SIMULAZIONE

Prof.ssa Amelia Giuseppina Nobile[1]

(a.a. 2023/2024)

19 maggio 2024

[1]Dipartimento di Informatica, Università degli Studi di Salerno

Capitolo 1

Sistemi di servizio

1.1 Introduzione

Un'area di grande interesse dell'informatica, dell'ingegneria e della matematica applicata è la *teoria delle file di attesa*, detta anche *teoria delle code*. La teoria delle file di attesa si propone di formulare e analizzare modelli matematici e di simulazione atti a descrivere sistemi reali in cui il generico utente richiede un particolare servizio e deve attendere in qualche tipo di coda (o fila di attesa) se il servitore non è immediatamente disponibile. Tipici esempi in cui si presentano file di attesa sono le chiamate ad un centralino telefonico, gli utenti in banca, alla posta o in un ospedale, i clienti in mense, in supermarket o in ristoranti, le persone in attesa di un taxi, le automobili ad un incrocio, gli aerei in attesa di decollare o di atterrare in un aeroporto, i pezzi in attesa di essere lavorati, le macchine in avaria in un'officina, ...

I risultati forniti dalla teoria delle file di attesa trovano applicazione in numerosi campi: sistemi di elaborazione, sistemi di comunicazione e di trasmissione dati, sistemi di trasporto, sistemi di produzione industriale, sistemi per la gestione di servizi pubblici e privati, ...

La teoria delle file di attesa è essenzialmente di natura probabilistica e fornisce una descrizione dei cambiamenti di stato nella lunghezza delle code, del tempo di permanenza di un utente nella fila di attesa, del tempo di attesa di un utente nel sistema, del periodo di occupazione e del periodo di ozio di un centro di servizio, ...

L'analisi effettuata mediante la teoria delle file di attesa si propone di stabilire la tipologia dei modelli più adatti a descrivere sistemi di servizio reali fornendo idonee misure di prestazione e individuando gli opportuni cambiamenti da apportare a tali sistemi per migliorare, se necessario, le loro prestazioni.

Un generico sistema di servizio può essere schematizzato come illustrato nella Figura 1.1 nella quale sono rappresentati:

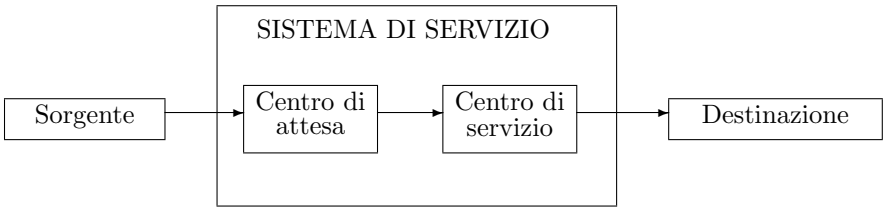

Figura 1.1: Rappresentazione di un sistema di servizio.

- *la sorgente*, ossia l'insieme dei potenziali utenti che si possono presentare al sistema di servizio;
- *il centro di attesa*, ossia l'insieme delle richieste di servizio che, non potendo essere immediatamente soddisfatte, restano in attesa di poter essere prese in considerazione;
- *il centro di servizio*, ossia l'insieme dei punti nei quali viene soddisfatta la richiesta;
- *la destinazione*, ossia l'insieme delle richieste di servizio che, essendo state soddisfatte, lasciano i punti di servizio.

Sorgente
La sorgente (o popolazione) contiene i potenziali utenti del sistema di servizio, ossia l'insieme da cui arrivano gli utenti. Essa può essere finita o infinita. Un sistema di servizio la cui sorgente è infinita è più facile da descrivere matematicamente di un sistema con sorgente finita. Ciò è dovuto alla circostanza che nel caso di sorgente finita il numero degli utenti nel sistema influenza i parametri di arrivo; infatti, se tutti i potenziali utenti sono già arrivati nel sistema i parametri di arrivo sono nulli. Spesso se la sorgente è finita e contiene numerosi potenziali utenti, si assume che sia infinita per rendere più semplice la trattazione matematica. Gli utenti possono anche provenire da diverse distinte sorgenti. Gli utenti che provengono da una stessa sorgente sono tra loro indistinguibili. Si suppone invece che esistano diverse sorgenti quando si desidera distinguere gli utenti per qualche ragione, ad esempio a causa di differenti livelli di priorità oppure a causa di differenti provenienze geografiche.

Gli utenti provenienti da una sorgente si inseriscono in un sistema di servizio per ricevere un determinato servizio. Il termine *utente* è inteso in senso generico: può essere un messaggio che deve essere trasmesso, una richiesta di servizio I/O, un programma che richiede servizi di CPU in un sistema multiprogrammato, ...

Centro di attesa
L'accesso ad un sistema di servizio può essere realizzato attraverso un *centro di attesa* (buffer) che può avere la possibilità di contenere un numero limitato o illimitato di utenti. La capacità del centro di attesa può essere quindi finita o infinita. Se il sistema possiede un centro di attesa a capacità limitata, il numero

degli utenti in attesa non può superare un certo limite caratteristico del sistema di servizio e pertanto una richiesta di servizio che si presenta quando il centro di attesa è saturo viene respinta. Un esempio di centro di attesa con capacità limitata è un centralino telefonico che può avere in attesa soltanto un numero finito di chiamate. Esistono sistemi, noti in letteratura come *loss systems*, che hanno un centro di attesa a capacità nulla; in essi se un utente arriva quando tutti i servitori sono occupati, la sua richiesta di servizio è respinta. Un esempio di centro di attesa con capacità nulla è un centralino telefonico in cui una chiamata in arrivo è accettata immediatamente oppure è rifiutata. Se, invece, il sistema possiede un centro di attesa a capacità illimitata nessuna richiesta di servizio viene perduta per quanto lunga possa essere la durata dell'attesa (a meno che gli individui in attesa decidano di allontanarsi spontaneamente dal sistema).

È evidente che non sempre le richieste di servizio entrano in attesa per poter essere soddisfatte; il fenomeno dell'attesa si presenta soltanto quando il sistema di servizio non ha risorse immediatamente disponibili per soddisfare le richieste.

Centro di servizio

Superato il centro di attesa gli utenti accedono al centro di servizio che può consistere di uno o più servitori. Il *servitore* è un'entità in grado di eseguire il servizio richiesto dall'utente. Ovviamente un sistema con più servitori può fornire simultaneamente servizio a più utenti. I servitori hanno caratteristiche identiche, lavorano in parallelo e non possono rimanere inattivi in presenza di utenti nella fila di attesa. Se tutti i servitori nel centro di servizio sono occupati l'utente, quando si inserisce nel sistema, deve mettersi in fila di attesa finché non si liberi uno dei servitori.

In un sistema di servizio si suppone che esista un unico centro di attesa anche in presenza di uno o più servitori che lavorano in parallelo. Quando ogni singolo servitore è dotato di un proprio centro di attesa (buffer) si preferisce parlare di una *rete di code* piuttosto che di un unico sistema di servizio.

Destinazione

Ogni utente lascia istantaneamente il sistema di servizio dopo aver completato il suo servizio. L'insieme delle richieste di servizio espletate sono instradate verso la destinazione.

Capacità del sistema

Con il termine *capacità del sistema* si intende il numero massimo di utenti (inclusi quelli in servizio) che possono essere contenuti nel sistema.

Disciplina di servizio

Il complesso di regole secondo le quali gli utenti in attesa passano dal centro di attesa al centro di servizio è detto *disciplina di servizio*. Essa specifica quale sarà il prossimo utente tra quelli in attesa che accede al centro di servizio non appena si libera uno dei servitori. La disciplina di servizio più comune è la disciplina *FIFO* (first-in, first-out) secondo la quale il primo arrivato è il primo ad essere servito. Esiste anche la disciplina di servizio *LIFO* (last-in, first-out) secondo la quale l'ultimo arrivato è il primo ad essere servito. Un'altra importante disciplina è la *SIRO* (service in random order) con la quale ogni utente nel centro di attesa ha la stessa probabilità di essere selezionato per il

servizio. La disciplina *PRI* (priority service) invece prevede che alcuni utenti abbiano un trattamento privilegiato; gli utenti sono in tal caso suddivisi in classi di priorità ed il sistema di coda attua una politica preferenziale nei riguardi di alcune classi di utenti.

In un sistema di servizio gli utenti ritengono fondamentale la riduzione dei tempi di attesa, mentre il gestore del sistema è solitamente interessato al massimo sfruttamento delle risorse (servitori) pur cercando di rispettare le esigenze degli utenti. Il progettista di un sistema di servizio deve quindi essere in grado in base alla struttura del sistema, alla frequenza di arrivo degli utenti, al numero di servitori e alla loro velocità di servizio di analizzare le prestazioni del sistema e di apportare, se necessario, opportuni cambiamenti alla struttura stessa.

Di fondamentale importanza per la descrizione di un sistema di servizio sono i meccanismi degli arrivi e delle partenze.

1.2 Meccanismo degli arrivi

Per descrivere il meccanismo degli arrivi occorre conoscere la funzione di distribuzione delle variabili aleatorie $T_1, T_2, \ldots$, descriventi i *tempi di interarrivo*. Il generico T_i rappresenta la lunghezza dell'intervallo di tempo che intercorre tra l'arrivo $(i-1)$-esimo e l'arrivo i-esimo.

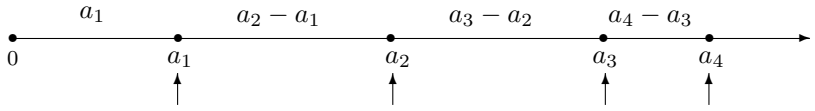

Figura 1.2: Una rappresentazione degli intervalli di interarrivo.

Ad esempio, in Figura 1.2 gli istanti di arrivo degli utenti sono denotati con $a_1, a_2, a_3, \ldots$ e gli intervalli di interarrivo hanno rispettive lunghezza a_1, $a_2 - a_1$, $a_3 - a_2$, ...

Di particolare importanza sono alcune caratteristiche numeriche di tali variabili aleatorie, quali i valori medi $E(T_i)$ e le varianze $\mathrm{Var}(T_i)$ dei tempi di interarrivo T_i $(i = 1, 2, \ldots)$.

Spesso si suppone che $T_1, T_2, \ldots$ sia una successione di variabili aleatorie indipendenti e identicamente distribuite (*iid*). In tal caso, se si denota con T una generica di tali variabili aleatorie, occorre specificare la sua funzione di distribuzione e la sua densità di probabilità.

In letteratura la funzione di distribuzione di T è solitamente denotata con $A(t) = P(T < t)$ e la sua densità di probabilità con $a(t)$. Alcune delle notazioni più frequentemente utilizzate per i tempi di interarrivo sono le seguenti:

D - tempi di interarrivo *iid* con funzione di distribuzione deterministica,

U - tempi di interarrivo *iid* con funzione di distribuzione uniforme,

M - tempi di interarrivo *iid* con funzione di distribuzione esponenziale,

E_k - tempi di interarrivo *iid* con funzione di distribuzione di Erlang di ordine k,

H_k - tempi di interarrivo *iid* con funzione di distribuzione iperesponenziale di ordine k,

GI - tempi di interarrivo *iid* con funzione di distribuzione generale.

Analizziamo ora più in dettaglio i vari meccanismi degli arrivi.

1.2.1 Meccanismo degli arrivi di tipo D

Il meccanismo degli arrivi più semplice che si possa immaginare è quello regolare (deterministico); esso è caratterizzato da una cadenza temporale costante degli arrivi. Supponiamo che un generico intervallo di interarrivo sia di lunghezza fissa $1/\lambda$. La lunghezza di tale intervallo può essere quindi descritta da una variabile aleatoria T degenere la cui funzione di distribuzione è

$$A(t) = P(T < t) = \begin{cases} 0, & t \leq 1/\lambda \\ 1, & t > 1/\lambda \,. \end{cases} \tag{1.1}$$

Il valore medio e la varianza del tempo di interarrivo sono rispettivamente:

$$E(T) = \frac{1}{\lambda}\,, \qquad \mathrm{Var}(T) = 0. \tag{1.2}$$

Meccanismi degli arrivi deterministici si possono presentare quando si considerano sistemi di servizio in cascata che prevedono due o più posti di lavoro nei quali l'uscita di un posto di lavoro costituisce l'ingresso per il successivo. Se il tempo di servizio del precedente posto di lavoro è costante, allora la distribuzione degli intervalli di interarrivo per il successivo posto di lavoro è quella deterministica. Occorre sottolineare che, eccetto nel caso di una catena di montaggio, nella realtà raramente si incontrano meccanismi degli arrivi regolari.

1.2.2 Meccanismo degli arrivi di tipo U

Nel meccanismo degli arrivi di tipo U i tempi di interarrivo sono indipendenti e identicamente distribuiti con funzione di distribuzione uniforme. Sia T una variabile aleatoria uniformemente distribuita nell'intervallo (a, b), descrivente la lunghezza di un generico tempo di interarrivo uniforme. La sua funzione di distribuzione è

$$A(t) = P(T < t) = \begin{cases} 0, & t \leq a \\ \dfrac{t-a}{b-a}\,, & a < t \leq b \\ 1, & t > b \end{cases} \tag{1.3}$$

e quindi la densità di probabilità è:

$$a(t) = \frac{dA(t)}{dt} = \begin{cases} \dfrac{1}{b-a}\,, & a < t < b \\ 0, & \text{altrimenti.} \end{cases} \tag{1.4}$$

In Figura 1.3 è rappresentata la funzione di distribuzione (1.3) e la densità di probabilità (1.4) della variabile aleatoria T uniformemente distribuita nell'intervallo (a, b).

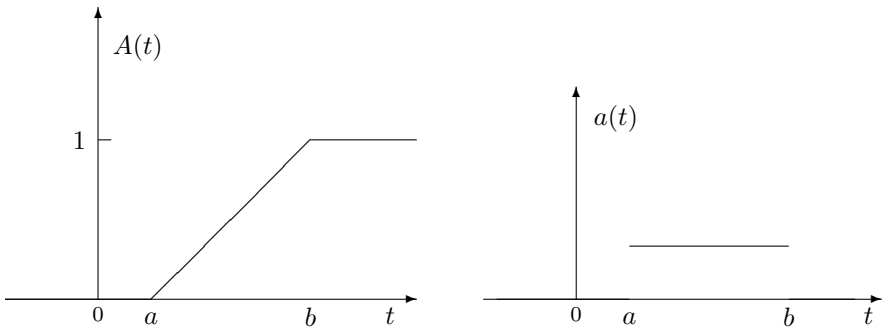

Figura 1.3: Funzione di distribuzione e densità di probabilità della variabile aleatoria T uniformemente distribuita nell'intervallo (a, b).

Il valore medio e la varianza del tempo di interarrivo sono rispettivamente:

$$E(T) = \frac{a+b}{2}\,, \qquad \mathrm{Var}(T) = \frac{(b-a)^2}{12}\,\cdot \tag{1.5}$$

Se si desidera che il tempo medio di interarrivo sia $E(T) = 1/\lambda$ (come nel caso deterministico), basta scegliere $a = 0$ e $b = 2/\lambda$. La variabile aleatoria T è allora uniformemente distribuita nell'intervallo $(0, 2/\lambda)$ con densità di probabilità:

$$a(t) = \begin{cases} \dfrac{\lambda}{2}\,, & 0 < t < \dfrac{2}{\lambda} \\ 0, & \text{altrimenti.} \end{cases} \tag{1.6}$$

La distribuzione uniforme è utilizzata in qualsiasi situazione in cui si sceglie un tempo di interarrivo "a caso" in un fissato intervallo, senza alcuna preferenza per valori inferiori, superiori o medi nell'intervallo.

1.2.3 Meccanismo degli arrivi di tipo M

Nel meccanismo degli arrivi di tipo M i tempi di interarrivo sono indipendenti e identicamente distribuiti con funzione di distribuzione esponenziale. La lettera M significa Markov ed indica la mancanza di memoria della distribuzione esponenziale. La funzione di distribuzione esponenziale è frequentemente utilizzata nella teoria delle file di attesa per le importanti proprietà di cui essa gode.

Sia T una variabile aleatoria esponenzialmente distribuita con valore medio $1/\lambda$, descrivente la lunghezza di un generico tempo di interarrivo esponenziale. La sua funzione di distribuzione è:

$$A(t) = P(T < t) = \begin{cases} 0, & t \le 0 \\ 1 - e^{-\lambda t}, & t > 0 \end{cases} \tag{1.7}$$

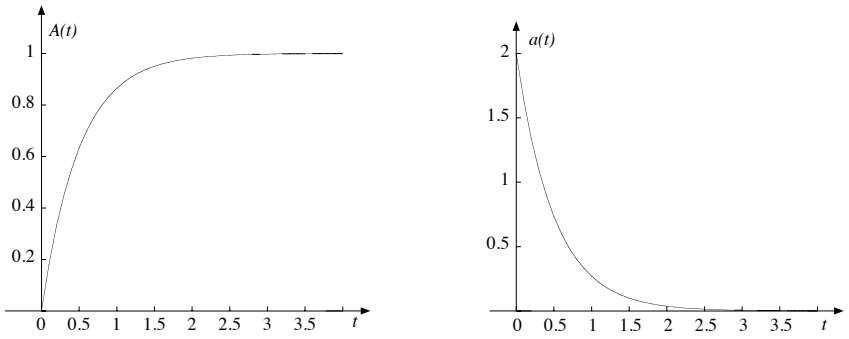

Figura 1.4: Funzione di distribuzione e densità di probabilità della variabile aleatoria T esponenzialmente distribuita con valore medio $1/\lambda$, con $\lambda = 2$.

e quindi la sua densità di probabilità è:

$$a(t) = \frac{dA(t)}{dt} = \begin{cases} \lambda\, e^{-\lambda\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{1.8}$$

In Figura 1.4 è rappresentata la funzione di distribuzione (1.7) e la densità di probabilità (1.8) della variabile aleatoria T esponenzialmente distribuita con valore medio $1/\lambda$.
La densità esponenziale è una funzione strettamente decrescente e quindi i valori più piccoli sono i più probabili. Il valore medio e la varianza del tempo di interarrivo sono rispettivamente:

$$E(T) = \frac{1}{\lambda}\,, \qquad \mathrm{Var}(T) = \frac{1}{\lambda^2}\cdot \tag{1.9}$$

Il parametro λ è l'inverso del tempo medio di interarrivo, ossia del tempo medio che intercorre tra l'arrivo di due utenti successivi, e può essere interpretato come la *frequenza media di arrivo degli utenti per unità di tempo.*

La funzione di distribuzione esponenziale riveste notevole importanza sia teorica che applicativa. Interviene spesso quando si considerano sistemi di servizio in cui si suppone che i tempi di interarrivo degli utenti (oppure i tempi di servizio) siano distribuiti esponenzialmente con un certo parametro. Interviene anche quando si considera la durata di vita, ovviamente aleatoria, di un componente elettronico o di una macchina.

Se si denota con T la lunghezza di un generico intervallo di interarrivo, la variabile aleatoria esponenziale gode (come accade per la distribuzione geometrica nel caso discreto) dell'importante proprietà

$$P(T > t + s \mid T > s) = P(T > t) \qquad (t > 0, s > 0), \tag{1.10}$$

che esprime il fatto che la probabilità condizionata che il tempo di interarrivo sia maggiore di $t+s$ dato che tale tempo è maggiore di s non dipende da quanto

si è già atteso, ossia da s. Questa circostanza esprime la *mancanza di memoria* della funzione di distribuzione esponenziale. La validità della (1.10) può essere facilmente dimostrata. Infatti, risulta:

$$P(T > t+s \mid T > s) = \frac{P(T > t+s,\ T > s)}{P(T > s)} = \frac{P(T > t+s)}{P(T > s)}$$
$$= \frac{e^{-\lambda\,(t+s)}}{e^{-\lambda\,s}} = e^{-\lambda\,t} = P(T > t).$$

Per la mancanza di memoria della funzione di distribuzione esponenziale, si ha inoltre che il *tempo di interarrivo residuo* ha la stessa distribuzione del tempo di interarrivo. Infatti, se denotiamo con T un generico tempo di interarrivo e con Z una variabile aleatoria che descrive il tempo di interarrivo residuo, ossia $Z = T - \tau$, se $t > 0$ si ha

$$P(Z \le t \mid Z > 0) = 1 - P(Z > t \mid Z > 0) = 1 - P(T - \tau > t \mid T > \tau)$$
$$= 1 - P(T > t + \tau \mid T > \tau) = 1 - e^{-\lambda\,t},$$

ossia Z è distribuita esponenzialmente con valore medio $1/\lambda$.

1.2.4 Meccanismo degli arrivi di tipo E_k

Consideriamo il sistema di servizio, schematizzato nella Figura 1.5, in cui l'ingresso al sistema è unico ed esiste un distributore che assegna ordinatamente a ciascuno delle k file di attesa gli arrivi. Alla prima fila di attesa sono così assegnati il primo arrivo, il $(k{+}1)$–esimo arrivo, il $(2k{+}1)$–esimo arrivo ed in generale il $(i\,k+1)$–esimo arrivo $(i = 0, 1, \ldots)$; alla generica j–esima $(j = 1, 2, \ldots, k)$ fila di attesa viene assegnato il j–esimo arrivo, il $(k+j)$–esimo arrivo ed in generale il $(i\,k+j)$–esimo arrivo $(i = 0, 1, \ldots)$.

Supponiamo che i tempi di interarrivo degli utenti che accedono al sistema siano indipendenti ed esponenzialmente distribuiti con valore medio $1/\varrho$. Denotiamo con T la lunghezza dell'intervallo di tempo che intercorre tra due arrivi in una generica delle k file di attesa. Poiché tra un arrivo ed il successivo in una delle k file di attesa intercorrono k intervalli di interarrivo esponenziali, la variabile aleatoria T può essere vista come la somma di k variabili aleatorie $T_1, T_2, \ldots, T_k$ indipendenti, ognuna distribuita esponenzialmente con valore medio $1/\varrho$, ossia $T = T_1 + T_2 + \ldots + T_k$. La somma di k variabili aleatorie indipendenti di tipo esponenziale con valore medio $1/\varrho$, è distribuita con densità di probabilità di Erlang di ordine k, ossia

$$a(t) = \begin{cases} \dfrac{\varrho^k}{(k-1)!}\, e^{-\varrho t}\, t^{k-1}, & t > 0 \\ 0, & t \le 0. \end{cases} \tag{1.11}$$

Il valore medio e la varianza del tempo di interarrivo sono rispettivamente:

$$E(T) = E(T_1 + T_2 + \ldots + T_k) = E(T_1) + E(T_2) + \ldots + E(T_k) = \frac{k}{\varrho}\,,$$

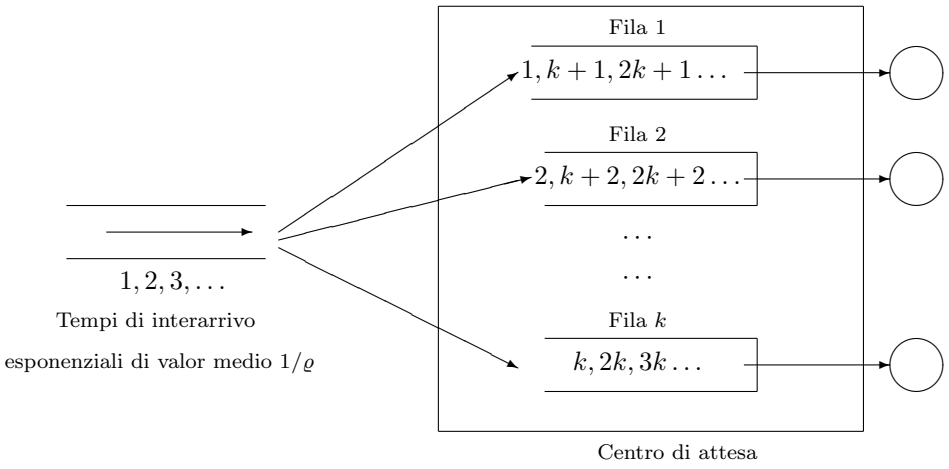

Figura 1.5: Tempi di interarrivo di tipo E_k in ognuna delle k file di attesa.

$$\mathrm{Var}(T) = \mathrm{Var}(T_1 + T_2 + \ldots + T_k) = \mathrm{Var}(T_1) + \mathrm{Var}(T_2) + \ldots + \mathrm{Var}(T_k) = \frac{k}{\varrho^2} \cdot \tag{1.12}$$

Se si desidera che il tempo medio di interarrivo sia $E(T) = 1/\lambda$ (come nei casi deterministico e esponenziale), basta scegliere $\varrho = k\,\lambda$. La variabile aleatoria T è allora caratterizzata dalla seguente densità di probabilità di Erlang di ordine k:

$$a(t) = \begin{cases} \dfrac{(k\,\lambda)^k}{(k-1)!}\, e^{-k\,\lambda t}\, t^{k-1}, & t > 0 \\ 0, & t \le 0. \end{cases} \tag{1.13}$$

e quindi il valore medio e la varianza del tempo di interarrivo sono $E(T) = 1/\lambda$ e $\mathrm{Var}(T) = 1/(k\,\lambda^2)$. Se si pone $k = 1$ nella (1.13) si ottiene la densità esponenziale (1.8). In Figura 1.6 è rappresentata la densità di probabilità (1.13) della variabile aleatoria T con densità di Erlang di ordine k con $\lambda = 2$ e per $k = 1, 2, 3, 4$.

1.2.5 Meccanismo degli arrivi di tipo H_k

Consideriamo il sistema di servizio, schematizzato nella Figura 1.7, che ha il compito di servire k differenti sorgenti. Ricordiamo che i potenziali utenti sono suddivisi in k diverse sorgenti a causa di differenti livelli di priorità loro assegnati oppure a causa di loro diverse provenienze geografiche. Supponiamo che i tempi di interarrivo degli utenti che accedono alla sorgente j–esima siano descritti da variabili aleatorie indipendenti e distribuite esponenzialmente con parametro λ_j $(j = 1, 2, \ldots, k)$. Il centro di attesa è provvisto di un ingresso unico che

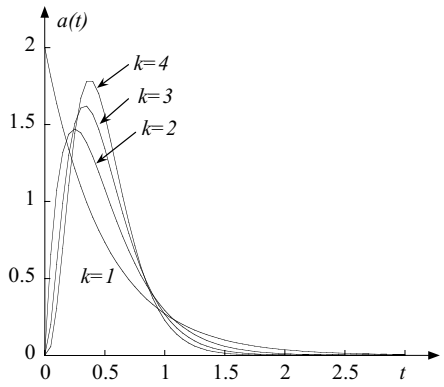

Figura 1.6: Densità (1.13) di Erlang di ordine k con $\lambda = 2$ e per $k = 1, 2, 3, 4$.

provvede a scegliere con probabilità p_j la sorgente j–esima ($j = 1, 2, \ldots, k$) ed ad avviare al centro di attesa la prima delle richieste di servizio relative alla sorgente selezionata. Assumiamo che

$$p_j \geq 0 \qquad (j = 1, 2, \ldots, k), \qquad \sum_{j=1}^{k} p_j = 1. \tag{1.14}$$

La scelta delle probabilità $p_1, p_2, \ldots, p_k$ dipende dalla priorità assegnata agli utenti delle varie sorgenti oppure dal numero di potenziali utenti provenienti da diverse località geografiche che accedono al centro di servizio.

Denotiamo con T la variabile aleatoria che descrive la lunghezza dell'intervallo di tempo tra due arrivi successivi al centro di attesa del sistema. Inoltre, denotiamo T_j la variabile aleatoria che descrive la lunghezza dell'intervallo di interarrivo degli utenti nella sorgente j–esima e con A_j l'evento "*è stata scelta la sorgente j–esima*" ($j = 1, 2, \ldots, k$). Si nota immediatamente che l'evento $\{T < t\}$ può essere così rappresentato:

$$\{T < t\} = \bigcup_{j=1}^{k} \Big[A_j \cap \{T < t\} \Big]. \tag{1.15}$$

Infatti, l'evento $\{T < t\}$ si realizza se si verifica uno qualunque dei k eventi incompatibili $A_1, A_2, \ldots, A_k$ ed inoltre $T < t$. Pertanto se $t > 0$, la probabilità della realizzazione dell'evento $\{T < t\}$ è data dalla somma delle probabilità p_j (associata all'evento A_j) per la probabilità dell'evento $\{T_j < t\}$, ossia

$$A(t) = P(T < t) = \sum_{j=1}^{k} P\Big[A_j \cap \{T < t\} \Big] = \sum_{j=1}^{k} P(A_j)\, P(T < t \mid A_j)$$

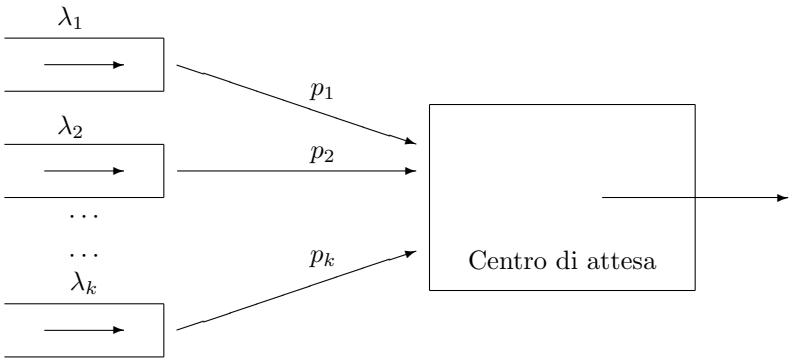

Figura 1.7: Tempi di interarrivo di tipo H_k nella fila di attesa.

$$= \sum_{j=1}^{k} p_j \, P(T_j < t) = \sum_{j=1}^{k} p_j \left(1 - e^{-\lambda_j \, t}\right).$$

Ricordando la (1.14) segue immediatamente che la funzione di distribuzione del tempo di interarrivo T è

$$A(t) = \begin{cases} 0, & t \leq 0 \\ 1 - \displaystyle\sum_{j=1}^{k} p_j \, e^{-\lambda_j \, t}, & t > 0 \end{cases} \tag{1.16}$$

e pertanto la sua densità di probabilità è

$$a(t) = \begin{cases} \displaystyle\sum_{j=1}^{k} p_j \, \lambda_j \, e^{-\lambda_j \, t}, & t > 0 \\ 0, & \text{altrimenti,} \end{cases} \tag{1.17}$$

ossia una densità iperesponenziale di ordine k relativa ai tempi di interarrivo. Ponendo $k = 1$, oppure $\lambda_1 = \lambda_2 = \ldots = \lambda_k = \lambda$, nella (1.17) si ottiene la densità esponenziale (1.8).
Dalla (1.17) è possibile ricavare immediatamente il valore medio e la varianza del tempo di interarrivo, ossia

$$\begin{aligned} E(T) &= \sum_{j=1}^{k} \frac{p_j}{\lambda_j} \, , \\ \mathrm{Var}(T) &= E(T^2) - [E(T)]^2 = 2 \sum_{j=1}^{k} \frac{p_j}{\lambda_j^2} - \left[\sum_{j=1}^{k} \frac{p_j}{\lambda_j}\right]^2 . \end{aligned} \tag{1.18}$$

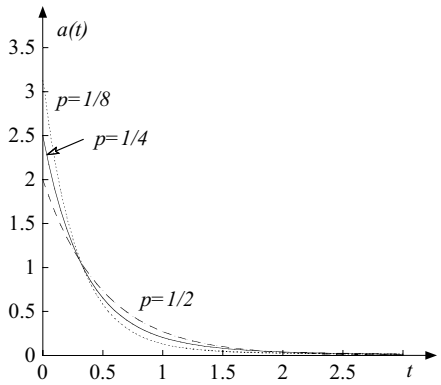

Figura 1.8: Densità iperesponenziale di ordine k con $\lambda = 2$, $k = 2$ e con $p = 1/2$, $p = 1/4$ e $p = 1/8$.

