Dalla FREQUENZA alla PROBABILITA'

Esempio clinico – variabile dicotomica

X : presenza/assenza di ipertensione, uomo 50-55 anni.

Si osservano 1000 individui, 380 sono ipertesi, si conclude che la **proporzione (frequenza) di ipertesi è 0.38=38%**

Questa è una **stima** della **probabilita' (ignota)** di ipertensione nell'intera popolazione - uomo 50-55 anni.

La probabilità di ipertensione 'vera' potrebbe essere

P(X = 'Iperteso') = 0.40 = 40%.

La Probabilità : un meccanismo che lavora dietro le quinte

Guida la realizzazione dei fenomeni aleatori ed in generale la manifestazione delle variabili sulle unità statistiche

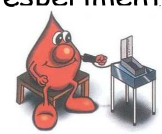
La performance di un balletto dipende dalle scelte del coreografo e da altri elementi casuali propri del momento

Nella statistica descrittiva si parla di variabili/rilevazioni e dati/modalità…

Nella teoria della Probabilità si parla di esperimenti (rilevazioni) ed eventi (modalità)

X : presenza/assenza di ipertensione, uomo 50-55 anni.

X è una variabile che ottengo con un semplice esperimento : valutazione della pressione del soggetto.



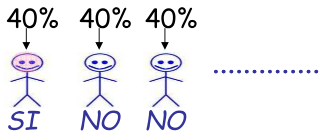
L'esito della valutazione puo' dare origine ai due eventi (**eventi semplici**)

X = 'iperteso' e X = 'non iperteso'

Questi sono i dati che verranno osservati.

La Probabilità agisce dietro le quinte sulla singola unità statistica nel determinare la realizzazione della variabile osservata

(a differenza della frequenza relativa)



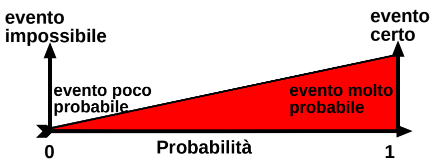
Globalmente il numero di SI che osservo sui 1000 analizzati tendera' ad essere uguale ad un 40%

Probabilità

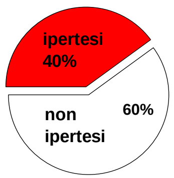
La PROBABILITÀ MISURA L'INCERTEZZA del verificarsi di un evento

Definizione di PROBABILITÀ : ASSIOMI

1. La probabilità di osservare un evento è compresa tra 0 (0%) e 1 (100%)



2. La somma delle probabilita' di tutti i possibili eventi è 1 (100%)



3. La probabilita' è una funzione additiva
   (eventi incompatibili tra di loro)

P(X = 'iperteso') + P(X = 'non iperteso') = 1 = 100%

Definizione di PROBABILITÀ :

- CLASSICA: "La probabilità di un evento A è pari al rapporto fra i casi favorevoli all'evento e tutti i casi ugualmente possibili"
- FREQUENTISTA : "La probabilità di un evento A è approssimata dalla frequenza relativa con cui l'evento si manifesta, qualora si ripeta il medesimo esperimento un numero molto elevato di volte"
- SOGGETIVISTA: "La probabilità di un evento A misura il grado di fiducia, espresso fra 0 e 1, che un individuo (coerente) attribuisce all'evento A, sulla base delle informazioni in suo possesso"

La Probabilità : un concetto che lavora dietro le quinte

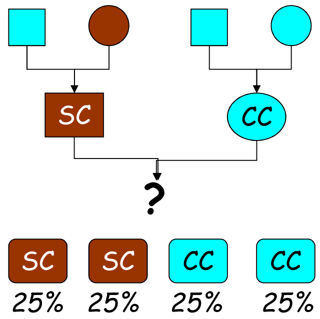
Figlio con mutazione genetica in funzione del patrimonio dei genitori (leggi della natura)

Risposta ad un farmaco (studio empirico)

Valutazione medica (leggi della natura, studio empirico, valutazione soggettiva)

Esempio genetica – UNA VARIABILE dicotomica

X : colore occhi di un figlio (C=chiaro/S=scuro)



Senza fare esperimenti di alcun tipo, conosco la 'legge che lavora dietro le quinte' del fenomeno: S è dominante.

