Statistica Medica: scienza che si occupa della raccolta, dell'analisi e dell'interpretazione di dati medici

Studi clinici es. EPIDEMIOLOGIA: distrubuzione e determinanti di malattie

si effettuano degli studi → dobbiamo creare un progetto per definire lo studio

1. si seleziona un campione di pazienti
2. possiamo effettuare 2 tipi di studi:
   - OSSERVAZIONALE: il ricercatore si limita ad osservare il campione senza interferire
   - SPERIMENTALE (di intervento): il ricercatore interviene sulle condizioni dei soggetti del camp.
     ↳ in genere l'intervento è RANDOMIZZATO per avere dei dati migliori e per etica

ovviamente si devono conoscere le condizioni dei pazienti

3. si analizzano i dati raccolti per trarre delle conclusioni (reali o concettuali) determinandone anche il grado di incertezza

STATISTICA →
- INTERFERENZIALE: conclusioni su tutta la popolazione da cui proviene il campione (mod. PROBABILISTICI)
- DESCRITTIVA: descrive i dati con tabelle, grafici e indici di sintesi

dai dati si estrapolano quantità che riassumano le caratteristiche essenziali
quantità = statistiche; caratteristiche essenziali = variabili

DATI
- qualitativi: classi distinte senza relazione numerica
  - ordinale (posso assegnare un ordine)
  - nominale (non sono ordinate, es. gr. sang.)
- quantitativi: ~~dati di tipo~~ numerico
  - DISCRETI: numeri interi
  - CONTINUI: range (campo di variazione)

i dati vengono raccolti in tabelle per sintetizzare le osservazioni
↳ da qui vedo la distribuzione di FREQUENZE (insieme delle frequenze di tutte le possibili ~~variabili~~ caratteristiche

frequenze
- ASSOLUTA (n° esatto)
- RELATIVA (rapporto)
- PERCENTUALE (%)

* se alla variabile posso dare un ordine "quantitativo" allora posso trovare anche le FREQUENZE CUMULATE (n° di individui il cui valore è ≤ a quello considerato)

Rappresentazione immediata dei dati = GRAFICI

• DIAGRAMMA a BARRE (BARPLOT): Per ogni categoria si disegna un rettangolo con base costante e altezza pari alla frequenza, staccati ed equidistanti tra loro
se la variabile è ordinata si rispetta l'ordine

• DIAGRAMMA a TORTA:

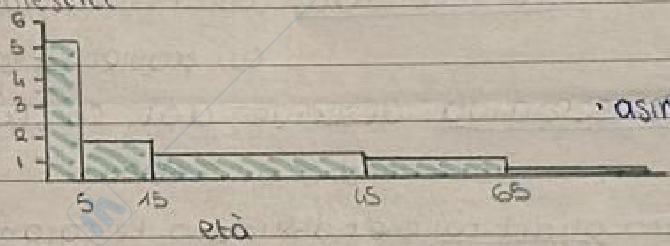
Per le variabili continue → DENSITÀ di frequenza: rapporto tra la frequenza e l'ampiezza della classe

$d_i = p_i / e_i$

nell'istogramma l'area totale sarà 1 e l'altezza = densità di frequenza relativa

es: vittime incidenti domestici

età	fr. rel. %	densità
0-4	25.3	5.6
5-14	18.9	1.89
15-44	30.3	1.01
45-64	13.6	0.68
65+	11.7	0.33



• asimmetrico positivamente

le caratteristiche più rilevanti vengono rappresentate da numeri = INDICI di SINTESI

indici di posizione

• MODA: è il valore più frequente per ogni tipo di dato (unimodale). quando ci sono più mode diverse = BIMODALE o MULTIMODALE

• MEDIA: è la somma di tutti i valori trovati divisa il n delle osservazioni

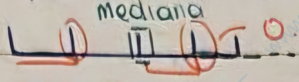
• MEDIANA: è il valore centrale della distribuzione che considero per cui al di sotto di essa trovo la metà delle osservazioni più piccole

• QUANTILE: valore che divide la distribuzione così che al di sotto di esso vengano lasciate un numero p di osservazioni
- ordino n osservazioni → la scala ha n+1 punti
- la porzione che sta sotto l'i-esima osservazione è $i/(n+1)$

