

Prof.CORTEGGIANI

PRIMA LEZIONE

La statistica è una scienza matematica che si occupa della raccolta, classificazione, sintesi ed interpretazione di dati numerici

La statistica biomedica applica i principi della statistica alle scienze biologiche e ai fenomeni legati alla salute degli esseri umani

Per chi ,come noi, lavora per diventare specialista in campo medico, è importante conoscere i principi di statistica per due motivi:

1.Perché siamo consumatori di ricerca.

Dobbiamo essere capaci ed autonomi nell'interpretazione ottimale dei risultati della ricerca. la nostra conoscenza è in evoluzione. La medicina si basa sulle conoscenze da applicare al meglio possibile sul paziente , per salvare più vite e fare meno danni. Per fare ciò bisogna cercare dalla letteratura le indicazioni nuove ed insieme associarle alle conoscenze pratiche .

2 .Perchè siamo co-produttori di ricerca . Dobbiamo essere capaci di collaborare con i colleghi, compresi lo statistico, per migliorare la qualità della progettazione degli studi.

L' informatizzazione ha portato ad avvicinare i professionisti sanitari ad utilizzare sempre più spesso tecniche di analisi statistiche .

Il problema si pone quando, l' applicazione di un test tramite un software nelle mani di un soggetto poco competente ,seppur faciliti l' esecuzione delle analisi, aumenta il rischio di applicazioni scorrette.

EBM - Evidence Based Medicine

È un movimento culturale, che «costituisce un approccio alla pratica clinica dove le decisioni cliniche risultano dall'integrazione tra l'esperienza del medico e l'utilizzo coscienzioso, esplicito e giudizioso delle migliori evidenze scientifiche disponibili, mediate dalle preferenze del paziente». " sapere scegliere il meglio per il paziente."

Saper interpretare e valutare le pubblicazioni

Dobbiamo imparare a tutelarci da interpretazioni fuorvianti delle pubblicazioni

Non sempre articoli pubblicati in riviste prestigiose sono garanzia di buona qualità del lavoro.

Per ovviare a questi possibili confondimenti bisogna saper interpretare correttamente i risultati della ricerca scientifica.

Teoria della comunicazione a doppio senso

Non è nostra responsabilità il processo di analisi statistica. Quest'ultimo è responsabilità dello statistico, persona di riferimento ogni volta che un processo di ricerca deve essere portato a compimento.

Tuttavia, poiché allo statistico manca la conoscenza della gestione clinica delle ipotesi della ricerca, noi lo supporteremo nel versante medico-clinico.

Serve la collaborazione tra più figure. Attenzione non si può fare tutto da soli!!

ASPETTI FONDAMENTALI DI UN PROTOCOLLO DI RICERCA(PDR)

Per produrre un valido protocollo di ricerca bisogna sviluppare con attenzione i sei punti riportati di seguito e in maniera ordinata(senza saltare da un punto all'altro).

1. Il quesito di ricerca
2. Il razionale della ricerca
3. Il disegno dello studio
4. Popolazione di riferimento e campione
5. Metodi di valutazione e misurazione delle variabili
6. Pianificazione dell'analisi statistica

Un PDR nasce da un quesito di ricerca, cioè dal confrontare, cercare associazioni o indagare un aspetto che riguarda la salute dei nostri pazienti.

Il fine ultimo è ottenere dei numeri e da questi interpretare la risposta al quesito di ricerca.

Il processo è circolare, perché i dati finali che abbiamo ricavato pongono le basi per un'altra ipotesi di ricerca, così da aggiungere sempre nuove conoscenze all'argomento.

Leggere un articolo scientifico e analizzarlo tramite questi sei punti standard, con preciso ordine, ci consente di comprendere la qualità del lavoro scientifico.

Teoria GIGO (garbage in, garbage out- spazzatura entra ,spazzatura esce)

Con questa espressione intendiamo che: raccogliere dati in maniera pessima, senza ragionarci o senza criterio, porterà certamente anche il migliore degli statistici ad avere risultati spazzatura. Allo stesso modo sarà vero il contrario ovvero: ottimi dati ma nelle mani di un pessimo statistico daranno risultati spazzatura.