Figlio sara S = SC o C = CC a seconda di cosa ereditera' dal padre.

$$P(S) = P(C) = 0.5 = 50\%$$

Se si osserva un campione di figli di padri e madri di questa tipologia, circa il 50% sara' C e circa il 50% sara' S.

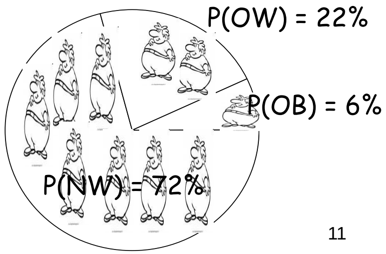
Esempio antropometria – una VARIABILE discreta

X : classe ponderale bambini in eta' scolare
(NW=normopeso, OW=sovrappeso, OB=obeso)

Si osservano 1000 bambini:

X	f(x)	Frequenza relativa
NW	700	0.70 = 70%
OW	230	0.23 = 23%
OB	70	0.07 = 7%

Queste frequenze **stimano le probabilita'** (ignote) relative all'intera popolazione, che potrebbero esser :



Esempio epidemiologia – UNA VARIABILE continua

PAD misurata in 1500 uomini 35 - 44 anni.

Istogramma ed approssimazione Gaussiana che **stimano l'andamento della probabilita'** (ignota) di questo fenomeno

Se l'ipertensione diastolica viene definita come il superamento di 80 mmHg, allora la **stima della probabilita'** (ignota) di avere ipertensione è 0.50=50%.

Simboli che utilizzeremo

- Unione $\cup$ (oppure)
- Intersezione $\cap$ (e)
- Condizionata | (dato che)
- $\bar{A}$ complementare di A

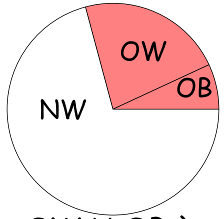
Eventi semplici - Unione

L'osservazione di X (esperimento) da luogo ad un insieme di valori (eventi semplici)

NW OW OB

Gli eventi **composti** sono l'unione di piu' eventi semplici, e.g.

OW U OB che in altre parole è OW oppure OB



L'evento composto OW U OB è osservato quando

OW oppure OB

Eventi semplici - Unione

P(OW U OB) = P(OW) + P(OB) = 0.23+0.07 = 0.30

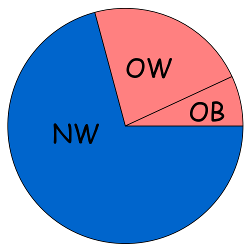
In alternativa, si puo' ragionare come segue:

1) Su 100 persone, i numeri stimati di OW e di OB sono
   100*P(OW) = 100*0.23 = 23 100*P(OB) = 100*0.07 = 7
2) In totale 23+7=30 persone sono OW oppure OB
3) Si ragionava su 100 persone quindi la stima di

   P(OW U OB) è 30/100 = 0.30

Evento complementare di un evento (semplice o composto)

La stessa probabilità, P(OW U OB), posso ottenerla indirettamente ricorrendo a tutti gli eventi che **non** avverano OW o OB -> NW
L'evento normopeso è il **complementare** dell'evento sovrappeso o obeso (OW U OB)



$\text{P(OW U OB)} = 1 - \text{P}(\overline{\text{OW U OB}})$
$= 1 - \text{P(NW)}$
$= 1 - 0.70 = 0.30 = 30\%$

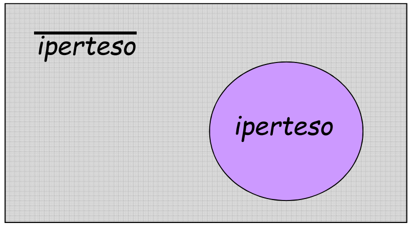
$\to \text{P}(\overline{\text{OW U OB}}) = \text{P(NW)}$

Evento complementare di un evento (semplice o composto)

In altre parole, il complementare di un evento è tutto ciò che non appartiene a quell'evento.

