Biostatistica

22 gennaio 2025

Indice

Parte I

Health economics

Capitolo 1

Introduzione

1.1 Economia sanitaria

L'economia sanitaria applica la teoria economica all'analisi del mercato sanitario, e dei suoi problemi di allocazione.
La giustificazione dell'intervento pubblico nel campo sanitario può derivare da motivi efficientistici ed equitativi:

1. mancanza di sovranità del consumatore: situazione di informazione imperfetta ed asimmetrica, interventi sanitari come experience good
2. presenza di forti esternalità positive insite nel bene salute
3. la spesa per la salute è connessa ingran parte alla contrazione di malattie, quindi ad eventi aleatori: sarebbe pertanto efficiente l'impiego di contratti di assicurazione per la copertura dei rischi. La realizzazione di un sistema privato di assicurazione potrebbe però essere compromessa da probabilità di evento non indipendenti (es epidemie), vicine all'unità (es malattie ereditarie), problemi di moral hazard (es caso del terzo pagante) ed averse selection. Altro problema di carattere etico consiste nella possibilità di creare cream skimming da parte delle agenzie assicurative
4. mancanza di una molteplicità di produttori, soprattutto se si oncentra l'attenzione in un contesto di piccole o medie dimensioni (provinciali/regionali)
5. disomogeneità del prodotto/prestazioni, anche se si considera la fornitura del medesimo servizio
6. forma di finanziamento delle prestazioni indiretta: fiscalità, non prezzi

Ovvio è (ma neanche troppo, da un punto di vista storico) che le decisioni allocative dello Stato debbano basarsi su studi comparativi delle alternative disponibili: ecco che entra in gioco la valutazione economica in sanità

1.2 Sistema sanitario e valutazione economica

Alla base della valutazione economica troviamo concetti classici dell'economia:

- **costo opportunità**: consiste nel valore dell'alternativa cui si rinuncia, impiegando le risorse in un dato modo. Nel campo sanitario esso è spesso rappresentato da outcome sanitari (quantificati in differenti unità di misura) raggiungibili mediante intestimenti in alternative.
  L'utilizzo del costo opportunità è alla base della valutazione economica

- **costo marginale**: variazione del costo totale derivante dall'aumentare di una unità la produzione di un prodotto/servizio; questa è variabile fondamentale di orientamento della scelta di produzione perché in linea di massima, l'aumento del prodotto dovrebbe essere effettuato solo quando il beneficio marginale supera il costo marginale

Un **sistema sanitario è efficiente** (Dirindin) quando:

1. le risorse vengono impiegate per prestazioni efficaci, ossia in grado di incidere beneficamente sulla patologia del paziente

2. tali prestazioni presentano un beneficio marginale uguale o maggiore rispetto a quello ottenuto con impieghi aternativi

3. si inizia nela produzione da quelle erogazioni che garantistcono benefici di outcome maggiori, passando via via a quelle minori;

4. nella produzione del servizio occorre scegliere la combinazione di fattori produttivi che minimizzi i costi

Questo garantirebbe efficienza (ossia la maggior produzione di benessere possibile), ma non necessariamente equità (ossia distribuzione del benessere secondo criteri preferiti da un punto di vista etico-politico).

1.3 Processo di valutazione economica in sanità

Il perseguimento di un sistema sanitario efficiente parte necessariamente dalle singole azioni. ogni scelta allocativa (un farmaco/intervento piuttosto che un altro) dovrebbe fondarsi su un processo di valutazione articolato nelle seguenti fasi:

1. definizione della domanda di studio e scelta delle alternative confrontabili: occorre specificare chiaramente la popolazione/su che pazienti, il trattamento da valutare, quello con cui compararlo (es pratica standard), su che pazienti, e la variabile utilizzata per giudicarne l'effetto clinico.

2. analisi dei dati di efficacia della terapia: vi sono diversi studi e metodi di analisi di efficacia

3. **scelta tecnica di valutazione economica**: gli esiti sanitari debbono essere trasformati in ouput economici.
   Vi sono diversi approcci impiegabili che dipendono dalla disponibilità dei dati

   - nell'analisi costo efficacia i benefici del trattamento sono misurati in unità naturali; il problema è che studi che adottano un differente indicatore di efficacia non possono esser confrontati

- nell'analisi costo utilità si pesano gli anni di vita mediante la qualità di vita dei pazienti. Si utilizzano questionario che riportano su una scala normalizzata la condizine di salute. Ogni anno si riceve un punteggio comreso tra 0 (per la morte) e 1 (in caso di perfetta salute); questi punteggio vengono detti QOL (quality of life): ad esempio un individuo di media salute (es qol=0.5) necessita di due anni di vita per guadagnarsi un QALY (quality adjusted life year). Mediante un confronto tra l'esito di tale indicatore per i trattati e per il gruyppo di controllo è possibile ottenere una stima della variazione della qualità della vita indotto dal trattamento
- nell'analisi costi benefici, si esprime il beneficio clinico in termini monetari, motivo per cui questo approccio è stato criticato

Per quanto riguarda i costi essi possono esser classificati in

- diretti: costi associati alla produzione del servizio sanitari (ricoveri, terapie, visite mediche) e non (assistenza domiciliare)
- indiretti: perdite di produttività del paziente e dei rispettivi famigliari

4. **scelta ottica di analisi valutativa** (es sistema sanitario pagante, società): ha impatto sui costi e benefici che si dovranno considerare. Ad esempio se fatto nell'ottica del sistema sanitario dobbiamo tenere conto degli stimendi dei medici, non delle perdite di produttività di pazienti e famigliari.

5. **svolgimento dell'analisi economica**: attualizzazione , calcolo di indicatori sintetici, e analisi di sensibilità. Si ha

   - attualizzazione di costi e benefici avviene con un tasso compreso tra il 3 e il 5
   - sia effetto che costo posono esser maggiori, minori o uguali rispetto all'alternativa comparata.
     La scelta allocativa non appare evidente se si hanno contemporaneamente maggiori benefici e maggiori costi, oppure minori benefici e minori costi.
     In questi casi (per analisi costo efficacia e costo utilità) occorre rapportare l'incremento dei costi con quello dei benefici. Se si è utilizzato un approccio costo-efficacia sarà
   - analisi di sensibilita: ve ne sono due tipi
     (a) si fanno variare i parametri utilizzati nel computo d icosti e benefici (es tasso di interesse, funzioni di distributzione di costi e benefici) per vedere come cambia il risultato finale. Facendo variare un parametro alla volta si ha una analisi univariata, più parametri contemporaneamente. I risultati sono robusti se le conclusioni non variano
     (b) si fanno best e worst case scenario: best case si considera efficacia maggiore e costi minori, worst case il contrario

1.4 Cost utility shit

Qol possono esser calcolati secondo due approcci:

- approccio *clinico*: si valuta la condizione di salute del paziente mediante questionari medici generici (SF36, EQ5D) o specifici per patologia. Gli strumenti generici hanno dalla loro il beneficio di versatilità d'uso, quelli specifici dell'approfondimento di indagine
- approccio *economico*: si chiede al paziente di valutare, in maniera diretta o meno, il proprio stato di salute (approccio che si richiama maggiormente al concetto di utilità. Vi sono tre tecniche principali:
  1. **VAS** o rating scale: si richiede al paziente di porre una X su una linea che va dalla perfetta salute alla morte per indicare quanto ci si ritiene in salute
  2. **standard gamble**: si chiede al paziente di scegliere tra un determinato stato di malattia e la partecipazione all'urna che da p probabilità di morire e $(1-p)$ di vivere in perfetta salute; qual è la p che rende indifferente la scelta.
  3. **time trade off**:viene chiesto al paziente quanti anni di vita nello stato di malattia sarebbe dispostio a rinunciare per vivere in salute. L'utilità è pari a

$$u = \frac{\text{anni di vita in condizioni perfette se si accetta}}{\text{anni di vita da vivere in malattia}} \tag{1.1}$$

Idea è integrare la valutazione di efficacia con quella dei costi, al fine di dire se qualcosa di sperimentale è il miglior risultato di salute ottenibile con le risorse a disposizione.

Prospettive di analisi Si possono adottare le prospettive del:

- paziente (e famiglia),
- del terzo pagante (SSN): qui non sono rilevanti i costi privati di pazienti e familiari
- più in generale della società:

Valutazione dei costi I costi di un programma/trattamento etc possono essere diretti, indiretti o intangibili, e sono misurati in unità monetarie:

- i costi diretti sono quelli imputabili alla malattia principale (+ complicanze da co-morbosità) del settore sanitario (personale, attrezzature, farmaci, materiali), di pazienti e familiari (trasporto, alloggio, assistenza domiciliare) e di altri settori (modifiche richieste dal programma)
- i costi indiretti (o sociali) sono quelli del tempo di lavoro perso (paziente, familiari) e riduzione della produttività
- i costi intangibili (es dolore, ansia per intervento) possono solo essere elencati (sono compresi nei QALY)

Valutazione delle conseguenze Le conseguenze possono essere misurate in differenti modi, dando origine a diversi tipi di analisi, quando si va a confrontare coi costi/benefici tra due o + alternative:

- unità fisiche (es decessi, infezioni) da origine a costo-efficacia
- utilità (QALY) e costo-utilità
- unità monetarie (benefici diretti, indiretti e intangibili) da origine ad analisi costi-benefici

1.5 Comparazione di costi ed effetti

Piano costi-efficacia La differenza tra costi ed esiti nei due o più programmi messi a confronto viene poi plottata su un diagramma cartesiano con asse x la differenza di efficacia e asse y la differenza di costi:

- nel quadrante sud-est il trattamento sperimentale ha esiti migliori e costi minori, pertanto è palesemente superiore al trattamento controllo
- nel quadrante nord-ovest il trattamento sperimentale ha efficacia minore e costo maggiore ed è palesemente dominato dal trattamento controllo
- nei quadranti nord-est e sud-ovest non vi è un trattamento manifestamente superiore ed è necessaria una valutazione ulteriore: a nord est costi ed effetti del trattamento sperimentale sono superiori, mentre in sud-ovest si ha una effetto minore ma anche un costo. Bisognerà in entrambi i casi confrontare il costo per unità di efficacia con una soglia, che costituisce la pendenza di una retta passante dal centro degli assi. Le soluzioni al di sotto della linea sono considerate accettabili.

ICER A questo punto è possibile calcolare l'ICER (incremental cost effectiveness ratio come

$$ICER = \frac{\Delta C}{\Delta E} = \frac{C_{exp} - C_{control}}{E_{exp} - E_{control}}$$

che dice il costo aggiuntivo per unità di effetto aggiuntivo. Le analisi di costo efficacia e costo utilità differiranno per il denominatore, dove avremo rispettivamente unità naturali (anni vita) o QALY (anni vita moltiplicati per QOL)

Il programma è considerato finanziabile se ha un ICER minore della disponibilità a pagare per unità di outcome aggiuntivoq

$$\frac{\Delta C}{\Delta E} < R_T \tag{1.2}$$

con R_t la soglia che manifesta la disponibilità a pagare per una unità di efficacia in più (es anno di vita/QALY in più).

Aggiunta della variabilità La stima dell'ICER ad ora è unica e non dispone di variabilità. Il modo più comune per aggiungerla è mediante bootstrap: mediante questo si avranno

- una nuvola di punti sul piano costo efficacia
- una stima dell'intervallo di confidenza dell'ICER (es BCA bootstrap)

Per valori differenti di soglia si può calcolare la percentuale di campioni che rispettano la 1.2, detta probabilità di costo efficacia.

NMB L'equazione 1.2 di scelta dell'ICER si può riarrangiare portando tutto a destra e riesprimendo il criterio di accettazione sulla base del Net Monetary Benefit come:

$$NMB = R_t \Delta E - \Delta C > 0$$

In questa formulazione il programma è accettabile rispetto alla controparte se ha un NMB positivo. Dato che queste valutazioni dipendono fortemente da R_t ossia la disponibilità a pagare del finanziatore si può calcolare:

- il NMB per vari livelli di soglia, per vedere dove diventa positivo
- nel caso in cui vi sia variabilità nelle stime di costi e ricavi, al variare di R_t, quale è la probabilità che l'$NMB > 0$

Regressioni: OLS e SUR Dato che ogni paziente ha una misurazione di efficacia, una di costo e si ha una disponibilità a pagare, la NMB può essere calcolata a livello individuale come

$$NMB_i = E_i \cdot R_t - C_i$$

questo apre spazio alla modellazione mediante OLS, es per quantificare l'effetto del trattamento sul NMB medio più aggiustamento per covariate varie, in setting osservazionali o anche sperimentali.
Metodi più recenti (SUR) si sono spostati sul definire due equazioni di regressione, una per costi e una pergli effetti e correlandone il termine d'errore (non considerato indipendente tra i due: tipicamente costi ed efficacia sono negativamente correlati, costi maggiori si hanno in condizioni di efficacia minore). L'approccio seemingly unrelated regression ha il vantaggio di permettere differenti covariate e forme funzionali nelle due equazioni. Vedere Willan Briggs e Hoch 2004: regression methods for covariate adjuisment and subgroup analusos for non censored cost-effectiveness data

Capitolo 2

Outcomes

2.1 Costi

Il metodo di racolta dati sui costi dipende se siamo all'interno di un trial, dove accesso/quantificazione a livello di singolo paziente è possibile oppure no

- nei trial un modo per quantificare le prestazioni è il questionario CSRI (Client Services Receipt Inventory)
- al di fuori ci sono valorizzazioni di DRG

Più info sulla valorizzazione dei costi in Glick *e altri* (2007)

2.1.1 Attualizzazione

Costi che si verifichino in diversi periodi debbono essere attualizzati, scontandoli ad un determinato tasso (Baio (2013) riporta che NICE consiglia il 3.5% ma meglio effettuare analisi di sensibilità tra lo 0 e il 6).

```
## esempio pag 18 baio
npv <- function(y, t, i){
    #y flow
    #t time of flow in years
    #i yearly interest rate
    num <- y
    den <- (1 + i)^t
    sum(num/den)
}

npv(rep(15000, 5), 0:4, 0.035)

## [1] 70096.19
```

2.2 Efficacia

2.2.1 Outcome generici vs specifici

La qualità della vita del paziente deve essere tradotta in una utilità, ossia una misura nell'intervallo 0-1 con 0 = morte e 1 = perfetta salute.
Per misurare la qualità della vita si possono utilizzare strumenti generici o specifici:

- decision maker tendono a preferire outcome generici, che permettono di comparare costo-efficacia tra varie malattie/aree;
- laddove vi sia preoccupazione che uno strumento generico possa non funzionare bene o non vi siano dati di comparazione si può ricorrere a uno strumento specifico, posto che sia

Tra gli strumenti generici si annoverano:

- SF-6D: 11 item presi dall'SF-36 e combinati per produrre uno score
- EQ-5D-5L: scala in due parti (la prima su aspetti specifici, la seconda una valutazione overall)
- FACT-8D

2.2.2 Trasformazione in utilità e calcolo dei QALY

2.2.2.1 Utilità

Approfondire la metodologia in R nei pacchetti `eq5d` o fact-8D (per il FACT-G).

2.2.2.2 QALY

Dopo il calcolo dell'utilità a diversi punti occorre calcolare i QALY: l'attesa di vita di 1 anno di salute perfetta è valorizzata 1, mentre morte è valorizzata con 0.
Occorre che vi sia una valutazione di benessere anche al baseline. Vedere comunque `https://github.com/Health-Economics-in-R/QALY` e `https://en.wikipedia.org/wiki/Quality-adjusted_life_year` per altre robe sui QALY

```
QALY <- function(u, t, i = 0){
    ## browser()
    ## u utility
    ## t times in years
    ## i interest rate
    if (t[1] != 0) stop("QALY needs baseline measurement for utility")
    durate <- diff(t)
    utility_couples <- cbind(u[-length(u)], u[-1])
    mean_u <- apply(utility_couples, 1, mean, na.rm = TRUE)
    sum(durate * mean_u * (1/(1 + i)^(t[-length(t)])))
}

                                        ##y flow
```

```
                                        ##t time of flow in years
                                        ##i yearly interest rate
    ## num <- y
    ## den <- (1 + i)^t


## baio pag 24
u <- c(0.656, 0.744, 0.85, 0.744, 0.744)
t <- c(0,         6,   12,    18,    24)/12
QALY(u, t)

## [1] 1.519

u <- c(0.656, 0.656, 0.656, 0.656, 0.744)
QALY(u, t)

## [1] 1.334
```

Attualizzare i QALY Diversi attualizzano il QALY applicando un tasso di sconto (comune ai costi).

```
## variazione con attualizzazione
QALY(u, t, i = 0.035)

## [1] 1.299712
```

Parte II

Introduzione

Capitolo 3

Studi biomedici

3.1 Classificazioni

Definition 3.1.1 (Oggetto dello studio)**.** Si hanno

- studi per *descrivere una casistica* in un dato momento (ad esempio osservazionali cross-section)
- studi ove si cercano *relazioni di causa effetto* (efficacia trattamento, studi prognostici, studi eziologici)
- studi *diagnostici*
- studi per lo *sviluppo di strumenti/tool* (ad esempio questionari)
- studi che *sintetizzino l'evidenza disponibile* (revisioni sistematiche e metanalisi)

Definition 3.1.2 (Ruolo del ricercatore)**.** In studi dove si indagano relazioni di causa-effetto possiamo distinguere(Bacchieri e Della Cioppa, 2004):

- **Osservazionali**: il ricercatore studia la relazione tra una *caratteristica* (fattore demografico, ambientale, marker genetico), ed una *variabile di interesse* (detta anche *outcome*, ad esempio insorgenza malattia o guarigione), senza intervenire in alcun modo sulle condizione in cui lo studio viene condotto: seleziona il campione e poi osserva eventuali associazioni tra possibili fattori di rischio/protezione e la malattia che possano *suggerire* una relazione di causa-effetto.
- **Sperimentali** studi nei quali il ricercatore controlla le condizioni di svolgimento; nello specifico il ricercatore studia la relazione tra un *fattore sperimentale*, assegnato dal ricercatore stesso, ed un *variabile di interesse*; tutto questo il più possibile al netto dell'effetto di *fattori sub-sperimentali* (caratteristiche demografiche o della malattia, trattamenti precedenti e concomitanti, centro) che possano influenzare autonomamente la variabile di interesse.

Definition 3.1.3 (Criteri causalità)**.** Quando vi è causalità tra una caratteristica/fattore sperimentale e una variabile di interesse? alcune condizioni (nessuna necessaria o sufficiente) che rafforzano l'evidenza di causalità (Woodward, 2004)

1. forte associazione tra caratteristica/fattore sperimentale e una variabile di interesse
2. l'esposizione a caratteristica/fattore sperimentale è temporalmente precedente alla manifestazione della variabile di interesse
3. vi è una plausibile spiegazione biologica
4. l'associazione è supportata da altre ricerche in setting differenti
5. reversibilità: vi dovrebbe essere evidenza che se la causa è rimossa, anche l'effetto dovrebbe scomparire
6. dose-effetto: vi dovrebbe essere evidenza che a maggiori livelli di esposizione maggiore è il manifestarsi della variabile di interesse
7. assenza di spiegazioni alternative convincenti per eventuali variazioni della variabile di interesse

3.2 Misurazione ed errori

Remark 1. Una misura può essere vista come la somma di diverse componenti

$$\text{valore misurato} = \text{valore reale} + \text{errore casuale} + \text{errore sistematico}$$

Il processo di misurazione di un qualsiasi fenomeno può essere riguardato da due tipi di errori e questi sono parte integrante della variabilità dei fenomeni come li conosciamo.

Definition 3.2.1 (Errore casuale)**.** Errore che produce oscillazioni che non hanno un andamento riproducibile

Definition 3.2.2 (Errore sistematico (*bias*))**.** Errore che produce risultati che differiscono dal valore vero sistematicamente sempre nella stessa direzione

Remark 2. Sia errori casuali che sistematici hanno un impatto sul risultato della misurazione tuttavia il loro effetto è diverso:

- gli errori casuali, all'aumentare del numero dellemnisurazioni, tendono ad avere un impatto minore sugli indici di tendenza centrale (media/mediana) perché gli errori in una direzione o nell'altra tendono a compensarsi
- i secondi invece non si mitigano all'aumentare del campione

Capitolo 4

Strumenti di aggiornamento e ricerca bibliografica

4.1 Cosa sappiamo ad oggi

4.1.1 Sintesi su malattie per professionisti

Per una sintesi su una malattia fare un mix delle seguenti:

- dynamed
- ClinicalKey; conviene selezionare `Clinical Overview` se si vuole una sintesi su una malattia
- UpToDate (sono sintesi fatte da esperti)

4.1.2 Risposte più approfondite

Il database da consultare (soprattutto per diagnosi e trattamento) è quello delle https://www.cochranelibrary.com/ che costituiscono il gold standard

4.1.3 Ricerca bibliografica

4.1.3.1 Dizionario

Definition 4.1.1 (MEDLINE)**.** Un database di studi contenente autori, abstract e link e indicizzati secondo MeSH

Definition 4.1.2 (PubMed)**.** Un motore di ricerca gratuito su MEDLINE

Definition 4.1.3 (MeSH)**.** Vocabolario di temi medici gerarchico utilizzato per indicizzare per topic gli articoli e permettere ricerca precisa

4.1.3.2 Costruzione della ricerca

Prima di effettuare una ricerca nel database PubMed preprocessa la stringa di ricerca; per controllare cosa effettivamente PubMed cerchi (in seguito a quanto inserito) andare a vedere la history sotto l'advanced search.

Quando si effettua una ricerca semplice in Pubmed di default ci vengono restituiti molti risultati perchè:

- viene ricercato questo termine in tutti i campi del database (quindi nel titolo, autore, affiliazioni, giornale, abstract, lingua, mesh, keyword fornite dagli autori degli articoli). Questo è il modo più inclusivo di effettuare ricerche ma il rischio di falsi positivi è alto
- se il termine non è posto tra apici il termine viene applicato il mapping ossia si ammettono sinonimi e parole simili

Se vogliamo essere più precisi possiamo:

- specificare i fields
- disattivare il mapping (magari una volta individuati mesh e fields adatti) ponendo tra apici
- combinare il tutto con gli operatori logici e parentesi tonde

Definition 4.1.4 (Operatori logici)**.** Vanno categoricamente maiuscoli e sono

- `AND`: recupera items che contengono entrambi i termini
- `OR`: recupera items che contengono almeno uno dei termini
- `NOT`: recupera documenti che contengono solo il primo dei due termini, escludendo il secondo o i documenti in cui ci sia compresenza dei due

Definition 4.1.5 (Parentesi tonde)**.** Indirizza la precedenza/sequenza della ricerca come in algebra, ad esempio

```
cancer AND (prognosis OR diagnosis)
```

cercherà prima le parole prognosis o diagnosis e in seguito matcherà con i reisultati che hanno anche cancer

Definition 4.1.6 (Field tags)**.** Si possono rinvenire nella ricerca avanzata, ma i più importanti sono:

- `[au]`: autore
- `[mh]` o `[mesh]`: mesh
- `[ti]`: cerca nel titolo. Utile per restringere di molto la ricerca
- `[tiab]`: cerca la parola nel titolo o nell'abstract
- `[tw]`: per essere più inclusivi rispetto a `tiab` cerca in titolo o abstract o mesh o nelle keyword fornite dall'autore più altro (ma non cerca in affiliazione, autore, giornale)

Remark 3. Per le revisioni sistematiche i più comuni sono `tiab` e `tw`

Remark 4. Il field tags si applicano a tutte le parole che precedono; nel seguente esempio `tiab` è applicato anche a `cancer`

```
cancer prognosis[tiab] survival
```

viene mappato a

```
"cancer prognosis"[Title/Abstract] AND ("mortality"[MeSH Subheading] OR "mortality"[A
```

Definition 4.1.7 (Troncamento)**.** L'utilizzo dell'asterisco funziona come carattere di globbing: es `osteo*` trova osteosarcoma, osteoarthritis etc

4.1.3.3 Identificazione dei termini da ricercare

Per ricercare con i termini corretti da utilizzare nella ricerca "definitiva":

- analizzare i risultati di una ricerca base (in advanced -> history)
- analizzare i dati/categorizzazioni/metadati di un articolo considerato rilevante
- utilizzare il database delle MeSH (dalla schermata iniziale andare su Explore -> Mesh Database): per farlo cercare parole e poi aggiungere i filtri per la mesh che meglio rappresentano

4.2 Cosa bolle in pentola

Protocolli:

- ClinicalTrials.gov per singoli studi (trial ma anche osservazionali)
- prospero per le revisioni sistematiche
- cercare protocol nel titolo su pubmed

Capitolo 5

Misure epidemiologiche assortite

5.1 Misure e test di associazione

5.1.1 Esposizione ed esito dicotomici

Nel caso sia l'esposizione (a un fattore di rischio o a un trattamento) che l'esito siano variabili dicotomiche le frequenze possono essere rappresentate in una tabella 2×2 analoga a 5.1.

5.1.1.1 Misure

Sulla base di essa sono definibili diverse misure di associazione tra esposizione ed esito.

Definition 5.1.1 (Rischio)**.** Definito come:

$$Risk = \frac{a+b}{n}$$

Definition 5.1.2 (Risk ratio (o relative risk))**.** Definito come:

$$RR = \frac{a/(a+c)}{b/(b+d)} = \frac{a(b+d)}{b(a+c)}$$

con al numeratore il rischio tra gli esposti, al denominatore quello dei non esposti.

	Esposti	Non Esposti	Tot
Malati	a	b	$a+b$
Non Malati	c	d	$c+d$
Tot	$a+c$	$b+d$	$n = a+b+c+d$

Tabella 5.1: Tabella per misure di associazione

Definition 5.1.3 (Risk difference (o absolute risk reduction))**.** Definita come:

$$RD = \frac{a}{a+c} - \frac{b}{b+d}$$

con al numeratore il rischio tra gli esposti, al denominatore quello dei non esposti.

Definition 5.1.4 (Odds ratio)**.** Definito come

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

con al numeratore l'odd tra gli esposti, al denominatore quello dei non esposti.

Remark 5. Costituisce una buona approssimazione al rischio relativo quando si è in presenza di una malattia con prevalenza rara. Ossia:

$$a + c \simeq c,\ b + d \simeq d \iff RR = \frac{a/(a+c)}{b/(b+d)} \simeq \frac{a/c}{b/d} = OR \tag{5.1}$$

TODO: relazione odds ratio e risk ratio

Relative risk e odds ratio Generalmente presentare i risultati in termini di rischio relativo può essere compreso maggiormente (essendo i rischi delle probabilità). Le argomentazioni a favore dell'odds ratio:

- in alcuni casi (es studi caso controllo) è l'unica misura che può essere stimata correttamente;
- più facilmente stimabile (regressione logistica).

5.1.1.2 Test

Per testare l'associazione tra i due fattori `fisher.test` o `chisq.test`, oppure costruire l'intervallo di confidenza degli stimatori e verificare che non includano la soglia di non differenza (per RD è 0, per OR e RR è 1).

5.1.2 Esposizione multinomiale, esito dicotomico

5.1.2.1 Misure

L'esposizione multinomiale può essere naturale o derivare da una discretizzazione di una variabile quantitativa.
Comunque sia, nel caso si sceglie un gruppo di esposizione a fungere da gruppo base e si comparano gli altri livelli con questo utilizzando le misure già definite.

5.1.2.2 Linear trend test

Un test che assume importanza in questo ambito è il test dei trend lineari (Armitage, 1955), dove si verifica che a livelli via via crescenti del fattore la percentuale si modifichi linearmente (in aumento o decremento):

```r
# x e n sono le frequenze nell'ordine del fattore di rischio crescente
x <- c(83, 90, 129, 70)
n <- c(86, 93, 136, 82)

## implementato in R con
prop.trend.test(x, n)

##
##  Chi-squared Test for Trend in Proportions
##
## data:  x out of n ,
##  using scores: 1 2 3 4
## X-squared = 8.2249, df = 1, p-value = 0.004132

## under the hood <U+00E8> un modello lineare sulle probabilit<U+00E0> con pesi particolari
## dei quali non ho capito la genesi/interpretazione per ora
score <- 1:4
p <- sum(x) / sum(n)
w <- n/p/(1 - p)
df <- data.frame(freq = x/n, score = score)
a <- anova(mod <- lm(freq ~ score, data = df, weights = w))
chisq <- c(`X-squared` = a["score", "Sum Sq"])
(p <- pchisq(as.numeric(chisq), 1, lower.tail = FALSE))

## [1] 0.004131897

## riproduzione test su table 2.2 (pag 41) effettuato a pag 145 woodward
## x e n sono le frequenze nell'ordine del fattore di rischio crescente
x <- c(100, 382, 183, 668, 279, 109)
n <- c(592, 2254, 1017, 3150, 1253, 415)
prop.trend.test(x, n)$statistic

## X-squared
##  33.63156
```

Qualora il test venga significativo mentre il chi quadrato non lo sia sta a significare che sebbene non vi sia evidenza che le proporzioni negli strati differiscono dalla proporzione media, all'aumentare dello strato di esposizione si nota un incremento della proporzione registrata.

5.1.2.3 Test di non linearità

Il test si calcola la differenza tra l astatistica chi quadrato e la statistica del trend test; tale differenza è comparata ai valori critici della ristribuzione chi quadrato con gradi di libertà dati dalla differenza delle due componenti

```r
prop.nonlinear.trend.test <- function(x, score = seq_len(ncol(x))){
    ## x is a 2 x l matrix (l is the number of groups
    ## todo farla anche per table
    ## e per due variabili (forse?)
    if (!is.matrix(x) || nrow(x) != 2) stop('x must be a 2 x l matrix')
```

```
    method <- 'Test for non-linear trend in Proportions'
    dname <- paste(paste(x[1,], collapse = ', '), "out of",
                   paste(colSums(x), collapse = ', '))
    dname <- paste(dname, "\n using scores:", paste(score, collapse = " "))
    chi <- chisq.test(x)
    ltt <- prop.trend.test(x  = x[1, ], n = colSums(x))
    test <- chi$statistic - ltt$statistic
    df <- chi$parameter['df'] - ltt$parameter['df']
    p_value <- pchisq(q = test, df = df, lower.tail = FALSE)
    list(chi, ltt, test, df, p_value)
    structure(list(statistic = test,
                   data.name = dname,
                   parameter = c("df" = df),
                   p.value = p_value,
                   method = method),
              class = 'htest')
}

## woodward pag 146

## table 2.2 pag 41
chd <- c(100,  382,  183,  668,  279, 109)
n   <- c(592, 2254, 1017, 3150, 1253, 415)
nochd <- n - chd
m <- rbind(chd, nochd)
colnames(m) <-  c("I", "II", "IIIn", "IIIm", 'IV', 'V')
m

##         I   II IIIn IIIm  IV   V
## chd   100  382  183  668 279 109
## nochd 492 1872  834 2482 974 306

## validazione esempio pag 146
prop.nonlinear.trend.test(m)

##
##  Test for non-linear trend in Proportions
##
## data:  100, 382, 183, 668, 279, 109 out of 592, 2254, 1017, 3150, 1253, 415
##  using scores: 1 2 3 4 5 6
## X-squared = 2.7677, df.df = 4, p-value = 0.5974

prop.trend.test(m[1,], colSums(m))

##
##  Chi-squared Test for Trend in Proportions
##
## data:  m[1, ] out of colSums(m) ,
##  using scores: 1 2 3 4 5 6
## X-squared = 33.632, df = 1, p-value = 6.66e-09
```

```
## ------------------------------------
## attributable risk (woodward pag 148)
## ------------------------------------
library(attribrisk)

## Error in library(attribrisk):  there is no package called 'attribrisk'

es3.1 <- matrix(c(31, 1386, 15, 1883), ncol = 2)
dimnames(es3.1) <- list('death' = c('Yes' , 'No'), 'smoke' = c('Yes', 'No'))
## calcolo a mano (facendo riferimento a therneau come notazione di formule)
addmargins(es3.1)

##       smoke
## death  Yes   No  Sum
##   Yes   31   15   46
##   No  1386 1883 3269
##   Sum 1417 1898 3315

pd <- 46/3315
pd_given_notf <- 15/1898
(ar <- (pd - pd_given_notf)/pd)

## [1] 0.4304646

## simuliamo una situazione di dataset reale
library(lbmisc)
es3.1_df <- table2df(as.table(es3.1))
## e' importante che i si e i no siano codificati bene (No gruppo base, Yessa
## l'evento per cui si vuole stimare il rischio attribuibile)
es3.1_df$death <- relevel(es3.1_df$death, ref = 'No')
es3.1_df$smoke <- relevel(es3.1_df$smoke, ref = 'No')
str(es3.1_df)

## 'data.frame': 3315 obs. of  2 variables:
##  $ death: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ smoke: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...

table(es3.1_df)

##       smoke
## death   No  Yes
##   No  1883 1386
##   Yes   15   31

## table(es3.1_df)
attribrisk(death  ~ expos(smoke), data = es3.1_df)

## Error in attribrisk(death ~ expos(smoke), data = es3.1_df):  could
not find function "attribrisk"
```

```r
## non vengono proprio uguali gli intervalli di confidenza (il
## pacchetto usa jackknife, non una formula chiusa), comunque la stima
## c'e

es_th <- matrix(c(938, 763, 384, 559), ncol = 2)
dimnames(es_th) <- list('stroke' = c('Yes' , 'No'), 'hbp' = c('Yes', 'No'))
addmargins(es_th)
```

```
##       hbp
## stroke  Yes  No  Sum
##    Yes  938 384 1322
##    No   763 559 1322
##    Sum 1701 943 2644
```

```r
es_th <- table2df(as.table(es_th))
es_th$stroke <- relevel(es_th$stroke, ref = 'No')
es_th$hbp <- relevel(es_th$hbp, ref = 'No')
(example1 <- attribrisk(death ~ expos(hbp), data = es_th))
```

```
## Error in attribrisk(death ~ expos(hbp), data = es_th):  could not
find function "attribrisk"
```

```r
## COMUNQUE FARSI UNA FUNZIONCINA BASE BASE PER CALCOLARLO A MANO CHE
## NON MI FIDO ECCESSIVAMENTE


## --------------------------
## rate
## --------------------------

## esesmpio woodward pag 155

## rate uomini (1080/612955) (per 1000 uomini)
do.call(c, poisson.test(x = 1080, T = 612955)[c("estimate", 'conf.int')])*1000
```

```
## estimate.event rate           conf.int1           conf.int2
##            1.761956            1.658427            1.870256
```

```r
## relative rate uomini vs donne
do.call(c, poisson.test(x = c(1080, 306),
                        T = c(612955, 634103))[c("estimate", 'conf.int')])
```

```
## estimate.rate ratio           conf.int1           conf.int2
##            3.651183            3.212988            4.158972
```

```r
## stessa cosa fatta con modello di poisson
relrate_test <- data.frame(ev = c(1080, 306),
                           group = c(1, 0),
                           midp = c(612955, 634103))
rate_mod <- glm(ev ~ group + offset(log(midp)),
                data = relrate_test,
                family = poisson)
exp(cbind('estimate' = coef(rate_mod), confint(rate_mod)))['group', ]
```

```
## Waiting for profiling to be done...

## estimate    2.5 %   97.5 %
## 3.651183 3.220635 4.151744

## il denominatore del rate in questo caso <U+00E8> la popolazione a met<U+00E0>
## anno ma possono anche essere il tempo complessivo di osservazione
```

Capitolo 6

Protocollo, raccolta dati e articolo

6.1 Scrittura protocollo

Idee:

- partire dalla fine: ossia dalla pubblicazione che si vuole sottomettere e in altre parole che risposta si vuol dare, suoi contenuti e tipo di studio
- trovare una guideline per quel tipo di pubblicazione (consort, strobe, stard ecc) nell'equator network
- iniziare a compilare *direttamente in inglese*, il protocollo aiutandosi (ossia dando risposta a) tutti/maggior parte degli elementi messi in evidenza dalla guideline di pubblicazione relativa

Scrivere il protocollo direttamente in inglese fa si che al momento della scrittura dell'articolo vero e proprio sezioni come Introduzione, Materiale e Metodi siano già state scritte e possano essere solo copiate e incollate; mancheranno solo i risultati derivanti dall'analisi statistica e la discussion.

Una volta scritto il protocollo e individuati i task da svolgere, individuare la lista degli autori della pubblicazione.

6.2 Dati e loro raccolta

6.2.1 Tipologia di variabili

- **qualitative**: le rispettive modalità sono rappresentate da *attributi* (generalmente parole o frasi)
  - *nominali* (o *sconnesse*): le modalità non assumono alcun ordine. L'unica operazione effettuabile tra due unità consiste nello stabilire se posseggono o meno la stessa modalità/attributo.

    Example 6.2.1. Genere

  - *ordinali*: variabile le cui modalità presentano un ordinamento, per cui è possibile stabilire tra le modalità di due unità una relazione di ordine rispetto alla variabile considerata.

    Example 6.2.2. Per il TITOLO DI STUDIO, *Licenza media* $<$ *Diploma* $<$ *Laurea*

- **quantitative**: le rispettive modalità sono rappresentate da numeri

  - *discrete*: le sue modalità (numero) possono essere poste in corrispondenza con l'insieme $\mathbb{N}$ o un suo sottoinsieme (ossia $\mathcal{Y}$ è finito o numerabile)

    Example 6.2.3. NUMERO DI FIGLI

  - *continue*: le sue modalità (numero) possono essere poste in corrispondenza con l'insieme $\mathbb{R}$ o un suo sottoinsieme.

    Example 6.2.4. ALTEZZA

    Remark 6. Alcune variabili concettualmente continue (es età, altezza) possono essere registrate mediante valori discreti a causa della limitatezza di precisione insita nel relativo strumento di misurazione.

Le variabili quantitative possono anche essere classificate in base alla presenza di uno zero convenzionale o meno:

- quantitativa *per intervallo*: variabili che hanno una unità di misura ma non dello 0 (inteso come assenza della quantità da misurare), che viene inteso come convenzionale/arbitrario.
  Variabili del genere permettono solo un confronto per differenza tra le modalità che i soggetti assumono, mentre non ci permettono di calcolare rapporti che abbiano un senso (perchè lo 0 della scala è arbitrario).
  Esempi: misurazioni della temperatura in Celsius o Fahrenheit; il tempo misurato su differenti calendari.

- quantitativa *per rapporto*: variabili per le quali è intrinseca/univoca la definizione dello zero, corrispondente all'*assenza della caratteristica* misurata. Valori negativi non dovrebbero esser possibili.
  Il fatto che l'origine sia condivisa permette di calcolare rapporti tra grandezze diverse.
  Esempi: altezza, peso età, calorie.

Remark 7. *I metodi* per l'analisi delle variabili misurate su scale per intervallo o per rapporto *non differiscono* tra loro viceversa differiscono quelli da usati per quantitative discrete o qualitative.

6.3 Scrittura articolo

6.3.1 Autorship

ICMJE introduce quattro criteri per l'authorship di un lavoro (ICMJE, 2014):

1. *contributo sostanziale* alla *concezione* o *disegno* del lavoro; o alla *acquisizione*, *analisi*, o *interpretazione* dei dati;

2. *scrittura* o *revisione* critica delle bozze per aspetti dal contenuto intellettuale rilevante;
3. *approvazione* finale della versione per la pubblicazione;
4. disponibilità ad essere *accountable* per tutti gli aspetti del lavoro, ad assicurare che domande relative ad accuratezza/integrità di qualsiasi parte del lavoro siano investigate e risolte.

La distinzione tra author e contributor raccomandata da ICMJE, avviene come segue:

authors coloro che rispettano tutti i 4 requisiti contemporaneamente;

contributors coloro ne rispettano solo alcuni, ma non tutti e 4[1]

Idealmente, come sempre ICMJE raccomanda, quando il lavoro di ricerca è condotto da un gruppo nutrito di persone, il gruppo stesso dovrebbe decidere *in anticipo* (prima che il lavoro inizi) chi figurerà come autore (incaricandolo del rispetto degli oneri dell'authorship) e confermare in sede di sottomissione chi lo debba essere veramente (alla luce del rispetto effettivo o meno degli oneri imposti dall'autorship di cui prima).

[1] Esempi di attività tipiche che, da sole, non sono sufficienti per l'authorship: supervisione di un gruppo di ricerca, supporto amministrativo, supporto di scrittura, editing tecnico, correzione bozze e lingua.

