

# **STATISTICA E PROBABILITÀ**

## **1. Misurazione e Misura**

**Misurazione:** Processo per determinare il valore di una grandezza, seguendo un metodo preciso.

**Misura:** Il valore ottenuto dalla misurazione (deve sempre avere un'unità di misura).

**Valore vero:** Il valore perfetto, ottenibile solo con una misurazione senza errori.

### **Metodi di Misurazione**

**Diretto:** Si misura una grandezza con uno strumento specifico (es. usare un termometro per la temperatura).

**Indiretto:** Si calcola il valore di una grandezza basandosi su altre misurazioni e formule matematiche (es. velocità = spazio/tempo).

### **Scale di Misurazione**

1. **Nominale:** Classifica senza ordine (es. gruppi sanguigni).
2. **Ordinale:** Classifica con ordine, ma senza distanze precise tra i valori (es. classifica di una gara).
3. **A intervalli:** Ha distanze precise tra valori, ma senza un punto zero assoluto (es. temperatura in °C).
4. **A rapporti:** Ha uno zero assoluto e i rapporti tra i valori hanno senso (es. altezza, peso).

### **Caratteristiche di una Buona Misura**

**Affidabilità:** I risultati sono stabili e ripetibili.

**Validità:** Misura effettivamente quello che deve misurare.

**Controllabilità:** Deve poter essere verificata.

## **2. Errori di Misura**

L'errore è la differenza tra la misura ottenuta e il valore vero.

Poiché il valore vero è sconosciuto, si stima **l'incertezza**.

### **Tipi di Errori**

1. *Strumentali:* Lo strumento non è perfetto o è calibrato male.
2. *Metodologici:* Il metodo usato per misurare è impreciso.
3. *Esecutivi:* L'operatore sbaglia nell'uso dello strumento.
4. *Casuali:* Variazioni imprevedibili nelle condizioni di misura.

### **Caratteristiche degli Strumenti di Misura**

**Intervallo di funzionamento:** Gamma di valori che può misurare.

**Prontezza:** Velocità con cui dà il risultato.

**Sensibilità:** Quanto è preciso nel rilevare piccole variazioni.

**Precisione:** Quanto è costante nel dare sempre lo stesso risultato.

**Accuratezza:** Quanto il valore misurato è vicino al valore vero.

## **3. Probabilità**

La probabilità misura quanto è probabile che un evento accada (probabilità relativa; da un punto di vista statistico i valori sono compresi tra 0 e 1)

Definizione Classica:

$$\Pr(A) = \frac{\text{casi favorevoli}}{\text{Casi possibili}}$$

-

**Probabilità a posteriori:** Si calcola dopo che l'evento è già accaduto.

Esempio: Se in un'epidemia 48 persone si ammalano su 192 esposte, la probabilità di ammalarsi è (25%)

**4. Probabilità di Eventi Complessi** gli eventi A e B possono essere presenti assieme (AeB) o in alternativa (AoB), si gestiscono con 2 regole:

a) Regola della Moltiplicazione (per eventi indipendenti):

$$\Pr(A \text{ e } B) = \Pr(A) \cdot \Pr(B)$$

b) Regola dell'Addizione (per eventi mutuamente esclusivi):

$$\Pr(A \text{ o } B) = \Pr(A) + \Pr(B)$$

Se gli eventi **NON sono esclusivi**, bisogna sottrarre la probabilità che si verifichino entrambi:

$$\Pr(A \text{ o } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ e } B)$$

**5. Legge dei Grandi Numeri**

Se ripetiamo molte volte un esperimento casuale, la frequenza relativa di un evento si avvicina alla sua probabilità teorica.

Esempio: Lanciando una moneta molte volte, la frequenza di testa tenderà a  $\frac{1}{2}$

In questa legge la differenza tra il valore osservato ed il valore teorico atteso tende a diminuire aumentano il numero di prove; più alto è il numero di prove maggiore è la precisione **CUT OFF**.

