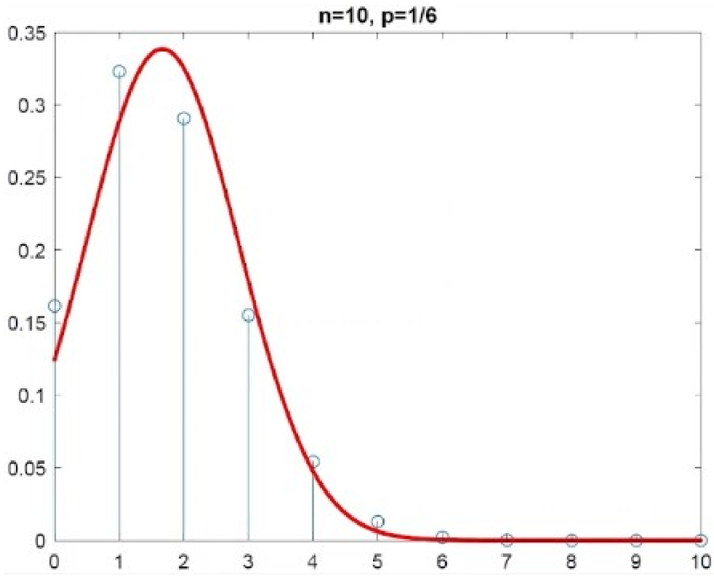
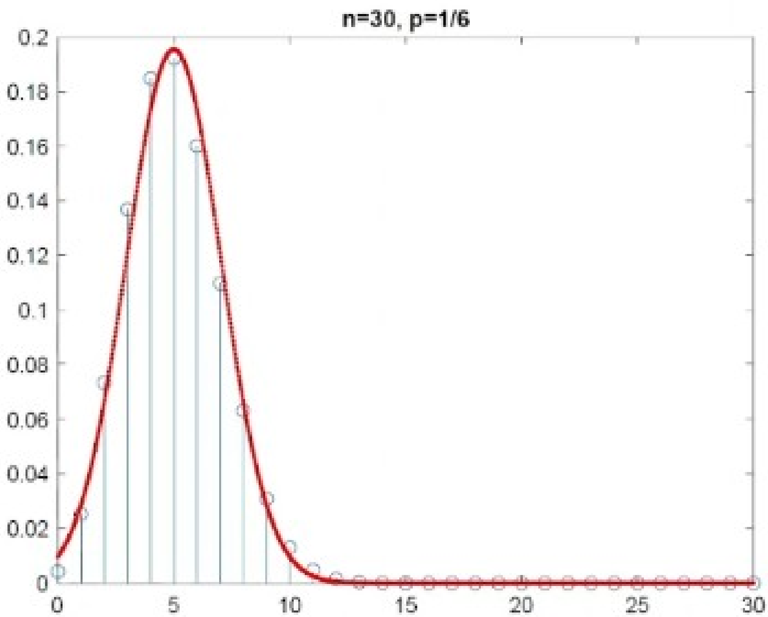
Differenza tra una binomiale e una gaussiana



In blu è rappresentata la binomiale con parametri n e p. In rosso è rappresentata una gaussiana di parametri np e npq.

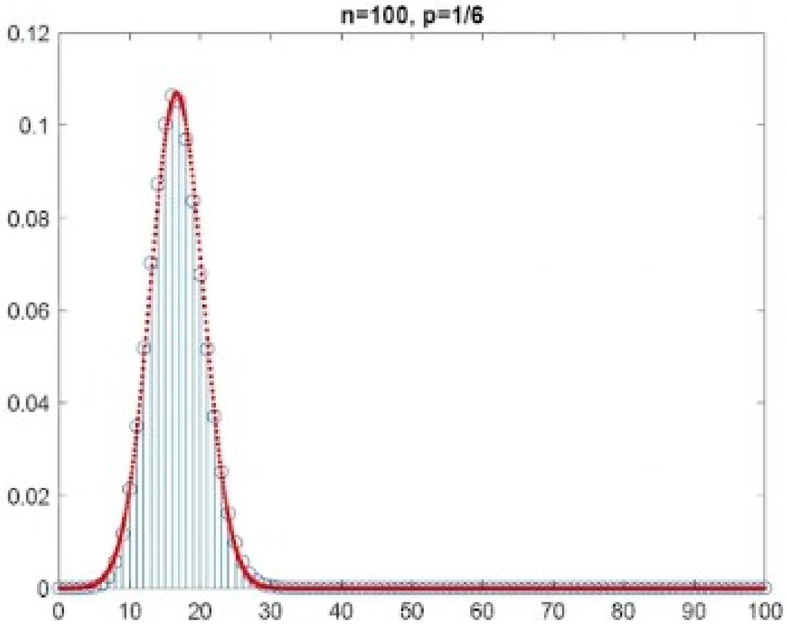
L'approssimazione è un po' grossolana, l'andamento è simile alla gaussiana, ma ci sono valori sottostimati o sovrastimati dalla funzione di densità della gaussiana.

