risponde} \mid i \text{ campionata}) = \pi_i \times p_{ri}.$$

In conclusione, il peso "aggiustato" per l'unità i sarebbe:

$$\tilde{\pi}_i^{-1} = \frac{1}{\pi_i \times p_{ri}}$$

.

Aggiustamento per MR (cntd)

In generale non si conoscono le probabilità di non risposta. Dunque in pratica si cerca di stimarle usando le informazioni ausiliarie disponibili. Nell'esempio della stima della media del fatturato abbiamo ipotizzato che la prob. dipenda solo dalla classe di addetti, e la abbiamo stimata con il tasso di risposta entro ciascuna classe:

$$p_{ri} = r_j/n_j$$

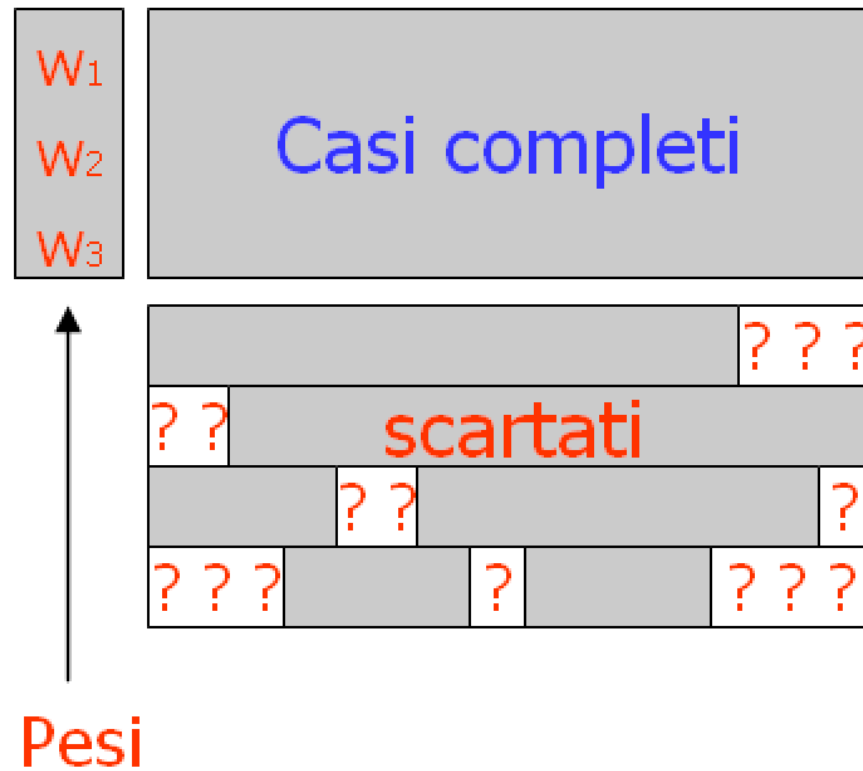
dove j è la classe di appartenenza dell'unità i . Se i pesi campionari sono tutti uguali (w), questa ipotesi ci conduce allo stimatore:

$$\left(\sum_i^n w\right)^{-1} \sum_j^4 \sum_i^{r_j} w \frac{n_j}{r_j} y_i = \frac{1}{n} \sum n_j \widehat{Y}_{r_j}$$

Analisi dati completi

- Nell'esempio visto, nel caso MAR, abbiamo usato la variabile ausiliaria "classe di addetti" per compensare la differenza tra rispondenti e non rispondenti in termini della variabile target (fatturato).
- se avessimo stimato il fatturato medio calcolando la media su tutta la popolazione anzichè separatamente per le diverse classi avremmo ottenuto una stima distorta. Ad es., se le aziende con fatturato minore rispondono meno, otteniamo una sovrastima del fatturato, perchè le aziende "rispondenti" sono quelle con fatturato maggiore.
- in generale, l'analisi che si basa solo sulle unità in cui tutte le variabili sono osservate si dice analisi dei **Casi Completi** (CC).

Analisi CC



Vantaggi analisi CC

- semplice (metodi e software standard)
- non usa dati artificiali (imputati)
- può essere appropriata in caso di pochi dati mancanti e con meccanismo MCAR
- comparabilità delle statistiche univariate, infatti esse sono calcolate su un campione con lo stesso numero di casi.

