FASI DI UN'INDAGINE STATISTICA: acquisizione dei dati, analisi dei dati, rappresentazione grafica dei dati - INDICI STATISTICI: media, moda, mediana, varianza

Prof. Pierangelo Veltri – 27/09/2023- Autori: Calisto, Panarello - Revisionatori: Panarello, Calisto

ESPERIMENTI (ACQUISIZIONE DEI DATI)

In statistica medica si esegue un'acquisizione dei dati per effettuare meccanismi di analisi e validazione. Per quanto riguarda l'acquisizione, può essere effettuata in diversi modi:

- ***A posteriori:*** recuperare le informazioni successivamente agli eventi.
- ***In parallelo***: i dati vengono acquisiti contemporaneamente allo svolgimento dell'esperimento.
- ***Studio osservazionale***: le misurazioni vengono effettuate senza intervenire (coinvolgere) sui soggetti.
- ***Esperimenti che riducono la presenza di variabili nascoste.***

Quando si esegue un'acquisizione dei dati bisogna fare attenzione a fenomeni "particolari", come quelli legati alla presenza di informazioni che apparentemente possono sembrare non significative, ma in realtà lo sono: sono le cosiddette *variabili nascoste*, che potrebbero avere un peso significativo non solo nell'acquisizione del dato ma anche nelle fasi successive dell'analisi.

STUDIO OSSERVAZIONALE

Lo studio osservazionale (*Figura 1.3*) osserva e misura ma non agisce sui soggetti. Le osservazioni possono essere effettuate: in un *intervallo temporale passato*, in un *istante preciso*, in un *intervallo temporale futuro.*

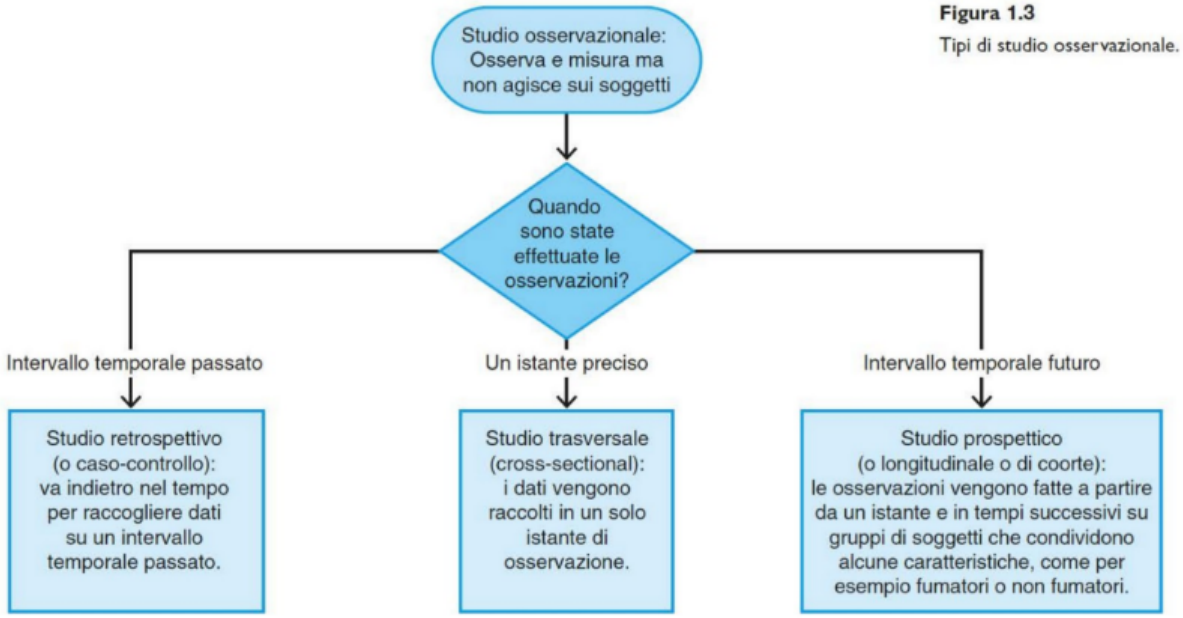


Figura 1.3
Tipi di studio osservazionale.