Capitolo 7

Confounding e interazione

```
## -------------------------------------------------------
## WOODWARD standardizzazione tassi (esempio pag 178 sgg)
## -------------------------------------------------------
evs <- list('I' = c(0,0,1,6,7,16,17,25),
            'II' = c(0,0,4,7,13,11,28,44),
            'III' = c(0,0,1,9,17,19,43,53),
            'IV' = c(0,1,5,10,15,24,28,56))
popns <- list('I' = c(4784, 4210, 3396, 3226, 2391, 2156, 2182, 2054),
              'II' = c(4972, 4045, 3094, 2655, 2343, 2394, 2597, 2667),
              'III' = c(4351, 3232, 2438, 2241, 2360, 2708, 2968, 2802),
              'IV' = c(4440, 3685, 2966, 2763, 2388, 2566, 2387, 2380))
## standardizzazione diretta (non serve ev della popolazione standard)
std_ev1 <- list('All' = rep(NA, 8))
std_pop1 <- list('All' = c(8,6,6,6,6,5,4,4))

## esempio per standardizzazione indiretta
std_ev2  <- list('All' = c(0,1,11,32,52,70,116,178))
std_pop2 <- list('All' = c(18547,15172,11894,10885,9482,9824,10134,9903))

stdrate <- function(ev = NA,        # n. events per strata our sample
                    pop = NA,       # n. pop (or exposure time) of our sample
                    std_ev = NA,    # n. of events standard/reference pop
                    std_pop = NA,   # n. pop. (or exposure time) of std pop
                    per = 1000,    # multiply rate per X
                    ser_per = 100)# multiply ser per X
{
    ## ref woodward pag 181
    if (length(pop) != length(std_pop))
        stop("pop and std_pop must be of the same length.")
    ## per il calcolo del crudo e standardizzazione diretta non serve std_ev
    ## crude calculation
    crude <- (sum(ev)/sum(pop)) * per
    ## direct standardization
    dstd <- sum((ev/pop)*std_pop ) / sum(std_pop)
```

```
    dstd_se <-
        (1 / sum(std_pop)) *
        sqrt(sum(ev * (std_pop / pop)^2))
    dstd_ci <- dstd + c(-1,1) * qnorm(0.975) * dstd_se
    dstd_res <- setNames(c(dstd, dstd_se, dstd_ci),
                         c('est','se',  'lower.ci', 'upper.ci')) * per
    ## ser/smr/sir ecc ecc ecc
    exp_ev <- sum((std_ev / std_pop) * pop)
    ser <- sum(ev) / exp_ev
    ser_se <- sqrt(sum(ev)) / exp_ev
    ser_res <- setNames(c(ser, ser_se), c('est', 'se')) * ser_per
    ## indirect standardization
    crude_pop <- sum(std_ev) / sum(std_pop)
    indstd <-  ser * crude_pop
    indstd_se <-  crude_pop * ser_se
    indstd_res <- setNames(c(indstd, indstd_se), c('est', 'se')) * per
    ## results
    list('crude' = crude,
         'direct_std' = dstd_res,
         'ser' = ser_res,
         'indirect_std' = indstd_res)
}

Map(stdrate, evs, popns, std_ev1, std_pop1)
```

```
## $I
## $I$crude
## [1] 2.950941
##
## $I$direct_std
##      est       se lower.ci upper.ci
## 3.276607 0.388698 2.514773 4.038441
##
## $I$ser
## est  se
##  NA  NA
##
## $I$indirect_std
## est  se
##  NA  NA
##
##
## $II
## $II$crude
## [1] 4.320265
##
## $II$direct_std
##       est        se  lower.ci  upper.ci
## 4.1990966 0.4153992 3.3849292 5.0132641
##
```

```
## $II$ser
## est  se
##  NA  NA
##
## $II$indirect_std
## est  se
##  NA  NA
##
##
## $III
## $III$crude
## [1] 6.147186
##
## $III$direct_std
##       est        se  lower.ci  upper.ci
## 5.2993494 0.4615262 4.3947748 6.2039241
##
## $III$ser
## est  se
##  NA  NA
##
## $III$indirect_std
## est  se
##  NA  NA
##
##
## $IV
## $IV$crude
## [1] 5.896076
##
## $IV$direct_std
##       est        se  lower.ci  upper.ci
## 5.7544597 0.4933616 4.7874887 6.7214307
##
## $IV$ser
## est  se
##  NA  NA
##
## $IV$indirect_std
## est  se
##  NA  NA
```

```
(res1 <- Map(stdrate, evs, popns, std_ev2, std_pop2))
```

```
## $I
## $I$crude
## [1] 2.950941
##
## $I$direct_std
##       est        se  lower.ci  upper.ci
```

```
## 3.3795540 0.4001367 2.5953005 4.1638074
##
## $I$ser
##       est        se
## 69.718302  8.216381
##
## $I$indirect_std
##       est        se
## 3.3462108 0.3943547
##
##
## $II
## $II$crude
## [1] 4.320265
##
## $II$direct_std
##       est        se  lower.ci  upper.ci
## 4.3245158 0.4183908 3.5044850 5.1445466
##
## $II$ser
##       est        se
## 90.291430  8.728802
##
## $II$indirect_std
##      est       se
## 4.333642 0.418949
##
##
## $III
## $III$crude
## [1] 6.147186
##
## $III$direct_std
##       est        se  lower.ci  upper.ci
## 5.4252307 0.4576616 4.5282304 6.3222310
##
## $III$ser
##        est         se
## 113.028861   9.485171
##
## $III$indirect_std
##       est        se
## 5.4249513 0.4552518
##
##
## $IV
## $IV$crude
## [1] 5.896076
##
```

```
## $IV$direct_std
##      est       se lower.ci upper.ci
## 5.914942 0.502030 4.930982 6.898903
##
## $IV$ser
##       est        se
## 123.45627  10.47142
##
## $IV$indirect_std
##       est        se
## 5.9254268 0.5025881

## res2 <- Map(stdrate, evs, popns, std_ev2, std_pop2, list(1), list(1))
## res1[1]
## res2[1]

## ----------------------------------------
## mantel hanzel odd ratio, esempio pag 189
## ----------------------------------------

dimn <- list('housing' = c('rented', 'owner'),
             'chd' = c('yes', 'no'))
no_smoke <- matrix(c(33, 48, 923, 1722), ncol = 2, dimnames = dimn)
smoke    <- matrix(c(52, 29, 898, 678), ncol = 2, dimnames = dimn)
library(lbmisc)
no_smoke_df <- table2df(as.table(no_smoke))
smoke_df <- table2df(as.table(smoke))
no_smoke_df$smoke <- 'No'
smoke_df$smoke <- 'Yes'

df <- rbind(smoke_df, no_smoke_df)
tab <- table(df)
mantelhaen.test(tab)

##
##  Mantel-Haenszel chi-squared test with continuity correction
##
## data:  tab
## Mantel-Haenszel X-squared = 2.5354, df = 1, p-value = 0.1113
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9537528 1.8203312
## sample estimates:
## common odds ratio
##          1.317629

## oppure, ma <U+00E8> forse pi<U+00F9> chiara la precedente ...
## with(df, mantelhaen.test(housing, chd, smoke))
```

Parte III

Studi sperimentali

Capitolo 8

Studi sperimentali

Definition 8.0.1 (Studio sperimentale)**.** Studio in cui i ricercatori *assegnano* un trattamento (di varia natura)

Example 8.0.1. È passibile di studio sperimentale un farmaco o un intervento chirurgico; non lo è il fumo (somministrazione non etica) o il genere (somministrazione impossibile)

8.1 Fasi degli studi sperimentali (farmacologici)

Definition 8.1.1 (Sperimentazione clinica (dlgs 1997))**.** Sperimentazione condotta *su soggetti umani* a verificare effetti clinici ... di un prodotto in sperimentazione con l'obiettivo di valutarne sicurezza ed efficacia.

Remark 8 (Fasi della sperimentazione)*.* Lo sviluppo clinico di un farmaco è generalmente suddiviso in fasi temporali (dalla 1 alla 4), i cui confini non sono tracciabili in modo rigido (Amadori, 2004):

1. primi studi su un nuovo principio attivo condotti sull'uomo: scopo primario consiste nell'ottenere una valutazione preliminare su sicurezza e dosaggio del farmaco da impiegare (nonché avere le prime info su farmacocinetica e farmacodinamica nell'uomo)
2. scopo è dimostrare l'attività del farmaco (nonché valutarne meglio la sicurezza)
3. dimostrare l'efficacia terapeutica del farmaco
4. studi post commercializzazione per valutazione di sicurezza nel mediolungo periodo

Definition 8.1.2 (Attività di un farmaco)**.** Capacità del trattamento di indurre le modificazioni della malattia grazie alle quali si presume che l'ammalato possa avere un beneficio.

Definition 8.1.3 (Efficacia di un farmaco)**.** Capacità del trattamento di indurre un beneficio clinico negli ammalati ai quali viene somministrato.

Capitolo 9

Fase 1

Costituiscono il primo passo nella sperimentazione sull'uomo dopo, il completamento degli studi preclinici (in vitro e successivamente in vivo, su animali).

9.1 Obiettivi

Obiettivo primario Determinare la dose di un farmaco da raccomandare per gli studi successivi, in relazione alla tossicità registrata.

Remark 9. Sebbene tossicità sia il termine comunemente utilizzato, "eventi avversi" è più appropriato dato che tossicità implica una relazione causale con il farmaco mentre "evento avverso" rimane più neutrale (Eisenhauer *e altri*, 2014)

Remark 10 (Criteri valutazione tossicità). Standard moderno è la classificazione CTCAE (Common Terminology for Adverse Events) che per ogni possibile evento avverso rilevato durante l'osservazione (ad esempio) definisce una scala di gravità crescente di severità da 1 (moderato) a 5 (decesso dovuto ad evento avverso).

Example 9.1.1. Per il CTCAE Term `Anemia`, il grado 1 corrisponde a `Hemoglobin (Hgb) <LLN - 10.0 g/dL; <LLN - 6.2 mmol/L; <LLN - 100 g/L`.

Obiettivi secondari I principali sono descrivere farmacocinetica e farmacodinamica del trattamento.

Definition 9.1.1 (Farmacocinetica (PK))**.** Modificazione nel tempo della concentrazione nel sangue di un farmaco.

Remark 11. È possibile costruire per una data dose e via di somministrazione una curva che rappresenti i valori di concentrazione plasmatica in funzione del tempo. Per tutte le vie di somministrazione ad eccezione dell'intravascolare l'andamento della curva prevede tre fasi: crescita (assorbimento farmaco), picco e decremento (metabolizzazione/smaltimento). Nel caso di somministrazione intravascolare non vi è assorbimento quindi il picco si ha all'inizio e la curva è monotona decrescente.

Definition 9.1.2 (Farmacodinamica (PD))**.** Studio degli effetti biochimici e fisiologici del farmaco.

Remark 12. L'effetto si descrive costruendo un grafico che rappresenta l'entità della risposta (in termini ad esempio di espressione di una data proteina) in funzione del logaritmo della dose somministrata (perché si presuppone una relazione dose-risposta).

9.2 Popolazione

Sono generalmente sono condotti su volontari sani.
In oncologia, data l'utilizzazione di farmaci potenzialmente tossici, la sperimentazione non è fattibile su volontario sano (per la tossicità insita nel trattamento) ma è condotta su pazienti affetti da tumore che forniscono il proprio consenso.

Remark 13. Recentemente si è assistita alla crescita di interesse nei confronti dei cosiddetti phase 0 trial (Eisenhauer *e altri*, 2014, p. 5) dove un *limitatissimo numero di soggetti sani può* essere arruolato per valutare microdosi di una nuova terapia al fine di rispondere a domande utili prima che una fase 1 vera e propria possa iniziare.

9.3 Definizioni

Definition 9.3.1 (Dose massima tollerata (MTD))**.** Dose da raccomandare per gli studi delle fasi successive, coincidente con la massima dose per la quale la tossicità del farmaco (eventuali eventi avversi) risulta accettabile.

Definition 9.3.2 (Dose massima somministrata (MAD))**.** Dose alla quale l'escalation cessa a causa dell'osservazione di un numero critico di DLT.

Definition 9.3.3 (Tossicità dose-limitante (DLT))**.** Ogni evento tossico così severo o irreversibile tale da impedire l'incremento della dose.

Remark 14. La DLT è spesso definita come l'occorrenza di tossicità severa: gradi 3 o 4 per eventi avversi non ematologici o gradi 4 per tossicità ematologiche (Eisenhauer *e altri*, 2014, p. 171).

Definition 9.3.4 (Recommended phase II dose (RP2D))**.** Dose raccomandata per la fase 2 (concide con la MTD)

9.4 Disegni

Uno studio di fase 1 è tipicamente disegnato come uno studio a dosi crescenti per la determinazione della MTD. In merito alla dose somministrata le decisioni da prendere sono:

- la dose (espressa in milligrammi per metro quadrato di superficie corporea)dalla quale partire: scelta comune consiste nel 10% della MELD10 (o LD_{10})

  Definition 9.4.1 (MELD10 (o LD_{10}))**.** Dose letale per il 10% dei topi sottoposti a tale dose.

- lo schema di incremento: trade off tra aumento troppo veloce (esporre pazienti a tossicità eccessive) o troppo lento (allungamento dei tempi di sviluppo di un farmaco potenzialmente utile). Ogni disegno (vedi in seguito) propone un diverso schema.

L'aumento della dose è possibile solo dopo che è trascorso un periodo di tempo sufficientemente prolungato per osservare l'eventuale effetto tossico nei pazienti inseriti a livello precedente.

I disegni di fase 1 possono essere classificati, in base allo schema di incremento, in due gruppi (Eisenhauer *e altri*, 2014):

rule-based o algoritmici determinano la dose attraverso un processo iterativo: i pazienti vengono assegnati a dose via via crescenti in base a regole prespecificate in protocollo, in relazione ad eventi di tossicità.

Example 9.4.1. Il disegno standard 3+3

model-based stima la relazione dose tossicità e procede alla scelta della dose

Example 9.4.2. Continual reassessment method

9.4.1 Disegno standard (3+3)

9.4.1.1 Funzionamento

Il disegno standard utilizza:

- un profilo di incrementi decrescenti del livello di dose secondo secondo lo schema di tabella 9.1; in particolare la dose non viene mai variata nel singolo paziente ma si considerano nuovi pazienti ai quali somministrare la nuova dose;
- un algoritmo di passaggio a dosaggi via via successivi come da diagramma 9.1.

La procedura, dal punto di vista statistico si basa sulle seguenti considerazioni:

- se almeno 2 su 3 pazienti trattati ad un particolare livello di dose mostrano una DLT, si può affermare con confidenza del 90% che la probabilità di DLT a quella dose è $> 20\%$

```
binom.test(2, 3, alternative = 'greater', conf.level = 0.9)$conf.int
```

```
## [1] 0.1958001 1.0000000
## attr(,"conf.level")
## [1] 0.9
```

- d'altra parte, se 0 pazienti mostrano una DLT si può affermare con confidenza del 90%, che la vera probabilità di DLT è $< 55\%$

```
binom.test(0, 3, alternative = 'less', conf.level = 0.9)$conf.int
```

```
## [1] 0.0000000 0.5358411
## attr(,"conf.level")
## [1] 0.9
```

Livello	Aumento	Esempio dose (mg/m^2)
1	10% del MELD10	1
2	100	2
3	67	3.3
4	50	5.0
5	40	6.7
6	33	8.8
7	33	11.8
8	33	15.7

Tabella 9.1: Schema incrementi di dose. Da Amadori (2004).

Nel caso lo studio non riesca a identificare la MTD si può procedere in due direzioni:

- proporre la dose più elevata come dose ottimale per la fase 2
- progettare una nuova fase 1 che preveda dosi ancor più elevate di farmaco

Nel caso si esca per MTD identificata alla dose iniziale occorre disegnare un nuovo studio di fase 1 che parta da una dose iniziale inferiore.

9.4.1.2 Critiche

Le principali (Amadori, 2004):

- troppi pazienti vengono trattati a basse dosi, con poca utilità da un punto di vista terapeutico
- incrementi lenti e numerosi, durata lunga per la definizione della MTD

9.4.1.3 Varianti dello schema

Dose iniziale maggiore Es impostando la dose iniziale al 20% della MELD invece che al 10%

Accelerated titration design L'idea è trattare un primo paziente a livelli di dose via via crescenti fino a quando non si osserva una tossicità di grado 2 o maggiore. Quando questo avviene si trattano i successivi pazienti partendo dalla dose precedente e seguendo lo schema classico.

9.4.2 Continual reassessment method (Adattamento continuo)

Si stima in maniera Bayesiana la relazione tra dose e tossicità per derivare il livello di dose che è associato ad una determinata frequenza di DLT (solitamente il 20-30%):

- si parte da una distribuzione a priori della probabilità di DLT in funzione della dose
- ogni nuovo paziente che entra nello studio viene trattato alla dose stimata come MTD (allo stato attuale)

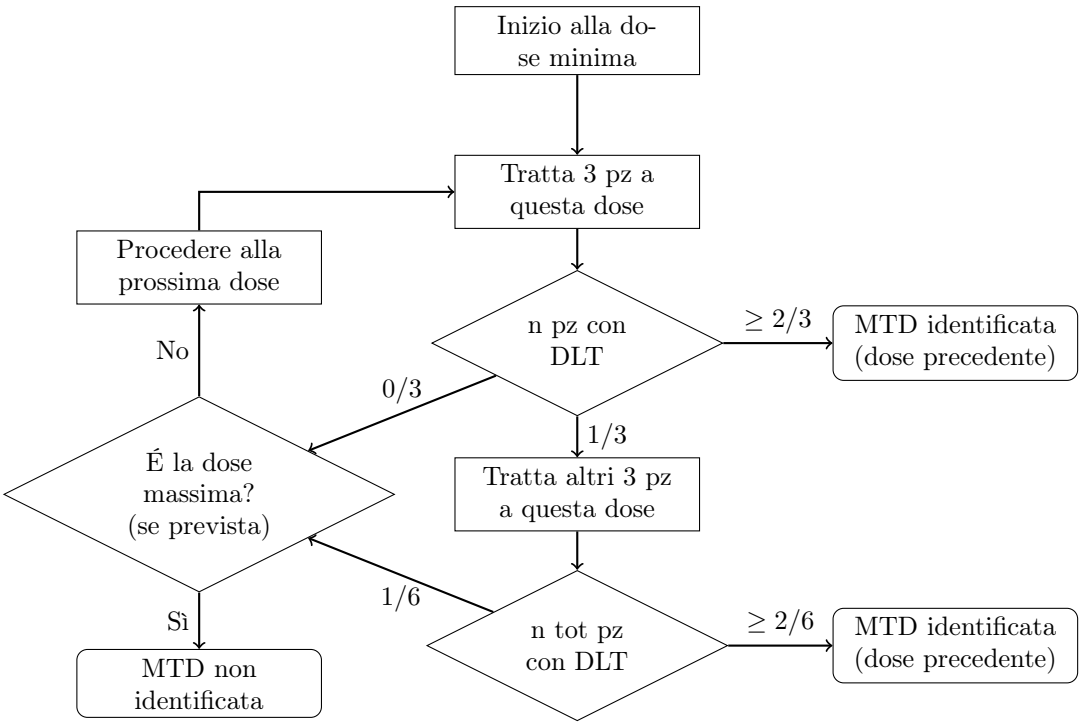


Figura 9.1: Flowchart schema classico 3+3. Da Eisenhauer *e altri* (2014).

- la risposta del paziente viene utilizzata per aggiornare il modello in maniera tale da aggiornare la distribuzione a posteriori della relazione dose/tossicità sulla base della quale effettuare la scelta di dose per il pz successivo

Con questo metodo:

- occorre inserire un criterio di uscita dallo studio ma inserendo nello studio 20-25 pazienti si ottiene una stima sufficientemente accurata di MTD
- ogni paziente viene trattato con una dose di farmaco sufficientemente alta (con un beneficio clinico per tutti non solo gli ultimi)

Per ulteriori approfondimenti: Eisenhauer *e altri* (2014, p. 178) e Wheeler *e altri* (2019).

9.4.3 Disegno per nuovi farmaci

Laddove non vi sia correlazione tra dose somministrata ed effetto cade il paradigma sul quale si fonda il disegno classico di fase 1 (utile per lo più per farmaci tipo i citotossici).
L'idea è che la dose ottimale da mandare in fase 2 è una non necessariamente tossica e procedere in maniera classica porterebbe ad un sovradosaggio del farmaco.

TODO: Verifica disegni per nuovi farmaci

Capitolo 10

Fase 2

Screenano farmaci poco attivi avviando a nuovi studi di più ampie dimensioni soltanto quelli potenzialmente efficaci.

10.1 Obiettivi

Obiettivo primario Valutare l'attività del trattamento intesa come capacità di indurre le modificazioni sulla malattia/paziente che fanno ben sperare per la sua prognosi.

Obiettivi secondari Principalmente tossicità, ulteriori analisi di farmacocinetica e farmacodinamica, valutazione effetti su specifici bersagli molecolari (se noti)

10.2 Aspetti da considerare nel disegno

Alcuni aspetti da chiarire nel disegno di uno studio di fase 2 (Brown *e altri* (2014) e (Amadori, 2004)) seguono.

10.2.1 Popolazione

Uno studio di fase 2 richiede pazienti:

- affetti da uno specifico tipo di neoplasia
- per i quali si abbia avuto progressione dopo una terapia standard
- per i quali sia possibile valutare la risposta della malattia al trattamento (es dimensione tumore o altri indicatori)
- per i quali il trattamento possa dare un beneficio

Per questo motivo:

- nel caso di farmaci citotossici sono esclusi che presentano una malattia non misurabile (poiché non è possibile definire una risposta parziale);

- possono essere esclusi pz con performance status basso e/o aspettativa di vita < 3 mesi;
- i pz non debbono avere gravi malattie concomitanti
- adeguata funzionalità renale, epatica, cardiaca e respiratoria.
- possono essere esclusi pz con precedenti trattamenti (non standard, altri studi): potrebbero debilitare il paziente rendendo difficile poter somministrare la dose piena del farmaco sperimentato;

10.2.2 Trattamento

Quale è il *meccanismo d'azione*? Questa avrà influsso su diversi aspetti, principalmente l'outcome adottato e l'impiego della randomizzazione o meno.
Il trattamento è (o è assimilabile a) un:

- **citotossico**: l'obiettivo del trattamento (o comunque come è ragionevole misurare il funzionamento del farmaco) è la contrazione della massa tumorale.
  Se la risposta al trattamento è misurabile attraverso la dimensione del tumore i criteri standard attuali sono i RECIST (Eisenhauer *e altri*, 2009). Data l'estrema rarità di una regressione tumorale spontanea, è giustificato ipotizzare che la risposta che si è eventualmente verificata possa essere solo la conseguenza del farmaco e ciò rende possibile la programmazione di studi a singolo braccio.
- **citostatico**: l'obiettivo del trattamento non è tanto una diminuzione del volume, piuttosto evitare/rallentarne la progressione.
  Nel caso di farmaci non citotossici il criterio dimensionale su cui si basa la valutazione della risposta dei trattamenti classici viene tipicamente a cadere. Si rende necessaria l'identificazione di nuovi parametri per valutare l'attività del trattamento (Amadori, 2004):
  - parametri clinici: valutazioni temporali quali tempo alla progressione, sopravvivenza, numero di pazienti vivi o liberi da progressione
  - criteri biologici: ad esempio parametri misurabili su campioni bioptici o su altro materiale biologico

  Diviene necessario uno studio controllato necessario laddove si utilizzano indicatori che possono modificarsi spontaneamente nel tempo.

10.2.3 Outcome

L'outcome deve essere un surrogato validato, ossia tipicamente correlato ad un outcome hard/di efficacia usato in fase 3. Outcome tipici sono:

- risposta dicotomica (es CR+PR e parziale vs SD+PD)
- risposta multinomiale (es CR+PR vs SD vs PD)
- quantitativa (es marker)
- tempi all'evento:

  - TTP: dalla randomizzazione alla progressione
  - TTF: dalla randomizzazione all'interruzione del trattamento (per progressione, tossicità, scelta del paziente o morte)
  - PFS: dalla randomizzazione alla progressione o morte (quale
  - tempo ad un evento rilevante (es al SAE per studi concentrati sulla tossicità, alla prima frattura per trattamenti contro metastasi ossee)

10.2.4 Randomizzazione

La randomizzazione è necessaria se:

- l'outcome scelto è soggetto ad evoluzione spontanea
- dati storici non sono disponibili o sono poco affidabili/confrontabili eccetera
- molteplici trattamenti oppure molteplici dose/schedule/sequenze di un singolo trattamento

10.2.5 Scopo (sottofase) dello studio

A volte si suddividono gli studi di fase II in due sottogruppi:

1. phase IIa (*proof of concept*): si valuta l'attività di un trattamento che ha completato la fase 1 oppure si valutano diverse dosi di un trattamento per valutare la dose-risposta (learning trial). Più tipicamente a singolo braccio.
2. phase IIb (*go/no-go*): studio nel quale si decide se procedere o meno con la fase 3, può comparare diversi dosaggi o un dosaggio verso placebo, tipicamente randomizzando.

A volte le due sottofasi vengono riunite in un unico studio.

10.2.6 Categorie disegni

Stadio unico Un campione prefissato di pazienti è reclutato, trattato e seguito sino all'ottenimento dell'outcome. Questi disegni:

- sono semplici: evitano complessità inerenti strategie di reclutamento in presenza di analisi ad interim
- non permettono flessibilità, es fermare il trial in anticipo a causa di livelli troppo bassi (o veramente alti) di attività

Sono utili se:

- il trattamento è molto efficace e si presume di concludere lo studio con pochi pazienti;
- vi è relativa sicurezza sulla tossicità tale da non rendere necessaria una analisi ad interim.

Due stadi Sono opportuni se l'attività del trattamento è sconosciuta e/o la tossicità non può essere sottovalutata (ci si lascia spazio per terminare in anticipo).
I pazienti sono reclutati in due fasi:

- al termine della prima si decide se continuare o fermarsi (mancanza di attività o forte presenza di attività, tossicità inaccettabile) oppure su quale trattamento portare alla seconda fase;
- al termine della seconda si fornisce la stima di attività

Data la natura del disegno il sample size complessivo non è fisso, pertanto di solito questi disegni forniscono un sample size massimo e un sample size medio (ASN, che contempli la probabilità di terminazione precoce e i sample size nei due stati del mondo).

Multistage Anche detti group-sequential, sono simili ai disegni a due stadi (ma ad un numero di stadi maggiori) e vengono specificati criteri di stop per ciascuno di essi.

Continuous monitoring La valutazione di attività è effettuata ogni volta che un nuovo paziente ha il dato di attività (anche qui lo si fa per early termination alla massima potenza).

Decision theoretic Valutano costi e guadagni associati ad effettuare decisioni sbagliate alla fine di una fase 2, incorporando funzioni di utilità associate a questi costi/guadagni. Valutazione più generale rispetto a quella meramente clinica, ma di difficile applicazione; raramente usati in oncologia.

Three outcomes Alla fine dello studio non vi sono solo due ipotesi (nulla e alternativa) ma si può ottenere anche uno studio non conclusivo

Phase 2/3 Sono usati quando una transizione dalla fase 2 alla 3 deve essere il più efficiente possibile: permettono di incorporare i dati generati dalla fase 2 nell'analisi di fase 3 (risparmiando così pazienti). Sono solitamente randomizzati con gruppo di controllo. La randomizzazione può essere usata per:

- decidere se continuare un singolo trattamento da portare in fase 3;
- selezionare il trattamento con la miglior risposta dei diversi confrontati, da portare in fase 3.

Per evitare problemi di molteplicità, nell'analisi inerente la fase 2 si usa un outcome diverso da quello di fase 3 (e si inizia a misurare l'outcome di fase 3 anche nei pazienti che iniziano con la 2; chiaramente facendo i due studi in maniera separata si può beneficiare della conoscenza accumulata durante la fase 2 nel disegno della fase 3 (es cambi ai criteri di eleggibilità, schema follow up).

Randomised discontinuation Nei trial a randomized discontinuation (o *enrichment* trial) i pazienti vengono trattati tutti con lo stesso trattamento e in seguito valutati per risposta

- quelli in progressione escono dallo studio
- quelli in risposta (parziale/completa) continuano a ricevere il trattamento
- quelli in stable disease sono randomizzati a continuare il trattamento, a interromperlo (e rimanere senza trattamento) oppure a ricevere un trattamento standard (in relazione alla domanda dello studio)

10.3 Disegni comuni

Nell'approccio frequentista si ha un test di ipotesi unilaterale del tipo

$$H_0 : p \leq p_0$$
$$H_1 : p \geq p_1$$

con:

- $p_1 > p_0$
- p_0 livello di risposta giudicato fallimentare (oppure quella del trattamento standard se disponibile)
- p_1 un livello di risposta auspicabile; p_1 è scelto in modo tale che $p_1 - p_0$ rappresenti il miglioramento auspicabile col nuovo trattamento

10.3.1 Livelli di errore nell'inferenza

Ponendoci in ottica di screening di potenziali candidati, riguardo agli errori:

- l'errore del primo tipo, α, può andare dal 5% tipico della fase 3 sino al 20%, al fine di risparmiare pazienti;
- l'errore di tipo 2, β, deve comunque rimanere al massimo 20% (meglio 10%); si vuole comunque evitare di avere dei falsi negativi in questa fase di screening, similmente a quanto avviene nella fase 3.

10.3.2 Stadio unico - A'Hern

Adottando una distribuzione binomiale dei successi si giudica la combinazione ampiezza campionaria/eventi osservati soddisfacente

$$\mathbb{P}\left(\text{reject } H_0 | \pi = p_0\right) \leq \alpha \, \mathbb{P}\left(\text{accept } H_0 | \pi = p_1\right) \leq \beta$$

A livello di algoritmo

- si scelgono i parametri α, β;
- si cicla sul sample size (diciamo da 1 a 2000);
- per ogni sample size si cicla sul numero di successi (da 0 a sample size) a 2000), considerati come soglia per rifiutare la nulla;

- si calcolando le probabilità associate alle binomiali sotto nulla e alternativa;
- si ritiene la combinazione sample size/eventi con sample size minore

Example 10.3.1.
```
lbss::ahern(p0 = 0.4, p1 = 0.55, alpha = 0.1, beta = 0.1)
## Error in loadNamespace(x):  there is no package called 'lbss'
```

Per impostare il trial come sopra che giudichi il farmaco utile per la fase 3 se consente un incremento di 15 punti percentuali nella risposta si ha che

- occorre reclutare 75 pazienti
- si può passare alla fase 3 se almeno 36 vanno in risposta; infatti in tal caso il limite inferiore dell'intervallo di confidenza (Clopper Pearson, al 90% ad una coda), è 0.4006 che è maggiore di p_0.

10.3.3 Due stadi - Simon

È un disegno:

- a due stadi
- per outcome binario
- che permette early termination per mancanza di attività (non è previsto early stop per eccesso di attività)

Se:

- n_1 sono i pazienti da reclutare al primo stage
- r_1 i successi che fungono da cutoff al primo stage (r_1 si termina $r_1 + 1$ si prosegue)
- n_2 i pazienti da reclutare al secondo stage
- r i successi che fungono da cutoff al secondo stage (se r lo studio non va in fase 3 se $r + 1$ sì)

Le seguenti sono la probabilità di early (al primo step) termination `PET`, il numero atteso di pazienti `EP`, la probabilità di rifiutare H_0 al termine dello studio (`R`) (serve per calcolare la potenza direi se)

$$
\begin{aligned}
PET(p) &= B(r_1; p, n_1) \\
EP(p) &= n_1 + (1 - PET(p)) \cdot n_2 \\
R(p) &= 1 - B(r_1; p, b_1) - \sum_{m=r_1+1}^{\min(n_1, r)} b(m; p, n_1) \cdot B(r - m; p), n_2)
\end{aligned}
$$

dove b è la PMF e B la cumulative distribution function della binomiale.

Le quantità di sopra possono essere calcolate sotto varie ipotesi di p ma di interesse nell'ipotesi che sia vera H_0 ($p = p_0$).
Del disegno di Simon (1989), una volta fissati α, β e indagando per forza bruta ve ne sono due varianti:

- **optimal**: minimizza i pazienti attesi sotto H_0 ($EN(p_0)$) nonchè i pazienti da reclutare al primo stage
- **minimax** minimizza i pazienti complessivi ($n_1 + n_2$) al termine delle due fasi (da preferire nel caso vi possano essere difficoltà di reclutamento, es tumori rari)

Example 10.3.2. Se $p_0 = 0.4$, $p_1 = 0.55$, $\alpha = \beta = 0.1$ e si considererà interessante il farmaco qualora permetta un incremento di .15 rispetto alla risposta standard.

```
# pu <U+00E8> p0, pa <U+00E8> p1, ep1 <U+00E8> alpha, ep2 <U+00E8> beta
clinfun::ph2simon(pu = 0.4, pa = 0.55, ep1 = 0.1, ep2 = 0.1)

## Error in loadNamespace(x):  there is no package called 'clinfun'
```

Lo studio si svolgerà come segue:

- durante la prima fase bisogna arruolare 38 pazienti con Optimal (45 con Minimax)
- si termina alla prima fase (farmaco *non* di interesse) se le risposte sono al massimo 16 (18 per minimax), altrimenti si prosegue alla seconda fase
- durante la seconda fase bisogna arruolare altri 50 pazienti (28 per minimax), per arrivare complessivamente a 88 pazienti (73 per minimax)
- il farmaco non è di interesse se le risposte sono complessivamente (fase 1 e 2) al massimo 40 (o 34 per minimax), altrimenti si può proseguire con studi di fase 3
- `PET`: probabilità di terminare lo studio allo step iniziale sotto H_0 ($p = p_0$) è 66% (56% per minimax) ed è calcolata mediante la cumulata della binomiale di r_1 risposte su n_1 trial sotto ipotesi che $p = p_0$ (terminiamo al primo stage se osserviamo r_1 risposte o meno), ossia

```
pbinom(16, 38, 0.4)

## [1] 0.6695864
```

- `EN`: il numero atteso di pazienti da reclutare sotto l'ipotesi H_0 è 54 (57 per minimax) ed è calcolato come

$$EN = n_1 + (1 - PET)n_2$$

con n_1 pazienti della fase iniziale, n_2 della seconda fase e `PET` probabilità di termine alla fine della prima fase.

```
38 + (1 - pbinom(16, 38, 0.4)) * 50

## [1] 54.52068
```

Example 10.3.3. Provando a riprodurre a mano l'algoritmo: fissato α, β, p_0, p_1

- for sample size tot: $n \in (2, \ldots, max\_samplesize)$
- for sample size tot: $n_1 \in (1, \ldots, n-1)$
- ottieni sample size fase2: $n_2 = n - n_1$
- for successi possibili fase1 $r_1 \in (0, n_1)$
- for successi possibili fase 2: $what$
- calcola la probabilità di rifiutare l'ipotesi nulla sotto ipotesi che sia vera (e verifica che tale valore sia inferiore a α) o sia vera H_1 (e che sia maggiore di $1 - \beta$)
- restituisci le combinazioni di parametri che rispettano $alpha, \beta$

10.3.4 Altri disegni

Si hanno:

- Ensign *e altri* (1994) facilita l'early stopping rispetto a Simon per mancata attività, mentre Chen (1997) rimuove alcune limitazioni
- Chang *e altri* (1987) permettono l'early stopping sia in caso di bassa che alta attività
- Bryant e Day (1995) considerano congiutamente l'analisi di attività e tossicità, i due obbiettivi principali di uno studio di fase II

Per un confronto e un aiuto nella scelta, complessivamente, Mariani e Marubini (1996).

10.3.5 Stima al termine di un multistage

Occorre adottare procedure adhoc come spiegato in Mariani e Marubini (1996)

10.4 Criteri RECIST

10.4.1 Classificazione delle lesioni e tumour burden

Al baseline lesioni/linfonodi sono classificati come:

- *misurabile*:
  - lesioni tumorali con diametro maggiore $\geq$ 20mm (se misurato con raggi) o $\geq$ 10mm (CT scan, calibro)
  - linfonodi con asse minore $\geq$ 15mm (CT scan)
- *non misurabile*: tutte le rimanenti lesioni

Il metodo di misurazione impiegato al baseline deve rimanere costante ai successivi follow-up.

Definition 10.4.1 (Measurable disease)**.** Presenza di almeno una lesione misurabile.

Remark 15. In protocolli (Eisenhauer *e altri*, 2009, p. 232):

- in cui la risposta è l'endpoint primario solo i pazienti con measurable disease al baseline dovrebbero essere inclusi in protocollo;
- in cui la progressione tumorale è l'outcome primario (es pfs o proporzione di pazienti con progressione ad un dato istante) il protocollo deve specificare se l'ingresso in analisi è ristretto a coloro che hanno measurable disease o tutti (anche a pazienti con nemmeno una lesione misurabile).

Remark 16. Al fine di determinare l'*overall tumor burden* al baseline, le lesioni vengono ulteriormente suddivise in lesioni target e non

Definition 10.4.2 (lesioni target)**.** lesioni misurabili, fino ad un massimo di 2 per organo e di 5 in totale nel singolo paziente, scelte in modo da risultare rappresentative di tutti gli organi interessati; debbono essere scelte in base alla dimensione del diametro maggiore e all'accessibilità prevedibile nelle successive valutazioni;

Definition 10.4.3 (lesioni non target)**.** tutte le lesioni rimanenti.

Definition 10.4.4 (Dimensione al baseline (overall tumor burden))**.** Somma dei diametri maggiori (asse minore per i linfonodi) di tutte le lesioni target, rilevata all'inizio del trattamento

Remark 17. La dimensione al baseline **costituisce il riferimento** per le successive misurazioni e per la valutazione della risposta. La misura deve essere monitorata lungo il follow up per la determinazione della risposta

10.4.2 Risposta

Remark 18. Per valutare la risposta globale del paziente si valuta:

- risposta nelle lesioni target
- risposta nelle rimanenti
- comparsa di nuove lesioni

Definition 10.4.5 (Risposte lesioni target)**.** Sono:

- *risposta completa (CR)*: scomparsa di tutte le lesioni target, tutti i linfonodi patologici (target o non target) con asse inferiore < 10mm;
- *risposta parziale (PR)*: diminuzione $\geq 30\%$ rispetto alla dimensione al baseline della somma dei diametri delle lesioni target
- *progressione (PD)*: aumento $\geq 20\%$ (ma comunque ≥ 5mm) rispetto rispetto alla dimensione minima registrata durante il follow-up (sia essa quella del baseline o meno);
- *stabilità di malattia (SD)*: ne riduzione in grado di dare PR, ne incremento tale da qualificare PD (caso residuale rispetto ai precedenti due).

Lesioni target	Lesioni non target	Nuove lesioni?	Overall response
CR	CR	No	CR
CR	Non-CR/non-PD	No	PR
CR	Non valutate	No	PR
PR	Non-PD o non tutte valutate	No	PR
SD	Non-PD o non tutte valutate	No	SD
Non tutte valutate	Non-PD	No	NE
PD	Qualsiasi	Sì o No	PD
Qualsiasi	PD	Sì o No	PD
Qualsiasi	Qualsiasi	Sì	PD

Tabella 10.1: Valutazione risposta globale (OR) nei pz con malattia misurabile

- nel caso non tutte le lesioni target abbiano misurazione ad un dato follow up la risposta si considera *non valutabile* (*NE*).

Definition 10.4.6 (Risposte lesioni non target)**.** Sono:

- *risposta completa (CR)*: scomparsa di tutte le lesioni non target e negativizzazione dei marcatori tumorali. Tutti i linfonodi non patologici (asse minore < 10 mm);
- *Non-CR/Non-PD*: persistenza di una o più lesioni non target e/o persistenza di livelli elevati dei marcatori tumorali;
- *progressione (PD)*: inequivocabile progressione di lesioni non target preesistenti e/o comparsa di 1 o più lesioni

Definition 10.4.7 (Risposta globale (overall response, OR))**.** Risposta complessiva del paziente al trattamento, che contempla sia lesioni target che non; nel caso di pazienti con malattia misurabile al baseline è definita sulla base di tabella 10.1. Per altri casi cfr Eisenhauer *e altri* (2009, p. 235).

10.4.3 Outcome derivabili

Definition 10.4.8 (Migliore risposta globale/complessiva (best overall response)**.** È la migliore risposta registrata dall'inizio sino alla fine del trattamento

Example 10.4.1. Un paziente che abbia SD alla prima valutazione, PR alla seconda e PD all'ultima ha come miglior risposta globale PR

Remark 19. La miglior risposta complessiva in genere è l'indicatore più utilizzato per definire l'attività di un farmaco in una fase 2.