Svantaggi analisi CC

- nel caso non MCAR le stime possono essere distorte
- non si usa tutta l'informazione disponibile, dunque anche in caso MCAR, le stime sono meno precise. La perdita di precisione può essere significativa se il tasso di mancata risposta è elevato.

Analisi sui dati completi: Esempio

Sia (X, Y) una normale bivariata. Sia disponibile un campione di n osservazioni. X è sempre osservata, Y ha $n - r$ mancate risposte (MCAR). Stimiamo μ_y , il valore atteso di Y . Consideriamo lo stimatore $\bar{Y}_r = \frac{1}{r} \sum_{i=1}^r y_i$ ottenuto come media degli r valori osservati di Y .

Se i dati fossero stati completi, avremmo usato lo stimatore $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n y_i$. Il rapporto delle varianze dei due stimatori è n/r , dunque, se il tasso di MR è del 50%, la varianza dello stimatore raddoppia.

Analisi sui dati completi: Esempio

Se invece usiamo tutta l'informazione disponibile (cioè anche i valori della variabile X), possiamo ottenere uno stimatore più efficiente di \bar{Y}_r . Ad esempio, lo stimatore MLE di μ_y è:

$$\hat{\mu}_{yML} = \hat{\mu}_y = \bar{Y}_r + \hat{\beta}_r(\bar{X}_n - \bar{X}_r)$$

dove β_r è la stima del coefficiente di regressione di Y su X .

In questo caso abbiamo:

$$Var(\hat{\mu}_{yML})/Var(\bar{Y}_r) \approx 1 - \frac{n-r}{n}\rho^2$$

Conclusione

Quindi se $\rho^2 \rightarrow 0$ le due varianze sono uguali, mentre se $\rho^2 \rightarrow 1$ il rapporto tende a r/n , cioè si ha lo stesso guadagno di efficienza che si ha nel passare dallo stimatore \bar{Y}_r allo stimatore \bar{Y}_n .

L'esempio mostra come l'utilizzo di informazione ausiliaria può migliorare molto la precisione dello stimatore se l'informazione è rilevante ai fini della stima di interesse (nell'esempio se ρ è vicina a 1).

Analisi sui dati completi: distorsione

Consideriamo la media totale della popolazione μ . Questa può essere scritta, scomponendola nella media della sottopopolazione dei rispondenti μ_r e quella dei non rispondenti μ_{nr} , come

