STATISTICA

La statistica descrittiva è utilizzata per raccogliere, analizzare e descrivere DATI con l'obbiettivo di determinare RELAZIONI E DIFFERENZE tra i gruppi. Scopo della raccolta dei dati è metterli a confronto. Bisogna aver ben chiaro l'obbiettivo dello studio e dell'analisi. Aver chiaro anche quale tipo di informazione devo raccogliere e rendere fruibile e facile la consultazione/lettura del dato ottenuto.

Le Tecniche di campionamento: estraggo dei CAMPIONI rappresentativi da una POPOLAZIONE numerosa, da essi traiamo delle conoscenze.

- Concetto di INFERENZE: generalizzazioni dei risultati osservati nel campione all'intera popolazione di riferimento.
- Concetto di MODELLO: il trattamento statistico del dato, se effettuato in maniera corretta, si avvicina quanto più verosimilmente al dato reale, ciò che otteniamo è un modello, un'idea della realtà, che NON è la realtà. Concetto di verità approssimativa della realtà. Es: modellino ferrari F1

E' importante semplificare i grandi dati se possibile, perché così facendo diventa di nostra più facile comprensione, il numero è molto più piccolo e quindi anche più facile da gestire

La raccolta dei dati porta spesso al fenomeno del BIG DATA: possesso di enormi quantità di dati NON sfruttati (data mining: strumenti e tecniche di campionamento mirate all'estrazione di conoscenza di grandi quantità di dati) per trarre conoscenza.

Lo sport può rappresentare uno scenario ideale per gli strumenti e le tecniche di analisi statistica, in grado di produrre modelli descrittivi e inferenziali. Nello sport, intorno all'atleta, ruotano una serie di figure professionali che hanno rapporti funzionali con il trattamento dei dati.

Quando si fa una ricerca viene fatta una raccolta dei dati. I criteri e le precauzioni per la completezza e l'accuratezza delle informazioni raccolte sono:

- Quali e quanti dati raccogliere
- Rilevazione completa sull'intera popolazione
- Il campione deve essere rappresentativo di tutta la popolazione
- Avere chiaro l'obbiettivo ultimo della ricerca
- Recuperare le informazioni
- Indentificare con chiarezza termini, concetti e definizioni usate
- Decidere il formato di rappresentazione del dato

Riassumendo:

1. RACCOLTA DEI DATI: concetto di PREVISIONE : i dati che raccolgo devono fornirmi una capacità PREDITTIVA in un comportamento futuro della popolazione
2. ANALISI E STUDIO DEI DATI RACCOLTI
3. DEDUZIONE DI UNA LEGGE/MODELLO
4. PRESENTAZIONE DEI DATI tramite TABELLE: che servono per una presentazione sintetica dei dati che devono essere significativi e non ambigui, messi in modo semplice e devono contenere un buon

numero di dati numerici, è il punto di partenza per la rappresentazione, lettura, interpretazione ed elaborazione successiva. Deve essere costituita da TITOLO, , COLONNE E COLONNA MADRE (prima colonna a sx), TESTATA, FINESTRA DI DIALOGO (=rappresenta l'interfacciabilità tra la tabella e l'utente), Nota e la FONTE DI PROVENIENZA DEI DATI (opzionale) Oppure GRAFICI: ci dà una rappresentazione del dato raccolto e processato, si coglie immediatamente il senso geometrico della figura e si ha una visualizzazione della frequenza con cui si presenta una modalità- deve essere accurato e competo di titoli. Si occorre al grafico quando i dati statisti che raccolgo non riesco a riproporli in una tabella.

DATI E MISURAZIONI E VARIABILI

Staile Holder: colui che è interessato in natura al risultato della ricerca.

Degli strumenti di misurazione statistica sono anche gli **INDICI DI POSIZIONE**:

- MEDIA ARTMETICA SEMPLICE che viene indicata con X (x segnato) restituisce l'ordine di grandezza del fenomeno (cioè sostituito ai valori di x osservati lascia invariata la somma). Nella maggior parte dei casi cade centralmente all'interno dell'insieme ordinato dei dati.

  Formula: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ somma tutti i valori numerici divisa per il numero di valori numerici considerati.

- MEDIA ARITMETICA PONDERATA: quando ho una notevole quantità di dati con una distribuzione di Frequenze. Fi rappresenta la frequenza (o l'importanza) di quel valore nella distribuzione (il numero di volte che quel valore compare nella distribuzione)

  Formula: $M_{a,pond} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$ è calcolata sommando tutti i valori delle modalità per un coefficiente che definisce il suo peso rispetto agli altri valori della distribuzione (cioè le Xi vengono moltiplicate alle fi. Prima faccio i prodotti, poi li sommo e poi divido per n)

  Per la MEDIA PONDERATA PER CLASSI devo calcolare il valore centrale della classe, cioè il valore minimo meno il valore massimo diviso 2, poi moltiplicarlo per la frequenza; fare poi tutte le somme dei valori calcolati e dividere per n, cioè per il numero delle osservazioni.

- MODA o NORMA: è la modalità di massima frequenza, ovvero il valore che compare più frequentemente e viene indicato con $V_{\circ}$ La distribuzione può essere UNImodale, se ammette un solo valore modale, Bimodale, se ne ammette due, TRImodale, tre etc..

- MEDIANA: la utilizzo molto quando ho a che fare con campioni non numerosissimi. Si vede come in piccoli gruppi dove il valore media influenza poco la mediana restituisce un valore più stabile. In una distribuzione di n valori Xi ordinati in modo crescente o decrescente, la mediana è il valore che si colloca a metà della distribuzione.

  Se n. è dispari : la mediana è il dato centrale di x.

  Se n è pari: la mediana è la media dei due valori centrali di x. La mediana BIPARTISCE la distribuzione in due sotto distribuzioni Att.ne: devo ordinare i valori di x (crescente o decrescente) è a metà della sequenza!!

INDICI DI DISPERSIONE: misurano la variabilità dei valori di x rispetto alla media o alla mediana.

DEVIANZA: si ottiene a partire dal concetto di scarto rispetto ad un numero centrale della distribuzione che è la MEDIA (lo scarto ci permette di valutare l'incertezza associata alla media, tanto più grande è lo scarto rispetto alla media in senso positivo o negativo tanto più la stima che abbiamo ottenuto è piuttosto incerta) ed è la sommatoria di tutti gli scarti al quadrato. Se raddoppia il numero dei dati, raddoppia anche la devianza anche se la variabilità dei dati rimane costante. E' la base delle misure di dispersione di tipo quantitativo. Da essa discende la varianza, la deviazione standard o scarto quadratico medio. Se il numero di dati raddoppia anche la devianza raddoppia, nonostante la variabilità dei dati si mantenga constante.

Data N la dimensione della popolazione e μ la media si definisce Devianza le somme di i da 1 a N - μ elevato al quadrato (cioè i valori di x- la media elevati al quadrato):

Devianza su Popolazione $\sum (Xi - \mu)^2$ (media sulla popolazione)

Devianza su Campione $\sum (Xi - X \text{ segnato})^2$ (media su campione)

Devianza su campione + frequenze $\sum (Xi - X \text{ segnato})^2 \times Fi$

VARIANZA:

E' la DEVIANZA MEDIA rapportata al numero di osservazioni. Non è direttamente confrontabile con gli indici di posizione, è la media degli scarti al quadrato (praticamente la devianza fratto n se riferito a popolazione, fratto n-1 se riferito a campione). La varianza è influenzata da eventuali osservazioni anomale e non è direttamente confrontabile con la media. La varianza è espressa nel quadrato dell'unità di misura utilizzata per la variabile per cui non è molto interpretabile dal punto di vista pratico perciò poi userò la deviazione standard che è appunto la varianza sotto radice che restituisce la stessa unità di misura della variabile e della media.

Le formule:

Su Popolazione $\sum (Xi - \mu)^2 \div n$ simbolo: sigma quadro per la popolazione

Su Campione $\sum (Xi - X \text{ segnato})^2 \div (n-1)$ simbolo: s quadro per il campione

La Varianza ponderata la uso quando i dati sono raggruppati in classi, per cui si tengono in considerazione le frequenze. Le formule:

$\sum (Xi - Xsegnato)^2 \times Fi \div (n-1) = \sum (FiXi)^2 - (FiXi)^2/n$ tutto $\div(n-1)$

NB: n-1 è anche il grado di libertà

VARIANZA PONDERATA: è la varianza quando presenta dati sono raggruppati in classi, si tengono in considerazione le frequenze.

SCARTO O DEVIAZIONE: Ci indica quanto la i-esima misura (cioè una singola osservazione) differenzia dalla media e ci permette di valutare l'INCERTEZZA di una stima.