DESCRIZIONE DI STATISTICA DESCRITTIVA/INFERENZIALE

-statistica DESCRITTIVA: offre il metodo per riassumere le informazioni in modo da renderle utilizzabili più facilmente. Riduce i dati sostituendo a molti numeri poche misure. Ovvero descrive un fenomeno misurandolo e sintetizzandolo.

-statistica INFERENZIALE : generalizza le informazioni, ricavando proprietà e leggi generali sulla base di dati rilevati solo su una parte(campione) della popolazione.

Definizioni di elementi di cui la statistica si compone

- ✓ unità statistica(o di analisi) : ogni singolo elemento che appartiene ad una popolazione oggetto di studio.

Ex. Studiare tutti gli ospedali della Sicilia, ogni ospedale è un unità statistica. Studiare specializzandi di primo anno, ogni specializzando è un unità statistica.

- ✓ popolazione: è l'insieme e la totalità delle unità statistiche che voglio studiare.

Ex tutti gli specializzandi di primo anno sono la mia popolazione. Tutti gli ospedali della Sicilia sono la mia popolazione.

- ✓ campione: è una parte rappresentativa della popolazione di cui studio le caratteristiche, sperando di carpire tramite questo le caratteristiche della popolazione generale .

ex. creo il mio campione con gli specializzandi di primo anno che hanno seguito tutte le lezioni di Gregoretti

- ✓ Campionamento : (dalla popolazione al campione).procedura attraverso la quale alcune unità di popolazione vengono selezionate per creare il campione da studiare ed osservare ,oggetto di studio.
- ✓ Inferenza : (dal campione alla popolazione). è il processo di generalizzazione dei risultati ricavati dal campione . Sarà tanto attendibile quanto più le caratteristiche della popolazione, all' interno del campione, saranno rappresentate in uguale proporzione.

La finalità, per la quale procediamo a campionamento ed inferenza , è quella di eseguire osservazioni, rilevare misure e ottenere risultati dal campione selezionato per trarre conclusioni valide per l'intera popolazione di riferimento.

- ✓ Errori di campionamento(EDC)

sono la differenza che c'è tra il risultato ottenuto dal campione e la vera caratteristica della popolazione.

Il campione deve avere delle caratteristiche e modalità ben precise così da essere il più rappresentativo possibile dell'intera popolazione, al fine di ridurre al minimo gli errori di campionamento. Quest'ultimo non può essere calcolato con esattezza ma può essere STIMATO.

- ✓ Stimare è calcolare la probabilità che il risultato del campione sia effettivamente quello della popolazione. Maggiore sarà il campione e maggiore sarà la probabilità che il risultato sia vero.

L'EDC si verifica per :

1. Variazione casuale

Dovuto al CASO. Da questo errore deriva un buon campionamento, perché può essere stimato

2. Selezione viziata del campione

In questo caso il campione è DISTORTO o BIASATO. Da questo ne deriva un cattivo campione, perché l'errore non può essere stimato.

Il campionamento può essere svolto selezionando/arruolando volontari, campioni di convenienza o mediante una scelta ragionata. In questi tre casi, i risultati saranno poco rappresentativi della popolazione.

Il campionamento più rappresentativo si ha quando ogni unità della popolazione ha la stessa probabilità di entrare a far parte del campione. Viene definito selezione casuale o RANDOMIZZAZIONE.

