

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

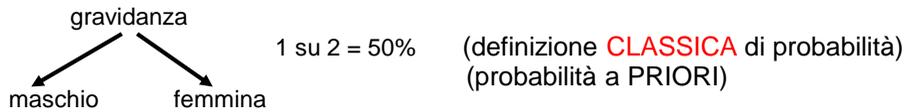
presenza di un'infezione virale

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

Qual è la probabilità che un neonato sia femmina?



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".

Campionato del Mondo di calcio 2014

Definizione classica

Definizione frequentista

Ci sono 32 squadre.
Pertanto la probabilità
che ogni squadra ha di
vincere è $1/32=3.125\%$

Anno	Vincitore	Tabella di frequenza al 2010	
		Freq Ass	Freq Rel
1930	Uruguay		
1934	Italia	Brasile 5	26%
1938	Italia	Italia 4	21%
1950	Uruguay	Germania 3	16%
1954	Germania	Argentina 2	11%
1958	Brasile	Uruguay 2	11%
1962	Brasile	Spagna 1	5%
1966	Inghilterra	Francia 1	5%
1970	Brasile	Inghilterra 1	5%
1974	Germania	Totale 19	100%
1978	Argentina		
1982	Italia		
1986	Argentina	Squadre in semifinale:	
1990	Germania	Germania I	
1994	Brasile	Argentina II	
1998	Francia	PaesiBassi III	
2002	Brasile	Brasile IV	
2006	Italia		
2010	Spagna		

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*:

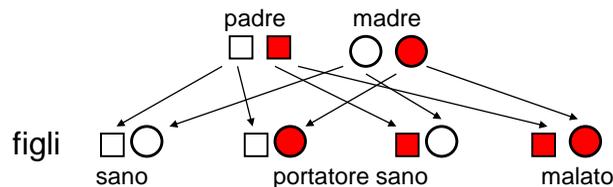
$$P(A) = \frac{n}{N}$$

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 (*legge dei grandi numeri*):

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

Frequenza relativa su un gran numero di prove

Nella concezione classica la probabilità è stabilita A PRIORI, prima di guardare i dati.

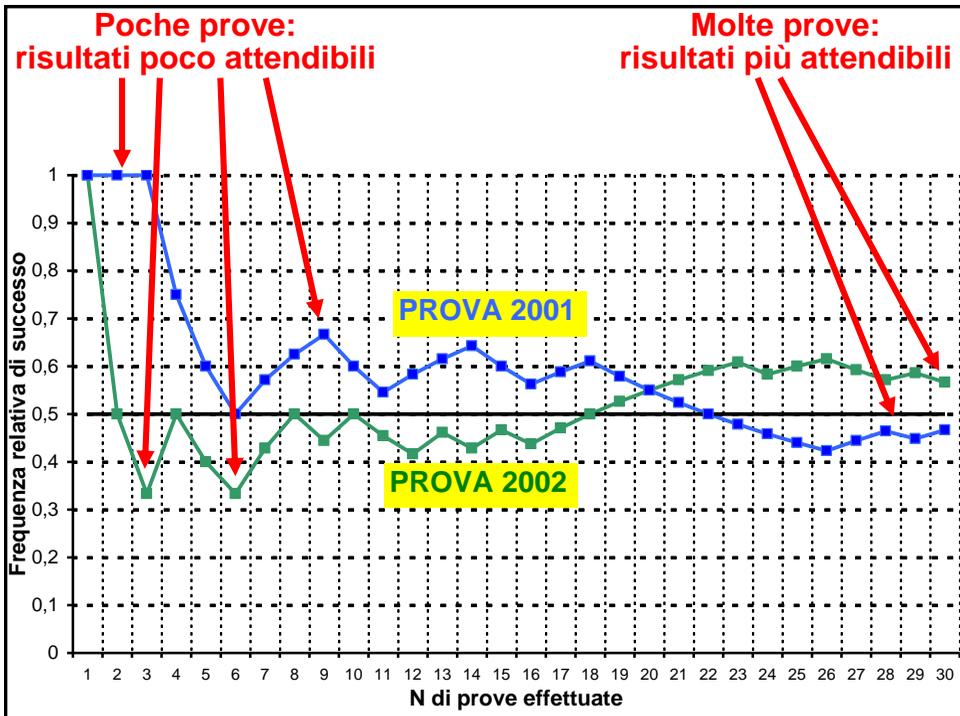
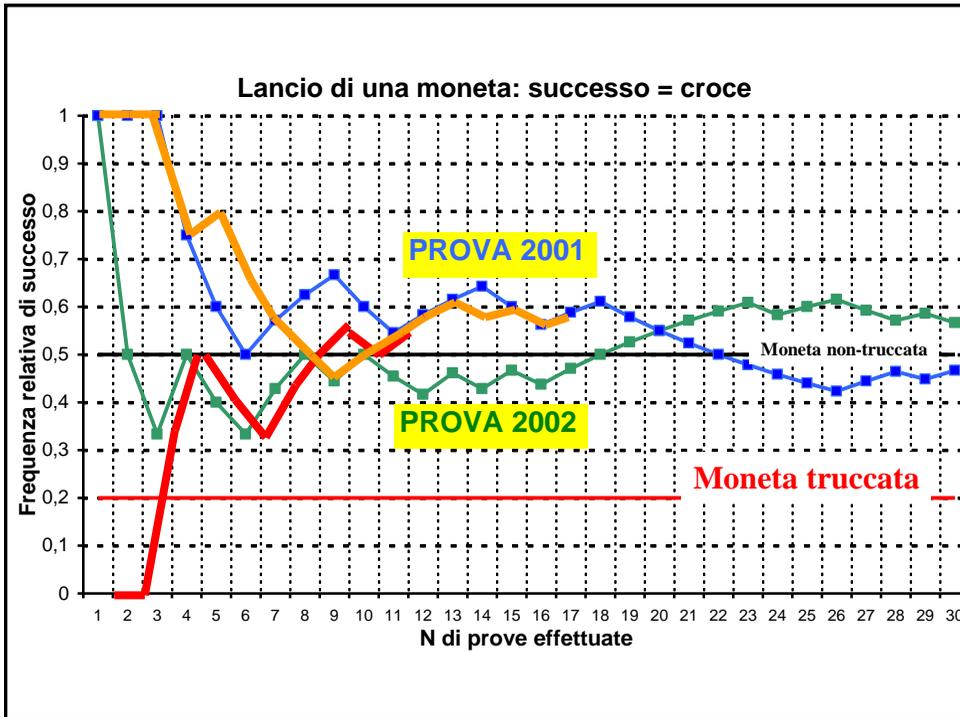
Nella concezione frequentista la probabilità è ricavata A POSTERIORI, dall'esame dei dati.