Associazione: grado di dipendenza statistica tra 2 o più eventi –

Associazione non è uguale a casualità perché non sempre sono legati a causa effetto.

Indicatori di sensibilità/specificità:

- Indicatori (markers): sostanza (parametro) la cui presenza indica lo stato di un organismo (malattia, esposizione, ecc)
- Sensibilità: probabilità che il markers sia positivo quando c'è la malattia cioè proporzione data da veri positivi: veri positivi (veri positivi+ falsi negativi)
- Specificità: probabilità che il markers sia negativo in assenza della patologia o sostanza cioè proporzione veri negativi +falsi positivi.

Frequenze:

frequenza assoluta: numero di volte che un avvenimento si verifica

frequenza relativa:  $\frac{\text{numero di prove in cui un evento si verifica}}{\text{numero di prove totali}}$

frequenza percentuale: frequenza relativa per 100.

**Errore standard:** stabilisce un intervallo entro il quale una percentuale osservata può cadere con una certa probabilità.

**6. Test Statistici**

È una di calcolo che permette di rifiutare un'ipotesi

**Chi-Quadrato:** Verifica se due percentuali sono diverse per caso o per una differenza reale.

Se la statistica test supera un valore soglia, si rifiuta l'ipotesi nulla (che dice che non c'è differenza). È utilizzabile quando il numero di osservazioni è maggiore di 30 ed il contenuto di ogni cella è maggiore di 5 (altrimenti si usa il test di Fischer).

Esempio: Se un dado è lanciato 2000 volte e il numero 1 esce molto più spesso degli altri, si può verificare se il dado è truccato usando il test chi-quadrato.

**Performance diagnostica:** corrisponde al grado di accuratezza.

Gli indici di accuratezza sono 5 :

- 1) Sensibilità: Percentuale di malati che risultano positivi al test.

- 2) Specificità: Percentuale di sani che risultano negativi al test.
- 3) Valore Predittivo Positivo (VPP): Probabilità che un test positivo indichi davvero la malattia.
- 4) Valore Predittivo Negativo (VPN): Probabilità che un test negativo significhi davvero assenza di malattia.
- 5) Accuratezza diagnostica

**Curva ROC:** Grafico che mostra il compromesso tra sensibilità e specificità. Più la curva è vicina all'angolo in alto a sinistra, più il test è efficace.

Esempio: Nel test del PSA per il cancro alla prostata, abbassare la soglia di positività aumenta la sensibilità (meno falsi negativi) ma riduce la specificità (più falsi positivi).

Conclusione

Questi concetti sono fondamentali per comprendere le misurazioni, la probabilità e i test statistici. Capire sensibilità, specificità e valore predittivo aiuta a valutare l'affidabilità dei test diagnostici e delle previsioni probabilistiche.

## **Modelli di Classificazione**

### **Cos'è la classificazione?**

Abbiamo un insieme di dati chiamato training set.

Ogni dato ha vari attributi, uno dei quali indica la classe di appartenenza.

Obiettivo: trovare una funzione che assegni la classe giusta a nuovi dati non classificati.

Per verificare il modello, usiamo un test set.

Classificazione vs Clustering

Classificazione (Supervised Learning):

I dati nel training set hanno già etichette di classe assegnate.

Il modello impara da questi dati e assegna etichette a nuovi dati.

Clustering (Unsupervised Learning):

Non sappiamo a priori a quale classe appartengano i dati.

Il modello cerca gruppi omogenei tra i dati, senza informazioni preesistenti.

### **Tecniche di Classificazione**

Rule-based Methods: usa regole per classificare.

Memory-based reasoning: classifica in base a dati simili nel passato.

Neural Networks: modelli ispirati al cervello umano.

Genetic Algorithms: simula l'evoluzione naturale per trovare soluzioni.

Naïve Bayes: usa probabilità per classificare.

Support Vector Machines (SVM): separa le classi in modo ottimale.