Formula: $d_i = x_i - \bar{x}$ (lo scarto è la differenza dei singoli valori meno la media cioè: x1 - media)

SCARTO QUADRATICO MEDIO\ **DEVIANZIONE STANDARD**: E' un indice di dispersione. Molto usato. Viene indicato con la lettera greca sigma(σ) per la popolazione e la lettera s per il campione.. Misura la dispersione dei dati intorno alla media, ha valore sempre positivo. E' la radice quadrata della varianza. Attraverso al MEDIA dei singoli scarti valuto l'incertezza di una stima. E' direttamente confrontabile con le MISURE DI POSIZIONE. Si utilizza appunto per misurare quanto sono lontane le unità statistiche dalla media: misura la distanza

"tipica" di ogni singola misura dalla media. Sono due le formule nel caso sia su POPOLAZIONE o su un CAMPIONE.

Formula su popolazione: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$ se è su un campione al denominatore c'è n-1.

La dev.st potrebbe non essere utile nel confronto delle variabilità all'interno di 2 gruppi di dati.

LA PERCENTUALE: Esprime il rapporto tra le due quantità, è il risultato di una proporzione:

$a : b = x : 100 \Leftrightarrow \frac{a}{b} = \frac{x}{100}$ mi serve per misurare la proporzione tra due insiemi.

COEFFICIENTE DI VARIAZIONE CV% o deviazione standard relativa:

E' un indice di dispersione espresso in PERCENTUALE consente di misurare la variabilità indipendentemente dalla grandezza e dalla scala di misura dei dati, non dipende dall'unità di misura. Lo uso per confrontare variabili misurate su gruppi diversi e che sono espresse con diverse unità di misura (es. Anni e Kg). Quantifica in modo oggettivo quanto sia grande il valore di una dev.st rispetto alla media.

CV% = deviazione standard σ ÷(media μ / x segnato) tutto ×100 (userò μ se è su popolazione e x segnato se su campione) per entrambi i gruppi e poi li confronto

CV=0 allora anche dv.st=0 per cui vuol dire che tutte le unità statistiche hanno lo stesso valore. Non parlo di variabilità ma di costante.

CV vicino allo 0 allora vuol dire che la dv.st è piccola rispetto alla media.

CV=0,5 è la soglia del CV, se inferiore a 0,5 la variabilità dei dati è contenuta per cui la media è un buon indicatore. Se è maggiore di 0,5 la variabilità è alta per cui la media potrebbe non essere un buon indicatore.

CONFRONTO TRA DATI STATISTICI: IL MODELLO DISTRIBUTIVO GAUSSIANO o distribuzione normale

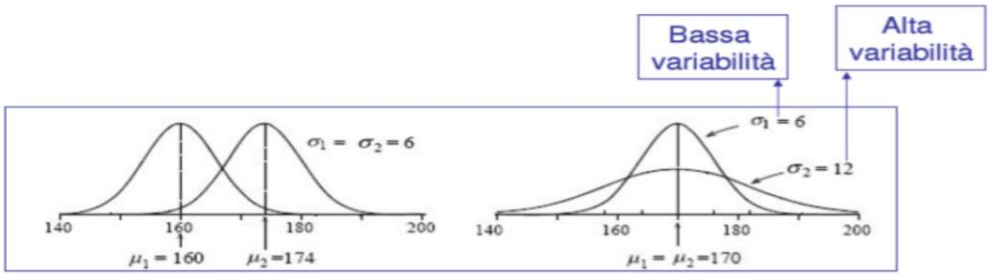
La distribuzione NORMALE o GAUSSIANA è una distribuzione continua molto utilizzata in statistica ad è molto importante. E' definita tra due parametri, la media (μ) e la varianza (σ^2), in cui la **media** ne definisce la **posizione** e la **varianza** la **forma**. E' una distribuzione simmetrica e graficamente ha una tipica forma a campana.

E' UNIMODALE e centrata sulla media.

1° caratteristica: distribuzione unimodale centrata sulla media cioè MEDIA=MEDIANA=MODA

2° caratteristica: l'area sottesa dalla curva (della campana) è = 1

3° caratteristica: esiste tra – infinito e + infinito (-∞ e +∞)



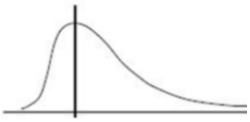
↕ ↕

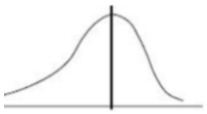
Media (posizione) diversa μ ≠μ | Media (posizione) uguale μ=μ

Dispersione (varianza) uguale σ=σ | Dispersione (varianza) diversa σ≠σ

La campana alta indica una bassa variabilità, quella bassa un'ampia variabilità. Se le varianze sono uguali tra i due dati anche se la media è diversa, hanno la stessa forma.

Gli indici di posizione e quelli di dispersione riassumono l'informazione statistica. La distribuzione di frequenza osservata è ricostruita a partire dai dati. Una distribuzione teorica di probabilità è definita da una funzione matematica che permette il calcolo della probabilità associata a ciascun valore o intervallo di valori. La funzione distribuzione teorica discreta è definita dalla probabilità che la variabile assuma un certo valore. La distribuzione teorica continua di una probabilità l'abbiamo per x variabili di tipo continuo, è una probabilità divisa per un intervallo, è una densità di probabilità.

DISTRIBUZIONE ASSIMENTRICA POSITIVA  MEDIA>MODA

DISTRIBUZIONE ASSIMETRICA NEGATIVA  MEDIA <MODA

Teorema del limite centrale

Una variabile costituita dalla somma di un numero elevato di fattori si distribuisce in maniera Gaussiana anche se il singolo fattore ha una distribuzione di frequenza, e se il numero è sufficiente, il teorema è valido anche se la distribuzione della variabile del singolo fattore NON normale, significa che all'aumentare del numero dei dati la distribuzione tende ad essere Normale.

TIPI DI VARIABILI STATISTICHE

Nella fase di analisi dei dati, la statistica studia le variabili e le loro relazioni. Scopo della raccolta dati è ricercare la relazione causa\effetto e interpretare e prevedere fenomeni. L'analisi può essere di tre tipi:

1. MONOVARIATA (studio di 1 variabile alla volta)
2. BIVARIATA
3. MULTIVARIATA

Tipi di variabili:

Variabili **QUALITATIVE**: i valori sono MODALITA', prive di significato numerico. Possono essere:

VARIABILI NOMINALI: il carattere assume stati DISCRETI NON ORDINABILI, non ha stati intermedi, le relazioni possibili sono ≠ o = (es, verde/rosso/giallo… oppure nomi delle persone, oppure "letale o NoN letale" non c'è una via di mezzo, lo stato civile assume stati discreti non ordinabili, non esiste un punto intermedio tra essere celibi e essere coniugati) non posso stabilire una relazione di < o >, non posso fare nessuna operazione aritmetica.

VARIABILI ORDINABILI: il carattere assume stati DISCRETI ORDINABILI, può essere gerarchica cioè = ≠ < > che sono le operazione consentite (es. molto soddisfatto/poco soddisfatto/soddisfatto.. oppure "età fino ai 40 /

maggiore di 40" oppure il titolo di studio con le modalità disposte, licenza media, diploma, laurea) è possibile calcolare la mediana (dato di mezzo) ma non la media e tutto quello collegato a essa.

Le variabili **QUANTITATIVE**: i valori sono grandezze numeriche, si esprimono in intervalli di classe (discrete=numeri interi risultato di un conteggio oppure Continue=numeri infiniti di valori in un intervallo esempio se trovo numeri con la virgola kg. 58,33 variabile quantitativa continua). Possono essere:

VARIABILI CARDINALI: la proprietà assume valori NUMERICI, valgono tutte le operazioni aritmetiche = ≠ < > + × ÷ - (es. età, peso…)

CLASSIFICAZIONI DELGI ERRORI DI MISURA

MISURAZIONE → MISURA → NUMERO REALE → VALORE VERO

La misurazione è il Procedimento che permette di assegnare una misura. La misura è l'attribuzione di un valore numerico ad una proprietà rispetto all'unità di misura. Il Numero reale è il risultato di una misurazione, il Valore vero è quello che noi produciamo, stime di Teta.

X=numero reale, risultato di una misurazione. 0= valore vero TETA, dal quale produciamo una stima.

La MISURA: assegnazione di un valore numerico ad una proprietà (unità di misura). La misurazione: procedimento che porta all'assegnazione della misura specifica.