Definition 10.4.9 (Durata della risposta complessiva)**.** Misurata dal momento in cui si ha CR o PR (quale che sia registrata prima) fino alla prima data in cui si sia documentata una PD o una ripresa di malattia

Definition 10.4.10 (Durata della risposta completa)**.** Dal momento in cui per la prima volta si ha CR fino alla prima data di recurrent disease

Definition 10.4.11 (Durata della stabilità di malattia)**.** Misurata dall'inizio del trattamento sino a che è registrata una PD

Capitolo 11

Feasibility/Pilot studies

11.1 Definizioni

In letteratura ci sono almeno due accezioni, soprattutto per fattibilità

- gli studi pilota/di fattibilità sono trial randomizzati piccoli che vengo fatti per vagliare la fattibilità di un trial futuro propriamente dimensionato. La *fattibilità* studiata è quella *dello studio*. L'estensione del consort Eldridge *e altri* (2016) va ovviamente in questa direzione

- nell'ambito della ricerca sulla implementazione di un trattamento si può studiare la fattibilità dello stesso. La *fattibilità* studiata è quella *del trattamento*

11.2 Approccio a Maglietta - (mail mia 12/1/23)

Cari Debora e Silvio, nell'intento di condividere qualcosa di eventualmente utile ai fini dei rapporti con lo statistico del CE (Maglietta, AO PR), vi riporto come sto impostando ultimamente studi (approvati e) dimensionati in base a fattibilità (sicuramente i retrospettivi, forse anche i prospettici, se es si vuole andare avanti al max tot tempo a reclutare).

Per quella che è la mia esperienza non è che al CE guardino necessariamente male/non accettino gli studi dimensionati con quel che c'è/riusciamo a fare; solo, affinché non siano contestati su sto punto, bisogna prestare particolare attenzione al wording degli obiettivi dello studio (a parità di analisi statistica eseguita).

In sintesi, il canovaccio che personalmente adotterò laddove figuri come responsabile statistico (fino a "fail"/prova contraria), è circa il seguente:

- Sample size: (solita frase) fatto secondo fattibilità di rilevazione/in assenza di ipotesi a priori da sottoporre a verifica, variata in base alle circostanze di studio/ispirazione giornaliera

- Obiettivo primario: usare i termini "descrizione"/"esplorazione", NO "analisi"/"valutazione"

- Analisi primaria: analisi inferenziale con magari maggiore enfasi su intervalli di confidenza più che su test/verifica di ipotesi

Cose che mi fanno pensare che questo possa essere un approccio utile:

- la mia esperienza a Parma (Maglietta è stato quello che mi ha sostituito (dopo un po' di prove con altra gente) quando sono venuto a Reggio) e la filosofia ivi dominante: a volte con un feticcio eccessivo per il formalismo/"wording", ben più di quanto un anglosassone (il referee, nostro target di riferimento) abbia in mente;

- due studi osservazionali approvati utilizzando questo approccio (caso1.pdf protocollo finale; caso2.pdf richieste modifiche e protocollo con track changes che è stato in seguito approvato) e uno studio sperimentale prospettico (caso3.pdf, cfr primo/terzo punto) in attesa di valutazione in cui la richiesta di modifica del CE è in linea. Ovviamente, non c'è bisogno di dirlo, ma vi allego il tutto per vs eventuale consultazione/utilizzo no condivisione bla bla bla..

Se può per caso esserVi eventualmente utile mi fa piacere.
un saluto, Luca

Capitolo 12

Fase 3

12.1 Obiettivi

Obiettivo *primario* è valutare l'efficacia del trattamento (nel prolungare la sopravvivenza rispetto allo standard attuale); obiettivi *secondari* sono:

- valutare ulteriormente la *risposta*
- valutare ulteriormente la *tollerabilità/eventi avversi*
- impatto sulla *QOL*

12.2 Classificazione di studi

12.2.1 Studi esplicativi e pragmatici

Remark 20. Questa classificazione presenta profondo legame con il concetto di validità esterna e dirette implicazioni nell'analisi statistica (ITT vs PP).

Definition 12.2.1 (Studi esplicativi (Explanatory trials))**.** Volti alla dimostrazione dell'efficacia del trattamento in *condizioni ideali*

Definition 12.2.2 (Studi pragmatici)**.** Mirati a valutare l'efficacia in un contesto assistenziale reale

12.3 Validità di uno studio

Remark 21. Uno studio è tanto più utile quanto più valido. Ve ne sono due accezioni

Definition 12.3.1 (Validità interna)**.** Grado con cui i risultati di uno studio sono vorretti per i pazienti che ne fanno parte (assenza di bias, dipende dalla conduzione)

Definition 12.3.2 (Validità esterna)**.** Generalizzabilità delle stime ad altri contesti

12.3.1 Validità interna

Remark 22. Uno studio è "valido internamente" se nella pianificazione e conduzione non si sono verificati tre tipi di bias (Sackett):

- Distorsione da selezione (selection bias): sbilanciamento tra i trattamenti nella distribuzione di fattori capaci di influenzare l'end-point.
- Distorsione di valutazione (assessment bias): sbilanciamento tra i trattamenti nel modo in cui i soggetti sono seguiti/valutati nel corso dello studio
- Distorsione di analisi (analysis bias): distorsione che avviene in fase di analisi dei dati

Remark 23. La validità interna è il punto di forza dei disegni sperimentali (soprattutto RCT) rispetto agli osservazionali.

12.3.2 Validità esterna

Remark 24. Uno studio è "valido esternamente" se i suoi risultati (siano essi distorti o meno) sono generalizzabili ad altri contesti.

Remark 25. E' la domanda che implicitamente si pone un clinico quando valuta un RCT altrui per scelte terapeutiche nei confronti di propri pazienti.

Remark 26. La generalizzabilità è legata a:

- criteri di inclusione/esclusione dei pazienti
- setting dove lo studio è condotto (Es Ospedali Hi-Tech vs. arruolamento "sul territorio")
- principi di analisi impiegati nella stima (ITT, PP)

12.4 PICO

12.4.1 Popolazione

12.4.1.1 Criteri di inclusione/esclusione

Definition 12.4.1 (Principio di incertezza (*equipoise*))**.** La scelta della popolazione deve essere fondato su questo: definire i criteri di inclusione/esclusione (su paziente/malattia) individuando coloro per i quali il medico è indeciso su quale possa essere il miglior trattamento.

Remark 27. La popolazione è definita in relazione agli obiettivi: un approccio pragmatico può portare ad allargare le maglie, mentre uno esplicativo a contrarle.

12.4.1.2 Popolazione d'analisi

Per quanto accuratamente condotta la ricerca, è quasi inevitabile un qualche scostamento dal protocollo per alcuni pazienti; ad esempio:

- pazienti inseriti che non rispettano i criteri di eleggibilità
- pazienti nel gruppo di trattamento che non hanno completato lo stesso, o pazienti del gruppo di controllo che hanno ricevuto trattamento sperimentale

In sede di analisi *di efficacia* si pone se includere o meno le info dei pazienti non aderenti al protocollo, e questo dipende dalla tipologia di approccio adottato nello studio (pragmatico o esplicativo)

Approccio pragmatico e popolazione ITT Se l'obiettivo è accertare il beneficio del trattamento in reali condizioni di pratica clinica, le deviazioni dal protocollo ricreano all'interno dello studio condizioni vicine alla pratica clinica; pertanto secondo questo approccio l'analisi deve essere condotta in base al principio dell'intenzione a trattare (ITT); l'insieme di tutti i soggetti randomizzati, aderenti o meno al protocollo, forma la cosiddetta popolazione intention-to-treat e su questa si concentrano le analisi.

Approccio esplicativo e popolazione PP Se si vuole valutare l'efficacia nelle condizioni ideali del trattamento l'analisi si deve limitare a quei pazienti che sono aderenti al protocollo, che nel loro insieme formano la popolazione per protocol. Vanno stabiliti i criteri minimi sufficienti per giudicare il trattamento ricevuto come adeguato.

Popolazione di safety Per quanto riguarda la valutazione *di tollerabilità* del trattamento ci si basa sull'insieme dei pazienti che hanno ricevuto almeno una dose, indipendentemente da qualunque altro fattore, e questa è la *safety population*.

12.4.2 Outcome

Gli outcome più utilizzati (soprattutto in ambito oncologico) sono:

- Overall Survival
- Progression Free Survival
- Time to progression

12.5 Randomizzazione

12.5.1 Alcuni concetti

Definition 12.5.1 (Bliding)**.** Consiste nella non conoscenza del braccio di allocazione da parte di un determinato gruppo di soggetti

Remark 28. Tradizionalmente gli studi possono essere classificati come:

- in *aperto*: nessun soggetto è cieco rispetto all'allocazione
- in *singolo cieco*: il paziente non è a conoscenza del braccio
- *doppio cieco*: sia paziente che medico che somministra/effettua il trattamento non sono a conoscenza del braccio
- *triplo cieco*: paziente, medico che tratta (es oncologo) e valutatore di outcome se differente (es radiologo) non sono a conoscenza dell'outcome

Spesso meglio esplicitare per esteso i soggetti che sottostanno a blinding

Remark 29. Serve per eliminare possibili bias nella valutazione d'outcome:

- il paziente potrebbe avere un effetto benefico non dovuto al trattamento nel sapere di essere stato trattato nel braccio sperimentale;
- chi somministra il trattamento (se non all'oscuro del braccio) potrebbe suggerirlo involontariamente al paziente;
- chi valuta potrebbe inconsciamente o meno biasare le valutazioni.

Remark 30. Negli studi in cui chi somministra il trattamento non è in cieco si pone comunque il nascondere la sequenza di allocazione fino a che non è effettivamente utilizzata

Definition 12.5.2 (Allocation concealment)**.** Mantenimento della segretezza della lista di randomizzazione sino alla richiesta di randomizzazione, per evitare allocazioni non casuali (es pazienti più gravi o amici/parenti)

Remark 31. Per garantirla la lista non deve essere in possesso degli sperimentatori, ma di soggetti terzi alla sperimentazione (ma di supporto alla stessa) oppure in sistemi automatici.

12.5.2 Tipologie

Definition 12.5.3 (Semplice)**.** Quella che si ottiene lanciando una moneta

Remark 32. Semplice da effettuare ma non vi è garanzia su

- bilanciamento complessivo dei bracci a fine studio: sbilanciamenti notevoli su
- bilanciamento in itinere durante lo studio: ad esempio si evita che grossa parte dei trattati (o dei controlli) vengano trattati all'inizio dello studio e i controlli alla fine (o viceversa), esponendo magari a differenti condizioni ambientali le due popolazioni
- bilanciamento tra trattati e controlli su fattori prognostici correlati dell'outcome

Definition 12.5.4 (A blocchi bilanciati)**.** Non si randomizza un singolo paziente ma un blocco bilanciato di pazienti (es in ratio 1:1, due o 4 o 6 pazienti equamente suddivisi tra i due bracci di trattamento)

Remark 33. È good practice effettuare blichi di dimensione variabile, per aumentare l'allocation concealment.

Remark 34. Rispetto alla randomizzazione semplice il blocking risolve lo sbilanciamento tra bracci complessivo (o quello in itinere durante lo studio)

Definition 12.5.5 (Stratificata)**.** Realizzazione di una lista di randomizzazione (semplice o a blocchi) per ogni strato definito da un fattore prognostico/predittivo

Remark 35. Quest'ultima garantisce il fatto che non vi siano sbilanciamenti grossi sui fattori prognostici/predittivi tra casi e controlli.
Nei multicentrici d'obbligo è farla utilizzando il centro come fattore di stratificazione: sia per evitare che tutti i trattamenti siano fatti in un centro e tutti i controlli in un altro, sia perché eventuali problemi con un centro che portino all'esclusione dello stesso non hanno ripercussioni sul bilanciamento complessivo

Remark 36. La considerazione di molteplici fattori di stratificazione contemporaneamente pone il rischio di sbilanciamenti numerici tra trattati e controlli, a maggior ragione se le liste non sono costruite mediante blocchi bilanciati.
Soprattutto su trial di piccole dimensioni questo conduce alla minimizzazione (o randomizzazione adattiva

Definition 12.5.6 (Adattata (minimization))**.** L'idea è, ad ogni nuovo paziente da reclutare, tenere conto delle sue caratteristiche per valutare quale assegnazione introduce minor sbilanciamento; l'assegnazione poi può avvenire in maniera deterministica o probabilistica (es assegnando

12.5.3 Altre questioni

Comparabilità dei gruppi randomizzati Altman (1985) sostiene come:

- a causa della randomizzazione non sia necessario valutare la comparabilità dei gruppi randomizzati (la p sarebbe la probabilità della differenza osservata tra gruppi nell'ipotesi non vi fosse differenza tra gruppi; ma effettivamente non vi è differenza tra gruppi nella popolazione, essendo avvenuta la randomizzazioner)

- può essere necessario aggiustare per sbilanciamenti, soprattutto se le variabili che in seguito alla randomizzazione risultino sbilanciate abbiano una associazione con l'outcome (es età). Il modo per procedere è aggiornare le stime di efficacia per fattori che per sensibilità clinica e buon senso appaiono sbilanciati.

12.6 Definizione dell'effetto del trattamento

Per definire univocamente il *segnale*, ovvero la grandezza attraverso la quale viene valutato, a livello di gruppo e in termini comparativi l'effetto del trattamento sperimentale che si intende studiare occorre effettuare una serie di passaggi (Bacchieri e Della Cioppa, 2004).

1. Definizione degli aspetti della malattia in studio su cui il trattamento vuole incidere; questi sono definiti come *livelli terapeutici*

2. Identificazione di una sola variabile (quindi di un dato processo di misurazione) utile ad identificare l'effetto del trattamento su ogni soggetto, *per ogni livello terapeutico* considerato. Questa variabile è chiamata **endpoint**, o variabile risposta. La sua scelta dovrebbe esser determinata primariamente da motivi clinici.

3. Definizione degli **indicatori di gruppo**: decidere come sintetizzare ciascun end point a livello di gruppo; la scelta di questo dipende dal tipo di end-point e dalla sua distribuzione (la scelta qui è primariamente statistica)

4. Per ciascun indicatore di gruppo bisogna definire il modo in cui vengono comparati matematicamente i gruppi a confronto (differenza o rapporto tra indicatori di gruppo). Questo è il **segnale** ovvero l'effetto complessivo del trattamento

5. Occorre **gerarchizzare i livelli terapeutici**, identificando di conseguenza il segnale primario, sulla quale si giudicherà il confronto dei trattamenti in studio

6. Definizione delle **soglie di rilevanza/non rilevanza** clinica. Nel caso di *studi di superiorità* (volti a dimostrare la superiorità del trattamento sperimentale) bisogna formulare una previsione dell'entità dell'effetto del gruppo sperimentale considerata sufficiente per dichiarare l'effetto biologicamente/clinicamente rilevante. Negli studi di equivalenza e non inferiorità bisogna definire di non rilevanza clinica o margine di equivalenza, cioè la massima differenza che si può tollerare tra gruppi per poter affermare che sono simili (studi di equivalenza) o che quello sperimentale sia non inferiore a quello di controllo (non inferiorità).

12.7 Disegni meno frequenti

Remark 37. Tipicamente i trial prevedono due bracci paralleli in allocazione 1:1; questo garantisce potenza maggiore per un confronto su una singola ipotesi.
Ciò non toglie che in determinate circostanze altri disegni possano tornare comodi o esser necessari.

12.7.1 Disegno a più bracci paralleli

Remark 38. Essendo molteplici le ipotesi da saggiare sono necessari più test statistici, tra loro non indipendenti; occorre pertanto gestire problemi legati alla numerosità campionaria e molteplicità.

12.7.2 Disegno fattoriale

Remark 39. Mediante questo si effettuano 2 o più confronti terapeutici diversi nella stessa ricerca, senza aumentare il numero totale di pazienti

Remark 40. L'assunto che sta alla base di questo disegno è che le terapie abbiano meccanismi d'azione diversi e che non interferiscano tra loro (ossia effetti

additivi, senza interazione); se viceversa si prevede di saggiare l'interazione è necessario dimensionare adeguatamente (cosa che solitamente porta a numerosità elevate)

Remark 41. Ad esempio in un fattoriale 2×2 (due trattamenti per entrambi di due livelli, presente o assente) la lista di randomizzazione contempererà quattro bracci.

12.7.3 Disegno cross-over

Remark 42. In questo disegno ogni partecipante è controllo di se stesso; nella forma più semplice il disegno prevede che ogni paziente riceva entrambi i trattamenti a confronto in successione, secondo un ordine casuale

Remark 43 (Condizioni di applicabilità). Quando:

- le terapie sono di durata breve
- gli effetti si manifestano in poco
- in assenza di trattamento la malattia rimane stabile nel periodo necessario per somministrare i due trattamenti
- gli effetti a lungo termine del primo trattamento sono nulli o comunque scompaiono dopo un periodo di *wash-out* adeguato

Remark 44. Trova scarsa applicazione in oncologia: può esser applicato nel campo delle terapie palliative (es terapia del dolore per confrontare due diversi analgesici)

12.8 Altri studi comparativi (metodologicamente inferiori)

12.8.1 Prima-dopo

Definition 12.8.1 (Disegno)**.** Ad un gruppo di pazienti somministriamo il trattamento e confrontiamo il risultato dopo con quello prima

Remark 45 (Pro). Sono:

- Disegno semplice (conduzione/comunicazione al pz.)
- Minori problemi etici se vi è presunzione di efficacia (tutti trattati, nessun escluso)
- Disegno parsimonioso (minor variabilità, test paired più potenti, sufficienza di un campione più contenuto rispetto ai disegni controllati, minor costo dello studio)

Remark 46 (Contro). Rischio di introduzione di **bias** (la stima di effetto del trattamento è sbagliata) in presenza di:

1. **Dinamica spontanea** della malattia: negli studi prima dopo si assume che in assenza dell'intervento non ci sarebbe stata una variazione della malattia. È un assunto non verificabile; si possono fare congetture (che dipendono dalla patologia in questione) ma rimangono tali.
   La direzione del Bias dipende dalla direzione della dinamica: es se la malattia "migliora" nel tempo e il trattamento mostra un effetto positivo, l'efficacia del trattamento in una stima prima- dopo sarà "gonfiato")
   Il problema è tanto maggiore (e il disegno prima-dopo è meno applicabile) tanto quanto: La malattia presenta una dinamica spontanea veloce e il trattamento è lento a produrre effetti

2. **Variazioni del contesto** dello studio: le variazioni tra il prima e il dopo che incidono sull'outcome (e/o sulla sua valutazione) oltre al trattamento. In primis: Cambio Outcome Assessor, Macchinari, Software di valutazione.
   Rischio più alto tanto è più lungo lo studio; difficile prevedere la direzione del bias.

3. **Regressione verso la media**:

4. **Effetto apprendimento**: se la valutazione dell'effetto del trattamento avviene mediante test di "abilità" del paziente, questi può imparare e performare meglio al test anche in assenza di efficacia del trattamento. (Una soluzione parziale potrebbe essere una fase di pratica pre valutazione basale)

5. **Effetto psicologico** (placebo): solo il fatto di saper di esser curati, ha effetti benefici sulla prognosi della malattia, indipendentemente dall'efficacia clinica del trattamento.
   L'effetto placebo tende ad incrementare la stima di efficacia di un trattamento in un confronto pre-post; a contrario del disegno con gruppo di controllo trattato a placebo, in un pre-post questo bias non riesce ad esser scorporato.

12.8.2 Trial controllati non randomizzati

Remark 47. Rientrano in questa categoria studi eterogenei con

- Controlli paralleli, non selezionati mediante il caso (es iniziale del cognome, giorni sett.)
- Controlli storici o da banche dati (pz trattati in passato)

Remark 48. Rimane dubbia la confrontabilità tra i due gruppi: pertanto considerati da diversi autori/istituzioni metodologicamente di qualità inferiore!

Remark 49. Ergo: se possibile non progettarli. Nel caso, l'analisi non può prescindere da modelli multivariati.

Capitolo 13

Fase 4

Sono gli studi sul farmaco, spesso osservazionali, svolti nell'ambito della pratica clinica dopo che questo è stato messo in commercio.

Razionale Negli studi di fase precedente la numerosità di pazienti nonché il loro follow up è relativamente limitato, pertanto di fatto si riescono a rilevare solamente gli eventi avversi più comuni e che si manifestano entro relativamente poco tempo.

Obiettivi Si pongono la valutazione di (Amadori, 2004):

- efficacia e tollerabilità del trattamento in popolazioni non selezionate
- interazioni con altri farmaci in commercio
- impiego a lungo termine del farmaco (es terapie croniche)
- implicazioni farmacoeconomiche
- impatto sulla qualità della vita

Alla base di diversi obiettivi vi è l'attività di *farmacosorveglianza/farmacovigilanza*, che si basa sulle segnalazioni spontanee delle reazioni avverse, note o impreviste

Disegno Tipicamente si tratta di disegni osservazionali, quindi coorte, caso-controllo o studi cross section. Non servono a dimostrare l'efficacia di un trattamento (nota al termine della fase III) ne hanno il rigore metodologico degli studi sperimentali nel dimostrare associazioni causali.

Parte IV

Studi osservazionali

Capitolo 14

Introduzione osservazionali

14.1 Tipi di studi

Definition 14.1.1 (Coorte)**.** Il ricercatore seleziona due gruppi di soggetti, con e senza la caratteristica in studio (es genere: maschi e femmine) ma senza l'evento di interesse (es tumore). I due gruppi vengono osservati per un dato periodo di tempo e viene confronta l'incidenza dell'evento di interesse.

Remark 50. A parte la caratteristica i due gruppi dovrebbero essere il più possibile omogenei: quindi ben venga prelevare dallo stesso bacino e laddove fattibile formare il campione mediante estrazione casuale dalle due popolazioni (con/senza fattore).

Definition 14.1.2 (Caso controllo)**.** Il ricercatore seleziona due gruppi di soggetti, rispettivamente con (casi) e senza (controlli) l'evento di interesse oggi e per ogni soggetto ricerca nel passato, per un dato periodo di osservazione, informazioni sull'esposizione alla caratteristica studiata. Si stima/confronta l'associazione tra esposizione ed esito.

Remark 51. Alcune considerazioni:

- anche in questo caso si può estrarre casualmente tra casi e controlli per avere rappresentatività delle popolazioni
- la scelta dei controlli è particolarmente critica

Definition 14.1.3 (Cross sectional)**.** Si rileva in un dato momento la presenza di un determinata condizione

Remark 52. Questi studi non sono adatti per inferire causalità in quanto nel caso di esposizione ed esito con una fotografia in un dato momento non si può stabilire cosa è venuto temporalmente/clinicamente prima.

14.2 Bias e confondimento

Di fronte ad un risultato che mostri un'assocaizione tra esposizione e malattia, il ricercatore si deve chiedere:

- può esser dovuta al caso? Nell'analisi statistica e nello specifico nella determinazione dell'ampiezza campionaria si stabilisce come gestire questo tipo di errore (di prima specie)
- può esser dovuta a **bias**/distorsione?
- può esser dovuta a **confondimento**?

Definition 14.2.1 (Bias)**.** Qualunque errore sistematico che porti ad una stima non corretta dell'associazione tra esposizione ed evento

Remark 53. Può verificarsi in qualunque fase dello studio e non riesce ad esser eliminato in sede di analisi statistica

Remark 54. Alcuni esempi sono

- distorsione da *selezione*: l'errore sistematico riguarda la selezione dei soggetti da includere nello studio
- distorsione di *osservazione*: difficoltà a ricordare, influsso dell'intervistatore, perdita di soggetti al follow up sistematicamente diversi da coloro che rimangono,

Definition 14.2.2 (Confondimento)**.** Si ha se l'entità e a volte la direzione dell'associazione tra caratteristica ed evento è modificata dalla presenza di un terzo fattore che è contemporaneamente:

- associato con la caratteristica
- distribuito in modo sbilanciato tra i gruppi a confronto

Remark 55. Nel bias lo sbilanciamento sistematico tra i gruppi è indotto dal disegno, mentre nel confondimento è insito nel fenomeno.

Remark 56. A volte il termine confondimento è usato impropriamente come sinonimo di associazione spuria, che dai medici è intesa come associazione "non vera" mentre dagli statistici per indicare una associazione tra due fattori in realtà causata da un terzo (es correlazione tra consumo di cioccolata a livello nazionale e premi nobel è una associazione spuria per la presenza del PIL procapite)

Capitolo 15

Coorte

In uno studio di coorte gli individui sono seguiti per un dato periodo di tempo per monitorare il loro stato di salute e nello specifico, uno specifico evento (es morte, ammalarsi ecc).

15.1 Disegni

Disegno base L'approccio più semplice consiste nel selezionare due gruppi di persone al baseline; il primo consiste in persone che possiedono un qualche attributo di interesse (es esposizione ad un fattore di rischio) mentre l'altro no, al fine di studiare l'eccesso di mortalità/morbidità associato al fattore di rischio. A volte si riesce a studiare la popolazione completa esposta al fattore di rischio; quando invece se ne si può studiare solo un campione è meglio sceglierlo in maniera casuale.
Ogni soggetto che entra nello studio deve essere libero da evento analizzato (morte/malattia) al baseline proprio perché si sta cercando di capire la causalità associata al fattore di rischio.

Disegno con campione unico Prendiamo tot pazienti in blocco unico ed analizziamo l'associazione in questi:

- pro l'informazione sulla stratificazione in pazienti con e senza non è necessaria a priori e in generale è logisticamente più facile da fare;
- pro: è un disegno comodo soprattutto se con un unico studio si stanno analizzando il contributo di più fattori di rischio contemporaneamente;
- pro: incidentalmente si possono fornire stime della prevalenza del fattore di rischio (se si prende un campione casuale della popolazione);
- contro: la distribuzione del fattore di rischio potrebbe essere notevolmente e originare confronti statistici poco potenti.

Disegno con confronto esterno Arruolare non due bracci ma solo il braccio di coloro che hanno l'esposizione di interesse e confrontare l'esito con una popolazione esterna, spesso e volentieri la popolazione generale. Costa meno

ma la possibilità di bias è indiscussa; la popolazione generale può differire dal campione non solo per il fattore considerato ma per molto altro che si fatica a controllare.

Coorte retrospettiva Magari svolto su dati già disponibili, è comunque necessario:

- garantire che all'inizio del follow up tutti i pazienti fossero disease free
- essere confidenti sul fatto che il dataset collezionato per altri fini possa rispondere anche a quello dell'indagine (es valutare definizioni adottate ed eventuali cambiamenti nel tempo, completezza dei dati eccetera)

15.2 Pro/contro

Vantaggi dello studio di coorte:

- la prospetticità, unita al fatto che i pazienti siano liberi da evento all'inizio del follow up, è ideale per suggerire la causalità delle associazioni
- l'effetto di un singolo fattore di rischio può essere valutato su più malattie (outcome)

Svantaggi:

- costosi (es se l'outcome necessita di esami) e lunghi (in particolar modo per malattie, outcome, lenti a svilupparsi) se condotti in maniera prospettica
- non adatti per outcome rari: in tal caso potremmo dover arruolare un campione enorme (al fine che si verifichino eventi) o monitorare per tempi molto lunghi. Per questo si possono scegliere outcome intermedi (surrogati) più frequenti accettandone tutte le critiche associate (validazione, non è un outcome hard eccetera);
- study effect (assenza di blinding): se il paziente sa di essere in uno studio potrebbe comportarsi diversamente dal caso in cui non lo sapesse. E le associazioni derivanti potrebbero risultare falsate o non rappresentative della realtà
- l'esposizione al fattore di interesse potrebbe variare nel corso dello studio (può essere considerato in analisi)
- pazienti che si ritirano o persi al fup: gestibili a patto che il censoring non sia informativo

15.3 Strategie d'analisi

Se ciascuno nella coorte viene arruolato nello stesso momento e viene seguito per lo stesso ammontare di tempo si possono analizzare i dati in maniera binaria classica (tipicamente). Questa è una **fixed cohort**.

Nel caso il follow up non sia uguale per tutti (**variable cohort**) meglio approcciarsi con survival analysis oppure analisi person-years (poisson)

Capitolo 16

Caso controllo

Parte V

Studi di diagnostica

Capitolo 17

Introduzione all diagnostica

17.1 Introduzione

La diagnostica è volta a raccogliere ed analizzare informazioni per determinare la *condizione del paziente*. La ricerca diagnostica è motivata dal fatto che non sempre un metodo di diagnosi efficace esista; altresì si può esser interessati a sviluppare nuove metodologie diagnostiche se quelle esistenti sono particolarmente costose, invasive, tossiche ecc.

Dal punto di vista statistico, la metodologia della diagnostica, volta prevalentemente alla classificazione, si applica anche ad altri problemi in medicina; ad esempio anche la prognosi può esser considerata una diagnosi (si cerca di indovinare l'esito, non la diagnosi). Allo stesso modo i test di screening sono test diagnostici (in popolazioni con prevalenze tipicamente rare).[1]

La performance di un test diagnostico può esser valutata a livelli progressivi; qualità del dato/immagine (aspetto tecnico), accuratezza diagnostica, effetto sulle decisioni di trattamento, impatto sull'outcome del paziente, costo/efficacia per la società (aspetti scientifici). In ogni modo per esser efficace a livelli superiori, deve esser efficace ad un livello inferiore. Qui studiamo il primo degli aspetti scientifici che si pongono ovvero lo studio dell'accuratezza diagnostica del test.

Lo sviluppo di un nuovo test, seguendo Pepe (2004), è utile in presenza delle seguenti condizioni (riguardanti la malattia, il suo trattamento, e il test stesso):

1. la malattia dovrebbe esser grave o potenzialmente grave (se il diagnosticarla non provoca un risparmio di quantità/qualità di vita, non è costo efficace)

2. la malattia dovrebbe esser abbastanza prevalente nella popolazione target (non rarissima)

3. la malattia dovrebbe esser trattabile (se no è inutile testare)

4. il trattamento dovrebbe esser disponibile per coloro che hanno un test positivo (es non troppo costoso, se no è comunque inutile)

[1] Si tratta di studi in cui si desidera, per diversi fini, classificare i pazienti; che la classificazione riguardi uno stato di malattia o un esito, dal punto di vista strettamente di calcolo è lo stesso.

5. il test non dovrebbe causare (troppi) effetti avversi al paziente
6. il test dovrebbe classificare accuratamente gli individui malati da quelli sani: questa è l'*accuratezza diagnostica*

17.1.1 Disegni di ricerca principali

Sebbene nel seguito ci si concentri sullo studio di accuratezza, esistono differenti tipologie di studi nell'ambito della diagnostica ed essi dipendono sostanzialmente dall'obiettivo/domanda della ricerca.

Se l'obiettivo è valutare la *capacità discriminatoria di un test nell'individuare sani e malati*, lo studio principale è quello di **accuratezza diagnostica** in cui viene indagata la relazione tra risultato del test e presenza della condizione. Hanno tipicamente un disegno cross-sectional.

Se si vuole valutare l'*impatto dell'impiego del metodo diagnostico nella pratica clinica e sulla prognosi* dei pazienti, esistono diversi disegni, tra i quali l'**RTC diagnostico** è metodologicamente quello più forte. Anche disegni di coorte (guardo chi è stato diagnosticato con che cosa, e ne seguo poi il fup) e caso controllo (guardo malati e non malati oggi e torno indietro nel ricercare procedure diagnostiche di mio interesse) sono impiegati

La *sintesi delle evidenze disponibili* può esser affrontata mediante **revisioni sistematiche** (che forniscono una valutazione globale della procedura diagnostica).
Sempre in questa categoria, esistono altresì **studi di economia sanitaria** volti alla costo efficacia delle procedure diagnostiche. Vengono infine svolti studi sullo sviluppo di **equazioni di predizione** (o *clinical prediction rules, CPR*) diagnostica/prognostica, che possano servire al clinico per effettuare delle decisioni nella propria pratica.

17.1.2 Architettura della ricerca diagnostica

Qui mettere buntinx pag 20

17.2 RCT Diagnostici

Knotterus capitolo 4, TODO

Capitolo 18

Studi di accuratezza diagnostica

18.1 Introduzione

Gli **studi di accuratezza diagnostica**, che costituiscono il nostro focus, sono ricerche volte a determinare la *capacità di un test di discriminare* tra pazienti con e senza una determinata condizione clinica.

Qualunque dato (caratteristiche del paziente, segni e sintimi, esami fisici, storia del pz oppure test di laboratorio) può esser in linea teorica considerato come "test", e verificarne la capacità discriminatoria.

Sostanzialmente un test accurato classifica i soggetti correttamente in relazione alla loro condizione. Test inaccurati fanno si che (troppi) individui malati vengano classificati come sani e viceversa. I primi non vengono trattati come dovrebbero, i secondi possono ricevere procedure non necessario, spesso invasive e costose.
Pertanto prima che un test possa esser utilizzato in pratica clinica, la sua accuratezza diagnostica deve esser valutata.

Per poter valutare la capacità classificatoria di un test, i suoi risultati debbono esser confrontati alternativamente con una diagnosi standard di riferimento; un **gold standard**, ovvero un altro test che classifichi in assoluta correttezza la condizione del pz non necessariamente esiste[1]. Nel caso, occorre impiegare un *reference standard* che approssimi il gold standard al meglio.
In merito alla presenza di un gold standard, va comunque precisato che gli studi di accuratezza diagnostica non sono meramente studi di "agreement" tra due misure. Interpretare ogni differenza tra lo standard e il test sperimentale come un fallimento di quest'ultimo non è necessariamente corretto (soprattutto se lo standard non è gold) . Può esser peraltro che due metodi misurino concetti lievemente differenti.

[1] Esempi di gold standard sono i report dall'autopsia, i rilievi chirurgici, risultati dell'analisi patologica su campioni

	Reference standard (D)	
Test (Y)	Presente (1)	Assente (0)
Presente (1)	$s_1 = Tp$	$r_1 = Fp$
Assente (0)	$s_0 = Fn$	$r_0 = Tn$

Tabella 18.1: Matrice 2x2

18.2 Misure di accuratezza diagnostica

18.2.1 Dati dicotomici

Se ci riferiamo ad un test che fornisce nativamente risultati in forma dicotomica (del tipo malattia assente o presente) e confrontando i suoi risultati con un gold o reference standard. Arriviamo a definire la tabella 18.1, e i valori (conteggi) di cui è composta:

- **Tp**: true positive. Casi in cui il test *individua (Y=1) correttamente (D=1)* la presenza di malattia (effettivamente *presente*).
- **Fp**: false positive. Casi in cui il test *individua (Y=1) fallacemente (D=0)* la presenza di malattia (effettivamente *assente*).
- **Fn**: false negative. Casi in cui il test *esclude (Y=0) erroneamente (D=1)* la malattia (effettivamente *presente*).
- **Tn**: true negative. Casi in cui il test *esclude (Y=0) correttamente (D=0)* la malattia (effettivamente *assente*).

18.2.1.1 Sensibilità, specificità

La **sensibilità** (*sensitivity* o *true positive fraction*, TPF) è l'abilità del test di individuare la malattia quando è presente:

$$Sens = TPF = P(Y = 1|D = 1) = \frac{Tp}{Tp + Fn} \tag{18.1}$$

La **specificità** (*specificity* o *true negative fraction*, TNF) è l'abilità del test di escludere la condizione in pazienti che non ne siano affetti.

$$Spec = TNF = P(Y = 0|D = 0) = \frac{Tn}{Tn + Fp} \tag{18.2}$$

Un test perfetto (che non commette falsi positivi o negativi) avrebbe:

$$Sens = Spec = 1 \tag{18.3}$$

Un test non in grado di discriminare (equivalente a tirare una moneta sia nel caso il paziente abbia o non abbia la malattia):

$$Sens = Spec = 0.5 \tag{18.4}$$

Allo stesso modo possono esser definite le seguenti misure:

- FNF, o *false negative fraction*, come

$$FNF = P(Y = 0|D = 1) = \frac{Fn}{Tp + Fn} = 1 - TPF \tag{18.5}$$

- FPF, o *false positive fraction*

$$FPF = P(Y = 1|D = 0) = \frac{Fp}{Tn + Fp} = 1 - TNF \tag{18.6}$$

La **prevalenza** della malattia nella popolazione è definita come:

$$Prev = \rho = P(D = 1) = \frac{Fn + Tp}{n} \tag{18.7}$$

dove $n = Fn + Tp + Tn + Fp$.
L'**accuratezza** è definita come la probabilità che il test azzecchi la diagnosi;

$$Acc = P(Y = D) = \frac{Tp + Tn}{n} \tag{18.8}$$

La **probabilità di missclassification** è il complemento a 1 dell'accuratezza e può esser scritta come funzione della prevalenza della malattia, di FNF e FPF

$$Miss = P(Y \neq D) = \rho(FNF) + (1 - \rho)FPF \tag{18.9}$$

Le misure complessive di accuratezza/missclassification non sono generalmente considerate una sintesi adeguata dell'accuratezza diagnostica di un test medico. Piuttosto, bisognerebbe riportare sia sensibilità e specificità (ovvero i complementi a 1 di FNF e FPF) separatamente poiché:

- costi e conseguenze dei due tipi di errori possono esser molto differenti (falsi negativi perdono le cure necessarie, falsi positivi si sottopongono a terapie non necessarie). Solitamente i falsi positivi sono giudicati meno gravi dei falsi negativi;

- inoltre mentre sensibilità e specificità non dipendono dalla prevalenza della malattia nella popolazione, misure di sintesi come accuratezza e missclassification vi dipendono largamente.

18.2.1.2 Valori predittivi

Guardando alla tabella per riga, come alternativa alle misure di accuratezza che fondano il proprio denominatore sullo stato della malattia vi sono quelle che lo fondano sul risultato del test, che pongono enfasi su quanto bene i risultati del test prevedano lo stato effettivo di malattia.
Il valore predittivo positivo (**positive predictive value** o PPV) esprime la probabilità che un test positivo riesca ad individuare correttamente un soggetto avente la condizione:

$$PPV = P(D = 1|Y = 1) = \frac{Tp}{Tp + Fp} \tag{18.10}$$

Il valore predittivo negativo (**negative predictive value** o NPV) esprime la probabilità che un test negativo riesca ad escludere correttamente che un soggetto sia senza la condizione:

$$NPV = P(D = 0|Y = 0) = \frac{Tn}{Tn + Fn} \tag{18.11}$$

Un test perfetto prevederà la presenza di condizione in maniera perfetta, avendo

$$PPV = NPV = 1 \tag{18.12}$$

Invece un test inutile che non porta informazione sulla presenza effettiva di malattia sarà tale che

$$PPV = \quad P(D = 1|Y = 1) = P(D = 1) = \rho \tag{18.13}$$
$$NPV = \quad P(D = 0|Y = 0) = P(D = 0) = 1 - \rho \tag{18.14}$$

I valori predittivi non sono usati per quantificare la performance intrinseca del test, perchè non hanno come denominatore la condizione effettiva e dipendono tra l'altro dalla prevalenza della malattia. Piuttosto sono impiegati per quantificare il valore clinico del test (a questi sono maggiormente interessati paziente e caregiver).

In generale i valori predittivi dipendono sia dalla performance intrinseca del test (sensibilità, specificità) che dalla prevalenza della malattia. Si può scrivere

$$PPV = \frac{\rho \cdot Sens}{\rho \cdot Sens + (1 - \rho) \cdot (1 - Spec)} \tag{18.15}$$

$$NPV = \frac{(1 - \rho) \cdot Spec}{(1 - \rho) \cdot Spec + \rho \cdot (1 - Sens)} \tag{18.16}$$

La dimostrazione è un'applicazione del teorema di Bayes.

