STATISTICA

Lezione 1 11/11/2024

La parola statistica deriva dalla parola stato infatti è la diversità nello stato raccontata dai numeri

Dati della popolazione —>Si prende un campione —>ordinati nel database —> distribuzione di frequenza —> interpretazione clinica.

THE SURVIVORSHIP (preso i sopravvissuti) BIAS (errore di ragionamento)

METODO SCIENTIFICO

1. Osservare il fenomeno
2. Formulare domande ——>vedere tramite ricerche se ci sono già risposte
3. Formulare ipotesi
4. Fare gli esperimenti
5. Registrare e analizzare dati
6. Trarre conclusioni

Statistica non è aritmetica, non è una scienza esatta e parla di probabilità.

- UGUAGLIANZA= non molto usata

  Nella proprietà transitiva nella statistica può capitare che: A=B, B=C, C≠A.

  (=)—>non è diverso. (≠) —>diverso

- DIVERSITA'= accetto che ci sia varianza nella biologia

Domanda di ricerca —>la risposta dà l'indirizzo per uscire dal labirinto

SORGENTI DI VARIABILITA':

- Biologia: variabilità tra soggetti. Fattori che rendono diversi i vari individui (età, sesso, ecc…). Tutti siamo diversi da tutti.
- Temporale: influenza la ripetibilità nel tempo. Variazioni nello stesso individuo da un momento all'altro. Cambiamo nel tempo.
- Errore di misura:possono essere sistematici, casuali o artefatti.

Popolazione: persone o oggetti che hanno qualcosa in comune

Campione: sottoinsieme di una popolazione che rappresenta l'intera popolazione. Il numero di membri che compare in essa è detto Dimensione del campione (N).

TIPI DI STUDIO

- Osservazione: misuro qualcosa che già succede. Prendere i dati che sono già stati misurati.
- Sperimentale: misuro qualcosa di nuovo che applico
- Prospettico: dati presi nello studio. Ho deciso tutto prima e poi inizio lo studio.
- Retrospettivo: analizzo dati che già esistevano
- Longitudinale: si misura pre e post qualcosa in un gruppo
- Cross-Sectional: Si misura una volta sola in più gruppi

Nella scienza c'è bisogno di controllo —> controllo trials (verificare che senza causa non c'è effetto)—> Randomized (a caso) —> Blind (paziente ceco in molti casi.

Meta-analisi: prendere varie analisi (vari studi) e si mettono insieme per visione completa

Lezione 2 18/11/2024

Termini sui dati:
- Drop-out: un paziente viene escluso da uno studio o si ritira dallo studio se dopo averlo iniziato non lo completa
- Intention to treat: Se un paziente permane nel gruppo in cui è stato assegnato anche se non completa il trattamento
- Outliers: un paziente che viene escluso a posteriori dallo studio perché risponde in modo diverso dagli altri.

VARIABILI: Grandezze che possono assumere diversi numeri
Tipi di Variabili:
- Nominali —> quelle per cui il loro valore è una categoria che non è possibile riordinare. ES: Genere —> maschio o femmina
- Ordinali —> hanno un ordine. ES: da 1 a 10
- Continue —> seguono l'ordine matematico e sono una misura.
- Discrete —> possono essere contate, fanno parte delle continue.

Scala Likert: Danno una frase e vanno a 5 o a 7 livelli e sono simmetriche rispetto al centro.
Le scale cliniche mettono i pazienti su gradini diversi
Misura ordinale: mette in confronto ciò che è troppo diverso. Effetto soffitto-pavimento.
Non segue regole aritmetiche
Misura Continua: sensibile, oggettiva, statistica parametrica.
Esistono diversi test da usare in base ai dati che si hanno.
La statistica può essere inferenziale e descrittiva (descrive le caratteristiche essenziali dei dati, facendo tabelle e grafici con dentro gli indici—>raccolta di dati in un unico numero, questo avviene tramite: frequenza, tendenza centrale, forma della distribuzione, dispersione). Inoltre può essere parametrica o non parametrica (seguono un ordine).
FREQUENZA:
- Frequenza assoluta: quante volte accade un evento
- Frequenza relativa: è la frequenza assoluta sul numero totale di casi o osservazione (si usa percentuale per non usare virgole)
- Frequenza percentuale: frequenza relativa moltiplicata per 100
- Frequenza cumulata: è la frequenza di una classe sommata alle frequenze di tutte le precedenti o successive.