Decision Tree

**Decision Tree:** struttura ad albero, semplice ed efficace, poco costosa, semplice, buona accuratezza)

Uno degli approcci migliori: facile da costruire, interpretare e integrare con database.

Funziona come un diagramma a blocchi:

Ogni nodo interno testa un attributo.

Gli archi rappresentano i risultati del test.

I nodi foglia assegnano una classe.

Per classificare un nuovo dato:

1. Si parte dalla radice.
2. Si segue il percorso basato sugli attributi del dato.
3. Si arriva a una foglia che assegna la classe.

Classification vs Regression Tree

*Classification Tree*: il risultato è una classe (es. "positivo" o "negativo").

*Regression Tree*: il risultato è un valore continuo (es. "temperatura media").

## **Valutazione dell'Accordo tra Osservatori**

**Cos'è la concordanza?** Misura quanto due test o due persone sono d'accordo nel valutare qualcosa.

Importante quando si analizzano dati come radiografie, ECG o diagnosi mediche.

**Obiettivo:** verificare se le valutazioni sono riproducibili e affidabili.

**Test di McNemar** (per dati dicotomici) Serve a verificare se due valutazioni sono statisticamente diverse; Si calcolano le frequenze dei risultati discordanti e si usa un test statistico.

Un p-value alto indica che non ci sono differenze significative.

## **Accordo tra valutatori (Inter-rater Reliability)**

Ipotesi di base:

1. Le categorie sono mutuamente esclusive.
2. I valutatori operano in modo indipendente.
3. I valutatori hanno lo stesso livello di competenza.
4. Non ci sono restrizioni sulle decisioni.

## **Kappa di Cohen**

Misura quanto due valutatori siano d'accordo, correggendo per il caso.

Valori:

$<0$  → nessun accordo

0-0.20 → scarso

0.21-0.40 → lieve

0.41-0.60 → moderato

0.61-0.80 → sostanziale

0.81-0.99 → quasi perfetto

Esempio: due medici classificano 200 pazienti. Se l'accordo è inferiore a quello atteso per caso, significa che i criteri non sono affidabili.

## **Metanalisi**

***Cos'è una metanalisi?*** Analizza i risultati di più studi su un argomento.

Serve per:

1. Confrontare studi con risultati discordanti.
2. Aumentare la potenza statistica combinando dati.
3. Ridurre errori e incertezze.

**Vantaggi:**

Riassume più studi.

Aumenta la dimensione del campione.

Permette di analizzare sottogruppi.

### **Svantaggi:**

Bias(modificazioni)di pubblicazione (studi negativi meno pubblicati).

Bias di selezione (criteri di inclusione/esclusione).

Eterogeneità (studi con metodologie diverse)

**Come si fa una metanalisi?** si seguono i seguenti passaggi:

1. **Definizione del protocollo**, si stabiliscono:

Criteri di inclusione/esclusione degli studi.

Endpoint (tipologia di misurazione) chiari (es. guarigione, sopravvivenza).

2. **Selezione degli studi:**

Si includono solo studi che rispettano i criteri.

3. **Analisi statistica:**

Si calcola un valore unico (es. odds ratio).

Si rappresenta graficamente con un Forest Plot ( o grafico a foresta)

Esempio:

Si confronta l'uso di antibiotici vs placebo nelle infezioni neonatali.

Linea verticale = parità di efficacia (odds ratio = 1).

Intervalli di confidenza = margine di errore di ogni studio.

Diamante = risultato combinato della metanalisi.

Se il diamante è lontano da 1, il trattamento è efficace.

### **Studi Clinici**

Fasi dello sviluppo di un farmaco

Tempo medio di sviluppo: 12 anni.

Costo medio: 800 milioni di dollari.

Solo 1 su 5000 molecole arriva sul mercato.

### **Fasi della sperimentazione clinica**

1. **Fase I (sicurezza e metabolismo)**

Test su volontari sani o pazienti senza alternative terapeutiche.

Obiettivo: metabolismo, tossicità, dose massima tollerata.