Tipi di errori:

- ERRORE SISTEMATICO: causato dalle condizioni di esecuzione del procedimento di misurazione, dipende dall'accuratezza e dalle stime di accuratezza. (es strumento calibrato male) Posso sovrastimare o sottostimare il vero valore di teta. Ci indica l'accuratezza di una misura. (σ=mu-vero valore di TETA)
- ERRORE CASUALE: Non è possibile stabilire se l'errore avrà grandezza sup o inf al dato reale. Si distribuisce intorno alla media o in difetto o in eccesso. Le stime di precisione sono basate su ro (lettera greca) del campione. E' causato da fenomeni aleatori (es errori di lettura, fermare il cronometro prima/dopo). Ci indica la precisione di una misura. (E=xi-mu cioè la differenza tra la singola misurazione e la media di tutte le misurazioni) Si riduce con la ripetizione della misurazione : Esempio, se ho più giudici che prendono le misure per lo stesso atleta, per avere una misura precisa farò la media di tutte le misure prese dai giudici e avrò così il valore più accurato.
- ERRORE TOTALE: è quello più attendibile. E' la differenza tra il valore misurato e quello vero. E' la somma dell'errore sistematico e quello casuale.

Per stimare l'accuratezza di una misura bisogna: fare la media (x segnato) poi calcolare l'errore sistematico che, se è positivo ho una sovrastima, se negativo ho una sottostima. Poi devo esprimere l'errore sistematico in percentuale. Poi, sottraendo il valore percentuale a 100, vedo di quanto è accurata la misura. (deviaz st. diviso media tutto moltiplicato per 100)

Per stimare la precisione della misura devo calcolare sigma (deviaz.stand) ma al posto di x segnato, devo usare il valore vero di teta. Il calcolo dell'imprecisione in percentuale è il valore dell'errore casuale fratto teta e moltiplicato per 100. Il valore poi di quanto è accurato lo trovo, sempre in percentuale, sottraendo a 100 il valore in percentuale dell'errore casuale stimato.

TABELLE E FREQUENZE ASSULTE, RELATIVE E CUMULATE

Le tabelle ad 1 entrata mi permettono di fare un'analisi MONOvariata, considera una variabile alla volta. Le tabelle a DOPPIA entrata mi permettono di fare un'analisi BIvariata, variabile di riga X e variabile di colonna Y,

mette a confronto le due variabili per valutare dipendenza/indipendenza e il tipo di relazione che esiste tra loro.

La distribuzione è l'insieme dei valori di una variabile e delle sue frequenze.

Nella tabella BIVARIATA abbiamo:

- La FREQUENZA ASSOLUTA (ni): il numero di volte in cui ricorre un valore (ogni modalità o intervallo di classe) in una distribuzione, praticamente il numero che trovo nella cella. La FREQUENZA ASSOLUTA è il numero di osservazioni che corrispondono ai diversi valori (modalità\intervalli di classe) della variabile. Calcolare Ni vuol dire contare quante volte ricorre un valore
- DISTRIBUZIONI MARGINALI: permettono l'analisi delle due variabili x e y separatamente. Sono i totali delle righe (contenute nell'ultima riga di x) e totali di colonna (nell'ultima colonna Y).
- DISTRIBUZIONI CONDIZIONATE: fissata una modalità della prima variabile (carattere), le distribuzioni condizionate sono le distribuzioni che assume la seconda variabile. Non possono essere confrontate tra loro perché si riferiscono a totali marginali diversi.
- DISTRIBUZIONI PERCENTUALI: non si sommano. Se sono distribuzioni percentuali condizionate di riga: devo dividere ogni x di una riga per il totale di quella riga marginale, tutto moltiplicato poi per 100. Se sono distribuzioni percentuali di cella sul totale devo dividere ogni valore, di tutta la tabella, per n e poi moltiplicare per 100.

Una distribuzione di frequenza è l'insieme dei valori di una variabile e del numero di volte (frequenza) con cui rincorrono nel campione. Per la costruzione di una distribuzione di frequenza bisogna definire un criterio di classificazione, ciò che osservo, e dare a ogni valore la frequenza corrispondente.

Il criterio di classificazione → Variabili qualitativa: tipo nominale o ordinale. Variabile quantitativa: è di tipo cardinale. La classificazione deve avere due caratteristiche; deve essere completa e priva di ambiguità (ogni dato deve appartenere a una classe, non può appartenere a +' classi)

La variabile qualitativa (che può essere NOMINALE o ORDINABILE) si esprime in MODALITA', sono prive di significato numerico e le scale di misura a cui sono soggette sono appunto nominale e/o ordinale.

Mentre le variabili Quantitative (è di tipo cardinale) si esprime in intervalli di classe e assumono valori numerici, si distinguono in Discrete (numeri interi) e Continue (infinito n° di valori all'interno di un intervallo, quantità che usiamo numeri con la virgola, es kg 51,52…). Le scale di misura appunto sono di tipo cardinale e ricorrono a intervalli di classe (per quelle continue, per esempio, usando le parentesi quadre, che include il valore, o tonde, che lo esclude)

La FREQUENZA RELATIVA (pi) consente il confronto in campioni di dimensioni diverse. E' il rapporto tra il numero di osservazioni che corrispondono ai diversi valori di Ni della variabile e la numerosità del campione (n). $Pi= Ni/N$. (le frequenze assegnate diviso il totale delle osservazioni) Per calcolare quanto un fenomeno si manifesta su una casistica di 100 osservazioni uso la formula: $Pi\%=Ni/N*100$ che è la frequenza relativa PERCENTUALE.

Le FREQUANZE CUMULATE SI DIVIDONO IN ASSOLUTA O RELATIVA: l'assoluta (fi) è il numero di osservazioni il cui valore è inferiore o uguale ad un determinato valore di Xi (a cascata).

Vale: $Fi(-\infty)=0$ \ $Fi(+\infty)= n$

Quella relativa vale la formula $Pi= Fi/N$ cioè il rapporto tra frequenza assoluta e n. Per la percentuale invece $Pi\%= Fi/n*100$

COME COSTRUIRE GLI INTERVALLI DI FREQUENZA:

1. Trovare il valore minimo e quello massimo cioè Xmin e Xmax
2. Determinare il campo di variazione o RANGE → Xmax-Xmin
3. Sapere qual è il numero degli intervalli (k) che stabilisce lo statista di quanti valori tener conto
4. Determinare AMPIEZZA DEGLI INTERVALLI che si ottiene dividendo il Range/k
5. Costruisco l'intervallo di classe facendo attenzione che abbiano intersezione vuota. Costruisco gli intervalli con le PARENTESI (parentesi quadra = intervallo chiuso includo quel valore, parentesi tonda = intervallo aperto escludo quel valore)
6. Posiziono gli individui nelle classi.

RAPPRESENTAZIONI GRAFICHE

Il Grafico rappresenta in modo sintetico la distribuzione di una variabile in una statistica. Ha una lettura immediata dei dati e delle relazioni tra essi. Il grafico ottimale è accurato, semplice e chiaro. I grafici sono utilizzati in due fasi 1) analisi dei dati 2) presentazione dei risultati.

Per le rappresentazioni di VARIABILI QUALITATIVE uso:

DIAGRAMMA A SETTORI CIRCOLARI O A TORTA: lo uso per variabili qualitative NOMNALI. L'area del cerchio rappresenta la frequenza totale, i settori solo le frequenze delle singole modalità. Le modalità non sono ordinate.

Posso usare anche il GRAFICO A BARRE/NASTRO: lo uso con variabili qualitative ORDINALI. Evita un'erronea impressione di un ordinamento. A barre metto nelle ascisse la modalità e nelle ordinate le frequenze. Su quello a nastro metto nelle ascisse le frequenze e nelle ordinate la modalità. Se uso il grafico SUDDIVISI posso rappresentare più distribuzioni contemporaneamente. Uso questo grafico solo se i valori non sono raggruppati in classi, oppure in classi di uguale ampiezza.

Il GRAFICO FIGURATIVO utilizza simboli che riproducono l'oggetto al quale si riferisce, la figura è la Modalità. E' meno preciso, ha scopi divulgativi.

Per le VARIABILI QUANTITATIVE uso i grafici a ISTOGRAMMA (tridimensionale). E' composto da una serie di barre rettangolari contigue. La loro area è proporzionale alla frequenza. Hanno classi di ampiezza diversa (basi diverse). La base del rettangolo è l'ampiezza dell'intervallo, l'altezza è la densità di frequenza.

Il POLIGONO DI FREQUENZA si ottiene dall'istogramma. Unisco i valori centrali superiori dei rettangoli dell'istogramma creando una LINEA POLIGONALE. Posso confrontare diverse distribuzioni in un unico grafico.

Il GRAFICO PER SPEZZATE: parto dal grafico per punti e li unisco, li congiungo. Evidenzia la continuità tra i valori (serie temporali). Ogni punto corrisponde ad un valore rilevato, posso analizzare più variabili nello stesso momento e associa variabili quantitative.

IL CAMPIONAMENTO

E' importante per le procedure statistiche.