Se si desidera che il tempo medio di interarrivo sia $E(T) = 1/\lambda$, basta scegliere $\lambda_j = k\, p_j\, \lambda$ $(j = 1, 2, \ldots, k)$. La variabile aleatoria T è allora caratterizzata dalla seguente densità di probabilità iperesponenziale di ordine k:

$$a(t) = \begin{cases} k\,\lambda \sum_{j=1}^{k} p_j^2\, e^{-k\, p_j\, \lambda\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{1.19}$$

In Figura 1.8 è rappresentata sia la densità di probabilità (1.19) della variabile aleatoria T con densità iperesponenziale di ordine k con $\lambda = 2$, $k = 2$ e con $p = 1/2$, $p = 1/4$ e $p = 1/8$.

1.2.6 Meccanismo degli arrivi di tipo GI

Nel meccanismo degli arrivi di tipo GI si suppone che i tempi di interarrivo siano indipendenti e identicamente distribuiti con distribuzione di tipo generale. Occorre ricercare caratteristiche generali del sistema di servizio per una qualsiasi distribuzione dei tempi di interarrivo. Ovviamente quando si specifica il tipo di distribuzione (D, U, M, E_k, H_k,...) è possibile ottenere maggiori informazioni sull'evoluzione del sistema considerato.

1.3 Meccanismo di servizio

Per descrivere il meccanismo di servizio occorre conoscere la funzione di distribuzione delle variabili aleatorie $S_1, S_2, \ldots$, rappresentanti i tempi di servizio per ognuno degli utenti. Il generico S_i descrive la lunghezza dell'intervallo

di tempo occorrente per servire l'utente i-esimo ($i = 1, 2, \ldots$). Di particolare importanza sono alcune caratteristiche numeriche di tali variabili aleatorie, quali i valori medi $E(S_i)$, le varianze $\mathrm{Var}(S_i)$ e i coefficienti di variazione $C(S_i) = \sqrt{\mathrm{Var}(S_i)}/E(S_i)$ dei tempi di servizio S_i ($i = 1, 2, \ldots$).

Spesso si suppone che $S_1, S_2, \ldots$ sia una successione di variabili aleatorie indipendenti e identicamente distribuite. In tal caso se si denota con S una generica di tali variabili aleatorie, occorre specificare la sua funzione di distribuzione e la sua densità di probabilità. In letteratura la funzione di distribuzione di S viene solitamente denotata con $B(t) = P(S < t)$ e la sua densità di probabilità con $b(t)$. Alcune delle notazioni più frequentemente utilizzate per i tempi di servizio sono le seguenti:

D - tempi di servizio *iid* con funzione di distribuzione deterministica,

U - tempi di servizio *iid* con funzione di distribuzione uniforme,

M - tempi di servizio *iid* con funzione di distribuzione esponenziale,

E_k - tempi di servizio *iid* con funzione di distribuzione di Erlang di ordine k,

H_k - tempi di servizio *iid* con funzione di distribuzione iperesponenziale di ordine k,

G - tempi di servizio *iid* con funzione di distribuzione generale.

Analizziamo ora in dettaglio i vari meccanismi di servizio.

1.3.1 Meccanismo di servizio di tipo D

Il meccanismo di servizio più semplice che si possa immaginare è quello regolare; esso è caratterizzato da una cadenza temporale costante del servizio. Se si suppone quindi che il generico tempo di servizio sia di lunghezza fissa $1/\mu$, allora tale tempo può essere descritto da una variabile aleatoria S degenere la cui funzione di distribuzione è

$$B(t) = P(S < t) = \begin{cases} 0, & t \le 1/\mu \\ 1, & t > 1/\mu. \end{cases} \tag{1.20}$$

Il valore medio, la varianza e il coefficiente di variazione del tempo di servizio sono rispettivamente:

$$E(S) = \frac{1}{\mu}, \qquad \mathrm{Var}(S) = 0, \qquad C(S) = 0. \tag{1.21}$$

Meccanismi di servizio di tipo D si possono presentare, ad esempio, in catene di montaggio in cui il tempo di produzione di un certo pezzo può ritenersi costante.

1.3.2 Meccanismo di servizio di tipo U

Nel meccanismo di servizio di tipo U i tempi di servizio sono indipendenti e identicamente distribuiti con funzione di distribuzione uniforme. Se supponiamo che la variabile aleatoria S sia uniformemente distribuita in (a, b), allora la funzione di distribuzione è

$$B(t) = P(S < t) = \begin{cases} 0, & t \le a \\ \dfrac{t-a}{b-a}, & a < t \le b \\ 1, & t > b \end{cases} \tag{1.22}$$

e quindi la sua densità di probabilità è

$$b(t) = \frac{dB(t)}{dt} = \begin{cases} \dfrac{1}{b-a}, & a < t < b \\ 0, & \text{altrimenti.} \end{cases} \tag{1.23}$$

Il valore medio, la varianza e il coefficiente di variazione del tempo di servizio sono rispettivamente:

$$E(S) = \frac{a+b}{2}, \qquad \mathrm{Var}(S) = \frac{(b-a)^2}{12}, \qquad C(S) = \frac{b-a}{\sqrt{3}\,(a+b)}\cdot \tag{1.24}$$

Si nota che il coefficiente di variazione $C(S)$ è sempre minore dell'unità.
Se si desidera che il tempo medio di servizio sia $1/\mu$ (come nel caso deterministico), basta scegliere $a = 0$ e $b = 2/\mu$. La variabile aleatoria S è allora uniformemente distribuita in $(0, 2/\mu)$ con densità di probabilità:

$$b(t) = \begin{cases} \dfrac{\mu}{2}, & 0 < t < \dfrac{2}{\mu} \\ 0, & \text{altrimenti.} \end{cases} \tag{1.25}$$

1.3.3 Meccanismo di servizio di tipo M

Nel meccanismo di servizio di tipo M i tempi di servizio sono indipendenti e identicamente distribuiti con funzione di distribuzione esponenziale. La lettera M significa Markov a causa della mancanza di memoria della funzione di distribuzione esponenziale. Sia S una variabile aleatoria esponenzialmente distribuita con valore medio $1/\mu$. La sua funzione di distribuzione è

$$B(t) = P(S < t) = \begin{cases} 0, & t \le 0 \\ 1 - e^{-\mu\, t}, & t > 0 \end{cases} \tag{1.26}$$

e quindi la sua densità di probabilità è

$$b(t) = \frac{dB(t)}{dt} = \begin{cases} \mu\, e^{-\mu\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{1.27}$$

Il valore medio, la varianza e il coefficiente di variazione del tempo di servizio sono rispettivamente:

$$E(S) = \frac{1}{\mu}\,, \qquad \mathrm{Var}(S) = \frac{1}{\mu^2}\,, \qquad C(S) = 1. \tag{1.28}$$

Il parametro μ è l'inverso del tempo medio di servizio, ossia del tempo medio necessario per servire un utente, e può essere interpretato come la frequenza media di partenza degli utenti per unità di tempo.

Per la mancanza di memoria della funzione di distribuzione esponenziale, si ha che il *tempo di servizio residuo* ha la stessa distribuzione del tempo di servizio.

La funzione di distribuzione esponenziale gode inoltre di un'altra interessante proprietà: il minimo di k variabili aleatorie $S_1, S_2, \ldots, S_k$ indipendenti e distribuite esponenzialmente con rispettivi valori medi $1/\mu_1, 1/\mu_2, \ldots, 1/\mu_k$ è ancora distribuito esponenzialmente con valore medio $1/(\mu_1 + \mu_2 + \ldots + \mu_k)$. Infatti, se si denota con

$$S = \min(S_1, S_2, \ldots, S_k),$$

allora quando $t > 0$ si ha:

$$\begin{aligned} P(S > t) &= P\{\min(S_1, S_2, \ldots, S_k) > t\} = P\{S_1 > t, S_2 > t, \ldots, S_k > t\} \\ &= P(S_1 > t)\, P(S_2 > t) \cdots P(S_k > t) = e^{-\mu_1 t} e^{-\mu_2 t} \cdots e^{-\mu_k t}, \end{aligned}$$

avendo utilizzato l'indipendenza delle variabili aleatorie e la loro distribuzione esponenziale. Quindi si ha

$$P(S < t) = \begin{cases} 1 - e^{-(\mu_1+\mu_2+\ldots+\mu_k)\, t} & t > 0 \\ \\ 0, & \text{altrimenti,} \end{cases}$$

ossia S è distribuita esponenzialmente con valore medio $1/(\mu_1 + \mu_2 + \ldots + \mu_k)$. Questa proprietà si rivela particolarmente utile quando si considera un sistema di servizio con k servitori identici che lavorano in parallelo i cui tempi di servizio sono distribuiti esponenzialmente con valore medio $1/\mu$. Se tutti i servitori sono occupati, il prossimo utente in fila di attesa per accedere al servizio dovrà attendere il minimo dei tempi residui di servizio dei k servitori. Tale tempo è quindi distribuito esponenzialmente con valore medio $1/(k\,\mu)$.

1.3.4 Meccanismo di servizio di tipo E_k

Consideriamo il sistema di servizio, schematizzato nella Figura 1.9, in cui il centro di servizio consiste di k identiche ed indipendenti fasi. Il tempo di servizio di una generica fase j ($j = 1, 2, \ldots, k$) è descritto da una variabile aleatoria esponenziale di valore medio $1/(k\,\mu)$. Un esempio tipico è quello di un centro di assistenza automobilistico che prevede varie operazioni elementari sulle auto (rifornimento, cambio dell'olio, controllo acqua, ingrassaggio, ...). Sia S una

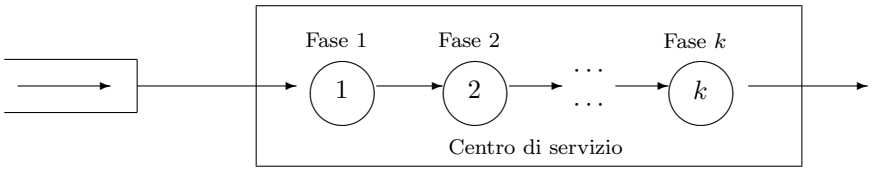

Figura 1.9: Tempi di servizio di tipo E_k nel caso in cui il centro di servizio prevede k successive fasi.

variabile aleatoria che descrive il tempo di servizio di un utente (ossia il tempo misurato dall'istante in cui l'utente entra nella prima fase fino a quando esce dalla k–esima fase). Inoltre sia S_j la variabile aleatoria che descrive il tempo di servizio alla stazione j–esima (ossia il tempo misurato dall'istante in cui l'utente entra nella fase j–esima fino a quando ne esce). Si nota che

$$S = S_1 + S_2 + \ldots + S_k, \tag{1.29}$$

ossia S è la somma di k variabili aleatorie indipendenti, ognuna distribuita esponenzialmente con valore medio $(k\mu)^{-1}$. Pertanto S è caratterizzata da una densità di Erlang di ordine k:

$$b(t) = \begin{cases} \dfrac{(k\,\mu)^k}{(k-1)!}\, e^{-k\,\mu\,t}\; t^{k-1}, & t > 0 \\ 0, & t \leq 0. \end{cases} \tag{1.30}$$

Il valore medio, la varianza e il coefficiente di variazione del tempo di servizio sono rispettivamente:

$$E(S) = \frac{1}{\mu}\,, \qquad \mathrm{Var}(S) = \frac{1}{k\,\mu^2}\,, \qquad C(S) = \frac{1}{\sqrt{k}}\,\cdot \tag{1.31}$$

Se si pone $k = 1$ nella (1.30) si riottiene la densità esponenziale (1.27). Inoltre, quando $k \to +\infty$ si nota che il coefficiente di variazione in (1.31) tende a zero come nel caso deterministico.
Si può pertanto affermare che quando viene applicata una distribuzione di Erlang di ordine k alla durata globale del servizio, ciò equivale ad immaginare il funzionamento del servizio organizzato in k fasi successive, a ciascuna delle quali è adibito un operatore specializzato il cui servizio ha una durata distribuita esponenzialmente; viceversa, quando un servizio viene svolto in k fasi successive ciascuna delle quali caratterizzata da una stessa distribuzione esponenziale, la durata globale del servizio si distribuisce secondo una distribuzione di Erlang di ordine k.

1.3.5 Meccanismo di servizio di tipo H_k

Consideriamo il sistema di servizio, schematizzato nella Figura 1.10, consistente in un centro di servizio costituito da un unico servitore che provvede a fornire k tipi di differenti servizi.

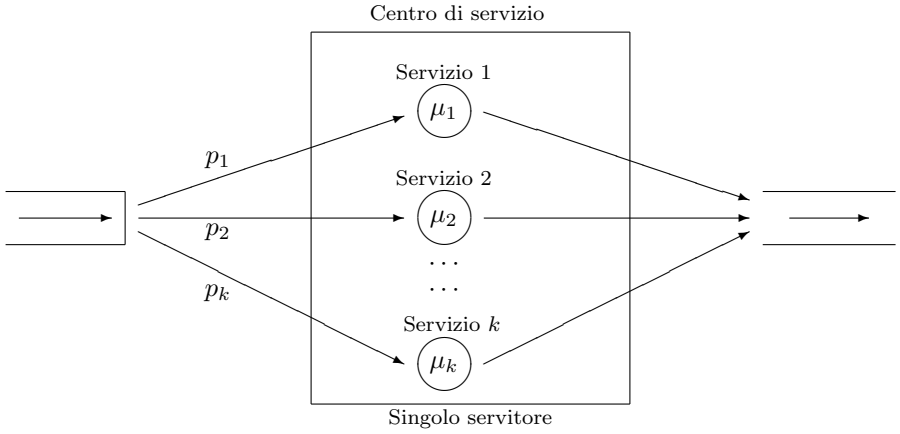

Figura 1.10: Tempi di servizio di tipo H_k per un singolo servitore che offre k differenti servizi.

Supponiamo che la probabilità che l'utente richieda un servizio di tipo j sia p_j per ogni $j = 1, 2, \ldots, k$, dove

$$p_j \geq 0 \qquad (j = 1, 2, \ldots, k), \qquad \sum_{j=1}^{k} p_j = 1. \tag{1.32}$$

Inoltre supponiamo che il j–esimo tipo di servizio ($j = 1, 2, \ldots, k$) sia caratterizzato da una durata del servizio distribuita esponenzialmente con valore medio $1/\mu_j$. Un esempio tipico è quello di un incaricato per l'assistenza al pubblico che fornisce k tipi di informazioni all'ingresso di un grande ufficio.

Denotiamo con S la variabile aleatoria che descrive il tempo di servizio, ossia il tempo necessario per soddisfare un tipo qualsiasi di richiesta fatta dall'utente. Inoltre, denotiamo S_j la variabile aleatoria che descrive il tempo di servizio degli utenti che richiedono il tipo j–esimo di servizio e con B_j l'evento "*è stata scelto dall'utente il j–esimo tipo di servizio*" ($j = 1, 2 \ldots, k$). Si nota immediatamente che l'evento $\{S < t\}$ può essere così scritto

$$\{S < t\} = \bigcup_{j=1}^{k} \Big[B_j \cap \{S < t\}\Big]. \tag{1.33}$$

Infatti, l'evento $\{S < t\}$ si realizza se si verifica uno qualunque dei k eventi incompatibili $B_1, B_2, \ldots, B_k$ ed inoltre $S < t$. Pertanto se $t > 0$, la probabilità della realizzazione dell'evento $\{S < t\}$ è quindi data dalla somma delle probabilità p_j (associata all'evento B_j) per la probabilità dell'evento $\{S_j < t\}$,

ossia

$$B(t) = P(S < t) = \sum_{j=1}^{k} P\Big[B_j \cap \{S < t\}\Big] = \sum_{j=1}^{k} P(B_j)\, P(S < t \mid B_j)$$
$$= \sum_{j=1}^{k} p_j\, P(S_j < t) = \sum_{j=1}^{k} p_j \left[1 - e^{-\mu_j\, t}\right].$$

Segue immediatamente che la funzione di distribuzione del tempo di servizio S è

$$B(t) = \begin{cases} 0, & t \le 0 \\ 1 - \displaystyle\sum_{j=1}^{k} p_j\, e^{-\mu_j\, t}, & t > 0 \end{cases} \tag{1.34}$$

e quindi la sua densità di probabilità è

$$b(t) = \begin{cases} \displaystyle\sum_{j=1}^{k} p_j\, \mu_j\, e^{-\mu_j\, t}, & t > 0 \\ 0, & \text{altrimenti,} \end{cases} \tag{1.35}$$

ossia una densità iperesponenziale di ordine k relativa alla durata del servizio. Ponendo $k = 1$ oppure $\mu_1 = \mu_2 \ldots = \mu_k = \mu$ nella (1.35) si ottiene la densità esponenziale (1.27).
Dalla (1.35) è possibile ricavare immediatamente il valore medio e la varianza del tempo di servizio, ossia

$$\begin{aligned} E(S) &= \sum_{j=1}^{k} \frac{p_j}{\mu_j}\,, \\ \mathrm{Var}(S) &= E(S^2) - [E(S)]^2 = 2 \sum_{j=1}^{k} \frac{p_j}{\mu_j^2} - \left[\sum_{j=1}^{k} \frac{p_j}{\mu_j}\right]^2. \end{aligned} \tag{1.36}$$

Il coefficiente di variazione è quindi:

$$C(S) = \sqrt{\frac{2 \displaystyle\sum_{j=1}^{k} \frac{p_j}{\mu_j^2}}{\left[\displaystyle\sum_{j=1}^{k} \frac{p_j}{\mu_j}\right]^2} - 1}\quad . \tag{1.37}$$

Proposizione 1.1 *$C(S) \ge 1$, con l'uguaglianza se e solo se $\mu_1 = \mu_2 = \ldots = \mu_k$.*

Dimostrazione Osserviamo in primo luogo che dalla seconda delle (1.36) segue

$$\mathrm{Var}(S) - [E(S)]^2 = 2\left[\sum_{j=1}^{k} \frac{p_j}{\mu_j^2} - \left(\sum_{j=1}^{k} \frac{p_j}{\mu_j}\right)^2\right]. \tag{1.38}$$

Vogliamo mostrare che $C(S) \geq 1$, ossia che $\mathrm{Var}(S) - [E(S)]^2 \geq 0$. La disuguaglianza di Cauchy afferma che

$$\left(\sum_{j=1}^{k} x_j\, y_j\right)^2 \leq \left(\sum_{j=1}^{k} x_j^2\right)\left(\sum_{j=1}^{k} y_j^2\right), \tag{1.39}$$

con l'uguaglianza se e solo se esistono due numeri reali a e b non entrambi nulli tali che $a\, x_j + b\, y_j = 0$ $(j = 1, 2, \ldots, k)$. Ponendo $x_j = \sqrt{p_j}$, $y_j = \sqrt{p_j}/\mu_j$ per ogni $j = 1, 2, \ldots, k$, la disuguaglianza di Cauchy diventa

$$\left[\sum_{j=1}^{k} \frac{p_j}{\mu_j}\right]^2 \leq \sum_{j=1}^{k} p_j \sum_{j=1}^{k} \frac{p_j}{\mu_j^2} = \sum_{j=1}^{k} \frac{p_j}{\mu_j^2} \tag{1.40}$$

e l'uguaglianza vale se e solo se esistono due numeri reali a e b non entrambi nulli tali che $a\sqrt{p_j} + b\sqrt{p_j}/\mu_j = 0$ per ogni $j = 1, 2, \ldots, k$, ossia se e solo se $\mu_1 = \mu_2 = \ldots = \mu_k$.
Facendo uso della (1.40) in (1.38) segue che $\mathrm{Var}(S) - [E(S)]^2 \geq 0$ e quindi $C(S) \geq 1$ con l'uguaglianza se e solo se $\mu_1 = \mu_2 = \ldots = \mu_k$. □

Se si desidera che il tempo medio di servizio sia $E(S) = 1/\mu$, basta scegliere $\mu_j = k\, p_j\, \mu$ $(j = 1, 2, \ldots, k)$. La variabile aleatoria S è allora caratterizzata dalla seguente densità di probabilità iperesponenziale di ordine k:

$$b(t) = \begin{cases} k\,\mu \displaystyle\sum_{j=1}^{k} p_j^2\, e^{-k\, p_j \mu\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{1.41}$$

1.3.6 Meccanismo di servizio di tipo G

Nel meccanismo di servizio di tipo G si suppone che i tempi di servizio siano indipendenti e identicamente distribuiti con distribuzione di tipo generale. Occorre ricercare caratteristiche generali del sistema di servizio per una qualsiasi distribuzione dei tempi di servizio. Ovviamente quando si specifica il tipo di distribuzione (D, U, M, E_k, H_k,...) è possibile ottenere maggiori informazioni sull'evoluzione del sistema in esame.

Occorre sottolineare che per una variabile aleatoria S, con $E(S) \neq 0$, il coefficiente di variazione

$$C(S) = \frac{\sqrt{\mathrm{Var}(S)}}{E(S)} \tag{1.42}$$

è un utile parametro per misurare il carattere della distribuzione di probabilità usata. Infatti, se S è deterministica $C(S) = 0$, se S è uniforme in $(0, 2/\mu)$ si ha $C(S) = 1/\sqrt{3}$, se S è esponenziale $C(S) = 1$, se S è caratterizzata da una distribuzione di Erlang di ordine k allora $C(S) = 1/\sqrt{k} \leq 1$ ed infine se S è caratterizzata da distribuzione iperesponenziale di ordine k allora $C(S) \geq 1$.

1.4 Notazione di Kendall

Per descrivere i sistemi di servizio si utilizza una speciale terminologia, introdotta nel 1953 dal matematico e statistico inglese David George Kendall (1918–2007), ossia

$$A/B/s/K/m/Z \tag{1.43}$$

dove

A - descrive la distribuzione dei tempi di interarrivo,

B - la distribuzione dei tempi di servizio per ognuno dei servitori,

s - il numero di servitori (che lavorano in parallelo),

K - la capacità del sistema (ossia il massimo numero di utenti che possono essere presenti nel sistema inclusi quelli in servizio),

m - il numero di potenziali utenti nella sorgente,

Z - la disciplina di servizio.

Spesso si adopera la notazione abbreviata

$$A/B/s, \tag{1.44}$$

intendendo che non ci sono limitazioni alla lunghezza della fila di attesa, la sorgente è infinita e la disciplina di servizio è quella *FIFO*. I simboli scelti da Kendall e tradizionalmente usati per A e B sono

D - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione deterministica,

U - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione uniforme,

M - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione esponenziale,

E_k - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione di Erlang di ordine k,

H_k - tempi di interarrivo (di servizio) *iid* con funzione di distribuzione iperesponenziale di ordine k,

GI - tempi di interarrivo *iid* con funzione di distribuzione generale,

G - tempi di servizio per servitore *iid* con funzione di distribuzione generale.

La notazione di Kendall $D/D/1$ significa che i tempi di interarrivo sono indipendenti e della stessa lunghezza, i tempi di servizio sono anche indipendenti e della stessa lunghezza, esiste un unico servitore, la sorgente è infinita, la capacità del sistema è infinita e la disciplina di servizio è quella *FIFO*. Invece, la notazione di Kendall $M/E_5/4/16/\infty/SIRO$ significa che i tempi di interarrivo sono indipendenti e identicamente distribuiti con legge esponenziale, i tempi di servizio sono indipendenti e identicamente distribuiti con legge di Erlang di ordine 5 per ognuno dei 4 servitori disponibili, la capacità del sistema è 16 (4 in servizio e 12 in fila di attesa), il numero di potenziali utenti nella sorgente è infinito e la disciplina di servizio è "service in random order". Inoltre, la notazione di Kendall $M/H_3/1/10$ mostra che i tempi di interarrivo sono indipendenti e identicamente distribuiti con legge esponenziale, i tempi di servizio sono indipendenti e identicamente distribuiti con legge iperesponenziale di ordine 3 (l'unico servitore offre tre differenti tipi di servizio), la capacità del sistema è 10 (massimo 9 utenti in fila di attesa e uno in servizio), il numero di potenziali utenti della sorgente è infinito e la disciplina di servizio è quella FIFO.

Quando si suppone che un sistema di servizio sia $M/G/s$, con tempi di interarrivo esponenziali e tempi di servizio generali per ognuno degli s servitori, si intende determinare delle relazioni valide per qualsiasi sistema di servizio di questo tipo. Pertanto tali relazioni debbono sussistere anche per il sistema di servizio $M/M/s$. Comunque, se si analizza il sistema di servizio $M/M/s$ si riescono ad ottenere maggiori informazioni rispetto al sistema di servizio generale $M/G/s$.

Occorre sottolineare che per quanto la notazione di Kendall sia molto utilizzata in letteratura, essa non permette di descrivere tutte le situazioni possibili, come ad esempio quello in cui si verificano arrivi degli utenti in gruppo e non singolarmente.

1.5 Esempi di sistemi di servizio con la notazione di Kendall

Vogliamo descrivere graficamente alcuni sistemi di servizio rappresentabili con la notazione di Kendall.

Sistema M/M/1

- tempi di interarrivo esponenziali di valore medio $1/\lambda$;

- tempi di servizio esponenziali di valore medio $1/\mu$;

- unico servitore; capacità infinita della fila di attesa; disciplina FIFO.

Per il sistema di servizio $M/M/1$ siamo interessati ad analizzare in quali condizioni il sistema si congestiona e a determinare i principali parametri prestazionali.

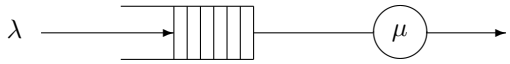

Figura 1.11: Sistema di servizio $M/M/1$.

Sistema M/M/1/1

- tempi di interarrivo esponenziali di valore medio $1/\lambda$;

- tempi di servizio esponenziali di valore medio $1/\mu$;

- unico servitore; capacità nulla della fila di attesa; disciplina FIFO.

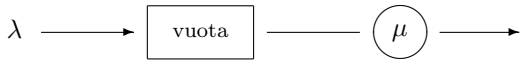

Figura 1.12: Sistema di servizio $M/M/1/1$.

Per il sistema di servizio $M/M/1/1$ siamo interessati ad analizzare la probabilità che un utente in arrivo sia rifiutato e a determinare i principali parametri prestazionali del sistema.

Sistema M/M/1/K

- tempi di interarrivo esponenziali di valore medio $1/\lambda$;

- tempi di servizio esponenziali di valore medio $1/\mu$;

- unico servitore; capacità $K-1$ della fila di attesa; disciplina FIFO.

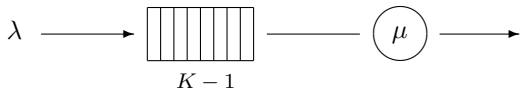

Figura 1.13: Sistema di servizio $M/M/1/K$.

Per il sistema di servizio $M/M/1/K$ siamo interessati ad analizzare la probabilità che un utente in arrivo sia rifiutato e a determinare i principali parametri prestazionali del sistema. Inoltre, desideriamo confrontarlo con il sistema $M/M/1$.

Sistema M/M/2

- tempi di interarrivo esponenziali di valore medio $1/\lambda$;

- tempi di servizio per servitore esponenziali di valore medio $1/\mu$;
- due servitori identici che lavorano in parallelo; capacità infinita della fila di attesa; disciplina FIFO.

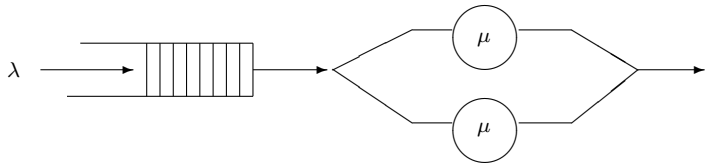

Figura 1.14: Il sistema di servizio $M/M/2$.

Per il sistema di servizio $M/M/2$ siamo interessati ad analizzare in quali condizioni il sistema si congestiona e a determinare i principali parametri prestazionali. Inoltre, desideriamo confrontarlo con alcuni particolari sistemi di tipo $M/M/1$.

Sistema M/M/s

- tempi di interarrivo esponenziali di valore medio $1/\lambda$;
- tempi di servizio per servitore esponenziali di valore medio $1/\mu$;
- s servitori identici che lavorano in parallelo; capacità infinita della fila di attesa; disciplina FIFO.

