Master di I Livello in Management per funzioni di coordinamento delle professioni sanitarie A.A. 2020/2021

Università degli Studi di Milano-Bicocca

Appunti Disegno di Studio in Epidemiologia

Marta Pelusi
@Marta629

Indice

1 Study design

La ricerca biomedica ha come **obiettivo** l'investigazione delle relazioni tra fattori di esposizione (caratteristiche del paziente/trattamenti) e una condizione di salute.
Il **disegno dello studio** è una visione di un prodotto finale e uno schema/ragionamento logico/organizzativo per provare a rispondere alle domande motivanti. Un ruolo fondamentale è la selezione dei campioni e strumenti statistici.

Target population: popolazione rispetto alla quale si vuole fare una generalizzazione rispetto ai risultati che si ottengono.
Popolazione campionaria: lo studio si costruisce a partire da un campione a partire da una sample population (popolazione campione).
Spesso il campione è un sottoinsieme della sample population.
Esposizione: fattori che potrebbero avere influenza sull'outcome.

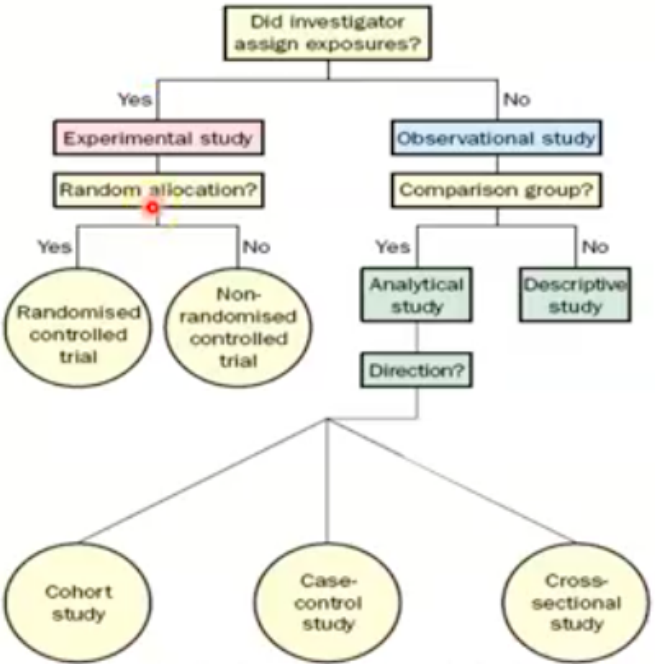

Figura 1: Esempio diagramma di studio.

La situazione migliore per avere uno studio più pulito è quella dello **studio randomizzato**. Grazie alla randomizzazione si avranno due gruppi a confronto che hanno più o meno la stessa organizzazione. Quello che differirà i due gruppi sarà la terapia di trattamento.

In questo diagramma si distinguono tre categorie di *studio analitico*, per cui è importante cogliere le differenze:

- **studi caso-controllo:** la selezione dei soggetti avviene rispetto alla conoscenze del loro outcome (casi: outcome positivo, controlli: outcome negativo). Qui si estrae dallo schema un campione di casi e un campione

di controlli per vedere in modo *retrospettivo*(guardando il passato) come si comportava il fattore di esposizione.

- **studi di coorte:** selezione del campione con un singolo campione estratto *in blocco* o con due campioni estratti dalla sottopopolazione di *esposti* e di *non esposti*. C'è una condizione di tempo (futuro) dove si attenderà l'esposizione dell'outcome.

- **studio fotografia:** il tempo si è fermato e si valuta l'associazione tra l'esposizione e outcome che sono copresenti al momento della valutazione. Come avviene per lo studio di coorte, può essere costituito o da un campione stato in blocco o da due campioni (esposti/non esposti). Il campione è estratto il più delle volte dalla sample population. Non c'è una dimensione di tempo perchè l'outcome è già disponibile.

1.1 Studi sperimentali

Negli **studi sperimentali** si è interessati a capire l'impatto sullo sviluppo di una condizione di salute.
Il vantaggio degli **studi randomizzati e controllati** è che grazie alla randomizzazione gli esposti e i non esposti condividono la stessa distribuzione di una serie di variabili che possono avere un impatto sull'outcome. La differenza tra i due gruppi è data dall'esposizione, che è propria dell'organizzazione dello sperimentatore. Grazie a questo forte impatto che ha lo sperimentatore a livello di organizzazione, hanno un'analisi statistica molto semplice in quanto c'è molto controllo dall'altro, quindi si analizza solo la variabilità dell'outcome (imputabili all'esposizione/non esposizione).
Questo tipo di studio migliora se l'esposizione viene data in modo *cieco*, ovvero il paziente non è consapevole della sua esposizione. Anche l'outcome è osservato in cieco, misurandolo e facendo una valutazione meno distorta e condizionata dal trattamento ricevuto dal paziente. Chi analizza i dati è avvantaggiato se non sa a quale trattamento corrisponde un outcome perchè si sarà meno tentati di dimostrare il corretto funzionamento di un trattamento (perchè non si conosce qual è).
Il **principio intention-to-treat** è un principio che sta alla base dell'analisi dei dati, e dice che il fattore di esposizione che il paziente ha ricevuto casualmente viene mantenuto fisso anche se il paziente, nel corso del tempo, decide di cambiare la sua esposizione. Rappresenta quello che succede nella pratica quando si *prescrive*, ovvero non ci si accerta che il trattamento sia davvero stato seguito dal paziente.

1.2 Studi osservazionali

Il ricercatore è interessato a capire l'impatto di un fattore di esposizione rispetto allo sviluppo di una certa condizione di salute in outcome. In questi studi non vi è randimizzazione e bisogna ragionare rispetto al **problema di confondimento**. Esso è il complementare del bilanciamento che si ha negli studi randomizzati. Si

ha un **confondimento** se i due gruppi esposti/non esposti hanno al loro interno un andamento diverso di alcune variabili. In modo più specifico si parla di confondimento quando c'è uno sbilanciamento su una variabile che è sia legata all'esposizione/non esposizione e che ha anche un impatto sull'outcome.
Questo sbilanciamento va tenuto conto in fase di analisi, altrimenti si rischia di sopravvalutare il ruolo della variabile sbilanciata rispetto a quella che rimane "nascosta". Questi studi hanno un'analisi statistica un po' più complessa rispetto allo studio randomizzato perchè dev'essere aggiustato in relazione a possibili *confounder*.

1.3 Differenze tra studio di coorte e studio randomizzato

Come già detto, c'è un elemento di confusione riguardo al fattore **tempo**, il quale nello studio cross-sectional non esiste. Il nostro obiettivo è capire se l'esposizione/non esposizione ha impatto sull'outcome, ma a volte potrebbe essere l'outcome ad avere impatto sull'esposizione/non esposizione. Un esempio di questo fenomeno può essere il fattore obesità.

Lo studio di coorte con l'aspetto del **tempo** ha degli elementi in comune con la dimensione di tempo che si ha negli studi controllati randomizzati. In questo tipo di studio si randomizza e si osserva *nel tempo* se vi è uno sviluppo dell'outcome in un certo orizzonte temporale. Lo studio di coorte e lo studio randomizzato controllato hanno quindi in comune l'aspetto del tempo che *guarda avanti* per vedere se si manifesta l'outcome. Nella nomenclatura, per questo motivo, si fa confusione. La differenza, però, tra le due situazioni è che nello studio randomizzato ho deciso io esposto/non esposto e c'è un bilanciamento tra le variabili, mentre nello studio di coorte non ho deciso io chi sono esposti/non esposti, ma li ho solamente osservati; ho campioni in partenza che non sono bilanciati e nell'analizzare l'outcome devo tenere in considerazione fattori di confondimento.

2 Quantità stimabili

Notazione: esposto=E, non esposto=NE, outcome positivo=O, outcome negativo=NO

Per quanto riguarda le quantità che si possono stimare dal punto di vista probabilistico con una strategia di campionamento **randomizzato controllato**, si valuterà nel tempo la probabilità condizionata sull'osservazione dell'outcome:

$$P(O|E),\ P(O|NE)$$

Per quanto riguarda le seguenti probabilità, invece, non sono stimabili da uno studio randomizzato controllato in quanto non si è scelto un campione di O e NO, si hanno avuti outcome in base a ciò che si è sviluppato nel tempo, mentre il mio campione era E e NE.

$$P(E|O),\ P(E|NO)$$

Per non sbagliare si considera la regola: si possono stimare le probabilità condizionate solo se sono condizionate ad un fattore che rappresenta la classificazione del mio campione. Questo è il caso degli studi randomizzati (campione di esposti/non esposti).
Dal rapporto di queste probabilità si ottiene il **rischio relativo**:

$$\text{rischiorelativo} = \frac{P(O|E)}{P(O|NE)}$$

Facendo la **differenza** tra le due probabilità di può anche calcolare la seguente probabilità:

$$\text{differenza} = P(O|E) - P(O|NE)$$

Considerando, invece, uno studio **cross sectional** si sa che si campiona da tutta la popolazione, si possono stimare tutte le quantità sopracitate.
Per lo studio **caso-controllo** si campiona dall'outcome, quindi le probabilità stimabili sono solamente

$$P(E|O),\ P(E|NO)$$

Non si può quindi stimare il rischio relativo, ma si possono stimare la **probabilità di esposizione relativa** o (**odds ratio**), ovvero l'approssimazione del rischio relativo ottenuta da uno studio di caso controllo che mima il rischio relativo ottenuto da uno studio cross sectional o da uno studio di coorte, a seconda del contesto.
Per lo **studio di coorte** il campionamento si fa separatamente da esposti e non esposti e si potranno stimare le probabilità in modo analogo a quello che posso fare in uno studio randomizzato.

In conclusione le probabilità

$$P(O|E),\ P(O|NE)$$

sono quelle che si prestano ad un'interpretazione migliore.

Esempio 1 (**coorte o cross sectional**)**.** *Considero una sample population di numero finito che rappresenta un contesto di* ***coorte o cross sectional****. Si considerano 20 esposti e 80 non esposti.*
La differenza tra questi due studi è che nello studio di coorte le conclusioni avranno un nesso di causalità, mentre nello studio cross sectional ci sono dei dubbi sulla causalità che a volte diventa causalità inversa.

	Outcome POS	**Outcome NEG**	
EXPOSED	50	60	110
NON EXPOSED	300	590	890
	350	650	1000

	Outcome POS	**Outcome NEG**	
EXPOSED			20
NON EXPOSED			80

Supponiamo di campionare 100 soggetti prendendone 20 dagli eposti e 80 dai non esposti (campioni casuali). Vogliamo calcolare, su 20 esposti, quanti soggetti con outcome positivo ci aspettiamo.

