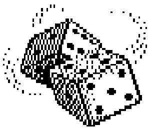
Cenni di calcolo delle probabilità e statistica



Dario Maio

http://bias.csr.unibo.it/maio/

Il concetto di probabilità

- Il termine **probabilità** è usato nel linguaggio quotidiano per denotare casi di non assoluta certezza:
  - probabilità che una macchina venga rubata;
  - probabilità che un televisore si rompa prima dello scadere della garanzia;
  - probabilità che una squadra vinca un incontro di calcio;
  - probabilità che si risolva una crisi politica o un conflitto;
  - probabilità che un titolo azionario aumenti di almeno il 5% entro i prossimi sei mesi;
  - probabilità che un nuovo prodotto riscuota il successo del mercato.
- La categoria del probabile è introdotta nelle argomentazioni quando viene meno la categoria del certo.
- L'origine stessa della parola (dal latino probàbilis, da probàre, provare) sta a indicare "degno di approvazione", "verosimile", "accettabile", "credibile", "ammissibile in base ad argomenti abbastanza sicuri".
- Il termine **probabilità** è quindi usato come **misura del grado di plausibilità di una affermazione**, ovvero del "*verificarsi di un certo evento*".

Alcune definizioni

- EVENTO : rappresenta il risultato di una osservazione o di un esperimento o la descrizione di un potenziale risultato o, in generale, qualsiasi affermazione della quale sia verificabile il contenuto di verità, almeno in linea di principio. Gli eventi vengono rappresentati con le lettere maiuscole A, B, C.
  - INTERSEZIONE di due eventi A, B ($A \cap B$) indica il verificarsi sia di A sia di B
  - UNIONE di due eventi A, B ($A \cup B$) indica il verificarsi dell'evento A o dell'evento B o di entrambi.
  - COMPLEMENTO di un evento A, $\overline{A}$ é l'evento contrario di A.
- Ω (U), SPAZIO DEGLI EVENTI (Universo, Spazio dei campioni) = l'insieme di tutti i risultati possibili.
  - Esempio: lancio di un dado $\Omega = \{1,2,3,4,5,6\}$.
- ***La probabilità è la misura del grado di fiducia che un evento si verifichi.***

Definizioni di Probabilità

- DEFINIZIONE FREQUENTISTICA - La probabilità di un evento è la frequenza relativa con cui l'evento si verifica in una lunga serie di prove ripetute in condizioni simili.
  - **Legge empirica del caso**: in una serie di prove ripetute n volte nelle stesse condizioni, ciascuno degli eventi possibili A si manifesta con una frequenza relativa m/n che si avvicina alla sua probabilità Pr(A).
  - L'approssimazione migliora con il numero delle prove.
- DEFINIZIONE CLASSICA - La probabilità di un evento é il rapporto fra il numero dei casi favorevoli m ed il numero dei casi possibili n, supposto che essi siano tutti equiprobabili.
  - La probabilità che un evento si verifichi assume un valore compreso tra 0 e 1.
  - Se tutti i casi possibili sono anche favorevoli m=n, cioè l'evento A è certo. Se al contrario non esistono casi favorevoli l'evento A é impossibile quindi m=0. La probabilità dell'evento complementare $\overline{A}$ corrisponde a [1-Pr(A)] .
- DEFINIZIONE ASSIOMATICA - La probabilità è una funzione che associa a ogni evento A dello spazio degli eventi U un numero reale in modo che siano soddisfatti i seguenti assiomi:
  - 1) $Pr(A) \geq 0$
  - 2) $Pr(U)=1$
  - 3) Se A e B sono incompatibili (ovvero se $A \cap B = \varnothing$) $Pr(A \cup B) = Pr(A)+Pr(B)$.

Esempi

- La probabilità che esca testa lanciando una moneta "regolare" è pari a 1/2 .
- La probabilità che esca un numero maggiore o uguale a 5 lanciando un dado "regolare" è 2/6=1/3 .
- La probabilità che compaia come risultato 11 in una coppia di dadi è 2/36 (infatti solo le combinazioni (5,6) e (6,5) sono favorevoli).
- La probabilità che esca un numero pari da un'urna contenente 15 palline numerate da 1 a 15 è 7/15 .
- La probabilità di estrarre almeno una pallina rossa su 2 da un'urna contenente 10 palline rosse e 20 bianche è 49/87 :
  - l'evento contrario a "escono 2 palline bianche"

$$1-\frac{\binom{20}{2}}{\binom{30}{2}}=\frac{49}{87}$$

Richiami di Calcolo Combinatorio

CAMPIONAMENTO IN BLOCCO

- **Permutazioni (Fattoriale):** $$n\,!=n\times(n-1)\times(n-2)\times\ldots2\times1 \qquad 0\,!=1$$
- **Disposizioni Semplici (*n* oggetti a gruppi di *r*) (ordinamento):**

$$D_{n,r}=\frac{n!}{(n-r)!}=n\times(n-1)\times\ldots\times(n-r+1)$$

- **Combinazioni Semplici (*n* oggetti a gruppi di *r*):**

$$C_{n,r}=\frac{D_{n,r}}{r!}=\binom{n}{r}=\frac{n!}{r!(n-r)!}=\frac{n(n-1)\ldots(n-r+1)}{r!}$$

