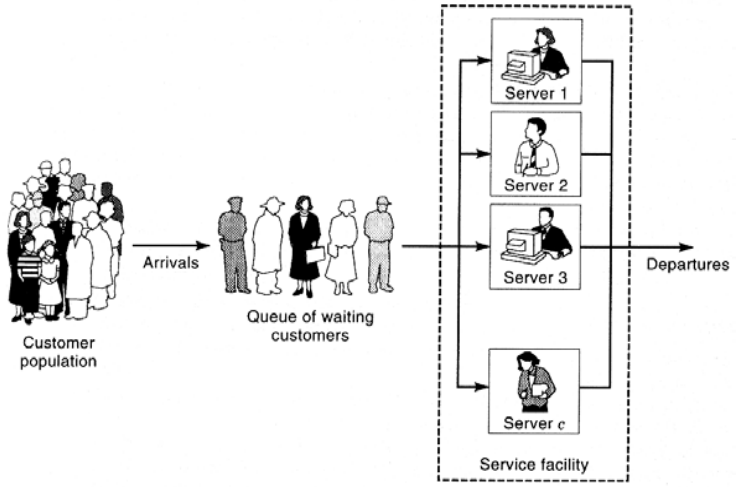
Elementi di *Queuing Theory*

Applicazione della teoria delle probabilità a sistemi in cui "**clienti**" si candidano per ricevere un "**servizio**"

Un sistema tipico può essere definito come uno in cui i clienti

- arrivano per richiedere un servizio
- aspettano di essere serviti
- partono dopo essere serviti



Cenno storico

Lavoro pionieristico di K. Erlang tra il 1909 e il 1920 per analizzare il traffico telefonico

A. A. Markov (1856-1922) formalizza la teoria delle catene markoviane

Molti sviluppi matematici tra il 1950 e il 1970

Esplosione di interesse pratico a partire dagli anni 70 per

- ricerca operazionale
- analisi dei modelli di migrazione di popolazioni in economia ed in biologia
- modelli di flusso in elettricità ed in dinamica dei fluidi
- sistemi di produzione, catene di montaggio
- controllo di inventario
- servizi (ospedali, aeroporti, traffico su strada, banche etc.)
- reti di comunicazione voce e dati, tanto a commutazione di pacchetti quanto a commutazione di circuiti
- calcolatori e sistemi di calcolatori, tanto in *time sharing* che in *processor sharing*

. ..e in fisica delle alte energie?

Varietà di sistemi e parametri

Sequenza di arrivo

I clienti arrivano ai tempi t_1, t_2, t_3 ... ordinati in modo crescente. Si defisce come intervallo tra due arrivi $y_i = t_{i+1} - t_i$.

I tempi di arrivo sono assunti di solito come generati da un processo stocastico e gli intervalli y_i sono indipendenti e distribuiti secondo la distribuzione di probabilità $A(\tau)=P(y_i \leq \tau)$.

I clienti possono arrivare individualmente o in gruppi di numero fisso o aleatorio.

Popolazione dei clienti e loro comportamento

La popolazione o sorgente di clienti può essere di dimensioni finite o infinite. Nel caso di dimensioni finite, il numero di clienti già nel sistema influenza la sequenza di arrivo.

Il comportamento dei clienti è pure importante: un cliente può rinunciare a mettersi in coda o abbandonare la coda dopo un certo tempo o ancora saltare da una coda ad un'altra.

Meccanismo di servizio

Il servizio è descritto dal numero di *servers* e dalla definizione dei tempi di servizio. Si assume normalmente che i tempi di servizio per clienti successivi x_1 , x_2 , x_3 ... sono indipendenti e distribuiti secondo la distribuzione di probabilità $B(\tau)=P(x_i \leq \tau)$.
I clienti possono essere serviti individualmente o in gruppi di numero fisso o aleatorio.

Numero massimo di clienti ammessi

Può essere infinito o limitato (anche a zero in coda, se non c'è sala d'aspetto).

Numero delle unità di sevizio *(servers)*

Uno, alcuni o moltissimi (=> ∞)

Disciplina della coda

›› First-Come First Served (FCFS) ovvero First-In First Out (FIFO)

›› Last-Come First Served (LCFS) ovvero Last-In First Out (LIFO)

›› Service In Random Order (SIRO) ovvero Random Selection for Service (RSS). Il prossimo cliente da servire è scelto secondo una distribuzione di probabilità discreta, di solito uniforme

›› Priority (PR o PRI)
La popolazione é divisa in due o più classi:
• Se la disciplina è *non-preemptive*, quando un cliente di alta priorità arriva è messo in testa alla coda.
• Se la disciplina è *preemptive*, quando un cliente di alta priorità arriva si sostituisce immediatamente a un cliente che riceve servizio: quest'ultimo è rimesso in coda e, quando sarà il suo nuovo turno può riprendere il servizio al punto in cui era stato interrotto *(preemptive resume)* o ricominciarlo daccapo *(preemptive repeat)*.

Notazione di Kendall

Espressione sintetica del tipo di coda: ***A/B/c/K/m/Z***

A tipo di distribuzione dei tempi di arrivo
B tipo di distribuzione dei tempi di servizio
c numero di *servers* (in parallelo)
K numero massimo di clienti nel sistema (in coda più in servizio)
m dimensione della popolazione dei clienti
Z descrizione della disciplina della coda

I simboli usuali per la prima e seconda posizione, tempi di arrivo e tempi di servizio, sono:

M distribuiti esponenzialmente
D deterministici, per esempio costanti
E_k secondo Erlang, a livello *k*
G generale

Le quantità in terza, quarta e quinta posizione sono un numero intero positivo.

La sesta posizione determina una delle discipline *FCRS, FIFO, LCFS, LIFO, SIRO, RSS, PR* and *PRI* o ancora *GD* per una disciplina generale.

Per esempio, un sistema *M/M/2/50/* ∞ */SIRO* ha intervalli di arrivo e tempi di servizio distribuiti esponenzialmente, 2 *servers,* una capacità massima di 50 (48 in coda e 2 in servizio), una popolazione di clienti molto grande e disciplina di servizio aleatoria.

Quando gli ultimi tre numeri non sono indicati si sottintende
$K = \infty$, $m = \infty$, $Z = FCFS$

Nota: per quanto la notazione di Kendall sia molto popolare non descrive tutti i casi possibili, per esempio arrivi in gruppo.

Processi stocastici

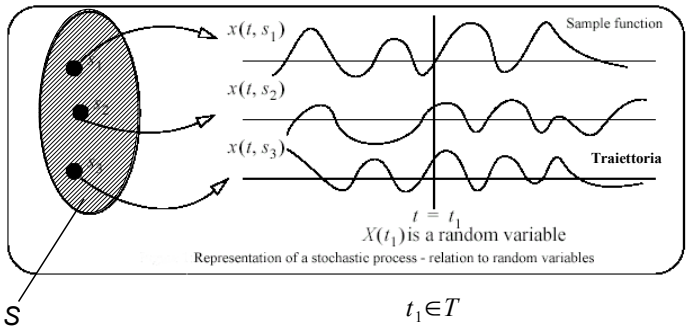
Due variabili aleatorie (o casuali) possono essere indipendenti o correlate.

Esistono situazioni in cui la dipendenza funzionale tra sistemi di variabili è più intricata di una semplice correlazione: questa dipendenza trova una descrizione nei processi stocastici.

Un processo stocastico è un insieme (o una famiglia) di variabili aleatorie che hanno in comune lo spazio di probabilità S nel quale sono definite.

In *queuing theory* il tempo è usato come parametro per identificare le variabili della famiglia: *X(t)* è il valore o lo **stato del sistema** al tempo *t*, ove *t* appartiene all'insieme T, un insieme discreto di valori, o continuo.

I valori che la variabile aleatoria *X(t)* può assumere nello spazio di probabilità *S* sono determinati statisticamente dalla sua funzione di distribuzione (o dalla densità di probabilità).



In un esperimento governato da una variabile aleatoria, ad ogni evento *s* appartenente ad *S* corrisponde un valore per la variabile.
In un processo stocastico ad ogni evento *s* appartenente ad *S* corrisponde una traiettoria o realizzazione (*sample function*) $x(t, s)$, funzione del tempo, per $X(t)$.

T ed S discreti: il processo è una sequenza aleatoria discreta.
Es.: X_n rappresenta l'*n*-esimo risultato ottenuto gettando un dado
$T = \{1,2,3.... n\}$ e $S = \{1,2,3,4,5,6\}$

T discreto, S continuo: il processo è una sequenza aleatoria continua.
Es.: X_n rappresenta la temperatura misurata in un punto ad ogni ora di un giorno $T = \{1,2,3.....24\}$; la temperatura può assumere ogni valore in un intervallo.

T continuo, S discreto: il processo è un processo aleatorio discreto.
Es.: $X(t)$ rappresenta il numero di telefonate ricevute tra 0 e t;
$S = \{0,1,2,3...\}$

T ed S continui: il processo è un processo aleatorio continuo.
Es.: $X(t)$ rappresenta la temperatura massima in un punto nell'intervallo da 0 a *t*.

Un **processo di Markov** è un processo stocastico di grande generalità che dipende in modo limitato dalla storia del sistema. Precisamente, lo stato presente $X(t_0)$, al tempo t_0, contiene tutta l'informazione relativa al passato necessaria a determinare le proprietà future del processo: si dice che **non ha memoria** (*memoryless*).
Un processo di Markov può essere discreto in *T* o in *S* (catena markoviana) o continuo: c'è confusione nella terminologia!

Esempio: **media aritmetica di variabili aleatorie indipendenti**

La media statistica X_n di *n* variabili aleatorie indipendenti Y_i aventi la stessa distribuzione è

$$X_n = \frac{Y_1 + Y_2 + ... + Y_n}{n} = \frac{n-1}{n} \frac{Y_1 + Y_2 + ... + Y_{n-1}}{n-1} + \frac{Y_n}{n}$$

$$X_n = \frac{1}{n}[(n-1) X_{n-1} + Y_n] \qquad n = 1, 2, ...$$

Processi stocastici matematicamente trattabili (attraenti) sono i **processi di Poisson** che si basano sulla distribuzione dello stesso nome.

Supponiamo che gli eventi si verifichino in modo aleatorio ad un tasso medio di λ eventi al secondo. Sia *N(t)* il numero di eventi nell'intervallo da 0 a *t*. *N(t)* è non decrescente, a incrementi interi e continua in *t*: *N(t)* rappresenta un processo stocastico di Poisson se la sua distribuzione di probabilità è

$$P(N(t)=k)=\frac{(\lambda t)^k}{k}!\,e^{-\lambda} \quad \text{per } k = 0, 1, 2, 3 \ ....$$

ossia una distribuzione di Poisson con valor medio λt , con $\lambda = costante$

Variabili del comportamento della coda

Analisi in regime stazionario ovvero in regime transitorio:
anche in regime stazionario i parametri sono delle variabili aleatorie, non delle costanti, e sono determinate dalla loro distribuzione di probabilità

Lunghezza della coda: numero di clienti in attesa di essere serviti
Numero di clienti nel sistema

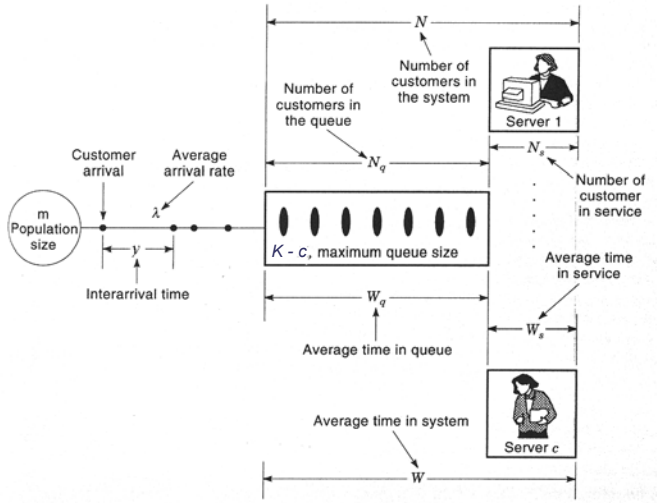
Tempo di attesa: tempo in sala d'aspetto
Tempo totale nel sistema

Jitter del tempo di attesa o nel sistema

Frazione di tempo in cui il servizio non è disponibile *(busy)*

Definizione delle variabili usuali

y	variabile aleatoria che rappresenta gli intervalli tra arrivi di clienti successivi
λ	frequenza media di arrivo dei clienti;
$1/\lambda$	intervallo medio tra due arrivi
x	variabile aleatoria che rappresenta i tempi di servizio
$1/\mu$	tempo medio di servizio;
μ	frequenza media di servizio
$\rho = \lambda/\mu$	intensità di traffico o di carico
$N(t) \Rightarrow N$	variabili aleatorie che rappresentano il numero di clienti nel sistema al tempo *t* e a regime
$N_q(t) \Rightarrow N_q$	variabili aleatorie che rappresentano il numero di clienti in sala d'aspetto al tempo *t* e a regime
$N_c(t) \Rightarrow N_c$	variabili aleatorie che rappresentano il numero di clienti in servizio al tempo *t* e a regime
L	valor medio del numero di clienti nel sistema a regime
L_q	valor medio del numero di clienti in sala d'aspetto a regime
L_c	valor medio del numero di clienti in servizio a regime
$w(x)$	variabile aleatoria che rappresenta la distribuzione di probabilità del tempo passato nel sistema *(waiting time)* a regime
W	valor medio del tempo passato dal cliente nel sistema
W_q	valor medio del tempo passato dal cliente in sala d'aspetto
W_c	valor medio del tempo passato dal cliente in servizio