Esistono vari tipi di tabelle che possono essere usate: Tabelle Open-ended, tabella di contingenza, tabella 2x2.
Distribuzione di frequenza classica: media centrale che allontanandosi scende o sale.
TENDENZA CENTRALE:
Queste misure servono per individuare il valore intorno al quale i dati sono raggruppati, è la misura più appropriata per sintetizzare l'insieme delle osservazione. Le misure sono 3:
- Media (centro aritmetico dei dati, sommatoria dei dati su n) —>ha sia vantaggi che svantaggi. E' il parametro più comune per valutare la tendenza centrale. Se si divide il campione di dati in sotto-campioni omogenei la media delle medie è uguale alla media dell'intero campione. La somma degli scarti tra i vari valori e la media è sempre uguale a 0. La media può essere un valore che non è presente nel dataset. E' molto sensibile alla presenza di outliers.
- Mediana (valore centrale dei dati, prende i dati e dopo averli messi in ordine si prende quello di mezzo —> se pari prendere il numero a destra della metà o media centrale) —>ha sia vantaggi che svantaggi. E' sempre un valore già presente nel dataset. Poco sensibile agli outliers. Se si divide il campione ordinato in 4 gruppi avremo i quartili come mediane di ogni sottogruppo di dati. E' il caso più rappresentativo ma non la miglior rappresentazione della distribuzione.
- Moda (valore più frequente) —>ha sia vantaggi che svantaggi. E' sempre un valore già presente. Non è sensibile agli outliers. Non descrive la distribuzione né l'ordine.

Variabile	Moda	Mediana	Media
Nominale	SI	NO	NO
Ordinale	Solo per scopi specifici	SI	Si se la distribuzione è normale
Continua	NO	Si se la distribuzione è normale	SI

DISPERSIONE DEI DATI:

La prima misura ad essere stata storicamente utilizzata per descrivere la dispersione o variabilità dei dati è il campo o intervallo di variazione, definito come la differenza tra il valore massimo e quello minimo.

Intervallo di variazione = Valore massimo - Valore minimo

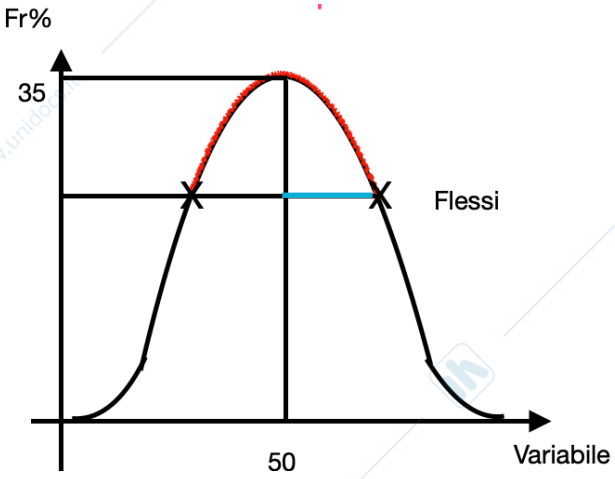
Inconvenienti:

- l'incapacità di sapere come i dati sono distribuiti entro l'intervallo;
- la sua dipendenza dal numero di osservazioni. All'aumentare del numero dei dati, cresce anche la probabilità di trovare un valore minore del minimo precedente ed uno maggiore di quello massimo precedente.

Varianza: somma scarti al quadrato diviso N.

Deviazione standard: radice varianza (σ=sigma) —>± numero deviazione standard.