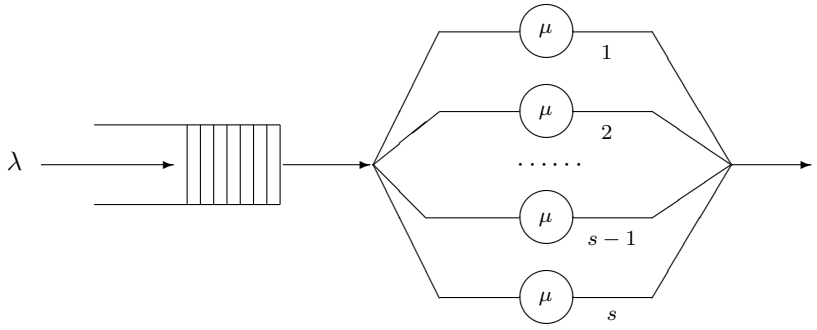

Figura 1.15: Sistema di servizio $M/M/s$.

Per il sistema di servizio $M/M/s$ siamo interessati ad analizzare in quali condizioni il sistema si congestiona e a determinare i principali parametri prestazionali. Inoltre, desideriamo stabilire il numero di servitori necessari e sufficienti per garantire l'efficienza del sistema.

Sistema M/M/s/s

- tempi di interarrivo esponenziali di valore medio $1/\lambda$;

- tempi di servizio per servitore esponenziali di valore medio $1/\mu$;

- s servitori identici che lavorano in parallelo; capacità nulla della fila di attesa; disciplina FIFO.

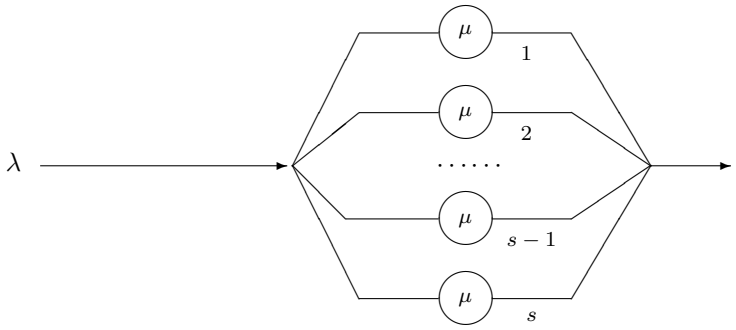

Figura 1.16: Sistema di servizio $M/M/s/s$.

Per il sistema di servizio $M/M/s/s$ siamo interessati ad analizzare la probabilità che un utente in arrivo sia rifiutato e a determinare i principali parametri prestazionali del sistema.

Sistema M/E_3/1

- tempi di interarrivo esponenziali;

- tempi di servizio di Erlang di ordine 3;

- un centro di servizio che prevede tre fasi successive; capacità infinita della fila di attesa; disciplina FIFO.

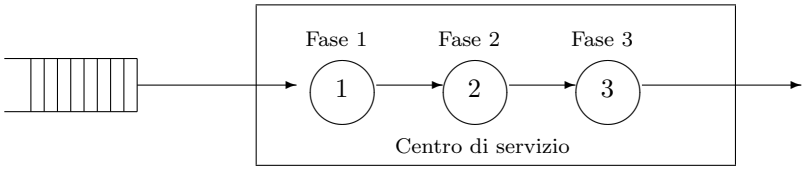

Figura 1.17: Sistema $M/E_3/1$ con centro di servizio che prevede 3 fasi successive.

Per il sistema di servizio $M/E_k/1$ siamo interessati ad analizzare in quali condizioni il sistema si congestiona. Inoltre, desideriamo confrontare i suoi parametri prestazionali con quelli di altri sistemi caratterizzati da differenti meccanismi di servizio.

Sistema M/H_3/1

- tempi di interarrivo esponenziali;
- tempi di servizio iperesponenziale di ordine 3;
- singolo servitore che offre tre differenti servizi; capacità infinita della fila di attesa; disciplina FIFO.

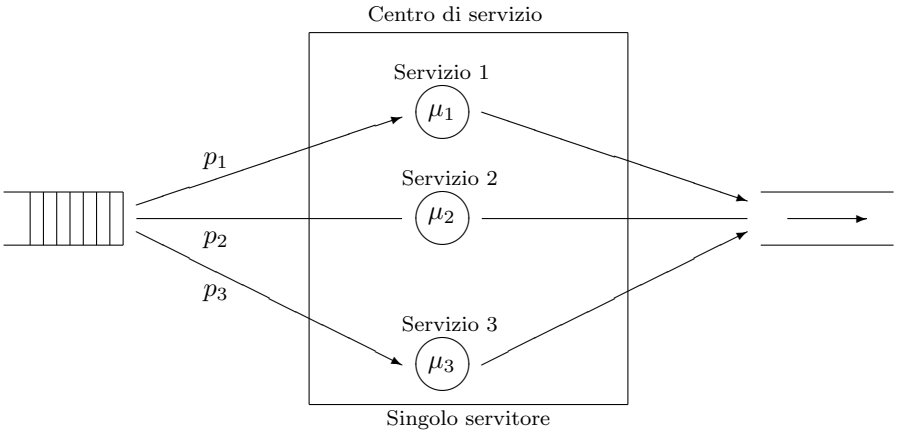

Figura 1.18: Sistema $M/H_3/1$ con un singolo servitore che offre 3 differenti servizi.

Per il sistema di servizio $M/H_k/1$ siamo interessati ad analizzare in quali condizioni il sistema si congestiona. Inoltre, desideriamo confrontare i suoi parametri prestazionali con quelli di altri sistemi caratterizzati da differenti meccanismi di servizio.

Sistema M/E3/2/16/∞/SIRO

- tempi di interarrivo esponenziali;
- tempi di servizio di Erlang di ordine 3 per servitore;
- due centri di servizio che prevedono tre fasi successive; capacità finita della fila di attesa (14 utenti); disciplina SIRO.

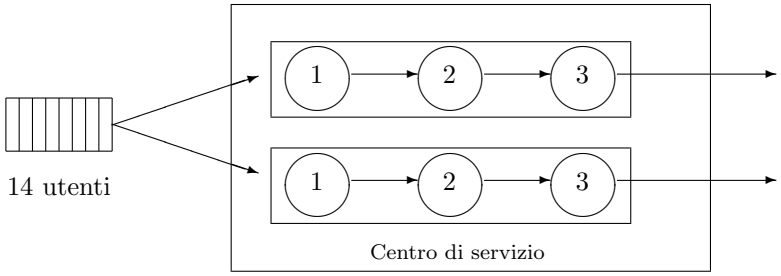

Figura 1.19: Sistema $M/E3/2/16/\infty/SIRO$ con centro di servizio che prevede 3 fasi successive per ognuno dei due servitori.

Sistema $M/H_3/1/10$

- tempi di interarrivo esponenziali;

- tempi di servizio iperesponenziale di ordine 3;

- singolo servitore che offre tre differenti servizi; capacità finita della fila di attesa; disciplina FIFO.

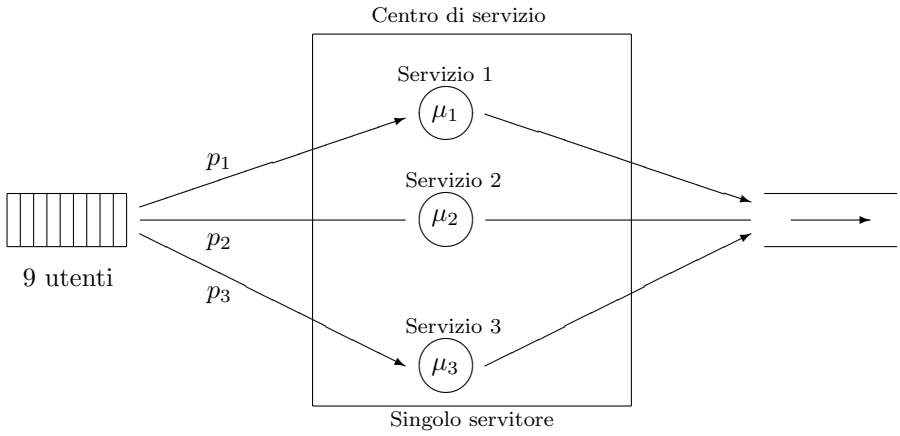

Figura 1.20: Sistema $M/H_3/1/10$ con un singolo servitore che offre 3 differenti servizi.

Desideriamo successivamente individuare opportuni algoritmi che ci consentano di simulare al computer tali sistemi di servizio generando opportune sequenze di tempi di interarrivo e di servizio e calcolando i principali parametri prestazionali del sistema.

Capitolo 2

Analisi del sistema

2.1 Introduzione

In questo capitolo introdurremo i parametri prestazionali di maggiore interesse nell'analisi dei sistemi di servizio.

Il numero di utenti presenti in un sistema di servizio può essere descritto da un processo stocastico $\{N(t), t \geq 0\}$ discreto nello spazio degli stati e continuo nel tempo. Per ogni fissato istante di tempo t, $N(t)$ è una variabile aleatoria descrivente il numero di utenti presenti nel sistema al tempo t. Tale variabile aleatoria assume valori in un insieme finito $\{0, 1, \ldots, K\}$ se la capacità del sistema di servizio è K, mentre assume valori nell'insieme numerabile $\{0, 1, \ldots\}$ se la capacità del sistema di servizio è infinita.

2.2 Alcune misure prestazionali

L'indagine effettuata tramite la teoria delle file di attesa permette di descrivere:

- **lo stato del sistema** Se si denota con $\{N(t),\, t \geq 0\}$ il processo stocastico che descrive il numero $N(t)$ di utenti presenti nel sistema al tempo t occorre, se possibile, determinare

$$p_n(t) = P\{N(t) = n\} \qquad (n = 0, 1, \ldots), \tag{2.1}$$

  ossia la probabilità che siano presenti n utenti nel sistema al tempo t ed alcune caratteristiche quali il valore medio $E[N(t)]$ e la varianza $Var[N(t)]$ del numero di utenti presenti nel sistema ad ogni fissato istante di tempo t. Occorre inoltre stabilire se il sistema raggiunge una *situazione di equilibrio statistico* ed in tal caso occorre calcolare la distribuzione di equilibrio

$$q_n = \lim_{t \to +\infty} p_n(t) \qquad (n = 0, 1, \ldots) \tag{2.2}$$

ed alcune caratteristiche quali il valore medio e la varianza del numero di utenti presenti nel sistema nella situazione di equilibrio statistico.

- **il tempo di permanenza nella fila di attesa di un utente** Il tempo di permanenza nella fila di attesa di un utente è una variabile aleatoria che descrive il tempo che un utente deve attendere in fila di attesa prima di essere servito.

- **il tempo di attesa di un utente nel sistema** Il tempo di attesa di un utente è una variabile aleatoria che descrive il tempo che un utente spende nel sistema, ossia il tempo che un utente deve attendere in fila di attesa più il suo tempo di servizio.

- **il periodo di occupazione** Il periodo di occupazione è una variabile aleatoria che descrive la lunghezza dell'intervallo di tempo che inizia con l'arrivo di un utente che trova l'unico servitore libero e continua fino a quando il servitore è per la prima volta nuovamente libero. Nel caso di più servitori il periodo di occupazione descrive la lunghezza l'intervallo di tempo che inizia con l'arrivo di un utente che trova tutti i servitori liberi e continua fino a che tutti i servitori sono per la prima volta nuovamente liberi. Il periodo di occupazione (*busy period*) quindi descrive la lunghezza dell'intervallo di tempo in cui il centro di servizio non è disponibile.

- **il tempo di ozio** Il tempo di ozio (*idle period*), detto anche periodo di soggiorno nello stato 0, è una variabile aleatoria che descrive la lunghezza dell'intervallo di tempo in cui il centro di servizio è inutilizzato.

Una tipica *realizzazione* di un sistema di servizio è una funzione a gradini, ossia una funzione costante a tratti con salti diretti verso il basso o verso l'alto ogni volta che accade un evento. Nella Figura 2.1 è rappresentata una realizzazione di un sistema di servizio con singolo servitore e disciplina FIFO e sono indicati gli istanti di arrivo $a_1, a_2, \ldots$ degli utenti, gli istanti di partenza $u_1, u_2, \ldots$ degli utenti, lo stato dei sistema $N(t)$ ai vari istanti di tempo t, i tempi di attesa nel sistema $W_1, W_2, \ldots$ degli utenti, i tempi di interarrivo $T_1, T_2, \ldots$ degli utenti, i tempi di servizio $S_1, S_2, \ldots$ per servitore degli utenti, i periodi di occupazione $B_1, B_2, \ldots$ del centro di servizio e i tempi di ozio $O_1, O_2, \ldots$ del centro di servizio.

Le notazioni fondamentali utilizzate nella teoria delle file di attesa sono le seguenti:

$N(t)$ - variabile aleatoria che descrive il numero di utenti presenti nel sistema (inclusi quelli in servizio) al tempo t;

$p_n(t)$ - probabilità che al tempo t siano presenti nel sistema n utenti (inclusi quelli in servizio);

N - variabile aleatoria che descrive il numero di utenti presenti nel sistema (inclusi quelli in servizio) nella situazione di equilibrio statistico;

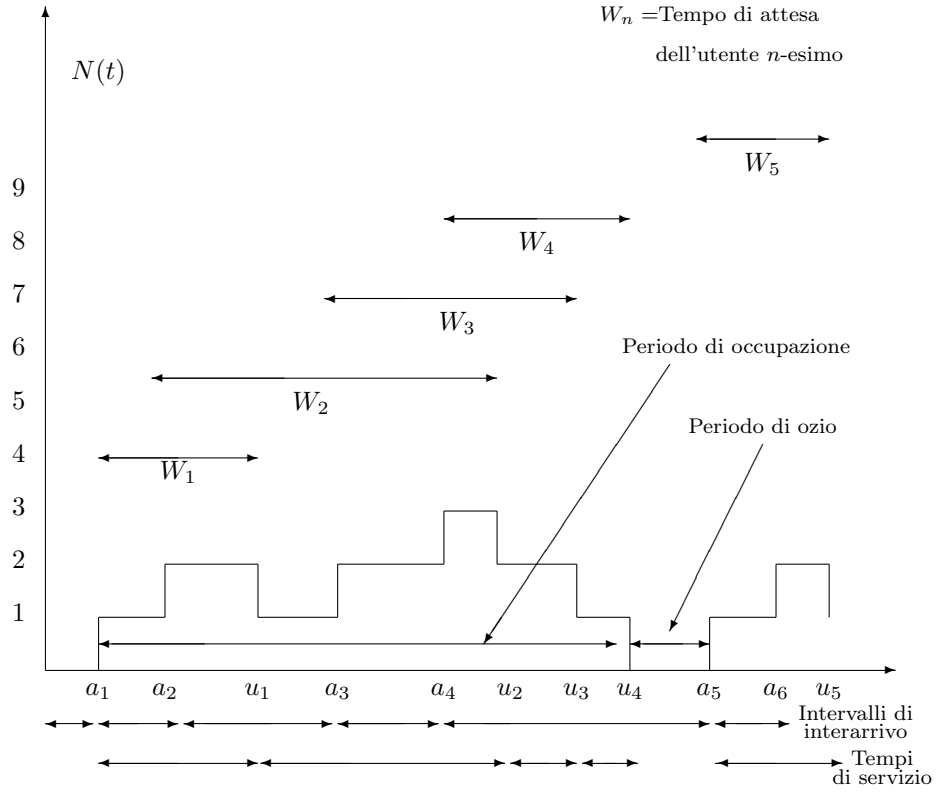

Figura 2.1: Una tipica realizzazione di un sistema di servizio.

q_n - probabilità che siano presenti nel sistema n utenti (inclusi quelli in servizio) nella situazione di equilibrio statistico;

$N_q(t)$ - variabile aleatoria che descrive il numero di utenti presenti nella fila di attesa al tempo t;

N_q - variabile aleatoria che descrive il numero di utenti presenti nella fila di attesa nella situazione di equilibrio statistico;

$N_s(t)$ - variabile aleatoria che descrive il numero di utenti in servizio al tempo t;

N_s - variabile aleatoria che descrive il numero di utenti in servizio nella situazione di equilibrio statistico;

T - variabile aleatoria che descrive il generico tempo di interarrivo nell'ipotesi in cui i tempi di interarrivo siano *iid*;

S - variabile aleatoria che descrive il generico tempo necessario ad un servitore per servire un utente nell'ipotesi in cui i tempi di servizio siano *iid*;

W - variabile aleatoria che descrive il tempo di attesa di un utente nel sistema incluso il suo tempo di servizio;

Q - variabile aleatoria che descrive il tempo che un utente spende nella fila di attesa prima di essere servito;

B - variabile aleatoria che descrive il periodo di occupazione del centro di servizio, ossia il periodo di tempo in cui esiste almeno un utente e quindi il centro di servizio è occupato da almeno un utente.

I - variabile aleatoria che descrive il tempo di ozio del centro di servizio, ossia il periodo di tempo in cui non ci sono utenti nel sistema ed il centro di servizio è inoperoso.

Un sistema di servizio alterna sempre periodi di occupazione e periodi di ozio. Come si evince dalla Figura 2.1 un periodo di ozio si presenta quando l'istante di partenza u_i dell'utente i–esimo è immediatamente seguito dall'istante a_{i+1} di arrivo dell'utente $(i+1)$–esimo creando il periodo di ozio (u_i, a_{i+1}). La Figura 2.2 mostra che i periodi di ozio possono essere riguardati come *tempi residui degli intervalli di interarrivo*.

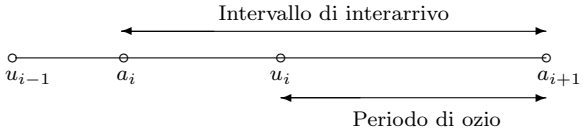

Figura 2.2: Un periodo di ozio del centro di servizio.

La Figura 2.3 mostra che sussiste la seguente relazione:

$$N(t) = N_q(t) + N_s(t) \qquad (t \geq 0), \tag{2.3}$$

ossia il numero di utenti presenti al tempo t è uguale alla somma del numero di utenti presenti nella fila di attesa al tempo t e del numero di utenti in servizio nello stesso istante di tempo. Nella situazione di equilibrio statistico, se esiste, si ha quindi:

$$N = N_q + N_s. \tag{2.4}$$

Sussiste inoltre la seguente relazione:

$$W = Q + S, \tag{2.5}$$

ossia il tempo di attesa di un utente nel sistema è uguale alla somma del suo tempo di attesa in coda Q e del suo tempo di servizio S.

Come suggerisce la notazione di Kendall, per valutare le misure prestazionali di un sistema di servizio occorre assumere che siano note alcune proprietà del

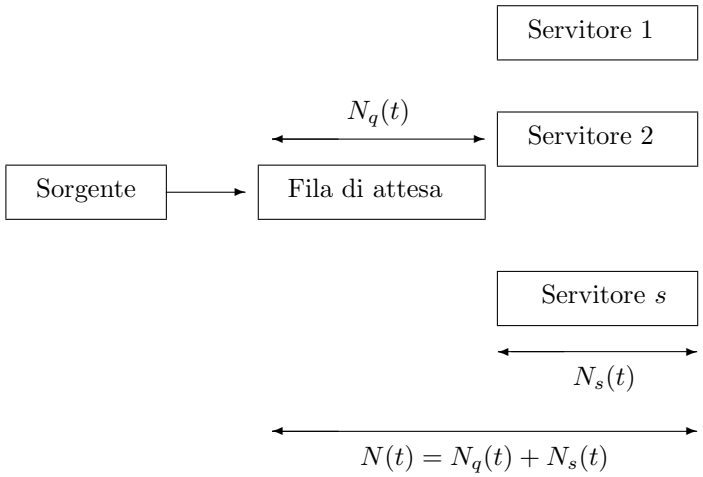

Figura 2.3: Numero di utenti presenti in un sistema di servizio al tempo t.

sistema stesso. Spesso si assume che siano note le distribuzioni dei tempi di interarrivo e dei tempi di servizio per ognuno dei servitori.

Se si denota con

λ^* : *frequenza media di arrivo per unità di tempo*

μ^* : *frequenza media di partenza da un generico servitore per unità di tempo*,

una misura prestazionale fondamentale in un sistema di servizio è rappresentata dall'*intensità di traffico*, così definita

$$a = \frac{\lambda^*}{\mu^*}. \tag{2.6}$$

Tale coefficiente rappresenta *l'intensità del lavoro che svolge il sistema di servizio nella situazione di equilibrio statistico.*

Il rapporto tra la frequenza media degli arrivi e la frequenza totale delle partenze che il sistema di servizio può realizzare lavorando a pieno regime, ossia il rapporto tra l'intensità di traffico ed il numero di servitori presenti nel centro di servizio

$$\varrho^* = \frac{\lambda^*}{s\,\mu^*} \tag{2.7}$$

prende il nome di *fattore di utilizzazione del sistema.* Tale coefficiente rappresenta *l'intensità del lavoro di ognuno dei servitori nella situazione di equilibrio statistico.*

Quando più ϱ^* si avvicina all'unità tanto più il sistema tende ad avere tutti i posti di lavoro occupati con il rischio di entrare in congestione permanente ($\varrho^* = 1$). Quando $\varrho^* \geq 1$, ossia quando nell'unità di tempo la frequenza media

degli arrivi è maggiore o uguale della frequenza media delle partenze, il sistema con capacità infinita non raggiunge una situazione di equilibrio statistico e la lunghezza della fila di attesa tende ad aumentare indefinitamente. Quindi, in un sistema di servizio a capacità infinita, il parametro $\varrho^* = \lambda^*/(s\mu^*)$ fornisce una *misura di congestione* del sistema: deve essere inferiore all'unità affinché il sistema con s servitori in parallelo non si congestioni.

2.3 Leggi di Little

Nella teoria delle file di attesa esistono delle relazioni che valgono sotto condizioni abbastanza generali. Tra queste relazioni rivestono un ruolo fondamentale le *leggi di Little*. Esse si applicano ad un qualsiasi *sistema di servizio in condizioni di equilibrio statistico*.

Se si denota con λ^* la frequenza media di arrivo nel sistema per unità di tempo, con $E(N)$ il numero medio di utenti nel sistema e con $E(W)$ il tempo medio di attesa di un utente nel sistema, la *prima legge di Little* afferma che

$$E(N) = \lambda^* E(W), \tag{2.8}$$

ossia il numero medio di utenti nel sistema è uguale al prodotto della frequenza media di arrivo nel sistema per unità di tempo e del tempo medio di attesa di un utente nel sistema.
La *seconda legge di Little* si applica alla fila di attesa e afferma che

$$E(N_q) = \lambda^* E(Q), \tag{2.9}$$

ossia il numero medio di utenti nella fila di attesa è uguale al prodotto della frequenza media di arrivo nel sistema per unità di tempo e del tempo medio di permanenza di un utente nel centro di attesa.
La legge di Little può essere formalizzata anche per il centro di servizio. Infatti, sottraendo membro a membro i termini delle relazioni (2.8) e (2.9), si ottiene

$$E(N) - E(N_q) = \lambda^* [E(W) - E(Q)].$$

Essendo $E(N) = E(N_q) + E(N_s)$ e $E(W) = E(Q) + E(S)$, la *terza legge di Little* afferma che

$$E(N_s) = \lambda^* E(S), \tag{2.10}$$

ossia il numero medio di utenti in servizio è uguale al prodotto della frequenza media di arrivo e del tempo medio di servizio.
Le tre leggi di Little rivestono una notevole importanza nelle teoria delle file di attesa poiché esse *non dipendono dalla distribuzione dei tempi di interarrivo e dei tempi di servizio, dal numero di servitori nel sistema e dalla disciplina di servizio*.

Una dimostrazione rigorosa delle relazioni (2.8) e (2.9) è stata fornita nel 1961 dal fisico statunitense John D. C. Little, e perciò tali relazioni sono note

come *formule di Little*[1][2] Little è stato docente del MIT Sloan School of Management ed è considerato il fondatore della scienza del marketing. La dimostrazione rigorosa di queste relazioni si rivela complessa dal punto di vista matematico; esistono comunque varie dimostrazioni di tipo euristico molto più semplici.

Una dimostrazione più semplice, che descriveremo nel seguito, delle relazioni (2.8) e (2.9) è stata fornita da nel 1969 da Eilon [3]. Eilon è stato docente di Management Science all'Imperial College of Science and Technology dell'Università di Londra. Essa non dipende *(i)* dalla distribuzione dei tempi di interarrivo e dei tempi di servizio, *(ii)* dal numero di servitori nel sistema e *(iii)* dalla disciplina di servizio.

2.3.1 Formula di Little per l'intero sistema

Dimostriamo la prima formula di Little, ossia la (2.8).

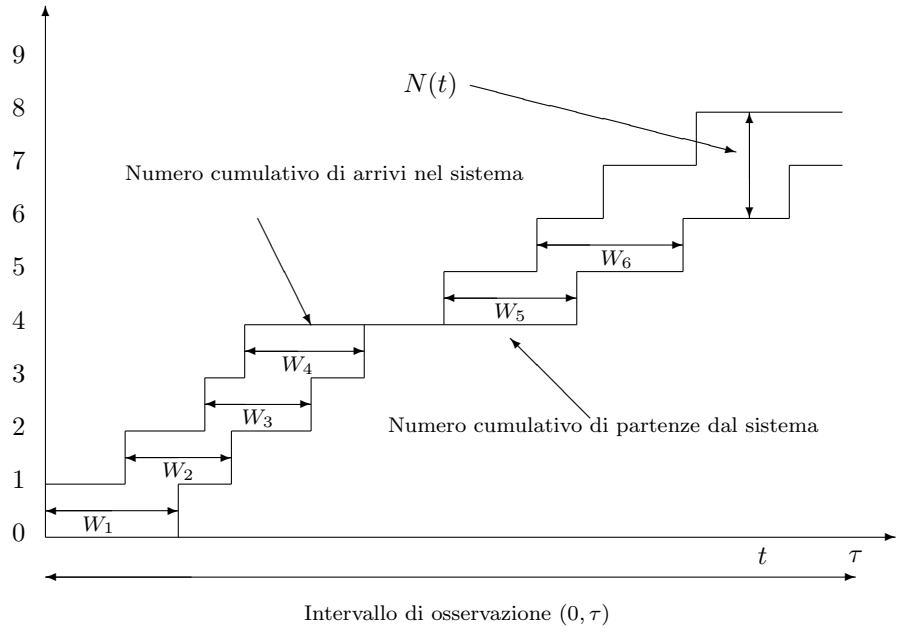

Figura 2.4: Numero cumulativo di arrivi e numero cumulativo di partenze dal sistema di servizio.

[1]Little, J.D.C. "A Proof for the Queuing Formula: $L = \lambda W$", Operations Research. 9 (3): 383–387 (1961).

[2]Little, J.D.C. "Little's Law as Viewed on Its 50th Anniversary", Operations Research 59 (3): 536–549 (2011).

[3]Eilon, Samuel, "A Simpler Proof of $L = \lambda W$", Operations Research. 17 (5): 915–917. (1969).

Nella Figura 2.4 la linea superiore descrive il *numero cumulativo di arrivi* e la linea inferiore il *numero cumulativo di partenze dal sistema.* La distanza verticale tra due linee fornisce il numero di utenti $N(t)$ presenti nel sistema ad un fissato istante t, mentre la distanza orizzontale denota il tempo di attesa nel sistema (tempo di permanenza in fila di attesa più tempo di servizio). Supponiamo che il sistema sia stato in funzione per un certo tempo e che successivamente abbia raggiunto una situazione di equilibrio statistico. Consideriamo un intervallo di tempo $(0, \tau)$ che può includere nessuno, uno o più periodi di occupazione. Denotiamo con

- $N_a(\tau)$ numero totale di arrivi durante l'intervallo $(0, \tau)$
- $\overline{\lambda}(\tau)$ frequenza media di arrivo per unità di tempo nell'intervallo $(0, \tau)$. Si nota che
$$\overline{\lambda}(\tau) = \frac{N_a(\tau)}{\tau}\,, \tag{2.11}$$
ossia frequenza media di arrivo per unità di tempo nell'intervallo $(0, \tau)$ è data dal rapporto tra il numero totale di arrivi durante l'intervallo $(0, \tau)$ e la lunghezza dell'intervallo.

  Siano inoltre

- $W_c(\tau)$ tempo totale di attesa (tempo cumulativo di attesa) nel sistema di tutti gli utenti che arrivano nell'intervallo $(0, \tau)$.

  È evidente che
$$W_c(\tau) = W_1 + W_2 + \ldots,$$
ossia il tempo totale di attesa nel sistema è la somma dei tempi di attesa dei vari utenti che sono arrivati nell'intervallo $(0, \tau)$. Osservando la Figura 2.4 si nota anche che
$$W_c(\tau) = 1 \times W_1 + 1 \times W_2 + \ldots = \int_0^\tau N(t)\; dt,$$
ossia $W_c(\tau)$ descrive l'area compresa tra le due linee nell'intervallo $(0, \tau)$. Indichiamo inoltre con

- $\overline{W}(\tau)$ media dei tempi di attesa nel sistema degli utenti arrivati durante l'intervallo $(0, \tau)$. Risulta che
$$\overline{W}(\tau) = \frac{W_c(\tau)}{N_a(\tau)} = \frac{1}{N_a(\tau)} \int_0^\tau N(t)\; dt, \tag{2.12}$$
ossia la media del tempo di attesa nel sistema degli utenti arrivati durante l'intervallo $(0, \tau)$ è uguale al rapporto tra il tempo totale di attesa nel sistema di tutti gli utenti che sono arrivati nell'intervallo $(0, \tau)$ ed il numero totale di arrivi in tale intervallo. Denotiamo infine con

- $\overline{N}(\tau)$ media per unità di tempo del numero di utenti nel sistema nell'intervallo $(0, \tau)$.

