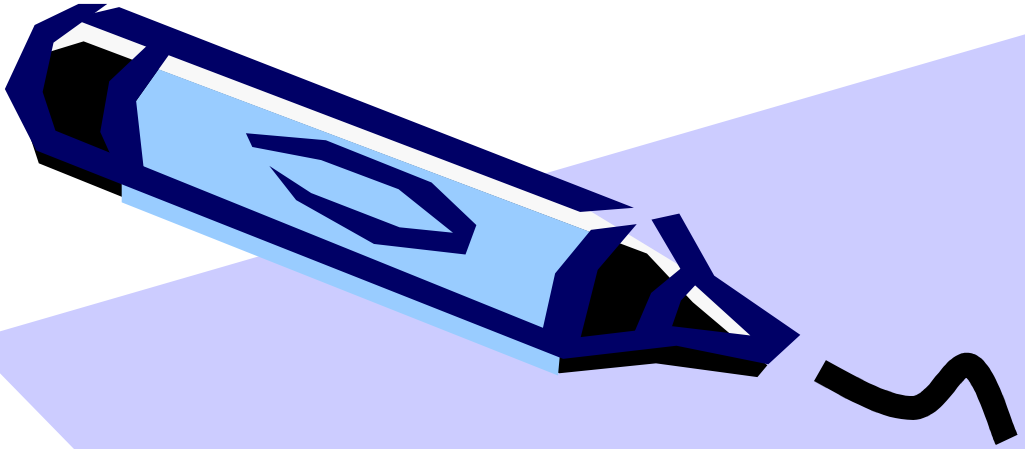
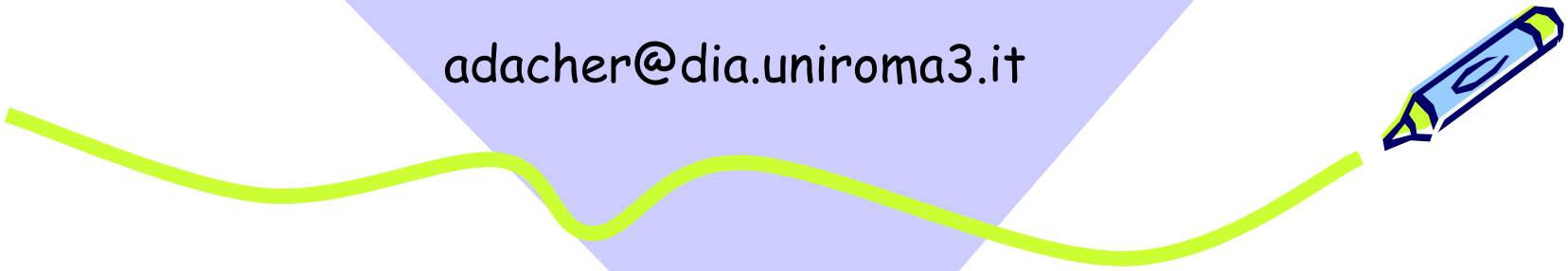
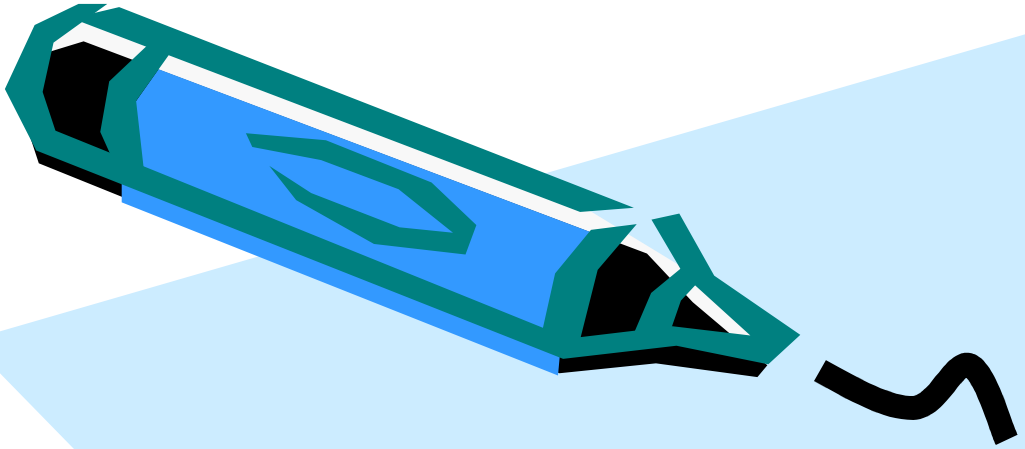


Simulazione dei Sistemi produttvi e logistci

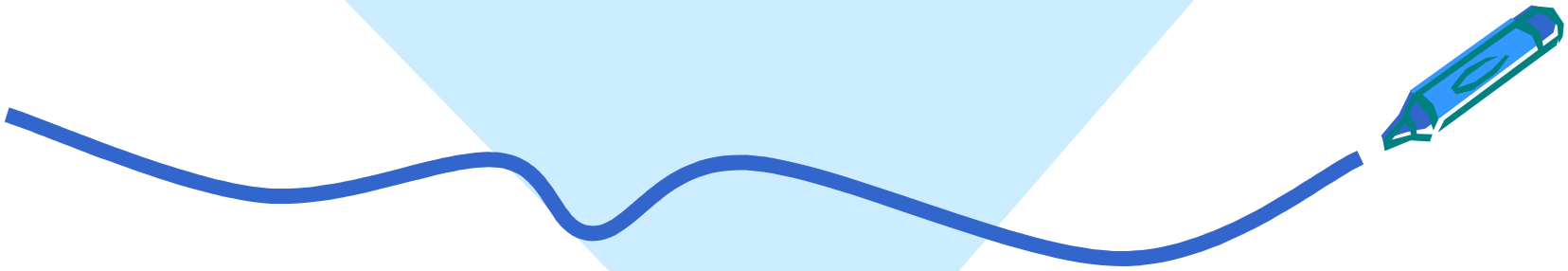
adacher@dia.uniroma3.it



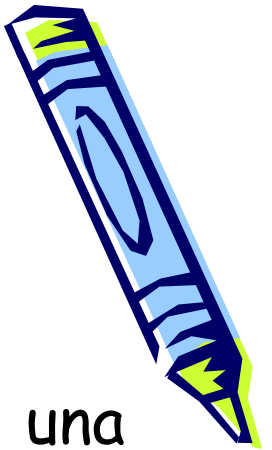


Introduzione

Sistemi e Modelli



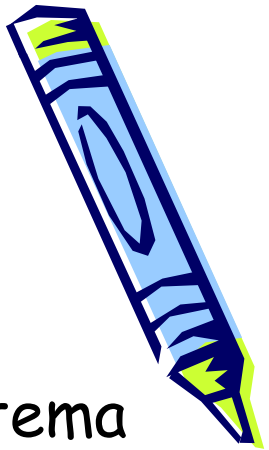
Sistemi e Modelli



Lo studio e l'analisi di sistemi tramite una rappresentazione astratta o una sua formalizzazione è utilizzato in molte differenti discipline scientifiche dall'informatica, alla fisica, dalla biologia all'economia.

Definiamo un **sistema** come un insieme di componenti (elementi, entità) interdipendenti e che interagiscono per raggiungere un determinato obiettivo.

Sistemi e Modelli



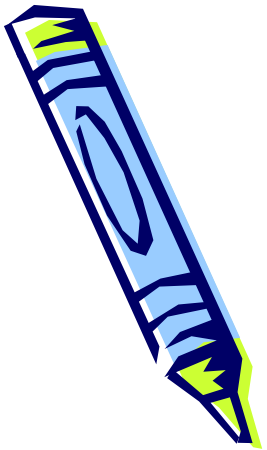
Lo studio e l'analisi del comportamento di un sistema e la sua valutazione in termini di costo e prestazioni è fondamentale durante tutto il ciclo di vita del sistema.

In particolare

- *nella fase di progettazione:*

  questo caso include il progetto di sistemi non esistenti, anche in una fase iniziale, quando occorre operare delle scelte fra configurazioni alternative valutandole senza avere a disposizione le relative implementazioni;

Sistemi e Modelli



- *nella fase di dimensionamento e acquisizione:*

  questa fase comprende le scelte fra diversi sistemi o componenti disponibili ed esistenti;

- *nella fase di evoluzione della configurazione e del carico:*

  in questo caso si considerano tutti gli aspetti e i problemi relativi alla modifica ed evoluzione di un sistema esistente, tipicamente per una sua espansione o un suo miglioramento, sia per variazioni della configurazione che per variazioni del carico di lavoro.

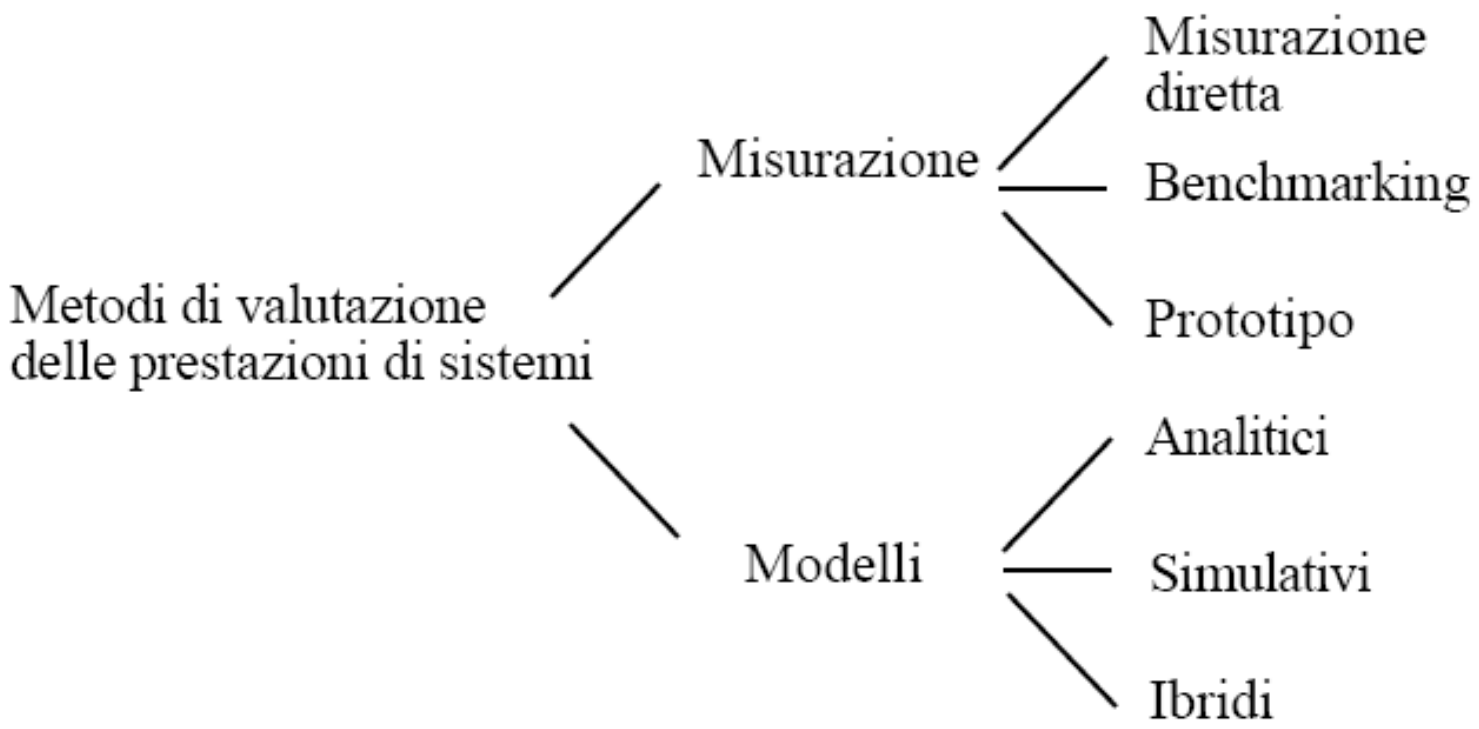
Sistemi e Modelli

Le metodologie per la valutazione delle prestazioni di sistemi possono essere distinte in due categorie principali

- tecniche di misurazione
- tecniche modellistiche

Le prestazioni di un sistema di elaborazione possono essere quantificate da figure di merito o **indici di prestazione** che descrivono l'efficienza dello svolgimento delle sue funzioni. Nel primo caso gli indici di prestazione del sistema vengono misurati, mentre nel secondo caso vengono calcolati, applicando e risolvendo modelli analitici, o stimati, utilizzando ed eseguendo modelli di simulazione

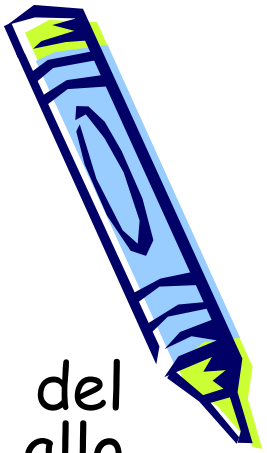
Sistemi e Modelli



Sistemi e Modelli

L'uso dei modelli per la valutazione e lo studio del comportamento dei sistemi diventa indispensabile nella fase di progetto di sistemi non esistenti (per cui le tecniche di misurazione diretta o artificiale non sono applicabili) e in particolar modo nei primi stadi di progetto in cui è importante poter discernere fra differenti alternative senza dover scendere ad un livello di dettaglio elevato.

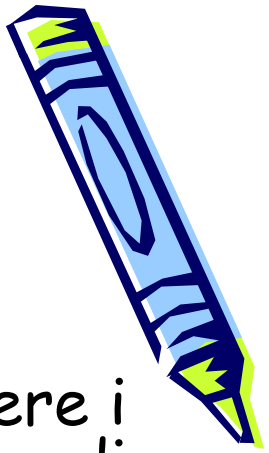
Sistemi e Modelli



Un **modello** è una rappresentazione astratta del sistema che include solo gli aspetti rilevanti allo scopo dello studio del sistema. Un modello è definito ad un determinato **livello di astrazione**, ovvero il sistema viene descritto con un certo livello di dettaglio, includendo nella rappresentazione solo quelle componenti e interazioni fra componenti che si ritengono necessarie allo scopo prefisso.

Alla definizione del modello segue la sua **parametrizzazione**, per poter considerare le alternative di studio, e la sua valutazione o soluzione per ottenere le informazioni relative allo studio del sistema.

Sistemi e Modelli

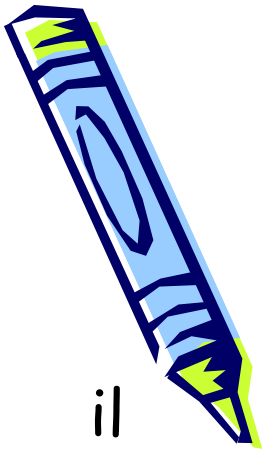


Fra le tecniche modellistiche si possono distinguere i modelli e i metodi analitici e i modelli e le tecniche di simulazione.

In un **modello analitico** le componenti e il carico del sistema sono rappresentate da variabili e parametri, e le interazioni fra le componenti da relazioni fra queste quantità.

La valutazione del sistema effettuata utilizzando il modello analitico richiede il calcolo della sua soluzione tramite metodi analitici o soluzioni numeriche.

Sistemi e Modelli



Un **modello di simulazione** riproduce il comportamento dinamico del sistema nel tempo rappresentando le componenti e le interazioni in termini di relazioni funzionali.

La valutazione di un sistema tramite un modello di simulazione richiede l'esecuzione (*run*) di un programma di simulazione, o simulatore che rappresenta l'evoluzione "temporale" del sistema e su cui si effettuano delle misure per stimare le grandezze di interesse.

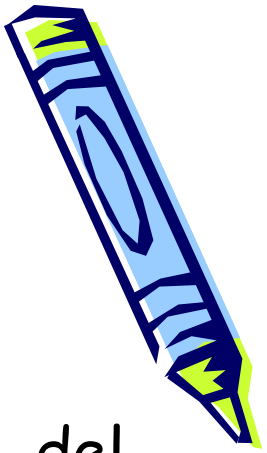
Sistemi e Modelli

Riassumendo, la definizione e l'impiego di un modello per lo studio di un sistema presenta diversi **vantaggi**, fra i quali:

- *aumento delle conoscenze* :

  la definizione di un modello aiuta ad organizzare le conoscenze teoriche e le osservazioni empiriche sul sistema, portando ad una maggiore comprensione del sistema stesso; infatti durante il processo di astrazione occorre identificare quali sono le componenti e le interazioni rilevanti allo scopo dello studio.

Sistemi e Modelli



- *analisi del sistema*:

  l'impiego di un modello facilita l'analisi del sistema;

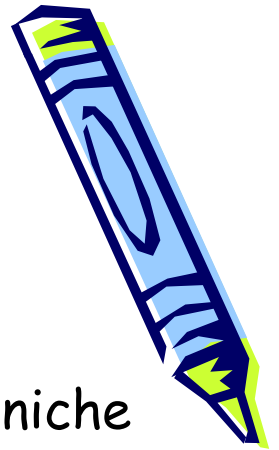
- *modificabilità*:

  il modello è maggiormente modificabile e manipolabile rispetto al sistema stesso permettendo la valutazione di diverse alternative, compatibilmente con la definizione e il livello di astrazione adottato;

- *diversi obbiettivi di studio:*

  l'impiego di diversi modelli dello stesso sistema permette la valutazione di diversi obiettivi.

Sistemi e Modelli



D'altro canto fra i limiti e gli svantaggi delle tecniche modellistiche notiamo:

- *scelta del modello*:

  la scelta del livello di astrazione appropriato può essere un compito non semplice; l'uso di un modello non appropriato può chiaramente portare ad errori di valutazione;

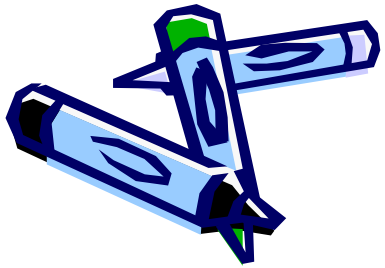
- *uso errato del modello*:

  vi è il rischio di utilizzare un modello oltre il suo campo di validità, ovvero anche quando le assunzioni e le ipotesi che hanno portato alla sua definizione non sono più verificate; in altre parole, occorre fare attenzione ad un uso improprio del modello dovuto all'estrapolazione dei risultati oltre il suo campo di applicabilità.

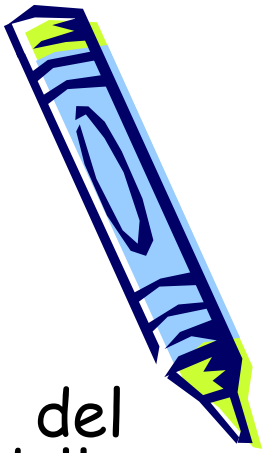
Classificazione dei Sistemi

L'evoluzione nel tempo di un sistema è descritta, ad ogni istante, dallo **stato** del sistema che ne rappresenta la condizione in quel particolare momento. Lo stato è espresso in termini di variabili di stato che descrivono le entità del sistema e i loro attributi.

Le **attività** delle componenti nel tempo e le interazioni fra le componenti sono descritte dalle regole di trasformazione fra stati.



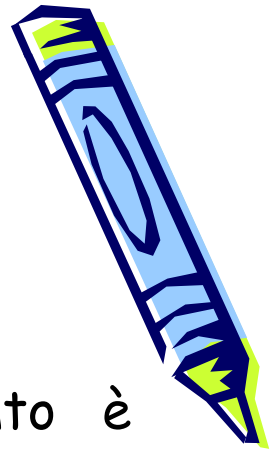
Classificazione dei Sistemi



La descrizione nel tempo del comportamento del sistema e della sua evoluzione è rappresenta dalla **storia degli stati**, ovvero dalla successione temporale degli stati del sistema.

Un sistema opera in un ambiente che può influenzare il comportamento del sistema stesso. Occorre quindi identificare senza ambiguità il sistema e la sua interfaccia rispetto all'ambiente esterno. Le variabili di stato si distinguono in variabili **endogene**, se il loro cambiamento è dovuto soltanto ad attività interne al sistema, e variabili **esogene** se sono influenzate dall'ambiente esterno al sistema.

Classificazione dei Sistemi

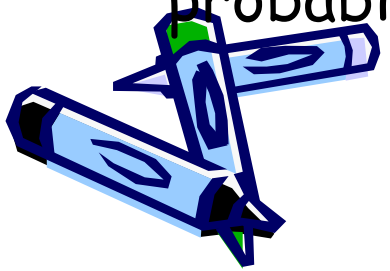


Un sistema è detto **chiuso** se il suo comportamento è completamente determinato da attività interne, cioè se non esistono variabili esogene. Al contrario, un sistema è **aperto** se interagisce con l'ambiente esterno, come viene espresso dalle variabili esogene.

I sistemi si distinguono in **continui** o **discreti** a seconda del tipo di cambiamento dei valori, continuo o discreto, delle variabili di stato. Ad esempio se la variabile di stato rappresenta la temperatura in un dato luogo, poiché i suoi cambiamenti sono graduali e continui, abbiamo un sistema continuo. Viceversa, se ad esempio il sistema è descritto dal numero di persone presenti in una stanza, i cambiamenti avvengono istantaneamente per passi discreti e quindi si osserva un sistema discreto.

Classificazione dei Sistemi

Il modo in cui avvengono le trasformazioni fra stati determina se un sistema è **deterministico** o **stocastico**. Nel primo caso le regole di trasformazione determinano univocamente il cambiamento di stato del sistema, mentre nel secondo caso da uno stato è possibile raggiungere diversi stati secondo una legge di probabilità associata alla regola di trasformazione. Esempi di sistemi deterministici si possono osservare in alcuni sistemi di produzione e di automazione I sistemi stocastici in cui le variabili di stato variano con casualità secondo leggi di distribuzione di probabilità si osservano in diversi campi.

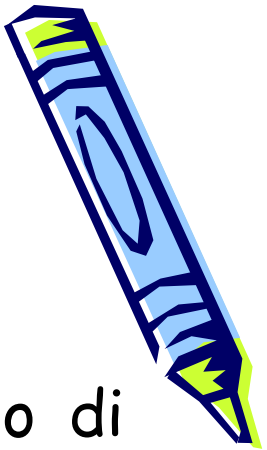


Sistemi e Modelli

La natura stocastica o deterministica, continua o discreta di un sistema non è una sua proprietà assoluta, ma dipende dalla visione da parte dell'osservatore del sistema stesso che è determinata dagli obiettivi e dal metodo di studio, così come dall'esperienza dell'osservatore.

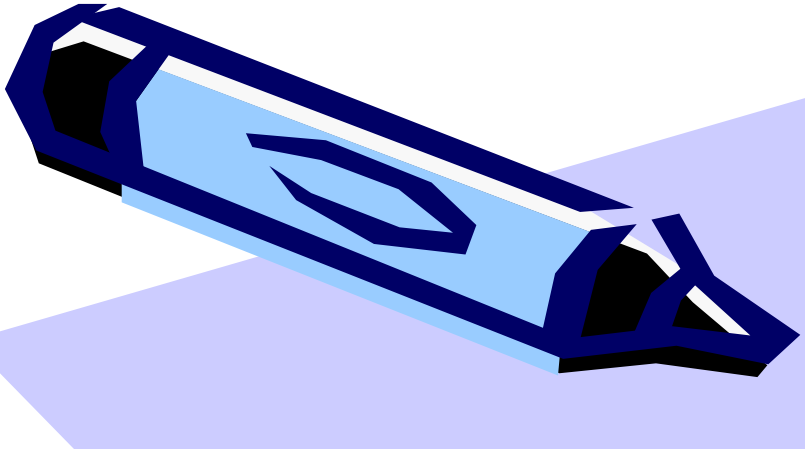
Analogamente ai sistemi, anche i modelli possono essere distinti in aperti e chiusi, continui e discreti, deterministici e stocastici. Non necessariamente il tipo di modello corrisponde al tipo di sistema rappresentato.

Modelli

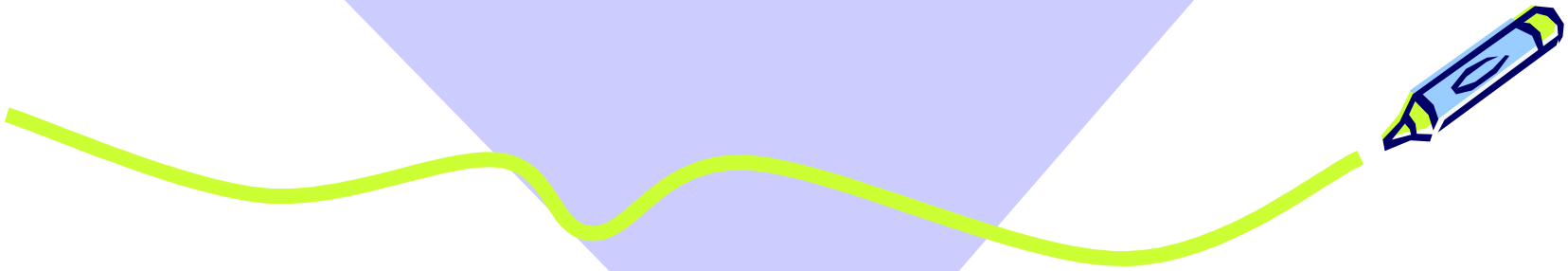


La natura del modello dipende non solo dal tipo di sistema studiato ma anche dal livello di astrazione impiegato e dall'obiettivo per il quale il modello è definito.

Infatti il modello deve riprodurre tutte quelle proprietà elementari delle componenti del sistema e le loro interazioni da cui dipendono le funzionalità, oggetto di studio, che si è interessati a rappresentare e a valutare.



Classificazione dei Modelli



Classificazione

- sistema statico: l'uscita y(t) non dipende dall'ingresso u(τ), τ < t.
- sistema dinamico: l'uscita y(t) dipende dall'ingresso u(τ), τ < t.

Un sistema dinamico è tipicamente espresso da un sistema di equazioni del tipo

$$\begin{cases} \underline{\dot{x}}(t) = \underline{f}(\underline{x}(t), \underline{u}(t), t) \\ \underline{y}(t) = \underline{g}(\underline{x}(t), \underline{u}(t), t) \end{cases}$$

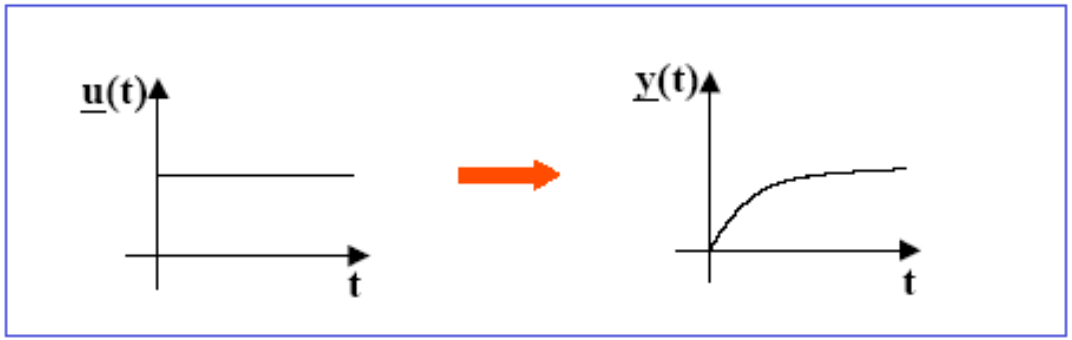
Classificazione

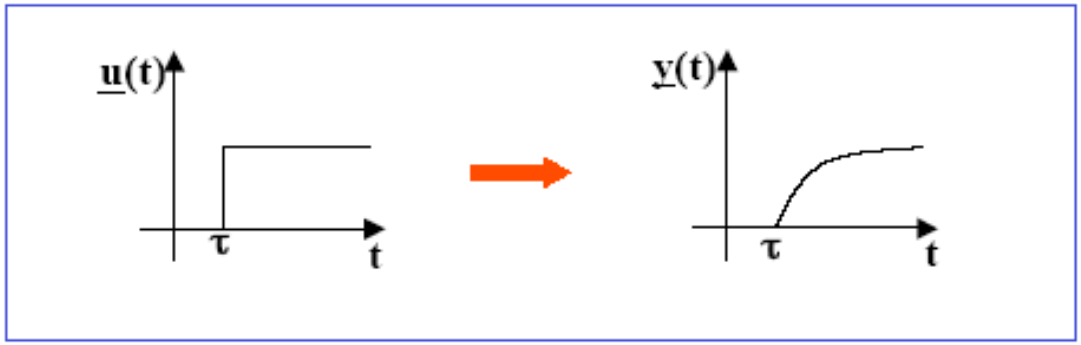
sistema a tempo continuo: ogni grandezza è definibile $t \varepsilon \Re$ (il tempo è una variabile continua).

sistema a tempo discreto: le grandezze sono definibili solo in istanti di tempo discreti, ossia l'insieme dei tempi è una sequenza di istanti equispaziati (l'intervallo di tempo tra due istanti consecutivi è costante ed è detto intervallo di campionamento). Il tempo è una variabile a valori interi.

Classificazione

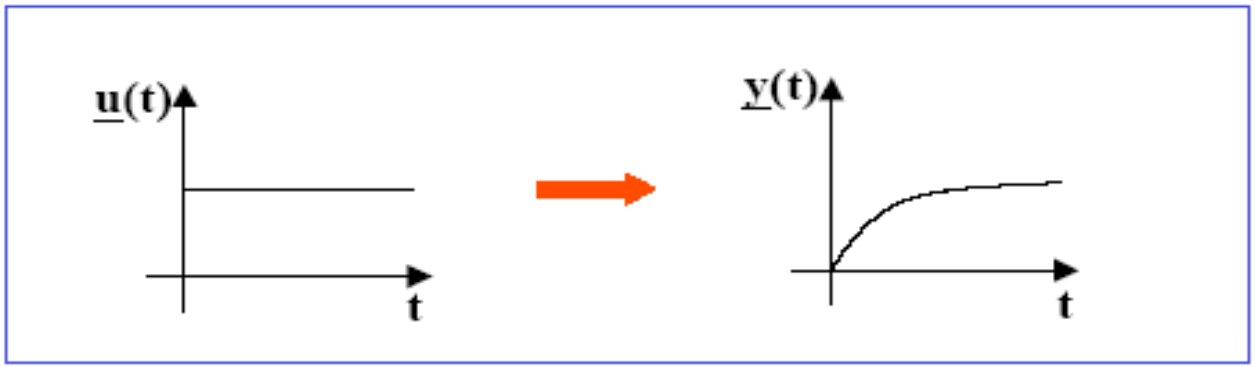
sistema dinamico tempo invariante (o stazionario): se l'ingresso u(t) applicato in t genera l'uscita y(t), lo stesso ingresso applicato in τ, ossia u(t-τ), genera l'uscita y(t- τ) per ogni τ

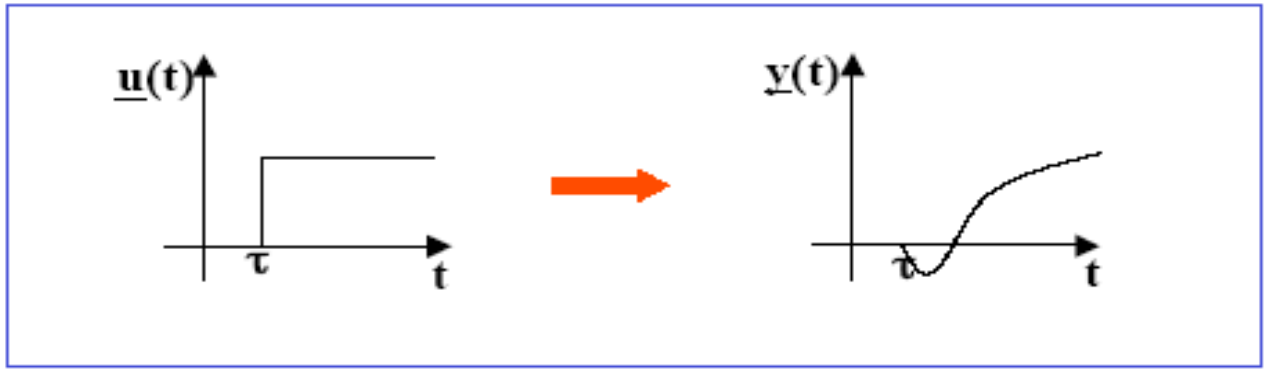




Classificazione

sistema dinamico tempo variante: la funzione che definisce la dipendenza dell'uscita dall'ingresso dipende esplicitamente dal tempo





Classificazione

sistema lineare: tutte le dipendenze sono lineari

sistema non lineare: esistono dipendenze non lineari

Un sistema dinamico lineare a tempo continuo è tipicamente espresso da un sistema di equazioni del tipo

$$\begin{cases} \dot{\underline{x}}(t) = A(t)\underline{x}(t) + B(t)\underline{u}(t) \\ \underline{y}(t) = C(t)\underline{x}(t) + D(t)\underline{u}(t) \end{cases}$$

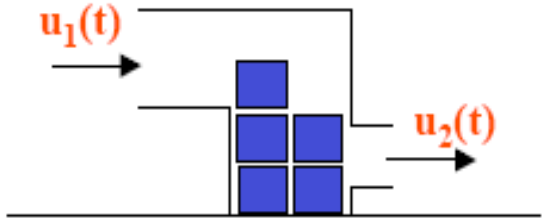
Se il sistema è stazionario, $A(t) \equiv A$, $B(t) \equiv B$, $C(t) \equiv C$, $D(t) \equiv D$.

Classificazione

sistema a stato continuo: lo spazio degli stati è un insieme continuo

sistema a stato discreto: lo spazio degli stati è un insieme discreto

Es.: magazzino di prodotti finiti in un sistema di produzione



$u_1(t)$: **prodotti in ingresso**

$u_2(t)$: **prodotti in uscita**

$x(t)$ = n° prodotti in magazzino

⇩

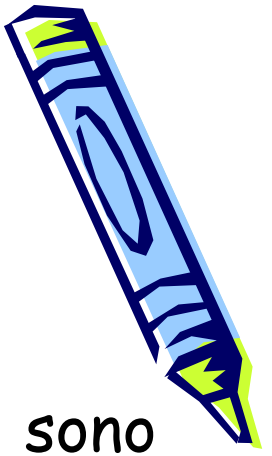
lo spazio degli stati è discreto

Classificazione

sistema guidato dal tempo: lo stato cambia in modo sincrono con il variare del tempo

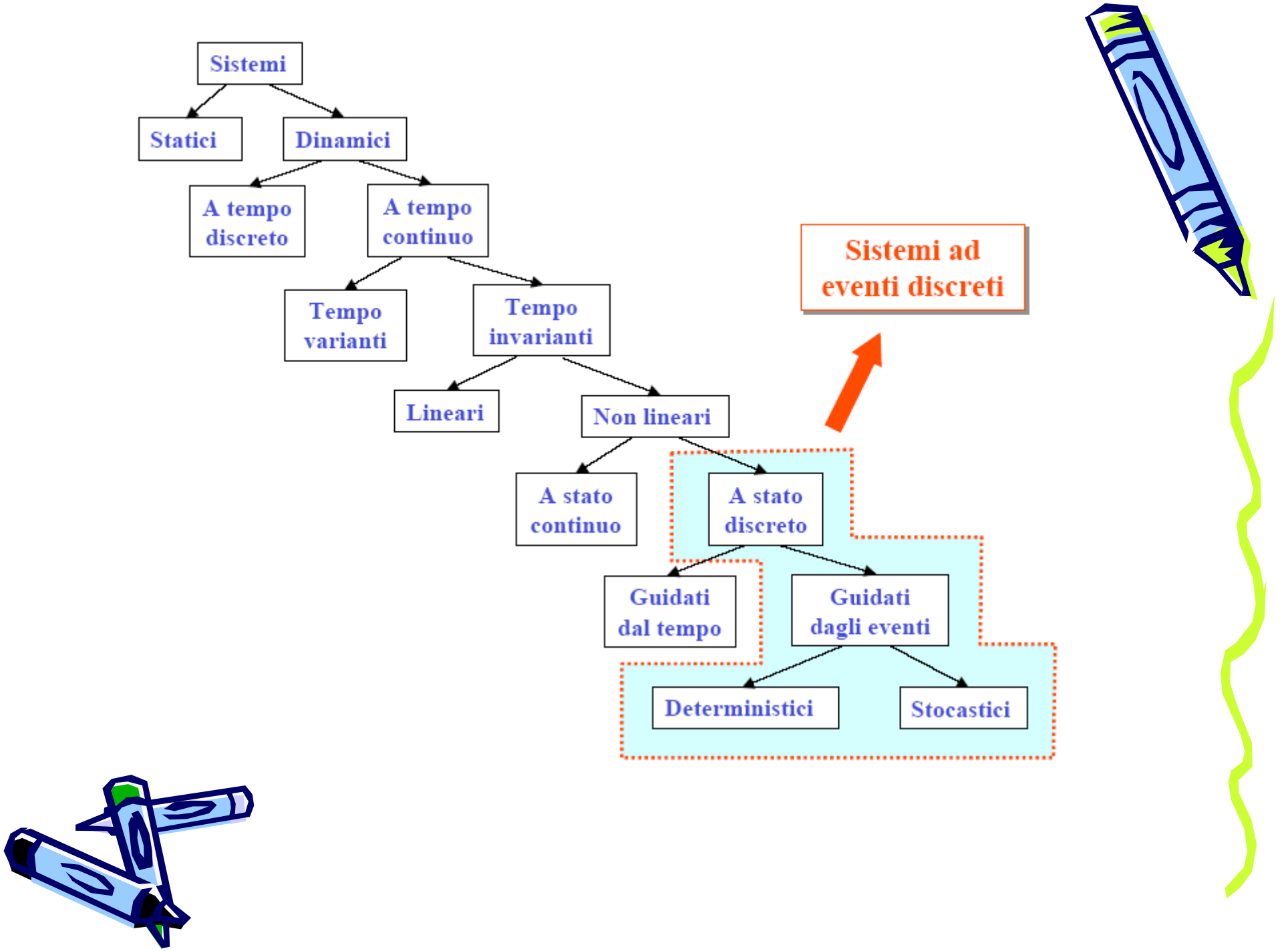
sistema guidato dagli eventi: lo stato cambia in modo asincrono in corrispondenza dell'occorrenza di condizioni particolari denominate **eventi**

Classificazione

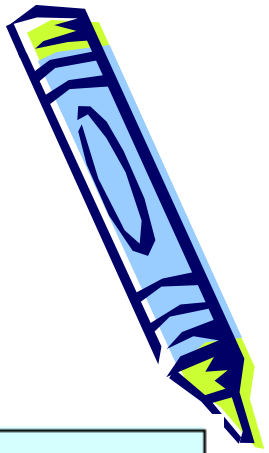


sistema deterministico: tutte le grandezze sono deterministiche

sistema stocastico: almeno una delle grandezze è una variabile stocastica (lo stato di un sistema dinamico stocastico è un processo stocastico che può essere definito e studiato in modo probabilistico)


Sistemi
Statici
Dinamici
A tempo discreto
A tempo continuo
Tempo varianti
Tempo invarianti
Lineari
Non lineari
A stato continuo
A stato discreto
Guidati dal tempo
Guidati dagli eventi
Deterministici
Stocastici
Sistemi ad eventi discreti

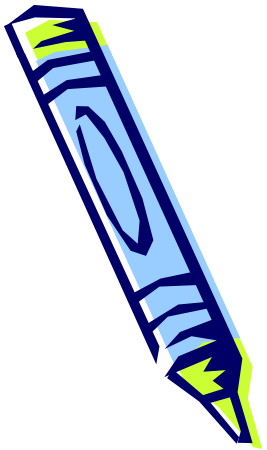
Sistemi ad eventi discreti (SED)



Definizione: Un ***sistema dinamico ad eventi discreti*** **(Discrete Event Dynamic System - DEDS)** è un sistema a stato discreto e guidato dagli eventi, ossia un sistema in cui i cambiamenti dello stato avvengono unicamente all'occorrenza di eventi asincroni.

Un evento può essere identificato:
con un'azione specifica (shut-down di un calcolatore in seguito ad uno specifico comando)
con un'occorrenza spontanea dovuta alla natura delle cose (shut-down di un calcolatore per motivi incomprensibili)

Esempi di SED

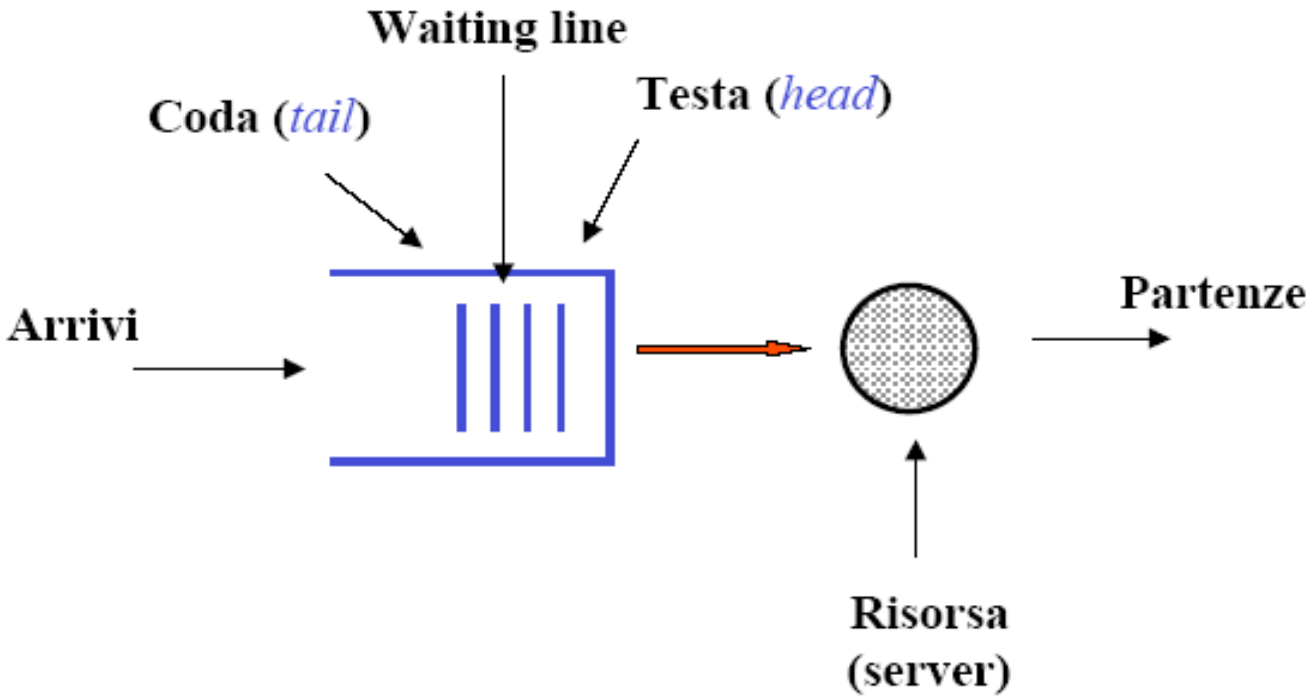


Sistemi a coda (***Queueing systems***)

Un sistema a coda include:

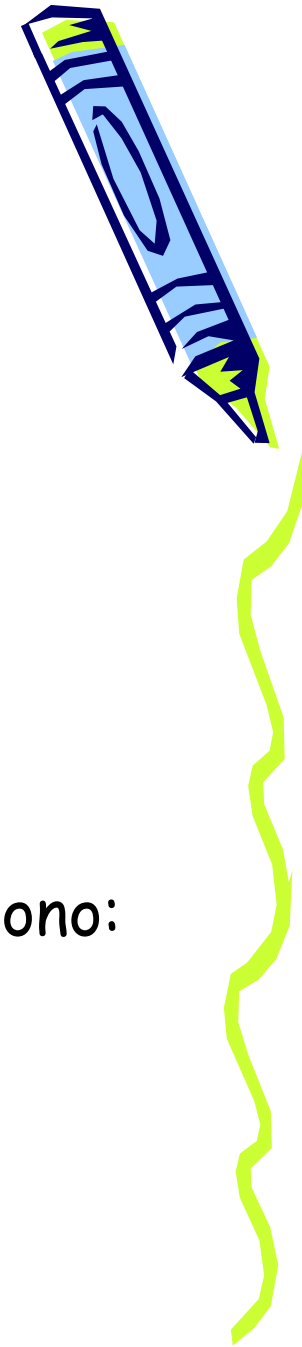
- una o più **risorse** condivise;
- un insieme di **clienti** che necessitano di un servizio da parte delle risorse;
- uno spazio fisico o virtuale, detto **coda**, dove i clienti attendono di ricevere il servizio.

Esempi di SED



Rappresentazione schematica di una coda

Sistemi a coda

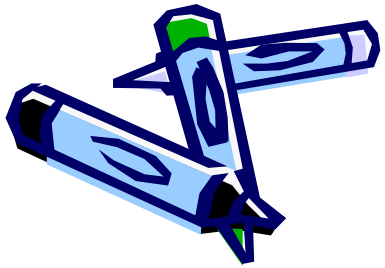


Lo **stato** di un sistema a coda è tipicamente rappresentato da:

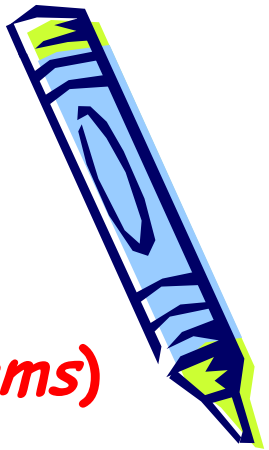
- lo stato del server (libero, occupato)
- lo stato di ogni cliente (in coda, in servizio)
- il numero di clienti in coda

Gli **eventi** che determinano transizioni di stato sono:

- arrivo di un cliente nel sistema
- partenza di un cliente dal sistema



Esempi di SED



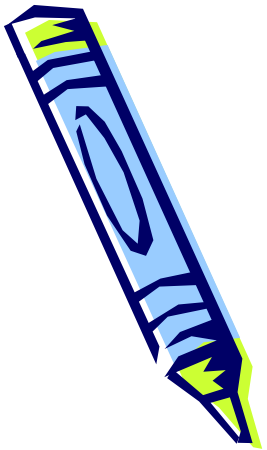
Sistemi di comunicazione (*Communication systems*)

I clienti di un sistema di comunicazione sono messaggi, pacchetti, chiamate, …

Un cliente di un sistema di comunicazione è generato da una sorgente, deve raggiungere una destinazione.

Per raggiungere la destinazione il cliente deve accedere ad una serie di servizi (stazioni di commutazione, computer, stazioni radio, satelliti,

Sistemi di Comunicazione



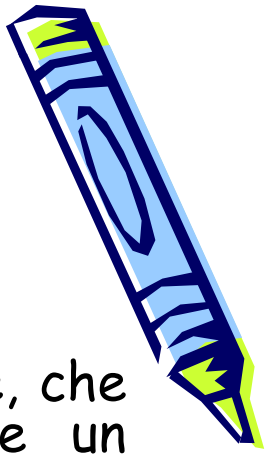
Un sistema di comunicazione deve garantire:

- accesso ai servizi efficiente e corretto ("fair")
- recapito del messaggio al destinatario in maniera corretta ed in tempi brevi

definizione ed utilizzo di ***protocolli* di comunicazione**

- **N.B.** Il progetto e la verifica funzionale di protocolli di comunicazione è un problema non banale

Sistemi di Comunicazione



Esempio

Due utenti, *A* e *B*, condividono un canale di comunicazione, che può servire un utente alla volta. Se *B* trasmette un messaggio al canale mentre *A* sta usando il canale, vi è una ***collisione*** ed il messaggio di *B* si perde.

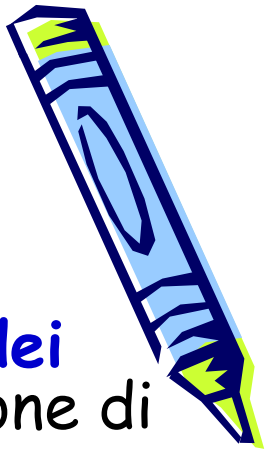
Il canale può trovarsi in 3 stati:

- I – Idle
- T – Trasmissione di un messaggio
- C – Trasmissione di due messaggi e collisione

Gli utenti possono trovarsi nei seguenti stati:

- I – Idle
- T – Trasmissione
- W – In attesa di trasmissione di un messaggio esistente

Sistemi di Comunicazione



Gli **eventi** che guidano il sistema sono gli **arrivi dei messaggi** ad A e B, le **azioni** di A e B (spedizione di messaggio sul canale), ed il **completamento di una trasmissione** da parte del canale

$$X = \{[x_A,\ x_B,\ x_{CH}]^T : x_A, x_B \in (I,T,W),\ x_{CH} \in (I,T,C)\}$$

$$E = \{a_A,\ a_B,\ t_A,\ t_B,\ t_{CH}\}$$

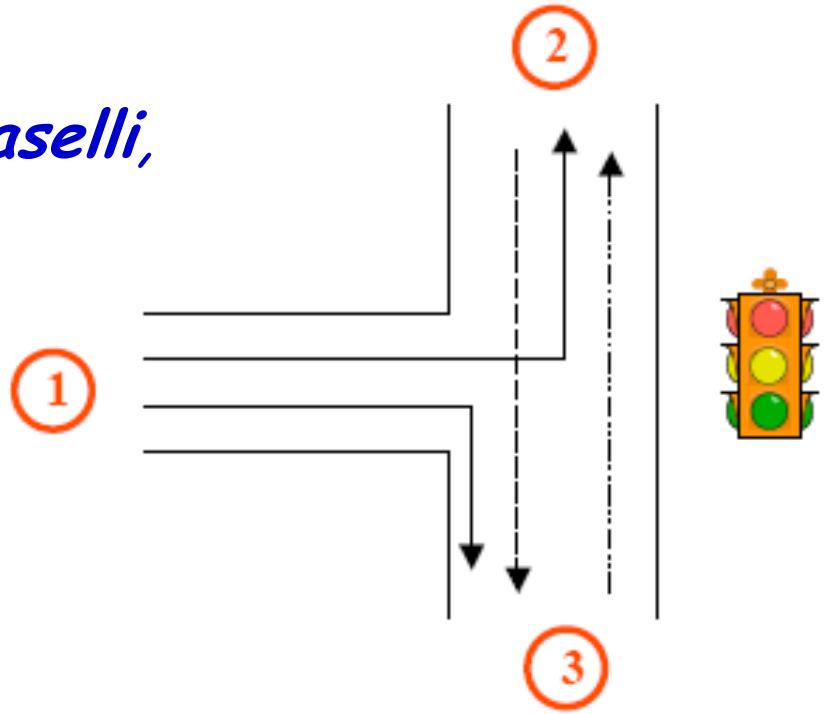
Problema

Quali regole di accesso per evitare continue collisioni?

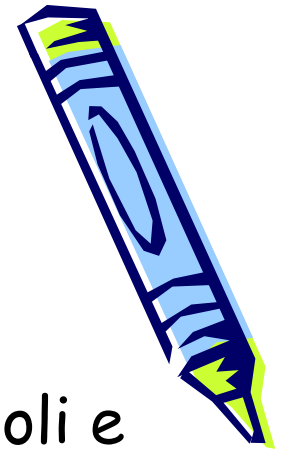
Esempi di SED

Sistemi di trasporto

- Clienti: ***veicoli***
- Risorse: ***semafori, caselli, strade***



Sistemi di trasporto

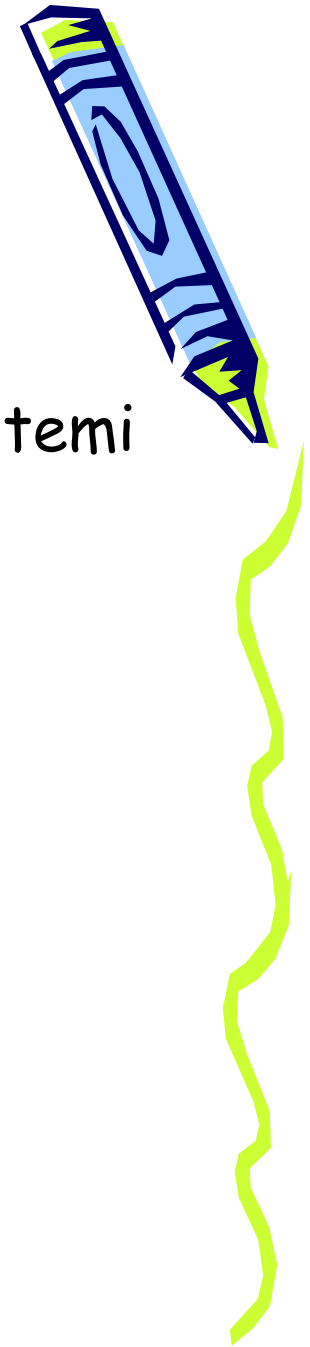


Spazio degli **stati**: lunghezza delle 4 code di veicoli e colore del semaforo

Spazio degli **eventi** $E = \{a_{12}, a_{13}, a_{23}, a_{32}, p_{12}, p_{13}, p_{23}, p_{32}, r, v\}$

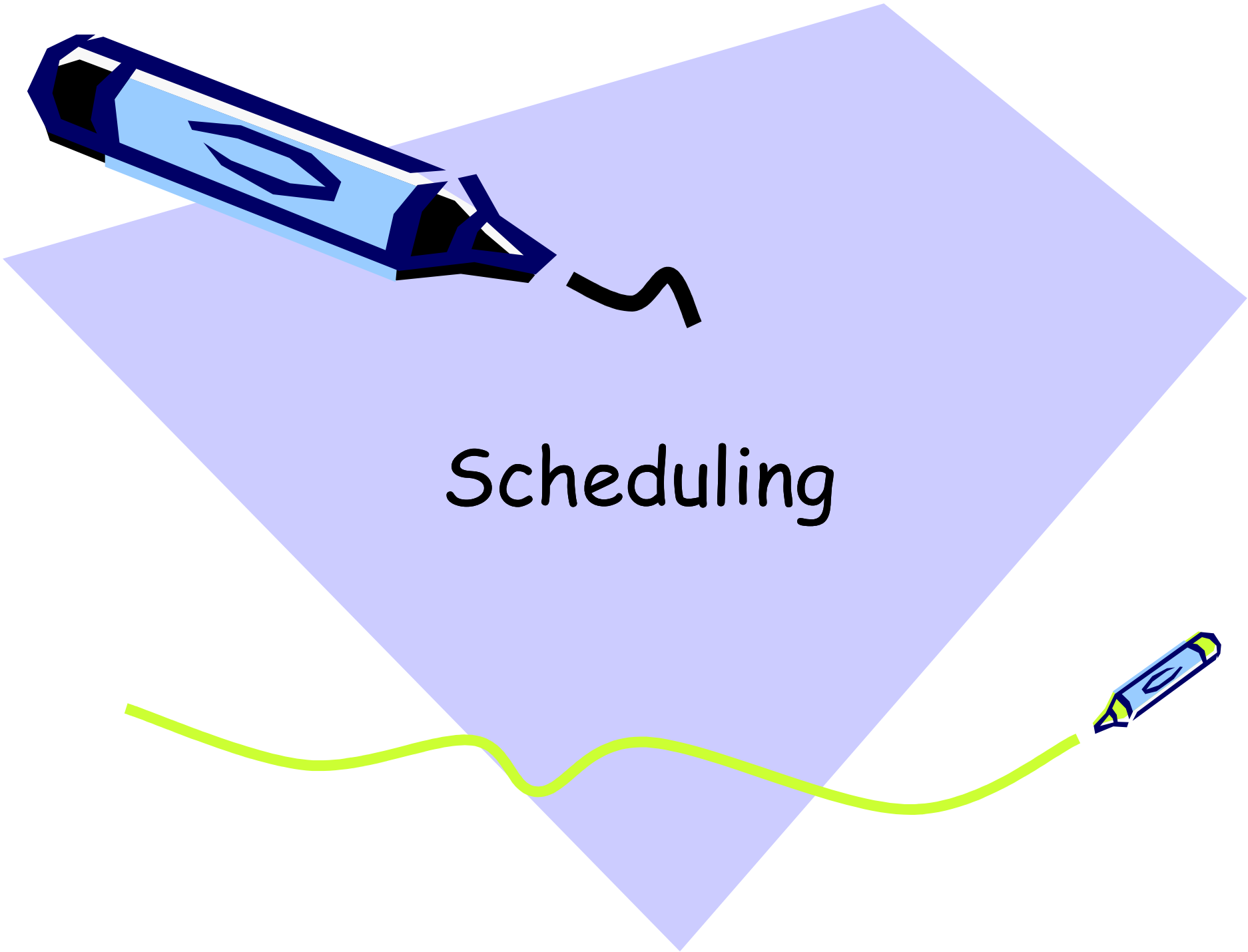
- a_{ij} : arrivo di un veicolo dal punto *i* diretto verso *j*
- p_{ij} : il veicolo proveniente da *i* lascia l'incrocio nella direzione *j*
- **v, r**: semaforo verde (rosso) per i veicoli nei versi (*23* e *32*)

Esempi di SED



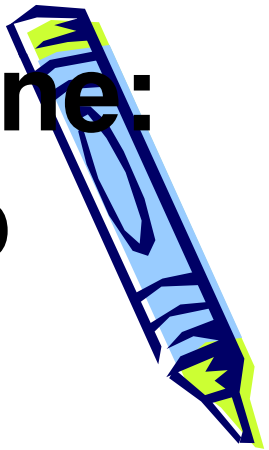
Altri sistemi tipicamente rappresentati come sistemi ad eventi discreti sono:

- Sistemi manifatturieri
- Sistemi di elaborazione


Scheduling

Organizzazione della produzione

Pianificazione della produzione: schedulazione di dettaglio



Sono dati:

- Un insieme di lavori (*job*): ognuno costituito da una o più operazioni
- Un insieme di risorse (*macchine*) che devono essere utilizzate per eseguire i lavori

Scheduling delle operazioni

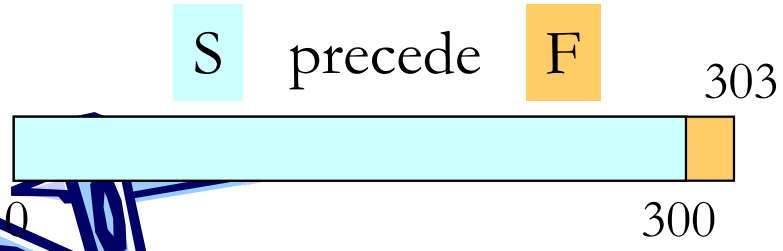
Scelta dei tempi di inizio e fine di ogni operazione su ogni macchina

Esempio: fotocopie

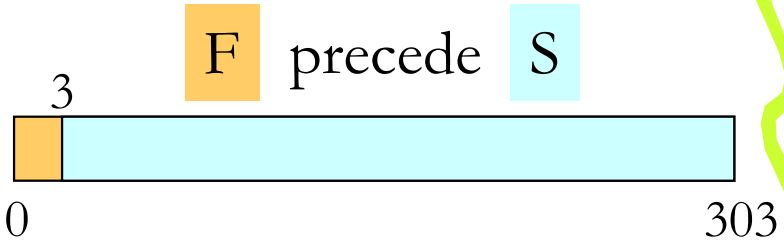
Silvia (S) e Francesca (F) giungono nello stesso momento ad una macchina fotocopiatrice. F deve fare 1 fotocopia, S ne deve fare 100.
Il galateo imporrebbe a S di lasciar passare per prima F, che impegna la macchina per un tempo breve e poi la lascia libera.

È questa una buona idea?

Analisi. Supponiamo che l'esecuzione di una fotocopia richieda 3 secondi. Due casi:



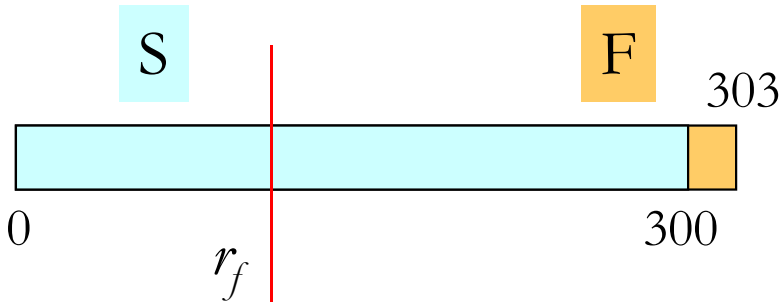
Attesa totale = 300 + 303 = 603



Attesa totale = 3 + 303 = 306

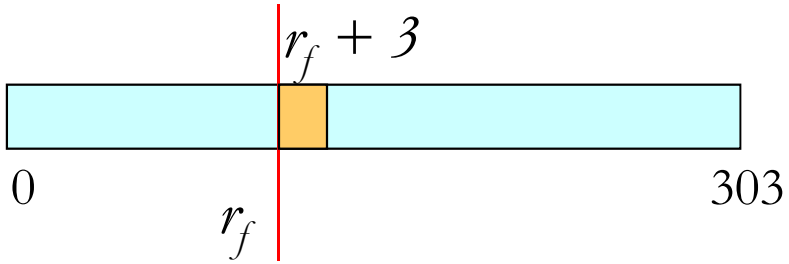
Esempio: fotocopie

E se Silvia (S) giungesse prima di Francesca (F) (che arriva all'istante r_f)?



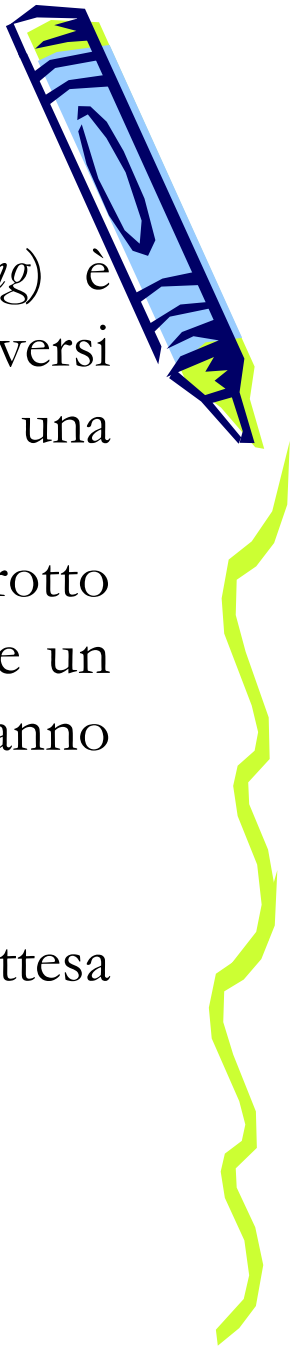
Attesa totale = $300 + 303 - r_f = 603 - r_f$

Di nuovo, il galateo imporrebbe a Silvia di interrompere le proprie copie in favore di Francesca:



Attesa totale = $3 + 303 = 306$

Esempio: CPU

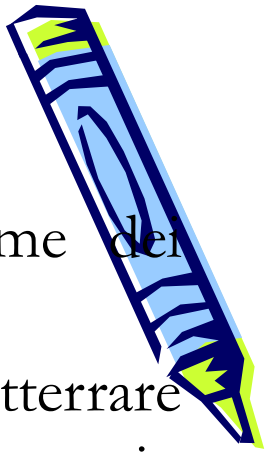


• uno dei compiti del sistema operativo (*multitasking*) è quello di disciplinare l'accesso alla CPU dei diversi programmi di calcolo (eventualmente associati ad una priorità).

• il processamento di un programma può essere interrotto per consentire il completamento di altri. Ciò evita che un programma "lungo" ne blocchi altri molto brevi che hanno priorità più bassa.

•L'obiettivo tipico consiste nel minimizzare l'attesa complessiva dei programmi.

Esempio: voli in atterraggio



• L'area terminale di un aeroporto contiene l'insieme dei velivoli prossimi all'atterraggio.

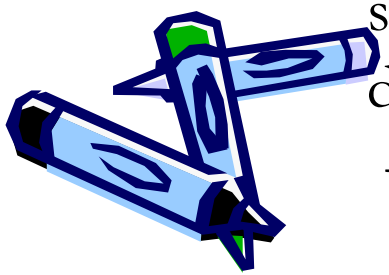
• Il sistema di controllo dell'aeroporto permette di far atterrare al più un volo alla volta e la durata della manovra di atterraggio è nota per ogni velivolo.

• Ad ogni velivolo è inoltre associato un tempo previsto di atterraggio

• Obiettivo del controllore può essere quello di decidere la sequenza in modo da minimizzare la somma (pesata) dei ritardi.