$$\mu = \pi_r \mu_r + (1 - \pi_r) \mu_{nr}$$

e quindi la distorsione della media calcolata sui rispondenti è

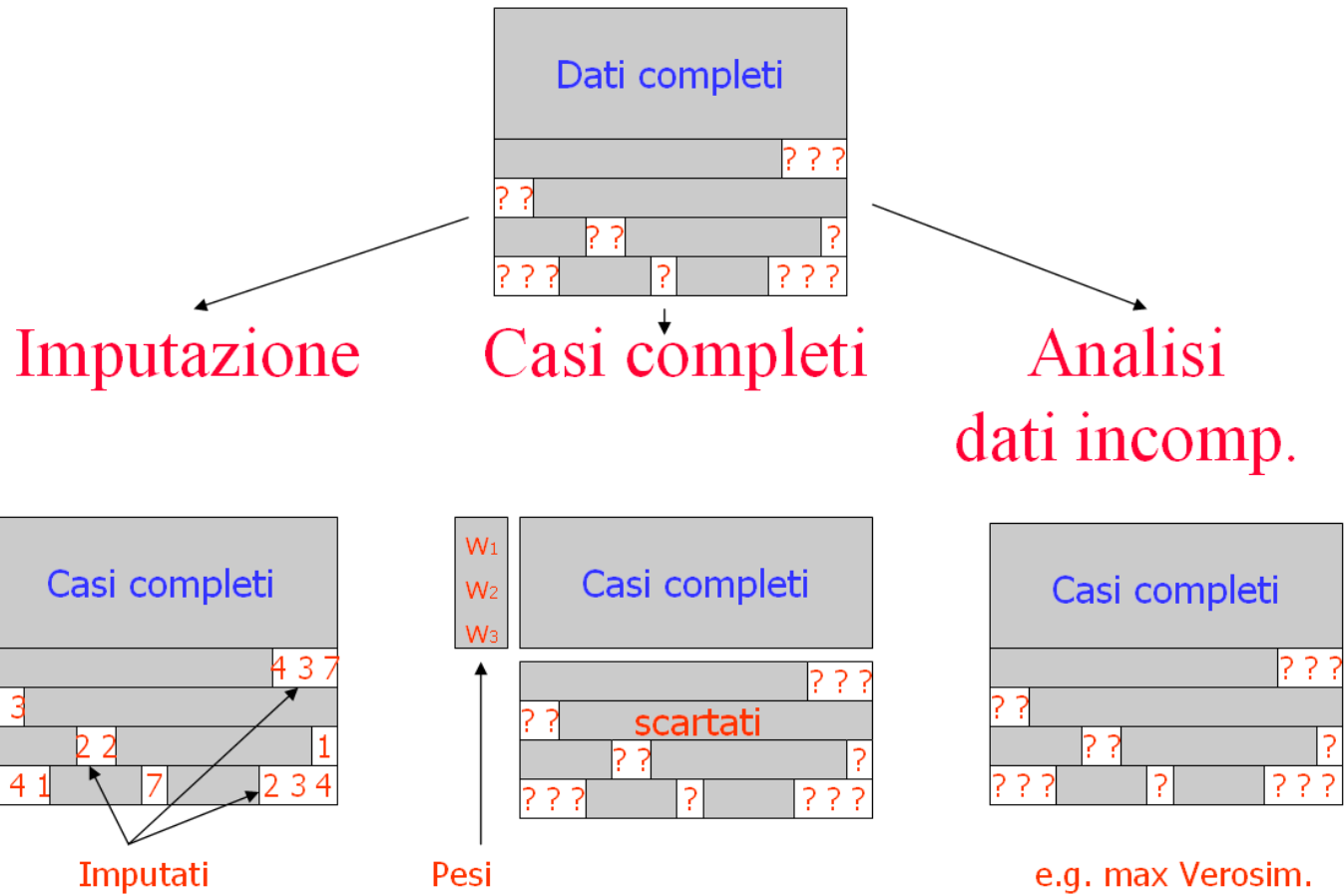
$$\mu_r - \mu = (1 - \pi_r)(\mu_r - \mu_{nr})$$

dove π_r è la proporzione dei rispondenti. Dunque in presenza di MR, la distorsione è nulla solo nel caso MCAR ($\mu_r = \mu_{nr}$)

Alternative

- analisi dati "disponibili" (AC: available cases)
- completamento del dataset mediante dati "artificiali" (imputazione)
- analisi basata sui dati incompleti (e.g. algoritmo EM)

Strategie



Analisi dati disponibili

- Consiste nell'usare, per ogni stima di interesse, tutte le unità in cui sono disponibili le variabili necessarie per la data stima. Ad esempio, per stimare la correlazione tra le variabili X e Y , si possono usare tutte le unità in cui entrambe queste variabili sono osservate.
- Il metodo è da evitare perchè può produrre stime non coerenti: ad esempio, se per ogni coppia di variabili si stima la covarianza indipendentemente, si può arrivare ad una matrice di varianze e covarianze che non è definita positiva.

Analisi su tutti i dati

Di questa classe di metodi fanno parte i metodi di stima di un modello con dati incompleti e l'imputazione.

- **Imputazione**
 - Viene ricostruito un dataset rettangolare completo (senza "buchi"), rimpiazzando i valori mancanti con valori "artificiali" plausibili.
- **Stima con dati incompleti**
 - Sotto ipotesi di un modello parametrico esplicito è possibile, in alcuni casi stimare (ad esempio mediante MLE) i parametri sulla base dei dati incompleti. Un metodo popolare è l' algoritmo EM.

Imputazione

I valori mancanti vengono sostituiti con valori “plausibili” e poi si possono utilizzare le analisi standard applicandole al dataset completato.

Vantaggi. Praticità e semplicità con cui dopo si possono condurre le analisi.

Svantaggi. Si introduce una ulteriore fonte di variabilità, quindi quando si analizza il dataset completo bisogna tenere in considerazione che i dati imputati non sono realmente osservati. Inoltre la valutazione di un metodo di imputazione può solo essere fatta relativamente ad una particolare stima di interesse.

Metodi di Imputazione delle MRP

- **non parametrici** (hot-deck, alberi di regressione...)
- **parametrici** (Regressione, EM,...)
- **misti** (PMM,...)