Si nota che

$$\overline{N}(\tau) = \frac{1}{\tau} \int_0^\tau N(t)\, dt = \frac{W_c(\tau)}{\tau}\,, \tag{2.13}$$

ossia la media per unità di tempo del numero di utenti nel sistema è uguale al rapporto tra tempo totale di attesa nel sistema di tutti i utenti che arrivano nell'intervallo $(0, \tau)$ e la lunghezza di tale intervallo.
Dalle relazioni (2.11), (2.12) e (2.13) segue che

$$\overline{N}(\tau) = \frac{W_c(\tau)}{\tau} = \frac{W_c(\tau)}{N_a(\tau)}\, \frac{N_a(\tau)}{\tau} = \overline{W}(\tau)\, \overline{\lambda}(\tau),$$

ossia

$$\overline{N}(\tau) = \overline{\lambda}(\tau)\, \overline{W}(\tau). \tag{2.14}$$

Supponiamo che quando $\tau \to +\infty$ esistano finiti i limiti di $\overline{\lambda}(\tau)$ e di $\overline{W}(\tau)$:

$$\lambda^* = \lim_{\tau \to +\infty} \overline{\lambda}(\tau)\,, \qquad E(W) = \lim_{\tau \to +\infty} \overline{W}(\tau). \tag{2.15}$$

In tali ipotesi, dalla (2.14) segue che esiste finito anche il limite di $\overline{N}(\tau)$ quando $\tau \to +\infty$ e risulta

$$E(N) = \lim_{\tau \to +\infty} \overline{N}(\tau).$$

La prima formula di Little, ossia la (2.8), segue quindi immediatamente dalla (2.14) procedendo al limite per $\tau \to +\infty$.

2.3.2 Formula di Little per la fila di attesa

Dimostriamo ora la seconda formula di Little per la fila di attesa, ossia la (2.9).

Nella Figura 2.5 la linea superiore descrive il *numero cumulativo di arrivi* e la linea inferiore il *numero cumulativo di partenze dalla fila di attesa*. La distanza verticale tra due linee fornisce il numero di utenti $N_Q(t)$ presenti nella fila di attesa ad un fissato istante t, mentre la distanza orizzontale denota il tempo di permanenza nella fila di attesa. Supponiamo che il sistema sia stato in funzione per un certo tempo e che successivamente abbia raggiunto una situazione di equilibrio statistico. Consideriamo un intervallo di tempo $(0, \tau)$ che può includere nessuno o più periodi di occupazione. Denotiamo nuovamente con $N_a(\tau)$ il numero totale di arrivi durante l'intervallo $(0, \tau)$ e con $\overline{\lambda}(\tau)$ la frequenza media di arrivo per unità di tempo nell'intervallo $(0, \tau)$. Risulta nuovamente essere valida la (2.11). Indichiamo inoltre con

- $Q_c(\tau)$ tempo totale di permanenza nella fila di attesa di tutti gli utenti che arrivano nell'intervallo $(0, \tau)$;

  Si nota che

$$Q_c(\tau) = Q_1 + Q_2 + \ldots,$$

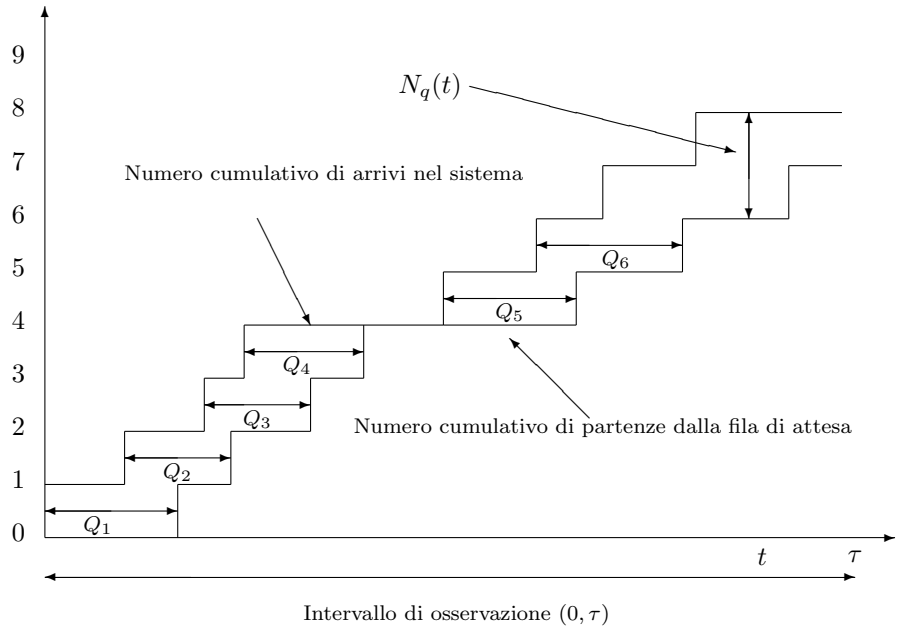

Figura 2.5: Numero cumulativo di arrivi e numero cumulativo di partenze dalla fila di attesa.

ossia il tempo totale di permanenza nella fila di attesa è la somma dei tempi di permanenza nella fila di attesa dei vari utenti che sono arrivati nell'intervallo $(0, \tau)$. Osservando la figura si nota che

$$Q_c(\tau) = 1 \times Q_1 + 1 \times Q_2 + \ldots = \int_0^\tau N_q(t)\, dt$$

ossia $Q_c(\tau)$ descrive l'area compresa tra le due linee nell'intervallo $(0, \tau)$. Indichiamo inoltre con

- $\overline{Q}(\tau)$ la media dei tempi di permanenza nella fila di attesa degli utenti arrivati durante l'intervallo $(0, \tau)$. Si nota che

$$\overline{Q}(\tau) = \frac{Q_c(\tau)}{N_a(\tau)}\,, \tag{2.16}$$

ossia la media del tempo di permanenza nel sistema degli utenti arrivati durante l'intervallo $(0, \tau)$ è uguale al rapporto tra il tempo totale di permanenza nella fila di attesa di tutti gli utenti che sono arrivati nell'intervallo $(0, \tau)$ ed il numero totale di arrivi in tale intervallo. Denotiamo infine con

- $\overline{N}_q(\tau)$ la media per unità di tempo del numero di utenti nella fila di attesa nell'intervallo $(0, \tau)$.

È evidente che

$$\overline{N}_q(\tau) = \frac{1}{\tau} \int_0^\tau N_q(t)\, dt = \frac{Q_c(\tau)}{\tau}\,, \tag{2.17}$$

ossia la media per unità di tempo del numero di utenti nella fila di attesa è uguale al rapporto tra tempo totale di permanenza nella fila di attesa di tutti gli utenti che arrivano nell'intervallo $(0, \tau)$ e la lunghezza di tale intervallo. Dalle relazioni (2.11), (2.16) e (2.17) segue che

$$\overline{N}_q(\tau) = \frac{Q_c(\tau)}{\tau} = \frac{Q_c(\tau)}{N_a(\tau)}\, \frac{N_a(\tau)}{\tau} = \overline{Q}(\tau)\, \overline{\lambda}(\tau),$$

ossia

$$\overline{N}_q(\tau) = \overline{\lambda}(\tau)\, \overline{Q}(\tau). \tag{2.18}$$

Supponiamo che quando $\tau \to +\infty$ esistano finiti i limiti di $\overline{\lambda}(\tau)$ e di $\overline{Q}(\tau)$:

$$\lambda^* = \lim_{\tau \to +\infty} \overline{\lambda}(\tau)\,, \qquad E(Q) = \lim_{\tau \to +\infty} \overline{Q}(\tau). \tag{2.19}$$

In tali ipotesi, dalla (2.18) segue che esiste finito anche il limite di $\overline{N}_q(\tau)$ quando $\tau \to +\infty$ e risulta

$$E(N_q) = \lim_{\tau \to +\infty} \overline{N}(\tau).$$

La seconda formula di Little, ossia la (2.9), segue quindi immediatamente dalla (2.18) procedendo al limite per $\tau \to +\infty$.

2.4 Periodi di occupazione e di ozio

Consideriamo un sistema di servizio con un unico servitore. Tale sistema alterna periodi di ozio (quando non ci sono utenti nel sistema e quindi il servitore è libero) e periodi di occupazione (quando esiste almeno un utente nel sistema ed il servitore è occupato).

Denotiamo con $I_1, I_2, \ldots$ e $B_1, B_2, \ldots$ rispettivamente le lunghezze dei periodi di ozio e dei periodi di occupazione. Nella situazione di equilibrio statistico la probabilità che il sistema sia asintoticamente vuoto può essere così calcolata:

$$q_0 = \lim_{n \to \infty} \frac{I_1 + I_2 + \ldots + I_n}{I_1 + I_2 + \ldots + I_n + B_1 + B_2 + \ldots + B_n}\,. \tag{2.20}$$

Se le successioni $I_1, I_2, \ldots$ e $B_1, B_2, \ldots$ sono entrambe costituite da variabili aleatorie indipendenti e identicamente distribuite, allora dividendo il numeratore ed il denominatore del rapporto in (2.20) per n ed applicando la legge dei grandi numeri si ha

$$\begin{aligned} q_0 &= \lim_{n \to \infty} \frac{(I_1 + I_2 + \ldots + I_n)/n}{(I_1 + I_2 + \ldots + I_n)/n + (B_1 + B_2 + \ldots + B_n)/n} \\ &= \frac{E(I)}{E(I) + E(B)}\,, \end{aligned} \tag{2.21}$$

ossia la probabilità che il sistema sia asintoticamente vuoto è uguale al rapporto tra il tempo medio di ozio e la somma del tempo medio di ozio e del tempo medio di occupazione del servitore. Dalla (2.21) si ricava:

$$1 - q_0 = \frac{E(B)}{E(I) + E(B)}, \tag{2.22}$$

ossia la probabilità che nel sistema sia presente almeno un utente è uguale al rapporto tra il tempo medio di occupazione e la somma del tempo medio di ozio e del tempo medio di occupazione del servitore. La (2.22) può anche essere così scritta

$$E(B) = \frac{(1 - q_0)\, E(I)}{q_0}\,, \tag{2.23}$$

Come mostra la Figura 2.2 i periodi di ozio del servitore possono essere riguardati come tempi residui dei tempi di interarrivo.

Se si ipotizza che i tempi di interarrivo sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$, allora ricordando la proprietà di mancanza di memoria della distribuzione esponenziale si ricava che anche i tempi residui sono caratterizzati dalla stessa distribuzione esponenziale. In tale ipotesi la densità di probabilità del periodo di ozio è quindi:

$$f_I(t) = \begin{cases} \lambda\, e^{-\lambda\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases} \tag{2.24}$$

Essendo $E(I) = 1/\lambda$, dalla relazione (2.23) si ottiene $E(B) = (1 - q_0)/(\lambda\, q_0)$.

In conclusione, nella situazione di equilibrio statistico, se i tempi di interarrivo sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$, allora il tempo medio di ozio e il tempo medio di occupazione del servitore sono rispettivamente:

$$E(I) = \frac{1}{\lambda}, \qquad E(B) = \frac{1 - q_0}{\lambda\, q_0}. \tag{2.25}$$

Capitolo 3

Processi di nascita morte

3.1 Introduzione

In questo capitolo siamo interessati ad introdurre alcuni processi stocastici discreti nello spazio degli stati e continui nel tempo, ossia il processo di Poisson e i processi di nascita morte. Per i processi di nascita morte individueremo le condizioni affinché si raggiunga l'equilibrio statistico e determineremo la distribuzione di equilibrio. I processi di nascita morte saranno utilizzati nei prossimi due capitoli per analizzare alcuni sistemi di servizio determinando i loro principali parametri prestazionali.

Sia $\{N(t),\ t \geq 0\}$ un processo stocastico continuo nel tempo e discreto nello spazio degli stati. Per ogni fissato $t \geq 0$, $N(t)$ è una variabile aleatoria discreta che assume un numero finito o al più numerabile di valori. Le realizzazioni di tale processo sono funzioni a gradino, ossia funzioni costanti a tratti con salti diretti verso il basso o verso l'alto ogni volta che si verifica un cambiamento di stato. Le situazioni fisiche da cui tali processi sorgono sono quelle in cui lo stato del sistema è caratterizzato da un numero intero di particelle (o individui, utenti) e i cambiamenti di stato rappresentano l'addizione o la sottrazione di particelle dal sistema in vario modo: nascite, morti, immigrazioni, emigrazioni,....

3.2 Processo stocastico di Poisson

Il più semplice processo stocastico continuo nel tempo e discreto nello spazio degli stati è il *processo di Poisson*. Tale processo si rivela utile nella descrizione di alcuni fenomeni di conteggio che evolvono nel tempo quali il numero di arrivi ad un sistema di servizio, il numero di chiamate ad un centralino telefonico, ... Una tipica realizzazione del processo di Poisson che descrive il numero di arrivi ad un centralino telefonico è illustrata in Figura 3.1

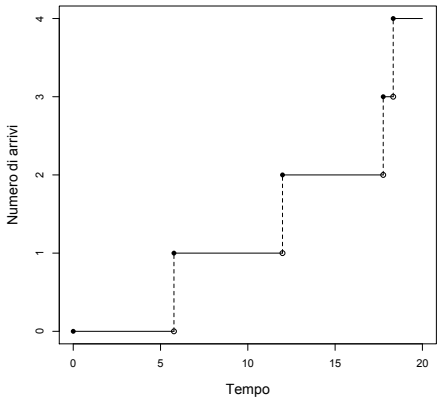

Figura 3.1: Una realizzazione del processo di Poisson.

Supponiamo di indicare con $N(t) \quad (t \geq 0)$ il numero di arrivi (ad esempio, chiamate che si presentano ad un centralino telefonico) nell'intervallo di tempo $(0, t)$ e con $N(t, t+\Delta t)$ il numero di arrivi nell'intervallo $(t, t+\Delta t)$. Ovviamente risulta $N(t, t + \Delta t) = N(t + \Delta t) - N(t)$.

Definizione 3.1 *Un processo stocastico* $\{N(t),\ t \geq 0\}$ *è detto di Poisson con parametro* ϱ $(\varrho > 0)$ *se si ha:*

(i) $N(0) = 0$,

(ii) il processo ha incrementi indipendenti e stazionari,

(iii) $P\big\{N(t, t + \Delta t) = 1\big\} = \varrho\ \Delta t + o(\Delta t)$,

(iv) $P\big\{N(t, t + \Delta t) > 1\big\} = o(\Delta t)$,
dove $o(\Delta t)$ *è un infinitesimo di ordine superiore rispetto a* Δt *e* ϱ *denota il parametro di arrivo con dimensione fisiche* $[tempo]^{-1}$.

La condizione *(i)* significa assumere che fino al tempo $t = 0$ non si sono verificati eventi. La condizione *(ii)* assicura che gli eventi che si verificano in intervalli di tempo disgiunti, ossia che non si sovrappongono, sono stocasticamente indipendenti (il processo ha incrementi indipendenti) e inoltre la distribuzione del numero di eventi che si verificano in ogni intervallo di tempo dipende soltanto dalla lunghezza dell'intervallo considerato (il processo ha incrementi stazionari). Le condizioni *(iii)* e *(iv)* invece assicurano che

$$P\big\{N(t, t + \Delta t) = 0\big\} = 1 - \varrho\ \Delta t + o(\Delta t).$$

Inoltre, la condizione *(iv)* mostra che in un piccolo intervallo di tempo $(t, t+\Delta t)$ gli eventi si verificano al più singolarmente. Dalla condizione *(ii)* di indipendenza

scaturisce che

$$P\{N(t,t+\Delta t)=1|N(t)=n\}=\varrho\,\Delta t+o(\Delta t)\qquad(n=0,1,\ldots),$$
$$P\{N(t,t+\Delta t)=0|N(t)=n\}=1-\varrho\,\Delta t+o(\Delta t)\qquad(n=0,1,\ldots),$$
$$P\{N(t,t+\Delta t)>1|N(t)=n\}=o(\Delta t)\qquad(n=0,1,\ldots).$$

Denotiamo con

$$p_n(t)=P\{N(t)=n\}\qquad(n=0,1,\ldots)$$

la probabilità che sia n il numero di arrivi fino al tempo t, ossia nell'intervallo di tempo $(0,t)$.

Proposizione 3.1 *Per un processo stocastico di Poisson* $\{N(t),\,t\geq 0\}$ *di parametro* $\varrho>0$*, si ha*

$$p_n(t)=\frac{(\varrho\,t)^n}{n!}\,e^{-\varrho\,t}\qquad(n=0,1,\ldots),\tag{3.1}$$

ossia per ogni fissato t *si ottiene una funzione di probabilità di Poisson di parametro* $\varrho\,t$*.*

Dimostrazione Come si evince dalla Figura 3.2, si nota che sussistono le

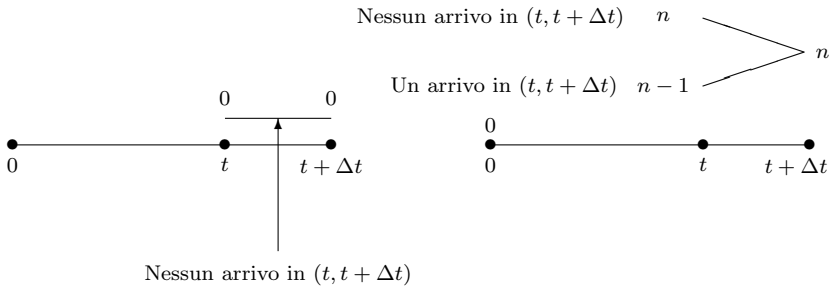

Figura 3.2: Cambiamenti di stato in $(t,t+\Delta t)$ nel processo di Poisson.

seguenti identità:

$$\begin{aligned}p_0(t+\Delta t)&=P\{N(t+\Delta t)=0\}=P\{N(t)=0,N(t,t+\Delta t)=0\}\\&=p_0(t)\,(1-\varrho\,\Delta t)+o(\Delta t)\\p_n(t+\Delta t)&=P\{N(t+\Delta t)=n\}=P\{N(t)=n,N(t,t+\Delta t)=0\}\\&\quad+P\{N(t)=n-1,N(t,t+\Delta t)=1\}+o(\Delta t)\\&=p_n(t)\,(1-\varrho\,\Delta t)+p_{n-1}(t)\,\varrho\,\Delta t+o(\Delta t)\qquad(n=1,2,\ldots),\end{aligned}$$

ossia

$$\frac{p_0(t+\Delta t)-p_0(t)}{\Delta t}=-\varrho\,p_0(t)+\frac{o(\Delta t)}{\Delta t},$$
$$\frac{p_n(t+\Delta t)-p_n(t)}{\Delta t}=-\varrho\,p_n(t)+\varrho\,p_{n-1}(t)+\frac{o(\Delta t)}{\Delta t}\qquad(n=1,2,\ldots).$$

Procedendo al limite quando $\Delta t \to 0$ si ricava:

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -\varrho\, p_0(t), \\ \frac{dp_n(t)}{dt} &= -\varrho\, p_n(t) + \varrho\, p_{n-1}(t) \qquad (n = 1, 2, \ldots). \end{aligned} \tag{3.2}$$

Abbiamo ottenuto un sistema di equazioni differenziali e alle differenze del primo ordine in n che, per l'ipotesi *(i)*, deve essere risolto con le condizioni iniziali

$$p_n(0) = P\{N(0) = n\} = \begin{cases} 1, & n = 0 \\ 0, & n = 1, 2, \ldots \end{cases} \tag{3.3}$$

Consideriamo la funzione generatrice di probabilità

$$G(z,t) = \sum_{n=0}^{+\infty} z^n\, p_n(t). \tag{3.4}$$

Moltiplicando ambo i membri della seconda delle (3.2) per z^n e sommando su $n = 1, 2, \ldots$ si ha:

$$\frac{\partial}{\partial t}\Big[G(z,t) - p_0(t)\Big] = -\varrho\,\Big[G(z,t) - p_0(t)\Big] + \varrho\, z\, G(z,t)$$

da cui, utilizzando la prima delle (3.2), si ottiene

$$\frac{\partial G(z,t)}{\partial t} = \varrho\,(z-1)\, G(z,t). \tag{3.5}$$

Ricordando (3.3) e (3.4) segue che l'equazione (3.5) deve essere risolta con la condizione iniziale:

$$G(z,0) = \sum_{n=0}^{+\infty} z^n\, p_n(0) = 1. \tag{3.6}$$

La soluzione della (3.5), con la condizione iniziale (3.6), è:

$$G(z,t) = \exp\{\varrho\, t\,(z-1)\} = e^{-\varrho\, t}\; e^{\varrho\, t\, z}. \tag{3.7}$$

Espandendo $e^{\varrho\, t\, z}$ in serie di potenze di z, la (3.7) diventa

$$G(z,t) = e^{-\varrho\, t} \sum_{n=0}^{+\infty} \frac{(\varrho\, t)^n}{n!}\; z^n. \tag{3.8}$$

Uguagliando uguali potenze di z in (3.4) e (3.8) segue immediatamente la (3.1). Per ogni fissato t abbiamo quindi ottenuto una funzione di probabilità di Poisson di parametro $\varrho\, t$. □

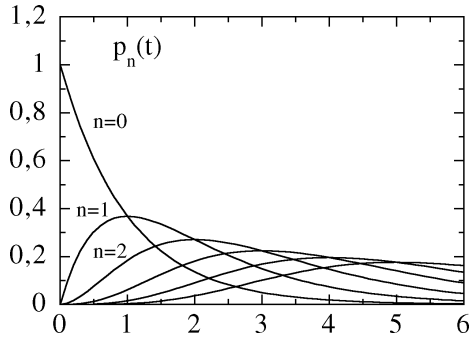

Figura 3.3: Le probabilità $p_n(t)$ $(n = 0, 1, \ldots, 5)$ del processo di Poisson in funzione di $\varrho\, t$.

Nella Figura 3.3 dall'alto verso il basso sono riportate le probabilità $p_0(t)$, $p_1(t)$, $p_2(t)$, $p_3(t)$, $p_4(t)$ e $p_5(t)$ in funzione di $\varrho\, t$. Si nota che mentre $p_0(t)$ è una funzione decrescente in $\varrho\, t$, le probabilità $p_n(t) \quad (n = 1, 2, \ldots)$ presentano un punto di massimo quando $\varrho\, t = n$. Il valore medio e la varianza del numero di arrivi nell'intervallo $(0, t)$ sono:

$$E[N(t)] = \sum_{n=1}^{+\infty} n\, p_n(t) = \varrho\, t, \qquad \mathrm{Var}[N(t)] = \varrho\, t. \tag{3.9}$$

Inoltre, il coefficiente di variazione è:

$$C[N(t)] = \frac{\sqrt{\mathrm{Var}[N(t)]}}{E[N(t)]} = \frac{1}{\sqrt{\varrho\, t}}\,.$$

Si nota che $\lim_{t \to +\infty} C[N(t)] = 0$, che evidenzia che al crescere del tempo il numero medio di arrivi diventa sempre più significativo.

Un'importante proprietà del processo stocastico di Poisson di parametro ϱ è che *i tempi di interarrivo*, ossia le lunghezze degli intervalli tra due arrivi successivi, *sono indipendenti e identicamente distribuiti con densità esponenziale di parametro* ϱ. Quindi, se supponiamo che gli arrivi si verifichino ai tempi T_1, $T_1 + T_2$, $T_1 + T_2 + T_3$, $\cdots$, dove T_n denota la lunghezza dell'intervallo aleatorio di tempo tra l'evento $(n-1)$–esimo e l'evento n–esimo, si ha:

$$f_{T_n}(t) = \begin{cases} \varrho\, e^{-\varrho\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases}$$

Il valore medio e la varianza dei tempi di interarrivo sono rispettivamente:

$$E(T_n) = \frac{1}{\varrho}\,, \qquad \mathrm{Var}(T_n) = \frac{1}{\varrho^2} \qquad (n = 1, 2, \ldots)$$

e il coefficiente di variazione è quindi unitario.

Inoltre, in un processo stocastico di Poisson di parametro $\varrho > 0$, la variabile aleatoria $T_1 + T_2 + \ldots + T_k$, che descrive la lunghezza l'intervallo di tempo fino all'arrivo k–esimo, è caratterizzata da densità di probabilità

$$f(t) = \frac{dP(T_1 + T_2 + \ldots + T_k < t)}{dt} = \begin{cases} \dfrac{\varrho^k}{(k-1)!}\, e^{-\varrho t}\, t^{k-1}, & t > 0 \\ 0, & \text{altrimenti,} \end{cases}$$

ossia una densità di Erlang di ordine k e di parametro ϱ.

Esempio 3.1 Supponiamo che gli arrivi ad un sistema di servizio si verifichino in accordo ad un processo di Poisson e che il tempo medio tra due arrivi successivi sia di 25 secondi. Determinare la funzione di distribuzione $P(T < t)$ dei tempi di interarrivo e la probabilità che nell'intervallo $(0, t)$ si siano verificati n arrivi, misurando il tempo in minuti. Calcolare infine la media e la varianza del numero di arrivi alla fine della prima ora.
Si nota che il tempo medio di interarrivo è

$$E(T) = \frac{1}{\lambda} = 25 \text{ secondi} = 25\,\frac{1}{60} \text{ minuti} = \frac{5}{12} \text{ minuti},$$

da cui segue che la frequenza di arrivo al minuto è $\lambda = 12/5 = 2.4$. La funzione di distribuzione dei tempi di interarrivo è esponenziale:

$$P(T < t) = 1 - e^{-\lambda\, t} = 1 - e^{-2.4\, t} \qquad (t > 0),$$

con il tempo t misurato in minuti. La probabilità che si verifichino n arrivi nell'intervallo $(0, t)$, con t misurato in minuti, è

$$p_n(t) = P\{N(t) = n\} = \frac{(\lambda\, t)^n}{n!}\, e^{-\lambda\, t} = \frac{(2.4\, t)^n}{n!}\, e^{-2.4\, t} \qquad (n = 0, 1, \ldots).$$

Per calcolare la media e la varianza del numero di arrivi alla fine della prima ora, basta porre $t = 60$ minuti e ricordare la (3.9):

$$E[N(60)] = \text{Var}[N(60)] = \lambda\, t = 2.4 \cdot 60 = 144,$$

ossia si hanno in media 144 arrivi dopo la prima ora con una deviazione standard di 12. ♢

3.3 Processi stocastici di nascita morte

I processi stocastici di nascita-morte sono di fondamentale importanza nella costruzione di vari modelli probabilistici atti a descrivere fenomeni in cui lo stato del sistema è caratterizzato da un numero intero di individui e in cui i cambiamenti di stato rappresentano l'addizione o la sottrazione di individui dal sistema in vari modi: nascite, morti, immigrazioni, emigrazioni, arrivi e partenze degli utenti da un sistema di servizio. In Figura 3.4 è rappresentata una realizzazione di un processo di nascita-morte.

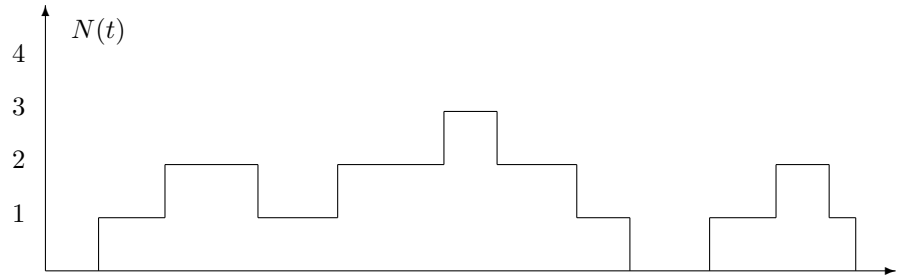

Figura 3.4: Una tipica realizzazione di un processo di nascita-morte.

I processi di nascita-morte sono i più noti processi stocastici discreti nello spazio degli stati e continui nel tempo e sono stati introdotti nel 1939 dal matematico croato, naturalizzato statunitense, William Feller (1906-1970), le cui pubblicazioni sono state e sono tuttora essenziali per la diffusione della teoria della probabilità nel mondo[1] [2]. I processi di nascita-morte sono utilizzati per costruire modelli di crescita di popolazione, di sistemi di servizio, in epidemiologia e in molte aree di interesse sia teorico che applicativo.

Con riferimento ai sistemi di servizio, è spesso ragionevole supporre che sia il parametro di arrivo dell'utente che entra nel sistema sia il parametro di partenza dell'utente che esce dal sistema (essendo stato servito) dipendano dal numero degli utenti presenti nel sistema. Spesso si suppone che il tempo T_n che intercorre tra l'arrivo $(n-1)$–esimo e l'arrivo n–esimo $(n = 0, 1, \ldots)$ sia distribuito esponenzialmente con valore medio $1/\lambda_n$ e che il tempo S_n occorrente per servire l'utente n–esimo $(n = 1, 2, \ldots)$ sia anche distribuito esponenzialmente con valore medio $1/\mu_n$; inoltre, spesso si assume che sia i tempi di interarrivo $T_1, T_2, \ldots$ sia i tempi di servizio $S_1, S_2, \ldots$ siano indipendenti tra loro. Gli insiemi $\{\lambda_n,\, n = 0, 1, \ldots\}$ e $\{\mu_n,\, n = 1, 2, \ldots\}$ contengono rispettivamente i parametri di arrivo (di nascita) e i parametri di partenza (di morte).

Definiamo ora un processo nascita-morte facendo riferimento principalmente alla teoria delle delle file di attesa.

Definizione 3.2 *Sia* $\{N(t),\, t \geq 0\}$ *un processo stocastico avente spazio degli stati* $0, 1, 2, \ldots$. *Supponiamo che questo processo descriva un sistema che si trova nello stato* E_n *al tempo* t *se e solo se* $N(t) = n$, *ossia se il numero di utenti (individui) presenti al tempo* t *è* n. *Tale processo stocastico è detto di nascita–morte se esistono dei parametri di arrivo (di nascita)* $\{\lambda_n,\, n = 0, 1, \ldots\}$ *e di partenza (di morte)* $\{\mu_n,\, n = 1, 2, \ldots\}$ *tali da soddisfare i seguenti postulati:*

(i) In un piccolo intervallo di tempo Δt *si possono avere cambiamenti di stato soltanto dallo stato* E_n *allo stato* E_{n+1} *oppure dallo stato* E_n *allo stato*

[1] An Introduction to Probability Theory and its Applications, Volume I, 3rd edition (1968)
[2] An Introduction to Probability Theory and its Applications, Volume II, 2nd edition (1971)

E_{n-1} se $n \geq 1$, mentre se $n = 0$ si può avere un cambiamento di stato soltanto dallo stato E_0 allo stato E_1.