•Vincoli aggiuntivi:

- a causa della turbolenza, si deve garantire una certa separazione in atterraggio fra un aeromobile di grandi dimensioni ed di piccole dimensioni
- precedenze fra voli



Scheduling delle operazioni

Consideriamo:

3 lavori e 3 macchine

Job	Sequenza delle operazioni operazione=(macchina, tempo)		
J_1	(M_1,10)	(M_2,5)	(M_3,6)
J_2	(M_2,5)	(M_1,8)	-
J_3	(M_1,2)	(M_3,10)	(M_2,4)

Diagramma di Gantt

Job	Sequenza Operazioni		
J_1	$(M_1,10)$	$(M_2,5)$	$(M_3,6)$
J_2	$(M_2,5)$	$(M_1,8)$	-
J_3	$(M_1,2)$	$(M_3,10)$	$(M_2,4)$

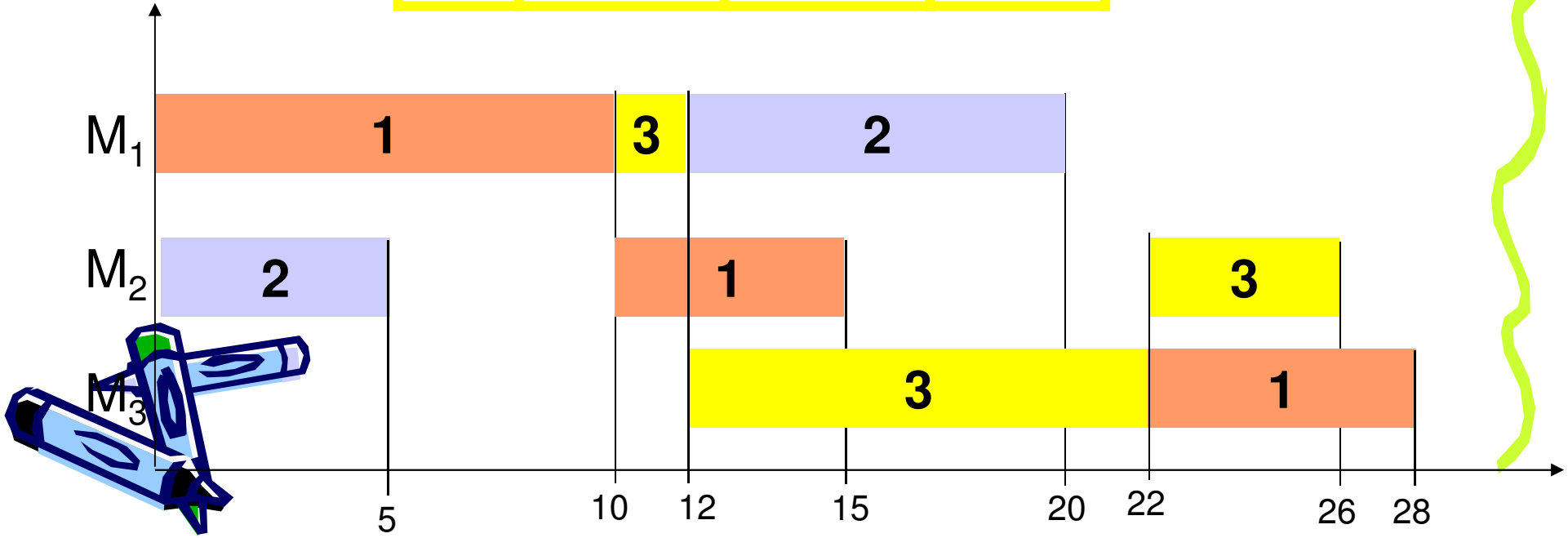
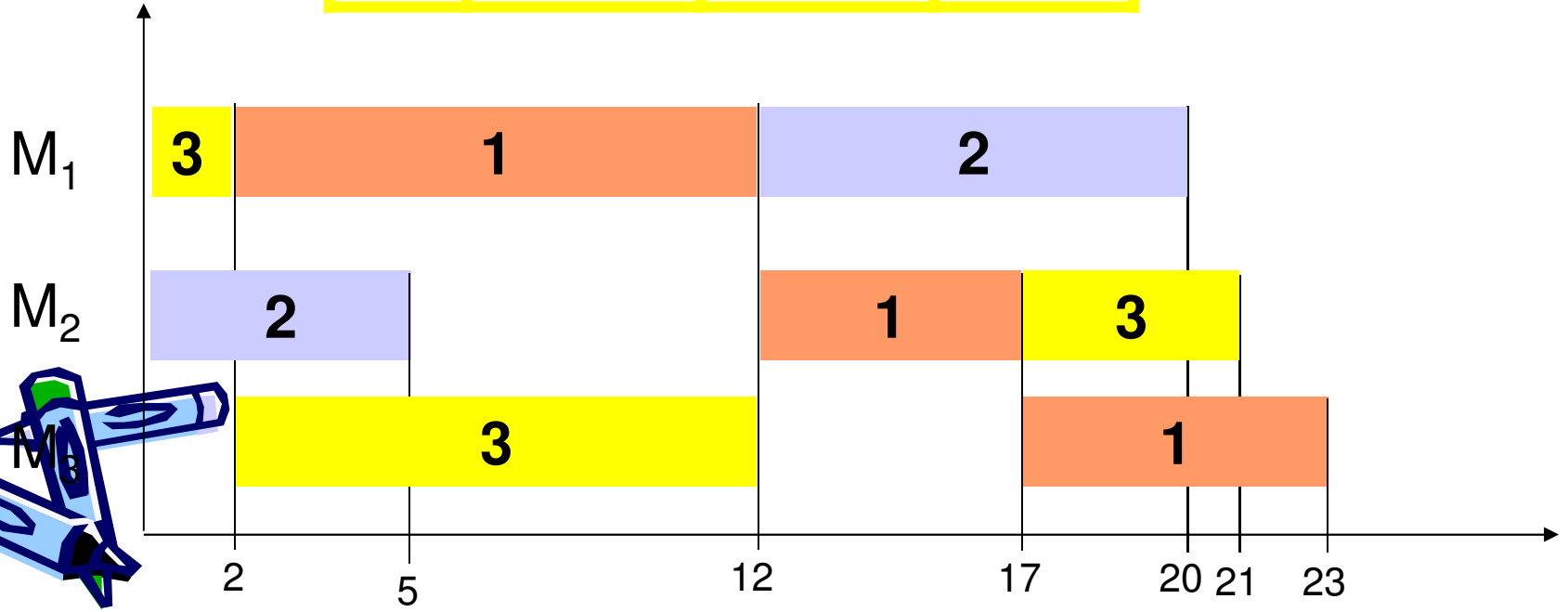


Diagramma di Gantt

Job	Sequenza Operazioni		
J_1	$(M_1,10)$	$(M_2,5)$	$(M_3,6)$
J_2	$(M_2,5)$	$(M_1,8)$	-
J_3	$(M_1,2)$	$(M_3,10)$	$(M_2,4)$



Esempio

Aldo (A), Bruno (B), Carlo (C), Duilio (D) condividono un appartamento. Ogni mattino ricevono 4 giornali: Financial Times (FT), Guardian (G), Daily Express (DE), Sun (S).
Ciascun lettore inizia la lettura ad una certa ora, ha la propria sequenza fissata di lettura dei gionali e legge ciascun giornale per un tempo prefissato:

lettore	Ora inizio	Sequenza lettura (tempo in min.)			
Aldo	8.30	FT(60)	G (30)	DE (2)	S(5)
Bruno	8.45	G(75)	DE(3)	FT(25)	S(10)
Carlo	8.45	DE(5)	G(15)	FT(10)	S(30)
Duilio	9.30	S(90)	FT(1)	G(1)	DE(1)

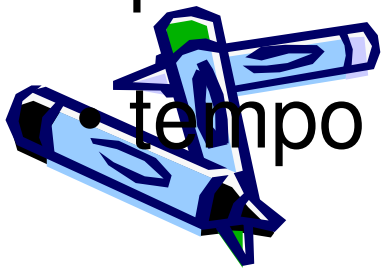
Esempio

- tutti i lettori preferiscono (eventualmente) aspettare che un giornale (il prossimo nella propria sequenza) sia disponibile anziché modificare la sequenza prefissata.
- nessun lettore rilascia un giornale prima di averlo letto completamente.
- ciascun lettore termina la lettura di tutti i giornali prima di uscire.
- i quattro lettori attendono che tutti abbiano terminato di leggere prima di uscire di casa

Problema:

Qual'è la minima ora in cui A, B, C, D possono uscire?

Classificazione dei problemi di *scheduling*



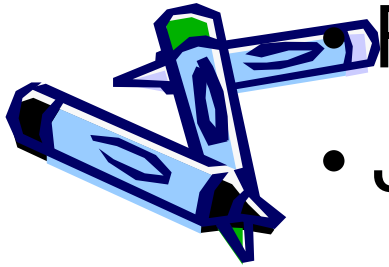
Caratterizzazione dei lavori:

- tempo di processamento p_j (p_{kj})
- data di consegna (*duedate* o *deadline*) d_j
- data di rilascio (*release date*) r_j
- peso del lavoro (priorità) w_j
- tempo di set-up tra due lavori s_{ij}

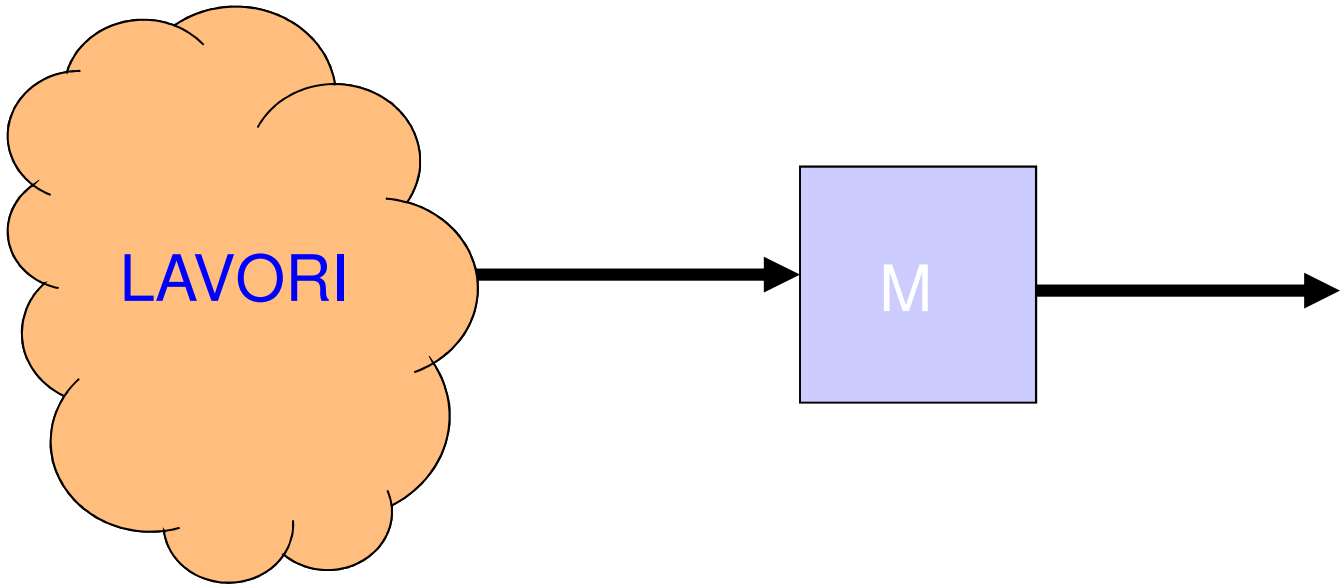
Classificazione dei problemi di *scheduling*

Caratterizzazione delle risorse e dell'ambiente produttivo:

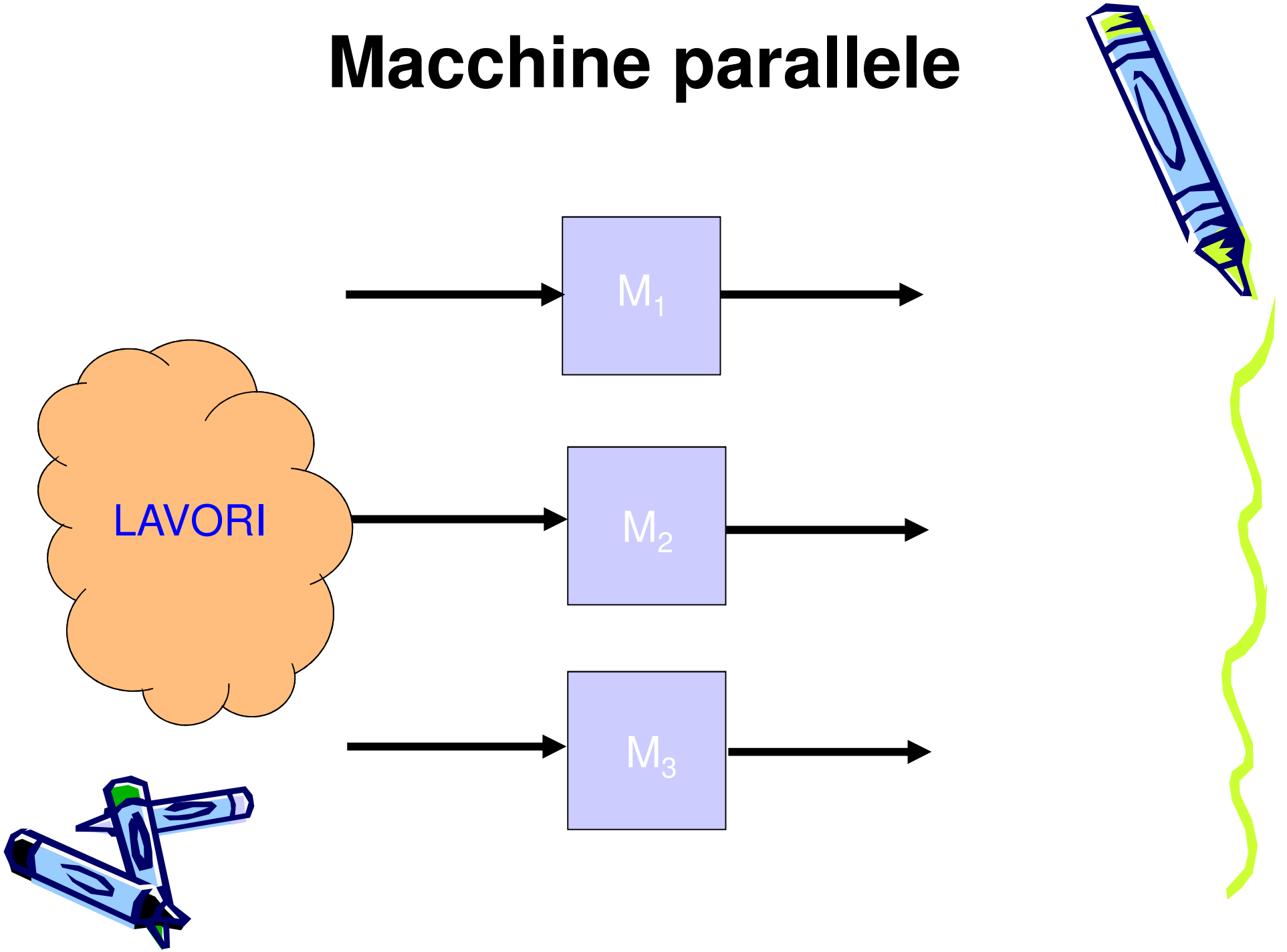
- macchina singola
- macchine parallele
  - *identiche*
  - *scorrelate*
  - *uniformi*
- Flow shop
- Job shop



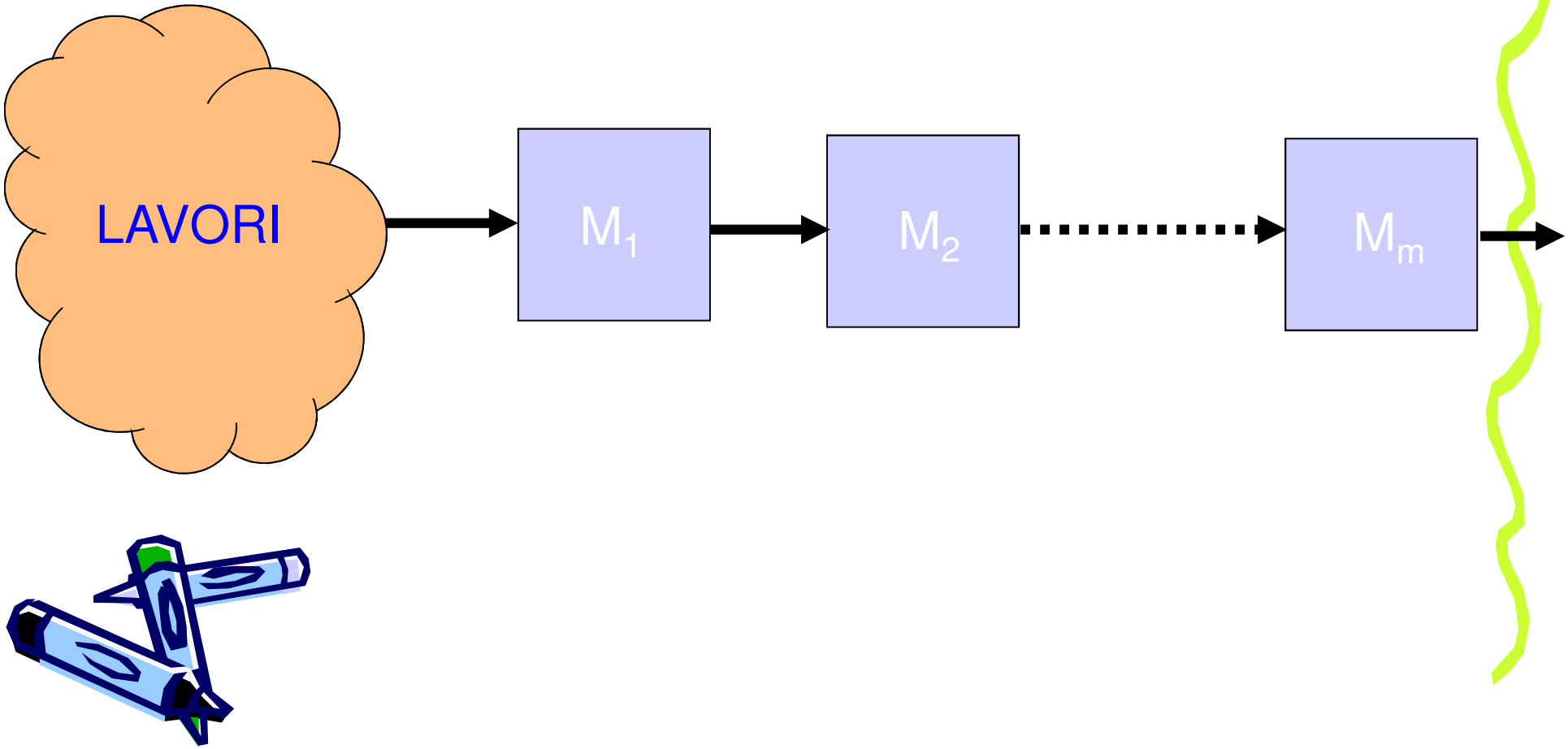
Macchina singola



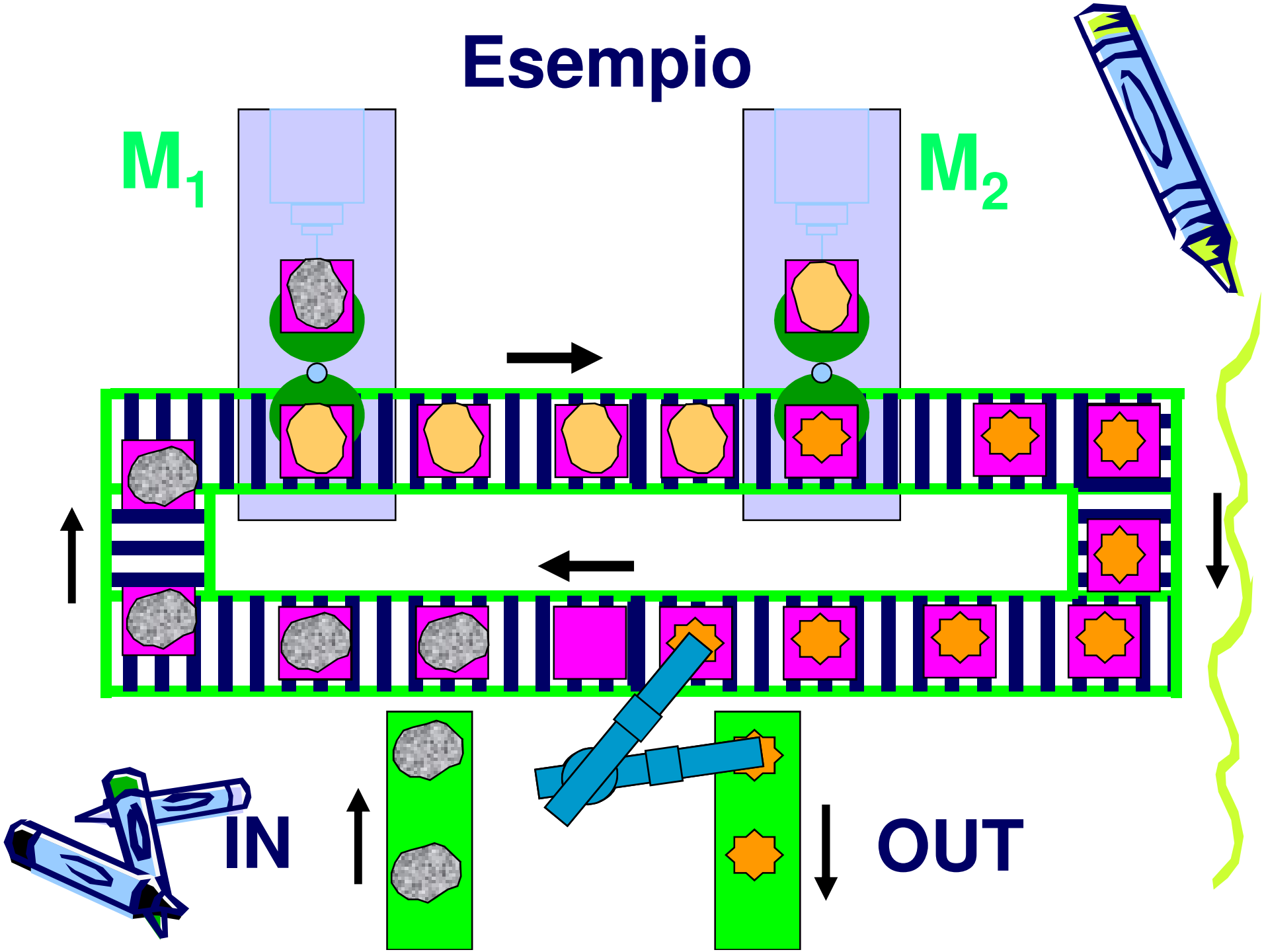
Macchine parallele



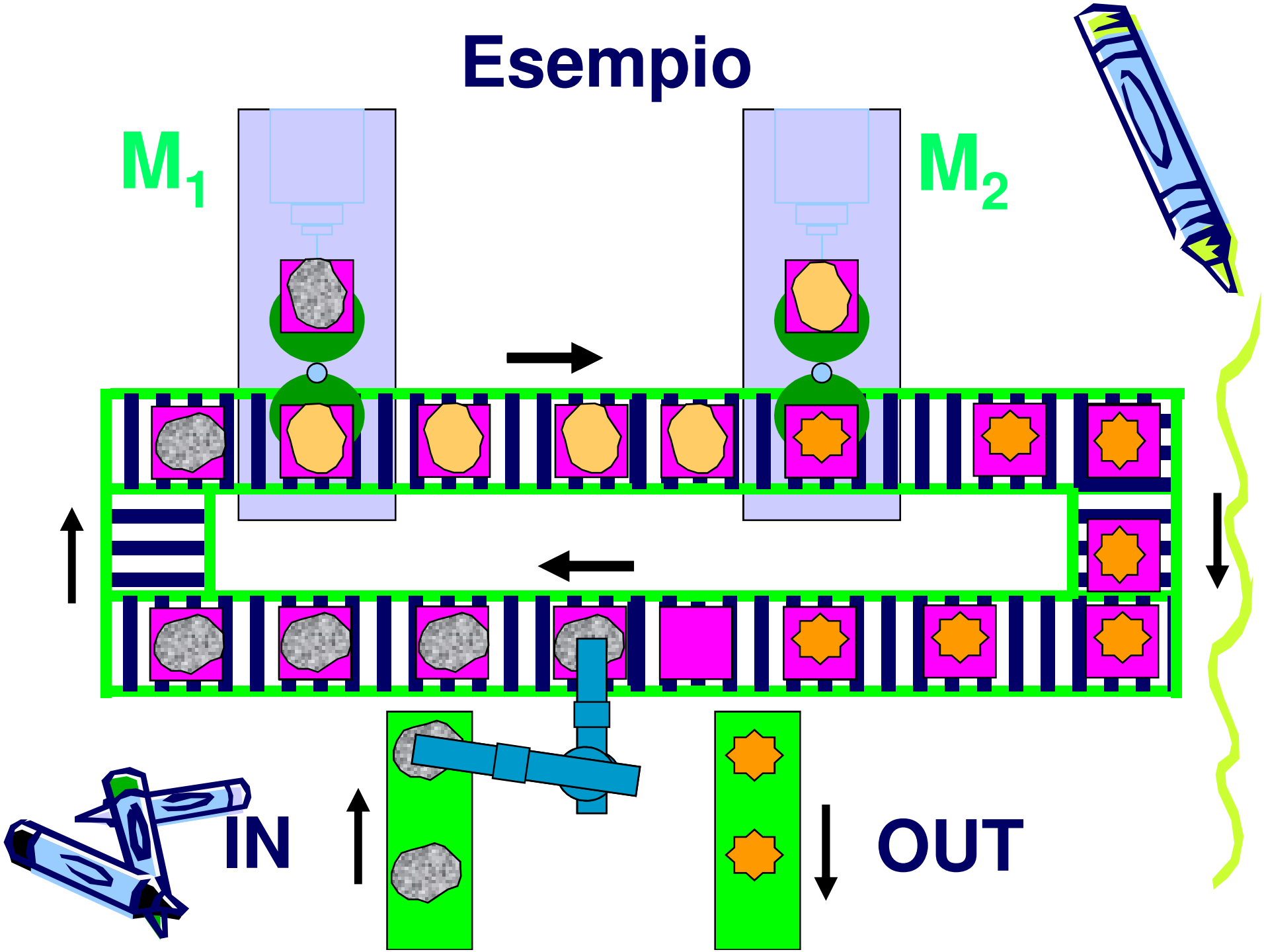
Macchine in linea
(Flow shop)



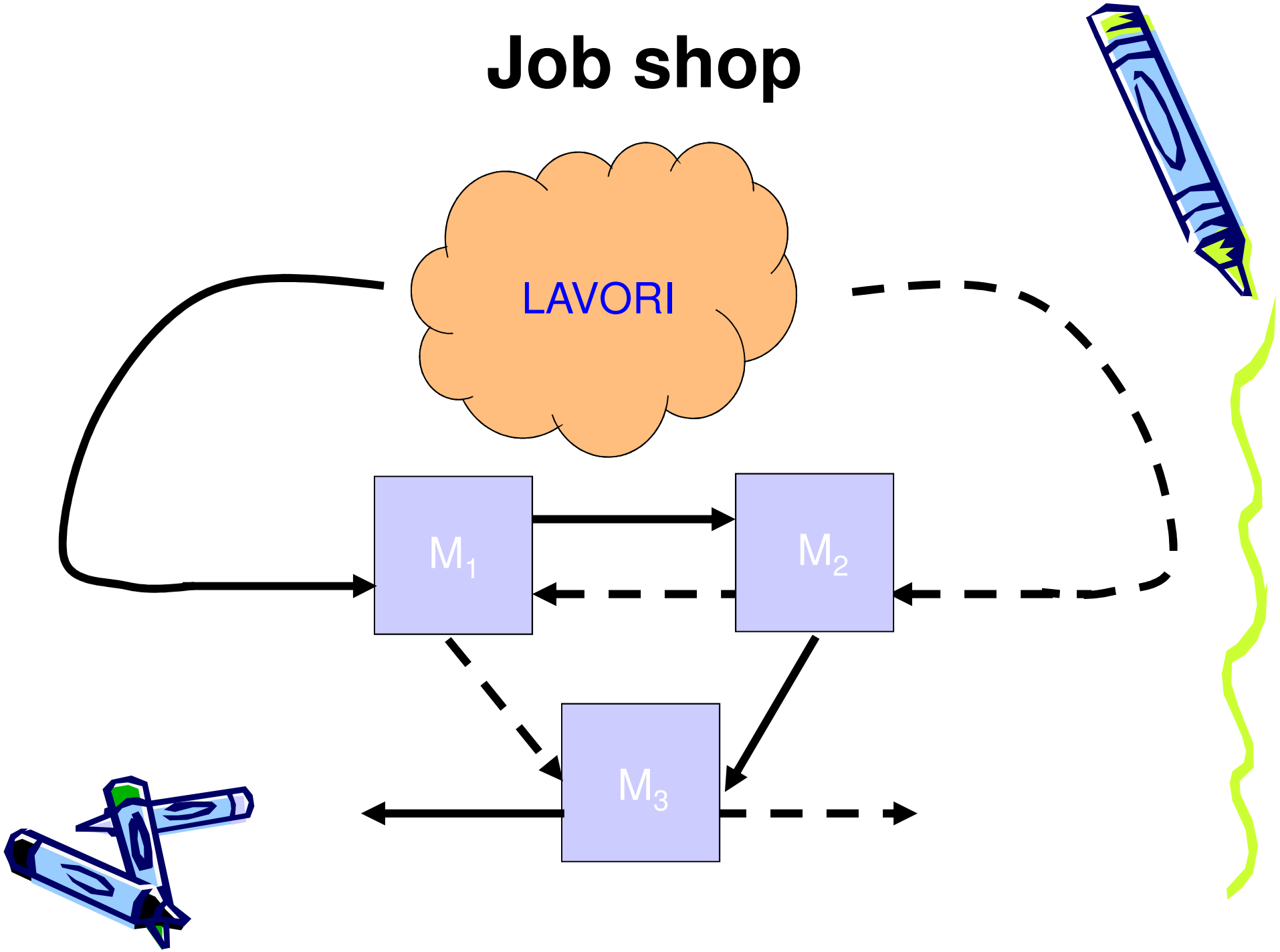
Esempio



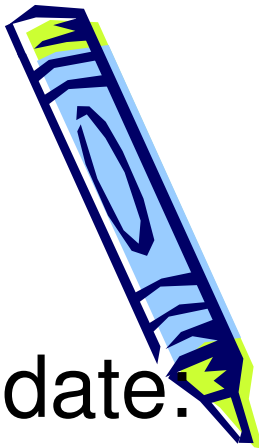
Esempio



Job shop



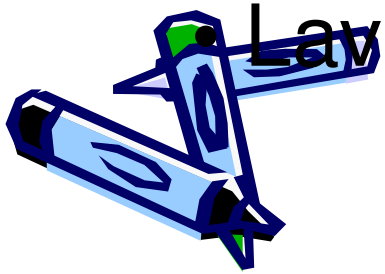
Misure di prestazione (lavori)



Dato il lavoro *i* con release date e duedate:

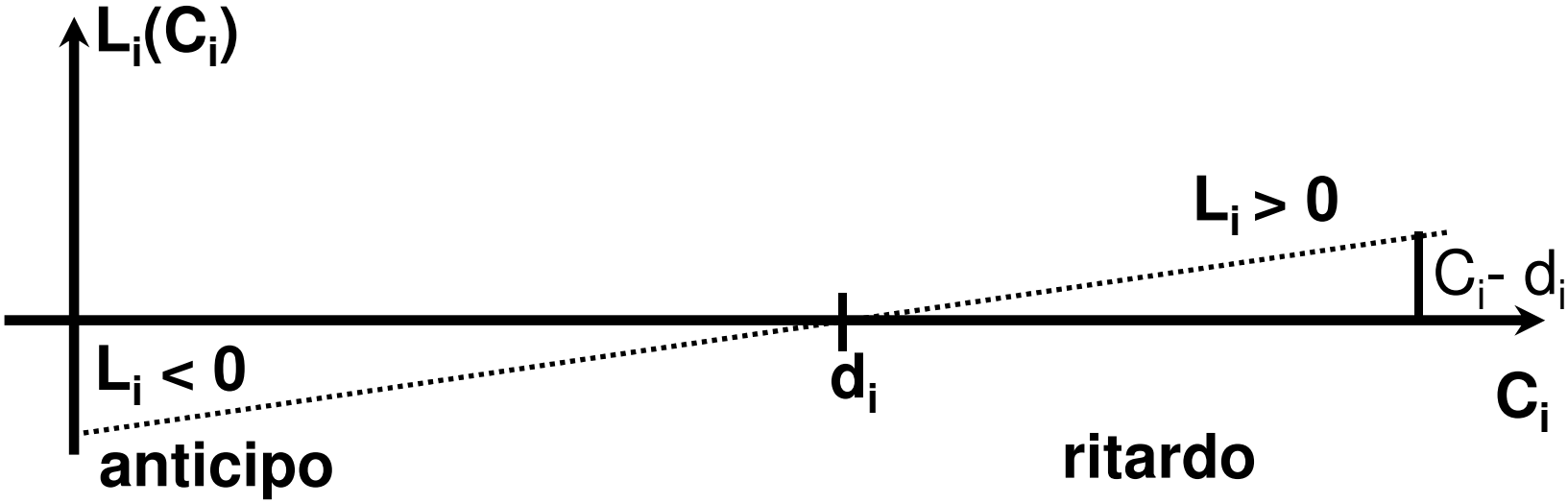
- tempo di completamento C_i
- tempo di attraversamento $F_i = C_i - r_i$
- *Lateness* $L_i = C_i - d_i$
- *Tardiness* $T_i = \max\{ 0, C_i - d_i \}$
- *Earliness* $E_i = \max\{ 0, d_i - C_i \}$
- Lavori in ritardo $U_i = 1$ se $C_i > d_i$

  $U_i = 0$ se $C_i \leq d_i$



Lateness (Ritardo)

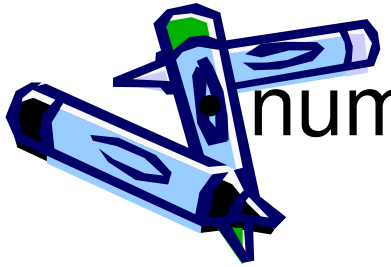
Ritardo del lavoro *i* : $L_i = C_i - d_i$



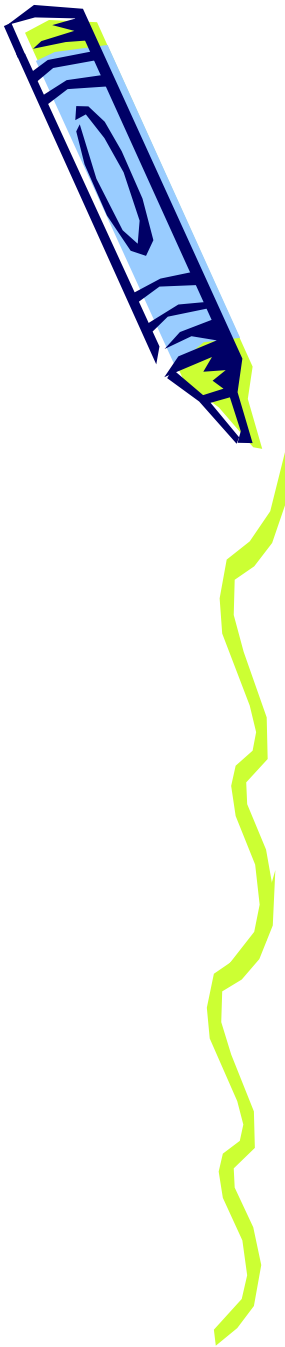
d_i: tempo di consegna (duedate) per il lavoro *i*

Misure di prestazione (sistema)

- somma dei tempi di completamento: $\Sigma_i C_i$
- *flow time* totale: $\Sigma_i F_i$
- massima *Lateness:* $L_{max} = \max_i L_i$
- massima *Tardiness:* $T_{max} = \max_i T_i$
- *Tardiness* totale pesata $\Sigma_i w_i T_i$
- makespan $C_{max} = \max_i C_i$
- numero di lavori in ritardo $\Sigma_i U_i$



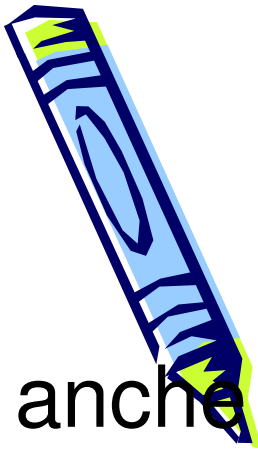
Misure di prestazione



Equivalenza tra misure

$$\sum_{i=1}^{n} L_i = \sum_{i=1}^{n} C_i - \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} F_i + \sum_{i=1}^{n} (r_i - d_i)$$

Misure di prestazione



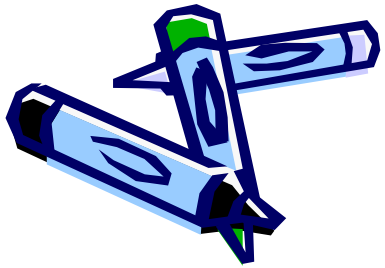
Una sol. che minimizza L_{max} minimizza anche T_{max} (ma, in generale, non è vero il viceversa):

$$T_{\max} = \max\{T_1,\ldots,T_n,0\} =$$
$$\max\{\max\{L_1,0\},\ldots,\max\{L_n,0\}\} =$$
$$= \max\{L_1,\ldots,L_n,0\} = \max\{L_{\max},0\}$$

Esempio

Lavori	1	2	3	4	5
p_i	8	16	10	7	2

Sequenza ottima (5, 4, 1, 3, 2)



Esempio

Aldo (A), Bruno (B), Carlo (C), Duilio (D) condividono un appartamento. Ogni mattino ricevono 4 giornali: Financial Times (FT), Guardian (G), Daily Express (DE), Sun (S).
Ciascun lettore inizia la lettura ad una certa ora, ha la propria sequenza fissata di lettura dei gionali e legge ciascun giornale per un tempo prefissato:

lettore	Ora inizio	Sequenza lettura (tempo in min.)			
Aldo	8.30	FT(60)	G (30)	DE (2)	S(5)
Bruno	8.45	G(75)	DE(3)	FT(25)	S(10)
Carlo	8.45	DE(5)	G(15)	FT(10)	S(30)
Duilio	9.30	S(90)	FT(1)	G(1)	DE(1)

Esempio

- tutti i lettori preferiscono (eventualmente) aspettare che un giornale (il prossimo nella propria sequenza) sia disponibile anziché modificare la sequenza prefissata.
- nessun lettore rilascia un giornale prima di averlo letto completamente.
- ciascun lettore termina la lettura di tutti i giornali prima di uscire.
- i quattro lettori attendono che tutti abbiano terminato di leggere prima di uscire di casa

Problema:

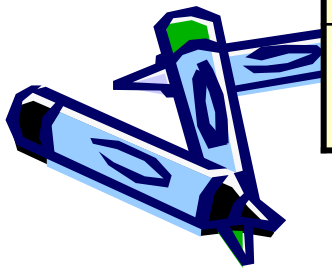
Qual'è la minima ora in cui A, B, C, D possono uscire?

Soluzioni

• Il problema equivale a: determinare in quale ordine ciascun giornale deve esser letto dai quattro lettori in modo da minimizzare il tempo di lettura totale.

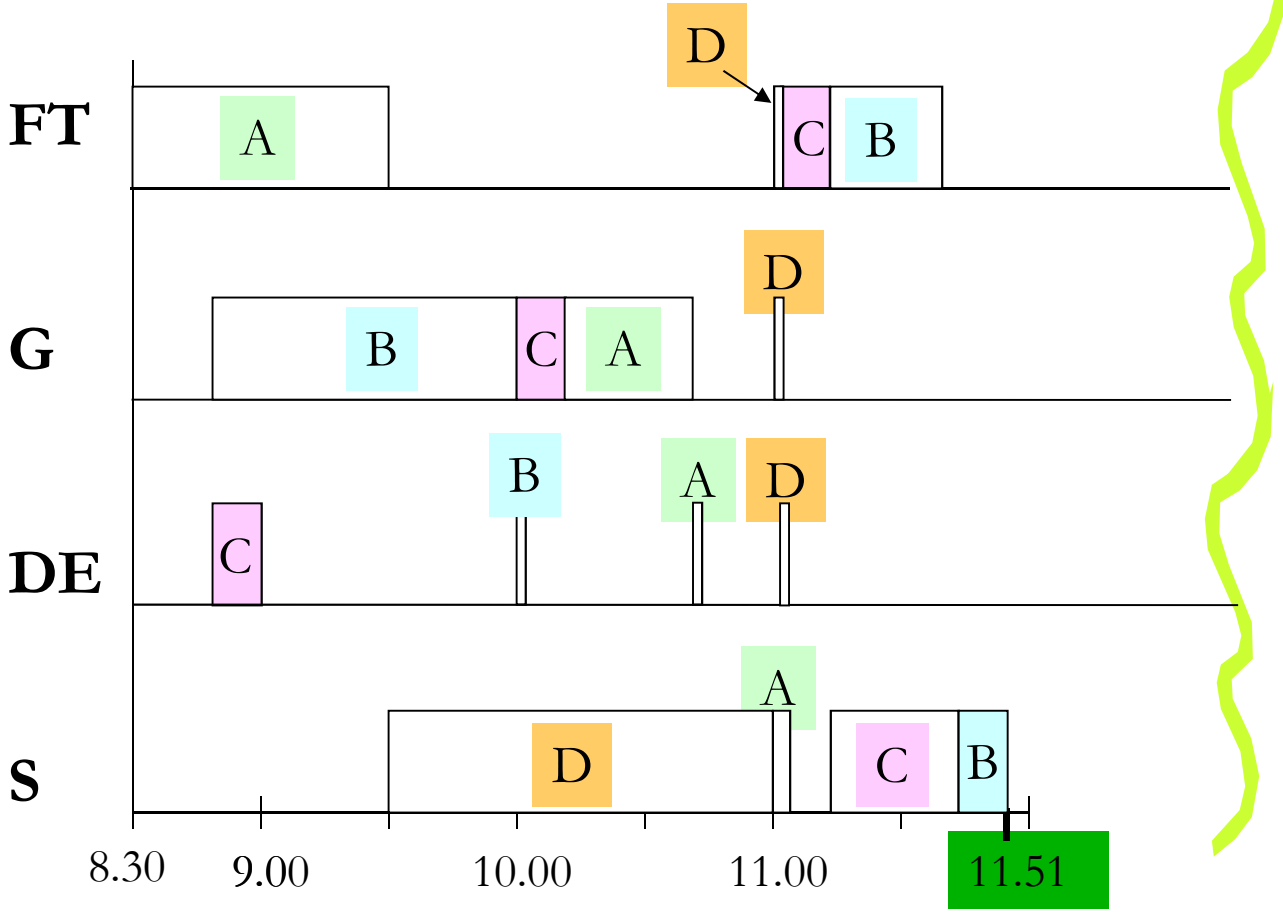
una possibile soluzione:

giornale	Lett. 1	Lett. 2	Lett. 3	Lett. 4
FT	A	D	C	B
G	B	C	A	D
DE	C	B	A	D
S	D	A	C	B



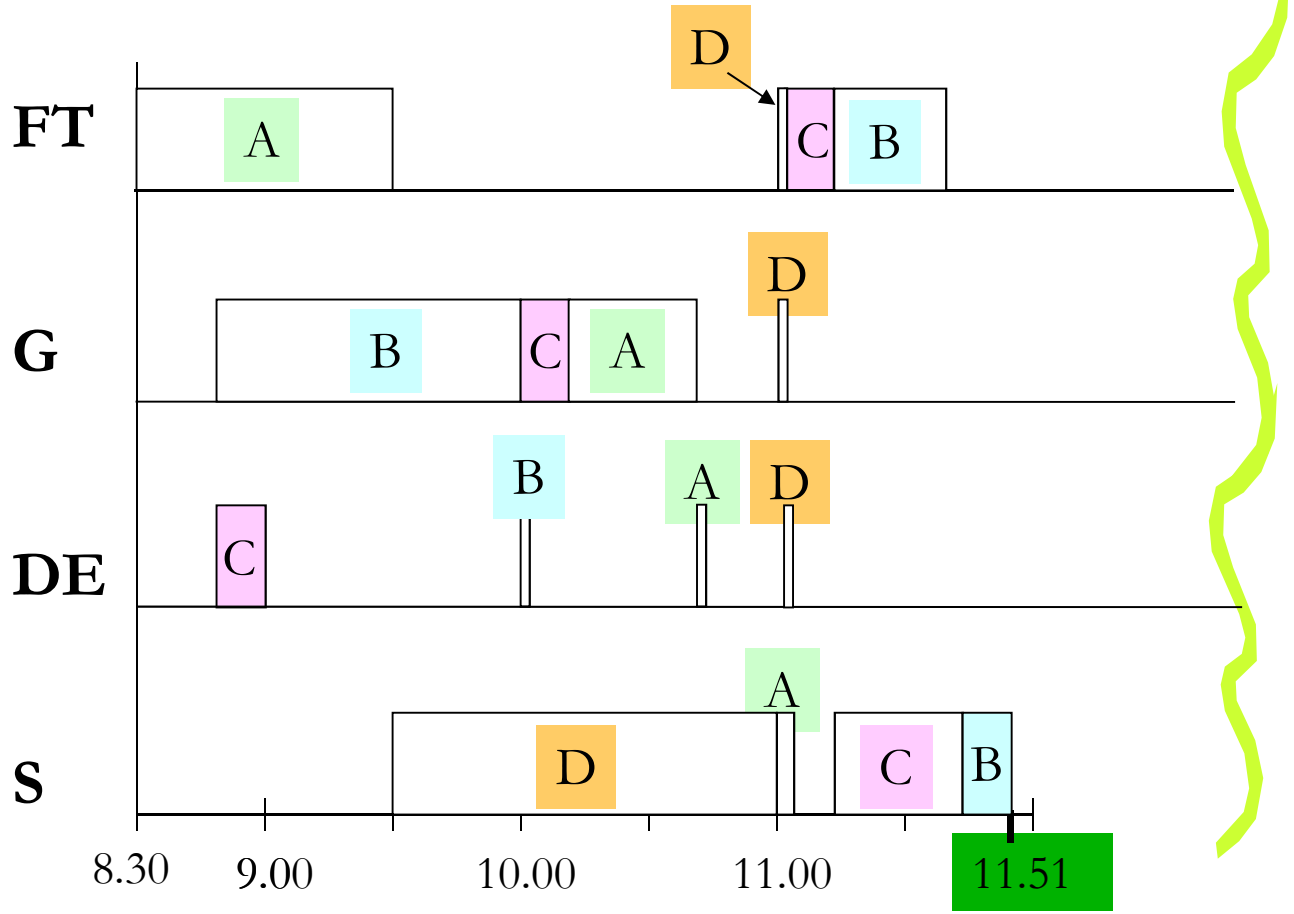
Giorn.	L1	L2	L 3	L4
FT	A	D	C	B
G	B	C	A	D
DE	C	B	A	D
S	D	A	C	B

Diagramma di Gantt

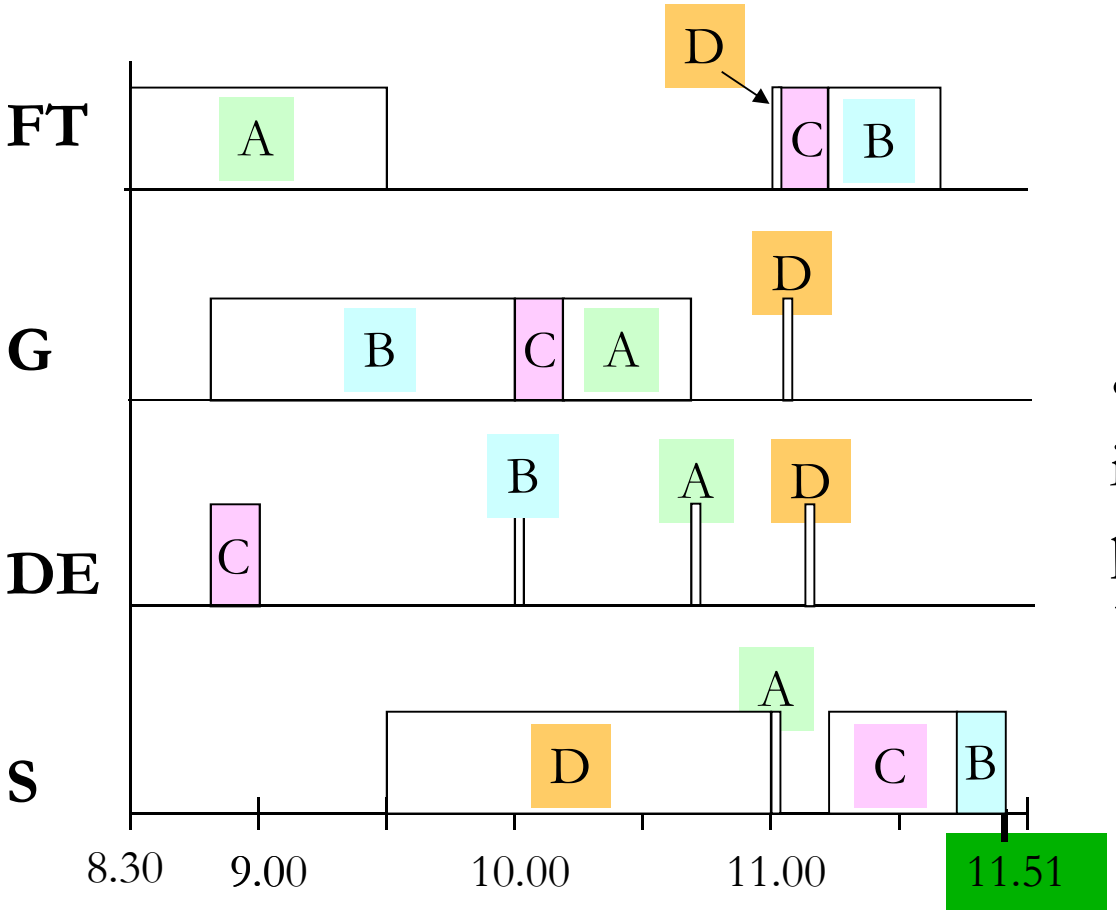


lettore	inizio	Sequenza lettura			
Aldo	8.30	FT	G	DE	S
Bruno	8.45	G	DE	FT	S
Carlo	8.45	DE	G	FT	S
Duilio	9.30	S	FT	G	DE

Vincoli tecnologici



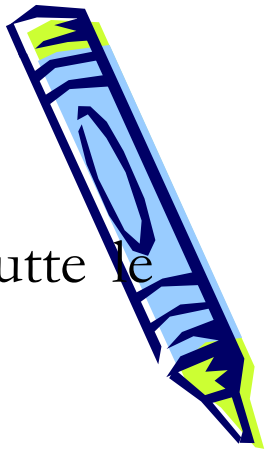
Commenti



• un lettore può essere inattivo (in *idle*) in certi periodi (C in [10.15, 11.01])

• ci può essere un giornale "fermo" anche se c'è un lettore disponibile che deve ancora leggerlo: B potrebbe avere **S** 11.05, ma questo violerebbe la sua sequenza; FT potrebbe avere C alle 9.30 ma viola la soluzione

Esercizi



• Determinare il valore del tempo di completamento di tutte le letture per la seguente soluzione:

giornale	Lett. 1	Lett. 2	Lett. 3	Lett. 4
FT	D	B	A	C
G	D	C	B	A
DE	D	B	C	A
S	A	D	C	B

• Determinare uno schedule migliore di quello proposto in precedenza

Enumerazione totale

- Il numero di soluzioni è $(4!)^4 = 331.776$
- per un problema con n lettori e m giornali diventa $(n!)^m$.
- esaminando 10.000 soluzioni al secondo (4 giornali e n lettori):

n	tempo
5	354 min
10	$2 \cdot 10^{17}$ giorni

Elementi di un problema di scheduling

- ***job***: attività da svolgere

  Fotocopia, esecuzione di un programma di calcolo, lettura di un giornale

Un job può rappresentare una singola attività o un insieme di attività (*task*) tecnologicamente legate

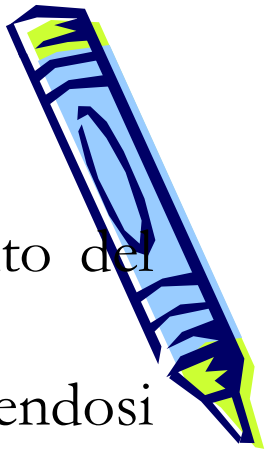
- ***macchine***: risorse che eseguono le attività

  Fotocopiatrice, CPU, giornali

L'esecuzione di un'attività su una macchina è detta ***operazione.***

I problemi di scheduling si classificano in base alle caratteristiche dei task e all'architettura delle macchine

Attributi dei job



- ***tempo di processamento*** p_{ij} : durata del processamento del job *j* sulla macchina *i*
- ***release date*** r_j : tempo in cui *j* arriva nel sistema, rendendosi disponibile al processamento
- ***due date*** d_j: tempo entro il quale si desidera che il job *j* sia completato (data di consegna)
- ***peso*** w_j: indica l'importanza del job j

Notazione a tre campi: $\alpha/\beta/\gamma$:

α descrive il sistema di macchine

β rappresenta vincoli e modalità di processamento (0, 1 o più componenti)

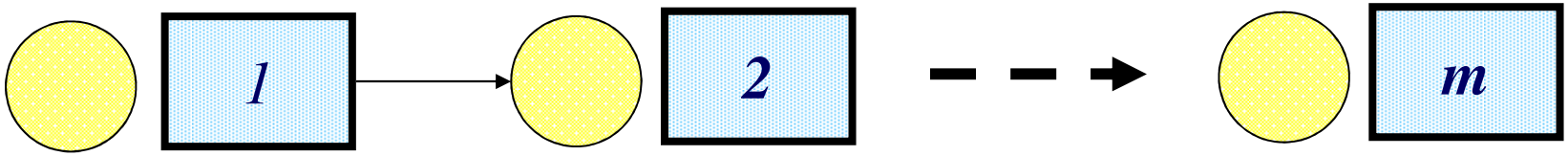
γ indica l'obiettivo

Campo α

• *Macchina singola (1)* : ciascun job richiede una singola operazione da eseguirsi sull'unica macchina disponibile

• *Macchine identiche parallele (P_m)* : ciascun job richiede una singola operazione da eseguirsi su una qualunque delle *m* macchine identiche.

• *Flow shop (F_m)* : *m* macchine in serie. Ciascun job deve essere processato su ciascuna di esse (*m* task). Tutti i job hanno lo stesso *routing*. Una volta terminata un'operazione il job viene posto nel buffer della macchina successiva



• *Job shop (J_m)* : ciascun job deve essere processato da un sottoinsieme di un insieme di *m* macchine, secondo un proprio routing.

Campo β

• *Release dates* (r_j) : il job j non può iniziare il processamento prima dell'istante r_j

• *Tempi di set-up* (s_{jk}) : tempo richiesto per il riattrezzaggio delle macchine fra i job j e k. Se dipende dalla macchina i, si esprime con s^i_{jk}

• *Preemption (prmp)* : è ammesso interrompere un'operazione su una macchina prima del suo completamento per iniziare una nuova operazione. Il processamento eseguito prima dell'interruzione non va perso: quando l'operazione viene ripresa, essa richiede solo il tempo *rimanente* di processamento

• *vincoli di precedenza (prec)* : un job può essere processato solo se tutti i job di un certo insieme sono stati completati. Grafo delle precedenze. Se ciascun job ha al più un predecessore e un successore β = *chain*.

• *breakdown (brkdwn)* : le macchine non sono sempre disponibili, ma hanno periodi fissati di interruzione del servizio

Campo γ

Definizioni.

C_j tempo di completamento del job j

$L_j = C_j - d_j$ ***lateness*** del job j

$T_j = max(L_j, 0)$ ***tardiness*** del job j

$U_j = 1$ se $C_j > d_j$ (*tardy* job) e 0 altrimenti

- ***Makespan*** (C_{max}) : max (C_1, …, C_n) tempo di completamento dell'ultimo job
- ***Massima Lateness*** (L_{max})
- ***Tempo totale pesato di completamento*** ($\Sigma w_j C_j$)
- ***Tardiness totale pesata*** ($\Sigma w_j T_j$)
- ***Numero di tardy job*** (ΣU_j)

Definizioni di base

- Una funzione obiettivo si dice **regolare** se è non-decrescente in $C_1, \ldots, C_n$

- Una permutazione dei job che definisce l'ordine con cui i job sono processati su una certa macchina si dice **sequenza**

- Una soluzione di un problema di scheduling si dice **schedule**. Consiste nell' assegnamento di un istante di inizio a ciascuno dei task di ogni job (se *prmp*, istanti di inizio di ciascuna delle parti in cui viene suddiviso un task)

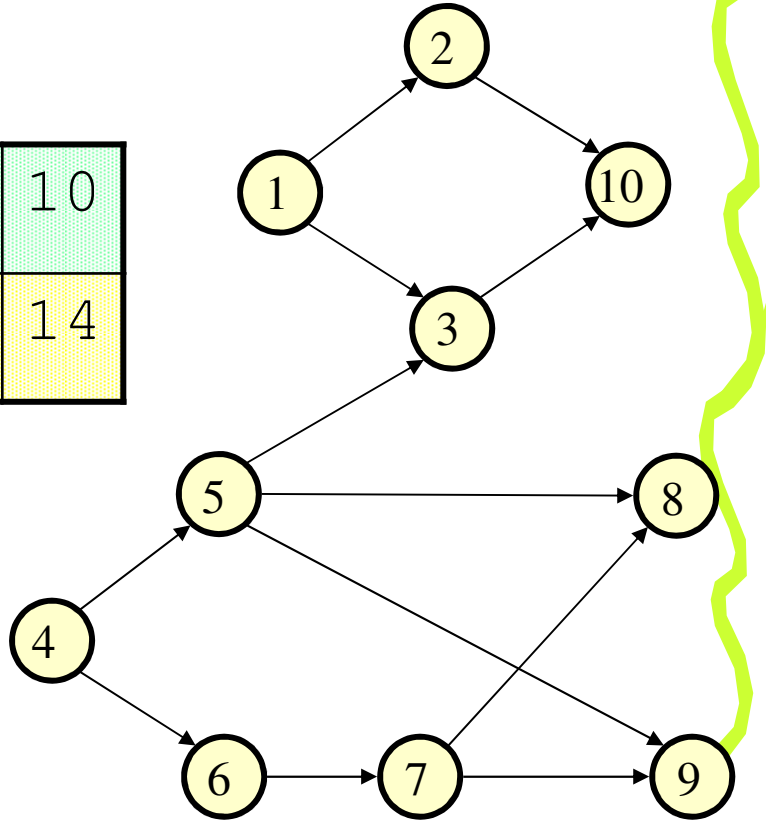
- uno schedule ammissibile è detto **nondelay** se nessuna macchina è ferma quando esiste un'operazione disponibile al processamento

Esempio: fermi macchina non forzati

• Richiedere che lo schedule sia privo di fermi macchina non forzati può portare a comportamenti controintuitivi.

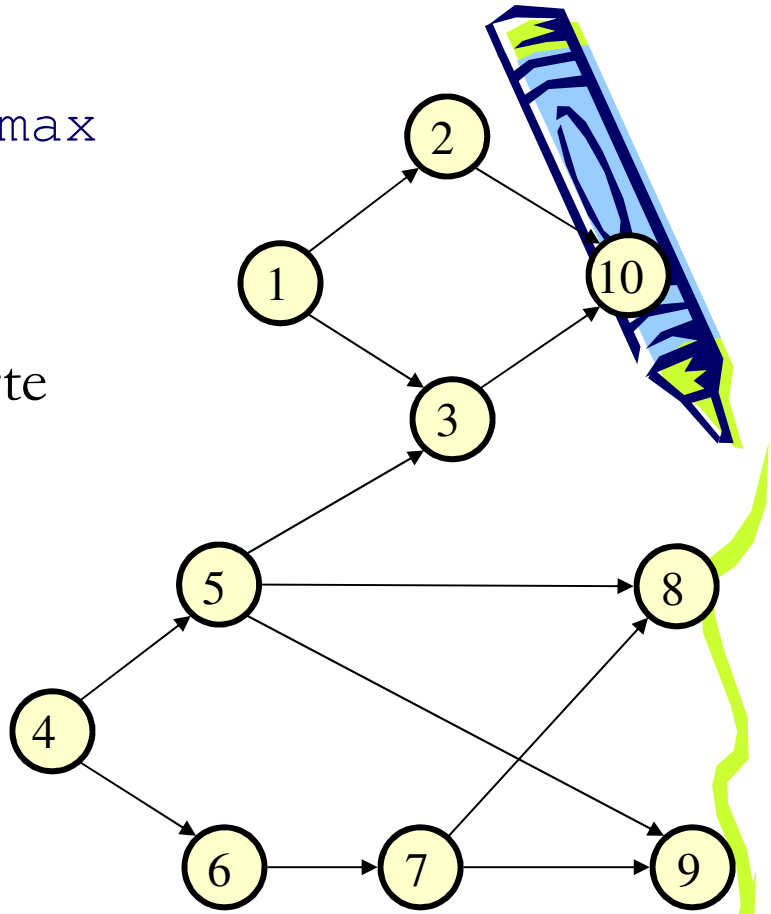
Esempio: $P_2/prec/C_{max}$

job	1	2	3	4	5	6	7	8	9	10
p_j	7	6	6	1	2	1	1	7	7	14

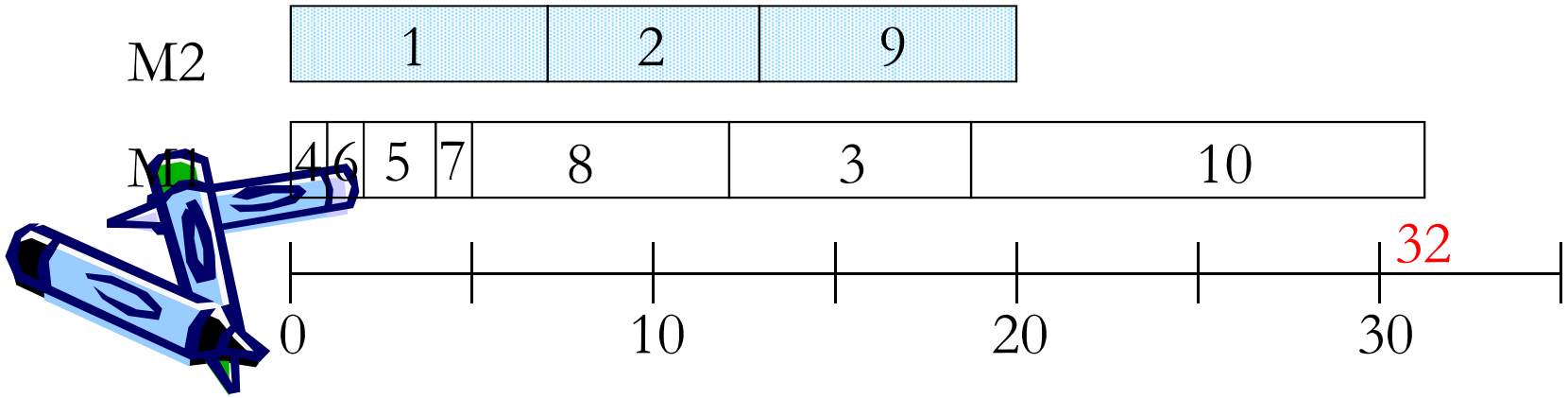


$P_2/prec/C_{max}$

Regola di buon senso: mandare una parte processabile sulla prima macchina libera.

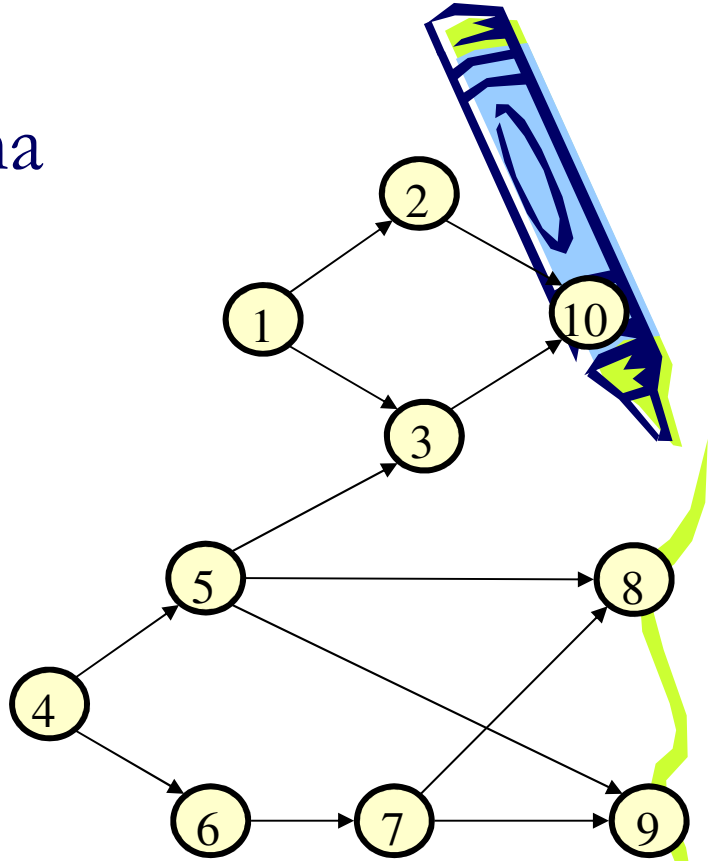


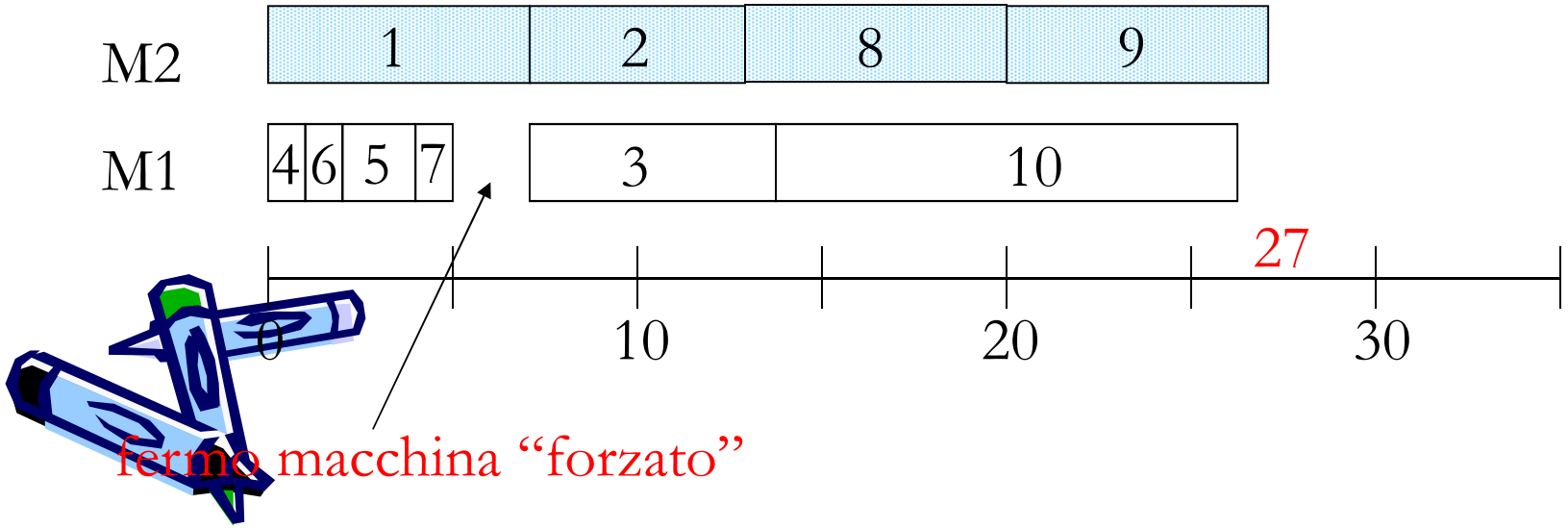
Soluzione:



Soluzione ottima

• Nonostante l'idea di non fermare le macchine se non necessario sia ragionevole, la seguente soluzione migliora quella ottenuta dalla regola precedente:





Scheduling su singola macchina

Descrizione del problema

Un insieme di *n* operazioni deve essere eseguito su una macchina

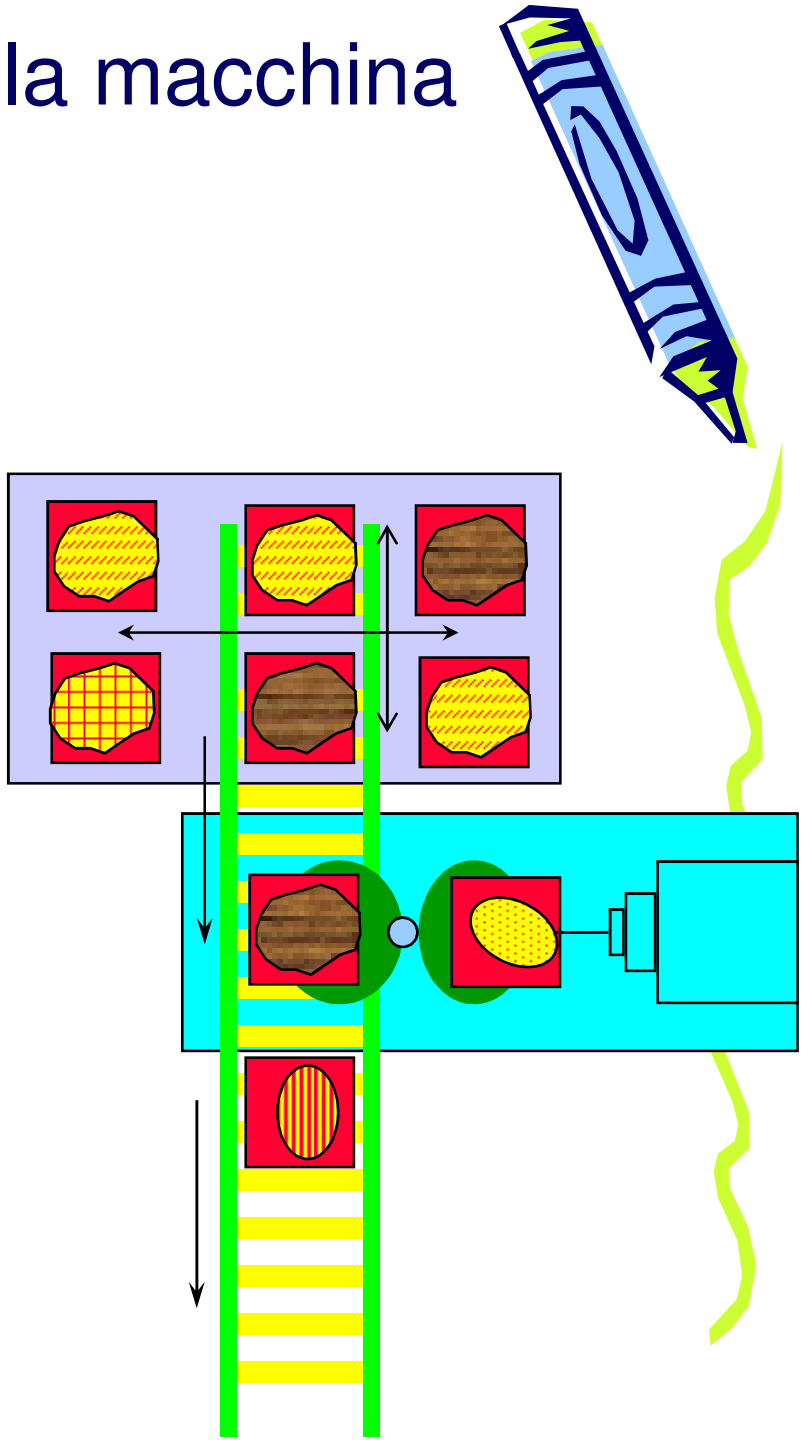
Dati

I tempi di *processamento* p_i, i=1,…,n, del lavoro i sulla macchina sono noti.