Relazioni ovvie
Per la lunghezza della coda

$$N(t) = N_q(t) + N_c(t)$$
$$N = N_q + N_c$$

dalla seconda, prendendo il valor medio

$$L = E(N) = E(N_q) + E(N_c) = L_q + L_c$$

Analogamente per i tempi di attesa

$$W = W_q + W_c$$

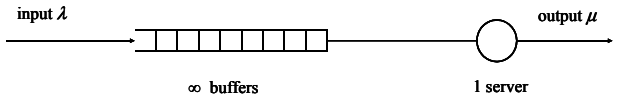
Formula di Little
Relazione molto generale, valida per un a vasta classe di distribuzioni di probabilità sia dei tempi di arrivo che dei tempi di servizio, tra il tempo medio passato nel sistema e il numero medio di clienti nel sistema

$$L = \lambda W$$

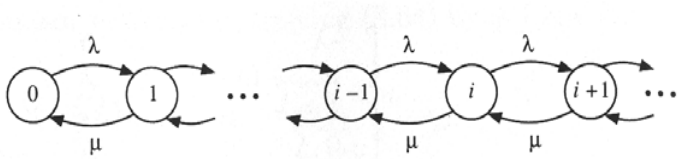
Si può applicare anche separatamente alla sala (alle sale) d'aspetto e al *server* (ai *servers*)

Sistema *M/M/*1

Si tratta di un sistema a un solo *server*, con intervalli di arrivo e tempi di servizio distribuiti esponenzialmente. Tanto la popolazione dei clienti quanto la dimensione della sala d'aspetto sono infinite e la disciplina è *FCFS*



La probabilità che il sistema contenga *i* clienti si calcola analiticamente usando il formalismo basato sui diagrammi di transizione di probabilità per i processi di nascita e morte (*birth and death)*:



I cerchi rappresentano stati ciascuno con un numero *i* di clienti nel sistema; le frecce rappresentano il flusso di probabilità entrante o uscente da uno stato.
Lo stato con zero clienti può rimanere tale solo se la probabilità che un cliente arrivi è compensata dalla probabilità che un cliente parta:

$$\lambda \pi_0 = \mu \pi_1$$

Lo stato con un solo cliente ubbidisce all'equazione di equilibrio:

$$\lambda \pi_0 + \mu \pi_2 = (\lambda + \mu) \pi_1$$

Gli stati con due o più clienti ubbidiscono a equazioni di equilibrio simili alla precedente.
Si possono ottenere le probabilità successive in funzione di π_0 .

Normalizzando la somma delle probabilità a uno si ottiene:

$$\pi_k = (1-\rho)\rho^k \quad \text{ove } \rho = \lambda/\mu$$

Il numero medio di clienti nel sistema e la lunghezza della coda sono:

$$L = E(N) = \sum_{k=0}^{\infty} k\,\pi_k = \frac{\rho}{1-\rho} \qquad L_q = \frac{\rho^2}{1-\rho}$$

Il tempo medio passato nel sistema, in coda o in servizio, si trova con la formula di Little

$$W = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}$$

La lunghezza della coda e il tempo di attesa esplodono quando l'intensità del traffico ρ si avvicina a 1

Sistema *M/M/1/K*

L'esempio precedente è utopistico in quanto nessun sistema può ricevere un numero infinito di clienti.
Il sistema ***M/M/1/K*** ammette K clienti; i clienti successivi sono persi.

La probabilità che il sistema contenga i clienti è:

$$\pi_i = \frac{\rho^i}{\sum_{j=0}^{K} \rho^j} = \frac{(1-\rho)\rho^i}{1-\rho^{K+1}} \qquad \text{per } i = 0, 1, \ldots, K \text{ quando } \rho < 1$$

$$\pi_i = \frac{1}{K+1} \qquad \text{per } i = 0, 1, \ldots, K \text{ quando } \rho = 1$$

Il numero medio di clienti nel sistema è, per $\rho < 1$:

$$L = \sum_{i=0}^{K} i\,\pi_i = \sum_{i=0}^{K} i \frac{1-\rho}{1-\rho^{K+1}}\rho^i = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$$

quindi sempre più piccolo che nel caso ***M/M/1*** ove era $L = \frac{\rho}{1-\rho}$

La probabilità che un cliente sia respinto è π_K .

I clienti arrivano nel sistema ad un tasso ridotto $\lambda_{eff} = (1-\pi_K)\lambda$

Il *server* è utilizzato soltanto a $(1-\pi_K)\rho$

In particolare, se il sistema può contenere un solo cliente, ***M/M/1/1***

$$\pi_0 = \frac{1}{1+\rho} \qquad \pi_1 = \frac{\rho}{1+\rho} = \frac{\lambda}{\lambda+\mu}$$

π_1 è il numero medio di clienti nel sistema e la frazione di tempo morto

cfr. la definizione di tempo morto di un rivelatore nel caso non estensibile: si tratta di un sistema ***M/D/1/1***, ove D = costante

con riferimento alla lezione 1: **Tempo Morto**

Tempo morto nel caso non estensibile, non paralizzabile

n vero tasso di conteggio entrante => λ
m tasso di eventi registrati $\rho = \lambda/\mu = n\tau$
τ tempo morto (servizio) => $1/\mu$

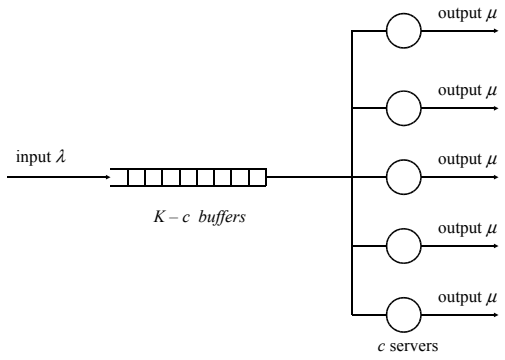
$$n = m + nm\tau \qquad n = \frac{m}{1-m\tau} \qquad m = \frac{n}{1+n\tau}$$

π_0 frazione di tempo vivo
π_1 frazione di tempo morto

$$\pi_0 = \frac{m}{n} = \frac{1}{1+n\tau} = \frac{1}{1+\rho} \qquad \pi_1 = \frac{n-m}{n} = m\tau = \frac{\rho}{1+\rho}$$

Sistema *M/M/c/K*

Ci sono c *servers* uguali e $K-c$ *buffers* $(c \le K)$: il numero massimo di clienti nel sistema è K