Ad esempio, il complementare dell'evento 'iperteso' è

'$\overline{\text{iperteso}}$' o 'non iperteso'



Esempio antropometria – DUE VARIABILI discrete

X : classe ponderale bambini in eta' scolare

(NW=normopeso, OW=sovrappeso, OB=obeso).

G : sesso del bambino (M=maschio, F=femmina)

Si osserva un campione di 1000 bambini:

	X=Classe Ponderale			
G=sesso	***NW***	***OW***	***OB***	***Totale***
F	*300*	*85*	*15*	*400*
M	*400*	*145*	*55*	*600*
Totale	700	230	70	1000

Probabilita' Condizionata – variabili discrete

La probabilita' che guida il comportamento di X lavora dietro le quinte nella stessa maniera per maschi è femmine?

	NW	OW	OB	
F	300	85	15	400
M	400	145	55	600
Totale	700	230	70	1000

Il fenomeno X si puo' descrivere separatamente per riga in M e F ottenendo le seguenti frequenze relative:

	X=Classe Ponderale			
G=sesso	***NW***	***OW***	***OB***	***Totale***
F	*0.75=300/400*	*0.21*	*0.04*	*1=100%*
M	*0.67=400/600*	*0.24*	*0.09*	*1=100%*
F U M	***0.70=700/1000***	***0.23***	***0.07***	***1=100%***

Probabilita' Condizionata – variabili discrete

	X=Classe Ponderale			
G=sesso	***NW***	***OW***	***OB***	***Totale***
F	*0.75*	*0.21*	*0.04*	*1=100%*
M	*0.67*	*0.24*	*0.09*	*1=100%*
F U M	***0.70***	***0.23***	***0.07***	***1=100%***

Le frequenze 0.75=75%, 0.21=21%, 0.04=4% sono stime della probabilita' di essere nelle classi ponderali **condizionata al fatto che si è F**

$P(NW|F)$ $P(OW|F)$ $P(OB|F)$

dato che ..

L'evento condizionante (F) riduce lo spazio degli eventi possibili al solo mondo femminile -> il denominatore non è più 1000 ma 400!

Probabilita' Condizionata – variabili discrete

G=sesso	X=Classe Ponderale			
	NW	OW	OB	Totale
F	0.75	0.21	0.04	1=100%
M	0.67	0.24	0.09	1=100%
F U M	0.70	0.23	0.07	1=100%

Analogamente, la seconda riga stima le probabilità

P(NW|M) P(OW|M) P(OB|M)

La probabilità che un bambino sia sovrappeso <u>dato che</u> è maschio è del 24%.

Probabilita' Condizionata – variabili discrete

G=sesso	X=Classe Ponderale			
	NW	OW	OB	Totale
F	0.75	0.21	0.04	1=100%
M	0.67	0.24	0.09	1=100%
F U M	0.70	0.23	0.07	1=100%

Siamo portati a sospettare che

P(X = …| F) sia diversa da P(X = …| M)

e quindi anche da P(X =…|F U M)

Le variabili osservate si dice che sono (forse) DIPENDENTI perché il sesso pare che influenzi il comportamento della classe ponderale

Cosa mi aspetterei di osservare nel caso in cui il sesso <u>non influenzi</u> la classe ponderale (X e G indipendenti) ?

G= sesso	X=Classe Ponderale			
	NW	OW	OB	Totale
F	400·0.7	400·0.23	400·0.07	400
M	600·0.7	600·0.23	600·0.07	600
F U M	700	230	70	1000

Tra le femmine mi aspetterei di osservare 280 normopeso, 92 sovrappeso e 28 obese …

NOTA: 280 è pari al prodotto dei totali di riga (400) e colonna (700) diviso per il totale (1000)

-> 280= (400*700)/1000

Probabilita' Condizionata – variabili discrete

Da un punto di vista matematico il fenomeno G=sesso si puo' descrivere separatamente per colonna in NW, OW e OB ottenendo le frequenze relative rappresentate sotto