Es: BI= 50 ± 5



Deviazione standard:

Distanza tra media e flesso della curva

Media:50, Moda:50 perché è la frequenza più alta, Mediana: 50 perché è il valore centrale(simmetrica).—> Non sempre coincidono

Distribuzione normale: Raccontata da media (μ o $\overline{x}$) e deviazione standard (σ)

Il 68% del range sono dati.

Le sette caratteristiche della distribuzione normale
1. Le distribuzioni normali sono simmetriche intorno alla loro media.
2. Media, mediana e moda coincidono nelle distribuzioni normali.
3. L'area sotto la curva è uguale alla dimensione campionaria (se $y=f_{abs}$), a 1 (se $y=f_{rel}$), o a 100% (se $y=f_{rel\%}$).
4. Le distribuzioni normali sono dense al centro e meno dense nelle due code.
5. Le distribuzioni normali sono definite da due soli parametri: la media (μ) e la deviazione standard (σ).
6. Il 68% dei dati (e quindi dell'area) sono inclusi nell'arco di una deviazione standard dalla media.
7. Approssimativamente il 95% dei dati è incluso nell'arco di circa due deviazioni standard dalla media (1.96σ).

Skewness—> misura quando una curva è asimmetrica.

Kurtosis—>misura gaussianità della curva. Quindi è una misura degli outliers presenti nella distribuzione e riguarda le code di distribuzione. Spesso è confusa con quanto la curva sia ripida o piatta.

- Platicurtica: (Kurtosis < 3): La distribuzione è bassa, le code sono più basse della distribuzione normale. Il picco è più basso e largo, ci sono poche code e mancano outliers.
- Mesocurtica: la distribuzione assomiglia a quella normale, con valori estremi distribuiti sulle code. La distribuzione normale ha una curtosi = 3.
- Leptocurtica (Kurtosis > 3): La distribuzione è più alta di quella normale, con code più grosse. Il picco è alto e sottile, i dati sono polarizzati dalle code e ci possono essere diversi outliers.

La mediana è accompagnata da quartini divisi in 4 equamente.

STATISTICA INFERENZIALE

- Probabilità
- Ipotesi Nulla
- t-test
- Analisi della Varianza
- Regressione e Correlazione
- Test non parametrici

Probabilità: studio delle possibilità è una branca della matematica che riguarda quanto spesso un fenomeno può accadere

P(x)= Nx / Ntot (Nx=numero possibilità che esca x, Ntot=numero totale degli eventi)

Tipi di probabilità:

- Probabilità classica: La probabilità di un evento è la proporzione tra le volte che l'evento può teoricamente accadere e tutte le possibili occorrenze di tutti i possibili eventi. È legata al modello matematico.
- Probabilità frequentista: La probabilità di un evento è la proporzione tra le volte che l'evento accade in una lunga serie (teoricamente infinita) di ripetizioni (indipendenti) dell'esperimento. È legata ai dati.
- Probabilità bayesiana: La probabilità di un evento è il grado di fiducia che un agente dotato di intelligenza razionale assegna alla possibilità che l'evento accada. È un misto tra modello e dati.

Ipotesi nulla: L'ipotesi nulla H_0 afferma che non ci sono reali differenze tra gruppi di dati osservati, le differenze apparenti sono solo dovute al caso

T-test: Se probabilità bassa (<5%) rigetto ipotesi nulla, più t è alta più le medie sono distanti. Abbiamo due gruppi, si calcolano entrambi le code, dati possono essere accoppiati o no.

$$t = \frac{(\bar{x} - \mu)}{\sigma_s / \sqrt{N}}$$

x=media campione, μ=media popolazione

σ=deviazione standard

Lezione 3 25/11/2024

Scale di valutazione:deve essere valida e ripetibile(intra-operatore o inter-operatore).