Come si riduce l'errore di campionamento per selezione viziata? Ci sono tanti tipi di campionamento, alcuni validi altri non validi. Uno studio può essere fatto su volontari sani, errore massimo di campionamento ma non si può fare niente, significa che lo studio ha delle caratteristiche applicabili solo su gente che dice 'va bene lo voglio fare'. In genere sono studi piccoli, fattibili che testano che una cosa è sicura, per esempio: l'utilizzo di una metodica per la misurazione di volumi del casco che abbiamo fatto tre mesi fa e su volontari sani ma perché volevamo soltanto vedere se un metodo era effettivamente funzionante e non volevamo avere dei dati sulla popolazione. Altra possibilità è il campione di convenienza: tutti i pazienti che passano dalla terapia intensiva del policlinico dal 2015 al 2019. Questo ultimo è un campione che a me fa comodo ed è facile da prendere anche se non è rappresentativo di tutti i pazienti

della terapia intensiva del mondo. Ricordate però che l'unico metodo di campionamento che riduce al massimo la selezione viziata è affidarsi al caso e questo è il concetto di randomizzazione. Se io affido al caso la selezione dei pazienti, al farmaco A e al farmaco B, riduco al massimo la possibilità di avere fatto un errore nella selezione del campione. Oppure voglio studiare qualcosa su tutti gli studenti dell'ateneo che sono più di 4000, faccio una lista di 4000 nomi e dico a caso, tramite il processo di selezione a caso (randomizzazione), quali di questi devo studiare. Questo è un processo di randomizzazione ed è il caso che stabilirà se devo studiare o non studiare quella persona.

Fricano: 'Quindi tra un campionamento che riguarda tutti i pazienti che nel 2018 passano dalla rianimazione ed un campionamento casuale di tutti i pazienti che sono passati dalla rianimazione nel 2018, ad esempio 200 su 4000 pazienti, quale dei due è più attendibile?'

Il caso è sempre la soluzione migliore.

Fricano: 'Quindi anche se il campione è più piccolo avendo utilizzato il caso è comunque più preciso del primo?'

La cosa ottimale sarebbe un grande campione ottenuto a caso, infatti gli studi che danno risposte a domande di enorme importanza clinica sono tutti studi grandi, multicentrici con la selezione dei pazienti randomizzata. Lo studio randomizzato piccolo su un numero di pazienti più piccolo ha sempre un vantaggio rispetto ad uno studio non randomizzato e se è fatto bene è l'unico metodo che permette di ridurre l'errore di campionamento. Ultima cosa, più grande è il numero meglio è e questo vale anche quando utilizziamo il campionamento casuale. Più campionamenti casuali facciamo migliore è la rappresentatività del nostro campione, infatti gli studi piccoli e randomizzati su 40 pazienti sono degli studi che devono essere poi confermati in campioni più grandi e mai definitivi. Quindi, morale della favola, quando leggete uno studio state attenti a come il campione è stato selezionato. Gli abstract e i poster ai congressi, presentati in genere come entry level della ricerca, non sono mai studi randomizzati perché lo studio randomizzato è più difficile, ci vogliono concezioni migliori di metodologia della ricerca però è il metodo migliore per ridurre la variabilità. Ultima cosa che vi dico, la randomizzazione è un enorme vantaggio che riduce l'impatto anche di fattori che noi non conosciamo, vi faccio un esempio: supponiamo che io voglia utilizzare un campione di convenienza del 2018 dei pazienti in terapia intensiva; nel 2018 c'è stato un boom di un'influenza da HA3N2 e avrò tanti pazienti che entrano per insufficienza respiratoria ma se io negli ultimi 5 anni randomizzo la selezione i fattori noti e non noti, a cui non sto pensando, si bilanceranno tra i gruppi che sto prendendo in studio. Immaginate la randomizzazione come una livella, il terreno può essere un po' più alto un po' più basso, io livello tutto e dico fattori noti e ignoti dovete stare tutti allo stesso modo e allora affido al caso, lancio della monetina e non posso sbagliare. Non possiamo controllare tutti i fattori, ce ne sono alcuni che non conosciamo ma se li affido al caso il caso lo farà per noi.

Altre definizioni

VARIABILE: È una caratteristica dell'unità statistica che io sto osservando in quel momento che per definizione è soggetta a variazione e che io sono intenzionato a misurare, un esempio: la vostra pressione arteriosa e quindi una variabile soggetta a variazione perchè non è uguale per tutti, ha una certa variabilità che io vado a misurare in ciascuna unità statistica che è ognuno di voi. La variabile è la pressione arteriosa, 120/60 mmhg è il dato. Quindi la variabile è l'etichetta, il numero è il dato che è infatti il risultato della misurazione.