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



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:

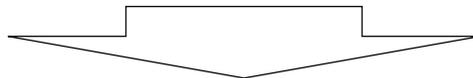
“Ich will die Wurzeln nicht von unten anschauen”

(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**

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) $P(S) = 1$ $p(\text{miglioramento}) + p(\text{stazionarietà}) + p(\text{peggioramento}) = 1$
 $p(\text{Rh negativo}) + p(\text{Rh positivo}) = 1$

La somma della probabilità di tutti gli eventi possibili è uno.

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

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

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

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

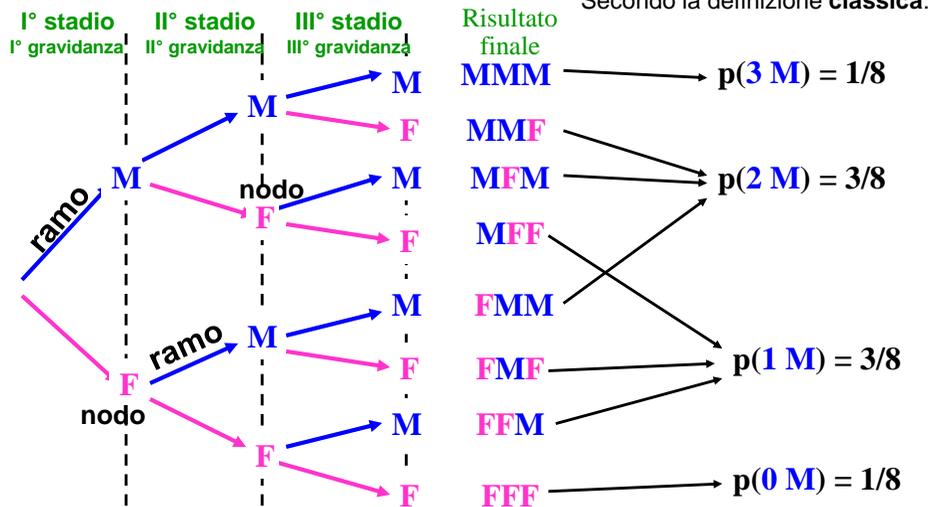
- 1) Il diagramma ad albero
- 2) Il diagramma di Eulero-Venn