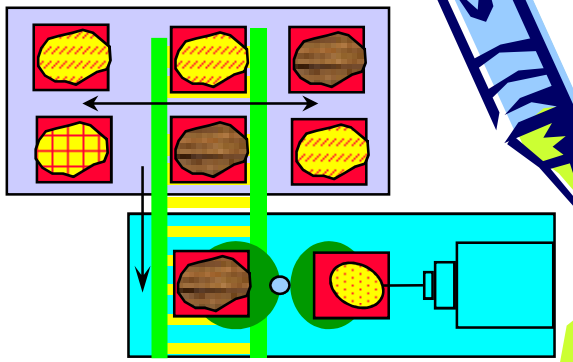
Obiettivo

Sequenziare le operazioni sulla macchina in modo da minimizzare la somma dei tempi di completamento.

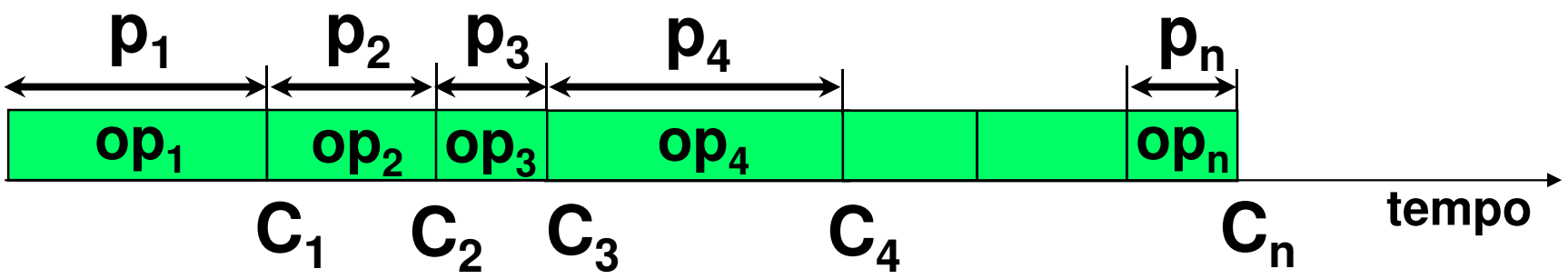
$$\min \Sigma_i C_i$$



Gantt del Sequenziamento



Sequenza S



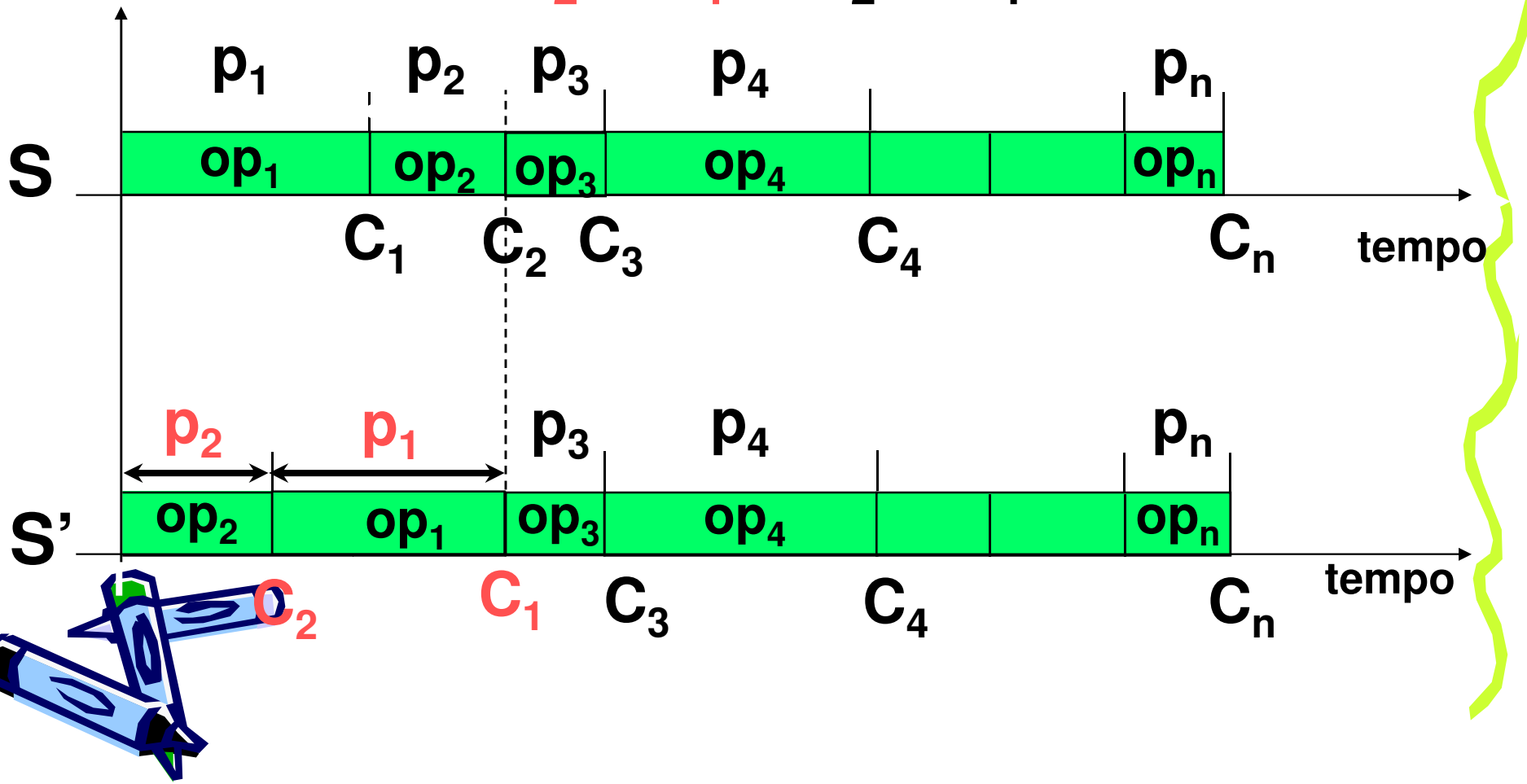
$C_n=\Sigma_i p_i$: tempo di completamento totale (*makespan*)

Obiettivo: min $\Sigma_i C_i$

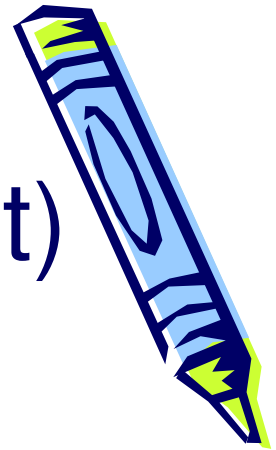
se $p_2 < p_1$ allora scambiando le op. *1* e *2* si ha

$C_2 < C_1$ e $C_1 = C_2$

$C_2 + C_1 < C_2 + C_1$



Regola *SPT* (shortest processing time first)

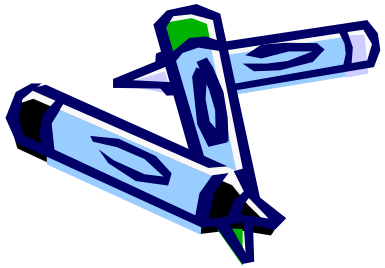


SPT: sequenzia prima le operazioni che hanno tempo di esecuzione più piccolo

Consente di minimizzare la somma dei tempi di completamento $\Sigma_i C_i$ di n operazioni (lavori) su una macchina

1.1 Tempo totale pesato di completamento

$$\Sigma w_j C_j$$



Tempo totale pesato di completamento

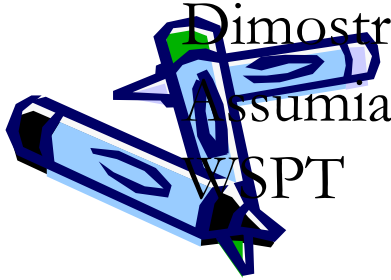
$1//\ \Sigma w_j C_j$

Definiamo WSPT una regola che ordina i job per valori non crescenti del rapporto w_j/p_j

Sussiste il seguente

Teorema 1.1 La regola WSPT calcola una soluzione ottima del problema $1//\ \Sigma w_j C_j$

Dimostrazione. (Per contraddizione)
Assumiamo che S sia uno schedule ottimo e che non rispetti WSPT

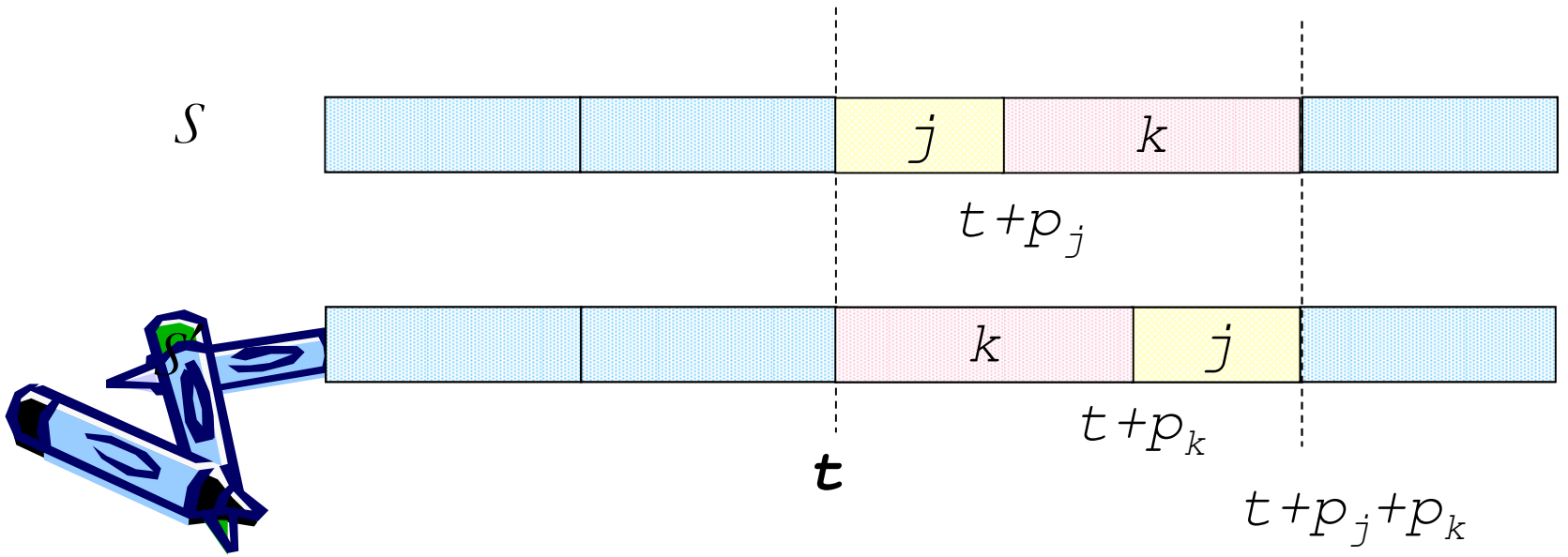


Dimostrazione

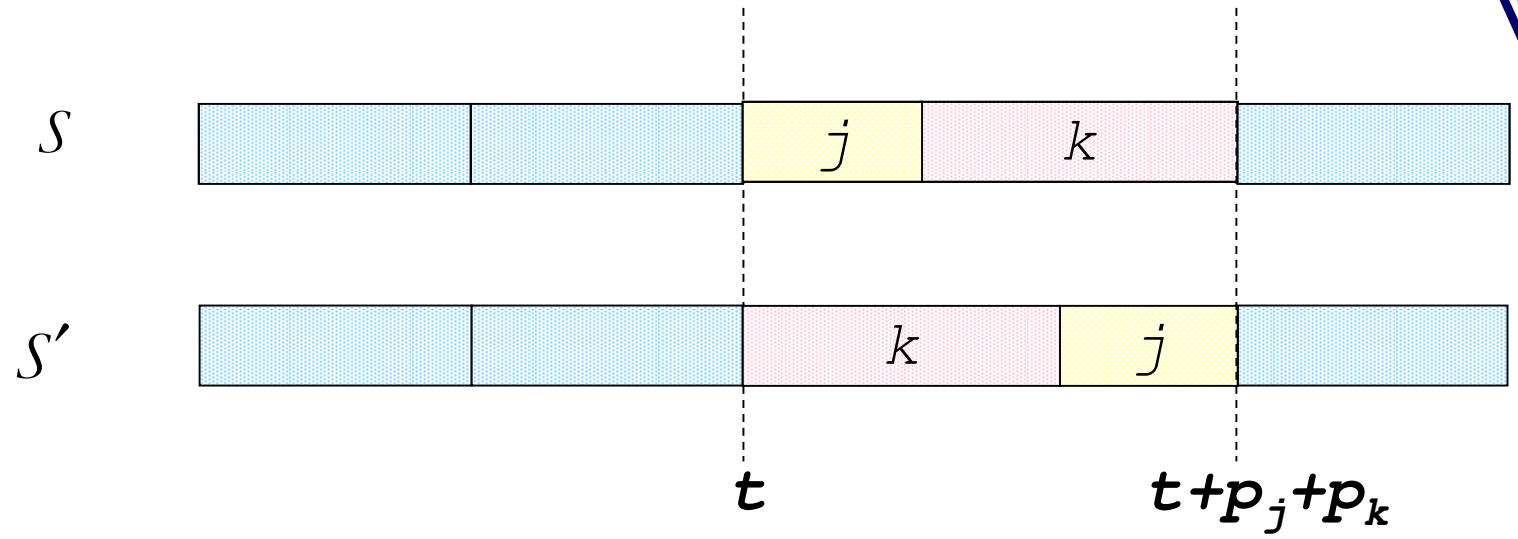
Allora devono esserci in S due job adiacenti, diciamo j seguito da k tali che

$$w_j/p_j \; < \; w_k/p_k$$

Assumiamo che il job j inizi all'istante t.
Eseguiamo uno scambio dei job j e k ottenendo un nuovo schedule S'



Dimostrazione



Il tempo totale (pesato) di completamento dei job che precedono e seguono la coppia (j,k) non è modificato dallo scambio. Il contributo dei job j e k in S è:

$$(t+p_j)\,w_j \;+\; (t+p_j+p_k)\,w_k$$

mentre in S' è:

$$(t+p_k)\,w_k \;+\; (t+p_k+p_j)\,w_j$$

Dimostrazione

in S] $(t+p_j)\,w_j + (t+p_j+p_k)\,w_k$

in S'] $(t+p_k)\,w_k + (t+p_k+p_j)\,w_j$

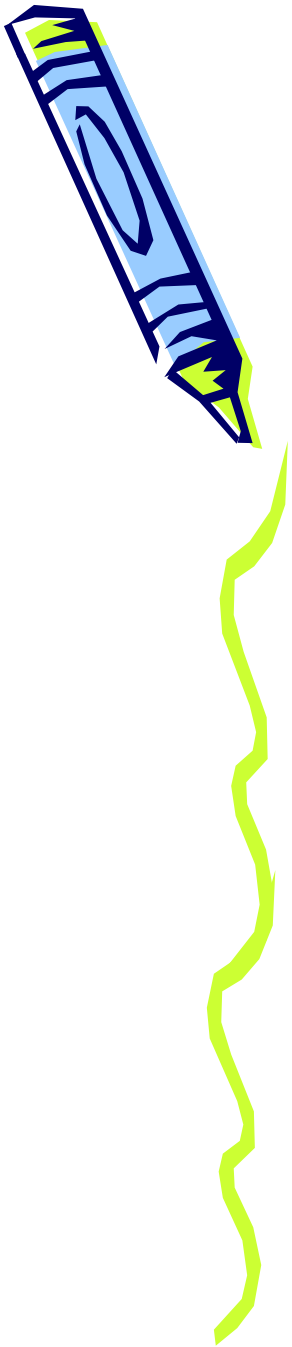
Eliminando i termini uguali:

in S] $p_j w_k$

in S'] $p_k w_j$

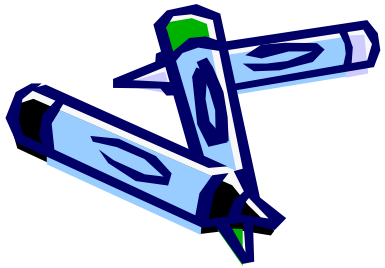
Quindi, se $w_j/p_j < w_k/p_k$, risulta $p_k\ w_j < p_j\ w_k$, cioè il tempo totale pesato in S' è strettamente inferiore a quello in S, contraddizione

□



Numero di tardy job

$1//\Sigma U_j$



Struttura delle soluzioni ottime

• ogni schedule ottimo è composto da un insieme di job che arrivano in tempo e da un secondo insieme di job in ritardo (rispetto alle proprie due date)

• il primo insieme rispetta EDD, in modo che L_{max} sia minimizzata (e minore o uguale a zero).

• l'ordine dei job del secondo insieme non ha impatto sulla funzione obiettivo

Algoritmo di *Moore*

- Costruisce lo schedule a partire dall'inizio
- J job già schedulati (schedule parziale, secondo EDD)
- J^d job già esaminati e fissati come tardy job
- J^c job ancora non esaminati

Inizializzazione:

$J = \emptyset$, $J^d = \emptyset$, $J^c = \{1, \ldots, n\}$.

Repeat:

1. Schedula il job j^* in J^c con la minima due date:
   aggiorna $J^c := J^c \setminus \{j^*\}$

2. Se j^* è in ritardo, cioè $\sum_{j \in J} p_j > d_{j^*}$

   elimina dallo schedule parziale il job più lungo, sia k^*.
   Aggiorna $J^d := J^d \cup \{k^*\}$

Until ($J^c \neq \emptyset$)

Correttezza

Teorema. L'algoritmo di Moore restituisce una soluzione ottima di $1//\Sigma U_j$

Dimostrazione. Senza perdita di generalità si può assumere $d_1 \leq d_2 \leq \ldots \leq d_n$ Sia J_k un sottoinsieme di job che soddisfa le seguenti proprietà:

1. ha il massimo numero N_k di job in tempo fra quelli di $\{1, \ldots, k\}$
2. fra tutti i sottoinsiemi di $\{1, \ldots, k\}$ con N_k job in tempo, J_k ha il minimo tempo di processamento

Per definizione, J_n corrisponde ad uno schedule ottimo.
Dimostriamo **per induzione** che l'algoritmo di Moore restituisce J_n

Induzione

- l'algoritmo costruisce J_1 in modo da rispettare (1) e (2).
- Assumiamo che (1) e (2) valgano per k, cioè che Moore costruisca J_k e mostriamo che esse valgono per $k+1$

Sussistono due casi:

Caso a. Il job k+1 è aggiunto all'insieme J_k ed è completato in tempo. Ovviamente, è impossibile avere un maggior numero di job in tempo fra quelli in $\{1,\ldots,k+1\}$, quindi vale (1).

Inoltre, il job $k+1$ deve appartenere all'insieme. Quindi, il tempo di processamento totale è minimo fra tutti gli insiemi che soddisfano (1). Quindi vale (2).

Induzione

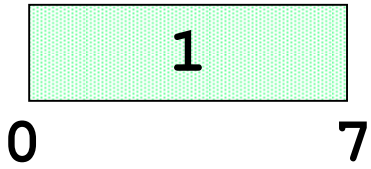
Caso b. Il job k+1 è aggiunto all'insieme J_k ed è completato in ritardo. Dato che J_k soddisfa (1) e (2) deve essere $N_k = N_{k+1}$.

Quindi, aggiungere il job k+1 a J_k non aumenta il numero di job in tempo. Ma agguingere k+1 e cancellare il job più lungo fra quelli di $J_k \cup \{k+1\}$ mantiene uguale il numero di job in tempo e riduce il tempo di processamento complessivo. Questo completa la prova.

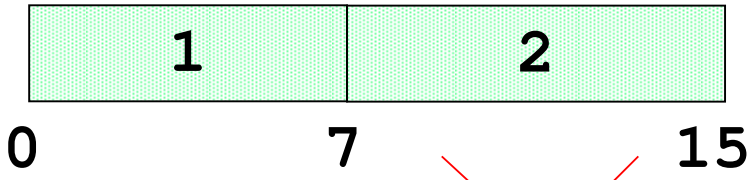
Esempio

job	1	2	3	4	5
p_j	7	8	4	6	6
d_j	9	17	18	19	21

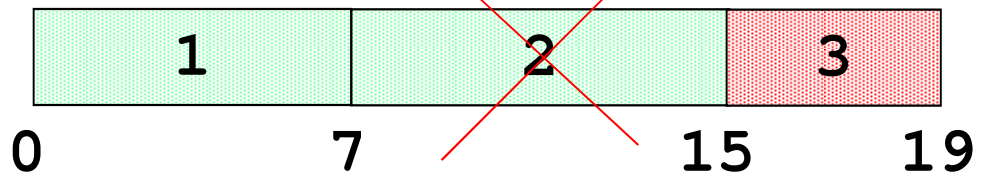
iter 1.

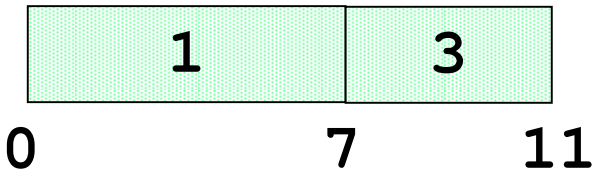


iter 2.



iter 3.

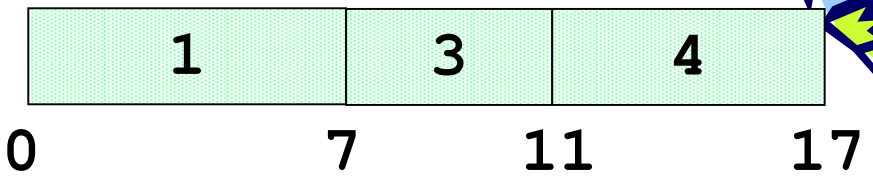




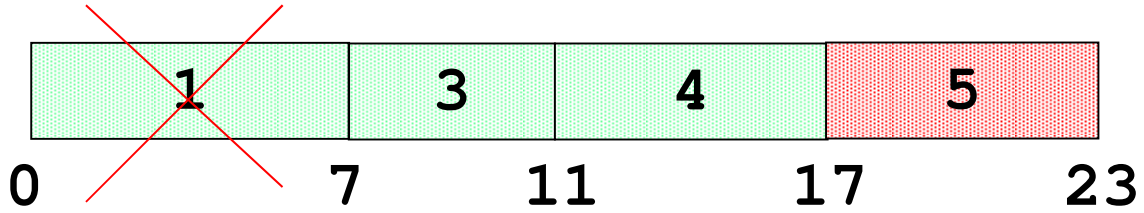
Esempio

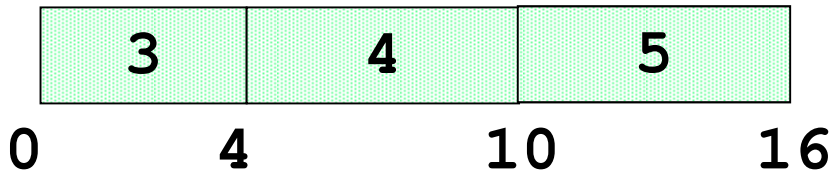
job	1	2	3	4	5
p_j	7	8	4	6	6
d_j	9	17	18	19	21

iter 4.

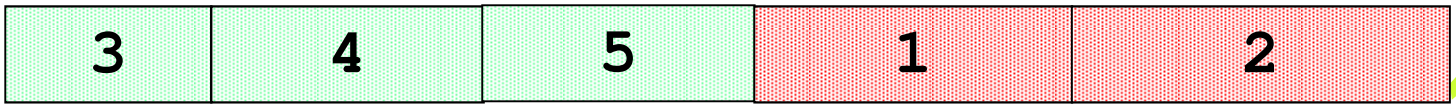


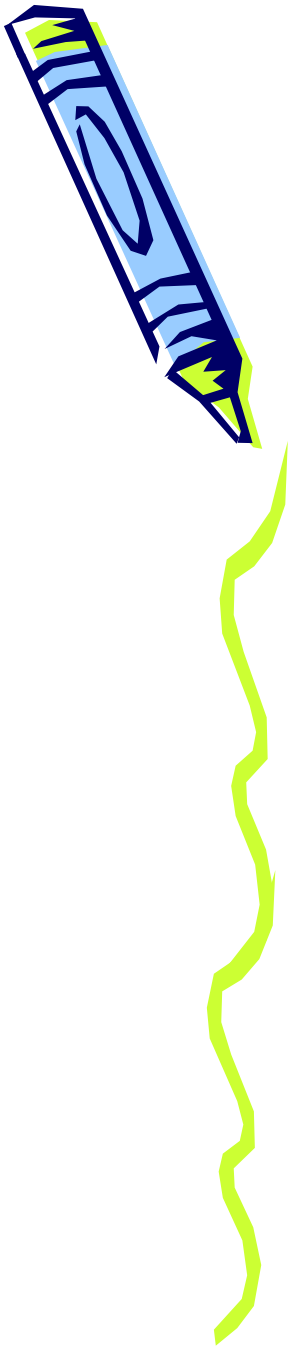
iter 5.





Soluzione





Massima Lateness

L_{max}

$1/prec/h_{max}$

Forma della funzione obiettivo:

$$h_{max} = max\ (h_1(C_1),\ \ldots,\ h_n(C_n))$$

In cui $h_j(C_j)$ è una arbitraria funzione non decrescente di C_j

Esempi:

$h_j(C_j) = C_j - d_j = L_j \Rightarrow h_{max} = L_{max}$

$h_j(C_j) = max\ (0, L_j) = T_j \Rightarrow h_{max} = T_{max}$

Precedenza: G_P generico grafo aciclico.

- Dato che le soluzioni ottime sono nondelay, l'ultimo job termina all'istante $C_{max} = \sum_{j=1}^{n} p_j$

Algoritmo di *Lawler*

- Costruisce lo schedule a partire dal fondo
- J insieme dei job già schedulati nell'intervallo $[C_{max} - \sum_{j \in J} p_j, C_{max}]$
- $J^c = \{1, \ldots, n\} - J$
- J' insieme dei job <u>schedulabili</u> immediatamente prima di J (= tutti i successori sono in J)

Inizializzazione:
$J = \emptyset$, $J^c = \{1, \ldots, n\}$, J' insieme dei job privi di successori

Loop: while ($J^c \neq \emptyset$)
Schedula in ultima posizione il job j^* tale che

$$h_{j^*}\left(\sum_{j \in J^c} p_j\right) = \min_{j \in J'} \left(h_j\left(\sum_{j \in J^c} p_j\right)\right)$$

$J := J \cup \{j^*\}$, $J^c := J^c \setminus \{j^*\}$, aggiorna J'
Endloop.

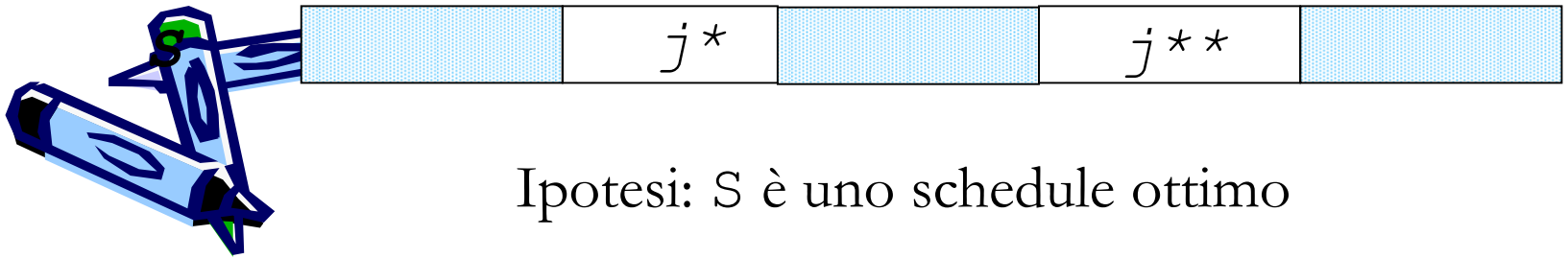
Corretezza

Teorema. L'algoritmo di Lawler restituisce uno schedule ottimo per $1/prec/h_{max}$

Dimostrazione. Per contraddizione: ad una generica iterazione è selezionato il job $j^{**} \in J$ che non ha il minimo costo di completamento

$$h_{j^*}\left(\sum_{j \in J^c} p_j\right)$$

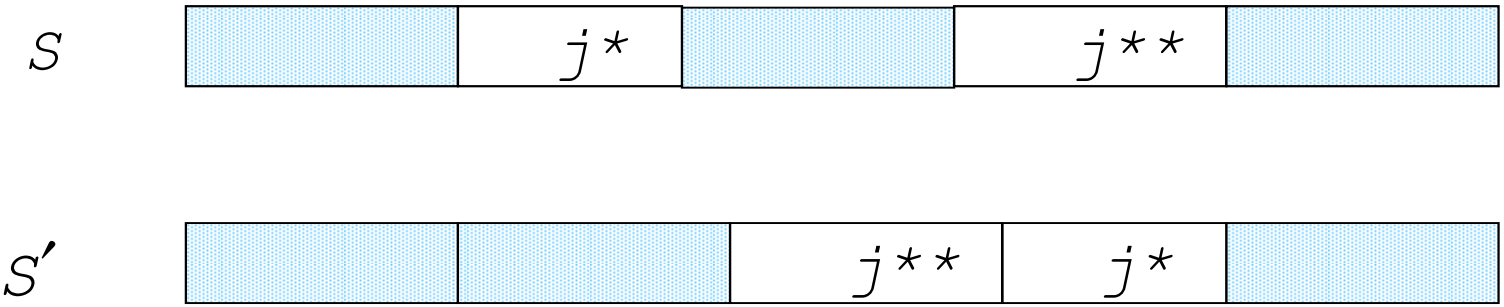
Quindi, j^* è schedulato prima di j^{**}



Ipotesi: S è uno schedule ottimo

Dimostrazione (continua)

Consideriamo un nuovo schedule S' ottenuto da S spostando il job j^{**} subito dopo j^{*}.



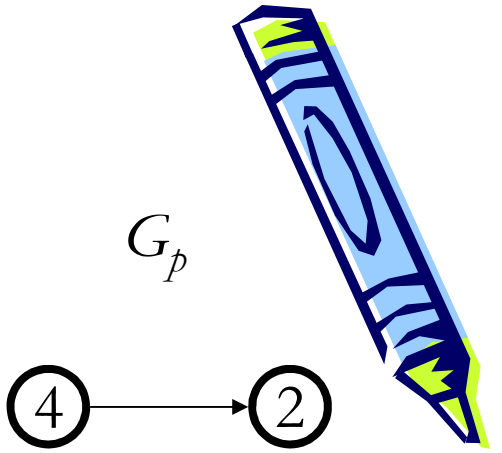
- L'unico job che peggiora il suo costo di completamento è j^{*}.

- Tuttavia, il suo costo di completamento in S' è, per definizione, non superiore al costo di j^{**} in S. Quindi S non è ottimo.

□

Esempio

job	1	2	3	4
p_j	3	4	2	6
h_j	10	$1+C_2$	$1.5\ C_3$	$2^{(C4/5)}$



Iter 1.

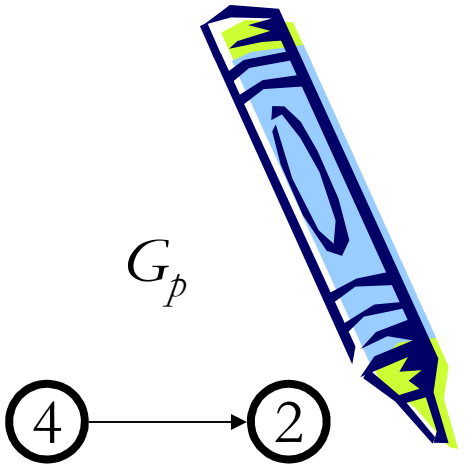
$$\sum_{j \in J^c} p_j = 15$$

J	$\varnothing$
J^c	{1,2,3,4}
J'	{1,2,3}

job	1	2	3
costo	10	16	22.5

Esempio

job	1	2	3	4
p_j	3	4	2	6
h_j	10	$1+C_2$	$1.5\ C_3$	$2^{(C4/5)}$

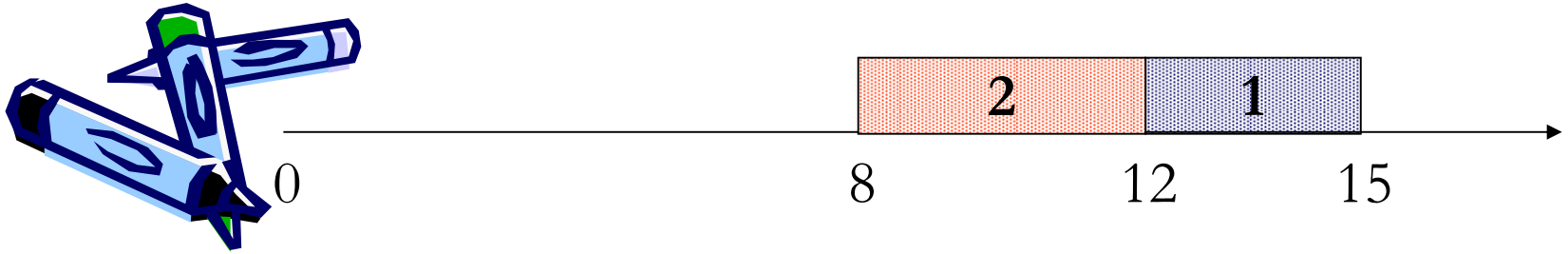


Iter 2.

$$\sum_{j \in J^c} p_j = 12$$

J	{1}
J^c	{2, 3,4}
J'	{2,3}

job	2	3
costo	13	18



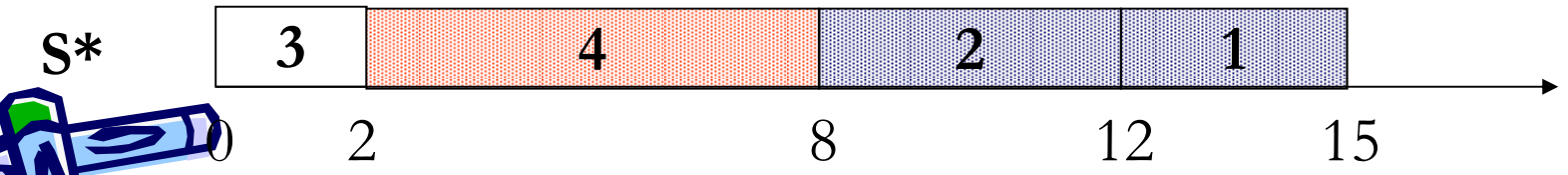
Esempio

job	1	2	3	4
p_j	3	4	2	6
h_j	10	$1+C_2$	$1.5\ C_3$	$2^{(C4/5)}$

Iter 3. $\sum_{j \in J^c} p_j = 8$

J	{1,2}
J^c	{3,4}
J'	{3,4}

job	3	4
costo	12	3.03



z(S*) = max(10, 13, 3.03, 3)

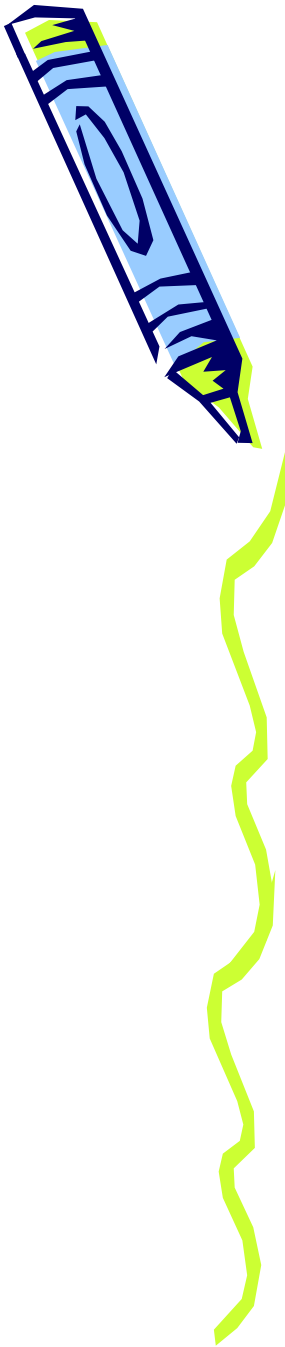
Complessità

L'algoritmo di Lawler si esegue in tempo $O(n^2)$

Infatti, assumendo che il calcolo del valore delle funzioni h_j richieda tempo costante, l'algoritmo richiede n passi, ciascuno basato su n confronti

- Nel caso $1//T_{max}$ l'algoritmo di Lawler si specializza:

Equivale alla regola **EDD** (Earliest Due Date)



$1/r_j/L_{max}$

Algoritmo branch-and-bound

Branch-and-bound: nozioni di base

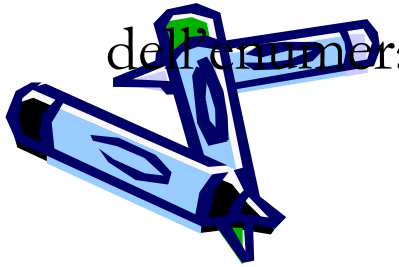
Dato un problema di ottimizzazione combinatoria: $P_0 = (\boldsymbol{z}, S_0)$

$$z^* = min \{\boldsymbol{z(x)} : \boldsymbol{x} \in S_0\}$$

$$S_0 = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m\}$$

• in teoria, è sempre possibile risolvere il problema mediante enumerazione totale. In pratica, dato il valore elevatissimo di m, questo è inapplicabile.

• il metodo branch-and-bound si basa sull'idea di enumerare tutte le soluzioni in S_0 in modo "intelligente" (cioè più efficiente dell'enumerazione totale)



Branching

Si costruisce una partizione $\{S_1, \ldots, S_k\}$ di S_0. Allora, posto

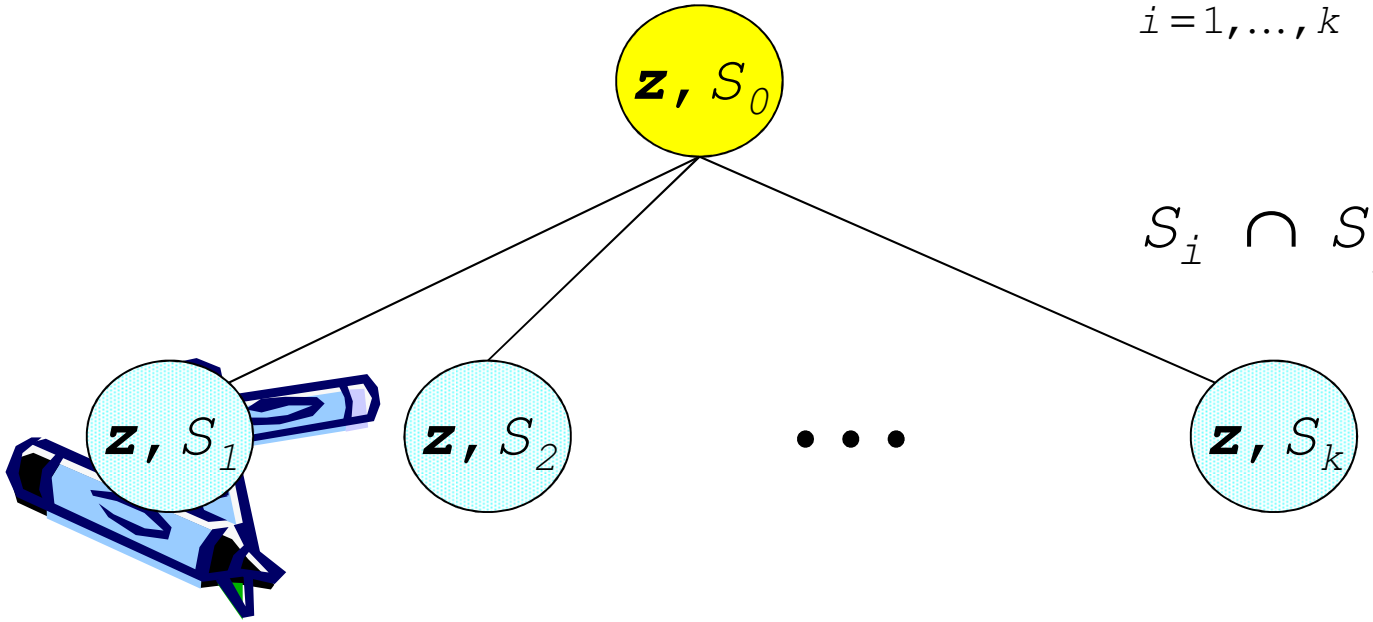
$$z_i^* = min \{z(x) : x \in S_i\}$$

Risulta:

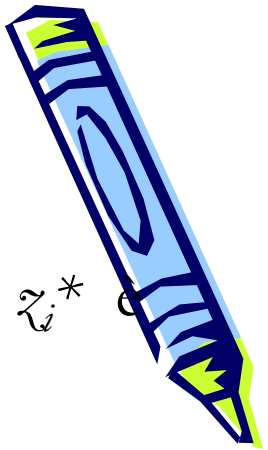
$$z^* = min (z_1^*, \ldots, z_k^*)$$

$$\bigcup_{i=1,\ldots,k} S_i = S_0$$

$$S_i \cap S_j = \emptyset, \; i \neq j$$



Enumerazione implicita

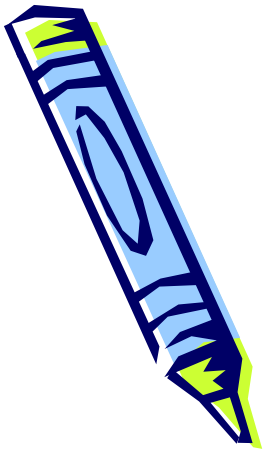


• Se $|S_i \cap S|=1$, $i=1,\ldots,k$, allora il calcolo di z_i^* è facile: enumerazione totale ($k = m$)

• Se k è polinomiale nelle dimensioni dell'input e il problema è NP-hard, allora calcolare z_i^* è NP-hard (se P ≠ NP) e, molto spesso, computazionalmente intrattabile in senso sperimentale.

•Quindi, anziché calcolare z_i^* si applica una tecnica ad-hoc nel tentativo di certificare che nessuna soluzione migliore della migliore soluzione ammissibile nota ($\bar{x}$, **ottimo corrente**) è contenuta in S_i (**valutazione del sottoproblema**). Se ha successo, il problema ($\mathbf{z}$, S_i) è scartato (***pruning***)

Valutazione del sottoproblema



- Calcola una **limitazione inferiore L_i^* di z_i^***.

Se $L_i^* \geq z(\overline{x})$

allora S_i non contiene soluzioni ammissibili migliori di $\overline{x}$ e il sottoproblema (z, S_i) è scartato

Altrimenti

Si costruisce una partizione di S_i costruendo nuovi sottoproblemi.

Calcolo di L_i^*

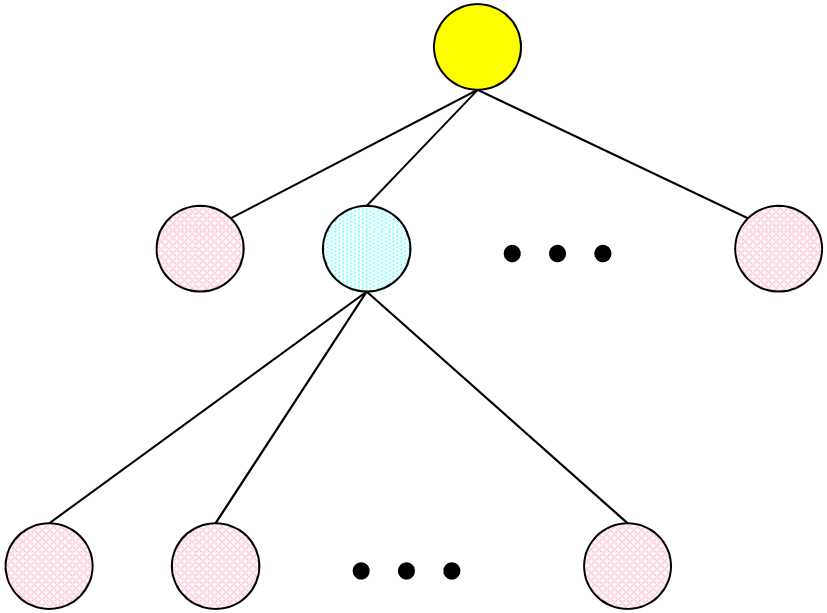
Dato un problema di ottimizzazione $P=(z, S)$ (di minimo), si definisce ***rilassamento*** di P un nuovo problema $RP=(w, \Phi)$ tale che:

(i) $S \subseteq \Phi$

(ii) $\forall x \in S$ risulta $w(x) \leq z(x)$

- Il calcolo di L_i^* si effettua risolvendo <u>in modo esatto</u> un opportuno rilassamento di (z, S_i).
- La scelta del rilassamento si basa su due esigenze, spesso contrastanti:
  - Ottenere buone approssimazioni di z_i^*
  - Richiedere tempi di calcolo non elevati

Albero di enumerazione



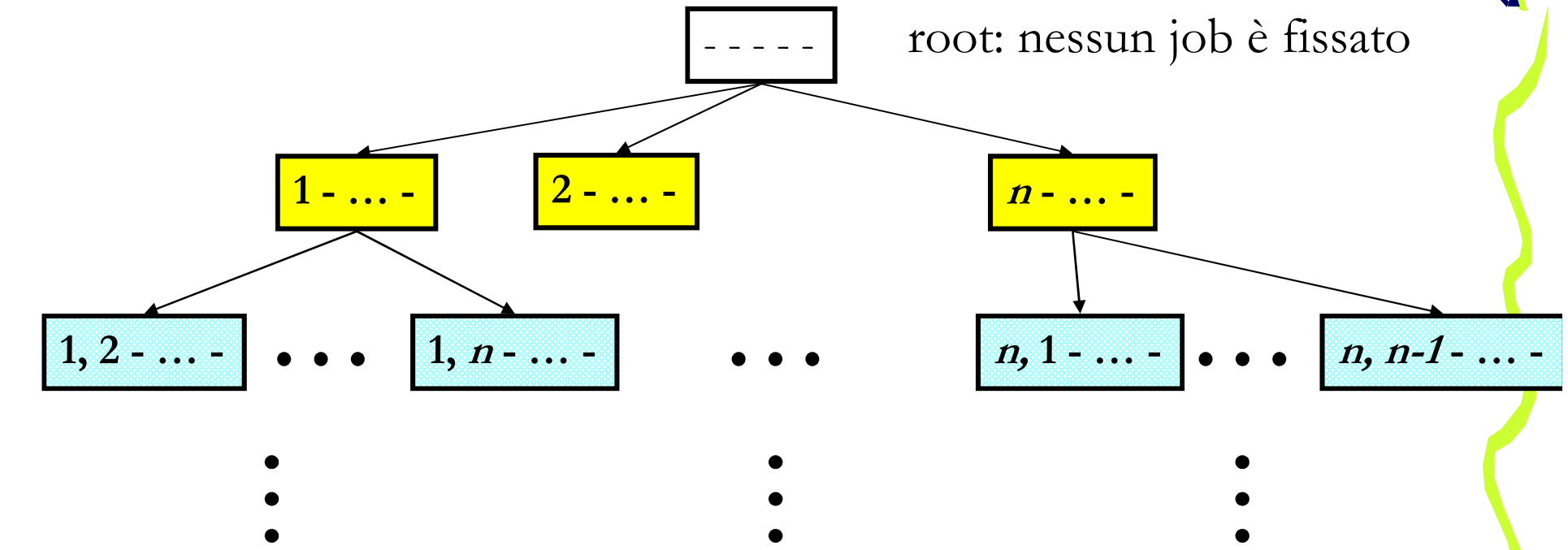
• Le foglie dell'albero di enumerazione forniscono, in ogni fase del suo sviluppo, una partizione di S_0 (***lista dei sottoproblemi candidati***).

• Arresto:

- la lista dei problemi candidati è vuota
- si ottiene una soluzione ammissibile di valore L_0

Branch-and-bound per $1/r_j/L_{max}$

branching: al livello h dell'albero di enumerazione si fissa in tutti i modi possibili il job in posizione h (**schedule *parziale***):



Livello n : ***n!*** **sottoproblemi**

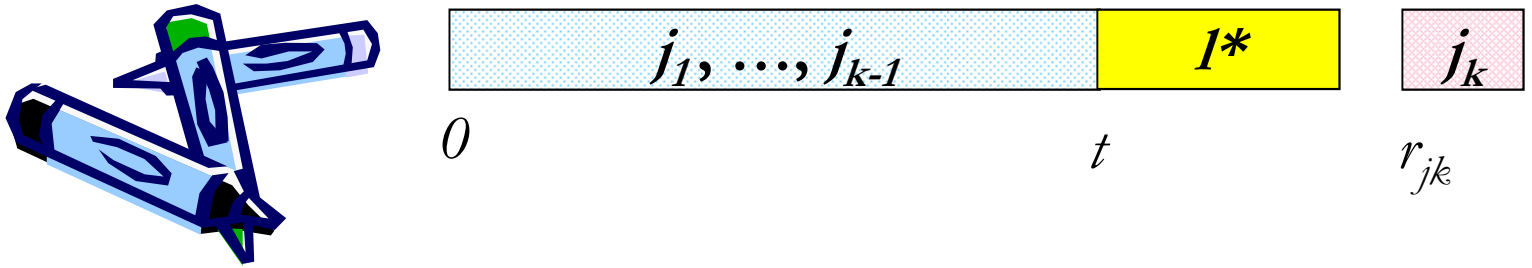
Valutazione del sottoproblema

• Rilassamento preemptivo: $1/r_j, prmp/L_{max}$ è risolto all'ottimo dalla regola PEDD (perché è un rilassamento?)

• Consideriamo un problema al livello k-1, in cui i job $j_1, \ldots, j_{k-1}$ sono fissati nelle prime k-1 posizioni: il sottoproblema $j_1, \ldots, j_{k-1}, \boldsymbol{j_k}$ deve essere generato solo se:

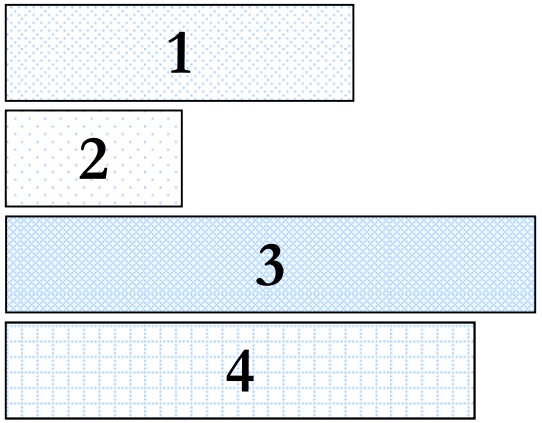
$$r_{j_k} < \min_{l \in J} (\max(t, r_l) + p_l) = \max(t, r_{l*}) + p_{l*}$$

altrimenti si otterrebbe un sottoproblema dominato da:

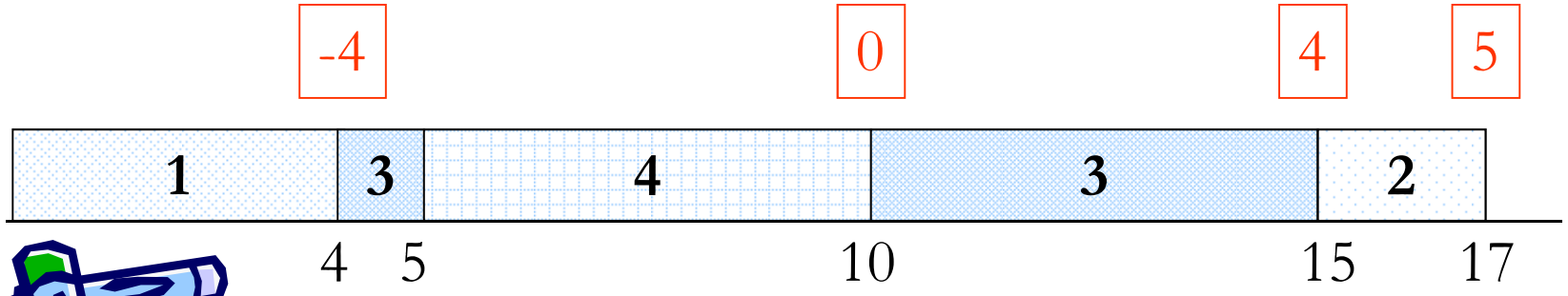


Esercizio

job	1	2	3	4
p_j	4	2	6	5
r_j	0	1	3	5
d_j	8	12	11	10



nodo radice:



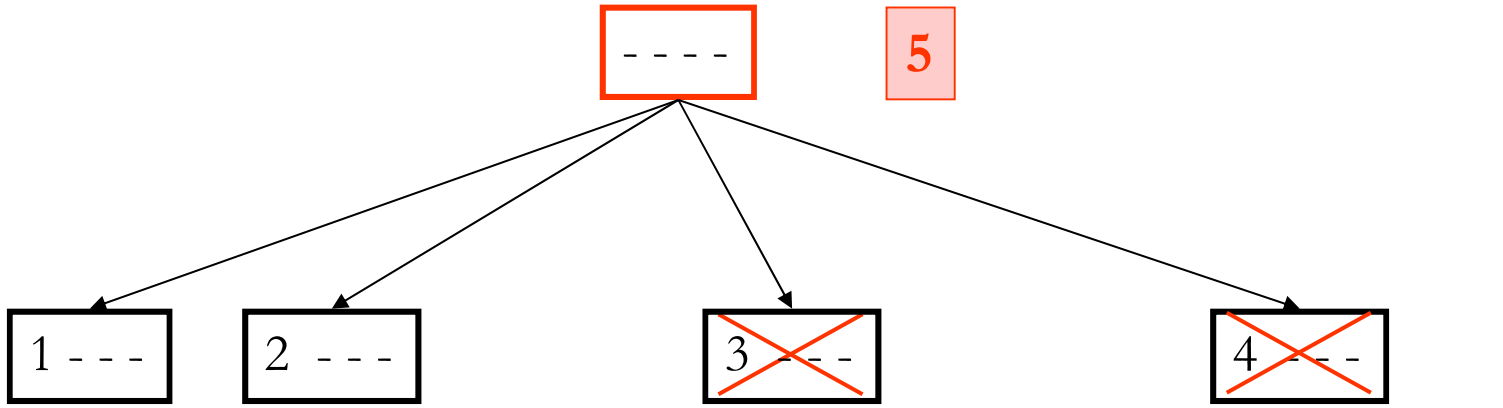
$L_0 = 5$

Branching

job	1	2	3	4
p_j	4	2	6	5
r_j	0	1	3	5
d_j	8	12	11	10

ottimo corrente: (1,2,3,4)

$\bar{z} = 7$



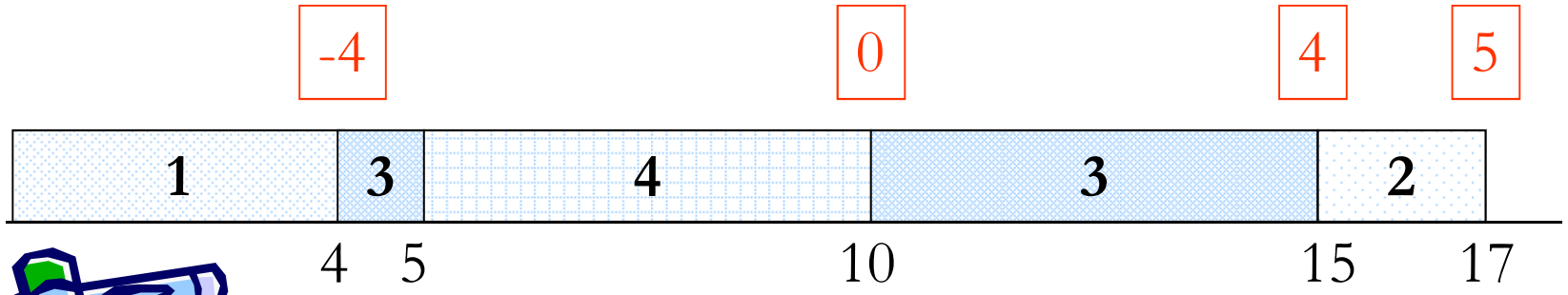
$r_3 > r_2 + p_2$

$r_4 > r_2 + p_2$

Eliminati per dominanza

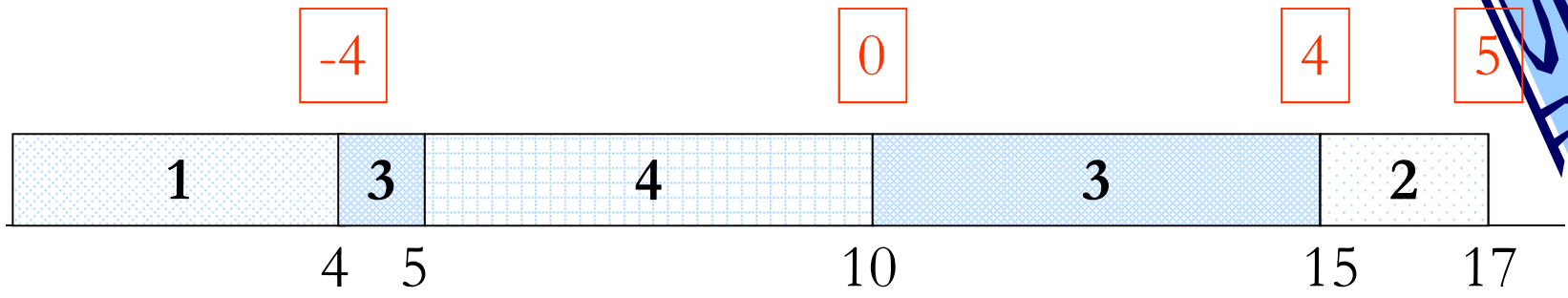
Valutazione di (1 - - -)

job	1	2	3	4
p_j	4	2	6	5
r_j	0	1	3	5
d_j	8	12	11	10



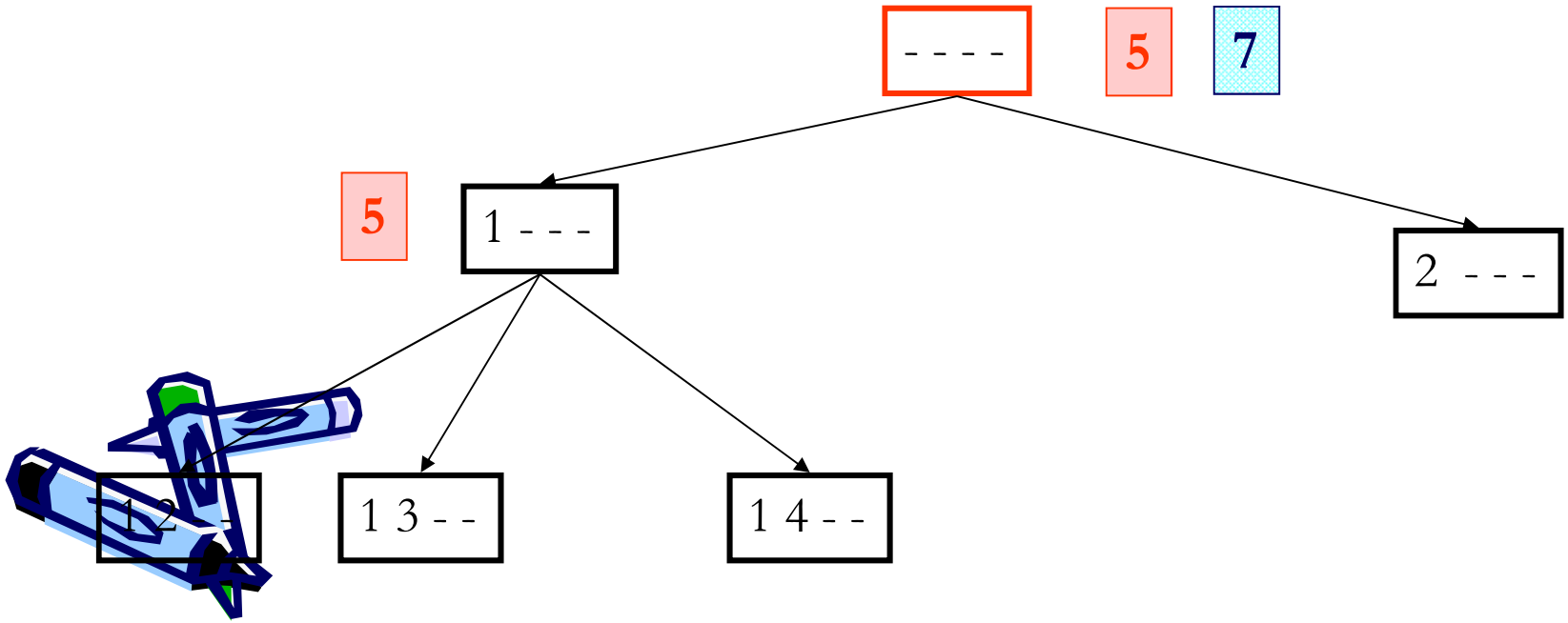
$L(1,---) = 5$

Valutazione di (1 - - -)



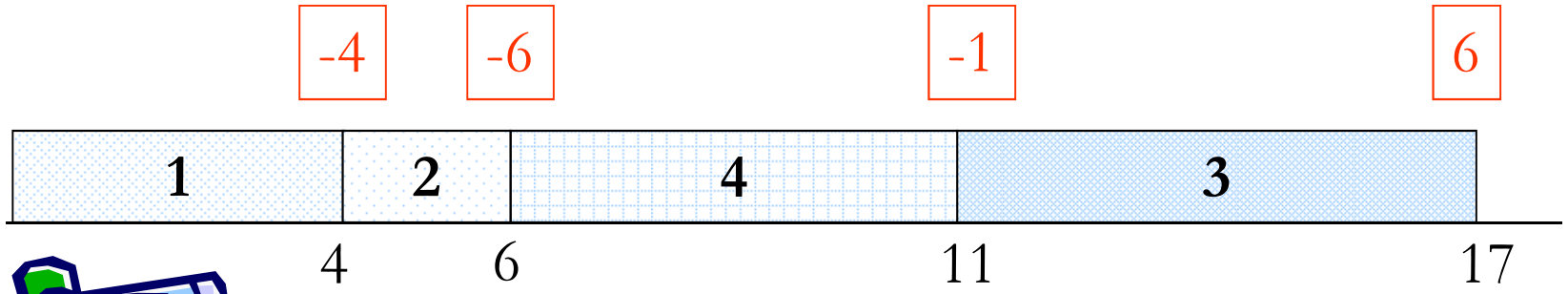
$L(1,---) = 5$

la condizione di pruning fallisce



Valutazione di (1 2 - -)

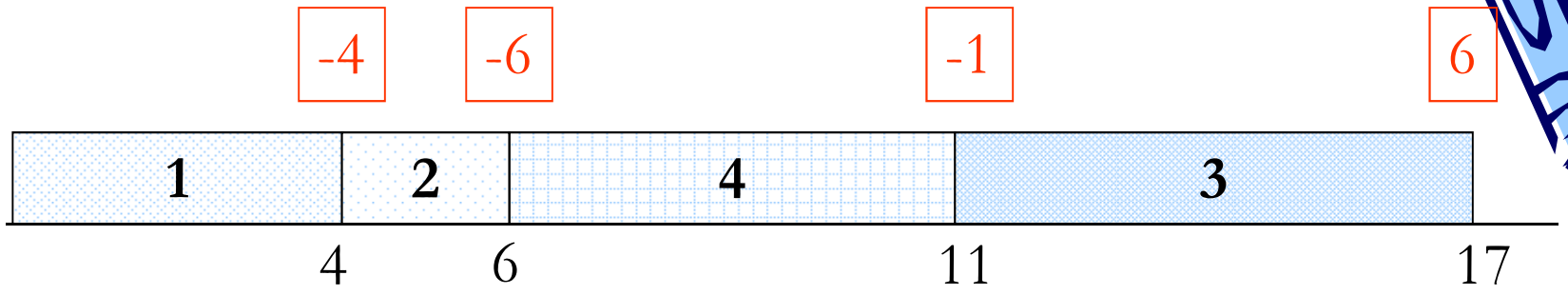
job	1	2	3	4
p_j	4	2	6	5
r_j	0	1	3	5
d_j	8	12	11	10



L(1,2--) = 6

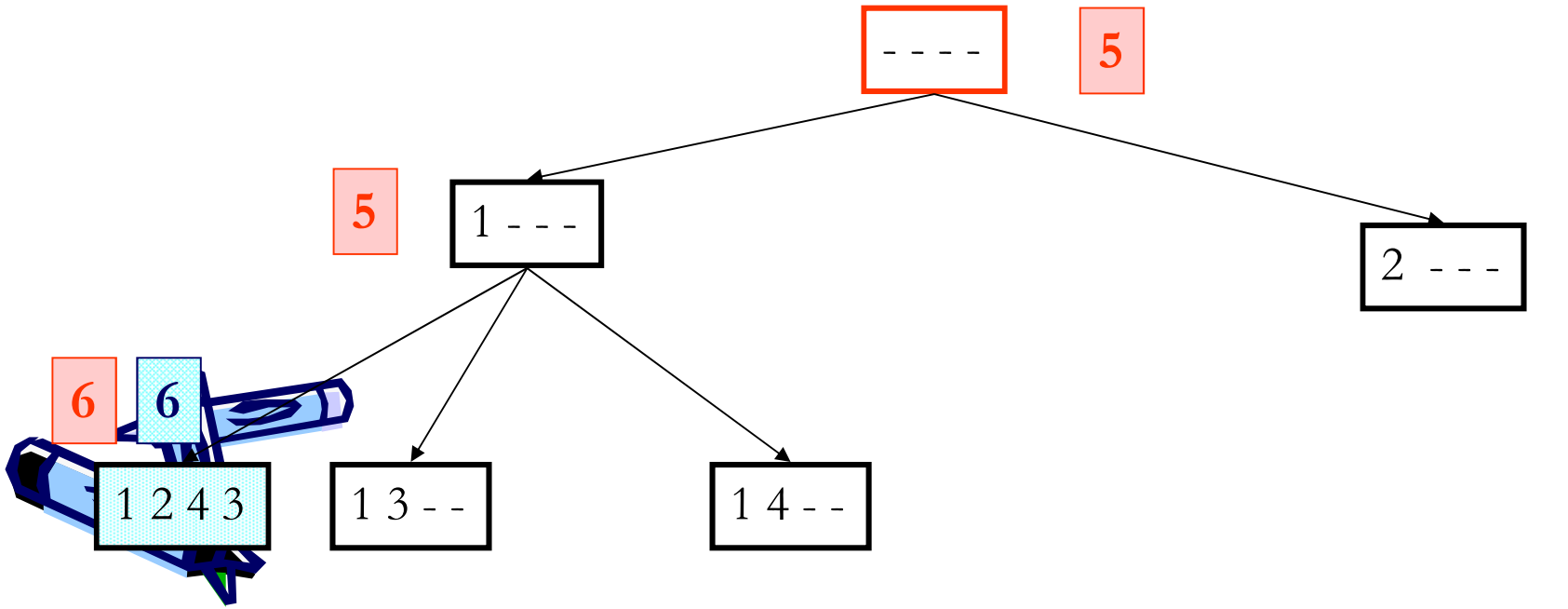
Soluzione ammissibile $\bar{z} = 6$

Valutazione di (1 2 - -)



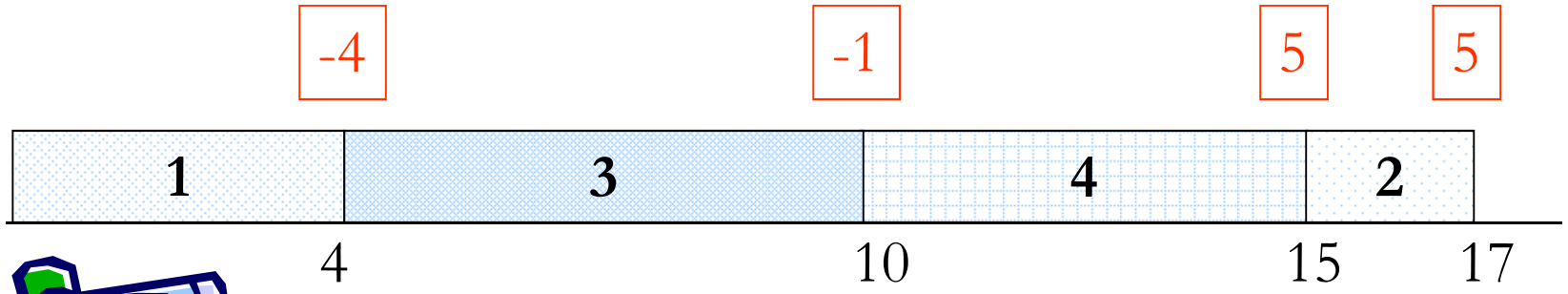
L(1,2--) = 6

sottoproblema chiuso per ottimalità



Valutazione di (1 3 - -)

job	1	2	3	4
p_j	4	2	6	5
r_j	0	1	3	5
d_j	8	12	11	10

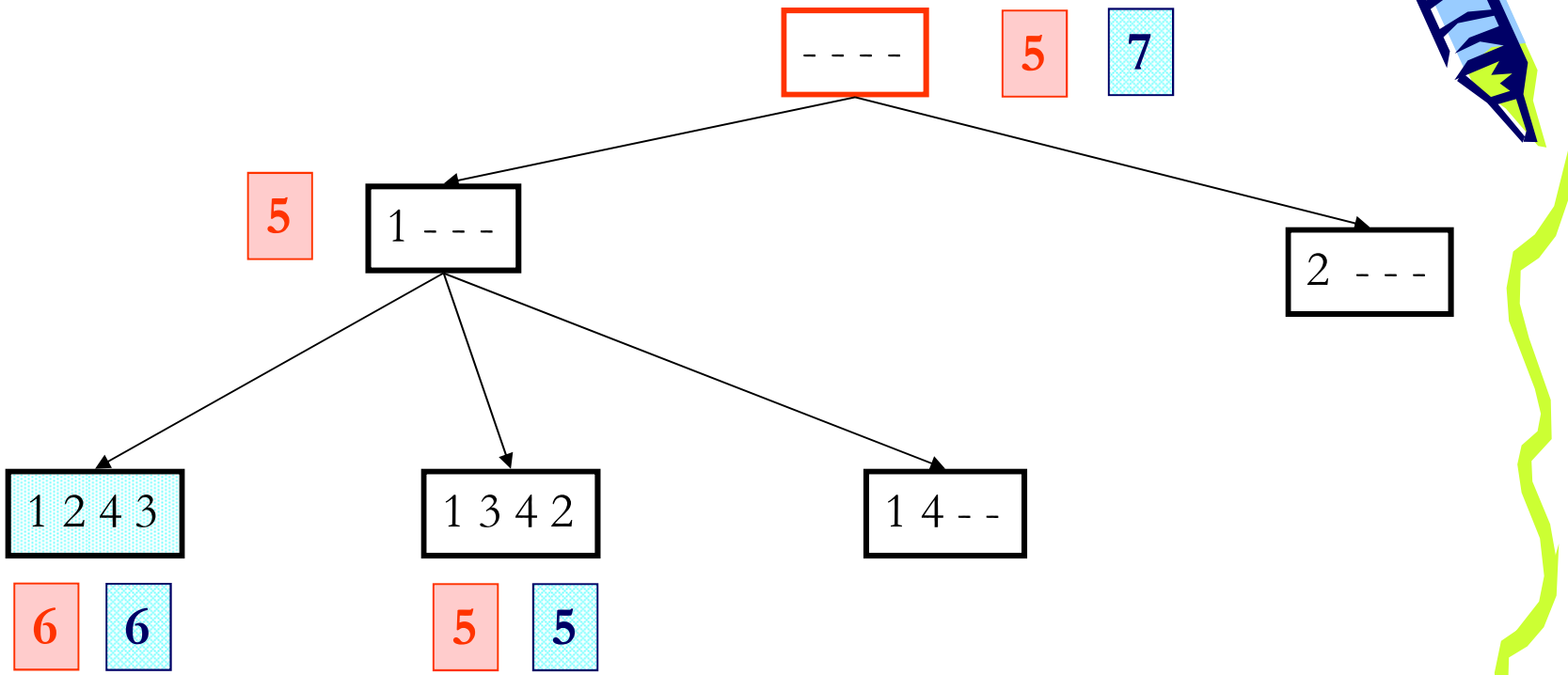


L(1,3--) = 5

Soluzione ammissibile

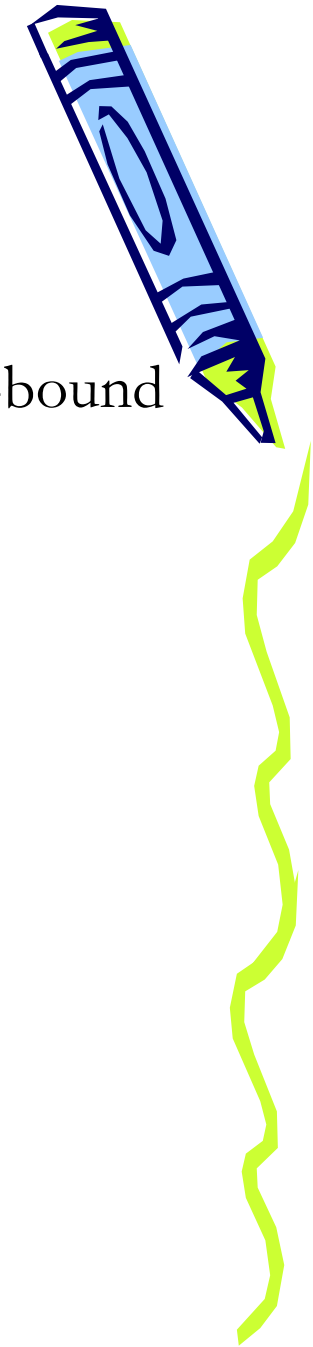
$\tilde{z} = 5$

Arresto



Soluzione ammissibile di valore pari al bound del nodo radice: ottima

Scelte implementative



• Riassumendo, l'implementazione di un branch-and-bound consiste delle seguenti scelte:

• regola di branching
• rilassamento
• regole di dominanza
• regola di selezione del sottoproblema da valutare
• euristiche (in particolare al nodo radice)

INTRODUZIONE

Negli ultimi vent'anni si è evidenziata la necessità di studiare i sempre più numerosi sistemi realizzati dall'uomo, tendenzialmente molto complessi, considerati non tradizionali rispetto alla trattazione classiche proprie della Teoria dei Sistemi e del Controllo.