Fricano: 'Quindi in questo caso l'unità siamo noi o è la nostra pressione?'

No, l'unità sei tu e su di te vado a misurare una variabile che è la pressione arteriosa e ottengo un dato che è il 120/60 mmhg.

Il dato non è detto che sia per forza un numero, può essere anche una modalità. Guardiamo gli occhi, ogni vostro paio di occhi è una unità statistica, il colore è la variabile e il verde, ad esempio, è il dato che in questo caso non è un numero ma una modalità

Mirasola: 'Ma il dato lo stabiliamo noi?'

Noi stabiliamo la variabile

Mirasola: 'Come faccio a capire il dato da attenzionare? Verde in questo caso, il colore degli occhi, come faccio a stabilire che è verde e non marrone?'

Perchè utilizzi una corretta unità di misurazione. Se io ti devo misurare la pressione la misuro in mmhg non la misuro in cmh2o perchè so che l'unità di misura della pressione arteriosa sono i mmhg. Se io devo catalogare il colore dei tuoi occhi mi rifaccio ad una scala colorimetrica internazionale per cui secondo la mia vista questo è marrone, questo è verde e questo è azzurro. La domanda che fai è giusta, infatti per ogni misurazione è importante utilizzare l'unità di misurazione corretta (cmh2o, mmhg ecc) a seconda di quello che dovete misurare.

Lara: 'E questo anche per rendere il mio dato confrontabile col tuo, giusto? '

Assolutamente sì. Non sono infrequenti gli studi in cui qualcuno carica i dati su un database e non dice a chi deve analizzare i dati che cosa è, tu pensi di aver confrontato una cosa e invece è tutt'altro.

Lara: 'Prima parlavi di variabile come di qualcosa di modificabile e invece la variabile indipendente? Ad esempio sesso o razza, è corretto parlare di variabile per cose che resteranno tali nel tempo?'

Sono variabili nella popolazione, non nella persona. Se io studio la vostra pressione arteriosa la definisco variabile perché tu l'avrai in un modo, lei in un altro e così via. Stessa cosa per il sesso. Poi che quella variabile all'interno della stessa persona possa mutare nel tempo è un altro discorso ma la definizione di variabile è perchè nell'insieme delle unità statistiche non ottengo

lo stesso risultato costantemente. Non esiste un fenomeno biologico che è uguale per tutti gli essere umani. Qualunque cosa voi immaginate per ognuno di noi è diverso, anche se di poco ma è diverso.

Cerchiamo di descrivere le variabili e ce ne sono di due tipi fondamentali, una è quantitativa e l'altra è qualitativa.

La quantitativa è facile, sono numeri che possono appartenere a qualunque insieme della scala dei numeri reali, da zero a infinito. Queste sono le variabili continue. Se cerco di contare gli elementi dentari di ciascuno di noi si tratterà sempre di una variabile quantitativa però non esiste un essere umano con 200 o più denti, il numero dei denti è finito nella anatomia normale umana e quindi il numero massimo di elementi dentari arriva ad un numero reale, non ne posso avere 128 e parleremo di variabile discreta. Dunque Continua da zero a infinito, quando invece la restringo un po' la chiamo discreta ma il concetto è sempre quello. Qualitative significa che quella variabile esprime una qualità, ad esempio maschio/femmina oppure verde/marrone/blu ecc, quando invece quella variabile di tipo qualitativo si può esprimere come acceso/spento, bianco/nero oppure 0/1 si chiama BULEANA cioè può assumere solo due modalità, ad esempio lo stato di accensione della radio è una variabile qualitativa di tipo buleano in quanto la radio può essere accesa o spenta. Mentre invece il vostro stato di veglia in questo momento non è acceso o spento perchè siete tutti in una zona grigia tra il confuso, lo stanco e lo schifato quindi io la descriverei come una variabile qualitativa di tipo ORDINALE perchè ci sono vari gradi, scale di grigio, e assegno a ciascuno vostro stato di veglia una di quelle.