CAMPIONAMENTO CON REINSERIMENTO

- **Disposizioni con Ripetizione:** $D^*_{n,r}=n^r$
- **Combinazioni con Ripetizione:** $C^*_{n,r}=\frac{n^r}{r!}$

Schema delle scelte successive e principio del prodotto delle possibilità

- Il numero totale di possibilità per un oggetto formato da diversi elementi per i quali vi siano k scelte è dato dal prodotto delle possibilità corrispondenti ai singoli elementi associati.
- **Esempi**

1) Quanti sono gli anagrammi della parola "MATRICOLE"?
   - Vi sono 9 lettere da disporre in 9 caselle; avremo 9 possibilità per la prima, 8 per la seconda, …, 1 per l'ultima.
   - 9*8*7*6*5*4*3*2*1 = 9! = 362880

2) A una gara partecipano 40 concorrenti. Quante sono le possibili classifiche per i primi 5 posti?
   - 40*39*38*37*36 = $D_{40,5}$ = 78960960

3) Quante diverse colonne si possono giocare al totocalcio?
   - Vi sono 13 caselle da riempire con i simboli 1, 2, X. Per la prima casella 3 possibilità, per la seconda ancora 3 e così fino all'ultima.
   - 3*3*3*…*3 = $D^*_{3,13}$ = 3^{13}

Schema delle scelte simultanee e coefficienti binomiali

- Sia S un insieme di n elementi e sia k un intero compreso fra 0 e n. Quanti sono i sottoinsiemi di S aventi k elementi? In questo caso occorre pensare agli oggetti come scelti "*simultaneamente*". Il problema si risolve con i coefficienti binomiali.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \binom{n}{1} = n, \qquad \binom{n}{0} = \binom{n}{n} = 1$$

- **Esempio**
  - Quante cinquine si possono estrarre da un'urna contenente i 90 numeri del lotto?

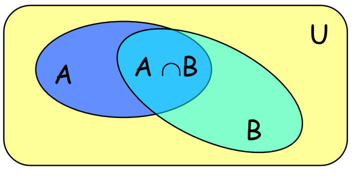
$$C_{(90,5)} = \binom{90}{5} = 43949268$$

Esempi

- La probabilità di trovare un record nell'i-esimo blocco di un file di NP blocchi(supponendo una distribuzione uniforme dei record nei blocchi) è 1/NP, la probabilità che quel blocco non ospiti il record cercato è (1-1/NP).
- La probabilità di non trovare nessuno di ER record nell'i-esimo blocco di un file di NP blocchi (in cui si suppone anche capacità infinità) è $(1-1/NP)^{ER}$. La probabilità che quel blocco ospiti almeno uno degli ER record è $1-(1-1/NP)^{ER}$.
- Dato un juke-box di N=100 dischi ottici e D=4 drive, la probabilità che una richiesta faccia riferimento a un disco già montato su drive (nel caso di distribuzione uniforme dei dati e delle richieste) è 4/100.
- Quanti sono gli anagrammi di ANAGRAMMA?
  - Scelgo le 4 caselle in cui sistemare le A: $\binom{9}{4} = 126$
  - Tra le 5 rimaste scelgo le 2 caselle per le M: $\binom{5}{2} = 10$
  - Nelle ultime 3 caselle pongo le lettere rimaste: 3! = 6
  - Complessivamente vi sono 126*10*6 = 7560 anagrammi possibili.

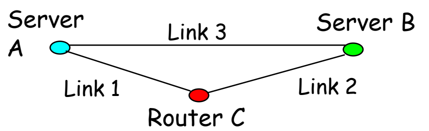
Probabilità della somma di eventi compatibili

- **Relazione di Boole**
  - $Pr(A \cup B) = Pr(A)+Pr(B) - Pr(A \cap B)$



- **Esempio**
  - Durante la fase di test di un sistema informativo si è rilevato che su 1000 operazioni effettuate, 50 hanno causato una terminazione anticipata della sessione di lavoro, 30 hanno determinato inconsistenza dei dati e fra queste 12 presentavano entrambi i difetti. Qual è il tasso di errore del sistema (=probabilità di riscontrare almeno uno dei difetti)?
  - Gli eventi A (terminazione anticipata della sessione) e B (inconsistenza dei dati) sono compatibili.
  - $Pr(A \cup B) = Pr(A)+Pr(B) - Pr(A \cap B) = 50/1000+30/1000-12/1000 = 6{,}8\%$.

Esercizio 1



- Si vuole trasferire un file dal server A al server B; i possibili cammini sono:
  Link3: collegamento diretto oppure Link1 fino al router C e Link2 da C a B

  Calcolare la probabilità che non si riesca a trasferire il file, essendo p la probabilità che un link sia interrotto (assumendo guasti indipendenti).