Analisi di varianza: quando ho più di 2 campioni. I dati sono curve gaussiana. Fare tanti t-test a gruppi aumenta errori di tipo ± quindi si fa l'Analisi di Varianza (ANOVA) e si calcola quanto sono grandi le medie rispetto alle varianze. Si raggruppano i dati per ciascun gruppo. Si calcola media per ogni gruppo e poi si fa la media totale (non si sa la varianza tra medie).

$$F_{(gl\ tra\ ,gl_{entro})} = \frac{Var_{tra}}{Var_{entro}}$$

Var tra: tra le medie

Var entro: nei gruppi

Test post-hoc —>'ipotesi nulla dell'ANOVA è relativa all'uguaglianza di tutte le medie M1 = M2 = M3 =….=Mn. Se $p<0.05$ una o più uguaglianze non è verificate, i post-hoc servono solo a capire quale di queste non è verificata.

Test a-priori —>decisi prima del test vanno fatti indipendentemente dal risultato dell'ANOVA.

Correlazione:

Regressione: $y=f(x)=a*x+b$ (all'aumentare di x aumenta y)
Regressione Multipla: $y=f(x_1,x_2,...)=a_1*x_1+a_2*x_2+...+b$
Regressione Logistica: y è una variabile binaria
Regressione Logistica Binaria: y e x_i sono binarie

Correlazione semplice—>Relazione: può essere curvilinea. Esistono 2 tipi:
positiva(quando all'aumentare di x aumenta y)
negativa(quando all'aumentare di x diminuisce y). Questi due tipi possono essere forti o deboli.

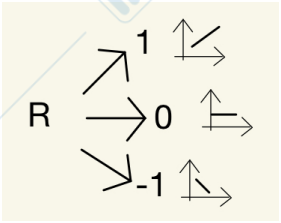
Regressioni lineari —> coefficiente di Pearson R(quando le due variabili possono essere correlati). Esso va da -1 a 1: Se vicino a 0 la correlazione è nulla, invece se ci troviamo agli estremi la correlazione è forte (vicino a -1 negativa quindi x aumenta e y diminuisce, vicino a 1 positiva quindi x aumenta e y aumenta).Coefficiente di correlazione di Spearman ρ(rho) per statistica non parametrica (è il coefficiente di Pearson applicato ai ranghi invece che ai valori)

Regressioni non lineari —>R^2 è il coefficiente di Determinazione e ci fa capire quanto sono fatti bene i nostri dati. In questo caso R non può essere calcolato.

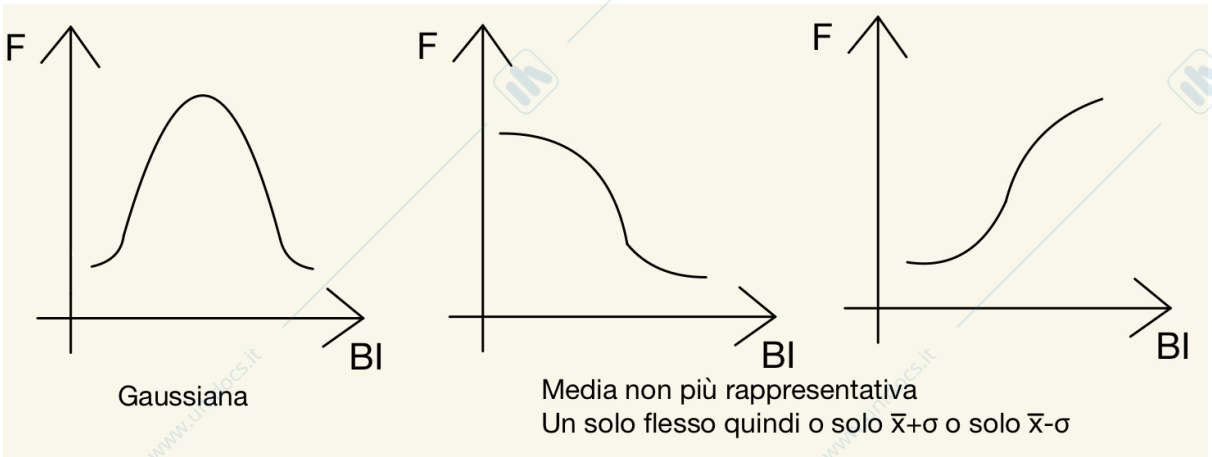
Storks deliver babies (cicogne consegnano bambini) (ρ=0,008)—>Si è visto che nelle zone in cui ci sono più nascite abbiamo più nidi di cicogne—> Questo ci fa capire che non sempre una correlazione significativa ammette una causa-effetto.

