www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Biostatistica 01
Prof.ssa Clelia Di Serio – 21 ottobre 2022

La statistica ha il compito di dare un linguaggio ai dati.

I dati sono rappresentati come una matrice: per riga rappresentano i record (soggetti), mentre per colonna i campi (caratteristiche). Il dato quindi a livello pratico sono le cartelle cliniche che vengono compilate durante la giornata in reparto o in laboratorio; inizialmente la raccolta dati era di tipo manuale, oggi invece è tutto tecnologico.
Il diario clinico è il diario del paziente e la più grossa difficoltà è l'interpretazione del diario.

L'analisi del dato si suddivide: dall'osservazione del dato del campione e dall'altra parte la popolazione.
L'inferenza è basata sull'induzione e l'induzione è passare dal particolare al generale -> detto ragionamento inferenziale.
La forza della mia ricerca fatta su un campione è tanto maggiore quanto più questo campione può somigliare veramente alla popolazione in generale.
La rappresentatività è la somiglianza non sempre è un campione rappresentativo, ma con la probabilità possiamo esprimere quanto sia simile alla popolazione di riferimento.

La statistica descrittiva descrive le caratteristiche del campione cercando di sintetizzare ed infine cerchiamo di trarre delle conclusioni; la statistica descrittiva la dividiamo: univariata, cioè descrive una caratteristica per volta o bivariata, cioè si mettono in relazione più dati.

Quando ci riferiamo a un campione intendiamo dati campionali -> parliamo di statistiche/indicatori
Quando ci riferiremo alla popolazione si parla di stime -> il dato della popolazione perciò non sarà mai noto, ma si può solo stimare se il campione della popolazione è rappresentativo

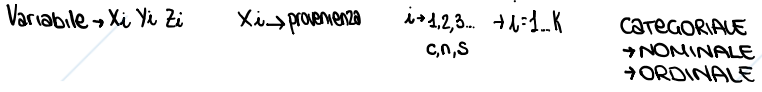
Classificazione di variabili:

- Variabile nominale → categorica → qualitativa —> ovvero una variabile nominale è anche categorica e qualitativa, ma non tutte le variabili qualitative sono nominali.
- Categoriale ordinale (secondo livello) —> variabili che indicano dei livelli in una successione ordinale (es: scala linkert, sono le scale di soddisfazione)
- Scala quantitativa che si distingue in:
    - continua -> l'età è una variabile continua, cioè la posso esprimere con continuità, ma per comodità di calcolo solitamente si considera l'età discreta o può diventare anche categoriale ordinale (es: fascia età)
    - discreta -> es: variabile di conteggio

Gli strumenti di sintesi cambiano a seconda del dato che abbiamo e principalmente sono:

- Grafici -> il tipo di grafico con cui si esprime il dato è fondamentale
- Tabelle -> permettono di sintetizzare il dato
- Indicatori -> sintetizzo l'informazione

Quando abbiamo una caratteristica del soggetto la chiamiamo **variabile** e solitamente si indica con una lettera latina che tipicamente sono: X/Y/Z e con un indicatore (i) che indica l'attribuzione e varia sui livelli.



www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Nel database ci sono due dimensioni:

- n -> numero soggetti, da 1 a n significa che si sta facendo ancora l'indice sui soggetti
- k -> numero categorie, significa che quel "i" va da 1 a k

Le scale quantitative dove X_i è associata al soggetto, quindi si dice che i va da 1 a n ($i \leqslant n$)

La variabile indicatore è una variabile che di solito indica uno stato e quindi devi dare come nome il nome della variabile (es: se la chiami "malato" una persona lo deve essere).

Quando si descrivono le caratteristiche di un dato, quelle caratteristiche le esaminiamo per due aspetti:

- tendenza centrale -> il dato come si colloca
- variabilità -> il dato quanto è variabile

La distribuzione di una variabile statistica è una relazione che associa alla variabile una frequenza, è una relazione in cui ho la variabile X_i con diversi valori (X_1 X_2 X_3 ec) che posso e essere categoriali o quantitativi. La frequenza si chiama n_i se è una frequenza assoluta oppure la chiamerò p_i se è relativa (espressa in percentuale). 1.13

Abbiamo tre misure di tendenza centrale:

- Media aritmetica -> si calcola solo per le variabili quantitative. La media è influenzata molto dai valori estremi
  Formula -> sommatoria : n
- Mediana -> è il valore che mi taglia al centro la distribuzione, mi lascia a destra e a sinistra il 50% e non è influenzata da valori estremi. La mediana la posso calcolare anche per le variabili ordinali (dal più piccolo al più grande)
  Formula: $\frac{n+1}{2}$
- Moda -> è il valore della massima frequenza (indica il picco), si può calcolare su tutte le variabili

Non c'è una tendenza centrale migliore, la soluzione più giusta è tenere in considerazione tutte e tre. Ad esempio quando ho la media uguale alla mediana ho un'informazione di dati distribuiti in modo simmetrico, simmetrico vuol dire che il valore che taglia a metà lascia esattamente 50% più alto e il 50% più basso a destra e sinistra ed è anche il valore medio. Invece, quando una media è più piccola della mediana significa che si hanno dei valori estremi a sinistra che spostano la media verso sinistra, quando è il contrario la curva si sposta a destra.

La variabilità è più importante della centralità.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari