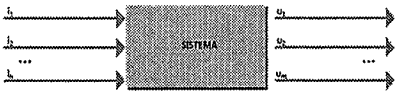
BIOSTATISTICA **02/03/20**

La Statistica è una scienza matematica che si occupa dell'elaborazione dei risultati dell'osservazione di caratteri posseduti dagli elementi di un insieme (fenomeno collettivo).
La Biostatistica è la statistica applicata alla biomedicina.
Iniziamo distinguendo due tipi di fenomeni:

- Fenomeni Aleatori o Casuali sono quei fenomeni che non posso descrivere con una equazione matematica in quanto non sono previsibili con precisione, uso equazioni per quantificare l'incertezza associata al fenomeno.
- Fenomeni Deterministici sono quei fenomeni di cui posso predirne con precisione l'avvenimento, posso inoltre definire una o più equazioni che caratterizzano il fenomeno.

Concezione Frequentista: guarda la frequenza con cui la cosa che stiamo osservando avviene sulla base di un esperimento ripetuto molte volte nelle stesse condizioni. Tale fenomeno può essere aleatorio come no, esistono esempi quotidiani di fenomeni che noi consideriamo aleatori anche quando non lo sono.
Il Sistema Dinamico è un sistema caratterizzato da equazioni che mettono in relazione l'ingresso con l'uscita, in ogni istante di tempo tale equazione mappa ingresso e uscita.
Le grandezze del sistema sono:
-variabili in ingresso I
- variabili uscita U
-variabili di stato S, che definiscono la configurazione interna



Esistono Sistemi Caotici in cui basta apportare una variazione infinitesima dell'ingresso per fare in modo che l'uscita diverga con andamento esponenziale. Tali sistemi prima di essere scoperti erano associali a quelli aleatori ma oggi sono una vera e propria classe di sistemi chiamati ***sistemi caotici deterministici***.

$$\mathcal{E}(t) = \mathcal{E}_0 e^{\lambda t}$$

Noi trattiamo i sistemi aleatori!
Sono definiti Sistemi Aleatori (Stocastici o Probabilistici) quei sistemi in cui non è possibile definire un'equazione che relaziona ingresso e uscita. Tuttavia tenteremo di definire il fenomeno partendo da dati concreti di osservazione e proveremo a dire qualcosa sul sistema ovvero proveremo a "caratterizzarlo"→ riusciremo solo a quantificare l'incertezza di tali sistemi.
Possiamo avere:

- dati di partenza o valori campionari che possono essere in una o più dimensioni
- serie temporali per definire un certo stato
- eventi che fanno parte delle serie temporali ma avvengono in certi tempi specifici che noi cerchiamo di caratterizzare (es. battito ciglia)
- immagini in due o tre dimensioni che possono essere nel tempo (es. tac, raggi x)

Esempi:
-Registrazioni Neuronali hanno un andamento che fluttua attorno a un certo valore, se succede qualcosa il segnale sale rispetto a questo valore e poi vi fa ritorno. C'è differenza tra considerarlo una serie temporale o un evento di cui vogliamo calcolare il tempo.

-EEG(elettroencefalogramma) sono elettrodi che sentono la variazione di campo magnetico in uno strato di corteccia cerebrale e colgono i segnali durante una specifica azione
-ECG(elettrocardiogramma), quando si è fermi a riposo il cuore batte più di quando si è in corsa o in movimento → impredicibile quando avverrà il prossimo battito però possiamo dire con una certa probabilità che avverrà entro un certo range di tempo

-Risonanza Magnetica verifica quali sono le attivazioni di certe aree cerebrali in determinati casi (es. immagino di muovere una mano)