Tipico disegno: dose escalation (si aumenta la dose finché non compaiono effetti collaterali).

2. **Fase II (attività e tollerabilità)**

Verifica se il farmaco ha un effetto biologico.

Confronto tra diverse dosi.

3. **Fase III (efficacia clinica)**

Confronta il farmaco con lo standard di cura.

Studi randomizzati e multicentrici.

Obiettivo: dimostrare che il farmaco funziona e ha effetti accettabili.

4. **Fase IV (monitoraggio post-mercato)**

Studia la sicurezza a lungo termine.

Valuta l'impatto nella pratica clinica reale.

Analizza anche costi ed efficacia economica.

## **Biosimilari e Comparabilità**

Cos'è un biosimilare? Copia di un farmaco biologico originatore, il cui brevetto è scaduto, Sviluppato per avere la stessa efficacia e sicurezza.

Come si verifica la comparabilità?

### **1. Studi di qualità comparativi:**

Analizzano la struttura e funzione del farmaco in laboratorio.

### **2. Studi non clinici comparativi:**

Testano l'effetto del farmaco su cellule e modelli animali.

### **3. Studi clinici comparativi:**

Non servono per dimostrare l'efficacia (già nota).

Servono a confermare che non ci siano differenze significative con il farmaco originale.

## **Analisi di Sopravvivenza**

**La statistica** è fondamentale nella ricerca biomedica perché permette di trarre conclusioni basate su un campione anziché su un'intera popolazione. Una volta selezionato un campione, è possibile testare ipotesi utilizzando la statistica descrittiva (che descrive i dati, come età, sesso, fumo) e la statistica inferenziale (che permette di fare previsioni o confronti).

**L'analisi di sopravvivenza** rientra nella statistica inferenziale e si distingue per mettere in relazione un evento con il fattore tempo. Ad esempio, se vogliamo studiare la sopravvivenza a due anni di un gruppo di persone, dobbiamo seguirle per 24 mesi e registrare i dati sotto forma di coppie, dove:

c indica la condizione del paziente (0 = vivo, 1 = morto).

t rappresenta il tempo di osservazione.

Le coorti di pazienti sono dinamiche, poiché il numero di partecipanti può variare nel tempo in base alla comparsa dell'evento d'interesse.

Concetti di base

### **1. Tempo di sopravvivenza**

Il "tempo di sopravvivenza" non si riferisce solo alla morte, ma a qualsiasi evento misurabile nel tempo, come:

Il tempo fino all'insorgenza di una malattia.

Il tempo prima di un evento cardiovascolare (infarto, ictus).

Il tempo fino al rigetto di un trapianto renale.

Il tempo fino al raddoppio di un valore biochimico anomalo (ad es. creatinina).

Se un paziente non ha sviluppato l'evento d'interesse, il suo tempo di sopravvivenza corrisponde alla durata dell'osservazione fino alla fine dello studio.

### **2. Probabilità condizionata**

La sopravvivenza è una probabilità condizionata: la probabilità di essere vivi a un certo tempo dipende dal fatto di essere sopravvissuti fino a quel momento.

Ad esempio, se:

La probabilità di sopravvivere il primo giorno è 0.80 (P1)

Quella di sopravvivere il secondo giorno è 0.68 (P2)

,Quella del terzo giorno è 0.55(P3),

allora la sopravvivenza cumulativa sarà (30%).

Un esempio analogo è il lancio di una moneta bilanciata. La probabilità di ottenere testa in tre lanci successivi è (12.5%).

Regola moltiplicatoria della probabilità data dal prodotto ossia  $P = P1.P2.P3$  (dal risultato si ricava la percentuale)

### 3. Rischio osservato

Il rischio osservato è misurato direttamente dall'esperienza. Supponiamo di monitorare 11 pazienti ipertesi:

4 hanno ipertrofia ventricolare sinistra (IVS), 7 non hanno IVS.