G=sesso	X=Classe Ponderale			
	NW	OW	OB	NW U OW U OB
F	0.43	0.37	0.21	0.40
M	0.57	0.63	0.79	0.60
F U M	1	1	1	1

Concludendo che siamo portati a sospettare che

P(G = …| NW) sia diversa da P(G = …| OW) …

Tuttavia è poco naturale pensare che la classe ponderale influenzi il comportamento del sesso…

Probabilita' Congiunta - Intersezione

Il fenomeno costituito dalla coppia di variabili G e X si puo' descrivere globalmente ottenendo le frequenze relative su due variabili rappresentate sotto

	X=Classe Ponderale			
G=sesso	***NW***	***OW***	***OB***	***Totale***
F	***0.30=300/1000***	***0.085***	***0.015***	***0.400***
M	***0.400***	***0.145***	***0.055***	***0.600***
Totale	***0.700***	***0.230***	***0.070***	***1=100%***

P(F ∩ NW) che in altre parole è P(F e NW)

Si noti che la tabella è additiva in riga e colonna

0.30 stima la probabilita' di essere femmina e normopeso

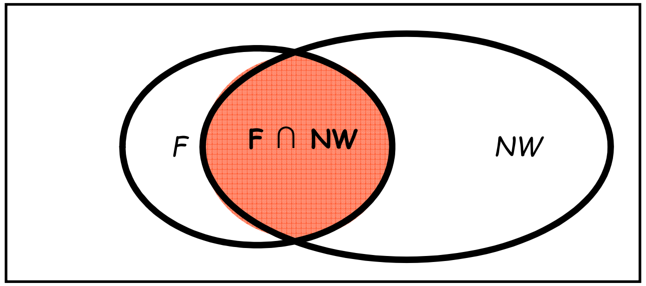
Probabilita' Congiunta - Intersezione

L'evento F ∩ NW

è un **evento composto** che coinvolge due variabili

intersezione di **due eventi semplici**,

uno per la variabile sesso, l'altro per la variabile classe ponderale

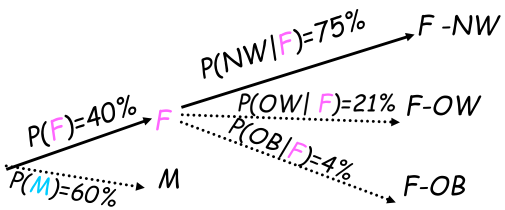


Probabilita' Congiunta - Intersezione

La probabilità congiunta si può ricavare anche tramite la probabilità condizionata, infatti

$$P(NW \mid F) = \frac{300}{400} = \frac{300/1000}{400/1000} = \frac{P(NW \cap F)}{P(F)}$$

P(F ∩ NW) = P(F) * P(NW| F) =
= 0.40 * 0.75 = 0.30



Oppure, analogamente :
P(F ∩ NW) = P(NW) *P(F|NW) = 0.7 * 0.43 = 0.30

Probabilita' Congiunta - Intersezione

In una scuola, le allieve femmine sono il **40%.** Sapendo che il **75%** delle femmine in età scolare è normopeso, che percentuale di femmine normopeso mi aspetto nella scuola?

1) Su 100 persone, il numero stimato di femmine è 100*P(F) = 100*0.40 = 40

2) Su queste 40 femmine, il numero stimato di NW è

40*P(NW|F) = 40*0.75 = 30

3) Si ragionava su 100 persone la stima di

P(F ∩ NW) è 30/100 = 0.30

Probabilita' Congiunta e Indipendenza

E se il sesso e la classe ponderale fossero ***indipendenti*** ?
Cioè, se il sesso non influisse sulla classe ponderale ?