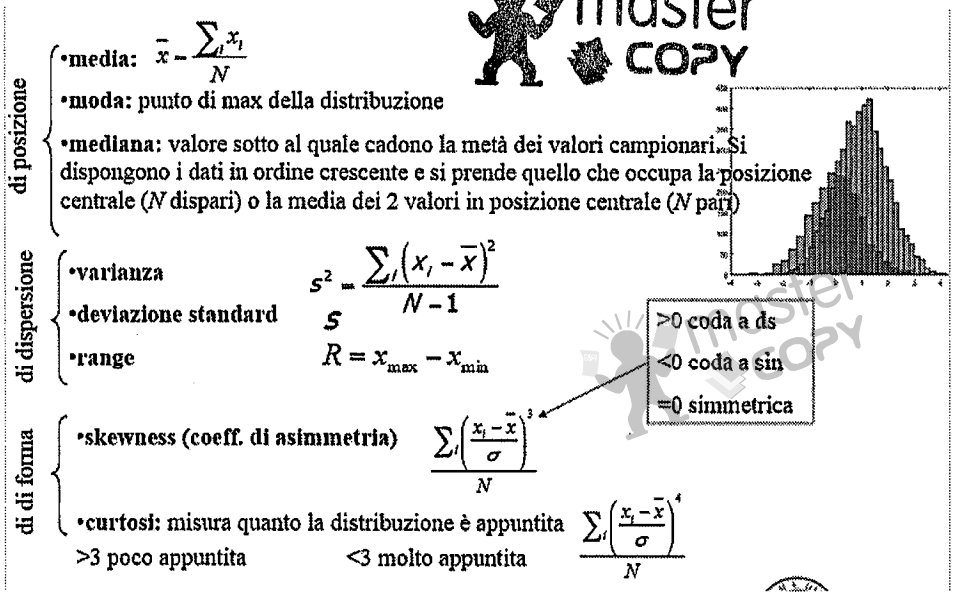
La statistica nasce da una quantificazione dell'informazione che si chiama Shannon Entropy

$$H = -\sum_i p_i \log_2(p_i)$$

Tale scienza tenta di quantificare la presenza di variabili aleatorie nei dati, un suo primo utilizzo appare nella Statistica Descrittiva la quale ha come scopo quello di descrivere i dati raccolti. Si parte da un insieme di dati o valori campionari che vanno descritti, es. sono stati presi da 410 cellule, tali dati vengono poi descritti perché non possono essere trattati tutti insieme. Le cose si complicano quando abbiamo più variabili e vogliamo caratterizzare uno stato per esempio: un medico vuole dire se persona è sana o no avendo tremila analisi, dunque prenderà in considerazione solo quelle con cui è più a suo agio e darà un risultato. La statistica descrittiva sintetizza i dati campionari che osserviamo e per descriverli possiamo farne un grafico che ci aiuta a livello intuitivo a trovare informazioni relative al campione.
Potremmo prendere un indice e valutarne i valori ma non è molto utile poiché l'ordine con cui piazziamo i dati è arbitrario.
Definiamo delle tabelle di frequenza: prendiamo un determinato valore o un range di valori e andiamo a contare quante volte osserviamo questi valori (quante volte compare il valore 55). Avremo così un grafico molto più leggibile.
Esiste una differenza fra frequenza assoluta (es. compare 4 volte il valore 55) e frequenza relativa (normalizziamo i valori trovati poiché io posso averne visti un tot su un campione di 400 individui e un'altra persona ne ha visti un altro tot su un campione da 500 individui)
Normalizzare significa prendere i campioni trovati e dividerli per il numero totale di campioni analizzati.

Usiamo grafici chiamati istogrammi o digrammi a bastoncino, tuttavia le osservazioni viste nel grafico sono puramente soggettive quindi ci servono dei quantificatori che sono forniti dalla statistica descrittiva.

Esempi:



Descrivere un campione è una cosa che va fatta con criterio, dobbiamo scegliere opportunamente quale quantificatore usare! (in genere ne servono due o tre).

Si dice campione un aggregato di unità appartenenti a una certa popolazione e selezionate mediante un esperimento di campionamento, i valori reali dell'esperimento sono i valori campionari del campione statistico. Definiamo **Popolazione** l'insieme che comprende ogni possibile e immaginabile relazione, è una **variabile aleatoria** e definiamo **unità statistiche** gli elementi della popolazione.

Esempio: voglio misurare il livello della glicemia di tutti i soggetti sani allora la mia popolazione è l'insieme di tutte le persone sane.
Se prendo la glicemia su 200 casi allora sto guardando un **campione** della popolazione (un suo sottoinsieme) ovvero una **realizzazione della variabile aleatoria**.
Descriviamo con la statistica descrittiva ma in realtà a nessuno importa dire qualcosa solo su quei soggetti ma interessa sapere qualcosa sulla popolazione di riferimento, parto dal campione e voglio dire qualcosa sulla popolazione entra nella **STATISTICA INFERENZIALE**:

descrittiva→ descrizione campione

inferenziale→ usa la statistica descrittiva per dire qualcosa sulla popolazione del campione (si fa inferenza) ovvero tramite le informazioni ottenute sul campione possiamo fare affermazioni sulle caratteristiche di tutta la popolazione.