18.2.1.3 Uso di R - Stima accuratezza

Esempio 2.1 pag 17 della pepe

```
cass <- read.csv("~/dataset/pepe/est1.csv")

## Warning in file(file, "rt"):  cannot open file '/home/l/dataset/pepe/est1.csv':
No such file or directory
## Error in file(file, "rt"):  cannot open the connection

names(cass) <- c("cad","est","cph")

## Error:  object 'cass' not found

## Cad: Coronary artery disease
## EST: Exercise Stress test
## CPH: Chest pain history
dim(cass)

## Error:  object 'cass' not found
```

```
head(cass)

## Error:  object 'cass' not found

library(lbdiag)

## Error in library(lbdiag):  there is no package called 'lbdiag'

da(test=cass$est, refstd=cass$cad)

## Error in da(test = cass$est, refstd = cass$cad):  could not find
function "da"
```

Il commento si può fare per colonne o per righe della tabella:

- per "colonne", la prevalenza della malattia (prev) è molto alta, quasi il 70%. La sensibilità è circa dell'80 mentre la specificità è del 74%. Il test manca il 20% dei malati e erroneamente identifica come malati il 26% dei sani. La decisione di utilizzare il test necessiterebbe di prendere in considerazione il rischio e il beneficio di procedure diagnostiche addizionali e/o trattamenti associati ad una diagnosi positiva, cosiccome le conseguenze delle mancate diagnosi.
- per "righe", circa il 63% della popolazione viene diagnosticata positiva al test. Tra i soggetti che hanno test positivo la stragrande maggioranza (88%) ha effettivamente la malattia. I positivi quindi sono verosimilmente all'ultima fase diagnostica dato che hanno una elevata probabilità di esser malati
  Al contrario il 39% dei soggetti che vengono diagnosticati negativi hanno in realtà la malattia. Pertanto potrebbe esser di interesse la ricerca di ulteriori tecniche diagnostiche da applicare a coloro che hanno un test negativo, al fine di identificare coloro che in realtà hanno la malattia.

18.2.1.4 Uso di R - Inferenza accuratezza

Esempio 2.3 pag 22 della pepe: intervalli di confidenza delle stime

```
## Sens 95CI
binom.test(815,1023)$conf.int

## [1] 0.7706868 0.8209461
## attr(,"conf.level")
## [1] 0.95

## Spec 95CI
binom.test(327,442)$conf.int

## [1] 0.6962660 0.7801285
## attr(,"conf.level")
## [1] 0.95

## PPV 95CI
binom.test(815,930)$conf.int
```

```
## [1] 0.8534513 0.8968204
## attr(,"conf.level")
## [1] 0.95

## NPV 95CI
binom.test(327,535)$conf.int

## [1] 0.5684514 0.6527429
## attr(,"conf.level")
## [1] 0.95

## Joint 95 Confidence interval Sens spec pag 23
binom.test(815,1023, conf.level = 0.975)$conf.int

## [1] 0.7669239 0.8242471
## attr(,"conf.level")
## [1] 0.975

binom.test(327,442, conf.level = 0.975)$conf.int

## [1] 0.6900029 0.7855267
## attr(,"conf.level")
## [1] 0.975
```

18.2.1.5 Molteplicità di focus diagnostici entro paziente

Nel caso in cui la **presenza di malattia possa esser diagnosticata più volte per ogni paziente** (es se si vuole diagnosticare se una lesione è maligna, ma le lesioni in un pz possano esser molteplici), allora si può costruire la tabella a livello di lesione/polipo, anzichè al livello di paziente. Per farlo i conti vanno fatti a livello di singola lesione, per cui un Tp sarebbe una lesione che il test ha correttamente individuato come positiva.
Nel caso invece si desideri di *passare da livello di lesione a quello di paziente*:

- per il pz i-esimo il reference standard è impostato a malato se il pz stesso ha *almeno una lesione maligna*
- per il pz i-esimo il reference standard è impostato a sano se il pz stesso non ha *neanche una lesione maligna*

18.2.2 Dati quantitativi

Molti test forniscono una misura numerica.
In questo setting, il valore restituito dal test varia da 0.03 a 0.58 nei pazienti con frattura, da 0 a 0.13 nei pazienti senza frattura.
Siamo interessati a diagnosticare la frattura sulla base del test numerico, meno invasivo

```
gap <- c(0.58, .41, .18, .15,.15,.10,.07,.07,.05,.03,
         .13,.13,.07,.05,.03,.03,.03,0,0,0)
```

```
fracture <- c(rep(1,10),rep(0,10))
hv <- data.frame(fracture,gap)
hv[hv$fracture==1 ,"gap"]

##  [1] 0.58 0.41 0.18 0.15 0.15 0.10 0.07 0.07 0.05 0.03

hv[hv$fracture==0 ,"gap"]

##  [1] 0.13 0.13 0.07 0.05 0.03 0.03 0.03 0.00 0.00 0.00
```

Per poter determinare sensibilità e specificità dobbiamo scegliere una **soglia** e *specificare se sopra di essa il test sia considerato positivo o negativo*. Nel nostro caso, data le conoscenze cliniche, specifichiamo che un valore di test superiore a 0.05 (decision threshold) sia considerabile come un test positivo (quindi predittivo della presenza di frattura).

Avremmo potuto scegliere qualsiasi valore come soglia per determinare l'esito del test. Questo fa si che a differenti soglie corrispondano differenti performance diagnostiche

18.2.3 Dati ordinali

Parte VI

Revisioni sistematiche

Capitolo 19

Introduzione

19.1 Definizioni e risorse utili

Remark 57. Qui ci si basa per lo più su Higgins *e altri* (2019)

Definition 19.1.1 (Revisione sistematica (RS))**.** Studio che sintetizza l'evidenza empirica che rispetta pre-determinati criteri di eleggibilità, al fine di rispondere ad una specifica domanda di ricerca

Remark 58. Le RS si contrappongono alle Revisioni Narrative

Definition 19.1.2 (Revisione narrativa)**.** Revisione dove l'autore non esplicita come ha scelto gli studi da includere o meno (scegliendo mediante esperienza personale o contatti/conoscenze)

Remark 59 (Caratteristiche di una revisione sistematica)*.* Rispettivamente:

- obittivo/set di obiettivi chiari, con criteri di eleggibilità degli studi predeterminati
- metodologia riproducibile
- ricerca sistematica che identifichi tutti gli studi che rispettino i criteri di eleggibilità
- valutazione validità dei risultati degli studi selezionati (es valutazione rischio di bias)
- sintesi sistematica di caratteristiche e risultati degli studi inclusi

Remark 60 (Fasi di una revisione sistematica)*.* Rispettivamente:

- Dichiarare gli obiettivi e i criteri di eleggibilità
- Ricerca di studi che sembrano rispettare i criteri di eleggibilità
- Porre in dataset le caratteristiche degli studi identificati e valutarne la qualità metodologica
- applicare i criteri di eleggibilità e giustificare ogni esclusione

- raccogliere i dati da analizzare
- analisi dei risultati degli studi eleggibili, usando meta-analisi se appropriato e possibile
- analisi di sensibilità e analisi sottogruppi, se appropriato e possibile
- preparazione del report

Definition 19.1.3 (Metanalisi (MA))**.** Metodo statistico di sintesi dei risultati di differenti studi.

Remark 61. Non necessariamente tutte le revisioni sistematiche presentano anche una metanalisi: può essere talvolta non utile o appropriato combinare quantitativamente le informazioni derivanti da studi fra loro troppo diversi o eterogenei.

Remark 62. L'analisi della coerenza e qualità di un insieme di studi è una delle caratteristiche più importanti di una RS

Remark 63. Si può ritenere appropriata la combinazione quantitativa dei dati di studi diversi (MA) quando:

- più di uno studio ha stimato l'effetto del trattamento/terapia
- le differenze fra gli studi in termini di pazienti, interventi e caratteristiche del setting sono minimo o comunque non permettono a priori, di ipotizzare un impatto sull'outcome
- l'outcome nei diversi studi è stato misurato in maniera simile
- gli autori degli studi primari riportano i dati numerici necessari per effettuare la combinazione

Remark 64 (Risorse per il lettore interessato). In alcun ordine in particolare:

- Cochrane Handbook for Systematic Reviews of Interventions: manuale per revisioni su studi di efficacia
- Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy: manuale per studi di accuratezza diagnostica
- Liberati *e altri* (2009): guideline scrittura revisione sistematica e metanalisi
- RevMan e RevMan Web: software per la conduzione di revisioni della Cochrane collaboration
- Altri tool cochrane

19.2 Preparazione e mantenimento di una review (Cochrane)

19.2.1 Protocollo

Dato che lo svolgimento di una revisione necessita di diverse decisioni, e volendo evitare che queste siano indirizzate dai risultati degli studi inclusi, analogamente

a quanto avviene per i singoli studi (es rct) la pubblicazione del protocollo (anch'esso fattibile mediante revman e pubblicabile nel Cochrane database of systematic reviews) di una review prima dello studio ne limita i possibili bias

19.2.2 Review team

È essenziale che la revisione sia preparata da almeno due persone, per limitare possibili errori nella selezione degli studi

19.3 Domanda della ricerca e criteri inclusione

Per circoscrivere la ricerca bisogna aver chiaro

- PICO degli studi: nel caso non siano interventistici (es braccio singolo), comunque i gruppi da includere
  - per la popolazione: occorre definire precisamente la malattia e i suoi criteri di diagnosi, oltre eventualmente ad altri aspetti quali eta, sesso, setting di cura ecc. Sui casi dubbi di studi si può evitare di escludere e decidere in un secondo momento, giustificando
  - intervento e comparazione
  - per gli outcome, vanno scelti quelli essenzioali per il decision making (non più di sette) che vanno divisi in primary (non più di tre) e secondary (i rimanenti).
    Inoltre il reporting dell'outcome o meno nello studio non dovrebbe essere una causa di esclusione dello stesso (perché non fornisce dati utilizzabili, es nella metanalisi)
- tipi di studi da sintetizzare: prospettico/retrospettivo, con/senza gruppo di controllo, sperimentale o osservazionale.

Remark 65. Sulla tipologia di studi ci si riferirà a quanto si usa in base al quesito clinico: es per l'eziologia si procederà con sintesi di caso controllo (o coorte), diagnosi mediante studi di coorte (o trasversali o rct), terapia mediante rct, prognosi mediante coorti)

Remark 66 (MA di studi osservazionali). Poiché le stime provenienti da studi osservazionali possono esser distorte, è verosimile che la loro combinazione sia una valutazione distorta. Ciononostante in letteratura esistono esempi di MA di studi osservazionali: i risultati sono da interpretare con cautela.
Al momento le MA di studi osservazionali non sono raccomandate in letteratura, mentre è assolutamente utile l'esecuzione di RS di studi con questo disegno; queste, pur non portando a combinazioni quantitative formali, possono fornire info sull'effetto del trattamento e possono esser utili per indirizzare la ricerca futura

19.4 Ricerca degli studi

Le revisioni richiedono un'approfondita e riproducibile ricerca per identificare più studi rilevanti possibili:

- coinvolgere personale esperto nella ricerca
- La ricerca solo su pubmed (MEDLINE) è generalmente considerata insufficiente
- EMBASE va considerato
- per i trial consultare il motore della cochrane (CENTRAL); sugli studi di efficacia basati su RCT ci si può forse limitare a questo
- se possibile, oltre ai primi 3, considerare altri db di letteratura: es specifici per soggetto, dottorati e dissertazioni, letteratura grigia (abstract congressi)

È importante *documentare* le ricerche effettuate; per questo e per maggiori info sulla ricerca vedere il capitolo 6 di Higgins *e altri* (2019).

19.5 Selezione degli studi e collezione dati

19.5.1 Selezione studi

Il tipico processo per la selezione degli studi è il seguente:

1. integrare i risultati delle ricerche (dai vari database) e rimuovere i record duplicati (stessi articoli)
2. esaminare titolo e abstract per rimuovere articoli irrilevanti (ma senza esagerare)
3. ottenere il pdf dell'articolo (per quelli rilevanti)
4. nel caso per un unico studio siano disponibili più articoli (determinarlo serve un po di lavoro da detective basato sui nomi degli autori, location e setting di studio, numero di partecipanti e dati di baseline, data/durata dello studio; mal che vada contattare gli autori) unificare il record per non considerare gli stessi risultati più volte
5. leggere gli articoli interi per determinare la complicance con i criteri di eleggibilità
6. decidere se includere o meno e procedere alla data collection sui risultati

È considgliabile che lo step 2 e a maggior ragione il 5 sia effettuato da 2 o più autori indipendentemente; consigliabile che nel team vi siano anche revisori non esperti della materia poiché gli specialisti possono avere bias sulla scelta di cosa includere o meno.

Eventuali disagreement su cosa includere o meno possono essere analizzati mediante K di Cohen (es se si classifica lo studio come includere si/no/dubbio) e nel caso di bassi valori (poco agreement) possono rivelare la n ecessità di rivedere assieme i criteri di eleggibilità. Comunque il disagreement può essere solutamente risolto mediante discussione del caso o dall'arbitrato di un terzo (o infine contattando l'autore nei casi dubbi di inclusione per verificare se l'articolo rispetta i criteri)

19.5.2 Dati da raccogliere

L'estrazione dati sarebbe meglio farla in doppio, autonomamente per evitare errori anche qui

19.5.2.1 Elementi

Vedere il manuale (capitolo 7 per approfondimenti sui singoli campi).
Source:

- Study ID (created by review author).
- Report ID (created by review author).
- Review author ID (created by review author).
- Citation and contact details.

Eligibility

- Confirm eligibility for review.
- Reason for exclusion.

Methods

- Study design.
- Total study duration.
- Sequence generation*.
- Allocation sequence concealment*.
- Blinding*.
- Other concerns about bias*.

Participants

- Total number.
- Setting.
- Diagnostic criteria.
- Age.
- Sex.
- Country.
- [Co-morbidity].
- [Socio-demographics].
- [Ethnicity].
- [Date of study].

Interventions

- Total number of intervention groups.
- For each intervention and comparison group of interest:
- Specific intervention.
- Intervention details (sufficient for replication, if feasible).
- [Integrity of intervention].

Outcomes

- Outcomes and time points (i) collected; (ii) reported*.

For each outcome of interest:

- Outcome definition (with diagnostic criteria if relevant).
- Unit of measurement (if relevant).
- For scales: upper and lower limits, and whether high or low score is good.

Results

- Number of participants allocated to each intervention group.

For each outcome of interest:

- Sample size.
- Missing participants*.
- Summary data for each intervention group (e.g. 2×2 table for dichotomous data; means and SDs for continuous data).
- [Estimate of effect with confidence interval; P value]z.
- [Subgroup analyses].

Miscellaneous

- Funding source.
- Key conclusions of the study authors.
- Miscellaneous comments from the study authors.
- References to other relevant studies.
- Correspondence required.
- Miscellaneous comments by the review authors.

19.5.2.2 Stime per misure dicotomiche

I quattro numeri della tabella 2×2.

19.5.2.3 Stime per variabili quantitative

Media, sd e numerosità di ciascun braccio.

19.5.2.4 Stime per analisi di sopravvivenza

Servono le stime dei log hazard ratio e i relativi standard error. Se hanno usato cox bene, alternativamente (spesso) necessario ottenere i dati originali (a meno che non si percorrano strade più complesse, vedi 7.7.6 per dettagli)

19.6 Valutazione del rischio di bias negli studi inclusi

Bias è la deviazione sistematica della stima dal valore vero (non va confuso con la imprecisione delle stime); includere studi con stime affette da bias inficia la revisione.

Uno studio potrebbe essere stato eseguito coi più alti standard possibili, ma ciò nonostante non essere immune a rischi di bias (es un trial dove non è possibile effettuare il doppio cieco).

19.6.1 Fonti di bias nei clinical trial

L'affidabilità dei risultati di uno studio randomizzato dipendono dal fatto che le fonti di bias siano state evitate:

- *selection* bias: differenze sistematiche tra le caratteristiche di baseline dei gruppi che sono comparati; limitato dalla randomizzazione, se funziona bene
- *performance* bias: differenze sistematiche tra gruppi nella cura fornita e/o nell'esposizione a fattori oltre all'intervento di interesse; limitato da blinding dei partecipanti e del personale coinvolto nello studio
- *detection* bias: differenze tra gruppi in come l'outcome è determinato; limitato da blinding dei valutatori
- *attrition* bias: differenze tra gruppi nei ritiri dallo studio e quindi nella completezza dei dati
- *reporting* bias: differenze tra risultati riportati e non riportati (i risultati non positivi vengono più difficilmente riportati e ciò inficia i dati utilizzabili nella revisione)

19.7 Mantenimento della Revisione

Le revisioni sistematiche debbono essere mantenute (indicativamente ogni 2 anni) al fine di mantenere l'evidenza più aggiornata sugli effetti degli interventi. Le ragioni per un aggiornamento:

- nuovi studi disponibili

- strumenti migliori per la caratterizzazione di sottogruppi (es genetica), nuovi trattamenti, nuove misure di outcome
- nuove metodologie per la conduzione di revisioni sistematiche

Capitolo 20

Analisi dei dati e metanalisi

20.1 Outcome e misure di efficacia

Remark 67. Per outcome:

- dicotomici si usa OR o RR
- quantitativi si usa la differenza delle medie tra bracci o la differenza delle medie (tra bracci) standardizzata (ossia divisa la deviazione standard complessiva dell'outcome), altre volte detto *effect size*. Si ricorre al primo se tutti gli studi usano la stessa scala di misura dell'outcome, o il secondo se l'efficacia è misurata su scale diverse in studi diversi
- per tassi si usa il rate ratio
- per analisi di sopravvivenza si usa l'hazard ratio di regressioni di cox univariate

Remark 68. Le misure di effetto espresse come rapporto (OR, RR, rate ratio e HR) sono solitamente log trasformate (per essere simmetriche e centrate sullo 0

Remark 69. Allo stesso modo il display grafico di meta analisi effettuate su misure a rapporto (non precedentemente logaritmizzate) usano solitamente una scala logaritmica per affinché gli intervalli di confidenza siano simmetrici

20.2 Eterogeneità

Remark 70. I risultati di una MA sono interpretabili e utili quando gli studi che sono stati combinati erano sufficientemente comparabili, ossia *poco eterogenei*.

Remark 71. Vi sono due componenti principali dell'eterogeneità

Definition 20.2.1 (Eterogeneità clinica)**.** Si analizzano studi con differenze su pazienti, trattamento, setting di studio e outcome

Definition 20.2.2 (Eterogeneità metodologica)**.** Differenze su disegni (sperimentale/osservazionale) qualità e tipo di analisi (ITT/per protocol)

Remark 72. In generale se gli intervalli di confidenza dei singoli studi mostrano poco overlap, può essere che vi sia eterogeneità. È una condizione poco auspicabile soprattutto se gli studi sono discordi sulla direzione dell'effetto (alcuni dicono che il trattamento sperimentale sia meglio, altri che sia peggio).

Remark 73 (Test). Formalmente si dispone di un test che ha come ipotesi nulla quella di omogeneità degli studi: se la nulla viene rifiutata ($p < 0.05$) siamo in una situazione di eterogeneità.
Dato che è un test poco potente, ogni tanto si abbassa la soglia accettando come eterogenea una situazione con $p < 0.1$.
Il test di eterogeneità (implementato da revman) è

$$Q = \sum W_i(\theta_i - \theta)^2$$

dove W_i è il peso del singolo studio, θ_i è la stima di efficacia del singolo studio (log OR, log RR) e θ è la stima pooled.
Sotto ipotesi nulla che non vi sia differenza negli effetti dell'intervento tra gli studi questa statistica segue una distribuzione chi quadrato con $k - 1$ gradi di libertà e k il numero di studi analizzati

Remark 74. Alcuni sostengono che l'eterogeneità è inevitabile ed è dunque (non si può escluderne tutte le componenti/fonti), quindi tanto vale evitare il prendere decisioni sulla base di un test. Misure alternative che quantifichino l'eterogeneità sono state proposte

Remark 75 (I^2).

$$I^2 = \left(\frac{Q - df}{Q}\right) \times 100\%$$

con Q è la statistica chi square del test di cui sopra e df sono i suoi gradi di libertà ($k-1$ con k studi). Questa indica la percentuale di variabilità nelle stime che è dovuta ad eterogeneità piuttosto che all'errore di campionamento (caso); una guida grezza alla sua interpretazione è

- da 0 a 40: eterogeneità potrebbe essere non importante
- da 30 a 60: potrebbe essere moderata
- da 50 a 90: potrebbe essere sostanziale
- da 70 a 100: potrebbe essere notevole

Remark 76. Nel caso di eterogeneità vi sono due approcci di analisi:

- analisi per sottogruppi o meta-regressioni
- stima dell'effetto del trattamento con modello ad effetti casuali (random effects model), soprattutto se quanto meno la direzione dell'effetto è abbastanza univoca

Remark 77 (Analisi per sottogruppi). Si può effettuare per subset di pazienti o studi

- suddivide gli studi in base ad alcune caratteristiche *pre-specificate* nel protocollo della revisione

- si effettua la meta-analisi in ogni gruppo (al fine di ridurre l'eterogeneità interna)

Per un test sulla presenza di eterogeneità si può effettuare uno test di eterogeneità sui risultati dei sottogruppi (invece che sui studi individuali).

Remark 78 (Meta-regressione). Può quantificare l'impatto sulle stime di diverse caratteristiche dei vari studi, dovrebbe essere considerata solo se vi sono 10 studi o più

20.3 Metanalisi in a nutshell

Remark 79. Quando si fa una metanalisi non si fa una semplice somma dei pazienti e degli eventi occorsi nei singoli studi (per evitare stime inficiate dal paradosso di Simpson); si preserva invece la loro individualità e si procede ad una media pesata (al numeratore sommatoria di effetto per peso, al denominatore sommatoria dei pesi) dell'effetto dell'intervento

Remark 80 (Fasi della metanalisi). Si procede

1. innanzitutto a calcolare una statistica per ogni studio per descrivere l'effetto del trattamento;
2. si effettua poi una stima complessiva dell'intervento calcolando una media pesata degli effetti stimati nei singoli studi

$$\frac{\sum Y_i W_i}{\sum W_i}$$

   con Y_i stima dell'effetto del singolo studio (per misure rapporto come OR o RR ecc qui si usa il logaritmo) e W_i peso associato allo studio; se tutti gli studi assumono lo stesso peso, la stima è semplicemente la media degli studi
3. si può ora assumere che l'intervento abbia una sua efficacia unica/intrinseca che tutti gli studi stanno cercando di stimare e che le variazioni nelle singole stime siano dovuti all'errore di campionamento: in questo caso si procede ad una *analisi ad effetti fissi*.
   Se viceversa si pensa che l'efficacia dell'intervento abbia effettivamente una distribuzione (e non un valore singolo) si procede ad una metanalisi *ad effetti random*.
4. si calcola l'errore standard della stima complessiva, per derivare un intervallo di confidenza

20.4 Metodi di calcolo dei pesi W_i

Remark 81 (Inverse variance method). I pesi associati a ciascuno studio sono spesso e volentieri l'inverso della varianza (o meglio del quadrato dell'errore standard) della stima per ciascuno studio

$$W_i = 1/SE^2$$

Pertanto i pesi dei vari studi saranno proporzionali alla loro dimensione: studi più grandi con errore standard tipicamente minore avranno più peso rispetto a studi più piccoli.

Example 20.4.1. Nella classica tabella 2×2 con trattamento in colonna ed evento in riga

$$SE^2(logOR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$
$$SE^2(logRR) = \frac{1}{a} + \frac{1}{b} - \frac{1}{a+c} - \frac{1}{b+d}$$

Example 20.4.2. Nel caso di outcome quantitativo, per la differenza di medie, tipicamente

$$SE^2(meandiff) = \frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}$$

con sd_1^2 e n_1 la varianza e la numerosità del primo braccio.
Per la differenza di medie standardizzate si una un'0altra formula

Example 20.4.3. Per il rate ratio

$$SE^2(lograteratio) = \frac{1}{E_1} + \frac{1}{E_2}$$

con E_1 gli eventi nel primo braccio ed E_2 nel secondo

Remark 82 (Nel caso di meta analisi con random effects). Gli errori standard studio specifici (gli SE) sono aggiustati incorporando una misura di eterogeneità tra gli effetti osservata negli studi, detta τ^2 (tau al quadrato)

Remark 83. In generale metodi ad effetti random ed effetti fissi daranno risultati analoghi quando non vi è eterogeneità tra gli studi.
Viceversa quando vi è eterogeneità gli intervalli di confidenza saranno più larghi nelle stime random effects rispetto a fixed effects e quindi più difficilmente si avrà un risultato overall statisticamente significativo

Capitolo 21

Effect size and precision

21.1 Overview

treatment effects misura di associazione tra due variabili derivante da un singolo studio che partecipa alla metanalisi in cui una rappresenta un fattore sperimentale

effect size misura di associazione tra due variabili di un singolo studio

single group summary sintesi di una variabile (quindi non una associazione di due variabili): ad esempio una prevalenza

Remark 84. Le metanalisi si possono analizzare qualsiasi tipo di misura, tra quelle elencate, e ai fini del metodo non cambia troppo. In generale qui si userà *effect size* in senso generico, intendendo il risultato desumibile dal singolo studio, e potendo intendere con esso anche treatment effect o single group summary

Remark 85. Una volta che abbiamo calcolato l'effect size e costruito mediante l'errore standard un intervallo di confidenza, le formule per calcolare un effetto complessivo, per la verifica dell'eterogeneità e così via sono le stesse indipendentemente dal tipo di effect size adoperato

21.2 Effect size basati su medie

Quando gli studi riportano medie e deviazioni standard l'effect size di elezione sono solitamente la differenza di medie grezze o standardizzate (o il response ratio ma mi sembra meno interessante).

21.2.1 Differenza di medie non standardizzate in gruppi indipendenti

Se μ_1 e μ_2 sono le medie nelle popolazione, la differenza in queste è $\Delta = \mu_1 - \mu_2$; lo stimatore di questo parametro è

$$D = \overline{X}_1 - \overline{X}_2$$

Se non assumiamo che le due deviazioni standard del carattere nei gruppi della popolazione siano differenti, l'errore standard di D è

$$SE_D = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

con S_1, S_2 le deviazioni standard campionarie e n_1, n_2 i sample size dei due gruppi

21.2.2 Differenza di medie standardizzate in gruppi indipendenti

Coincide con una d di cohen e si usa tipicamente se:

- l'outcome analizzato è meno conosciuto/standard
- se studi diversi usano outcome quantitativi differenti, al fine di omogeneizzare

è definita nella popolazione come

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

(dove abbiamo ipotizzato che le due popolazioni siano accomunate dalla deviazione standard del carattere $\sigma_1 = \sigma_2 = \sigma$).

Nel campione lo stimatore è

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S_{within}}$$

con

$$S_{within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 + 2}}$$

con X_1, X_2 medie campionarie, n_1, n_2 i sample size e S_1, S_2 le deviaizoni standard nei due gruppi. Un errore standard è

$$SE_d = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}}$$

Emerge come in realtà lo stimatore d di Cohen sia lievemente biasato e tenda a sovrastimare δ in *campioni piccoli*; il

lo stimatore corretto è il g di Hedges, che viene calcolato a partire da d e da una correzione (quella più utilizzata in pratica è la seguente)

$$J = 1 - \frac{3}{4df - 1}$$

con df sono i gradi di libertà usati per stimare S_{within} che per due gruppi indipendenti è $n_1 + n_2 - 2$. Si ha che lo stimatore corretto è allora

$$g = J \times d$$

e

$$SE_g = \sqrt{J^2 \times V_d}$$

	Trattati	Controlli
Eventi	A	C
Non eventi	B	D
Tot	n_1	n_2

Tabella 21.1: Tabella 2 x 2

21.2.3 Response ratio

21.3 Effect size basati su dati binari (tabelle 2×2)

21.3.1 Risk ratio

Per i risk ratio i conti (intervallo di confidenza) sono fatti su scala logaritmica (che normalizzano la distribuzione dello stimatore) e riportato sulla scala normale. Lo stimatore

$$RR = \frac{A/n_1}{C/n_2}$$

il log risk ratio è

$$LogRiskRatio = \log(RR)$$

che ha errore standard pari a

$$SE_{LogRiskRatio} = \sqrt{\frac{1}{A} - \frac{1}{n_1} + \frac{1}{C} - \frac{1}{n_2}}$$

Questo serve per costruire l'intervallo di confidenza in scala logaritmica

$$LL_{LogRiskRatio} = LogRiskRatio - 1.96 \cdot SE_{LogRiskRatio}$$

$$UL_{LogRiskRatio} = LogRiskRatio + 1.96 \cdot SE_{LogRiskRatio}$$

Per riportarlo poi il tutto in scala normale

$$LL_{RiskRatio} = \exp LL_{LogRiskRatio}$$

$$UL_{RiskRatio} = \exp UL_{LogRiskRatio}$$

21.3.2 Odds ratio

Anche qui si procede in scala logaritmica

$$OR = \frac{AD}{CB}$$

$$LogOddsRatio = \log(OR)$$

avente errore standard

$$SE_{LogOddsRatio} = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$$

ed intervallo di confidenza

$$LL_{LogOddsRatio} = LogOddsRatio - 1.96 \cdot SE_{LogOddsRatio}$$

$$UL_{LogOddsRatio} = LogOddsRatio + 1.96 \cdot SE_{LogOddsRatio}$$

Per riportarlo poi il tutto in scala normale

$$LL_{OddsRatio} = \exp LL_{LogOddsRatio}$$

$$UL_{OddsRatio} = \exp UL_{LogOddsRatio}$$

21.3.3 Risk difference

Qui a differenza dei precedenti i calcoli su scala normale

$$RiskDiff = \frac{A}{n_1} - \frac{C}{n_2}$$

con errore standard approssimato a

$$SE_{RiskDiff} = \sqrt{\frac{AB}{n_1^3} + \frac{CD}{n_2^3}}$$

21.3.4 Considerazioni sulla scelta

Fare i calcoli in RR o OR poi predirre la risk difference per vari risk del gruppo baseline (controllo)

21.4 Effect size basati su correlazioni

I conti si fanno effettuando la trasformata Z di Fisher (per avere uno stimatore approssimativamente normale); se ρ è il coefficiente di Pearson si ha la trasformata come

$$z = 0.5 \cdot \log\left(\frac{1+\rho}{1-\rho}\right)$$

che ha errore standard

$$SE_z = \sqrt{\frac{1}{n-3}}$$

Calcolato il CI si riporta all'unità di correlazione mediante la trasformata inversa

$$\rho = \frac{e^{2z}-1}{e^{2z}+1}$$

21.5 Conversione tra effect size

Si può fare un minestrone di misure di efficacia con dati binari, continui o di correlazione (in genere riportando in termini di d di Cohen); vedere il capitolo 7 di Borenstein *e altri* (2011)

Capitolo 22

Modelli ad effetti fissi e ad effetti random

22.1 Introduzione

La maggior parte delle metanalisi si basa su due modelli statistici, il modello ad effetto fisso e quello ad effetti random

Effetto fisso si assume che vi sia un unico, vero, effect size associato al trattamento che sottosta a tutti gli studi dell'analisi considerata e che le differenze osservate tra studi siano dovuti al campionamento

Effetti random si ipotizza che l'effetto del trattamento (nella popolazione) possa effettivamente variare da studio a studio, quindi la variabilità che si osserva tra gli studi è dovuta sia al campionamento che al fatto che estrariamo campioni da urne differenti. Se fosse possibile effettuare infiniti studi si otterrebbe la distribuzione dell'effect size; siamo per lo più interessati ad un valore centrale (media/valore atteso) di questi effect size

Remark 86. La differenza principale che ne deriva riguarda la varianza, per il resto come si vedrà le formule dei due modelli sono uguali

Remark 87. Nella discussione che segue conviene tenere ben distinto l'*effetto vero* (quello dell'intervento nella popolazione) da quello *osservato* (ossia stimato nel campione)

Remark 88. In generale, di default, non vi sono motivazioni per assumere che gli effetti sottostanti (di studi differenti, in setting lievemente differenti ecc) siano coincidenti; pertanto, di solito, random effects all the way!

L'analisi ad effetti fissi ha senso se due condizioni sono rispettate: tutti gli studi della metanalisi sono funzionalmente identici e il nostro obiettivo è stimare l'effect size comune per questo tipo di popolazione e non si cerca di generalizzare ad altre popolazioni. Se viceversa i ricercatori stanno accumulando dati da studi che sono stati eseguiti da ricercatori che hanno operato indipendentemente, sarebbe inverosimile che tutti gli studi fossero funzionalmente equivalenti.

Alcuni effettuano una analisi ad effetti random dopo avere effettuato un test di eterogeneità ma non è la strada che si consiglia qui; viceversa se uno fa

l'analisi a effetti fissi poi l'eterogeneità ne esce, allora l'ipotesi su cui si è basato forse non regge.

22.2 Effetto fisso

Sotto effetto fisso ipotizziamo che vi sia un unico effect size associato al trattamento, θ e che ogni effect size rilevato in ciascuno studio incluso nell'analisi Y_i differisca di un errore casuale ε_i dovuto al campione estratto

$$Y_i = \theta + \varepsilon_i \tag{22.1}$$

In una analisi ad effetti fissi ad ogni studio viene assegnato un peso pari a

$$W_i = \frac{1}{V_{Y_i}} \tag{22.2}$$

con V_{Y_i} la cd varianza (il quadrato dell'errore standard dello stimatore considerato applicato allo studio). La stima dell'effetto complessivo per una metanalisi con k studi è dato da

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i} \tag{22.3}$$

La varianza di questa stima complessiva è data da

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i} \tag{22.4}$$

quindi è di fatto legata direttamente alla varianza dei singoli studi. L'errore standard dello stimatore complessivo è la radice di questa

$$SE_M = \sqrt{V_M} \tag{22.5}$$

Ottenute queste stime si può costruire l'intervallo di confidenza dello stimatore complessivo

$$LL = M - 1.96 \cdot SE_M \qquad UL = M + 1.96 \cdot SE_M \tag{22.6}$$

o procedere a test Z sulla presenza di effetto come abitualmente mediante

$$Z = \frac{M}{SE_M} \tag{22.7}$$

e confrontando con i quantili soglia della normale standard.

22.3 Effetti random

Come detto si ipotizza che vi sia una distribuzione di effect size piuttosto che un singolo effect size unico.
In una analisi ad effetti random si ipotizza che l'effect size di un singolo studio si discosti dal valore medio/atteso degli effect size per questo tipo di intervento per due componenti:

- innanzitutto l'effect size della popolazione/trattamento considerata si discosta dal valore atteso degli effect size per una componente ζ_i (se avessimo un numero di studi infiniti potremmo calcolare l'effect size atteso complessivo)
- l'effect size osservato nel campione si discosta dall'effect size della popolazione considerata sempre per un errore casuale (se avessimo un campione infinito l'errore sarebbe nullo)

In altre parole si ha che l'effect size osservato sia la somma di tre componenti:

$$Y_i = \underbrace{\mu + \zeta_i}_{\theta_i} + \varepsilon_i \tag{22.8}$$

con μ il valore atteso degli effect size complessivamente e θ_i l'effect size della popolazione considerata.
La distanza di θ_i da μ dipende dalla variabilità della distribuzione degli effetti tra gli studi, indicata con τ (se si pensa alla deviazione standard) o τ^2 (se si pensa alla varianza. Essendo un valore unico per tutto la metanalisi, rientra nei conteggi della variabilità di tutti gli studi considerati, come si vedrà.

Uno stimatore di τ^2 (Dersimonian e Laird) è il seguente:

$$T^2 = \frac{Q - df}{C} \tag{22.9}$$

dove

$$\begin{aligned} Q &= \sum_{i=1}^{k} W_i Y_i^2 - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i} \\ df &= k - 1 \\ C &= \sum_i W_i - \frac{\sum W_i^2}{\sum W_i} \end{aligned}$$

con k il numero di studi.

Una volta ottenuta la stima, la stima della varianza per il singolo studio diviene

$$V_{Y_i}^* = V_{Y_i} + T^2$$

ossia si aggiunge questa componente a quella già adottata per il modello a effetto fisso. I calcoli poi procedono analogamente in tutto e per tutto a quanto già visto in tal caso.

Si approfondirà il discorso quando di parlerà di eterogeneità

22.4 Un confronto

Remark 89. A parità di altre condizioni, i modelli ad effetti per la presenza di una componente comune di varianza (positiva, che viene sommata ad un'altra componente positiva) rende complessivamente le varianze dei vari studi più simili tra loro, dunque i pesi assegnati più omogenei e complessivamente rispetto ad un'analisi ad effetto fisso pesa maggiormente gli studi più piccoli e meno quelli

grossi; a parole dato che ogni studio fornisce informazioni su differenti effect size, vogliamo assicurarci che tutti siano rappresentati nella stima complessiva e che non vogliamo penalizzare alcuni effect size solo perché i relativi studi siano piccoli (come si fa in uno studio ad effetto fisso); per lo stesso ragionamento non diamo eccessivo peso a studi con molti pazienti.

Remark 90. In merito all'interpretazione dell'intervallo di confidenza, nel caso dell'effetto fisso si tratta dell'intervallo di confidenza di questo, mentre in quello di effetti random si tratta dell'intervallo di confidenza del valore atteso dei vari effect size

Remark 91. Sull'interpretazione del test, per l'effetto fisso l'ipotesi nulla è che l'effetto sia nullo, mentre in quello degli effetti random è che il valore atteso/medio degli effetti random sia nullo (ma non si esclude che vi possano essere popolazioni/setting in cui l'effect size sia effettivamente diverso)

22.5 Esempi

Qui si riproducono gli esempi del capitolo 14 di Borenstein *e altri* (2011) mediante il pacchetto `metafor` di R. Dato che la differenza principale si ha solamente nel calcolo degli effect size si approfondisce per esteso il caso di dati dicotomici, facendo vedere quello che varia nella stima per i continui e correlazionali.

22.5.1 Dati dicotomici

Procediamo alla stima degli effect size di ciascuno studio mediante `escalc` e poi alla metanalisi mediante `rma` (o `rma.uni`, sinonimo)

```
library(lbdatasets)
library(metafor)
metabin

##     study te tne  tn ce cne  cn
## 1   Saint 12  53  65 16  49  65
## 2   Kelly  8  32  40 10  30  40
## 3 Pilbeam 14  66  80 19  61  80
## 4    Lane 25 375 400 80 320 400
## 5  Wright  8  32  40 11  29  40
## 6     Day 16  49  65 18  47  65

## questo fornisce il calcolo degli effect size per
## il caso degli odds-ratio (stime sulla log scale)
## cfr pag93 tab 14.5
(es_bin <- escalc(measure = 'OR',  # log odds-ratio
                  ai = te,         # eventi nei trattati
                  bi = tne,        # non eventi nei trattati
                  n1i = tn,        # per check totale trattati
                  ci = ce,         # eventi nei controlli
                  di = cne,        # non eventi nei controlli
                  n2i = cn,        # per check totale controlli
                  data = metabin)) # data.frame di riferimento
```

```
##
##     study te tne  tn ce cne  cn      yi     vi
## 1   Saint 12  53  65 16  49  65 -0.3662 0.1851
## 2   Kelly  8  32  40 10  30  40 -0.2877 0.2896
## 3 Pilbeam 14  66  80 19  61  80 -0.3842 0.1556
## 4    Lane 25 375 400 80 320 400 -1.3218 0.0583
## 5  Wright  8  32  40 11  29  40 -0.4169 0.2816
## 6     Day 16  49  65 18  47  65 -0.1595 0.1597

## fixed effect meta analysis
(bin_fe <- rma(yi = yi, vi = vi, data = es_bin, method = 'FE'))

##
## Fixed-Effects Model (k = 6)
##
## I^2 (total heterogeneity / total variability):   52.61%
## H^2 (total variability / sampling variability):  2.11
##
## Test for Heterogeneity:
## Q(df = 5) = 10.5512, p-val = 0.0610
##
## Model Results:
##
## estimate      se     zval    pval    ci.lb    ci.ub
##  -0.7241  0.1539  -4.7068  <.0001  -1.0257  -0.4226  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## random effects meta analysis (Tau^2 stimato secondo dersimonian e laird)
(bin_re <- rma(yi = yi, vi = vi, data = es_bin, method = 'DL'))

##
## Random-Effects Model (k = 6; tau^2 estimator: DL)
##
## tau^2 (estimated amount of total heterogeneity): 0.1729 (SE = 0.2148)
## tau (square root of estimated tau^2 value):      0.4158
## I^2 (total heterogeneity / total variability):   52.61%
## H^2 (total variability / sampling variability):  2.11
##
## Test for Heterogeneity:
## Q(df = 5) = 10.5512, p-val = 0.0610
##
## Model Results:
##
## estimate      se     zval    pval    ci.lb    ci.ub
##  -0.5663  0.2388  -2.3711  0.0177  -1.0344  -0.0982  *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Una volta ottenuta la stima di metanalisi possiamo indagare l'oggetto; concentrandoci sul caso degli effetti random ...