Metodi semplici (1)

- *Imputazione con media.* I valori mancanti di una variabile vengono imputati con la media della variabile sulle unità rispondenti. Spesso la popolazione è partizionata in strati (*celle di imputazione*) e per ogni unità con valore mancante la media è calcolata all'interno dello strato di appartenenza dell'unità. L'imputazione per media nelle celle è analoga alla riponderazione all'interno delle celle.
- *Imputazione tramite regressione.* Si imputa con il valore di regressione rispetto a una o più variabili ausiliarie. L'imputazione per media è un caso particolare della regressione, dove le variabili esplicative sono le variabili dummy che indicano le celle di imputazione. Alla media predittiva si può eventualmente aggiungere un residuo.

Metodi semplici (2)

- *Imputazione hot-deck*. Si sostituisce un valore preso da un rispondente “simile” della stessa indagine. Questo è un metodo molto usato nella statistica ufficiale.
- *Imputazione cold-deck*. Analogo al precedente, ma i valori imputati sono presi da una diversa indagine (o da una diversa occasione della stessa indagine).

Imputazione con la media

Questo metodo consiste nell'imputare i valori mancanti con la media calcolata sui valori osservati. E' semplice e può essere soddisfacente se le stime di interesse sono medie o totali e se la mancata risposta è MCAR (rispondenti e non rispondenti hanno le stesse proprietà). Sia \bar{y}_r la media degli r rispondenti su un campione di n unità. allora la stima della media di y dopo l'imputazione per media diventa:

$$\frac{1}{n} \left(\sum_i^r y_i + \sum_{r+1}^n \bar{y}_r \right) = \frac{1}{n} (r\bar{y}_r + (n - r)\bar{y}_r) = \bar{y}_r$$

e poichè il meccanismo è MCAR \bar{y}_r è una stima non distorta di \bar{y}_n

Imputazione con la media (cntd)

Se invece devo stimare qualche parametro non lineare, come per esempio la varianza, questo non è soddisfacente, infatti la stima della varianza sui dati imputati è

$$s_r(n_r - 1)/(n - 1).$$

Se s_r è una stima consistente della varianza della popolazione, la stima che si ottiene con l'imputazione per media è distorta (sottostimata) di un fattore pari a $(n_r - 1)/(n - 1)$. Infatti la distribuzione è stata "appiattita" sulla media.

Imputazione con medie condizionate

Lo sfruttamento di informazione ausiliaria può migliorare la precisione delle stime basate su dati imputati. In presenza di covariate con forte potere esplicativo si può imputare la media condizionata. A seconda che le var. ausiliarie siano categoriche o quantitative otteniamo rispettivamente:

- Imputazione con media per strati
- Imputazione con regressione senza residuo

Imputazione con media per strati (1)

Se abbiamo J strati, e \bar{y}_{jr} è la media della variabile Y sugli r_j rispondenti nello strato j , la stima della media di Y è:

$$\bar{y}_{cm} = \frac{1}{n} \sum_{j=1}^J \left(\sum_{i=1}^{r_j} y_{ij} + \sum_{i=r_j+1}^{n_j} \bar{y}_{jr} \right) = \frac{1}{n} \sum_{j=1}^J n_j \bar{y}_{jr}$$

Se (in media) $\bar{y}_{jr} = \bar{y}_j$ (MCAR) tale stimatore è corretto. Si noti l'analogia con il campionamento stratificato.

Imputazione con media per strati (2)

Il guadagno che si ha confrontando tale stimatore con quello che si avrebbe nel caso di imputazione con media (non condizionata) si può ricavare dalle formule note per la teoria dei campioni:

$$V(\bar{Y}) - V(\bar{Y}_{cm}) = n^2 \frac{1 - r/n}{r} \sum_{j=1}^J \frac{n_j}{n} (\bar{Y}_j - \bar{Y})^2$$

Questa differenza è tanto più grande quanto più sono differenti le medie degli strati dalla media generale.

Imputazione con media all'interno di strati (2)

Come formare gli strati?

Non esiste una ricetta universale. Tuttavia devono essere soddisfatti i seguenti criteri:

1. mancata risposta (approssimativamente) MCAR all'interno degli strati
2. massima omogeneità **entro** gli strati e diversità **tra** gli strati

Dunque è importante scegliere una variabile di stratificazione con forte potere esplicativo relativamente alla variabile di interesse.

