ELEMENTI DI TEORIA DELLA PROBABILITA'

Sezione di Epidemiologia e Statistica Medica
Università degli Studi di Verona

ELEMENTI DI TEORIA DELLA PROBABILITA'

La **TEORIA DELLA PROBABILITA'** ci permette di studiare e descrivere gli eventi aleatori.

(alea =dado in latino; alea iacta est = il dado è tratto).

***DEFINIZIONE*:** un **evento** è **aleatorio** quando non si può prevedere con certezza se si avvererà o meno.

Esempi:
numero estratto al lotto / faccia di una moneta / schedina del totocalcio
presenza di un'infezione virale
nascita di un figlio sano
incidente stradale in un adolescente che sta imparando a usare il motorino
sopravvivenza dopo una mastectomia radicale per tumore alla mammella

CONCEZIONE CLASSICA DELLA PROBABILITÀ

La probabilità di un evento A è il rapporto tra il numero di *casi favorevoli* al verificarsi di A (*n*) e il numero di *casi possibili* (*N*), purché tutti i casi siano *equi-probabili*:

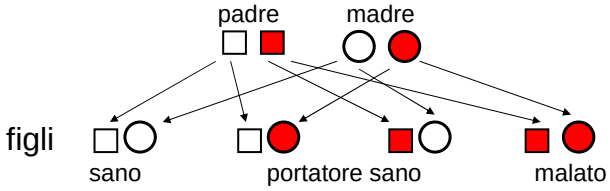
$$P(A) = \frac{n}{N}$$

Tale definizione vale se i possibili risultati sono equi-probabili (gioco d'azzardo)

Esempi: *probabilità di estrarre un asso da un mazzo di 52 carte = 4/52 = 0.08*
probabilità di ottenere testa nel lancio di una moneta = 1/2 = 0.5

⇒ scarsamente applicabile in medicina

Malattie genetiche *(Se entrambi i genitori sono portatori sani del gene della talassemia o della fibrosi cistica, la probabilità di avere un figlio malato è una su quattro).*



CONCEZIONE FREQUENTISTA DELLA PROBABILITÀ

La probabilità di un evento A è la *frequenza relativa di successo* (avverarsi di A) in una serie di prove tendenti all'infinito, ripetute sotto identiche condizioni:

$$P(A) = \lim_{N \to \infty} \frac{n}{N}$$

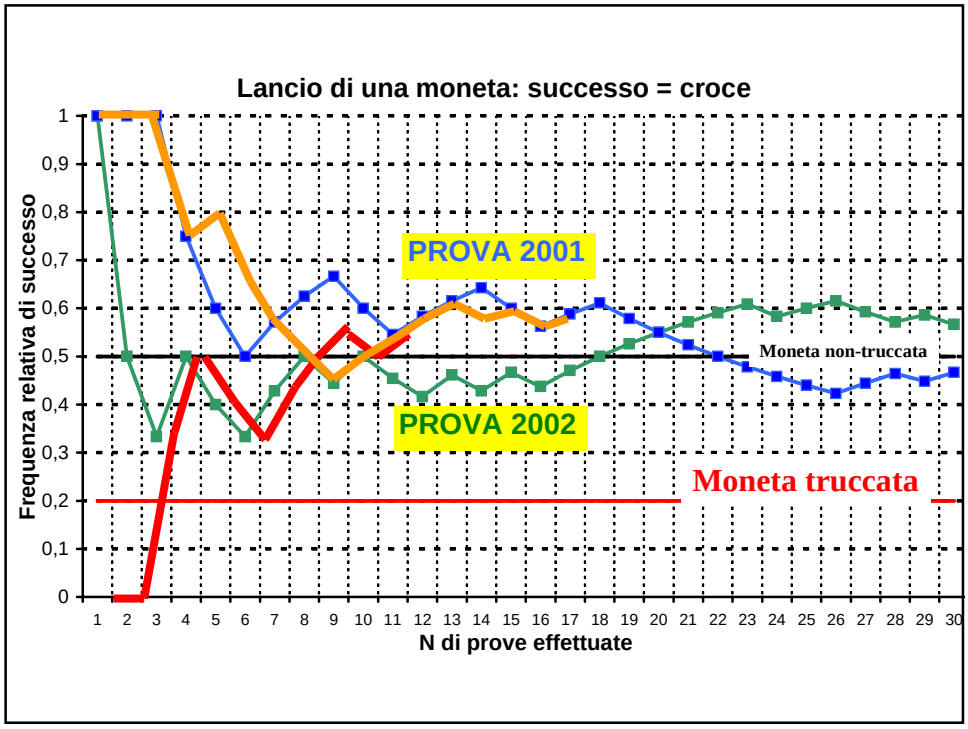
Frequenza relativa su un gran numero di prove