Un campione è un gruppo di partecipanti, trattamenti o situazioni su cui viene condotta una ricerca. Si costruisce un mondo reale ma più piccolo, una semplificazione della realtà. Fornisce un modello semplificato della realtà. Sul campione c'è un metodo di selezione, un concetto di rappresentatività e la generalizzabilità perfetta del modello che ho trovato a una realtà più ampia che è quella della popolazione.

Con POPOLAZIONE O UNIVERSO ci riferiamo ad un insieme di unità di analisi simili tra di loro per una o più caratteristiche che rappresentano l'oggetto di indagine. La popolazione può essere FINTA, tipo la popolazione

dell'india, o INFINITA, tipo i lanci effettuabili con un dado. La popolazione sono quante ne vogliamo identificare, tutto il mondo, chi ha gli occhi blu, chi ha i capelli neri, etc. Popolazione o Universo sono un insieme di unità di analisi simili tra loro per 1 o più caratteristiche. In statistica uso lettere greche per riferimenti alla popolazione (media della popolazione:mu).

Per la Statistica usiamo l'alfabeto latina (Media M) mentre per i parametri ottenuti usiamo l'alfabeto greco (Media μ)

L'obbiettivo di uno studio su campione fare una STIMA (è una procedura, non è una verità assoluta, assegnazione di uno o più valori numerici ad un parametro ignoto che caratterizza la popolazione – su base di dati campionari-) le tecniche di stima sono basate su indici, GLI STIMATORI (che approssimano alla realtà), ottenuti dai dati del campione, che possono essere CORRETTI- se il valore medio corrisponde al valore del parametro della popolazione- O DISTORTI – se il valore non corrisponde, discosta da quello della popolazione-) di una campione della popolazione. Il CAMPIONE è un gruppo di partecipanti o situazioni su cui viene condotta una ricerca. Viene selezionato da un gruppo più grande definito popolazione e deve essere rappresentativo. In statistica uso lettere latine per riferirmi al campione (es. m=media campionaria)

Il campionamento può essere: PROBABILISTICO o NON PROBABILISTICO.

Quello PROBABILISTICO ha un alto grado di rappresentatività. Si estraggono dei numeri dalla Tavola dei Numeri Casuali. Nel campionamento casuale semplice (RANDOM) il campione dei partecipanti viene selezionato da un gruppo più grande definito POPOLAZIONE. I risultati che si ottengono possono essere APPLICATI, INFERITI O GENERAZLIZZATI tra i valori trovati nel campione e quelli della popolazione. Il campionamento probabilistico può avvenire in modo:

- Casuale semplice (RANDOM): il campione viene scelto casualmente estraendo numeri dalla tavola dei numeri casuali, selezione casuale di un campione di minimo 30 unità da un gruppo definito popolazione. Ha un alto grado di inferenza statistica rispetto alla popolazione di studio. E' un buon modello. Più il numero che campiono è alto più sarà precisa la statistica. Si usa anche Excel per estrarre il campionamento. C'è la funzione =casuale.tra() si può fare a uno o più stadi.
- Stratificato: a strati (gradi)
- Per clauster (aree)
- Sistematico

Il campionamento NON probabilistico è meno potente dal punto di vista statistico perché si sceglie su cosa campionare. Può avvenire:

- Per quote
- Per scelta ragionata
- Per testimoni privilegiati

LE MISURE DI TENDENZA CENTRALE sono: Media (limiti operativi), Mediana, Moda. Danno informazione non esaustiva del fenomeno, sono un valore ipotetico. Valori intuitivi, valori di posizione (mediana). La moda è la frequenza con cui si manifesta un valore.

LE MISURE DI DISPERSIONE sono: Devianza, Varianza, Deviazione Standard, Coefficiente di variazione. Ci fanno capire come un campione attorno alla media, mediana, moda si distribuisce a dx , a sx osservando il grafico.

LE MISURE DI VARIABILITA'

La STATISTICA DESCRITTIVA studia i fenomeni che variano all'interno della POPOLAZIONE STATISTICHE che è formata da unità statistiche. L'unità statistica (persone, imprese, famiglie, oggetti..) è l'elemento base della popolazione è definita in termini di contesto, spazio, tempo e su di essa viene rilevata la CARATTERISTICA oggetto di studio.

La STATISTICA INFERENZIALE interviene sulla presentazione dei dati raccolti. Generalizza le informazioni raccolte nel campione per individuare le PROPRIETA' GENERALI della popolazione.

SCHEMA:

1-POPOLAZIONE → STATISTICA DESCRITTIVA → 2-CAMPIONE →CAMPINAMENTO PROBABILISTICO(ad uno stadio o a più stadi) → INTERVALLI DI CONFIDENZA→ 3- PRESENTAZIONE DEI DATI →STATISTICA INFERENZIALE → 4-STUDIO DELLE PROPRIETA' DELLA POPOLAZIONE

Per cui scelgo un campione, scelta di tipo probabilistico, estratto a sorte tramite randomizzazione. Deve essere rappresentativo della popolazione di riferimento. Il modo in cui si sceglie di estrarre le unità statistiche influenza la precisione delle stime, in cui avviene l'intervallo di confidenza, in cui trovo i valori della popolazione di partenza. Le caratteristiche del campionamento sono: completezza, conoscenza della probabilità della selezione, efficienza. Ci sono 2 tipi di campione: a 1 stadio e a 2 o più stadi. A 1 stadio si ha una sola estrazione di unità campione (più semplice da trattare). 2 o più stadi avvengono almeno 2 estrazioni di unità campione, gerarchicamente ordinabili (es. regione, provincia, comune, famiglia..) solo l'ultima estrazione interessa per l'unità statistica dell'indagine. (è il più usato).

VARIABILITA' A DUE DIMENSIONI

Si usa tabella a doppia entrata: permette la rappresentazione congiunta di 2 variabili e le mette in relazione. Passaggi: 1- divido le unità statistiche in modalità o classi per entrambi i caratteri in modo indipendente 2 – riporto queste modalità/classe nella testata della tabella x per il primo carattere e nella colonna madre per il secondo carattere. 3 – riporto i valori dei due caratteri nelle celle.

Distribuzione: insieme dei valori di una variabile e delle sue frequenze

Distribuzione bivariata – (analizza due variabili contemporaneamente) si usa o l'istogramma o il diagramma di dispersione

Distribuzioni marginali: solo i totali delle righe (contenute nell'ultima riga x) e delle colonne (nell'ultima colonna Y)

Distribuzioni parziali: frequenze riportate o nella colonna o nella riga. Se indico solo una delle due

Quando le caselle sono eccessivamente numerose per essere riportate in una tabella, occorre raggruppare in classi almeno una varabile.

LA RELAZIONE TRA DUE CARATTERI

Lo scopo della raccolta dei dati è la ricerca delle relazioni di tipo ca usa-effetto tra fenomeni per interpretare e prevedere.

In una distribuzione BIVARIATA o multipla per studiare la reazione bisogna tener conto di due aspetti fondamentali: la natura delle variabili, se qualitativa, quantitativa etc e il tipo di relazione che voglio rilevare. Ci sono 2 tipologie di relazione di dipendenza di 2 caratteri che studio: logica cioè tra i 2 caratteri esiste una relazione di tipo causa\effetto oppure statistica, cioè tra i due caratteri esiste una regolarità tra le associazioni, tra le diverse modalità, quindi ad es. esiste una periodicità nell'associazione delle diverse modalità dei caratteri. L'indipendenza logica NON implica indipendenza statistica. Gli obbiettivi sono: stabilire la dipendenza tra x e y per cui non una dip logica ma di tipo statistico, valutare l'intensità della dipendenza, capire come raggiungere gli obbiettivi.

Riassumendo - La dipendenza di 2 caratteri può essere:

- Dipendenza logica (causa-effetto- c'è una relazione di causa effetto tra i due caratteri)
- Dipendenza statistica (regolarità nell'associazione tra modalità – la modalità di una dei due caratteri influenza la modalità assunta dall'altro carattere)
- Dipendenza Parametrica: c'è una dipendenza in media tra variabili quantitative e non è simmetrica
- Dipendenza Funzionale: studia una variabile in funzione dell'altra (andamento sul piano cartesiano della dipendenza); quale forma funzionale ha la dipendenza → concorde o discorde
- Indipendenza: quando la distribuzione condizionata X non cambia al variare delle modalità di Y (e/o viceversa) l'indipendenza si dice simmetrica quando X è indipendente in distribuzione da Y e viceversa.

Attenzione!! → Se esiste una dipendenza logica NON per forza esiste anche una dipendenza statistica. L'una NON implica l'altra.