Dopo un anno, osserviamo che:

3 pazienti su 4 con IVS hanno avuto un infarto.

1 paziente su 7 senza IVS ha avuto un infarto.

Il rischio relativo è , il che significa che i pazienti con IVS hanno un rischio di eventi cardiovascolari 5.4 volte maggiore rispetto a quelli senza IVS.

### Analisi di Kaplan-Meier

Il metodo di Kaplan-Meier è lo strumento principale per costruire curve di sopravvivenza, che mostrano la probabilità di sopravvivenza nel tempo. Questo metodo permette di:

1. Stimare la funzione di sopravvivenza (es. probabilità di sopravvivenza a 3 o 5 anni).
2. Confrontare la sopravvivenza tra gruppi di pazienti sottoposti a trattamenti diversi.
3. Valutare il potere prognostico di diverse variabili, infatti per fare un'analisi di sopravvivenza sono fondamentali 2 **variabili**:

- Variabile censor (troncamento) indica se il paziente ha avuto o meno l'evento terminale.
- Variabile survival (sopravvivenza) indica il periodo di osservazione

*La sopravvivenza nel periodo* (SP) è la percentuale di pazienti vivi in quel periodo e si calcola con la formula:

$$SP = \frac{\text{numero di eventi in un intervallo } t}{\text{Numero di persone a rischio nell'intervallo } t}$$

### **Costruzione della curva di sopravvivenza**

Supponiamo di seguire 10 pazienti con IVS per 240 giorni per monitorare l'incidenza di infarto fatale. Se un paziente non ha un infarto durante lo studio o viene perso al follow-up, si dice che è censurato., essi rimangono nell'analisi fino al momento in cui ci sono dati certi sulla loro salute perchè questi dati sono importanti nell'analisi.

#### **Per costruire la curva:**

Si divide il periodo di studio in intervalli determinati dagli eventi osservati.

Si calcola, per ciascun intervallo, la probabilità di sopravvivenza e la sopravvivenza cumulativa.

Si rappresentano questi valori in un grafico con:

L'asse X che indica il tempo.

L'asse Y che indica la probabilità cumulativa di sopravvivenza.

Confronto tra due curve di sopravvivenza

Quando vogliamo confrontare la sopravvivenza di due gruppi (ad esempio, fumatori vs non fumatori), utilizziamo il **log-rank test**, che verifica se esiste una differenza statisticamente significativa tra le curve.

Se però le curve si incrociano, il log-rank test non è più adeguato e si usa il test di

**Wilcoxon-Breslow-Gehan**, che dà più peso agli eventi nelle fasi iniziali della curva.

**Modello di Cox** Il **modello di regressione di Cox** è un metodo avanzato che analizza il rischio di un evento tenendo conto di più fattori contemporaneamente. Ad esempio, può stimare il rischio di metastasi considerando:

Data della diagnosi.

Data della comparsa della metastasi.

Altri fattori di rischio (età, sesso, trattamenti ricevuti).

### **Hazard Ratio (HR)**

L'Hazard Ratio confronta il rischio tra due gruppi:

$HR > 1$  → aumento del rischio nel gruppo sperimentale.

$HR < 1$  → diminuzione del rischio nel gruppo sperimentale.

$HR = 1$  → nessuna differenza tra i gruppi.

Esempio: se  $HR = 2$ , significa che i pazienti trattati hanno un rischio doppio rispetto a quelli non trattati.

### **Riepilogo**

1. L'analisi di sopravvivenza studia il tempo fino al verificarsi di un evento.
2. Kaplan-Meier è il metodo più usato per stimare la sopravvivenza e confrontare gruppi.
3. Il log-rank test verifica se le differenze tra due curve di sopravvivenza sono significative.
4. La regressione di Cox permette di valutare più fattori contemporaneamente.
5. L'Hazard Ratio misura il rischio relativo tra gruppi.

Questo metodo statistico è ampiamente utilizzato in studi clinici per analizzare l'efficacia di trattamenti e identificare fattori di rischio.