Questi sistemi dinamici, i cui stati assumono diversi valori logici o simbolici, piuttosto che numerici, in corrispondenza dell'occorrenza di eventi, non sempre possono essere descritti in termini numerici.

INTRODUZIONE

Ne sono esempi significativi i processi produttivi, le reti di elaboratori elettronici, di trasporto, di comunicazione e sistemi formati per integrazione delle suddette tipologie di sistemi.

Esempi di eventi sono: l'arrivo di un cliente nel sistema o la sua partenza da esso, il completamento di una lavorazione o il guastarsi di una macchina in un sistema di produzione, la trasmissione/ricezione di un pacchetto di dati in una rete di telecomunicazioni, il verificarsi di un disturbo o il cambiamento del segnale di riferimento in un complesso sistema di controllo

INTRODUZIONE

L'evoluzione nel tempo di un sistema con tali caratteristiche sembra essere descritta da sequenze di occorrenze di cambiamenti discreti e qualitativi del sistema, ignorando i micro cambiamenti che avvengono continuamente.

SISTEMA (qualitativa)

Ente fisico che risponde alle sollecitazione esercitata da una certa azione producendo una reazione.

Per sviluppare tecniche di progetto, di controllo e/o di valutazione delle prestazioni di un sistema sulla base di specifiche predefinite è necessaria una definizione QUANTITATIVA :**MODELLO FORMALE**.

INTRODUZIONE

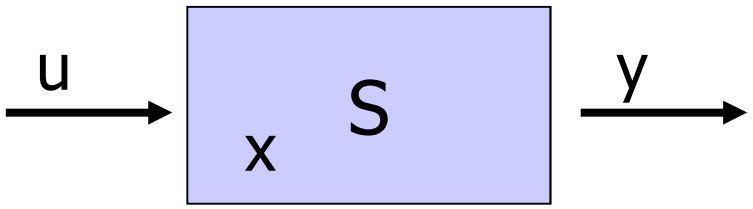
Variabili che evolvono nel tempo

- CAUSE ESTERNE AL SISTEMA (INGRESSI)

  Grandezze il cui andamento nel tempo può essere indipendente dal tipo di sistema

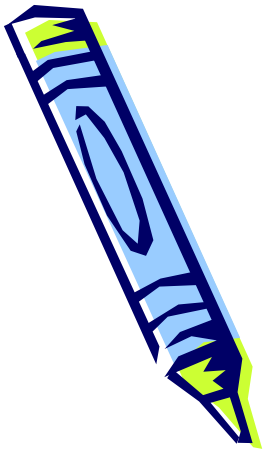
- EFFETTI (USCITE)

  Grandezze il cui andamento nel tempo dipende, almeno in parte dal tipo di sistema e dalle cause esterne



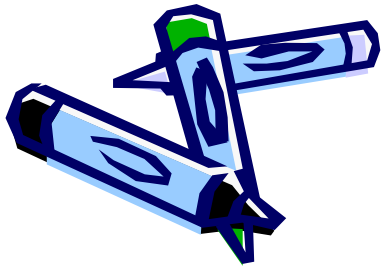
In generale, non è immediato legare in modo semplice l'uscita con l'ingresso, cioè realizzare la dipendenza ingresso/uscita.

INTRODUZIONE



STATO

Rappresenta il comportamento del sistema ad un dato istante di tempo, concentrando in sé l'informazione sul passato e sul presente del sistema X(t) è lo stato all'istante t



INTRODUZIONE

Un sistema ad eventi discreti (SED) si puo definire come un sistema dinamico il cui comportamento è caratterizzato dall'occorrenza di eventi istantanei con un cadenzamento irregolare non necessariamente noto.

Le caratteristiche indiscusse possedute da un SED sono fondamentalmente legate all'evoluzione dinamica basata sull'occorrenza **asincrona** degli eventi, anziché sull' avanzamento sincrono del tempo, e al fatto che almeno alcune delle variabili che descrivono il comportamento di un SED sono discrete.

INTRODUZIONE

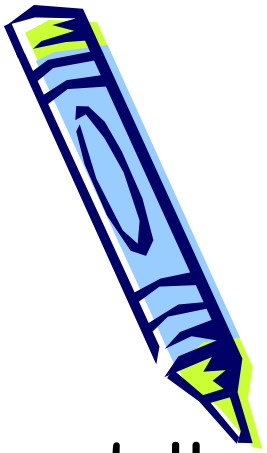
Dal punto di vista formale, un SED può essere considerato come un sistema dinamico, con un opportuno spazio di stato e un proprio meccanismo di transizione di stato.

Un sistema ad eventi discreti è un sistema il cui comportamento dinamico è caratterizzato dall'accadimento **asincrono** di eventi che individuano lo svolgimento di attività di durata non necessariamente nota.

Formalmente, un sistema ad eventi discreti è caratterizzato da:

- un insieme **E** degli eventi accadibili;
- spazio di stato costituito da un insieme discreto **X**
- **evoluzione** dello stato event-driven, cioè regolata dagli eventi: lo stato evolve nel tempo in dipendenza dell'accadimento di eventi asincroni, appartenenti all'insieme E

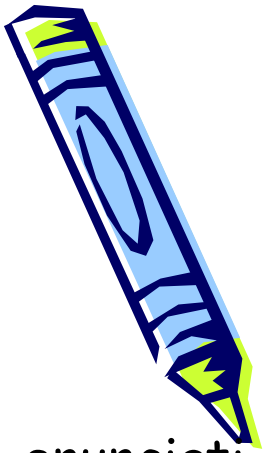
INTRODUZIONE



L'equazione che descrive l'evoluzione dello stato a partire dallo stato iniziale x_0 è:

- $x_{k+1} = f(x_k, e_k)$ $k \in \mathbb{N}$
- x_{k+1} è lo stato del sistema dopo l'accadimento del k-esimo evento
- e_k è il k-esimo evento accaduto dall'istante iniziale considerato, che fa transire lo stato da x_k a x_{k+1}
- $f: X \times E \rightarrow X$ è la funzione di transizione di stato

INTRODUZIONE

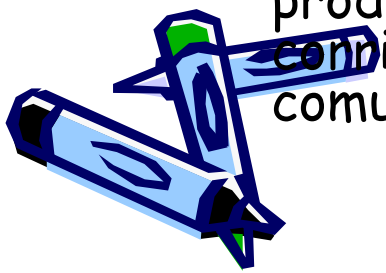


L'esempio più semplice per esplicare i concetti fin qui enunciati riguardo ai SED, è senza dubbio il sistema a coda o ad accodamento. Un sistema di questo tipo può essere considerato come il blocco elementare con cui costruire le rappresentazioni di molte tipologie di SED.

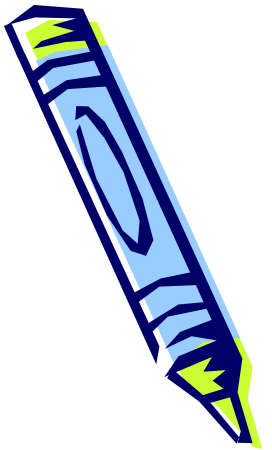
Un sistema a coda si fonda su tre componenti fondamentali:

- le entità che attendono per utilizzare le risorse, dette clienti
- le risorse per cui ci si accoda, detti serventi o servitori
- lo spazio in cui si attende, che è la coda vera e propria

I clienti possono essere persone, messaggi in reti di telecomunicazioni, task in computer, semilavorati in sistemi di produzione, veicoli in reti di trasporto, ecc. Esempi di serventi corrispondenti sono invece ancora persone, canali di comunicazione, processori, macchine, semafori,ecc.



INTRODUZIONE



Visto come un SED, il sistema a coda è caratterizzato dall'insieme di eventi

E = {a,p} con

- a: evento di arrivo di un cliente;
- p: evento di partenza di un cliente.

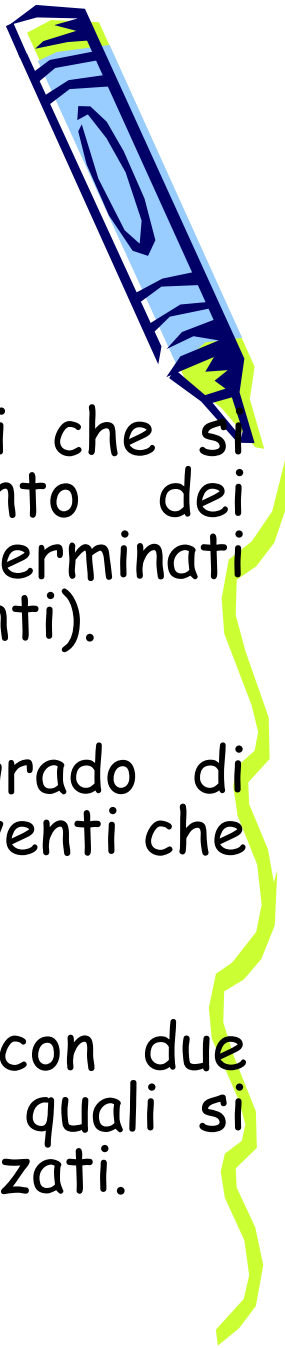
La variabile di stato più intuitiva è il numero di clienti in coda; in questo caso si pone X = {1,2,3,…}

Per specificare completamente le caratteristiche di un sistema a coda bisogna ancora definire:

- La capacità della coda, cioè il numero di clienti che possono accodarsi (spesso considerato illimitato);
- La disciplina di accodamento, cioè la regola con cui si sceglie il prossimo cliente da servire tra quelli in coda.

Collegando tra loro più blocchi elementari coda si costruiscono reti di code.

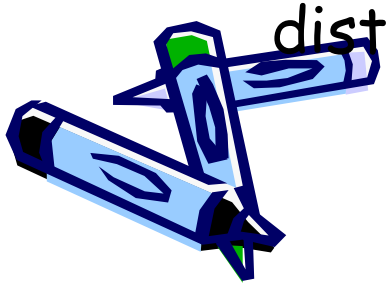
INTRODUZIONE



E' attraverso i Modelli ad Eventi Discreti che si effettua un'astrazione del comportamento dei sistemi, registrando l'occorrenza di determinati eventi discreti (traccia/traiettoria degli eventi).

Un MED è un modello matematico in grado di rappresentare l' insieme delle tracce degli eventi che possono essere generate da un sistema.

Le tracce possono essere rappresentate con due diversi livelli di astrazione, sulla base dei quali si distinguono : modelli logici e modelli temporizzati.

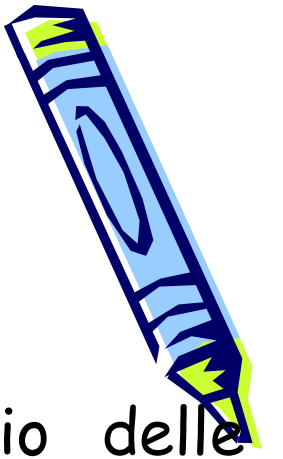


INTRODUZIONE

Nei **MODELLI LOGICI** la traccia degli eventi è costituita semplicemente da una sequenza di eventi $\{e_1, e_2,\ldots\}$, in ordine di occorrenza, senza alcuna informazione circa i tempi di occorrenza degli eventi; dato uno stato iniziale x_0, la traiettoria dello stato verrà costruita nel tempo la sequenza di stati $\{x_0,x_1,x_2,\ldots\}$, risultanti dall'accadimento della sequenza di eventi, ma non è possibile specificare gli istanti di tempo in cui avvengono le transizioni di stato.

Nei **MODELLI TEMPORIZZATI** invece la traccia degli eventi è costituita da una sequenza di coppie $\{e_1\ t_1,\ e_2t_2\ ,e_3t_3,\ldots\}$,dove ogni evento e_i è accoppiato al suo tempo di accadimento, t_i, eventualmente stocastico: dato uno stato iniziale x_0, la traiettoria dello stato verrà costruita nel tempo la sequenza di stati $\{x_0,x_1,x_2,\ldots\}$, risultanti dall'accadimento della sequenza di eventi, si sa che le transizioni di stato avvengono negli istanti di occorrenza degli eventi.

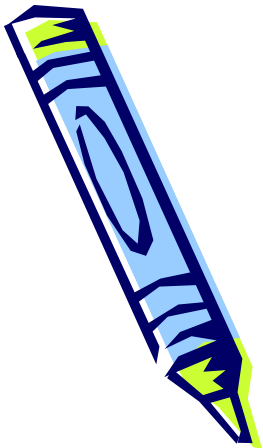
INTRODUZIONE



I **modelli logici** rendono agevole lo studio delle proprietà qualitative del sistema e consentono quindi di effettuare l'analisi strutturale di un SED, mentre i **modelli temporizzati** permettono di studiare i diversi comportamenti nel tempo del sistema, pertanto sono indispensabili qualora si voglia effettuare l'analisi prestazionale di un SED.

Nella formulazione del modello logico è fondamentale specificare l'insieme delle traiettorie ammissibili, ossia le sequenze di eventi fisicamente realizzabili. A questo scopo può essere adottato uno dei noti formalismi sviluppati per rappresentare le transizioni di stato in un SED, come gli AUTOMI, o le RETI DI PETRI.

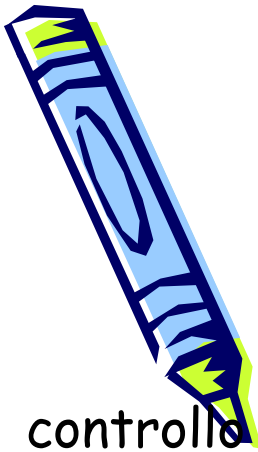
INTRODUZIONE



La necessità di modelli per descrivere il funzionamento dei sistemi è una costante di tutti i problemi di ingegneria: non è possibile progettare alcunché se non si dispone di un modello adeguato.

Peraltro, il tipo di modello che serve può essere molto diverso, a seconda dell'uso che se ne deve fare. Per esempio, il modello dinamico di un sistema che si presta per il progetto di un sistema di controllo è generalmente molto più semplice di un simulatore dello stesso sistema.

INTRODUZIONE



Cosa vogliamo descrivere con i modelli nel contesto del controllo logico?

Vogliamo descrivere il funzionamento di impianti molto complessi ed eterogenei:

celle robotizzate, centri di lavorazione meccanica, impianti batch/chimici, ecc.

Ciascuno di questi può essere visto come un insieme di sotto-sistemi, dispositivi, macchinari, ecc. opportunamente interconnessi, ognuno dei quali può essere modellizzato con molto dettaglio (v. modello motore, serbatoio, ecc.).

A noi interessa studiare questi processi ad un livello di astrazione più elevato, in cui si evidenzino le *sequenze di operazioni*, con i relativi problemi di sincronizzazione, parallelismo, ecc.

INTRODUZIONE

Ci poniamo domande come:

- Che operazione devo svolgere dopo l'operazione *X*?
- Le operazioni *X* e *Y* possono essere svolte in parallelo?
- In quali condizioni non devo eseguire l'operazione *X*?
- Ci sono risorse sufficienti per svolgere le operazioni che mi servono?

Questo modo di ragionare è tipico dei sistemi manifatturieri, le cui caratteristiche macroscopiche sono descrivibili con *condizioni logiche di funzionamento discrete*, senza valori numerici, come ad es. "macchina pronta per la lavorazione", "macchina in attesa", "macchina guasta".

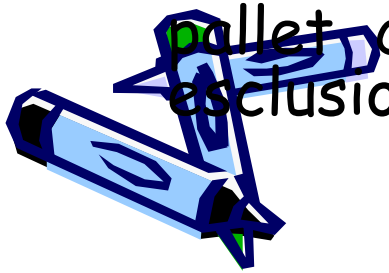
Tali condizioni logiche cambiano in modo istantaneo da un valore all'altro, ad es. con un *comando* "accendi la macchina", oppure con un *segnale* di "fine corsa raggiunto". Normalmente, non è noto a priori né quale sia il nuovo valore, né l'istante temporale in cui avviene il cambiamento.

INTRODUZIONE

L'evoluzione di tali sistemi può allora essere studiata in termini di cambiamento delle condizioni logiche di funzionamento discrete, per effetto di sequenze di comandi/segnali.

Si evidenziano così alcuni funzionamenti tipici, come:

- evoluzione parallela e asincrona (macchine in parallelo → vanno sincronizzate)
- presenza di scelte (bivio in una linea)
- condivisione di risorse (magazzino di utensili o pallet condiviso da più macchine, vincoli di mutua esclusione nell'allocazione delle risorse)



INTRODUZIONE

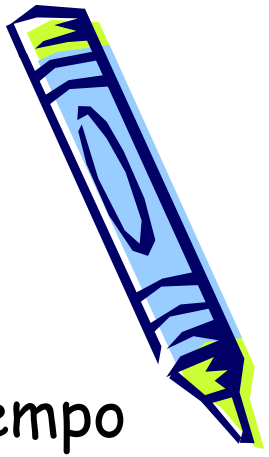
Che tipo di modello matematico ci serve per descrivere questi fenomeni?

Osservando il parallelismo che sussiste fra i concetti seguenti:

- condizione logica di funzionamento ↔ stato (discreto)
- sequenza comandi/segnali ↔ sequenza di ingressi

è facile capire che lo strumento che ci serve è una qualche forma di *sistema dinamico*.

INTRODUZIONE



I sistemi dinamici che conosciamo (sistemi a tempo continuo o discreto), però, hanno alcune caratteristiche che non si prestano bene alla descrizione dei fenomeni che abbiamo citato in precedenza:

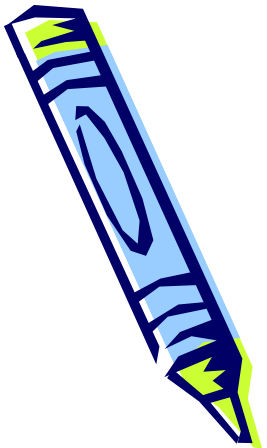
- lo spazio di stato è continuo, ovvero le variabili variano in modo continuo sull'asse reale, mentre a noi interessa esprimere concetti come "serbatoio pieno" o "serbatoio vuoto" (invece di "il serbatoio contiene *X* litri di acqua") → ***sistemi a stato discreto***
- l'evoluzione è guidata dal "tempo", ovvero lo stato può variare ad ogni istante; nel nostro caso lo stato cambia solo in certi istanti, con transizioni istantanee (da "macchina occupata" a "macchina libera") → ***sistemi ad eventi***: lo stato varia quando si verifica un evento (istantaneo)

INTRODUZIONE

Nei sistemi guidati dagli eventi (event-driven) la modellizzazione e l'analisi sono rese complicate dal fatto che occorre specificare i meccanismi asincroni di occorrenza degli eventi nel tempo.

Tra sistemi time-driven e event-driven c'è concettualmente la stessa differenza che sussiste in un calcolatore tra le operazioni sincronizzate dal clock e quelle gestite tramite interrupt.

INTRODUZIONE

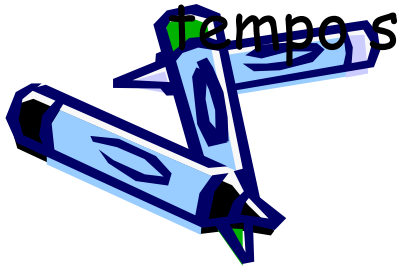


I sistemi dinamici si dividono in:

- sistemi dinamici a tempo continuo o discreto l'evoluzione è guidata dal "tempo"

- sistemi dinamici ad eventi discreti l'evoluzione è guidata dall'accadimento degli "eventi", considerati per semplicità istantanei, che accadono ad intervalli irregolari non noti a priori

- sistemi ibridi l'evoluzione è determinata sia dal tempo sia da eventi



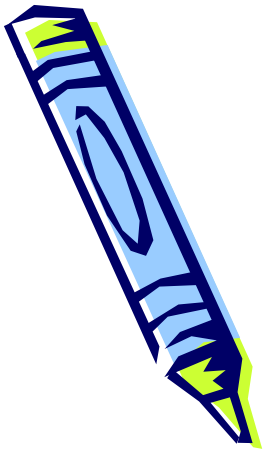
INTRODUZIONE

Un sistema ad eventi discreti è caratterizzato dai seguenti elementi:

- le variabili di stato assumono valori numerici discreti (cioè una quantità finita o numerabile di valori) o sono descrivibili in termini simbolici (parole, stringhe, ecc.);

- gli stati cambiano in corrispondenza dell'occorrenza di eventi, i quali anch'essi possono essere descritti in termini non numerici.

INTRODUZIONE

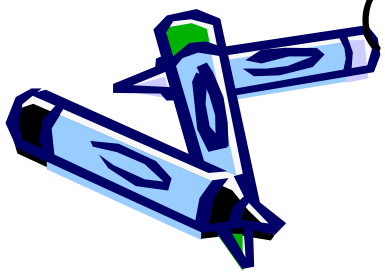


ESEMPIO:

Lo stato assume solo i quattro valori x_1, x_2, x_3 e x_4, e cambia solo in alcuni istanti (t_1, t_2, t_3, t_4), in corrispondenza degli eventi e_1, e_2, e_3, e_4.

Assumendo che il sistema sia deterministico (nel senso che la legge che determina lo stato successivo in corrispondenza dell'occorrenza di un evento sia unica), l' informazione completa è fornita dalla sequenza (temporizzata) di eventi:

$(t1,e1)\ (t2,e2)\ (t3,e3)\ (t4,e4)$



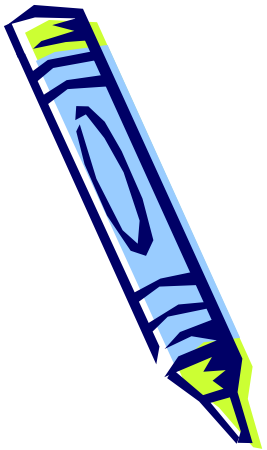
INTRODUZIONE

Spesso, tuttavia, non ci interessa *quando* il sistema entra in un determinato stato o *quanto a lungo* il sistema rimane nel medesimo stato, ma piuttosto l'ordinamento degli eventi (e quindi di transizioni):

e_1, e_2, e_3, e_4.

Ci interessa cioè se un evento accade *prima* o *dopo* un altro. Eliminando la temporizzazione, stiamo di fatto modellizzando il *comportamento logico* del sistema.

INTRODUZIONE



Un modello logico di questo tipo consente di:

- distinguere le sequenze di eventi che sono compatibili con delle specifiche di comportamento
- verificare se un determinato stato è raggiungibile, e con quale sequenza di eventi
- verificare se il sistema si blocca in uno stato

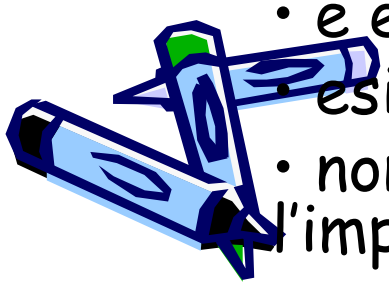
INTRODUZIONE

Chiaramente, non può servire per un'analisi delle prestazioni del sistema.

Non consente di rispondere a domande come le seguenti:

- quanto tempo spende il sistema in un determinato stato?
- qual è il tempo minimo in cui uno stato può essere raggiunto?
- può una sequenza di eventi essere completata in un tempo assegnato?
- quanto dura il tempo di ciclo di una sequenza di lavorazioni?

INTRODUZIONE

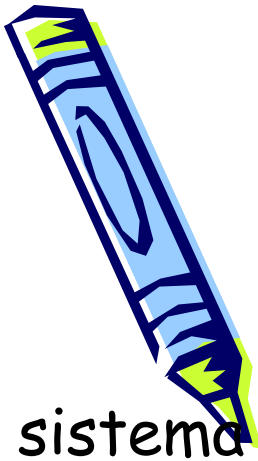


In entrambi i modelli visti è difficile cogliere parallelismo, condivisione di risorse,scelte, asincronia, ecc. → vanno bene per descrivere un determinato legame ingresso/uscita, ma non ci aiutano a capire la struttura interna del sistema.

Un terzo modo di rappresentare i sistemi ad eventi discreti consiste nell'utilizzare un linguaggio di programmazione, che ha le seguenti importanti caratteristiche:

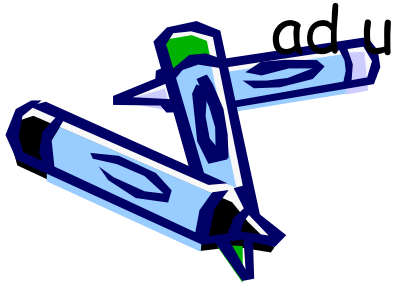
- è un modello formale (e analizzabile in modo formale)
- è eseguibile (su una determinata macchina)
- esistono supporti (debug, case, ecc.)
- non ha problemi di traduzione, ma è pronto per l'implementazione del controllo!

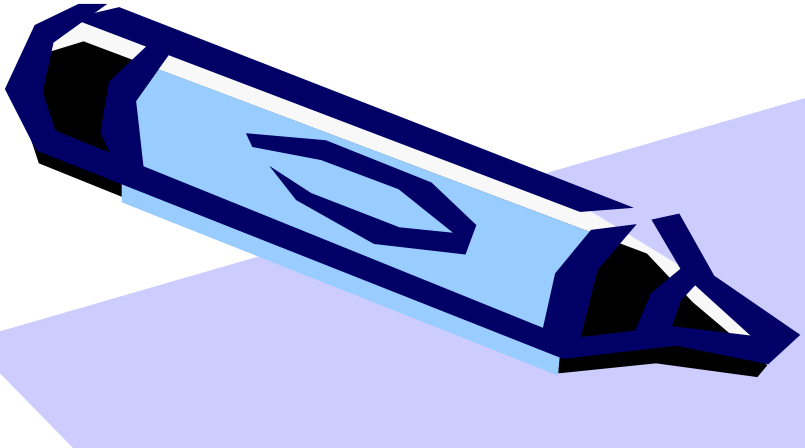
INTRODUZIONE



Purtroppo, la descrizione e lo studio di un sistema tramite linguaggio di programmazione:

- è un lavoro complesso e dettagliato
- non favorisce l'astrazione dei concetti principali
- non ha strutture modellistiche standard
- i relativi "modelli" non sono portabili da un sistema ad un altro





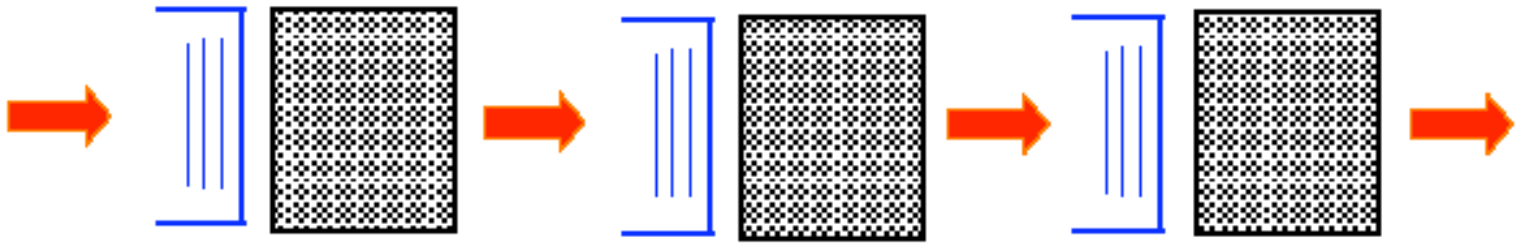
Modelli di processi produttivi

Livello: operativo, decisionale, di controllo

Produzione discreta: classificazione storica

Dal punto di vista storico i sistemi di produzione discreta sono stati classificati in:

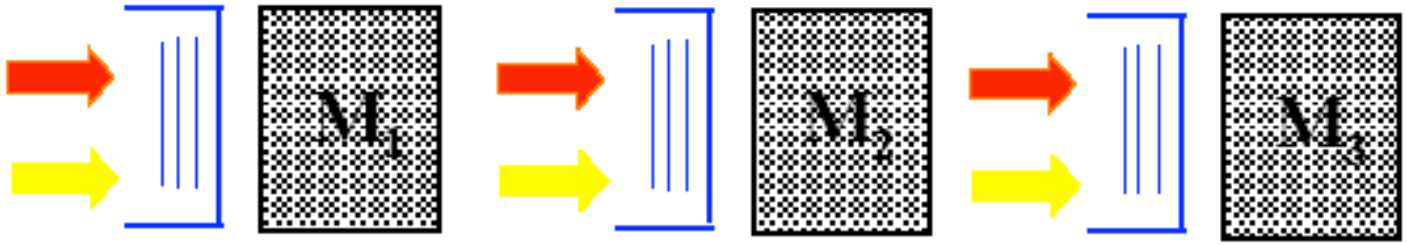
- **Linea di trasferimento (o di montaggio) – *Transfer line***



- un solo prodotto
- produzione di massa

Produzione discreta: classificazione storica

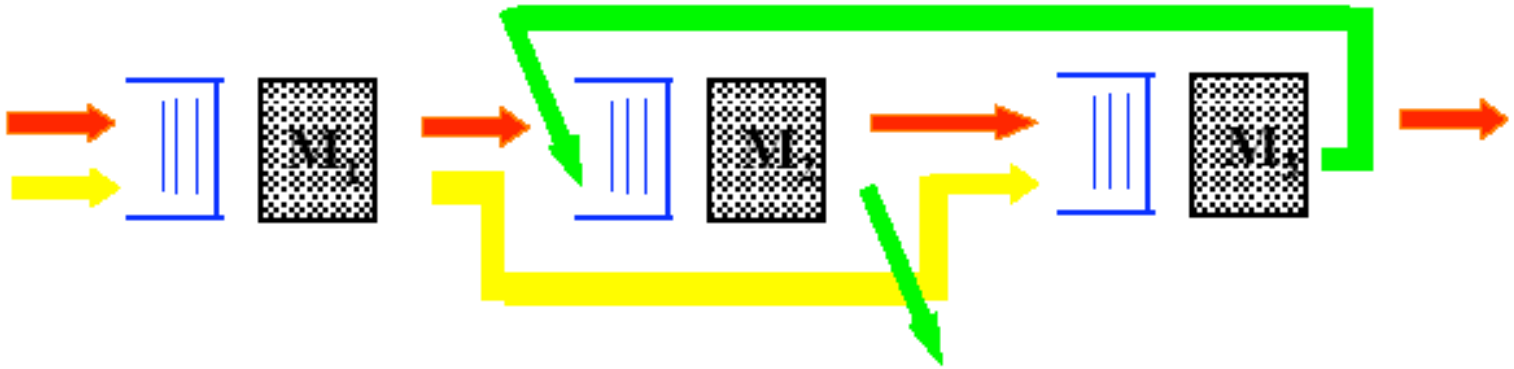
Modelli Flow-shop



- diversi prodotti
- stessa sequenza di operazioni sulle stesse risorse
- problemi di sequenziamento delle singole risorse

Produzione discreta: classificazione storica

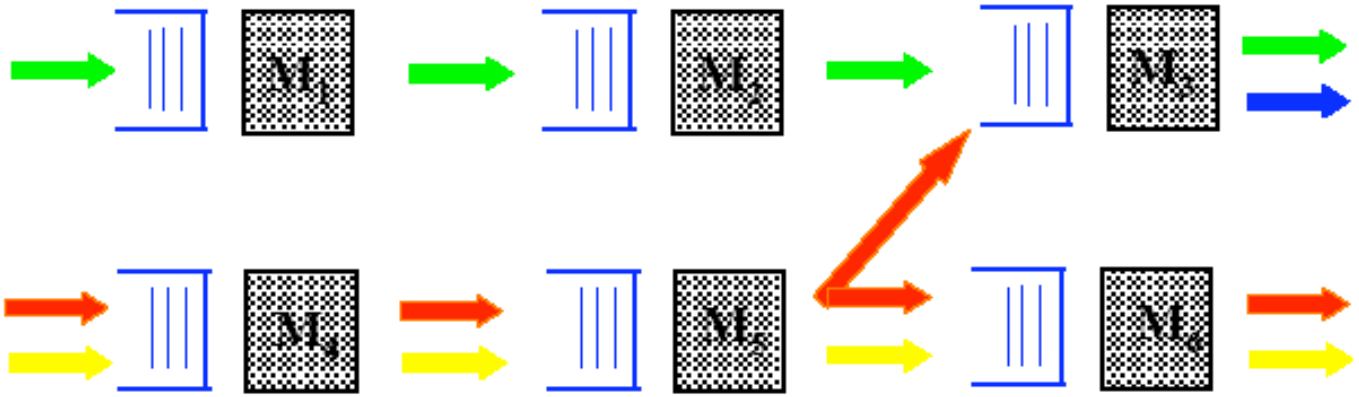
Modelli Job-shop



- diverse categorie di prodotti
- sequenze di operazioni diverse
- problemi di sequenziamento delle singole risorse

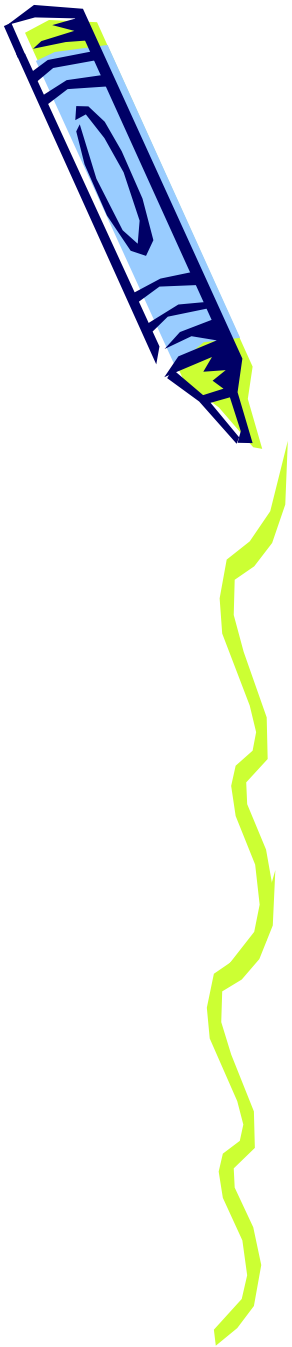
Produzione discreta: classificazione storica

Sistemi flessibili di lavorazione (*Flexible manufacturing systems –FMS*)



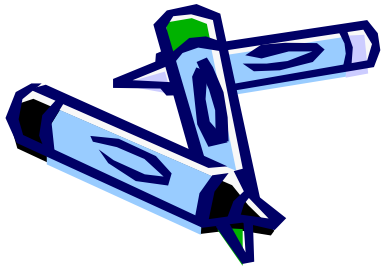
- diversi prodotti
- sequenze di lavorazione diverse
- lavorazioni eseguite su più di una risorsa
- problemi di assegnazione delle operazioni alle risorse, con conseguente gestione dei flussi (dimensionamento, routing)
- problemi di sequenziamento locale delle risorse

Modello generale



Le entità presenti nel modello sono:

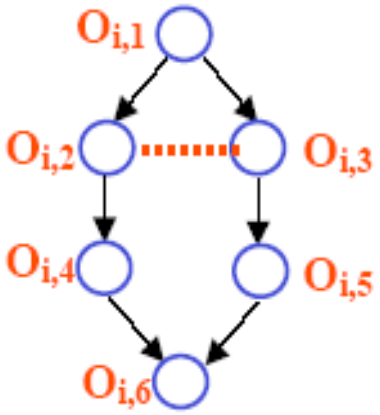
- prodotti e semilavorati
- risorse produttive
- risorse di trasporto
- risorse di immagazzinamento



Prodotti e semilavorati

p classi di prodotti $P_1, \ldots, P_p$

Un *grafo di precedenza delle operazioni* è associato ad ogni classe di prodotti



$O_{i,j}$ = j-esima operazione necessaria per realizzare prodotti di classe P_i

..... vincolo di incompatibilità tra le operazioni

Grafo di precedenza delle operazioni

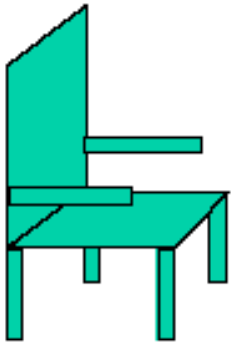
Il grafo di precedenza delle operazioni per la generica classe di prodotti Pi:

- è un grafo orientato
- è composto da $\mathbf{n_i}$ nodi (n_i = n° operazioni necessarie per produrre prodotti di classe Pi)
- generalmente è un albero - non prevede cicli
- generalmente è un in-albero (*in-tree*) - esiste un'unica operazione finale
- comprende vincoli di incompatibilità tra le operazioni

Prodotti e semilavorati

Per ogni classe di prodotti P_i viene definito anche il ***grafo di realizzazione del prodotto*** che rappresenta quali componenti di base/semilavorati servono per ottenere ogni semilavorato/prodotto finito.

Esempio:

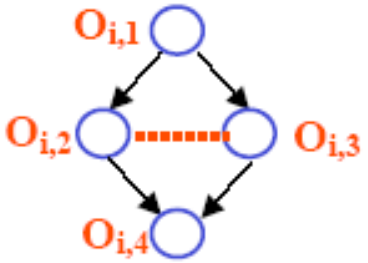


Componenti di base:

- 4 gambe
- 1 sedile
- 1 schienale
- 2 braccioli

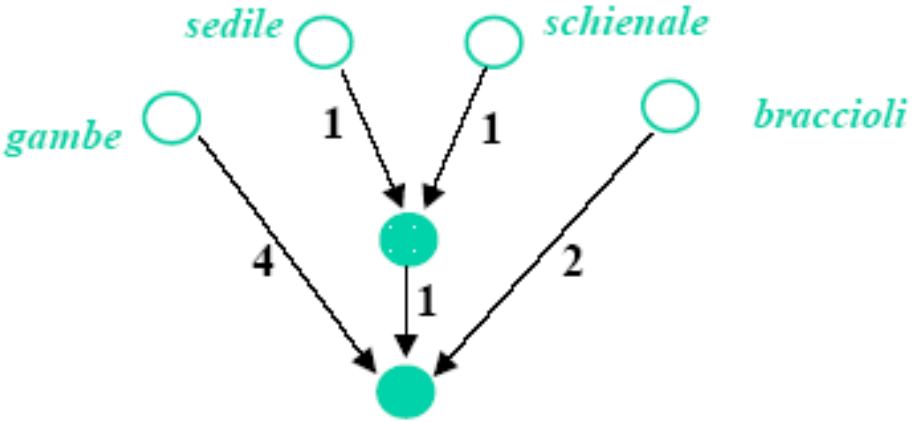
Grafo di precedenza delle operazioni:

- $O_{i,1}$: "saldatura" tra schienale e sedile
- $O_{i,2}$: montaggio delle gambe
- $O_{i,3}$: montaggio dei braccioli
- $O_{i,4}$: verniciatura



Grafo di realizzazione del prodotto

Grafo di realizzazione del prodotto



componenti di base

componenti intermedi/ semilavorati/parti

prodotto finito

Il grafo di realizzazione del prodotto per la generica classe di prodotti P_i:

- è un grafo orientato
- è un albero
- è un in-albero

Le operazioni

In riferimento all'operazione $O_{i,j}$ (j-esima operazione per la i-esima classe di prodotti) si definiscono:

- $\varphi_{i,j}$: funzione fisica
- $R_{i,j}$: insieme delle risorse che possono eseguire la funzione $\varphi_{i,j}$
- $s_{i,j}$: quantità di servizio
- $t_{i,j,k}$: tempo di esecuzione di $O_{i,j}$ sulla k-esima risorsa $\in R_{i,j}$
- $v_{i,j,k}$: velocità di esecuzione di $O_{i,j}$ sulla k-esima risorsa $\in R_{i,j}$

$$t_{i,j,k} = s_{i,j} / v_{i,j,k}$$

$t_{i,j,k}$ è una grandezza deterministica o stocastica

➔ è possibile definire un tempo di attesa massimo e minimo tra l'esecuzione di due operazioni consecutive

Risorse produttive

Le risorse produttive si classificano in

- risorse non condivisibili
- risorse multiservizio
- risorse condivisibili (*non considerate nel presente modello*)

r risorse produttive $\mathbf{R_1}, \ldots, \mathbf{R_r}$

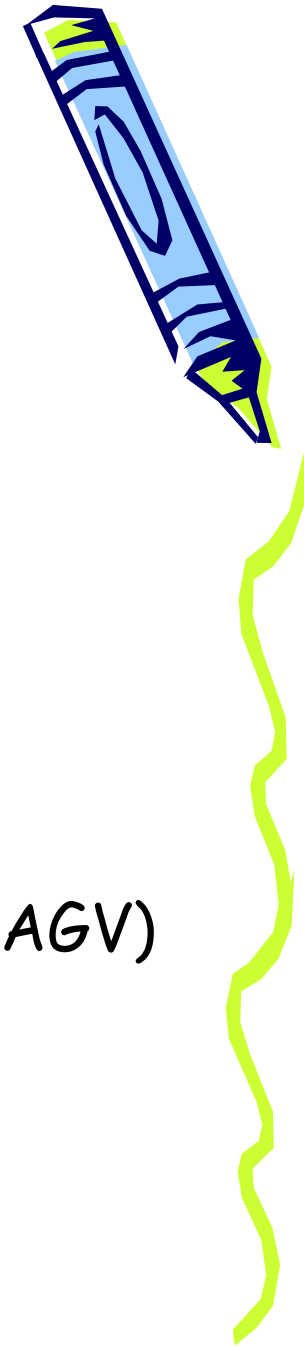
Si definiscono

- $\mathbf{v_{i,j,k}}$ = velocità di esecuzione di $\varphi_{i,j}$
- $sut_k(l,n)$ = ***tempo di attrezzaggio*** per passare dall'esecuzione della l-esima funzione fisica all'esecuzione della n-esima funzione fisica

$$sut_k(l,n) \neq sut_k(n,l)$$

$$sut_k(l,l) \neq 0$$

Risorse di trasporto

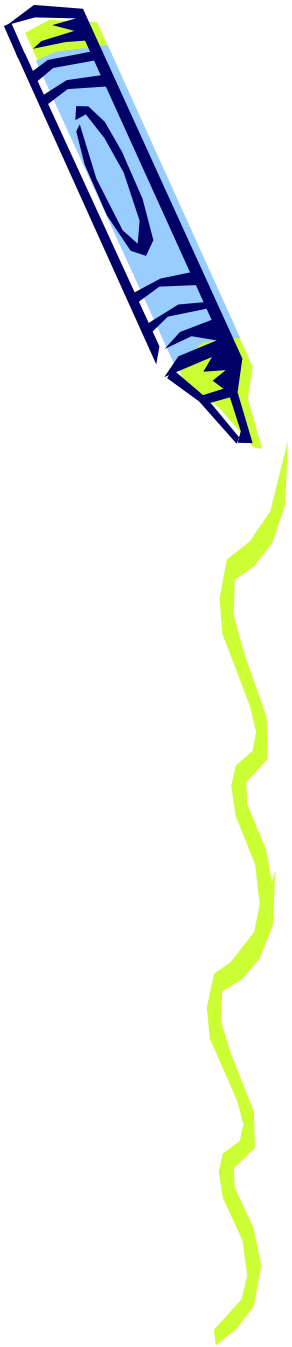


Le risorse di trasporto si classificano in

trasporto su supporto fisso
- ¬ percorso fissato
- ¬ capacità "infinita"

trasporto su veicoli (Automated Guided Vehicle-AGV)
- ¬ percorso non fissato
- ¬ capacità finita

Risorse di immagazzinamento

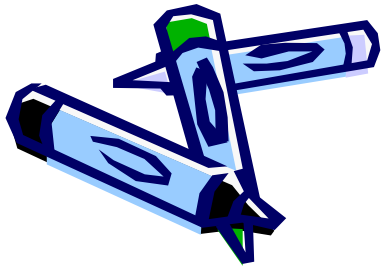


Le risorse di immagazzinamento si classificano in

magazzini

- di ingresso
- per semilavorati
- di uscita

buffer



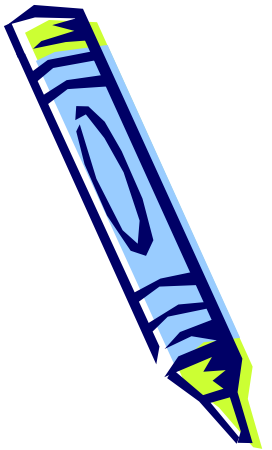
Indici di prestazione

Indici di prestazione legati alle commesse **(produzione make-to-order)**

Definizioni di base:

- $J_{i,j}$ → *job* relativo alla j-esima commessa della classe di prodotti Pi;
- $L_{i,j}$ → dimensione del lotto (*lot-size*) del job $J_{i,j}$;
- $dd_{i,j}$ → data di consegna (*due-date*) del job Ji,j ;
- $dl_{i,j}$ → *dead-line* del job Ji,j ;
- $c_{i,j}$ → istante di completamento (*completion time*) del job $J_{i,j}$;

Indici di prestazione



- $T_{i,j} \rightarrow$ ***tardiness*** del job $J_{i,j}$
  $T_{i,j} = \max\{c_{i,j} - dd_{i,j}, 0\}$

- $E_{i,j} \rightarrow$ ***earliness*** del job $J_{i,j}$
  $E_{i,j} = \max\{dd_{i,j} - c_{i,j}, 0\}$

- $a_{i,j} \rightarrow$ istante di arrivo (*arrival time*) dei componenti di base/semilavorati necessari per il job $J_{i,j}$;

- $F_{i,j} \rightarrow$ tempo di "esecuzione" (*flow time*) del job $J_{i,j}$, supponendo che il job non preveda assemblaggi:
  $F_{i,j} = c_{i,j} - a_{i,j}$

Indici di prestazione

Indici di prestazione legati al tempo di completamento delle commesse

Tardiness massima:

- $T_{i,max} = \max_{j=1,\dots,N} \{ \alpha_{i,j} \; T_{i,j} \}$

dove:

- **N** è il numero di job (commesse) previsti per la classe di prodotti Pi
- $\alpha_{i,j}$, j=1,…,N, sono pesi (coefficienti costanti)

La tardiness massima è indicativa della massima insoddisfazione del cliente.

Indici di prestazione

Indici di prestazione legati al tempo di completamento delle commesse

Earliness massima:

$$E_{i,max} = \max_{j=1,\dots,N} \{ \beta_{i,j} E_{i,j} \}$$

dove:

- **N** è il numero di job (commesse) previsti per la classe di prodotti Pi
- $\beta_{i,j}$, j=1,…,N, sono pesi (coefficienti costanti)

La earliness massima è indicativa del tempo massimo di immagazzinamento dei prodotti finiti.

Indici di prestazione

Indici di prestazione legati al tempo di completamento delle commesse

➔ **Tardiness media:**

$$\overline{T_i} = \frac{1}{N} \sum_{j=1}^{N} \alpha_{i,j} T_{i,j}$$

➔ **Earliness media:**

$$\overline{E_i} = \frac{1}{N} \sum_{j=1}^{N} \beta_{i,j} E_{i,j}$$

N.B.: Gli indici di prestazione legati a tardiness e earliness si possono calcolare solo se sono noti i completion times dei job

Indici di prestazione

Indici di prestazione legati al tempo di completamento delle commesse

➔ **Flow-time medio:**

$$\overline{F_i} = \frac{1}{N} \sum_{j=1}^{N} F_{i,j}$$

➔ **Flow-time massimo:**

$$F_{i,max} = \max_{j=1,..,N} \{F_{i,j}\}$$

N.B.: Il calcolo del flow-time si può estendere anche al caso in cui un job preveda degli assemblaggi riferendosi al tempo di lavorazione di ogni singolo semilavorato presente prima dell'assemblaggio.

Indici di prestazione

Indici di prestazione legati al tempo di completamento delle commesse

➔ **Completion-time medio:**

$$\overline{C_i} = \frac{1}{N} \sum_{j=1}^{N} C_{i,j}$$

➔ **Completion-time massimo (*makespan*):**

$$C_{i,max} = \max_{j=1,..,N} \{C_{i,j}\}$$

N.B.: Il makespan che, considerando una singola classe di prodotti, viene inteso come l'istante in cui l'impianto ritorna vuoto, viene spesso considerato come orizzonte temporale del problema di ottimizzazione da risolvere.

Indici di prestazione

Indici di prestazione legati all'utilizzo delle risorse

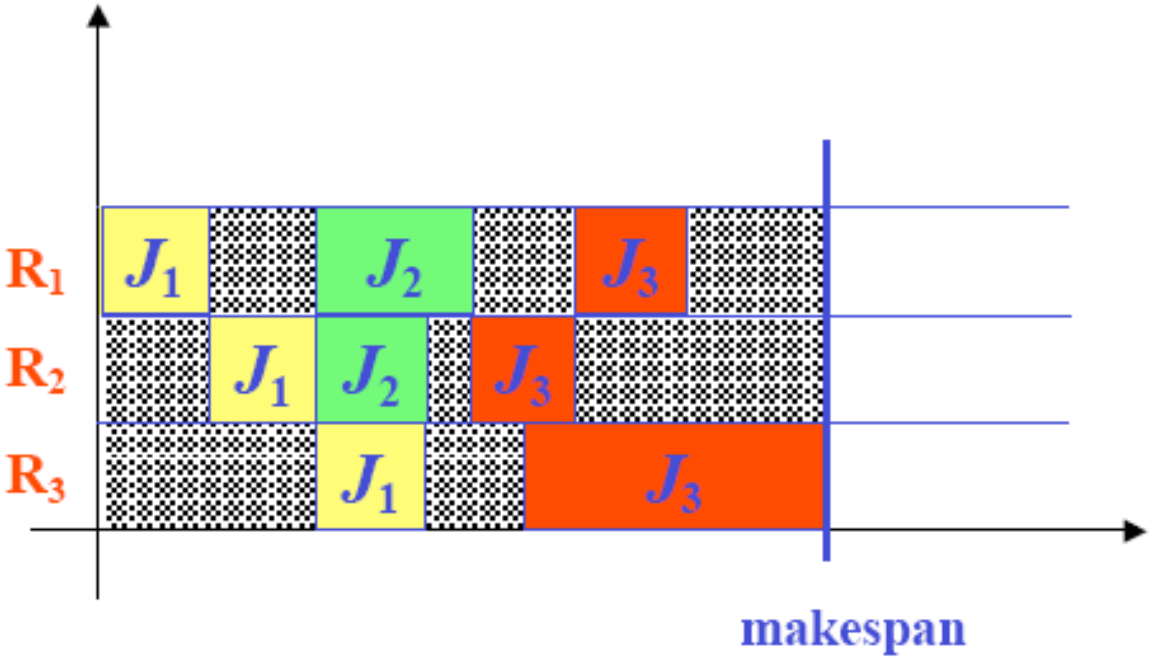


Diagramma di Gantt

inattività (*idleness*) della risorsa

Indici di prestazione

Indici di prestazione legati all'utilizzo delle risorse

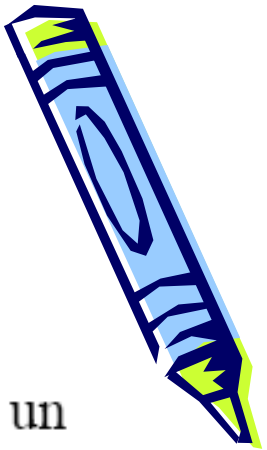
Sia

I_k = tempo totale durante il quale R_k è rimasta inattiva

- $I_{max} = \max_{k=1,..,K} \{I_k\}$
- $\bar{I} = \frac{1}{K} \sum_{k=1}^{K} I_k$

N.B.:
Questi stessi indici di prestazione possono essere espressi in termini percentuali definendo I_k come (tempo di inattività/makespan).

Indici di prestazione



Indici di prestazione legati all'inventory

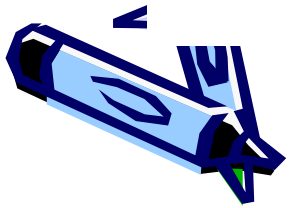
L'*inventory* o *work-in-process* o *work-in-progress* (*w.i.p.*)in un particolare istante, è l'insieme dei componenti di base, semilavorati e prodotti presenti in quell'istante nel sistema, inteso come l'impianto e tutti i magazzini.

Sia

k_i = numero di componenti di base necessari per realizzare prodotti di classe P_i;

$\gamma_{i,j}$ = costo unitario del j-esimo componente necessario per realizzare prodotti di classe P_i;

$Q_{i,j}(t)$ = quantità di items del j-esimo componente necessario per realizzare prodotti di classe P_i, presente nel sistema all'istante t (nel caso della sedia, ad esempio, è il numero di gambe presenti nel sistema all'istante t);



Indici di prestazione

Indici di prestazione legati all'inventory

➔ **inventory**

$$w.i.p.(\mathrm{t}) = \sum_{\mathrm{i}=1}^{\mathrm{P}} \sum_{\mathrm{j}=1}^{\mathrm{k_i}} \gamma_{\mathrm{i,j}} Q_{\mathrm{i,j}}(\mathrm{t})$$

➔ **costo di inventory complessivo**

$$C_{\mathrm{inv}} = \frac{1}{C_{\max}} \int_0^{C_{\max}} w.i.p.(\mathrm{t})\,\mathrm{dt}$$

Indici di prestazione

Indici di prestazione per sistemi di produzione di tipo make-to-stock

N.B. Nei sistemi di produzione make-to-order l'ottimizzazione avviene su orizzonte finito, nei sistemi di produzione make-to-stock o, addirittura, nella produzione continua, le decisioni vengono prese, invece, su orizzonte infinito

Definizioni di base:

$\overline{w}_i$ = inventory medio per la classe di prodotti P_i

$\overline{F}_i$ = flow – time medio per la classe di prodotti P_i

X_i = ***produttività*** (***throughput***) per la classe di prodotti P_i

Il throughput indica la quantità di prodotto di classe P_i, realizzata nell'unità di tempo.

Indici di prestazione

Indici di prestazione per sistemi di produzione di tipo make-to-stock

Se il sistema è in equilibrio (esauriti i transitori), esiste la ***legge di Little***:

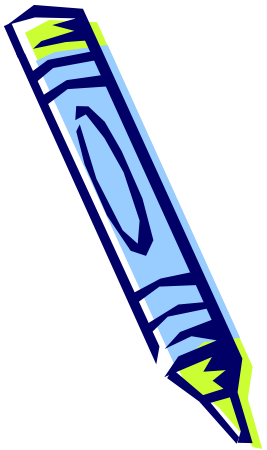
$$\overline{w}_i = \overline{F}_i X_i$$

Indici di prestazione:

- throughput
- work-in-progress medio
- flow-time medio

N.B. Si definiscono, in questo caso, anche vincoli sui cosiddetti **"mix" di produttività**, ossia sui rapporti tra i throughput delle diverse classi di prodotto (X_i / X_j)

Indici di prestazione

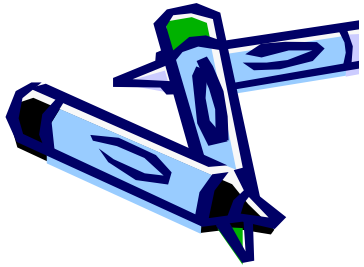


Altri costi legati alla produzione

- costo di **produzione** vero e proprio
- costo di "**lavoro straordinario**"

Altri costi

- costi legati alla **qualità del prodotto**
- costi legati all'**affidabilità del processo produttivo**
- costi legati al **progetto del processo di lavorazione**
- costi legati al **progetto dell'impianto produttivo**



Modello di produzione Discreto

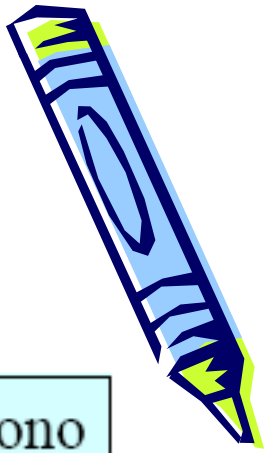
- **assegnazione** delle operazioni alle risorse
- **sequenziamento** delle operazioni sulle risorse
- **temporizzazione** dell'esecuzione delle operazioni

N.B. Nel modello make-to-stock, la definizione delle decisioni di cui sopra consente di tracciare il diagramma di Gantt.

Altri gradi di libertà

- istanti di arrivo dei lotti di semilavorati
- dimensione dei lotti (eventuale suddivisione delle commesse in "sottocommesse")
- ordinamento sequenziale di operazioni non compatibili

Formulazione di SED



I parametri e le variabili di stato associati alle entità possono essere sia grandezze deterministiche che variabili stocastiche

Es.: il tempo di servizio di un cliente su una risorsa in un sistema a coda può essere una costante, ma anche una variabile aleatoria con caratteristiche stocastiche definite a priori

Formulazione di SED

Ad ogni evento (tipo di evento) deve essere associata una procedura, detta ***transizione di stato***, che definisce come lo stato del sistema, ossia le variabili di stato associate alle entità, evolvono a causa dell'evento.

Nota: l'identificazione dei tipi di evento che caratterizzano il comportamento dinamico di un sistema avviene proprio con la verifica di quali condizioni facciano variare almeno una variabile di stato di una entità.

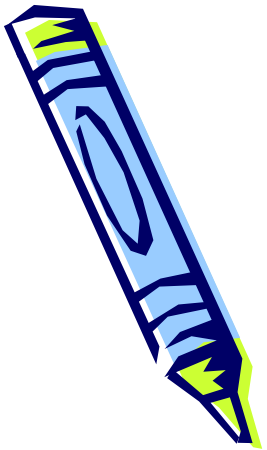
Formulazione di SED

Altre variabili necessarie in un modello ad eventi discreti sono:

variabili esogene: sono variabili **esterne** che costituiscono gli ingressi al modello. Si dividono in variabili controllabili (note e deterministiche) e non controllabili (stocastiche)

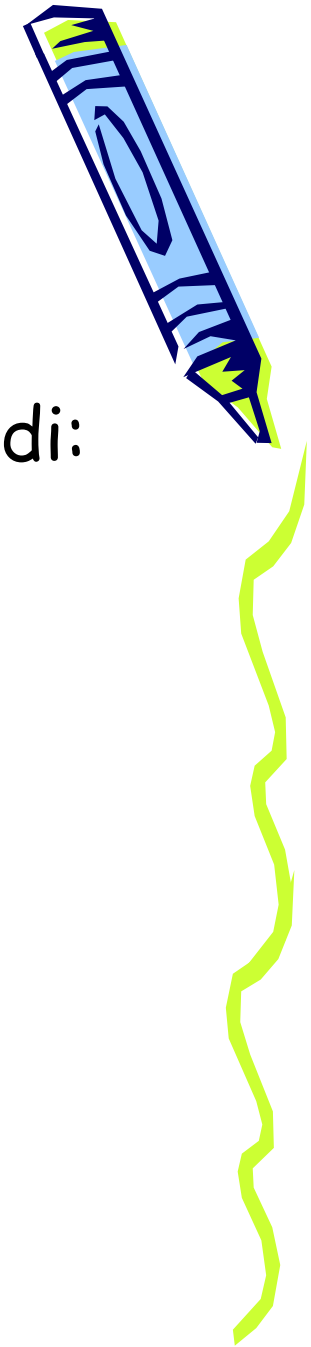
variabili endogene: sono variabili **interne** al modello, ossia generate dal modello. Sono necessarie per determinare le uscite del modello.