Se la distribuzione condizionata x/y non cambia al variare delle modalità di Y allora si dice che la variabile X è indipendente in distribuzione da Y. E' simmetrica per cui se X è indipendente da Y allora anche Y lo è da x. Le variabili sono connesse se non si è in grado di costruire la frequenza congiunta di almeno 1 elemento a partire da quelle marginali. Si parla di dipendenza assoluta dei due caratteri. Per studiare e misurare la dipendenza tra due caratteri occorre studiare le CONTINGENZE Cij che è lo scarto tra la frequenza osservata in una cella e la frequenza teorica che si osserverebbe se le 2 variabili fossero completamente indipendenti. Due caratteri sono connessi se esiste una cella per cui il valore di Cij è diverso da 0. L'indice CHI-QUADRO (χ^2) SI USA PER STABILIRE IL GRADO DI CONNESSIONE TRA DUE VARIABILI QUALITATIVE, è la media ponderata delle contingenze al quadrato, è un indice simmetrico che misura la connessione reciproca tra x e y. Con il CHI QUADRO non è possibile studiare il livello di connessione tra 2 variabili in un campione e quello delle stesse variabili in un altro insieme di dati per cui si usa V DI CRAMER che è l'indice chi quadro normalizzato.

INDIPENDENZA DI UNA DISTRIBUZIONE

Se la distribuzione condizionata x\y non cambia al variare delle modalità di y (ovvero restituisce valori tutti uguali), allora la variabile x è indipendente in distribuzione da y.

L'indipendenza è simmetrica: se x indipendente da y anche y indipendente da x

INDICI DI DIPENDENZA TRA 2 CARETTERI SONO:

- Contingenza
- Indice di Mortara
- Chi-quadro
- V di Cramer
- Dipendenza parametrica
- Dipendenza Funzionale

LE CONTINGENZE: le sono indicano con C_{ij}

Come lo calcolo? E' lo scarto tra ogni frequenza osservata in una cella e la frequenza teorica che osserverei se le variabili fossero indipendenti (freq. Osservata-freq. Teorica) ; dove la frequenza teorica viene ad essere calcolata come = frequenza marginale di riga*frequenza marginale di colonna/ totale

se Cij = 0 variabili indipendenti

se Cij > 0 connessione positiva(se aumenta uno aumenta anche l'altro)

se Cij< 0 connessione negativa(uno aumenta l'altro diminuisce)

E' un indice percentuale. F.osservata-F.teorica= Cij → grado di indipendenza dei due caratteri

NB: due caratteri sono contingenti se e solo se esiste una cella per cui cij è diversa da 0.

INDICE DI MORTARA: media ponderata dei rapporti di contingenza in valore assoluto (da 0 a 2) → somma contingenze cij. Se più vicino a 0=0 variabili indipendenti se più vicino a 2=2 variabili dipendenti. Tutte le contingenze sommate (in modulo → positiva +) ÷ n

$$M = \sum_{i=1}^{m}\sum_{j=1}^{s}\left|\frac{(f_{ij} - f_{i.}f_{.j})}{f_{i.}f_{.j}}\right| f_{i.}f_{.j} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{s}\left|c_{ij}\right| = \begin{cases} 0 & \text{Indipendenza dei due caratteri} \\ 2 & \text{Dipendenza dei due caratteri} \end{cases}$$

INDICE CHI-QUADRO χ^2: si usa per stabilire il grado di connessione tra 2 variabili qualitative, ci fornisce una misura della dipendenza tra due caratteri (x e y). E' una media ponderata delle contingenze al quadrato. E' un indice simmetrico che ci fornisce la misura della dipendenza tra x e y.

Formula:

χ^2 = ∑ (frequenza osservata – frequenza teorica)² ÷ frequenza teorica

Freq.Teorica= frequenza Marginale di riga × frequenza Marginale di colonna ÷ totale osservazioni

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{s}\frac{(n_{ij} - \dot{n_{ij}})^2}{\dot{n_{ij}}} = n\left[\sum_{i=1}^{m}\sum_{j=1}^{s}\left(\frac{n_{ij}^2}{n_{i.}n_{.j}}\right) - 1\right] = \begin{cases} 0 & \text{Indipendenza dei due caratteri} \\ \text{non nullo} & \text{Dipendenza dei due caratteri} \end{cases}$$

Nel caso di dipendenza tra due caratteri:

$$\max \chi^2 = n \cdot \min[(m-1);(s-1)]$$

V di CRAMER: è l'indice CHI-QUADRO normalizzato (quanto è forte la dipendenza, da 0 a 1) Se =0 Indipendenza; se diversa da 0 dipendenza

Max χ^2 = totale × (n° delle variabili X-1)

Freq.assoluta × (tulle le variabili di colonna-1) × (tutte le variabili di riga -1)

$$V = \frac{\chi^2}{\max \chi^2} = \frac{\chi^2}{n \cdot [(m-1);(s-1)]} = \begin{cases} 0 & \text{Indipendenza dei due caratteri} \\ 1 & \text{Dipendenza dei due caratteri} \end{cases}$$

n=numero del campione che viene moltiplicato per il valore più basso tra : conto il numero di colonne della tabella e sottraggo 1, poi conto il numero delle righe della tabella e sottraggo 1 e scelgo il numero più basso tra i due.

COVARIANZA E CORRELAZIONE

Il chi quadro è l'indice di dipendenza assoluta. Altri tipi di dipendenza sono: Dip. Parametrica e dip. Funzionale.

Dipendenza parametrica → DIPENDENZA IN MEDIA (modello più diffuso) → NON E' SIMMETRICA a differenza della dipendenza assoluta

La dipendenza in media è la misura della connessione di un carattere (variabili) quantitativo y con un altro carattere qualunque di x.

Se la media (Y/Xi)=M (Y) per ogni Xi allora Y è indipendente in media da X.

Se M(x/Yi)=M(x) per ogni Y allora X è Indipendente in media da Y

M=media aritmetica

VEDI ESEMPIO SLIDE PAG 5 delle lezioni in piattaforma

Ci sono degli indici che permettono di misurare il tipo di dipendenza, o meglio, il livello e il grado di dipendenza di due caratteri.

INDICE DI PEARSON (per la dipendenza in media): Può ESSERE 0 O 1

Per misurare la dipendenza in media si ricorre al rapporto di CORRELAZIONE detto anche INDICE ETA QUADRATO DI PEARSON: si fa riferimento ai concetti di devianza= la devianza di una variabile Y rispetto all'altra variabile x, che si può scomporre in devianza interna, esterna e totale.

Devianza interna: $\sum(Xi\text{-f.media di x})^2 \times$ F.assoluta

Devianza Esterna : $\sum$(f.media di una riga- F.media tot)$^2 \times$ F.marginale

Devianza Totale: Devian. Interna + Devianza Esterna

L'indice di Pearson si calcola: Devianza esterna fratto devianza Totale oppure 1- devianza Interna fratto devianza totale, è il rapporto tra dev.est e dev.tot.

Se la devianza esterna di Y è nulla si ha l'indipendenza in media di Y da X. Se la devianza interna di Y è nulla si ha la dipendenza in media massima di Y da X (lo stesso vale per x in funzione di y)

La Proprietà di r è → $-1 \leq r \leq 1$ se -1 x e y sono dipendenti\lineare discorde se 0 x e y sono indipendenti. Non lineare se +1 x e y sono dipendenti\lineare concorde . Più il valore di r si avvicina a 0 più x e y sono indipendenti, più si avvicina a -1, +1, più sono dipendenti. L'obbiettivo è <u>stabilire se esiste una correlazione tra l'andamento delle due variabili</u>, ossia se al variare di una la corrispondente variazione dell'altra segue un andamento di tipo concorde o discorde, o se i due andamenti sono +/- dipendenti.

DIPENDENZA FUNZIONALE: studia una variabile in funzione dell'altra.

E' una dipendenza di tipo lineare e si stabilisce solo tra caratteri quantitativi (a differenza della dipendenza assoluta che si pu stabilire tra caratteri di qualsiasi tipo). Si considera un diagramma di dispersione in cui l'origine degli assi sia stata traslata sul baricentro. Classifichiamo le coordinate dei punti nei 4 quadranti. Nel primo quadrante si ha gli Xi e gli Yi sono positivi e nel terzo quadrante sono entrambi negativi → coordinate e valori sono cooncordanti in segno mentre; mentre nel II e le IV quadrante sono discordi in segno.

CODEVIANZA: E' la somma dei prodotti XiYi e sintetizza la distribuzione dei punti nei 4 quadranti. La codevianza dipende dalla numerosità del campione e dalle Unità di misura.