Popolazione

Calcoliamo la probabilità di outcome positivo sapendo che il soggetto è esposto:

$$P(outP|E) = \frac{50}{110} = 0.45$$

Avendo campionato 20 soggetti dalla popolazione di 110, il numero di esposti che mi aspetto è pari a $0.45 \cdot 20 = 9.09 \approx 9$. *Per differenza posso calcolare l'outcome negativo facendo* $20 - 9 = 11$.
Lo stesso ragionamento lo si può fare per i non esposti calcolando la probabilità di outcome positivo sapendo che il soggetto è non esposto:

$$P(outP|NE) = \frac{300}{890} = 0.34$$

Se nello studio reale ho campionato 20 non esposti, mi aspetto circa $0.34 \cdot 80 = 26.9 \approx 27$. *Per differenza si ha* $80 - 27 = 53$ *outcome negativi.*
La tabella completa è la seguente:

	Outcome POS	**Outcome NEG**	
EXPOSED	**9**	**11**	20
NON EXPOSED	**27**	**53**	80

Voglio ora calcolare il rischio relativo di popolazione:

$$RR_p = \frac{P(outP|E)}{P(outP|NE)} = \frac{0.45}{0.34} = 1.348 \approx 1.35$$

Calcolo la risk difference, ovvero la differenza tra i rischi:

$$\text{RISK DIFF}_\text{p} = P(outP|E) - P(outP|NE) = 0.11747 \approx 0.12$$

Stima

Ora stimo la probabilità di outcome positivo e negativo a partire dal mio studio:

$$P(outP|E) = \frac{9}{20} = 0.45$$

$$P(outP|NE) = \frac{27}{80} = 0.34$$

Calcolo nuovamente il rischio relativo di stima:

$$RR_s = \frac{P(outP|E)}{P(outP|NE)} = \frac{0.45}{0.34} = 1.33$$

Calcolo la risk difference, ovvero la differenza tra i rischi:

$$\text{RISK DIFF}_\text{s} = P(outP|E) - P(outP|NE) = 0.11250 \approx 0.11$$

Nota: *sebbene i calcoli sono gli stessi, il valore dei due rischi è diverso per una questione di approssimazione; il RR_p ha più cifre decimali approssimate rispetto al RR_s.*

Tutto questo si potrebbe estendere anche ad uno ***studio randomizzato*** *(se i campioni fossero randomizzati): non posso campionare dall'outcome e ho un aspetto di tempo per osservare l'outcome. In questo caso fisso il numero di esposti uguale al numero di non esposti (che sono quindi bilanciati) e poi vado a vedere l'outcome.*

Esempio 2 (**caso controllo**)**.** *Considero una sample population di numero finito che rappresenta un contesto di* ***caso controllo****. Si considerano 20 controlli e 80 casi.*

	Outcome POS	**Outcome NEG**	
EXPOSED	50	60	110
NON EXPOSED	300	590	890
	350	650	1000

	Outcome POS	**Outcome NEG**
EXPOSED		
NON EXPOSED		
	80	20
	casi	**controlli**

Popolazione

Queste probabilità vengono calcolate esattamente come mostrato nell'esempio 1:

$$P(outP|E) = 0.45455 \quad P(outP|NE) = 0.33708$$

$$RR_p = 1.35 \quad \text{RISK DIFF}_\text{p} = 0.11747$$

Supponiamo di organizzare uno studio caso controllo con 80 casi e 20 controlli e si vuole calcolare la loro esposizione, quindi devo calcolare il numero di soggetti esposti. Voglio calcolare quindi

$$P(E|outP) = \frac{50}{350} = 0.1428$$

$$P(E|outN) = \frac{60}{650} = 0.0923$$

Vado quindi a calcolare il numero di esposti come $80 \cdot 0.1428 = 11.42 \approx 11$ *e si fa la stessa cosa per il controlli, ovvero* $20 \cdot 0.1428 = 1.8461 \approx 2$.
Per differenza si calcolano i non esposti: $80 - 11 = 69$ *per i casi e* $20 - 2 = 18$

per i controlli.
La tabella risultante, quindi, risulta la seguente:

	Outcome POS	**Outcome NEG**	
EXPOSED	**11**	**2**	**13**
NON EXPOSED	**69**	**18**	87
	80	20	
	casi	**controlli**	

Stima

Ovviamente è possibile calcolare le probabilità di outcome positivo per esposti e non esposti:

$$P(outP|E) = \frac{11}{13} = 0.85 \quad P(outP|NE) = \frac{69}{87} = 0.79$$

$$RR_s = \frac{P(outP|E)}{P(outP|NE)} = 1.07 \quad \text{RISK DIFF}_\text{s} = P(outP|E) - P(outP|NE) = 0.05305$$

L'errore che si fa nello stimare queste probabilità campionando dall'outcome, ci porta a due valori che sono estremamente diversi dai valori ottenuti in popolazione (sono più del doppio). Quindi non bisogna calcolare mai il rischio di outcome (positivo o negativo) partendo da un caso controllo.