Le probabiltà che ci siano 0 e *n* clienti nel sistema sono,
se $\rho = \lambda / \mu$ (μ relativo a <u>un solo</u> *server*)

$$\pi_0 = \frac{1}{\sum_{l=0}^{c-1} \frac{\rho^l}{l!} + \sum_{l=c}^{K} \frac{\rho^l}{c!\, c^{l-l}}}$$

$$\pi_n = \pi_0 \rho^n / n! \qquad 0 \le n < c$$

$$\pi_n = \pi_0 \rho^n / (c!\, c^{n-K}) \qquad c \le n \le K$$

Per $n = K$ tutti i *buffers* sono occupati, si perdono clienti e si ha tempo morto. La frazione di tempo morto è $\pi_K = \pi_0 \rho^K / c!$

Il numero di clienti che entrano effettivamentenel sistema è minore di λ

$$\lambda_{eff} = (1 - \pi_K)\lambda$$

Il sistema ***M/M/c*** è rappresentato dalle stesse formule quando $K \Rightarrow \infty$

Il *buffer* è infinito e il sistema è simile a ***M/M/1***

Le probabiltà che ci siano 0 e *n* clienti nel sistema sono, con $\rho = \lambda / \mu$
(μ relativo a <u>un solo</u> *server*)

$$\pi_0 = \frac{1}{\sum_{l=0}^{c-1} \frac{\rho^l}{l!} + \frac{\rho^c}{c!(1-\rho)}}$$

$$\pi_n = \pi_0 \rho^n / n! \qquad 0 \le n < c$$

$$\pi_n = \pi_0 \rho^n / (c!\, c^{n-c}) \qquad c \le n$$

La probabilità che non ci siano servers liberi e che si debba aspettare in coda è

$$P_{coda} = \pi_0 \frac{\rho^c}{c!\left(1 - \frac{\rho}{m}\right)}$$ formula *C* di Erlang

Sistema ***M/M/c/c***
Non ci sono *buffers*

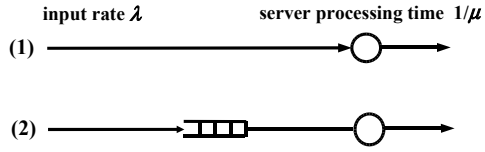
$$\pi_0 = \frac{1}{\sum_{l=0}^{c} \frac{\rho^l}{l}}$$

$$\pi_n = \pi_0 \rho^n / n! \qquad 0 \le n \le c$$

La probabilità di non trovare un server libero (frazione di tempo morto) è

$$\pi_c = \frac{\frac{\rho^c}{n}!}{\sum_{l=0}^{c} \frac{\rho^l}{l}}$$ formula *B* di Erlang

Esempi
a) - quant'è il tempo morto del sistema **(1)** quando la $\lambda = \mu$?



Il sistema **(1)** è ***M/M/1/1***; la probabilità che sia occupato (frazione di tempo morto) e il numero di clienti nel sistema sono, per ρ = 1:

$$\pi_1 = \frac{\rho}{1+\rho} = \frac{\lambda}{\lambda+\mu} = 0.5 \qquad L = \sum_{i=0}^{1} i\,\pi_i = \frac{\rho}{1+\rho} = 0.5$$

b) – come cambia il tempo morto se si aggiunge un *buffer* (***derandomizer***) di profondità 4 come in **(2)** ? La coda si allunga?

Il sistema **(2)** è ***M/M/1/K*** con $K = 4 + 1$
La probabilità che il sistema contenga 5 clienti è, per $\rho = 1$:

$$\pi_5 = \frac{\rho^5}{\sum_{i=0}^{5} \rho^i} = 1/6$$

La frazione di tempo morto si riduce da 1/2 a 1/6 grazie ai 4 *buffers*
Ovviamente ora c'è coda. Il numero medio di clienti nel sistema è, per $\rho = 1$:

$$L = \sum_{i=0}^{5} i\,\pi_i = \frac{\sum_{i=0}^{5} i\,\rho^i}{\sum_{i=0}^{5} \rho^i} = 15/16$$

c) - quale velocità deve avere il processore in **(1)** per ottenere lo stesso tempo morto come in **(2)**?

In un sistema ***M/M/1/1*** come il sistema **(1)**, la frazione di tempo morto è

$$\pi_1 = \frac{\rho}{1+\rho} = \frac{\lambda}{\lambda+\mu}$$

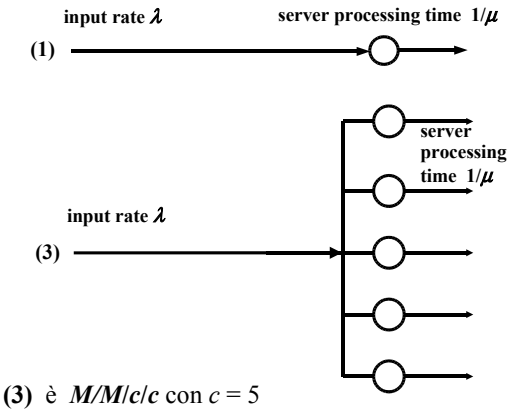
Essa vale 1/6 se $\rho = 1/5$ o $\mu = 5\lambda$: quindi è necessario un processore cinque volte più veloce, ma poco sfruttato

Il numero medio di clienti nel sistema ***M/M/1/1*** con $\mu = 5\lambda$ è solo

$$L = \sum_{i=0}^{1} i\,\pi_i = \frac{\rho}{1+\rho} = 1/6$$

I *buffers* costano meno dei processori ma allungano il tempo di permanenza nel sistema (latenza)

d) – è meglio avere un processore rapido come in **(1)** ovvero tanti processori lenti come in **(3)**? Come cambiano tempo morto e numero di clienti nel sistema?
Il sistema **(3)** ha la stessa potenza di calcolo di **(1)**, per il quale $\rho = 1/5$



Il sistema **(3)** è ***M/M/c/c*** con $c = 5$

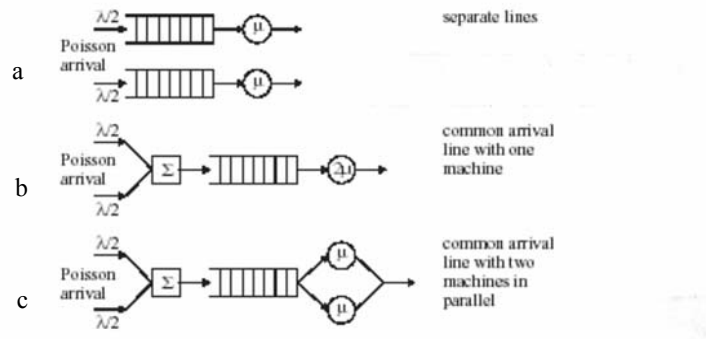
Supponiamo di avere 5 processori lenti, ciascuno con $\rho = 1$

$$\pi_n = \frac{\frac{\rho^n}{n!}}{\sum_{l=0}^{5} \frac{\rho^l}{l!}} \qquad 0 \le n \le 5 \qquad \pi_5 = \frac{\frac{\rho^5}{5!}}{\sum_{l=0}^{5} \frac{\rho^l}{l!}} = 1/32$$