```
## cosa c'<U+00E8>
names(bin_re)
```

```
##  [1] "b"            "beta"         "se"           "zval"         "pval"
##  [6] "ci.lb"        "ci.ub"        "vb"           "tau2"         "se.tau2"
## [11] "tau2.fix"     "tau2.f"       "I2"           "H2"           "R2"
## [16] "vt"           "QE"           "QEp"          "QM"           "QMdf"
## [21] "QMp"          "k"            "k.f"          "k.eff"        "k.all"
## [26] "p"            "p.eff"        "parms"        "int.only"     "int.incl"
## [31] "intercept"    "allvipos"     "coef.na"      "yi"           "vi"
## [36] "X"            "weights"      "yi.f"         "vi.f"         "X.f"
## [41] "weights.f"    "M"            "outdat.f"     "ni"           "ni.f"
## [46] "ids"          "not.na"       "subset"       "slab"         "slab.null"
## [51] "measure"      "method"       "model"        "weighted"     "test"
## [56] "dfs"          "ddf"          "s2w"          "btt"          "m"
## [61] "digits"       "level"        "control"      "verbose"      "add"
## [66] "to"           "drop00"       "fit.stats"    "data"         "formula.yi"
## [71] "formula.mods" "version"      "call"         "time"
```

```
## alcune stime puntuali presentate tra pag 96 e 97
with(bin_re, c('stima' = b,
               'se' = se,
               'ci.low' = ci.lb,
               'ci.up' = ci.ub,
               'z' = zval,
               'two_tail_p' = pval))
```

```
##       stima          se      ci.low       ci.up           z  two_tail_p
## -0.56629590  0.23883443 -1.03440279 -0.09818902 -2.37108149  0.01773612
```

```
## per ottenere i pesi relativi (sommano a 100)
weights(bin_re)
```

```
##        1        2        3        4        5        6
## 15.93285 12.33370 17.36382 24.67247 12.54918 17.14797
```

```
## stime e ci di alcuni pararametri stimati
confint(bin_re)
```

```
##
##        estimate  ci.lb   ci.ub
## tau^2    0.1729 0.0000  0.9656
## tau      0.4158 0.0000  0.9826
## I^2(%)  52.6118 0.0000 86.1112
## H^2      2.1102 1.0000  7.2001
```

Infine procediamo alla visualizzazione del forestplot (sulla scala degli odds-ratio e cercando di ottimizzare un minimo) dell'analisi ad effetto fisso in figura

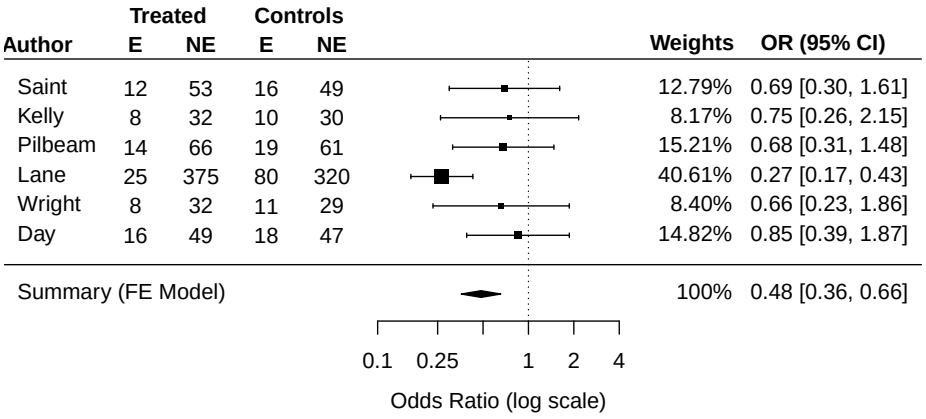


Figura 22.1: Forestplot metanalisi fixed effect

22.1 mentre quella con effetti random in figura 22.2.

```
## ----------------------------
## Fixed effect MA forest plot
## ----------------------------
par(mar = c(5,1,1,1))
data_matrix <- with(metabin, cbind(te, tne, ce, cne))
data_matrix_col_pos <- seq(-6, -3)
ticks_pos <- log(c(0.1, 0.25, 0.5, 1, 2, 4))
xlim <- c(-8, 6)
forest(x = bin_fe,      # fornire la stima degli effect size
       atransf = exp,  # per ottenere gli OR invece dei log (OR)
       at = ticks_pos, # posizionamento dei ticks rispettando la scala
       xlim = xlim,
       ## matrice dei dati a sinistra e posizionamento delle colonne
       ilab = data_matrix,
       ilab.xpos = data_matrix_col_pos,
       slab = metabin$study,    ## etichette di riga
       showweights = TRUE,      ## display dei pesi
       mlab = 'Summary (FE Model)')
## intestazioni
par(font = 2)
title_first_row  <- 8.5
title_second_row <- 7.5
text(x = data_matrix_col_pos, y = title_second_row,
     labels = rep(c("E", 'NE'), 2))
text(x = -7.5, y = title_second_row, 'Author')
text(x = c(-5.5, -3.5), y = title_first_row, c('Treated', 'Controls'))
text(x = c(2.5, 4.5), y = title_second_row, c('Weights', 'OR (95% CI)'))
```

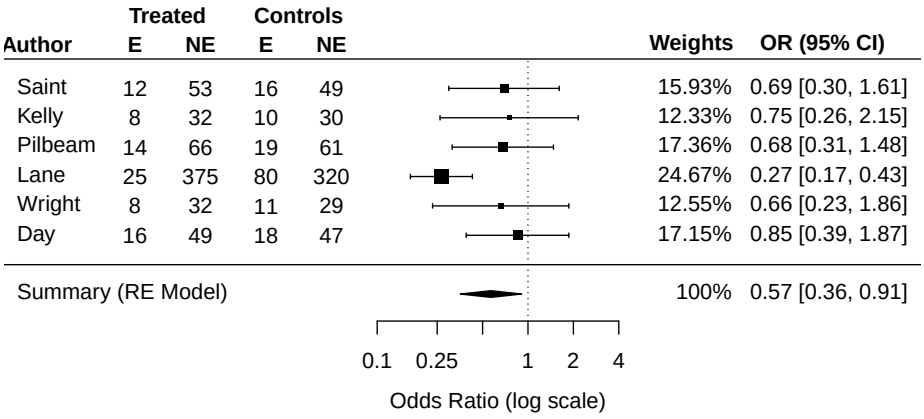


Figura 22.2: Forestplot metanalisi random effects

```
## ----------------------------
## Random effects MA forest plot
## ----------------------------
par(mar = c(5,1,1,1))
forest(x = bin_re,      # fornire la stima degli effect size
       atransf = exp,  # per ottenere gli OR invece dei log (OR)
       at = ticks_pos, # posizionamento dei ticks rispettando la scala
       xlim = xlim,
       ## matrice dei dati a sinistra e posizionamento delle colonne
       ilab = data_matrix,
       ilab.xpos = data_matrix_col_pos,
       slab = metabin$study,    ## etichette di riga
       showweights = TRUE,      ## display dei pesi
       mlab = 'Summary (RE Model)')
## intestazioni
par(font = 2)
text(x = data_matrix_col_pos, y = title_second_row,
     labels = rep(c("E", 'NE'), 2))
text(x = -7.5, y = title_second_row, 'Author')
text(x = c(-5.5, -3.5), y = title_first_row, c('Treated', 'Controls'))
text(x = c(2.5, 4.5), y = title_second_row, c('Weights', 'OR (95% CI)'))
```

22.5.2 Dati continui

```
metacon

##     study tm tsd  tn cm csd  cn
## 1 Carroll 94  22  60 92  20  60
## 2   Grant 98  21  65 92  22  65
## 3    Peck 98  28  40 88  26  40
```

```
## 4   Donat 94  19 200 82  17 200
## 5 Stewart 98  21  50 88  22  45
## 6   Young 96  21  85 92  22  85

## cfr tab 14.2 pag 88
(es_con <- escalc(measure = 'SMD',    # SMD
                  m1i = tm,           # media trattati
                  sd1i = tsd,         # sd trattati
                  n1i = tn,           # n trattati
                  m2i = cm,           # media controlli
                  sd2i = csd,         # sd controlli
                  n2i = cn,           # n controlli
                  data = metacon))    # data.frame di riferimento

##
##     study tm tsd  tn cm csd  cn     yi     vi
## 1 Carroll 94  22  60 92  20  60 0.0945 0.0334
## 2   Grant 98  21  65 92  22  65 0.2774 0.0311
## 3    Peck 98  28  40 88  26  40 0.3665 0.0508
## 4   Donat 94  19 200 82  17 200 0.6644 0.0106
## 5 Stewart 98  21  50 88  22  45 0.4618 0.0433
## 6   Young 96  21  85 92  22  85 0.1852 0.0236

## ## fixed effect meta analysis
## (con_fe <- rma(yi = yi, vi = vi, data = es_con, method = 'FE'))
## ## random effects meta analysis (Tau^2 stimato secondo dersimonian e laird)
## (con_re <- rma(yi = yi, vi = vi, data = es_con, method = 'DL'))
```

22.5.3 Correlazioni

```
metacor

##     study  cor   n
## 1   Fonda 0.50  40
## 2  Newman 0.60  90
## 3   Grant 0.40  25
## 4 Granger 0.20 400
## 5 Milland 0.70  60
## 6   Finch 0.45  50

## vedere per confronto table 14.8 pag 98
(es_cor <- escalc(measure = 'ZCOR', # trasformata di Fisher su correlazione
                  ri = cor,         # correlazioni raw
                  ni = n,           # n per gruppo
                  data = metacor))  # data.frame di riferimento

##
##     study  cor   n     yi     vi
## 1   Fonda 0.50  40 0.5493 0.0270
```

```
## 2  Newman 0.60  90 0.6931 0.0115
## 3   Grant 0.40  25 0.4236 0.0455
## 4 Granger 0.20 400 0.2027 0.0025
## 5 Milland 0.70  60 0.8673 0.0175
## 6   Finch 0.45  50 0.4847 0.0213

## ## fixed effect meta analysis
## (cor_fe <- rma(yi = yi, vi = vi, data = es_cor, method = 'FE'))
## ## random effects meta analysis (Tau^2 stimato secondo dersimonian e laird)
## (cor_re <- rma(yi = yi, vi = vi, data = es_cor, method = 'DL'))
```

Capitolo 23

Eterogeneità

23.1 Quantificazione

Per eterogeneità intendiamo la variabilità della distribuzione dell'effect size nella popolazione: sotto un modello fixed essa si assume essere 0 mentre sotto un modello random un valore positivo.

Nel caso vi sia eterogeneità, la variabilità complessiva degli effect size può/deve essere spacchettata in due componenti:

- variabilità dovuta all'eterogeneità
- variabilità dovuta al campionamento casuale

Un primo indice di eterogeneità di k studi è

$$Q = \sum_{i=1}^{k} W_i (Y_i - M)^2 = \sum_{i=1}^{k} \left(\frac{Y_i - M}{S_i} \right)^2 \tag{23.1}$$

con

- $W_i = 1/V_i$ peso dello studio e V_i sua varianza (e S_i errore standard)
- Y_i l'effect size del singolo studio
- M l'effect size complessivo derivante dai k studi

Si tratta di una somma di scarti, standardizzati al quadrato e misura la variabilità totale presente.

23.1.1 Test di eterogeneità

Se vogliamo un test di eterogeneità, nell'ipotesi nulla di assenza di eterogeneità (es se $Y_i = M + \varepsilon_i$ con ε_i normale), il Q si distribuisce come un χ^2 con $df = k - 1$ gradi di libertà. Anche qui, se il test è non significativo non vuol necessariamente dire che non vi sia eterogeneità, potrebbe essere low power (per questo spesso si usa invece di 0.05 0.1 come soglia decisionale

23.1.2 Scarto di eterogeneità

Possiamo poi produrre tutto un altro range di indicatori partendo dalla differenza tra il valore osservato di variabilità totale Q e quello atteso in assenza di eterogeneità.
Nel caso in cui tutti gli studi condividano un effect size comune, il valore atteso di Q è una somma di scarti (determinati dall'errore di campionamento casuale) pari ai gradi di libertà

$$df = k - 1$$

e la misura

$$Q - df$$

se positiva indicherà l'eccesso di variazione dovuto alla eterogeneità, mentre se negativa indicherà una scarsissima eterogeneità (minore di quella ragionevolmente assumibile come casuale).

23.1.3 Stima di τ^2

Con τ^2 si intende la varianza degli effetti associati al trattamento nella popolazione; dato che non la possiamo stimare direttamente, ricorriamo ad uno stimatore T^2 basato sugli studi a nostra disposizione che si ottiene come

$$T^2 = \frac{Q - df}{C} \tag{23.2}$$

con

$$C = \sum_i W_i - \frac{\sum W_i^2}{W_i} \tag{23.3}$$

T^2 è usata per il calcolo dei pesi nei modelli random effect come

$$W_i^* = \frac{1}{V_{Y_i}^*} = \frac{1}{V_{Y_i} + T^2} \tag{23.4}$$

23.1.4 I^2

È un indicatore della percentuale di variabilità complessiva dovuta ad eterogeneità e calcola come

$$I^2 = \frac{Q - df}{Q} \times 100 \tag{23.5}$$

dove al numeratore abbiamo un indicatore di τ^2 al denominatore un indicatore della varianza totale. Vi sono anche intervalli di confidenza per I^2, volendo. Higgins suggerisce (pag 119 libro) di considerare 25, 50 e 75 come eterogeneità bassa, media ed elevata

23.1.5 Applicazioni in R

In metafor queste misure vengono calcolate (alcune soltanto nel caso di modelli modelli random effects) mediante `rma`, come si è visto negli esempi precedenti

23.2 Prediction intervals

Per le stime globali associate ad una metanalisi possiamo provvedere un intervallo di confidenza del valore centrale, che ha come scopo quello di descrivere un set di valori verosimili per l'effetto complessivo (stima fixed) o dell'effetto medio (stima random).

Nel caso di modelli random, ove abbiamo una stima di τ^2, possiamo anche fornire un intervallo di predizione ossia (specularmente ad un intervallo di predizione su una retta di regressione di uno studio primario) un range di valori verosimili per il valore di effetto di un nuovo studio. Si costruisce come

$$\mu \pm z_{1-\alpha/2}\sqrt{\tau^2}$$

ma se non si hanno i valori della popolazione qui richiesti si rimedia con

$$M^* \pm t_{1-\alpha/2,df}\sqrt{T^2 + V_{M^*}}$$

dove M^* è l'effect size medio e V_{M^*} la sua varianza.

Graficamente l'intervallo di predizione si espande come una barretta oltre il diamante dell'intervallo di confidenza della media

23.3 Analisi per sottogruppi

Capitolo 19 libro

TODO: fixme

23.4 Metaregressione

La regressione può esser utilizzata anche avendo come unità di analisi il singolo studio (nello specifico il suo effect size, spesso e volentieri logaritmizzato per renderlo simmetrico attorno allo zero). Valgono anche qui considerazioni sul rapporto tra covariate adottate e studi/pazienti disponibili (anche se non vi sono regole scritte nella pietra).

Si sudia l'esempio del vaccino BCG nel prevenire la tubercolosi, analisi fatta sui risk ratio (si riproduce solamente la stima con random effects)

```
options(width = 100)
library(metafor)
(db <- dat.bcg)  # dataset impiegato, disponibile in metafor
```

```
##    trial               author year tpos  tneg cpos  cneg ablat      alloc
## 1      1              Aronson 1948    4   119   11   128    44     random
## 2      2    Ferguson & Simes 1949    6   300   29   274    55     random
## 3      3     Rosenthal et al 1960    3   228   11   209    42     random
## 4      4   Hart & Sutherland 1977   62 13536  248 12619    52     random
## 5      5 Frimodt-Moller et al 1973   33  5036   47  5761    13  alternate
## 6      6     Stein & Aronson 1953  180  1361  372  1079    44  alternate
## 7      7    Vandiviere et al 1973    8  2537   10   619    19     random
## 8      8          TPT Madras 1980  505 87886  499 87892    13     random
## 9      9    Coetzee & Berjak 1968   29  7470   45  7232    27     random
## 10    10     Rosenthal et al 1961   17  1699   65  1600    42 systematic
```

```
## 11    11       Comstock et al 1974  186 50448  141 27197    18 systematic
## 12    12   Comstock & Webster 1969    5  2493    3  2338    33 systematic
## 13    13       Comstock et al 1976   27 16886   29 17825    33 systematic

rr <- escalc(measure = 'RR', # log
             ai = tpos,
             bi = tneg,
             ci = cpos,
             di = cneg,
             data = db)
rr <- rr[with(rr, order(yi)), ]
rr[, names(rr) %without% c("alloc")]

##
##    trial               author year tpos  tneg cpos  cneg ablat      yi     vi
## 7      7     Vandiviere et al 1973    8  2537   10   619    19 -1.6209 0.2230
## 2      2     Ferguson & Simes 1949    6   300   29   274    55 -1.5854 0.1946
## 4      4    Hart & Sutherland 1977   62 13536  248 12619    52 -1.4416 0.0200
## 10    10      Rosenthal et al 1961   17  1699   65  1600    42 -1.3713 0.0730
## 3      3      Rosenthal et al 1960    3   228   11   209    42 -1.3481 0.4154
## 1      1              Aronson 1948    4   119   11   128    44 -0.8893 0.3256
## 6      6      Stein & Aronson 1953  180  1361  372  1079    44 -0.7861 0.0069
## 9      9     Coetzee & Berjak 1968   29  7470   45  7232    27 -0.4694 0.0564
## 11    11       Comstock et al 1974  186 50448  141 27197    18 -0.3394 0.0124
## 5      5 Frimodt-Moller et al 1973   33  5036   47  5761    13 -0.2175 0.0512
## 13    13       Comstock et al 1976   27 16886   29 17825    33 -0.0173 0.0714
## 8      8           TPT Madras 1980  505 87886  499 87892    13  0.0120 0.0040
## 12    12   Comstock & Webster 1969    5  2493    3  2338    33  0.4459 0.5325

## Stima overall FE, compliant col libro
(ma <- rma(yi = yi, vi = vi, method = 'DL', data = rr))

##
## Random-Effects Model (k = 13; tau^2 estimator: DL)
##
## tau^2 (estimated amount of total heterogeneity): 0.3088 (SE = 0.2299)
## tau (square root of estimated tau^2 value):      0.5557
## I^2 (total heterogeneity / total variability):   92.12%
## H^2 (total variability / sampling variability):  12.69
##
## Test for Heterogeneity:
## Q(df = 12) = 152.2330, p-val < .0001
##
## Model Results:
##
## estimate      se     zval    pval    ci.lb    ci.ub
##  -0.7141  0.1787  -3.9952  <.0001  -1.0644  -0.3638  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Metaregressione FE: si pone il rhs della formula in mods (ablat = latitude)
(mr <- rma(yi = yi, vi = vi, mods = ~ ablat , method = 'DL', data = rr))

##
## Mixed-Effects Model (k = 13; tau^2 estimator: DL)
##
## tau^2 (estimated amount of residual heterogeneity):     0.0633 (SE = 0.0548)
## tau (square root of estimated tau^2 value):             0.2516
## I^2 (residual heterogeneity / unaccounted variability): 64.21%
## H^2 (unaccounted variability / sampling variability):   2.79
## R^2 (amount of heterogeneity accounted for):            79.50%
##
## Test for Residual Heterogeneity:
## QE(df = 11) = 30.7331, p-val = 0.0012
##
## Test of Moderators (coefficient 2):
## QM(df = 1) = 18.8452, p-val < .0001
##
## Model Results:
##
##          estimate      se     zval    pval    ci.lb    ci.ub
## intrcpt    0.2595  0.2323   1.1172  0.2639  -0.1958   0.7149
## ablat     -0.0292  0.0067  -4.3411  <.0001  -0.0424  -0.0160  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Parte VII

Dimensionamento campionario

Capitolo 24

Introduzione al dimensionamento campionario

24.1 Approcci e ambiti di dimensionamento

24.1.1 Errori nei test di ipotesi

Distinguiamo:

- errore di primo tipo: indicato con α è la probabilità di rifiutare l'ipotesi nulla quando essa in realtà è vera (trattamento non funziona ma concludiamo erroneamente che è efficace);
- errore del secondo tipo: indicato con β è la probabilità di non rifiutare la nulla quando essa è in realtà falsa (il trattamento è efficace ma concludiamo erroneamente il contrario)

24.1.2 Giustificazione del dimensionamento

L'attività di dimensionare propriamente uno studio è volta ad evitare:

- studi sottodimensionati: con poche probabilità di dimostrare una differenza, qualora sia presente, sono di fatto poco etici poiché si sottopongono pazienti ad uno studio senza che ve ne sia un beneficio scientifico;
- studi sovradimensionati: oltre ad essere costosi sono poco etici poiché se con meno pazienti potessimo dimostrare l'efficacia del trattamento non esporremmo i pazienti in eccesso del gruppo di controllo ad un trattamento subottimale.

24.1.3 Approcci al dimensionamento

Due approcci possibili, complementari e che rispondono a necessità differenti, al dimensionamento(Chow *e altri*, 2007):

- *precision analysis*: si vuole determinare quanti pazienti sono necessari al fine di ottenere stime con la precisione (es ampiezza intervallo di confidenza)

reputata ragionevole/di interesse (maggior enfasi sul controllo dell'errore di primo tipo);

- *power analysis* si vuole determinare quanti soggetti sono necessari (in un campione) per avere una determinata probabilità (es 80%) di identificare (ossia test statisticamente significativi) una data differenza, reputata clinicamente rilevante, qualora essa esista effettivamente (nella popolazione). (maggior enfasi sul controllo dell'errore del secondo tipo)

24.1.4 Ambiti di dimensionamento

Dimensionamento può essere fatto in 4 ambiti differenti:

- *stima*: caso classico dove si calcolano i soggetti necessari in base alla domanda
- *giustificazione*: dove si fornisce una giustificazione per un campione selezionato a priori (es sulla base di fattibilità)
- *adjustment*: dove il sample size stimato in precedenza viene aggiustato per fattori come dropout o covariate, al fine di fornire un numero sufficiente per l'analisi da condurre
- *re-estimation*: nell'ambito degli studi con analisi ad interim si provvede ad aggiustamento basato sulle informazioni raccolte

24.2 Ipotesi a confronto e disegni

Per fornire un calcolo affidabile deve essere scelto in anticipo un test per l'ipotesi di interesse, determinato in base al disegno dello studio e alla domanda alla quale vuole rispondere. Spesso il focus principale è l'efficacia o la *safety*, e per entrambe si possono configurare confronti di (Chow *e altri*, 2007):

- *eguaglianza*: le ipotesi al confronto sono:

$$H_0 : \mu_T = \mu_C \quad \text{vs} \quad H_1 : \mu_T \neq \mu_C \tag{24.1}$$

  con μ_C, μ_T la risposta nella variabile outcome per controlli (o valore teorico) e trattati rispettivamente. Il rifiuto della nulla suggerisce vi sia differenza tra trattati e controlli;

- *non inferiorità di un margine*: le ipotesi al confronto sono

$$H_0 : \mu_C - \mu_T \geq \delta \quad \text{vs} \quad H_1 : \mu_C - \mu_T < \delta \tag{24.2}$$

  dove δ è una differenza di interesse clinico ritenuta importante. Il rifiuto della nulla suggerisce che la differenza tra trattati e controlli sia inferiore ad una differenza rilevante δ, e quindi il trattamento sperimentale sia efficace quanto la terapia standard (comune in trial di efficacia dove il trattamento sperimentale è meno tossico dello standard, più facile da amministrare o meno costoso);

- *superiorità di un margine*: le ipotesi al confronto sono

$$H_0 : \mu_T - \mu_C \leq \delta \quad \text{vs} \quad H_1 : \mu_T - \mu_C > \delta \tag{24.3}$$

  Nel caso si rifiuti la nulla la differenza tra trattati e controlli è maggiore di una soglia reputata rilevante e in questo senso il trattamento sperimentale è superiore rispetto alla terapia standard.
  Da notare che le ipotesi di cui sopra testano la cosiddetta *superiorità clinica*; nel caso $\delta = 0$ alle ipotesi di sopra ci si riferisce come di *superiorità statistica*

- *equivalenza*; le ipotesi confrontate sono

$$H_0 : |\mu_T - \mu_C| \geq \delta \quad \text{vs} \quad H_1 : |\mu_T - \mu_C| < \delta \tag{24.4}$$

  Nel caso la nulla venga rifiutata, si conclude che la differenza tra trattati e controlli non sia clinicamente rilevante

Le ipotesi di confronto (eguaglianza, ecc) debbono essere ben chiare all'atto del dimensionamento poiché il dimensionamento (formule) dipende da esse.

24.3 Considerazioni assortite

24.3.1 Test a una o due code

Nell'ambito della power analysis lo studio può prevedere un test ad una o due code:

- il beneficio del test ad una coda, a parità di altre condizioni è che richiede tipicamente un campione inferiore
- lo svantaggio è che ci si espone maggiormente ad un rischio di falsi positivi; nel caso non vi sia efficacia del trattamento commettiamo un errore del primo tipo nel 5% dei casi con un test ad una coda, solo nel 2.5% dei casi per un test a due code. Nella realtà se per l'approvazione di un trattamento occorrono 2 trial indipendenti, il rischio di falsi positivi nel caso questi trial usino test a una coda è $0.05^2 = 0.0025 = 0.25\%$, mentre nel caso di due trial che impieghino test a due code il rischio di falsi positivi scende a $0.025^2 = 0.000625 = 0.0625\%$.

Alcuni ricercatori reputano 0.25% un rischio comunque accettabile, giustificando l'impiego di test a una coda, l'FDA sembra preferire test a due code

24.3.2 Aggiustamenti per dropouts

In presenza di dropout e dati per l'outcome principale

Example 24.3.1. Se il calcolo del sample size ci suggerisce 86 pazienti, ma si reputa verosimile un dropout del 20% sarà necessario arruolare

$$(86/(100 - 20)) \cdot 100 = 86/0.8 = 107.5 \approx 108$$

pazienti

24.3.3 Pacchetti R

Pacchetti utili:

- `TrialSize` implementa le funzioni per (Chow *e altri*, 2007)
- `presize` calcola il campione sulla base di stima e ampiezza dell'intervallo di confidenza (oppure l'ampiezza garantita da un campione)
- `CRTSize` sample size per cluster randomized trials
- `clinfun` ha funzioni per dimensionamento e analisi di studi di fase 2, test esatto di Fisher
- `CRM` ha un Continual Reassessment Method per le fasi 1

Capitolo 25

Un gruppo

25.1 Precision analysis - casi base

In questi casi si determina il campione basandosi sull'errore di primo tipo e utilizzando gli approcci degli intervalli di confidenza.

La precisione di una stima dipende dall'ampiezza del suo intervallo di confidenza:

1. direttamente dal livello di confidenza ad $(1-\alpha)\cdot 100\%$
2. direttamente dalla variabilità del fenomeno
3. inversamente dal numero di soggetti impiegati nella stima

Essendo tipicamente i primi due parametri considerati fissi/dati si agisce sul terzo al fine di avere una precisione di stima (ampiezza dell'intervallo di confidenza) accettabile.

Definition 25.1.1 (Errore massimo (Maximum error))**.** Si chiama così la semiampiezza (ampiezza/2) massima dell'intervallo di confidenza che si è disposti ad accettare.

25.1.1 Stima di una media

Nel caso n iid normali $y_1, \ldots, y_n$ con media μ e varianza σ^2.

25.1.1.1 Intervallo a due code

Varianza della popolazione nota Qualora la varianza sia nota un intervallo di confidenza a due code per μ, con un livello di confidenza pari a $(1-\alpha)\cdot 100\%$ è dato da:

$$\hat{\mu} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{25.1}$$

L'errore massimo che siamo disposti a commettere è

$$E = |\hat{\mu} - \mu| = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{25.2}$$

e risolvendo per n si ottiene l'ampiezza campionaria in grado di garantirlo

$$n = \frac{z^2_{1-\alpha/2} \cdot \sigma^2}{E^2} \tag{25.3}$$

Si noti come in questo approccio non si fa uso dell'errore β.

TODO: dire meglio qui

Example 25.1.1. Vogliamo determinare il campione necessario per avere il 95% di probabilità che l'errore nella stima effettuata sia meno del 10% della deviazione standard del fenomeno (ossia 0.1σ). Si ha che l'errore massimo è

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 0.1\sigma$$

E pertanto

$$n = \frac{z^2_{1-\alpha/2}\sigma^2}{E^2} = \frac{(1.96)^2\sigma^2}{(0.1\sigma)^2} = 384.2 \approx 385$$

Varianza ignota Nel caso in cui la varianza non sia conosciuta e occorra stimarla, la formula dell'intervallo si modifica come segue

$$\hat{\mu} \pm t_{1-\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}} \tag{25.4}$$

da cui

$$E = t_{1-\alpha/2,n-1}\frac{\sigma}{\sqrt{n}} \tag{25.5}$$

e specularmente per il calcolo dell'ampiezza:

$$n = \frac{t^2_{1-\alpha/2,n-1} \cdot \hat{\sigma}^2}{E^2} \tag{25.6}$$

25.1.1.2 Intervallo a una coda

Nel caso di intervallo ad una coda, si prende come misura di errore massimo la distanza tra la stima e il valore dell'intervallo calcolato (non quello infinito, chiaramente). Ad esempio il limite inferiore di un intervallo di confidenza ad una coda (facciamo l'esempio di varianza nota per brevità di notazione)

$$L = \hat{\mu} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \tag{25.7}$$

con l'intervallo che va da L a $+\infty+$. L'errore massimo che vorremo commettere sarà dunque paria a

$$E = z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \tag{25.8}$$

e risolvendo per n la formula per il dimensionamento è

$$n = \frac{z^2_{1-\alpha} \cdot \sigma^2}{E^2} \tag{25.9}$$

Example 25.1.2. Per un intervallo ad una coda al 95% con errore non superiore al 10% della deviazione standard: si ha che che l'errore massimo è

$$z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 0.1\sigma$$

E pertanto

$$n = \frac{z_{1-\alpha}\sigma^2}{E^2} = \frac{(1.65)^2\sigma^2}{(0.1\sigma)^2} = 272.2 \approx 273$$

25.1.2 Stima di una proporzione

25.1.2.1 Intervallo a due code

La formula dell'intervallo asintotico di una proporzione è

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

da cui l'errore

$$E = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \tag{25.10}$$

e la formula del campione

$$n = \frac{z^2_{1-\alpha/2} \cdot p \cdot (1-p)}{E^2} \tag{25.11}$$

Example 25.1.3. Se si desidera nell'ipotesi che $p = 0.5$ che l'intervallo di confidenza abbia una semiampiezza di 0.05 si ha

$$\frac{1.96^2 \cdot 0.5 \cdot (1-0.5)}{0.05^2} = 384.1 \approx 385$$

Remark 92. A parità di livello di significatività ed ampiezza massima di errore della stima, la numerosità dipende dalla variabilità $p(1-p)$ al numeratore; cautelativamente, qualora non si abbia una minima idea di quale possa essere la prevalenza, adottare $p = 0.5$

Example 25.1.4. In figura 25.1 la semi ampiezza di un intervallo di confidenza al 95% per un campione di 150 pazienti al variare della prevalenza

```
e <- function(p, alpha = 0.05, tails = 2, n = 150){
    z <- qnorm(1 - alpha/tails)
    z * sqrt( p * (1-p) / n )
}

ps <- seq(0.1, 0.9, by = 0.01)
plot(x = ps, y = e(p = ps), xlab = 'p', ylab = 'E', pch = NA, ylim = c(0, 0.1))
lines(x = ps, y = e(p = ps))
```

25.2 Power analysis

Dato che un errore del primo tipo è solitamente considerato più importante/grave, un approccio tipico al test di ipotesi è di controllare α ad un livello accettabile e cercare di minimizzare β scetliendo un sample size adeguato.

25.2.1 Test per una media

Qui supponiamo di avere dati di una variabile quantitativa su n soggetti $x_1, \ldots, x_n$; media e varianza campionaria sono rispettivamente

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{25.12}$$

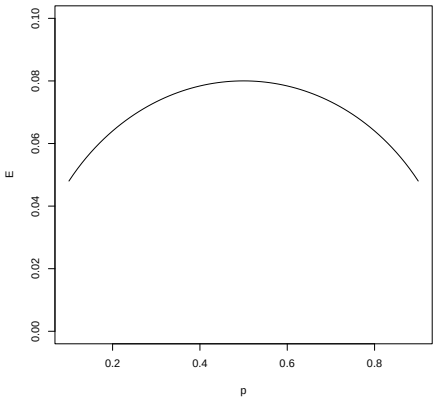


Figura 25.1: Semiampiezza intervallo di confidenza per 150 pz e varie prevalenze

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{25.13}$$

25.2.1.1 Equivalenza

Vogliamo verificare se la risposta media della popolazione μ sia differente o meno da un valore di riferimento μ_0 e denominiamo $\epsilon = \mu - \mu_0$ la differenza. Chiaramente:

$$\epsilon = 0 \iff \mu = \mu_0$$

Per verificare se vi sia una differenza tra la risposta media e il valore di riferimento le ipotesi poste a confronto sono:

$$H_0 : \epsilon = 0, \quad H_1 : \epsilon \neq 0$$

Varianza conosciuta Ipotizzando di conoscere la deviazione standard σ del carattere nella popolazione, il test da applicare è lo z:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

si ha che:

- sotto ipotesi nulla che $\epsilon = 0$, z si distribuisce come $N(\epsilon, 1) = N(0, 1)$, da cui deriva che con test bilaterale rifiutiamo H_0 se $|z| > z_{1-\alpha/2}$, e se $\alpha = 0.05$ per $|z| > 1.96$;

- sotto ipotesi alternativa z si distribuisce come $N(\epsilon^*, 1)$ con

$$\epsilon^* = \frac{\epsilon}{\sigma/\sqrt{n}} = \frac{\epsilon\sqrt{n}}{\sigma}$$

con $\epsilon \neq 0$. La potenza di tale test è la probabilità di ottenere un risultato oltre la soglia rifiuto nell'ipotesi che sia vera l'alternativa, ossia:

$$\begin{aligned}
\mathbb{P}\left(|N(\epsilon^*, 1)| > z_{1-\alpha/2}\right) &= \mathbb{P}\left(N(\epsilon^*, 1) > z_{1-\alpha/2}\right) + \mathbb{P}\left(N(\epsilon^*, 1) < -z_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(N(0, 1) > z_{1-\alpha/2} - \epsilon^*\right) + \mathbb{P}\left(N(0, 1) < -z_{1-\alpha/2} - \epsilon^*\right) \\
&\overset{(1)}{=} \mathbb{P}\left(N(0, 1) < \epsilon^* - z_{1-\alpha/2}\right) + \mathbb{P}\left(N(0, 1) < -z_{1-\alpha/2} - \epsilon^*\right) \\
&= \Phi(\epsilon^* - z_{1-\alpha/2}) + \Phi(-\epsilon^* - z_{1-\alpha/2}) \\
&= \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma} - z_{1-\alpha/2}\right) + \Phi\left(-\frac{\epsilon\sqrt{n}}{\sigma} - z_{1-\alpha/2}\right)
\end{aligned}$$

dove in (1) abbiamo sfruttato la simmetria della normale. Ignorando una piccola parte di potenza ($\leq \alpha/2$), possiamo dire che la potenza è approssimativamente pari a

$$\Phi\left(\frac{\epsilon\sqrt{n}}{\sigma} - z_{1-\alpha/2}\right) + \Phi\left(-\frac{\epsilon\sqrt{n}}{\sigma} - z_{1-\alpha/2}\right) \approx \Phi\left(\left|\frac{\epsilon\sqrt{n}}{\sigma}\right| - z_{1-\alpha/2}\right) = \Phi\left(\frac{|\epsilon|\sqrt{n}}{\sigma} - z_{1-\alpha/2}\right)$$

Affinché la potenza sia $1 - \beta$ con β scelto a piacere, si deve avere

$$\Phi\left(\frac{|\epsilon|\sqrt{n}}{\sigma} - z_{1-\alpha/2}\right) = 1 - \beta$$

ossia

$$\frac{|\epsilon|\sqrt{n}}{\sigma} - z_{1-\alpha/2} = z_{1-\beta}$$

e risolvendo per n si giunge a

$$\begin{aligned}
\sqrt{n}\,|\epsilon| &= (z_{1-\beta} + z_{1-\alpha/2})\sigma \\
n &= \frac{(z_{1-\beta} + z_{1-\alpha/2})^2\sigma^2}{\epsilon^2}
\end{aligned}$$

In merito a quest'ultima alcuni libri scrivono z_β al posto di $z_{1-\beta}$ e $z_{\alpha/2}$ al posto di $z_{1-\alpha/2}$, ma è corretto come si è fatto qui.

Example 25.2.1. Riproducendo, se $\alpha = 0.05$, $\beta = 0.2$, $\sigma = 1$, $\epsilon = 0.5$ si ha

```
num <- (qnorm(1-0.05/2) + qnorm(1 - 0.2))^2 * 1^2
den <- 0.5^2
num / den

## [1] 31.39552
```

Varianza ignota Quando σ^2 è sconosciuta può esser rimpiazzata dalla varianza campionaria data in 25.13 e l'ipotesi H_0 viene rifiutata se

$$\left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right| > t_{1-\alpha/2,n-1}$$

dove $t_{1-\alpha/2,n-1}$ è il quantile $1 - \alpha/2$ della distribuzione t con $n - 1$ gradi di libertà

```r
beta <- function(theta, alpha = 0.05, max_beta = 0.2, N = 2:45){
    # per ogni campione indagato, fai il calcolo di beta ottenuto
    res <- lapply(N, function(n){
        df <- n - 1
        ncp <- sqrt(n) * theta
        t <- qt(1 - alpha/2, df = df, ncp = 0)
        beta <- pt(q = t, df = n-1, ncp = ncp)
        data.frame(n = n, beta = beta, ok = beta < max_beta)
    })
    do.call(rbind, res)
}

beta(theta = 0.5)
```

```
##     n       beta    ok
## 1   2 0.94690414 FALSE
## 2   3 0.92106852 FALSE
## 3   4 0.89191676 FALSE
## 4   5 0.86154723 FALSE
## 5   6 0.83067614 FALSE
## 6   7 0.79962458 FALSE
## 7   8 0.76859262 FALSE
## 8   9 0.73773267 FALSE
## 9  10 0.70717143 FALSE
## 10 11 0.67701744 FALSE
## 11 12 0.64736426 FALSE
## 12 13 0.61829219 FALSE
## 13 14 0.58986950 FALSE
## 14 15 0.56215341 FALSE
## 15 16 0.53519106 FALSE
## 16 17 0.50902038 FALSE
## 17 18 0.48367096 FALSE
## 18 19 0.45916479 FALSE
## 19 20 0.43551710 FALSE
## 20 21 0.41273703 FALSE
## 21 22 0.39082828 FALSE
## 22 23 0.36978977 FALSE
## 23 24 0.34961622 FALSE
## 24 25 0.33029864 FALSE
## 25 26 0.31182487 FALSE
## 26 27 0.29418002 FALSE
## 27 28 0.27734685 FALSE
## 28 29 0.26130620 FALSE
## 29 30 0.24603728 FALSE
## 30 31 0.23151800 FALSE
## 31 32 0.21772526 FALSE
## 32 33 0.20463518 FALSE
## 33 34 0.19222331  TRUE
## 34 35 0.18046489  TRUE
## 35 36 0.16933492  TRUE
```

```
## 36 37 0.15880842  TRUE
## 37 38 0.14886050  TRUE
## 38 39 0.13946650  TRUE
## 39 40 0.13060209  TRUE
## 40 41 0.12224334  TRUE
## 41 42 0.11436681  TRUE
## 42 43 0.10694963  TRUE
## 43 44 0.09996950  TRUE
## 44 45 0.09340477  TRUE
```

Per cui si nota che all'aumentare del campione l'errore di secondo tipo diminuisce e ad un sample size di 34 si ottiene un $\beta < 0.2$ mentre per una potenza del 90% occorrono 44 soggetti.