$p_{max} = 1$
n = n° osservaz.
i = osservazione

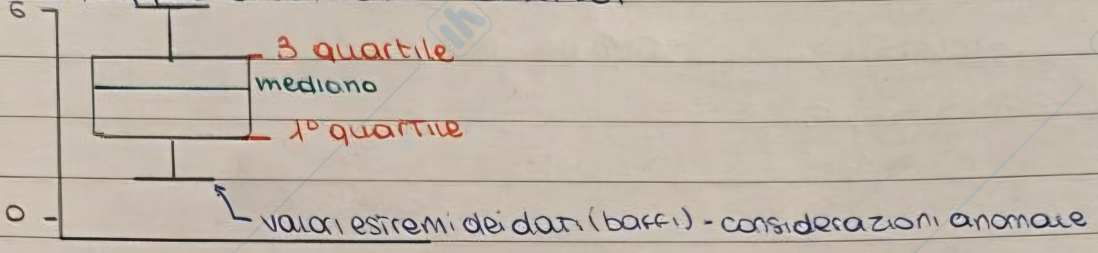
$i = p(n+1)$ trovo l'osservazione che corrisponde al quantile
se i è intero ok, se non è intero il quantile starà tra l'intero (j) e il n° succ.
$x_j + (x_{j+1} - x_j) \times (i - j)$

è più utile dividere il campione (distribuzione) in parti uguali

· in 4 parti = quartile 

· in 5 parti: quintile 

Per la rappresentazione = BOXPLOT



indici di dispersione

· RANGE: intervallo tra il valore minimo e il valore massimo (non è molto aff.)
meglio usare il range interquartile ovvero l'intervallo tra il 1° e il 3° q.

· VARIANZA: calcolo la deviazione della media = differenza tra ogni osservazione e la media
→ elevo ogni risultato al quadrato così da eliminare i problemi di segno
e faccio la Σ = somma dei quadrati intorno alla media
→ li divido per $n-1$ (gradi di libertà)

perciò VARIANZA $S^2 = \frac{1}{n-1} \cdot \Sigma (x_i - x)^2$ oppure $S^2 = \frac{1}{n-1}\left(\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right)$

· DEVIAZIONE STANDARD: per tornare alla stessa unità di misura delle osservazioni
faccio la radice della varianza: $SD = \sqrt{S^2}$
ovvero $SD = \sqrt{\frac{1}{n-1} \cdot \Sigma (x_i - x)^2}$

STATISTICA INFERENZIALE

estende i risultati a tutta la popolazione da cui è stato preso il campione

↳ USA LE PROBABILITÀ: probabilità con cui un determinato evento accada in determinate circostanze

- insieme dei possibili risultati = spazio campionario (es dado {1.2.3.4.5.6})

es A{1.3.5} B{4.5.6}
- evento unione: $A \cup B$ {1.3.4.5.6}
- evento intersezione: $A \cap B$ {5}
- evento complementare $\bar{A}$ non A {2.4.6}

▷ la FREQUENZA RELATIVA è il rapporto tra il numero di volte in cui è possibile il risultato e il numero dei risultati possibili $n(A)/n$ è sempre compreso tra 0 e 1

▷ Se EVENTI MUTUAMENTE ESCLUSIVI allora $A \cap B = \emptyset$ e $A \cup B = P(A) + P(B)$

es dado $p = 6$. la probabilità di avere 1 o 3? $P(1) = 1/6$ $P(3) = 1/6 \rightarrow 1/6 + 1/6 = 1/3$

▷ Se eventi INDIPENDENTI allora $A \cup B = P(A) \cdot P(B)$

es dado $n = 6$ ma lo lancio 2 volte. la probabilità di avere 1 e 3? $P(1)\ 1/6$ $P(3)\ 1/6 \rightarrow 1/36$

- VARIABILE ALEATORIA (CASUALE):
  - DISCRETA = numero finito ≥ 0 e per cui $\Sigma p = 1$
  - CONTINUA = intervallo per cui $P(a<x<b) = \int_a^b f(x)\,dx$ ← densità di probabilità

- MEDIA = μ (valore atteso): è la Σ del valore assunto · probabilità che accada

es. 2 monete, quante volte esce testa?