- Soluzione: l'evento congiunto

  (Link3 interrotto) $\cap$ ((Link1 interrotto) $\cup$ (Link2 interrotto))

  ha probabilità di verificarsi pari a:

$$p \times \left(p + p - p^2\right) = p \times \left(2p - p^2\right) = p^2 \times (2 - p)$$

Probabilità condizionata

- **Probabilità condizionata**
  - Rappresenta la probabilità di un evento, valutata sapendo che si è verificato un altro evento.
  - Es. : Probabilità che la somma dei lanci di 2 dadi dia 5 sapendo che il risultato del lancio del primo dado è 6: Pr(E)=0 .
  - Pr(B|A)= Pr(A $\cap$ B) / Pr(A)

- **Esempio**
  - Si trovi la probabilità che lanciando un dado una sola volta si verifichi un numero minore di 4 nei seguenti due casi: (a) non si dispone di altre informazioni, (b) è noto che il risultato è un numero pari.
  - (a) Sia *B* l'evento {numero minore di 4}: Pr(B)=1/2 .
  - (b) Sia *A* l'evento {numero pari}, risulta quindi che Pr(A) = 1/2. Inoltre per l'evento congiunto (pari, minore di 4) si calcola Pr(A $\cap$ B)=1/6 (infatti solo il risultato 2 è un numero pari minore di 4). Da cui si ottiene: Pr(B|A)= Pr(A $\cap$ B) / Pr(A) = 1/3 .
  - Quindi l'informazione (b) aggiunta ad (a) restringe la probabilità da 1/2 a 1/3: esiste un solo numero pari inferiore a 4 su tre numeri pari compresi tra 1 e 6.

Teorema di Bayes

- **Teorema di Bayes (o delle probabilità composte)**
  - La probabilità del prodotto di due eventi è uguale al prodotto della probabilità di uno degli eventi per la probabilità condizionata dell'altro calcolata a condizione che il primo abbia luogo.
  - $Pr(B|A) = Pr(B)\ Pr(A|B)/Pr(A)$
- **Eventi indipendenti**
  - Se A e B sono eventi indipendenti si ha: $P(B|A) = P(B)$ e $P(A|B) = P(A)$.
  - La probabilità del prodotto di due eventi indipendenti è uguale al prodotto delle probabilità di questi eventi: $P(A \cap B) = P(A)\ P(B)$.
- **Nel caso generale di N eventi**
  - Sia A un evento di probabilità non nulla e B_j una partizione di Ω ovvero una famiglia di eventi a 2 a 2 indipendenti e tali che $\bigcup_{j=1}^{N} B_j = \Omega$

$$Pr(B_k | A) = \frac{Pr(A | B_k)Pr(B_k)}{\sum_{j=1}^{N} Pr(A | B_j)Pr(B_j)}$$

Esercizio 2

- **Test di collaudo di un processo produttivo**
  - Un sistema automatico per il controllo di qualità di hard disk garantisce che se un pezzo è difettoso viene eliminato con probabilità 0.995; anche un pezzo non difettoso può essere eliminato, con probabilità 0.001. La probabilità che un pezzo sia difettoso è 0.2. Qual è la probabilità che un pezzo non eliminato sia difettoso?

Soluzione

- Sia E="il pezzo è eliminato" ; D= "il pezzo è difettoso"; $Pr(D|\overline{E})$?
- $Pr(E|D)=0.995$ $Pr(E|\overline{D})=0.001$ $Pr(D)=0.2$
- $Pr(D|\overline{E})= Pr(D)\ Pr(\overline{E}|D)\ /\ (Pr(D)\ Pr(\overline{E}|D) + Pr(\overline{D})\ Pr(\overline{E}|\overline{D})\ )$
- Osserviamo che: $Pr(\overline{E}|D) = 1- Pr(E|D)=0.005$
  $Pr(\overline{E}|\overline{D}) = 1- Pr(E|\overline{D})=0.999$
  $Pr(\overline{D}) = 1- Pr(D)=0.8$
- $Pr(D|\overline{E})= 0.2*0.005/\ (0.2*0.005 + 0.8*0.999\ ) \cong 0.125\%$

Variabile statistica

- Indipendentemente dal tipo di rilevazione, ciascuna informazione che verrà rilevata sarà chiamata **Variabile**.
  - A differenza delle variabili in Fisica che indicano sempre qualcosa di misurabile, in Statistica si possono rilevare anche dati non misurabili (Sesso, Stato Civile, Livello di istruzione, Area geografica,…)
  - Vengono definite **modalità** di una variabile statistica le sue "categorie".
    - » Es. : le modalità della variabile Sesso sono 'Maschio/ Femmina', le modalità della variabile Stato civile sono 'celibe/coniugato/…'
- Le variabili possono essere divise in due grandi gruppi:
  - **Qualitative**: contengono modalità 'nominali' (es: Stato Civile) o 'ordinali' (es: Livello di Istruzione).

  Per una più efficace gestione dei dati ciascuno di questi 'valori' può essere codificato, cioè a ciascuna modalità è attribuito un numero che non ha però significato di numero in termini aritmetici, ma di semplice codice.
    - » Es. : se codifichiamo la variabile Sesso con 1=Maschio e 2=Femmina, 1 e 2 non hanno significato aritmetico, cioè la relazione 1 < 2 non avrà in questo caso alcun senso.
  - **Quantitative**: contengono valori numerici interpretabili come tali (es: Età, Numero di figli).