Usiamo questo esempio per introdurre il concetto di ***odds ratio****. L'odds ratio è calcolato nel seguente modo:*

$$\text{odds ratio} = \frac{\text{E in eventi}}{\text{E in non eventi}} = \frac{\text{E in out P}}{\text{E in out N}}$$

*L'**odd** è definito come il rapporto tra l'odd di esposizione in outcome positivo, ovvero la probabilità di essere esposto sapendo di avere un outcome positivo e 1 meno tale probabilità, ovvero:*

$$\text{odd} = \frac{P(E|outP)}{1 - P(E|outP)}$$

Calcoliamo $P(E|outP) = 0.1375$, $P(E|outN) = 0.100$, *quindi*

$$\text{odd}_\text{P} = \frac{P(E|outP)}{1 - P(E|outP)} = 0.1594 \quad \text{odd}_\text{N} = \frac{P(E|outN)}{1 - P(E|outN)} = 0.1111$$

*L'**odds ratio**, in definitiva, è il rapporto in ratio tra i due odds, ovvero:*

$$\text{odds ratio} = \frac{\text{odd}_\text{P}}{\text{odd}_\text{N}} = 1.43478$$

*Questo numero è il rapporto di due numeri con dubbia interpretazione. Si dimostra formalmente che l'**odds ratio** è una stima buona del rischio relativo che si*

avrebbe se invece di uno studio caso controllo si fosse fatto uno studio che si presta alla stima del rischio relativo, ovvero è una buona stima del rischio relativo. Questo è confermato dal fatto che 1.43 è molto più vicino al rischio relativo=1.35 (popolazione) di quanto invece non lo sia 1.07 (stima).
L'odds ratio è l'unica quantità utile che si può evincere da un caso controllo che può stimare il rischio relativo.

3 Articolo Women and Birth

è importante conoscere cosa dice l'intero articolo
clicca qui per scaricarlo

Il fattore di esposizione, secondo questo articolo, è l'ospedalizzazione. L'**obiettivo** di questo studio è stato cercare di capire se esiste una relazione tra il momento di inizio dell'ospedalizzazione di una donna che ha il travaglio cercando di capire come questo fattore di esposizione abbia un impatto sia sulla tipologia di parto che sugli eventuali interventi ultra-parto. L'obiettivo è anche di studiare una causalità, ovvero l'idea è che l'inizio di ospedalizzazione possa avere un impatto su come andrà il parto e su eventuali interventi ultra-parto.
Questo si classifica come **studio di coorte** in quanto il campione è un'intera sottopopolazione che afferisce ad un particolare centro, e questa scelta viene fatta in blocco basandosi su criteri di reggibilità, non viene fatta rispetto all'outcome e al momento dell'identificazione si può fare una classificazione rispetto ad esposti/non esposti e poi serve tempo per valutare il manifestarsi dell'outcome. Per sua natura il follow up di studi di questo tipo è estremamente breve. Per semplificare, si può pensare che questo sia uno studio cross sectional, ma è anche vero che è un po' atipico perchè non ha quel dubbio di fondo che è quello di non avere una certezza sulla direzione della causalità (che in questo caso c'è, ovvero l'ingresso in ospedale può avere impatto sul tipo di parto).

Per quanto riguarda la **popolazione**, si ha la popolazione target e la sample population (1446 donne), c'è una specifica su criteri che queste donne devono soddisfare: criteri del basso e criteri di esclusione.
Una volta che si ha a disposizione la sample population si fa una categorizzazione in esposte alla fase latente (684 donne) ed esposte alla fase attiva (762 donne). Questo è un contesto non randomizzato (perchè i gruppi non sono bilanciati) e non si sa niente sull'outcome.

Per quanto riguarda le **variabili**, sono descritte (nomi, decisioni prese per codifica delle variabili, ...) nel paragrafo 2.3 dell'articolo.

La parte dei **metodi statistici** la si comprende meglio dopo che si sono letti i risultati. Questa parte è presente negli articoli perchè è necessario descrivere i metodi che hanno portato a determinati risultati.
Il **modello di regressione pragmatico** ha come obiettivo andare a vedere l'impatto che l'esposizione all'ospedalizzazione ha sull'intervento intraparto sen-

za tenere conto degli interventi intraparto, mentre il **modello di regressione allargato** considererà anche l'impatto degli interventi intraparto, quindi mostrerà che l'ospedalizzazione precoce ha impatto sull'intervento intraparto e che l'intervento intraparto ha a sua volta impatto su altro.

4 Articolo Worls Health Organization reference values for human semen characteristics

è importante conoscere cosa dice l'intero articolo
clicca qui per scaricarlo

Questo articolo ha come **obiettivo** definire degli intervalli di riferimento per i parametri che si utilizzano per definire la fertilità dell'uomo. In altre parole l'obiettivo è individuare dei *cut points* per l'anormalità di certi parametri. Questo è uno **studio descrittivo osservazionale**, ovvero non c'è esposizione e non c'è una comparazione come domanda cardine.

Sono state considerate due **popolazioni**: una è quella degli uomini fertili sicuramente, uomini con fertilitià sconosciuta e uomini selezionati fertili rispetto ai criteri antecedenti a questo studio e l'altra è quella degli uomini che ci hanno impiegato meno di 12 mesi per raggiungere il concepimento. Quest'ultima è stata scelta come popolazione di riferimento per costruire gli intervalli di reference per quanto riguarda i vari parametri. Si può fare un'osservazione importante sul fatto che tra le varie popolazioni disponibili venga scelta proprio quella degli uomini fertili; in relazione all'obiettivo dello studio può essere criticabile questa scelta. Potrebbe essere sensato utilizzare come popolazione target un mix di uomini fertili e non fertili dando un cut point per età. In questo caso, invece, verranno inviduati i meno fertili tra i fertili e questo verrà utilizzato per mettere "campanelli di allarme" sui parametri quando un uomo fa controlli per la sua fertilità.