possibilità =	testa-testa	testa-croce	croce-testa	croce-croce
	$x=2$	$x=1$	$x=1$	$x=0$

per cui $x=2$ $P = 1/4$

$x=1$ $P = 2/4$ media $= 2 \cdot 1/4 + 1 \cdot 1/2 + 0 \cdot 1/4 = 1/2 + 1/2 = 1$

$x=0$ $P = 1/4$

- VARIANZA = σ^2 valore atteso del quadrato degli scarti

$x=0$ $\sigma = x - \mu = -1 \rightarrow |-1| = 1$

$x=1$ $\sigma = 1-1 = 0$ varianza $= (0-1)^2 \cdot 1/4 + (1-1)^2 \cdot 1/2 + (2-1)^2 \cdot 1/4 = 1/2$

$x=2$ $\sigma = 2-1 = 1$

con le variabili discrete "piccole" uso la DISTRIBUZIONE BINOMIALE

$0! = 1$

es. soggetti guariti dopo aver trattato un campione di n pazienti

$n! = n(n-1)\cdot(n-2)\ldots 1$

n: numero di risultati possibili

r = numero di successi

p = probabilità di successo

$$\text{PROB}(r) = \frac{n!}{r!(n-r)!}\cdot p^r \cdot (1-p)^{n-r}$$

es. ~~con [illegible] la probabilità di avere [illegible]~~

in una popolazione in cui il 20% è fumatore, prendendo un campione di 7 uomini quale è la probabilità che 4 fumino?

$r = 4$

$p = 20\% = 0.2$

$n = 7$

$$\frac{7!}{4!(3)!}\cdot 0.2^4 \cdot (1-0.2)^3 = 0.29$$

con le variabili continue uso la DISTRIBUZIONE NORMALE

- è simmetrica rispetto alla media e μ = valore massimo e moda
- $\mu-\sigma$ e $\mu+\sigma$ punti di flesso
- → ha dei valori standard z
- per cui $P(X > x_1) = P\left(z > \frac{x_1-\mu}{\sigma}\right)$ → per cui $= 1 - P\left(z \leq \frac{x_1-\mu}{\sigma}\right)$

es $\mu = 120$ $\sigma = 13$ $P(x > 140)$

$$P(x>140) = P\left(z > \frac{140-120}{13}\right) \to P(z > 1.54) = 1 - P(z \leq 1.54) = 1 - 0.9382 = 0.0618$$

↳ guardo sulla tabella

- se in un range:

$$P(x_1 \leq X \leq x_2) = P\left(\underbrace{\frac{x_1-\mu}{\sigma}}_{z_1} \leq z \leq \underbrace{\frac{x_2-\mu}{\sigma}}_{z_2}\right) \text{ per cui } P(z \leq z_2) - P(z < z_1)$$

es di prima ma $100 \leq X \leq 130$: $P\left(z \leq \frac{130-120}{13}\right) - P\left(z < \frac{100-120}{13}\right) = P(z \leq 0.77) - P(z < -1.54)^*$

inverso, da P voglio x

$\mu = 120$ $\sigma = 13$ $X > x_1 \sim P(X \geq x_1) = 5\% = 0.05$

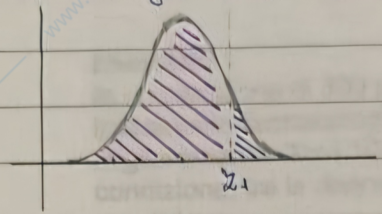
$$P\left(z > \frac{x_1-120}{13}\right) = 0.05 \to 0.05 = 1 - P\left(z < \frac{x_1-120}{13}\right)$$

$$+0.95 = +P\left(z < \frac{x_1-120}{13}\right) = z = 1.654$$

$$1.654 = \frac{x - 120}{13} =$$

NB

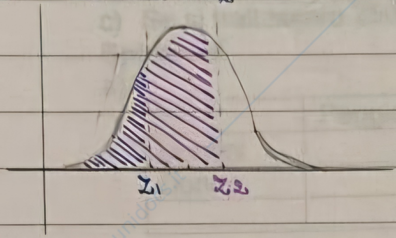
quando devo trovare la densità di frequenza → Paragono la curva alla NORMALE STANDARD per cui ad ogni valore di Z corrisponde l'area della parte $Z < z_1$



per cui $P(X > x_1) = P(Z > z_1)$ $\quad z_1 = \frac{x_1 - \mu}{\sigma}$

trovo z_1 e guardo sulla tabella a quale area mi corrisponde $P(Z > z_1) = 1 - P(Z < z_1)$

Se $x_1 < X < x_2$, $P(x_1 < X < x_2) = P(z_1 < Z < z_2) \rightarrow = P(Z < z_2) - P(Z < z_1)$



INTERVALLI DI CONFIDENZA

▸ misuro l'incertezza per cui la μ_0 della popolazione possa essere la stessa $\bar{x}$ del camp.