(ii) Se al tempo t il sistema è nello stato E_n, la probabilità che nell'intervallo di tempo $(t, t+\Delta t)$ avvenga una transizione dallo stato E_n allo stato E_{n+1} è $\lambda_n \Delta t + o(\Delta t)$, mentre la probabilità che nell'intervallo di tempo $(t, t+\Delta t)$ avvenga una transizione dallo stato E_n allo stato E_{n-1} è $\mu_n \Delta t + o(\Delta t)$.

(iii) Se al tempo t il sistema è nello stato E_n, la probabilità che nell'intervallo di tempo $(t, t+\Delta t)$ avvenga più di una transizione è $o(\Delta t)$.

Il postulato *(i)* mostra che se nel sistema è presente almeno un utente in un piccolo intervallo di tempo può verificarsi al più una transizione (un arrivo oppure una partenza), mentre se il sistema è vuoto non vi possono essere uscite dal sistema. Il postulato *(ii)* fornisce le probabilità di transizione, ossia le probabilità di arrivo o di partenza in un piccolo intervallo di tempo $(t, t+\Delta t)$ quando il numero degli utenti nel sistema al tempo t è n. L'ultimo postulato mostra che la probabilità che in un piccolo intervallo di tempo si verifichi più di una transizione è trascurabile.

Se indichiamo con $N(t)$ $(t \geq 0)$ il numero degli utenti presenti nel sistema al tempo t e con $N(t, t+\Delta t)$ l'incremento nel numero di utenti nell'intervallo $(t, t+\Delta t]$, le ipotesi precedenti possono essere così formulate:

$$
\begin{aligned}
&P\big\{N(t, t+\Delta t) = 1 | N(t) = n\big\} = \lambda_n \, \Delta t + o(\Delta t) \quad (n = 0, 1, \ldots), \\
&P\big\{N(t, t+\Delta t) = -1 | N(t) = n\big\} = \mu_n \, \Delta t + o(\Delta t) \quad (n = 1, 2, \ldots), \\
&P\big\{N(t, t+\Delta t) = 0 | N(t) = n\big\} = \begin{cases} 1 - \lambda_0 \, \Delta t + o(\Delta t), & n = 0 \\ 1 - (\lambda_n + \mu_n) \, \Delta t + o(\Delta t), & n = 1, 2, \ldots, \end{cases}
\end{aligned}
\tag{3.10}
$$

dove $o(\Delta t)$ è un infinitesimo di ordine superiore rispetto a Δt, $\lambda_0, \lambda_1, \ldots$ denotano i parametri di arrivo e $\mu_1, \mu_2, \ldots$ denotano i parametri di partenza. Tali parametri di arrivo e di partenza hanno dimensioni fisiche $[tempo]^{-1}$. Dalle (3.10) segue immediatamente che la probabilità che in un piccolo intervallo di tempo si verifichi più di una transizione è trascurabile. Denotiamo con

$$
p_n(t) = P\big\{N(t) = n\big\} \qquad (n = 0, 1, \ldots),
$$

la probabilità che sia n il numero degli utenti presenti nel sistema al tempo t. Come si evince dalla Figura 3.5, si ha:

$$
\begin{aligned}
p_0(t+\Delta t) &= P\big\{N(t+\Delta t) = 0\big\} = p_0(t) \, [1 - \lambda_0 \, \Delta t] + p_1(t) \, \mu_1 \, \Delta t + o(\Delta t), \\
p_n(t+\Delta t) &= P\big\{N(t+\Delta t) = n\big\} = p_{n-1}(t) \, \lambda_{n-1} \, \Delta \, t + p_n(t) \, [1 - (\lambda_n + \mu_n) \, \Delta t] \\
&\quad + p_{n+1}(t) \, \mu_{n+1} \, \Delta t + o(\Delta t) \qquad (n = 1, 2, \ldots),
\end{aligned}
$$

ossia

$$
\frac{p_0(t+\Delta t) - p_0(t)}{\Delta t} = -\lambda_0 \, p_0(t) + \mu_1 \, p_1(t) + \frac{o(\Delta t)}{\Delta t},
$$

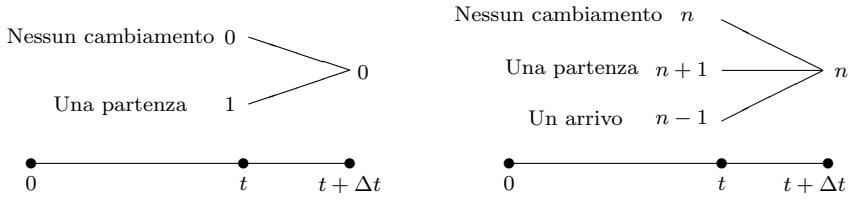

Figura 3.5: Cambiamenti di stato in $(t, t+\Delta t)$ nel processo di nascita morte.

$$\frac{p_n(t+\Delta t) - p_n(t)}{\Delta t} = \lambda_{n-1}\, p_{n-1}(t) - (\lambda_n + \mu_n)\, p_n(t) + \mu_{n+1}\, p_{n+1}(t) + \frac{o(\Delta t)}{\Delta t} \qquad (n = 1, 2, \ldots).$$

Procedendo al limite quando $\Delta t \to 0$ si ottiene il seguente sistema di equazioni differenziali e alle differenze del secondo ordine in n:

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -\lambda_0\, p_0(t) + \mu_1\, p_1(t), \\ \frac{dp_n(t)}{dt} &= \lambda_{n-1}\, p_{n-1}(t) - (\lambda_n + \mu_n)\, p_n(t) + \mu_{n+1}\, p_{n+1}(t) \qquad (n = 1, 2, \ldots). \end{aligned} \tag{3.11}$$

Se supponiamo che inizialmente nel sistema di servizio siano presenti i utenti, ossia $P\{N(0) = i\} = 1$, occorre risolvere il sistema (3.11) con le condizioni iniziali:

$$p_n(0) = \begin{cases} 1, & n = i \\ 0, & n \neq i. \end{cases} \tag{3.12}$$

Esempio 3.2 Supponiamo che

$$\begin{aligned} \lambda_n &= \varrho \qquad n = 0, 1, 2, \ldots \\ \mu_n &= 0 \qquad n = 1, 2, \ldots \end{aligned}$$

e inoltre

$$p_n(0) = \begin{cases} 1, & n = 0 \\ 0, & n = 1, 2, \ldots. \end{cases}$$

In questo caso il processo di nascita morte si riduce a un processo di Poisson di parametro ϱ. Abbiamo precedentemente mostrato che $\{p_0(t), p_1(t), \ldots\}$ è una distribuzione di Poisson di parametro $\varrho\, t$ per ogni $t \geq 0$. Il processo di Poisson può essere allora visto come un processo di pura nascita. $\diamondsuit$

In un processo stocastico di nascita–morte per determinare le probabilità di avere n utenti $(n = 0, 1, \ldots)$ nel sistema al tempo t $(t > 0)$ occorre quindi risolvere il sistema di equazioni differenziali (3.11) con le condizioni iniziali (3.12). Si può dimostrare che in condizioni abbastanza generali questo sistema di equazioni ammette un'unica soluzione. Comunque, eccetto in casi particolarmente semplici, la distribuzione $\{p_0(t), p_1(t), \ldots\}$ è difficile da determinare risolvendo il

sistema (3.11). Risulta anche notevolmente complesso calcolare la soluzione approssimata del sistema (3.11) ricorrendo a metodi numerici. Nella maggior parte dei casi per stimare le probabilità nel transiente di un processo di nascita–morte occorre ricorrere a metodi di simulazione.

3.4 Equilibrio statistico

La distribuzione di probabilità $\{p_0(t), p_1(t), \ldots\}$ di un processo di nascita-morte permette di descrivere il comportamento del sistema. Come già precedentemente sottolineato è in generale molto difficile ottenere la distribuzione di probabilità nel transiente, ossia per ogni fissato istante di tempo t. Vogliamo quindi determinare delle condizioni sui parametri di arrivo $\{\lambda_n,\ n = 0, 1, \ldots\}$ e di partenza $\{\mu_n,\ n = 1, 2, \ldots\}$ che conducano ad una situazione di equilibrio statistico del sistema. A tal fine denotiamo con

$$q_n = \lim_{t \to +\infty} p_n(t) \qquad (n = 0, 1, \ldots) \tag{3.13}$$

la probabilità di avere n utenti nel sistema nella situazione di equilibrio statistico.

Se i limiti nella (3.13) *esistono e non dipendono dalle condizioni iniziali* per ogni $n = 0, 1, \ldots$, diremo che *il sistema raggiunge una situazione di equilibrio statistico* descritta dalla distribuzione di equilibrio $\{q_0, q_1, \ldots\}$.

Osserviamo che se i limiti in (3.13) esistono, allora al crescere del tempo $dp_n(t)/dt$ tende a zero per ogni $n = 0, 1, \ldots$ Il sistema (3.11) quindi diventa:

$$\begin{aligned} &-\lambda_0\, q_0 + \mu_1\, q_1 = 0, \\ &\lambda_{n-1}\, q_{n-1} - (\lambda_n + \mu_n)\, q_n + \mu_{n+1}\, q_{n+1} = 0 \qquad (n = 1, 2, \ldots). \end{aligned} \tag{3.14}$$

Le equazioni alle differenze (3.14) possono essere ricavate in modo alternativo costruendo il *grafo di transizione*, illustrato in Figura 3.6. In questo grafo ogni stato E_n è rappresentato da un cerchietto (nodo) etichettato con il numero n. Gli archi che collegano i nodi mostrano quali sono le possibili transizioni di stato e sono etichettati con i parametri di transizione (ossia i parametri di arrivo o di servizio).

Se un processo nascita–morte ha raggiunto la situazione di equilibrio statistico, allora per ogni stato E_n del sistema $(n = 0, 1, \ldots)$ assumiamo che valga il *principio di bilanciamento* che afferma: "*il flusso medio che entra nel nodo n deve uguagliare il flusso medio che esce da tale nodo*". Le equazioni che esprimono tale principio sono dette *equazioni di bilanciamento*.

Ricaviamo ora il sistema di equazioni alle differenze (3.14) utilizzando il principio di bilanciamento. Nella situazione di equilibrio statistico il grafo di transizione di un processo di nascita–morte è il seguente: Per il nodo E_0 si ha che il flusso medio entrante è $\mu_1\, q_1$ e il flusso medio uscente è $\lambda_0\, q_0$; pertanto per il principio di bilanciamento si deve avere

$$\mu_1\, q_1 = \lambda_0\, q_0. \tag{3.15}$$

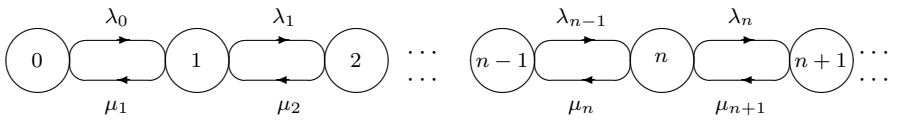

Figura 3.6: Grafo di transizione di un processo di nascita–morte in condizioni di equilibrio statistico.

Invece, per un generico nodo $E_n \quad (n = 1, 2, \ldots)$ si ha che il flusso medio entrante è $\lambda_{n-1}\, q_{n-1} + \mu_{n+1}\, q_{n+1}$ e il flusso medio uscente è $\lambda_n\, q_n + \mu_n\, q_n$; pertanto per il principio di bilanciamento si deve avere

$$\lambda_{n-1}\, q_{n-1} + \mu_{n+1}\, q_{n+1} = \lambda_n\, q_n + \mu_n\, q_n \qquad (n = 1, 2, \ldots). \tag{3.16}$$

Le equazioni (3.15) e (3.16) corrispondono a quelle del sistema (3.14). Ci proponiamo ora di determinare la soluzione del sistema (3.14).

Proposizione 3.2 *Un processo nascita–morte* $\{N(t), t \geq 0\}$ *ammette una distribuzione di equilibrio statistico* $\{q_0, q_1, \ldots\}$ *se e solo se*

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} < +\infty \tag{3.17}$$

e si ha

$$\begin{aligned} q_0 &= P(N = 0) = \left[1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}\right]^{-1}, \\ q_n &= P(N = n) = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \qquad (n = 1, 2, \ldots). \end{aligned} \tag{3.18}$$

Dimostrazione Se si pone

$$g_n = \lambda_{n-1}\, q_{n-1} - \mu_n\, q_n, \qquad (n = 1, 2, \ldots), \tag{3.19}$$

la seconda delle (3.14) si può così scrivere:

$$g_{n+1} = g_n \qquad (n = 2, 3, \ldots),$$

ossia un'equazione alle differenze lineare del primo ordine la cui soluzione è una costante reale c. Si deve quindi avere

$$g_n = \lambda_{n-1}\, q_{n-1} - \mu_n\, q_n = c \qquad (n = 2, 3, \ldots). \tag{3.20}$$

Imponendo che la (3.20) sia soddisfatta anche per $n = 1$ si ottiene

$$\lambda_0\, q_0 - \mu_1\, q_1 = c,$$

da cui, per la prima delle (3.14), si ha $c = 0$. Ponendo nella (3.20) $c = 0$ si ottiene:

$$q_n = \frac{\lambda_{n-1}}{\mu_n}\, q_{n-1} = \frac{\lambda_{n-1}\lambda_{n-2}}{\mu_n\mu_{n-1}}\, q_{n-2} = \ldots\ldots = \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}\, q_0. \tag{3.21}$$

La probabilità q_0 può essere determinata imponendo che l'insieme $\{q_0, q_1, \ldots\}$ sia una distribuzione di probabilità, ossia

$$q_n \geq 0 \qquad (n = 0, 1, \ldots), \qquad\qquad \sum_{n=0}^{+\infty} q_n = 1. \tag{3.22}$$

Facendo uso di (3.21) nella seconda delle (3.22) si ricava:

$$1 = \sum_{n=0}^{+\infty} q_n = q_0 + q_0 \sum_{n=1}^{+\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n} = q_0\left[1 + \sum_{n=1}^{+\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}\right],$$

ossia

$$q_0 = P(N = 0) = \left[1 + \sum_{n=1}^{+\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}\right]^{-1}.$$

Si nota che affinché il processo nascita–morte ammetta una distribuzione di equilibrio statistico $\{q_0, q_1, \ldots\}$ è necessario che la serie

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}$$

converga. In un processo nascita–morte si può dimostrare che tale condizione è anche sufficiente. □

Se la serie (3.17) diverge, ciò indica che il sistema di servizio è *instabile* nel senso che in media gli arrivi si verificano più frequentemente delle partenze. Pertanto la (3.17) è chiamata "condizione di equilibrio" o "condizione di stabilità" di un processo di nascita morte.

In un processo di nascita morte in condizioni di equilibrio statistico la frequenza media di arrivo per unità di tempo può essere così definita:

$$\lambda^* = \sum_{n=0}^{+\infty} \lambda_n\, q_n, \tag{3.23}$$

ossia è la media pesata delle frequenze di arrivo λ_n $(n = 0, 1, \ldots)$. Inoltre, la frequenza media di partenza per unità di tempo per ognuno dei servitori è data da:

$$\mu^* = \frac{1}{1 - q_0} \sum_{n=1}^{+\infty} \mu_n\, q_n = \sum_{n=1}^{+\infty} \mu_n\, \frac{q_n}{1 - q_0}\,, \tag{3.24}$$

ossia è la media pesata delle frequenze di partenza μ_n $(n = 1, 2, \ldots)$. Quindi, λ^* e μ^* si possono rispettivamente interpretare come i valori medi dei parametri di

arrivo $\lambda_0, \lambda_1, \ldots$ e di partenza $\mu_1, \mu_2, \ldots$, ottenuti utilizzando la distribuzione di equilibrio $\{q_0, q_1, \ldots\}$.

L'intensità di traffico, ossia l'intensità di lavoro che svolge il sistema di servizio nella situazione di equilibrio statistico, è $a = \lambda^*/\mu^*$. Se si denota con s il numero di servitori presenti nel sistema di servizio, il fattore di utilizzazione del sistema, ossia l'intensità di lavoro per servitore nella situazione di equilibrio statistico, è quindi $\varrho^* = a/s = \lambda^*/(s\,\mu^*)$.

Esempio 3.3 Consideriamo un sistema di servizio con unico servitore e con una fila di attesa che può contenere al massimo un unico utente.

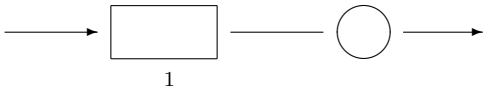

Figura 3.7: Sistema di servizio con unico servitore e fila di attesa con unico utente.

Il numero di utenti presenti in tale sistema può essere 0 (nessun utente sia in fila di attesa che in servizio), 1 (nessun utente in fila di attesa e un unico utente in servizio) e 2 (un solo utente in fila di attesa e un solo utente in servizio). Gli utenti possono accedere al sistema soltanto se ci sono 0 utenti oppure 1 solo utente (già in servizio). Assumiamo che $\lambda_0 = 7$, $\lambda_1 = 2$. Gli utenti possono uscire dal sistema soltanto se ci sono 1 utente (in servizio) oppure 2 utenti (uno in servizio e uno in fila di attesa). Assumiamo che $\mu_1 = 3$ e $\mu_2 = 6$.

Il grafo di transizione in condizioni di equilibrio è illustrato in Figura 3.8 e contiene tre stati $0, 1, 2$ poiché nel sistema possono essere presenti al più 2 utenti.

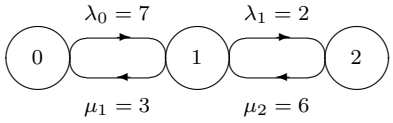

Figura 3.8: Grafo di transizione in condizioni di equilibrio statistico.

Le equazioni di bilanciamento sono:

$$7\,q_0 = 3\,q_1, \qquad 7\,q_0 + 6\,q_2 = 5\,q_1, \qquad 2\,q_1 = 6\,q_2,$$

da cui si ricava

$$q_1 = \frac{7}{3}\,q_0, \qquad q_2 = \frac{1}{3}\,q_1 = \frac{7}{9}\,q_0.$$

Ricordando che (q_0, q_1, q_2) è una distribuzione di probabilità:

$$q_0 \left(1 + \frac{7}{3} + \frac{7}{9}\right) = 1.$$

La distribuzione di probabilità in condizioni di equilibrio è:

$$q_0 = \frac{9}{37}, \qquad q_1 = \frac{21}{37}, \qquad q_2 = \frac{7}{37}.$$

Tale distribuzione può essere ottenuta direttamente utilizzando le (3.18):

$$q_0 = \left(1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0\lambda_1}{\mu_1\mu_2}\right)^{-1} = \frac{9}{37}, \quad q_1 = \frac{\lambda_0}{\mu_1} q_0 = \frac{21}{37}, \quad q_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2}\, q_0 = \frac{7}{37}.$$

Dalle (3.23) e (3.24) otteniamo le frequenze medie di arrivo e di partenza degli utenti:

$$\lambda^* = \lambda_0\, q_0 + \lambda_1\, q_1 = 7\,\frac{9}{37} + 2\,\frac{21}{37} = \frac{105}{37} = 2.84,$$
$$\mu^* = \frac{1}{1-q_0}\,(\mu_1\, q_1 + \mu_2\, q_2) = \frac{37}{28}\left(3\,\frac{21}{37} + 6\,\frac{7}{37}\right) = \frac{105}{28} = 3.75,$$

da cui è possibile ricavare l'intensità di traffico del sistema, ossia l'intensità di lavoro che svolge il sistema di servizio in condizione di equilibrio:

$$\varrho^* = \frac{\lambda^*}{\mu^*} = \frac{28}{37} = 0.76.$$

La conoscenza delle probabilità q_0, q_1, q_2 consente di ottenere il numero medio di utenti presenti nel sistema in condizione di equilibrio, ossia

$$E(N) = 0 \cdot q_0 + 1 \cdot q_1 + 2 \cdot q_2 = \frac{35}{37} = 0.95$$

Quindi, in media nel sistema esiste un solo utente. Inoltre, in condizioni di equilibrio, utilizzando la prima legge di Little è possibile calcolare il tempo medio di attesa nel sistema di un utente ossia

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{35}{105} = 0.33.$$

♢

Esempio 3.4 Consideriamo un sistema di servizio $M/M/1$, in cui i tempi di interarrivo sono esponenziali con valore medio $1/\lambda$, i tempi di servizio sono esponenziali con valore medio $1/\mu$, la capacità del sistema è infinita ed esiste un unico servitore. Le frequenze di arrivo e di servizio sono costanti, ossia

$$\lambda_n = \lambda \quad (n = 0, 1, \ldots), \qquad\qquad \mu_n = \mu \quad (n = 1, 2, \ldots).$$

Tale sistema di servizio è illustrato in Figura 3.9. Gli arrivi sono descritti da un processo di Poisson con frequenza λ.
Il grafo di transizione del sistema $M/M/1$ in condizioni di equilibrio è illustrato in Figura 3.10.

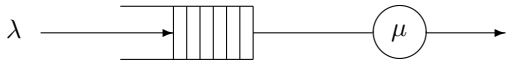

Figura 3.9: Sistema di servizio $M/M/1$.

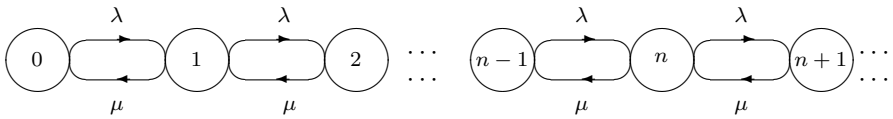

Figura 3.10: Grafo di transizione del sistema $M/M/1$ in condizioni di equilibrio.

Ponendo $\varrho = \lambda/\mu$, dalla (3.17) si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \sum_{n=0}^{+\infty} \Big(\frac{\lambda}{\mu}\Big)^n = \sum_{n=0}^{+\infty} \varrho^n,$$

ossia una serie geometrica che converge a $(1-\varrho)^{-1}$ se e solo se la ragione $\varrho = \lambda/\mu < 1$. Quindi, il sistema di servizio $M/M/1$ raggiunge una situazione di equilibrio statistico se e solo se $\varrho = \lambda/\mu < 1$ e si ha:

$$\begin{aligned} q_0 &= P(N=0) = 1-\varrho, \\ q_n &= P(N=n) = q_0 \, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = (1-\varrho)\, \varrho^n \qquad (n = 1, 2, \ldots). \end{aligned}$$

La conoscenza di tali probabilità consente di ottenere il numero medio di utenti presenti nel sistema di servizio:

$$E(N) = 0 \cdot q_0 + 1 \cdot q_1 + 2 \cdot q_2 + \ldots = \sum_{n=0}^{+\infty} n \, q_n = \frac{\varrho}{1-\varrho}.$$

Affinché il sistema di servizio $M/M/1$ non si congestioni occorre che $\lambda < \mu$. In condizione di equilibrio statistico, dalle (3.23) e (3.24) otteniamo le frequenze medie di arrivo e di partenza degli utenti:

$$\begin{aligned} \lambda^* &= \sum_{n=0}^{+\infty} \lambda_n \, q_n = \lambda \sum_{n=0}^{+\infty} q_n = \lambda, \\ \mu^* &= \frac{1}{1-q_0} \sum_{n=1}^{+\infty} \mu_n \, q_n = \frac{\mu}{1-q_0} \sum_{n=1}^{+\infty} q_n = \frac{\mu}{1-q_0} \, (1-q_0) = \mu, \end{aligned}$$

che coincidono con λ e μ. Quindi, l'intensità di lavoro svolta dal sistema di servizio nella situazione di equilibrio è $\varrho^* = \lambda^*/\mu^* = \lambda/\mu$.

Nel prossimo capitolo riconsideremo il sistema di servizio $M/M/1$ e determineremo le principali misure prestazionali del sistema. ♢

Capitolo 4

Modelli con singolo servitore

4.1 Introduzione

In questo capitolo analizzeremo i principali sistemi di servizio con singolo servitore, ossia i sistemi $M/M/1$, $M/M/1/1$, $M/M/1/K$ e $M/G/1$. Inoltre, confronteremo il sistema $M/M/1$ con un sistema di servizio adattivo di interesse nella teoria delle file di attesa. Lo scopo è quello di individuare i principali parametri prestazionali dei vari sistemi e di analizzare il loro comportamento in situazioni di equilibrio statistico.

4.2 Sistema di servizio $M/M/1$

Supponiamo che gli utenti arrivino ad un sistema di servizio secondo un processo di Poisson di parametro λ. I tempi tra due successivi arrivi (tempi di interarrivo) sono quindi indipendenti e esponenzialmente distribuiti con valore medio $1/\lambda$. Il sistema ha un unico servitore e utilizza la disciplina di servizio FIFO. La capacità del sistema è infinita. Dopo l'arrivo, ogni utente è immediatamente servito se il servitore è libero, mentre se il servitore è occupato l'utente si mette in fila di attesa. Quando il servitore termina di servire un utente, si ha una partenza dal sistema e un nuovo utente nella fila di attesa, se ne esiste almeno uno presente, può accedere al servizio. Supponiamo che i tempi di servizio degli utenti siano indipendenti e esponenzialmente distribuiti con valore medio $1/\mu$.

Tale sistema, illustrato in Figura 4.1, è noto in letteratura come *sistema di servizio* $M/M/1$ (o equivalentemente come *sistema di servizio* $M/M/1/\infty$). La prima M significa che i tempi di interarrivo sono indipendenti e distribuiti esponenzialmente (proprietà di Markov legata alla mancanza di memoria della

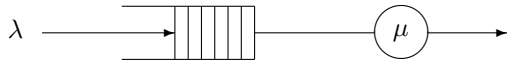

Figura 4.1: Sistema di servizio $M/M/1$.

funzione di distribuzione esponenziale); la seconda M significa che i tempi di servizio sono indipendenti e distribuiti esponenzialmente; il simbolo 1 si riferisce all'unico servitore disponibile, il simbolo ∞ indica che la capacità del sistema è illimitata.

Sia $N(t)$ il numero di utenti presenti nel sistema al tempo t. Il processo stocastico $\{N(t),\, t \geq 0\}$ può essere descritto facendo ricorso ad un processo di nascita–morte caratterizzato da parametri

$$\lambda_n = \lambda \quad (n = 0, 1, \ldots), \qquad\qquad \mu_n = \mu \quad (n = 1, 2, \ldots). \tag{4.1}$$

Vogliamo ora vedere sotto quali condizioni il sistema $M/M/1$ raggiunge una situazione di equilibrio statistico. Facendo uso di (4.1) in (3.17) e ponendo $\varrho = \lambda/\mu$, si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \sum_{n=0}^{+\infty} \Big(\frac{\lambda}{\mu}\Big)^n = \sum_{n=0}^{+\infty} \varrho^n,$$

ossia una serie geometrica che converge a $(1 - \varrho)^{-1}$ se e solo se la ragione $\varrho = \lambda/\mu < 1$. Quindi, il sistema di servizio $M/M/1$ raggiunge una situazione di equilibrio statistico se e solo se $\varrho = \lambda/\mu < 1$ e si ha:

$$\begin{aligned} q_0 &= P(N = 0) = 1 - \varrho, \\ q_n &= P(N = n) = q_0 \, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = (1 - \varrho)\, \varrho^n \qquad (n = 1, 2, \ldots). \end{aligned}$$

Proposizione 4.1 *Il sistema di servizio $M/M/1$ raggiunge una situazione di equilibrio statistico se e solo se $\varrho = \lambda/\mu < 1$ e si ha:*

$$q_n = P(N = n) = (1 - \varrho)\, \varrho^n \qquad (n = 0, 1, \ldots), \tag{4.2}$$

ossia una funzione di probabilità geometrica di parametro ϱ.

Nella situazione di equilibrio statistico il valore medio, il momento del secondo ordine e la varianza del numero di utenti presenti nel sistema sono quindi:

$$\begin{aligned} E(N) &= \sum_{n=0}^{\infty} n\, q_n = (1 - \varrho) \sum_{n=1}^{\infty} n\, \varrho^n = \frac{\varrho}{1 - \varrho}, \\ E(N^2) &= \sum_{n=0}^{\infty} n^2\, q_n = (1 - \varrho) \sum_{n=1}^{\infty} n^2\, \varrho^n = \frac{\varrho\,(1 + \varrho)}{(1 - \varrho)^2}, \\ \mathrm{Var}(N) &= E(N^2) - [E(N)]^2 = \frac{\varrho}{(1 - \varrho)^2}, \end{aligned} \tag{4.3}$$

dove si è fatto uso delle seguenti identità:

$$\sum_{n=1}^{\infty} n\, z^n = \frac{z}{(1-z)^2}, \qquad \sum_{n=1}^{\infty} n^2\, z^n = \frac{z\,(1+z)}{(1-z)^3}\,. \qquad (|\, z\,| < 1)$$

Se $\varrho \geq 1$ il sistema di servizio è *instabile* e il numero di utenti nel sistema è destinato a crescere indefinitamente.

Se denotiamo con T la variabile aleatoria esponenziale che descrive un generico tempo di interarrivo e con S la variabile aleatoria esponenziale che descrive un generico tempo di servizio, per il sistema $M/M/1$ si ha:

$$E(T) = \frac{1}{\lambda}, \qquad E(S) = \frac{1}{\mu}, \qquad \varrho = \frac{E(S)}{E(T)} = \frac{\lambda}{\mu}\,.$$

La condizione $\varrho < 1$ è quindi equivalente a richiedere che $E(S) < E(T)$, ossia il sistema $M/M/1$ raggiunge una situazione di *equilibrio statistico* se e solo se *il tempo medio di servizio è minore del tempo medio di interarrivo.*

Nella Tabella 4.1 sono indicati il valore medio, la varianza e il coefficiente di variazione del numero di utenti presenti nel sistema nella situazione di equilibrio statistico per alcune scelte dell'intensità di traffico ϱ. Si nota che il valore medio e la varianza del numero di utenti sono funzioni crescenti in ϱ e tendono all'infinito quando $\varrho \to 1$.