$$L = \sum_{0}^{5} i\,\pi_i = \frac{\sum_{0}^{5} i\,\frac{\rho^i}{i!}}{\sum_{0}^{5} \frac{\rho^i}{i!}} = 325/32$$

	capacità di processo del sistema	$\rho = \lambda/\mu$ **per server**	**frazione di tempo morto**	**numero di clienti nel sitema**
a - M/M/1/1	μ	1	1/2	1/2
b - M/M/1/5	μ	1	1/6	15/6
c - M/M/1/1	5μ	0.2	1/6	1/6
d - M/M/5/5	5μ	1	1/326	325/326

f) – **confronto tra tre topologie**



Total time in system: $W^{(b)} < W^{(c)} < W^{(a)}$

Code più generali

Sistema ***M/G*/1**
Gli arrivi seguono un processo di Poisson con con tasso medio λ ma il servizio avviene secondo una distribuzione diversa dall'esponenziale.
Sia $1/\mu$ il valor medio della distribuzione dei tempi di servizio
σ^2 la varianza della distribuzione dei tempi di servizio

La formula di Pollaczek-Khinchin permette di calcolare il tempo medio di attesa in coda in funzione di σ^2 e di $\rho = \lambda/\mu$:

$$W_q = \frac{\lambda(\sigma^2 + 1/\mu^2)}{2(1-\rho)}$$ formula P - K

Il tempo totale nel sistema è:

$$W = \frac{1}{\mu} + \frac{\lambda(\sigma^2 + 1/\mu^2)}{2(1-\rho)}$$

Usando la formula di Little si ottengono il numero medio di clienti in coda

$$L_q = \lambda W_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}$$

e il numero medio di clienti nel sistema

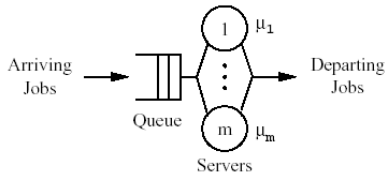
$$L = \lambda W = \rho + \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}$$

Nel caso *M/M*/1 in cui i tempi di servizio sono distribuiti esponenzialmente $\sigma^2 = 1/\mu^2$ e si ritrova $L = \frac{\rho}{1-\rho}$

Nel caso di tempi di servizio costanti come in *M/D*/1, $\sigma^2 = 0$: il numero medio di clienti nel sistema è più piccolo $L = \frac{\rho}{1-\rho}(1 - \frac{\rho}{2})$

Sistemi eterogenei

Un sistema simile a ***M/M/m,*** in cui gli ***m*** *servers* hanno tempi di servizio $1/\mu_k$ diversi, si presenta in molti casi pratici.

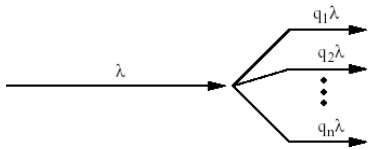


La scelta della strategia di distribuzione del carico (*load balancing*) permette di utilizzare il sistema nel modo migliore.Tra le strategie più comuni:

- Approssimazione ***Heavy Traffic***

Nel caso in cui i processori siano molto occupati, ossia quando un cliente in arrivo trova al massimo un processore libero, si può fare l'ipotesi

che il flusso entrante di clienti si distribuisca tra i vari processori secondo una probabilità q_k proporzionale alla capacità di servizio dei processori stessi.



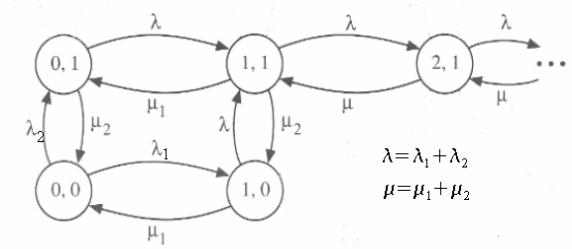
$$q_k = \frac{\mu_k}{\sum_{i=1}^{m} \mu_i} \qquad \lambda_k = \lambda q_k = \lambda \frac{\mu_k}{\sum_{i=1}^{m} \mu_i} \qquad \rho_k = \frac{\lambda_k}{\mu_k} = \frac{\lambda}{\sum_{i=1}^{m} \mu_i}$$

Il carico ρ_k su ciascun processore è costante

- Approssimazione ***Random***
- Approssimazione ***Fastest Free Server, FFS:*** il prossimo cliente è servito dal processore più veloce disponibile

Limitiamo l'analisi al sistema ***M***/***M***/2 eterogeneo.
Il sistema è rappresentato da due stati. Il primo stato k_1 rappresenta l'occupazione del primo *server* e della coda: esso può assumere i valori $k_1 = 0, 1 \ldots . \infty$. il secondo stato k_2 rappresenta l'occupazione del secondo *server*: esso può assumere i valori $k_2 = 0, 1$

Il diagramma nascite-morti per questo sistema prevede un preambolo, nel quale i due processori sono in competizione, secondo la strategia scelta, seguito da una sequenza in cui i due processori collaborano.



Il flusso entrante λ di clienti si divide in due correnti λ_1 e λ_2
Si applicano le equazioni di equilibrio ai nodi (0,0), (0,1) e (1,0) ottenendo le probabilità $\pi(0,1)$, $\pi(1,0)$ e $\pi(1,1)$ in funzione di $\pi(0,0)$.

A destra dello stato (1,1) la coda si comporta, eccetto per la normalizzazione, come una coda ***M***/***M***/2 con due *servers* di cui la capacità si sommano: quindi i termini successivi si possono esprimere in funzione di $\pi(1,1)$ che corrisponde a π_2 nella notazione ***M***/***M***/2. I termini successivi a π_2 si esprimono in funzione soltanto del carico del sistema distibuito su entrambi i *servers* pienamente occupati, definito da $a = \lambda/(\mu_1 + \mu_2)$

Nel preambolo:

- *heavy traffic* $\lambda_1 = a\,\mu_1 \quad \lambda_2 = a\,\mu_2$
- *random* $\lambda_1 = \lambda_2 = \lambda/2$
- ***FFS*** $\lambda_1 = \lambda \quad \lambda_2 = 0$

La tabella seguente presenta $\bar{\rho}$, definito da $\frac{1}{\bar{\rho}} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\rho_k}$, (carico medio o coefficiente di utilizzazione media del sistema), $\rho_k = \frac{\lambda_k}{\mu_k}$ (carico medio per ciascun processore) e L (numero medio di clienti nel sistema) per le varie strategie nel caso eterogeneo ***M***/***M***/2 con:

$$\lambda = 0.2 \quad \mu_1 = 0.5 \text{ e } \mu_2 = 0.25$$

	HT	***FFS***	***random***
$\bar{\rho}$	0.267	0.234	0.275
ρ_1	0.267	0.306	0.230
ρ_2	0.267	0.189	0.341
L	0.575	0.533	0.615

Quando il numero dei processori è grande la complicazione algebrica cresce e può essere preferibile ricorrere a simulazione.