Variabili quantitative

- Le variabili quantitative si distinguono in:
  - **Variabili discrete**: possono assumere solo alcuni dei valori compresi in un intervallo.
    - » Numero di sigarette fumate al giorno, Numero di figli sono variabili discrete (per queste hanno senso solo dei valori interi).
  - **Variabili continue**: possono assumere tutti i valori contenuti in un determinato intervallo.
    - » Es. : Altezza, Peso, Tempo (una volta definito l'intervallo, tutti i valori compresi tra il valore minimo e il valore massimo saranno valori 'assumibili').
- Trasformazione di variabili quantitative
  - A volte alcune variabili quantitative possono fornire un'informazione interessante anche tramite una descrizione in termini qualitativi. Si parla in questo caso di **Trasformazione di Variabili da quantitative a qualitative**, e consiste nel raggruppare in classi i valori delle variabili quantitative.
    - » Es. : La variabile Altezza contiene valori numerici come 180, 167, 173, ….. è possibile raggrupparla in classi: ($h \leq 160$, $160<h \leq 180$, $180<h \leq 200$, $h>200$) .

Variabili aleatorie discrete

- Lancio ripetuto di una moneta
  - Indicando con t il risultato "testa" e con c il risultato "croce", per ciascun lancio, nel caso del doppio lancio di una moneta lo spazio dei campioni risulta $\Omega = \{tt, tc, ct, cc\}$.
  - Si supponga ora di introdurre la variabile X per rappresentare il numero delle volte in cui si verifica l'evento "croce". A ogni punto campione è possibile assegnare un numero come valore di X:

Punto Campione	*tt*	*tc*	*ct*	*cc*
Variabile Casuale X	0	1	1	2

- **Variabile Aleatoria Discreta**
  - è una variabile che può assumere un insieme di valori **finito** o **numerabile**, in dipendenza del verificarsi di eventi aleatori che costituiscono una partizione di un universo fissato;
  - è una funzione dello spazio Ω degli eventi che associa a ogni evento, appartenente a una partizione di Ω, un numero reale.

$$X : \Omega \to \Re$$

Distribuzione di probabilità per una variabile discreta

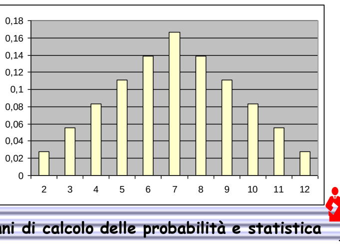
- **Legge (distribuzione) di una v.a.**
  - Si definisce distribuzione di probabilità della v.a. X l'applicazione che associa a ogni intervallo I di $\Re$ la probabilità $\Pr(X \in I)$.
- **Densità (discreta) di una v.a.**
  - Sia X una v.a. e siano $\{x_k\}_{k=1..inf}$ i valori che essa può assumere. Si dice **densità di X** la funzione p_X che a ogni valore assunto da X associa la probabilità che X assuma quel valore: $p_X(x_k) = \Pr(X = x_k)$.

$$\sum_{k=1}^{\infty} p_X(x_k) = 1$$

Densità discreta della v.a. "numero di croci al lancio di 2 monete"

Valori	0	1	2
Probabilità	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Eventi	*tt*	*tc, ct*	*cc*

Densità discreta della v.a. "somma dei punti di due dadi"

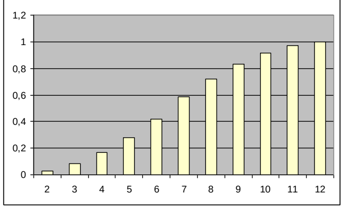


Funzione di ripartizione

- **Funzione di ripartizione o Distribuzione cumulativa**
  - Esprime, per ogni valore x_i la probabilità che la variabile aleatoria assuma un valore non superiore a x_i: $F(x_i)=\Pr(X \le x_i)$.
  - Il valore di $F(x_i)$ si ottiene sommando le probabilità $p_X(x_k)$, per ogni $k<=i$.

$$F(x_i) = \sum_{k=1}^{i} p_X(x_k)$$

Funzione di ripartizione della v.a. "somma dei punti di due dadi"



Funzione di ripartizione della v.a. "numero di croci al lancio di 2 monete"

Valori	0	1	2
Probabilità	$\frac{1}{4}$	$\frac{3}{4}$	1
Eventi	*tt*	*tc, ct, tt*	*tc, ct, tt, cc*

Esercizio 3

- **File system**
  - Un file system è strutturato in modo che la richiesta di un blocco possa essere soddisfatta indifferentemente su uno di due dischi mantenuti costantemente allineati. **In corrispondenza della richiesta di accesso a un blocco viene scelto il disco per il quale il tempo di posizionamento è più breve**; cioè l'indirizzo della traccia richiesta viene comparato con l'indirizzo relativo all'ultima richiesta in coda al primo e al secondo disco, e viene scelto il disco per il quale risulta minimo lo spostamento del braccio.
  - Si assumano indipendenti le richieste di accesso e uniforme la distribuzione degli accessi sulle tracce; è inoltre nota la distribuzione cumulativa di probabilità $F(t_{seek})$ del tempo di seek per ognuna delle due code di servizio:

$\Pr(t_{seek} \le 0 \text{ msec}) = 0.004$ $\quad\Pr(t_{seek} \le 3 \text{ msec}) = 0.036$

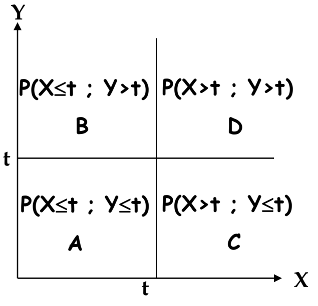
$\Pr(t_{seek} \le 7 \text{ msec}) = 0.2$ $\quad\Pr(t_{seek} \le 13 \text{ msec}) = 1$

$t_{savg} = 11.644$ msec

  - Si determini il **valor medio del tempo di seek** con riferimento al complesso dei due dischi.