CODEV. (X,Y) = $\sum_{i=1}$(xi-x segnato)×(Yi-Y segnato) che è il prodotto degli scarti di tutti gli X con quelli di Y oppure n×covarianza

Per il calcolo del coeff. Di correlazione ci sono da fare due passi - **1° passo**:

COVARIANZA (devo dividere la codevianza per la dimensione campionaria): serve per eliminare la dipendenza dalla numerosità del campione nel calcolo della correlazione tra due variabili. E' un numero che fornisce una misura di quanto le due variabili varino assieme (la loro dipendenza). E' il valore atteso dei prodotti delle loro distanze dalla media: formula

> σxy= codev (x,y) fratto n cioè la sommatoria di tutti : (x1-media di x) per (y1-media di y) diviso il n osservazioni att.ne lo faccio per ogni singolo fattore poi li sommo

La covarianza è già di per se un indice di tipo di correlazione. Ad esempio se è di segno negativo si dice che esiste una dipendenza di tipo discorde (se una aumenta l'altra diminuisce) tra le due variabili.

Se =0 allora i due punti sono uniformemente distribuiti sui 4 quadranti

Se >0 prevalgono i punti del I e III quadrante, i caratteri sono concordanti

Se <0 prevalgono i punti di II e IV quadrante, caratteri discordanti

Poi il **2° passo** rendo la covarianza indipendente dalle unità di misura, per cui divido la covarianza per gli scostamenti quadratici medi (o dev.st) delle due variabili e ottengo:

IL COEFFICENZE DI CORRELAZIONE DI BRAVAIS-PEARSON (r) : serve per eliminare la dipendenza dalle unità di misura nel calcolo della correlazione e ci indica se esiste CORRELAZIONE è FORTE o DEBOLE. Date due variabili x e y, l'indice di correlazione di Pearson è definito come la covarianza diviso il prodotto delle deviazioni standard. Delle due variabili. E' un indice che esprime una eventuale relazione di linearità tra le due variabili. Indica se una correlazione è forte o debole.

r = Covarianza diviso (deviazione standard (x) * deviazione standard (Y)) = σxy/σx*σy

$-1 \leq r \leq +1$

R = 1 (si avvicina a 1) la relazione tra x e y è lineare e concorde : MASSIMA CONCORDANZA retta crescente

R = 0 x e y sono indipendenti : ASSENZA di concordanza (i punti nel grafico sono disposti casualmente)

R = -1 (si avvicina a -1) la relazione tra x e y è lineare e discorde : MASSIMA DISCORDANZA retta decrescente

Concordanza: al valore più grande di x corrisponde il valore più grande di y o viceversa, al valore più piccolo di x corrisponde il valore più piccolo di y

Discordanza: al valore più piccolo di x corrisponde il valore più grande di y e viceversa.

REGRESSIONE LINEARE

Se le variabili sono correlate tra loro è possibile individuare una retta di regressione: Il procedimento della regressione lineare si usa per effettuare delle previsioni su futuri risultati. Si individuano dei parametri b0 e b1 che definiscono una retta di regressione che descrive l'andamento dei valori che ho. Si usa l'equazione della retta per scopi predittivi.

Y= b0+b * x b0= y + b1 * x b1= codev (x,y)/ dev (x)

INTRODUZIONE ALLA STATISTICA INFERENZIALE

Cerca di generalizzare le proprietà relative al campione estendendole all'intera popolazione, cioè verificare se è possibile trasferire i risultati ottenuti per un campione ad una popolazione più estesa (estendo alla popolazione i dati raccolti con la statistica descrittiva che estrae dalla popolazione, in modo casuale, un campione e su di esso si fanno dei calcoli in base alle diverse operazione viste. In base a questi calcoli vengono rilevate sul campione tutta una serie di proprietà. Poi con la statistica inferenziale cerco di generalizzare queste proprietà alla popolazione. Rimane importane il concetto di rappresentatività del campione rispetto alla popolazione, di quanto potremmo cioè ritrovare nella popolazione in generale di ciò che abbiamo trovato nel particolare del campione). Ci sono dei problemi per la statistica inferenziale: il campione non è mai perfettamente rappresentativo della popolazione per cui ci possono essere errori di campionamento. L'affidabilità del campione dipende dall'errore sistematico che ci fornisce una misura di affidabilità del campione. Il campione, inoltre, non è mai perfettamente rappresentativo della popolazione, è semplicemente un modello che tenta di ricostruire ciò che avviene nel generale (nella popolazione), ci possono essere errori di campionamento.

Per misurare l'effetto del trattamento dei dati che avviene sul campione (e non sulla popolazione) si prende in considerazione la DIFFERENZA tra la media campionaria e il VALORE VERO della media calcolata rispetto all'intera popolazione.

Punto di partenza dell'inferenza statistica è il calcolo degli indici forniti dalla media aritmetica e dalla varianza. Questi indici, media e varianza campionaria, vengono anche detti STIMATORI.

2 importanti proprietà:

1- se il valore medio dello stimatore è uguale al vero valore ignoto della popolazione = **correttezza** del valore

2 – se all'aumentare dei dati del campione il valore dello stimatore converge al vero valore ignoto della popolazione = allora la verifica delle ipotesi è **consistente**.

La verifica delle ipotesi o il test statistico è un procedimento decisionale che, a partire dai dati statistici osservati, consente di stabilire se una determinata IPOTESI è vera o falsa. Serve per conoscere e ridurre l'incertezza del processo decisionale (stima).

L'ipotesi H0 si riferisce sempre ad un parametro della popolazione e non a una statistica campionaria (come la media campionaria) e contiene il segno di uguaglianza relativo al parametro della popolazione.

L'ipotesi alterativa H1 non contiene mai un segno di uguaglianza relativa al valore specificato del parametro della popolazione, ma un segno di diversità. E' inoltre la negazione dell'ipotesi nulla H0.

Considerando una serie di campioni indipendenti provenienti da una certa popolazione, si calcola, per ciascun campione, la MEDIA campionaria. Si può costruire la distribuzione di frequenza facendo valere il teorema del limite centrale: a campioni numerosi corrisponde un grado di approssimazione migliore tanto più la distribuzione di probabilità della popolazione è simmetrica rispetto alla media.

Intervallo di confidenza di livello α della media della popolazione è quella regione che con probabilità $1-\alpha$ conterrà il vero valore della media della popolazione. (z=valore critico)

Formula: (x segnato $- z * \sigma/\sqrt{n}$; x segnato $+ z * \sigma/\sqrt{n}$) $\sqrt{n}$=radice quadrata di n

Z= funzione inversa della funzione di ripartizione della normale, cioè il valore x totale che l'area sotto la normale sia pari ad un ammontare prefissato

Alfa α = livello di significatività del test (α solitamente 0,05) è la probabilità di accettare H0 anche se non è vera. Quando il ricercatore è disposto a sopportare questa accettazione anche nel caso l'ipotesi H0 NON SIA VERA.

σ = lo standard error del campione

n = numerosità del campione

x segnato = media campionaria

La verifica delle ipotesi serve per : conoscere e ridurre l'incertezza del processo decisionale ; controllare il rischio della decisione sulla base delle statistiche campionarie.

Tutti i test statistici hanno degli elementi comuni e un analogo processo decisionale.

- Ipotesi nulla H0 → è l'ipotesi che il ricercatore vuole sottoporre a verifica, si ritiene essere lo status Quo.
- Ipotesi alternativa H1 → è l'ipotesi che sarà accettata in caso di rifiuto dell'ipotesi nulla, spesso accettata come prova di nuova teoria.
- Livello di significatività (α) → probabilità di accettare l'ipotesi nulla, anche nel caso essa sia non vera, che il ricercatore è disposto ad accettare.

CRITERIO DI DECISIONE

Si seleziona un valore critico di significatività α posto a 0,05 che determina una regione critica della distribuzione dove è improbabile che H0 sia vera.

Se H0 è vera → probabilità coincide con la media del campione in una determinata regione della distribuzione.

Processo di decisione.

1 – si definisce il criterio (H0 o H1)

2 – si raccolgono i dati

3 – si calcola la media

4 – si calcola z (media campione – media popolazione / errore standard)

Decisione:

a). si rifiuta H0 (la media del campione si colloca nella regione critica) per cui esiste una forte ≠ tra media campione e media popolazione

b). si accetta H0 (la media del campione si colloca vicina alla media della popolazione)

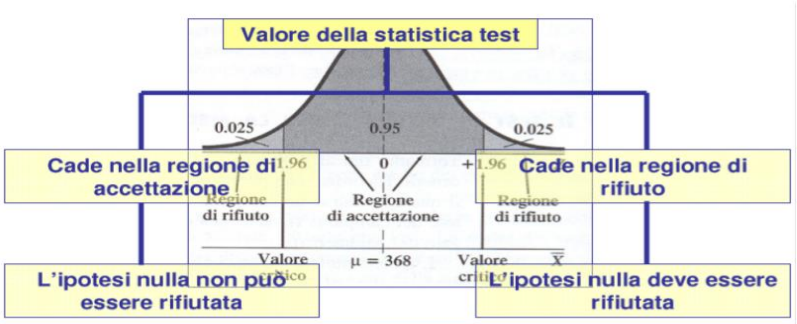
Il TEST STATISTICO

Misura quanto è vicino il valore campionario all'ipotesi NULLA, più è vicino, più accetto H0.