Però non è per sempre perchè quando prendo dei dati e quindi misuro la variabile posso trasformare una variabile quantitativa in una qualitativa. Ad esempio l'età che tipo di variabile è? È Quantitativa. Però se io decido che 0-1, 0-2 sono lattanti, fino ai 13 anni sono adolescenti, dai 13 ai 60 sono adulti e poi ci sono gli anziani che cosa ho fatto? Ho trasformato una variabile quantitativa in una qualitativa. Poi posso decidere di fare una cosa molto giornalistica e divido la popolazione in giovani e vecchi e allora, come acceso/spento, la variabile diventa BOLEANA.

Ora io di tutti questi numeri che cosa faccio? Li devo mettere assieme e creare il database ovvero un magazzino dove inserisco tutto quello di cui abbiamo parlato. Questa fase è critica perchè spesso viene data in mano a quelli più giovani che combinano dei disastri inimmaginabili però siccome voi avete seguito questa lezione vi do una dritta molto semplice: nelle colonne troviamo le variabili, sempre, in tutte le tabelle del mondo; nelle righe troviamo le unità statistiche, quindi leggendo quante righe posso sapere quante unità statistiche ci sono. Se io invece mi leggo le intestazioni delle colonne scopro che cosa sono andato a misurare. Non c'è altro modo per creare un database, è sempre lo stesso. Vi ho detto che questo passaggio è critico perchè poi questo va in mano a chi deve fare l'analisi statistica e solo se ha seguito queste regole si può arrivare alla fine, altrimenti è un gran casino e quindi quando leggete dei risultati tutto quello che riguarda l'ipotesi statistica è passato per una cosa del genere. L'ultima cosa che vi dico su questo è che a volte per semplificare il processo di creazione del database io

posso scegliere di assegnare dei codici, come voglio: maschio/femmina per me può diventare 0/1, l'età inferiore a 25, tra 25 e 50 oppure sopra 50 può diventare 0,1,2,3 non è importante e sono io che decido come codificare. Questo significa in statistica codificare. Consiglio che vi do, se mai farete una cosa del genere e credo che lo farete, ricordatevi i codici.

Bonanno: 'Ma noi quando facciamo una codifica, quando diamo dei codici dobbiamo partire sempre dal presupposto che zero è la cosa migliore e il numero più alto è la cosa peggiore?'

No, teoricamente no però è deduttivo il pensiero. Dipende quello che tu vuoi vedere, come se tu vedi il bicchiere mezzo pieno o mezzo vuoto. In genere cambia a seconda di chi carica. Ad esempio chi carica 0/1 per maschio/femmina ne riconosce subito il sesso. Nessuno mette al proprio sesso lo zero a meno che non ha esperienze diverse però in genere funziona così.

Adesso parliamo di statistica descrittiva. Abbiamo capito che cosa è però dobbiamo anche capire cosa riguarda.

La statistica descrittiva è quel processo per cui noi otteniamo misure di tendenza centrale, misure di variabilità, tabelle e grafici. L'essere padroni di principi di statistica descrittiva vi fa diventare, all'improvviso, nella comunità medica una sorta di santone. Tu dai le caratteristiche dei pazienti con un grafico e la gente ti guarda come se avessi fatto una magia e in realtà hai soltanto, forse, 3-4 concetti di excel o di un programma ancora più scarso.

Quindi misure di tendenza centrale, parliamo in questo caso di variabili continue, sono delle misure che ci danno un numero che sta più o meno in mezzo ai dati del mio campione. Per intenderci, media, mediana e moda dunque ci dà un numero che ci fa capire dove sta la parte centrale dei miei dati. Ecco perché si chiamano anche di posizione. Quello principe è la media aritmetica, non devo dire a voi cosa è però sappiate che è una misura di tendenza centrale.