	X=Classe Ponderale			
G=sesso	NW	OW	OB	Totale
F				0.4
M				0.6
F ∪ M	0.70	0.23	0.07	1

Dovrei considerare che la distribuzione della classe ponderale nelle Femmine e nei Maschi sia uguale a quella marginale

$P(NW|F) = P(NW|M) = P(NW)$

$P(F \cap NW) = P(F) * P(NW)$
$= 0.40*0.70 = 0.28 \neq 0.30$

Definizione di Indipendenza

Definizione:

Due eventi/variabili A e B sono dette indipendenti se:

il verificarsi di uno non dice nulla riguardo al verificarsi dell'altro

- $P(NW|F)=P(NW)$
- $P(F|NW)=P(F)$

Proprietà di simmetria

- $P(NW \cap F) = P(F) * P(NW|F) = P(F) * P(NW)$ *oppure*
- $P(NW \cap F) = P(NW) * P(F|NW) = P(NW) * P(F)$

Definizione di Indipendenza

Definizione:

Due eventi/variabili A e B sono dette indipendenti se:

- $P(A|B)=P(A)$
- $P(B|A)=P(B)$

Proprietà di simmetria

- $P(A \cap B) = P(B) * P(A|B) = P(B) * P(A)$ *oppure*
- $P(A \cap B) = P(A) * P(B|A) = P(A) * P(B)$

…In Conclusione…

La probabilità dell'intersezione di due eventi A e B è pari a:

$$\boxed{P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)}$$

Se A e B non sono indipendenti

$$\boxed{P(A \cap B) = P(A)P(B)}$$

Se A e B sono indipendenti

Dalla probabilità congiunta alla probabilità condizionata …

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

$$\Downarrow$$

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

Unione – 2 variabili discrete

P=	X=Classe Ponderale			
ipertensione	NW	OW	OB	Totale
Sì	14	21	10	45
No	686	229	40	955
Totale	700	250	50	1000

Stimare la probabilità di essere iperteso oppure di essere obeso.

P(Sì U OB) che in altre parole è P(Sì o OB)

Si tratta di un 'o' (oppure) non esclusivo.

Ragionamento 1)

Stimare la probabilità di essere iperteso oppure di essere obeso.

P=	X=Classe Ponderale			
ipertensione	NW	OW	OB	Totale
Sì	14	21	10	45
No	686	229	40	955
Totale	700	250	50	1000

Sembrerebbe naturale sommare le stime delle corrispondenti probabilità' (P(Sì)=4.5% + P(OB)=5%), poiché entrambe le condizioni avvalorano il fatto di essere o iperteso o obeso!

Tuttavia questo porterebbe a duplicare la presenza della condizione di Sì e OB che va quindi sottratta:

P(Sì U OB) = P(Sì)+P(OB)-P(Sì∩OB)
= 0.045+0.050-0.010 = 0.085

Ragionamento 2)

Stimare la probabilità di essere iperteso oppure di essere obeso.

P=	X=Classe Ponderale			
ipertensione	NW	OW	OB	Totale
Sì	14	21	10	45
No	686	229	40	955
Totale	700	250	50	1000

Ipertesi
P(Sì)=4.5%

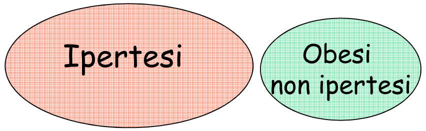
Non ipertesi obesi
P (No ∩ OB) = 4%

Entrambe le condizioni avvalorano il fatto di essere o iperteso o obeso
-> è naturale sommare le stime delle corrispondenti probabilità :
P(Sì U OB)=P(Sì)+P(No∩OB)=0.045+0.040=0.085

Unione – 2 variabili discrete

Quale differenza c'è tra il ragionamento 1) e 2) ?

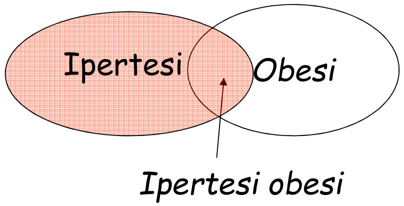
- Il ragionamento 2) considera eventi INCOMPATIBILI (o mutuamente esclusivi)



$P(Sì \cap OB) = \varnothing$

- Il ragionamento 1) no!

-> Gli eventi iperteso e obeso sono COMPATIBILI



$P(Sì \cap OB) \neq \varnothing$

Unione – 2 variabili discrete

Stimare la probabilità di essere iperteso oppure di essere obeso.