La tabella seguente permette di confrontare il coefficiente di utilizzazione media del sistema $\bar{\rho}$ e la lunghezza media della coda $L_q = L - \bar{\rho}$ per un sistema eterogeneo ***M*/*M*/5** nel caso approssimato *heavy traffic* e nelle simulazioni *FFS* e *Random*

λ	μ_k	$\bar{\rho}(HT)$	$\bar{\rho}(FFS)$	$\bar{\rho}(rand)$	$L_q(HT)$	$L_q(FFS)$	$L_q(rand)$
73	10,15,20,20,25	0.811	0.799	0.819	2.472	2.367	2.495
73	16,17,18,19,20	0.811	0.807	0.812	2.472	2.476	2.500
73	5,10,18,22,35	0.811	0.808	0.844	2.472	2.443	2.626
81.1	10,15,20,25,30	0.811	0.799	0.824	2.472	2.424	2.523
81.1	8,14,20,26,32	0.811	0.801	0.830	2.472	2.442	2.569
81.1	8, 9,20,31,32	0.811	0.807	0.839	2.472	2.456	2.606
81.1	4, 8,16,32,40	0.811	0.826	0.858	2.472	2.549	2.713

Modelli più complessi

Sia data una coda ***D*/*M*/1** in cui i clienti arrivano in modo deterministico, per esempio ad intervalli regolari. In questo caso la previsione del comportamento della coda non dipende solo dalla situazione presente, ma anche dal ricordo di quando arrivò l'ultimo cliente. Sono necessari metodi più generali di quelli presentati (*imbedded Markov processes*).

Se la disciplina della coda comporta priorità, i modelli analitici si complicano considerevolmente.

Una variazione di questa situazione si ha quando molte code sono servite da un solo server secondo un certo algoritmo per distribuire le risorse.

Esempio di coda *G*/*D*/1
I clienti entrano nel sistema con distribuzione uniforme dei tempi di arrivo compresa tra 2 e 4 secondi, quindi mediamente ogni 3 secondi.
Il tempo di servizio è fisso a 1 secondo.

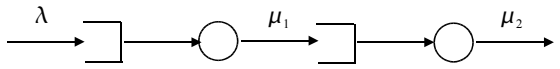
Il numero di clienti in coda L_q è sempre zero.
Il numero medio di clienti nel sistema L è 1/3

Una coda M/D/1 con $\lambda = 1/3$, $\mu = 1$, $\rho = 1/3$ avrebbe un numero medio di clienti in coda (formula di P – K) $L_q = \frac{\rho^2}{2(1-\rho)} = 1/12$ e nel sistema $L = 5/12$

Una coda M/M/1 con $\rho = 1/3$ avrebbe un numero medio di clienti in coda di $L_q = \frac{\rho^2}{(1-\rho)} = 1/6$ e nel sistema $L = 1/2$

Reti di Code

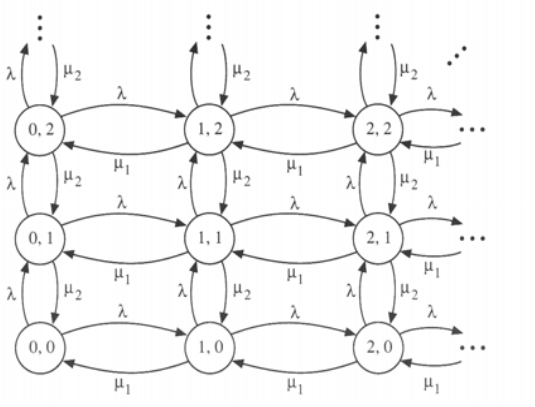
Due code ***M*/*M*/1** in tandem non sono riducibili a una sola coda e neppure a due code scorrelate



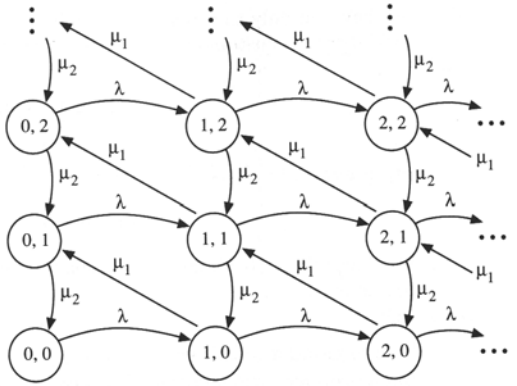
Per quanto i tempi tra due arrivi e i tempi di servizio siano distribuiti esponenzialmente, c'è una struttura sequenziale asimmetrica .

Si può dimostrare che l'uscita della prima coda è poissoniana e quindi anche la seconda coda lo è.

Nel formalismo dei diagrammi di transizione di probabilità due code scorrelate entrambe con frequenza d'ingresso λ e con tempi di servizio $1/\mu_1$ e $1/\mu_2$ rispettivamente, sono rappresentate dal diagramma



Se le due code sono in tandem il diagramma è



Product form

In un sistema costituito da due code poissoniane, la probabilità che la prima coda contenga n_1 clienti e che la seconda coda contenga n_2 clienti sia $\pi(n_1, n_2)$

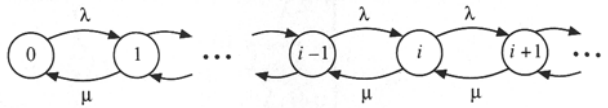
Nel caso di due code scorrelate $\pi(n_1, n_2) = \pi(n_1)\pi(n_2)$.

Questa condizione (*product form*) può verificarsi anche per le code componenti di sistemi complessi, per esempio le due code M/M/1 in tandem. In tal caso sono possibili soluzioni analitiche per le probabilità in stato stazionario.

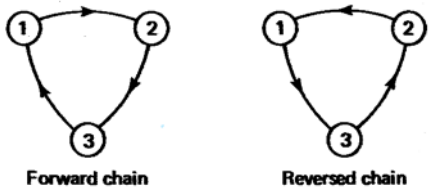
Reversibilità

Una catene markoviana che continua ad essere una catena colle stesse proprietà quando si scambia t con $\tau - t$, ove $\tau = costante$, è detta reversibile.

Per esempio, in una coda M/M/1 quando si cambia la direzione del tempo è come scambiare λ con μ e le equazioni di equilibrio restano le stesse.



Questo non è il caso in generale.



Nel caso in cui le code componenti un sistema complesso sono reversibili sono possibili soluzioni analitiche per le probabilità in stato stazionario

Due code M/M/1 in tandem non sono reversibili.