Formulazione di SED



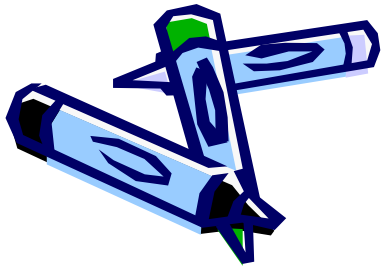
Nota: le variabili esogene non dipendono da altre variabili, le variabili di stato dipendono dalle variabili esogene, le variabili endogene dipendono sia dalle variabili di stato che dalle variabili esogene.

Formulazione di SED

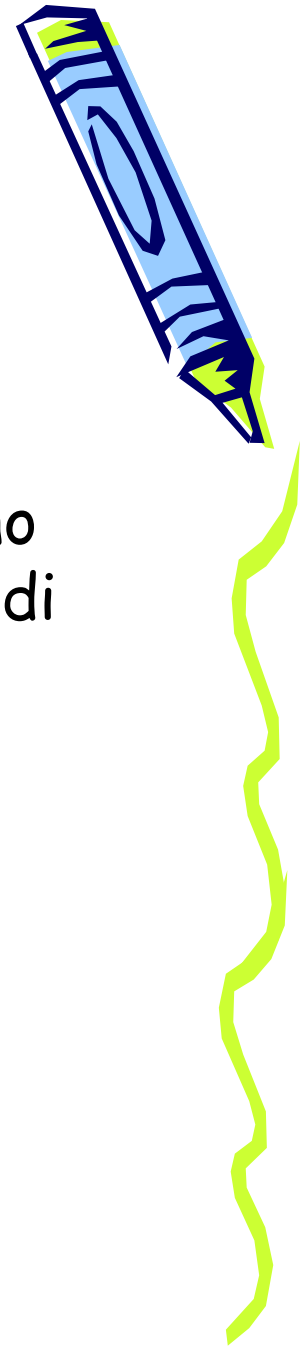


Un modello ad eventi discreti si compone di:

- **entità**: componente del sistema che richiede esplicita rappresentazione nel modello
- **eventi**: condizioni asincrone che determinano le transizioni di stato



Formulazione di SED

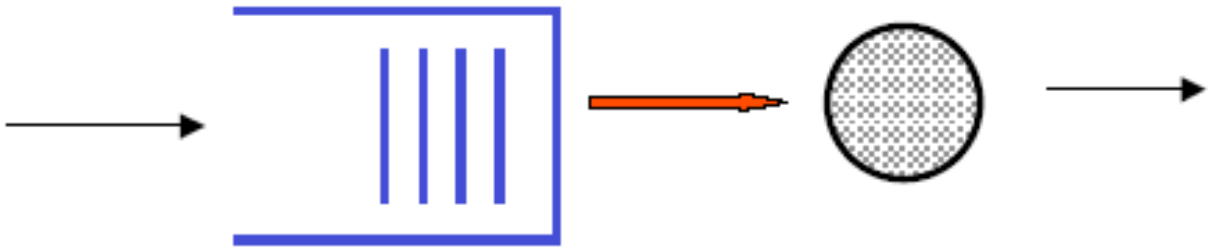


Ad ogni entità devono essere associati:

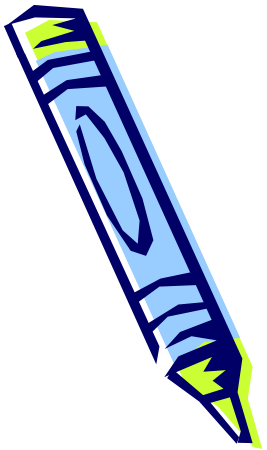
- **parametri**: grandezze **statiche** che definiscono caratteristiche fisiche, strutturali, operative di un'entità;
- **variabili di stato**: grandezze **dinamiche** che caratterizzano l'evoluzione dell'entità in conseguenza degli eventi.

Esempio
Sistema a coda

Il sistema è costituito da una risorsa (server) che deve servire un insieme di clienti. Il server è dotato di una coda con politica di gestione di tipo ***First In First Out*** (**FIFO**)



Esempio
Sistema a coda



Il modello del sistema prevede due entità:

cliente

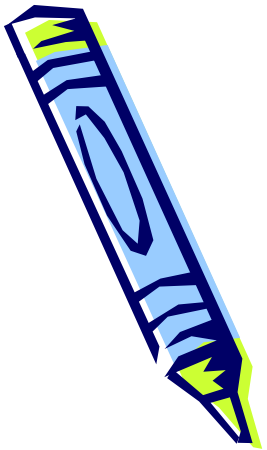
server

e due eventi:

arrivo cliente (**evento tipo 1**)

partenza cliente (**evento tipo 2**)

Esempio
Sistema a coda



Entità cliente:

parametri: tempo di interarrivo U(0.0,2.0)

var. stato: posizione {assente; in coda; in servizio}

Entità server:

parametri: tempo di servizio U(0.0,2.0)

var. stato: stato {libero; occupato}

n° clienti presenti (in coda e in servizio) {numeri interi}

lunghezza della coda {numeri interi}

Esempio
Sistema a coda

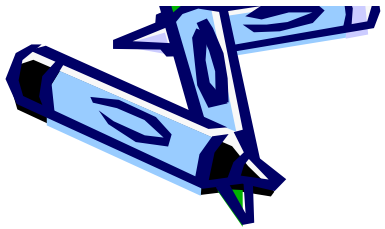
Transizione di stato evento tipo 1 (arrivo cliente)

server.nclienti := server.nclienti+1;
IF server.stato = “libero” THEN
    server.stato := “occupato”;
    cliente.posizione := “in servizio”;
    tserv := <generazione var. stocastica U(0.0,2.0)>;
    <schedulazione evento tipo 2 in t = tempo attuale+tserv>
ELSE
    cliente.posizione := “in coda”;
    server.lcoda := server.lcoda+1;
END
tarr := <generazione var. stocastica U(0.0,2.0)>;
<schedulazione evento tipo 1 in t = tempo attuale+tarr>.

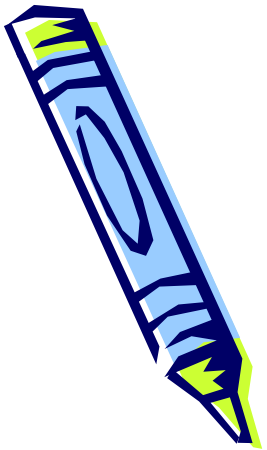
Esempio
Sistema a coda

Transizione di stato evento tipo 2 (partenza cliente)

```
server.stato := “libero”;
cliente.posizione := “assente”;
IF (server.lcoda > 0) THEN
        server.stato := “occupato”;
        cliente:= <primo cliente in coda>;
        cliente.posizione := “in servizio”;
        server.lcoda := server.lcoda-1;
        tserv := <generazione var. stocastica U(0.0,2.0)>;
        <schedulazione evento tipo 2 in t = tempo attuale+tserv>;
END.
```



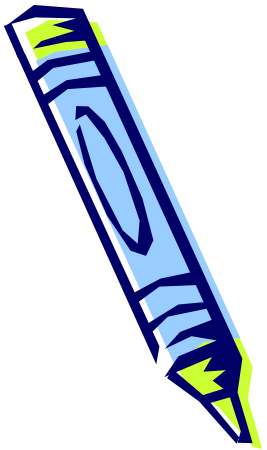
Individuazione degli eventi



analisi longitudinale: si descrive la dinamica del sistema in termini di flussi di entità transienti (es. flusso pezzi in un sistema di produzione)

analisi trasversale: si descrive la dinamica del sistema in termini di cicli di entità residenti (es. macchine in una sistema di produzione)

Individuazione degli eventi



Analisi longitudinale (per processi)

La sequenza degli eventi è descritta dal punto di vista delle entità transienti, ossia vengono indicate le azioni subite da queste. Tipicamente, ogni volta che viene completata un'attività, gli attributi delle entità transienti permettono di stabilire gli eventi futuri e modificano lo stato delle entità residenti, viste come risorse.

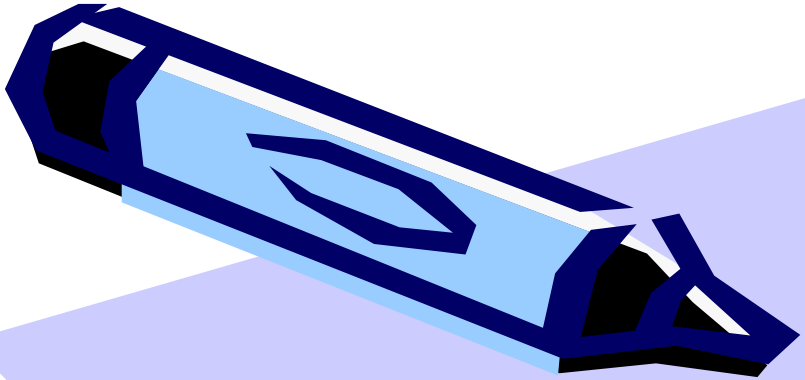
Analisi trasversale (per attività)

La sequenza degli eventi è descritta dal punto di vista delle entità residenti, ossia vengono indicati la sequenza dei cicli che ognuna di esse esegue ed i metodi di interazione con le altre entità del sistema.

Tipicamente, ogni volta che viene completato un ciclo, gli attributi dell'entità residente permettono di stabilire gli eventi futuri e modificano lo stato delle entità transienti.

Individuazione degli eventi

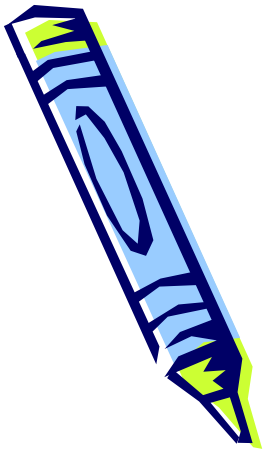
Conviene sempre eseguire entrambi i tipi di analisi (longitudinale e trasversale); il procedimento può essere laborioso, ma ha il vantaggio di rappresentare le relazioni **causa-effetto** all'interno del sistema da analizzare.



Analisi delle prestazioni tramite l'approccio simulativo:

i concetti fondamentali della simulazione ad eventi discreti

Vantaggi della simulazione

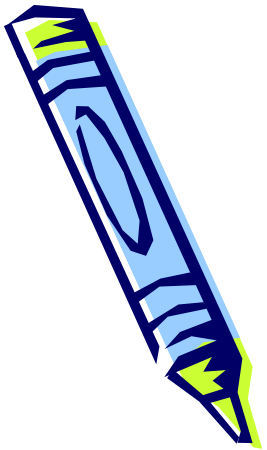


- lo stesso modello può essere utilizzato più volte;
- lo studio che precede la simulazione permette di acquisire conoscenza sul sistema;
- i modelli simulativi possono essere più semplici da utilizzare e non richiedono assunzioni semplificative;
- a volte i modelli simulativi sono gli unici applicabili.

Svantaggi della simulazione

- realizzare e validare un simulatore è molto costoso in termini di tempo e denaro;
- ottenere risultati statisticamente significativi può richiedere personale esperto e l'esecuzione di lunghe e numerose simulazioni;
- i modelli analitici consentono di analizzare in modo esatto il comportamento del sistema.

Scopo della simulazione



- Analisi di sistemi "ipotetici"
- Confronto tra due o più sistemi
- Determinare valore ottimale di parametri
- Determinare i punti critici (bottlenecks)
- Capacity planning
- Predire le prestazioni del sistema nel futuro

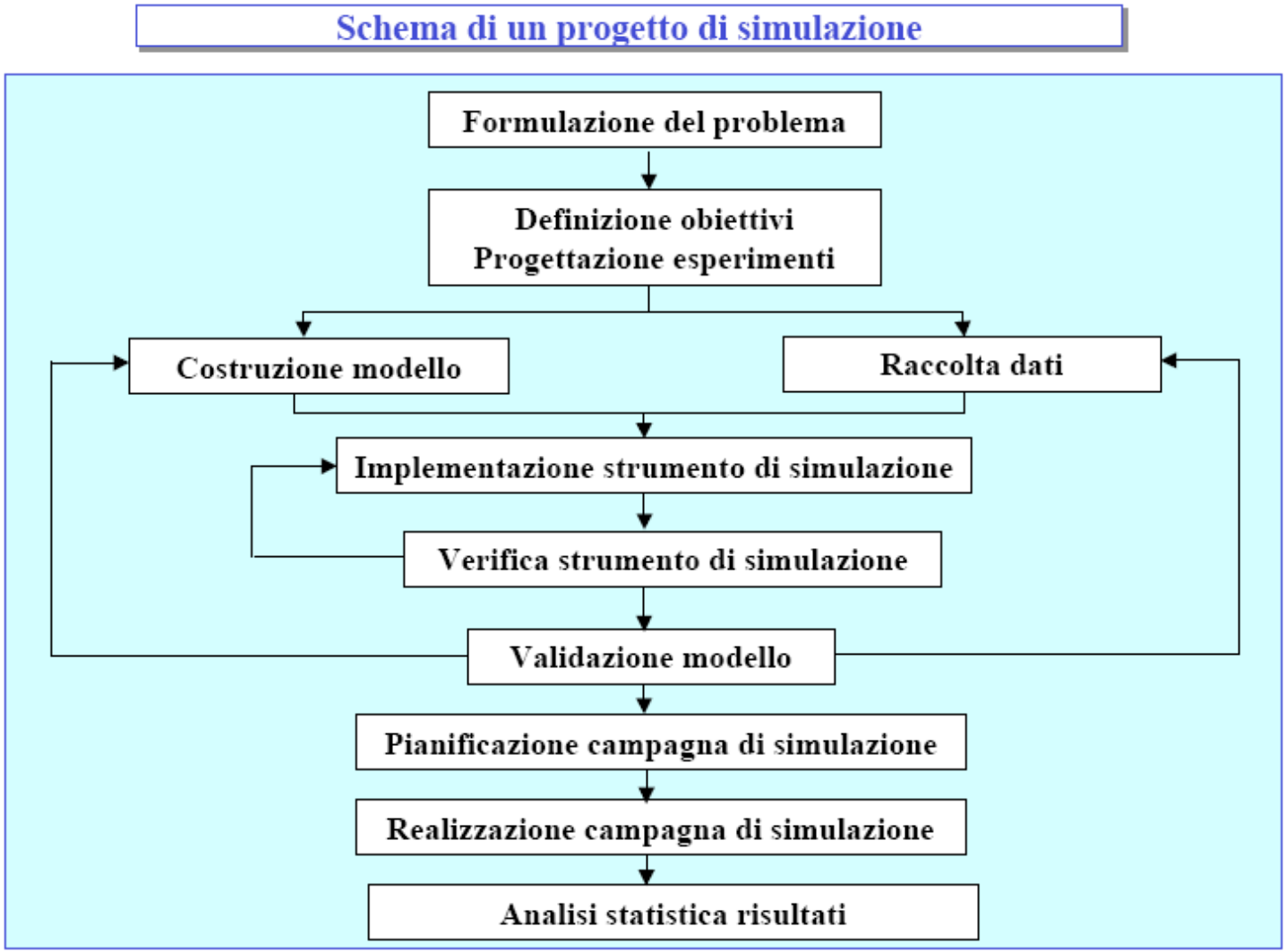
Esempi di applicazione

Sistemi di produzione: determinazione prestazioni del sistema (throughput, earliness/tardiness media e massima, ecc.), gestione magazzini e distribuzione (politiche di riordino, spedizione e consegna ottimali), pianificazione della produzione (determinazione dimensione macchine, buffer, squadre operai, sequenze di lavorazione e manutenzione, ecc.)

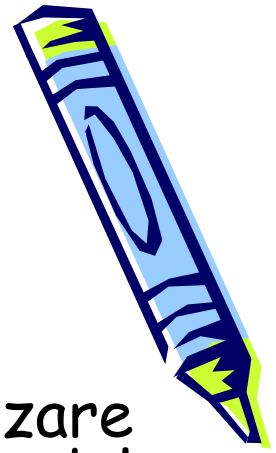
Sistemi di trasporto: determinazione prestazioni del sistema, definizione orari/cadenzamenti, ecc.

Aziende: definizione politiche di gestione ottimali

Progetto di Simulazione

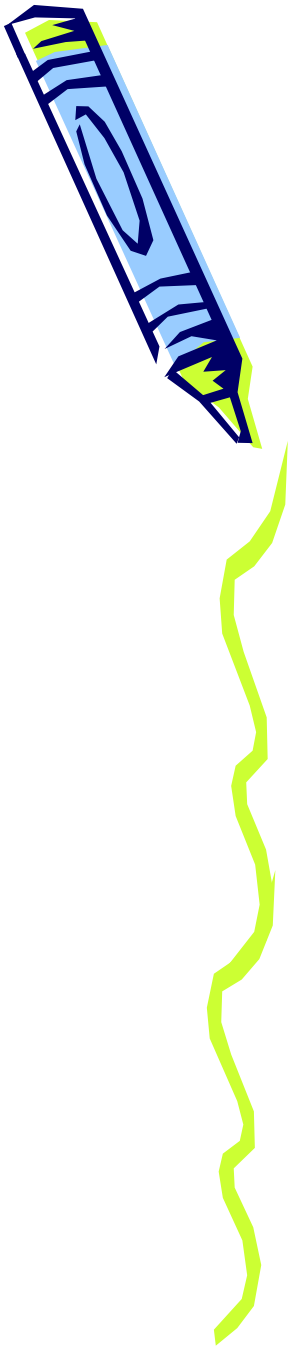


Costruzione del modello



- Individuazione del tipo di modello da utilizzare per rappresentare il comportamento dinamico del sistema reale
- Sviluppo del modello del sistema
- Individuazione delle componenti stocastiche all'interno del modello
- Identificazione delle caratteristiche statistiche delle componenti stocastiche inserite nel modello

Implementazione



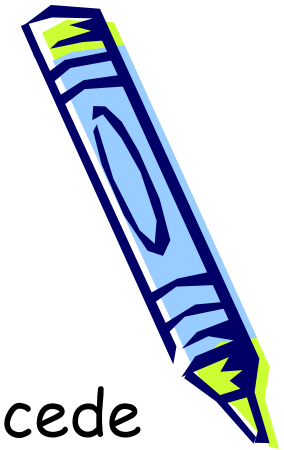
Linguaggi di programmazione

- costi inferiori
- tempo di sviluppo maggiore
- velocità maggiore

Strumenti di simulazione

- costi superiori
- tempo di sviluppo inferiore
- velocità minore

Schema simulativo "event-oriented"

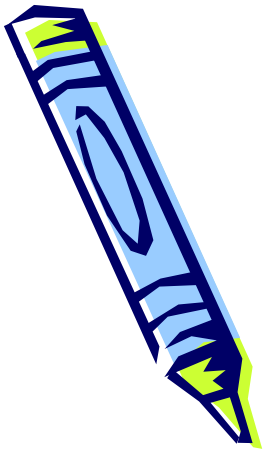


Lo schema simulativo **"event-oriented"** procede producendo una lista di *"fotografie"* del sistema in istanti di tempo discreti. Tali istanti di tempo sono gli **istanti di occorrenza degli eventi**.

Ogni fotografia del sistema contiene:

- lo stato del sistema
- una lista di tutti gli eventi già schedulati (*lista eventi attivi*)
- i contatori progressivi per il calcolo delle statistiche (var. endogene)

Schema "event-oriented"



Ad ogni istante di occorrenza di un evento si definisce:

Istante tempo	Stato entità 1	...	Stato entità n	Lista eventi attivi (LEA)	Var. endogene

Il meccanismo per fare avanzare correttamente la simulazione e garantire che tutti gli eventi avanzino nell'ordine cronologico corretto, si basa sulla costruzione e sulla scansione della lista degli eventi attivi.

Schema "event-oriented"

La **lista degli eventi attivi (*Scheduled Event List-SEL*)** all'istante t contiene tutti gli eventi già schedulati per accadere nel futuro con i corrispondenti tempi di occorrenza.

LEA(t)

t_1	t_2	t_3	...
e_1	e_2	e_3	...

N.B.
La lista degli eventi attivi è ordinata per tempi di occorrenza degli eventi non decrescenti
(e.g. $t \leq t_1 \leq t_2 \leq t_3 \leq \ldots$)

Schema "event-oriented"

La procedura di simulazione consiste nella ripetizione iterativa dei seguenti passi:

1 Rimozione del primo elemento in LEA (t1,e1)

2 Aggiornamento del tempo di simulazione a t1

3 Aggiornamento dello stato del sistema a x' = f(x, e1)

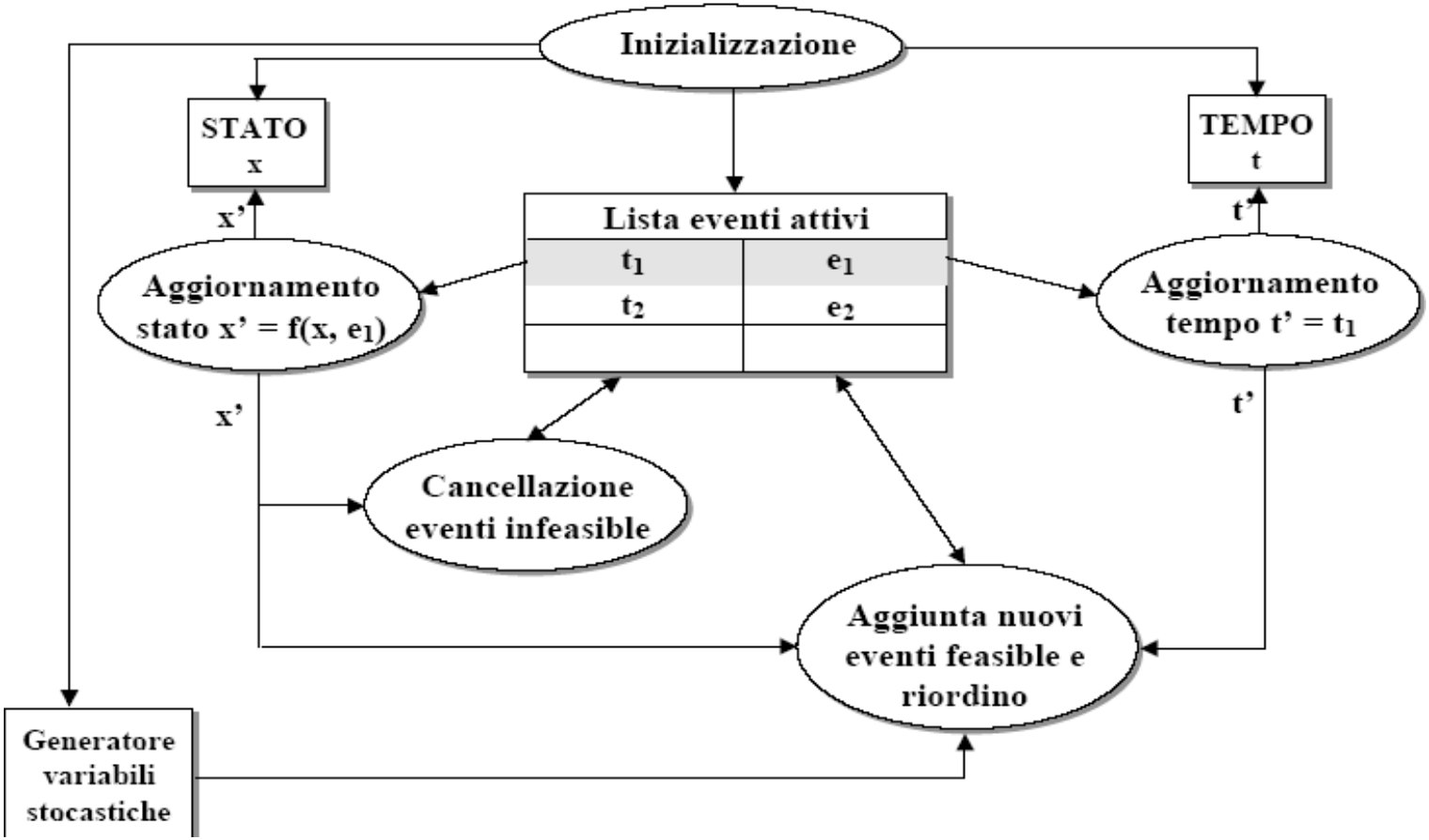
(f rappresenta la transizione di stato relativa all'evento e1)

4 Cancellazione da LEA di eventi resi non più feasible dall'evento e1

5 Aggiunta in LEA di eventi resi feasible dall'evento e1

6 Ordinamento della LEA in ordine cronologico crescente

Schema "event-oriented"



Commenti

- La fase di inizializzazione della LEA consiste nell'inserimento nella lista di tutti gli eventi che possono essere schedulati all'inizio della simulazione

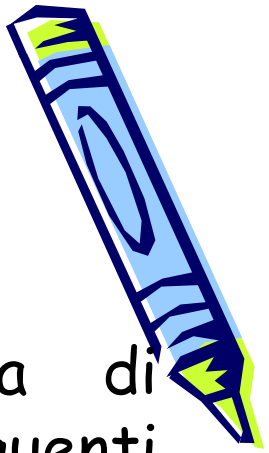
- La procedura di simulazione termina quando incontra nella LEA l'evento ***fine simulazione***

- L'evento fine simulazione può essere inserito in LEA nella fase di inizializzazione (si conosce a priori la durata della simulazione) oppure durante la simulazione quando si verificano particolari condizioni (es.: è stato processato un numero di clienti predefinito oppure il sistema ha raggiunto una situazione di regime)

Commenti

L'implementazione della LEA deve essere eseguita garantendo la massima efficienza delle operazioni di: inserimento di un elemento in lista, ricerca di un elemento in lista, cancellazione di un elemento dalla lista, riordino della lista.

Componenti principali



Uno strumento di simulazione con schema di avanzamento event-oriented si compone dei seguenti elementi:

- **stato** : struttura dati di memorizzazione dello stato
- **tempo** : variabili di memorizzazione del tempo reale e del tempo simulato
- **lista degli eventi attivi** : struttura dati di implmenetazione della lista ordinata degli eventi attivi e dei loro tempi di occorrenza
- **registri di dati** : variabili o liste di memorizzazione delle grandezze necessarie per il calcolo delle uscite della simulazione (variabili endogene)

Componenti

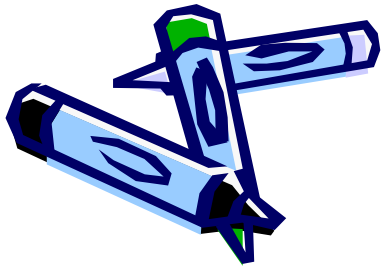
- **procedura d'inizializzazione**: inizializza variabili e registri, inizializza generatore di variabili stocastiche, definisce LEA(0)
- **procedura aggiornamento tempo**
- **procedura aggiornamento stato**
- **generatore variabili stocastiche**: insieme di procedure che trasformano i numeri random generati dal calcolatore stocastiche con opportune distribuzioni di probabilità
- **modulo di generazione stime di uscita e report**
- **modulo principale**: implementa il ciclo di avanzamento simulazione e coordina il programma di simulazione

Schema: process-oriented

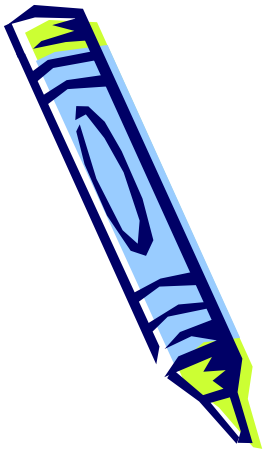
Lo schema simulativo "**process-oriented**" interpreta l'evoluzione di un sistema ad eventi discreti come l'insieme delle attività delle entità che compongono il sistema.

Più precisamente lo schema simulativo per processi definisce, per ogni entità di tipo transiente, un "**processo**" inteso come **sequenza di attività** (servizi, oppure attese in coda) di tale entità.

Ogni attività è costituita da una **sequenza di eventi separati da intervalli di tempo** (ad es. l'attività servizio è la sequenza degli eventi inizio servizio e fine servizio separati dalla durata del servizio stesso).



Schema: process-oriented

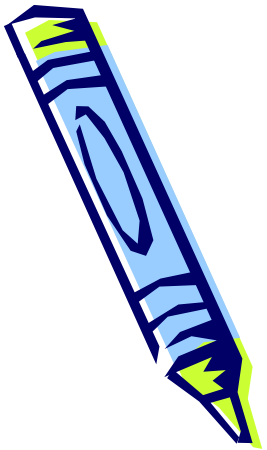


- Un processo è una sequenza di attività

Le attività si distinguono in:

- **attività logiche**: sono azioni istantanee (verifica di condizioni, aggiornamento di strutture dati, ecc.)
- **attività temporizzate**: sono attività caratterizzate da una durata. Si distinguono in:
  - attività con durata predefinita: la durata dell'attività è prefissata (può anche essere una variabile stocastica)
  - attività con durata non predefinita: la durata dell'attività dipende dallo stato del sistema (es. tempo di attesa in coda)

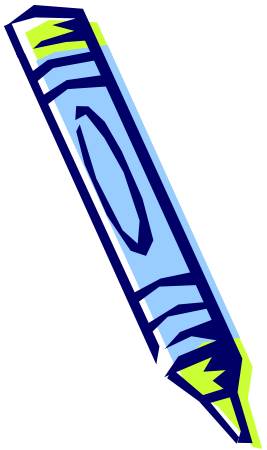
Schema: process-oriented



Gli elementi che compongono un modello ad eventi discreti da utilizzare in uno schema simulativo per processi sono:

- **entità**: sono intese, in questo caso, come gli elementi che richiedono un servizio;
- **attributi, variabili di stato**: hanno lo stesso significato che nel caso dello schema event-oriented;
- **attività o funzioni di processo**: sono le azioni istantanee o temporizzate che le entità eseguono;
- **risorse**: sono gli elementi che forniscono i servizi richiesti dalle entità;
- **code**: locazioni fisiche o virtuali dove le entità attendono servizi da parte delle risorse.

Schema: process-oriented



Il flusso di esecuzione di un processo in esecuzione emula il flusso di un oggetto attraverso il sistema

- L'esecuzione procede finchè il flusso non viene bloccato o entra in una nuova attività
- Attesa in coda, servizio (ritardo)
- Quando il flusso di un'entità viene bloccato, il tempo simulato avanza al tempo di inizio previsto dalla prima successiva entità in esecuzione

Esempio: coda

- Il sistema si compone di un unico server con una sola classe di clienti.
- La simulazione inizia a t = 0 e con il sistema vuoto (server libero, lunghezza della coda = 0)
- La simulazione termina quando sono stati processati *N* clienti (N è predefinito)
- Obiettivo della simulazione è la stima delle seguenti quantità:
  - tempo medio atteso di permanenza nel sistema (***expected average system time***)
  - probabilità che il tempo di permanenza di un cliente nel sistema ecceda una "deadline" *d*
  - utilizzo del server
  - lunghezza media della coda

Esempio: coda

Il tempo medio di permanenza dei clienti nel sistema è una variabile stocastica che dipende dal tempo di interarrivo dei clienti e dalle sequenze di servizio associate ad essi. Sia W_N tale variabile.

In ogni esecuzione (***run***) della simulazione si misurano i tempi di permanenza di ogni cliente nel sistema, indicati con W_1, W_2, …, W_N e si calcola un'istanza dello stimatore: la media aritmetica delle N osservazioni

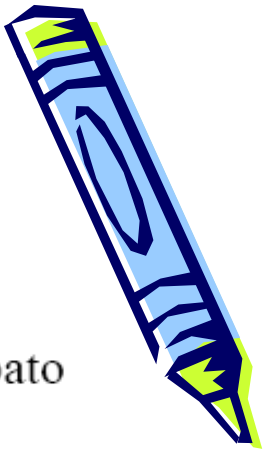
Esempio: coda

La probabilità che il tempo di permanenza di un cliente nel sistema ecceda una deadline *dd* è una misura del grado di soddisfazione del cliente (o del livello di servizio del sistema).

Tale grandezza è indicata con p_N^d e, indicando con n_N il numero di clienti il cui tempo di permanenza nel sistema supera d, può essere stimata con lo stimatore

$$\hat{p}_N^d = \frac{n_N}{N}$$

Esempio: coda



Con utilizzo del server si intende la probabilità che il server sia occupato durante il tempo globale di servizio degli N clienti.

Tale grandezza è indicata con ρ_N ed è equivalente alla frazione di tempo durante la quale il numero di clienti nel sistema è positivo.

Definendo:

N(i) = tempo totale osservato durante il quale il numero di clienti nel sistema è i, i ≥0.

$$N_N = \sum_{i=0}^{\infty} N(i)$$

Lo stimatore di ρ_N è calcolabile come

$$\hat{\rho}_N = \frac{\sum_{i=1}^{\infty} N(i)}{N_N} = 1 - \frac{N(0)}{N_N}$$

Esempio: coda

Sia Q_N la lunghezza media della coda nel tempo necessario a servire gli N clienti e sia $p_N(i)$ la probabilità che la lunghezza della coda sia i durante lo stesso intervallo di tempo.

Allora:

$$Q_N = \sum_{i=0}^{\infty} i\, p_N(i)$$

Definendo:

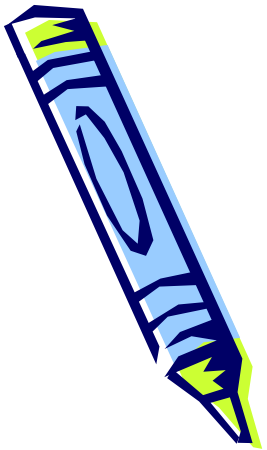
T(i) = tempo totale osservato durante il quale la lunghezza della coda è i, i ≥0.

$$T_N = \sum_{i=0}^{\infty} T(i)$$

Uno stimatore di $p_N(i)$ è

$$\hat{p}_N(i) = \frac{T(i)}{T_N} \quad \rightarrow \quad \hat{Q}_N = \frac{1}{T_N} \sum_{i=0}^{\infty} i\, T(i)$$

Esempio: coda



Dati simulazione:

- N=5;
- $d = 2.0$;
- tempi di interarrivo:

  $y_1 = 0.4, y_2 = 0.3, y_3 = 0.4, y_4 = 1.7, y_5 = 1.7, y_6 = 0.5;$

- tempi di servizio:

  $z_1 = 1.6, z_2 = 0.5, z_3 = 1.0, z_4 = 0.3, z_5 = 0.8.$

Esempio: coda

Procedura di inizializzazione:

(Oss.: consideriamo come stato in modo esplicito solo il numero di clienti presenti nel sistema)

tempo
0.0

stato
0

0.4
evento tipo 1

← LEA

Tempi di arrivo	
1	
2	
3	
4	
5	

Tempi di permanenza s_i	
1	
2	
3	
4	
5	

n_N
0

N(i)	
0	
1	
2	
3	
4	

T(i)	
0	
1	
2	
3	
4	

Esempio: coda

t=0.4: evento tipo 1 → arrivo di un cliente nel sistema

tempo
0.4

stato
1

0.7	2.0
evento tipo 1	evento tipo 2

← LEA

Tempi di arrivo	
1	0.4
2	
3	
4	
5	

Tempi di permanenza s_i	
1	
2	
3	
4	
5	

n_N
0

N(i)	
0	0.4
1	
2	
3	
4	

T(i)	
0	0.4
1	
2	
3	
4	

Esempio: coda

t=0.7: evento tipo 1 → arrivo di un cliente nel sistema

tempo
0.7

stato
2

1.1	2.0
evento tipo 1	evento tipo 2

← LEA

Tempi di arrivo	
1	0.4
2	0.7
3	
4	
5	

Tempi di permanenza s_i	
1	
2	
3	
4	
5	

n_N
0

N(i)	
0	0.4
1	0.3
2	
3	
4	

T(i)	
0	0.7
1	
2	
3	
4	

Esempio: coda

t=1.1: evento tipo 1 → arrivo di un cliente nel sistema

tempo	stato
1.1	3

2.0	2.8
evento tipo 2	evento tipo 1

← LEA

	Tempi di arrivo
1	0.4
2	0.7
3	1.1
4	
5	

	Tempi di permanenza s_i
1	
2	
3	
4	
5	

n_N
0

	N(i)
0	0.4
1	0.3
2	0.4
3	
4	

	T(i)
0	0.7
1	0.4
2	
3	
4	

Esempio: coda

t=2.0: evento tipo 2 → partenza di un cliente dal sistema

tempo	stato
2.0	2

2.5	2.8
evento tipo 2	evento tipo 1

← LEA

	Tempi di arrivo
1	0.4
2	0.7
3	1.1
4	
5	

	Tempi di permanenza s_i
1	1.6
2	
3	
4	
5	

n_N
0

	N(i)
0	0.4
1	0.3
2	0.4
3	0.9
4	

	T(i)
0	0.7
1	0.4
2	0.9
3	
4	

Esempio: coda

t=2.5: evento tipo 2 → partenza di un cliente dal sistema

tempo	stato
2.5	1

2.8	3.5
evento tipo 1	evento tipo 2

← LEA

	Tempi di arrivo
1	0.4
2	0.7
3	1.1
4	
5	

	Tempi di permanenza s_i
1	1.6
2	1.8
3	
4	
5	

n_N
0

	N(i)
0	0.4
1	0.3
2	0.9
3	0.9
4	

	T(i)
0	0.7
1	0.9
2	0.9
3	
4	

Esempio: coda

t=2.8: evento tipo 1 → arrivo di un cliente nel sistema

tempo
2.8

stato
2

3.5	4.5
evento tipo 2	evento tipo 1

← LEA

Tempi di arrivo	
1	0.4
2	0.7
3	1.1
4	2.8
5	

Tempi di permanenza s_i	
1	1.6
2	1.8
3	
4	
5	

n_N
0

N(i)	
0	0.4
1	0.6
2	0.9
3	0.9
4	

T(i)	
0	1.0
1	0.9
2	0.9
3	
4	

Esempio: coda

t=3.5: evento tipo 2 → partenza di un cliente dal sistema

tempo	stato
3.5	1

3.8	4.5
evento tipo 2	evento tipo 1

← LEA

	Tempi di arrivo
1	0.4
2	0.7
3	1.1
4	2.8
5	

	Tempi di permanenza s_i
1	1.6
2	1.8
3	2.4
4	
5	

n_N
1

	N(i)
0	0.4
1	0.6
2	1.6
3	0.9
4	

	T(i)
0	1.0
1	1.6
2	0.9
3	
4	

Esempio: coda

t=3.8: evento tipo 2 → partenza di un cliente dal sistema

tempo	stato
3.8	0

4.5
evento tipo 1

← LEA

	Tempi di arrivo
1	0.4
2	0.7
3	1.1
4	2.8
5	

	Tempi di permanenza s_i
1	1.6
2	1.8
3	2.4
4	1.0
5	

n_N
1

	N(i)
0	0.4
1	0.9
2	1.6
3	0.9
4	

	T(i)
0	1.3
1	1.6
2	0.9
3	
4	

Esempio: coda

t=4.5: evento tipo 1 → arrivo di un cliente nel sistema

tempo
4.5

stato
1

5.0	5.3
evento tipo 1	evento tipo 2

← LEA

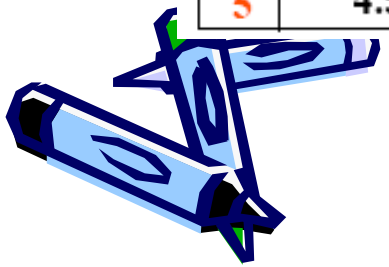
Tempi di arrivo	
1	0.4
2	0.7
3	1.1
4	2.8
5	4.5

Tempi di permanenza s_i	
1	1.6
2	1.8
3	2.4
4	1.0
5	

n_N
1

N(i)	
0	1.1
1	0.9
2	1.6
3	0.9
4	

T(i)	
0	2.0
1	1.6
2	0.9
3	
4	



Esempio: coda

t=5.0: evento tipo 1 → arrivo di un cliente nel sistema

tempo	stato
5.0	2

5.3	5.9
evento tipo 2	evento tipo 1

← LEA

Tempi di arrivo	
1	0.4
2	0.7
3	1.1
4	2.8
5	4.5

Tempi di permanenza s_i	
1	1.6
2	1.8
3	2.4
4	1.0
5	

n_N
1

N(i)	
0	1.1
1	1.4
2	1.6
3	0.9
4	

T(i)	
0	2.5
1	1.6
2	0.9
3	
4	

Esempio: coda

t=5.3: evento tipo 2 → partenza di un cliente dal sistema

tempo	stato
5.3	1

5.3	5.9
fine sim.	evento tipo 1

← LEA

Tempi di arrivo	
1	0.4
2	0.7
3	1.1
4	2.8
5	4.5

Tempi di permanenza s_i	
1	1.6
2	1.8
3	2.4
4	1.0
5	0.8

n_N
1

N(i)	
0	1.1
1	1.4
2	1.9
3	0.9
4	

T(i)	
0	2.5
1	1.9
2	0.9
3	
4	

Esempio: coda

t=5.3: fine simulazione

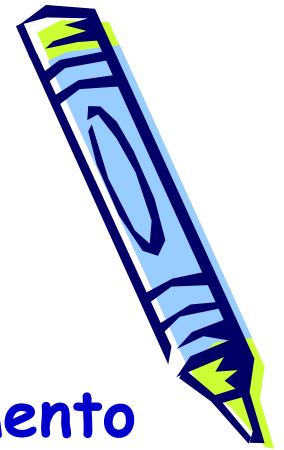
$$W_5 = \frac{(1.6+1.8+2.4+1.0+0.8)}{5} = 1.52$$

$$\hat{p}_5^d = \frac{1}{5} = 0.2$$

$$\hat{\rho}_5 = 1 - \frac{1.1}{5.3} = 0.79$$

$$\hat{Q}_5 = \frac{(1*1.9+2*0.9)}{5.3} = 0.699$$

Il tempo nella simulazione



Esistono tre diverse connotazioni di **avanzamento del tempo** nella simulazione:

- tempo reale: è il tempo dal punto di vista del sistema reale (è una variabile continua)
- tempo simulato: è il tempo dal punto di vista del modello (avanza per eventi)
- tempo di esecuzione: è il tempo dal punto di vista del processore, è il tempo di esecuzione del programma "simulatore"

Aspetti statistici nella simulazione

Richiami di statistica

Probabilità di un evento

Dato uno spazio campione $\Omega : \{ \omega_1 \ \omega_2 \ ,\ldots, \ \omega_m \}$, di cardinalità $|\Omega| = m$.

La probabilità è una funzione:

- $P( ) : \Omega \Rightarrow [0,1]$

che associa ad ogni evento elementare $\omega_i \in \Omega$ la sua probabilità di accadimento $P(\omega_i)$.

Sotto l'assunzione che gli elementi/eventi ωi possano o no avvenire, la probabilità $P(\omega_i)$, che viene associata all'evento ω_i, rappresenta la percentuale di volte che ci si aspetta che l'evento elementare accada.

Variabile aleatoria

- Ogni singolo evento ω di uno spazio campione delle prove Ω può essere associato in modo biunivoco a un numero attraverso una particolare funzione matematica:

$$X(\ ):\Omega \rightarrow \Re$$

- Tale corrispondenza si chiama variabile aleatoria è discreta se assume solo un insieme finito o almeno numerabile di valori (continua o mista altrimenti).

Esempio
lancio del dado

La v.a. corrispondente è una variabile X che può assumere solo valori discreti (o meglio interi positivi) nell'intervallo [1,6].

Per questo tipo di variabili aleatorie si definisce la probabilità di assumere una valore x_n come:

$$\Pr\{X = x_n\} = p(x_n) = p_n$$

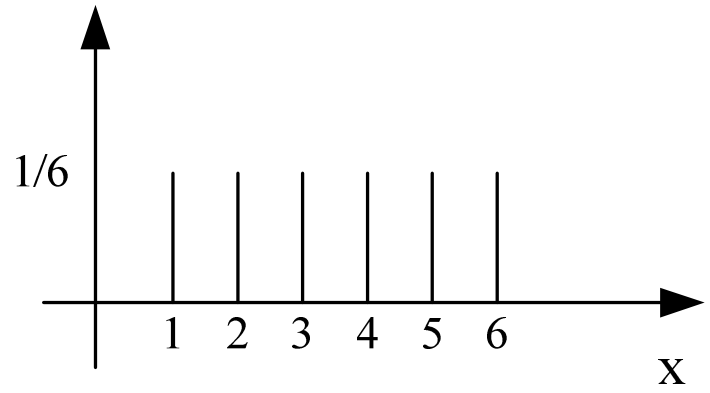
Chiaramente per tutti i valori che può assumere l'indice *n* vale la relazione di congruenza:

$$\sum_n p_n = 1$$

Esempio
lancio del dado

E' possibile rappresentare tutte le probabilità p_n mediante un grafico ed una funzione detta *funzione massa di probabilità* della v.a. discreta X.

Ad esempio per l'esperimento del lancio del dado si ha:

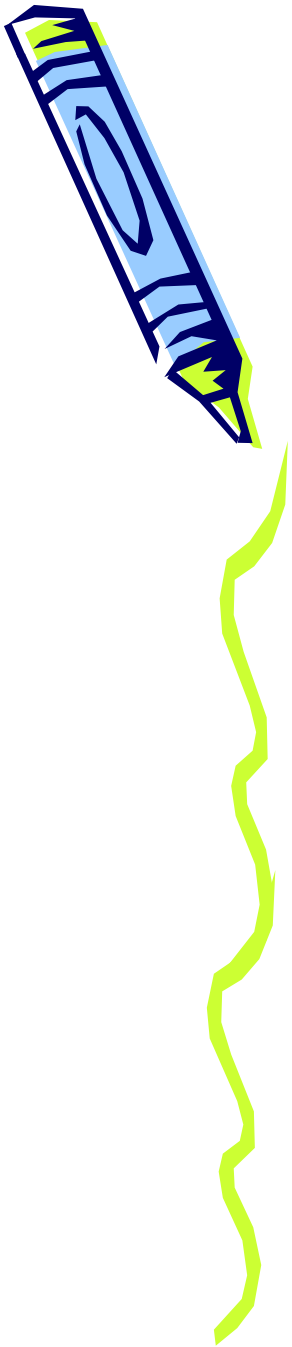


Distribuzione di probabilità

La funzione distribuzione di probabilità $F_X(x)$, definita sia per le variabili discrete che per quelle continue, molto spesso anche indicata con l'acronimo *cpf* (*Cumulative Probability Function*) e descrive come è distribuita sull'asse reale la probabilità dei valori assunti dalla v.a., essa viene definita dalla relazione:

$$F_X(x) = \Pr\{X \leq x\}$$

Distribuzione di probabilità



Per cui valgono le seguenti proprietà:

$$0 \le F_X(x) \le 1$$

$$F_X(-\infty) = \lim_{x \to -\infty} F_X(x) = 0$$

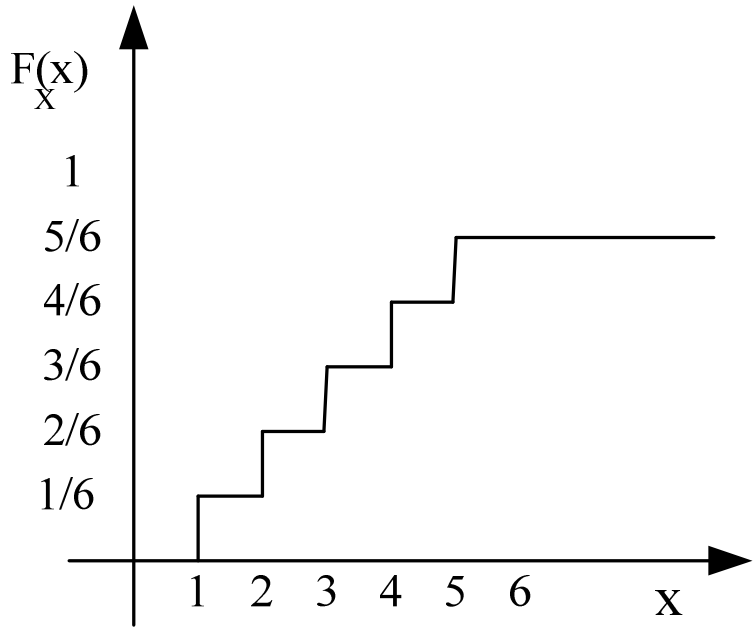
$$F_X(\infty) = 1$$

$$F_X(x_2) - F_X(x_1) = \Pr\{x_1 < X \le x_2\}$$

Distribuzione di probabilità

La funzione distribuzione di probabilità si ottiene dalla funzione massa di probabilità mediante la relazione:

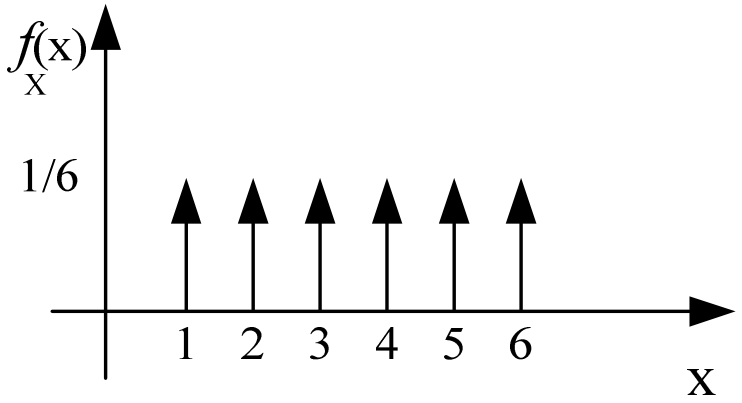
$$F_X(x) = \sum_n p_n u(x - x_n)$$



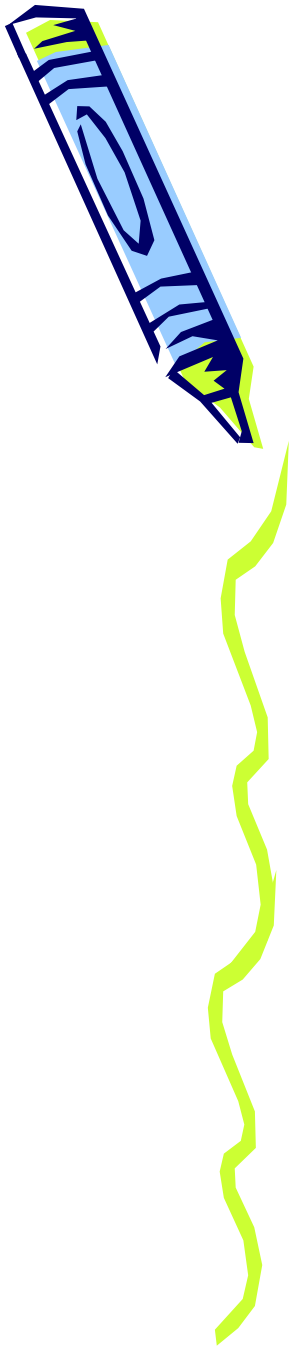
Funzione densità di probabilità

Per poter definire una funzione densità di probabilità anche per le v.a. discrete occorre ricorrere alla funzione delta di Dirac $\delta(x)$, per la quale si ha:

$$f_X(x) = \sum_n p_n \delta(x - x_n)$$



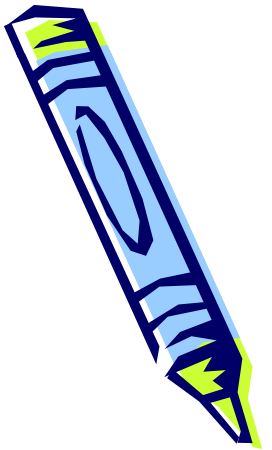
Distribuzione di probabilità



Lancio di un dado, con
x = ``numero impresso sulla faccia superiore'':

$$f(x) = 1/6 \qquad x = 1, 2, \ldots, 6.$$

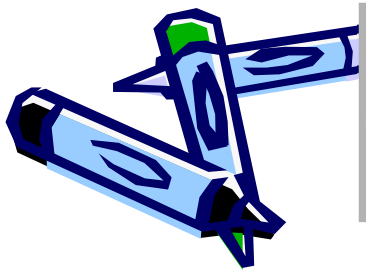
Distribuzione di probabilità



Lancio di due dadi, con
x = ``somma dei due risultati''

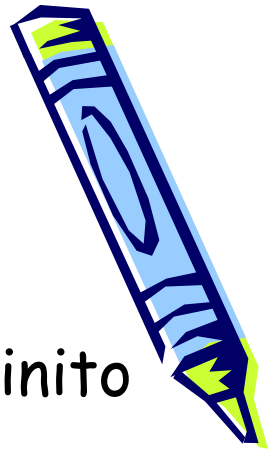
$$f(2) = 1/36 \qquad f(3) = 2/36$$

Volendo un'espressione sintetica (forma chiusa) per le probabilità (non è assolutamente necessario e non sempre fattibile) si può scrivere



$$f(x) = \frac{6 - |7 - x|}{36} \qquad x = 2, 3, \ldots, 12\,.$$

Valore atteso



è il *valor medio* che per una v.a. discreta è definito da:

$$E[X] = \sum_n p_n x_n$$

Anche se può non coincidere con nessuno dei valori assunti dalla v.a. è un parametro molto importate perché aiuta a capire il comportamento della funzione distribuzione di probabilità

Varianza

la *varianza*, detta anche indice di dispersione, offre un indice dell'addensamento dei valori della v.a. attorno al valor medio:

$$Var[X] = E[(X - E[X])^2] = \sum_n p_n (x_n - E[x])^2$$

La radice quadrata della varianza è anche denominata deviazione standard σ_X

Funzione generatrice di probabilità

Un utile ausilio nel calcolo dei momenti di una v.a. discreta è data dalla trasformata Z . Infatti è possibile definire la funzione generatrice di probabilità *P(z)* come:

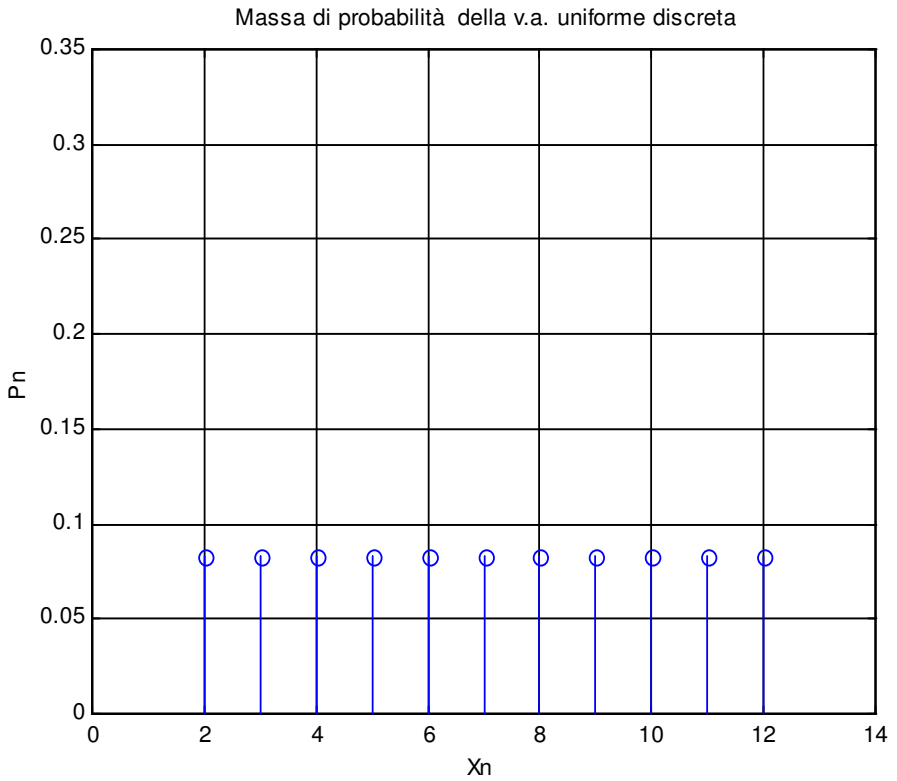
$$P(z) = E[z^X] \qquad P(z) = \sum_n p_n z^n$$

E' immediato verificare la corrispondenza tra la funzione generatrice di probabilità e la trasformata Z della sequenza numerica della funzione massa di probabilità. La funzione generatrice permette di calcolare il valor medio e tutti i momenti di una v.a.

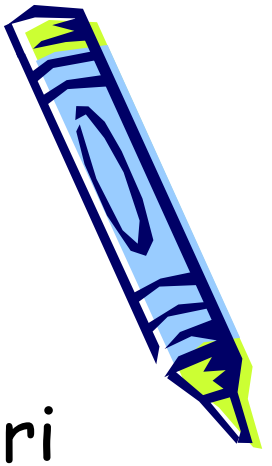
Variabile aleatoria uniforme discreta

La v.a. discreta uniforme ha la seguente massa di probabilità:

($k = 1, M = 11$)



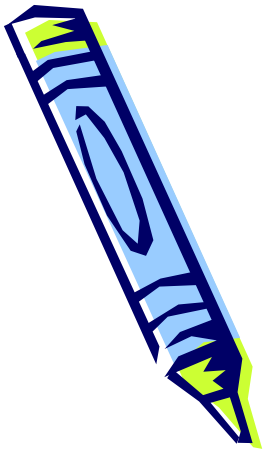
Variabile aleatoria uniforme discreta



E' definita su di un intervallo finito di valori $x_n \in [k+1, k+M]$, con valore $1/M$ su tutto l'intervallo. Per il calcolo del valor atteso si ha:

$$E[X] = \sum_n p_n x_n = \frac{1}{M} \sum_{n=k+1}^{k+M} n = k + \frac{1}{M} \sum_{n=1}^{M} n = k + \frac{(M+1)}{2}$$

Variabile aleatoria uniforme discreta



La varianza vale:

$$Var[X] = \sum_n p_n x_n^2 - E^2[X] = \frac{M^2 - 1}{12}$$

La funzione generatrice di probabilità:

$$P(z) = E[z^X] = \frac{z^{(k+1)}(1 - z^M)}{M(1 - z)}$$

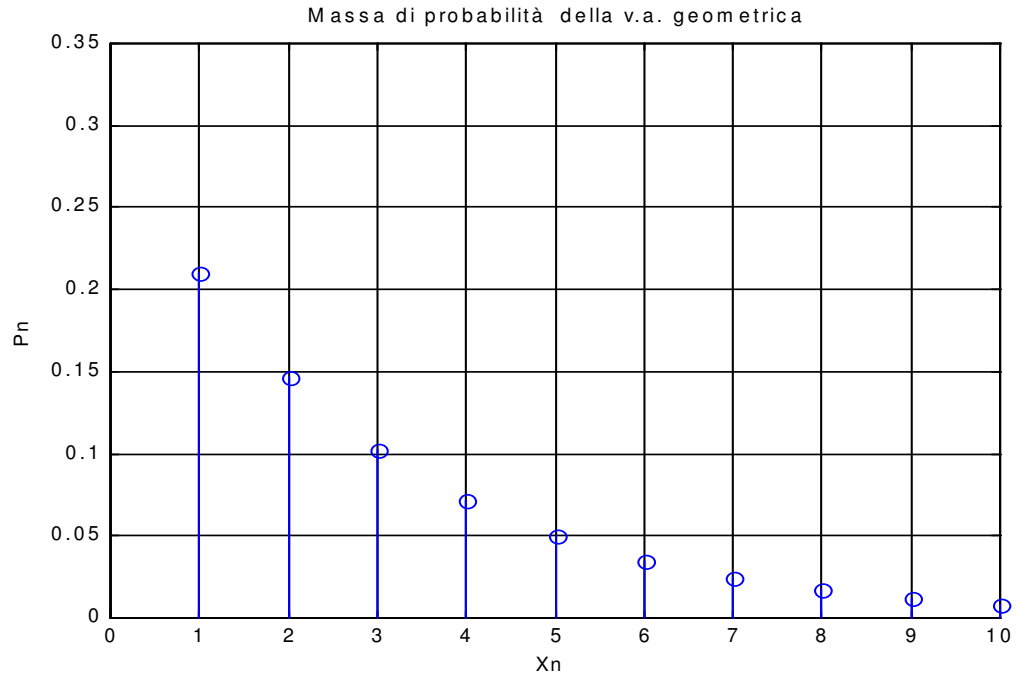
Variabile aleatoria geometrica

Per descrivere la v.a. geometrica supponiamo di effettuare un esperimento casuale, in cui è di interesse un evento specifico che si verifica con probabilità *p*. Questo evento viene anche definito *successo* dell'esperimento. La probabilità che questo evento si verifichi dopo *k* prove è

$$\Pr(successo \quad dopo \quad k \quad prove) = p(1-p)^k$$

Variabile aleatoria geometrica

Distribuzione geometrica ($k = 10$, $p = 0.3$)



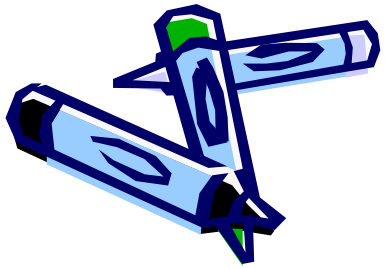
Variabile aleatoria geometrica

Il valor medio della distribuzione geometrica è:

$$E[X] = \sum_n p_n x_n = \sum_{n=0}^{+\infty} np(1-p)^n = \frac{1-p}{p}$$

Mentre la varianza e la funzione generatrice di probabilità sono:

$$Var[X] = \sum_n p_n x_n^2 - E^2[X] = \sum_n n^2 p(1-p)^n - \left(\frac{1-p}{p}\right)^2 = \frac{1-p}{p^2}$$



$$P(z) = E[z^X] = \sum_n z^n p(1-p)^n = \frac{p}{1-z(1-p)}$$

Variabile aleatoria binomiale

Consideriamo un semplice esperimento che può avere solo due risultati: *successo* con probabilità p ed insuccesso con probabilità $q = 1 - p$. Ad esempio il lancio di una moneta con p la probabilità che si abbia testa e q la probabilità che si abbia croce. Considerati N lanci indipendenti, la probabilità che si abbia testa n volte è data da:

$$\Pr\{n \quad volte \quad testa \quad su \quad N \quad lanci\} = \binom{N}{n} p^n q^{(N-n)} \qquad p + q = 1$$

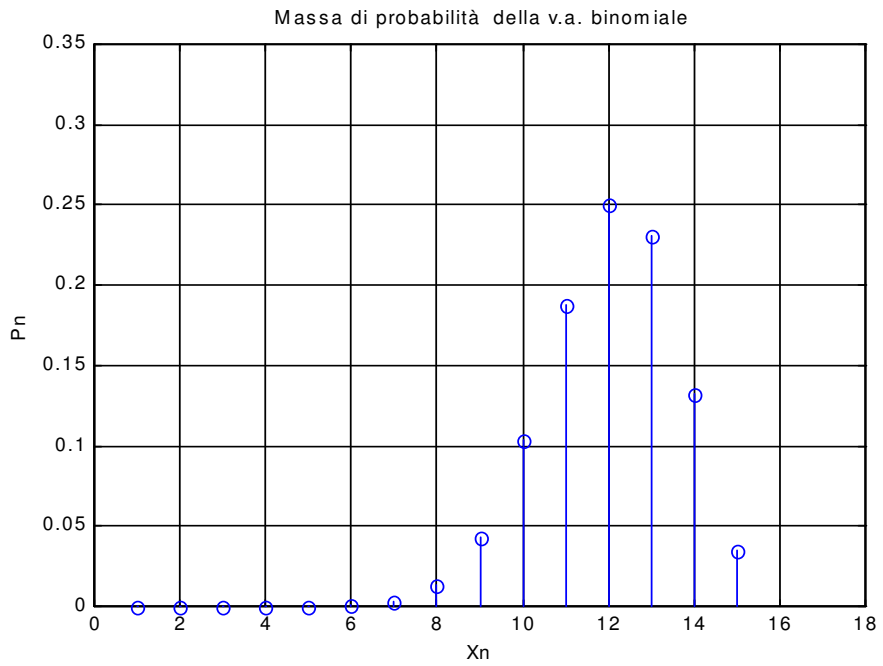
Se effettuiamo una sola prova $N = 1$, otteniamo la v.a. di Bernoulli, una v.a. aleatoria che può assumere solo due valori (0 ,1) rispettivamente con probabilità $1 - p$ e p.

Variabile aleatoria binomiale

La variabile aleatoria che descrive il problema delle prove ripetute è la v.a Binomiale la cui funzione massa di probabilità vale:

$$p_n = \binom{N}{n} p^n q^{(N-n)}$$

N =15, p =0.8



Variabile aleatoria binomiale

Il calcolo del valor medio:

$$E[n] = \sum_{n=0}^{N} np_n = \sum_{n=0}^{N} n \binom{N}{n} p^n q^{(N-n)} = Np$$

La varianza vale:

$$Var[n] = E[n^2] - E^2[n] = Npq$$

mentre la funzione generatrice di probabilità vale

$$P(z) = E[z^n] = \sum_{n=0}^{N} z^n \binom{N}{n} p^n q^{(N-n)} = (1 - p + zp)^N$$

v.a. di Poisson

Sempre in riferimento all'esperimento delle prove ripetute, supponiamo di poter ripetere le prove un numero infinito di volte. La probabilità che si verifichi *k* volte l'evento favorevole può essere calcolata mediante l'operazione di limite:

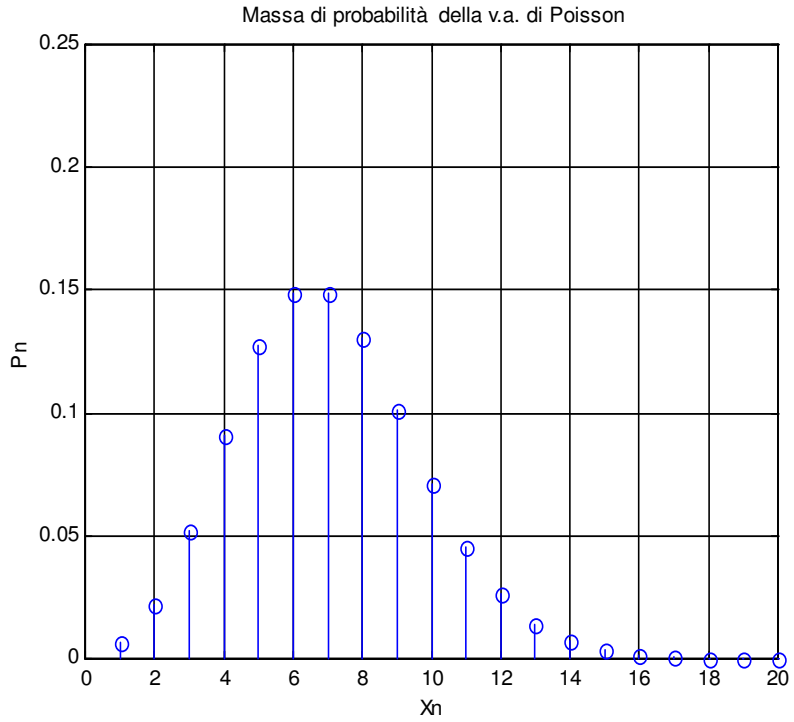
$$\Pr\{k \quad successi \quad in \quad \infty \quad prove \quad\} = \lim_{\substack{N\to\infty \\ p\to 0}} \binom{N}{k} p^k q^{(N-k)} = \frac{\lambda^k}{k!} e^{-\lambda}$$

v.a. di Poisson

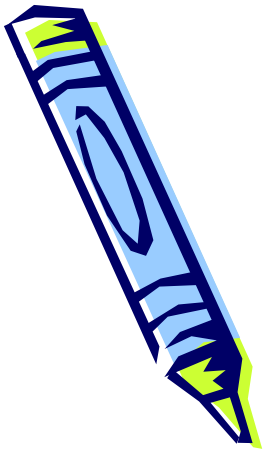
Dove λ è un valore che rimane costante (poiché il numero di prove tende all'infinito e la probabilità a zero). La nuova variabile aleatoria ottenuta è detta v.a. di Poisson, ed una v.a. di conteggio molto usata:

$$p_n = \frac{\lambda^n}{n!} e^{-\lambda}$$

con $\lambda = 7$



v.a. di Poisson



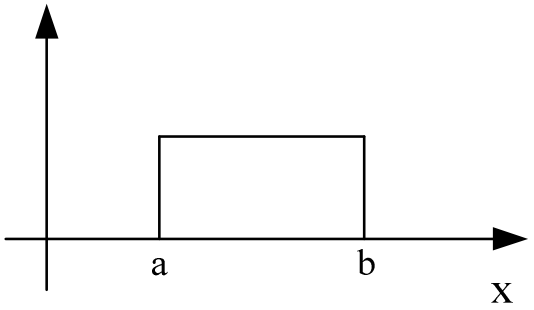
valor medio e varianza coincidono e sono pari a λ.