$IC = \mu_0 \pm z_{1-\alpha/2} \cdot SE$ se ho $IC = 95\%$ $\alpha = 0.05$ per cui $z_{1-\alpha/2} = \pm 1.96$

$IC = 99\%$ $\alpha = 0.01$ ± 2.58

$IC = 99\%$ $\alpha = 0.005$ ± 2.62

1 solo campione $\mu = \mu_0$ e $SE = \sigma/\sqrt{n}$

1 sola proporzione $\mu_0 = p$ e $SE = \sqrt{\frac{p(1-p)}{n}}$ vedo se $\bar{x}$ cade nella regione critica

2 campioni:

- differenza tra 2 medie $\mu_0 = \bar{x}_1 - \bar{x}_2$ e $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- differenza tra 2 proporzioni $\mu_0 = p_1 - p_2$ e $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

~~[illegible]~~

▸ se il campione è piccolo uso t student → sulla tavola cerco colonna $= n-1$ e riga $= \alpha/2$

e $IC = x_1 - x_2 \pm t_{\alpha/2} \cdot SE$

QUANDO VOGLIO FARE 1 PARAGONE uso STATISTICA TEST

TEST D'IPOTESI H_0 = IPOTESI NULLA → NON CI SONO DIFFERENZE

H_A = IPOTESI ALTERNATIVA → CI SONO DIFFERENZE

▸ $\bar{x} \simeq \mu_0$ popolazione

- IC con $\alpha/2$ e vedo se $\bar{x}$ è nella regione critica → RIFIUTO H_0
- standardizzo $\bar{x} \to z = \frac{\bar{x} - \mu_0}{SE} = z_{OBS}$ $|z_{OBS}| > z_{1-\alpha/2}$ → RIFIUTO H_0
- PVALUE: $2P(z > |z_{OBS}|) \to 2 \cdot (1 - P(z < |z_{OBS}|))$ → $P_V < \alpha$ RIFIUTO H_0
- se σ o μ sono ignoti uso t student

▸ confronto tra 2 campioni

- diff 2 medie con n_1 e $n_2 > 30$

$z_{OBS} = \frac{\bar{x}_1 - \bar{x}_2}{SE}$ e $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- diff. tra proporzioni

$$z_{OBS} = \frac{p_1 - p_2}{SE} \qquad SE = \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \qquad \text{di cui } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

- campioni indipendenti con σ ignota

$$t_{OBS} = \frac{\bar{x}_1 - \bar{x}_2}{SE} \qquad SE = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \quad \text{e} \quad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}$$

- se i campioni sono appaiati (es stesso campione prima e dopo)

$$t_{OBS} = \frac{\bar{d}}{SE} \quad \text{dove } \bar{d} = (\Sigma\, \text{prima-dopo})/n \rightarrow \text{calcolo } \sigma \text{ e } SE$$

Intervallo di confidenza:

se devo stimare la media di una popolazione μ

la miglior stima è la media del campione che esamino $\bar{x}$ (più il campione è grande, meglio)

→ posso misurare l'incertezza: ERRORE STANDARD = $SE = \sqrt{\frac{P(\text{probabilità successo}) \cdot (1-P}{n(\text{campione})}}$

$\bar{X} = \bar{x}_1 - \bar{x}_2$ diff. tra 2 medie $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

$SE = \sqrt{\frac{P(1-P)}{n}}$ $P = P_1 - P_2$ tra 2 proporz. $SE = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$

intervallo di confidenza = livello di fiducia = $(1-\alpha) \cdot 100\%$

NB. nelle distribuzioni campionarie $\sigma = \sigma/\sqrt{n} = SE$ e $z = \frac{\bar{x} - \mu}{SE}$

