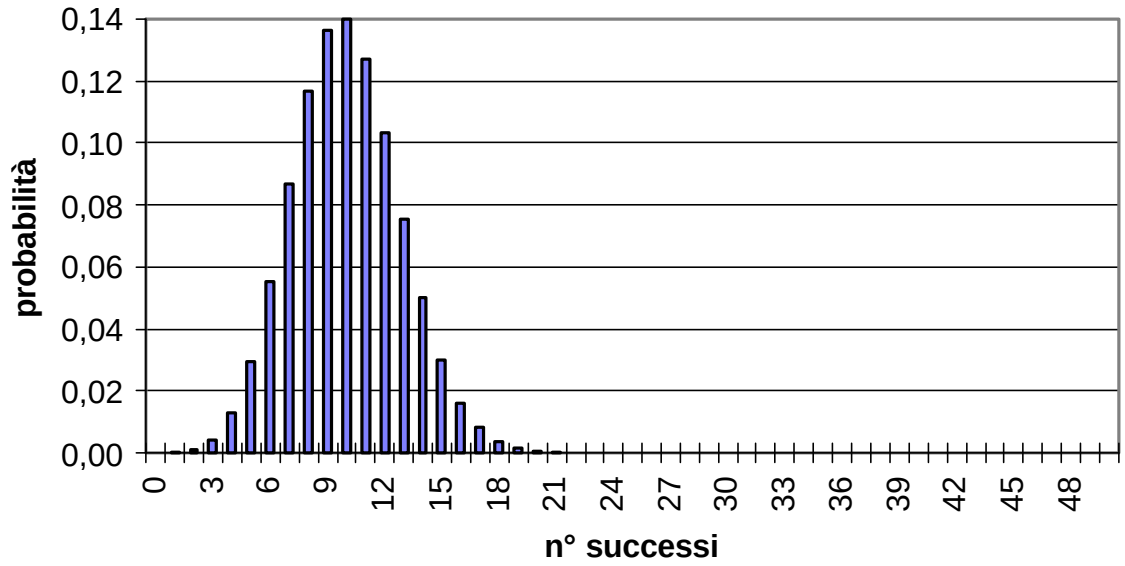
Se si conosce il valore di p.

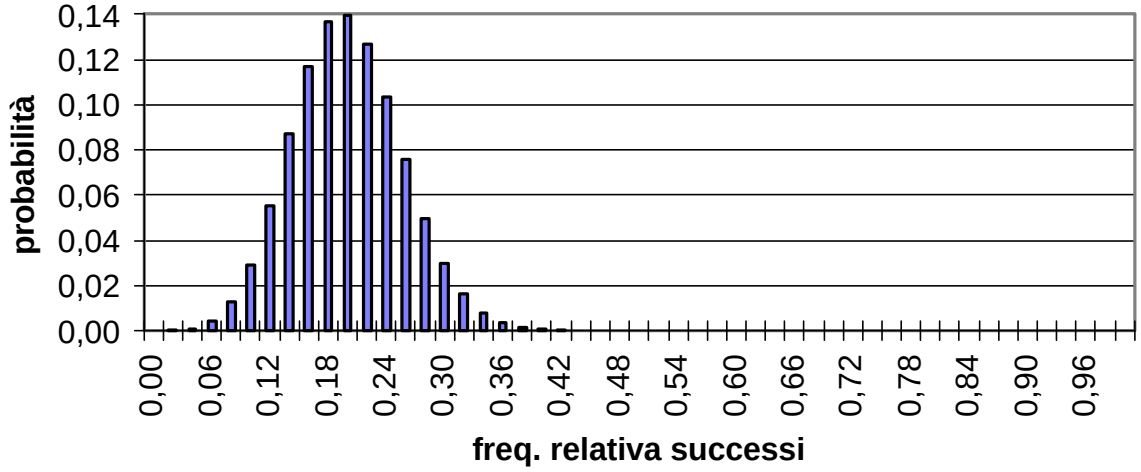
Sia p la probabilità (nota) del verificarsi di un dato evento in una prova ripetibile infinite volte (popolazione infinita) ; q=1-p è la probabilità che l'evento non si verifichi.

Preso un campione di **n** prove la variabile aleatoria "n° di eventi ("successi") verificati" è variabile fra 0 e n, con distribuzione binomiale.

Per esempio con p=0,2 e **n**=50 :



Meglio utilizzare la frequenza relativa di successi (es 30 su 50 = 60%).