Diagramma ad albero

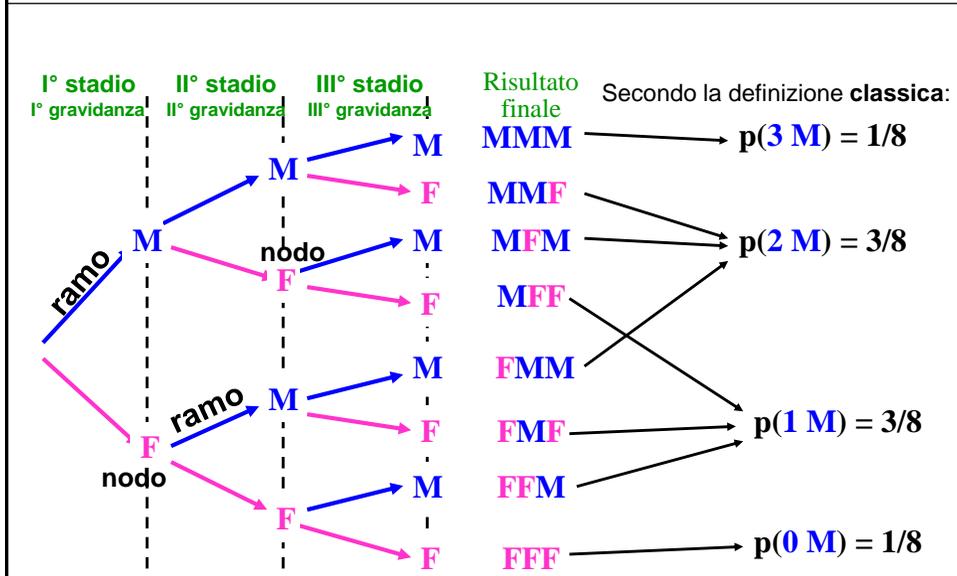
Se un esperimento è a più stadi, il problema di descrivere i possibili risultati può essere semplificato mediante l'uso di **diagrammi ad albero**.

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

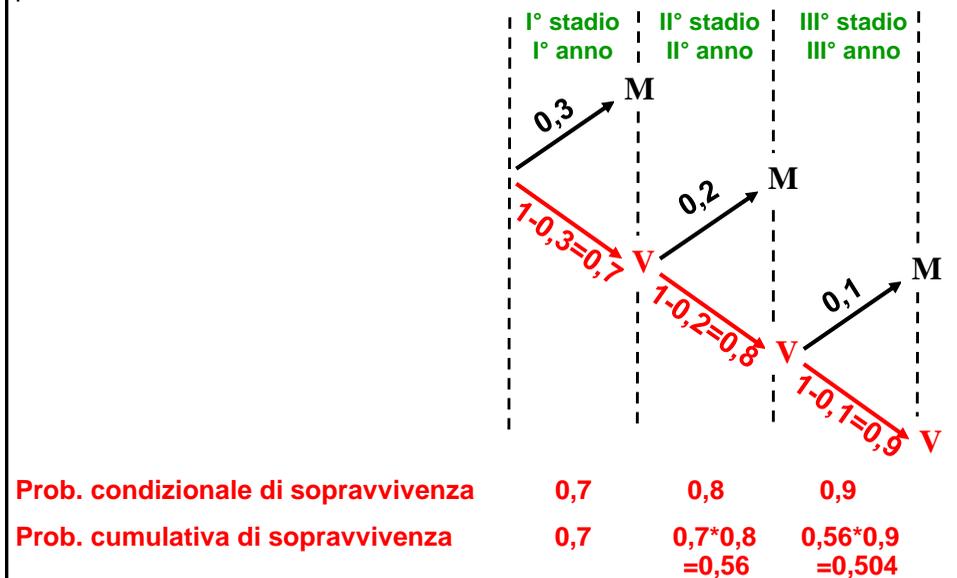
Secondo la definizione **classica**:



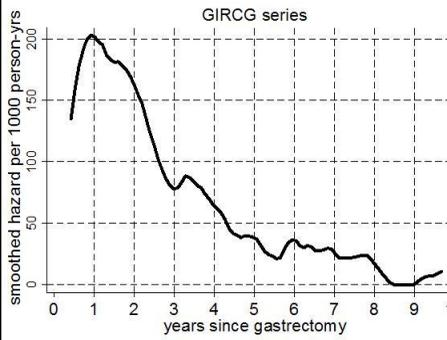
- per ogni stadio ci sono tanti rami quante sono le possibilità
- il numero totale di percorsi rappresenta il numero totale di eventi possibili
- ad ogni percorso è associata la probabilità corrispondente all'evento



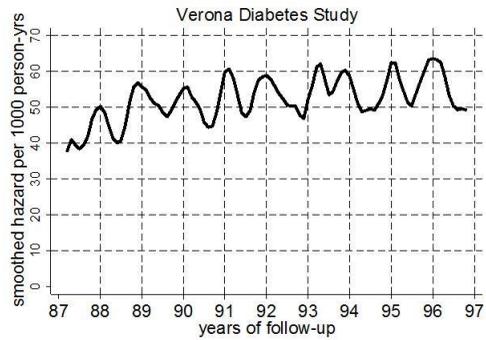
In un paziente affetto da un determinato tipo di tumore, la probabilità di morire nel I anno dalla diagnosi è del 30%, se arriva vivo alla fine del I anno la probabilità di morire nel II anno è del 20% e se arriva vivo alla fine del II anno la probabilità di morire nel III anno è del 10%.