Long Range Dependence, Auto-similarità, *Heavy Tails*

Il formalismo descritto finora si applica con successo ai processi stazionari, con distribuzioni di arrivi e tempi di servizio abbastanza regolari.
Il traffico nelle reti di trasmissione dati (locali e non), comprende di solito diverse componenti (TCP/IP, http, ftp, video, voce etc.) e spesso presenta picchi di richieste e ritardi di trasmissione. Per queste reti sono stati osservati due tipi di comportamento delle distribuzioni di traffico: ***long range dependence*** (***LRD***) e **auto-similarità** .

In un processo stocastico stazionario $X(s)$ il coefficiente di auto-correlazione tra il tempo s e il tempo $s+t$ è:

$$R(X(s), X(s+t)) = \frac{cov(X(s), X(s+t))}{\sigma^2_{X(s)}}$$

Non esiste ***LRD*** se $\int_0^\infty |R(X(s), X(s+t))| dt$ è finito,

ovvero, secondo un'altra definizione, se il coefficiente di auto-correlazione decresce come $t^{-\alpha}$ con $0<\alpha<1$.

L'**auto-similarità** (*self-similarity*) è la proprietà tipica dei frattali: se guardati ad "ingrandimenti" diversi, gli oggetti presentano sempre la stessa struttura.
Nel caso delle reti, il traffico ha dei *bursts* che si ripetono su scale variabili dal millisecondo all'ora.

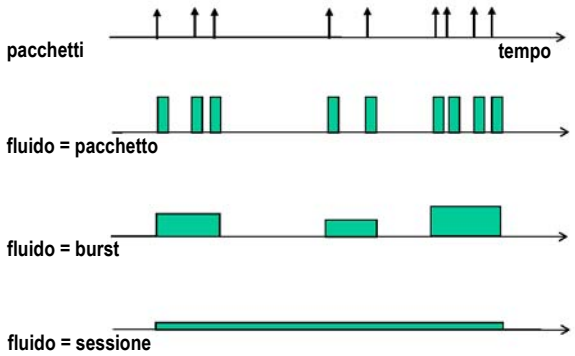
Un processo stocastico $X(t)$ è auto-simile con parametro (di Hurst) H se per ogni $a>0$ il processo $a^{-H} X(at)$ ha le stesse proprietà statistiche di $X(t)$, tipicamente:

Valor medio	$E[X(t)] = E[X(at)]/a^H$
Varianza	$Var[X(t)] = Var[X(at)]/a^2$
Auto-correlazione	$R(X(s), X(s+t)) = R(X(as), X(as+at))/a^2$

Quando si verificano ***LRD*** e auto-similarità le densità di probabilità presentano di solito code molto significative (***heavy tails***) che forzano la varianza verso valori elevati.

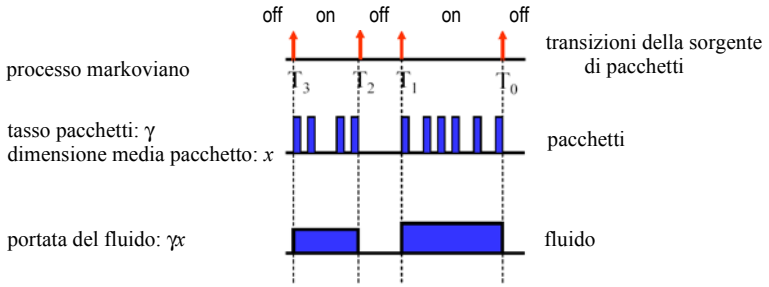
Modello *Fluid Flow*

Si basa sull'approssimazione che gli arrivi individuali possono essere riuniti in un flusso medio durante un intervallo.



Intuitivamente la rappresentazione come fluido sembra più semplice, se riferita a un tempo abbastanza lungo

Di solito ci si riferisce a un ***burst***, definito come il tempo in cui la sorgente di pacchetti emette in modo continuo.

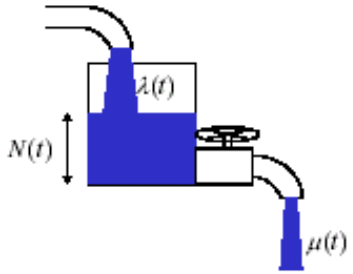


Si rappresenta un *burst* di pacchetti con la portata istantanea di fluido

Se $N(t)$ rappresenta il numero di clienti nel sistema, l'approssimazione del modello flusso di fluido stabilisce che

$$\frac{d}{dt}N(t) \approx \lambda(t) - \mu(t)$$

Il modello flusso di fluido può essere utilizzato per studiare i fenomeni transitori, ma non può descrivere lo stato stazionario.



Simulazione

Il formalismo matematico markoviano, anche in casi abbastanza semplici di sistemi deterministici, sistemi con priorità o reti di code, non permette in generale di trovare una soluzione analitica. Si deve quindi ricorrere alla simulazione.
Si deve ricorrere alla simulazione anche in casi in cui il comportamento a regime è calcolabile, ma si desidera conoscere il comportamento transitorio.

La generazione delle distribuzioni dei gli intervalli tra due arrivi e dei tempi di servizio è relativamente semplice, partendo da densità di probabilità definite analiticamente o da dati sperimentali.
Semplice ma delicata è in generale la descrizione degli algoritmi di priorità.
La descrizione delle interconnessioni tra *buffers* e *servers* e tra le code in rete si fa di solito usando linguaggi del tipo di **VHDL** (Very High Speed Integrated Circuits (**V**HSIC) **H**ardware **D**escription **L**anguage) o **VERILOG**, come per la progettazione di circuiti elettronici.

I tempi di simulazione possono essere **molto lunghi**.
Si ricorre perciò a soluzioni ibride, in cui si usano la formulazione matematica degli elementi più trattabili , mentre gli altri sono simulati.

Per ridurre i tempi di simulazione può essere necessario rinunciare a trattare gli arrivi individuali, approssimandone il comportamento collettivo come nel modello *fluid flow*.

Esistono programmi di simulazione che realizzano buona parte di queste funzioni, per esempio MOSEL e PEPSY (Un. Erlangen)

Referenze

La letteratura che tratta di *Queuing Theory* e delle sue applicazioni è molto vasta.
Una presentazione breve, chiara e rigorosa si trova nel capitolo 3 di:

D. P. Bertsekas and R. G. Gallager, Data Networks, 2nd ed., Prentice Hall, 1992

Referenze (cont.)