VARIABILI CONTINUE. STATISTICA PARAMETRICA.

- Per una curva abbiamo grafici di solito $x \pm \sigma$ (media più o meno deviazione standard)
- Per due curve: t-test $t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{N_1} - \frac{\sigma_2^2}{N_2}}}$ t sale e ρ scende
- Per più curve : $F_{(gl\ tra\ ,gl_{entro})} = \frac{Var_{tra}}{Var_{entro}}$ F sale e ρ scende
- Relazione x,y:



VARIABILI ORDINALI. STATISTICA NON PARAMETRICA.



Gaussiana

Media non più rappresentativa
Un solo flesso quindi o solo $\overline{x}+\sigma$ o solo $\overline{x}-\sigma$

VARIABILI NOMINALI

	Cancro		No Cancro	
Amianto	3	a	7	b
Biscotti	1	c	19	d

Ia=3/10 (Incidenza della fabbrica di amianto)

Ib=1/20 (Incidenza della fabbrica di biscotti)

RR=Ia/Ib = $\frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{\frac{3}{10}}{\frac{1}{20}} = 6$ 6 volte rischio più alto, per chi lavora l'amianto, di avere il cancro rispetto a Chi fabbrica biscotti

Numero positivo. Se R=1 abbiamo fattori uguali.

$\chi^2 = \rho(a,b,c,d)$

OR (Odds ratio)= (a/b)/(c/d) —>Semplifica RR,ha un fattore neutro, è l'esponenziale di β

Test:	**+**	**-**
Sano	FP	VN
Malato	VP	FN

V=vero F=falso P=positivo N=negativo.

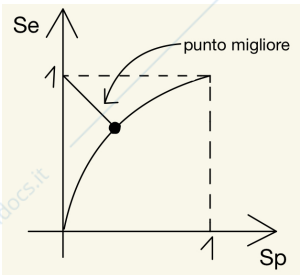
Se(sensibilità) =VP/(VP+FN)

Sp(Specificità)= VN/(VN+FP)

Acc(Accuratezza)= (VP+VN)/(VP+VN+FP+FN)

- Se=100% allora Sp scende
- Sp=100% allora Se scende

Curva Roc:



Prevalenza: La proporzione delle persone che hanno una certa condizione e la popolazione a rischio
P = numero di casi / soggetti a rischio

Incidenza: La proporzione delle persone che in un dato periodo sviluppano una certa condizione e la popolazione a rischio
I = nuovi casi in un periodo / soggetti a rischio

John Snow—> padre dell'epidemiologia. Scopre che il colera si diramava tramite l'acqua

Meta-Analisi—> unire più studi fatti da diverse persone in diversi modi. Uniamo rendendo tutto in valore nominale—>OR. (Nei grafici il numero di soggetti presi in studio viene rappresentato da un quadratino: più è grande il quadratino e più saranno i soggetti.)

STATISTICA NON PARAMETRICA.

Variabili ordinali (4 non è il doppio di 2 ma è maggiore). Le scale devono essere valide, ripetibili, consistenti e complete.