Imputazione con regressione senza residuo

In questo caso si sostituisce al valore mancante il valore stimato con la regressione. Caso semplice: 2 variabili (X, Y):

$$y_i = \hat{\alpha} + \hat{\beta}x_i$$

dove $\hat{\alpha}$ e $\hat{\beta}$ sono le stime dei minimi quadrati calcolate sulle unità in cui sono osservate entrambe le variabili.

Vale la pena notare che nel caso in cui le variabili sono “dummy” che indicano gli strati, si ottiene l’imputazione con media all’interno degli strati.

Imputazione con regressione senza residuo

In entrambi i metodi si imputano i valori mancanti con un valore medio (o più valori medi) e quindi è intuitivo che, sebbene lo stimatore per la media sia corretto, la varianza viene sottostimata (come mostrato nel caso per la media). In particolare, la sottostima della varianza di Y nei dati imputati è dovuta al fatto che non viene stimata nel modello la v varianza residua.

Imputazione per regressione con residuo

Per mantenere la variabilità si può aggiungere al valore medio imputato con la regressione un valore per il residuo ϵ estraendo da una normale con media zero e varianza pari alla varianza residua stimata dai dati completamente osservati (sia X che Y).

Le stime basate su dataset imputato con questo metodo sono consistenti per tutti i parametri della popolazione. In effetti i valori imputati sono generati dalla distribuzione "predittiva" (termine improprio) dei missing.

I casi con pattern di mancata risposta più complicati verranno trattati in seguito.

Ora ci occuperemo di metodi di imputazione che si basano su "un modello implicito" dei dati.

Imputazione tramite hot-deck (donatore)

I metodi *hot-deck* consistono nel prelevare il valore (o i valori) da imputare da un'altra unità della stessa indagine (donatore) in cui questi valori sono osservati.

L'origine del termine hot-deck deriva dal fatto che quando il computer era gestito tramite schede, hot-deck era appunto una scheda (record) calda, appena passata nel computer, e quindi appartenente alla stessa indagine, in contrapposizione al cold-deck.

Questa classe di metodi include il **donatore casuale** e il **donatore di minima distanza**. Esistono diverse varianti: ad es., i valori da imputare possono esser presi dallo stesso donatore (*imputazione congiunta*) o da donatori diversi (*imputazione sequenziale*).

Vantaggi imputazione hot-deck casuale

- Le stime dei parametri relativi alle distribuzioni univariate sono approssimativamente corrette anche per parametri non lineari, come per esempio la varianza.
- il donatore è spesso usato nella pratica delle indagini complesse perchè non richiede particolari ipotesi sulla distribuzione dei dati (applicabile anche nel caso di distribuzioni semicontinue)

Svantaggi imputazione hot-deck casuale

L' imputazione tende ad attenuare le associazioni tra le variabili non osservate e quelle osservate se i valori di quest'ultime non vengono tenuti in considerazione nel processo. Si può attenuare il problema sfruttando le covariate:

- Donatore casuale all'interno di strati
- Donatore di distanza minima

Per quanto riguarda la formazione delle celle, valgono le considerazioni precedenti. Analizziamo il donatore di distanza minima (NND)

Imputazione donatore di distanza minima

In questo approccio i valori mancanti di una osservazione vengono sostituiti con i valori dell'osservazione più simile. La similitudine è definita tramite una distanza calcolata sulla base di opportune covariate dette **variabili di matching**. Distanze comuni:

- $l_p : d_p(i, j) = \sum_k |x_{ik} - x_{jk}|^p$ ($p = 1$ Manhattan, $p = 2$ euclidea)
- deviazione massima $d(i, j) = \max_k |x_{ik} - x_{jk}|$
- Mahalanobis $d(i, j) = (x_i - x_j)^T S_{xx}^{-1} (x_i - x_j)$

Osservazioni e problemi

- Il metodo è 'asintotico'(richiede un elevato numero di donatori)
- Scelta delle variabili da includere nella funzione di distanza
- Ponderazione dei diversi contributi delle variabili
- Standardizzazione delle variabili
- Uso di un unico donatore per imputare più variabili per osservazione
- Imputazione condizionata ai vincoli logici

Metodi più complessi

Finora abbiamo esaminato metodi semplici applicabili quando il pattern di mancata risposta è particolarmente semplice (ad esempio una variabile con missing e una covariata sempre osservata). Spesso i dati hanno un aspetto irregolare e i "buchi" possono avere le configurazioni più svariate. Sono quindi necessari metodi che generalizzano quelli trattati fin qui.