25.2.1.2 Superiority/Non-inferiority

Disegni di non inferiorità e superiorità possono essere unificati dal seguente test a una coda

$$H_0 : \epsilon \leq \delta, \quad H_1 : \epsilon > \delta$$

con δ detto margine di superiorità o non inferiorità. Quando:

- $\delta = 0$: il disegno è una superiorità classica (in senso statistico), ossia confronta $H_0 : \mu \leq \mu_0$ con $H_1 : \mu > \mu_0$
- $\delta > 0$: disegno di superiorità che verifica che la differenza tra media della popolazione e valore teorico sia superiore ad un dato valore δ
- $\delta < 0$: il disegno è di non inferiorità; mira a verificare che il valore della popolazione sia entro una certa distanza δ dal valore ipotizzato μ_0.
  In altre parole se si verifica la nulla $\mu - \mu_0 \leq \delta$, ossia $\mu \leq \mu_0 + \delta$ (con $\delta < 0$); se si verifica l'alternativa $\mu - \mu_0 > \delta$ ossia $\mu_0 - \mu < -\delta$ con $-\delta > 0$

Varianza nota Ipotizzando che σ^2 sia nota il test che impieghiamo per la scelta è

$$\frac{\overline{x} - \mu_0 - \delta}{\sigma/\sqrt{n}}$$

Pertanto è come se si passasse dal test sul valore puntuale vs valore teorico al test della differenza $(\overline{x} - \mu_0)$ vs una sorta di valore teorico della differenza δ. Il test rispetto ad una z normale non cambia (perché è come se incrementassimo la costante data dall'ipotesi nulla al numeratore), quindi si distribuisce sempre con $N(0, 1)$. Essendo un test ad una coda sul valore scarto ϵ vs δ la zona di rifiuto di H_0 è tutta a destra, per cui H_0 è rifiutata se

$$\frac{\overline{x} - \mu_0 - \delta}{\sigma/\sqrt{n}} > z_{1-\alpha}$$

Il test ha una distribuzione normale centrata sul valore δ; sotto ipotesi alternativa, se $\epsilon > \delta$ la distribuzione dello stimatore non è centrata su δ ma su $\epsilon - \delta$, che in unità di misura della distribuzione nulla sono

$$\frac{\epsilon - \delta}{\sigma/\sqrt{n}}$$

quindi sotto ipotesi alternativa il test si distribuisce come

$$z \sim N\left(\frac{\epsilon - \delta}{\sigma/\sqrt{n}}, 1\right)$$

Siamo interessati alla potenza del test, ossia a:

$$\begin{aligned}\mathbb{P}\left(N\left(\frac{\epsilon - \delta}{\sigma/\sqrt{n}}, 1\right) > z_{1-\alpha}\right) &= \mathbb{P}\left(N(0,1) > z_{1-\alpha} - \frac{\epsilon - \delta}{\sigma/\sqrt{n}}\right) \\ &= \mathbb{P}\left(N(0,1) < \frac{\epsilon - \delta}{\sigma/\sqrt{n}} - z_{1-\alpha}\right) \\ &= \Phi\left(\frac{\epsilon - \delta}{\sigma/\sqrt{n}} - z_{1-\alpha}\right)\end{aligned}$$

Impostandola al livello di potenza desiderato:

$$\Phi\left(\frac{\epsilon - \delta}{\sigma/\sqrt{n}} - z_{1-\alpha}\right) = 1 - \beta$$

da cui

$$\frac{\epsilon - \delta}{\sigma/\sqrt{n}} - z_{1-\alpha} = z_{1-\beta}$$

e risolvendo per n si giunge a

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{(\epsilon - \delta)^2}$$

Varianza sconosciuta

25.2.1.3 Equivalence

25.2.2 Test per una proporzione

Parte VIII

Questionari

Capitolo 26

Introduzione

Qui ci concentra sulla traduzione e validazione di questionari in italiano; si assume che il questionario validato in inglese esista già. Le fasi:

- traduzione
- adattamento culturale
- validazione/valutazione

Ci si basa sulla revisione di Sousa e Rojjanasrirat (2011) e su Arafat (2016) .

26.1 Concetti introduttivi

SL source language (inglese)

TL target language (italiano)

26.1.1 Definizioni

reliability capacita del questionario di fornire risultati coerenti e riproducibili: viene analizzata per lo più in termini di

internal consistency capacità di uno sturmento di avere item suoi componenti interrelati; misurata mediante alfa di Cronbach

test-retest consistency capacità degli score di uno strumento di essere riproducibili se utilizzati sullo stesso paziente mentre le sue condizioni non sono cambiate; per testarla si applica lo stesso strumento agli stessi responder dopo un tot di tempo, solitamente due settimane e si valuta agreement nel tempo in condizioni idealmente di non variazione del concetto misurato

inter-rater reliability si valuta agreement tra differenti rater che giudicano una stessa cosa/problematica

validity si tratta della capacità di uno strumento di misurare quel che deve misurare (e non altro). Attiene/commisura aspetti differenti (la valutazione di alcuni piuttosto che altri dipende dal progetto):

face validity capacità di uno strumento di essere comprensibile e rilevante per la popolazine targettata. Per questa serve la fase di back translation durante la traduzione/adattamento, la revisione critica di un panel di esperti (su fattibilità leggibilita/chiarezza, coerenza di stile e formattazione);

content validity capacità di riflettere il dominio di interesse e la definizione concettuale di un costrutto; assicurata dalla fase di back translation, dlala revisione di letteratura ed opinione di panel di esperti e dal calcolo di indici di content validity (content validity index)

construct validity il grado in cui una misura e legata ad una variabile specificata, in accordo con una determinata teoria; viene valutata sulla base di un'analisi fatoriale e/o comparando con altri strumenti simili

divergent validity item in una sottoscala non dovrebbero correlare troppo con item esterni alla stessa o con lo score complessivo di un'altra sottoscala

discriminative validity capacità di distinguere tra gruppi per i quali ci si attende una differenza sulla base di una diagnosi clinica o altre caratteristiche

criterion validity valutazione di uno strumento (da prendere come proxy/surrogato) contro il il valore vero o uno standard accettato come valore vero

convergent validity il grado in cui gli score della misura si associano con score altri score disponibili che intendono valutare un costrutto simile; convergent e concurrent possono esser valutate comparando lo strumento con altri similari con misure di correlazione

concurrent validity associazione dello strumento con standards accettati

26.1.2 Sample size

La validazione di un questionario è uno studio multi-attività dove vengono analizzate varie caratteristiche dello strumento; nello specifico non guidato da una singola ipotesi inquadrabile in un test, per il quale si desidera avere una determinata potenza e anche per tali motivi si può affermare che non esista un consenso sui metodi di determinazione dell'ampiezza campionaria paragonabile a quello rinvenuto in altre aree di studio (es diagnostica/interventistica) Anthoine *e altri* (2014).
Per questi motivi le ampiezze campionarie per ciascuna attività/fase del presente progetto sono determinate contemperando da un lato *good-practices*/indicazioni di *guidelines* provenienti dalla letteratura (Sousa e Rojjanasrirat, 2011) e dall'altro la fattibilità di progetto.

26.2 Traduzione

26.2.1 Stadio 1: traduzione iniziale (forward)

Forward translation (dal SL a TL) svolta da due traduttori (o se possibili due team) madrelingua italiani con distinti background (team accomunati per caratteristiche):

- il primo conscio della terminologia medica e del settore nel quale lo strumento viene adoperato (*informed* translator), tipicamente un clinico
- il secondo non dovrebbe essere del settore ma piuttosto esperto delle peculiarità linguistiche e culturali dell'italiano (*uninformed* translator)
- entrambi dovrebbero essere mediamente conosci sia della cultura italiana che inglese (*bilingual* e *bicultural* translators)

Vengono prodotte le traduzioni TL1 (più medica) e TL2 (più linguaggio comune)

26.2.2 Stadio 2: comparazione delle traduzioni e sintesi

- un terzo soggetto indipendente confronta le due traduzioni (Istruzioni, item e formati di risposta) separatamente con il questionario originale, dopodiché confronta delle due traduzioni (TL1 e TL2) tra loro per identificare ambiguità e discrepanze (di parole, frasi o significati);
- Qualsiasi ambiguità/discrepanza deve essere risolta da un comitato composto da: i primi due translator, quest'ultimo che ha messo in luce le discrepanze, investigator e altri membri del team di ricerca

Alla fine del processo viene generata la versione preliminare iniziale della traduzione (PI-TL)

26.2.3 Stadio 3: back translation

La PI-TL viene ritradotta in inglese da due altri soggetti (o team)

- medesime qualifiche di quelli dello step 1: uno di ambito medico (conosce il lessico) e l'altro no
- madrelingua inglesi
- non essere assolutamente a conoscenza dello strumento originale (mai visto prima)

Verranno così prodotti due versioni inglesi, la B-TL1 e B-TL2

26.2.4 Stadio 4: comparazione delle traduzioni e sintesi

Si

- forma un team composto da un metodologo (es investigator o membro del research team) un clinico con esperienza nel contesto e tutti i quattro traduttori usati negli step 1 e 3

- lo sviluppatore dello strumento originale è un plus (per consigli o chiarimenti sui costrutti del questionario), così come avere almeno un membro del team/committee solamente inglese
- il team compara le back translation B-TL1 e B-TL2 con lo strumento originale e tra loro;
- se discrepanze nn risolvibili potrebbe esser necessario risvolgere gli step 1-4 con nuovi traduttori/soggetti, in toto o sugli item difficili

Eventuali discrepanze, a quanto si comprende, servono per modificare la PI-TL iniziale (dal quale B-TL1 e B-TL2 hanno origine) e derivare una versione pre finale in italiano (P-FTL), utile per i primi test psicometrici.

26.2.5 Stadio 5: test pilota e revisione panel esperti

Si fa:

- test pilota di P-FTL tra pazienti italiani madrelingua, utilizzando un campione tra i 10 e i 40 (in base alla disponibilità): ogni partecipante valuta dicotomicamente (chiaro, non chiaro) le istruzioni del questionario gli item e formati di risposta, e se non chiaro di fornire suggerimenti. Cose che sono reputate non chiare da almeno il 20% dovrebbero essere riviste
- raccomandato formare un gruppo di 6-10 esperti (clinici dell'ambito, italiani, esperti di dove lo strumento deve essere impiegato) danno rate chiaro/non chiaro (sempre su istruzioni item e risposte) e se non chiaro consigli. Anche qui cose che sono reputate non chiare da almeno il 20% dovrebbero essere riviste
- raccomandato che gli esperti valutino la content equivalence degli item sia a livello di singolo item (si calcola I-CVI) che a livello di scala (per calcolare S-CVI, e nello specifico la variante S-CVA/Ave) (Sousa e Rojjanasrirat, 2011), come anche Kappa di Cohen; Facendo uso di 10 esperti I-CVI deve essere almeno 0.78 e il S-CVA/Ave almeno il 0.90, il kappa di 0.60
  Item che non raggiungono score minimi debbono essere rivisti e rivalutati e iterativamente si deve raggiungere questi standard.

26.2.6 Stadio 6: testing preliminare con un campione bilingue

Qualora si abbia un campione della popolazione target che sia bilingue si può fare questa valutazione. Alternativamente passare alla fase 7. Si fa che:

- almeno 5 soggetti per item
- ai partecipanti viene somministrato il P-FTL (inglese), al quale rispondono senza vedere lo strumento originale (sempre in inglese)
- dopodiché ad essi viene somministrato lo strumento originale, magari con item ordinati diversamente
- le risposte vengono comparate (agreement o correlazione)

26.2.7 Stadio 7: full psychometric testing

Si fa una valutazione di:

- internal consistency reliability (sensitiivity/specificity)
- stability reliability (test-retest)
- homogeneity
- construct-related validity (convergent/divergent validity)
- criterion-related validity (concuyrrent and/or predictive validity)
- factor structure dello strumento (dimensionality)
- model fit

Per il sample size:

- almeno 10 soggetti per item se factor analysis esplorativa, correlazione di pearson, scale e item analysis
- 300-500 se confirmatory, oppure fare power analysis

26.3 Validazione psicometrica

Qui ci si basa/traduce per lo più su Terwee *e altri* (2007).

26.3.1 Internal consistency

La consistenza interna è la misura nella quale gli item appartenenti ad una scala o sottoscala sono correlati, quindi misuranti lo stesso concetto.
Una scala internamente consistente (omogenea/unidimensionale) si ottiene attraverso la buona definizione di costrutti, buoni item e in seguito l'analisi componenti principali o l'analisi fattoriale esplorativa, seguita da un'analisi fattoriale confermatoria. Quando la consistenza interna è rilevante, l'anbalisi componenti principali o l'analisi fattoriale dobvrebbero essere applicate per determinare se gli item formano solo una scala/dimnensione o più di una:

- nel caso non vi siano ipotesi a priori riguardanti la dimensionalità del questionario, un'analisi esplorativa tipo componenti principali o analisi fattoriale può essere applicata
- ma se vi è una chiara ipotesi riguardante la struttura (es per l'esistenza di un modello teorico o perchèp la struttura dei fattori è stata determinata previously, l'analisi fattoriale confermatoria dovrebbe essere usata

La valutazione di consistenza interna si applica:

- al questionario nel suo complesso se questo misura un solo aspetto/scala
- ai vari item che corrispondono ai vari concetti/subscale che il questionario misura

e si misura mediante l'alpha di Cronbach; se questo è

- troppo basso indica una mancanza di correlazione tra gli item che dovrebbero definitre uno stesso concetto e che la sommarizzazione non ha molto senso
- uno troppo alto indica una alta correlazione tra gli intem e quindi il rischio di ridondanza

Si ha che Terwee *e altri* (2007) consigliano una alpha di Cronbach tra 0.70 e 0.95

26.3.2 Criterion validity

Come lo strumento si relazione ad un gold standard: correlazione di almeno 0.7 è un argomento convincente secondo Terwee *e altri* (2007). Occhio e croce (a mio parere e ignorando i ties) potrebbe essere valida una correlazione lineare di Pearson se il confronto è considerato gold standard (es se si vuole la relazione lineare/sostituitiva) o Spearman altrimenti

26.3.3 Construct validity

Lo strumento si comporta come dovrebbe fare, in linea teorica, sulla base di differenze che dovrebbero esistere. I comportamenti attesi debbono essere specificati in anticipo e in seguito testati.

Terwee *e altri* (2007) giudicano positivamente il questionario rispetto a questo parametro se le ipotesi sono specificate in anticipo e almeno il 75% dei risultati sono in linea con le ipotesi, in gruppi/sottogruppi di almeno 50 pazienti

26.3.4 Reproducibility

non convincente su Terwee *e altri* (2007)

26.3.5 Responsiveness

Si tratta di variazione nel tempo dello score in seguito alla variazione nelle condizioni materiali che lo score vuole misurare; ottimo articolo per questo è Deyo e Centor (1986).

Analogamente alla construct validity, dovrebbe essere valutata testando predefinite ipotesi, es correlazioni attese tra cambiamenti nelle misure o differenze attese nei cambiamenti tra gruppi conosciuti.

Un altro metodo è l'AUC di una ROC proposta da Deyo e Centor (1986), misura dell'abilità di un quoestionario di distinguere pazienti che hanno e non hanno cambiato according to un criterio esterno. Un AUC di almeno 0.7 è adeguato secondo Terwee *e altri* (2007)

26.3.6 Floor/ceiling effect

Sono considerati presenti se più del 15% dei rispondenti ottiene il punteggio minimo o massimo rispettivamente. Se cio avviene è possibile che la content validity dello strumento ne risenta, che questi pz non possano essere distinti e che la reposnsività a variazioni nel tempo sia bassa.
Terwee *e altri* (2007) giudicano positiva una scala se non si ha floor o ceiling in un campione di almeno 50 pazienti.

26.3.7 Interpretability

Meh

26.4 Statistiche di interesse

26.4.1 Cronbach's α

Ci si basa su DeVellis (2012), l'articolo da citare è Cronbach (1951). L'indice è utilizzato per la verifica del legame degli item tra loro; l'idea è che un legame/correlazione alta degli item sia indice che gli item stiano misurando lo (o siano determinati dallo) stesso costrutto latente. Alcuni punti:

- se un set di item misura più costrutti/sottoscale, il calcolo dell'α va effettuato per ciascuna sottoscala separatamente
- per item dicotomici è utilizzata la forma KR20 di Kuder Richardson

Se lo **score è la somma di tutti gli item** allora la variabilità dello score è la somma degli elementi della matrice di varianza/covanza:

$$\sigma_y^2 = \sum \sigma_i^2 + \sum \sigma_{i,j}$$

con σ_y^2 la varianza della scala ottenuta dalla somma di i item, σ_i^2 sono le varianze di quest'ultimi e $\sigma_{i,j}$ le covarianze desumibili dalla matrice di varianza/covarianza.
L'indice alpha mira a calcolare la quota di variabilità di y dovuta alle covarianze (escludendo dunque quella dovuta alle varianze) per avere una stima di quanta variabilità della variabile è dovuta al covariare dei suoi elementi costituenti (item). La formula è

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_y^2}\right)$$

In quest'ultimo caso $\frac{k}{k-1}$ è un mero fattore di normalizzazione per fare si che l'indice tra 0 e 1.
L'indice ha anche un'altra specificazione, conosciuta come Spearman-Brown prophecy formula che fa uso del numero totale di item k e della correlazione media tra le coppie di item $\overline{r}$

$$\alpha = \frac{k\overline{r}}{1 + (k-1)\overline{r}}$$

la quale mostra come l'alpha/reliability sia funzione diretta del numero di item della scala e della correlazione media. Piu item una scala contiene e maggiore la correlazione media tra questi, maggiore sarà l'alpha.

26.4.2 ICC

26.4.3 Cohen's κ

È una statistica che si pone di misurare l'agreement tra *due* rater (o entro rater in due tempi) più robusta della percentuale di casi in cui le valutazioni sono in accordo perché cerca di scorporare la quota di agreement che avverrebbe anche casualmente (stimando una sorta di eccesso di agreement).

	Yes	No
Yes	a	b
No	c	d

Tabella 26.1: Tabella 2×2 con conteggi frequenze ($a+b+c+d$ = item/soggetti valutati)

26.4.3.1 Definizione ed interpretazione

La definizione è:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{26.1}$$

Dove:

- p_o è l'agreement osservato (probabilità, ossia la percentuale di casi che risiedono sulla diagonale principale della tabella bivariata)
- p_e è la probabilità di agreement attesa dovuta al mero caso (ossia ipotizzando che i rater giudichino ogni giudizio con percentuali uguali alla marginale)
- $p_o - p_e$ è una sorta di eccesso di agreement sperimentato (oltre a quello dovuto al caso)
- $1 - p_e$ è la differenza tra agreement massimo raggiungibile (1 se tutte le valutazioni risiedono sulla diagonale principale) e quello dovuto al caso (funge da denominatore "normalizzante")

Pertanto l'indice è una stima della quota di agreement possibile, oltre a quello casuale, che è spiegato dai rater. L'indice ha range tra -1 e $+1$ e:

- il valore è negativo se un agreement casuale avrebbe fatto addirittura meglio
- il valore è 0 se non si va oltre un agreement anche spiegabile sulla base del caso
- il valore è tanto più vicino a 1 tanto più l'agreement è perfetto e tutte le valutazioni si concentrano sulla diagonale principale

Facciamo un esempio con due modalità (tabella 26.1) di rate Yes o No e vediamo come calcolare le componenti dell'indice Si ha che

$$p_o = \frac{a+d}{a+b+c+d}$$
$$p_e = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} + \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d}$$

26.4.3.2 Cutoff interpretazione

Un modo di interpretare il coefficiente (arbitrario) è quello fornito da Landis e Koch (1977), riportato in tabella 26.2.

Valore	Interpretazione
< 0	Poor agreement
$0.01 - 0.20$	Slight agreement
$0.21 - 0.40$	Fair agreement
$0.41 - 0.60$	Moderate agreement
$0.61 - 0.80$	Substantial agreement
$0.81 - 1.00$	Almost perfect agreement

Tabella 26.2: Cutoff di κ proposti in Landis e Koch (1977)

26.4.3.3 Esempi

Nel seguito alcuni esempi che partono dal caso di perfetto agreement (100 item classificati ugualmente da due rater, 50 classificati entrambi A e 50 entrambi B) al caso di perfetto disagreement (100 item classificati in maniera opposta, ove ci sono 50 item che sono classificati A dal primo rater e B dal secondo e altri 50 viceversa):

```
cases <- list(c(50, 0, 0, 50),
              c(40, 10, 10, 40),
              rep(25, 4),
              c(10, 40, 40, 10),
              c(0, 50, 50, 0))

tmp <- lapply(cases, function(x){
    tab <- as.table(matrix(x, ncol = 2, byrow = TRUE))
    colnames(tab) <- rownames(tab) <- c('A', 'B')
    kappa <- suppressWarnings(lbagree::cohen_k(tab)$unweighted)
    msg <- sprintf("La seguente tabella ha agreement (kappa) pari a: %+.2f (%+.2f to %+.2f).\n\
                   kappa[1],
                   kappa[2],
                   kappa[3])
    cat(msg)
    print(addmargins(tab))
    cat('\n')
})
```

```
## Error in loadNamespace(x):  there is no package called 'lbagree'
```

26.4.3.4 Versione pesata

Nel caso in cui i giudizi non siano nominali, bensì ordinali (e in misura maggiore di 2) si può pesare il disagreement per il "livello" dello stesso, adottando la weighted κ.
Tre matrici sono impiegate. Quella delle frequenze osservate, quella delle frequenze attese sotto agreement casuale e quella dei pesi del disagreement.
Ora ci sono diverse costruzioni possibili a seconda che la matrice dei pesi sia costruita in un modo o nell'altro. Qui seguiamo wikipedia, che a sua volta segue lo sviluppo di Cohen stesso (altrimenti vedi fleiss che è più intuitivo forse). La

matrice dei pesi è simmetrica, ha 0 sulla diagonale principale e le rimanenti indicano la "gravità del disagreement" (spesso la misura è lineare e consiste nel numero di step necessari per spostarsi sulla diagonale principale, ma può essere anche quadratica o altro).
La definizione è:

$$\kappa = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} e_{ij}w_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} o_{ij}w_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} e_{ij}w_{ij}}$$

dove k è il numero di giudizi possibili, dove w_{ij} son gli elementi nella matrice dei pesi o_{ij} delle frequenze osservate e_{ij} delle frequenze attese.

26.4.4 Fleiss κ

26.4.4.1 Definizione

Può essere pensata come una la generalizzazione ad n rater della K di Cohen. La definizione è analoga:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{26.2}$$

Si tratta di vedere come cambiano p_o e p_e in presenza di molteplici rater.
che vi siano:

- N pazienti (o item di un questionario), $i = 1, \ldots, N$ giudicati su
- k ($j = 1, \ldots, k$) categorie possibili da
- n rater

sia n_{ij} il numero di rater che hanno assegnato al paziente i il giudizio j; questa è la matrice delle frequenze dalla quale si calcola la *kappa*.
Calcoliamo la proporzione p_j di tutte le valutazioni fatte nella j-esima categoria come

$$p_j = \frac{\sum_{i=1}^{N} n_{ij}}{Nn}$$

dove al denominatore abbiamo il prodotto dei valutati per le valutazioni procapite. Chiaramente si ha che $\sum_{j=1}^{k} p_j = 1$. Sotto ipotesi di agreement casuale, l'agreement sarebbe

$$p_e = \sum_{j=1}^{k} p_j^2$$

ora mi riesce difficile interpretarlo, so far, ma è tipo l'indice di herfindal ergo qualcosa che sta tra $1/j$ (se i rater equidistribuiscono mediamente le valutazioni) e 1 (se i rater danno valutazioni su una sola categoria).
Vediamo ora su ciascun paziente/item quanto i rater concordino calcolando P_i: ossia calcoliamo quante coppie rater/rater sono in agreement rispetto al numero di coppie rater/rater possibili

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$

al numeratore abbiamo la somma di tutte le coppie formabili che sono in agreement (ciclando sul singolo giudizio), al denominatore fuori somma il totale di coppie formabili da n rater differenti.
L'agreement ottenuto è semplicemente una media degli agreement sui singoli pazienti/item

$$p_o = \frac{1}{N}\sum_{i=1}^{N} P_i$$

26.4.4.2 Cutoff interpretazione

Per i cutoff di interpretazione si può fare riferimento a quelli del κ di Cohen proposti da Landis e Koch (1977) e riportati in tabella 26.2

26.4.5 Lin's CCC

Alcuni casi di studio (abbastanza paradigmatici e poco reali) sono presentati in figura **??**, per due rater con misurazioni esclusivamente positive (passando da una situazione di perfetto disagreement ad una di perfetto agreement), giusto per capire l'andamento del coefficiente in situazioni differenti.
Negli esempi di figura **??** è schematizzato l'andamento in relazione alla pendenza; per vedere come cambiano stima e intervallo di confidenza se ad una situazione di agreement perfetto (Lin = 1) aggiungiamo via via sempre più "rumore" e variabilità, si fa riferimento a figura **??**.
Si notano due cose: non solo l'intervallo di confidenza si allarga (a testimoniare che la nuvola di punti la possiamo approssimare con un ventaglio sempre più ampio di rette, che a loro volta stanno a indicare un differente agreement) ma per effetto della variabilità anche il coefficiente di agreement stimato diminuisce (a testimoniare comunque una diminuzione di agreement tra i due rater rispetto alla situazione di linea perfetta).

```
## Error in loadNamespace(x):  there is no package called 'lbagree'
```

```
## Error in loadNamespace(x):  there is no package called 'lbagree'
```

26.4.6 OCCC

Interpretazione è analoga al coefficiente di Lin ma fornisce un giudizio di agreement complessivo per i n rater (come la fleiss kappa per la cohen's one)

26.5 Intro ad analisi fattoriale

26.5.1 Esplorativa

Secondo DeVellis (2012) serve a diverse cose:

- determinare il numero di variabili latenti, e darne una interpretazione, che un questionario sta misurando

- spiegare la variabilità di un set numerosi di item con un numero compatto di variabili nuove (fattori)
- identificare gli item che performano meglio o peggio all'interno di un questionario (es item che non si legano bene ad alcuno dei costrutti più evidenti possono esser considerati per l'eliminazione).

26.5.2 Confermativa

Capitolo 27

Analisi dei fattori

27.1 Introduzione alla metodologia

L'analisi fattoriale consiste nell'applicazione di una serie di tecniche d'analisi multivariata, che consentono di semplificare e sintetizzare un insieme notevole di indicatori (variabili di partenza) in un sottoinsieme più ristretto di altri indicatori (variabili latenti, o fattori) più significativi, garantendo una limitata perdita d'informazione.
Questo processo di creazione di nuove variabili viene condotto (generalmente si fa mediante un calcolatore) studiando le correlazioni delle variabili di partenza fra di loro, e raggruppando quelle fortemente correlate in una nuova variabile.

Vi sono due principali tipi di analisi fattoriale: esplorativa e confermativa.

esplorativa si applica quando il ricercatore no conosce quanti fattori son necesari per spiegare le relazioni tra un set di variabili; pertanto il ricercatore usa l'analisi dei fattori per espolrare le dimensioni sottostanti ad un costrutto di interesse. E' quello che faremo nel seguito

confermativa è utilizzata quando il ricercatore ha conoscenza della struttura sottostante del costrutto sotto analisi ed è interessato a sapere a che livello il set di fattori identificati fitti i dati a disposizione (questo si fa mediante tecniche di *structural equation modelling*

Adottando la *prima visuale*, l'analisi dei fattori ha come scopo l'identificazione di una struttura sottostante ad un insieme di variabili correlate fra loro.

Da n variabili iniziali X_i che hanno un certo legame fra loro (es media voto, n° ore studiate, n°esami per sessione) arriviamo a m (con $m < n$) nuove variabili Y_i (detti **fattori latenti**), non correlate fra di loro, che sintetizzano in buona parte le informazioni contenute nelle n variabili iniziali.

I concetti base che occorre tener presenti nell'analisi fattoriale sono:

- per valutare la parte comune di due variabili (vedremo meglio tra poco) si può usare la misura della correlazione lineare
- per ottenere da più variabili un fattore, si può usare una loro combinazione lineare

- per conoscere quanto un fattore sia esplicativo del comportamento delle sue variabili, si può valutare il legame tra questo e le n variabili originarie, mediante al correlazione multipla)

L'analisi fattoriale si svolge sostanzialmente attraverso tre fasi:

1. determinazione della matrice di correlazione
2. estrazione dei fattori iniziali (per ridurre il numero di variabili esplicative)
3. rotazione degli assi (per migliorare la capacità esplicativa delle nuove variabili)

Se ipotizziamo una relazione lineare fra le variabili osservate e i fattori sintetici, si può definire un sistema di equazioni del tipo:

$$Xi = k_{i1} \cdot F_1 + k_{i2} \cdot F_2 + \ldots + k_{im} \cdot F_m \tag{27.1}$$

dove i è l'indicatore della variabile iniziale, $j = 1, \ldots, m$ è l'indicatore dei fattori considerati e i k sono i pesi fattoriali che visualizzano la correlazione tra gli F e le X variabili iniziali.
Ciascuna variabile è descritta come una combinazione di nuove variabili non correlate fra loro. Abbiamo visualizzato la singola variabile come una combinazione lineare di diversi fattori latenti: la singola variabile è influenzata da due tipi di fattori:

- fattori che influenzano tutte le variabili (fattori di *comunalità*):es. F_1 può influenzare fortemente sia X_1 che X_2
- fattori specifici di questa variabile (fattori di *unicità*): es. F_2 può influenzare fortemente X_3 e non altre variabili

Qui esponiamo il metodo delle **componenti principali**: è un metodo di trasformazione matematica di un insieme di variabili in un nuovo insieme di variabili, che spiegano la maggior parte della variabilità dei dati. All'interno della famiglia dell'analisi fattoriale però, è bene ricordarlo, esistono numerose tecniche. Il metodo delle componenti principali è solo una.

27.1.1 Analisi dei fattori in ambito di rilevazione dati

Non tutta la scienza è test di ipotesi; a volte siamo interessati alla struttura di un particolare fenomeno (es psicologico, legato alla qualità della vita) e vogliamo quantificare quel fenomeno. Il nostro obiettivo è sviluppare uno strumento di raccolta dati che misuri adeguatamente e rifletta la struttura del costrutto di nostro interesse (ad esempio "paura connessa la sottoporsi ad un test")

Nel prosieguo ci porremo nell'ottica di aver costruito un possibile questionario per la rilevazione di un particolare costrutto e di analizzarne le proprietà. Nello specifico l'esempio è basato su un questionario che desidera quantificare la paura connessa al sottoporsi ad un test genetico per la determinazione se si ha o meno un fattore di rischio per lo sviluppo di tumori. A 205 individui sono stati presentati un set di 20 domande (appendice 1 pag 242) e sono state registrate anche altre alcune variabili socio demografiche (age, race, marital, educlevl,insured).

```
library(lbdatasets)
ms <- msofa
dim(ms)

## [1] 205  25

head(ms, n = 5)

##   c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 age race marital ed
## 1  4  4  4  4  4  2  4  5  1   3   2   1   2   1   4   5   2   2   1   3  21    1       1
## 2  2  3  5  3  3  2  3  4  4   2   4   1   4   4   3   2   4   3   3   2  22    1       1
## 3  3  5  5  4  4  4  5  5  2   4   1   1   5   1   3   5   2   2   2   5  23    1       1
## 4  3  5  4  5  4  4  4  5  3   4   2   4   3   4   4   1   3   5   5   4  24    1       1
## 5  2  3  5  3  4  4  2  5  1   3   2   3   4   2   3   4   4   3   2   4  25    1       1
##   insured
## 1       1
## 2       1
## 3       1
## 4       1
## 5       1
```

Gli item sottoposti ai pazienti sono le variabili C1-C20, e sono rilevati su una scala di Likert che va da 1 (assolutamente non preoccupato di quell'aspetto) a 5 (molto preoccupato di quell'aspetto).

27.2 Analisi della correlazione tra le variabili

A questo punto una volta che abbiamo i dati, per procedere nell'analisi fattoriale bisogna costruire la matrice di correlazione tra le variabili: ossia la tabella che indica la correlazione di ogni variabile con l'altra. Questa ci serve per avere una idea di quali siano le variabili iniziali che vengono influenzate da fattori comuni, e quindi possano esser espresse congiuntamente da un variabile latente sintetica.

Rispetto al nostro esempio analizziamo solo la matrice di correlazione delle prime 8 variabili. La forma più comune di correlazione utilizzata in analisi fattoriale è quella di Pearson, impiegata nel seguito

```
small.ms <- ms[1:8]
cor.mat <- cor(small.ms)
round(cor.mat, digits=2)

##       c1   c2   c3   c4    c5   c6   c7   c8
## c1  1.00 0.06 0.34 0.26 -0.04 0.17 0.45 0.11
## c2  0.06 1.00 0.25 0.56  0.39 0.50 0.19 0.28
## c3  0.34 0.25 1.00 0.24  0.09 0.38 0.61 0.21
## c4  0.26 0.56 0.24 1.00  0.38 0.39 0.35 0.25
## c5 -0.04 0.39 0.09 0.38  1.00 0.47 0.06 0.33
## c6  0.17 0.50 0.38 0.39  0.47 1.00 0.35 0.19
## c7  0.45 0.19 0.61 0.35  0.06 0.35 1.00 0.07
## c8  0.11 0.28 0.21 0.25  0.33 0.19 0.07 1.00
```

ρ_{xy}	R^2	Associazione
0-.29	0-.08	Debole
.3-.49	.09-.24	Bassa
.5-.69	.25-.48	Moderata
.7-.89	.49-.8	Forte
.9-1	.81-1	Molto forte

Tabella 27.1: Regole del pollice per valutare la forza dell'associazione

(alcune differenze di arrotondamento dovute ad R che si conforma ad IEC 60559).
La correlazione di Pearson ρ_{xy} assume valori tra -1 e 1, con valori piu grandi in valure assoluta indicanti una più forte leazione; un valore positiva indica una relazione diretta. Nella matrice non vediamo numerose correlazioni negative, ovvero chi tende ad avere maggiore paura in un item, la avrà anche negli altri. Una guida per l'interpretazione della forza dell'associazione (lineare) descritta da coefficienti di correlazione è riportato in tabella 27.1.
L'ipotesi è che su variabili molto correlate fra loro agiscano i medesimi fattori latenti che siamo interessati ad estrarre. I test sulle correlazioni, portano ai seguenti p-value:

```
library(lbstat)
cor.tests <- cor.test_p(small.ms, alternative = "greater")
round(cor.tests, digits = 3)
```

```
##       c1    c2    c3 c4    c5    c6    c7    c8
## c1 0.000 0.177 0.000  0 0.696 0.008 0.000 0.066
## c2 0.177 0.000 0.000  0 0.000 0.000 0.003 0.000
## c3 0.000 0.000 0.000  0 0.101 0.000 0.000 0.001
## c4 0.000 0.000 0.000  0 0.000 0.000 0.000 0.000
## c5 0.696 0.000 0.101  0 0.000 0.000 0.181 0.000
## c6 0.008 0.000 0.000  0 0.000 0.000 0.000 0.003
## c7 0.000 0.003 0.000  0 0.181 0.000 0.000 0.170
## c8 0.066 0.000 0.001  0 0.000 0.003 0.170 0.000
```

I test di correlazione aiutano ad individuare le associazioni presenti e quelle che potrebbero esser dovute al caso. Ad esempio la correlazione tra C_1 e C_2 non si riesce a rifiutare l'ipotesi nulla che la correlazione tra i due elementi sia nulla.

La matrice di correlazione e la tabella dei p-value ad essa associata ci fornisce indizi su quali item potrebbero raggrupparsi. La correlazione significativa tra gli item C1, C3 e C7 suggerisce che questi item potrebbero combinarsi in una sorta di sottoscala. La non significativa correlazione di C5 e C8 con i precedenti suggerisce che probabilmente questi elementi non finiranno nella sottoscala.

Prima di procedere con l'estrazione dei fattori bisogna:

- controllare il determinante della matrice di correlazione per evitare situazioni limite
- testare che la matrice di correlazione sia differente da una matrice di identità (che consiste in una matrice completamente nulla, ad eccezione degli elementi sulla diagonale principale, uguali ad 1). In tal caso infatti non si

riuscirebbe ad individuare fattori comuni alle variabili presentate, poichè queste sono tutte incorrelate tra loro.

27.2.1 Controllo dei determinanti

Il determinante di una matrice è un numero unico associato ad una data matrice quadrata; il determinante è critico per la risoluzione di sistemi di equazioni in quanto "determina" se una data matrice quadrata ha una inversa.
Non tutte le matrici quadrate hanno una inversa: per determinarlo è necesario calcolare il determinante. Se questo è uguale a 0 allora non vi sono inverse e ulteriori elaborazioni della matrice sono sconsigliabili.
In generale il valore di un determinante può variare tra $-\infty, \infty$; tuttavia valori di determinanti di matrici di correlazione variano solamente tra 0 e 1:

- quando tutti gli elementi al di fuori della diagonale sono 0, il determinante della matrice sarà 1. Questo significa che la matrice di correlazioneè una matrice di identità e in tal caso sarebbe poco saggio proseguire con l'analisi dei fattori poichè si estrarrebbero tanti fattori quante sono le variabili di origine
- se il determinante è 0, si dice che la matrice è singolare e ciò significa che vi è almeno una dipendenza lineare nella matrice, ovvero che una o più colonne (o righe) della matrice può essere ottenuta dalla trasformazione lineare di altre colonne (righe)
- se molto prossimo a 0 significa non vi è una dipendenza perfetta, ma ci si trova in una situazione di alta correlazione (*ill conditioned matrix*. In tal caso la matrice dovrebbe condurre a stime instabili

Il determinante della matrice di correlazione è

```
det(cor.mat)

## [1] 0.1046724
```

Dato che nel nostro caso il determinante non è uguale a 0, la matrice di correlazione non è singolare.
Se effettuassimo lo stesso test sulla matrice completa di dati

```
full.ms <- ms[1:20]
det(cor(full.ms))

## [1] 1.995333e-05
```

prossimo a 0 e ciò suggerisce che vi siano diversi item con forte correlazione tra loro.
In presenza di una matrice singolare o ill-conditioned (det prossimo o uguale a 0), i seguenti passi sono consigliabili:

- controlla le correlazioni fra item: se vi sono diversi item che hanno correlazioni superiori a .8 esamina l'item piu approfonditamente per l'utilità clinica ed elimina uno o più dall'analisi (ovvero riduci il numero di domande fatte)

- esamina il dataset per eventuali duplicazioni delle righe (pazienti); se due pazienti differenti hanno dato le stesse risposte, forse bisognerebbe eliminarne uno
- controlla che vi siano abbastanza soggetti per item: troppo pochi soggetti per item (inferiore ai 10-15) possono portare determinanti bassi

27.2.2 Check sulla matrice di correlazione

Durante questa fase esplorativa, è importante determinare se vi è un sufficiente numero di correlazioni significative tra gli item per giustificare l'analisi fattoriale. Infatti se le correlazioni tra item non sono significative, non sarà possibile ottenere un set parsimonioso di fattori che rappresentino i numerosi item della scala proposta, ma vi potrebbero esser tanti fattori quanti item.

Vi sono diversi test per farlo:

- test di sfericità di Bartlett
- Test di Kaiser-Meyer-Olkin
- Individual Measures of Sampling Adequacy (MSA)

27.2.2.1 Test di Bartlett

Il test si sfericità di bartlett testa l'ipotesi nulla che la matrice di correlazione sia una matrice identità (ipotesi nulla). Valori piu alti del test indicano grande verosimiglianza che la matrice di correlazione non sia identità (rifiuto della nulla).

Il test di bartlett è un test chi-quadrato che assume la seguente forma:

$$\chi^2 = -\log(\det R) \cdot \left[(N-1) - \left(\frac{2k+5}{6}\right)\right] \sim \chi^2_{\frac{k(k-1)}{2}} \tag{27.2}$$

dove:

- N è il sample size
- k è il numero di variabile nella matrice
- $\det R$ è il determinante della matrice di correlazione
- il numero di gradi di libertà rappresenta il numero di correlazioni sopra o sotto la diagonale principale

Il test calcolato ad 8 variabili è

```
n <- nrow(small.ms)
k <- ncol(small.ms)
## Statistica test
(chi2 <- -log(det(cor.mat))* (n-1-(2*k+5)/6))

## [1] 452.5124
```

```
## Gradi di libert<U+00E0> della distribuzione teorica
dof <- k*(k-1)/2
## Valore soglia
qchisq(p=.95, df=dof)

## [1] 41.33714

## p-value
pchisq(chi2, dof, lower.tail=F)

## [1] 1

# Now a function from
# http://minato.sip21c.org/swtips/factor-in-R.pdf let us calculate

# Bartlett test for full dataset
bartlett.sphericity.test <- function(x)
{
    # Source: http://minato.sip21c.org/swtips/factor-in-R.pdf
    method <- "Bartlett's test of sphericity"
    data.name <- deparse(substitute(x))
    x <- subset(x, complete.cases(x)) #Omit
    n <- nrow(x)
    p <- ncol(x)
    chisq <- (1-n+(2*p+5)/6)*log(det(cor(x)))
    df <- p*(p-1)/2
    p.value <- pchisq(chisq, df, lower.tail=FALSE)
    names(chisq) <- "X-squared"
    names(df) <- "df"
    return(structure(list(statistic=chisq,
                          parameter=df,
                          p.value=p.value,
                          method=method,
                          data.name=data.name),
                     class="htest"))
}

bartlett.sphericity.test(full.ms)

##
##  Bartlett's test of sphericity
##
## data:  full.ms
## X-squared = 2126.5, df = 190, p-value < 2.2e-16
```

Pertanto stando al test che rifiuta la nulla, possiamo effettuare la PCA.
In genere il test di bartlett è influenzato dal sample size (se N cresce, il test aumenta), pertanto dovrebbe esser utilizzato come *standard minimo* per valutare la qualità della matrice di correlazione.