Io esprimo dove sta la parte di mezzo del mio range di numeri/dato con un numerino solo. Facciamo finta che la media dell'altezza qua dentro è 170 cm, con un numero sintetizzo la variabile altezza di tutti voi, della mia popolazione.

Giammanco: 'Quando parliamo di variabili, stiamo parlando di variabili quantitative continue qui però quando parliamo di variabili quantitative discrete ad esempio il numero dei denti io posso comunque effettuare la misura di tendenza centrale?'

Assolutamente sì, non cambia niente. Solo che per una delle due sai dove si chiude il cerchio dell'altro no. Non cambia nulla.

Giammanco: 'Conoscendo il limite massimo della nostra variabile la misurazione assume un valore diverso, un significato diverso, una potenza statistica diversa?'

No, sarà solo più precisa, il concetto di precisione lo avremo dopo, perché oltre alla misura di tendenza centrale c'è un'altra misura di cui parliamo dopo e in quel caso ti aiuta quando hai un

insieme infinito e non finito. Però una cosa voglio dirla sulla media anche se è un concetto che dovrete sapere, se voi volete descrivere la variabile di un campione con un numero, tendenza centrale, sappiate che se qui dentro arriva un giocatore di basket dell' NBA che ha un'altezza totalmente al di fuori delle nostre comuni, la media ne risente tanto e non sarà la cosa più precisa, il parametro più preciso. Ricordatevi sempre che la media risente degli estremi, più strani sono gli estremi peggiore è il mestiere della media. Non è fatta per descrivere un campione in cui ci sono strani estremi.

Darisssh: 'E in questi casi come facciamo? Scartiamo i valori estremi?'

Puoi utilizzare altre misure di tendenza centrale oppure puoi cercare di rendere il campione più omogeneo ma non ci pensate in questo momento se no diventa un casino. Anche il concetto di mediana è abbastanza semplice: è quel valore che divide a metà la distribuzione dei dati della mia variabile (il 50% stanno su e il 50 % stanno giù). La mediana la dovete vedere come una cosa più robusta della media, la media è più elegante, graziosa, si veste bene, ha gusto e quando parla si capisce invece la mediana è una cosa grezzissima che però riesce sempre a darti delle informazioni qualunque siano i valori che stanno dentro il tuo campione. La mediana non è influenzata dagli estremi, riesce a darti delle informazioni anche se qui dentro entra il giocatore di basket. La moda è il numero che ricorre più volte. Sicuramente vi ricordate dai vostri studi che cosa è una distribuzione normale e sapete che tutti i fenomeni biologici nella popolazione degli esseri viventi segue la distribuzione normale, il problema è che i campioni no.

Quindi ricordatevi sempre, questo è molto importante quando leggete uno studio, che la distribuzione del campione guiderà la scelta delle misure di tendenza centrale che dobbiamo utilizzare. Quando il campione è distribuito normalmente (senza valori estremi) la scelta di media, moda o mediana è indifferente. Quando le cose cambiano e la distribuzione non ha più quella bella forma a campana bisogna stare attenti a cosa scegliere, perché la media di questi numeri qua non è rappresentativa...di mezzo, non lo può essere. Quello che voglio dirvi è attenzione alle caratteristiche del campione quando scegliamo l'uno o l'altro. Regoletta: quando vedete studi piccoli che vi danno medie prendete e accartocciate quello che avete stampato e lo buttate nella spazzatura, cancellate anche il pdf che avete scaricato.

Fricano: 'In questo caso preferiamo la mediana?'

Sì, potremmo preferire la mediana

Sicuramente vi ricordate cosa è una distribuzione normale e sapete che tutti i fenomeni biologici nella popolazione di tutti gli esseri viventi segue la distribuzione normale. Il problema è che i campioni no. Quindi **quando leggete uno studio ricordatevi che la distribuzione del campione guiderà la misura centrale che dobbiamo scegliere.** Quando il campione è distribuito normalmente come nella figura, visto che non ci sono outliers o dici moda mediana o media è la stessa identica cosa. Però quando le cose cambiano e non ha più la distribuzione a campana bisogna stare attenti a cosa scegliere, perché la media di questi numeri (figura) non è

rappresentativa del punto di mezzo, non lo può essere. Quindi attenzione alle caratteristiche del campione quando scegliamo l'uno o l'altro. Come regola generale se in uno studio piccolo vi danno come valore la media buttatelo, quando non c'è simmetria la media funziona male e potremmo considerare la mediana.