Passiamo a n = 30 e p sempre uguale a 1/6.



I valori sono un po' migliori.

Mandando n a 100, l'approssimazione si fa ancora migliore.



Mettiamo in grafico la funzione di massa di probabilità della binomiale e la confrontiamo con la funzione di densità della gaussiana, troviamo pian piano un risultato sempre migliore.

Ripasso

L'idea è che abbiamo una popolazione accomunata rispetto a una caratteristica dalla stessa funzione di distribuzione di probabilità che descrive la relativa variabile casuale. È una caratteristica misurabile (sì o no e altre). Lo stimatore è una variabile casuale, ogni volta che faccio un esperimento, la stima cambia di valore. Se l'esperimento è fatto in maniera corretta, la variazione di valore non sarà grande. Si sceglie di non utilizzare la stima secca, ma costruire un range di valori in cui confidiamo, abbiamo fiducia, speriamo che si trovi il parametro che vogliamo stimare. Perché non parliamo direttamente di probabilità? Riprendendo i calcoli, capiremo perché si parla di intervallo di confidenza e non probabilità.

Intervallo di confidenza bilaterale per la media mu di una popolazione gaussiana con sigma quadro nota

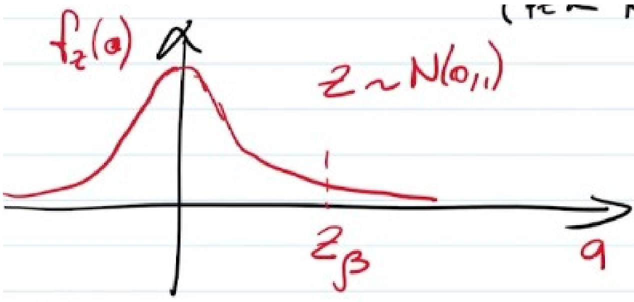
Bilaterale vuol dire che l'intervallo che vogliamo costruire è simmetrico rispetto al valore centrale. Si può costruire un intervallo unilaterale, quando vogliamo stimare che il parametro mu sia maggiore o uguale o minore o uguale di un certo valore, senza preoccuparsi che mu appartenga a un valore simmetrico.

Supponiamo la popolazione gaussiana, indipendentemente dalla numerosità del campione.

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

INDIPENDENTEMENTE DALLA NUMEROSITÀ DEL CAMPIONE (PER RIPRODUCIBILITÀ)

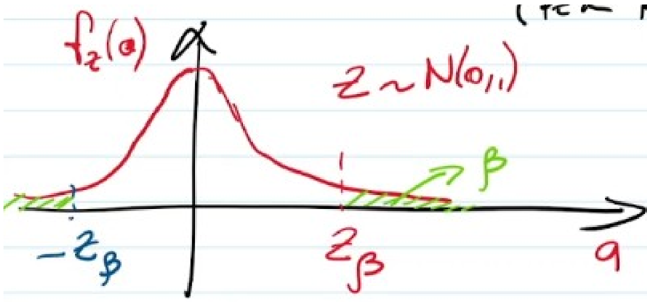
Ricordiamoci il valore critico, introducibile anche per la normale standard.



$$P(Z > z_\beta) = \beta$$

L’area tra la coda della funzione di densità della normale standard e l’asse delle a, da zbeta a infinito, vale beta.

Per simmetria, vale la stessa cosa se consideriamo zbeta nell’alto quadrante verso -infinito.



$$P(Z > z_\beta) = \beta$$
$$P(Z < -z_\beta) = \beta$$

$$P(-z_\beta \leq Z \leq z_\beta) = 1 - \left(P(Z > z_\beta) + P(Z < -z_\beta)\right)$$

Conosciamo tutte le probabilità.

$$= 1 - 2\beta$$

Finora, sono tutti passaggi di probabilità. Stiamo scartando le code della gaussiana che hanno area beta.