ALTRI TIPI DI ESPERIMENTO

Quando si parla di tipo di esperimento ci riferiamo all'acquisizione del dato, che avviene in diversi modi:

- ***Cieco***: il soggetto (paziente) non conosce appartenenza;
- ***Doppio cieco***: medico e paziente non conoscono la loro appartenenza al gruppo;
- ***Ripetizione***: esperimento condotto da più individui;
- ***Randomizzazione:*** soggetti assegnati a gruppi diversi un modo randomatico.

CAMPIONAMENTO

In generale, il campionamento si riferisce al processo di selezione di un sottoinsieme. Esso può essere distinto in:

- ***Campionamento casuale:*** ogni individuo della popolazione ha la stessa probabilità di essere scelto. Spesso si fa ricorso ad algoritmi per generare numeri telefonici casuali.
- ***Campionamento casuale semplice:*** viene estratto un campione di dimensione *n*, in modo che ogni campione di dimensione *n* abbia la stessa probabilità di essere scelto.
- ***Campionamento sistemico:*** si sceglie un punto di partenza, da cui viene selezionato ogni *k-esimo* elemento (per esempio, ogni 50esimo) della popolazione.
- ***Campionamento di convenienza:*** si usano dati di facile raccolta.
- ***Campionamento stratificato:*** si suddivide la popolazione in almeno due diversi sottogruppi (o strati), quindi viene estratto un campione da ciascun sottoinsieme.

ANALISI DEI DATI

Dopo la raccolta dei dati, si procede all'analisi statistica, in cui vengono letti i valori e si risponde a domande relative all'analisi di nostro interesse; il fine è quello di ottenere una rappresentazione della distribuzione di frequenza del dato (rappresentazione oggettiva che può essere, ad esempio, su un calcolo percentuale).

NB. statistica inferenziale: generazione sull'intera popolazione dalla quale sono estratti i dati.

RAPPRESENTAZIONE DEI DATI

Una forma di distribuzione dei dati è l'***istogramma***. Un istogramma è un grafico a barre che riporta sull'asse orizzontale le classi in cui sono stati suddivisi i dati e sull'asse verticale le frequenze. L'altezza delle barre corrisponde ai valori delle frequenze e le barre sono adiacenti le une alle altre. Si veda il seguente esempio relativo ai punteggi di QI di un gruppo di individui esposti a basso livello di piombo.

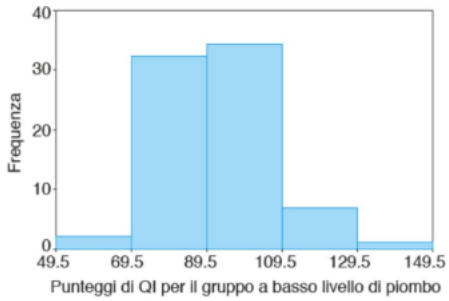


Figura 2.2
stogramma.

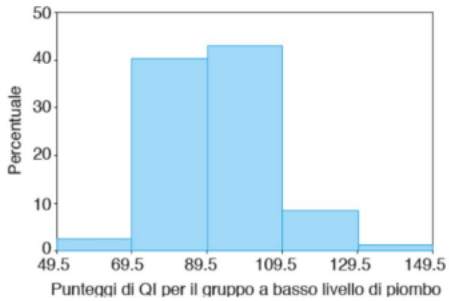
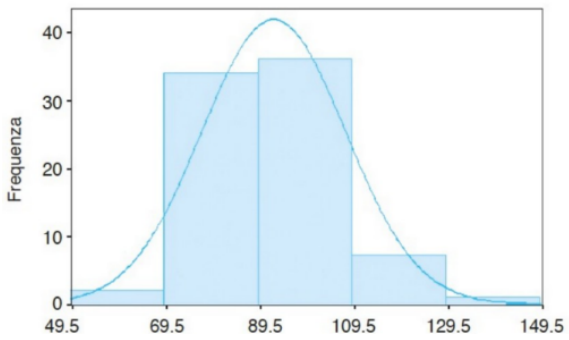


Figura 2.3
Istogramma delle frequenze relative.

Forme rilevanti di distribuzione dei dati

Esistono altre distribuzioni di cui le più note sono quelle in cui sull'asse delle ordinate colloco le frequenze e sull'asse delle ascisse colloco i dati di studio (il campione di studio). Il tipo di grafico utilizzato dipende dall'intervallo campionario.