Mortalità nel cancro gastrico

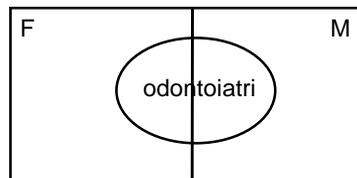


Mortalità nel diabete di tipo 2

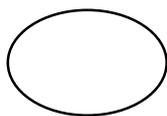


Verlato et al, World J Gastroenterol 2014

Diagramma di Venn: operazione sugli insiemi

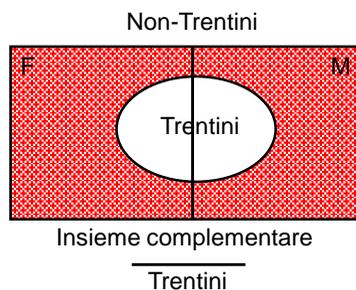
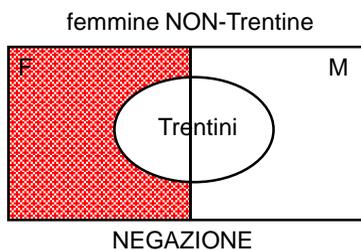
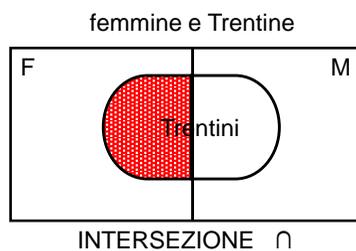
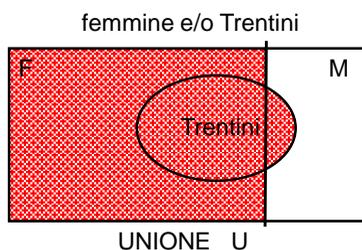


Intera popolazione in studio



sottogruppi

Diagramma di Venn: operazione sugli insiemi



REGOLE della PROBABILITA'

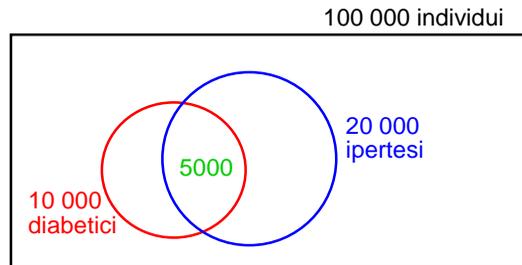
Probabilità dell'UNIONE di eventi ----→ Regola dell'ADDIZIONE

Probabilità della NEGAZIONE di eventi → Regola della SOTTRAZIONE

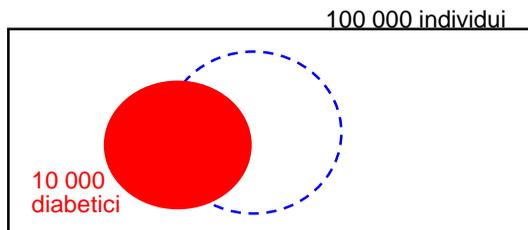
Probabilità dell'INTERSEZIONE di eventi → Regola della MOLTIPLICAZIONE

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.

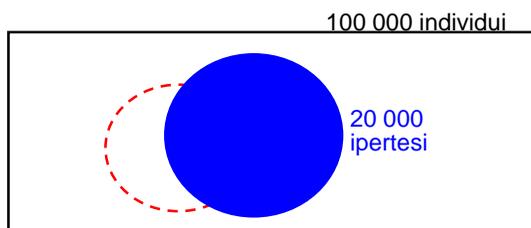


Qual è la probabilità di avere il diabete in quella popolazione?