Donatore di minima distanza

Nel caso del donatore di minima distanza la generalizzazione è semplice: si utilizzano come donatori i record completi e **si calcola la distanza sulla base delle sole variabili che sono osservate nel record da imputare di turno**. Ad esempio, se la distanza scelta è la variabile euclidea nello spazio delle variabili X_1, X_2, X_3, X_4 , e nell'unità i sono presenti solo le variabili di matching X_1 e X_3 , solo queste due saranno usate per calcolare la distanza rispetto a tutti i potenziali donatori j :

$$d_2(i, j) = \text{radq}[(x_{i1} - x_{j1})^2 + (x_{i3} - x_{j3})^2]$$

Per i metodi parametrici è richiesta qualche nozione supplementare.

Stima tramite modello - Metodi parametrici (1)

In questo ambito, i metodi di imputazione si basano sull'assunzione di un modello parametrico esplicito (e parsimonioso) che sintetizza le ipotesi dello statistico. Il modello può essere usato a due livelli:

- **Regressione**: alcune variabili sono considerate variabili *esplicative* ed altre variabili *risposta*
- **Stima della distribuzione congiunta**: si modella la distrib. congiunta (densità nel caso di variabili continue) di tutte le variabili di analisi.

In ogni caso lo scopo è quello di sostituire i valori mancanti con valori "appropriati" che possono essere dedotti usando il modello assunto.

Metodi parametrici (2)

I metodi basati sulla regressione sono utili quando il pattern di mancata risposta è univariato, o una generalizzazione del pattern univariato, cioè quando l'insieme delle variabili può essere splittato in un insieme di variabili sempre osservate (X) e un insieme di variabili missing su un certo sottoinsieme comune (Y) di unità non rispondenti. In tal caso l'imputazione è effettuata mediante la regressione di Y su X .

Nel caso di pattern generali è più utile stimare la distribuzione congiunta di tutte le variabili.

Stime sui soli dati completi

Come per i casi semplici visti in precedenza, si potrebbero effettuare le stime dei parametri θ della distribuzione congiunta utilizzando solo i record completamente osservati. Come si è visto però, questo è sub-ottimale per due motivi:

1. maggiore varianza delle stime
2. rischio di distorsione se il meccanismo di MR non è MCAR

Stime con dati incompleti

Dunque vogliamo usare tutta l'informazione disponibile: dati completi e dati incompleti. Consideriamo le stime dei parametri ottenute massimizzando la funzione di verosimiglianza (MLE). Ma qual è la funzione di verosimiglianza da massimizzare in quando i dati sono incompleti? Quali sono le distribuzioni di probabilità in gioco? Abbiamo due meccanismi aleatori:

$f(\mathbf{Y}|\theta)$: distribuzione dei dati completi

$f(\mathbf{M}|\mathbf{Y}, \psi)$: distribuzione della mancata risposta (\mathbf{M}) condizionata ai dati

θ e ψ sono insiemi di parametri che supponiamo *distinti*.

Stime di ML

Siamo interessati alla stima dei parametri θ . I parametri ψ sono di "disturbo". Possiamo dire qualcosa su θ disinteressandoci dei parametri ψ e dunque *ignorando* il meccanismo di mancata risposta? La risposta è sì se il meccanismo di MR è MAR:

$$f(\mathbf{M}|\mathbf{Y}, \psi) = f(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}; \psi) = f(\mathbf{M}|\mathbf{Y}_{obs}; \psi)$$

cioè se la probabilità di non risposta, ***condizionatamente ai dati osservati*** non dipende dai valori non osservati.