Parto dalla descrizione dei campioni, applico un metodo inferenziale e poi cerco di dire qualcosa sulla popolazione → non posso dire cose certe ma posso quantificarle con un quantificatore di inferenza che è la PROBABILITÀ. Non faremo quantificazioni puntuali ma intervallari, a ogni intervallo di popolazione attribuiremo una probabilità. Quantifichiamo l'incertezza con la funzione di probabilità, importante è l'intuizione ma bisogna stare attenti perché può ingannare.

Probabilità di osservare 1 lanciando una volta un dato : 1/6

Probabilità di osservare 1 se lo lancio due volte: (1-5/6)^2=11/36

Esiste una **applicazione** della definizione di probabilità: è applicabile solamente se gli eventi sono equiprobabili e indica che la probabilità è data dal numero di casi favorevoli fratto il numero di casi totali.
Esempi banali:

$$P[i] = \lim_{N \to \infty} \frac{n_i}{N}$$

- Probabilità di avere il diabete in popolazione di 100'000 individui su 10'000 individui
  → p(diabete) = 10'000/100'000= 10%
- Probabilità di avere il diabete e l'ipertensione
  → p(diabete&ipertensione)=5'000/100'000=5%
  prendo quelli che hanno sia diabete sia ipertensione
- Probabilità di avere diabete o ipertensione
  → p(diabete e ipertensione)= (10'000+20'000-5'000)/100'000=25%
  devo togliere quelli che li hanno entrambi senno vengono contati due volte, se ipertensione e diabete fosse indipendenti (non posso averli entrambi insieme) allora la probabilità dell'unione sarebbe la somma delle singole probabilità
- 5 carte da un mazzo da 52 (13 carte per ognuno dei 4 semi) devo ottenere una coppia e un tris. Assumendo che il mazzo non sia truccato e ogni carta abbia la stessa probabilità di essere osservata allora posso usare la formula sopra e devo contare quante sono le combinazioni di casi favorevoli.

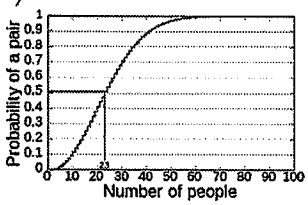
Operatore binomiale:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \quad \text{for } 0 \le k \le n,$$

- Problema del compleanno: calcola la probabilità che, date n persone, due di queste facciano il compleanno lo stesso giorno.

NB: Non esiste una sola soluzione perché dipende dalle assunzioni alla base
Assunzioni: eventi equiprobabili ovvero su 365gg l'anno la probabilità di nascere sia la stessa allora otterrei che:
-se n>365 la probabilità sarebbe 1
-se n<365 devo quantificare l'incertezza, l'intuizione ci farebbe fare una proporzione ma è sbagliato poiché tale probabilità su gruppo di 160 persone è 99,9%, è una proporzione fattoriale

$$\bar{p}(n) = 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{n-1}{365}\right)$$
$$= \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}$$
$$= \frac{365!}{365^n (365 - n)!} = \frac{n! \cdot \binom{365}{n}}{365^n} = \frac{{}_{365}P_n}{365^n}$$
$$p(n) = 1 - \bar{p}(n).$$



n	$\bar{p}(n)$
5	2.7%
10	11.7%
20	41.1%
23	50.7%
30	70.6%
40	89.1%
50	97.0%
60	99.4%
70	99.9%
100	99.99997%
200	99.99999999999999999999999998%
300	$(100 - (8\times10^{-60}))\%$
350	$(100 - (3\times10^{-125}))\%$
365	$(100 - (1.45\times10^{-155}))\%$
366	100%

- Gioco tre porte (Monty Knows): dietro una porta ho un premio e dietro le altre ho altre cose

Assunzione: il presentatore sa dove è la macchina dunque il concorrente sceglie una porta e il presentatore apre una porta che sicuramente non ha il premio dieto, successivamente il presentatore chiede se vuole cambiare la scelta ed è dimostrato col problema delle probabilità totali che la vincita col cambio è a 2/3 di probabilità e la vincita restando con la stessa scelta è 1/3 di probabilità.

- Problema della rovina dei giocatori d'azzardo: andiamo al casinò con I monete e N è la vincita, P è la nostra probabilità di vincere e Q è quella di perdere quindi P+Q=1
  Gioco assume iterazioni successive, a ogni vincita faccio i+1 e vinco se i tende a n e perdo se i tende a zero. Se P=Q=0.5 allora la probabilità di vincere è sempre i/n ma in realtà il casinò ha sempre più probabilità di vincere.

$$P_i = \begin{cases} \frac{1-(\frac{q}{p})^i}{1-(\frac{q}{p})^N}, & \text{if } p \neq q; \\ \frac{i}{N}, & \text{if } p = q = 0.5. \end{cases}$$

DEFINIZIONE DI PROBABILITÀ: funzione matematica in cui abbiamo un insieme Ω chiamato spazio degli eventi, su questo insieme costruiamo l'insieme degli eventi dato da σ-algebra (comprende tutte le parti di omega compreso insieme nullo e omega stesso).