Avendo tutte le informazioni in tabella:

P= ipertensione	X=Classe Ponderale			
	NW	OW	OB	Totale
Sì	14	21	10	45
No	686	229	40	955
Totale	700	250	50	1000

$P(Sì \cup OB) = P(Sì \cap NW) + P(Sì \cap OW) + P(Sì \cap OB) + P(No \cap OB) =$

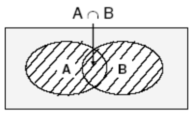
$= 14/1000 + 21/1000 + 10/1000 + 40/1000 =$

$= 85/1000 = 0.085 = 8.5\%$

…In Conclusione…

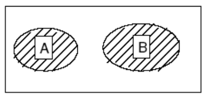
La probabilità dell'unione di due eventi A e B è pari a:

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



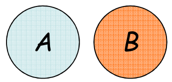
Se A e B sono compatibili

$$\boxed{P(A \cup B) = P(A) + P(B)}$$

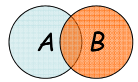


Se A e B sono incompatibili

Riepilogando



Dati due eventi A e B



$A \cap B = \varnothing$
- $P(A \cup B) = P(A) + P(B)$
- $P(A \cap B) = 0$

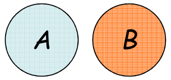
$A \cap B \neq \varnothing$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  dove $P(A \cap B)$

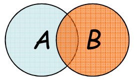
- $P(A \cap B) = P(A) \cdot P(B \mid A)$
  $= P(B) \cdot P(A \mid B)$

Se A e B sono indipendenti
- $P(A \cap B) = P(A) \cdot P(B)$

Riepilogando



Dati due eventi A e B



$A \cap B = \emptyset$

- $P(A \cup B) = P(A) + P(B)$
- $P(A \cap B) = 0$

$A \cap B \neq \emptyset$

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

dove $P(A \cap B)$

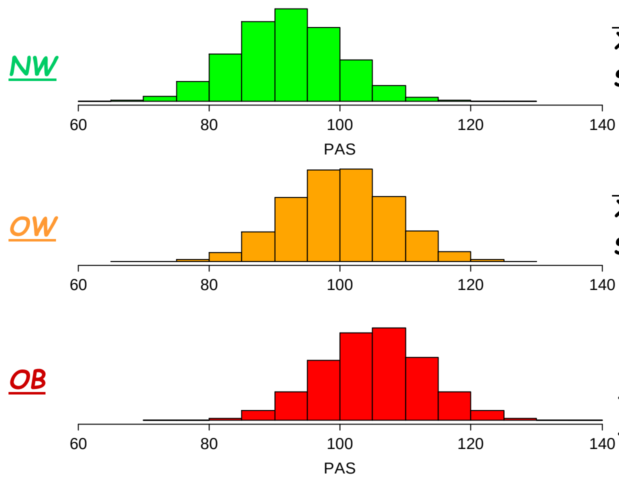
- $P(A \cap B) = P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$