Soluzione esercizio 3

- Si considerano due variabili aleatorie X e Y che rappresentano rispettivamente il tempo di seek per il primo e per il secondo disco.
- La variabile aleatoria che interessa è $Z = \min(X,Y)$.



L'evento $Z \le t$ si verifica ogni volta che è vera l'espressione:

$(X \le t$ and $Y \le t)$ or $(X \le t$ and $Y > t)$ or $(X > t$ and $Y \le t)$

quindi si può scrivere che:

$\Pr(Z \le t) = \Pr(A) + \Pr(B) + \Pr(C) = 1 - P(D)$

Dall'ipotesi di indipendenza si ha:

$\Pr(D) = \Pr(X>t ; Y>t) = \Pr(X>t)*\Pr(Y>t) =$
$=(1-\Pr(X \le t))*(1-\Pr(Y \le t)) = (1-F(t)) * (1-F(t))=$
$=[1-F(t)]^2$

da cui

$\Pr(Z \le t) = 1 - [1-F(t)]^2$

- La distribuzione cumulativa relativa al complesso dei due dischi diventa:

$\Pr(t_{seek2} \le 0 \text{ msec}) = 0.008$ $\quad \Pr(t_{seek2} \le 3 \text{ msec}) = 0.071$

$\Pr(t_{seek2} \le 7 \text{ msec}) = 0.360$ $\quad \Pr(t_{seek2} \le 13 \text{ msec}) = 1$

$t_{savg2} = (1-0.360)*13 + (0.360-0.071)*7 + (0.071-0.008)*3 = 10.532$ msec

Momenti delle v.a. discrete

- **Valore atteso** o Valor medio o Media o Speranza Matematica
  - rappresenta il valore previsto che si potrà ottenere in un gran numero di prove

$$E(X) = \sum_k x_k \cdot p_X(x_k) = \mu$$

- **Proprietà**

$$E(a \cdot X + b) = a \cdot E(X) + b$$
$$E(X + Y) = E(X) + E(Y)$$
$$E(X \cdot Y) = E(X) \cdot E(Y)$$

se X e Y sono indipendenti

- **Varianza e scarto quadratico medio**
  - Scarto: differenza tra un valore x_i della v.a. e la media μ .
  - **Varianza** $$Var(X) = E\left[(X-\mu)^2\right] = \sum_k (x_k - \mu)^2 \cdot p_X(x_k)$$
  - La varianza rappresenta la misura del rischio di una distribuzione, la dispersione attorno alla media.
  - **Scarto quadratico medio** $$\sigma(X) = \sqrt{Var(X)}$$

Momenti centrali delle v.a. discrete

- **Proprietà della varianza**

$$Var(X) = E(X^2) - E(X)^2$$

$$Var(a \cdot X + b) = a^2 \cdot Var(X)$$

$$Var(X + Y) = Var(X) + Var(Y)$$ se X e Y sono indipendenti

- **Proprietà del valor medio**
  - Il valor medio è quel valore che rende minima la media dei quadrati degli scarti: $E\left[(X-k)^2\right]$ è minimo se $k = \mu$.

- **Momenti centrali di ordine n**

$$M_n(X) = E\left[(X-\mu)^n\right] = \sum_k (x_k - \mu)^n \cdot p_X(x_k)$$

Esercizio 4

- **La regola 80-20**
  - Si consideri un file ad accesso sequenziale; sia $K=\{k_1, k_2, \ldots k_{5n}\}$ l'insieme dei $5 \cdot n$ valori di chiave dei record memorizzati nel file. Un blocco del file ospita per ipotesi un solo record. L'80% delle richieste di accesso interessa il 20% dei record, che sono allocati casualmente nei primi n blocchi, ciascuno di essi con probabilità:

$$\Pr(k_i) = \frac{0.8}{n} \quad \text{per } i = 1, \ldots, n$$

    Il restante 20% delle richieste di accesso interessa l'80% dei record allocati blocchi da $n+1$ a $5 \cdot n$, con probabilità:

$$\Pr(k_i) = \frac{0.2}{4 \cdot n} \quad \text{per } i = n+1, \ldots, 5 \cdot n$$

  - Assumendo che il costo di accesso al record memorizzato nel blocco i del file sia $C(i)=i$, si determini il valor medio C_m del costo di accesso a un record.