ϱ	$E(N)$	$\mathrm{Var}(N)$	$C(N)$
0.1	0.11111	0.12346	3.1623
0.2	0.25000	0.31250	2.2361
0.3	0.42857	0.61224	1.8257
0.4	0.66667	1.1111	1.5811
0.5	1.0000	2.0000	1.4142
0.6	1.5000	3.7500	1.2910
0.7	2.3333	7.7778	1.1952
0.8	4.0000	20.000	1.1180
0.9	9.0000	90.000	1.0541
0.99	99.000	9900.0	1.0050
0.999	999.01	9.9903×10^5	1.0005

Tabella 4.1: Valore medio, varianza e coefficiente di variazione del numero di utenti nel sistema $M/M/1$ in condizioni di equilibrio per alcune scelte di ϱ $(0 < \varrho < 1)$.

Facendo uso di (3.23) e (3.24), la frequenza media di arrivo e la frequenza media di partenza per unità di tempo nel sistema $M/M/1$ sono:

$$\lambda^* = \sum_{n=0}^{+\infty} \lambda_n\, q_n = \lambda, \qquad \mu^* = \frac{1}{1-q_0} \sum_{n=1}^{+\infty} \mu_n\, q_n = \mu.$$

Il *fattore di utilizzazione* del sistema $M/M/1$ (che coincide con l'intensità di traffico) è quindi $\varrho^* = \lambda/\mu = \varrho$ e fornisce l'intensità di lavoro svolta dal sistema di servizio nella situazione di equilibrio. Infatti, il fattore di utilizzazione

del sistema $M/M/1$ coincide con *la probabilità di avere almeno un utente nel sistema*, ossia con

$$P(N \geq 1) = 1 - q_0 = \varrho.$$

Inoltre, se $\varrho < 1$, dalla (4.2) segue che la probabilità che in condizioni di equilibrio siano presenti nel sistema un numero maggiore o uguale a k di utenti è:

$$P(N \geq k) = \sum_{n=k}^{+\infty} q_k = (1-\varrho) \sum_{n=k}^{+\infty} \varrho^n = (1-\varrho)\,\varrho^k \sum_{n=k}^{+\infty} \varrho^{n-k} = \varrho^k.$$

Dalla prima legge di Little $E(N) = \lambda^* E(W)$ ricaviamo che nella situazione di equilibrio statistico il tempo medio di attesa di un utente nel sistema è

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{\varrho}{\lambda\,(1-\varrho)} = \frac{1}{\mu\,(1-\varrho)} = \frac{1}{\mu - \lambda}.$$

Il tempo medio di permanenza di un utente nella fila di attesa è:

$$E(Q) = E(W) - E(S) = \frac{1}{\mu\,(1-\varrho)} - \frac{1}{\mu} = \frac{\varrho}{\mu\,(1-\varrho)} = \frac{\varrho}{\mu - \lambda}\,.$$

Dalla seconda legge di Little segue che nella situazione di equilibrio statistico il numero medio di utenti nella fila di attesa è:

$$E(N_q) = \lambda^*\,E(Q) = \lambda\,\frac{\varrho}{\mu\,(1-\varrho)} = \frac{\varrho^2}{1-\varrho}.$$

Si noti che $E(N_q)$ si può anche valutare nel seguente modo:

$$E(N_q) = \sum_{n=1}^{\infty} (n-1)\,q_n = E(N) - (1-q_0) = \frac{\varrho}{1-\varrho} - \varrho = \frac{\varrho^2}{1-\varrho}.$$

La terza legge di Little mostra infine che il numero medio di utenti in servizio

$$E(N_s) = \lambda^*\,E(S) = \frac{\lambda}{\mu} = \varrho$$

è uguale all'intensità di traffico del centro di servizio, che coincide anche con il *fattore di utilizzazione del sistema*.

Nel sistema $M/M/1$, i periodi di ozio del servitore, che possono essere riguardati come tempi residui dei tempi di interarrivo, sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$, ossia con densità

$$f_I(t) = \begin{cases} \lambda\,e^{-\lambda\,t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases}$$

Ricordando la (2.25), risulta che

$$E(B) = \frac{(1-q_0)\,E(I)}{q_0} = \frac{\varrho}{\lambda\,(1-\varrho)} = \frac{1}{\mu - \lambda}$$

Nella situazione di equilibrio statistico, per il sistema $M/M/1$ risulta che il tempo medio di occupazione del servitore coincide con il tempo medio di attesa nel sistema, ossia $E(B) = E(W)$.

Nella Tabella 4.2 sono riportati i principali parametri prestazionali del sistema $M/M/1$.

$$\lambda_n = \lambda \quad (n = 0, 1, \ldots), \qquad \mu_n = \mu \quad (n = 1, 2, \ldots)$$

$$\varrho = \lambda/\mu < 1 \qquad \text{(condizione di equilibrio statistico)}$$

$$q_n = P(N = n) = (1 - \varrho)\,\varrho^n \qquad (n = 0, 1, \ldots)$$

$$\lambda^* = \lambda, \qquad \mu^* = \mu, \qquad \varrho^* = \frac{\lambda}{\mu} = 1 - q_0 = P(N \geq 1)$$

$$E(N) = \frac{\varrho}{1 - \varrho}\,, \qquad E(W) = \frac{E(N)}{\lambda^*} = \frac{1}{\mu - \lambda}$$

$$E(N_q) = \frac{\varrho^2}{1 - \varrho}\,, \qquad E(Q) = \frac{\varrho}{\mu - \lambda}$$

$$E(N_s) = \frac{\lambda}{\mu} = \varrho, \qquad E(S) = \frac{1}{\mu}$$

$$E(T) = \frac{1}{\lambda}\,, \qquad E(I) = \frac{1}{\lambda}\,, \qquad E(B) = \frac{1}{\mu - \lambda}$$

Tabella 4.2: Parametri prestazionali del sistema di servizio $M/M/1$.

Esempio 4.1 Supponiamo di considerare una linea di comunicazione che può trasmettere messaggi di 8 bit. Il tempo medio necessario per trasmettere un messaggio di 8 bit è di 0.004 secondi e il tempo medio di interarrivo tra i messaggi di 8 bit è di 0.005 secondi. Se un sistema di servizio $M/M/1$ modella tale linea di comunicazione, dire se il sistema raggiunge una situazione di equilibrio e calcolare i principali parametri prestazionali.

In questo caso il tempo medio di interarrivo tra i messaggi e il tempo medio per trasmettere un messaggio sono:

$$\frac{1}{\lambda} = 0.005 \text{ secondi}, \qquad \frac{1}{\mu} = 0.004 \text{ secondi},$$

da cui

$$\lambda^* = \lambda = 200 \text{ messaggi al secondo}, \qquad \mu^* = \mu = 250 \text{ messaggi al secondo}.$$

Quindi il fattore di utilizzazione del sistema è:

$$\varrho^* = \frac{\lambda^*}{\mu^*} = \varrho = \frac{\lambda}{\mu} = \frac{200}{250} = \frac{4}{5} = 0.8 < 1.$$

Il tempo di interarrivo tra messaggi successivi è distribuito esponenzialmente con valore medio $E(T) = 1/\lambda = 1/200 = 0.005$ secondi e il tempo per trasmettere un messaggio è distribuito esponenzialmente con valore medio $E(S) = 1/\mu = 1/250 = 0.004$ secondi. Poiché $\varrho = E(S)/E(T) = 0.8 < 1$, il sistema di comunicazione considerato non si congestiona.

In condizioni di equilibrio statistico, il numero medio di messaggi nel sistema e in attesa di essere trasmessi sono:

$$E(N) = \frac{\varrho}{1-\varrho} = 4 \text{ messaggi}, \qquad E(N_q) = \frac{\varrho^2}{1-\varrho} = \frac{16}{5} = 3.2 \text{ messaggi}.$$

Inoltre, il tempo medio di attesa nel sistema e nella fila di attesa sono

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{4}{200} = 0.02 \text{ secondi}, \quad E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{3.2}{200} = 0.016 \text{ secondi},$$

e il numero medio di messaggi in trasmissione e il tempo medio di trasmissione sono:

$$E(N_s) = \varrho = 0.8 \text{ messaggi}, \qquad E(S) = \frac{1}{\mu} = 0.004.$$

La probabilità che nel sistema siano presenti un numero maggiore o uguale a k messaggi è $P(N \geq k) = \varrho^k$, da cui segue che

$$P(N \geq 1) = \varrho = 0.8, \qquad P(N \geq 2) = \varrho^2 = 0.64, \qquad P(N \geq 3) = \varrho^3 = 0.512,$$
$$P(N \geq 4) = \varrho^4 = 0.4096, \qquad P(N \geq 5) = \varrho^5 = 0.32768, \ldots\ldots$$

♢

4.3 Sistema di servizio con svendita

Consideriamo un sistema di servizio a capacità infinita con unica fila di attesa e singolo servitore in cui gli utenti sono attratti da una lunga coda e il servitore accelera il suo servizio all'aumentare della lunghezza della coda.

Un sistema di servizio di questo tipo può essere descritto mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri:

$$\begin{aligned} \lambda_n &= \lambda\,(n+1) \qquad (n = 0, 1, \ldots) \\ \mu_n &= \mu\, n \qquad (n = 1, 2, \ldots). \end{aligned} \tag{4.4}$$

Tale modello si presta a descrivere alcune situazioni reali quali una grossa svendita in cui gli utenti sono attratti dalla lunga coda e il servitore cerca di accelerare il servizio per vendere la maggior parte della merce in deposito. Vogliamo determinare in quali condizioni tale sistema raggiunge una situazione di equilibrio statistico e calcolare i principali parametri prestazionali del sistema. Inoltre, desideriamo confrontare il sistema di servizio con svendita con il sistema $M/M/1$ con frequenze di arrivo λ e frequenze di partenza μ.

Facendo uso di (4.4) in (3.17) e denotando con $\varrho = \lambda/\mu$, si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \sum_{n=1}^{+\infty} \frac{\lambda\, 2\lambda \cdots n\,\lambda}{\mu\, 2\,\mu \cdots n\,\mu} = \sum_{n=0}^{+\infty} \Big(\frac{\lambda}{\mu}\Big)^n = \sum_{n=0}^{+\infty} \varrho^n,$$

ossia una serie geometrica che converge se e solo se la ragione $\varrho = \lambda/\mu < 1$, fornendo come somma $(1-\varrho)^{-1}$.

Proposizione 4.2 *Il sistema di servizio con svendita raggiunge una situazione di equilibrio statistico se e solo se $\varrho = \lambda/\mu < 1$ e si ha:*

$$q_n = P(N = n) = (1-\varrho)\,\varrho^n \qquad (n = 0, 1, \ldots).$$

Si nota che il sistema di servizio con svendita ammette la stessa distribuzione di equilibrio del sistema $M/M/1$, ossia una distribuzione geometrica di parametro ϱ. Nella situazione di equilibrio statistico, il valore medio e la varianza del numero di utenti presenti nel sistema sono quindi gli stessi del sistema $M/M/1$, forniti in (4.3).

I modelli $M/M/1$ e quello con svendita, anche se caratterizzati dalla stessa distribuzione di equilibrio, sono fondamentalmente diversi nella fase transiente e inoltre hanno alcuni parametri prestazionali differenti. Infatti, nel sistema di servizio con svendita le frequenza medie di arrivo e di partenza per unità di tempo sono differenti da quelle del sistema $M/M/1$ e risulta:

$$\lambda^* = \sum_{n=0}^{+\infty} \lambda_n\, q_n = \lambda \sum_{n=0}^{+\infty} (n+1)\, q_n = \lambda\,[E(N)+1] = \lambda\left(\frac{\varrho}{1-\varrho} + 1\right) = \frac{\lambda}{1-\varrho}$$

$$\mu^* = \frac{1}{1-q_0} \sum_{n=1}^{+\infty} \mu_n\, q_n = \frac{\mu}{1-q_0} \sum_{n=1}^{+\infty} n\, q_n = \frac{\mu}{1-q_0}\, E(N) = \frac{\mu}{\varrho}\,\frac{\varrho}{1-\varrho} = \frac{\mu}{1-\varrho}.$$

Si nota che $\lambda^* > \lambda$ e $\mu^* > \mu$, ossia le frequenze medie di arrivo e di partenza del sistema di servizio con svendita sono maggiori di quelle del sistema $M/M/1$. Il fattore di utilizzazione del sistema (che coincide con l'intensità di traffico) è

$$\varrho^* = \frac{\lambda^*}{\mu^*} = \frac{\lambda}{\mu}$$

e coincide con quello del sistema $M/M/1$. Si nota che

$$\varrho^* = P(N \geq 1) = 1 - q_0 = \varrho.$$

Dalla prima legge di Little ricaviamo che nella situazione di equilibrio statistico il tempo medio di attesa di un utente nel sistema è

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{1-\varrho}{\lambda}\,\frac{\varrho}{1-\varrho} = \frac{1}{\mu}\,.$$

Il numero medio di utenti in fila di attesa può essere così ottenuto:

$$E(N_q) = \sum_{n=1}^{+\infty} (n-1)\, q_n = E(N) - (1 - q_0) = \frac{\varrho^2}{1-\varrho}\,,$$

che coincide con quello del sistema di servizio $M/M/1$. Dalla seconda legge di Little si ottiene:

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{\varrho^2}{1-\varrho}\,\frac{1-\varrho}{\lambda} = \frac{\varrho}{\mu}.$$

Il numero medio di utenti in servizio è

$$E(N_s) = E(N) - E(N_q) = \frac{\varrho}{1-\varrho} - \frac{\varrho^2}{1-\varrho} = \varrho,$$

che coincide con la probabilità che nel sistema sia presente almeno un utente. Dalla terza legge di Little si ricava quindi

$$E(S) = \frac{E(N_s)}{\lambda^*} = \varrho\,\frac{1-\varrho}{\lambda} = \frac{1-\varrho}{\mu},$$

ossia $E(S) = 1/\mu^*$. Nella Tabella 4.3 sono riportati i principali parametri prestazionali del sistema di servizio con svendita e unico servitore.

$\lambda_n = \lambda\,(n+1) \quad (n = 0, 1, \ldots), \qquad \mu_n = \mu\, n \quad (n = 1, 2, \ldots)$
$\varrho = \lambda/\mu < 1 \qquad$ (condizione di equilibrio statistico)
$q_n = P(N = n) = (1-\varrho)\,\varrho^n \qquad (n = 0, 1, \ldots)$
$\lambda^* = \dfrac{\lambda}{1-\varrho}, \qquad \mu^* = \dfrac{\mu}{1-\varrho}, \qquad \varrho^* = \dfrac{\lambda}{\mu} = 1 - q_0 = P(N \geq 1)$
$E(N) = \dfrac{\varrho}{1-\varrho}\,, \qquad E(W) = \dfrac{E(N)}{\lambda^*} = \dfrac{1}{\mu}$
$E(N_q) = \dfrac{\varrho^2}{1-\varrho}\,, \qquad E(Q) = \dfrac{\varrho}{\mu}$
$E(N_s) = \dfrac{\lambda}{\mu} = \varrho, \qquad E(S) = \dfrac{1-\varrho}{\mu}, \qquad E(T) = \dfrac{1-\varrho}{\lambda}$

Tabella 4.3: Parametri prestazionali del sistema di servizio con svendita e unico servitore

Come si evince dalle Tabelle 4.2 e 4.3, entrambi i sistemi di servizio $M/M/1$ e con svendita sono caratterizzati dalla stessa distribuzione di equilibrio, dallo stesso numero medio di utenti nel sistema, nella fila di attesa e nel centro di

servizio e dalla stessa intensità di traffico; risultano invece differenti la frequenza media di arrivo e di partenza per unità di tempo, i tempi medi di attesa degli utenti nel sistema, i tempi medi di permanenza nella fila di attesa e i tempi medi di servizio. In particolare, per il sistema di servizio con svendita si nota che il tempo medio di attesa, il tempo medio di permanenza in fila di attesa e il tempo medio di servizio sono inferiori rispetto a quelli del sistema $M/M/1$.

Esempio 4.2 Consideriamo un sistema di servizio con singolo servitore a capacità infinita. Desideriamo confrontare i parametri prestazionali del sistema $M/M/1$ e del sistema con svendita scegliendo in entrambi i modelli $\lambda = 0.8$ utenti al minuto e $\mu = 1$ utenti al minuto.
Entrambi i sistemi di servizio non si congestionano essendo $\varrho = \lambda/\mu = 0.8 < 1$. Tali sistemi sono caratterizzati dalla stessa distribuzione di probabilità in equilibrio:

$$q_n = P(N = n) = (1 - \varrho)\,\varrho^n = 0.2 \cdot 0.8^n, \qquad n = 0, 1, \ldots$$

I numeri medi di utenti nel sistema, in fila di attesa e in servizio sono uguali per entrambi i sistemi:

$$\begin{aligned} E(N) &= \frac{\varrho}{1-\varrho} = \frac{0.8}{0.2} = 4 \text{ utenti al minuto,} \\ E(N_q) &= \frac{\varrho^2}{1-\varrho} = \frac{0.8^2}{0.2} = 3.2 \text{ utenti al minuto,} \\ E(N_s) &= \varrho = 0.8 \text{ utenti al minuto.} \end{aligned}$$

I tempi di attesa, di permanenza in fila di attesa e nel sistema sono invece differenti per i due sistemi. Per il sistema $M/M/1$ si ha;

$$\begin{aligned} &\lambda^* = \lambda = 0.8, \qquad \mu^* = \mu = 1, \qquad \varrho^* = \frac{\lambda^*}{\mu^*} = 0.8, \\ &E(W) = \frac{E(N)}{\lambda^*} = \frac{4}{0.8} = 5 \text{ minuti}, \quad E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{3.2}{0.8} = 4 \text{ minuti}, \\ &E(S) = \frac{E(N_s)}{\lambda^*} = \frac{0.8}{0.8} = 1 \text{ minuto.} \end{aligned}$$

Invece per il sistema con svendita si ottiene:

$$\begin{aligned} &\lambda^* = \frac{\lambda}{1-\varrho} = \frac{0.8}{0.2} = 4, \qquad \mu^* = \frac{\mu}{1-\varrho} = \frac{1}{0.2} = 5, \qquad \varrho^* = \frac{\lambda^*}{\mu^*} = \frac{4}{5} = 0.8, \\ &E(W) = \frac{E(N)}{\lambda^*} = \frac{4}{4} = 1 \text{ minuto}, \quad E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{3.2}{4} = 0.8 \text{ minuti}, \\ &E(S) = \frac{E(N_s)}{\lambda^*} = \frac{0.8}{4} = 0.2 \text{ minuti.} \end{aligned}$$

Come ci si aspettava, nel sistema con svendita si ha che il tempo medio di attesa, il tempo medio di permanenza in fila di attesa e il tempo medio di servizio sono inferiori rispetto a quelli del sistema $M/M/1$. ♢

4.4 Sistema di servizio $M/M/1/1$

Il sistema $M/M/1/1$ descrive un centralino telefonico con un'unica linea disponibile, con fila di attesa ha capacità nulla in cui le chiamate che arrivano e trovano il centralino occupato sono perse. Supponiamo che gli utenti arrivino al sistema di servizio secondo un processo di Poisson di parametro λ. Se nel sistema è già presente un utente (in servizio) il nuovo utente in arrivo non può accedere al sistema. Quando il servitore termina di servire un utente, tale utente lascia il sistema e il nuovo utente in arrivo può usufruire del servizio. Supponiamo che i tempi di servizio degli utenti siano indipendenti e esponenzialmente distribuiti con valore medio $1/\mu$. Tale sistema, illustrato in Figura 4.2, è noto in letteratura come *sistema di servizio* $M/M/1/1$.

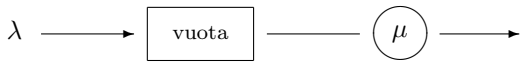

Figura 4.2: Sistema di servizio $M/M/1/1$.

La prima M significa che i tempi di interarrivo sono distribuiti esponenzialmente; la seconda M significa che i tempi di servizio sono anche distribuiti esponenzialmente; il simbolo 1 si riferisce all'unico servitore e l'ultimo simbolo indica che la capacità del sistema è unitaria. Nel sistema $M/M/1/1$ sono nulli sia il numero medio di utenti nella fila di attesa che il tempo medio di permanenza nella fila di attesa. Inoltre, il tempo medio di attesa di un utente coincide con il tempo medio di servizio, ossia $E(W) = E(S) = 1/\mu$, ed il numero medio di utenti nel sistema coincide con il numero medio di utenti in servizio, ossia $E(N) = E(N_s)$.

Denotiamo con $N(t)$ il numero di utenti presenti nel sistema $M/M/1/1$ al tempo t. Il processo stocastico $\{N(t),\, t \geq 0\}$ può essere descritto mediante un processo di nascita–morte caratterizzato da parametri:

$$\lambda_n = \begin{cases} \lambda, & n = 0 \\ 0, & n = 1, 2, \ldots \end{cases} \qquad \mu_n = \begin{cases} \mu, & n = 1 \\ 0, & n = 2, 3, \ldots \end{cases} \tag{4.5}$$

A differenza del sistema $M/M/1$, il sistema $M/M/1/1$ è a capacità finita e raggiunge una situazione di equilibrio anche quando $\varrho \geq 1$. Vogliamo ora determinare la distribuzione di equilibrio. Facendo uso di (4.5) in (3.17) e ponendo $\varrho = \lambda/\mu$ si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \frac{\lambda}{\mu} = 1 + \varrho$$

Proposizione 4.3 *Per il sistema $M/M/1/1$ in condizioni di equilibrio statistico si ha:*

$$q_0 = P(N = 0) = \frac{\mu}{\lambda + \mu} = \frac{1}{1 + \varrho}, \qquad q_1 = P(N = 1) = \frac{\lambda}{\lambda + \mu} = \frac{\varrho}{1 + \varrho}$$

e

$$E(N) = E(N_s) = 0 \cdot q_0 + 1 \cdot q_1 = \frac{\varrho}{1+\varrho}.$$

La *probabilità che un utente in arrivo sia rifiutato* corrisponde alla probabilità che il servitore sia occupato (non essendo presente fila di attesa), ossia è uguale a $q_1 = P(N = 1)$. Dalla Tabella 4.4 si evince che la probabilità che un utente in arrivo sia rifiutato è maggiore di 0.5 quando si ha $\varrho > 1$. Inoltre, la frequenza

ϱ	q_1	ϱ	q_1
0.2	0.166667	1.2	0.545455
0.4	0.285714	1.4	0.583333
0.6	0.375	1.6	0.615385
0.8	0.444444	1.8	0.642857
1.0	0.5	2.0	0.666667

Tabella 4.4: Probabilità che un utente in arrivo sia rifiutato nel sistema $M/M/1/1$.

media di servizio è $\mu^* = \mu$ e dalla terza legge di Little si ricava

$$\lambda^* = \frac{E(N_s)}{E(S)} = \frac{\varrho}{1+\varrho}\,\mu = \frac{\lambda}{1+\varrho} = \lambda\, q_0,$$

che mostra che *la frequenza media di arrivo è il prodotto della probabilità che un utente in arrivo non trovi utenti nel sistema e della frequenza di arrivo λ al sistema.* Quindi, l'intensità di traffico, che coincide con il fattore di utilizzazione del sistema, è

$$\varrho^* = \frac{\lambda^*}{\mu^*} = \frac{\varrho}{1+\varrho} = q_1 = 1 - q_0,$$

che coincide con la probabilità che il sistema sia occupato. I risultati per il sistema $M/M/1/1$ sono riportati in Tabella 4.5.

4.5 Sistema di servizio $M/M/1/K$

Supponiamo che gli utenti arrivino ad un sistema di servizio secondo un processo di Poisson di parametro λ. Dopo l'arrivo, ogni utente è immediatamente servito se il servitore è libero, mentre se il servitore è occupato l'utente si mette in fila di attesa se il numero di utenti presenti nel sistema è minore di K, mentre se K utenti sono già presenti nel sistema un nuovo utente in arrivo non può accedere al sistema. Quando il servitore termina di servire un utente, tale utente lascia il sistema e il nuovo utente nella fila di attesa (se ne esiste almeno uno presente) può usufruire del servizio. Supponiamo che i tempi di servizio degli utenti siano indipendenti e esponenzialmente distribuiti con valore medio $1/\mu$. In questo sistema la fila di attesa è limitata, ossia al più K utenti (incluso quello in servizio) possono essere presenti nel sistema e la disciplina di servizio è quella FIFO.

$$\lambda_n = \begin{cases} \lambda, & n = 0 \\ 0, & n = 1, 2, \ldots \end{cases} \qquad \mu_n = \begin{cases} \mu, & n = 1 \\ 0, & n = 2, 3, \ldots \end{cases}$$

$$\varrho = \lambda/\mu < +\infty$$

$$q_0 = P(N = 0) = \frac{1}{1+\varrho}, \qquad q_1 = P(N = 1) = \frac{\varrho}{1+\varrho}$$

$$\lambda^* = \frac{\lambda}{1+\varrho} \qquad \mu^* = \mu \qquad \varrho^* = \frac{\varrho}{1+\varrho}$$

$$E(N) = \frac{\varrho}{1+\varrho}, \qquad E(W) = \frac{E(N)}{\lambda^*} = \frac{1}{\mu}$$

$$E(N_q) = 0, \qquad E(Q) = 0, \quad E(N_s) = \frac{\varrho}{1+\varrho} = \varrho^*, \quad E(S) = \frac{1}{\mu}$$

Tabella 4.5: Parametri prestazionali del sistema di servizio $M/M/1/1$

Tale sistema, illustrato in Figura 4.3, è noto in letteratura come *sistema di servizio* $M/M/1/K$.

La prima M significa che i tempi di interarrivo sono distribuiti esponenzialmente; la seconda M significa che i tempi di servizio sono anche distribuiti esponenzialmente; il simbolo 1 si riferisce all'unico servitore e il simbolo K indica la capacità del sistema. In particolare, se si pone $K = 1$, si ottiene il sistema

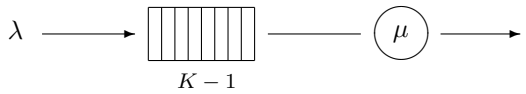

Figura 4.3: Sistema di servizio $M/M/1/K$.

di servizio $M/M/1/1$.

Denotiamo con $N(t)$ il numero di utenti presenti nel sistema $M/M/1/K$ al tempo t. Il processo stocastico $\{N(t),\, t \geq 0\}$ può essere descritto mediante un processo di nascita–morte caratterizzato da parametri:

$$\lambda_n = \begin{cases} \lambda, & n = 0, 1, \ldots, K-1 \\ 0, & n = K, K+1, \ldots \end{cases}$$

$$\mu_n = \begin{cases} \mu, & n = 1, 2, \ldots, K \\ 0, & n = K+1, K+2, \ldots \end{cases} \tag{4.6}$$

Poiché il sistema di servizio $M/M/1/K$ è a capacità finita, raggiunge sempre una situazione di equilibrio statistico.

Vogliamo ora determinare la distribuzione di equilibrio. Facendo uso di (4.6) in (3.17) e ponendo $\varrho = \lambda/\mu$ si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \sum_{n=0}^{K} \Big(\frac{\lambda}{\mu}\Big)^n = \sum_{n=0}^{K} \varrho^n = \begin{cases} \dfrac{1-\varrho^{K+1}}{1-\varrho}\,, & \varrho \neq 1 \\ K+1, & \varrho = 1. \end{cases}$$

Proposizione 4.4 *Per il sistema $M/M/1/K$ in condizioni di equilibrio statistico si ha*

$$q_0 = P(N=0) = \begin{cases} \dfrac{1-\varrho}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{1}{K+1}\,, & \varrho = 1 \end{cases}$$

e per $n = 1, 2, \ldots, K$ risulta: (4.7)

$$q_n = P(N=n) = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \begin{cases} \dfrac{1-\varrho}{1-\varrho^{K+1}}\, \varrho^n\,, & \varrho \neq 1 \\ \dfrac{1}{K+1}\,, & \varrho = 1. \end{cases}$$

A differenza del sistema $M/M/1$, il sistema $M/M/1/K$ raggiunge una situazione di equilibrio anche quando $\varrho \geq 1$. In particolare, quando $\varrho = 1$ la distribuzione di equilibrio è quella equiprobabile, ossia una distribuzione che assegna uguali probabilità a tutti i $K+1$ stati del sistema di servizio. Se $\varrho \neq 1$, dalle (4.7) si ricava

$$\frac{q_n}{q_{n-1}} = \varrho \qquad (n = 1, 2, \ldots)$$

Tale relazione mostra che

- se $\varrho < 1$ si ha $q_n < q_{n-1} \Rightarrow q_0 > q_1 > \ldots > q_K$;
- se $\varrho > 1$ si ha $q_n > q_{n-1} \Rightarrow q_K > q_{K-1} > \ldots > q_0$;
- se $\varrho = 1$ si ha $q_n = q_{n-1} \Rightarrow q_K = q_{K-1} = \ldots = q_0$.

Quindi, se $\varrho < 1$ è più probabile trovare il sistema $M/M/1/K$ vuoto, mentre se $\varrho > 1$ è più probabile trovare il sistema $M/M/1/K$ saturo.

Per evitare la congestione di un sistema di servizio (come nel sistema $M/M/1$ quando $\varrho \geq 1$) non è quindi conveniente ridurre la capacità del sistema. Infatti, se $\varrho > 1$, essendo più probabile trovare il sistema $M/M/1/K$ saturo, si impedisce a molti utenti di accedere al sistema di servizio.