Ci ricordiamo che in statistica abbiamo a che fare con una normale standard.

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Posso al posto di Z sostituire la media campionaria – mu su sigma radice di n.

$$P\left(-z_\beta \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_\beta\right) = 1-2\beta$$

Abbiamo introdotto qualcosa di statistico, la media campionaria.

E alla fine otterremo:

$$P\left(\overline{X}-z_\beta \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}+z_\beta \frac{\sigma}{\sqrt{n}}\right) = 1-2\beta$$

Per ragioni di comodità:

$$\begin{array}{l} 1-2\beta \\ \quad \parallel \quad \Rightarrow 2\beta=\alpha \\ 1-\alpha \quad \text{ovvero} \\ \qquad\quad \beta=\frac{\alpha}{2} \end{array}$$

Va fatto anche dall'altra parte.

$$P\left(\overline{X}-z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}+z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = \underbrace{1-\alpha}_{\text{CONFIDENZA}}$$

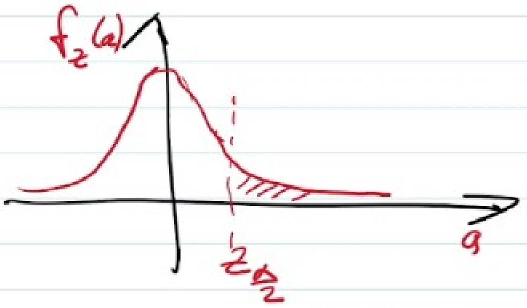
Possiamo parlare di probabilità finché descriviamo il metodo generale di costruire un intervallo, ma questi strumenti li usiamo nella pratica. Con un set di esperimenti, al posto della media campionaria vista come variabile casuale, mettiamo un numero, risultato dalle misurazioni compiute. Z di alfa mezzi è un numero (valore critico tabulato della normale standard), sigma è un numero, la radice quadrata della numerosità del campione è un numero. Quindi l'intervallo diventa un intervallo numerico. Nel momento in cui diventa un intervallo numerico, non è più questione di probabilità, mu non lo conosciamo e anch'esso è un numero, o c'è o non c'è. Messi insieme tutti i passaggi di questo ragionamento, non possiamo più parlare di probabilità, possiamo parlare di fiducia o speranza, di confidenza nel credere che effettivamente il valore da stimare sia dentro l'intervallo definito. È importante distinguere la confidenza dalla probabilità perché il passaggio dalla teoria su come costruire gli intervalli alla pratica io uso questa teoria in un caso pratico, arrivando a calcolare un intervallo numerico per il range di valori che mu potrebbe assumere, con confidenza 1 – alfa.

$$\mu \in \left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] \text{ con confidenza } 1-\alpha$$

Comincia la questione di come scegliere la confidenza.

Da un lato vorremmo un range di valori molto vicino, dall'altro più piccolo prendo l'intervallo di confidenza, più piccola è la confidenza.

$1-\alpha$



N.B. Al diminuire di $z_{\frac{\alpha}{2}}$ diminuisce la ampiezza dell'intervallo (vantaggio), ma decresce anche la confidenza (svantaggio)

Dobbiamo bilanciare una confidenza vicina a 1, con un intervallo sufficientemente piccolo come ampiezza.

Per ridurre l'intervallo lasciando $1-\alpha$ inalterata occorre aumentare n
$\Rightarrow$ per ottenere un buon intervallo di confidenza è necessario lavorare con un campione numeroso.

È un altro modo per vedere la legge dei grandi numeri.