Un libro che tratta molte applicazioni a calcolatori e sistemi di calcolatori, e che ha una buona bibliografia, è:

R. Nelson, Probability, Stochastic Processes and Queuing Theory, Springer Verlag, 1995

Moltissime lezioni si trovano sul web; per esempio:

W. Ukovich, Teoria delle code o file di attesa, dispense preparate da R. Pesenti, Consorzio Nettuno, Università di Trieste
www.citam.unibo.it/Tecnopolo/EsercRete/ Mate3/Code/Dispense.pdf
G Bolch, Modellierung und Leistungsbewertung von Rechnensystemen
http://www4.informatik.uni-erlangen.de/Lehre/WS02/V_MLR/Skript/

Per la simulazione, oltre alla referenza precedente:

Ptolemy Project, Heterogeneous Modeling and Design, UC Berkeley ECCS, http://ptolemy.eecs.berkeley.edu/

Appendice: Probabilità

Distribuzioni di probabilità

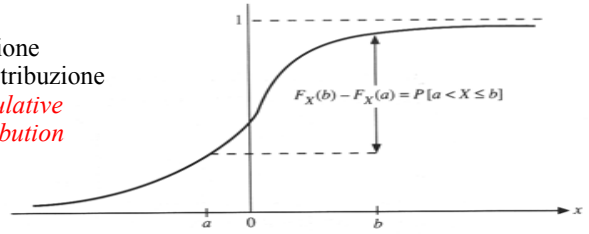
Una variabile aleatoria X (definita in un certo spazio di probabilità) si può visualizzare come una corrispondenza che produce un valore reale $X(\omega)$ per ogni "evento" ω appartenente allo spazio di probabilità. Variabili aleatorie discrete possono assumere solo valori numerabili (in numero finito o infinito); altrimenti sono variabili continue.

Il valore di una variabile aleatoria X può essere specificato soltanto in modo probabilistico, usando una funzione di distribuzione $F_X(x)$, definita come la probabilità $P(X<x)$, o la densità corrispondente $f_X(x)$.

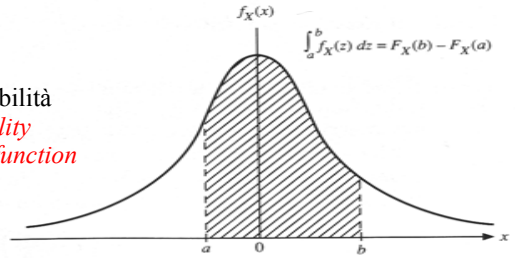
discreta $$F_X(x)=P(X<x)=\sum_{i\le x} f_X(i)$$

continua $$F_X(x)=P(X<x)=\int_{-\infty}^{x} f_X(x)\,dx$$

Funzione
di distribuzione
Cumulative distribution



Densità
di probabilità
Probability density function



Nota – I fisici chiamano di solito la densità di probabilità "distribuzione" (es. distribuzione di massa invariante). In statistica, distribuzione è riservato all'integrale della densità di probabilità.

Valore di attesa *(Expectation)*

Le densità di probabilità sono usate come fattore ponderale per ottenere informazioni sulle variabili aleatorie.

Se $g(X)$ è una funzione di una variabile aleatoria X, il valore di attesa di $g(X)$ è il numero

$$E[g(X)]=\int_{\Omega} g(x) f_X(x)\,dx$$

ove il dominio di integrazione è l'intero spazio di definizione di X.

L'operatore che calcola il valore di attesa è lineare

Valor medio $$\mu=E[X]$$

discreta $$\mu=\sum_{i=-\infty}^{\infty} i\, f_X(i)$$

continua $$\mu=\int_{-\infty}^{\infty} x\, f_X(x)\,dx$$

Varianza $$\sigma^2=E[(X-\mu)^2]=E[X^2]-\mu^2$$

discreta $$\sigma^2=\sum_{i=-\infty}^{\infty} i^2 f_X(i)$$

continua $$\sigma^2=\int_{-\infty}^{\infty} x^2 f_X(x)\,dx-\mu^2$$

Covarianza e correlazione

Se $f_{XY}(x, y)$ è la densità di probabilità congiunta di due variabili aleatorie X e Y, si può generalizzare il valore di attesa a una funzione di due variabili.

Valori medi e varianze sono:

$$\mu_X = E[X] = \int\int x f_{XY}(x, y)\,dx\,dy = \int x\,dx \int f_{XY}(x, y)\,dy$$
$$\mu_Y = E[Y] = \int\int y f_{XY}(x, y)\,dx\,dy$$
$$\sigma_X^2 = E[(X-\mu_X)^2]$$
$$\sigma_Y^2 = E[(Y-\mu_Y)^2]$$

Si definiscono la covarianza

$$cov(X, Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - E[X]E[Y$$

e il coefficiente di correlazione

$$corr(X, Y) = R(X, Y) = \frac{cov(X,}{\sigma_X \sigma_Y}$$

Esempio 1: **distribuzione binomiale**, discreta

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad F_X(x) = \sum_{k \le x} \binom{n}{k} p^k (1-p)^{n-k}$$

$$E[X] = n\,p$$

Esempio 2: **distribuzione di Poisson**, discreta

Limite della precedente quando $p = \lambda/n$, $n \Rightarrow \infty$

$$f_X(i) = \frac{\lambda^i e^{-\lambda}}{i} \qquad F_X(i) = \sum_{k \le i} \frac{\lambda^k e^{-\lambda}}{k}. \qquad i = 0, 1, 2, \ldots$$

$$E[X] = \lambda$$

Esempio 3: **distribuzione esponenziale**, continua

$$f_X(x) = \lambda e^{-\lambda x} \qquad F_X(x) = 1 - e^{-\lambda x} \qquad 0 \le x$$

valor medio $E[X] = 1/\lambda$ varianza $\sigma^2 = 1/\lambda^2$

Esempio 4: **distribuzione di Erlang**, continua

Sommando k variabli aleatorie esponenziali tutte di valor medio $1/k\lambda$ si ottiene una distribuzione di Erlang E_k di ordine (o fattore di forma) k.

Per $k = 1$ vedi Esempio 3.

Per $k = 2$

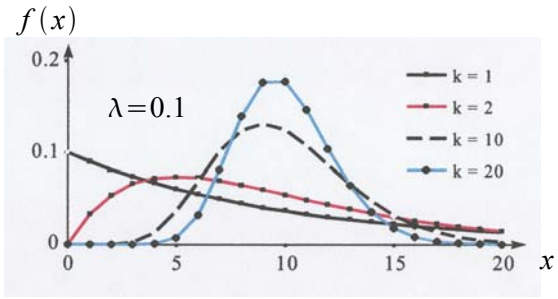
$$f_X(x) = \lambda^2 x e^{-\lambda x} \qquad F_X(x) = 1 - (1 + 2\lambda x) e^{-2\lambda x} \qquad 0 \le x$$

valor medio $E[X] = 1/\lambda$ varianza $\sigma^2 = 1/2\lambda^2$

Per $k \ge 1$

$$f_X(x) = \frac{(k\lambda)^k x^{k-1}}{(k-1)!} e^{-k\lambda x} \qquad F_X(x) = 1 - e^{-k\lambda x} \sum_{j=0}^{k-1} \frac{(k\lambda x)^j}{j!} \qquad 0 \le x$$

valor medio $E[X] = 1/\lambda$ varianza $\sigma^2 = 1/k\lambda^2$



Distribuzione di Erlang