Soluzione esercizio 4

- Indicando con B_i il valore della chiave del record in posizione i nell'allocazione che rispetta la cosiddetta "legge 80-20", il valor medio del costo di accesso a un record è:

$$C_m = \sum_{i=1}^{5\cdot n} \left(C_i(i)\cdot p(B_i)\right) = \sum_{i=1}^{5\cdot n} \left(i \cdot p(B_i)\right)$$

- che può essere riscritta come:

$$C_m = \sum_{i=1}^{n} \left(i \cdot \frac{0.8}{n}\right) + \sum_{i=n+1}^{5n} \left(i \cdot \frac{0.2}{4\cdot n}\right) = n + 0.5$$

Distribuzione uniforme

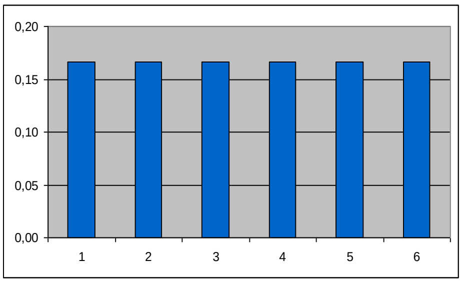
- **è la distribuzione di eventi equiprobabili**

$$p_X(x_k) = \frac{1}{n} \qquad k = 1 \ldots n$$

- **Valore atteso**

$$E(X) = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Densità discreta della v.a.
"risultato del lancio di un dado"



Distribuzione Binomiale

- **Esperimento Bernoulliano**
  - esperimento aleatorio che può avere solo 2 esiti: **successo** o **insuccesso**, rispettivamente con probabilità p e ($1-p$).
  - Es.: il lancio di 1 moneta.
- **Processo di Bernoulli**
  - sequenza di esperimenti di uguale parametro p, tra loro indipendenti.
  - Es.: il lancio **ripetuto** di 1 moneta.
- **Binomiale di parametri n e p**
  - variabile aleatoria che conta il numero di successi ottenuti su n prove
  - densità $X \sim B(n,p)$

$$p_X(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \qquad k = 0 \ldots n$$

- p è la probabilità di 1 successo
- p^k è la probabilità di k successi
- $(1-p)^{n-k}$ è la probabilità di $n-k$ insuccessi
- $\binom{n}{k}$ sono i possibili modi di avere k successi su n prove

Distribuzione Binomiale

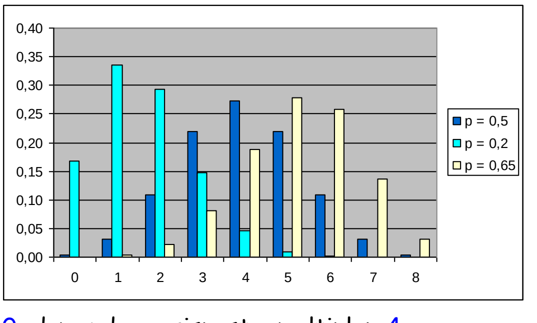
- **Valore atteso**

$$E(X) = np$$

- **Varianza**

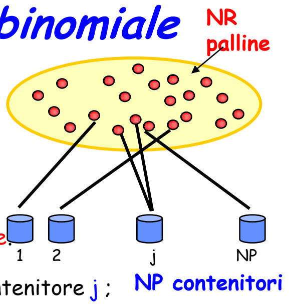
$$Var(X) = np(1-p)$$

Densità discreta della v.a. Binomiale $X \sim B(8,p)$



- **Esempio**
  - Un test consiste di 10 domande a risposta multipla: 4 risposte possibili, di cui 1 sola esatta. Per superare il test occorre rispondere correttamente ad almeno 8 domande. Qual è la probabilità di superare il test rispondendo a caso?
  - Il test è un processo di Bernoulli di legge $X \sim B(10, 0.25)$
  - $\Pr(X \geq 8) = p_X(8) + p_X(9) + p_X(10) \cong 0.04\%$

Esempio distribuzione binomiale



- **Distribuzione di palline in contenitori**
  - Si vogliono distribuire NR palline in NP contenitori (**di capacità infinita**).
  - Supponendo equiprobabile l'inserimento di una pallina in ciascun contenitore, determinare la legge della v.a. X che regola il numero di palline assegnate a un contenitore

$p = 1/NP$ probabilità che una pallina i sia assegnata a un contenitore j ;

$\left(\frac{1}{NP}\right)^{k}$ probabilità che k palline siano assegnate tutte al contenitore j ;

$\left(1-\frac{1}{NP}\right)^{NR-k}$ probabilità che le restanti NR-k palline non siano assegnate a j .

- considerando tutte le possibili combinazioni delle palline si nota che $X \sim B(NR, 1/NP)$

$$\Pr\{X = k\} = p_X(k) = \binom{NR}{k}\left(\frac{1}{NP}\right)^{k}\left(1-\frac{1}{NP}\right)^{NR-k}$$

$$E(X) = \frac{NR}{NP}$$

Caso ideale della distribuzione di record nei blocchi di un file (NR record in NP blocchi)

Distribuzione Geometrica

- **Distribuzione geometrica di parametro p**
  - in un processo di Bernoulli illimitato di parametro p, la v.a. che conta il numero di prove necessarie per ottenere il primo successo si chiama v.a. geometrica di parametro p, $X \sim G(p)$

$$p_X(k) = p\cdot(1-p)^{k-1} \qquad k = 1,2,\ldots$$

  » $(q)^{k-1}$ è la probabilità di k-1 insuccessi ($q = 1-p$)

  » p è la probabilità di un successo (dopo i k-1 insuccessi)