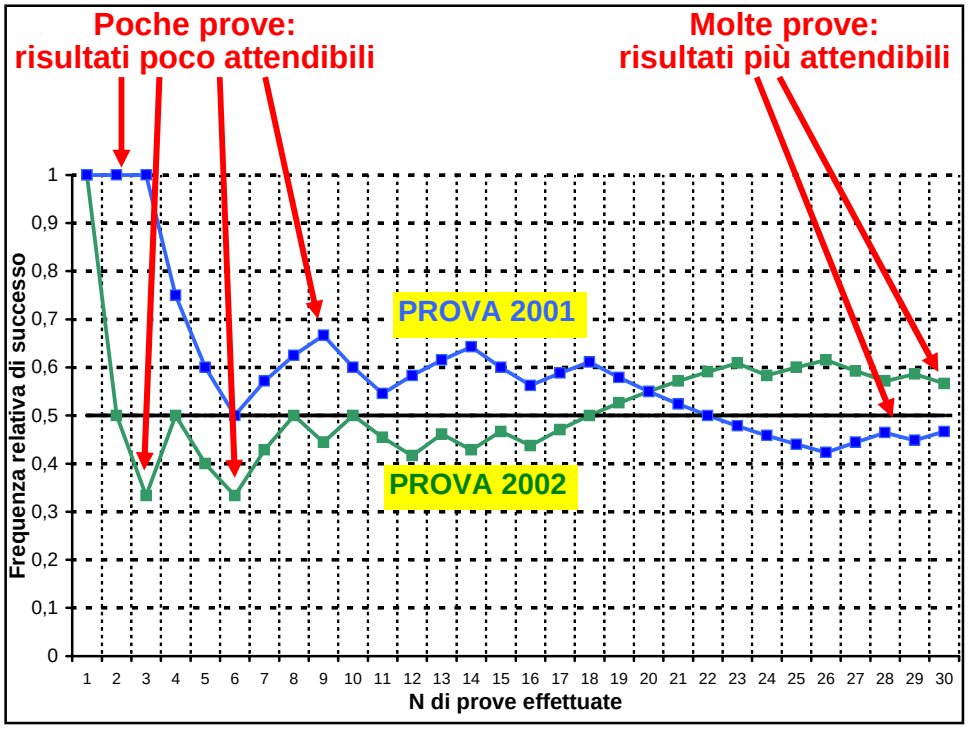
Nella concezione classica la probabilità è stabilita A PRIORI, prima di guardare i dati.
Nella concezione frequentista la probabilità è ricavata A POSTERIORI, dall'esame dei dati.
Nel caso della concezione frequentista, la probabilità viene assegnata sulla base dei risultati di un esperimento ripetuto molte volte nelle stesse condizioni o sulla base di situazioni che possono essere ricondotte a tale contesto concettuale (ad esempio, *utilizzo di statistiche correnti*).

ESEMPIO: *Qual è la mortalità post-operatoria dopo gastrectomia per cancro gastrico?*

Tra il 1988 e il 1998 a Verona, Siena e Forlì ci sono stati 30 morti su 933 resecati.

Frequenza relativa = 30/933 = 3,22% = Probabilità di mortalità post-operatoria


Lancio di una moneta: successo = croce
Frequenza relativa di successo
N di prove effettuate
PROVA 2001
PROVA 2002
Moneta non-truccata
Moneta truccata


Poche prove:
risultati poco attendibili
Molte prove:
risultati più attendibili
Frequenza relativa di successo
N di prove effettuate
PROVA 2001
PROVA 2002

Sapere che la mortalità post-operatoria per cancro gastrico nel periodo 1988-1998 nei 3 centri italiani era di 3,22% è un dato importante per effettuare ad esempio dei confronti internazionali.

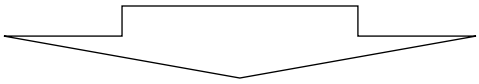
Tuttavia è plausibile che la mortalità postoperatoria dopo gastrectomia per cancro gastrico sia rimasta costante tra il 1988 e il 1998?

Non tutti gli eventi, pur valutabili in termini di probabilità, possiedono il requisito della **ripetitività sotto le stesse condizioni.**

Prima di un intervento neurochirurgico una paziente mi diceva:

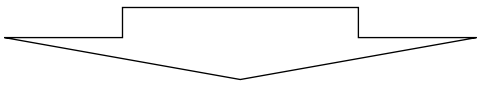
"Ich will die Wurzeln nicht von unten anschauen"

(Non voglio vedere le radici da sotto)



CONCEZIONE SOGGETTIVISTA DELLA PROBABILITÀ

Non tutti gli eventi, pur valutabili in termini di probabilità, possiedono il requisito della **ripetitività sotto le stesse condizioni.**



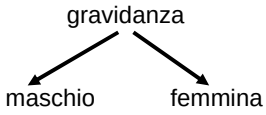
CONCEZIONE SOGGETTIVISTA DELLA PROBABILITÀ

La probabilità di un evento A è la valutazione del grado di fiducia che un individuo o un gruppo di individui coerentemente attribuisce all'avverarsi di A, in base alle proprie opinioni e informazioni

→ TEORIA BAYESIANA

- Riguarda quei fenomeni per i quali l'attesa o la convinzione rispetto all'esito influisce sull'evento stesso (*interventi chirurgici; eventi che dipendono dalla propria volontà, capacità, ...*)
- Riguarda per lo più eventi unici o irripetibili

Qual è la probabilità che un neonato sia femmina?



1 su 2 = 50% (definizione CLASSICA di probabilità)
(probabilità a PRIORI)

Però nel mondo, in assenza di interventi dell'uomo (aborti o infanticidi selettivi, omessa denuncia) nascono 1057 maschi ogni 1000 femmine.

1000 / (1000+1057) = 48,6% (definizione FREQUENTISTA di probabilità)
(probabilità a POSTERIORI)

L'ecografista, alla decima settimana di gravidanza, dice ai genitori che 80 su 100 il neonato è femmina. (definizione SOGGETTIVISTA di probabilità)

L'ecografista, secondo le sue opinioni ed informazioni, esprime coerentemente il suo grado di fiducia nell'avverarsi dell'evento "nascita di una femmina".

Quale approccio, dunque?

Nel contesto delle scienze sperimentali e/od osservazionali, quali la **medicina** e la **biologia** e di conseguenza **l'epidemiologia**, predominano i casi di eventi ripetibili, in condizioni almeno approssimativamente analoghe o simili, pertanto di norma si fa ricorso all'impostazione **frequentista** della probabilità.

Tuttavia quando si approccia il **singolo paziente**, è meglio utilizzare l'impostazione **soggettivista.**

Teoria assiomatica della probabilità

Qualsiasi sia la definizione di probabilità, per probabilità (P) si intende una funzione a valori reali definita sullo spazio campionario S che soddisfa le seguenti condizioni:

1) per qualsiasi evento A che appartiene ad S, si ha che $0 \leq P(A) \leq 1$

(in particolare, P(A) = 1 se A è l'evento certo

P(A) = 0 se A è l'evento impossibile)

2) $P(S) = 1$

p(miglioramento) + p(stazionarietà) + p(peggioramento) = 1
p(Rh negativo) + p(Rh positivo) = 1
La somma della probabilità di tutti gli eventi possibili è uno.

3) se $\{A_1, A_2, \ldots, A_i, \ldots\}$ sono una sequenza finita o infinita di eventi mutuamente esclusivi (o disgiunti) di S, allora

$$P(A_1 \cup A_2 \cup \ldots \cup A_i \cup \ldots) = P(A_1) + P(A_2) + \ldots + P(A_i) + \ldots$$

SPAZIO CAMPIONARIO = insieme di tutti i possibili risultati di un esperimento

Per affrontare i problemi di probabilità disponiamo di 2 ausili grafici importanti:

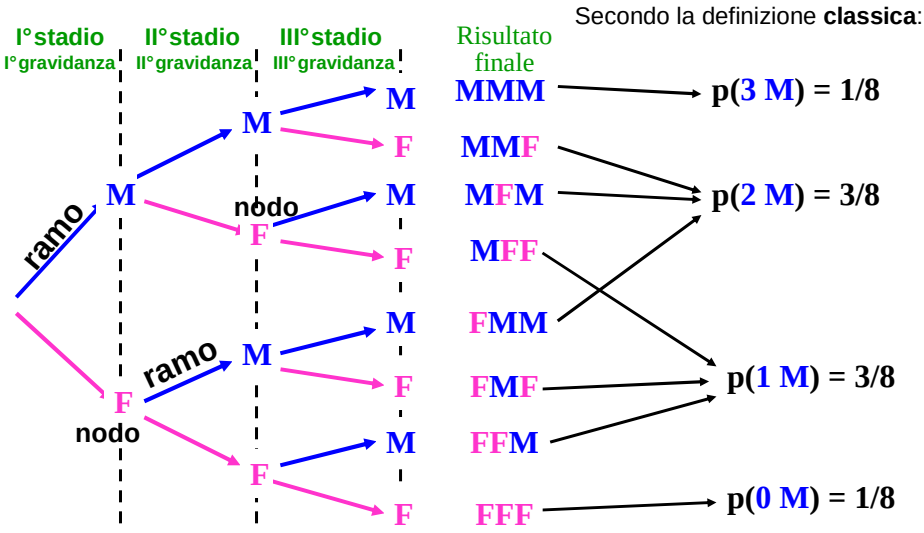
1) Il diagramma ad albero

2) Il diagramma di Eulero-Venn

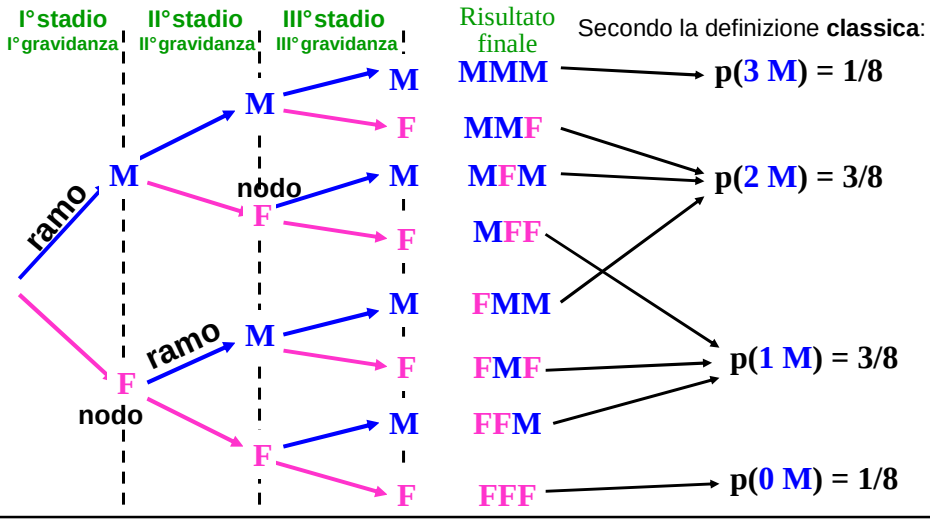
Diagramma ad albero

Se un esperimento è a più stadi, il problema di descrivere i possibili risultati può essere semplificato mediante l'uso di diagrammi ad albero.

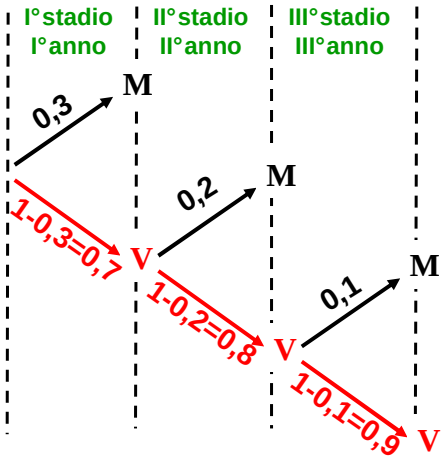
Esempio: *Quanti figli maschi possono nascere su 3 gravidanze?*



- per ogni stadio ci sono tanti rami quante sono le possibilità
- il numero totale di percorsi rappresenta il numero totale di eventi possibili
- ad ogni percorso è associata la probabilità corrispondente all'evento

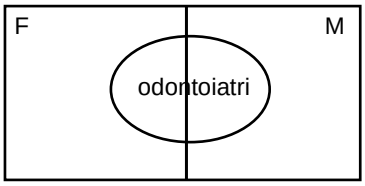


In un paziente affetto da un determinato tipo di tumore, la probabilità di morire nel I anno dalla diagnosi è del 30%, se arriva vivo alla fine del I anno la probabilità di morire nel II anno è del 20% e se arriva vivo alla fine del II anno la probabilità di morire nel III anno è del 10%.

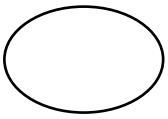


Prob. condizionale di sopravvivenza	**0,7**	**0,8**	**0,9**
Prob. cumulativa di sopravvivenza	**0,7**	**0,7*0,8 =0,56**	**0,56*0,9 =0,504**

Diagramma di Venn: operazione sugli insiemi



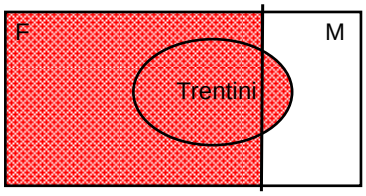
Intera popolazione in studio



sottogruppi

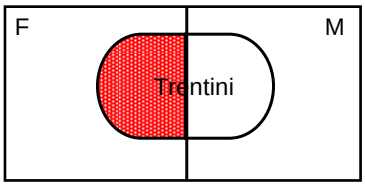
Diagramma di Venn: operazione sugli insiemi

femmine e/o Trentini



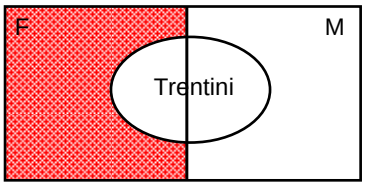
UNIONE U

femmine e Trentine



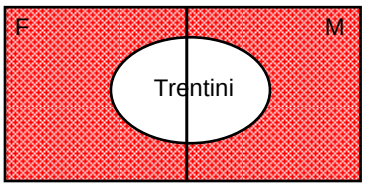
INTERSEZIONE ∩

femmine NON-Trentine



NEGAZIONE

Non-Trentini

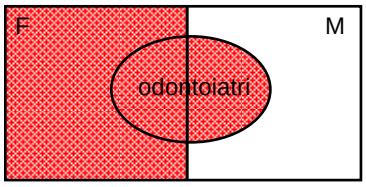


Insieme complementare

$\overline{\text{Trentini}}$

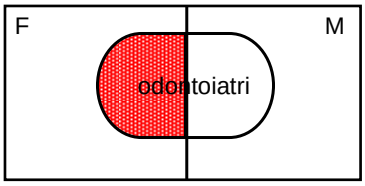
Diagramma di Venn: operazione sugli insiemi

femmine e/o odontoiatri



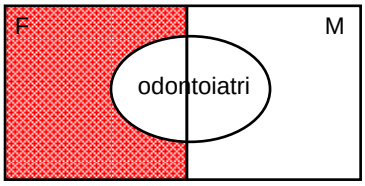
UNIONE U

femmine e odontoiatri



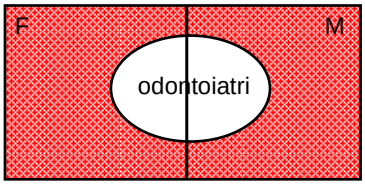
INTERSEZIONE ∩

femmine NON-odontoiatre



NEGAZIONE

Non-odontoiatri

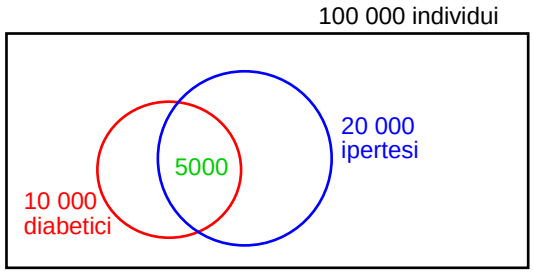


Insieme complementare

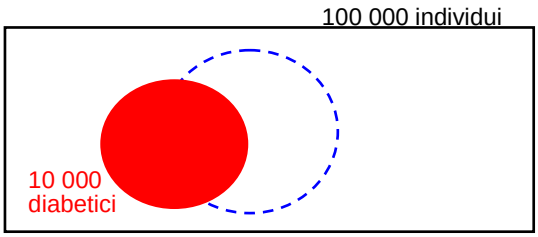
$\overline{\text{odontoiatri}}$

ESERCIZIO: CALCOLO DELLE PROBABILITA'

In una popolazione di 100 000 individui vi sono:
10 000 diabetici (e 90 000 non-diabetici)
20 000 ipertesi (e 80 000 non-ipertesi).
5000 persone che hanno sia il diabete che l'ipertensione.

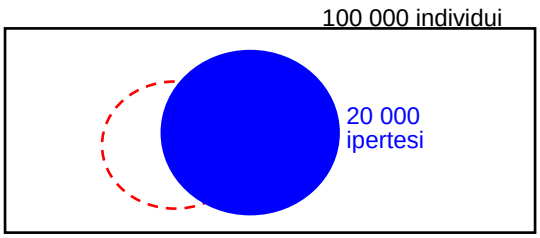


Qual è la probabilità di avere il diabete in quella popolazione?



p (diabete) = 10 000 / 100 000 = 0,1 = 10%

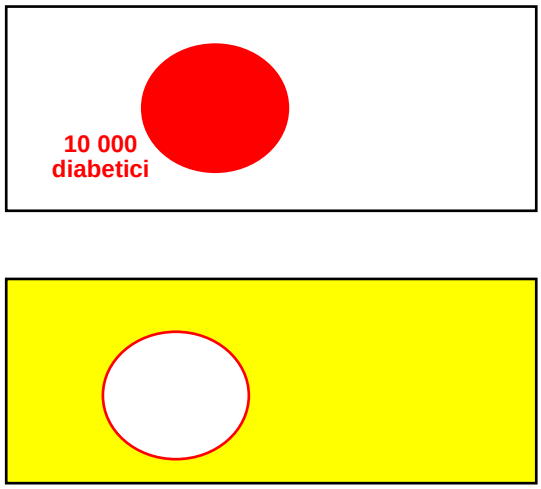
Qual è la probabilità di avere l'ipertensione in quella popolazione?



p (ipertensione) = 20 000 / 100 000 = 0,2 = 20%

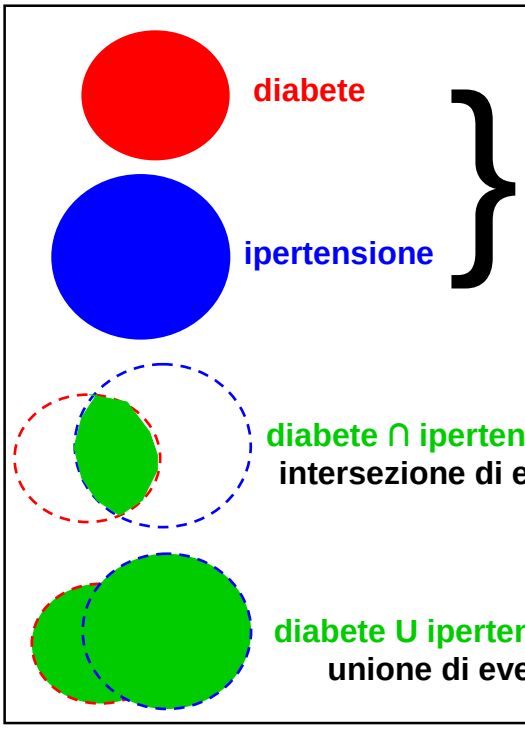
N.B. E' stato usato l'approccio frequentista: la probabilità è stata stimata dalla frequenza relativa

INSIEME COMPLEMENTARE



p (diabete) = 10 000 / 100 000 =
0,1 = 10%

p (non-diabete) =
90 000 / 100 000 = 0,9 = 90%



diabete

ipertensione

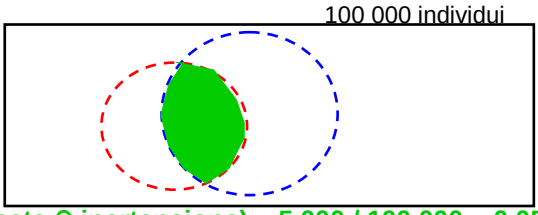
Eventi semplici

diabete ∩ ipertensione
intersezione di eventi

diabete U ipertensione
unione di eventi

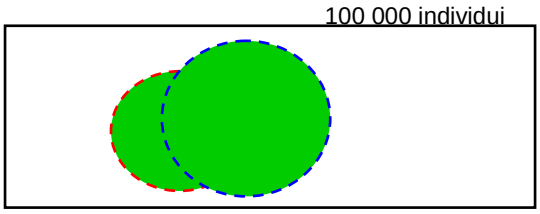
Eventi composti

Qual è la probabilità di avere il diabete **e** l'ipertensione
(**sia il diabete che l'ipertensione**)?



p (diabete ∩ ipertensione) = 5 000 / 100 000 = 0,05 = 5%

Qual è la probabilità di avere il diabete **e/o** l'ipertensione
(**solo il diabete o solo l'ipertensione o entrambi**)?



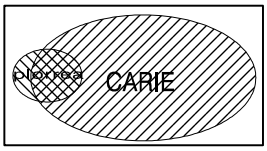
p(diabete U ipertensione) = (10000+20000-5000)/100000 = 25000/100000 =0,25=25%
p(diabete U ipertensione) = p(diabete) + p(ipertensione) - p(diabete ∩ ipertensione)
= 10% + 20% - 5% = 25%

Somma di probabilità

100 000 ab. = POPOLAZIONE TOTALE
80 000 ab. = AFFETTI DA CARIE
6 000 ab. = AFFETTI DA PIORREA

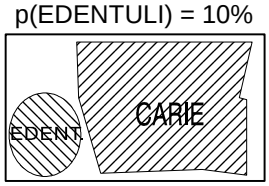
p(CARIE) = 80%
p(PIORREA) = 6%
p(CARIE ∪ PIORREA) = ?

CARIE e/o PIORREA



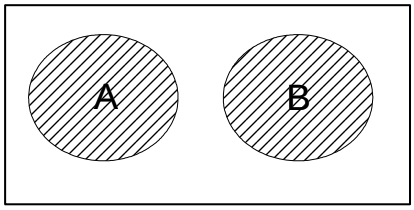
p(CARIE ∩ PIORREA) = 4%
p(CARIE ∪ PIORREA)=0.80+0.06-0.04=
=0.82=82%

10 000 ab. = EDENTULI
p(EDENTULI) = 10%

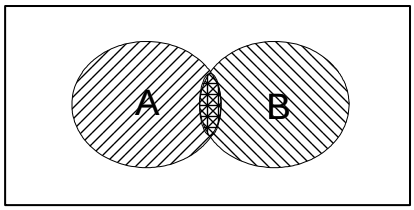


p(CARIE ∪ EDENTULI)=0.80+0.10=0.90

REGOLA DELL'ADDIZIONE



FORMA SEMPLICE:
$P(A \cup B) = P(A) + P(B)$
evento composto



FORMA GENERALE
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

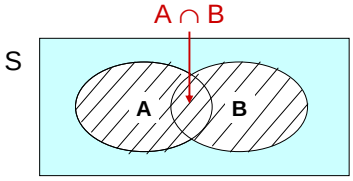
Regole del calcolo della probabilità

Il calcolo della probabilità è estremamente utile per stabilire sia la probabilità associata ad un evento, sia la probabilità associata ad un insieme di eventi.

Regola dell'addizione

Se A e B sono due eventi in S tali che $A \cap B \neq \varnothing$ (eventi non disgiunti):

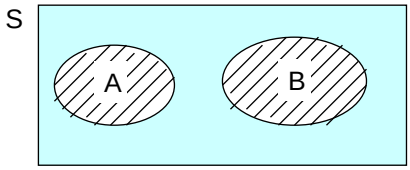
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Esercizio: *calcolare la probabilità di estrarre una carta rossa o una figura da un mazzo di 52 carte*

Se A e B sono due eventi in S tali che $A \cap B = \varnothing$ (eventi disgiunti):

$$P(A \cup B) = P(A) + P(B)$$



Esercizio: *calcolare la probabilità di estrarre una figura o una carta compresa tra 3 e 7 da un mazzo di 52 carte*

Se $\overline{A}$ è il complemento di A in S:

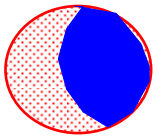
$$P(A \cup \overline{A}) = P(A) + P(\overline{A}) = 1$$

$$P(\overline{A}) = 1 - P(A)$$

Esercizio: *se la probabilità di morire nel I° anno dalla diagnos i per un paziente affetto da tumore al polmone è pari a 0.30, qual è la probabilità di sopravvivere al I° anno?*

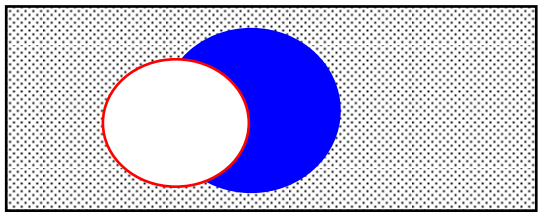
PROBABILITA' CONDIZIONALE

Finora nel calcolo delle probabilità abbiamo messo al denominatore la popolazione globale (100 000 persone). D'ora in poi useremo come denominatore dei sottogruppi particolari della popolazione.



Qual è la probabilità di ipertensione nei diabetici?

p (ipertensione/diabete) = 5 000 / 10 000 = 0,5 = 50%

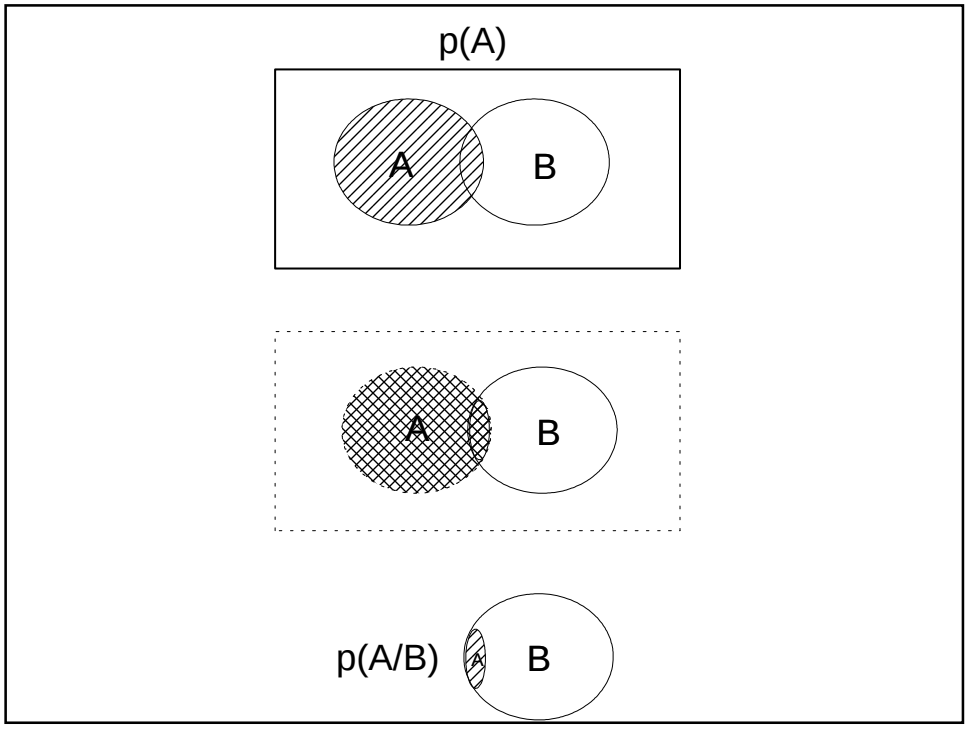


Qual è la probabilità di ipertensione nei non-diabetici?

p (ipertensione/non-diabete) = 15 000 / 90 000 = 0,167 = 16,7%

La probabilità di ipertensione è maggiore fra i diabetici (50%) rispetto ai non-diabetici (16,7%).

Il diabete è un fattore di rischio per l'ipertensione, e le due condizioni sono associate nell'ambito della sindrome plurimetabolica.



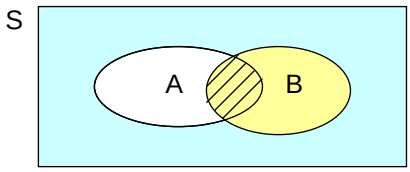
Probabilità condizionale

Talvolta è molto utile conoscere la probabilità di un evento A in S quando si è verificato un altro evento B in S → PROBABILITA' CONDIZIONALE

Esempio: *probabilità di uscita del 7 di quadri dato che è uscita una carta di quadri*
probabilità di avere un tumore al polmone dato che si fuma
probabilità di avere il colera data la presenza di una gastroenterite acuta

Se A e B sono due eventi dello spazio campionario S, si definisce **probabilità condizionale di A dato B**:

$$P(A|B) = P(A \cap B) / P(B)$$



N.B.: *lo spazio campionario dell'evento B diviene il nuovo spazio campionario*.

REGOLA della MOLTIPLICAZIONE di PROBABILITA'

p (diabete) = 10 000 / 100 000 = 0,1 = 10%

p (ipertensione) = 20 000 / 100 000 = 0,2 = 20%

Qual è la probabilità di avere sia il diabete che l'ipertensione?

$p(A \cap B) = P(A) \bullet P(B \mid A)$

p(diabete ∩ ipertensione) = p(diabete) * p(ipertensione/diabete) = 0,1 * 0,5 = 0,05

oppure

$p(A \cap B) = P(B) \bullet P(A \mid B)$

p(diabete ∩ ipertensione) = p(ipertensione) * p(diabete/ipertensione) = 0,2 * 0,25 = 0,05

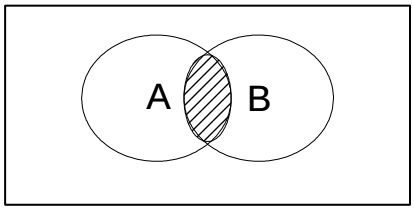
Se i due eventi fossero indipendenti, la probabilità sarebbe 0,1*0,2 = 0,02 = 2%

Pertanto i soggetti con il diabete E l'ipertensione dovrebbero essere
100 000 *0,02 = 2000 (ATTESI sotto l'ipotesi di indipendenza)

Ma i soggetti che hanno entrambe le condizioni sono 5000 (OSSERVATI)

Gli osservati sono molti di più degli attesi:
le variabili diabete e ipertensione non sono statisticamente indipendenti.

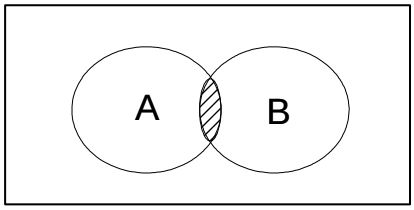
REGOLA DELLA MOLTIPLICAZIONE



FORMA SEMPLICE:

$P(A \cap B) = P(A) * P(B)$

eventi indipendenti



FORMA GENERALE:

$P(A \cap B) = P(A) * P(B/A)$

prob. congiunta — prob. condizionale

Dalla definizione di probabilità condizionale segue la REGOLA DELLA MOLTIPLICAZIONE:

$$P(A \cap B) = P(B) \bullet P(A \mid B)$$
$$= P(A) \bullet P(B \mid A)$$

Se il verificarsi di B non condiziona la probabilità del verificarsi di A, segue che:

$$P(A \mid B) = P(A)$$

e i due **eventi** sono detti **indipendenti**, ovvero: $P(A \cap B) = P(A) \bullet P(B)$

Prodotto di probabilità e sindrome plurimetabolica

Nello studio di Brunico (Bonora et al, Diabetes 47: 1643-1649, 1998):

N = 888

	Prevalenza
ridotta tolleranza glucidica	16,6%
dislipidemia	29,2%
iperuricemia	15,4%
ipertensione	37,3%

Se queste condizioni fossero indipendenti, la probabilità dell'intersezione (avere tutti e 4 i disturbi simultaneamente) sarebbe pari a:

0,166*0,292*0,154*0,373 = 0,0028 = 0,28%

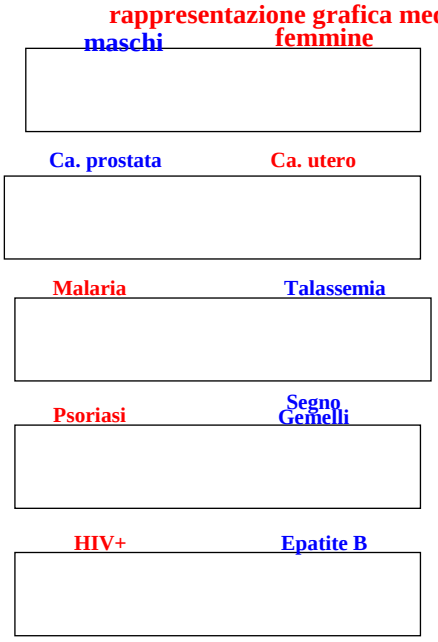
Gli attesi (soggetti con tutte e 4 le malattie sotto l'ipotesi di indipendenza dovrebbero essere) = N * p = 888*0,0028 = 2,5.

Invece se ne osservano 21.

Dal momento che gli osservati (21) sono molti di più degli attesi (2,5) si conclude che queste patologie non si riscontrano per caso negli stessi soggetti, ma rappresentano le diverse espressioni di una stessa patologia, la sindrome plurimetabolica.

Sindrome = insieme di sintomi e segni, in apparenza non collegati tra loro

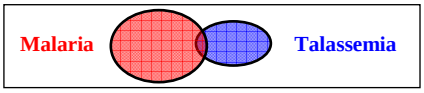
Dipendenza e indipendenza statistica
rappresentazione grafica mediante diagramma di Venn



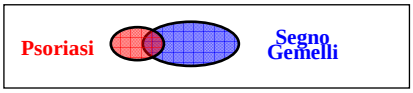
Dipendenza e indipendenza statistica
rappresentazione grafica mediante diagramma di Venn

Massima dipendenza negativa: eventi mutuamente esclusivi ed esaustivi

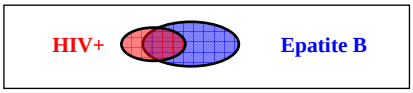
Massima dipendenza negativa: eventi mutuamente esclusivi (non esaustivi)



Dipendenza negativa (parziale): la talassemia protegge dalla malaria



Eventi statisticamente indipendenti: p(psoriasi/Gemelli) = p(psoriasi/altri_segni)



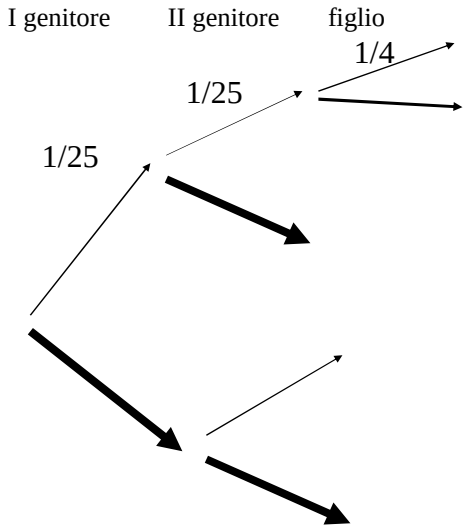
Dipendenza positiva: infezione da HIV e epatite B hanno la stessa modalità di trasmissione

La mucoviscidosi o fibrosi cistica (del pancreas) è una delle malattie genetiche più diffuse.

In Italia un adulto su 25 è portatore sano.

La malattia è di tipo autosomico recessivo.

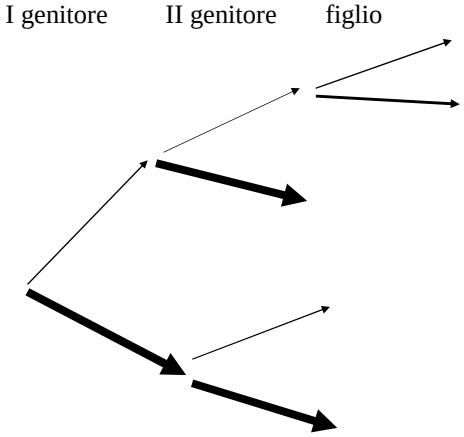
Qual è la probabilità che nasca un bambino affetto da mucoviscidosi?



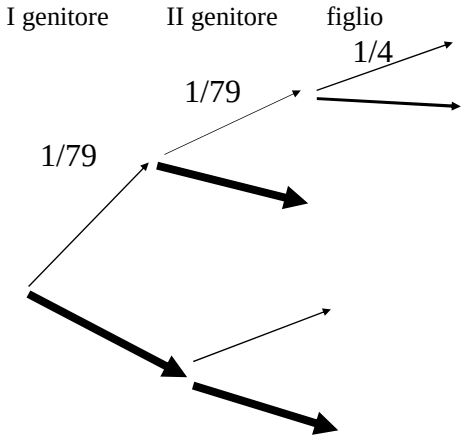
Probabilità che nasca un bambino con la mucoviscidosi

p(padre portatore sano) * p(madre portatrice sana) * p(figlio malato da 2 portatori sani) = (1/25) * (1/25) * (1/4) = 1/2500

In Finlandia la prevalenza della sindrome pluriendocrina è di 1 su 25000. Sapendo che questa malattia ha una trasmissione autosomica recessiva come la mucoviscidosi, qual è la prevalenza di portatori sani ?



Relazione tra numero di portatori sani e numero di affetti nella sindrome pluriendocrina in Finlandia



p(nascita di un individuo affetto da sindrome pluriendocrina) = p(padre portatore sano) * p(madre portatrice sana) * p(figlio malato da 2 portatori sani) = (1/79) * (1/79) * (1/4) = 1/24964

CALCOLO COMBINATORIO: CENNI

Cenni di calcolo combinatorio

Mettete in ordine descrescente i seguenti attori sulla base delle vostre preferenze:
Leonardo Di Caprio, Harrison Ford, Tom Cruise, Sean Connery, Ezio Greggio, Michele Placido

	I posto	II posto	III posto	IV posto	V posto	VI posto
	Di Caprio Ford Cruise Connery Greggio Placido	Di Caprio Ford Cruise Greggio Placido	Di Caprio Ford Cruise Greggio	Di Caprio Cruise Greggio	Di Caprio Greggio	Di Caprio
Scelte possibili	6	5	4	3	2	1

In tutto le scelte possibili sono 6*5*4*3*2*1 = 6! = 720.
Posso formare 720 gruppi che differiscono per l'ORDINE degli oggetti, ovvero n! PERMUTAZIONI.

Tra i seguenti 6 attori, sceglietene due, il primo e il secondo sulla base della vostra simpatia:
Leonardo Di Caprio, Harrison Ford, Tom Cruise, Sean Connery, Ezio Greggio, Michele Placido

	I posto	II posto	III posto	IV posto	V posto	VI posto
	Di Caprio Ford Cruise Connery Greggio Placido	Di Caprio Ford Cruise Greggio Placido	Di Caprio Ford Cruise Greggio	Di Caprio Cruise Greggio	Di Caprio Greggio	Di Caprio
Scelte possibili	6	5	4	3	2	1

In tutto le scelte possibili sono (6*5*4*3*2*1)/(4*3*2*1) = = 6!/4! = 30.
Con 6 oggetti posso formare 30 gruppi di 2 oggetti, che differiscano per l'ORDINE e il TIPO degli oggetti stessi, ovvero n!/(n-k)! DISPOSIZIONI.

Tra i seguenti 6 attori, scegliete i 2 che vi sono più simpatici:
Leonardo Di Caprio, Harrison Ford, Tom Cruise,
Sean Connery, Ezio Greggio, Michele Placido

	I posto	II posto
	Di Caprio Ford Cruise Connery Greggio Placido	Di Caprio Ford Cruise Greggio Placido
Scelte possibili	6	5

Se non si tiene conto dell'ORDINE,
il numero dei gruppi si dimezza:
Connery Placido = Placido Connery

In tutto le scelte possibili sono (6*5*4*3*2*1)/[(4*3*2*1)*2] =
= 6!/(4! 2!) = 15.
Con 6 oggetti posso formare 15 gruppi di 2 oggetti, che
differiscono fra loro per il TIPO degli oggetti stessi,
ovvero n!/[(n-k)!*k!] COMBINAZIONI.