Questo risultato è molto importante, ed è il motivo per il quale nella pratica la v.a. di Poisson viene impiegata come v.a. di conteggio dei processi di arrivo regolari. Considerate due v.a. di Poisson *X*, *Y* rispettivamente con parametri λ eμ la v.a. *W* ottenuta dalla somma è ancora una v.a. di Poisson con parametro $(\lambda + \mu)$

V. a. Uniforme

Anche per le variabili aleatorie continue possiamo definire una distribuzione uniforme:

L'espressione della funzione di densità di probabilità, definita nell'intervallo [*a,b*] è:



$$f_X(x) = \begin{cases} \dfrac{1}{(b-a)} & a \le x \le b \\ 0 & altrove \end{cases}$$

V. a. Uniforme

Il valor medio e la varianza si calcolano dalla definizione

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)dx = \frac{1}{(b-a)} \int_a^b x dx = \frac{(a+b)}{2}$$

$$Var[X] = E[X^2] - E^2[X] = \frac{1}{(b-a)} \int_a^b x^2 dx - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}$$

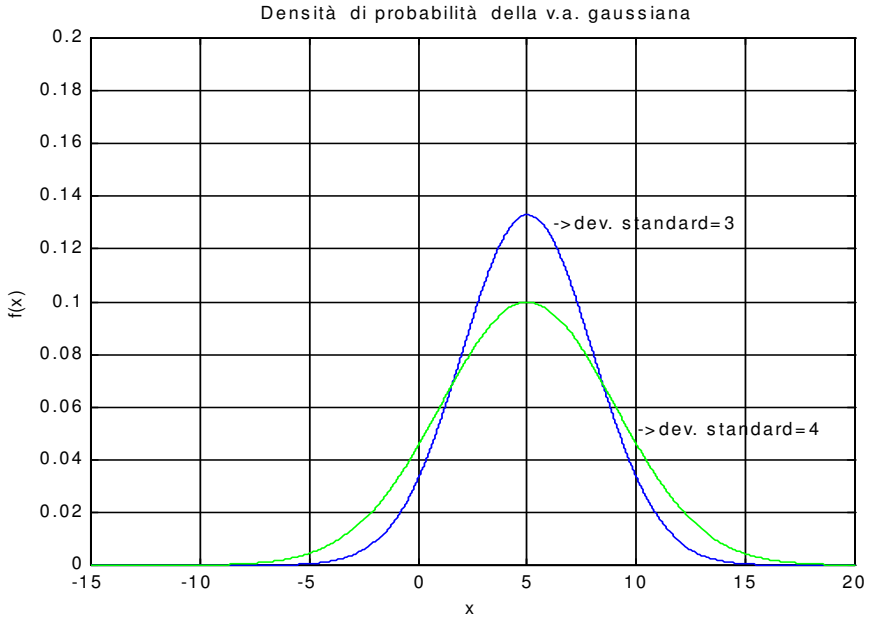
V. a. Gaussiana

La v.a. continua che sicuramente trova più larga applicazione è la v.a. gaussiana detta anche distribuzione normale definita su tutto l'asse reale]-∞, +∞[.

Il valor medio indicato con μ e la varianza $\sigma^2 \Rightarrow N(\mu,\sigma^2)$

($\mu = 5$, $\sigma^2 = 9$)

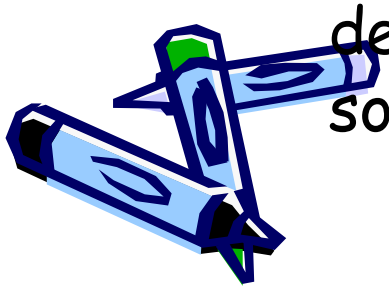
($\mu = 5$, $\sigma^2 = 16$)



V. a. Gaussiana

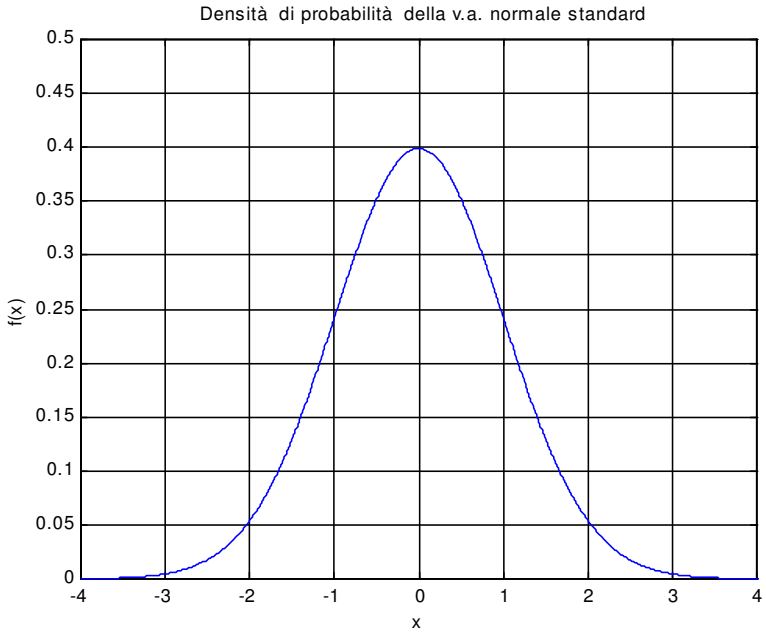
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

la funzione gaussiana è stata largamente studiata e tabulata, e trova larga applicazione nel campo della statistica, soprattutto viene impiegata per definire la precisione delle stime, e in tutti quei casi in cui si può definire un valore desiderato e si vuol vedere come i dati osservati sono distribuiti attorno a questo valore.



V. a. normale standard

Una curva gaussiana che trova largo impiego è quella ottenuta dalla coppia $\mu = 0$, $\sigma^2 = 1$, detta anche distribuzione normale standard



V. a. normale standard

L'espressione della distribuzione normale standard è:

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

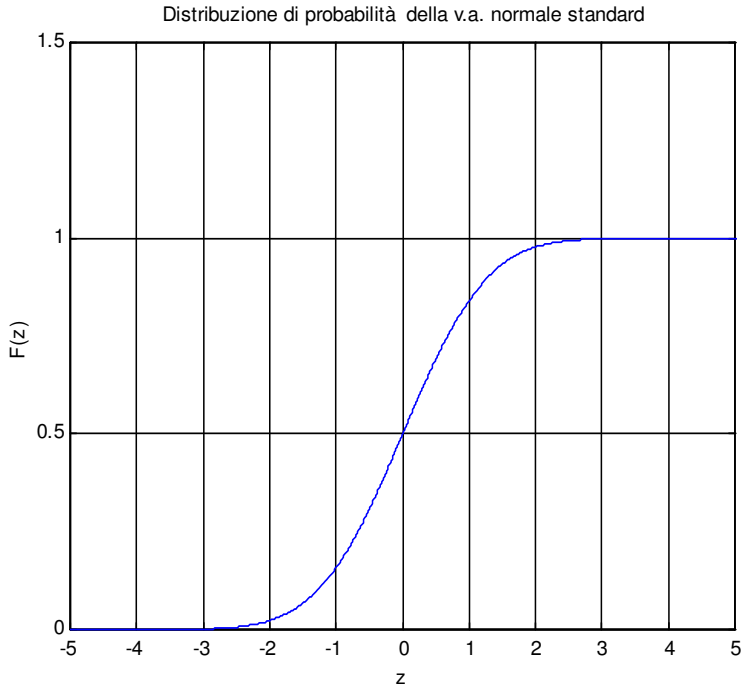
Molto spesso, nelle applicazioni pratica si usa la funzione di distribuzione standard piuttosto che la funzione di densità di probabilità. Ricordiamo che la densità di probabilità è la derivata della funzione di distribuzione:

$$f_X(x) = \frac{dF_X(x)}{dx} \Rightarrow F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

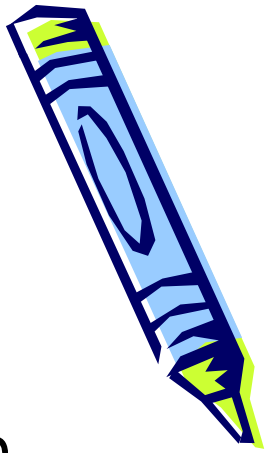
V. a. normale standard

La distribuzione di probabilità

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx$$



V. a. normale standard



Il motivo per cui si utilizza la $\Phi(z)$, è molto semplice, infatti dalla funzione di distribuzione si può ricavare direttamente la probabilità di un evento. Pertanto per il calcolo delle probabilità è opportuno ricordare che:

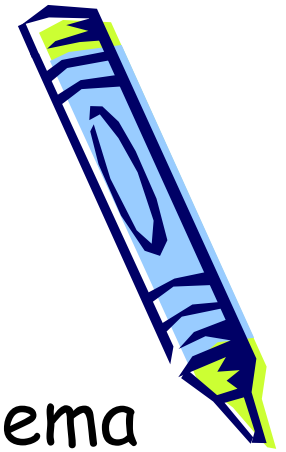
$$\Pr\{z < \alpha\} = \Phi(\alpha)$$

$$\Pr\{\alpha < z < \beta\} = \Phi(\beta) - \Phi(\alpha)$$

$$\Phi(-z) = 1 - \Phi(z)$$


Teoria della stima e della decisione statistica
Inferenza statistica

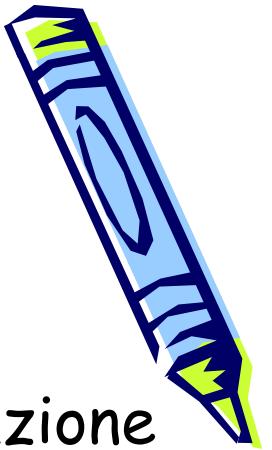
Inferenza statistica



Per effettuare una simulazione di un sistema che presenta elementi stocastici è necessario specificare le distribuzioni di probabilità che regolano i processi che caratterizzano il sistema stesso.

Se è possibile raccogliere dati reali (osservazioni) sulle variabili aleatorie di interesse, essi possono essere utilizzati per determinare queste distribuzioni facendo uso di tecniche di inferenza statistica

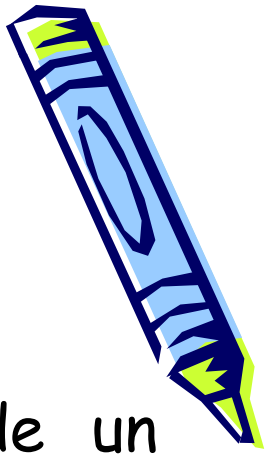
Inferenza statistica



Una volta stabilite tali distribuzioni, la simulazione procede generando valori casuali da queste distribuzioni, ovvero, durante ogni esecuzione, la simulazione genera osservazioni casuali di variabili aleatorie distribuite secondo particolari distribuzioni di probabilità.

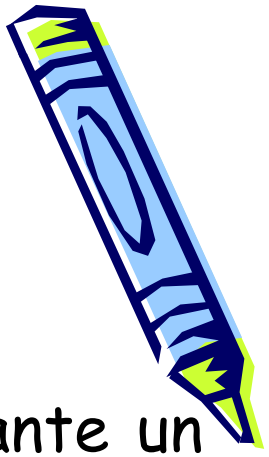
Oltre che per progettare una simulazione, è necessario l'uso di tecniche statistiche anche per interpretare i risultati ottenuti da una simulazione

Inferenza statistica



Con il termine *inferenza statistica* si intende un procedimento induttivo (trarre dal particolare conoscenza dell'universale), mediante il quale un risultato sperimentale, ottenuto dall'osservazione di alcuni elementi di un insieme, è utilizzato come strumento di conoscenza di alcune proprietà di tutto l'insieme.

Inferenza statistica

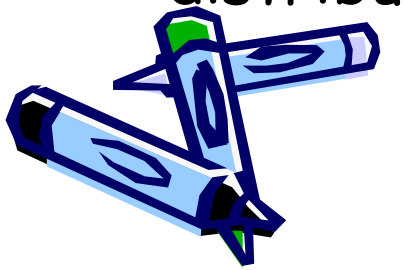


In generale, nello studio di un fenomeno riguardante un insieme di elementi (popolazione) che presenta caratteristiche aleatorie, molto spesso si dispone solo di informazioni su una parte di essi (campione) e si vogliono dedurre proprietà generali riguardanti l'intera popolazione.

Inferenza statistica

Solitamente viene fatta l'assunzione che esiste una distribuzione di probabilità della popolazione nel senso che se da essa vengono estratti casualmente alcuni elementi, ad essi sono associate variabili aleatorie indipendenti identicamente distribuite secondo tale distribuzione.

In questo senso, un insieme di variabili aleatorie $X_1,\ldots,X_n$ di variabili aleatorie indipendenti tutte con la stessa distribuzione si dice campione di questa distribuzione.

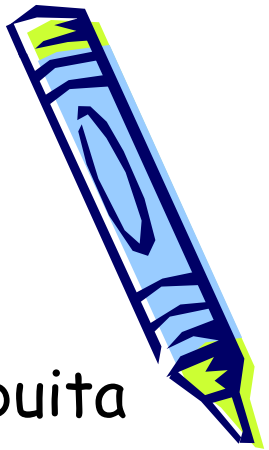


Inferenza statistica

L'interesse principale risiede nella possibilità di dedurre caratteristiche della distribuzione non nota sulla base dei dati a disposizione.

Naturalmente ci sono casi in cui della distribuzione della popolazione non si conosce nulla (se non il fatto che essa è discreta o continua), mentre in altri casi la distribuzione è nota ma non sono noti alcuni suoi parametri.

Stima di parametri

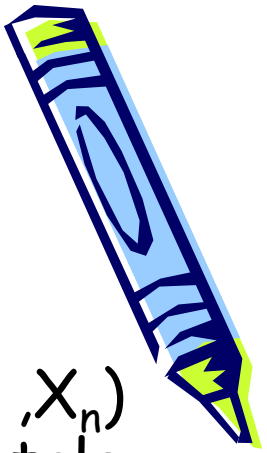


Supponiamo ora che la popolazione sia distribuita secondo una distribuzione di probabilità nota, ma caratterizzata da uno o più parametri incogniti.

Siamo in questo caso interessati a determinare tali parametri incogniti sulla base di un campione $X_1,\ldots,X_n$.

Si tratta di un problema di stima di parametri che consiste nel determinare, sulla base del campione $X_1,\ldots,X_n$, un valore per ciascuno dei parametri in modo che essi costituiscano la migliore approssimazione dei parametri incogniti.

Stima di parametri

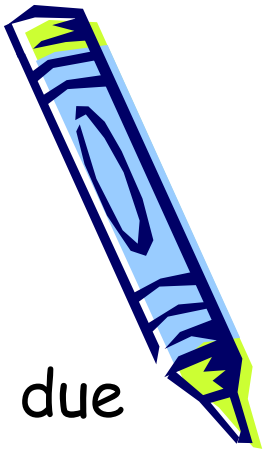


Uno stimatore è quindi una funzione $f(X_1, \ldots, X_n)$ delle osservazioni campionarie e il valore che tale funzione assume in corrispondenza di una particolare realizzazione del campione è detto stima.

Se θ e un parametro incognito, si indicherà con $\hat{\theta}$ la stima di θ.

In alcuni casi si determina un unico valore $\hat{\theta}$ come migliore approssimazione possibile del parametro θ e tale valore viene detto *stima puntuale*.

Stima di parametri



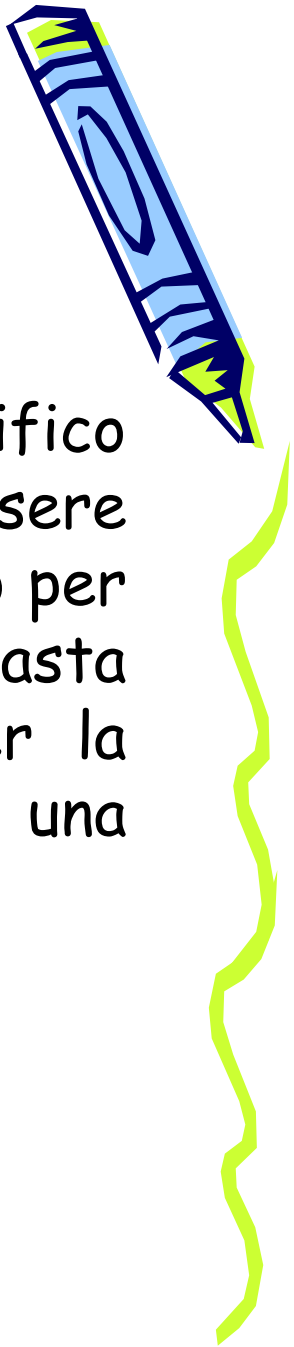
In altri casi, può essere preferibile calcolare due valori dello stimatore ovvero

$\theta_1 = h_1(X_1,\ldots,X_n)$ e $\theta_2 = h_2(X_1,\ldots,X_n)$

che definiscono un intervallo $[\theta_1, \theta_2]$ tale che, in un campionamento ripetuto, il valore incognito θ apparterrà all'intervallo in una determinata percentuale di casi (detta confidenza dell'intervallo).

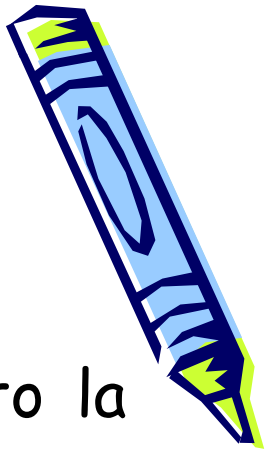
In questo caso si parla di *stima per intervalli*.

Stimatore



In molti casi reali si ipotizza che uno specifico comportamento di una popolazione possa essere descritto da una **curva gaussiana**, in questo caso per ricavare una stima della densità di probabilità, basta ricavare delle stime per il valor medio e per la varianza (parametri sufficienti per ricavare una curva gaussiana).

Stimatori



Si definisce valore dell'*errore* di campionamento la differenza $\hat{\theta}-\theta$.

Si chiama *distorsione* di uno stimatore h la differenza $E(h)-\theta$.

L' *errore quadratico medio* dello stimatore h è dato da $EQM(h) = E(h - \theta)^2$.

Stimatori

Uno stimatore $h = h(X_1,\ldots,X_n)$ si dice stimatore

corretto del parametro θ se risulta $E(h) = \theta$.

Se invece si ha $E(h) \neq \theta$ si dice che h è uno stimatore distorto per θ.

stimatore efficiente del parametro θ se

i) $E(h) = \theta$

ii) $Var(h) \leq Var(h_1)$ per ogni h_1 stimatore corretto di θ.

Stimatori

Il grado di bontà delle stime fornite dallo stimatore sono: *correttezza, consistenza* ed *efficienza*.

Uno stimatore si dice ***corretto*** se in media uguaglia il valore vero.

Uno stimatore si dice ***consistente*** quando migliora la stima al crescere delle dimensioni del campione misurato, e per $N \to \infty$, tende al valore vero.

L'***efficienza*** si misura confrontando due stimatori, e assegnando maggiore efficienza a quello con varianza minore.

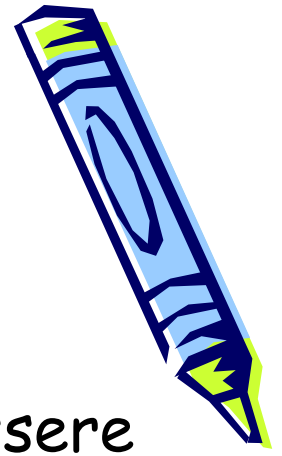
Stimatore

Se ad esempio vogliamo calcolare il valor medio del campione di dati raccolti possiamo effettuare la seguente media aritmetica:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Alla quale si dà il nome di *media campionaria*,

Verifica dello stimatore

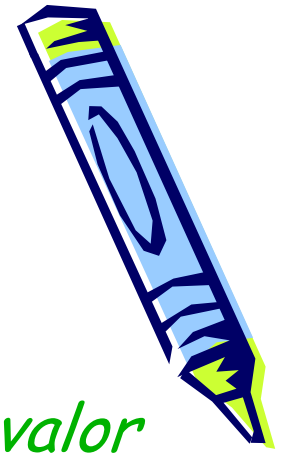


Per verificare se la media campionaria può essere uno stimatore corretto del parametro che si vuole descrive bisogna verificare che:

$$E[\overline{X}] = E\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \theta$$

Dove θ è il valore vero del parametro da stimare.

Stimatore corretto

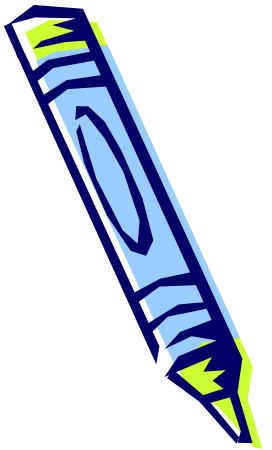


Supponiamo che il parametro *da stimare sia il valor medio* della variabile aleatoria con la quale si vuole descrivere il comportamento di una popolazione.

Utilizziamo come *stimatore la media campionaria*, si dimostra che la media campionaria è uno stimatore corretto

$$E[\overline{X}] = E\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \theta$$

Stimatore consistente



Inoltre:

$$Var[\bar{X}] = E[(\bar{X} - \theta)^2] = E\left[\frac{1}{N}\sum_{n=1}^{N}(X_n - \theta)^2\right] = \frac{1}{N^2}\sum_{n=1}^{N}Var[X_n] = \frac{Var(X_n)}{N}$$

Dunque la media campionaria è anche uno stimatore consistente, infatti la varianza dello stimatore scelto diminuisce all'aumentare di N (dimensione del campione).

Stimatore

Per stimare il valor medio si usa come stimatore la media campionaria, e verrebbe spontaneo utilizzare *la varianza campionaria*, come stimatore della varianza: La varianza campionaria S^2 viene ricavata dai dati misurati sul campione nel seguente modo:

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_N - \overline{X})^2}{N}$$

Stimatore

Purtroppo la varianza campionaria non è uno stimatore corretto infatti

$$S^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \overline{X})^2 \qquad E[S^2] = \frac{1}{N}\sum_{n=1}^{N}E[X_n^2] - E[\overline{X}^2] = \frac{N-1}{N}\sigma_X^2$$

il valor medio della varianza campionaria non coincide con la varianza da stimare, quindi non può essere usata come stimatore della varianza.

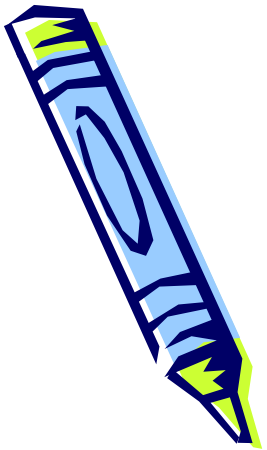
Stimatore

$$E[S^2] = \frac{1}{N}\sum_{n=1}^{N} E[X_n^2] - E[\overline{X}^2] = \frac{N-1}{N}\sigma_X^2$$

In questo caso il problema della scelta di uno stimatore corretto è semplice infatti basta utilizzare la seguente relazione per ottenere uno stimatore corretto:

$$\hat{S}^2 = \frac{N}{N-1}S^2$$

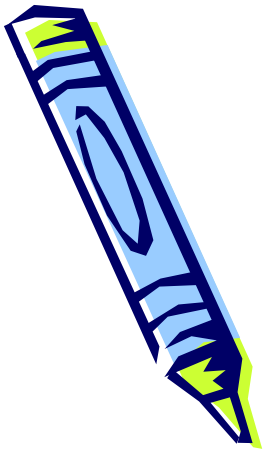
Metodo della massima verosimiglianza



Tale tecnica di stima più usata per le determinazione dei *parametri costanti*

I parametri costanti o deterministi sono dei valori numerici che esprimono alcune caratteristiche di una popolazione, ad esempio il valore medio. Altri esempi di parametri costanti che possono essere oggetto di stima sono la varianza, ed in generale i momenti di qualsiasi ordine

Metodo della massima verosimiglianza

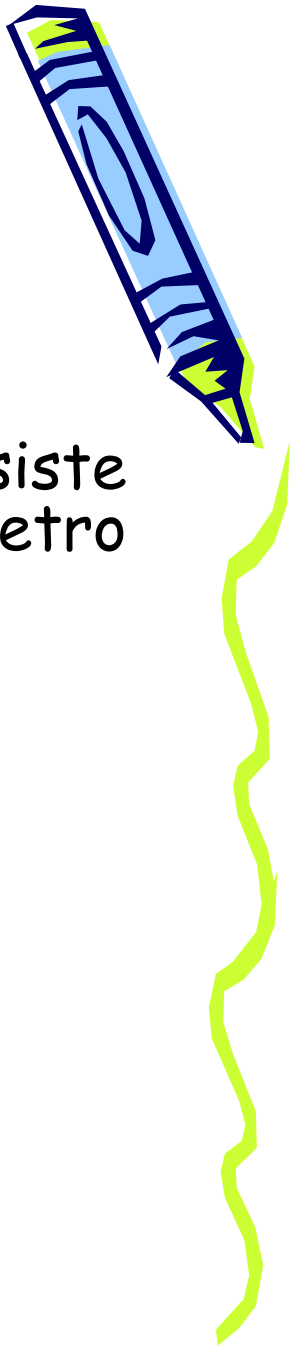


Date n osservazioni $X_1,\ldots,X_n$, assumiamo che esse siano ottenute da una distribuzione di probabilità avente densità $f_\theta(x)$, dove θ e un parametro che caratterizza la distribuzione.

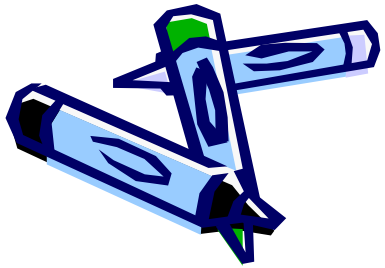
Nell'ipotesi che le osservazioni Xi siano indipendenti, Una misura della probabilità di aver ottenuto quelle osservazioni proprio da quella distribuzione (se θ e il valore del parametro incognito) è data dalla funzione di verosimiglianza

$$L(\theta) = \prod_{i=1}^{N} f(x_i;\theta)$$

Metodo della massima verosimiglianza

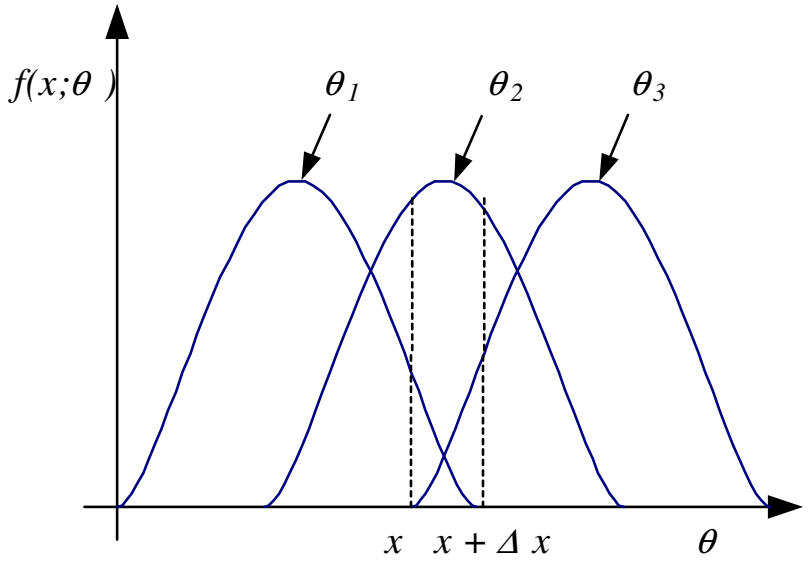


Il metodo della massima verosimiglianza consiste nello scegliere come stimatore del parametro incognito θ il valore che massimizza $L(\theta)$.

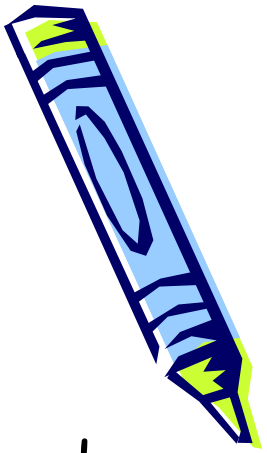


Metodo della massima verosimiglianza

La densità di probabilità $f_\theta(x)$ varia a seconda del valore del parametro, se l'osservazione cade nell'intervallo $[x, x+\Delta x]$ allora la tecnica della massima verosimiglianza sceglierà come stima del parametro θ, il valore θ_2.



Metodo della massima verosimiglianza

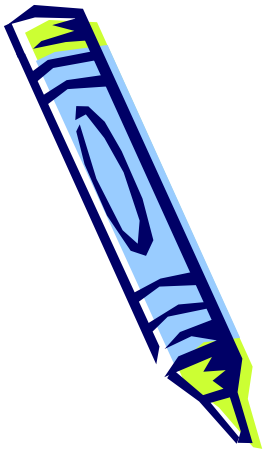


In molte applicazioni pratiche può essere più comodo calcolare il logaritmo della funzione di verosimiglianza, dato che il prodotto si trasforma in una somma:

$$\ln[L(\theta)] = \sum_{i=1}^{N} \ln[f(x_i;\theta)]$$

Poiché il logaritmo è una funzione crescente dell'argomento, per procedere nella stima a massima verosimiglianza, basta semplicemente verificare che $\ln[L(\theta)]$ sia derivabile, calcolare la derivata prima e ricavare il valore di θ per cui si annulla.

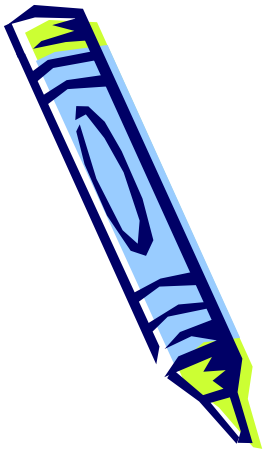
Esercizio



Date n osservazioni $X_1,\ldots,X_n$ della distribuzione esponenziale, determinare con il metodo della massima verosimiglianza il parametro λ della distribuzione.
La funzione di massima verosimiglianza è

$$L(\lambda) = \prod_{i=1}^{N} \lambda e^{-\lambda Xi} = \lambda^n \mathrm{e}^{-\lambda_n \overline{X_n}}.$$

Esercizio



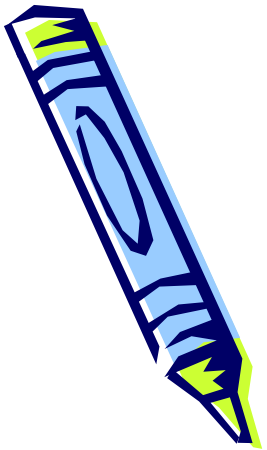
Uguagliando a zero la derivata (rispetto a λ) si ha

$$d(\lambda^n e^{-\lambda_n \overline{X_n}})/d\lambda = n\lambda^{n-1} e^{-\lambda_n \overline{X_n}} (1 - \lambda \overline{X_n}) = 0.$$

$$\hat{\lambda} = \frac{1}{\overline{X_n}}$$

è un punto di massimo per la funzione L(λ) visto che la derivata seconda è minore di zero.

Esercizio



Il valore ottenuto per la stima non ci sorprende perchè la media campionaria è uno stimatore corretto della media della distribuzione che è $1/\lambda$.

Allo stesso risultato si poteva arrivare utilizzando i logaritmi.

Stima per intervalli

Nei metodi di stima puntuale è sempre presente un errore $\hat{\theta}-\theta$ dovuto al fatto che la stima di θ in genere non coincide con il parametro θ.

Sorge quindi l'esigenza di determinare una misura dell'*errore commesso*.

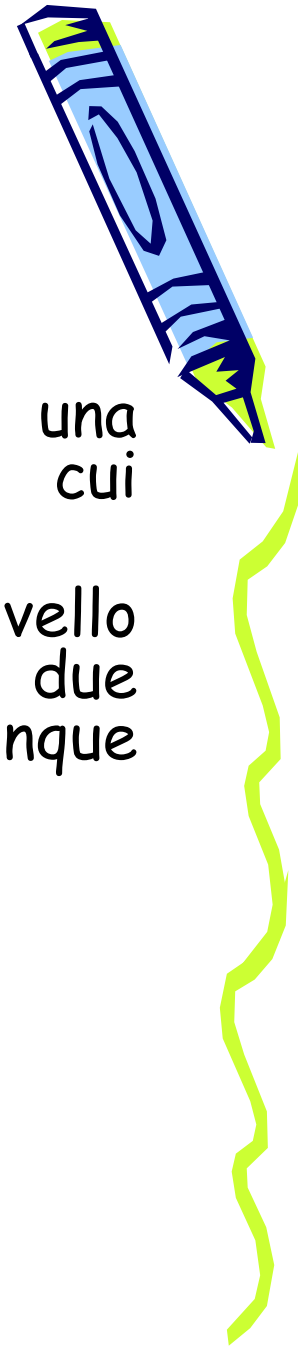
Inoltre, dato il campione $X_1, \ldots, X_n$ estratto da una distribuzione di probabilità caratterizzata da un parametro incognito θ, qualunque sia lo stimatore $h(X1, \ldots, Xn)$ scelto per stimare θ, esso dipende dal campione, ovvero lo stimatore fornirà stime diverse in corrispondenza di campioni diversi

Stima per intervalli

Queste due osservazioni fanno nascere l'esigenza di considerare stime per intervalli.

Infatti, sulla base dei valori di stima ottenuti considerando un campione casuale X1, . . . ,Xn, si può definire un intervallo in cui sono compresi i valori più probabili per il parametro θ, secondo un "livello di confidenza" fissato.

Stima per intervalli



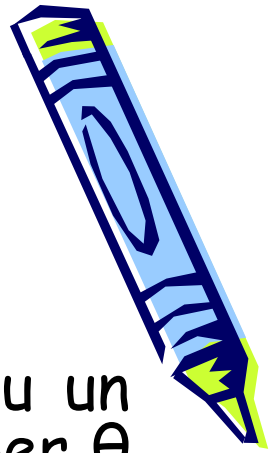
Si può procedere indirettamente utilizzando una statistica campionaria $g(X_1, \ldots, Xn)$ la cui distribuzione sia nota e non dipendente da θ.

Naturalmente, visto che la g è nota, fissato un livello di confidenza $(1 - \alpha)$, è possibile determinare due valori $g1$ e $g2$, indipendenti da θ tali che, comunque scelto $\alpha \in (0, 1)$

- $P(g1 \leq g \leq g2) = 1 - \alpha$.

.

Stima per intervalli



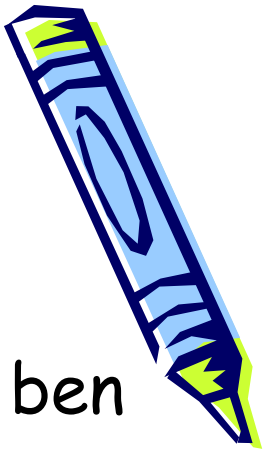
Lo scopo è quello di tradurre una probabilità su un intervallo per g in una probabilità su intervallo per θ in modo da poter avere

$$P(h1 \leq \theta \leq h2) = 1 - \alpha$$

ovvero in modo tale che h1 e h2 rappresentino gli estremi dell'intervallo per θ.

Le distribuzioni note alle quali si fa di solito riferimento sono la distribuzione Normale, la distribuzione t di Student e la distribuzione Chi-quadro.

Stima per intervalli

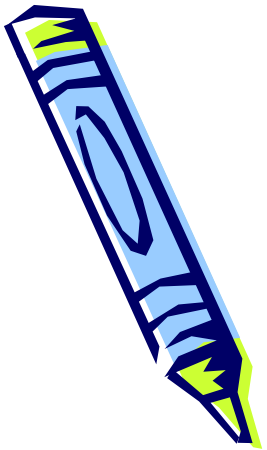


Lo scopo sarà quello di avere valore di **α** ben superiore a 0.5 in modo che

la probabilità che il parametro θ appartenga all'intervallo [h1, h2] sia tale da assicurare all'evento $h1 \leq \theta \leq h2$ (evento che si verifica nel 100(1 – **α**)% dei casi) una caratteristica di "sistematicità",

mentre all'evento complementare (che si verifica nel 100**α**% dei casi) una caratteristica di "accidentalità".

Intervallo di confidenza



Dato un campione

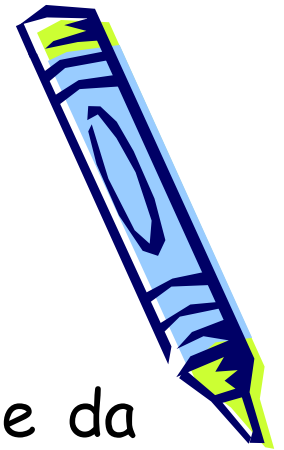
X1, . . . ,Xn, dato $\alpha \in (0, 1)$ e date le statistiche

$h1 = h1(X_1,\ldots,X_n)$ e $h2 = h2(X_1, \ldots,Xn)$ $h1 < h2$, per le quali

- $P(h1 \leq \theta \leq h2) = 1 - \alpha$,

l'intervallo [h1, h2] si dice intervallo di confidenza per θ con livello di confidenza pari ad $(1 - \alpha)$.

Stima per intervalli di una MEDIA



Siano date le osservazioni X1, . . . ,Xn estratte da una distribuzione di probabilità a media μ e varianza σ^2. Assumiamo inizialmente che la media μ sia incognita mentre la varianza sia nota.

Dal teorema del limite Centrale sappiamo che per n sufficientemente grande, la variabile aleatoria campionaria è distribuita approssimativamente secondo la distribuzione Normale standard indipendentemente dalla distribuzione delle Xi.

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{N}}$$

Stima per intervalli di una MEDIA

La variabile Z è una v.a. nota di cui si conosce la densità di probabilità che si trova tabulata in tutti i libri di statistica, quindi è facile calcolare l'intervallo in cui abbiamo probabilità di trovare Z.

Stima per intervalli di una MEDIA

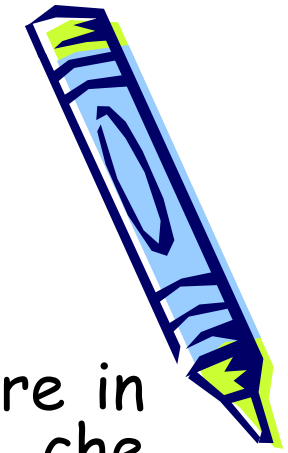
Ad esempio, per determinare un intervallo di confidenza al 95% per una media, dato un campione X1, . . . ,Xn, si trova il punto critico

$z_{1-a/2} = z_{0.975} = 1.96$

dalle tabelle della distribuzione Normale standard facilmente si ricava tale intervallo

$$\left[\overline{x_N} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{x_N} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

Stima per intervalli



*La **stima per intervalli*** consiste nel determinare in base a valori del campione, un intervallo che contenga il parametro da stimare, con un certo grado di incertezza ritenuto accettabile.

Data una v.a. gaussiana $N(\mu,\sigma^2)$ con varianza nota pari a $\sigma^2 = 9$, si vogliono ottenere informazioni sul valor medio μ utilizzando un campione X di 4. Con un intervallo di confidenza del 95%

Esempio

La media aritmetica del campione (media campionaria), è anch'essa una v.a. gaussiana con valor medio pari a μ e varianza pari σ^2 /N dove N è la dimensione del campione.

Consideriamo la variabile gaussiana standard Z espressa dalla relazione

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{N}} = \frac{2}{3}\left(\overline{X} - \mu\right)$$

Esempio

La variabile Z è una v.a. nota di cui si conosce la densità di probabilità che si trova tabulata in tutti i libri di statistica, quindi è facile calcolare l'intervallo in cui abbiamo il 95% di probabilità di trovare Z:

$$\Pr\{-1.96 < Z < 1.96\} = 0.95$$

$$\Pr\{-2.94 < (\overline{X} - \mu) < 2.94\} = 0.95$$

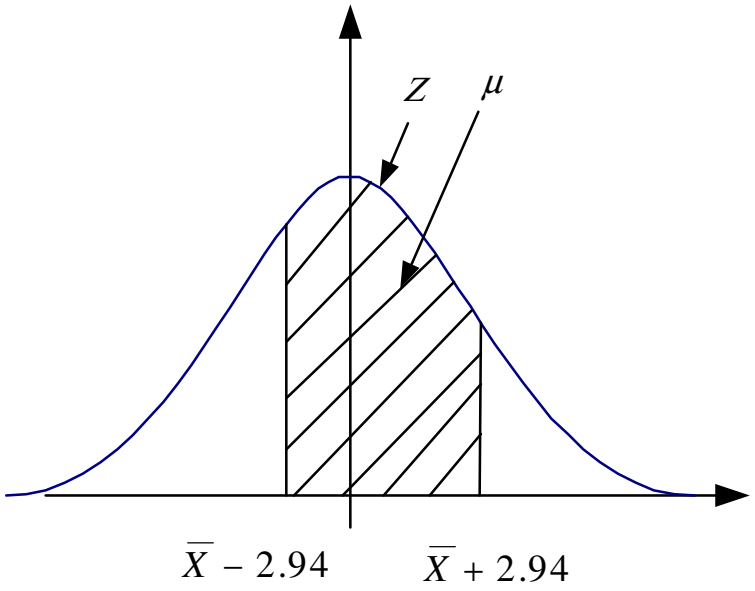
$$\Pr\{\overline{X} - 2.94 < \mu < \overline{X} + 2.94\} = 0.95$$

Sostituendo alla media campionaria il valore ottenuto dai dati osservati si ricava l'intervallo a cui appartiene μ con un incertezza pari al 5%

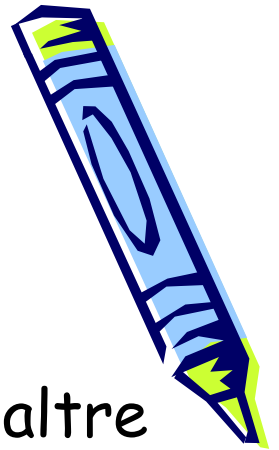
Esempio

Se ad esempio il campione osservato ha valori (1.2, 3.4, 0.6, 5.6) la media campionaria assume valore 2.7, e con la tecnica della stima per punti avremmo detto che il valore stimato di μ è proprio 2.7.

Invece con la tecnica della stima per intervalli, diciamo che μ appartiene all'intervallo [-0.24, 5.6] con un livello di incertezza del 5%.



Varianza non nota



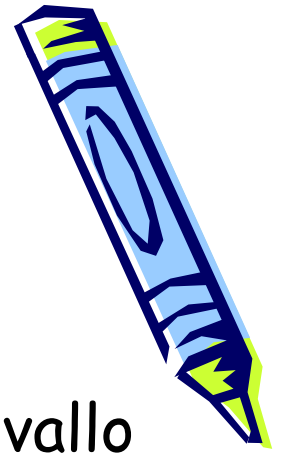
Se σ^2 non si conosce sono necessarie altre considerazioni.

Se il numero di osservazioni è elevato, per il Teorema del Limite Centrale, è possibile approssimare la varianza con la varianza campionaria, e l'approccio rimane lo stesso:

$$Z = \frac{\overline{X} - \mu}{s / \sqrt{N}}$$

Per n sufficientemente elevato possiamo ancora considerare la distribuzione Normale standard

Varianza non nota



La difficoltà nell'utilizzare questo tipo di intervallo di confidenza per μ sta nel fatto che esso ha valore asintotico, ovvero per n sufficientemente grande e quindi risulta approssimato.

Ovvero per valori piccoli di n si può utilizzare una definizione alternativa dell'intervallo di confidenza che fa riferimento alla distribuzione t di Student a n–1 gradi di libertà.

Varianza non nota

Se le Xi sono variabili Normali, la variabile

$$T = \sqrt{N-1}\,\frac{\overline{X}-\mu}{S}$$

ha distribuzione t di Student con n – 1 gradi di libertà per ogni n > 1. Poiché T dipende solo da μ (parametro da stimare) e dal valore del campione può essere utilizzata per determinare l'intervallo di fiducia, in particolare

$$\Pr\left\{-t_1 < \frac{\overline{X}-\mu}{S}\sqrt{N-1} < t_2\right\} = 1-\gamma$$

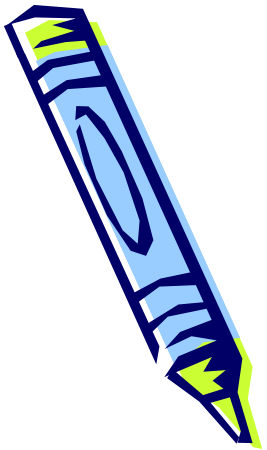
Varianza non nota

$$\Pr\left\{\bar{X}-\frac{t_1}{\sqrt{N-1}}S<\mu<\bar{X}+\frac{t_2}{\sqrt{N-1}}S\right\}=1-\gamma$$

Analogamente per determinare l'intervallo di fiducia per la stima della varianza utilizziamo una v.a. $\chi 2$

$$\Pr\{x_1<\chi^2<x_2\}=1-\gamma \qquad \Pr\left\{x_1<\frac{NS^2}{\sigma^2}<x_2\right\}=1-\gamma$$

Stime per intervalli



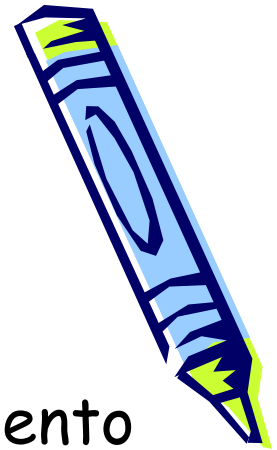
Ricordiamo le proprietà utilizzate:

Data una popolazione Normale, se si normalizza la media campionaria Xn sottraendo la sua media μ

dividendo per la sua deviazione standard $\sigma/\sqrt{n}$, si ottiene una variabile aleatoria *Normale standard*;

dividendo per $s_n/\sqrt{n}$, si ottiene una variabile aleatoria con distribuzione t di Student con n – 1 gradi di libertà.

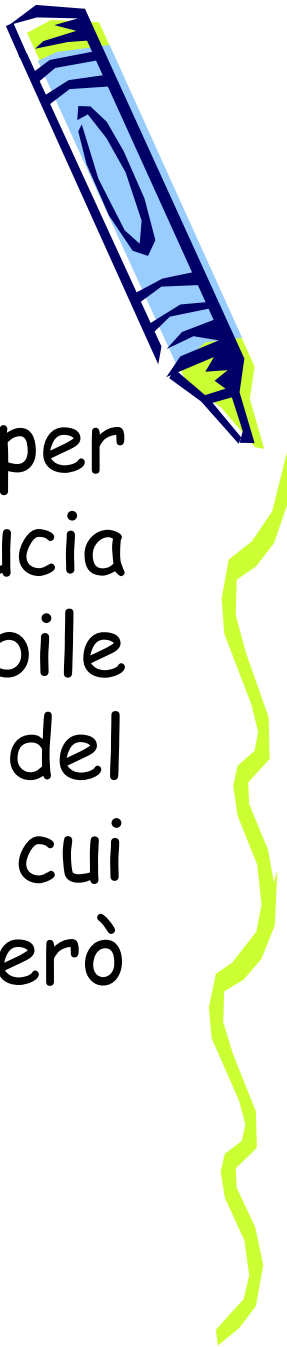
Stime per intervalli



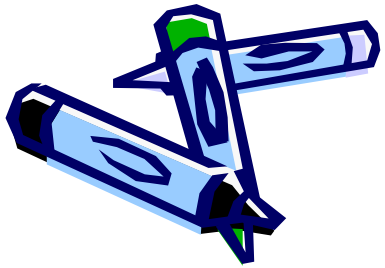
L' intervallo di confidenza definito in riferimento alla distribuzione Normale standard è basato sul Teorema del Limite Centrale e la copertura dipende dalla scelta di n.

L'intervallo di confidenza definito in riferimento alla distribuzione t di Student è approssimato perchè influenzato dalla distribuzione delle Xi che in generale non sono Normali; tuttavia questo secondo tipo di intervallo di confidenza ha maggiore copertura dell'altro.

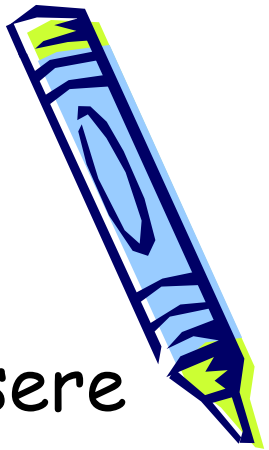
Stime per intervalli



In generale la tecnica impiegata per costruire un intervallo di fiducia consiste nel costruire una variabile aleatoria Z funzione del campione e del parametro che si vuole stimare, la cui legge di distribuzione sia però indipendente dal parametro.



Stime per intervalli



- La variabile normale standard può essere utilizzata soltanto quando è nota la varianza
- La variabile T student dipende solo dal valor atteso e dal valore del campione, quindi può essere utilizzata per determinare l'intervallo di fiducia per la stima del valor atteso.
- La variabile χ^2 serve per determinare l'intervallo di fiducia per la stima della varianza

Distribuzione di chi-quadro

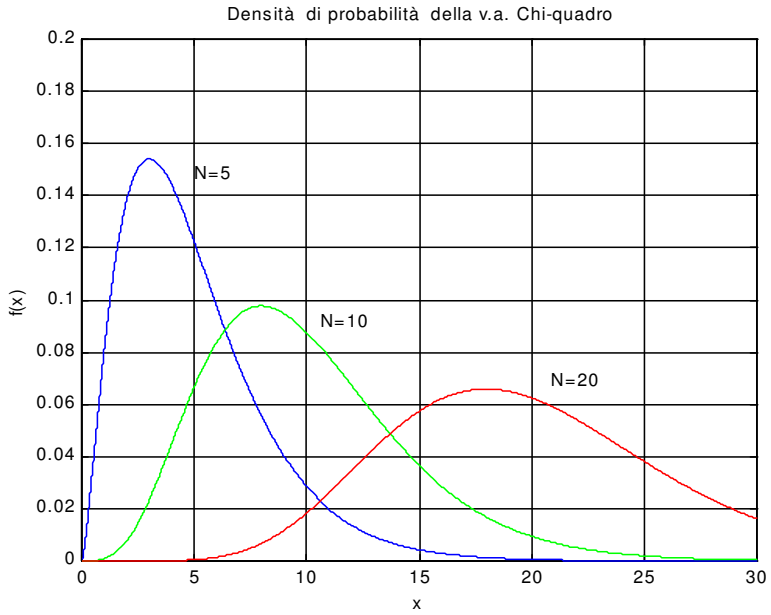
Consideriamo la seguente successione di v.a. normali standard indipendenti ed identicamente distribuite: $(X_1,\ldots,X_N)$, e consideriamo la v.a. X ottenuta come:

$$X = \sum_{i=1}^{N} X_i^2$$

La v.a. X prende il nome di variabile del χ^2 (Chi-quadro), può assumere solo valori positivi o nulli, e la funzione densità di probabilità dipende evidentemente da N (numero di gradi di libertà)

Distribuzione di chi-quadro

$$f_N(x) = \frac{e^{-\frac{x}{2}}}{2^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)} x^{\frac{N}{2}-1} \qquad x \geq 0$$



$$E[X] = N$$

$$Var[X] = 2N$$

Valor medio e varianza sono ottenuti direttamente dai gradi di libertà. Poiché la v.a. del χ^2 è ottenuta come somma di v.a. indipendenti per $N \to \infty$ converge ad una v.a. gaussiana di parametri N (N, $2N$).

Distribuzione di t student

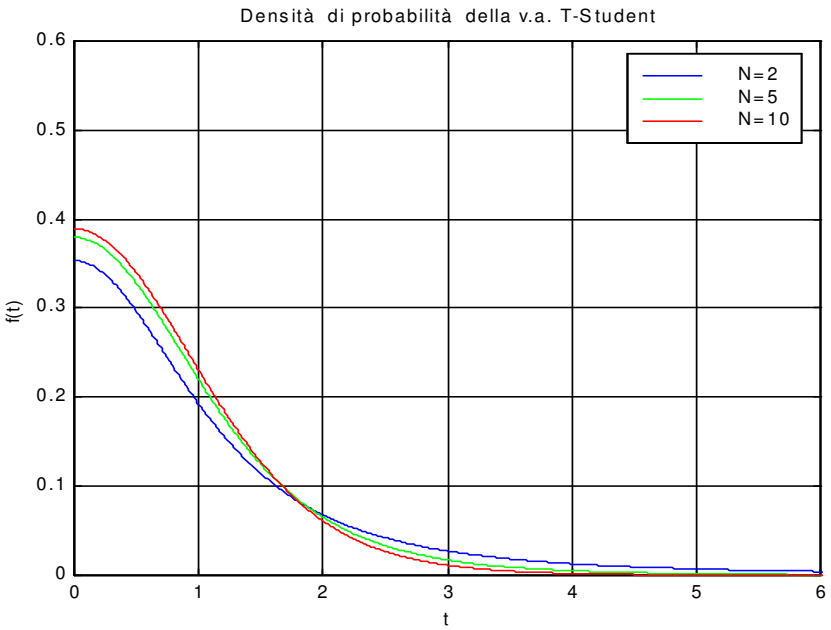
Un'altra v.a. molto usata in statistica che si ricava dalla v.a. gaussiana è la v.a. T-Student. Consideriamo una sequenza di $N+1$ (X, X_1, X_2,...,X_N), v.a. normali indipendenti, se escludiamo dalla sequenza la prima v.a. X, con le restanti N costruiamo una v.a. del χ^2 ad N gradi di libertà. La v.a. ottenuta dalla seguente trasformazione:

$$T = \sqrt{N}\,\frac{X}{\sqrt{Y}}$$

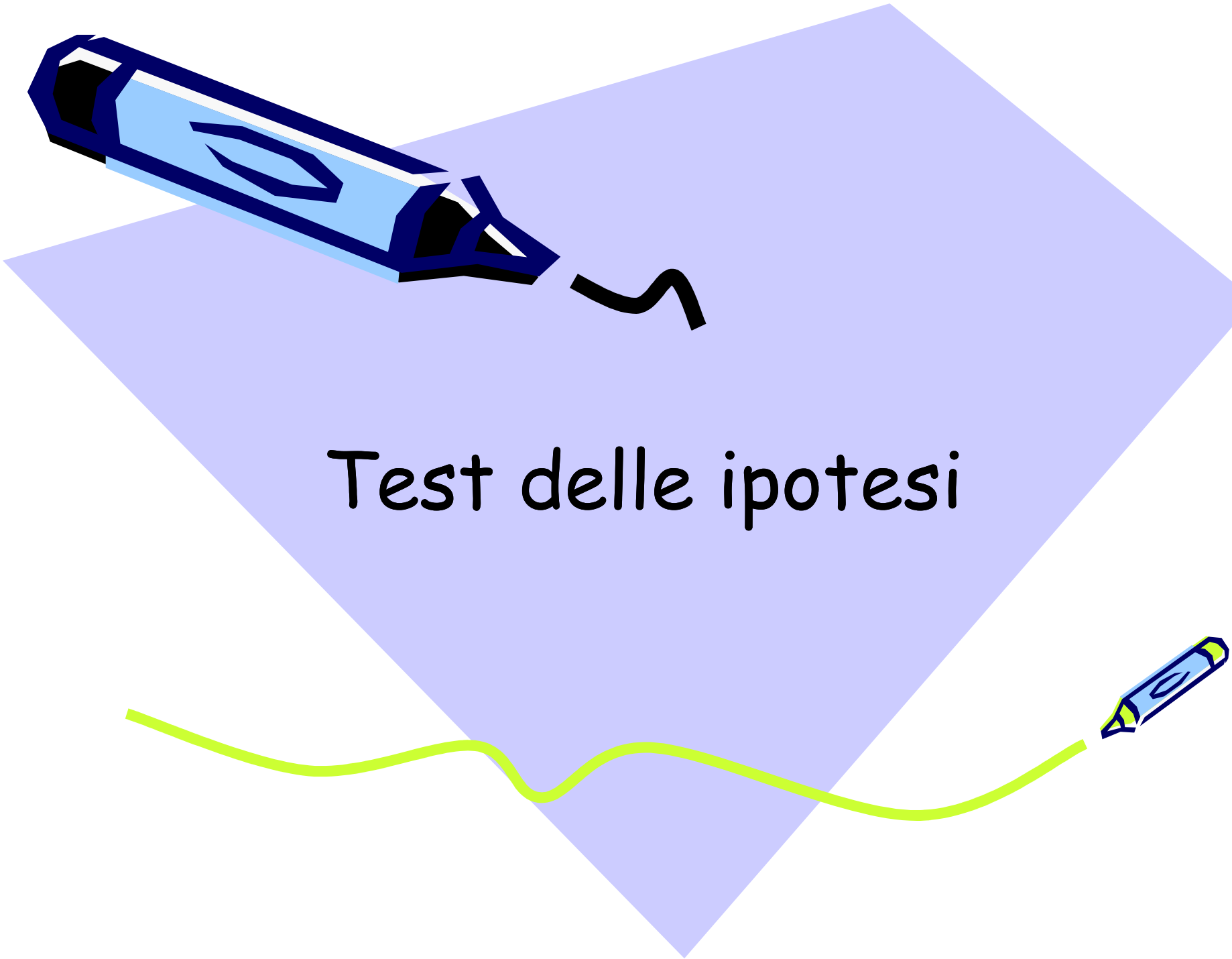
è detta v.a. T di Student, ad N gradi di libertà

Distribuzione di t student

$$f_N(t) = \frac{1}{\sqrt{N\pi}} \frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{N}{2}\right)} \left(1 + \frac{t^2}{N}\right)^{-\frac{N}{2}}$$



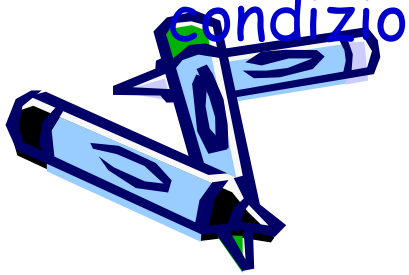
Per $N>1$ il valor medio è nullo, mentre per $N>2$ la varianza vale $\mathrm{Var}[T] = N/(N-2)$


Test delle ipotesi

Test delle ipotesi

Nel cercare di costruire un legame tra dati osservati e ipotesi teoriche sulle caratteristiche dell'intera popolazione si deve, in genere, prendere una decisione per il raggiungimento di tale conclusione generale e nasce il problema di esprimere un giudizio di plausibilità di un'ipotesi che si è specificata per la popolazione.

Per verificare *la coerenza* tra osservazioni e ipotesi fatta si fa uso di test statistici che prendono nome di test delle ipotesi. Possiamo dire che tali test devono confrontare i valori osservati e i corrispondenti valori teorici attesi condizionatamente all'ipotesi fatta.



Test delle ipotesi

Le differenze che vengono riscontrate possono essere ovviamente ricondotte a due possibilità:

1- l'ipotesi specificata è corretta e la differenza riscontrata è puramente casuale;

2- l'ipotesi specificata è errata e quindi non ci si può aspettare che i due valori siano "vicini".

Test delle ipotesi

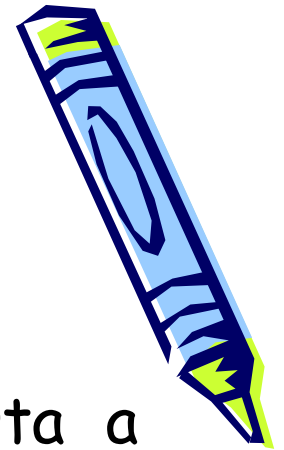
Il procedimento consiste nel confrontare due ipotesi: l'ipotesi da sottoporre a verifica e il suo complemento.

H_0 (ipotesi nulla) sottoinsieme dei valori individuati dall'ipotesi da sottoporre a verifica

H1 (ipotesi alternativa) è il suo complemento.

Se un test di ipotesi non scarta l'ipotesi H_0, questo non vuol dire che H_0 è accettata come vera, ma che può essere considerata possibile.

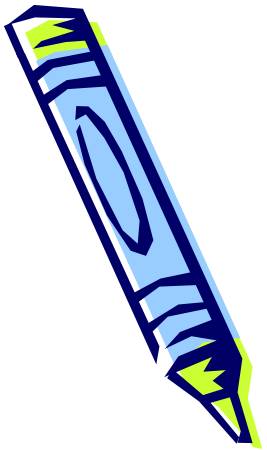
Test delle ipotesi



Si parla di *errore di I specie* se il test porta a rifiutare un'ipotesi H_0 quando questa è corretta e di *errore di II specie* se il test porta ad accettare H_0 quando questa è falsa.

Si noti che l'obiettivo non è quello di dire se l'ipotesi fatta è vera o falsa, ma piuttosto di verificare se l'ipotesi fatta sia compatibile con i dati. In genere, c'è un ampio margine di tolleranza nell'accettare H_0, mentre per rifiutarla occorre che i dati siano veramente poco probabili

Test delle ipotesi



Per testare le ipotesi, si specifica un valore, detto *livello di significatività* e si impone che il test sia tale che, quando l'ipotesi H_0 `e corretta, la probabilità che essa venga scartata è non superiore ad α .

Quindi un test con livello di significatività pari ad α deve essere tale che una probabilità di commettere un errore di I specie è minore o uguale ad α .

Esempio

Facciamo l'esempio di un campione gaussiana di varianza nota σ^2, di cui abbiamo stimato il valor medio con il valore μ_0 e di cui vogliamo verificare la correttezza.

Dunque l'ipotesi statistica da verificare è $\mu = \mu_0$

$$\Pr\left\{\mu_0 - \frac{z_\gamma \sigma}{\sqrt{N}} < \overline{X} < \mu_0 + \frac{z_\gamma \sigma}{\sqrt{N}}\right\} = 1 - \alpha$$

Esempio

Se per un particolare campione si ottiene un valore esterno a tale intervallo si può rifiutare l'ipotesi

($\mu = \mu_0$) con un'incertezza pari a α.

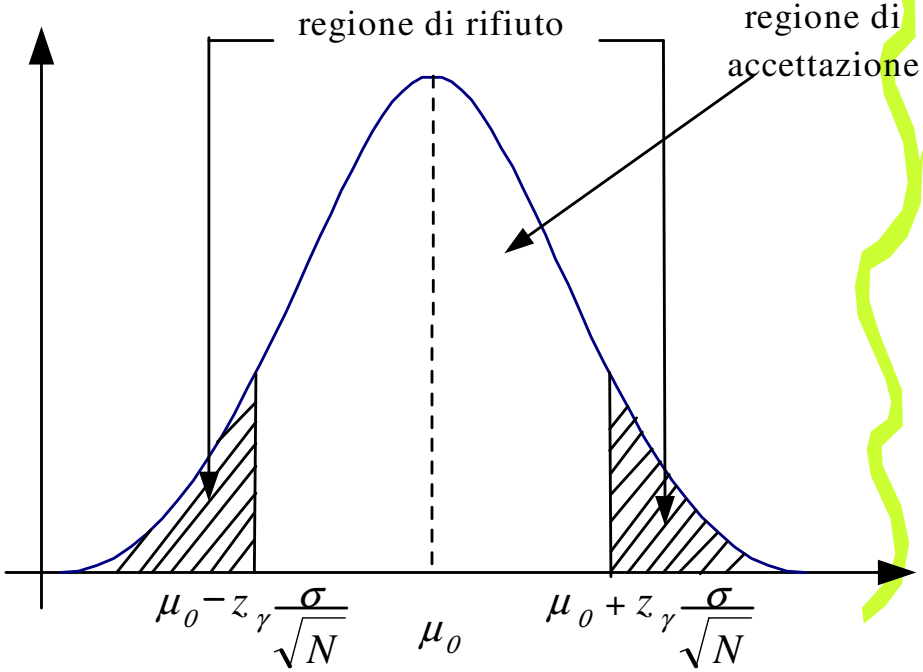
Questa incertezza corrisponde in pratica alla probabilità di rifiutare l'ipotesi anche se in effetti è vera (*errore del primo tipo*).

Se d'altro canto il valore ottenuto dal campione cade all'interno dell'intervallo non vi è motivo di rifiutare l'ipotesi

Esempio
(errore di I specie)

Dalla figura appare chiaro che diminuendo il valore di α, quindi diminuendo l'incertezza sul test, le due aree tratteggiate diminuiscono, e quindi aumenta la regione di accettazione

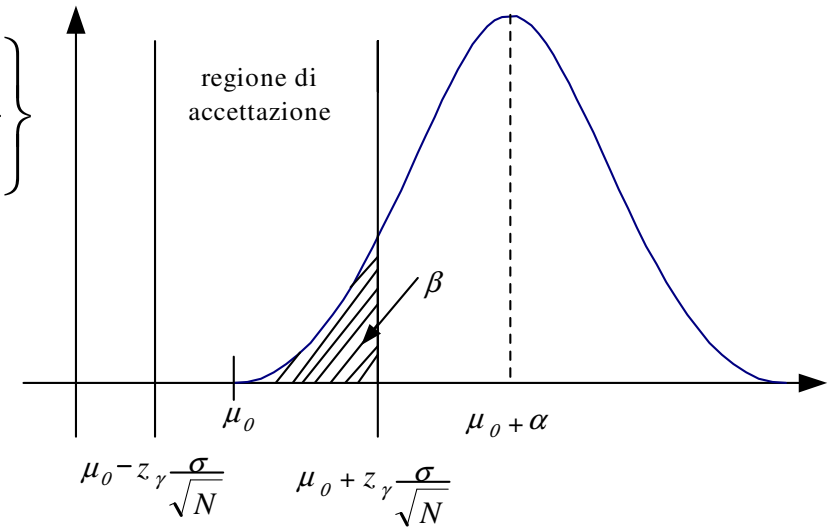
Per ovviare a tale inconveniente si può introdurre un fattore β che indica la probabilità di accettare l'ipotesi del test anche quando questa è falsa



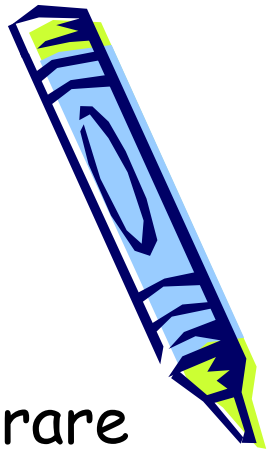
Esempio
(errore di II specie)

Quindi è la probabilità di accettare come vera l'ipotesi del test ($\mu = \mu_0$) quando in realtà $\mu = \mu_0 + \alpha$, quindi rappresenta l'errore di secondo tipo

$$\beta(\alpha) = \Pr\left\{\mu_0 - \frac{z_\alpha \sigma}{\sqrt{N}} < \overline{X} < \mu_0 + \frac{z_\alpha \sigma}{\sqrt{N}}\right\}$$



Test delle ipotesi



Si vuole utilizzare un test delle ipotesi per avvalorare o smentire un'ipotesi fatta sulla distribuzione di probabilità che meglio rappresenta i dati.