Valore KMO	Interpretazione
$< .5$	inaccettabile
$[0.5, 0.6)$	miserable
$[0.6, 0.7)$	mediocre
$[0.7, 0.8)$	intermedia
$[0.8, 0.9)$	meritoria
$[0.9, 1.0)$	meravigliosa

Tabella 27.2: Soglie KMO

27.2.2.2 Kaiser-Meyer-Olkin

Un secondo indicatore di forza di relazione tra item è il *coefficiente di correlazione parziale*: esso rappresenta la correlazione tra ogni paio di item, dopo aver rimosso l'effetto di altre covariate. Ad esempio il coefficente di correlazione parziale tra X e Y dato un set di variabili di controllo Z_i può esser ottenuto come correlazione tra i residui provenienti dalle regressioni lineari $X = f(Z)$ e $Y = f(Z)$ rispettivamente.
Se gli item condividono fattori comuni, sarà ragionevole attendersi che il coefficiente di correlazione parziale sia basso quando l'effetto delle variabili per cui si controlla è stato rimosso (qui si vuole che le variabili non siano incorrelate!).

KMO compara la grandezza del coefficiente di correlazione con il coefficiente di correlazione parziale:

$$KMO = \frac{\sum_{i \neq j} \sum \rho_{ij}^2}{\sum_{i \neq j} \sum \rho_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2} \tag{27.3}$$

dove:

- $\sum \sum$ è la somma su tutti gli items della matrice tra loro differenti (non sulla diagonale principale)
- ρ_{ij}^2 coefficiente di correlazione di Pearson tra l'item i-esimo e il j-esimo
- a_{ij}^2 è il coefficiente di correlazione parziale tra i e j

Il range di valori del KMO è $[0, 1]$ con un valore desiderabile maggiore di 0.7; nello specifico si veda tabella 27.2:

Oltre al KMO una misura di adeguatezza può esser calcolata per ogni item, utilizzando semplicemente la correlazione totale (con cosa, con se stessa?) e la parziale (sempre con cosa): si giunge alla MSA per la singola variabile.

$$MSA_i = \frac{\sum_{i \neq j} \sum \rho_{ij}^2}{\sum_{i \neq j} \sum \rho_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2} \tag{27.4}$$

i criteri per giudicare i valori associati alla singola variabile sono gli stessi di KMO, tab 27.2; idealmente vorremmo avere KMO e MSA superiori a 0.7 (almeno 0.6).

Il calcolo è svolto come segue

```
kmo <- function(x)
{
    ## http://minato.sip21c.org/swtips/factor-in-R.pdf

    ## Omit missing values
    x <- subset(x, complete.cases(x))
    ## Correlation matrix
    r <- cor(x)
    ## Squared correlation coefficients
    r2 <- r^2
    ## Inverse matrix of correlation matrix
    i <- solve(r)
    ## Diagonal elements of inverse matrix
    d <- diag(i)
    ## Squared partial correlation coefficients
    p2 <- (-i/sqrt(outer(d, d)))^2
    ## Delete diagonal elements
    diag(r2) <- diag(p2) <- 0

    KMO <- sum(r2)/(sum(r2)+sum(p2))
    MSA <- colSums(r2)/(colSums(r2)+colSums(p2))
    return(list(KMO=KMO, MSA=MSA))
}

kmo(small.ms)

## $KMO
## [1] 0.7174703
##
## $MSA
##        c1        c2        c3        c4        c5        c6        c7        c8
## 0.7374707 0.7290574 0.6945459 0.7334237 0.7011782 0.7646038 0.6706227 0.7101830

kmo(full.ms)

## $KMO
## [1] 0.762799
##
## $MSA
##        c1        c2        c3        c4        c5        c6        c7        c8        c9
## 0.8217862 0.5838181 0.8231222 0.7761811 0.6613024 0.7793981 0.8126636 0.6600160 0.4432129 0.
##       c11       c12       c13       c14       c15       c16       c17       c18       c19
## 0.7031584 0.8328215 0.8508683 0.6944658 0.7521610 0.8122778 0.8662241 0.7374312 0.7380344 0.
```

Cosa succe de se diversi coefficienti MSA sono bassi e le corrispondenti correlazioni parziali alte? se l'MSA di alcuni item è mediocre (<.6) ci potrebbero non esser fattori sottostanti che possano riassumere la relazione tra item. Per risovere, il ricercatore dovrebbe identificare gli item con la più bassa MSA, rimuoverli dalla lista analizzata e rifar girare il test della matrice di correlazione. Il processo dovrebbe esser continuato fino a che gli MSA individuali sono nei

range accettabili; se questo non funziona il ricercatore dovrebbe incrementere il campione o riconsiderare la fattibilità di un'analisi fattoriale.

In generale, la fase precedente all'estrazione dei fattori è **iterativa nell'applciazione dei test spiegati** in precedenza: l'algoritmo da seguire prima di arrivare all'estrazione dei fattori è riportato nel libro a pag 83.

Per la scelta dei fattori da eventualmente eliminare si può controllare la matrice di correlazione più in dettaglio per identificare elementi che sono troppo correlati tra loro ($\rho > .8$) o poco ($< .3$). Se gli item sono troppo correlati si potrebbero aver problemi di multicollenarità e la necessita di eliminare uno o più degli elementi molto correlati fra loro. Se al contrario gli item non sono correlati a sufficienza non vi sarà molta varianza comune che possa esser sintetizzata da un unico fattore, portando alla moltiplicazione di questi e a una minore capacità di sintesi. Anche in questo caso eliminare partendo dai valori più bassi di correlazione.

27.3 Estrazione dei fattori

Arrivati a questo punto siamo interessati a indentificare i fattori e le loro relazioni con le variabili di parte. Il processo generalmente si svolge in due fasi:

1. estrazione iniziale dei fattori
2. scelta del numero di fattori da estrarre effettivamente
3. rotazione dei fattori per migliorare l'interpretazione

Nel seguito ci concentriamo sulle prime due fasi; nella prossima sezione sull'ultima.

Il nostro obiettivo è condensare la varianza comune tra gli item per determinare fattori che rappresentino sufficientemente le variabili di partenza.
Le fonti di varianza degli score possono esser pensate come composte da tre componenti:

la varianza comune indicata con h^2 rappresenta la variabilità condivisa tra un set di item, la quale può esser spiegata da un set di fattori comuni

varianza specifica è una componente della varianza specifica di una particolare variabile, che non è condivisa con altri item posti nell'analisi

varianza dovuta all'errore della misurazione può esser valutata esaminando la consistenza interna degli item o *reliability* utilizzando il coefficiente α di Cronbach. Più è affidabile un set di item, minore è il loro errore di misurazione

La varianza comune, è il focus dell'analisi dei fattori. Il processo di estrazone *inizia fornendo una stima iniziale della variabilità dei singoli item spiegata dai fattori che siamo interessati ad estrarre*, detta **comunalità**. Questa stima della comunalità può variare tra 0 e 1 con valori più alti indicanti che i fattori estratti spiegano piu variabilità degli item di partenza. Vi sono due approcci

- **principal component analysis**: soluzione più semplice spiega tutta la variabilità dei fattori di partenza (quindi comunalità=1), mediante l'identificazione di *fattori o componenti*. Ognuno dei fattori è una combinazione lineare degli item inclusi nell'analisi e i fattori estratti sono ortogonali fra loro (ovvero non correlato). Critica dell'approccio è che determinando fattori latenti che spiegano tutta la variabilità ci si concentra anche su componenti della varianza che non sono strettamente quella comune a piu item.

- **common factor analysis** si focalizza solamente sulla varianza comune che gli item condividono, spiegandola da un numero di fattori sottostanti. I fattori estratti non sono combinazioni lineari degli item esaminati ma sono piu fattori ipotetici stimati dagli item esamitnati (comunalità<1). Sulla scelta della comunalità da introdurre, spesso si ricorre all'R^2 della regressione che vede l'item analizzato come variabile dipendente e tutte le altre variabili come indipendenti. Questo approccio è quello del *principal axis factoring* (PAF), quello di cui parleremo tra i metodi della common factor analysis

Per un'analisi fattoriale esplorativa gli autori consigliano di partire da una PCA e confrontare i risultati con una PAF, scegliendo quella che fitta meglio e che è piu intuitiva.

Analizzaimo in maggior dettaglio le due tecniche.

27.3.1 Analisi delle componenti principali

L'analisi fu sviluppada da Persona con l'obiettivo di riassumere le relazioni tra un set di variabili originali in termini di un numero minore di componenti principali tra loro incorrelate, che sono una combinazione lineare delle variabili originali.

Dato che il metodo è fortemente dipendete dalla quantità di variabilità totale, la PCA generalmente richiede che le variaibli esaminate siano sotto unità di misura simili. Pertanto è necessario standardizzare le variabili cosicche le medie siano nulle e le varianze unitarie.

Poichè la variabile standardizzata ha 1 di varianza, la **varianza totale da spiegare** è uguale a 1·numero variabili. La PCA assume che vi è tanta variabilità quante sono le variabili in analisi e che tuta la varianza di un singolo item piò esser spiegata dai fattori estratti.

Si introducono i concetti fondanti

- un *autovalore* λ (*eigenvalue*) rappresenta la varianza di tutti gli elementi complessivamente che può essere spiegata da un singolo fattore estratto considerato. In PCA il valore massimo che un autovalore può prendere è pari alla varianza totale da spiegare, come determinata in precedenza[1].

- un *autovettore* è un vettore di coefficienti, detti *pesi fattoriali* che costituiscono la correlazione tra un singolo fattore e un singolo item (si avrà un peso per ogni item). Il numero di autovettori equivale al numero di fattori estratti

[1] In teoria gli autovalori possono variare tra $-\infty, +\infty$, ma nella PCA sono generalmente >0. Valori negativi implicano il fatto che che generalmente >0la varianza spiegata negli items sia negativa, e questo avviene quando vi è forte multicollinearità

Ipotizzando una situazione in cui vi siano 3 item di partenza, per l'estrazione dei fattori si configura un sistema di equazioni del tipo:

$$\begin{cases} z_{C_1} = a_{C_1,I} \cdot PC_I + a_{C_1,II} \cdot PC_{II} + a_{C_1,III} \cdot PC_{III} \\ z_{C_2} = a_{C_2,I} \cdot PC_I + a_{C_2,II} \cdot PC_{II} + a_{C_2,III} \cdot PC_{III} \\ z_{C_3} = a_{C_3,I} \cdot PC_I + a_{C_3,II} \cdot PC_{II} + a_{C_3,III} \cdot PC_{III} \end{cases} \tag{27.5}$$

dove:

- z_{C_1} corrisponde all'item C_1 standardizzato
- PC_I, PC_{II} e PC_{III} sono le componenti principali estraibili
- i pesi fattoriali $a_{C_1,I}, a_{C_2,I}, a_{C_3,I}$ (prima colonna di pesi fattoriali) costituiscono l'autovettore del primo componente.
- l'autovalore λ_I associato a PC_I è ottenibile da:

$$\lambda_I = a^2_{C_1,I} + a^2_{C_2,I} + a^2_{C_3,I} \tag{27.6}$$

- $h^2_{C_3}$ ovvero ad esempiola proporzione di variabilità di C_3 spiegata dalle tre componenti si ricava come segue:

$$h^2_{C_3} = a^2_{C_3,I} + a^2_{C_3,II} + a^2_{C_3,III} \tag{27.7}$$

Si può dimostrare che la correlazione tra due item nella matrice di correlazione originale (es C1 e C3) è la somma dei prodotti dei pesi fattoriali:

$$\rho_{C_1,C_3} = a_{C_1,I} \cdot a_{C_3,I} + a_{C_1,II} \cdot a_{C_3,II} + a_{C_1,III} \cdot a_{C_3,III} \tag{27.8}$$

Il processo è iterativo e disegnato in maniera tale che la prima componente in PCA è una combinazione lineare delle variabili originali che spiega il massimo ammontare di variabilità tra le variabili originali, e via via le seguenti spiegano minor variabilità.
La prima componente principale è ottenuta dalla matrice di correlazione originaria; dalla seconda in poi si lavora su una matrice residuale ottenuta una volta che l'influenza di PC_1 è stata rimossa (e in tal modo è incorrelata alla prima). La seconda componente è quella che estrae in ordine la seconda maggiore variaiblità.

Nella PCA vi possono essere tante componenti principali quante sono le variabili originali, e in tal caso la variabilità spiegata è uguale alla somma delle varianze delle variabili originarie.
Quando il numero di variabili è largo, tuttavia, la maggior parte di variabilità si racchiude in un numero minore di comonenti; cio significa che gli autovalori delle prime componenti sara largo e quello delle ultime decisamente piccolo.

```
library(psych)

##
## Attaching package:  'psych'
## The following object is masked from 'package:lbmisc':
##
##     table2df
```

```
(pca.unrot.fit <- principal(small.ms,
                           nfactors = ncol(small.ms),
                           rotate = "none"))

## Principal Components Analysis
## Call: principal(r = small.ms, nfactors = ncol(small.ms), rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##     PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8 h2       u2 com
## c1 0.43  0.60  0.21  0.43  0.37  0.27  0.13  0.02  1 -8.9e-16 4.5
## c2 0.69 -0.37 -0.20  0.16 -0.38  0.27  0.18  0.24  1 -8.9e-16 3.5
## c3 0.64  0.48  0.07 -0.42 -0.22 -0.03  0.29 -0.23  1 -8.9e-16 3.8
## c4 0.72 -0.15 -0.11  0.50 -0.15 -0.30 -0.10 -0.26  1 -4.4e-16 2.9
## c5 0.55 -0.58  0.00 -0.14  0.46 -0.26  0.24  0.08  1 -1.8e-15 3.9
## c6 0.74 -0.16 -0.29 -0.28  0.22  0.33 -0.30 -0.13  1 -2.2e-16 3.0
## c7 0.63  0.59 -0.11 -0.11 -0.02 -0.30 -0.20  0.31  1  2.2e-16 3.3
## c8 0.45 -0.29  0.82 -0.08 -0.12  0.04 -0.16  0.03  1 -2.2e-16 2.0
##
##                        PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8
## SS loadings           3.05 1.54 0.87 0.76 0.63 0.50 0.35 0.31
## Proportion Var        0.38 0.19 0.11 0.10 0.08 0.06 0.04 0.04
## Cumulative Var        0.38 0.57 0.68 0.78 0.86 0.92 0.96 1.00
## Proportion Explained  0.38 0.19 0.11 0.10 0.08 0.06 0.04 0.04
## Cumulative Proportion 0.38 0.57 0.68 0.78 0.86 0.92 0.96 1.00
##
## Mean item complexity =  3.4
## Test of the hypothesis that 8 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

La matrice superiore è quella dei pesi fattoriali, quella inferiore da alcune misure di spiegazione della variabilità

27.3.2 Common factor analysis - Principal axis factoring

Tutte le tecniche di estrazione presuppongono che i fattori estratti inizialmente siano incorrelati fra loro. Per l'estrazione dei fattori a differenza della PCA, la CFA non ipotizza che i fattori siano una combinazione lineare perfetta degli item posti in analisi ma che siano invece costrutti ipotetici che vengano stimati dagli item. Dato ce questi fattori ipotetici sono generati dalla varianza comune (non dallatotale) la comunalit degli item (h^2) iniziale è meno di 1. Come risultato vi saranno un numero di fattori estratti inferiore agli item.

La difficoltà nella CFA è determinare quale valore di comunalità adottare per ogni item analizzato: diversi metodi sono stati sviluppati e i piu comuni si basano sull'R^2 di regressione di un item considerato (dipendente) rispetto a tuti gli altri (indipendenti).

In seguito alla determinazione della comunalità, l'analisi procede in maniera analoga alla PCA. L'esito saranno comunque fattori non correlati fra loro ma

che non riusciranno a spiegare il 100% della variabilità degli item di partenza (per costruzione). A parte questo i concetti introdotti per la PCA (autovalori, autovettori, pesi fattoriali) sono i medesimi anche per CFA/PAF.

```
(paf.unrot.fit <- fa(cor(small.ms),
                     nfactors = 7,
                     fm = "pa",
                     rotate="none"))

## Warning in sqrt(eigens$values[1:nfactors]):  NaNs produced
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate
= rotate, :  imaginary eigen value condition encountered in fa
##  Try again with SMC=FALSE
##  exiting fa
## Warning in cor.smooth(model):  I am sorry, there is something seriously
wrong with the correlation matrix,
## cor.smooth failed to  smooth it because some of the eigen values
are NA.
## Are you sure you specified the data correctly?
## Warning in max(R2, na.rm = TRUE): no non-missing arguments to max;
returning -Inf
## Warning in cov2cor(t(w) %*% r %*% w):  diag(V) had non-positive
or NA entries; the non-finite result may be dubious
## Error in l[id]:  subscript out of bounds
```

Il risultato non torna con quello di SPSS, forse dovuto all'alert che si stanno richiedendo troppi fattori (per cui si comporta ugualmente alla PCA, ponendo variabilità a 1, l'SMC è lo Squared multiple correlation ovvero R^2)

Sebbene il PAF sia il metodo piu comune in CFA ve ne sono anche altri: metodi di massima verosimiglianza, minimi quadrati non pesati e generalized least square

27.3.3 Quanti fattori estrarre

Vogliamo determinare il numero di fattori ottimale da estrarre. Più fattori teniamo piu spieghiamo variaiblità, ma chiaramente più si pone un problema di parsimonia.
Non vi è una regola unica a questo e le strade che si possono intraprendere sono molteplici:

- tenere i fattori che hanno $\lambda_i > 1$: si tratta di un criterio di economicità ce dice che un fattore è utile se spiega più di una quantità di variabilità pari a quella prodotta da una variabile (standardizzata). Nel caso di numerosi item posti in analisi, altrettanto numerosi fattori potrebbero rispettare questo criterio, quindi per sintesi si potrebbe voler impiegare altri criteri che più direttamente riguardano la quota di variaiblità progressivamente spiegata

- tenere tanti fattori quanti ne servono per spiegare una data percentuale di variabilità complessiva (si guardi la colonna della variabilità cumulata),

prefissata a priori (valori tipici sono 75-80%, sebbene non vi siano guidelines definitive per stabilire la soglia). La scelta è comunque discrezionale e va variata in relazione della tecnica di estrazione impiegata (PAF o PCA)

Nell'esempio gli autori scelgono di tenere 2 fattori e di confrontare le stime derivanti dai due metodi di estrazione

```
(pca.unrot.fit2 <- principal(small.ms, nfactors = 2, rotate = "none"))

## Principal Components Analysis
## Call: principal(r = small.ms, nfactors = 2, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##     PC1   PC2   h2   u2 com
## c1 0.43  0.60 0.54 0.46 1.8
## c2 0.69 -0.37 0.62 0.38 1.5
## c3 0.64  0.48 0.63 0.37 1.9
## c4 0.72 -0.15 0.55 0.45 1.1
## c5 0.55 -0.58 0.64 0.36 2.0
## c6 0.74 -0.16 0.57 0.43 1.1
## c7 0.63  0.59 0.75 0.25 2.0
## c8 0.45 -0.29 0.29 0.71 1.7
##
##                        PC1  PC2
## SS loadings           3.05 1.54
## Proportion Var        0.38 0.19
## Cumulative Var        0.38 0.57
## Proportion Explained  0.66 0.34
## Cumulative Proportion 0.66 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.1
##  with the empirical chi square  125  with prob <  2.1e-20
##
## Fit based upon off diagonal values = 0.9

(paf.unrot.fit2 <- fa(small.ms, nfactors = 2, fm = "pa", rotate = "none"))

## Factor Analysis using method =  pa
## Call: fa(r = small.ms, nfactors = 2, rotate = "none", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##     PA1   PA2   h2   u2 com
## c1 0.37 -0.37 0.27 0.73 2.0
## c2 0.63  0.35 0.53 0.47 1.6
## c3 0.59 -0.37 0.48 0.52 1.7
## c4 0.64  0.16 0.44 0.56 1.1
## c5 0.48  0.48 0.46 0.54 2.0
## c6 0.67  0.16 0.48 0.52 1.1
## c7 0.66 -0.58 0.77 0.23 2.0
## c8 0.35  0.19 0.16 0.84 1.5
```

```
##
##                        PA1  PA2
## SS loadings           2.54 1.05
## Proportion Var        0.32 0.13
## Cumulative Var        0.32 0.45
## Proportion Explained  0.71 0.29
## Cumulative Proportion 0.71 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  28  with the objective function =  2.26 with Chi Square =  452.51
## df of  the model are 13  and the objective function was  0.24
##
## The root mean square of the residuals (RMSR) is  0.05
## The df corrected root mean square of the residuals is  0.07
##
## The harmonic n.obs is  205 with the empirical chi square  27.65  with prob <  0.01
## The total n.obs was  205  with Likelihood Chi Square =  47.42  with prob <  8.2e-0
##
## Tucker Lewis Index of factoring reliability =  0.824
## RMSEA index =  0.114  and the 90 % confidence intervals are  0.08 0.15
## BIC =  -21.78
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                PA1  PA2
## Correlation of (regression) scores with factors   0.92 0.85
## Multiple R square of scores with factors          0.85 0.73
## Minimum correlation of possible factor scores     0.69 0.45
```

27.4 Rotazione dei fattori

Avendo ottenuto la soluzione non ruotata ai fattori si pone il problema di interpretare i cluster di item attorno ad uno o di un'altro fattore. Non sempre le soluzioni non ruotate fornicono interpretazioni facili dei cluster di item necessarie per identificare il significato dei fattori estratti.
Geometricamente se pensiamo a due fattori estratti potremmo plottare le righe associate ad un determinato item (i pesi fattoriali dei fattori estratti) in un piano a due dimensioni. Una situazione ottimale per l'interpretazione è quella in cui ogni coppia ha un valore dei due è alto e l'altro è basso, ovverosia prossimo allo 0. In tale situazione tutti i punti si concentrano nelle vicinanze di uno dei due assi che rappresentano i fattori.
Ma se cio non avviene è possibile far ruotare gli assi affinchè i pesi fattoriali si agglomerino meglio ad essi.
Per terminare l'interpretazione geometrica se si pensa a fattori scalati con media 0, la correlazione di due punti può essere vista come il coseno dell'angolo che formato: se l'angolo è 0 (ovvero i due punti vanno nella stessa direzione sopra o

sotto la media) la correlazione 1, se di 90 gradi la correlazione è 0, a 180 gradi si ha una correlazione pari a -1, a 270 ancora a 0.

La rotazione dei fattori è il processo di ruotazione degli atti dei fattori attorno all'origine al fine di ottenere una *struttura semplice* e dal punto di vista teorico una soluzione piu significativa.
La matrice su cui ci si focalizza è quella dei pesi fattoriali. Una struttura semplice si raggiunge se:

- ogni riga contiene almeno uno 0 (o valore prossimo)
- ogni colonna (autovettore) dovrebbe avere almeno tanti 0 quanti sono i fattori estratti
- altri criteri... vedi pag 132

Vi sono due classi generali di rotazioni: ortogonali e oblique. Entrambe condividono l'obiettivo di giungere ad una struttura semplice ma:

- in una rotazione ortogonale i fattori ruotati sono (continuano ad essere) indipendenti tra loro (incorrelati).
  Dato che in una rotazione orgogonale i fattori risultanti sono incorrelati, la matrice de pesi fattoriali non è solamente la semplice correlazione degli item con i fattori ma anche i simil beta di una regressione standardizzata che possono esser sati per stimare il contributo unico di ogni fattore alla varianza spiegata di un item.
- in una rotazione obliqua non necessariamente è imposta questa cosa, ma si accetta che tra fattori risultanti (e concetti sottostanti) vi possa essere correlazione. Una rotazione pressochè ortogonale potrà comunque esser raggiunta, nel caso.
  A differenza del caso precedente vi saranno due differenti matrici: la factor pattern matrix (matrice di pesi fattoriali che sono coefficienti di regressione standardizzati parziali) e una factor structure matrix (matrice di semplici correlazioni deigli item con i factor

Nella scelta tra rotazioni ortogonali e non, il ricercatore pensi al costrutto che deve esser misurato; se si crede che una certa correlazione esista (es tra i .2 e i .5 in valore assoluto) potrebbe esere sensato applicare rotazioni oblique. Se la correlazione ipotetica è inferiore (tra -.2 e +2) gli autori suggeriscono una trasformazione ortogoanale. Se superiore in valore assoluto a .5 vi è da pensare di eliminare qualce fattore. Ancora meglio applicare entrambe le tipologie e confrontare i risultati.

27.4.1 Rotazioni ortogonali

Come detto le rotazioni ortogonali mantengono la non correlazione dei fattori risultanti; è necessario imporre una rotazione che mantenga a 90° l'angolo tra i due assi. Di quanto poi questi debbano ruotare viene calcolato ottimizzando dalla procedura. Vi sono tre approcci maggiori alle rotazioni ortogonali: Varimax, Quartimax ed Equamax. Varimax è la procedura standard che trattiamo: essa mira a semplifiare le colonne delle soluzioni non ruotate massimizzando la variabilità dei pesi all'interno di un fattore, ovvero aumentando i pesi già alti e diminuendo quelli bassi.

Nel seguito estraiamo due fattori e ruotiamo la soluzione con la Varimax:

```
(pca.varimax.fit2 <- principal(small.ms, nfactors = 2, rotate = "varimax"))

## Principal Components Analysis
## Call: principal(r = small.ms, nfactors = 2, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC1   RC2   h2   u2 com
## c1 -0.02  0.74 0.54 0.46 1.0
## c2  0.78  0.12 0.62 0.38 1.1
## c3  0.22  0.77 0.63 0.37 1.2
## c4  0.67  0.32 0.55 0.45 1.4
## c5  0.79 -0.13 0.64 0.36 1.1
## c6  0.68  0.32 0.57 0.43 1.4
## c7  0.14  0.85 0.75 0.25 1.1
## c8  0.53  0.04 0.29 0.71 1.0
##
##                        RC1  RC2
## SS loadings           2.49 2.09
## Proportion Var        0.31 0.26
## Cumulative Var        0.31 0.57
## Proportion Explained  0.54 0.46
## Cumulative Proportion 0.54 1.00
##
## Mean item complexity =  1.1
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.1
##  with the empirical chi square  125  with prob <  2.1e-20
##
## Fit based upon off diagonal values = 0.9
```

Si nota la differenza rispetto a `pca.unrot.fit2`, ottenuta ugualmente ad eccezione della presenza di rotazione.

Nonostante il cambio nei pesi fattoriali dovuto alla rotazione, la proporzione di variabilità di un dato item spiegata (comunalità) rimane la stessa.

27.4.2 Rotazioni Oblique

Le rotazioni ortogonali si fondano sulla assunzione critica che i fattori siano incorrelati fra loro, assunzione raramente rispettata nelle scienze sociali dove i vari aspetti di una problematica sono speso interconnessi.

La differenza pratica in una rotazione obliqua è che non viene richiesto che un angolo di 90 gradi venga rispettato nella rotazione degli assi; al contrario ogni fattore originale è ruotato separatamente per un proprio ammontare.

Le tecniche di rotazione di questa famiglia più note sono Oblimin e Promax.

In generale nelle rotazioni oblique la somma dei quadrati dei pesi di ogni riga non necessariamnte risulterà h^2; e la somma dei quadrati dei coefficienti per colonna non eguagliera la varianza spiegata da unfattore, se non per caso e cosi via le correlazioni originali tra le variabili non possono esser riprodotto

come visto in precedenza. Vi è convergenza nel dire che la correlazione tra fattori non deve essere settata come troppo bassa o alta (tramite i parametri delle procedure)

27.4.2.1 Oblimin

Oblimin (a volte detta Direct Oblimin) cerca di soffisfare il procipio di struttura semplcie attraverso la specificazione di un parametro δutilizzato per controllare il grado di correlazione ammesso tra i fattori. δ può esser positivo o negativo:

- valori negativi sempre più grandi diminuiscono la correlazione tra fattori rendendoli più ortogonali (una situazione più o meno ortogonale la si ha quando $\delta = -4$
- valori positivi sempre più grandi aumenteranno la correlazione tra i fattori, ma valori più alti di .8 producono correlazioni molto alte che possono causare problemi alla risoluzione/convergenza (e non ha senso avere fattori molto correlati)
- se pari a 0 la rotazione viene detta Quartimin: in SPSS è il valore di default utilizzato

Non vi è una opinione condivisa in letteratura su quanto debba esser settato il valore assoluto di δ; gli autori suggeriscono che se il ricercatore ipotizza che la correlazione tra i fattori sia approssimativamente .30, valori di delta tra -.5 e +.5 generalmente ottengono questo risultato.

```
(pca.oblimin.fit2 <- principal(small.ms, nfactors = 2, rotate = "oblimin"))

## Loading required namespace:  GPArotation

## Principal Components Analysis
## Call: principal(r = small.ms, nfactors = 2, rotate = "oblimin")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      TC1   TC2   h2   u2 com
## c1 -0.12  0.76 0.54 0.46 1.1
## c2  0.78  0.04 0.62 0.38 1.0
## c3  0.11  0.76 0.63 0.37 1.0
## c4  0.64  0.25 0.55 0.45 1.3
## c5  0.82 -0.22 0.64 0.36 1.1
## c6  0.65  0.25 0.57 0.43 1.3
## c7  0.03  0.86 0.75 0.25 1.0
## c8  0.54 -0.02 0.29 0.71 1.0
##
##                        TC1  TC2
## SS loadings           2.48 2.11
## Proportion Var        0.31 0.26
## Cumulative Var        0.31 0.57
## Proportion Explained  0.54 0.46
## Cumulative Proportion 0.54 1.00
##
##  With component correlations of
```

```
##      TC1  TC2
## TC1 1.00 0.25
## TC2 0.25 1.00
##
## Mean item complexity =  1.1
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.1
##  with the empirical chi square  125  with prob <  2.1e-20
##
## Fit based upon off diagonal values = 0.9
```

27.4.2.2 Promax

E' una rotaizone obliqua che inizia con un ortogonale, solitamente varimax i pesi ortogonali sono inizialmente elevati ad una determinata potenza κ, con valore solitamente pari a 2,4 o 6 (SPSS default=3, mentre SAS=3), poi la soluzione è ruotata per permettere la correlazione tra i fattori.

Elevando i pesi si fa andare più velocemente giu i pesi passi di quanto non si faccia con i mesi alti. Potenze maggiori portano a maggiore corrleaizone tra i fattori; l'obiettivo è ottenere una soluzione che fornisce una buona struttura utilizzando una potenza per quanto possibile bassa.

```
(pca.promax.fit2 <- principal(small.ms, nfactors = 2, rotate = "promax"))
```

```
## Principal Components Analysis
## Call: principal(r = small.ms, nfactors = 2, rotate = "promax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC1   RC2   h2   u2 com
## c1 -0.16  0.77 0.54 0.46 1.1
## c2  0.79  0.00 0.62 0.38 1.0
## c3  0.08  0.76 0.63 0.37 1.0
## c4  0.64  0.22 0.55 0.45 1.2
## c5  0.84 -0.26 0.64 0.36 1.2
## c6  0.65  0.23 0.57 0.43 1.2
## c7 -0.01  0.87 0.75 0.25 1.0
## c8  0.55 -0.05 0.29 0.71 1.0
##
##                        RC1  RC2
## SS loadings           2.49 2.10
## Proportion Var        0.31 0.26
## Cumulative Var        0.31 0.57
## Proportion Explained  0.54 0.46
## Cumulative Proportion 0.54 1.00
##
##  With component correlations of
##      RC1  RC2
## RC1 1.00 0.33
## RC2 0.33 1.00
```

```
##
## Mean item complexity =  1.1
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.1
##  with the empirical chi square  125  with prob <  2.1e-20
##
## Fit based upon off diagonal values = 0.9
```

27.5 Evaluating and refining the factors

27.6 Interpretare i fattori e generare gli score

Parte IX

Missing data e multiple imputation

Capitolo 28

Introduzione ai dati mancanti

Qui ci si riferisce per lo più a VanBuuren (2018).

28.1 Identificazione della missingness

La funzione `md.pattern` che stampa le statistiche (e volendo il plot) dei dati mancanti nel dataset: nella tabella gli 1 significano dato presente mentre 0 mancante, ad inizio riga vengono fornite le frequenze del pattern, l'ultima colonna conta le variabili missing, l'ultima riga i totali di colonna (sotto ad una variabile il numero di missing per la stessa, nell'ultima colonna i missing complessivi del dataset. Righe e colonne sono ordinate per grado maggiore di missingness

```
library(mice)
mice::md.pattern(airquality, plot = FALSE)

##     Wind Temp Month Day Solar.R Ozone
## 111    1    1     1   1       1     1  0
## 35     1    1     1   1       1     0  1
## 5      1    1     1   1       0     1  1
## 2      1    1     1   1       0     0  2
##        0    0     0   0       7    37 44

## per riprodurre alcune statistiche
colSums(is.na(airquality))

##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0

sum(is.na(airquality))

## [1] 44

data <- airquality[, c("Ozone", "Solar.R")]
```

Pertanto ci sono complessivamente 44 dati mancanti nel dataset, le variabili interessate sono `Solar.R` (7 casi missing) e `Ozone` (37) e vi sono 35 righe con

`Ozone` mancante, 5 con `Solar.R` mancante e 2 con entrambi mancanti.
Se vogliamo estrarre le righe con dati incompleti usiamo `mice::ic`, per le righe tutte complete `mice::cc`.

28.2 Tipi di missingness

Per Rubin (1976) ogni dato (variabile su un pz) ha una data probabilità di essere missing; il processo che genera i dati mancanti può essere:

MCAR la probabilità di essere missing è la medesima per tutti i casi, i dati sono detti *missing completely at random*, ossia la causa della missingness non è legata al valore assunto. Questa ipotesi è spesso poco realistica.

MAR la probabilità di essere missing è la stessa all'interno di gruppi definiti da dati effettivamente osservati, i dati sono detti *missing at random*. È un'assunzione più realistica e i metodi statistici attuali di solito partono da questa

MNAR ossia *missing not at random*, è la condizione residuale rispetto alle precedenti, ossia la probabilità di essere missing dipende da ragioni (variabili) che ci sono sconosciute; è il caso più complesso e le strategie per gestirla sono cercare le cause della missingness o effettuare analisi di sensibilità/scenario what-if

28.3 Soluzioni ad-hoc

28.3.1 Listwise deletion

È l'approccio classicamente impiegato e consiste nel tenere i casi che hanno le variabili necessarie per l'analisi tutte complete.
Se siamo sotto MCAR, a parte la perdita di potenza (ed intervalli di confidenza più larghi) possiamo ignorare problemi di bias (inesistenti) nelle stime e procedere con la complete case analysis.
Se però i dati non sono MCAR a fare la complete case analysis rischiamo di biasare di brutto le stime (in ragione sia della distanza tra dati osservati e dati missing che della percentuale di dati missing).
Non vi sono proposte della percentuale di missingness al di sotto della quale si può ignorare il problema e andare di listwise senza pensarci.
In R si ottiene usando `na.omit` e `complete.cases`.

28.3.2 Pairwise deletion

Si cerca di utilizzare quello che c'è: ad esempio su una analisi bivariata le medie sono prese da ciascuna variabile per quello che ha e la varianza/covarianza considerando le coppie completamente disponibili.

28.3.3 Mean imputation

Si rimpiazzano i dati mancanti con la media dei dati presenti (o es la moda per dati categorici). Conduce a risultati non biasati solo sotto MCAR (ma che

rispetto a non fare nulla non perdono potenza).
Con mice si usa la funzione `mice` per imputare un dataset e crearne uno nuovo senza dati mancanti. Qui imputiamo il dataset `airquality`.

```
imp <- mice::mice(airquality, method = "mean", m = 1, maxit = 1, seed = 1456)

##
##  iter imp variable
##   1   1  Ozone  Solar.R
```

Nella chiamata abbiamo specificato:

- `method`, ossia il metodo di imputazione mediante
- un `seed` per riproducibilità
- mediante `m` di fornirci 1 dataset imputato
- `maxit` il numero massimo di iterazioni ad 1 (nessuna iterazione)

Qui potevamo lasciare `m` e `maxit` ai loro default con poche variazioni. La funzione ci comunica di aver effettuato imputazioni nelle variabili. L'oggetto restituito ha classe `mids` (per multiple imputed dataset).

```
class(imp)

## [1] "mids"

names(imp)

##  [1] "data"            "imp"             "m"               "where"           "blocks"
##  [6] "call"            "nmis"            "method"          "predictorMatrix" "visitSequence"
## [11] "formulas"        "post"            "blots"           "ignore"          "seed"
## [16] "iteration"       "lastSeedValue"   "chainMean"       "chainVar"        "loggedEvents"
## [21] "version"         "date"

## per estrarre il dataset completo da un oggetto mids
head(complete(imp, action = 1L))

##      Ozone  Solar.R Wind Temp Month Day
## 1 41.00000 190.0000  7.4   67     5   1
## 2 36.00000 118.0000  8.0   72     5   2
## 3 12.00000 149.0000 12.6   74     5   3
## 4 18.00000 313.0000 11.5   62     5   4
## 5 42.12931 185.9315 14.3   56     5   5
## 6 28.00000 185.9315 14.9   66     5   6

## action <U+00E8> tra 1 e imp$m per scegliere quale dataset imputato ritornare
```

28.3.4 Regression imputation

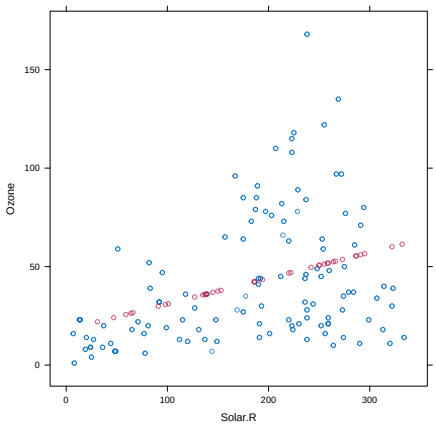
Si imputa sfruttando la relazione stimata tra una variabile ed un'altra (sui dati disponibili), ossia imputando con i valori predetti dalla regressione se il dato

originale è missing.
Si giunge a risultati non biasati se siamo sotto MCAR, se i beta non hanno bias (es no omitted variable bias); ma rispetto a mean imputation le stime non hanno bias se siamo sotto MAR (e abbiamo azzeccato il modello).
Tuttavia le correlazioni sono biasate verso l'alto nel dataset imputato, mentre la variabilità è sottostimata; di quanto sottostimata dipende dalla quantità di varianza spiegata (R^2) e dalla proporzione di missingness.

Chiaramente l'imputazione con questo metodo può funzionare bene (sotto MAR) se le predizioni sono vicine alla perfezione, ma di base le imputazioni basate su regression (cosiccome le incarnazioni moderne di machine learning) sono *probabilmente il metodo più pericoloso* tra quelli proposti. Potremmo pensare di "conservare" la relazione tra variabili ma quello che si va a fare è rafforzare artificialmente le relazioni nei dati, le correlazioni, a sottostimare le variabilità (risultati troppo belli per esser veri) e a condurci volentieri in falsi positivi e in relazioni spurie.

Per ottenerla in `mice` si usa il `method norm.predict`

```
imp <- mice(data, method = "norm.predict", seed = 1, m = 1, print = FALSE)
xyplot(imp, Ozone ~ Solar.R)
```

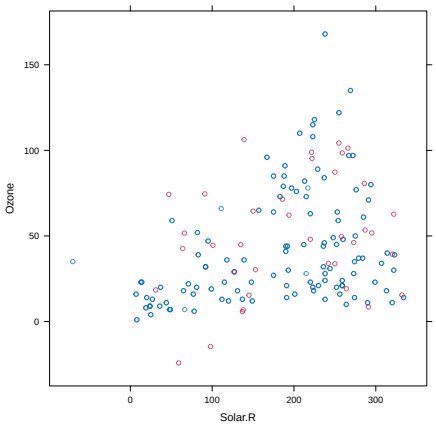


In questo caso notiamo che nel grafico i valori reali sono plottati in blu, mentre i predetti/imputati in rosso (e stanno tutti perfettamente sulla linea di regressione).

28.3.5 Stochastic regression imputation

È un fix della precedente che cerca di mettere la pezza al bias di correlazione aggiungendo noise alle predizioni

```
imp <- mice(data, method = "norm.nob", seed = 1, m = 1, print = FALSE)
xyplot(imp, Ozone ~ Solar.R)
```



Con `norm.nob` si richiede di applicare un metodo di regressione stocastica non-bayesiano; il metodo stima intercetta, pendenza e varianza dei residui sotto il modello lineare; dopodiché calcola le predizioni per ogni valore mancante e aggiunge uno scostamento casuale pescato dall'urna che descrive la distribuzione dei residui.
È un miglioramento rispetto alla precedente: l'idea di aggiungere noise dai residui è potente e sta alla base di metodi più avanzati

28.3.6 LOCF e BOCF

L sta per last mentre B sta per baseline; sono entrambe per dati specificamente longitudinali e i missing sono sostituiti rispettivamente dall'ultima osservazione disponibili o dalla prima. LOCF è la più usata e conosciuta.

Chiaramente LOCF può essere applicata senza problemi se sappiamo di variabili che non si muovono nel tempo; ma per gli outcome dei trial la LOCF è dubbia; la FDA l'ha storicamente vista come il metodo di analisi preferito (considerandolo conservativo e meno prono a bias rispetto alla listwise deletion). Tuttavia è stato dimostrato che può portare a bias anche sotto MCAR; se la si applica deve essere usata un metodo di analisi ad hoc che distingua tra dataset reale ed imputato [1]. LOCF e BOCF vanno usate solo se le assunzioni alla loro base sono giustificate[2]

28.3.7 Indicator method

Supponendo si voglia effettuare una regressione ma si hanno covariate con missing, allora considerando una singola covariata si sostituisce uno 0 al posto dei missing e nella regressione si aggiunge una dummy con valore 1 dove si è operata la sostituzione (questo per ogni variabile con missing).

[1] Molenberghs, G., and M. G. Kenward. 2007. Missing Data in Clinical Studies. Chichester: Wiley.
Kenward, M. G., and G. Molenberghs. 2009. "Last Observation Carried Forward: A Crystal Ball?" Journal of Biopharmaceutical Statistics 19 (5): 872–88.

[2] National Research Council. 2010. The Prevention and Treatment of Missing Data in Clinical Trials. Washington, D.C.: The National Academies Press.