Proposizione 4.5 *In condizioni di equilibrio statistico il numero medio di utenti presenti nel sistema $M/M/1/K$ è:*

$$E(N) = \begin{cases} \dfrac{\varrho}{1-\varrho} - \dfrac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{K}{2}\,, & \varrho = 1. \end{cases} \tag{4.8}$$

Dimostrazione Se $\varrho = 1$ si ha:

$$E(N) = \sum_{n=1}^{K} n\, q_n = \frac{1}{K+1} \sum_{n=1}^{K} n = \frac{1}{K+1}\, \frac{K\,(K+1)}{2} = \frac{K}{2}\,,$$

mentre se $\varrho \neq 1$ risulta

$$E(N) = \sum_{n=1}^{K} n\, q_n = \frac{1-\varrho}{1-\varrho^{K+1}} \sum_{n=1}^{K} n\, \varrho^n = \frac{\varrho}{1-\varrho} - \frac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}}\,,$$

dove si è fatto uso dell'identità:

$$\sum_{n=1}^{K} n\, z^n = \begin{cases} \dfrac{z\,(1-z^{K+1}) - (K+1)\,z^{K+1}(1-z)}{(1-z)^2}, & z \neq 1 \\ \dfrac{K\,(K+1)}{2}, & z = 1. \end{cases}$$

□

La (4.8) mostra che se $\varrho < 1$ il numero medio di utenti presenti nel sistema $M/M/1/K$ è inferiore al numero medio $E(N) = \varrho/(1-\varrho)$ di utenti presenti nel sistema $M/M/1$.

La *probabilità che un utente in arrivo sia rifiutato* è $P(N = K) = q_K$, dove

$$q_K = \begin{cases} \dfrac{1-\varrho}{1-\varrho^{K+1}}\, \varrho^K\,, & \varrho \neq 1 \\ \dfrac{1}{K+1}\,, & \varrho = 1. \end{cases}$$

Tale probabilità è una funzione decrescente in K. Quando $\varrho \leq 1$, la probabilità q_K tende a zero quando K cresce, mentre se $\varrho > 1$ la probabilità q_K tende a $(\varrho - 1)/\varrho$ quando K cresce.

Ad esempio, in Fig. 4.4 sono rappresentate le probabilità q_K per $\varrho = 0.9$ al variare di K. Invece, in Fig. 4.5 sono rappresentate le probabilità q_K per $\varrho = 2$ al variare di K. Si nota che $(\varrho - 1)/\varrho = 0.5$ e quindi la probabilità che un utente in arrivo sia rifiutato non può essere scelta arbitrariamente piccola come in Fig. 4.4.

Quando $\varrho \leq 1$ un progettista di sistemi può determinare come scegliere il più piccolo valore della capacità K del sistema affinché la probabilità di rifiutare utenti sia minore di un fissata probabilità α (scelta dal progettista molto piccola), ossia tale che $q_K < \alpha$. Ad esempio, in Fig. 4.4 per $\varrho = 0.9$ e $\alpha = 0.1$ si nota che il più piccolo valore di K tale che $q_K < \alpha = 0.1$ è $K = 7$. Quindi, per $\varrho = 0.9$ un sistema di servizio $M/M/1/7$ garantisce che la probabilità di rifiutare utenti sia minore di $\alpha = 0.1$.

La *probabilità che il sistema di servizio sia non saturo*, ossia la probabilità che altri utenti possano accedere al sistema, è

$$P(N < K) = 1 - q_K = \begin{cases} \dfrac{1-\varrho^K}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{K}{K+1}\,, & \varrho = 1. \end{cases}$$

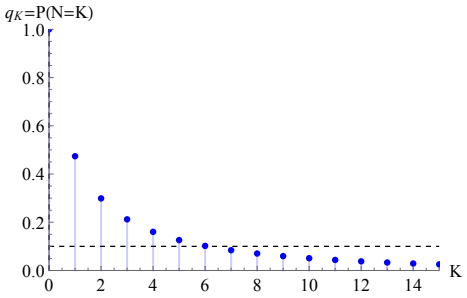

Figura 4.4: Le probabilità q_K sono rappresentate al variare di K per $\varrho = 0.9$. La linea orizzontale è relativa alla probabilità $\alpha = 0.1$.

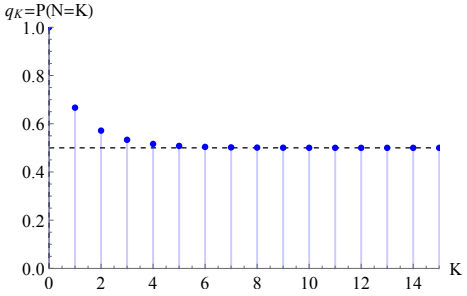

Figura 4.5: Le probabilità q_K sono rappresentate al variare di K per $\varrho = 2$. La linea orizzontale è relativa al valore asintotico $(\varrho - 1)/\varrho = 0.5$.

Nel sistema $M/M/1/K$ le frequenze medie di arrivo e di partenza per unità di tempo sono:

$$\lambda^* = \lambda\,(1 - q_K), \qquad \mu^* = \mu,$$

ossia:

$$\lambda^* = \sum_{n=0}^{+\infty} \lambda_n q_n = \lambda \sum_{n=0}^{K-1} q_n = \lambda\,(1 - q_K) = \begin{cases} \dfrac{\lambda\,(1-\varrho^K)}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{\lambda\,K}{K+1}\,, & \varrho = 1, \end{cases}$$

$$\mu^* = \frac{1}{1-q_0} \sum_{n=1}^{+\infty} \mu_n\, q_n = \frac{\mu}{1-q_0} \sum_{n=1}^{K} q_n = \frac{\mu}{1-q_0}\,(1-q_0) = \mu.$$

Si nota inoltre che *la frequenza media di partenza per unità di tempo è la stessa del sistema $M/M/1$, mentre la frequenza media di arrivo per unità di tempo dipende dalla capacità K del sistema ed è minore di quella del sistema $M/M/1$.*

Il fattore di utilizzazione del sistema, coincidente con l'intensità di traffico, è:

$$\varrho^* = \frac{\lambda^*}{\mu^*} = \frac{\lambda\,(1-q_K)}{\mu} = \varrho\,(1-q_K) = \begin{cases} \varrho\,\dfrac{1-\varrho^K}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{K}{K+1}\,, & \varrho = 1. \end{cases}$$

Si nota immediatamente che

$$\varrho^* = 1 - q_0,$$

ossia il *fattore di utilizzazione coincide con la probabilità che nel sistema sia presente almeno un utente*, ossia $\varrho^* = P(N \geq 1)$.

Proposizione 4.6 *Nella situazione di equilibrio statistico il tempo medio di attesa di un utente nel sistema $M/M/1/K$ è quindi:*

$$E(W) = \frac{E(N)}{\lambda^*} = \begin{cases} \dfrac{1}{\mu}\Big[\dfrac{1}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\Big], & \varrho \neq 1 \\ \dfrac{K+1}{2\,\mu}\,, & \varrho = 1. \end{cases} \tag{4.9}$$

Dimostrazione Dalla prima legge di Little ricaviamo il tempo medio di attesa di un utente nel sistema. Se $\varrho \neq 1$ si ha:

$$\begin{aligned} E(W) = \frac{E(N)}{\lambda^*} &= \frac{1-\varrho^{K+1}}{\lambda\,(1-\varrho^K)}\Big[\frac{\varrho}{1-\varrho} - \frac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}}\Big] \\ &= \frac{1}{\mu\,(1-\varrho^K)}\Big[\frac{1-\varrho^{K+1}}{1-\varrho} - (K+1)\,\varrho^K\Big] \\ &= \frac{1}{\mu\,(1-\varrho^K)}\,\frac{1-\varrho^K - K\,\varrho^K\,(1-\varrho)}{1-\varrho} \\ &= \frac{1}{\mu}\Big[\frac{1}{1-\varrho} - \frac{K\,\varrho^K}{1-\varrho^K}\Big], \end{aligned}$$

mentre se $\varrho = 1$ si ottiene:

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{K}{2}\,\frac{K+1}{\lambda\,K} = \frac{K+1}{2\,\lambda} = \frac{K+1}{2\,\mu}\,.$$

□

Se $\varrho < 1$, dalla (4.9) segue che il tempo medio di attesa nel sistema $M/M/1/K$ è inferiore al tempo medio di attesa $E(W) = 1/[\mu(1-\varrho)]$ nel sistema $M/M/1$.

Proposizione 4.7 *Nella situazione di equilibrio statistico il numero medio di utenti in fila di attesa nel sistema $M/M/1/K$ è:*

$$E(N_q) = \begin{cases} \dfrac{\varrho\,(1-\varrho^K)}{1-\varrho^{K+1}}\,\Big[\dfrac{\varrho}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\Big], & \varrho \neq 1 \\ \dfrac{K\,(K-1)}{2\,(K+1)}\,, & \varrho = 1. \end{cases} \tag{4.10}$$

Dimostrazione Il numero medio di utenti in fila di attesa è:

$$E(N_q) = \sum_{n=1}^{K} (n-1)\, q_n = \sum_{n=1}^{K} n\, q_n - \sum_{n=1}^{K} q_n = E(N) - (1 - q_0) \tag{4.11}$$

Dalla (4.11), se $\varrho = 1$ si ha

$$E(N_q) = \frac{K}{2} - \frac{K}{K+1} = \frac{K\,(K-1)}{2\,(K+1)}\,,$$

mentre se $\varrho \neq 1$ risulta:

$$E(N_q) = \frac{\varrho}{1-\varrho} - \frac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}} - \frac{\varrho\,(1-\varrho^k)}{1-\varrho^{k+1}} = \frac{\varrho}{1-\varrho} - \frac{\varrho\,(K\,\varrho^k + 1)}{1-\varrho^{k+1}}$$
$$= \frac{\varrho}{1-\varrho^{k+1}} \Big(\frac{1-\varrho^{k+1}}{1-\varrho} - K\,\varrho^k - 1 \Big) = \frac{\varrho\,(1-\varrho^K)}{1-\varrho^{K+1}} \Big[\frac{\varrho}{1-\varrho} - \frac{K\,\varrho^K}{1-\varrho^K} \Big].$$

□

Se $\varrho < 1$, dalla (4.10), il numero medio di utenti nella fila di attesa del sistema $M/M/1/K$ è inferiore al numero medio $E(N_q) = \varrho^2/(1-\varrho)$ di utenti nella fila di attesa del sistema $M/M/1$.

Dalla seconda legge di Little scaturisce che il tempo medio di permanenza di un utente nella fila di attesa è

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \begin{cases} \dfrac{1}{\mu} \Big[\dfrac{\varrho}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K} \Big], & \varrho \neq 1 \\ \dfrac{K-1}{2\,\mu}\,, & \varrho = 1. \end{cases} \tag{4.12}$$

Se $\varrho < 1$, dalla (4.12) segue che il tempo medio di permanenza nella fila di attesa nel sistema $M/M/1/K$ è inferiore al tempo medio di permanenza nella fila di attesa $E(Q) = \varrho/[\mu(1-\varrho)]$ nel sistema $M/M/1$.

Nella situazione di equilibrio si può valutare anche il numero medio di utenti in servizio, ossia

$$E(N_s) = E(N) - E(N_q) = \begin{cases} \dfrac{\varrho\,(1-\varrho^K)}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \dfrac{K}{K+1}\,, & \varrho = 1, \end{cases}$$

che coincide con l'intensità di traffico e anche con il fattore di utilizzazione ϱ^* del sistema. Se $\varrho < 1$ si ha anche che il numero medio di utenti in servizio nel sistema $M/M/1/K$ è inferiore al numero medio $E(N_s) = \varrho$ di utenti in servizio nel sistema $M/M/1$. Dalla terza legge di Little segue anche l'identità $E(S) = E(N_s)/\lambda^* = 1/\mu$.

Nella Tabella 4.6 sono riportati i principali parametri prestazionali del sistema di servizio $M/M/1/K$.

$$\lambda_n = \begin{cases} \lambda, & n = 0, 1, \ldots, K-1 \\ 0, & n = K, K+1, \ldots \end{cases} \qquad \mu_n = \begin{cases} \mu, & n = 1, 2, \ldots, K \\ 0, & n = K+1, K+2, \ldots \end{cases}$$

$$\varrho = \lambda/\mu < +\infty$$

$$q_n = P(N=n) = \begin{cases} \dfrac{1-\varrho}{1-\varrho^{K+1}}\,\varrho^n\,, & \varrho \neq 1 \\ 1/(K+1)\,, & \varrho = 1 \end{cases} \qquad (n = 0, 1, \ldots, K)$$

$$\lambda^* = \lambda\,(1-q_K) = \begin{cases} \dfrac{\lambda\,(1-\varrho^K)}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ \lambda\,K/(K+1)\,, & \varrho = 1 \end{cases}, \qquad \mu^* = \mu \qquad \varrho^* = \varrho\,(1-q_K)$$

$$E(N) = \begin{cases} \dfrac{\varrho}{1-\varrho} - \dfrac{(K+1)\,\varrho^{K+1}}{1-\varrho^{K+1}}\,, & \varrho \neq 1 \\ K/2\,, & \varrho = 1 \end{cases}, \quad E(W) = \begin{cases} \dfrac{1}{\mu}\left[\dfrac{1}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\right], & \varrho \neq 1 \\ (K+1)/(2\,\mu)\,, & \varrho = 1 \end{cases}$$

$$E(N_q) = \begin{cases} \varrho^*\left[\dfrac{\varrho}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\right], & \varrho \neq 1 \\ K\,(K-1)/[2\,(K+1)]\,, & \varrho = 1 \end{cases}, \quad E(Q) = \begin{cases} \dfrac{1}{\mu}\left[\dfrac{\varrho}{1-\varrho} - \dfrac{K\,\varrho^K}{1-\varrho^K}\right], & \varrho \neq 1 \\ (K-1)/(2\,\mu)\,, & \varrho = 1 \end{cases}$$

$$E(N_s) = \varrho\,(1-q_K) = \varrho^*, \qquad E(S) = 1/\mu$$

Tabella 4.6: Parametri prestazionali del sistema di servizio $M/M/1/K$

Esempio 4.3 Si consideri un centralino telefonico con un'unica linea disponibile che consenta l'attesa di due chiamate. Ulteriori chiamate quando nel sistema sono presenti tre chiamate (una in servizio e due in attesa) saranno rifiutate. Si supponga che la frequenza di arrivo è di 120 chiamate all'ora e che la durata media di una telefonata è di 20 secondi. Se un sistema di servizio $M/M/1/3$ modella il centralino telefonico considerato, calcolare i principali parametri prestazionali del sistema.

Si nota che $E(S) = 1/\mu = 20\,\text{secondi} = 20/60\,\text{minuti}$. Inoltre,

$$\lambda = \frac{120}{60} = 2\ \text{chiamate al minuto}, \qquad \mu = 60\ \frac{1}{20} = 3\ \text{chiamate al minuto},$$

e quindi $\varrho = \lambda/\mu = 2/3$. La distribuzione di equilibrio risulta essere:

$$q_n = P(N=n) = \frac{1-\varrho}{1-\varrho^4}\,\varrho^n = \frac{1-2/3}{1-(2/3)^4}\left(\frac{2}{3}\right)^n = \frac{27}{65}\left(\frac{2}{3}\right)^n, \quad n = 0, 1, 2, 3,$$

da cui segue che

$$q_0 = \frac{27}{65}, \qquad q_1 = \frac{18}{65}, \qquad q_2 = \frac{12}{65}, \qquad q_3 = \frac{8}{65}.$$

La probabilità che un utente in arrivo sia rifiutato è $q_3 = 0.1230769$. La frequenze medie di arrivo e di partenza sono:

$$\lambda^* = \lambda\,(1 - q_3) = 2\left(1 - \frac{8}{65}\right) = \frac{114}{65} = 1.754 \text{ chiamate al minuto,}$$
$$\mu^* = \mu = 3 \text{ chiamate al minuto.}$$

e quindi il fattore di utilizzazione del sistema è

$$\varrho^* = \frac{\lambda^*}{\mu^*} = \frac{114}{65} \cdot \frac{1}{3} = \frac{38}{65} = 0.585,$$

che coincide con $E(N_s)$. Il numero medio di chiamate presenti nel centralino telefonico è

$$E(N) = \frac{\varrho}{1-\varrho} - \frac{4\,\varrho^4}{1-\varrho^4} = \frac{66}{65} = 1.015 \text{ chiamate}$$

e il tempo medio di attesa nel centralino (permanenza in fila di attesa e in servizio) è

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{33}{57} = 0.579 \text{ minuti.}$$

Inoltre, il numero medio di chiamate in fila di attesa è:

$$E(N_q) = E(N) - E(N_s) = \frac{28}{65} = 0.431 \text{ chiamate}$$

e quindi il tempo medio di permanenza in fila di attesa è

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{14}{57} = 0.245 \text{ minuti.}$$

♢

4.6 Sistema di servizio $M/G/1$

Lo stato di un sistema di servizio di tipo $M/M/1$ può essere modellato utilizzando un processo stocastico di nascita morte $\{N(t), t \geq 0\}$, dove $N(t)$ rappresenta il numero di utenti presenti nel sistema al tempo t. Per studiare tale sistema ci siamo avvalsi soprattutto dell'ipotesi che i tempi di interarrivo ed anche quelli di servizio sono distribuiti esponenzialmente. Per la proprietà di Markov, legata all'assenza di memoria della densità di probabilità esponenziale, i tempi di interarrivo e di servizio residui sono distribuiti con la stessa legge di probabilità dei rispettivi tempi di interarrivo e di servizio. Quindi nel sistema $M/M/1$ non è necessario avere memoria sia del tempo trascorso dall'ultimo arrivo quando un nuovo cliente arriva nel sistema sia del tempo di servizio già speso dal cliente che sta ricevendo attualmente il servizio.

A differenza del sistema di servizio $M/M/1$, il sistema di servizio $M/G/1$ deve essere modellato con un processo stocastico non–markoviano. Infatti per

poter descrivere lo stato del sistema di servizio $M/G/1$ occorre specificare per ogni istante temporale t non soltanto il numero $N(t)$ di utenti presenti nel sistema al tempo t, ma anche il tempo di servizio $Y(t)$ che ha già ricevuto il cliente attualmente in servizio.

Vogliamo ora analizzare il sistema $M/G/1$ in cui gli arrivi si verificano secondo un processo di Poisson di parametro λ ed in cui i tempi di servizio sono indipendenti ed identicamente distribuiti con funzione di distribuzione di tipo generale, esiste un unico servitore e la capacità del sistema è infinita. Supponiamo che $E(S) = 1/\mu$.

Analogamente al sistema $M/M/1$, il sistema $M/G/1$ raggiunge una situazione di equilibrio statistico se $\varrho = \lambda/\mu < 1$ e la probabilità che nel sistema non siano presenti utenti è ancora

$$q_0 = P(N = 0) = 1 - \varrho.$$

Inoltre, il numero medio di utenti nel sistema in condizioni di equilibrio è:

$$E(N) = \varrho + \frac{\varrho^2\,[1 + C^2(S)]}{2\,(1 - \varrho)} \qquad (\varrho < 1) \tag{4.13}$$

dove $C(S) = \sqrt{\mathrm{Var}(S)}/E(S)$ denota il coefficiente di variazione della variabile aleatoria S. La (4.13) è detta *formula di Pollaczek–Khintchine*.

In particolare, se si considera il sistema $\mathbf{M/M/1}$, allora $C(S) = 1$ e la (4.13) diventa $E(N) = \varrho/(1-\varrho)$, che corrisponde all'espressione direttamente calcolata per il sistema $M/M/1$. Invece se si considera il sistema $\mathbf{M/D/1}$, allora $C(S) = 0$ e la (4.13) diventa $E(N) = \varrho + \varrho^2/[2(1 - \varrho)]$.

Vogliamo ora determinare gli altri parametri prestazionali del sistema di servizio $M/G/1$. Utilizzando la prima legge di Little risulta:

$$E(W) = \frac{E(N)}{\lambda} = \frac{1}{\mu} + \frac{\varrho\,[1 + C^2(S)]}{2\,\mu\,(1 - \varrho)}$$

Pertanto

$$E(Q) = E(W) - E(S) = \frac{\varrho\,[1 + C^2(S)]}{2\,\mu\,(1 - \varrho)},$$

da cui applicando la seconda legge di Little segue che

$$E(N_q) = \lambda\,E(Q) = \frac{\varrho^2\,[1 + C^2(S)]}{2\,(1 - \varrho)}.$$

Inoltre si ha:

$$E(N_s) = E(N) - E(N_q) = \varrho,$$

che coincide con l'intensità di traffico del sistema di servizio $M/G/1$.

Notiamo ora che nel modello $M/G/1$ i periodi di ozio, che possono essere riguardati come intervalli residui degli intervalli di interarrivo, sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$. Pertanto, si ha:

$$E(I) = \frac{1}{\lambda}, \qquad E(B) = \frac{(1 - q_0)\,E(I)}{q_0} = \frac{\varrho}{\lambda\,(1 - \varrho)} = \frac{1}{\mu - \lambda}.$$

Quindi, il tempo medio di ozio e di occupazione coincidono con quelli ottenuto per il modello $M/M/1$.

Occorre infine osservare che fissati il tempo medio di interarrivo $E(T) = 1/\lambda$ ed il tempo medio di servizio $E(S) = 1/\mu$, all'aumentare del coefficiente di variazione $C(S)$ aumenta il tempo medio di permanenza nella fila di attesa, il tempo medio di attesa nel sistema, il numero medio di utenti nella fila di attesa ed il numero medio di utenti nel sistema. Pertanto, a parità di tempo medio di interarrivo $E(T) = 1/\lambda$ e di tempo medio di servizio $E(S) = 1/\mu$, il sistema $\mathbf{M/D/1}$ (con coefficiente di variazione $C(S) = 0$) ha parametri prestazionali migliori rispetto al sistema $\mathbf{M/E_k/1}$ (con coefficiente di variazione $C(S) = 1/\sqrt{k}$ che a sua volta ha parametri prestazionali migliori rispetto al sistema $\mathbf{M/M/1}$ (con coefficiente di variazione $C(S) = 1$). Le prestazioni peggiori si manifestano nel sistema $\mathbf{M/H_k/1}$ che ha coefficiente di variazione $C(S) \geq 1$.

ϱ	$M/D/1$	$M/M/1$	$M/E_2/1$	$M/E_3/1$
0.1	0.10556	0.11111	0.10833	0.10741
0.2	0.22500	0.25000	0.23750	0.23333
0.3	0.36429	0.42857	0.39642	0.38571
0.4	0.53333	0.66667	0.60000	0.57778
0.5	0.75000	1.0000	0.87500	0.83333
0.6	1.05000	1.5000	1.27500	1.20000
0.7	1.51667	2.3333	1.92500	1.78889
0.8	2.40000	4.0000	3.20000	2.93333
0.9	4.95000	9.0000	6.97500	6.30000
0.99	49.9950	99.000	74.4975	66.3300
0.999	499.999	999.01	749.499	666.333

Tabella 4.7: Numero medio di utenti per alcune scelte di ϱ ($0 < \varrho < 1$) con tempi medi di interarrivo $E(T) = 1/\lambda$ e tempi medi di servizio $E(S) = 1/\mu$.

Nella Tabella 4.7 sono confrontati il numero medio di utenti presenti nel sistema $M/D/1$, $M/M/1$, $M/E_2/1$ e $M/E_3/1$ per differenti valori di $0 < \varrho < 1$. In questi sistemi di servizio con singolo servitore, i tempi di interarrivo sono esponenziali M con valore medio $1/\lambda$; i tempi medi di servizio sono deterministici D (tempi di servizio costanti di lunghezza $1/\mu$), esponenziali M (singola fase esponenziale di valore medio $1/\mu$), di Erlang di ordine due E_2 (servizio organizzato in due fasi esponenziali successive di valore medio $1/(2\mu)$) e di Erlang di ordine tre E_3 servizio organizzato in tre fasi esponenziali successive di valore medio $1/(3\mu)$) con tempo medio di servizio $1/\mu$. Si è utilizzata la formula di Pollaczek–Khintchine (4.13) con $C(S) = 0$ per il sistema $M/D/1$, $C(S) = 1$ per il sistema $M/M/1$, $C(S) = 1/\sqrt{2}$ per il sistema $M/E_2/1$ e $C(S) = 1/\sqrt{3}$ per il sistema $M/E_3/1$. Si nota che il numero di utenti nel sistema diminuisce all'aumentare del numero di fasi esponenziali nel servizio. Quindi, orgamizzare il servizio in più fasi successive permette di ottenere migliori prestazioni: diminuisce sia il numero di utenti nel sistema che il loro tempo di attesa.

Capitolo 5

Modelli con più servitori

5.1 Introduzione

In questo capitolo analizzeremo i principali sistemi di servizio con più servitori che lavorano in parallelo, ossia i sistemi $M/M/2$, $M/M/s$, $M/M/s/s$ e $M/M/\infty$, oltre ad altri sistemi di servizio di tipo adattivo di utilità nella teoria delle file di attesa. Lo scopo è quello di determinare i principali parametri prestazionali dei vari sistemi di servizio e di individuare, se necessario, idonee politiche atte ad evitare la congestione del sistema.

5.2 Sistema di servizio $M/M/2$

Un ovvio rimedio per un sistema di servizio che presenta congestione consiste nell'aumentare il numero di servitori. Consideriamo quindi un sistema di servizio a capacità infinita con un'unica fila di attesa e due servitori identici che lavorano in parallelo, rappresentato in Figura 5.1.

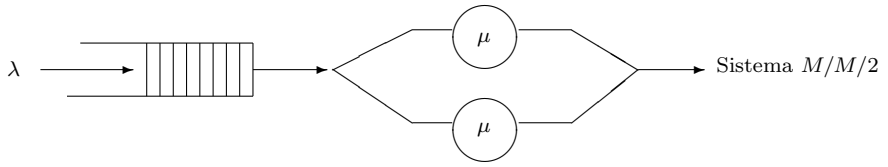

Figura 5.1: Il sistema di servizio $M/M/2$.

Supponiamo che gli utenti arrivino al sistema di servizio secondo un processo di Poisson di parametro λ. La prima M significa che i tempi di interarrivo sono

indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$ e la seconda M significa che i tempi di servizio per ognuno dei due servitori sono indipendenti e distribuiti esponenzialmente con valore medio $1/\mu$. Se un utente arriva e trova entrambi i servitori occupati si mette in fila di attesa, se trova un servitore occupato e l'altro libero sceglie il servitore libero e se entrambi i servitori sono liberi sceglie a caso uno di essi per essere servito. Tale sistema di servizio, noto in letteratura come $M/M/2$, è descrivibile mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri:

$$\begin{aligned} &\lambda_n = \lambda \qquad (n = 0, 1, \ldots) \\ &\mu_n = \mu \, \min(n, 2) = \begin{cases} \mu & n = 1 \\ 2\,\mu, & n = 2, 3, \ldots \end{cases} \end{aligned} \tag{5.1}$$

Vogliamo vedere in quali condizioni il sistema $M/M/2$ raggiunge una situazione di equilibrio statistico. Facendo uso di (5.1) in (3.17) si ha:

$$\begin{aligned} 1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} &= 1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu\,(2\,\mu)} + \frac{\lambda^3}{\mu\,(2\,\mu)^2} + \frac{\lambda^4}{\mu\,(2\,\mu)^3} + \ldots \\ &= 1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{\mu\,(2\,\mu)} \sum_{k=0}^{+\infty} \Big(\frac{\lambda}{2\,\mu}\Big)^k. \end{aligned}$$

Se si pone

$$\varrho_2 = \frac{\lambda}{2\,\mu},$$

si nota che la serie converge se e solo se $\varrho_2 < 1$ e si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + 2\,\varrho_2 + 2\,\varrho_2^2 \, \frac{1}{1 - \varrho_2} = \frac{1 + \varrho_2}{1 - \varrho_2}.$$

Proposizione 5.1 *Se $\varrho_2 = \lambda/(2\,\mu) < 1$, il sistema $M/M/2$ raggiunge una situazione di equilibrio e risulta:*

$$\begin{aligned} &q_0 = P(N = 0) = \frac{1 - \varrho_2}{1 + \varrho_2}\,, \\ &q_n = P(N = n) = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}\, q_0 = 2\, \frac{1 - \varrho_2}{1 + \varrho_2}\, \varrho_2^n \quad (n = 1, 2, \ldots). \end{aligned} \tag{5.2}$$

Se si denota con $E(T)$ il tempo medio di interarrivo e con $E(S)$ il tempo medio di servizio per servitore, la condizione di equilibrio $\varrho_2 < 1$ equivale a richiedere che $E(S)/E(T) < 2$.

Nel sistema $M/M/2$ in condizioni di equilibrio riveste particolare importanza la *probabilità che un utente in arrivo debba attendere in fila di attesa*; ciò si verifica se nel sistema sono presenti almeno due utenti, implicando che entrambi

i servitori sono occupati. La probabilità che in condizioni di equilibrio un utente in arrivo debba attendere in fila di attesa è

$$P(N \geq 2) = \sum_{n=2}^{+\infty} q_n = 2\,\frac{1-\varrho_2}{1+\varrho_2}\,\sum_{n=2}^{+\infty} \varrho_2^n = 2\,\frac{1-\varrho_2}{1+\varrho_2}\,\varrho_2^2\,\sum_{n=2}^{+\infty} \varrho_2^{n-2} = \frac{2\,\varrho_2^2}{1+\varrho_2}.$$

Tale probabilità è nota come "*formula C di Erlang*" ed è indicata con $C[2, \lambda/\mu]$, ossia:

$$C[2, \lambda/\mu] = P(N \geq 2) = \frac{2\,\varrho_2^2}{1+\varrho_2},$$

con $\varrho_2 = \lambda/(2\mu) < 1$. Essa riveste un ruolo fondamentale nella progettazione dei sistemi di servizio con più servitori. In Tabella 5.1 sono calcolate le probabilità che un utente in arrivo debba attendere in fila di attesa nel sistema $M/M/2$ al variare di λ/μ.