La distribuzione della statistica test è divisa in regione di rifiuto/accettazione, separate dal valore critico z.

Valore della statistica test se cade nella regione di accettazione l'ipotesi nulla H0 NON può essere rifiutata. Se cade nella regione di rifiuto, l'ipotesi nulla DEVE essere rifiutata e accetto l'ipotesi alternativa H1. Es: se test a due code: regione centrale di accettazione H0, due code laterali come rifiuto di H0. Se test a una coda: una sola regione di accettazione per H0, una regione (coda) di rifiuto H0, maggiore sensibilità, maggiori probabilità di errori di tipo 1.

Il Schema logico del processo di decisione



ATTENZIONE!! Per prendere una decisione occorre determinare il valore critico. Questo valore separa la ragione di accettazione (intervallo di confidenza) da quella del rifiuto.

H0 si riferisce sempre a parametri della popolazione. Contiene segno di uguaglianza relativo al parametro della popolazione.

H1 è la negazione dell'ipotesi NULLA. Contiene segno di DISuguaglianza rispetto al valore della popolazione.

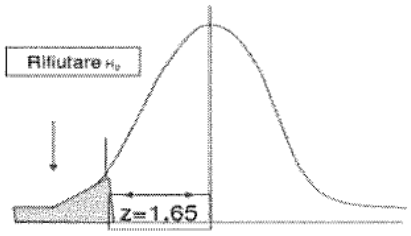
ERRORI INFERENZIALI

Errore di tipo 1 α → rifiuto H0, ma H0 è vera→ coefficiente di confidenza (1 – α) errore PIU' GRAVE

Errore di tipo 2 β → accetto H0 ma H0 è falsa→ potenza di un test (capacità di rifiutare H0 quando è falsa) (1 – β) è meno grave perché le statistiche ammettono un margine di errore.

α è più GRAVE di β perché se rifiuto H0 ipotesi NULLA garantisce la capacità del test statistico di essere in grado di rilevare la situazione in cui H1 sia vera! Il valore limite è pari a 1

Test a una coda: consente di rifiutare H0 anche quando Xsegnato ≠ μ di poco. C'è però un maggiore rischio di errore di tipo 1 (il più grave).



Nel test a 1 coda l'ipotesi H1 è monodirezionale cioè ha UNA SOLA CODA

IL TEST STATISTICO SU UN CAMPIONE

Per applicare questi metodi di INFERENZA PARAMETRICI devono esserci queste tre condizioni:

1. Indipendenza dei gruppi campionari (il campione deve essere estratto casualmente)
2. Normalità delle distribuzioni (solo se la distribuzione è di tipo gaussiano è avvenuta casualmente)
3. Omogeneità delle varianze (uso varianza per il calcolo dell'errore standard)

TEST Z per la media (sigma noto)

- Campione estratto casualmente – ha una maggiore variabilità
- Le osservazioni sono indipendenti tra loro
- La varianza della popolazione è NOTA → σ, μ è noto

Serve per testare se il campione, mostra una media che approssima bene quella della popolazione. E' un test statistico, ammette H0 e H1. La distribuzione è Gaussiana secondo H0. Se Z è vicino allo 0 accetto H0, se è lontano dallo 0 rifiuto H0.

Z= Xsegnato – μ (media popolazione) tutto diviso σ (sigma)/√n errore standard= σ/√N
errore stand. Stimato= s/√n dove s è la deviazione st campionaria e √n = radice quadrata di n

P-VALUE: (livello di significatività osservato). Probabilità che il possibile rifiuto dell'ipotesi nulla H0 sia solo dovuta al caso, dovuto la casualità del campione che ho estratto dall'universo e che non rispecchia perfettamente la situazione reale di quest'ultimo. Perciò indica la probabilità di osservare valori più estremi della statistica del test osservata. Rappresenta quanto H0 è vera, la probabilità di osservare un valore della statistica test uguale o più estremo del valore calcolato a partire dal campione. P-VALUE (P) è compreso tra 0 e 1.

Se (P) ≥ α accetto H0 se (P) < α rifiuto H0 (P)=p-value

Il T-value è il p-value nel test t di student

TEST t di Student (sigma NON noto): quando non ho la varianza della popolazione ed utilizzo quella del campione (S). La statistica t è formalmente identica alla statistica Z, ma stima la varianza della popolazione in base alla varianza campionaria di S. Quando conosco σ (varianza della popolazione) → TEST Z Quando non conosco σ → uso il test t

T= x segnato - μ tutto fratto s/√n s è la deviaz st. campionaria che è = radice quadrata $\sqrt{\sum(Xi-X)^2/(n-1)}$

NB: può essere ad una o due code

Altri test: - Test sulla normalità: verifica se il campione esaminato mostra o meno la propensione a seguire il modello probabilistico proprio della distribuzione di probabilità normale - Test 1 – proportion: testa se in certo campione il numero di elementi con una certa caratteristica rispettano statisticamente una proporzione fissata. Test non sulla media ma su un campione che rispetta una proporzione.

IL TEST SATISTICO SU DUE O PIU' CAMPIONI

Serve per il confronto tra 2 o più campioni, sia per misure indipendenti su <u>campioni diversi</u>, sia per <u>misure ripetute</u> (appaiate) sullo stesso campione. Test misure Indipendenti: campioni diversi, es. campione di uomini e campione di donne, oppure squadre diverse. Test a Misure dipendenti (ripetute): stesso campione visualizzato in due tempi diversi, es. stesso campione analizzato nel 2011 e nel 2015, oppure la capacità di svolgere un compito prima e dopo l'allenamento.

Popolazione 1: μ1 σ1 Popolazione 2: μ2 σ2

SCHEMA DI IPOTESI: H0: μ1= μ2(μ1-μ2=0) H1: μ1?μ2 (μ1-μ2≠0)

Test Z per campioni indipendenti:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Test t per campioni indipendenti:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Verifica delle ipotesi si basa sulla differenza tra le due medie campionarie con il test t. Con il test z serve per verificare la differenza tra le due medie ed è basato su varianza campionaria di tipo ponderato.

Processo decisionale:

1. Fissare α
2. Misuro differenze tra medie campioni (x1-x2) (attenzione x segnato, cioè media di x1 e x2)
3. Calcolo differenza tra questa e quella prevista da H0 – (μ1-μ2)
4. Calcolo t\z
5. Decido se accettare o rifiutare H0

VARIANZA PONDERATA S^2p= varianza 1+varianza 2 tutto fratto (n1-1) + (n2-1)

Vantaggi del test a misure ripetute: si può considerare un numero minore di soggetti rispetto ad un test con misure indipendenti su campioni. Considerando misure ottenute dallo stesso campione si eliminano gli effetti delle differenze individuali tra soggetti, per cui le misure sono più attendibili. Inoltre si ha una riduzione della varianza campionaria con un aumento della rappresentatività del campione.

Svantaggi del test a misure ripetute: Persistenza: si confrontano gli effetti dei trattamenti sullo stesso gruppo, quindi non si cambia mai campione. Inoltre l'effetto del primo trattamento persiste alterando i risultati relativi al secondo trattamento. Effetti progressivi: si può avere un peggioramento o un miglioramento del punteggio dovuto alla stanchezza o a effetti non specifici di apprendimento.

Analisi della varianza - **ANOVA**

Si utilizza questo test per verificare se due campioni dimostrino o meno di avere la stessa varianza. La popolazione deve essere distribuita normalmente. Le osservazioni indipendenti e i campioni con la stessa varianza. L'analisi della varianza consente la valutazione delle differenze tra i valori MEDI per due o più trattamenti, o due o più popolazioni. E' più efficiente del test t in quanto permette di analizzare più di 2 campioni. Inoltre si può trattare un'Anova a più fattori se si vuole confrontare ad es l'effetto di un nuovo metodo di allenamento rispetto ad un metodo standard.

L'analisi della varianza si divide in: Anova a un solo fattore oppure a due fattori.

L'anova valuta l'effetto dei fattori (variabili indipendenti e quasi dipendenti) sulla variabile dipendente.

Nota: le variabili indipendenti = sottoposte a manipolazione sperimentale da parte del ricercatore; le variabili quasi dipendenti = usata per distinguere tra diversi gruppi di risultati; variabili dipendente = quando il valore è determinato dai due fattori precedenti ed è una variabile il cui valore è determinato da quello dei fattori o variabili indipendenti (anova valuta l'effetto dei fattori indipendenti sulla variabile dipendente, cioè quanto le variabili indipendenti impattano sulle variabili dipendenti)

Anova a UN SOLO fattore con misure indipendenti: si confrontano due stime indipendenti della varianza della popolazione attraverso il test F.