Nei **risultati** verranno dati dei cut points inferiori in modo che se il parametro scende sotto il cut point si avrà un "campanello di allarme".

5 Sintesi e rappresentazione grafica dei dati

5.1 Distribuzioni di frequenza

vedere gli esempi sulle slide

Una **distribuzione di frequenza** è una tabella che è utile ad organizzare i riepilogare i dati quando si lavora con set di dati di grandi dimensioni. Una **tabella di frequenza** dove in una prima colonna c'è un elenco di valori/modalità che la nostra variabile può assumere; subito a destra si avrà il conteggio del numero di volte in cui quella variabile è stato osservato; questo prende il nome di *frequenza assoluta*. Ogni volta che si fa una tabella di frequenza raggruppata in

classi è importante decidere come costruire le classi e si deve specificare se gli estremi sono inclusi o non inclusi. Una buona regola è creare classi della stessa dimensione/cardinalità.
Gli estremi inferiori degli intervalli vengono chiamati **limiti inferiori**, analogamente per il **limite superiore**. Questo vale solamente se gli estremi sono *inclusi*, in quanto altrimenti il limite superiore non appartiene alla classe. Le classi sono anche caratterizzate da una **ampiezza**; per ottenere l'ampiezza si fa

$$\text{estr dx} \; - \; \text{estr sx} + 1$$

Ciascuna classe ha anche un proprio **valore centrale**, ad esempio l'intervallo $[60, 69]$ ha come valore centrale 64.5 ottenuto come $(60 + 69)/2$. Questo è considerato il *rappresentante* della classe e ha un vero significato solamente se si ha ben chiara l'ampiezza della classe. Una cosa interessante è creare la tabella di frequenza con le *frequenze relative percentuali*; queste non sono altro che le percentuali corrispondenti alle varie frequenze assolute.

Esempio 3. *Supponendo che per l'intervallo [60,69] si ha come frequenza assoluta 12 e supponendo che la frequenza totale sia 40, per ottenere la frequenza relativa percentuale si dovrà fare il seguente calcolo:* $(12/40) \cdot 100 = 30\%$

Per ottenere, invece, la vera *frequenza relativa*, devo fare la frazione di soggetti concentrata in una determinata classe, quindi prendendo come riferimento l'esempio precedente, per calcolare la frequenza relativa per l'intervallo $[60, 69]$ si dovrà fare il seguente calcolo: $12/40 = 0.3$.

Si usano più frequentemente le frequenze relative percentuali perchè, per quanto riguarda la comunicazione, una percentuale cattura maggiormente l'attenzione. La somma delle frequenze relative percentuali fa 100 per definizione. Allo stesso modo la somma delle frequenze relative fa 1. La somma delle frequenze assolute, invece, deve risultare pari al numero dei soggetti analizzati.

$$\text{freq relativa} = \frac{\text{freq classe}}{\text{somma freq}}$$

$$\text{freq relat \%} = \frac{\text{freq classe}}{\text{somma freq}} \cdot 100$$

Un'altra tipologia di frequenza, non particolarmente utilizzata, è la **frequenza cumulata**. Questa associa alla classe la conta dei soggetti che appartengono a quella classe oppure alle classi precedenti. Facciamo un esempio:

Rate	**Frequenza**
60-69	12
70-79	14
80-89	11

Rate	**Freq cumulata**
<70	12
<80	26
<90	37

Nell'ultima cella della frequenza cumulata, per definizione, avrò il valore pari al numero dei soggetti totali. Quando analizzo dati raccolti in periodi diversi

potrebbe essere utile fare la tabella delle frequenze cumulate rispetto alla scala temporale (giorni/mesi/anni).
Ovviamente esistono anche le **frequenze cumulate percentuali** calcolate dividendo per 100 il valore della frequenza cumulata ottenuta.

5.2 Istogrammi

Un **istogramma** è un grafico tipico utilizzato per la rappresentazione di dati quantitativi, che ha su scala orizzontale le *classi* dei valori dei dati che stiamo analizzando, mentre in verticale ha le frequenze.

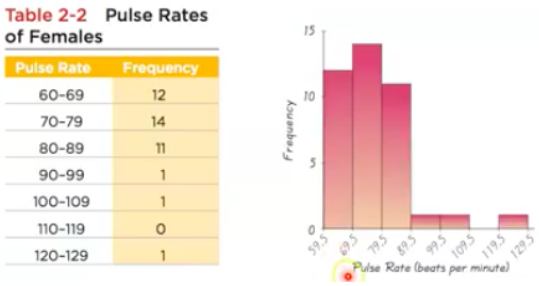

Table 2-2 Pulse Rates of Females

Pulse Rate	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

Figura 2: Esempio di istogramma relativo alle frequenze assolute.

L'istogramma deve avere sempre una partizione dell'asse delle x *disgiuntiva ed esclusiva*, ovvero non ci devono essere "buchi" tra le barre. L'ampiezza delle barre è l'ampiezza dell'intervallo/classe mentre l'altezza è il valore della frequenza.
È possibile fare anche l'istogramma che abbina all'ampiezza della classe un'altezza pari alla frequenza relativa percentuale. Questo tipo di istogramma è molto più utilizzato rispetto al precedente.

5.3 Altri grafici

vedere esempi su slide

Un grafico non particolarmente utilizzato è il **poligono di frequenza**, che cerca di dare un'idea di *andamento* attraverso una "funzione" continua con punti di non derivabilità. È costruito con segmenti che vengono collegati in riferimento alla singola classe in corrispondenza del punto medio della classe e del valore della frequenza assoluta. Non aggiunge niente rispetto all'istogramma. È possibile farlo anche per le frequenze relative e anche per le frequenze relative percentuali.