Intervallo di confidenza bilaterale per la media mu di una popolazione gaussiana con sigma quadro incognito

INTERVALLO DI CONFIDENZA BILATERALE PER LA MEDIA μ DI UNA POPOLAZIONE GAUSSIANA CON σ^2 INCOGNITO

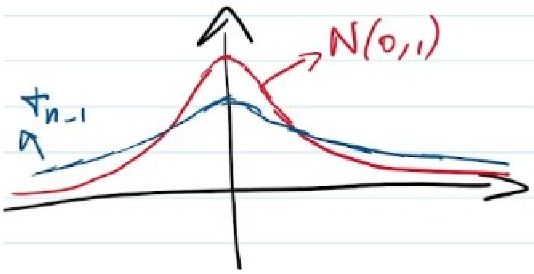
$\bar{X}$ STIMATORE DI μ

S^2 STIMATORE DI σ^2

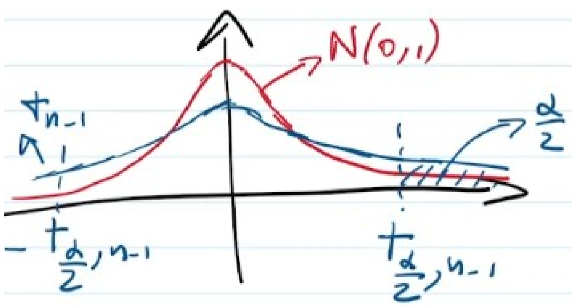
$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Stimiamo sigma quadro con la varianza campionaria. Siccome c'è un errore con la stima, non è più una gaussiana l'espressione scritta, è una T di Student a n – 1 gradi di libertà.

Qual è la differenza tra una gaussiana e la T di Student? Immaginiamo la curva in rosso la gaussiana e in blu la T di Student.



Sono tutte e due simmetriche rispetto all'asse y, ma la T di Student è una curva più schiacciata e larga, rispetto alla gaussiana. Anche per la T di Student ci sono i valori critici, anch'essi simmetrici. Dipendono dall'area a destra del valore critico e il numero di gradi di libertà caratteristici di quel problema. n è sempre la numerosità del campione (se n = 30, la T di Student avrà 29 gradi di libertà).



$$T \sim t_{n-1}$$

$$P\left(-t_{\frac{\alpha}{2},n-1} \le T \le t_{\frac{\alpha}{2},n-1}\right) = 1 - \left(P\left(T > t_{\frac{\alpha}{2},n-1}\right) + P\left(T < -t_{\frac{\alpha}{2},n-1}\right)\right)$$

Stiamo scartando le code della T di Student.

$$= 1 - \left(\frac{\alpha}{2} + \frac{\alpha}{2}\right) = 1 - \alpha$$

E in questa maniera costruiamo l'intervallo di confidenza.

$$\Downarrow$$

$$P\left(-t_{\frac{\alpha}{2},n-1} \le \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \le t_{\frac{\alpha}{2},n-1}\right) = 1 - \alpha$$

Si fanno gli stessi passaggi di prima.

$$\Downarrow$$

$$P\left(\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Il procedimento è lo stesso di prima, cambia solo il fatto che se cerco il valore critico, non lo cerco nella tabella della normale standard, ma nella tabella della T di Student.

Tutto tranne mu diventano dei numeri. E chiaramente valgono gli stessi ragionamenti fatti con sigma quadro nota.

$$\mu \in \left[\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right]$$

con confidenza $(1-\alpha)$

Se la T di Student è più schiacciata e larga, a parità di alfa/2, il valore critico della T di Student sarà più grande del valore critico della normale standard, perché l'area sotto la curva della T di Student in generale è un pochino più grande a parità di valore rispetto a quella della normale standard.

OSS. Dato che in generale le code della normale standard decrescono più velocemente di quelle della t_{n-1}, per $\frac{\alpha}{2}$ fissato

$$z_{\frac{\alpha}{2}} < t_{\frac{\alpha}{2}, n-1}$$

quindi immaginando che $S^2 \simeq \sigma^2$

l'intervallo di confidenza con σ^2 INCOGNITO sarà più ampio dell'intervallo con σ^2 noto

Proprio a causa di questa caratteristica della T di Student rispetto alla normale standard.

Ci possono essere anche differenze legate al fatto che non avendo una numerosità abbastanza ampia, anche la varianza campionaria potrebbe dare dei risultati diversi dalla sigma quadro, sono fluttuazioni statistiche.