Stime sotto hp MAR

Usando la formula di Bayes è facile verificare che la condizione MAR equivale al fatto che la distribuzione dei dati mancanti condizionata a quelli osservati sia la stessa per i rispondenti ($M = 0$) e i non rispondenti ($M = 1$):

$$f(\mathbf{Y}_{mis}|\mathbf{M}, \mathbf{Y}_{obs}) = \frac{f(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})}{f(\mathbf{M}|\mathbf{Y}_{obs})} = f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$$

Verosimiglianza osservata

Nel caso MAR la distribuzione congiunta di \mathbf{Y}_{obs} e \mathbf{M} può fattorizzarsi nel prodotto di una funzione che dipende solo dai parametri θ e una che dipende solo dai parametri ψ :

$$\begin{aligned} f(\mathbf{Y}_{obs}, \mathbf{M} | \theta, \psi) &= \int f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} | \theta) f(\mathbf{M} | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi) d\mathbf{Y}_{mis} = \\ &= f(\mathbf{M} | \mathbf{Y}_{obs}, \psi) \int f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} | \theta) d\mathbf{Y}_{mis} = \\ &= f(\mathbf{M} | \mathbf{Y}_{obs}, \psi) f(\mathbf{Y}_{obs} | \theta) \propto L(\theta | \mathbf{Y}_{obs}) \\ &L(\theta | \mathbf{Y}_{obs}) \quad \text{verosimiglianza osservata} \end{aligned}$$

Verosimiglianza osservata e algoritmo EM

Le stime di ML possono dunque essere ottenute massimizzando la verosimiglianza osservata $L(\theta|Y_{obs})$.

L'algoritmo **EM (Expectation-Maximization)** consente di effettuare stime di ML in presenza di dati mancanti attraverso tecniche standard usate per dati completi. L'algoritmo formalizza la procedura (in generale errata) che consiste nella applicazione iterata dei due passi seguenti:

1. rimpiazzare i valori mancanti con valori stimati in base alle stime correnti dei parametri (**E-step**)
2. ottenere nuove stime dal data-set completo così ottenuto (**M-step**)

Imputazione via EM

Pertanto, nel contesto dell'imputazione, l'EM è utilizzato come passo intermedio per stimare la distribuzione da cui si assume siano generati i dati. Il passo successivo consiste nel ricavare, dalle stime dei parametri della distribuzione congiunta dei dati, le stime (di ML) dei parametri delle distribuzioni condizionate corrispondenti ai vari pattern di MR.

Imputazione con o senza residuo

Stimata la distribuzione condizionata $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$ corrispondente ad un dato pattern di MR, l'imputazione può essere effettuata in due modi:

1. utilizzando le medie condizionate $E(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$
2. introducendo un residuo casuale, cioè generando dalla distribuzione $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$.

Con il primo metodo le stime di quantità lineari nei dati saranno più precise (minore varianza). Se tuttavia si è interessati a preservare le caratteristiche distribuzionali il secondo metodo è preferibile.

Modelli

Il modello più utilizzato per variabili quantitative è il modello normale. Per la distribuzione normale infatti sono disponibili metodi e software che consentono di effettuare inferenze in presenza di dati incompleti.

Spesso l'assunzione di normalità richiede che le variabili siano preliminarmente trasformate (e.g. trasformazioni logaritmiche).

In taluni casi tuttavia, l'ipotesi di normalità risulta inadeguata \Rightarrow è necessario ricorrere ad altri approcci.

Predictive Mean Matching

Un metodo che potrebbe essere più robusto rispetto all'assunzione di normalità è quello del Predictive Mean Matching (PMM). In questo approccio il modello viene usato solo ad uno stadio intermedio per calcolare i valori attesi dei missing condizionatamente ai valori osservati. Le medie condizionate però non sono imputate direttamente. Piuttosto, tramite le medie condizionate è definita una funzione di similarità da utilizzare nell'imputazione NND.

La funzione di similarità è funzione delle covariate osservate. Il metodo dipende ovviamente dal modello assunto.

Il PMM può essere utile tuttavia quando si rende necessario un approccio "parsimonioso" (ad esempio per esiguità del numero di osservazioni), ma nel contempo è difficile trovare un modello totalmente adeguato per i dati.

PMM: una variabile risposta

Nel caso di una variabile risposta Y affetta da mancata risposta e un insieme di p covariate (X_1, \dots, X_p) sempre osservate, il PMM consiste nel:

- determinare, per ogni osservazione u_i , il valore atteso condizionato $y_i^* = E(Y|x_{i1}, x_{i2}, \dots, x_{ip})$ calcolato con i parametri stimati dalla regressione di Y su X_1, X_2, \dots, X_p
- per ogni unità u_i con Y mancante, imputare il valore y_i prelevandolo dal donatore u_j , la cui media predittiva y_j^* è la più prossima a y_i^*

Nel caso $p = 1$, il PMM coincide con il NND.