Un altro grafico utilizzato è il **grafico delle sequenze cumulate - ogiva**, che associa i valori dell'asse delle x i valori di frequenza cumulata. All'estremo di destra della singola classe associa la frequenza cumulata. Per sua natura questo tipo di grafico tenderà a crescere verso l'alto rappresentando una "funzione" non decrescente fino ad arrivare al suo massimo che corrisponde al valore totale dei soggetti coinvolti nello studio.

Un altro grafico è il **dot plot**, cioè il "grafico dei puntini". In caso di variabili discrete (non per variabili continue) si impilano tanti pallini quanti sono i valori osservati. Questo tipo di grafico può essere abbastanza utile per i dati in laboratorio perchè hanno poche osservazioni (pochi pallini) per poter fare un display pieno dei miei dati fornendo, a chi lo guarda, i valori esatti.

Il grafico chiamato **stemplot o diagramma di stelo e foglia** permette di rappresentare dati quantitativi separando i valori in due parti. Questo è più un grafico complicativo veramente poco utilizzato. Questo è un modo alternativo di rendere disponibili i dati.

Il **grafico a barre** è molto utilizzato e ha sull'asse delle x la caratteristica qualitativa/quantitativa discreta che sto analizzando.

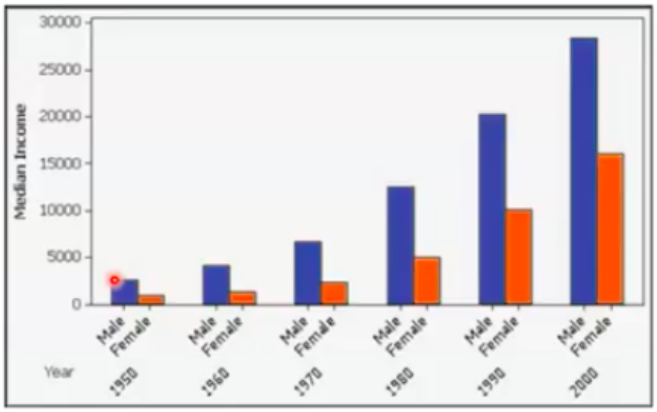

Figura 3: Esempio di grafico a barre.

Un grafico a barre più classico è il **diagramma di Pareto**, dove si rappresenta l'andamento solo di una variabile qualitativa/quantitativa discreta e sull'asse delle x si mettono dei segmenti con testo, mentre sull'asse delle y una particolare percentuale che caratterizza quella particolare variabile. La somma delle altezze delle barre fa 100 per definizione.

Il **grafico a torta** è molto utilizzato per dati qualitativi o per dati quantitativi discreti che assumono poche modalità e ad ogni modalità si associa una fetta di torta relativa alla sua frequenza. Ha senso mantenere un ordinamento delle modalità nelle varie fette di torta. Questo tipo di grafico non aggiunge nessuna informazione rispetto ad una tabella di frequenza.

Un altro tipo di grafico è il **grafico a dispersione/nuvola o scatter plot** ed è usato per rappresentare la relazione tra due variabili. Si rappresentano le osservazioni con un singolo punto che ha coordinate in asse delle x e in asse delle y. Più i punti sono disordinati e più si evince che tra le due variabili non c'è relazione.

Per dati di serie temporali si usa il **grafico delle serie temporali**. Non si presta a dati di laboratorio.

Un'altra tipologia di grafico sono i **pictogram**, grafici con oggetti tridimensionali. Questi grafici possono creare delle false impressioni perchè distorciono i dati e perchè non vengono fatti rispettando certe regole. Se lavoro con un grafico tridimensionale dovrei variare i volumi per spostarmi da una frequenza all'altra, quindi non vengono mantenute le proporzioni. È meglio non utilizzare questo tipo di grafico per evitare distorsioni agli occhi di chi legge.

5.4 Principi importanti

- per piccoli dataset si può decidere sempre di utilizzare una tabella. Un difetto dei dati di laboratorio è forzare a fare grafici quando invece il modo più semplice di renderli è di fare tabelle. I grafici vanno fatti se è richiesta una sintesi che semplifica un insieme di dati; questi devono permettere di concentrarsi sulla vera natura dei dati.
- *"Quasi tutto l'inchiostro in un grafico dovrebbe essere utilizzato per i dati, non per gli altri elementi di progettazione."*
- non usare aree e volumi per dati che sono di natura unidimensionale.
- non pubblicare mai grafici a torta perchè non hanno una scala appropriata. L'unica scala di un grafico a torta è un angolo, che rischia di assumere un peso diverso e non dà niente di più rispetto ad una tabella di frequenza.

6 Statistiche per la descrizione, l'esplorazione e il confronto dei dati

6.1 Misure di centro

Le **misure di centro** sono delle misure che sintetizzano l'ordine di grandezza di un fenomeno.
La misura di centro più nota è la **media aritmetica**, calcolata come

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

e tale valore si posizionerà in mezzo ad un insieme di dati (ad esempio rappresentati graficamente). La media si calcola solo per variabili quantitative di tipo continuo o discreto.
I *vantaggi* sono che è un indicatore affidabile che prende in considerazione tutti i dati che si stanno sintetizzando. Questo può essere visto anche come uno *svantaggio*, perchè la media aritmetica può essere molto molto influenzata a valori estremi, appunto perchè è sensibile a tutti i dati che si stanno analizzando.