```
imp <- mice(airquality, method = "mean", m = 1, maxit = 1, print = FALSE)
airquality2 <- cbind(complete(imp), oz_miss = is.na(airquality[, "Ozone"]))
fit <- lm(Wind ~ Ozone + oz_miss, data = airquality2)
```

Qua al posto di 0 si è usata la media e quello che va a cambiare è il coefficente di `oz_miss`, ma comunque l'idea è la stessa.
Questo è stato un metodo popolare in passato. An advantage is that the indicator method retains the full dataset. Also, it allows for systematic differences between the observed and the unobserved data by inclusion of the response indicator, and could be more efficient. White and Thompson (2005) pointed out that the method can be useful to estimate the treatment effect in randomized trials when a baseline covariate is partially observed. If the missing data are restricted to the covariate, if the interest is solely restricted to estimation of the treatment effect, if compliance to the allocated treatment is perfect and if the model is linear without interactions, then using the indicator method for that covariate yields an unbiased estimate of the treatment effect. This is true even if the missingness depends on the covariate itself. Additional work can be found in Groenwold et al. (2012; Sullivan et al. 2018). It is not yet clear whether the coverage of the confidence interval around the treatment estimate will be satisfactory for multiple incomplete baseline covariates.

The conditions under which the indicator method works may not be met in practice. For example, the method does not allow for missing data in the outcome, and generally fails in observational data. It has been shown that the method can yield severely biased regression estimates, even under MCAR and for low amounts of missing data (Vach and Blettner 1991; Greenland and Finkle 1995; Knol et al. 2010). The indicator method may have its uses in particular situations, but fails as a generic method to handle missing data.

28.3.8 Sintesi

In tabella 28.1 sintesi di assunzioni, correttezza di stima (assenza di bias) e correttezza delle stime di variabilità. Per l'assenza di bias in Reg Weight incomplete variable as dependent. Ad esempio la prima riga si legge

- Listwise deletion produces an unbiased estimate of the mean provided that the data are MCAR;
- Listwise deletion produces an estimate of the standard error that is too large.

Il segno - indica che il metodo non può produrre stime senza bias (es

28.4 Multiple imputation in a nutshell

La MI crea $m > 1$ dataset completi (spesso m è un valore alto); ciascuno di essi è analizzato da una procedura di analisi standard, dopodiché gli m risultati sono pooled in una stima finale alla quale è associato un errore standard.

Un esempio minimale con un modello di regressione è il seguente.

	Mean	Unbiased Reg Weight	Correlation	Std. error
Listwise	MCAR	MCAR	MCAR	Too large
Pairwise	MCAR	MCAR	MCAR	Complicated
Mean	MCAR	-	-	Too small
Regression	MAR	MAR	-	Too small
Stochastic	MAR	MAR	MAR	Too small
LOCF	-	-	-	Too small
Indicator	-	-	-	Too small

Tabella 28.1: Assunzioni e performance metodi ad-hoc

```
## con MI
imp <- mice(airquality, seed = 1, m = 20, print = FALSE)
fit <- with(imp, lm(Ozone ~ Wind + Temp + Solar.R))
summary(pool(fit))

##          term     estimate   std.error statistic       df      p.value
## 1 (Intercept) -65.87829658 23.09377412 -2.852643 69.97033 5.696702e-03
## 2        Wind  -3.01897171  0.66252377 -4.556775 70.51194 2.125022e-05
## 3        Temp   1.63483547  0.25107557  6.511328 75.99913 7.203792e-09
## 4     Solar.R   0.05861581  0.02267832  2.584662 90.10797 1.135441e-02

## stima diretta
fit <- lm(Ozone ~ Wind + Temp + Solar.R, data = airquality)
coef(summary(fit))

##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -64.34207893 23.05472435 -2.790841 6.226638e-03
## Wind         -3.33359131  0.65440710 -5.094063 1.515934e-06
## Temp          1.65209291  0.25352979  6.516366 2.423506e-09
## Solar.R       0.05982059  0.02318647  2.579979 1.123664e-02
```

Si nota come i beta siano molto simili (la maggior parte dei missing è nell'outcome Ozone), gli errori standard della MI sono lievemente più piccoli

Capitolo 29

Multiple imputation

29.1 Notazione

- $\mathbf{Y}$ una matrice $n \times p$ che contiene dati su p variabili per n unità del campione (un suo elemento indicato con y_{ij})
- $\mathbf{R}$ è il *response indicator*, ossia una $n \times p$ matrice di 0 o 1 (un suo elemento indicato con r_{ij}) che indica se i corrispondenti elementi di $\mathbf{Y}$ sono osservati ($r_{ij} = 1$) o mancanti ($r_{ij} = 0$)
- I dati osservati sono denotati da $\mathbf{Y}_{obs}$, quelli mancanti da $\mathbf{Y}_{mis}$ (gestita da $\mathbf{R}$), quindi $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. $\mathbf{Y}_{mis}$ ha valori reali ma che noi non possiamo vedere. Se fosse $\mathbf{Y} = \mathbf{Y}_{obs}$ (ossia tutto il campione è osservato, potremmo fare inferenza normalmente
- Il meccanismo che genera R potrebbe dipendere da $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ sia per il disegno (es alcuni dati a volte debbono essere missing per forza) che per casualità, e questa cosa è definita dal *missing data model*. Sia φ il set di parametri di tale modello; allora l'espressione generale che esprime il missing data model è
$$\mathbb{P}(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \varphi)$$
- m è il numero di imputazioni (dataset generati dalla MI)

I dati sono detti

- MCAR se
$$\mathbb{P}(\mathbf{R} = \mathbf{0}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \varphi) = \mathbb{P}(\mathbf{R} = \mathbf{0}, \varphi) \tag{29.1}$$
Ossia la probabilità che i dati siano mancanti non dipende dal valore assunto dagli stessi; bensì dipenden solamente da qualche parametro φ, la probabilità overall di essere missing.
- MAR se
$$\mathbb{P}(\mathbf{R} = \mathbf{0}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \varphi) = \mathbb{P}(\mathbf{R} = \mathbf{0}|\mathbf{Y}_{obs}, \varphi)$$
ossia la missingness dipende dall'informazione osservata oltre ad un fattore basale

- MNAR se
$$\mathbb{P}\left(\mathbf{R} = \mathbf{0} | \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \varphi\right)$$
non si semplifica, la probabilità di avere mancanti dipende anche dai valori assunti nei dati non osservati.

Come detto MCAR è poco realistica e ci sono metodi ad-hoc per gestirla; la MI può gestire sia MAR che MNAR

Differenti test sono stati proposti per valutare MCAR vs MAR; non sono molto utilizzati e il loro valore pratico non è chiaro. Non è possibile valutare MAR vs MNAR dato ce l'informazione necessaria per tale testing non è disponibile.

29.2 Quando non usare MI

Quando è utile utilizzare la complete case analysis (avendo a mente un modello di regressione):

- se la missingness è sull'outcome, senza pensarci due volte, qui i risultati di analisi normale e analisi MI è uguale
- se la probabilità di avere un dato mancante nelle covariate non dipende dal valore dell'outcome i coefficienti sono unbiased
- The second special case holds only if the complete data model is logistic regression. Suppose that the missing data are confined to either a dichotomous Y or to X, but not to both. Assuming that the model is correctly specified, the regression coefficients (except the intercept) from the complete-case analysis are unbiased if the probability to be missing depends only on Y and not on X (Vach 1994). This property provides the statistical basis of the estimation of the odds ratio from case-control studies in epidemiology. If missing data occur in both Y and X the property does not hold.

Comunque la complete case analysis dovrebbe essere una scelta esplicita; altre applicazioni particolari (es la mean, stochastic regression ecc) possono essere una alternativa anche se nessuna è generale come la MI.

29.3 Quante imputazioni (m)

Alcuni lavori empirici consigliano di porre m tra i 20 e i 100; comunque partire con un m basso nel setup del modello ed incrementarlo alla fine.

Capitolo 30

Univariate missing data

Qui chiamiamo la singola (ipotizzando che sia unica) variabile da imputare, *variabile target*. Ci poniamo sotto questa ipotesi.

Definition 30.0.1 (Ignorability)**.** le probabilità (distribuzione proprio) e relazioni entro le parti mancanti siano simili a quelle nelle parti osservate

Remark 93. Vedremo alcune alternative quando l'ipotesi non regge.

30.1 Variabile target quantitativa

Di base se vogliamo sfruttare la relazione con un'altra variabile:

- regression imputation: usare le predizioni di un modello lineare applicato ai dati (dove la variabile missing è la y): in `mice` implementato mediante `norm.predict`
- stochastic regression imputation: usare le predizioni e aggiungere un po' di noise: in `mice` è `norm.nob`
- usare predizioni con noise sia per il residuo che incorporando l'incertezza dei beta di stima; per fare questo bi sono metodi Bayesiani (in `mice`, `norm`)che pescano dalla posteriori dei parametri, oppure metodi bootstrap per costruire la distribuzione dei beta (in `mice` `norm.boot`)
- come la precedente ma aggiungiamo covariate al modello rendendo le stime auspicabilmente più precise
- *predictive mean matching*: calcoliamo la predizione prevista ma a questo punto per imputare usiamo un dato "vicino", selezionandolo casualmente tra un pool di candidati ragionevoli (in un intorno circolare del punto predetto, ossia confrontando le predizioni del missing con le predizioni dei dati disponibili).

Comunque `norm` e `norm.boot` performano meglio di quelli che vengono prima (a maggior ragione se il modello non è univariato)

30.1.1 Sul PMM e la scelta dei donor

In `mice` implementato mediante il method `pmm`, è un metodo facile da usare e relativamente versatile (gestisce anche variabili nominali o ordinali con la definizione di distanza adottata tipicamente .
Il parametro da settare è d il numero di candidati possibili *donor*, ossia il pool dei possibili donatori del valore dai quali estrarre. Il rischio a scegliere

TODO: Da approfondire sta distanza, cioè può essere un valore assoluto per predizioni numeriche, ma per qualitative si basa su differenza o ordinamento direi.

- $d = 1$, basso, è di scegliere lo stesso donor molte volte; il problema è maggiore in campioni piccoli o se ci sono più missing che disponibili e/o molti ties.
- d ad un valore alto, es $d = n/10$ può alleviare il problema ties ma si può andare a prendere unità anche molto differenti

Studi mostrano che:

- $d = 5$ e $d = 10$ funzionano bene; il default di `mice` è $d = 5$
- $d = 5$ potrebbe essere troppo elevato per dimensioni $n < 100$
- $d = 1$ è ok per studi con campione piccolo

Sono stati studiate strategie che pesano la probabilità di essere estratti per la distanza rispetto all'imputazione, facendo sì che si possa prenda tutto il campione, non specificando d; questo è disponibile in `mice` come metodo `midastouch`

Per concludere il metodo in generale funziona meglio in campioni grandi e fornisce imputazioni buone, non può darne fuori dal range osservato (a differenza es di predizioni basate su regressioni). Potrebbe non performare benissimo in piccoli campioni ma è una soluzione generale ben funzionante.

30.2 Variabile target qualitativa

Si va di regressione logistica (metodo `logreg` in mice) o multinomiale (unordered con il metodo `polyreg` o per dati ordinali con polr);

- per la logistica standard la probabilità predetta viene confrontata con un'estrazione uniforme tra 0 e 1 (se quest'ultima è sotto si predice evento, altrimenti non evento);
- per le multinomiali si procede analogamente mediante estrazione e vedendo dove ricade tale estrazione (costruendo dei ticks di probabilità che battezzano lo stato sulla base delle probabilità predette).

30.3 Variabile count

Le predizioni debbono essere non negative e per queste si può usare:

- predictive mean matching
- imputazioni per dati categorici ordinati (ossia trattare la variabile come un'ordinale)
- (zero-inflated) poisson regression

- (zero-inflated) negative binomial regression

Vedere il libro se si vogliono adottare le ultime due per pacchetti aggiuntivi/metodi da selezionare.

30.4 Semi continuous

Hanno densità di un punto specifico molto alta (spesso 0) e una distribuzione continua sui rimanenti valori (es fumo è a 0 per i non smoker mentre ha una distribuzione positiva con coda a destra per i fumatori).
Spesso questi dati sono modellati semplicemente come quantitativi:

- mediante pmm
- mediante un modello a due stadi; prima si determina se i valori imputati siano 0 o no (logistica), poi tra i secondi si imputa il valore (modello lineare)

Ci sono confronti tra i due metodi, vedere il libro per approfondimenti

30.5 Variabile censored

Ignoriamo per il momento

30.6 Nonignorable missing data

I precedenti metodi si fondano sull'assuzione di ignorabilità. Tuttavia non possiamo sapere se la missingess è di questa tipologia.
Vedi libro per approfondimenti.

Capitolo 31

Multivariate missing data

31.1 Pattern di missingness

Indichiamo con

- $\mathbf{Y}$, matrice $n \times p$ la matrice con tutti i dati (tutti sia disponibili che missing)
- ; indichiamo con $\mathbf{Y}_j$ la j variabile e con $\mathbf{Y}_{-j}$ tutte la matrice ad eccezione della j-esima variabile.
- $\mathbf{R}$ la matrice di 1 (presente) e 0 (missing) di dimensione $n \times p$ gemella a $\mathbf{Y}$

È utile distinguere tra diversi pattern di missingness (il pattern della missingness influenza l'ammontare di info che può essere trasferito da altre variabili in imputazione):

- *univariato* e *multivariato*: univariato se $\mathbf{Y}$ vi è solo una variabile con missing data
- *monotono* e *non* (generale): monotono se la variabile $\mathbf{Y}_j$ può essere ordinata in maniera tale che se Y_j è missing allora tutte le variabili $\mathbf{Y}_k$ con $k > j$ lo sono; se il pattern non è monotono si dice generale;
- *connected* e *unconnected*: detto connected se qualsiasi dato osservato/presente può essere raggiunto da qualsiasi altro dato osservato mediante una sequenza di movimenti orizzontali o verticali

Alcune funzioni e statistiche per statistiche sui dati mancanti:

- `md.pattern` calcola la frequenza dei patterns ed è utilizzabile per lo più per dataset con un numero basso di variabili; lo testiamo con un dataset di esempio

```
library(mice)
md.pattern(pattern4, plot = FALSE)
```

```
##   A B C
## 2 1 1 1 0
## 3 1 1 0 1
## 1 1 0 1 1
## 2 0 0 1 2
##   2 3 3 8
```

- analizzando il dataset a coppie di variabili si può usare la funzioni `md.pairs` che stampa sotto `rr` laddove entrambe le variabili sono presenti, `rm` dove solo la seconda è missing, `mr` dove solo la prima è missing, `mm` laddove entrambe le variabili sono missing

```
(p <- md.pairs(pattern4))

## $rr
##   A B C
## A 6 5 3
## B 5 5 2
## C 3 2 5
##
## $rm
##   A B C
## A 0 1 3
## B 0 0 3
## C 2 3 0
##
## $mr
##   A B C
## A 0 0 2
## B 1 0 3
## C 3 3 0
##
## $mm
##   A B C
## A 2 2 0
## B 2 3 0
## C 0 0 3
```

Ad esempio per la coppia di variabili (A, B) ci sono 5 casi in cui entrambe le variabili sono complete (in `rr`), un caso in cui solo A è mancante (`mr`) e 2 casi in cui entrambe lo sono (`mm`).

31.1.1 Inbound e outbound statistics

Sulla base delle statistiche legate a coppie di variaibli possiamo definire alcuni indicatori:

- *inbound statistic*: la proporzione di casi utilizzabili per imputare la variabile $\mathbf{Y}_j$ dalla $\mathbf{Y}_k$ è definita come

$$I_{jk} = \frac{\sum_{i=1}^{n}(1 - r_{ij})r_{ik}}{\sum_{i=1}^{n} 1 - r_{ij}} \tag{31.1}$$

  e può essere interpretata la proporzione di casi in cui $\mathbf{Y}_j$ è mancante e $\mathbf{Y}_k$ è osservata (sul numero complessivo di mancanti di $\mathbf{Y}_j$); vale 1 se $\mathbf{Y}_k$ è sempre osservata laddove $\mathbf{Y}_j$ è missing.
  Questa statistica può essere utile per selezionare velocemente quali predittori usare per imputare (preferendo quelli con valore maggiore). Per calcolarla sul dataset `pattern4`:

```
p$mr / (p$mr + p$mm)
```

```
##           A B C
## A 0.0000000 0 1
## B 0.3333333 0 1
## C 1.0000000 1 0
```

  la prima riga contiene $I_{AA} = 0$, $I_{AB} = 0$ e $I_{AC} = 1$, quindi ad esempio B non è rilevante per imputare A, mentre C è osservata per entrambi i casi in cui A è missing.
  Questa è una *inbound statistic* che misura come le missingess nella variabile j-esima sono connesse al resto dei dati

- l'*outbound statistic* O_{jk} misura come i dati presenti nella variabile Y_j sono connessi al resto delle variabili

$$O_{jk} = \frac{\sum_{i=1}^{n} r_{ij}(1 - r_{ik})}{\sum_{i=1}^{n} r_{ij}}$$

  è la proporzione di casi con Y_j osservato e Y_k mancante diviso il numero di casi in cui Y_j è osservato; è uguale a 1 se Y_j è osservata in tutti i record dove Y_j è missing e può essere usata per decidere se Y_j è un potenziale predittore per imputare Y_k.
  Per calcolarlo

```
p$rm/(p$rm+p$rr)
```

```
##     A         B   C
## A 0.0 0.1666667 0.5
## B 0.0 0.0000000 0.6
## C 0.4 0.6000000 0.0
```

  pertanto A è potenzialmente più utile per imputare C (3/6) rispetto a B (1/6).

31.1.2 Coefficienti influx e outflux

Il *coefficiente di influx* I_j, definito come

$$I_j = \frac{\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{i=1}^{n}(1 - r_{ij})r_{ik}}{\sum_{k=1}^{p}\sum_{i=1}^{n} r_{ik}} \tag{31.2}$$

è uguale alla proporzione di casi in cui Y_j è missing e Y_k è osservata diviso il numero complessivo di dati osservati: una variabile osservata completamente ha indice pari a 0, una completamente missing pari a 1

Il *coefficiente di outflux* è definito

$$O_j = \frac{\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{i=1}^{n} r_{ij}(1-r_{ik})}{\sum_{k=1}^{p}\sum_{i=1}^{n} 1-r_{ij}} \tag{31.3}$$

è la proporzione di righe con Y_j osservata e Y_k mancante diviso il numero complessivo di dati mancanti; è un indicatore dell'utilità di Y_j per imputare altre variabili; una variabile completamente osservata lo ha uguale a 1 mentre completamente mancante pari a 0

In `mice` per calcolare i coefficientid di influx e outflux

```
flux(pattern4)[, 1:3]

##    pobs influx outflux
## A 0.750  0.125   0.500
## B 0.625  0.250   0.375
## C 0.625  0.375   0.625
```

le righe corrispondono alle variabili, mentre le colonne le percentuali di dati osservati su ciascuna variabile, e gli indici I_j e O_j. Di base le variabili che sono piazzate in alto nella videata sono più complete e potenzialmente più utili per l'imputazione

Gli indici di influx e outflux sono di aiuto (a parità d'altro le variabili con alto influx e outflux sono da preferire), ma non garantiscono l'effettiva utilità (es se non sono correlate con la variabile da imputare). Alcuni approcci stanno considerando congiuntamente (es mediante moltiplicazione) questi indici e l'R^2 della regressione.

31.2 Questioni da considerare nella MI

La maggior parte dei modelli di MI per Y_j utilizzano tutte le rimanenti Y_{-j} come predittori (per preservare le relazioni) ma facendo si che possano sorgere vari problemi come:

- predittori missing
- dipendenza circolare
- con elevato numero di predittori e basso n multicollinearità
- le relazioni possono essere non lineari
- si possono creare situazioni strane tipo maschio = si e incinta = si

Vi sono tre strategie generali per imputare dataset multivariati:

- *monotone data imputation*: per pattern di missingness monotoni, le imputazioni sono create da una sequenza di metodi univariati

- *joint modeling*: per pattern generali (non monotoni), le imputazioni sono create da un modello multivariato (??MANOVA like)
- *fully conditional specification*: anche detta *chained equations* e *sequential regression*, per pattern generali, un modello è implicitamente specificato da un set di modelli univariati condizionali. L'imputazione è creata pescando da sti modelli.

31.3 Monotone data imputation

Se ipotizziamo che fino a $Y_1, \ldots, Y_p$ siano le variabili in progressione monotona di missingness e prima di esse ve ne siano X complete, usiamo le X per imputare Y_1, X più Y_1 integrata per imputare Y_2 e così via. Tutti i metodi di imputazione sono univariati (es se Y_1 è dicotomica può essere imputata mediante logistica). L'implementazione di questo tipo di imputazione in `mice` si ha mediante `visit` a `monotone`: vediamo un esempio

```
library(mice)
data <- nhanes2[, 1:3]
md.pattern(data, plot = FALSE)

##    age hyp bmi
## 16   1   1   1  0
## 1    1   1   0  1
## 8    1   0   0  2
##      0   8   9 17

imp <- mice(data, visit = "monotone", maxit = 1, m = 2, print = FALSE)
```

`md.pattern` fa vedere la struttura monotona della missingness di questo subset.

31.4 Joint modeling

L'assunzione è che i dati possano esser descritti da una distribuzione multivariata; assumendo l'ipotesi di ignorabilità le imputazioni sono create dalla distribuzione stimata come verosimile dai dati.

31.5 Fully conditional specification

Imputa i missing variabile per variabile; richiede la specifica di un modello per ogni variabile incompleta e crea imputazioni per ciascuna variabile in maniera iterativa. È quello che fa `mice`

Capitolo 32

Analisi di dati imputati

Il processo è questo

1. ottenere gli m dataset
2. stimare i parametri di interesse (sui dati completi) per ciascuno di essi
3. fare il pooling dei risultati per l'inferenza

Il workflow classico in `mice` prevede rispettivamente l'utilizzo delle funzioni

- `mice` per imputare
- `with` per effettuare le analisi
- `pool` per il pooling

Alcune caratteristiche delle funzioni in tabella 32.1.
In questo capitolo ci si concentra sulla fase di analisi; nel prossimo in quella di pooling.

```
library(mice)
imp <- mice(nhanes, seed = 123, print = FALSE)
fit <- with(imp, lm(chl ~ age + bmi + hyp))
est1 <- pool(fit)
```

32.1 Analisi

Per ripetere le analisi sui dataset imputati si usa `with` che prende in input due argomenti;

Funzione	Cosa produce	(aka di classe)
`mice`	multiply imputed dataset	`mids`
`complete`	multiply imputed list of data	`mild`
`with`	multiple imputation repeated analyses	`mira`
`pool`	multiple imputation pooled results	`mipo`

Tabella 32.1: Funzioni di `mice`

1. il primo è un `mids` prodotto da `mice::mice`
2. il secondo è una espressione che viene applicata a ciascun dataset

Alcune indicazioni sull'uso di `with`:

- i risultati delle stime si trovano in `analyses`

```
fit <- with(imp, lm(chl ~ bmi + age))
class(fit$analyses)

## [1] "list"

length(fit$analyses)

## [1] 5

fit$analyses[[1]]

##
## Call:
## lm(formula = chl ~ bmi + age)
##
## Coefficients:
## (Intercept)          bmi          age
##      56.374        3.158       32.093
```

- come detto si può passare una espressione generica che faccia uso dei `data.frame` (ipotizziamo di essere dentro uno di essi);

```
head(nhanes)

##   age  bmi hyp chl
## 1   1   NA  NA  NA
## 2   2 22.7   1 187
## 3   1   NA   1 187
## 4   3   NA  NA  NA
## 5   1 20.4   1 113
## 6   3   NA  NA 184

## un esempio fittizio di media del bmi del dataset imputato,
## moltiplicata per 2
nonsense <- expression(
    mu <- mean(bmi),
    mu * 2
)
fit <- with(imp, eval(nonsense))
unlist(fit$analyses)

## [1] 54.992 53.064 53.120 53.960 53.040
```

Da notare nell'esempio precedente

1. l'uso di `expression` per evitare che l'espressione venga valutata subito producendo errore (cercando bmi nel workspace)
2. la separazione mediante virgola degli statement sotto espressione
3. il fatto che possiamo assumere che `bmi` non sia missing (perché fa uso di un dataset imputato
4. infine l'uso di `eval` nella chiamata a `with`.

32.2 Pooling dei parametri

32.2.1 Inferenza scalare

32.2.1.1 Stimatori normali

Se si può assumere che lo stimatore $\widehat{Q}$ abbia una distribuzione normale (attorno al valore vero della popolazione Q con varianza U si applica la regola di Rubin per il pooling delle stime.

La difficoltà primaria risiede nel trovare i gradi di libertà per le distribuzioni t ed F Da finire qui quando hai più tempo .

TODO: Finire qui

Molti tipi di stime sono approssimativamente normally distribuited (es medie deviazioni standard, coefficienti di regressione, proporzioni, e predittori lineari), e la regola di Rubin può essere applicata direttamente a questi.

32.2.1.2 Stimatori non normali

Come combinare risultati con distribuzioni non normali: esempio coefficienti di correlazione, odds ratio, rischi relativi, hazard ratio, R^2 ecc?
La risposta è trasformazioni verso la normalità, calcolo dell'intervallo di confidenza mediante approssimazione e ritrasformazione del tutto nella scala originaria. Una

Example 32.2.1 (Correlazione)**.** Per la correlazione una trasformazione comune è la z di Fisher

$$z_l = \frac{1}{2} \log \frac{1+\rho_l}{1-\rho_l}$$

dove ρ_l è il coefficiente di correlazione dell'l-esimo dataset imputato. Per campioni grandi, la distribuzione di z_l è normale con varianza $\sigma^2 = 1/(n-3)$. Una volta effettuata la trasformazione si calcola la stima pooled $\overline{z}$ (e suo intervallo di confidenza) mediante la regola di Rubin, dopodiché di backtrasforma in scala originale sia stima che CI con la trasformazione inversa della z di Fisher ossia

$$\overline{\rho} = \frac{e^{2\overline{z}} - 1}{e^{2\overline{z}} + 1}$$

L'insieme delle trasformazioni suggerite per vari tipi di risultato è riassunta in 32.2.

TODO: Da finire con le citazioni

Statistica	Trasformazione	Source
Correlazione	Fisher z	
Odds ratio	log	
Relative risk	log	
Hazard ratio	log	
R^2	Fisher z on root	
Survival probability	Complementary log-log	
Survival distribution	log	

Tabella 32.2: Suggested transformations toward normality for various types of statistics. The transformed quantities can be pooled by Rubin's rules.

32.2.1.3 Distribuzioni sconosciute o complesse

Vi sono quantità per le quali la distribuzione è complessa o sconosciuta: ad esempio la C di Cramer, l'indice di discriminazione ecc. Attualmente conviene cercare una trasformazione che riconduca la distribuzione dello stimatore approssimativamente normale (e la sua inversa) e procedere come in precedenza per fare inferenza.

32.2.2 Inferenza vettoriale

Nel caso in cui siamo interessati ad un set di parametri congiuntamente, vi possono essere tre tipi di test multi-parametro: D_1 (test di Wald multivariato), D_2 (combinazione delle statistiche) o D_3 (likelihood ratio test). Vediamo il primo, consigliato.

32.2.2.1 D_1 test di Wald multivariato

Si testa se $Q = Q_0$ con Q_0 un vettore di k elementi sotto l'ipotesi nulla (solitamente tutti 0). Ad esempio possiamo testare i beta di una covariata qualitativa a più livelli per vedere se tenerla nel modello o no; nel seguente caso vediamo se tenere le classi di età

```
imp <- mice(nhanes2, m = 10, print = FALSE, seed = 71242)
m2 <- with(imp, lm(chl ~ age + bmi))
pool(m2)

## Class: mipo    m = 10
##          term  m  estimate        ubar           b          t dfcom        df
## 1 (Intercept) 10  6.131277 2474.600837 1433.877979 4051.86661    21  9.813976 0.63
## 2    age40-59 10 46.828507  255.560889  168.478122  440.88682    21  9.153159 0.72
## 3    age60-99 10 66.074544  297.082833  200.173522  517.27371    21  9.042895 0.74
## 4         bmi 10  5.885952    2.866251    1.366672    4.36959    21 10.828772 0.52
##         fmi
## 1 0.4845916
## 2 0.5157392
## 3 0.5210555
## 4 0.4389138
```

```
m1 <- with(imp, lm(chl ~ bmi))
summary(D1(m2, m1))

##
## Models:
##  model         formula
##      1 chl ~ age + bmi
##      2       chl ~ bmi
##
## Comparisons:
##    test statistic df1      df2 dfcom    p.value       riv
##  1 ~~ 2  5.021108   2 11.85501    21 0.02632666 0.6283236
##
## Number of imputations:  10   Method D1
```

Quindi dato che il test di Wald è significativo, la rimozione dell'età dal modello riduce il suo potere predittivo

32.2.3 Selezione Variabili

Non è ovvio come porre assieme i modelli generati da una procedura di selezione variabili (in alcuni casi si possono selezionare alcune variabili, in altri altre) e sono stati proposti diversi approcci

32.2.3.1 Stepwise

L'idea è stimare un modello con le variabili che sono state selezionate (es mediante stepwise da almeno metà delle procedure/campioni imputati

```
## creiamo un subset di dati biometrici
data <- boys[boys$age >= 8, -4]
## imputiamo i missing
imp <- mice(data, seed = 28382, m = 10, print = FALSE)
## parametri della stepwise (su tv, testicular volume)
scope <- list(upper = ~ age + hgt + wgt + hc + gen + phb + reg,
              lower = ~ 1)
expr <- expression(f1 <- lm(tv ~ 1),
                   f2 <- step(f1, scope = scope))
## stimiamo i modelli stepwise
fit <- with(imp, expr)
## estraiamo la conta delle variabili selezionate estraendo la formula
formulas <- lapply(fit$analyses, formula)
## ... o meglio le covariate dipendenti
terms <- lapply(formulas, terms)
votes <- unlist(lapply(terms, labels))
## vediamo le conte delle selezioni
table(votes)
which(table(votes) > 5)
## procediamo a stima finale
fit <- with(imp, lm(tv ~ age + gen + reg + phb + hgt))
```

```
## lui comunque alla fine non da la stima seguente
## summary(pool(fit))
```

Per approfondire:

- Wood, A. M., I. R. White, and P. Royston. 2008. "How Should Variable Selection Be Performed with Multiply Imputed Data?" Statistics in Medicine 27 (17): 3227–46.
- Vergouwe, Y., P. Royston, K. G. M. Moons, and D. G. Altman. 2010. "Development and Validation of a Prediction Model with Missing Predictor Data: A Practical Approach." Journal of Clinical Epidemiology 63 (2): 205–14.

32.2.3.2 Lasso

MI-LASSO method by Chen and Wang tests the coefficients across all the stacked datasets, thus ensuring model consistency across different imputations. Da approfondire:

- Zhao, Y., and Q. Long. 2017. "Variable Selection in the Presence of Missing Data: Imputation-Based Methods." Wiley Interdisciplinary Reviews: Computational Statistics 9 (5).
- Chen, Q., and S. Wang. 2013. "Variable Selection for Multiply-Imputed Data with Application to Dioxin Exposure Study"

Capitolo 33

Aspetti pratici

33.1 Scelta dei predittori della missingness

Si consiglia di porre

1. includere tutte le variabili (outcome e covariate) che verranno poste nel modello finale stimato sui dati completi
2. includere variabili che sono rilevanti/correlate con la non risposta (cfr correlazioni con l'indicatore di risposta della variabile che deve essere imputata) scegliendo una certa soglia al di sopra della quale includere la variabile
3. aggiungere variabili che spiegano buona percentuale di varianza (correlazione con l'outcome)
4. di eliminare degli step 2 e 3 le variabili con molti missing entro il sottogruppo dei casi incompleti (es sulla base della percentuale dei casi osservati in questo sottogruppo, la percentuale di casi utilizzabili).

Dare un occhio alla funzione `quickpred` (che dovrebbe essere un test di `cor` applicato ad un data.frame, sopra una certa soglia)

33.2 Altre cose utili/interessanti trovate qua e la

Qui:

- variabili carattere vanno eliminati o trasformati in factor
- alte proporzioni di missing: in fase esplorativa, eliminare variabili con più del 50% di missingness dall'imputazione
- alta multicollinearità: la MI non ama le variabili troppo correlate tra loro. La maggior parte delle volte ci pensa `mice` ad eliminarle ma se ci dovessero essere errori potrebbe essere utile partire con solamente un subset di variabili e incrementare il numero delle incluse fino a che si trova la problematica. Settando `print=TRUE` in `mice::mice` si riesce a vedere dove l'algoritmo ha difficoltà/si blocca nell'imputazione

- controllare i dataset imputati che non abbiano dati mancanti: se ci sono l'imputazione non ha avuto buon fine, provare ad aggiungere predittori?

Per plottare i dati imputati vs i dati reali si può usare `densityplot` sull'oggetto creato da `mice`

Bibliografia

Altman D. G. (1985). Comparability of Randomised Groups. *The Statistician*, **34**(1), 125.

Amadori D. (2004). *Sperimentazione clinica in oncologia.* Poletto Editore.

Anthoine E.; Moret L.; Regnault A.; Sbille V.; Hardouin J. B. (2014). Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures.

Arafat S. (2016). Cross Cultural Adaptation and Psychometric Validation of Instruments: Step-wise Description. *International journal of psychiatry*, **1**, 4.

Armitage P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**(3), 375–386.

Bacchieri A.; Della Cioppa G. (2004). *Fondamenti di ricerca clinica (Italian Edition).* Springer, 2004 edizione.

Baio G. (2013). *Bayesian methods in health economics (Chapman & Hall/CRC Biostatistics Series).* Chapman and Hall/CRC, 1 edizione.

Borenstein M.; Hedges L.; Higgins J.; Rothstein H. (2011). *Introduction to Meta-Analysis.* Wiley.

Brown S.; Gregory W.; Twelves C.; Brown J. (2014). *A Practical Guide to Designing Phase II Trials in Oncology.* Statistics in Practice. Wiley.

Bryant J.; Day R. (1995). Incorporating Toxicity Considerations Into the Design of Two-Stage Phase II Clinical Trials. *Biometrics*, **51**(4), 1372.

Chang M. N.; Therneau T. M.; Wieand H. S.; Cha S. S. (1987). Designs for Group Sequential Phase II Clinical Trials. *Biometrics*, **43**(4), 865.

Chen T. T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, **16**(23), 2701–2711.

Chow S.; Wang H.; Shao J. (2007). *Sample Size Calculations in Clinical Research, Second Edition.* Chapman & Hall/CRC Biostatistics Series. Taylor & Francis.

Cronbach L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**(3), 297–334.

DeVellis R. (2012). *Scale Development: Theory and Applications*. Applied Social Research Methods. SAGE Publications.

Deyo R. A.; Centor R. M. (1986). Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis*, **39**(11), 897–906.

Eisenhauer E.; Twelves C.; Buyse M. (2014). *Phase I Cancer Clinical Trials: A Practical Guide*. Oxford University Press.

Eisenhauer E. A.; Therasse P.; Bogaerts J.; Schwartz L. H.; Sargent D.; Ford R.; Dancey J.; Arbuck S.; Gwyther S.; Mooney M.; Rubinstein L.; Shankar L.; Dodd L.; Kaplan R.; Lacombe D.; Verweij J. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer*, **45**(2), 228–247.

Eldridge S. M.; Chan C. L.; Campbell M. J.; Bond C. M.; Hopewell S.; Thabane L.; Lancaster G. A. (2016). CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ*, **355**, i5239.

Ensign L. G.; Gehan E. A.; Kamen D. S.; Thall P. F. (1994). An optimal three-stage design for phase II clinical trials. *Statistics in Medicine*, **13**(17), 1727–1736.

Glick H. A.; Doshi J. A.; Sonnad S. S.; Polsky D. (2007). *Economic Evaluation in Clinical Trials (Handbooks in Health Economic Evaluation)*. Oxford University Press, USA, 1 edizione.

Higgins J. P.; Thomas J.; Chandler J.; Cumpston M.; Li T.; Page M. J.; Welch V. A., (A cura di) (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley, 2 edizione.

ICMJE (2014). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals.

Landis J. R.; Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

Liberati A.; Altman D. G.; Tetzlaff J.; Mulrow C.; Gotzsche P. C.; Ioannidis J. P.; Clarke M.; Devereaux P. J.; Kleijnen J.; Moher D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, **339**, b2700.

Mariani L.; Marubini E. (1996). Design and Analysis of Phase II Cancer Trials: A Review of Statistical Methods and Guidelines for Medical Researchers. *International Statistical Review / Revue Internationale de Statistique*, **64**(1), 61.

Pepe M. S. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction (Oxford Statistical Science Series)*. Oxford University Press, USA, 1 edizione.

Rubin D. B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Simon R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, **10**(1), 1–10.

Sousa V. D.; Rojjanasrirat W. (2011). Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *J Eval Clin Pract*, **17**(2), 268–274.

Terwee C. B.; Bot S. D.; de Boer M. R.; van der Windt D. A.; Knol D. L.; Dekker J.; Bouter L. M.; de Vet H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*, **60**(1), 34–42.

VanBuuren S. (2018). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, Taylor & Francis Group.

Wheeler G. M.; Mander A. P.; Bedding A.; Brock K.; Cornelius V.; Grieve A. P.; Jaki T.; Love S. B.; Odondi L.; Weir C. J.; Yap C.; Bond S. J. (2019). How to design a dose-finding study using the continual reassessment method. *BMC Medical Research Methodology*, **19**(1).

Woodward M. (2004). *Epidemiology: Study Design and Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 edizione.