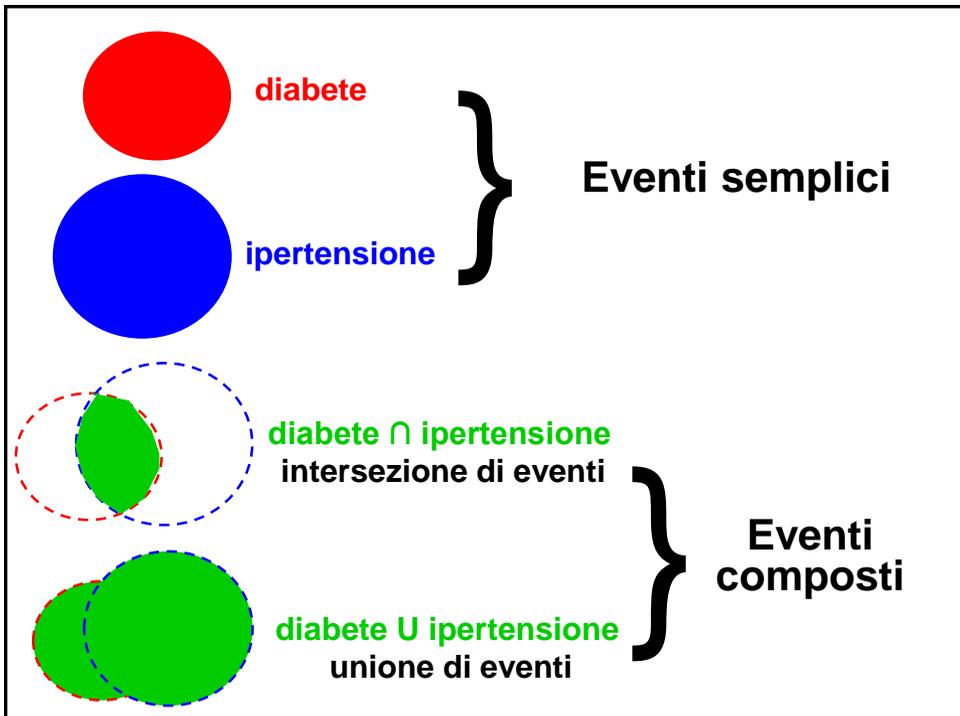
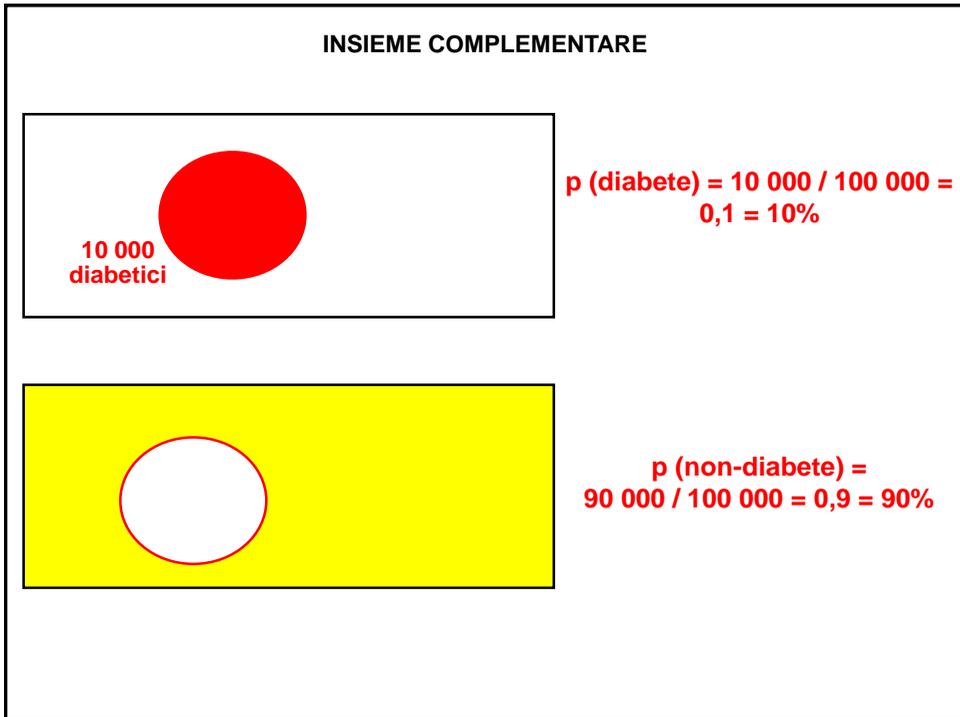