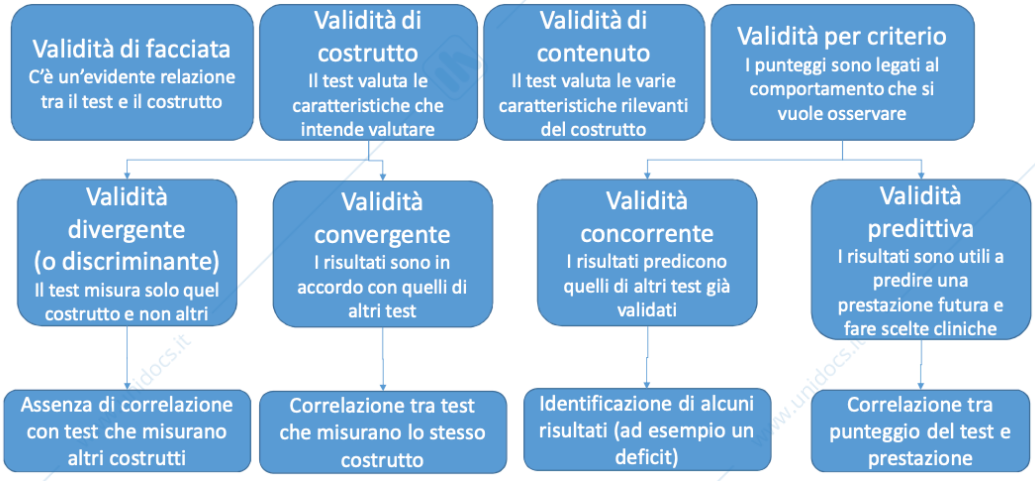
Differenza tra misura e stima: la misura deve essere precisa, la stima è una deduzione dalle misure.

Costrutto:è un abilità, un'attitudine o una capacità che è basata su uno o più teorie già stabilite, non è direttamente osservabile, ma può essere indirettamente stimato. Tramite operalizzazione (trovare il collegamento causale con qualcosa che può essere osservato) troviamo la misura.

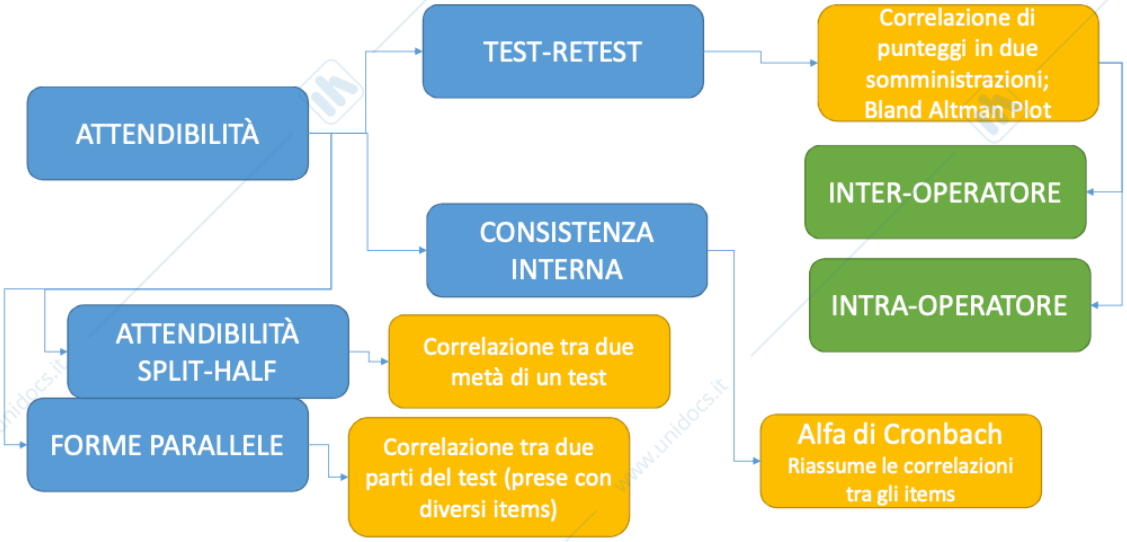
Esistono 2 tipi di indicatori:

- Formativi: indicatori che causano il costrutto.
- Riflessivi: indicatori causati dal costrutto.

Validità: è quanto bene un test misura quella che vuole misurare (il valor medio si avvicina al valore vero). Esistono tanti tipi di validità.



Attendibilità: è quanto un test è ripetibile tra diverse ripetizioni nel tempo (intra-operatore) o tra persone (inter-operatore), è una misura della precisione del test.



Item: analisi che serve per vedere se vanno nella stessa direzione vedo le varie correlazioni. Spearman-Brown:

$$\alpha = \frac{N \cdot \bar{R}}{1 + \bar{R} \cdot (N-1)}$$