Intervallo di confidenza bilaterale per la varianza sigma quadro di una popolazione gaussiana

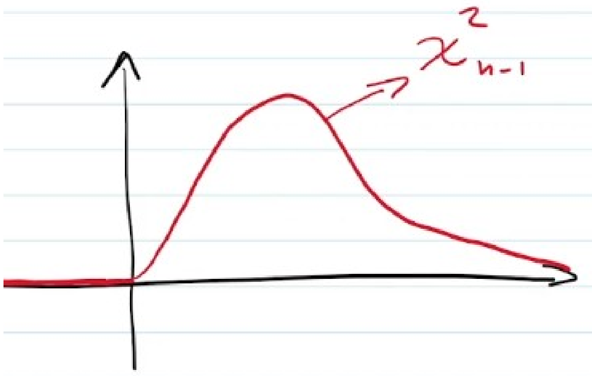
È irrilevante il fatto che conosciamo mu.

S^2 STIMATORE DI σ^2

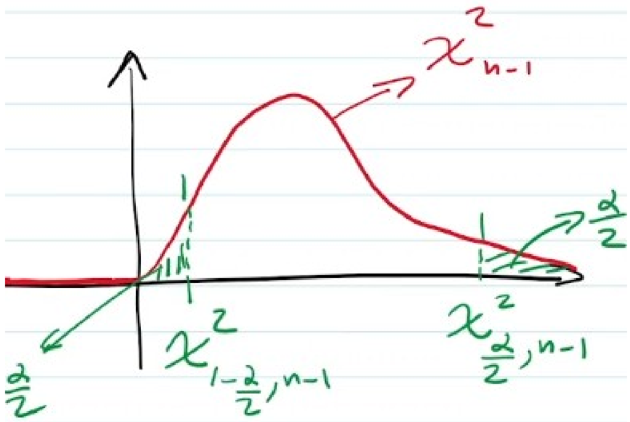
$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Sfrutteremo la funzione di densità di chi-quadro e i valori critici, presenti nelle tabelle.

La funzione di densità della chi-quadro non è simmetrica.



Bisogna cercare dei valori per cui a sinistra del primo e a destra del secondo, l'area sia alfa/2.



$$C \sim \chi^2_{n-1}$$

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-1} \leq C \leq \chi^2_{\frac{\alpha}{2},n-1}\right) = 1 - \left(P\left(C > \chi^2_{\frac{\alpha}{2},n-1}\right) + P\left(C < \chi^2_{1-\frac{\alpha}{2},n-1}\right)\right) = 1 - \left(\frac{\alpha}{2} + \frac{\alpha}{2}\right) = 1-\alpha$$

I passaggi sono sempre gli stessi, dalla probabilità alla confidenza.

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

$$P\left(\frac{1}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) = 1-\alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) = 1-\alpha$$

Da cui si ricava l'intervallo di confidenza. È sempre un range di valori, ma non è simmetrico come prima, cioè il valore centrale è la stima, però ha la caratteristica di un intervallo bilaterale.

$$\sigma^2 \in \left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right]$$

CON CONFIDENZA $(1-\alpha)$

Visione del lavoro assegnato a casa

$$C \sim \chi^2_n$$

$$Var(C)?$$

$$C = \sum_{k=1}^{n} Z_k^2 \quad \text{con } Z_k \sim N(0,1)$$
INDIPENDENTI

$$Var(C) = Var\left(\sum_{k=1}^{n} Z_k^2\right) = \sum_{k=1}^{n} Var(Z_k^2)$$

È possibile grazie all'indipendenza.

$$Var(C) = Var\left(\sum_{k=1}^{n} Z_k^2\right) \underset{\substack{\downarrow \\ \text{INDIPENDENZA}}}{=} \sum_{k=1}^{n} Var(Z_k^2)$$

Zk sono tutte identicamente distribuite, quindi una vale l'altra per la varianza.

$$Var(Y) = E[Y^2] - E^2[Y]$$

$$Var(Z_k^2) = E[Z_k^4] - E^2[Z_k^2]$$

$$E[Z_k^2] = Var(Z_k) + E^2[Z_k] \underset{\substack{\downarrow \\ Z_k \sim N(0,1)}}{=} 1 + 0$$

Per il momento di Zk alla quarta usiamo la funzione generatrice dei momenti.