$$p(\text{diabete}) = 10\,000 / 100\,000 = 0,1 = 10\%$$

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

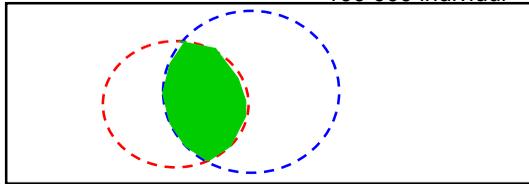


$$p(\text{ipertensione}) = 20\,000 / 100\,000 = 0,2 = 20\%$$



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

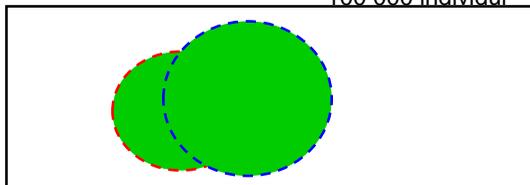
100 000 individui



$$p(\text{diabete} \cap \text{ipertensione}) = 5\,000 / 100\,000 = 0,05 = 5\%$$

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

100 000 individui



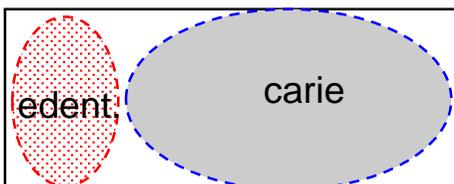
$$p(\text{diabete} \cup \text{ipertensione}) = (10\,000 + 20\,000 - 5\,000) / 100\,000 = 25\,000 / 100\,000 = 0,25 = 25\%$$

$$p(\text{diabete} \cup \text{ipertensione}) = p(\text{diabete}) + p(\text{ipertensione}) - p(\text{diabete} \cap \text{ipertensione}) = 10\% + 20\% - 5\% = 25\%$$

Somma di probabilità

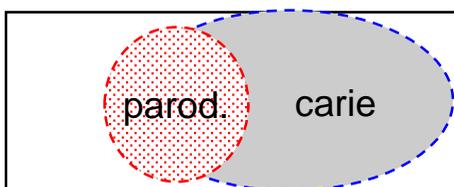
Popolazione totale = 100.000
Affetti da carie = 70.000
Edentuli (senza denti) = 15.000
Affetti da parodontite = 20.000

$p(\text{carie}) = 70\%$
 $p(\text{edentulo}) = 15\%$
 $p(\text{parodontite}) = 20\%$



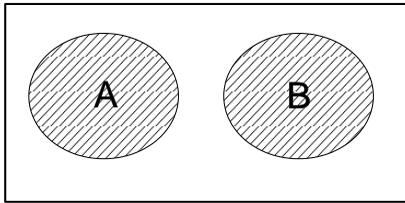
$$p(\text{carie} \cup \text{edentulo}) = p(\text{carie}) + p(\text{edentulo}) = 70\% + 15\% = 85\%$$

Carie + parodontite = 16.000
 $p(\text{carie} \cap \text{parodontite}) = 16\%$



$$p(\text{carie} \cup \text{parodontite}) = p(\text{carie}) + p(\text{parodontite}) - p(\text{carie} \cap \text{parodontite}) = 70\% + 20\% - 16\% = 74\%$$

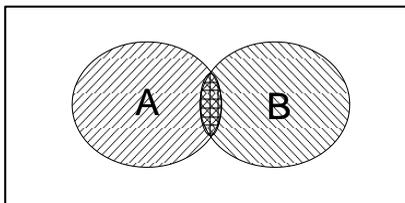
REGOLA DELL'ADDIZIONE



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

I due eventi A e B sono disgiunti (mutuamente esclusivi)

$$\begin{aligned} P(A \cup B) &= \\ P(A) + P(B) - P(A \cap B) &= \\ P(A) + P(B) - 0 &= \\ P(A) + P(B) & \end{aligned}$$



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

I due eventi A e B hanno un'intersezione

Regola della sottrazione (negazione)

La probabilità, che si verifichi un evento A, è uguale a 1 meno la probabilità che l'evento A non si verifichi.

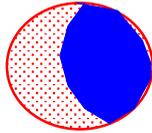
$$P(A) = 1 - P(\bar{A})$$



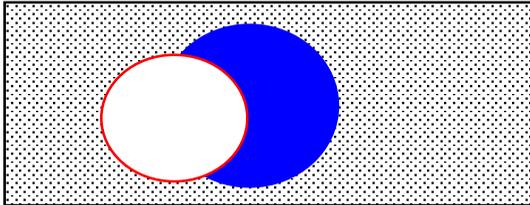
$$P(\text{vivo}) = 1 - P(\text{morto})$$

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?
 $p(\text{ipertensione}/\text{diabete}) = 5\ 000 / 10\ 000 = 0,5 = 50\%$



Qual è la probabilità di ipertensione nei non-diabetici?

$p(\text{ipertensione}/\text{non-diabete}) = 15\ 000 / 90\ 000 = 0,167 = 16,7\%$

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.