- *Grafico a campana e distribuzione di Gauss*

Un "grafico a campana" o "istogramma a campana" è un tipo di grafico utilizzato nella statistica per visualizzare la distribuzione di una variabile numerica continua. Questo tipo di grafico è anche noto come "istogramma

di Gauss" o "istogramma normale" perché spesso rappresenta una distribuzione normale o gaussiana, che ha una forma a campana.

Nella distribuzione a campana individuiamo due code agli estremi del campo (estremi della distribuzione) che rappresentano una piccola parte della popolazione.

Esempio: se abbiamo una popolazione con soggetti per la gran parte sani, un ragazzo di 22 anni con un infarto miocardico si collocherà agli estremi del campo (code).

- *Grafico dei quartili*

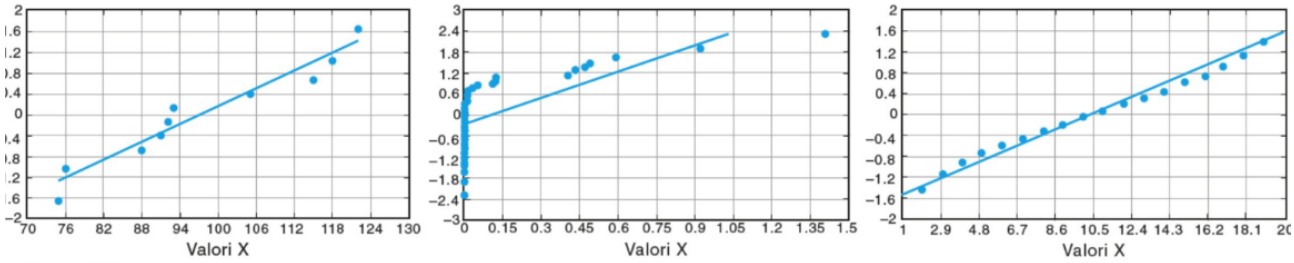


Figura 2.6

Distribuzione normale: I punti sono ragionevolmente vicini alla retta e non si osservano deviazioni sistematiche che non siano lineari.

Distribuzione non-normale: i punti non sono vicini alla retta.

Distribuzione non-normale: I punti mostrano un andamento sistematico e non lineare.

- *Dot plot* (rappresentazione a punti)

Figura 2.7
Dotplot della frequenza cardiaca di un campione di individui maschi.

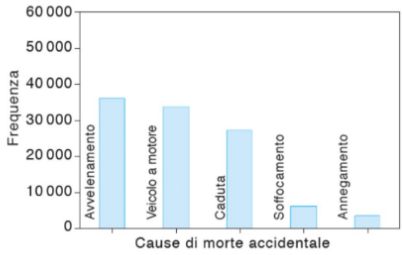


Figura 2.8
Dotplot della frequenza cardiaca di un campione di individui maschi.

- *Grafico a torta*

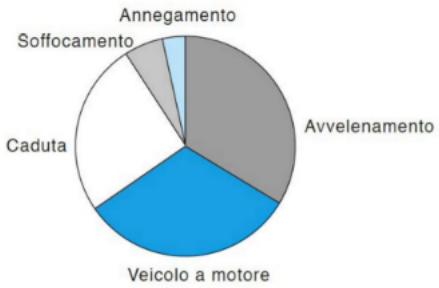


Figura 2.9
Grafico a torta per le cause di morte accidentale

Scatterplot osservazioni con correlazione tra oggetti (si parla di campioni bi-variati e si introduce il concetto di correlazione)