Il grafico ci informa che facendo 50 prove, l'evento "capitano 12 successi, pari al 24%" ha la probabilità di poco superiore al 10% (osservare !) ; quanto vale circa la prob che capitino 15 successi su 50 ? e 30 successi ?

Nota.
E' intuitivo paragonare l'istogramma (a gradini) con la curva della Normale ; è dimostrabile che per n (n° prove) tendente a infinito l'istogramma binomiale (n, p) converge a una Normale avente media e varianza della binomiale : np e npq.
In pratica se np>10 et n>50 la differenza fra bin e norm è già piccola.

Il caso dell'esempio è proprio al limite di accettabilità.
Provare a calcolare la probabilità che il n° di successi in 50 prove sia compreso fra 6 e 14 inclusi, prima con la binomiale, poi con la normale (attenzione alla correzione di continuità).
Confrontare i risultati.

Un caso reale.

Vi sono malattie, le cui cause non sono ben conosciute, per cui il fatto che una certa persona ne venga colpita ci appare come un fenomeno casuale, cioè imprevedibile. Di alcune si conosce il n° di nuovi casi verificatisi nel corso di un anno in una data popolazione, per un certo periodo di anni (serie storiche di dati).

Tale conoscenza deriva dal fatto che la malattia è curabile solo in ospedale, dove ogni nuovo caso viene registrato ; il Ministero della Sanità centralizza i dati e su tale base può dire di conoscere la probabilità p che un individuo residente in Italia contragga la malattia nel corso del prossimo anno (definizione frequentista della probabilità).

La probabilità, ricavata dalle serie storiche, che un italiano contragga nel corso di un anno una neoplasia è attualmente pari a 1,35/1000 ; il valore atteso di neoplasie in un anno in tutta la popolazione (circa 56 milioni) è dunque di circa 75.600 nuovi casi ; il valore che registreremo effettivamente sarà intorno a questo.

Essendo np = 75.600 e n = $56*10^6$ possiamo utilizzare al posto della binomiale la distribuzione normale con media μ=np e varianza npq = 75.600*(1-0,00135) = 75.498, da cui σ = 274,77. Al 99% il valore cadrà nell'intervallo : $\mu \pm k_{.995}*\sigma$ con $k_{.995}$ = 2,57 (vedere tavole di N) cioè ci aspettiamo con alto grado di fiducia che i casi siano compresi nell'intervallo (74.894 ; 76.306). La popolazione italiana è stata cosiderata come un campione di ampiezza gigantesca.

Preso invece un campione casuale di 39.450 italiani, il "valore atteso" è di 53,3 casi di insorgenza della malattia, ma il numero effettivamente riscontrato in un anno potrebbe essere inferiore o superiore. Sappiamo che al 99% il valore cadrà nell'intervallo :

$$\mu \pm k_{.995}*\sigma$$

μ = np = 39500*1,35/1000 = 53,3 >10 (si può sostituire la normale alla binomiale);

$\sigma = (npq)^{0.5} = (53,3*998,65/1000)^{0.5} = 7,3$

$k_{.995}$ = 2,57

da cui : $P(34,5< f < 72,1) = 99\%$

Si nota che il primo intervallo è ampio 1412 unità su 56.000.000 (2,5 per mille).
Il secondo è ampio 37,6 unità su 39.450 (9,5 per cento).

Vediamo in modo evidente che l'ampiezza dell'intervallo, a parità di livello di significatività, dipende dal valore di n (ampiezza del campione).

Si considera ora un sottoinsieme particolare della pop italiana, quello dei 39.450 militari che sono stati in Bosnia e Kossovo. Secondo quanto scrive il quotidiano La Stampa il 20 marzo 2001, la Commissione governativa, che indaga sulla questione, ha rilevato 28 casi di insorgenza di neoplasie fra di essi.

Se ne deduce che da un punto di vista statistico non c'è motivo per sospettare che l'essere stati in missione nei Balcani abbia favorito l'insorgere di tumori fra il personale dell'esercito.

Perfino se si trovassero in un anno 70 casi di neoplasia si direbbe che tale *frequenza non si discosta significativamente* dal valore atteso di 53,3 casi ; mentre se si trovassero 75 casi lo scostamento susciterebbe allarme.