Quindi, date le osservazioni $X_1, \ldots, X_n$,
l'ipotesi da sottoporre a verifica `e la seguente:

H_0 = X1, . . . ,Xn sono variabili aleatorie indipendenti, identicamente distribuite con funzione di distribuzione F

Test delle chi-quadro

Per effettuare la prova si divide in s parti l'intervallo di definizione della X ed in base alla legge di distribuzione ipotizzata si calcolano le probabilità $(p_1, \ldots, p_S)$ che la variabile assume nella prima, nella seconda … nella s-esima parte

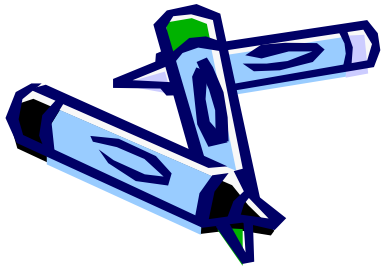
R_i la variabile che rappresenta il numero di elementi del campione che assumono un valore compreso nella i-esima parte, il sistema $(R_1, R_2, \ldots, R_S)$

$R_1 + R_2 + \ldots + R_S = N$

Test delle chi-quadro

I parametri del sistema appena descritto sono N la dimensione del campione e ($q_1,\ldots,q_S$) le probabilità che la variabile X assuma un valore nella prima, nella seconda … nella s-esima parte.

Se l'ipotesi fatta sulla distribuzione è vera allora queste probabilità coincideranno con le ($p_1,\ldots,p_S$) preventivamente calcolate



Test delle chi-quadro

Per la conduzione del test si costruisce la seguente variabile (statistica del test):

$$V = \sum_{i=1}^{S} \frac{(R_i - Np_i)^2}{Np_i}$$

E' quindi facile intuire che il valore assunto da *V* in corrispondenza di un particolare campione, può fornire una indicazione sull'accettabilità dell'ipotesi, nel senso che tanto più grande è *V* e tanto più l'ipotesi appare infondata

Test delle chi-quadro

V ha per $Np_i >> 1$, una distribuzione che approssimativamente è del χ^2 con s-1 gradi di libertà.

Pertanto fissato un certo livello di incertezza γ, dalle tabelle della densità di probabilità del χ^2 con s-1 gradi di libertà, si può ricavare $x\gamma$ tale che

$$\Pr\{\chi^2 \geq x_\gamma\} = \gamma$$

Test delle chi-quadro

A questo punto possiamo formulare un test di ipotesi statistiche del tipo:

$$\Pr\{0 < V < x_\gamma\} = 1 - \gamma$$

dove l'intervallo [0, $x\gamma$] definisce la regione di accettazione. L'ipotesi verrà rifiutata se il valore di V ottenuto da un particolare campione, è esterno alla regione di accettazione. L'errore del primo tipo che si commette è proprio uguale a γ. Se V cade all'interno della regione di accettazione non vi è motivo di rifiutare l'ipotesi del test, che quindi viene accettata

Test delle chi-quadro

L'errore del secondo tipo che si commette è difficile da esprimere.

Tutto il test si basa su un'ipotesi fondamentale, e cioè che la variabile V, abbia una legge di distribuzione del χ^2, affinché questo sia vero deve essere $N\,p_i >> 1$, e poiché le $p_i \in [0,1]$,

questa ipotesi è verifica se la dimensione del campione è abbastanza grande, da dividere l'intervallo di definizione in molte parti piccole, ed inoltre in ogni parte cada un numero elevato di valori del campione.

Esercizio

Gli incidenti di auto avvenuti in un anno su un tratto di strada sono indicati per ogni mese nella seguente tabella

Mese	Gen	Feb	Mar	Apr	Mag	Giu	Lug	Ago	Set	Ott	Nov	Dic.
Incid.	19	16	20	22	33	30	34	35	22	18	30	21

Si vuole provare con un livello di incertezza $\gamma = 0.01$, l'ipotesi che la probabilità evento incidente non dipenda dal particolare mese in cui accada

Esercizio

L'ipotesi che il numero di incidenti non dipenda dal mese, si traduce definendo una v.a. discreta X, che sia uniformemente distribuita nell'intervallo [1,12], con vettore delle probabilità: $p_i = 1/12$.

Per verificare se il numero degli incidenti segua questa legge di distribuzione utilizziamo il test del χ^2 definendo la variabile V

Esercizio

L'intervallo di definizione della distribuzione in esame, viene suddiviso chiaramente in dodici parti s=12. Al posto di *Ri* sostituiamo il relativo numero di incidenti nel relativo mese, mentre la dimensione del campione si ottiene sommando tutti i valori della tabella

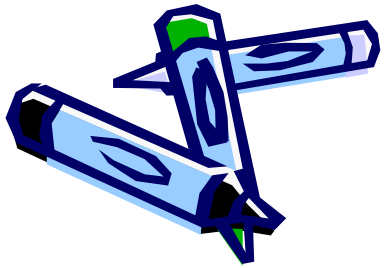
$$N = \sum_{i=1}^{12} R_i = 300 \Rightarrow Np_i = 25 \qquad V = \sum_{i=1}^{12} \frac{(R_i - 25)^2}{25} \cong 20.8$$

Esercizio

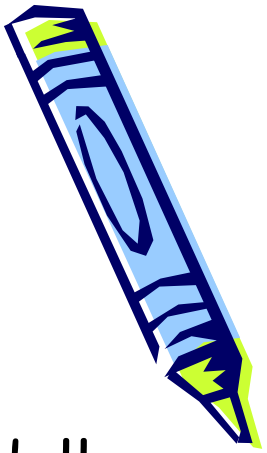
Poiché abbiamo visto che se l'ipotesi fatta è vera, V è approssimabile con una variabile del $\chi 2$ con $s-1 = 11$ gradi di libertà, dalle tabelle del $\chi 2$ ricaviamo $x\gamma$ tale che:

$$\Pr\{\chi^2 \geq x_\gamma\} = \gamma = 0.01 \Rightarrow x_\gamma = 24.72$$

Poiché $V < x\gamma$, l'ipotesi fatta è vera con un livello di fiducia pari al 99%.



Test di Kolmogorov – Smirnov



Questo tipo di test si concentra sulla forma della densità di probabilità del campione osservato e la confronta con quella ipotizzata. In realtà più che analizzare la densità di probabilità il test ricava la funzione di distribuzione procedendo in questo modo

Test di Kolmogorov – Smirnov

1- I valori del campione osservato vengono ordinati in funzione dell'ampiezza, mettendo al primo posto il valore con ampiezza minore e si passa da $(x_1, \ldots, x_N) \Rightarrow (x_{(1)}, \ldots, x_N)$;

2- Si costruisce la funzione di distribuzione osservata, che chiameremo $F_0(x)$, mediante la relazione $F_0[x(i)] = i/N$.

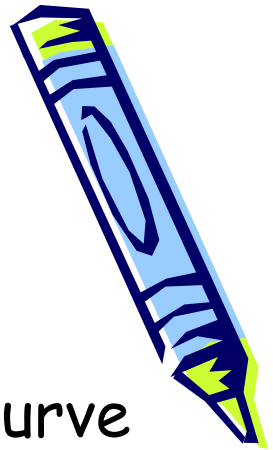
3- Si ricava una curva continua connettendo tutti i punti determinati al passo precedente

Test di Kolmogorov – Smirnov

Dopo aver costruito la funzione di distribuzione osservata, si mette a confronto con quella ipotizzata costruendo la seguente variabile *D* che verrà impiegata come statistica del test

$$D = \max_{i=1}^{N} \left\{ \left| F^0(X_{(i)}) - F_X(X_{(i)}) \right| \right\}$$

Test di Kolmogorov – Smirnov



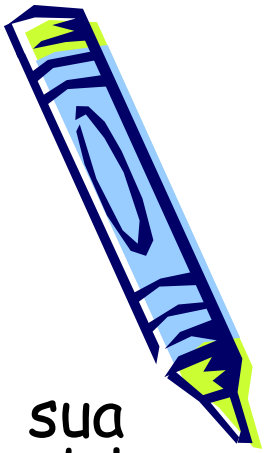
Praticamente *D* indica di quanto le due curve discostano nei punti osservati. A questo punto il test procede analogamente al test del χ^2, si fissa il livello di incertezza γ e da opportune tabelle specifiche per il test di Kolmogorov – Smirnov (K-S) si ricava il valore d_γ da usare nel test

$$\Pr\{(K-S) \geq d_\gamma\} = \gamma$$

Infine si confronta la statistica del test con d_γ e l'ipotesi fatta sulla funzione di distribuzione viene accettata se

$$\Pr\{0 < D < d_\gamma\} = 1-\gamma$$

Test di Kolmogorov – Smirnov



Il vantaggio di impiegare questo test è nella sua semplicità di implementazione, e soprattutto dal fatto che l'esito del test non è condizionato dalla dimensione del campione

Però per calcolare la statistica del test occorre conoscere la densità di probabilità o la funzione di distribuzione della curva ipotizzata dal test.

Le tabelle (K-S) sono calcolate su distribuzioni completamente note.

Solo in alcuni casi (distribuzione ipotizzata gaussiana) si riesce ad utilizzare questo test quando i parametri della distribuzione non sono noti

Esercizio

Dato il campione X = (30.1, 30.5, 28.7, 31.6, 32.5, 29, 27.4, 29.1, 33.5, 31), cerchiamo di verificare se si tratta di un campione gaussiana, mediante il test di Kolmogorov – Smirnov.

Poiché stiamo facendo l'ipotesi di distribuzione gaussiana, possiamo utilizzare dei valori stimati per descrivere la funzione di distribuzione teorica

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{10}\sum_{i=1}^{10} x_i = 30.34$$

$$\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu}_{ML})^2 = \frac{1}{10}\sum_{i=1}^{10} (x_i - 30.3)^2 = 3.142$$

Esercizio

I valori della distribuzione ipotizzata si possono ricavare dalle tabelle per la gaussiana N(30.34, 3.142),

Poiché abbiamo a disposizione la tabella della gaussiana standard $\Phi(z)$ si procede nel seguente modo:

$$F_X(x) = \Phi\left(\frac{x - \hat{\mu}_{ML}}{\sqrt{\hat{\sigma}^2_{ML}}}\right)$$

Esercizio

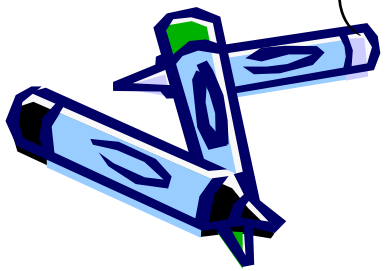
$$F_X(27.4) = \Phi\left(\frac{27.4-30.34}{\sqrt{3.142}}\right) = 0.0486 \quad F_X(28.7) = \Phi\left(\frac{28.7-30.34}{\sqrt{3.142}}\right) = 0.1774$$

$$F_X(29) = \Phi\left(\frac{29-30.34}{\sqrt{3.142}}\right) = 0.2248 \quad F_X(29.1) = \Phi\left(\frac{29.1-30.34}{\sqrt{3.142}}\right) = 0.2421$$

$$F_X(30.1) = \Phi\left(\frac{30.1-30.34}{\sqrt{3.142}}\right) = 0.4462 \quad F_X(30.5) = \Phi\left(\frac{30.5-30.34}{\sqrt{3.142}}\right) = 0.5360$$

$$F_X(31) = \Phi\left(\frac{31-30.34}{\sqrt{3.142}}\right) = 0.6452 \quad F_X(31.6) = \Phi\left(\frac{31.6-30.34}{\sqrt{3.142}}\right) = 0.7614$$

$$F_X(32.5) = \Phi\left(\frac{32.5-30.34}{\sqrt{3.142}}\right) = 0.8885 \quad F_X(33.5) = \Phi\left(\frac{33.5-30.34}{\sqrt{3.142}}\right) = 0.9627$$



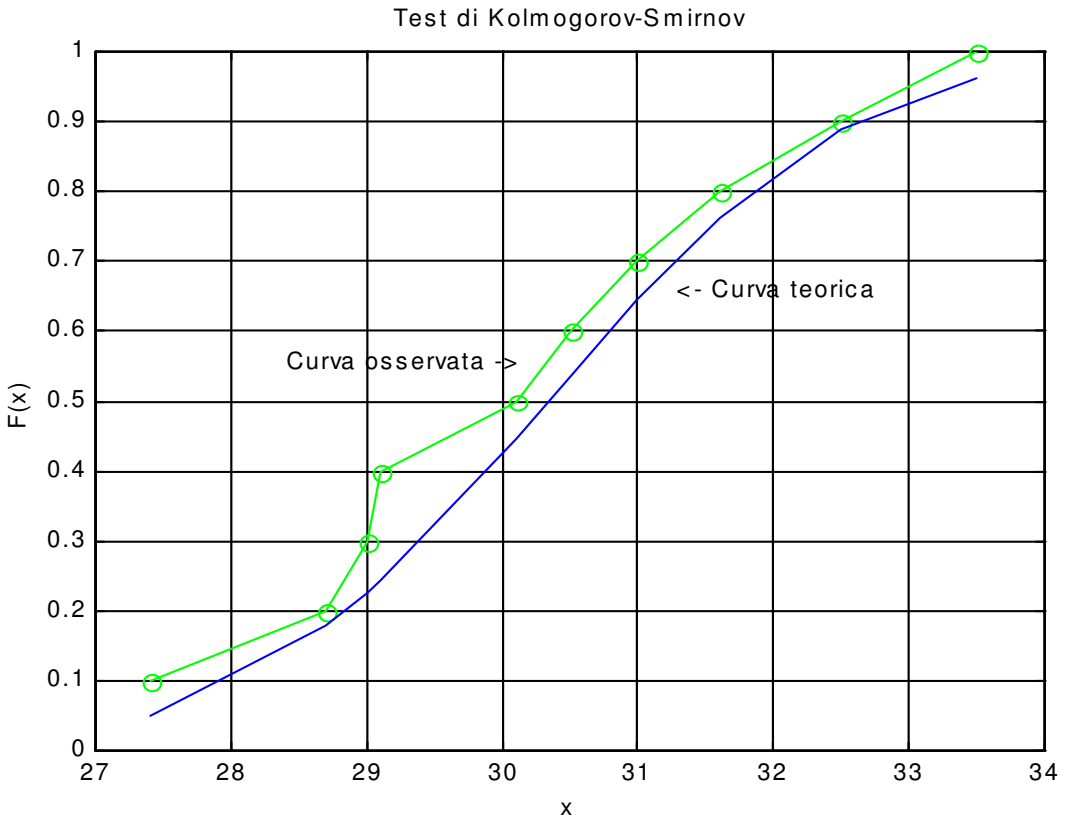
Esercizio

mentre la funzione distribuzione osservata la ricaviamo nel seguente modo

$$F^0(27.4)=0.1, \quad F^0(28.7)=0.2 \quad F^0(29)=0.3 \quad F^0(29.1)=0.4 \quad F^0(30.1)=0.5$$

$$F^0(30.5)=0.6, \quad F^0(31)=0.7 \quad F^0(31.6)=0.8 \quad F^0(32.5)=0.9 \quad F^0(33.5)=1.0$$

Esercizio


Test di Kolmogorov-Smirnov
F(x)
x
Curva osservata ->
<- Curva teorica
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
27
28
29
30
31
32
33
34

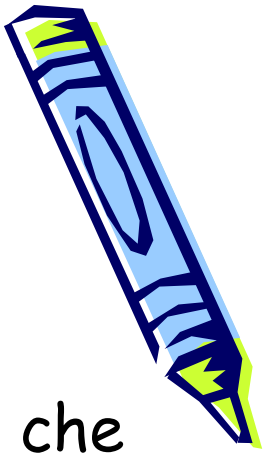
Esercizio

Il punto di massima distanza dalla curva ipotizzata è $D = 0.1579$.

Se poniamo un livello di incertezza pari a $\gamma = 0.05$, le tabelle (K – S) per la curva gaussiana forniscono per $N = 10$, un valore $d\gamma = 0.41$, poiché la statistica del test D è minore di questo valore, l'ipotesi fatta (campione gaussiano) può essere accettata


Analisi dei dati di input

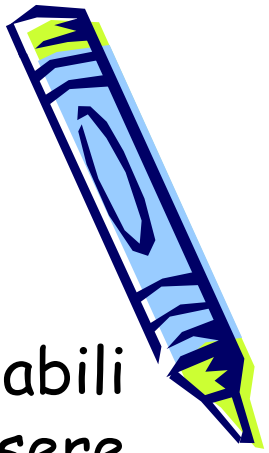
Analisi dati di input



Per condurre una simulazione di un sistema che presenta elementi stocastici è necessario specificare le distribuzioni di probabilità che regolano i processi che caratterizzano il sistema.

Una volta stabilite tali distribuzioni la simulazione procede generando valori casuali da queste distribuzioni

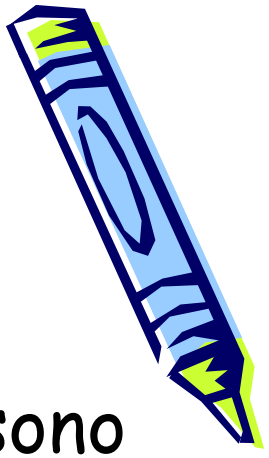
Analisi dati di input



Se è possibile raccogliere dati reali sulle variabili aleatorie di interesse, essi possono essere utilizzati per determinare una distribuzione secondo tre metodi

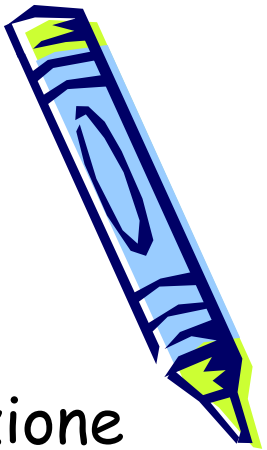
1. I dati sono usati *direttamente* nella simulazione
2. I dati sono raccolti per generare una *distribuzione empirica*
3. I dati sono utilizzati per definire una *distribuzione teorica*

"trace driver simulation"



Ha senso solamente quando si possono raccogliere grandi quantità di dati rappresentativi del funzionamento del sistema; ha l'ovvio difetto di rappresentare il "passato" ed è usato raramente; può essere utile per effettuare una validazione del modello, ovvero per confrontare il modello con il sistema reale, ma non permette un'analisi previsionale.

Distribuzione empirica

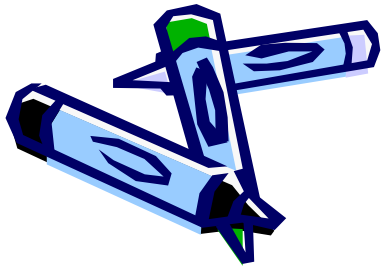


I dati sono raccolti per generare una distribuzione empirica, ovvero per definire una funzione di distribuzione empirica che verrà usata per produrre l'input della simulazione

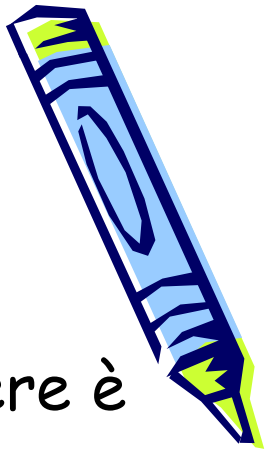
Tale approccio elimina inconveniente di non poter fare previsioni, poiché, almeno per distribuzioni continue, può essere ottenuto ogni valore compreso tra il minimo e il massimo osservati.

Distribuzione teorica

I dati raccolti sono utilizzati per definire una distribuzione teorica. Vengono utilizzate tecniche statistiche per analizzare se una distribuzione teorica tra quelle note sia adatta a rappresentare i dati, effettuando i test di ipotesi per verificare la rappresentatività della distribuzione ipotizzata (problema del "fitting").



Distribuzione teorica



I motivi per cui una distribuzione teorica in genere è preferibile a una empirica sono i seguenti:

- le distribuzioni empiriche possono avere irregolarità (specialmente se i dati sono scarsi) mentre le distribuzioni teoriche sono più "smooth", nel senso che tendono a regolarizzare i dati e rappresentano un comportamento generale;

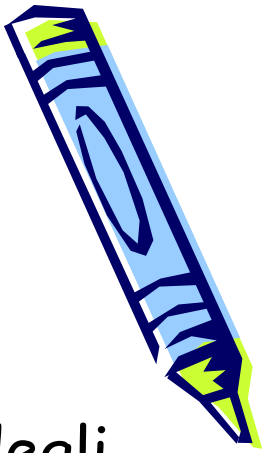
- le distribuzioni empiriche non permettono di generare valori al di fuori del range di valori osservati, mentre le misure di prestazione possono, a volte,dipendere anche da eventi "eccezionali" che corrispondono a valori fuori da tale range;

Distribuzione teorica

- Le distribuzioni teoriche sono un modo compatto di rappresentare un insieme di valori, mentre in una distribuzione empirica, se ci sono n dati disponibili, si ha bisogno di 2n valori per rappresentarla: il dato e le corrispondenti probabilità cumulative (si hanno quindi grandi quantità di dati da memorizzare);

- Le distribuzioni teoriche si possono variare più facilmente.

Distribuzione teorica

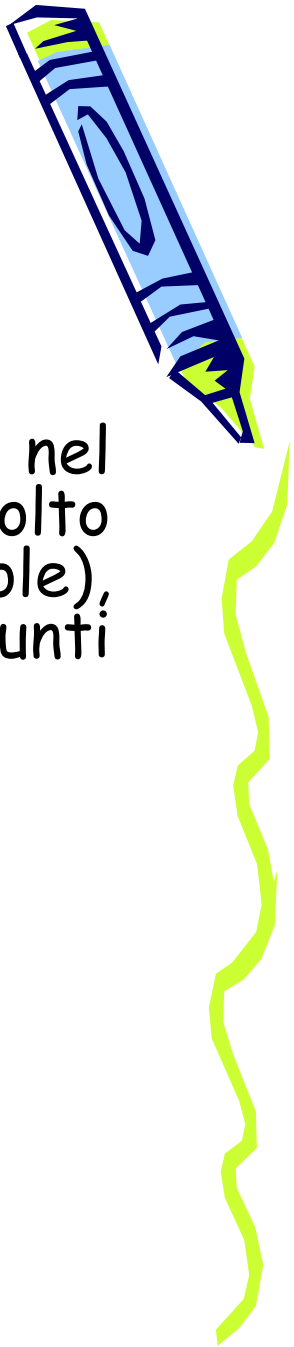


Ad esempio se la distribuzione esponenziale degli arrivi di un sistema di code ha media pari a

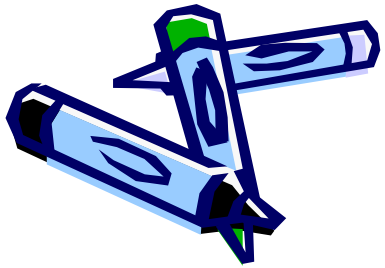
$\lambda = 5$, per effettuare una diminuzione del 20% sarà sufficiente considerare $\lambda = 4.9$

Tuttavia esistono situazioni in cui nessuna distribuzione teorica si adatta ai dati osservati e allora in questo caso si deve usare una distribuzione empirica.

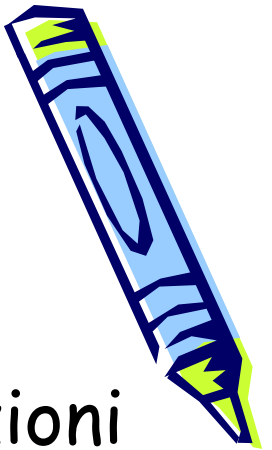
Distribuzione teorica



Un difetto dell'uso di distribuzioni teoriche sta nel fatto che esse possono generare anche valori molto grandi (anche se con probabilità molto piccole), quando nella pratica questi non vengono mai assunti realmente



Distribuzione empirica

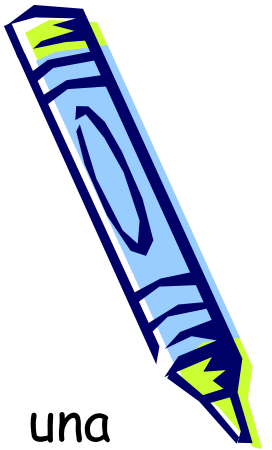


Supponiamo di disporre di n osservazioni $X_1,\ldots,X_n$ di una variabile aleatoria e di voler costruire, a partire da esse, una distribuzione continua.

Ordiniamo le Xi per valori crescenti e sia X(i) l'i-esima osservazione in ordine crescente, ovvero risulti

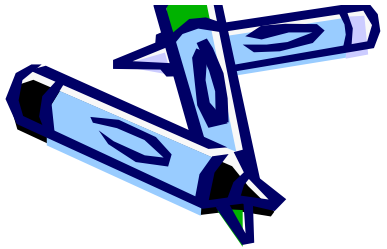
$$X(1) \leq X(2) \leq \ldots \leq X(n).$$

Distribuzione empirica

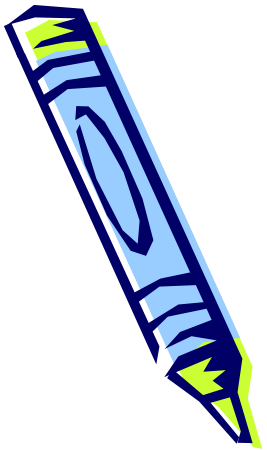


Si può costruire la distribuzione empirica come una distribuzione continua lineare a tratti, così definita:

$$F(x) = \begin{cases} 0 & \text{se} \quad x < X_{(1)} \\ \dfrac{i-1}{n-1} + \dfrac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})} & \text{se} \quad X_{(i)} \le x < X_{(i+1)}, \\ & \qquad i = 1, \ldots, n-1 \\ 1 & \text{se} \quad X_{(n)} \le x \end{cases}$$



Distribuzione empirica



Si osservi che per ogni i vale

$$F(X_{(i)}) = \frac{i-1}{n-1}$$

che è approssimativamente (per n grande) la proporzione dei campioni che sono minori di X(i).

Uno svantaggio nell'utilizzare una distribuzione empirica è che le variabili aleatorie generate da essa durante un'esecuzione di una simulazione non possono essere mai più piccole di X(1) o più grandi di X(n).

Distribuzione empirica

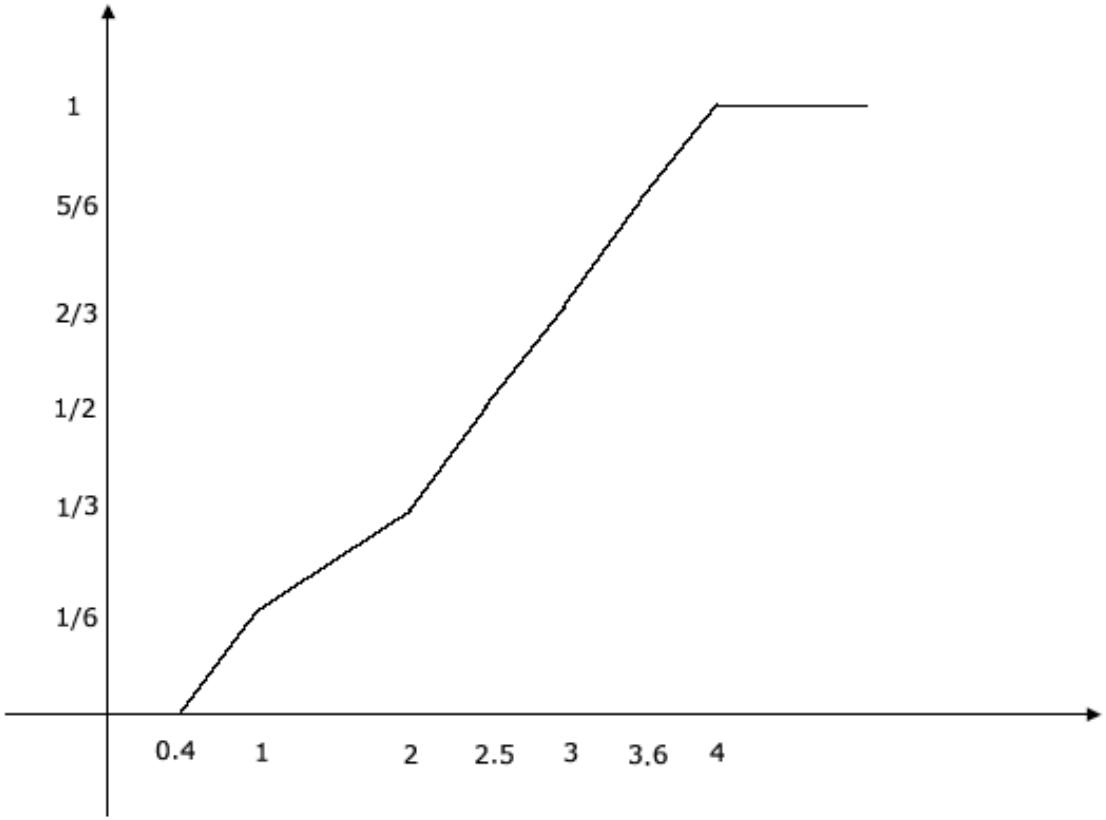
Analogamente si possono costruire distribuzioni empiriche per distribuzioni discrete; infatti, `e sufficiente per ogni x definire p(x) come "proporzione" delle Xi che sono uguali ad x.

ESEMPIO

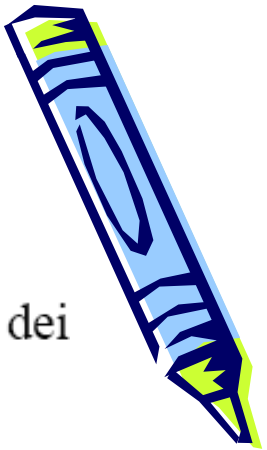
Disponendo dei seguenti valori osservati: 1, 0.4, 4, 2, 2.5, 3.6, 3 costruire il grafico della distribuzione empirica

Esempio

Dopo aver ordinato le osservazioni si ottiene il grafico della F(x)



Passi dell'analisi



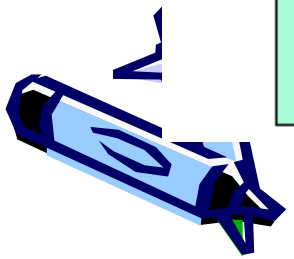
Analisi preliminare (già svolta a questo livello) per l'identificazione dei componenti del sistema e dei reciproci nessi causali.

Definizione dati di input:campagna dati per l'identificazione dei parametri caratterizzanti le entità del sistema:

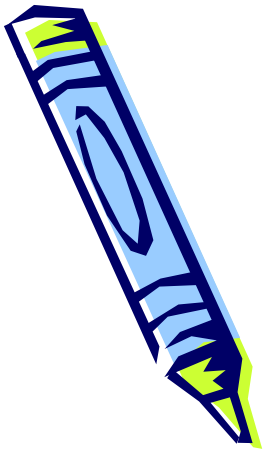
- parametri deterministici
- parametri stocastici

N.B. Sono errori comuni:

- ⇨assumere due parametri indipendenti quando non lo sono (si interviene quindi erroneamente su uno di essi ritenendo che non influenzi l'altro)
- ⇨assumere due parametri dipendenti quando non lo sono (si interviene quindi erroneamente su uno di essi ritenendo di potere modificare di conseguenza anche l'altro)



Indipendenza delle osservazioni

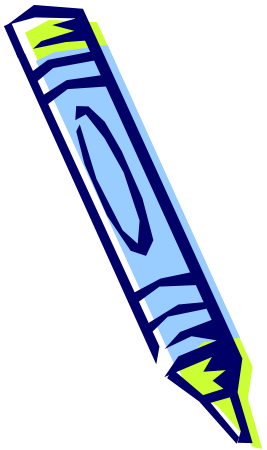


Un primo strumento di analisi è basato su una tecnica grafica. Siano $X_1,\ldots,X_n$ le osservazioni elencate così come sono state osservate nel tempo; un modo possibile per avere un'idea informale sull'indipendenza consiste nel valutare la correlazione fra diverse osservazioni. Sia la stima del coefficiente di correlazione di Xi e Xi+j , ovvero di due osservazioni distanti j.

$$\widehat{\rho}_j = \frac{\sum_{i=1}^{n-j}(X_i - \bar{X}_n)(X_{i+j} - \bar{X}_n)}{(n-j)s_n^2}$$

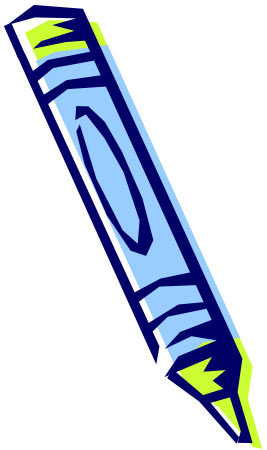
Se le osservazioni sono indipendenti allora il coefficiente di correlazione è nullo

Indipendenza delle osservazioni



Tuttavia poiché è una stima anche nel caso di osservazioni indipendenti potrebbe essere non nullo.
Ci si aspetta, comunque che esso sia prossimo a zero, e quindi possiamo dire che se è diverso da zero in maniera significativa, allora le Xi non sono indipendenti.

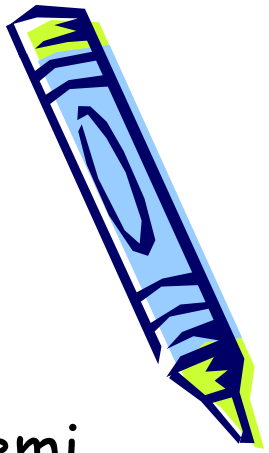
Indipendenza delle osservazioni



Ci sono due modi grafici per verificare informalmente se le Xi sono indipendenti: il grafico della stima ρj al variare di j e il diagramma di dispersione delle osservazioni X1, . . . ,Xn, ovvero le coppie (Xi,Xi+1) con i = 1, …,n – 1.

In caso di osservazioni indipendenti i punti dovrebbero risultare distribuiti casualmente sul piano, altrimenti, in presenza di correlazioni, essi saranno concentrati intorno a rette.

Passi dell'analisi



Definizione di parametri deterministici.

Teoricamente non si pongono particolari problemi, se non eventualmente quelli associati alla presenza di rumore nella misura di tali parametri.

Praticamente molte aziende hanno informazioni poco affidabili; sono quasi sempre necessarie campagne di misura, inventario o catalogazione.

Definizione di parametri stocastici.

Teoricamente e ***praticamente*** si pongono notevoli problemi la cui soluzione richiede generalmente un significativo investimento di tempo e denaro.

Passi dell'analisi

L'identificazione delle caratteristiche statistiche dei parametri stocastici presenti nel modello in esame si compie a partire da un vasto insieme di misurazioni reali.
Per ognuno dei parametri stocastici inseriti nel modello è necessario eseguire una campagna di raccolta di misure reali e, quindi, utilizzare la procedura basata sui passi seguenti.

Passi dell'analisi

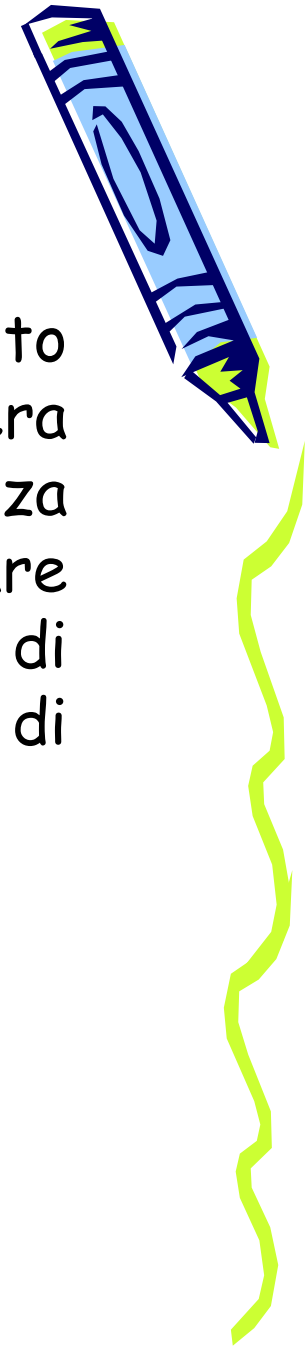
1) costruzione della "**distribuzione di frequenza**" o "**istogramma delle frequenze**" dei campioni raccolti per il parametro in esame;

2) **selezione** di una particolare **funzione di densità di probabilità** per il parametro in esame;

3) **stima dei parametri** della funzione di densità di probabilità scelta;

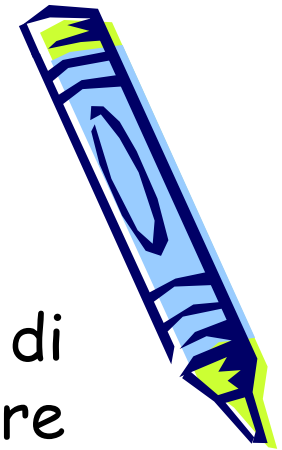
4) **validazione** della distribuzione scelta per mezzo di un test "goodness-of-fit".

Istogramma delle frequenze



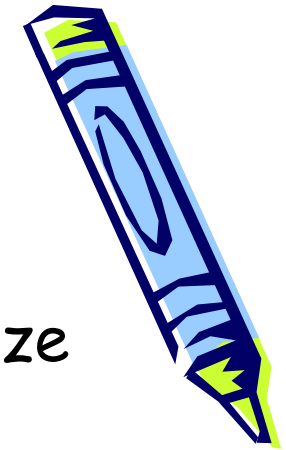
L'istogramma delle frequenze è una tecnica molto semplice che permette di verificare in maniera qualitativa, la densità di probabilità della sequenza generata. Soprattutto è semplice da implementare ed ha un forte impatto visivo perché permette di disegnare un grafico approssimato della densità di probabilità.

Istogramma delle frequenze



L'idea si basa sulla definizione di densità di probabilità,(analoga al test del χ^2), cioè suddividere l'intervallo di definizione della variabile in s sottointervalli uguali e contare quanti valori della sequenza generata cadono in ogni intervallo, in questo modo si verifica quanto è denso di valori ogni intervallino (da qui discende l'approssimazione della densità di probabilità). Gli intervallini più densi di valori corrisponderanno a zone della legge di distribuzione con più alta probabilità di avere un valore della variabile aleatoria, mentre quelli meno densi, corrisponderanno a zone con probabilità più basse.

Istogramma delle frequenze



Per la costruzione dell'istogramma delle frequenze sono necessari i passi seguenti:

- divisione del range di variazione dei dati in sottointervalli di uguale ampiezza
- etichettatura dell'asse orizzontale con i sottointervalli selezionati;
- determinazione delle frequenze di occorrenza della variabile in ogni intervallo;
- etichettatura dell'asse verticale con i valori di frequenza individuati;
- disegno delle frequenze nel piano definito.

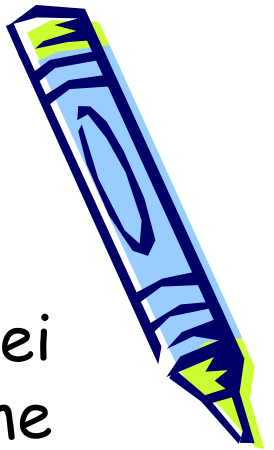
Istogramma delle frequenze

La scelta dell'ampiezza dei sottointervalli è cruciale infatti:

- se l'intervallo è troppo ampio, l'istogramma risulta troppo aggregato e non consente di individuare una funzione di densità di probabilità

- se l'intervallo è troppo piccolo, l'istogramma evidenzia troppo eventuali picchi negativi e positivi risultando troppo "brusco".

Esempio

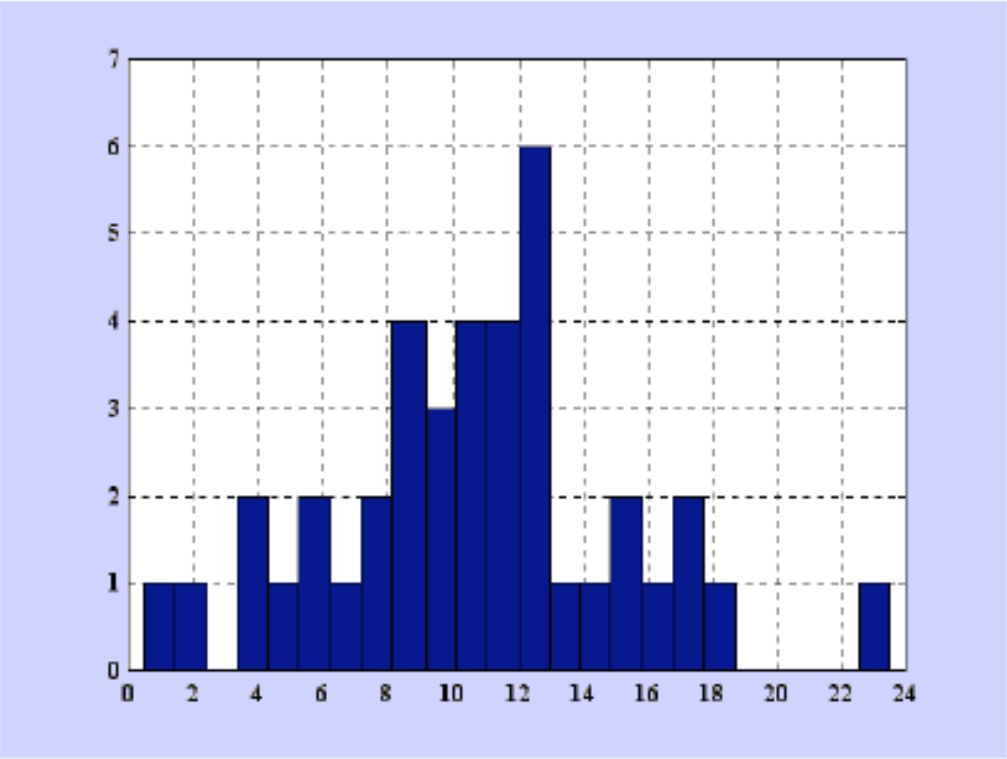


Si vuole disegnare l'istogramma delle frequenze dei dati relativi alla durata dei viaggi aerei di un insieme di passeggeri.
I campioni raccolti sono (unità di misura=ora):

0.5, 2.1, 4.1, 4.6, 5.7, 6.2, 6.6, 7.8, 8.1, 8.3, 8.4, 8.6,
8.9, 9.2, 9.8, 10.0, 10.3, 10.5, 10.6,
10.8, 11.2, 11.3, 11.6, 11.7, 12.1, 12.5, 12.6, 12.8, 12.9,
12.9, 13.2, 14.4, 15.0, 15.5, 16.3,
17.0, 17.3, 18.5, 23.5.

Costruzione dell'istogramma

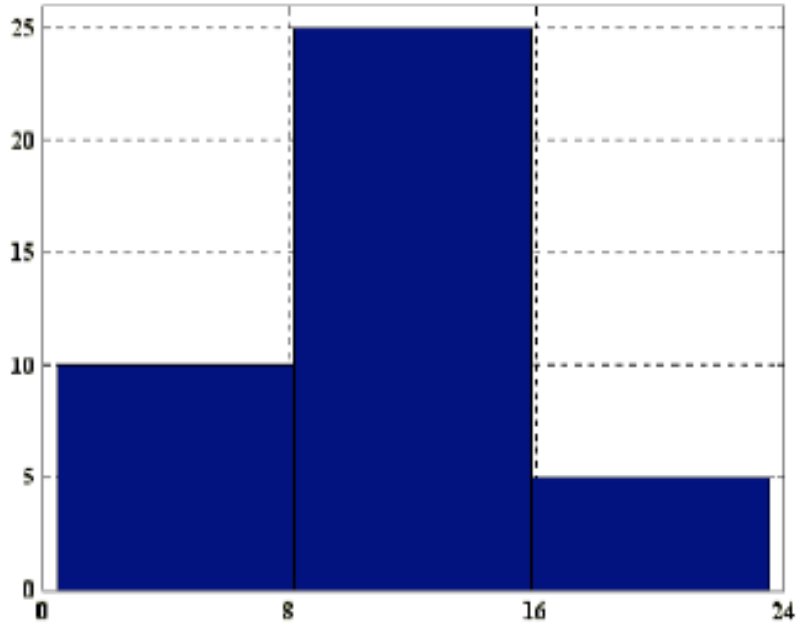
Scegliendo intervalli di durata 1 ora, l'istogramma risulta essere il seguente.



➔ L'ampiezza dei sottointervalli è troppo piccola

Costruzione dell'istogramma

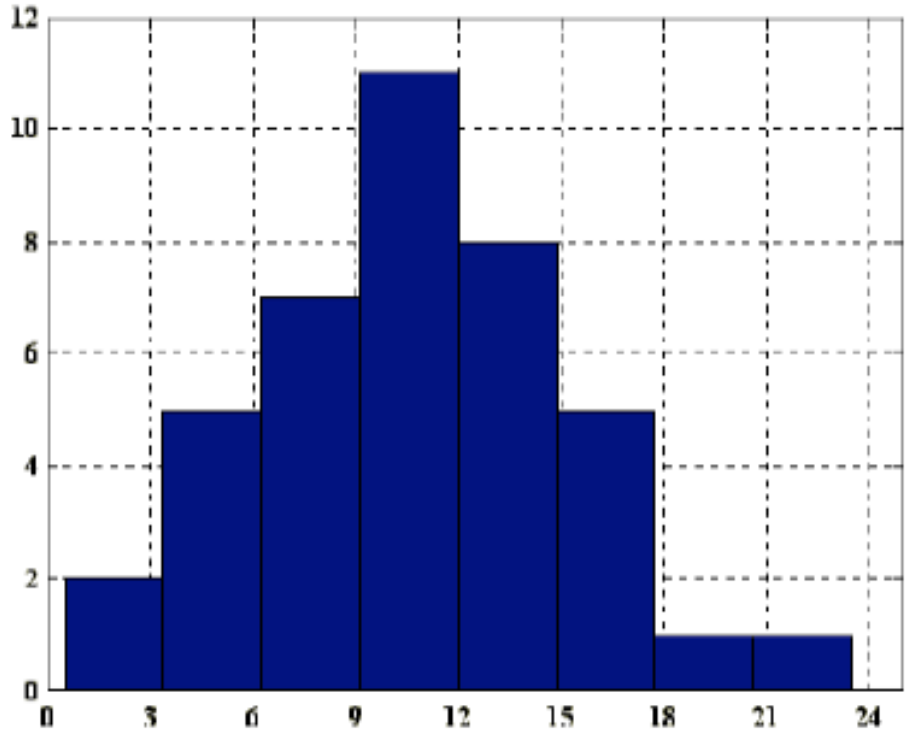
Scegliendo intervalli di durata 8 ore, l'istogramma risulta essere il seguente.



→ L'ampiezza dei sottointervalli è troppo grande

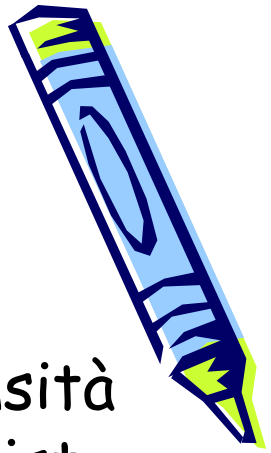
Costruzione dell'istogramma

Scegliendo intervalli di durata 3 ore, l'istogramma risulta essere il seguente.



→ L'ampiezza dei sottointervalli è corretta

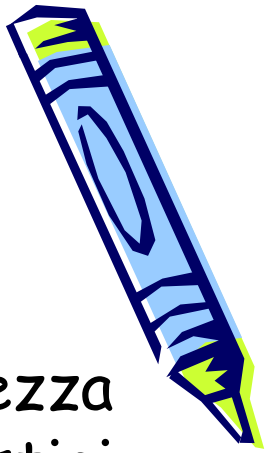
Analisi dei dati



Il passo 2), ossia la selezione di una funzione di densità di probabilità è molto semplice e consiste nell'individuare tale funzione sulla base della forma dell'istogramma delle frequenze (che, quindi, deve essere eseguito in maniera attenta e corretta).

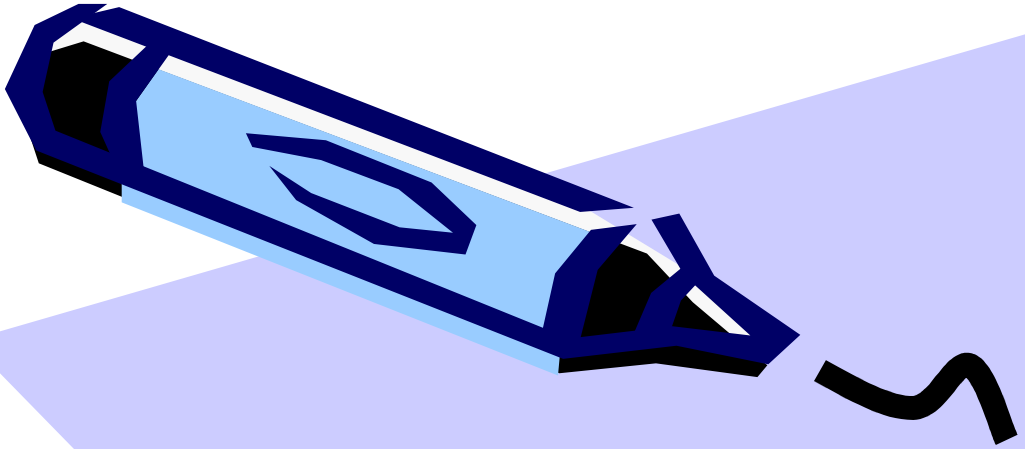
Il passo 3) riguarda, invece, la stima dei parametri della funzione di densità scelta. E' necessario, innanzitutto, calcolare le media campionaria e la varianza campionaria; quindi, sulla base di tali valori si consultano delle tabelle che contengono i cosiddetti "stimatori a massima verosimiglianza".

Analisi dei dati



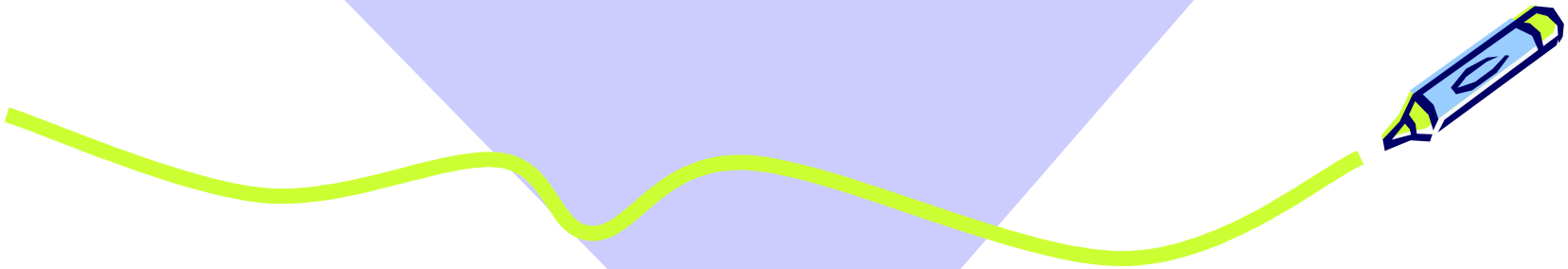
Il passo 4) prevede la verifica di correttezza dell'ipotesi distribuzionale eseguita. I test statistici più diffusi per questo scopo sono il test "**chi-quadro**" (o "chi-square") e il test **Kolmogorov-Smirnov**.

Una volta determinate le distribuzioni di input, la simulazione dovrà generare durante ogni esecuzione osservazioni casuali di variabili aleatorie distribuite secondo particolari distribuzioni di probabilità



Generazione di numeri random

Distribuzioni uniformi



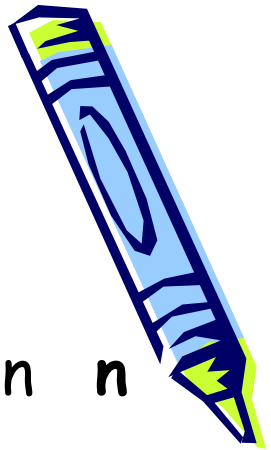
I numeri random

Per **numero random** (o numero casuale) si intende una variabile aleatoria distribuita in modo **uniforme tra 0 e 1**.

Le proprietà statistiche che una sequenza di numeri random deve possedere sono:

- uniformità
- indipendenza

I numeri random



Si supponga di dividere l'intervallo [0,1] in **n** sottointervalli di uguale ampiezza.

Conseguenza della proprietà di uniformità è:

- Se si eseguono **N** osservazioni di un numero casuale, il numero di osservazioni in ogni sottointervallo è pari a **N/n**.

Conseguenza della proprietà di indipendenza è:

- la probabilità di ottenere un valore in un particolare intervallo è indipendente dai valori precedentemente ottenuti.

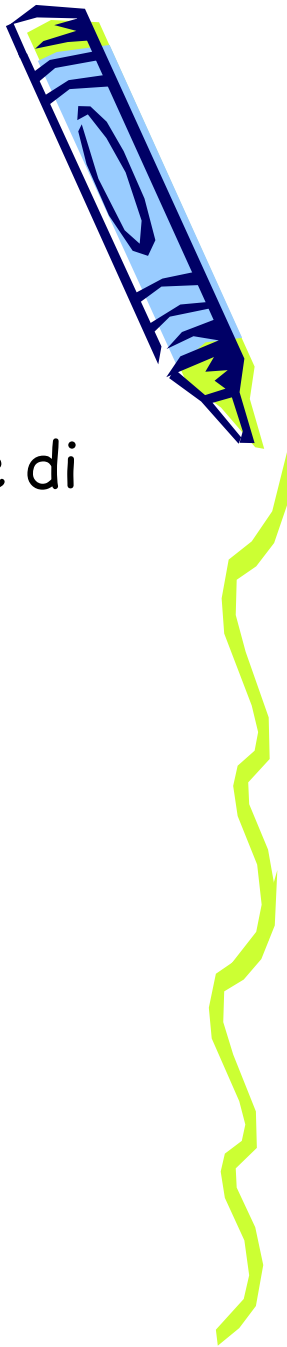
Generazione di numeri random

I numeri generati da una **routine di generazione** di numeri casuali, sono, in realtà numeri *pseudo-casuali.*

Una routine di generazione di numeri casuali deve:

- essere veloce
- avere un ciclo (periodo) sufficientemente lungo
- non presentare larghi gap (intervalli tra due numeri generati)
- essere replicabile
- generare numeri con proprietà statistiche più vicine possibile a quelle ideali.

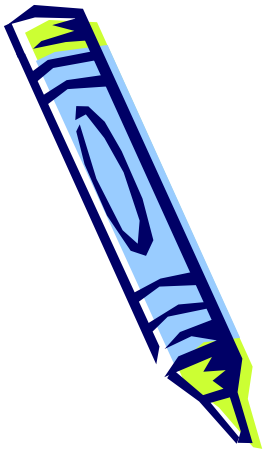
Generazione di numeri random



I difetti più comuni di una routine di generazione di numeri casuali sono:

- numeri non uniformemente distribuiti
- discretizzazione dei numeri generati
- media o varianza non corrette
- presenza di variazioni cicliche.

Tecnica di congruenza lineare (Lehmer, 1951)

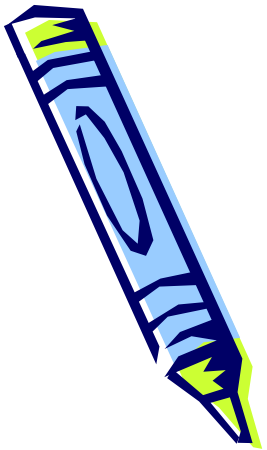


Uno dei metodi più utilizzati per la generazione di numeri casuali è la **Tecnica di congruenza lineare** (***Linear Congruential Method***).

•La relazione ricorsiva alla base di tale tecnica è:

$$x_{k+1} = (ax_k + c) \bmod m$$

Tecnica di congruenza lineare



dove:

- a → moltiplicatore
- c → incremento
- m → modulo
- x_0 → valore iniziale detto *seme*

a, c, m e x_0 sono numeri interi non negativi

l'operazione **mod** rappresenta il resto della divisione.

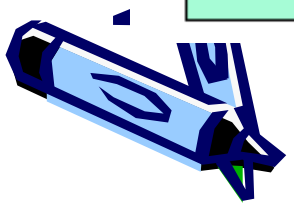
Ogni singola istanza di numero casuale è ottenuta come $u_k = x_k / m$

Tecnica di congruenza lineare

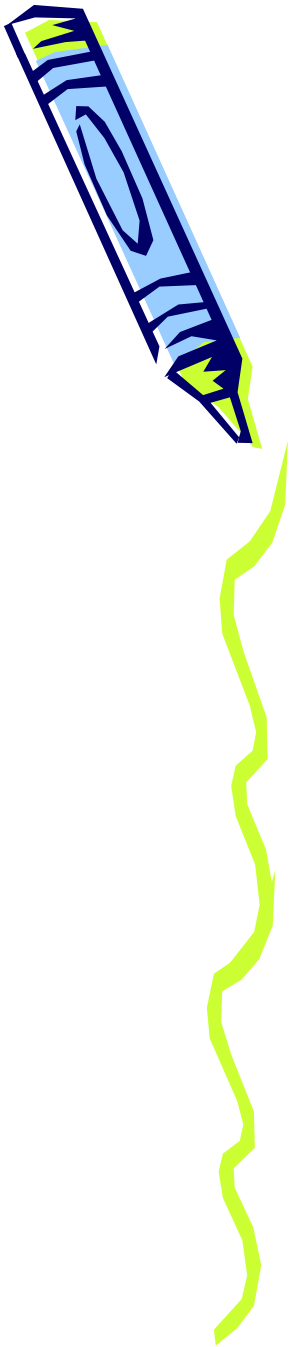
La tecnica di congruenza lineare presenta le seguenti caratteristiche:

- ☹ è ciclica (con periodo circa pari a m)
- ☹ i numeri generati sono discretizzati, infatti u_k può assumere solo i valori $0,1/m,2/m, \ldots, (m\text{-}1)/m$ (in realtà u_k potrebbe assumere ogni valore nell'intervallo, ad esempio, [0.5/m, 0.6/m], con probabilità 0.1/m, ma la probabilità che questo avvenga è 0).

N.B. Per utilizzare in maniera efficace la tecnica di congruenza lineare è necessario scegliere **valori molto grandi di *m***

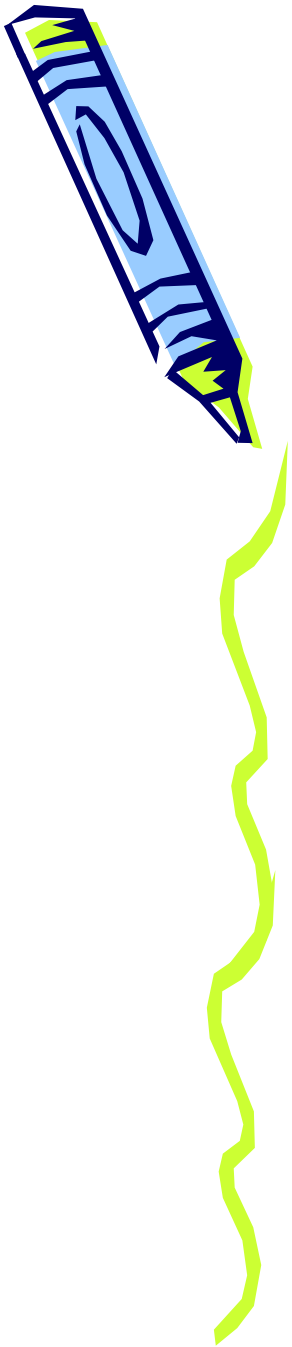


Esempio



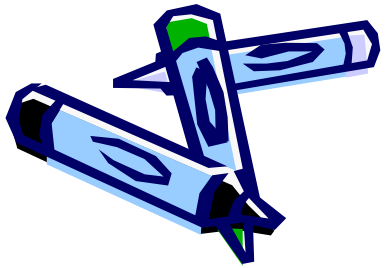
$a = 1, c = 5, m = 4, x_0 = 2$

Esempio

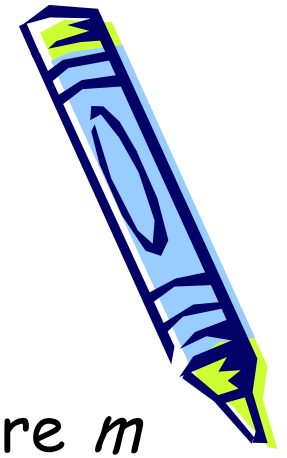


$a = 1, c = 5, m = 4, x_0 = 2$

x1 = 3
x2 = 0
x3 = 1
x4 = 2
x5 = 3



Distribuzione Uniforme



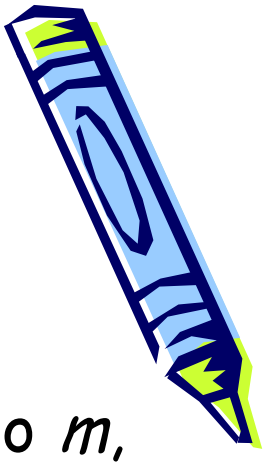
E' facile verificare che al più si possono generare *m* numeri interi *Xn* distinti nell'intervallo [0, *m* – 1].

In particolare se *c* = 0, il generatore viene detto *moltiplicativo*.

Una sequenza di numeri uniformante distribuita tra [0,1] può essere ottenuta nel seguente modo:

$$U = \frac{X_n}{m}$$

Distribuzione Uniforme



La sequenza ottenuta è periodica al più di periodo *m*, in particolare si dice che ha periodo pieno se il suo periodo è proprio *m*, e ciò si verifica quando sono verificate le seguenti condizioni:

- Se *m* e c sono primi tra loro;
- Se *m* è divisibile per un numero primo *b*, per il quale deve essere divisibile anche *a* – 1;
- Se *m* è divisibile per 4, allora anche *a* – 1 deve essere divisibile per 4.

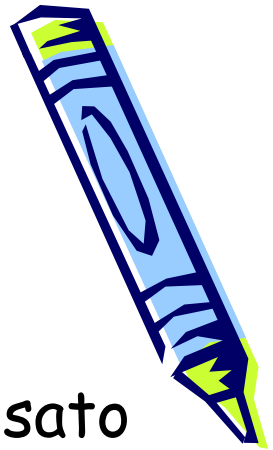
Osservazioni

Scegliendo un valore di *m* abbastanza grande si può ridurre sia il fenomeno della periodicità sia il fatto di generare numeri razionali.

Inoltre non è necessario ai fini della simulazione che vengano generati tutti i numeri tra [0,1], anche perché questi sono infiniti, ma è sufficiente che quanti più numeri possibili all'interno dell'intervallo abbiamo la stessa probabilità di essere generati.

Generalmente un valore di m è ($m \geq 10^9$) in modo che i numeri generati Ui costituiscono un sottoinsieme denso dell'intervallo [0, 1).

Esempi di generatori

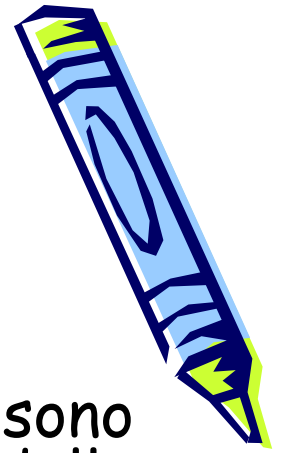


Un esempio di generatore moltiplicativo molto usato nei calcolatori a 32 bit è:

- (c = 0, a = 75, m = 2^{31} - 1) noto anche come generatore di Learmouth - Lewis.

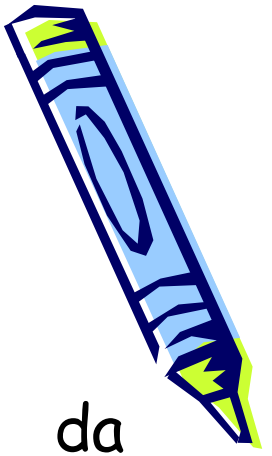
- un esempio di generatore non puramente moltiplicativo è (c = 453806245, a = 314159269 m = 2^{15}).

Osservazioni



Il confronto tra i diversi generatori che possono essere utilizzati va effettuato, sull'analisi della *periodicità*, la bontà dell'*uniformità* dei numeri generati e la *semplicità computazionale*, perché la generazioni di numeri troppo grandi può portare ad un impiego oneroso delle risorse del calcolatore, inoltre se i numeri *Xn* diventano troppo grandi, vengono troncati, e questo può causare una perdita delle statistiche di uniformità desiderate.

Generatori additivi



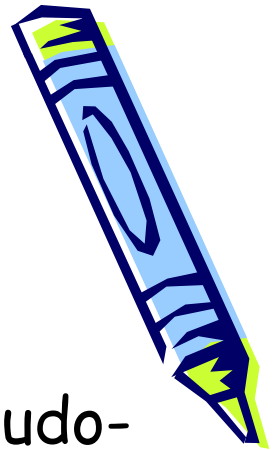
Esistono anche generatori più semplici da implementare, che però forniscono prestazioni inferiori.

E' il caso dei generatori puramente *additivi* basati sulla serie di Fibonacci:

$Xn = (X_{n-1} + X_{n-2})$ mod (m)

Un esempio è il generatore con $X1 = X0 = 1$, e $m = 2^{32}$

Test di uniformità

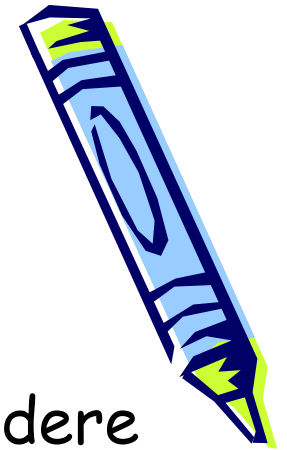


Dopo aver generato la sequenza numerica pseudo-casuale, occorre verificare la bontà della sequenza ottenuta.

Chiaramente si tratta di verificare se la sequenza ottenuta (che costituisce un campione casuale dell' esperimento) segue una distribuzione uniforme.