Un altro indicatore di posizione è la **mediana**. Anche questo dà l'idea dell'ordine di grandezza di un dato fenomeno. È calcolata mettendo i dati in ordine *crescente* e si va a selezionare l'osservazione che sta *in centro* nella successione ordinata dei dati ed è indicata con $\widetilde{x}$. A differenza della media, la mediana *non* è influenzata dai valori estremi. Questo è sia un vantaggio che uno svantaggio: non si sta dando peso al fatto che ci sono delle grandi eterogeneità.
Si predilige la mediana piuttosto che la media aritmetica nel caso in cui l'istogramma corrispondente ai valori da analizzare sia fortemente asimmetrico oppure nel caso si debbano analizzare pochi dati.
Se il totale dei miei valori è un numero dispari, la mediana è il valore esattamente in centro; se il totale dei miei valori è un numero pari, la mediana è calcolata come la media dei due valori centrali.

Un'altro indicatore usato per variabili qualitative o qualitative è la **moda**, che indica il valore che si verifica con massima frequenza (qualsiasi tipo di frequenza). *"La moda è il valore che va più di moda"*. È interessante osservare come il valore di moda possa avere una frequenza che *non* supera il 50%. Esistono situazioni in cui si possono avere più di una moda o nessuna moda.

Il **midrange** è il punto di mezzo di un segmento che si ottiene unendo il valore minimo e il valore massimo. In altre parole si calcola facendo la media aritmetica tra il valore minimo e il valore massimo. Questo, dunque, è calcolabile solo per valori quantitativi, ma è poco utilizzato. Questo ha uno grandissimo *svantaggio* ovvero che è altamente sensibile a dati estremi in quanto è calcolato solo col minimo e col massimo.

Sono importanti anche il concetto di **simmetria e asimmetria** di una variabile quantitativa continua. Immaginiamo di vedere un esempio con delle funzioni (raffigurate in blu) che approssimano l'andamento di un istogramma di un dato fenomeno:

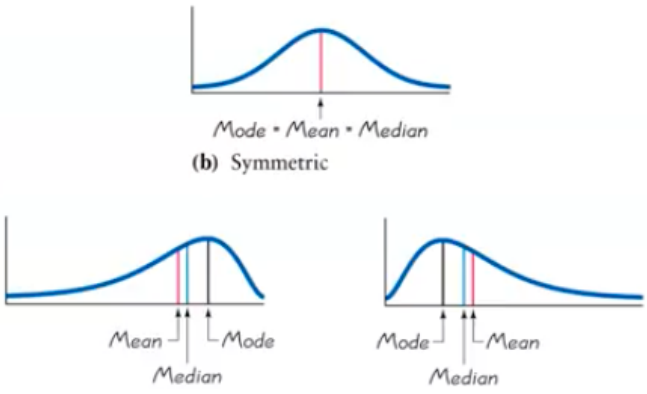

Figura 4: Esempio di simmetria (immagine in alto) e asimmetria (immagini in basso).

Nel caso di un **istogramma simmetrico** la media, moda e mediana coincidono. Nell'**istogramma asimmetrico**, invece, questi tre indicatori *non* coinci-

dono. Nell'immagine (**figura 4**) precedente descrive a sinistra una asimmetria *negativa*, mentre a destra una asimmetria *positiva*.

6.2 Misure di variazione

Attorno ad un unico valore che ha la pretesa di dare l'ordine di grandezza di un fenomeno, in realtà si hanno tutti gli altri dati che ruotano attorno a questo valore. È utile, quindi, dire quanta eterogeneità c'è attorno a quel dato valore. Per farlo rende molto bene l'esempio dell'insieme di voti sul libretto (considerati per semplificità senza cfu). Se si ha una media aritmetica pari a 26 questa è pienamente compatibile sia con voti pari a 26 sia con variabilità, ad esempio è compatibile sia con 20 che con 30. Lo scopo è capire come misurare questa variabilità.

Un modo molto semplice di misurare questa variabilità è il **range**, che è definito come la distanza tra il valore massimo e il valore minimo:

$$\text{range} = \text{v}_{\text{max}} - \text{v}_{\text{min}}$$

Questo è un indicatore "rozzo" perchè risente dei difetti che avrebbe il midrange nell'utilizzarlo come misura di centro, ed è dipendente da due valori che per loro natura sono valori estremi.

Un indicatore molto utilizzato per istogrammi abbastanza simmetrici dove la media aritmetica è un buon valore di centro è la **deviazione standard**, definita anche col nome di scarto quadratico medio. È calcolata nel seguente modo

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Questa è una somma molto rappresentativa dell'eterogeneità che i dati hanno con la media aritmetica, ovvero equivale all'eterogeneità che i dati hanno *tra loro*. Modificando leggermente la formula, si ha la seguente formula modificata:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

. Questa è una media delle distanze al quadrato, che rappresenta quanto più c'è variabilità tra i voti sul libretto.
La deviazione standard ha diverse **proprietà**:

- coinvolge tutti i dati
- è positiva
- può aumentare notevolmente con l'inclusione di uno o più valori anomali, ovvero valori molto lontani da tutti gli altri
- ha la stessa unità di misura dei valori dei dati originali

Esiste un modo di usare sia la media aritmetica che la deviazione standard per ottenere degli *intervalli* che in modo approssimato comprendono al loro interno delle percentuali approssimate ma fisse che tengono abbastanza bene delle percentuali di dati. Questo è l'uso veramente interessante della media e della deviazione standard.