Il suffisso σ serve per dire che rimango nell'insieme quando considero quello nullo o quello dello spazio campionario

Assioma 1= la probabilità dello spazio campione è 1

Assioma 2= se voglio calcolare la probabilità dell'unione di due eventi disgiunti allora tale probabilità è la somma delle probabilità dei singoli eventi

Probabilità: terna $(\Omega, \mathcal{A}, P)$ dove
Ω è un insieme
$\mathcal{A}$ è una σ-algebra di parti di Ω
P è una probabilità su $\mathcal{A}$

$\mathcal{A}$ è una famiglia di **parti** dell'insieme Ω , cioè è l'insieme di tutti i sottoinsiemi di Ω

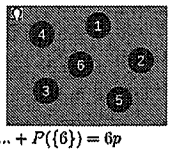
es. se $\Omega = \{a, b, c\}$ allora $\mathcal{A} = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \Omega\}$

σ-algebra: serie di operazioni (di misure) sui sottoinsiemi, e in particolare se:
$A_1, \ldots, A_n \in \mathcal{A}$
allora: $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$ e $\bigcap_{n=1}^{\infty} A_n \in \mathcal{A}$

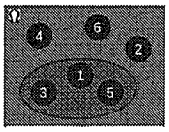
probabilità $P : \mathcal{A} \to \mathbb{R}^+$ tale che: $P(\Omega) = 1$ e se $\{A_n\}_n$ è una successione di elementi di $\mathcal{A}$ a due a due disgiunti, allora

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$$

es. estrazione di 6 palline da un'urna
$\Omega = \{1, 2, 3, 4, 5, 6\}$ insieme di tutti i **possibili** risultati
$\mathcal{A} = \mathcal{P}(\Omega)$ tutte le parti di Ω



$P(\{1\}) = P(\{2\}) = \ldots = P(\{6\}) = p$
$P(\Omega) = P(\{1\} \cup \{2\} \cup \ldots \cup \{6\}) = P(\{1\}) + P(\{2\}) + \ldots + P(\{6\}) = 6p$
ma:
$P(\Omega) = 1$
dunque:
$p = \frac{1}{6}$



L'evento $A \in \mathcal{A} : A = \{1, 3, 5\}$
si può scrivere come unione disgiunta $\{1\} \cup \{3\} \cup \{5\}$ e dunque:
$P(A) = P(\{1\}) + P(\{3\}) + P(\{5\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$



master COPY

Impostazione assiomatica delle probabilità (kolmogorov)

$$1)\ P(A) \geq 0$$
$$2)\ P(\Omega) = 1$$
$$3)\ P(A \cup B) = P(A) + P(B)$$
$$se\ A \cap B = \emptyset$$

P(A|B)→Probabilità di A condizionato B→ probabilità che si verifichi A conoscendo che sia verificato l'evento B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Metodo scientifico: partiamo da ipotesi iniziale, la probabilità è dovuta al caso, faccio i calcoli e cerco di confutare la mia ipotesi iniziale.
La statistica e la teoria delle probabilità servono per:

master COPY

-pianificare correttamente gli esperimenti di verifica

-descrivere e misurare aspetti della natura basandosi su campioni

-stimare una grandezza incognita di una popolazione partendo dalle misure effettuate su campioni

-ottenere un criterio oggettivo per accettare l'ipotesi nulla oppure l'ipotesi alternativa

TEORIA DELLA PROBABILITÀ 09/03/20

master COPY

Il concetto di probabilità è intrinseco a tutti noi e possiamo solo quantificarne l'incertezza. Definizione di funzione di probabilità è applicabile solo se gli eventi sono equiprobabili:

n= numero totale di prove effettuate

master COPY

$$Pr\{A\} = \lim_{n \to \infty} \frac{n_A}{n}$$

n_a=numero di volte in cui si è verificato l'evento A

Pr{A} è il limite della frequenza relativa dell'evento al tendere all'infinito del numero di prove.

Risulterà quindi molto importante imparare a contare quanti sono gli eventi di nostro interesse