Attenzione però : l'ampiezza dell'intervallo dipende dal valore della significatività che si è deciso di adottare. Per rendersene conto basta ripetere il calcolo con le impostazioni :

$$P(\min < f < \max) = 95\ \%$$
$$P(\min < f < \max) = 99{,}9\ \%$$

e determinare min e max.

Dunque non esiste un criterio oggettivo per fissare il valore del livello di significatività .

Fra i vari tumori, una particolare categoria (linfomi Hodkin) ha probabilità di 9,7/100.000 ; il valore atteso in un campione casuale di 39.450 è 3,81 ; casi effettivamente riscontrati dalla Commissione fra i militari italiani : 9. Poichè $np<10$ non si può utilizzare la normale per calcolare gli estremi dell'intervallo.

Usiamo allora la binomiale, eseguendo i calcoli con un foglio elettronico :

p(malattia)=	0,000097	casi malattia	prob.cumulata
1 - p =	0,999903	0	0,021778
n =	39450	1	0,105125
		2	0,264605
		3	0,468040
		4	0,662662
		5	0,811612
		6	0,906605
		7	0,958532
		8	0,983368
		9	0,993927
		10	0,997967
		11	0,999372
		12	0,999820
		13	0,999952
		14	0,999988

Scelto un livello di significatività del 5% già 7 casi di malattia suscitano il dubbio di possibile connessione fra l'essere stato militare nei Balcani e il linfoma di Hodgkin. Con livello del 2% l'allarme comincia con 8 casi ; con livello dell'1% comincia con 8 casi.

Dal quotidiano sopra citato si trascrive un brano di intervista all'ematologo F. Mandelli, membro della Commissione di inchiesta :

Domanda : I casi di tumore maligno sono stati 28. Un numero preoccupante ?

Risposta : Dall'analisi dei dati emerge che vi è un numero di casi **significativamente inferiore** a quello atteso. Ciò significa che è stato fatto un rapporto fra l'incidenza [frequenza relativa] dei tumori nella popolazione italiana e l'incidenza tra i militari italiani impegnati nei Balcani.

Domanda : Ma c'è un eccesso di linfomi di Hodgkin e di leucemia linfatica acuta....

Risposta : **Un eccesso statisticamente non significativo**, che merita comunque di essere valutato attentamente, ...

In base alle risposte del dott. Mandelli, quale valore di significatività è stato scelto dalla Commissione ? Si tratta di una scelta prudente ?

Se non si conosce il valore di p.

Se p è incognito **il problema è diverso dal precedente**, e si chiama **problema inverso** .

Ripetiamo l'impostazione matematica (ammettiamo che sia possibile usare la normale invece della binomiale) :

$$P\left(np - z_a\sqrt{np(1-p)} < f < np + z_a\sqrt{np(1-p)}\right) = 2a \qquad (1)$$

dove : f = frequenza assoluta nel campione (nota dopo aver eseguito le prove)

np = valore atteso (ignoto)

$np(1-p)$ = varianza (ignota)

z_a = valore critico della variabile con distribuzione normale

Nota : scrivere z_a e 2α equivale a $z_{a/2}$ e α

Decidiamo di effettuare 2000 prove, che ci danno 4 successi, e di adottare $2\alpha = 2\%$ (due code di area 1% ciascuna) ; allora l'incognita è solo p e si deve risolvere il sistema delle due disequazioni irrazionali (procedimento facile ma assai noioso, che termina con la determinazione dell'intervallo delle soluzioni, i cui estremi sono esprimibili con una formula).

Invece di esprimerci con la frequenza assoluta f è utile usare la relativa f/n :

$$P\left(p - z_a\sqrt{\frac{p(1-p)}{n}} < \frac{f}{n} < p + z_a\sqrt{\frac{p(1-p)}{n}}\right) = 2a \qquad (2)$$

questo modo di scrivere evidenzia il fatto che **al crescere di *n*** l'intervallo in cui con probabilità 2α cade la **frequenza relativa** di successi in un campione di n **unità diminuisce di ampiezza (la precisione della stima aumenta)**.

→ Per esercizio fare i passaggi di calcolo dalla (1) alla (2).

In definitiva il sistema è risolto da un intervallo di valori centrato su f/n ;

detto $r(n, f, z_a)$ il suo raggio, si avrà :

$$P(\, f/n - r(n, f\,, z_a) < p < f/n + r(n, f\,, z_a\,)) = 2\alpha$$

che esprime la **stima intervallare della probabilità incognita, sulla base del n° di successi riscontrato in un campione casuale**.