La regola empirica

Avendo un istogramma con aree interpretabili come sequenze relative, si ha complessivamente questa situazione:

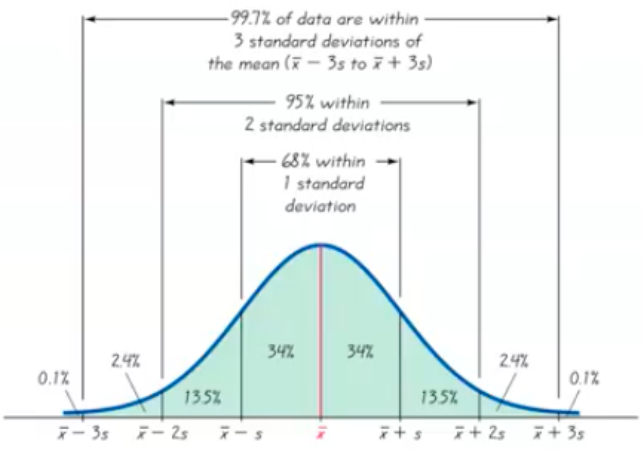

Figura 5

L'ultimo intervallo ci permette di capire come il fenomeno scorre complessivamente un intervallo di 6 deviazioni standard sull'asse delle x.

Un altro concetto molto importante è il concetto di **percentile**, che sono delle particolari misure di posizione che ci danno l'idea dell'andamento di una certa misura quantitativa continua. Sono 99 e sono denotati con

$$P_1,\ P_2,\ \ldots\ P_{99},\ \text{in generale } P_\alpha$$

La *mediana* viene anche chiamata il **50esimo percentile**, ed è $P_\alpha = P_0.5$ con $\alpha\%$ a sinistra (valori inferiori a P_α) e $(1-\alpha)\%$ a destra (valori superiori a P_α). Questo concetto può essere generalizzato a degli α diversi da 0.5, ad esempio $P_1 = P_{0.01}$ è il valore che lascia a sinistra lo 1% dei soggetti e a destra il 99% dei soggetti. I percentili dividono sempre *in 2* l'andamento di un fenomeno.
Esistono altri percentili particolare che, insieme alla mediana, prendono il nome di **quartili**. I quartili sono 3 e si chiamano così perchè ci permettono di costruire 4 intervalli. Sono chiamati Q_1 *primo quartile*, Q_2 *secondo quartile* e Q_3 *terzo quartile*. Questi quartili sono i *percentili*, rispettivamente, con un $\alpha\%$ pari a 25, 50 e 75. Quindi

- $Q_1 = P_{25} = P_{0.25}$, ovvero lascia il 25% delle osservazioni alla sua sinistra e il 75% alla sua destra

- $Q_2 = P_{50} = P_{0.5}$, ovvero lascia il 50% delle osservazioni alla sua sinistra e il 50% alla sua destra
- $Q_3 = P_{75} = P_{0.75}$, ovvero lascia il 75% delle osservazioni alla sua sinistra e il 25% alla sua destra

I quartili sono *3 percentili tipo* che permettono la costruzione di 4 macro intervalli:

Figura 6

I tre quartili utilizzati spesso per piccoli insiemi di dati (ad esempio dati di laboratorio) oppure per grandi insiemi di dati fortemente asimmetrici, si possono rappresentare graficamente tramite un grafico chiamato **boxplot**, anche chiamato *grafico a scatola e baffi*. Permette di rappresentare i tre quartili e il minimo e il massimo per dare una visualizzazione grafica dei 4 intervalli. Spesso il boxplot è rappresentato in verticale.

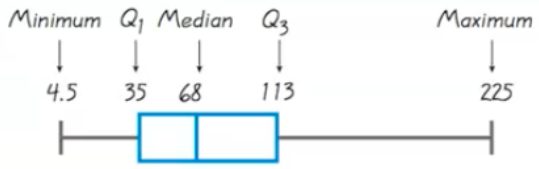

Figura 7: Esempio di boxplot.

Spesso in presenza di valori estremi, chiamati **outlier**, esiste un modo per metterli in luce definendo cosa sia un "valore estremo"

Definizione 1 (**outlier**). *Un valore è estremo o outlier se*

- *è superiore al terzo quartile* Q_3 *di un valore numerico maggiore di* $1.5 \cdot IQR$
- *è inferiore al primo quartile* Q_1 *di un valore numerico maggiore di* $1.5 \cdot IQR$

IQR è la differenza tra il terzo e il primo quartile, cioè $Q_3 - Q_1$ *(ampiezza della scatola)*

Gli outlier sono rappresentati tramite dei punti singoli, come mostrato nella figura di seguito:

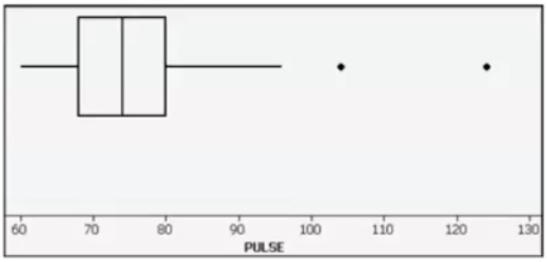

Figura 8: Esempio boxplot con outlier.

6.3 Revisione e anteprima

non ha detto niente riguardo a questa parte