La probabilità condizionale di A dato B è la probabilità che l'evento A abbia luogo, dato che si è già verificato l'evento B.

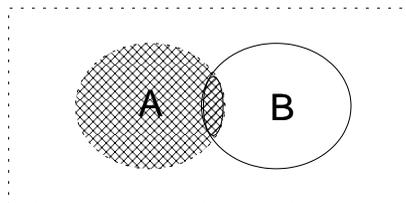
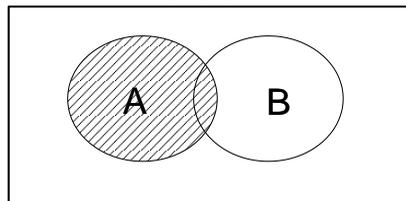
Esempi:

Probabilità di avere il cancro del polmone dato che uno ha fumato 20 pacchetti-anno.

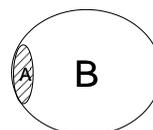
Probabilità di sviluppare l'asma dato che uno soffre già di rinite allergica.

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

$p(A)$



$p(A/B)$



La **REGOLA della MOLTIPLICAZIONE** può essere derivata dalla definizione di probabilità condizionale:

$$P(A | B) = P(A \cap B) / P(B) \quad \text{Probabilità condizionale}$$



$$P(A \cap B) = P(B) \cdot P(A | B) \quad \text{Regola della moltiplicazione}$$
$$= P(A) \cdot P(B | A)$$

Se i due eventi sono **indipendenti**: $P(A | B) = P(A)$

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



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

REGOLA della MOLTIPLICAZIONE di PROBABILITA'

$$p(\text{diabete}) = 10\,000 / 100\,000 = 0,1 = 10\%$$

$$p(\text{ipertensione}) = 20\,000 / 100\,000 = 0,2 = 20\%$$

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

$$p(A \cap B) = P(A) \cdot P(B | A)$$

$$p(\text{diabete} \cap \text{ipertensione}) = p(\text{diabete}) * p(\text{ipertensione}/\text{diabete}) = 0,1 * 0,5 = 0,05$$

oppure

$$p(A \cap B) = P(B) \cdot P(A | B)$$

$$p(\text{diabete} \cap \text{ipertensione}) = p(\text{ipertensione}) * p(\text{diabete}/\text{ipertensione}) = 0,2 * 0,25 = 0,05$$

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

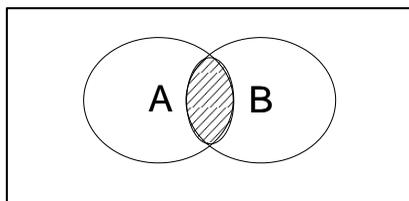
Pertanto i soggetti con il diabete E l'ipertensione dovrebbero essere

$$100\,000 * 0,02 = 2000 \quad (\text{ATTESI sotto l'ipotesi di indipendenza})$$

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

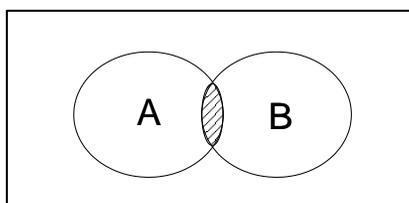
Gli osservati sono molti di più degli attesi:
le variabili diabete e ipertensione non sono statisticamente indipendenti.

REGOLA DELLA MOLTIPLICAZIONE



FORMA SEMPLICE:
 $P(A \cap B) = P(A) * P(B)$
eventi indipendenti

I due eventi A e B sono statisticamente indipendenti. In altre parole, un evento non modifica la probabilità che si verifichi l'altro evento.



FORMA GENERALE:
 $P(A \cap B) = P(A) * P(B/A)$
prob. congiunta prob. condizionale

Questa formula può essere usata per qualsiasi coppia di eventi, sia indipendenti che dipendenti.

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.

Dipendenza e indipendenza statistica
rappresentazione grafica mediante diagramma di Venn

maschi **femmine**



Ca. prostata **Ca. utero**



Malaria **Talassemia**



Psoriasi **Segno Gemelli**



HIV+ **Epatite B**



Dipendenza e indipendenza statistica
rappresentazione grafica mediante diagramma di Venn

maschi **femmine**



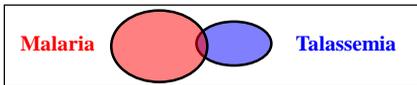
Massima dipendenza negativa: eventi mutuamente esclusivi ed esaustivi

Ca. prostata **Ca. utero**



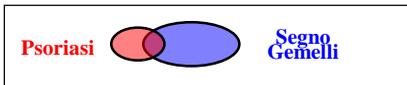
Massima dipendenza negativa: eventi mutuamente esclusivi (non esaustivi)