Test F → H0: che le medie μ dei gruppi in analisi siano uguali tra loro

H1: che almeno due medie delle popolazioni dei trattamenti che siano diverse tra loro

F= varianza delle medie / varianza ipotizzata → varianza tra campioni/varianza interna dei campioni

F≈1→ trattamento non ha effetto F≠1 → trattamento ha un effetto significativo

Il valore F è un rapporto tra le varianze (sempre positive).

Il rapporto F restituisce e stesse informazioni del test T con la differenza che mentre il test T si basa sulla differenza ra due medie il rapporto F si basa sulla varianza di un insieme di 2 o più medie.

Procedimento Anova:

1) Deviazione quadratiche (popolazione/tra campioni/nei campioni)
2) Individuo gradi di libertà
3) Varianza
4) Calcolo F
5) Decido (accetto o rifiuto H0)

1-. Procedimento per il calcolo delle deviazione quadratiche:

1) Devianza quadratica totale ∑(Xi-Xsegnato)² lo faccio per tutti i dati poi li sommo (x segnato è la media totale)
2) Devianza quadratica interna ∑(X1-X1segnato)² + ∑ (X2-X2segnato)²….(devianza dei singoli gruppi poi li sommo) x1 prima riga, x2 seconda riga..
3) Devianza quadratica fra campioni → Dev. Quandratica totale – Dev. quadratica interna

2-. Gradi di libertà: I **gradi di libertà di** una variabile aleatoria o **di** una **statistica** in genere esprimono il numero minimo **di** dati sufficienti a valutare la quantità d'informazione contenuta nella **statistica**. Infatti, quando un dato non è indipendente, l'informazione che esso fornisce è già contenuta implicitamente negli altri.

1) GDL Tolate: N-1 (numerosità totale del campione – 1) (conto tutti i valori di tutte le osserv. E tolgo 1)
2) GDL interni = GDL1 (gdl1-1) + GDL2 (gdl2-1)+…(numerosità di ogni gruppo – 1, sommati) (per ogni campione conto le osservazioni e tolgo 1 poi i risultati ottenuti li sommo)
3) GDL fra campioni = GDL Totale – GDL Interni

3-. Varianza: devianza /GDL

1) Varianza fra campioni= deviaz.fra/GDL fra
2) Varianza Interna= Devia.interna/GDL interni

4-. Calcolo F = varianza fra/varianza interna (se simile a 1 il tratt ha avuto effetto se diversa non ha avuto eff)

5-. α = 0,05 → F critico = 3,89 tabella→ GDL/α trovo il valore critico

F > valore critico tabella rifiuto H0 F< valore critico tabellare accetto H0

Validità di Anova a un solo fattore: occorre assumere che la popolazione sia distribuita normalmente; le osservazioni fatte devono essere tra loro indipendenti; i campioni devono avere la stessa varianza.

ANOVA A DUE FATTORI:

Studia quanta parte della varianza dipenda dal primo fattore , dal secondo fattore e dalla loro interazione (questi due effetti prendono il nome di effetti principali). L'anova a due fattri si compone di due passi fondamentali:

Primo passo: è lo studio degli essetti principali

-Analisi della Varianza per effetto di A – test F su A

- Analisi della varianza per effetto di B – test F su B

Secondo passo: è lo studio della presenza o meno di effetti dovuti all'interazione tra i due fattori A e B:

- Presenza di interazioni → il fattore A esercita il suo effetto solo in presenza del fattore B

- Assenza di iterazioni → i fattori A e B esercitano il loro effetto in modo autonomo l'uno dall'altro, cioè tutte le osservazioni sono spiegate dagli effetti principali.

Formulazione ipotesi: H0: tutti i valori osservati possono essere spiegati esclusivamente in termini degli effetti principali. H1: esiste almeno un valore che non può essere spiegato solo in termini di effetti principali, ma da interventi dovuti ai fattori A e B.

PROCEDIMENO DECISIONALE:

1- Analisi della varianza per effetto di A

2-Analisi della varianza per effetto di B

3-Analisi della varianza per interazione degli effetti di A e di B

Alla fine decido se accettare H0 o rifiutarla e accettare H1.

Condizioni di validità per Anova a due fattori: sono le stesse per l'anova a un fattore. Cioè:

-Occorre assumere che la popolazione sia distribuita normalmente; che le osservazioni fatte siano indipendenti tra loro; che i campioni abbiano la stessa varianza.

METODI NON PARAMETRICI

Non fanno parte della statistica descrittiva e inferenziale. Hanno bisogno di ipotesi meno restrittive della statistica classica. I Metodi NON parametrici: funzionano quando non è possibile ricorrere a metodi parametrici della statistica classica. Nella statistica non parametrica i test non dipendono dalla forma della distribuzione. Non testano parametri della popolazione. Hanno il vantaggio di essere utilizzati su qualsiasi tipo di dato, sono più economici, ma hanno lo svantaggio di essere difficili da usare su campioni numerosi. Sfruttano in modo meno completo l'informazione contenuta nei dati. Per la differenza tra due proporzioni si usa il test z o il test chi quadro. Il test z serve per verifiche di ipotesi direzionali (si basa sulla differenza tra proporzioni); mentre il test chi quadro viene ad essere applicato su un caso di c campioni (su tabelle 2x c; tabella a doppia entrata con gruppi e le modalità della variabile di interesse)

Nella statistica classica i test verificano i parametri ipotizzati. Le ipotesi su cui si basano i test sono FORTI (NORMALITà, CASUALITà, INDIPENDENZA).

Mentre nei metodi NON parametrici i test Non testano parametri della popolazione. NON hanno bisogno di ipotesi moto forti. I vantaggi di questi test sono: si possono usare su qualsiasi tipo di dato; sono efficaci; facili

da usare su piccoli campioni; economici e veloci. Gli SVATAGGI sono: difficile da usare su campioni molto numerosi; sono meno completi.

I test NON parametrici consentono la verifica di ipotesi relative a :

- Variabili NON numeriche (nominali, ordinabili)
- Distribuzioni diverse da quella normale oppure ignote
- Variabili qualitative

CONFRONTO TRA POPOLAZIONI: differenza

$$Z \simeq \frac{(p_{s_1} - p_{s_2}) - (p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Test Z

Ps: proporzione di successi campione

X= numero successi nel campione

P= proporzione successi popolazione → rapporto del numero successi fratto campionario.

Il test Z si basa sulla differenza tra proporzioni campionarie.

Test χ^2 → χ^2= ∑ (frequenze osservate - freq. Attese)²/ Freq.attese F.attese= proporzione media dei successi

Questo metodo. Il chi quadro, si basa su una tabella a doppia entrata. Le colonne rappresentano i gruppi di dati, le righe la modalità della variabile d'interesse (che assume due valori: successo o insuccesso). I dati corrispondono alla distribuzione di frequenza di una variabile categorica (es. sesso dei soggetti di un campione), quindi una variabile di tipo qualitativo. Il test chi quadro lo uso anche come test di indipendenza. H0=le variabili sono indipendenti H1=le variabili sono dipendenti

H0 → distribuzione attesa. Assenza di differenze rispetto a popolazione.

H1 → distribuzione diversa da quella attesa

χ^2 se ha valori alti allora sono lontani da quelli attesi, se il valore è basso allora i valori osservato sono vicini a quelli attesi.

ATTENZIONE!! I due test portano allo stesso risultato MA

Test Z → ipotesi direzionale

Test χ^2 → tabelle di dimensione 2C

MATCH ANALYSIS (analisi della competizione)

La match analisi negli sport di situazione è una branca della pedagogia sportiva e delle scienze motorie. Essa si focalizza sul comportamento degli atleti in gara e consiste nell'analisi, sotto differenti punti di vista, degli eventi che si verificano all'interno di un match; con la finalità di individuare fattori comuni nelle varie categorie giovanili che sono correlate alla vittoria dell'incontro. Permette di descrivere, classificare, spiegare e predire (su basi probabilistiche) l'andamento del match.

Il match analysis esplica tre azioni principali:

1) Osservazione: visione dei match e raccolta dati.
2) Elaborazione: analisi dei dati con la produzione di report cartaceo
3) Applicazione: dati prodotti utilizzati per produrre strategie o correttivi sulle prestazioni

Esiste di due tipi:

- QUALITATIVA: ci permette di valutare la qualità degli eventi, valutandoli sulla base di parametri definiti.
- QUANTITATIVA: ci permette di capire la quantità degli eventi che si sono verificati all'interno di un match (es. numero risposte vincenti)