Se io dispongo i valori delle vostre altezze in un file posso ordinare le vostre altezze e dividerle per quantili, il più piccolo assegno zero il massimo assegno 1 e divido i rimanenti in quantili (0,1-0,2...). Nell'utilizzo di quantili quello che in assoluto troverete più spesso è l'utilizzo dei quartili, cioè divido i numeri che ho in quattro parti. Questo è un altro modo per descrivere i dati che ho.

Però se ho un campione, non basta sapere quanto è il numero di mezzo, ma anche quanto varia il mio campione intorno a questo numero: ecco il concetto di misure di variabilità o dispersione (ho i miei dati ci metto un paletto in mezzo, ok, ma devo sapere come variano intorno a sto paletto i dati e se so quanto si disperdono in base al centro so tutto). Non guardate le formule!!! La varianza è una misura di dispersione.

La deviazione standard si associa alla media, quando riporto la media come valore centrale ci associo la deviazione standard.

Se io descrivo l'altezza più alta e la più bassa questo valore si chiama range, che è quella cosa che comprende tutto il resto, numero massimo e numero minimo.

Se io scelgo di riportare il terzo e il primo ho il range interquartile che è una misura di dispersione che si associa sempre alla mediana e che è un numero grezzo, forte e robusto.

Le associazioni sono: Media con deviazione standard e il range interquartile con la mediana.

Ricordatevi che bisogna accertarsi sempre della simmetria dei miei valori perché la mediana e la deviazione standard siano rappresentativi del mio campione.

Come fai a accertarti della simmetria del tuo campione? Devi guardare come si dispongono i valori e se c'è qualcosa che assomigli alla curva di simmetria cioè alla distribuzione normale posso essere ragionevolmente certo che il campione sia simmetrico. Altro trucco è vedere quante unità statistiche hai: quante meno sono tanto meno è la probabilità che tu possa usare media e deviazione standard. Maggiore è il numero di unità statistiche maggiore sarà la probabilità di poterle usare.

Fino adesso abbiamo parlato di valori quantitativi. Adesso possiamo parlare di misure di sintesi delle **variabili qualitative**.

Frequenza assoluta è una proporzione cioè quanto si ripete quella caratteristica dell'unità statistica sul totale (occhi verdi sulle unità statistiche, cioè il paio di occhi)

Frequenza relativa è la frequenza assoluta messa in relazione al numero totale del campione (tabella esplicativa: dove gruppo sanguigno, frequenza e il $32/55=0,58$ e la percentuale è il 58%. Ricordatevi che il totale nelle tabelle come queste sta sempre in fondo)

Frequenze e percentuali sono le misure di sintesi delle variabili qualitative.

Il tipo di grafico deve essere idoneo alla tipologia di variabile che tu stai trattando.

Ortogramma o diagramma a barre. E' utilizzato per variabili di tipo qualitativo e l'altezza è correlata con la frequenza. Per creare una cosa del genere devo riportare la frequenza dei dati del mio database di screening, è molto semplice ed è informativo (grafico pap test e PSA che sono stati più eseguiti nella popolazione rispetto ad altri test)

(figura) In questo caso in ascissa ho un valore quantitativo di valore della colesterolemia in cui i valori li ho scissi in cosiddetti ranghi, da un certo numero ad un altro numero, vi ricordate il processo di codifica (0,1) e il numero delle unità statistiche è sulle ordinate.

Diagramma lineare. La funzione è osservare la modifica nel tempo di quel dato che ho deciso di esaminare (in ascissa il tempo e il valore in ordinata) quindi metto in correlazione una variabile con il tempo. Può essere fatto per tutte le unità statistiche.