Malaria **Talassemia**



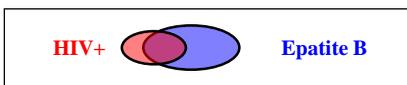
Dipendenza negativa (parziale): la talassemia protegge dalla malaria

Psoriasi **Segno Gemelli**



Eventi statisticamente indipendenti:
 $p(\text{psoriasi/Gemelli}) = p(\text{psoriasi/altri_segni})$

HIV+ **Epatite B**



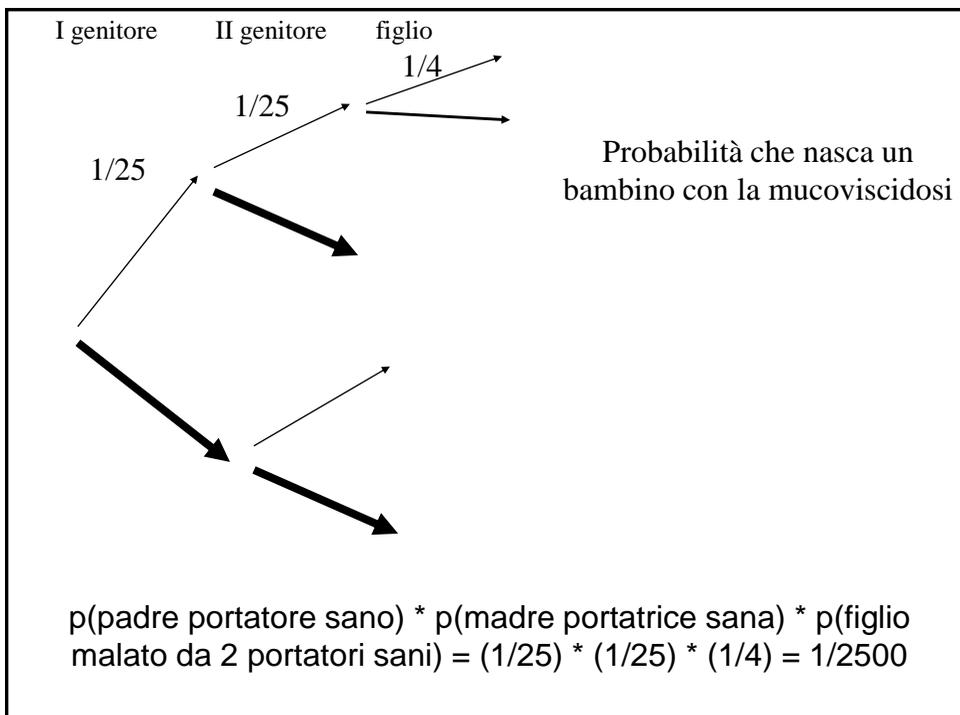
Dipendenza positiva: infezione da HIV e epatite B hanno la stessa modalità di trasmissione

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

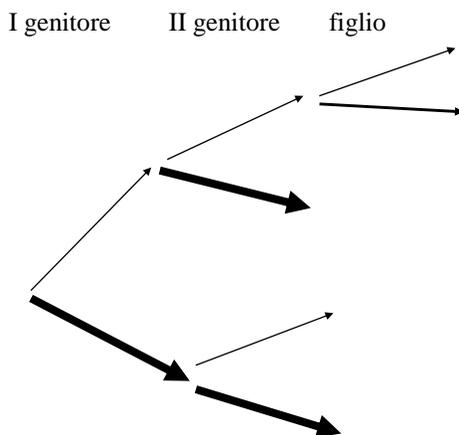
In Italia un adulto su 25 è portatore sano.

La malattia è di tipo autosomico recessivo.

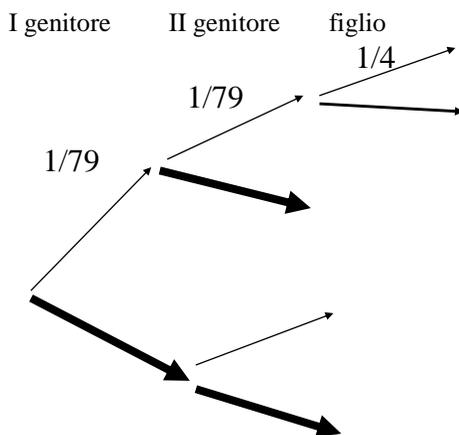
Qual è la probabilità che nasca un bambino affetto da mucoviscidosi?



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



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

Equilibrio di Hardy-Weinberg

p e q sono rispettivamente le frequenze alleliche degli alleli A e B

I genitore

II genitore