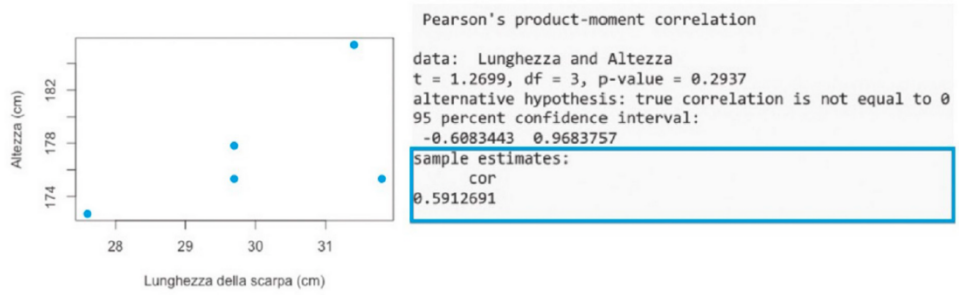
SCATTERPLOT E CORRELAZIONE

In caso di campioni bivariati (di cui si raccolgono coppie di variabili, ad esempio peso ed altezza) si introduce il concetto di **correlazione**, che viene rappresentata tramite un grafico detto **scatterplot**.

- Esiste correlazione tra due variabili quando i valori di una variabile sono in qualche modo associati ai valori dell'altra variabile.
- Esiste correlazione lineare tra due variabili quando c'è correlazione e i punti corrispondenti alle coppie di osservazioni disegnano una "nuvola" che può essere ben descritta da una retta. In questo caso si ricerca il cosiddetto *coefficiente di correlazione lineare*.
- Il coefficiente di correlazione lineare è in genere indicato con r e misura la forza dell'associazione lineare tra due variabili.
- Uno scatterplot è un grafico per dati bivariati quantitativi (x,y) con un asse orizzontale e un asse verticale. L'asse orizzontale è usato per la prima variabile (x) e l'asse verticale per la seconda variabile (y).

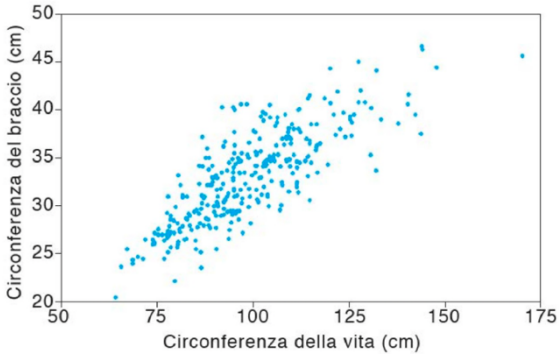
Tabella 2.6 Lunghezza della scarpa e altezza in individui maschi.

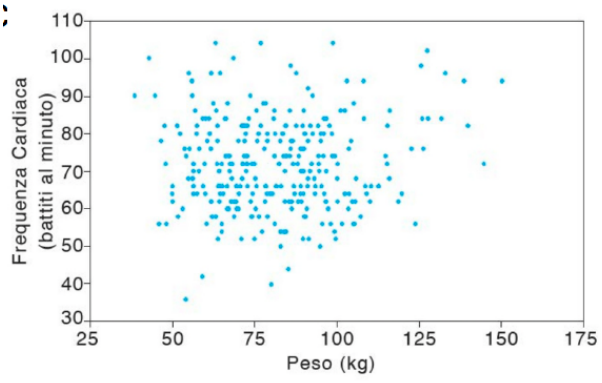
Lunghezza	29.7	29.7	31.4	31.8	27.6
Altezza	175.3	177.8	185.4	175.3	172.7



<u>Correlazione implica relazione ma non necessariamente causalità.</u>

Ad esempio, nello scatterplot a destra si osserva uno studio sulla correlazione tra la dimensione della vita e quella del braccio. In particolare, la forma approssimativamente rettilinea della nuvola di punti suggerisce che ci sia correlazione tra le due variabili in esame.





Nell'esempio a sinistra, invece, in cui si ricerca nel campione la correlazione tra peso e frequenza cardiaca, si osserva che la nuvola di punti non mostra una forma chiara e suggerisce che non ci sia correlazione tra il peso e le pulsazioni.

REGRESSIONE

Data una collezione di dati appaiati, la **retta di regressione** (o **retta di migliore approssimazione o retta dei minimi quadrati**) è la retta che "meglio" approssima lo scatterplot dei dati. Il criterio per determinare tale approssimazione è il criterio dei minimi quadrati.
Il **metodo dei minimi quadrati** (in inglese OLS: Ordinary Least Squares) è una tecnica di ottimizzazione (o regressione) che permette di trovare una funzione, rappresentata da una curva ottima (o curva di regressione), che si avvicini il più possibile ad un insieme di dati (tipicamente punti del piano). In particolare, la funzione trovata deve essere quella che minimizza la somma dei quadrati delle distanze tra i dati osservati e quelli della curva che rappresenta la funzione stessa. L'utilizzo più frequente è la deduzione dell'andamento medio in base ai dati sperimentali per l'estrapolazione fuori dal campo di misurazione.

Un esempio in ambito medico che prevede il calcolo di tale retta è la gestione del numero di posti letto in una struttura ospedaliera. In particolare, si calcola la retta di regressione e il collegamento (correlazione) tra numero di posti letto e numero di casi. Il grafico che rappresenta le due variabili singolarmente risulta di tipo esponenziale (predizione esponenziale); costruendo un grafico che ha come ascisse il numero di casi d'infezioni e nelle ordinate il numero di posti letto disponibili, si ottiene un grafico con andamento non esponenziale che permette di stimare approssimativamente (predire) il numero di posti letto disponibili ed infetti in un dato giorno.

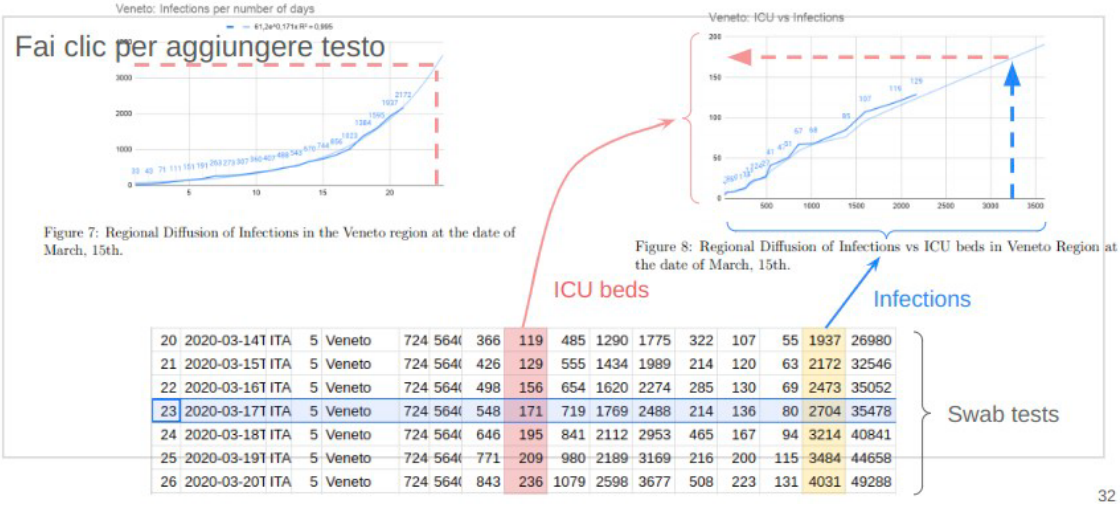


Figure 7: Regional Diffusion of Infections in the Veneto region at the date of March, 15th.

Figure 8: Regional Diffusion of Infections vs ICU beds in Veneto Region at the date of March, 15th.

20	2020-03-14T	ITA	5	Veneto	724	564(	366	119	485	1290	1775	322	107	55	1937	26980
21	2020-03-15T	ITA	5	Veneto	724	564(	426	129	555	1434	1989	214	120	63	2172	32546
22	2020-03-16T	ITA	5	Veneto	724	564(	498	156	654	1620	2274	285	130	69	2473	35052
23	2020-03-17T	ITA	5	Veneto	724	564(	548	171	719	1769	2488	214	136	80	2704	35478
24	2020-03-18T	ITA	5	Veneto	724	564(	646	195	841	2112	2953	465	167	94	3214	40841
25	2020-03-19T	ITA	5	Veneto	724	564(	771	209	980	2189	3169	216	200	115	3484	44658
26	2020-03-20T	ITA	5	Veneto	724	564(	843	236	1079	2598	3677	508	223	131	4031	49288

32

RED CAP

Sistema che consente di costruire un ambiente per l'acquisizione dei dati condivisibile e facilmente consultabile tramite codici di accesso. Tale sistema agevola la fase di acquisizione dei dati, di fondamentale importanza e relativamente complessa in quanto necessita tempo, risorse ed un campione adatto e disponibile.
LINK REDCAP: https://projectredcap.org/wp-content/themes/rcap/map/map_fullscreen.php

MISURE DI CENTRALITÀ

Non sempre i dati del dataset sono correlabili tramite coefficiente di correlazione. Un esempio è la misura della frequenza cardiaca nell'uomo e nella donna: per ragioni genetiche ed anatomiche, tali variabili risultano differenti in condizioni analoghe (esempio stessa età, stesso stile di vita, ecc.). In questi casi si utilizzano delle misure dette di centralità.