Per effettuare questa verifica è possibile utilizzare il test del $\chi 2$

Test di uniformità



La prima operazione da effettuare è dividere l'intervallo [0,1] in s sottointervalli della stessa lunghezza. Successivamente si contano quanti numeri della sequenza cadono nell'i-esimo intervallino:

$$R_i = \left| \left\{ x_j \middle| x_j \in s_i, j = 1 \ldots N \right\} \right|$$

Test di uniformità

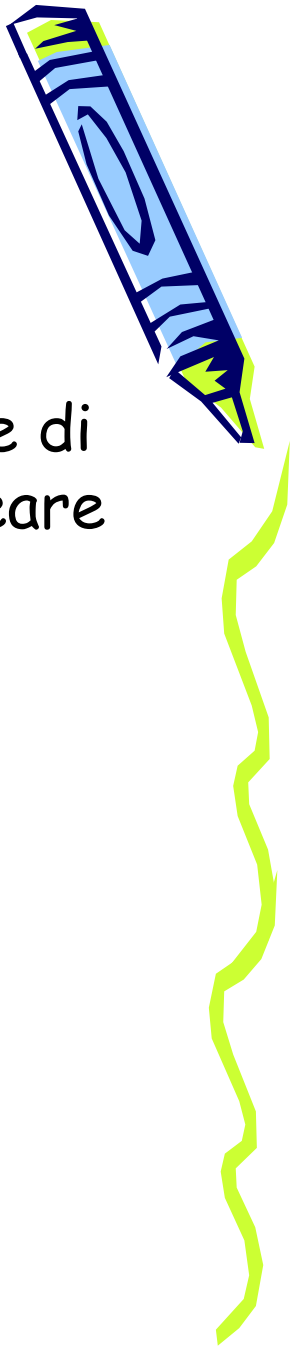
I valori *Ri* dovrebbero essere quanto più prossimi al valore *N/s*.

Infatti se la sequenza fosse perfettamente uniforme in ogni sottointervallo cadrebbero lo stesso numero di campioni della sequenza.

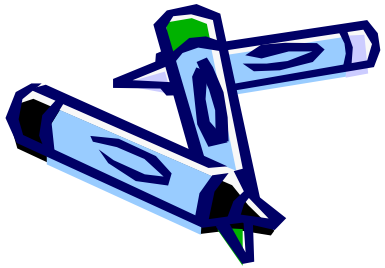
La variabile *V* per eseguire il test si calcola semplicemente:

$$V = \sum_{i=1}^{S} \frac{(R_i - N/s)^2}{N/s}$$

Esempio

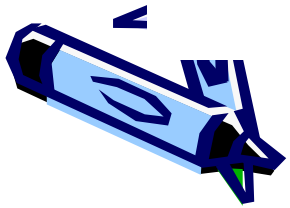


Generiamo una sequenza numerica pseudo-casuale di 100 valori mediante il generatore congruente lineare ($c = 0$, $a = 75$, $m = 2^{31} - 1$):

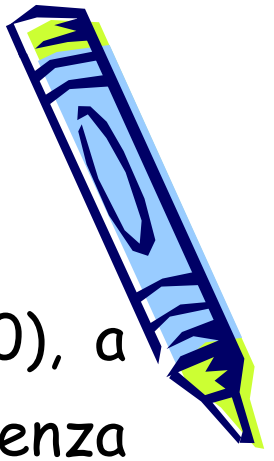


Esempio

0.0001	0.0071	0.0119	0.0452	0.0465	0.0632	0.077	0.0921
0.0927	0.0954	0.1032	0.1208	0.1327	0.1382	0.1539	0.16
0.1629	0.1629	0.1749	0.1896	0.1942	0.2266	0.2541	0.2693
0.2782	0.2823	0.2954	0.297	0.2975	0.3044	0.3097	0.3154
0.3163	0.3264	0.3277	0.3319	0.3457	0.3469	0.3608	0.3622
0.3641	0.3641	0.3777	0.3777	0.4098	0.4151	0.4598	0.4617
0.4704	0.4704	0.4746	0.5008	0.5102	0.5152	0.5336	0.5346
0.5392	0.5561	0.5641	0.5865	0.5926	0.5926	0.6056	0.6151
0.622	0.6245	0.6445	0.6534	0.6649	0.6649	0.6651	0.6651
0.6684	0.6825	0.6932	0.702	0.7115	0.7269	0.7271	0.7491
0.7491	0.7708	0.7773	0.7773	0.7886	0.793	0.8255	0.8342
0.835	0.8414	0.8471	0.8652	0.8677	0.8857	0.8898	0.9104
0.9111	0.9196	0.9766	0.9946				



Esempio



Suddividiamo l'intervallo [0,1] in 20 parti (s = 20), a questo punto contiamo quanti valori della sequenza cadono in ogni intervallino di ampiezza 0.05

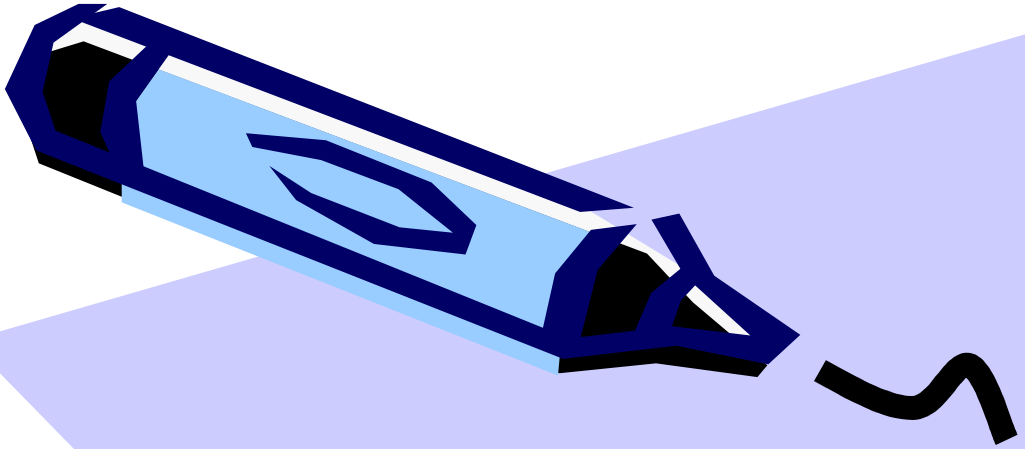
R =[5,3,4,8,1,8,9,4,2,6,6,4,5,7,5,5,4,7,5,2];

$$V = \sum_{i=1}^{S} \frac{(R_i - N/s)^2}{N/s} = \sum_{i=1}^{20} \frac{(R_i - 5)^2}{5} = 17{,}2$$

γ = 0.01, (χ^2 con 19 gradi di libertà)

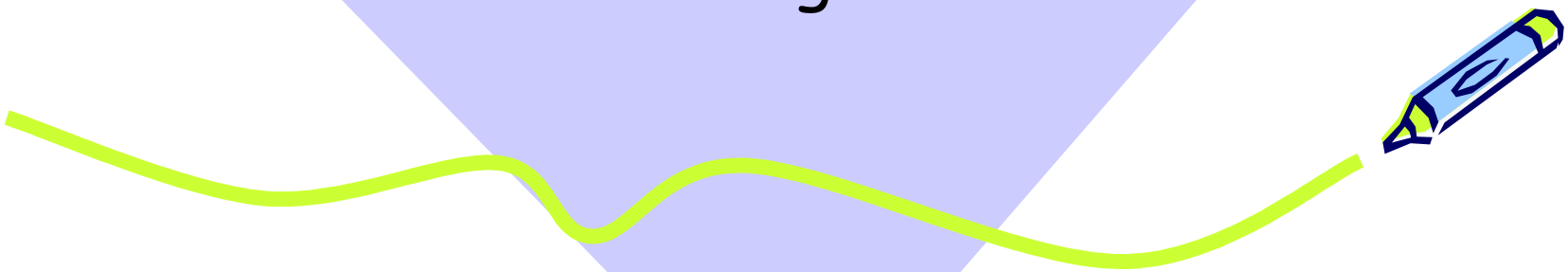
$x\gamma$ = 36,19.

la statistica del test *V* è minore di *possiamo accettare l'ipotesi di uniformità della sequenza generata.*

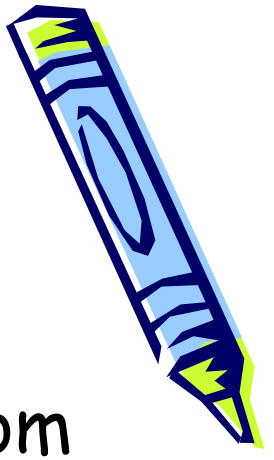


Generazione di numeri random

Distribuzioni generiche



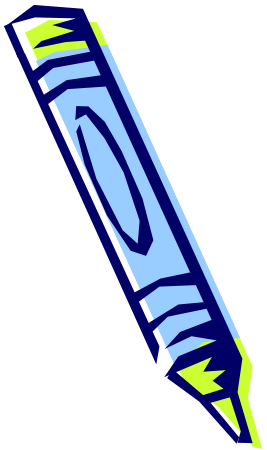
Generazione distribuzione generica



A partire da una sequenza di numeri random (U(0,1)) opportunamente generati, i metodi per la generazione di variabili aleatorie con distribuzione generica, sono:

- tecnica di **trasformazione inversa**
- metodo di **accettazione/rifiuto**
- metodo di **composizione**

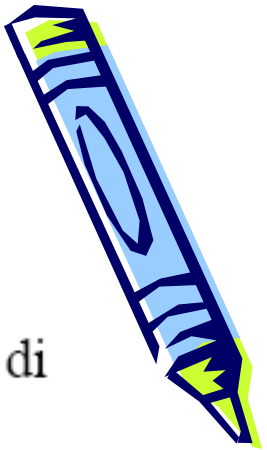
Generazione distribuzione generica



Una routine di generazione di variabili aleatorie deve:

- essere veloce
- avere un ciclo sufficientemente lungo
- non presentare larghi gap
- essere replicabile
- generare numeri con proprietà statistiche più vicine possibile a quelle ideali
- utilizzare poche volte la routine di generazione di numeri random

Trasformazione Inversa



Si vuole generare una variabile aleatoria X con funzione di densità di probabilità $f_X(x)$.

1) si calcola la funzione di distribuzione di probabilità o funzione cumulativa di probabilità

$$F_X(x) = \int_{-\infty}^{x} f_X(\tau) d\tau$$

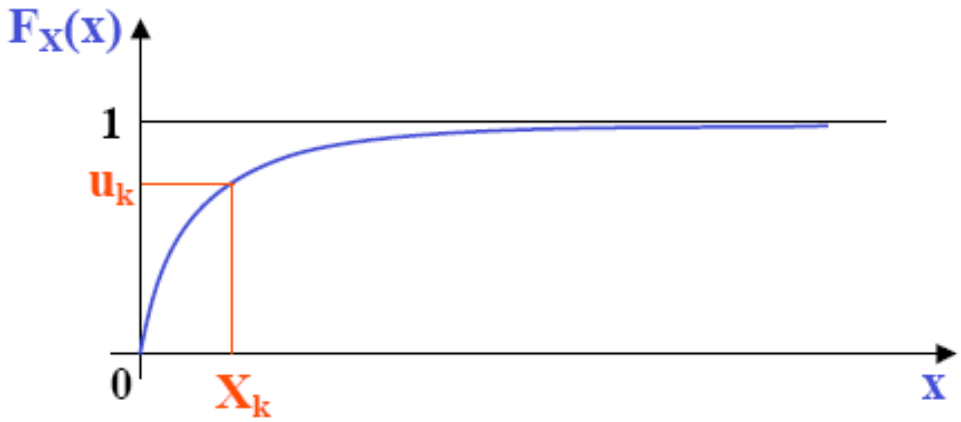
Tale funzione (qualora sia possibile calcolarla in forma chiusa) è continua, monotona crescente ed è sempre compresa tra 0 e 1 (per definizione $F_X(x) = P[X \leq x]$).

Trasformazione Inversa

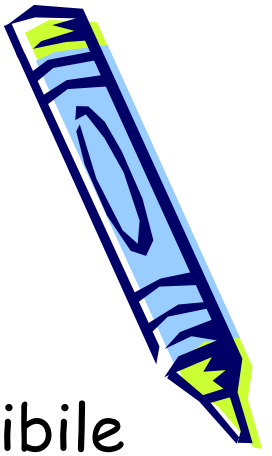
2) si pone u= $F_X(x)$ con u numero random (u ~ U(0,1))

3) si risolve X= $F_X^{-1}(u)$ e la variabile aleatoria X è distribuita secondo $f_X(x)$ (X ~ $f_X(x)$).

Graficamente:



Esempio

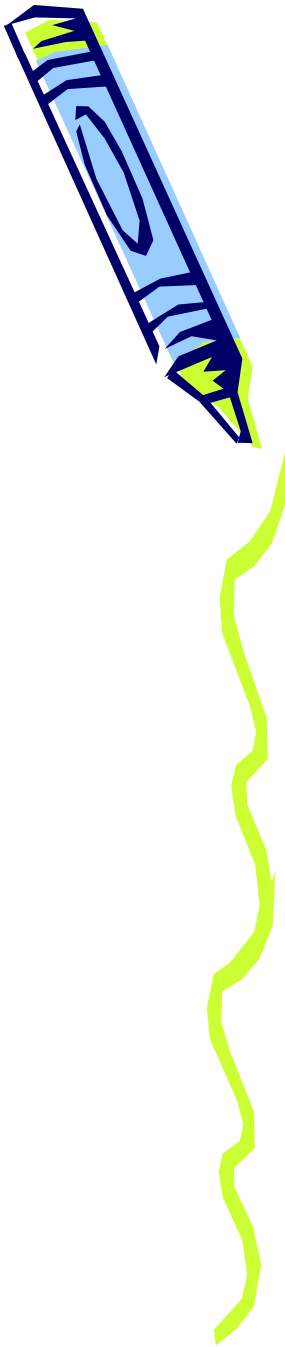


Per applicare quanto detto, vediamo come è possibile ottenere una v.a. esponenziale partendo da una v.a. uniforme U:

Supponiamo di voler costruire una successione di numeri pseudocasuali come osservazioni dalla distribuzione esponenziale ovvero con funzione di distribuzione

$$F(x) = 1 - e^{-\lambda x}$$

Esempio



Determiniamo F^{-1}:

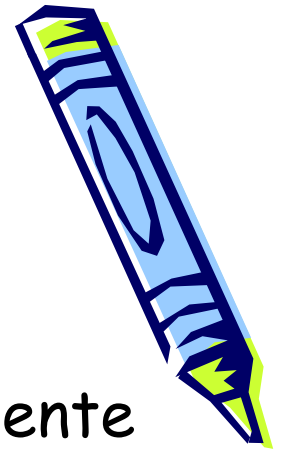
da $u = F(x) = 1 - e^{-\lambda x}$ si ricava $1 - u = e^{-\lambda x}$

$\ln (1 - u) = \ln (e^{-\lambda x})$ quindi $x = - \ln (1 - u)/\lambda$

ovvero

$$F^{-1}(u) = - \ln (1 - u)/\lambda$$

Esempio



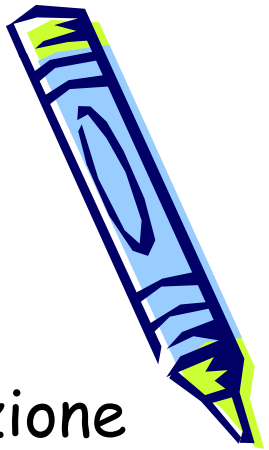
Quindi se U è una variabile aleatoria uniformemente distribuita in [0, 1),

$$X = F^{-1}(U) = - \ln (1 - u)/\lambda$$

è una variabile aleatoria con distribuzione esponenziale con media $1/\lambda$

Quindi, data una successione di numeri pseudocasuali con distribuzione uniforme in [0, 1), possiamo ottenere una successione di numeri pseudocasuali con distribuzione esponenziale.

Trasformazione Inversa

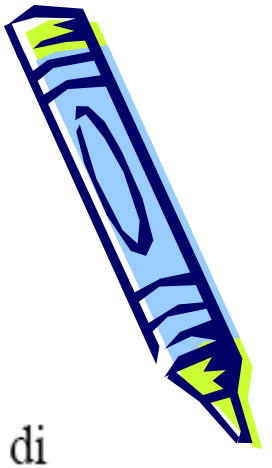


Se una variabile aleatoria U ha distribuzione uniforme in [0, 1), anche 1 – U ha distribuzione uniforme in [0, 1)

e quindi si può sostituire nell'argomento del logaritmo (1 – U) con U.

Tuttavia, questo cambiamento potrebbe indurre un cambiamento nella correlazione delle variabili X generate.

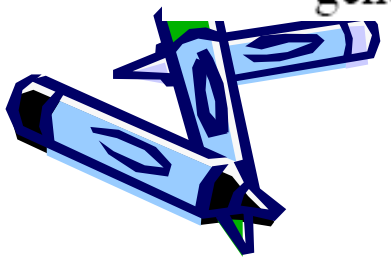
Trasformazione Inversa



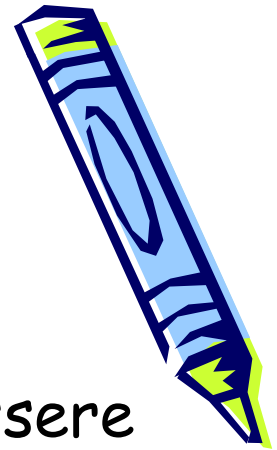
Osservazioni

La routine di generazione di una variabile aleatoria X con funzione di densità di probabilità esponenziale:

- chiama una sola volta, per ogni istanza di X, la routine (RAND) di generazione di numeri random;
- ha lo stesso ciclo di RAND;
- ha gap crescenti per X crescenti;
- è replicabile se lo è RAND;
- genererebbe numeri con proprietà statistiche ideali se RAND generasse numeri random ideali.



Trasformazione Inversa



Il metodo della trasformazione inversa può essere esteso ed utilizzato anche nel caso di distribuzioni discrete, ovvero quando si assume che la variabile X sia una variabile aleatoria discreta.

Supponiamo quindi che X assuma i valori x1, x2, . . . e supponiamo che essi siano ordinati, ovvero $x1 < x2 <\ldots$

Data una variabile U uniformemente distribuita in [0, 1) si definisce la variabile X nel seguente modo:

$$X(U) = \max \{xi \mid U \in [F(xi-1), F(xi)]\}$$

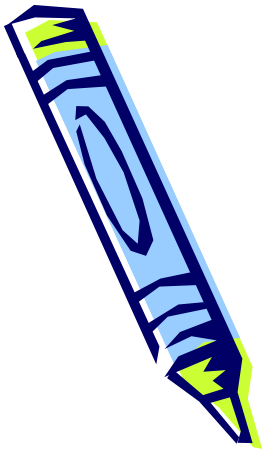
Trasformazione Inversa

ovvero si determina il più piccolo intero positivo $\underline{K}$ tale che $U \leq F(x_{\underline{k}})$ e si pone

$X = x_{\underline{k}}$.

Dobbiamo ora dimostrare che effettivamente la X così generata è quella desiderata, ovvero che risulta $P(X = xi) = p(x_i)$ per ogni i.

Trasformazione Inversa



Infatti si ha:

•per i = 1 risulta $X = x_1$ se e solo se $U \leq F(x1)$,
ma $F(x1) = p(x1)$ perché le xi sono ordinate.

Ora, poiché la U è uniformemente distribuita in [0, 1), si ha $P(X = x1) = P(U \leq F(x1)) = F(x1) = p(x1)$

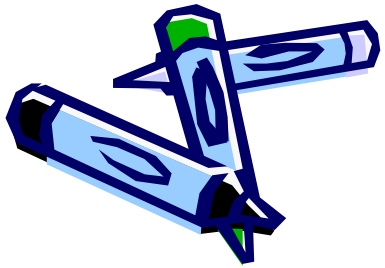
• per $i \geq 2$ risulta $X = xi$ se e solo se $F(xi-1) < U \leq F(xi)$ per come scelto i.

Inoltre, pochè la U è uniformemente distribuita in [0, 1) si ha

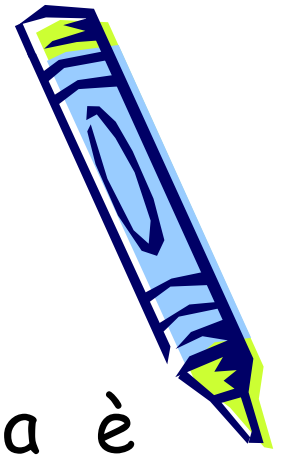
$$P(X = xi) = P(F(xi-1) < U \leq F(xi)) = F(xi) - F(xi-1) = p(xi)$$

Trasformazione Inversa

Il metodo della trasformazione inversa nel caso discreto ha una giustificazione molto intuitiva: si divide l'intervallo [0, 1) in sottointervalli contigui di ampiezza p(x1), p(x2), . . . e si assegna X a seconda del fatto che questi intervalli contengano la U che è stata generata



Metodo dell'accettazione -reiezione



Il metodo della trasformazione inversa è basato sul calcolo della trasformazione inversa F^{-1} che non sempre può essere calcolata o comunque non in maniera efficiente.

Per questa ragione sono stati sviluppati altri metodi fra i quali il metodo che esaminiamo in questo paragrafo detto "acceptance-rejection" o anche "metodo del rigetto".

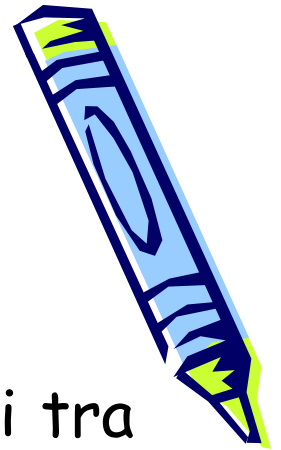
Metodo dell'accettazione -reiezione

Nel caso di leggi distribuzioni definite su intervalli finiti [*a,b*] si utilizza *il metodo della reiezione-accettazione*.

Supponiamo di conoscere la densità di probabilità della v.a. *X* che intendiamo generare: $f_X(x)$, definita su un intervallo finito [*a,b*], e l' immagine è definita sul codominio [0,*c*].

In pratica la funzione $f_X(x)$ è tutta contenuta all'interno del rettangolo [*a,b*]x[0,*c*].

Metodo dell'accettazione -reiezione



Generiamo due sequenze pseudo-casuali uniformi tra [0,1] : *U1* e *U2*.

Successivamente deriviamo, altre due sequenze numeriche uniformi secondo la seguente regola

$$\begin{cases} X = a + (b-a)U_1 \\ Y = cU_2 \end{cases}$$

ad ogni coppia di valori (*u1,u2*) corrisponderà una coppia (*x,y*) appartenente al rettangolo [*a,b*]x[0,*c*].

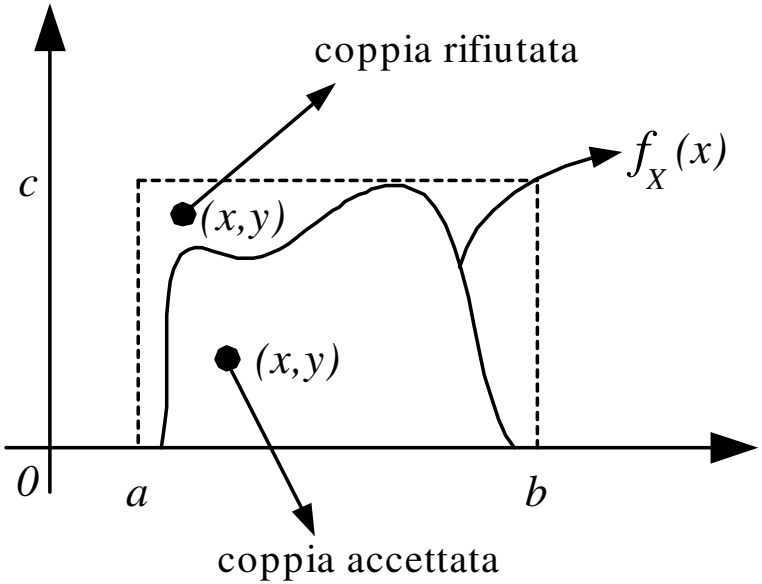
Metodo dell'accettazione -reiezione

Se la coppia (x,y) cade all'interno dell'area della funzione $f_X(x)$ viene accettata e sarà successivamente utilizzata per creare la sequenza pseudo-casuale desiderata, altrimenti viene scartata. In questo ultimo caso la procedura viene ripetuta fino a trovare una nuova coppia contenuta nell'area di $f_X(x)$.

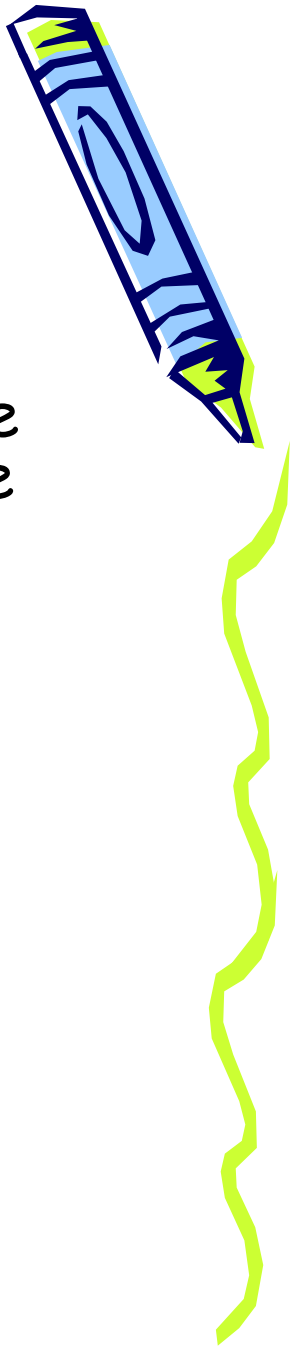
La sequenza di valori X così ottenuta è una sequenza pseudo-casuale che segue la legge di distribuzione $f_X(x)$, (infatti abbiamo scelto solo valori che cadono nella sua area).

Metodo dell'accettazione -reiezione

Questo metodo è molto efficiente quando l'area di $f_X(x)$, copre quasi tutto il rettangolo $[a,b]x[0,c]$, (in questo caso il numero di coppie scartato è molto esiguo).



Esercizio



Applichiamo il metodo dell'accettazione-reiezione per generare osservazioni casuali da una variabile aleatoria avente densità di probabilità

$f(x) = 20x(1-x)^3,\ 0 < x < 1$

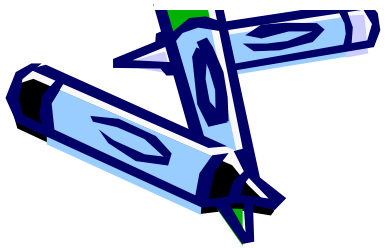
Metodo accettazione-reiezione

Si vuole generare una variabile aleatoria X con funzione di densità di probabilità $f_X(x)$ su un intervallo [a,b].

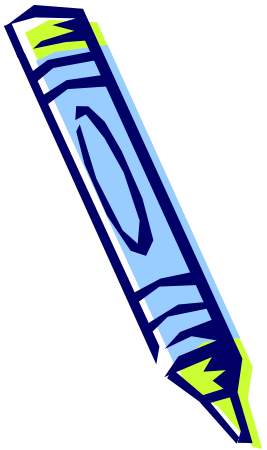
Algoritmo accettazione-rifiuto:

1) si genera un'istanza di una variabile R distribuita in modo uniforme nell'intervallo [a,b] (U(a,b))

2) si accetta tale valore con probabilità pari a $f_X(R)/ \max f_X(x)$, lo si rifiuta con probabilità $1- f_X(R)/\max f_X(x)$.



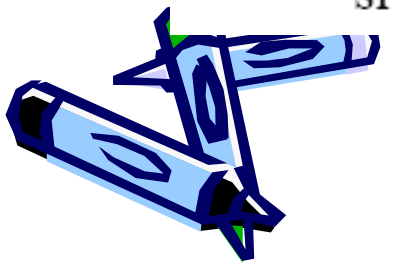
Metodo accettazione-reiezione



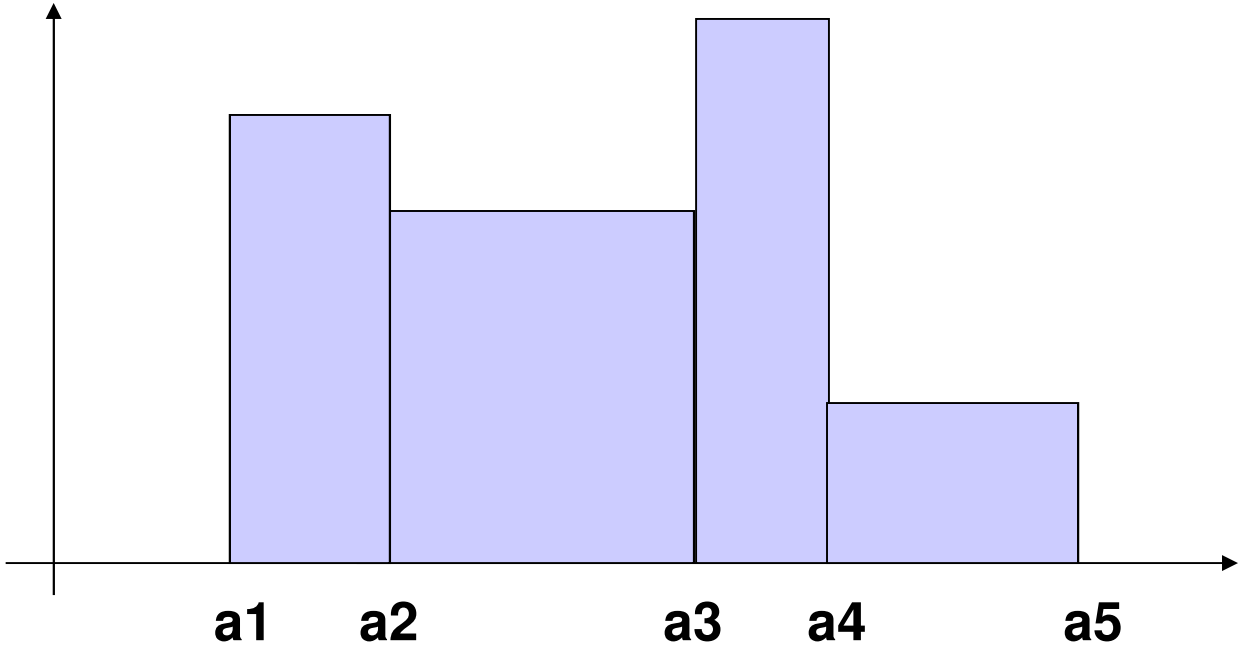
Osservazioni

La routine di generazione di una variabile aleatoria X con il metodo di accettazione-rifiuto:

- chiama almeno due volte, per ogni istanza di X, la routine RAND di generazione di numeri random; una volta per generare le variabile R, una volta per decidere se accettare o rifiutare il valore (esistono, però, metodi più specializzati e più efficienti);
- ha ciclo e gap che dipendono da RAND*RAND;
- è replicabile se lo è RAND;
- si utilizza solo in assenza di altri metodi.

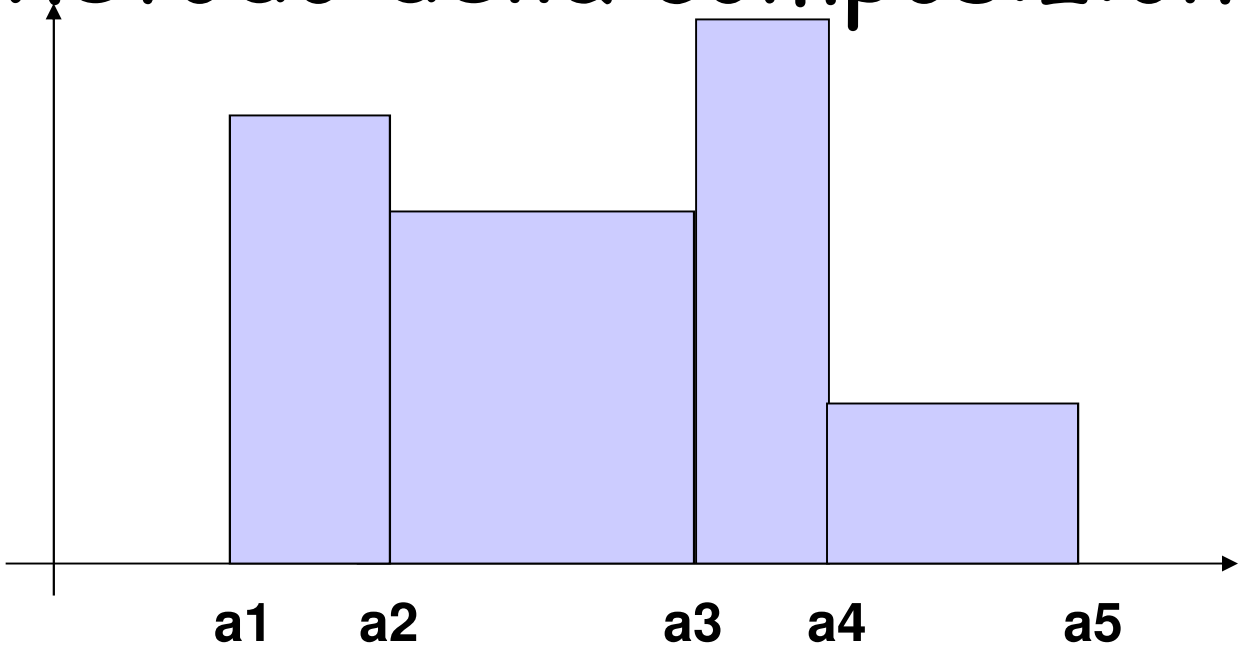


Metodo della composizione



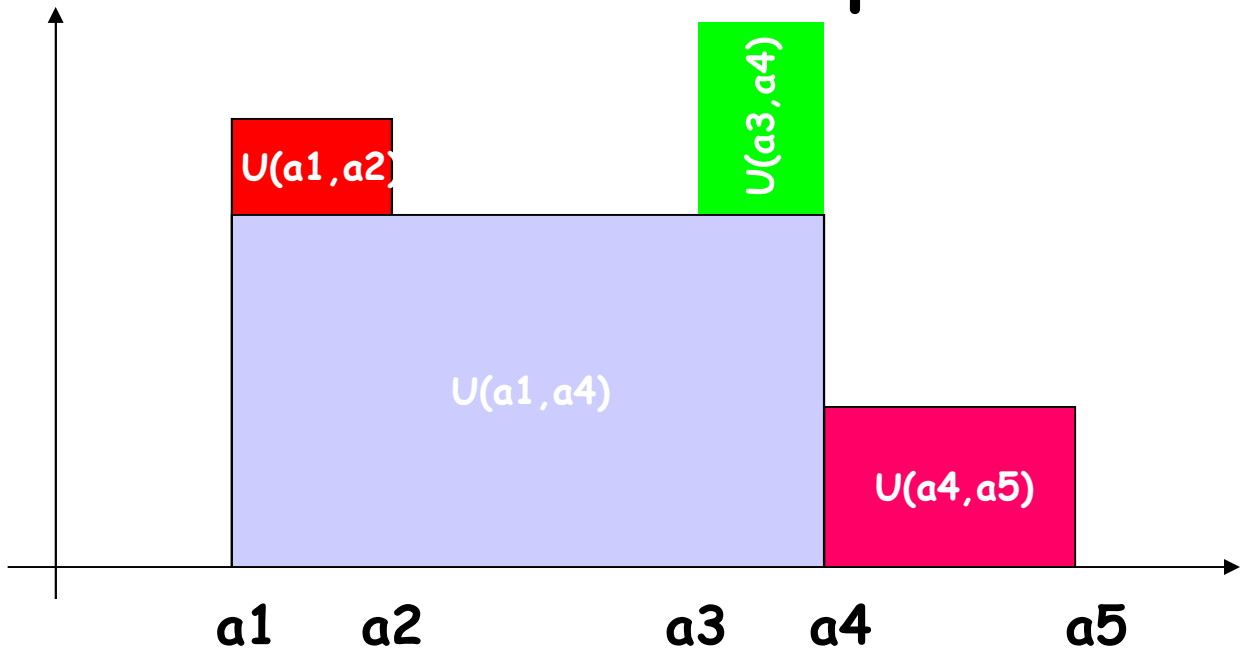
Questa distribuzione può essere vista come risultato della composizione di 4 uniformi: U(a1,a2), U(a2,a3), U(a3,a4), U(a4,a5).

Metodo della composizione

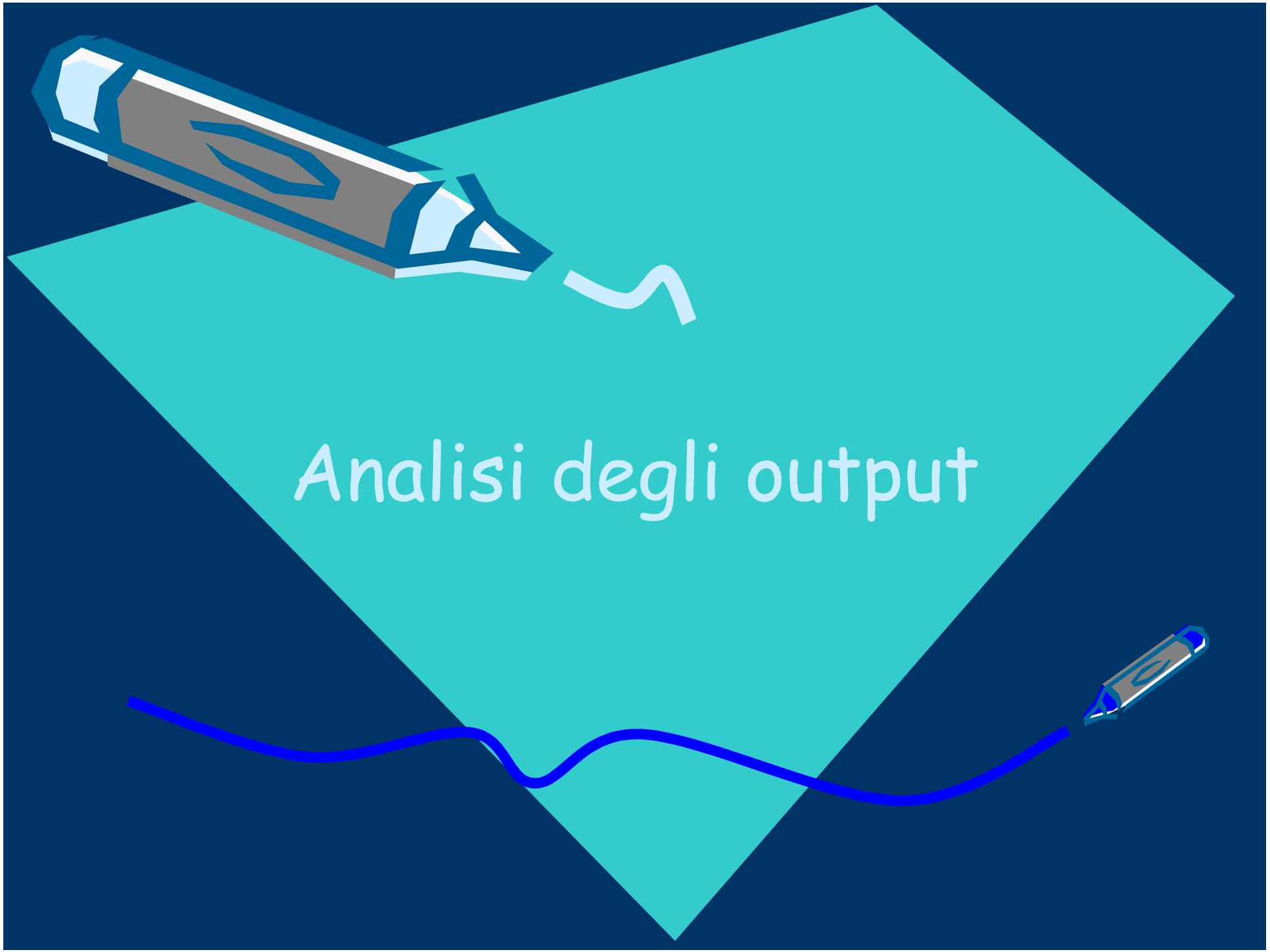


Siano p1, p2, p3 e p4 le aree dei quattro rettangoli: per generare un valore tratto da questa distribuzione scegliamo un rettangolo utilizzando le pi, poi generiamo un numero uniformemente distribuito in U(ai, ai+1)

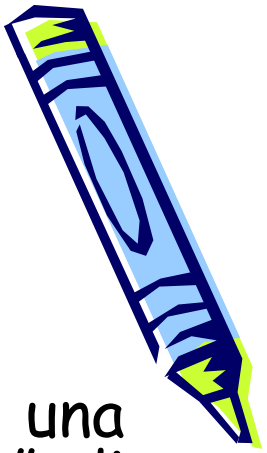
Metodo della composizione



Sono possibili anche scomposizioni alternative: in questo caso la distribuizione di prima viene scomposta in 4 uniformi,
U(a1,a2), U(a1,a4), U(a3,a4), U(a4,a5), ciascuna con probabilità uguale all'area del rettangolo corrispondente.


Analisi degli output

Osservazioni

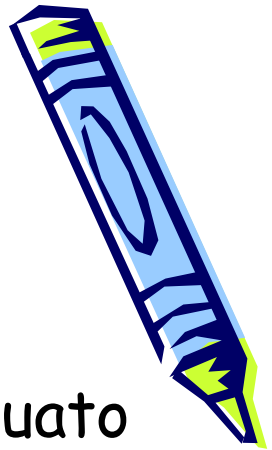


L'errore tipico che si commette nell'effettuare una simulazione è quello di eseguire un singolo "run" di lunghezza arbitraria e di prendere i risultati che si ottengono come stime delle caratteristiche del vero modello.

Poiché in una simulazione si utilizzano osservazioni casuali da una o più distribuzioni di probabilità, queste stime sono solamente realizzazioni particolari di variabili aleatorie che possono anche avere varianza molto grande.

Questo implica che, in un particolare run di una simulazione, queste stime possono anche differire di molto dalle corrispondenti reali caratteristiche del modello.

Osservazioni



Affinché i risultati di uno studio effettuato attraverso la simulazione abbiano senso è necessario l'uso di tecniche statistiche per progettare e analizzare gli esperimenti di una simulazione.

Nel fare ciò si manifesta subito il problema derivante dal fatto che i processi di output della simulazione sono, in generale, autocorrelati e non stazionari e questo rende inapplicabili direttamente le tecniche statistiche classiche che sono invece basate su osservazioni indipendenti, identicamente distribuite.

Osservazioni

Un altro problema esistente nell'ottenere stime precise delle caratteristiche del vero modello è dato dal tempo di calcolo necessario per raccogliere la quantità necessaria di dati di output per effettuare un'analisi statistica efficace, che potrebbe essere anche molto elevato.

Dati di output

Siano Y1, Y2, . . . i dati di output di un singolo run di una simulazione; ciascuna Yi può essere vista come una variabile aleatoria e quindi la collezione di variabili aleatorie {Yi, i = 1, 2, . . .} è un processo stocastico.

Ad esempio, in un sistema di code, le Yi possono rappresentare il tempo di attesa in coda dell'i-esimo utente.

In generale, le variabile aleatorie Yi non sono nè indipendenti, nè identicamente distribuite e quindi nell'analisi di questi dati non possono essere applicati direttamente i metodi di analisi statistica

Dati di output

Per ovviare a questo inconveniente, si effettuano più repliche della simulazione, ciascuna di lunghezza m, e si basa l'analisi sulle varie repliche.

Siano y11, y12, . . . y1m la realizzazione delle variabili aleatorie Y1, . . . , Ym ottenute con la prima replica. Nella seconda replica si avranno differenti realizzazioni delle variabili aleatorie Y1, . . . , Ym; siano esse y21, y22, . . . y2m.

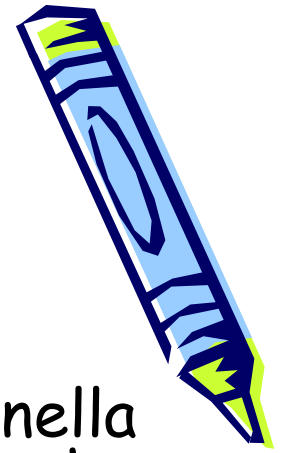
In generale, supponiamo di aver effettuato n repliche indipendenti di lunghezza m.

Dati di output

Le realizzazioni di una stessa replica non sono indipendenti, identicamente distribuite, ma se per ogni i = 1, . . . ,m consideriamo le osservazioni y1i, y2i, . . . yni, allora esse costituiscono osservazioni indipendenti, identicamente distribuite della variabile aleatoria Yi.

Quindi l'analisi statistica è applicabile alle osservazioni y1i, y2i, . . . yni, per ogni fissato i = 1, . . . , n.

Analisi dati di output

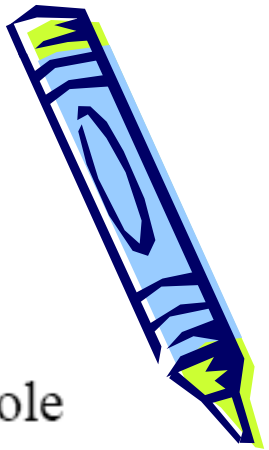


L'analisi di uscita della simulazione consiste nella determinazione (stima) delle grandezze che costituiscono gli obiettivi della simulazione.

La stima di tali grandezze si basa su tecniche di *stima di parametri*.

Le grandezze da stimare sono *parametri statistici di processi stocastici* che dipendono dalle condizioni iniziali della simulazione, dalle sequenze di numeri random utilizzati e dalla durata della simulazione.

Analisi dati di output



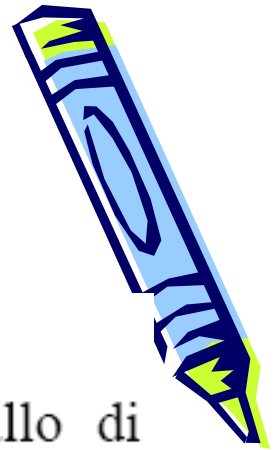
Sia θ una delle grandezze (parametri) da determinare.

➔ Date $X_1, X_2, \ldots, X_n$ osservazioni di tale grandezza, si vuole determinare una **stima di θ**.

La stima può essere:

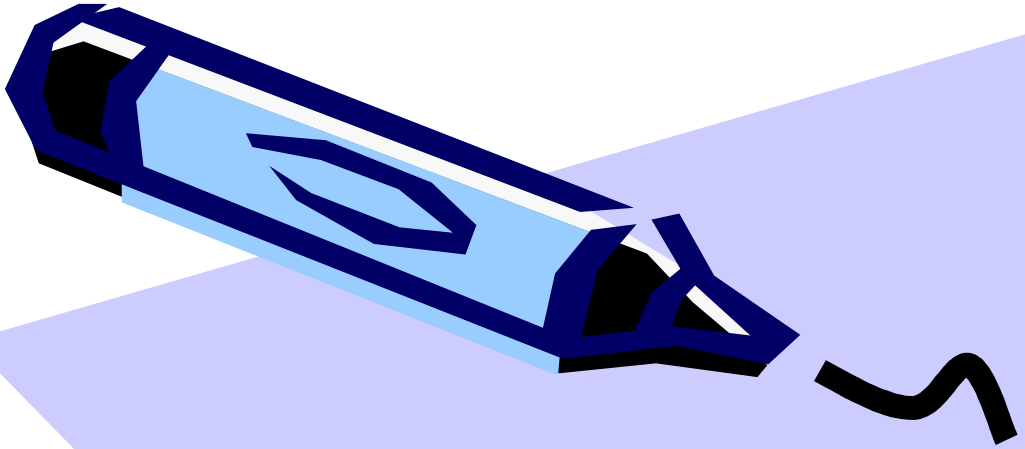
- ***stima per punti*** (o ***del valore*** o ***puntuale***): si desidera una stima del **valore "più plausibile"** di θ;
- ***stima ad intervallo*** (o ***intervallo di confidenza***): si desidera una stima dell'**intervallo di valori** in cui θ cade con una probabilità definita a priori.

Analisi dati di output



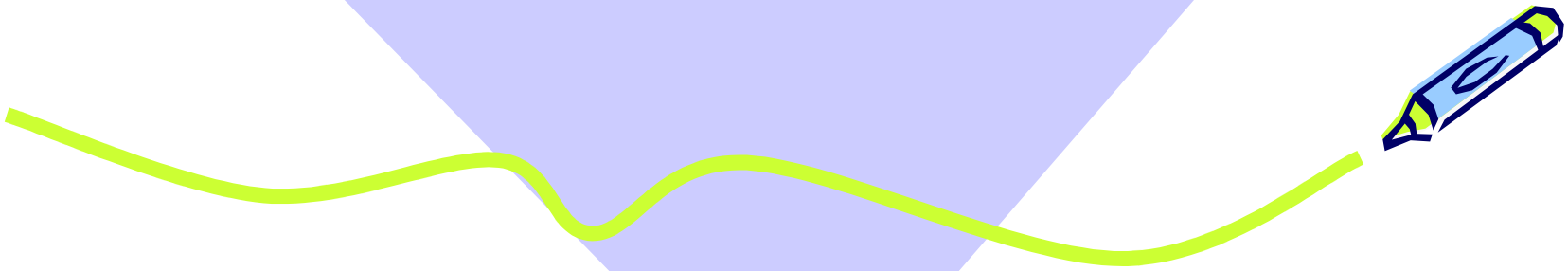
Osservazioni:

- ☹ se lo stimatore puntuale è deviato, si ottiene un intervallo di confidenza "centrato" su un valore errato
- ☹ se i campioni sono caratterizzati da una autocorrelazione positiva e si usa la varianza campionaria per stimare la varianza dello stimatore puntuale, si ottiene un intervallo di confidenza più stretto di quello reale
- 😐 se i campioni sono caratterizzati da una autocorrelazione negativa e si usa la varianza campionaria per stimare la varianza dello stimatore puntuale, si ottiene un intervallo di confidenza più largo di quello reale



Analisi output

Simulazione con e senza termine



Simulazione con e senza termine

Una **simulazione "con termine"** (***terminating***) ha durata associata ad un evento definito a priori (e.g. istante di tempo, verifica di una condizione sullo stato del sistema, ecc.)

Una **simulazione "senza termine"** (***non terminating***) ha durata associata al raggiungimento, da parte del sistema, di un comportamento di regime (durata non definibile a priori).

La simulazione con termine è utilizzata per analizzare ***comportamenti transitori*** del sistema, la simulazione senza termine è utilizzata per valutare le ***prestazioni a regime*** del sistema.

Simulazione con termine

L'analisi di uscita di simulazioni terminating si realizza con il **Metodo delle repliche indipendenti**.

Sia θ una delle grandezze da stimare.

- ad ogni run di simulazione si raccolgono M campioni che, opportunamente elaborati determinano una osservazione di θ
- si eseguono R repliche della simulazione, ottenendo R osservazioni di θ
- si esegue una **stima di θ** utilizzando la stima per punti e/o ad intervallo sulle R osservazioni.

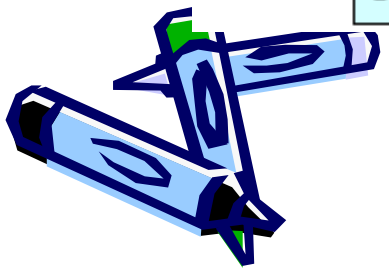
Simulazione con termine

In modo più specifico, siano:

- $X_{1,i}, X_{2,i}, \ldots, X_{M_i,i}$, i campioni raccolti all'i-esimo run di simulazione
- $L(\cdot)$ la funzione che caratterizza il calcolo della grandezza di interesse, ossia $L(X_{1,i}, X_{2,i}, \ldots, X_{M_i,i}) = L_i$ sia l'i-esima osservazione di θ.

Con il metodo delle repliche indipendenti, si raccolgono R campioni di θ, $L_1, L_2, \ldots, L_R$, sui quali si può applicare la stima per punti e/o ad intervallo.

N.B. il numero di campioni raccolti è differente ad ogni run di simulazione ed è esso stesso variabile stocastica



Simulazione con termine

Se la stima da calcolare è un valor medio, si calcola la **stima per punti** sulla base di $L_1, L_2, \ldots, L_R$, come

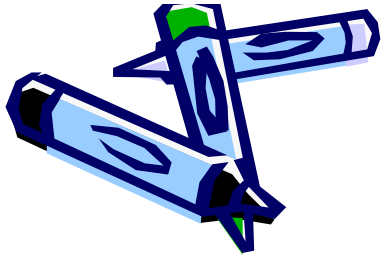
$$\hat{\theta}_R = \frac{1}{R}\sum_{i=1}^{R} L_i$$

e l'intervallo di confidenza come

$$\hat{\theta}_R - t_{\alpha/2,R-1}\sqrt{S^2/R} \le \theta \le \hat{\theta}_R + t_{\alpha/2,R-1}\sqrt{S^2/R}$$

con

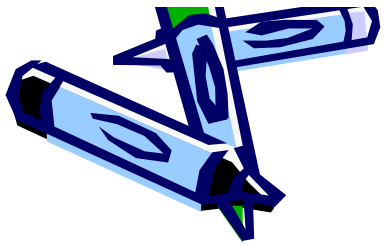
$$S_R^2 = \frac{1}{R-1}\sum_{i=1}^{R}(L_i - \hat{\theta}_R)^2$$



Simulazione con termine

Osservazioni:

- si può ottenere una stima non deviata se $L_1, L_2, \ldots, L_R$ sono **i.i.d.**, ossia se le repliche della simulazione sono eseguite:
  - variando i semi di inizializzazione dei generatori di variabili stocastiche;
  - con condizioni iniziali ottenute casualmente secondo distribuzioni opportune (lo stato iniziale "nullo" tipicamente non si presenta in realtà);
- l'intervallo di confidenza è **funzione del numero R di run eseguiti** e consente di determinare proprio un numero opportuno di repliche (si eseguono 4-5 repliche, si esegue la stima ad intervallo e si verifica come varierebbe l'intervallo aumentando il numero di repliche).



Simulazione senza termine

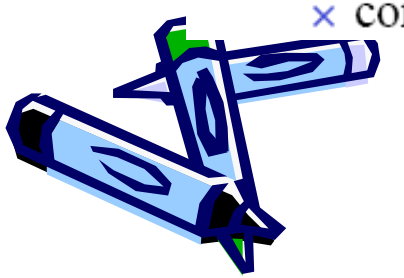
Con simulazioni senza termine si vuole determinare θ come **valore medio di regime** di una particolare grandezza di cui si raccolgono campioni X_i, ossia

$$\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i$$

(**n.b.** il regime esiste se il processo è stazionario).

Problemi:

- come si capisce di avere raggiunto una situazione di regime per una particolare grandezza? (→determinazione della ***durata del warm-up*** della simulazione)
- × come si raccolgono campioni i.i.d.?



Simulazione senza termine

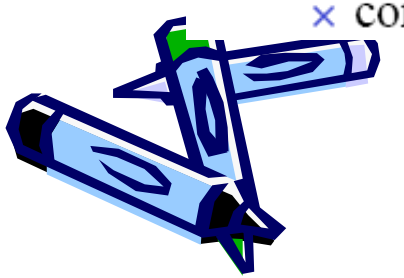
Con simulazioni senza termine si vuole determinare θ come **valore medio di regime** di una particolare grandezza di cui si raccolgono campioni X_i, ossia

$$\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i$$

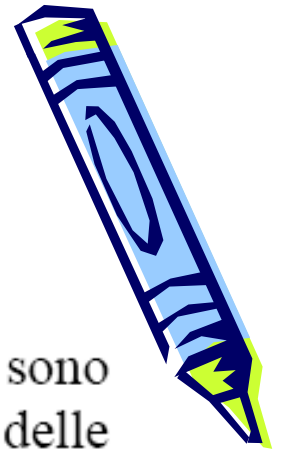
(**n.b.** il regime esiste se il processo è stazionario).

Problemi:

- come si capisce di avere raggiunto una situazione di regime per una particolare grandezza? (→determinazione della ***durata del warm-up*** della simulazione)
- × come si raccolgono campioni i.i.d.?



Simulazione senza termine



Le stime ottenute con la simulazione senza termine sono necessariamente di durata finita e sono deviate in funzione delle condizioni iniziali.

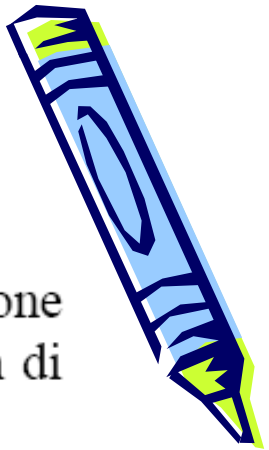
Per ridurre tale deviazione:

- si scelgono condizioni iniziali reali (analisi approfondita e costosa del sistema)
- si scelgono condizioni iniziali realistiche ottenibili da modelli matematici semplificativi del sistema reale
- si parte da uno stato iniziale arbitrario e non si raccolgono dati per una fase transitoria (***warm-up***)

Determinazione warm-up

Per determinare il punto di cancellazione si possono utilizzare **metodi statistici** molto complessi e, quindi, poco diffusi o **metodi empirici**.

Determinazione warm-up



I **metodi empirici** si basano sulla esecuzione di **R** run di simulazione (circa 10) e sulla suddivisione della fase di raccolta dati in **n** batch di uguale durata.

Siano:

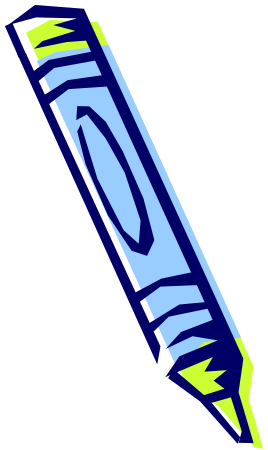
Y_{ri}= media sul batch i-mo della r-ma replicazione;

$$\overline{Y}_i = \frac{1}{R}\sum_{j=1}^{R} Y_{ji} = \text{media dei batch i - mi}$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} \overline{Y}_i = \text{media cumulata}$$

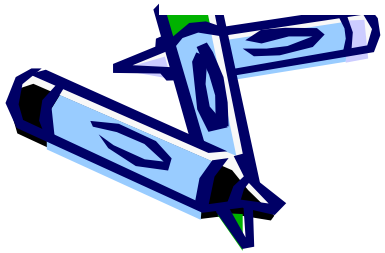
$$\overline{Y}(n,d) = \frac{1}{n-d}\sum_{i=d+1}^{n} \overline{Y}_i = \text{media cumulata con d cancellazioni}$$

Determinazione warm-up

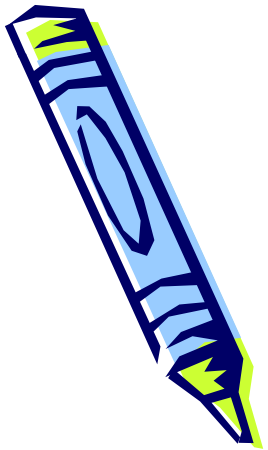


I **metodi empirici** si basano sui seguenti passi:

- esecuzione di R run di simulazione
- suddivisione della fase di raccolta dati in n batch di uguale durata
- per ogni gruppo di batch corrispondenti, calcolo della media dei batch
- calcolo della media cumulata eliminando progressivamente un batch alla volta a partire da quello iniziale
- se la media cumulata non risente "significativamente" dell'ultimo batch eliminato, definizione del punto di cancellazione come istante iniziale di tale batch

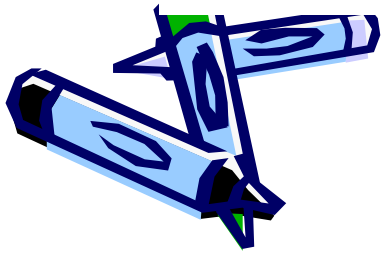


Determinazione warm-up

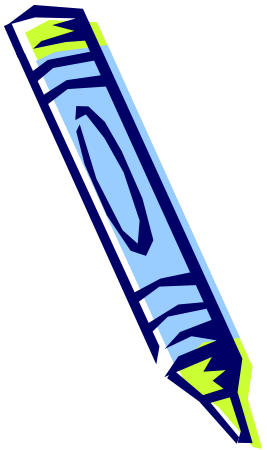


I **metodi empirici** si basano sui seguenti passi:

- esecuzione di R run di simulazione
- suddivisione della fase di raccolta dati in n batch di uguale durata
- per ogni gruppo di batch corrispondenti, calcolo della media dei batch
- calcolo della media cumulata eliminando progressivamente un batch alla volta a partire da quello iniziale
- se la media cumulata non risente "significativamente" dell'ultimo batch eliminato, definizione del punto di cancellazione come istante iniziale di tale batch



Determinazione warm-up

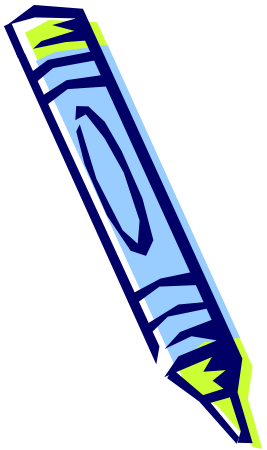


Il fatto che le variazioni non siano più significative si valuta:

- ad occhio
- verificando che le variazioni siano inferiori all'1%-2%
- calcolando l'intervallo di confidenza sulle medie dei batch corrispondenti
- calcolando l'intervallo di confidenza sulle medie cumulate
- con metodi statistici ad hoc

N.B.
I primi due metodi sono poco solidi statisticamente, l'ultimo metodo è molto complesso (quindi, poco diffuso)

Determinazione warn-up

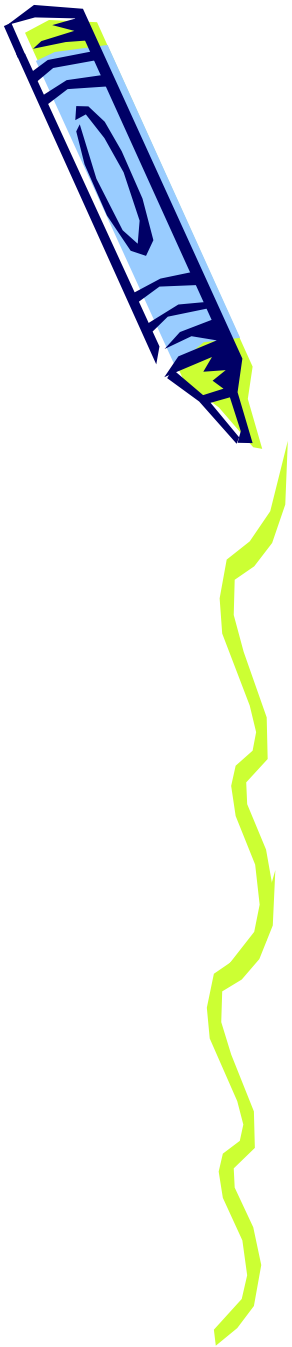


Osservazioni:

- La deviazione dovuta alle condizioni iniziali è affetta dal punto di cancellazione non dal numero di repliche

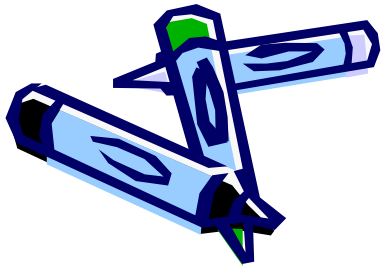
- Il punto di cancellazione dipende dal parametro da stimare e deve, quindi, essere ricalcolato per ogni parametro

Simulazione senza termine



Analisi degli output:

- Metodo delle **repliche con cancellazione**
- Metodo delle **medie batch**

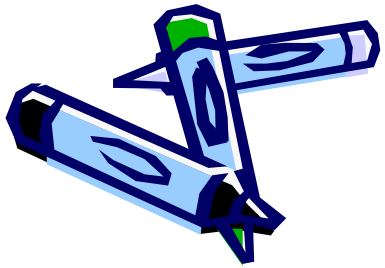


Repliche con cancellazione

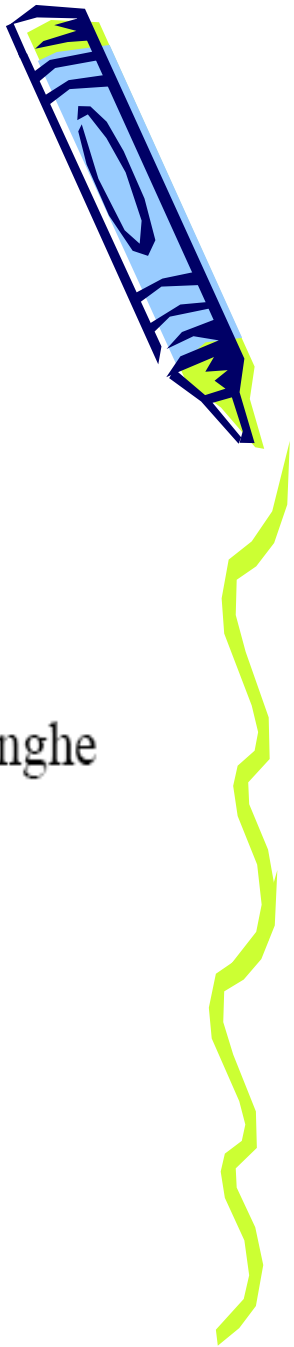
Il metodo delle repliche con cancellazione è analogo al metodo delle repliche indipendenti per simulazioni con termine.

Quindi, si eseguono numerosi run di simulazione e, per ognuno di essi:

- si calcola il punto di cancellazione;
- si calcola la media dei campioni della variabile di interesse utilizzando solo campioni raccolti dopo il punto di cancellazione;(tale media costituisce un'osservazione del parametro da stimare);
- si esegue una stima per punti e/o ad intervallo della grandezza di interesse utilizzando le medie calcolate al punto precedente.



Repliche con cancellazione



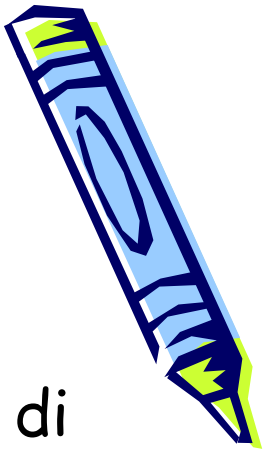
Vantaggi:

- ☺ i campioni utilizzati per la stima sono i.i.d.

Svantaggi:

- ☹ si devono eseguire molti run di simulazioni tipicamente molto lunghe (rispetto alle simulazioni con termine)
- ☹ si deve stimare il transitorio per ogni replica della simulazione

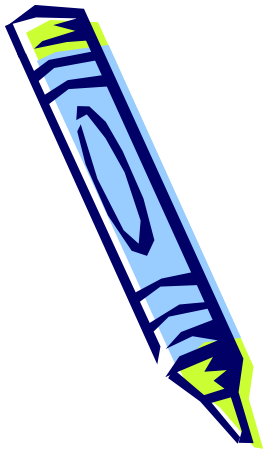
Metodo delle medie batch



Il metodo delle medie batch utilizza un unico run di simulazione per il quale:

- si calcola il punto di cancellazione;
- si divide la fase di raccolta dati in batch di uguale durata;
- si calcolano le medie sui batch (i valori Yri con r=1); ogni media costituisce un'osservazione del parametro da stimare;
- si esegue una stima per punti e/o ad intervallo della grandezza di interesse utilizzando le medie calcolate al punto precedente.

Metodo delle medie batch



Vantaggi:

- ☺ si esegue un unico run di simulazione
- ☺ si stima il transitorio una volta sola

Svantaggi:

- ☹ i campioni della grandezza di interesse calcolati sui batch rischiano di non essere i.i.d., ad esempio è facile che siano autocorrelati (l'autocorrelazione diminuisce considerando batch lunghi)