Se A e B sono indipendenti

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

Probabilita' Condizionata – variabili continue

PAS e classe ponderale in bambine/bambini di 4 anni

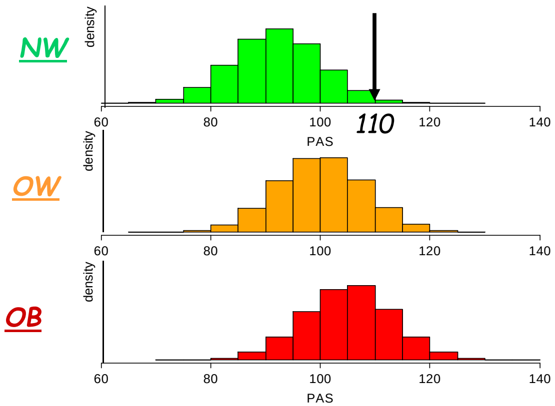


NW: $\overline{x}$=92 mmHg, s =8 mmHg

OW: $\overline{x}$ =100 mmHg, s =8 mmHg

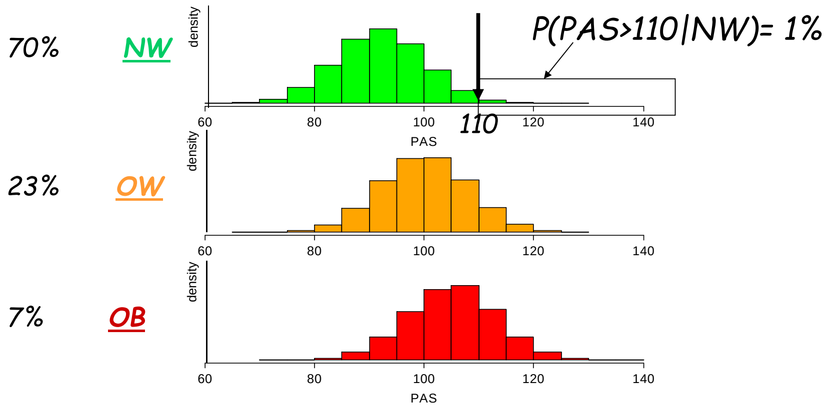
OB: $\overline{x}$ =105 mmHg, s =8 mmHg

Probabilita' Congiunta - Intersezione



Qual è la probabilità che un bambino sia NW e con PAS > 110 mmHg?

Probabilita' Congiunta - Intersezione



Qual è la probabilità che un bambino sia NW e con PAS > 110 mmHg?

P(NW ∩ PAS>110) = P(NW) * P(PAS>110| NW)
= 0.7 * 0.01 = 0.007 = 0.7%

A sample of 35 patients diagnosed with colon carcinoma in Finland during 1985-94, followed up 10 years

Time (yy)	$S(t_{i+1}\|t_i)$
[0-1)	0.77143
[1-2)	0.92308
[2-3)	0.76190
[3-4)	0.85185
[4-5)	1.00000
[5-6)	1.00000
[6-7)	1.00000
[7-8)	1.00000
[8-9)	0.55556
[9-10)	1.00000

Probabilità di sopravvivere al primo anno dopo la diagnosi

P(2|1)=P(2 ∩ 1)/P(1)
Probabilità di giungere vivi al secondo anno | si è sopravvissuti al primo

*P(2 ∩ 1)=P(2|1)*P(1)*

P(2)=P(2 ∩ 1)
Probabilità di sopravvivere fino al secondo anno

A sample of 35 patients diagnosed with colon carcinoma in Finland during 1985-94, followed up 10 years

Time (yy)	$S(t_{i+1}\|t_i)$
[0-1)	0.77143
[1-2)	0.92308
[2-3)	0.76190
[3-4)	0.85185
[4-5)	1.00000
[5-6)	1.00000
[6-7)	1.00000
[7-8)	1.00000
[8-9)	0.55556
[9-10)	1.00000

Qual è la probabilità di essere ancora vivi dopo 5 anni dalla diagnosi?

*P(5)=P(5|4)*P(4)=*

*P(5|4)*P(4|3)*P(3)=*

*P(5|4)*P(4|3)*P(3|2)*P(2|1)*P(1)=*

0.46216

A sample of 35 patients diagnosed with colon carcinoma in Finland during 1985-94, followed up 10 years

Time (yy)	$S(t_{i+1}\|t_i)$
[0-1)	0.77143
[1-2)	0.92308
[2-3)	0.76190
[3-4)	0.85185

Qual è la probabilità di essere ancora vivi dopo 10 anni dalla diagnosi, dato che si è sopravvissuti i primi 5?

P(10|5) = P(10∩5)/P(5) = P(10)/P(5)=

*P(10|9)*P(9|8)*P(8|7)*P(7|6)*P(6/5)*P(5|4)*P(4|3)*P(3|2)*P(2|1)*P(1)/*

*P(5|4)*P(4|3)*P(3|2)*P(2|1)*P(1)=*

*P(10|9)*P(9|8)*P(8|7)*P(7|6)*P(6/5)= 0.55556*