· se la confidenza = 95% → $\alpha = 0.05$ $\alpha/2 = 0.0025$ $z = \pm 1.96$

$IC = \mu \pm z \cdot SE$ se $\bar{x}$ tale che $\left| \frac{\bar{x} - \mu}{SE} \right| > z_{1-\alpha/2}$

$IC = 95\%$ → $\alpha = 0,05$ $z = \pm 1.96$

$IC = 90\%$ → $\alpha = 0,1$ $z = \pm 1,64$

$IC = 99\%$ → $\alpha = 0,01$ $z = \pm 2,58$

quando il campione è piccolo uso t student → colonna $= n-1$ e riga $= \alpha/2$

poi $IC = \bar{x} \pm t_{obs} \cdot SE$

Per trarre dei risultati su tutta la popolazione, estendo ad essa il risultato del campione perciò

μ campione = μ popolazione $SE = \sigma \text{campione}/\sqrt{n}$ se $n > 30$

lo standardizzo per cui $z = \frac{\bar{x} - \mu\text{campione}}{SE}$

TEST D'IPOTESI

genero un'affermazione circa 1 o più parametri di una popolazione circa la probabilità della distribuzione di un parametro

- H_0 = ipotesi nulla → non ci sono differenze (da confutare)
- H_A = ipotesi alternativa → ci sono differenze (da dimostrare)

in genere pongo:

$$\begin{cases} H_0: \mu_{campione} = \mu_0\, popolazione \\ H_A: \mu \neq \mu_0 \end{cases} \rightarrow \text{per cui } \bar{x} \simeq N(\mu_0 ; SE)$$

mi ricavo un IC in genere 95% per cui $\alpha = 0.05$ e controllo $\bar{x}$ dove cade

→ se cade nella regione critica = RIFIUTO H_0 (5%)

→ se cade nella regione di non rifiuto = NON RIFIUTO H_0 (95%)

NB posso standardizzare per cui $z_{OBS} = \frac{\bar{x} - \mu_0}{SE}$ e $|z_{OBS}| > z_{1-\alpha/2}$ RIFIUTO H_0

$|z_{OBS}| < z_{1-\alpha/2}$ NON RIFIUTO H_0

Posso usare anche P-VALUE

$P.VALUE = 2 \cdot P(Z > |z_{OBS}|) = 2 \cdot (1 - P(Z < z_{OBS}))$

→ se $2 \cdot (1 - P(Z < z_{OBS})) < \alpha$ RIFIUTO H_0

→ se $2 \cdot (1 - P(Z < z_{OB})) > \alpha$ NON RIFIUTO H_0

se ho μ_0 o SE(σ) o entrambe sconosciute con un campione $n < 30$

uso T student:

$t_{OBS} = \frac{\bar{x} - \mu_0}{SE}$ → poi cerco il grado di libertà sulla tavola T

colonna = $n-1$ (o $\Sigma n - nc$) riga = $\alpha/2$

→ se $t_{OBS} > t_{1-\alpha/2}$ RIFIUTO H_0 con Lv significatività α

→ se $t_{OBS} < t_{1-\alpha/2}$ non RIFIUTO H_0 con Lv significatività α

NB. se ho π_0 successi $z_{OBS} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$ e $|z_{OBS}| > z_{1-\alpha/2}$ RIFIUTO H_0 Lv signif. α

$|z_{OBS}| < z_{1-\alpha/2}$ NON RIFIUTO H_0 Lv S. α

NB se i campioni sono appaiati es stesso campione prima e dopo faccio la diff. prima e dopo, le sommo tutte e divido per n, poi mi calcola σ e SE e $t_{OBS} = \bar{d}/SE$

($\bar{d}$)

- se ho 2 campioni indipendenti con μ_2 e μ_1 e $\sigma_1^2 = \sigma_2^2$ ign

$t_{OBS} = \frac{\bar{x}_1 - \bar{x}_2}{SE}$ $SE = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}$ e

dif. tra medie

se n_1 e $n_2 > 30$

$z_{OBS} = \frac{\bar{x}_1 - \bar{x}_2}{SE}$ $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- se proporzioni dif. proporzioni

$z_{OBS} = \frac{p_1 - p_2}{SE}$ $SE = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ e $p =$