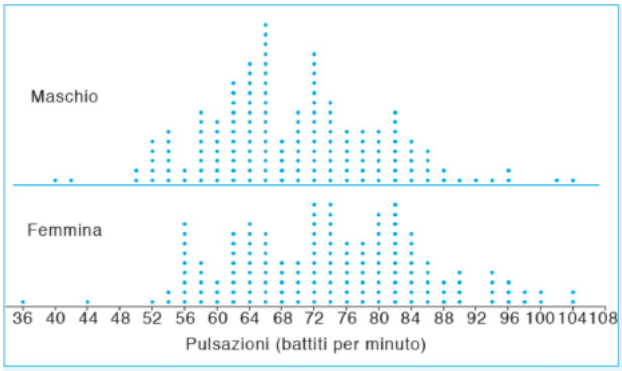


Figura 2.15
Dotplot della frequenza cardiaca in soggetti maschili e femminili.

Una **misura di centralità** è un valore che si trova al centro dei dataset e servono ad indicare un valore rappresentativo al centro dei dati. Esse sono:

- **Media aritmetica**: misura ottenuta sommando tra loro tutti i valori e dividendo poi il totale per il numero di tali valori. Spesso viene chiamata semplicemente *media* e si indica con il simbolo $\bar{x}$.
- **Media ponderata:** si ottiene moltiplicando ogni valore per il suo punteggio associato, sommando tali prodotti e dividendo il risultato per la somma dei pesi.
- **Midrange:** è costituita dal valore posto a metà tra il minimo e il massimo dei dati; viene determinato sommando il massimo al minimo e dividendo il risultato per 2.
- **Mediana:** è il valore centrale della distribuzione e può corrispondere alla media quando i valori sono distribuiti in ordine crescente.
- **Moda:** è il valore con frequenza più elevata. Se vi sono due differenti valori con la stessa frequenza massima, sono entrambi mode e la distribuzione è detta **bimodale**. Se tali valori sono più di due si dice distribuzione **multimodale**. Se nessun valore è ripetuto si dice che non vi è moda.

MISURE DI VARIAZIONE

Un indice di variabilità rispetto ad un centro misura la presenza o meno di una certa stabilità dei valori assunti dalle unità rispetto alla misura di tendenza centrale. Gli indici maggiormente diffusi si basano sul concetto di "scarto" (o scostamento) delle variabili rispetto alla media (intesa come media aritmetica). Tali indici, dette **misure di variazione**, sono 5:

- **Range (campo di variazione):** è il più semplice e il più immediato indice di variabilità e si calcola sottraendo dal punteggio più alto il punteggio più piccolo.
- **Differenza interquartile:** viene calcolata semplicemente sottraendo il punteggio corrispondente alla posizione del venticinquesimo percentile (denominato primo quartile o Q1) dal punteggio corrispondente alla posizione del settantacinquesimo percentile (il terzo quartile o Q3).
- **Deviazione media:** somma degli scarti di ogni punteggio dalla media divisa N (numero di valori) , dove lo scarto è la differenza tra il valore stesso e la media aritmetica della distribuzione.
- **Varianza:** somma degli scarti della media al quadrato diviso N.
- **Deviazione standard (scarto quadratico medio):** misura la variazione di tutti i valori rispetto alla media aritmetica; maggiore è la deviazione standard, maggiore è la variazione dei dati. La deviazione standard aumenta se vi sono outliers (valori fuori dal range del dataset). La deviazione standard si calcola con le formule in figura.

Formula 2.3 $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$ deviazione standard campionaria

Formula 2.4 $s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n - 1)}}$ formula utile per il calcolo della deviazione standard campionaria