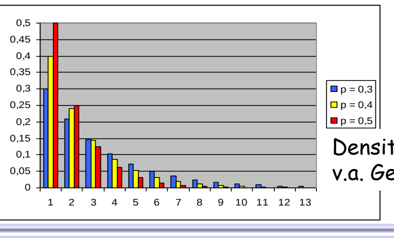
x_i	1	2	3	...	k	...
p_i	p	qp	q^2p		$q^{k-1}p$	

- **Valore atteso**

$$E(X) = \frac{1}{p}$$

- **Varianza**

$$Var(X) = \frac{q}{p^2}$$



Densità discreta della v.a. Geometrica $X \sim G(p)$

Esempio distribuzione geometrica

- Si lancia un dado più volte e si conta quante volte esce il 6.
  - Qual è la probabilità che per ottenere 6 occorra lanciarlo esattamente 6 volte?
  - Sia X il numero di lanci necessari per ottenere il primo 6.
  - **$X \sim G(1/6)$**

$$\Pr(X = 6) = p(1-p)^{6-1} = \tfrac{1}{6} \cdot \left(\tfrac{5}{6}\right)^{5} \cong 0.067$$

  - Qual è la probabilità di ottenere 6 al 6° lancio?

$1/6 \cong 0.167$

  - Qual è la probabilità che per ottenere 6 occorra lanciarlo più di 6 volte?

$$\Pr(X > 6) = \sum_{k=7}^{\infty} p(1-p)^{k-1} = 1 - \sum_{k=1}^{6} \tfrac{1}{6} \cdot \left(\tfrac{5}{6}\right)^{k-1} \cong 0.335$$

- **Esercizio**
  - Si lancia una moneta fino a quando non esce "testa". Supponendo Pr(T)= Pr(C) = $\frac{1}{2}$ calcolare la probabilità che esca per la prima volta testa dopo un numero dispari di lanci.

Campionamento senza reimmissione

- Un giocatore di poker riceve 5 carte da un mazzo di 52:
  - qual è la probabilità di avere esattamente k assi (con k=1,2,3,4)?

$$p(k) = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}$$

2 gruppi di carte: 4 assi, 48 restanti. Estrazione di 5 carte senza rimpiazzo

  - qual è la probabilità di avere almeno 2 assi?

$A_2 \cup A_3 \cup A_4$ $\quad p(2)+p(3)+p(4)=0.042$

  - qual è la probabilità di ricevere un colore servito?

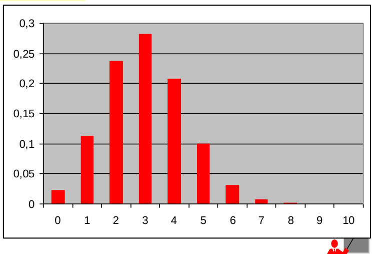
$$p(colore) = 4 \cdot p(\langle\rangle) = 4 \cdot \frac{\binom{13}{5}\binom{39}{0}}{\binom{52}{5}} = 4 \cdot 4.95 \cdot 10^{-4}$$

Distribuzione Ipergeometrica

- **Distribuzione ipergeometrica di parametri (N, K, n)**
  - supponendo di estrarre, senza reimmissione, n oggetti da un insieme che ne contiene N di cui K "neri" e N-K "bianchi", la v.a. che conta il numero di oggetti neri che si trovano nel campione estratto si chiama v.a. ipergeometrica di parametri N, K, n, $X \sim G(N,K,n)$.

$$p_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \qquad \begin{cases} 0 \le k \le n \\ k \le K \\ (n-k) \le (N-K) \end{cases}$$

Densità discreta della v.a. Ipergeometrica $X \sim G(100,30,10)$



- **Valore atteso**

$$E(X) = n\tfrac{K}{N}$$

- **Varianza**

$$Var(X) = n\tfrac{K}{N}\left(1-\tfrac{K}{N}\right)\left(\tfrac{N-n}{N-1}\right)$$

Esempio distribuzione ipergeometrica

- Un computer ha 30 unità disco; in ognuna sono archiviati 100 file. Un programma di esecuzione richiede di accedere a 40 file, tutti diversi. Qual è la probabilità che sia necessario accedere all'unità 1?
  - Sia Z il numero di file necessari all'esecuzione del programma che si trovano nell'unità 1.

$$\Pr(Z = i) = \frac{\binom{100}{i}\binom{2900}{40-i}}{\binom{3000}{40}} \qquad i = 0,1,\dots,40$$

  - $\Pr(Z > 0)$ è la probabilità che sia necessario accedere all'unità 1.

$$\Pr(Z > 0) = \sum_{i=1}^{40} \Pr(Z = i) = 1 - \Pr(Z = 0) = 1 - \frac{\binom{100}{0}\binom{2900}{40}}{\binom{3000}{40}} = 0.745$$

Distribuzione di Poisson

- **Distribuzione di Poisson di parametro λ**
  - la v.a. che approssima leggi binomiali per valori grandi di n e piccoli di p si chiama v.a. di Poisson di parametro λ, $X \sim P_0(\lambda)$

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!} \qquad k = 0,1,2,\ldots$$