PMM: caso generale

Nel caso generale di pattern di mancata risposta arbitrario, il metodo consiste nei seguenti passi:

- stima dei parametri della distribuzione congiunta dei dati mediante algoritmo EM
- per ogni record incompleto u_i , calcolo della media condizionata (in generale multi-dimensionale) $y^* = E(\mathbf{Y}_{mis,i} | \mathbf{y}_{obs,i})$
- per ogni record incompleto u_i , imputazione NND utilizzando la metrica di Mahalanobis basata sulla matrice di varianza e covarianza residua della regressione di \mathbf{Y}_{mis} su \mathbf{Y}_{obs} .

Il problema della stima della varianza

Il trattamento dei dati incompleti mediante l'imputazione ha il vantaggio di fornire un data-set rettangolare completo sulla base del quale possono essere effettuate inferenze usando metodi e software standard.

Tuttavia, ogni stima effettuata su data-set imputato è caratterizzata da una variabilità che è maggiore di quella che le competerebbe se tutti i dati fossero effettivamente osservati.

Ignorare il problema (cioè trattare i dati imputati come se fossero realmente osservati), porta dunque a sovrastimare la precisione delle stime (stime della varianza negativamente distorte, intervalli di confidenza troppo stretti, test non validi,...)

Approcci al problema

- Bootstrap (Efron)
- Jackknife (Rao)
- Model-assisted (Sarndal)
- Imputazione multipla (Rubin)
-

Imputazione multipla (Rubin, 1987)

In sostanza, il metodo consiste nell' eseguire più imputazioni dello stesso data-set, generando i valori da imputare dalla distribuzione predittiva dei valori mancanti condizionatamente a quelli osservati.

La variabilità risultante nelle inferenze ottenute sui diversi data-set completati dovrebbe così riflettere l'incertezza associata alla non risposta.

La metodologia, trova la sua collocazione più naturale in ambito bayesiano.

inferenza basata sull'imputazione multipla(1)

Notazioni e assunzioni:

Q = quantità di interesse nella popolazione

\hat{Q} = stimatore di Q basato su dati completi con varianza U ($V(Q) = U$)

$(\hat{Q} - Q) \approx N(0, U)$

\hat{U} = stimatore di U basato sui dati completi

supponiamo di avere ottenuto per imputazione m data-set completi e indichiamo con \hat{Q}_k la stima di Q , e con \hat{U}_k la stima della varianza di \hat{Q}_k sulla base del k -esimo data-set imputato ($k = 1, \dots, m$):

$\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_m$

$\hat{U}_1, \hat{U}_2, \dots, \hat{U}_m$

inferenza basata sull'imputazione multipla(2)

Lo stimatore di Q basato sugli m data set imputati è:

$$\hat{Q}^{(m)} = \frac{\sum_{k=1}^m \hat{Q}_k}{m}$$

La stima della varianza di $\hat{Q}^{(m)}$ è invece data da:

$$V(\hat{Q}^{(m)}) = \frac{\sum_{k=1}^m \hat{U}_k}{m} + \frac{\sum_{k=1}^m (\hat{Q}_k - \hat{Q}^{(m)})^2}{m - 1}$$

Vantaggi e Svantaggi dell'IM

- vantaggi:
 - possibilità di stimare la componente della variabilità delle stime associata alla MR
- svantaggi:
 - necessità di gestire m data-set completati
 - difficoltà di definire una procedura di imputazione adeguata in caso di disegno complesso

Approcci “euristici”

Quando la struttura dei dati è complessa ed è difficile trovare una forma parametrica esplicita per la distribuzione congiunta di tutte le variabili di interesse, una strategia molto diffusa consiste nel modellizzare separatamente diversi gruppi di variabili, oppure imputare *sequenzialmente* i dati mancanti, stimando *indipendentemente* le diverse distribuzioni condizionate.

Quest'ultimo approccio è adottato ad esempio nel software *Iveware* (*Raghunathan et al.*), che effettua l'imputazione (eventualmente multipla) fittando una sequenza di modelli di regressione (lineare, logistica, Poisson,...) e generando i dati dalle corrispondenti distribuzioni predittive.

Riferimenti base

Little, R. J. A. and Rubin, D. B. (2002). “Statistical Analysis with Missing Data” 2nd edition, Wiley.

Shafer, J.L. (1997). “Analysis of Incomplete Multivariate Data” New York: CRC Press.