$$E[Z_k^4] = \frac{d^4 \phi_{Z_k}(t)}{dt^4}\bigg|_{t=0} = \frac{d^4\left(e^{\frac{t^2}{2}}\right)}{dt^4}\bigg|_{t=0}$$

$$= \left(\frac{d^3}{dt} \left(t\, e^{\frac{t^2}{2}} \right) \right) \Big|_{t=0} = \left(\frac{d^2}{dt} \left(e^{\frac{t^2}{2}} + t \left(t e^{\frac{t^2}{2}} \right) \right) \right) \Big|_{t=0} =$$

$$= \left(\frac{d}{dt} \left(t e^{\frac{t^2}{2}} + 2t\, e^{\frac{t^2}{2}} + t^2\, t e^{\frac{t^2}{2}} \right) \right) \Big|_{t=0} =$$

$$= \left(e^{\frac{t^2}{2}} + t\, t e^{\frac{t^2}{2}} + 2 e^{\frac{t^2}{2}} + 2t\, t e^{\frac{t^2}{2}} + 3t^2 e^{\frac{t^2}{2}} + t^3\, t e^{\frac{t^2}{2}} \right) \Big|_{t=0}$$

$$= 3$$

$$E[Z_K^4] = 3 \quad , \quad E[Z_K^2] = 1$$

$$Var(Z_K^2) = E[Z_K^4] - E^2[Z_K^2] = 3 - 1 = 2$$

$$Var(C) = \sum_{K=1}^{n} Var(Z_K) = \sum_{K=1}^{n} 2 = 2n$$

La varianza di una chi-quadro a n gradi di libertà è 2n, la sua media è n.

Esercizi di ripasso

es. 1 Dati gli eventi A e $B \subset S$ con $P(A) = 0.5$, $P(A \cup B) = 0.6$ determinare $P(B)$ nei casi seguenti.

1) A e B sono indip.
2) A e B sono incompatibili
3) $P(A|B) = 0.4$

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

1) $P(A \cap B) = P(A)\,P(B)$ per INDIP.

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

$$P(B)\left(1 - P(A)\right) = P(A \cup B) - P(A)$$

$$P(B)\left(1 - 0.5\right) = 0.6 - 0.5$$

$$P(B) = \frac{0.1}{0.5} = \frac{1}{5}$$

2) A e B INCOMPATIBILI $\Rightarrow A \cap B = \emptyset$

Possiamo fare riferimento agli assiomi di Kolmogorov.

$P(A \cup B) = P(A) + P(B)$ per terzo assioma Kolmogorov

$$P(B) = P(A \cup B) - P(A) = 0.6 - 0.5 = 0.1$$

3) $P(A|B) = 0.4$

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

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

$$\Downarrow$$

$$P(A \cup B) - P(A) = P(B)\left(1 - P(A|B)\right)$$

$$0.6 - 0.5 = P(B)\left(1 - 0.4\right) \Rightarrow P(B) = \frac{0.1}{0.6} = \frac{1}{6}$$

Re Artù convoca 6 cavalieri. Tireranno a turno una moneta equilibrata e partirà per la prossima impresa il cavaliere il cui lancio darà per primo croce (C). Calcolare la probabilità che Lancillotto parta se Lancillotto lancia per primo.

L = "Lancillotto parte"

$$P(L) = P(C_1 \cup T_1 \cap T_2 \cap T_6 C_7 \cup T_1 \cap \dots T_{12} C_{13} \cup \dots)$$

Gli eventi sono disgiunti.

$$P(L) = P(C_1 \cup T_1 \cap T_2 \cap T_6 C_7 \cup T_1 \cap \dots T_{12} \cap C_{13} \cup \dots) =$$

$$= P(C_1) + P(T_1 \cap \dots T_6 \cap C_7) + P(T_1 \cap \dots T_{12} \cap C_{13}) + \dots =$$

$$= \frac{1}{2} + \left(\frac{1}{2}\right)^6 \cdot \frac{1}{2} + \left(\frac{1}{2}\right)^{12} \frac{1}{2} + \left(\frac{1}{2}\right)^{18} \frac{1}{2} + \dots =$$

$$= \sum_{k=0}^{+\infty} \frac{1}{2} \left(\frac{1}{2^6}\right)^k = \frac{1}{2} \sum_{k=0}^{+\infty} \left(\frac{1}{2^6}\right)^k$$

$$= \frac{1}{2} \frac{1}{1-\frac{1}{2^6}} = \frac{1}{2} \frac{1}{1-\frac{1}{64}} = \frac{1}{2} \frac{64}{63} = \frac{32}{63}$$

C’è la serie geometrica, è più complesso, l’importante è saperlo impostare.