Diagramma di dispersione (variabili quantitative) da come si disperdono i dati all'interno del grafico posso capire anche qualitativamente come si disperdono (figura: peso e altezza ogni puntino corrisponde ai due valori). Più frequente, più comune e di più facile lettura.

Diagramma a torta. Un colore assegnato una frequenza, tutta la torta è il 100% ,il pezzo ha un valore diverso associato ad una determinata modalità della variabile. E' molto frequente. Le torte con fettine minuscole è assurdo!!Non è per questo che viene creato, è sbagliato.

Scatola a baffi o box slot. E' molto frequente (es. con pressione arteriosa diastolica) la linea dentro la scatola è la mediana, le altre due linee che chiudono lo scatolo sono il primo e il terzo quartile, i baffi sono i limiti del range cioè numero massimo e il numero minimo. Dà idea della tendenza centrale e del range del nostro campione.

Più facile è il grafico e migliore sarà la riuscita di ciò che volevi fare.

Statistica inferenziale. Parto da una popolazione ne prendo una parte grazie al processo di campionamento, guardo le caratteristiche, misuro le variabili, e poi voglio descrivere le caratteristiche del campione riferendomi a quelle della popolazione con un certo margine di errore. Concetto di inferenza significa stimare la probabilità del risultato che ha quel campione sia veramente rappresentativo della popolazione.

La popolazione è l'insieme delle unità statistiche.

(es. io mi trovo all'interno di un bosco dove la luce non c'è e io posso dire in questo bosco ci sono faggi, ma non è così! Se in una zona c'è il sole, e non c'è tanta umidità, ci sono altri alberi. Quindi io devo esplorare tante zone per poter poi campionare il bosco. Se io fossi rimasta nella prima zona avrei detto che nel bosco c'era il 100% di faggi ma non è vero!)

(es. fai una biopsia cioè campioni una lesione ma non è detto che in quella lesione ci siano solo la tipologia delle cellule che tu hai prelevato. Il concetto di inferenza esprime la probabilità che le cellule che hai prelevato con la biopsia siano rappresentative della lesione stessa)

Teoria della verifica delle ipotesi

Qualunque protocollo di ricerca che si rispetti deve obbligatoriamente passare per la teoria della verifica delle ipotesi. La teoria della verifica delle ipotesi significa poter, in termini probabilistici, fare una affermazione relativa alla popolazione, sulla base dei risultati che ho ottenuto dal campione.

(es. ricordatevi il bosco. Proporzione degli alberi in base a quanto bosco esploro e assumere una probabilità che quegli alberi che ho visto siano un campione di tutti quelli presenti nel bosco)

La teoria ha degli step. Immaginate diverse teorie: incidenza di una certa condizione in una popolazione, un particolare farmaco versus placebo. E' così che si disegna un quesito di ricerca.

Si stabiliscono due ipotesi e si parte da H_0 o ipotesi nulla che dice sempre il contrario di quello che volete dimostrare, è sempre espressa al negativo. Cioè se il farmaco A è non è più efficace del farmaco B.

H_1 invece è al positivo ed è quello che volete dimostrare, es. il farmaco A è più efficace del farmaco B.

Es: mi chiedo se esiste una differenza del voto dell'esame di statistica per chi ha un diploma di maturità scientifica e chi no.

H_0 : la differenza tra il voto dei due gruppi, chi ha maturità scientifica e chi no, non c'è

L'intervento in studio è la maturità scientifica; il voto è la variante; il voto, il numero è cioè quello che andrò a registrare.

H_1 c'è l'effetto tra chi ha preso la maturità scientifica e chi no nel voto dell'esame di statistica.

Ho misurato le mie variabili, descritto le mie ipotesi, creato il database e ora devo decidere in fase analitica che ricetta cioè le modalità di lettura da utilizzare per ottenere la risposta alla mia ipotesi. Il test statistico è la procedura da utilizzare per smentire l'ipotesi nulla e confermare l'ipotesi sperimentale.

Da qui parte il concetto della probabilità.

The end.