  - è la legge limite per la distribuzione binomiale

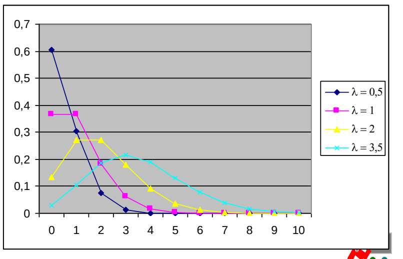
$$\lim_{n \to \infty} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \frac{np^k \cdot e^{-np}}{k!} = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \qquad con \ \lambda = np$$

- **Valore atteso**

$$E(X) = \lambda$$

- **Varianza**

$$Var(X) = \lambda$$



Densità discreta della v.a. di Poisson $X \sim P_0(\lambda)$

Esempio distribuzione di Poisson

- In una centrale telefonica arrivano chiamate con una densità di k in un'ora. Il numero di chiamate in qualsiasi intervallo è distribuito secondo la legge di Poisson. Trovare la probabilità che in 2 minuti arrivi almeno una chiamata.
  - Occorre calcolare il parametro λ relativo al periodo di interesse:

$$k : 60 = \lambda : 2 \longrightarrow \lambda = \frac{2k}{60}$$

  - Sia X il numero di chiamate nell'intervallo di interesse, si ha:

$$\Pr(X > 1) = 1 - P_0(0) = 1 - e^{-\lambda} = 1 - e^{-\frac{k}{30}}$$

Esercizio 5

- Con riferimento a un provider di servizi di collegamento internet, si assume che una percentuale pari a 1.4% delle segnalazioni di disservizio sia stata inviata dagli utenti per errore.

  Quesito A. Calcolare la probabilità che su 10 segnalazioni non ve ne siano di errate.

  Quesito B. Utilizzare l'approssimazione di Poisson per calcolare la probabilità che vi siano 2 segnalazioni errate sulle 500 ricevute.

- **Soluzione Quesito A**: il numero di segnalazioni effettuate per errore è una variabile casuale binomiale $X \sim B(n,p)$ *con n=10, p=0.014 da cui:*

$$p_X(0) = \binom{10}{0} \cdot 0.014^{0} \cdot (1 - 0.014)^{10-0} = 0.8685$$

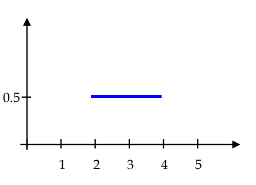
- **Soluzione Quesito B** : in questo caso *n=500, p=0.014 e np=500*0.014=7.* Il numero di segnalazioni effettuate per errore si può approssimare con una variabile casuale di Poisson Y di parametro $\lambda = 7$.

$$p_Y(2) = e^{-7} \frac{7^2}{2!} \cong 2.2341 \times 10^{-2}$$

Variabili aleatorie continue

- **Variabile Aleatoria Continua**
  - una v.a. continua X è definita assegnando una funzione di densità di probabilità (pdf) $f_X : \Re \rightarrow \Re$ con le seguenti proprietà:
  - $f_X(t) \geq 0$ per ogni $t \in \Re$
  - $\int_{\Re} f_X(t)dt = 1$

densità continua della v.a. uniforme su (2, 4)

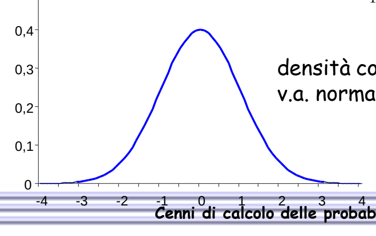


- **pdf uniforme**

$$f_X(t) = \frac{1}{b-a} I_{(a,b)}(t) \qquad \text{dove } I_{(a,b)}(t) = \begin{cases} 1 & t \in (a,b) \\ 0 & t \notin (a,b) \end{cases}$$

- **pdf Normale**
  - $X \sim N(0,1)$

$$f_X(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}$$

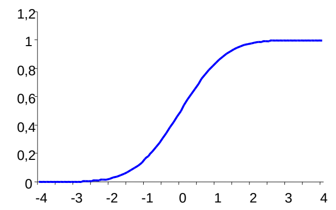


densità continua della v.a. normale standard

Proprietà delle variabili continue

- **Funzione di ripartizione**
  - Si definisce funzione di ripartizione v.a. X la funzione monotona crescente $F_X : \Re \rightarrow [0, 1]$ definita da:
  - $F_X(t) = \Pr(X \leq t)$ per ogni $t \in \Re$

$$F_X(t) = \int_{-\infty}^{t} f_X(y)dy$$



Funzione di ripartizione della v.a. normale standard

- **Valore atteso**

$$E(X) = \int_{\Re} t \cdot f_X(t)dt$$

  - **Distribuzione uniforme** $E(X) = (a+b)/2$
  - **Distribuzione normale standard** $E(X)=0$

- **Varianza**

$$Var(X) = \int_{\Re} (t - E(X))^2 \cdot f_X(t)dt = \int_{\Re} t^2 \cdot f_X(t)dt - \left[\int_{\Re} t \cdot f_X(t)dt\right]^2$$

Bibliografia

- M. Bramanti:
  Calcolo delle Probabilità e Statistica
  Progetto Leonardo, Ed. Esculapio, Bologna, 1997

- http://www-zeus.roma1.infn.it/~agostini/PRO/