λ/μ	$C[2, \lambda/\mu]$	λ/μ	$C[2, \lambda/\mu]$
0.1	0.0047619	1.1	0.390323
0.2	0.0181818	1.2	0.45
0.3	0.0391304	1.3	0.512121
0.4	0.0666667	1.4	0.576471
0.5	0.1	1.5	0.642857
0.6	0.138462	1.6	0.711111
0.7	0.181481	1.7	0.781081
0.8	0.228571	1.8	0.852632
0.9	0.27931	1.9	0.925641
1.0	0.333333	1.99	0.992506

Tabella 5.1: Formula $C[2, \lambda/\mu] = P(N \geq 2)$ di Erlang.

Storicamente, nel 1908 il matematico e statistico danese *Agner Krarup Erlang* (1878–1929) venne assunto dalla compagnia telefonica di Copenaghen e nel 1909 pubblicò il suo lavoro "*The theory of probability and telephone conversations*", che costituisce il primo studio dettagliato sul traffico telefonico. Nel 1917 tale studioso introdusse la variabile aleatoria di Erlang e le due formule (B e C) usate quasi subito da tutte le altre compagnie telefoniche nella progettazione dei centralini telefonici per prevenire sovraccarichi della linea o per calcolare il numero di linee e di personale necessari. Su proposta del matematico e statistico inglese David George Kendall (1918–2007) l'unità di misura di base del traffico, nell'ambito delle telecomunicazioni, è l'Erlang (che misura l'intensità di occupazione nell'unità di tempo scelta).

Per il sistema $M/M/2$ in condizioni di equilibrio si ha che la media e la varianza del numero di utenti nel sistema sono

$$E(N) = \sum_{n=1}^{+\infty} n\,q_n = 2\,\frac{1-\varrho_2}{1+\varrho_2}\sum_{n=1}^{+\infty} n\,\varrho_2^n = \frac{2\,\varrho_2}{1-\varrho_2^2},$$

$$\mathrm{Var}(N) = \frac{2\,\varrho_2\,(1+\varrho_2^2)}{(1-\varrho_2^2)^2}.$$

La frequenza media di arrivo è $\lambda^* = \lambda$ e la frequenza media di servizio per servitore è $\mu^* = \mu$. Il fattore di utilizzazione del sistema è quindi

$$\varrho^* = \frac{\lambda^*}{2\mu^*} = \frac{\lambda}{2\mu}$$

e rappresenta *l'intensità di lavoro svolta da ognuno dei due servitori per unità di tempo.*

Applicando la prima legge di Little risulta:

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{E(N)}{\lambda} = \frac{1}{\mu\,(1-\varrho_2^2)}.$$

Poiché il tempo medio di servizio per servitore è $E(S) = 1/\mu$, segue che

$$E(Q) = E(W) - E(S) = \frac{1}{\mu\,(1-\varrho_2^2)} - \frac{1}{\mu} = \frac{\varrho_2^2}{\mu\,(1-\varrho_2^2)},$$

da cui applicando la seconda legge di Little si ricava:

$$E(N_q) = \lambda^*\,E(Q) = \lambda\,E(Q) = \frac{2\,\varrho_2^3}{1-\varrho_2^2}.$$

Infine, utilizzando la terza legge di Little, si ottiene il numero medio di utenti in servizio:

$$E(N_s) = \lambda^* E(S) = \frac{\lambda}{\mu},$$

che coincide con *l'intensità di lavoro svolta dal centro di servizio.* I parametri prestazionali del sistema di servizio $M/M/2$ sono stati elencati in Tabella 5.2.

5.3 Confronti tra i sistemi $M/M/1$ e $M/M/2$

5.3.1 Primo confronto

Una domanda che ci si può porre è la seguente: *a parità del processo degli arrivi è più conveniente scegliere un sistema con due servitori, ognuno dei quali ha tempo medio di servizio pari a $1/\mu$, oppure un sistema con unico servitore avente tempo medio di servizio dimezzato rispetto al tempo medio di servizio dei due servitori del precedente sistema.*
In Figura 5.2 sono illustrati i due sistemi $M/M/1$ e $M/M/2$ considerati; il sistema $M/M/1$ ha un unico servitore doppiamente più veloce rispetto ad ognuno dei singoli servitori del sistema $M/M/2$. Entrambi i sistemi non si congestionano se $\varrho_2 = \lambda/(2\,\mu) < 1$.

Per rispondere alla precedente domanda consideriamo in primo luogo un sistema di servizio $M/M/1$ con parametri di arrivo $\lambda_n = \lambda \quad (n = 0, 1, \ldots)$ e con parametri di partenza $\mu_n = 2\,\mu \quad (n = 1, 2, \ldots)$. Abbiamo precedentemente mostrato che se $\varrho_2 = \lambda/(2\,\mu) < 1$ esiste la distribuzione di equilibrio $(q_0^{(1)}, q_1^{(1)}, \ldots)$ e si ha

$$q_n^{(1)} = P(N^{(1)} = n) = (1-\varrho_2)\,\varrho_2^n \qquad (n = 0, 1, \ldots).$$

$$\lambda_n = \lambda \quad (n = 0, 1, \ldots), \qquad \mu_1 = \mu, \qquad \mu_n = 2\,\mu \quad (n = 2, 3, \ldots)$$

$$\varrho_2 = \frac{\lambda}{2\,\mu} < 1 \qquad \text{(condizione di equilibrio statistico)}$$

$$q_0 = P(N = 0) = \frac{1 - \varrho_2}{1 + \varrho_2}\,, \qquad q_n = P(N = n) = 2\,\frac{1 - \varrho_2}{1 + \varrho_2}\,\varrho_2^n \qquad (n = 1, 2, \ldots)$$

$$\lambda^* = \lambda, \qquad \mu^* = \mu, \qquad a = \frac{\lambda^*}{\mu^*} = \frac{\lambda}{\mu}\,, \qquad \varrho^* = \frac{a}{2} = \frac{\lambda}{2\,\mu} = \varrho_2$$

$$C[2, \lambda/\mu] = P(N \geq 2) = \frac{2\,\varrho_2^2}{1 + \varrho_2} \quad \text{(formula C di Erlang)}$$

$$E(N) = \frac{2\,\varrho_2}{1 - \varrho_2^2}, \qquad E(W) = \frac{1}{\mu\,(1 - \varrho_2^2)}$$

$$E(N_q) = \frac{2\,\varrho_2^3}{1 - \varrho_2^2}\,, \qquad E(Q) = \frac{\varrho_2^2}{\mu\,(1 - \varrho_2^2)}$$

$$E(N_s) = \frac{\lambda}{\mu}, \qquad E(S) = \frac{1}{\mu}$$

Tabella 5.2: Parametri prestazionali del sistema di servizio $M/M/2$.

Consideriamo anche un sistema di servizio $M/M/2$ con parametri di arrivo $\lambda_n = \lambda \quad (n = 0, 1, \ldots)$ e con parametri di partenza $\mu_1 = \mu$, $\mu_n = 2\,\mu$ $(n = 2, 3, \ldots)$. Tale sistema ammette la distribuzione di equilibrio $(q_0^{(2)}, q_1^{(2)}, \ldots)$ quando $\varrho_2 = \lambda/(2\,\mu) < 1$:

$$q_0^{(2)} = P(N^{(2)} = 0) = \frac{1 - \varrho_2}{1 + \varrho_2},$$
$$q_n^{(2)} = P(N^{(2)} = n) = 2\,\frac{1 - \varrho_2}{1 + \varrho_2}\,\varrho_2^n \qquad (n = 1, 2, \ldots).$$

Pertanto, quando $\varrho_2 = \lambda/(2\,\mu) < 1$ si ha:

$$\frac{q_n^{(1)}}{q_n^{(2)}} = \begin{cases} 1 + \varrho_2, & n = 0 \\ \dfrac{1 + \varrho_2}{2}\,, & n = 1, 2, \ldots \end{cases} \tag{5.3}$$

Si nota immediatamente che

$$q_0^{(1)} > q_0^{(2)}; \qquad q_n^{(1)} < q_n^{(2)} \quad (n = 1, 2, \ldots). \tag{5.4}$$

Nella situazione di equilibrio statistico è quindi più probabile trovare il sistema vuoto nel sistema $M/M/1$ piuttosto che nel sistema $M/M/2$; inoltre è meno probabile trovare nel sistema $M/M/1$ un numero n di utenti $(n = 1, 2, \ldots)$

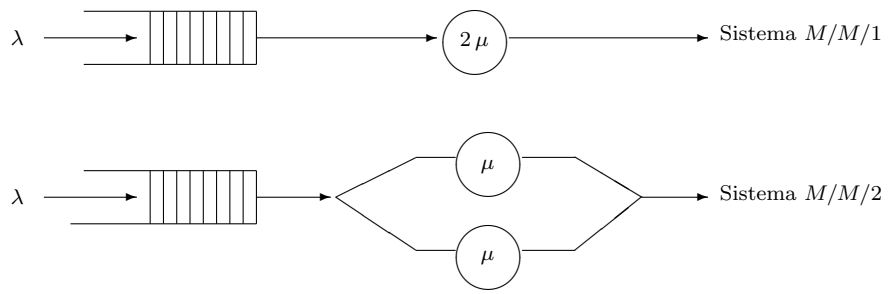

Figura 5.2: Due particolari sistemi di servizio $M/M/1$ e $M/M/2$.

piuttosto che nel sistema $M/M/2$. Facendo uso delle (5.3) si nota anche che quando ϱ_2 si approssima all'unità si ha che $q_0^{(1)} \sim 2\, q_0^{(2)}$, mentre $q_n^{(1)} \sim q_n^{(2)}$ per $n = 1, 2, \ldots$. Ciò significa che se i sistemi sono molto utilizzati, mantenendosi però in condizioni di non congestione, la probabilità di avere il sistema vuoto nel sistema $M/M/1$ è approssimativamente il doppio di quella del sistema $M/M/2$ mentre la probabilità di avere un certo numero di utenti è approssimativamente uguale nei due sistemi.

$M/M/1$	$M/M/2$
$\lambda_n = \lambda \quad (n = 0, 1, \ldots)$ $\mu_n = 2\,\mu \quad (n = 1, 2, \ldots)$	$\lambda_n = \lambda \quad (n = 0, 1, \ldots)$ $\mu_1 = \mu, \quad \mu_n = 2\,\mu \quad (n = 2, 3, \ldots)$
$\varrho_2 = \lambda/(2\,\mu) < 1$ $\lambda^* = \lambda, \quad \mu^* = 2\,\mu, \quad \varrho^* = \dfrac{\lambda}{2\,\mu}$	$\varrho_2 = \lambda/(2\,\mu) < 1$ $\lambda^* = \lambda, \quad \mu^* = \mu, \quad \varrho^* = \dfrac{\lambda}{2\,\mu}$
$E(T) = \dfrac{1}{\lambda}, \qquad E(S) = \dfrac{1}{2\,\mu}$	$E(T) = \dfrac{1}{\lambda}, \qquad E(S) = \dfrac{1}{\mu}$
$q_n = (1 - \varrho_2)\,\varrho_2^n \quad (n = 0, 1, \ldots)$	$q_0 = \dfrac{1 - \varrho_2}{1 + \varrho_2}, \quad q_n = 2\,\dfrac{1 - \varrho_2}{1 + \varrho_2}\,\varrho_2^n$ $(n = 1, 2, \ldots)$
$E(N) = \dfrac{\varrho_2}{1 - \varrho_2}$	$E(N) = \dfrac{2\,\varrho_2}{1 - \varrho_2^2}$
$\mathrm{Var}(N) = \dfrac{\varrho_2}{(1 - \varrho_2)^2}$	$\mathrm{Var}(N) = \dfrac{2\,\varrho_2\,(1 + \varrho_2^2)}{(1 - \varrho_2^2)^2}$

Tabella 5.3: Confronto tra i parametri prestazionali di un sistema $M/M/2$ e di un sistema $M/M/1$ il cui tempo medio di servizio è dimezzato rispetto ai tempi medi di servizio dei singoli servitori del sistema $M/M/2$.

La Tabella 5.3 riassume i principali risultati ottenuti dal confronto tra i

sistemi $M/M/1$ e $M/M/2$, con il primo sistema avente tempo medio di servizio dimezzato rispetto ai tempi medi di servizio dei singoli servitori del sistema $M/M/2$.

Per comprendere meglio le differenze tra i due sistemi è anche utile analizzare il valore medio e la varianza del numero di utenti. Nella situazione di equilibrio statistico, se denotiamo con $E(N^{(s)})$ e $\mathrm{Var}(N^{(s)})$ rispettivamente il valore medio e la varianza del numero di utenti presenti nel sistema $M/M/s$ $(s = 1, 2)$ si può mostrare che

$$\begin{aligned} E(N^{(1)}) &= \frac{\varrho_2}{1-\varrho_2}\,, \qquad \mathrm{Var}(N^{(1)}) = \frac{\varrho_2}{(1-\varrho_2)^2} \\ E(N^{(2)}) &= \frac{2\,\varrho_2}{1-\varrho_2^2}\,, \qquad \mathrm{Var}(N^{(2)}) = \frac{2\,\varrho_2\,(1+\varrho_2^2)}{(1-\varrho_2^2)^2}. \end{aligned} \tag{5.5}$$

Quindi, se $\varrho_2 < 1$ si ha:

$$E(N^{(1)}) < E(N^{(2)}), \qquad \mathrm{Var}(N^{(1)}) < \mathrm{Var}(N^{(2)}).$$

Tali disuguaglianze mostrano che in media si hanno meno utenti nel sistema $M/M/1$ piuttosto che nel sistema $M/M/2$ con una dispersione dal valore medio inferiore. Inoltre, dalle (5.5) segue:

$$\frac{E(N^{(1)})}{E(N^{(2)})} = \frac{1+\varrho_2}{2}\,, \qquad \frac{\mathrm{Var}(N^{(1)})}{\mathrm{Var}(N^{(2)})} = \frac{(1+\varrho_2)^2}{2\,(1+\varrho_2^2)}$$

che mostrano che quando ϱ_2 è prossimo all'unità si ha $E(N^{(1)}) \sim E(N^{(2)})$ e $\mathrm{Var}(N^{(1)}) \sim \mathrm{Var}(N^{(2)})$. Si nota quindi che se i due sistemi di servizio sono molto utilizzati (ϱ_2 prossimo a uno), in condizioni di non congestione il numero medio di utenti sarà approssimativamente lo stesso in entrambi i sistemi.

In conclusione, tra i due sistemi considerati, *il sistema $M/M/1$ è il più efficiente.*

5.3.2 Secondo confronto

Ci proponiamo di confrontare i due sistemi di servizio illustrati Figura 5.3. Il primo sistema consiste di due sistemi di servizio $M/M/1$ in parallelo indipendenti con frequenze di arrivo $\lambda/2$ e frequenza di partenza μ. Entrambi i sistemi $M/M/1$ non si congestionano se $\varrho_2 = \lambda/(2\,\mu) < 1$. Il secondo sistema consiste di un sistema di servizio $M/M/2$ con frequenze di arrivo λ e frequenze di partenza $\mu_1 = \mu$ e $\mu_n = 2\mu$ $(n = 2, 3, \ldots)$. Tale sistema non si congestiona se $\varrho_2 = \lambda/(2\,\mu) < 1$, essendo presenti due servitori. In condizioni di equilibrio statistico, ossia quando $\varrho_2 < 1$, si desidera stabilire quale dei due sistemi sia più efficiente in base al tempo medio di attesa degli utenti nel sistema.

Denotiamo con $N^{(1)}$ il numero di utenti presenti nel primo sistema (costituito dai due sistemi $M/M/1$ in parallelo) e con $q_n^{(1)}$ la probabilità di avere n utenti complessivamente nel sistema. Inoltre, denotiamo con $N^{(2)}$ il numero di utenti

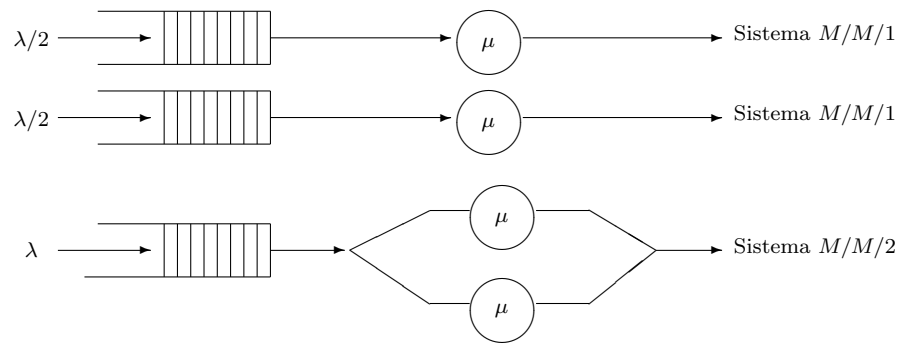

Figura 5.3: Confronto tra il sistema di servizio $M/M/2$ e un sistema di servizio costituito da due sistemi $M/M/1$ indipendenti.

presenti nel sistema $M/M/2$ e sia $q_n^{(2)}$ la probabilità di avere n utenti in tale sistema di servizio.

Essendo i due sistemi $M/M/1$ indipendenti, si ha:

$$q_n^{(1)} = P(N^{(1)} = n) = \sum_{k=0}^{n} q_k\, q_{n-k} = \sum_{k=0}^{n} \left[(1-\varrho_2)\varrho_2^k\right] \left[(1-\varrho_2)\varrho_2^{n-k}\right]$$

$$= (1-\varrho_2)^2 \sum_{k=0}^{n} \varrho_2^n = (n+1)(1-\varrho_2)^2\, \varrho_2^n \qquad (n = 0, 1, \ldots).$$

Si nota che

$$\frac{q_0^{(1)}}{q_0^{(2)}} = (1-\varrho_2)^2\, \frac{1+\varrho_2}{1-\varrho_2} = (1-\varrho_2)(1+\varrho_2) = 1-\varrho_2^2 < 1,$$

che mostra che nella situazione di equilibrio è più probabile trovare il sistema $M/M/2$ vuoto piuttosto che nel sistema costituito da due sistemi di servizio $M/M/1$ indipendenti. Il numero medio di utenti presenti nel primo sistema è

$$E(N^{(1)}) = \sum_{n=1}^{+\infty} n\, q_n^{(1)} = \frac{2\varrho_2}{1-\varrho_2},$$

mentre dalla Tabella 5.2 per il sistema $M/M/2$ segue che

$$E(N^{(2)}) = \frac{2\varrho_2}{1-\varrho_2^2}.$$

Quindi il numero medio di utenti nel sistema di servizio $M/M/2$ è inferiore al numero medio di utenti nel sistema costituito da due sistemi $M/M/1$ indipendenti. Utilizzando poi la prima legge di Little il tempo di attesa di un utente nel primo sistema è

$$E(W^{(1)}) = \frac{E(N^{(1)})}{\lambda} = \frac{1}{\mu(1-\varrho_2)},$$

mentre dalla Tabella 5.2 segue che il tempo medio di attesa nel sistema $M/M/2$ è

$$E(W^{(2)}) = \frac{1}{\mu\,(1-\varrho_2^2)}.$$

Si ha quindi $E(W^{(2)}) < E(W^{(1)})$, ossia il tempo medio di attesa di un utente nel sistema di servizio $M/M/2$ è inferiore al tempo medio di attesa di un utente nel sistema costituito da due sistemi $M/M/1$ indipendenti. Poiché l'utente sceglie casualmente il sistema $M/M/1$ a cui accedere, il tempo medio di attesa nel primo sistema può essere anche così calcolato:

$$E(W^{(1)}) = \frac{1}{2}\,\frac{1}{\mu-\lambda/2} + \frac{1}{2}\,\frac{1}{\mu-\lambda/2} = \frac{1}{\mu(1-\varrho_2)}.$$

In conclusione, *è più efficiente il sistema di servizio* $M/M/2$. Ciò è dovuto alla circostanza che un utente in arrivo nel primo sistema sceglie a caso una delle due file non tenendo conto del numero di utenti già presenti nei due sistemi $M/M/1$ indipendenti.

5.4 Sistema di servizio $M/M/s$

Consideriamo ora un sistema di servizio a capacità infinita con un'unica fila di attesa e s servitori identici illustrato in Figura 5.4.

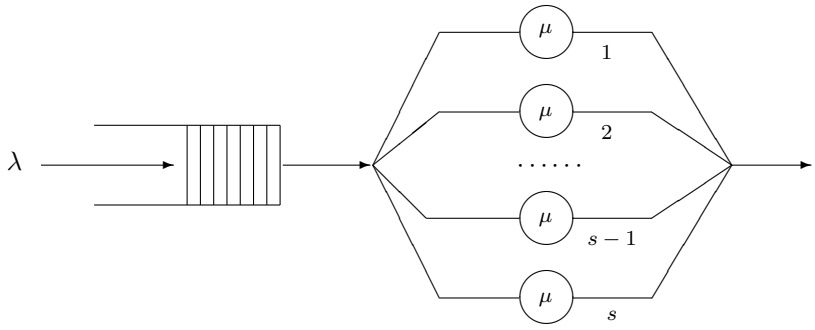

Figura 5.4: Sistema di servizio $M/M/s$.

Supponiamo che i tempi di interarrivo siano indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$ e che i tempi di servizio per ogni servitore siano indipendenti e distribuiti esponenzialmente con valore medio $1/\mu$. Se un utente arriva e trova tutti i servitori occupati si mette in fila di attesa, mentre se trova dei servitori liberi sceglie a caso uno di essi per essere servito. Tale sistema di servizio, noto in letteratura come $M/M/s$, è descrivibile mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri:

$$\lambda_n = \lambda \qquad (n = 0, 1, \ldots)$$

$$\mu_n = \mu \, \min(n, s) = \begin{cases} n\,\mu, & n = 1, 2, \ldots, s \\ s\,\mu, & n = s+1, s+2, \ldots \end{cases} \tag{5.6}$$

In particolare, se $s = 1$ si ha il sistema di servizio $M/M/1$ con unico servitore mentre se $s = 2$ si ottiene il sistema di servizio $M/M/2$ con due servitori.

In primo luogo vediamo in quali condizioni tale sistema raggiunge una situazione di equilibrio statistico. Facendo uso di (5.6) in (3.17) si ha:

$$\begin{aligned} 1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} &= 1 + \sum_{n=1}^{s-1} \frac{1}{n!} \Big(\frac{\lambda}{\mu}\Big)^n + \frac{\lambda^s}{s!\,\mu^s} \sum_{k=0}^{+\infty} \Big(\frac{\lambda}{s\,\mu}\Big)^k \\ &= \sum_{n=0}^{s-1} \frac{s^n}{n!} \Big(\frac{\lambda}{s\,\mu}\Big)^n + \frac{s^s}{s!} \Big(\frac{\lambda}{s\,\mu}\Big)^s \sum_{k=0}^{+\infty} \Big(\frac{\lambda}{s\,\mu}\Big)^k. \end{aligned}$$

Se si pone

$$\varrho_s = \frac{\lambda}{s\,\mu}\,, \tag{5.7}$$

si nota che la serie converge se e solo se $\varrho_s < 1$ e si ha

$$\begin{aligned} 1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} &= \sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{s^s}{s!} \frac{\varrho_s^s}{1-\varrho_s} \\ &= \sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!\,(1-\varrho_s)}\,. \end{aligned}$$

Proposizione 5.2 *Il sistema di servizio $M/M/s$ raggiunge quindi una situazione di equilibrio statistico se e solo se $\varrho_s = \lambda/(s\,\mu) < 1$ e risulta:*

$$q_0 = P(N=0) = \Big[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!\,(1-\varrho_s)}\Big]^{-1},$$

$$q_n = P(N=n) = q_0 \, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \begin{cases} q_0 \, \dfrac{(\lambda/\mu)^n}{n!}, & n = 1, 2, \ldots, s \\ q_0 \, \dfrac{(\lambda/\mu)^n}{s^{n-s}\, s!}, & n = s+1, s+2, \ldots. \end{cases} \tag{5.8}$$

Si noti che ponendo $s = 2$ nella (5.8) si ottiene la (5.2) per il sistema $M/M/2$.

Il parametro $\varrho_s = \lambda/(s\,\mu)$ fornisce una misura di congestione del sistema $M/M/s$. Infatti, se $\varrho_s \geq 1$ il sistema di servizio è *instabile* nel senso che il numero di utenti in fila di attesa è destinato a crescere indefinitamente.

Per decidere il numero di servitori necessari e sufficienti affinché il sistema sia stabile occorre osservare il rapporto λ/μ determinando l'intero positivo s tale che

$$s - 1 \leq \frac{\lambda}{\mu} < s. \tag{5.9}$$

Se $s-1 \leq \lambda/\mu < s$, allora s servitori sono necessari per raggiungere la situazione di equilibrio statistico poiché in tal caso $\varrho_s = \lambda/(s\mu) < 1$ e inoltre sono anche sufficienti nel senso che sarebbe poco economico considerare più di s servitori.

Poiché nel sistema $M/M/s$ il tempo medio di interarrivo è $E(T) = 1/\lambda$ e il tempo medio di servizio per ognuno dei servitori è $E(S) = 1/\mu$, si ha

$$\lambda^* = \frac{1}{E(T)} = \lambda, \qquad \mu^* = \frac{1}{E(S)} = \mu.$$

Quindi l'intensità di traffico relativa al centro di servizio è

$$a = \frac{\lambda^*}{\mu^*} = \frac{\lambda}{\mu}$$

e il fattore di utilizzazione del sistema è:

$$\varrho^* = \frac{\lambda^*}{s\,\mu^*} = \frac{\lambda}{s\,\mu} = \varrho_s\,.$$

Un parametro molto importante per il sistema di servizio $M/M/s$ è rappresentato dalla *probabilità che un utente in arrivo debba attendere nella fila di attesa prima di ricevere il servizio*. È evidente che ciò si verifica se e solo se vi sono almeno s utenti già presenti nel sistema. Tale probabilità, detta *formula C di Erlang*, può essere così ottenuta:

$$\begin{aligned} C[s,\lambda/\mu] = P(N \geq s) &= \sum_{n=s}^{+\infty} q_n = q_0 \sum_{n=s}^{+\infty} \frac{(\lambda/\mu)^n}{s^{n-s}s!} = \frac{q_0}{s!}\left(\frac{\lambda}{\mu}\right)^s \sum_{n=s}^{+\infty} \varrho_s^{n-s} \\ &= q_0\,\frac{(\lambda/\mu)^s}{s!\,(1-\varrho_s)}. \end{aligned} \tag{5.10}$$

Sostituendo q_0, la (5.10) può essere così riscritta:

$$C[s,\lambda/\mu] = P(N \geq s) = \frac{\dfrac{(\lambda/\mu)^s}{s!}}{\dfrac{(\lambda/\mu)^s}{s!} + (1-\varrho_s)\displaystyle\sum_{n=0}^{s-1}\frac{(\lambda/\mu)^n}{n!}}.$$

Per il sistema $M/M/1$ si ha $s = 1$ e $C[1,\lambda/\mu] = P(N \geq 1) = \varrho$, mentre per il sistema $M/M/2$ si ha $s = 2$ e $C[2,\lambda/\mu] = P(N \geq 2) = 2\,\varrho_2^2/(1+\varrho_2)$.

Come vedremo nel seguito i principali parametri prestazionali del sistema $M/M/s$ coinvolgono la formula C di Erlang. Si può infatti mostrare che

$$E(N) = \sum_{n=1}^{+\infty} n\,q_n = \frac{\lambda}{\mu} + \frac{\varrho_s}{1-\varrho_s}\,C[s,\lambda/\mu]\,. \tag{5.11}$$

Dalla prima legge di Little possiamo ricavare il tempo medio di attesa nel sistema

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{1}{\mu} + \frac{1}{s\,\mu}\,\frac{(s\,\varrho_s)^s}{s!\,(1-\varrho_s)^2}\,q_0 = \frac{1}{\mu} + \frac{1}{s\,\mu\,(1-\varrho_s)}\,C[s,\lambda/\mu]. \tag{5.12}$$

Inoltre, in condizioni di equilibrio statistico il tempo medio di permanenza nella fila di attesa è

$$E(Q) = E(W) - E(S) = E(W) - \frac{1}{\mu} = \frac{1}{s\,\mu}\ \frac{(s\,\varrho_s)^s}{s!\,(1-\varrho_s)^2}\ q_0$$
$$= \frac{1}{s\,\mu\,(1-\varrho_s)}\ C[s, \lambda/\mu], \tag{5.13}$$

e dalla seconda legge di Little segue che il numero medio di utenti nella fila di attesa è:

$$E(N_q) = \sum_{n=s}^{+\infty}(n-s)\,q_n = \lambda^*\,E(Q) = \frac{\lambda}{s\,\mu}\ \frac{(s\,\varrho_s)^s}{s!\,(1-\varrho_s)^2}\ q_0$$
$$= \frac{\varrho_s\,(s\,\varrho_s)^s}{s!\,(1-\varrho_s)^2}\ q_0 = \frac{\lambda}{s\,\mu\,(1-\varrho_s)}\ C[s, \lambda/\mu] = \frac{\varrho_s}{1-\varrho_s}\ C[s, \lambda/\mu]. \tag{5.14}$$

Infine, applicando la terza legge di Little si ottiene il numero medio di utenti in servizio (o equivalentemente il numero medio di servitori occupati):

$$E(N_s) = \lambda^*\,E(S) = \frac{\lambda}{\mu} < s.$$

coincide con l'intensità di lavoro svolta dal centro di servizio.

Nella Tabella 5.4 sono riportati i principali parametri prestazionali del sistema $M/M/s$.

Risultati per il sistema M/M/1 e confronti

Supponiamo ora che $s = 1$, ossia riconsideriamo il sistema $M/M/1$. Per $s = 1$ si ha $C[s, \lambda/\mu] = C[1, \lambda/\mu] = \varrho = \lambda/\mu$, che coincide con la probabilità che almeno un utente sia presente nel sistema. Inoltre, per $s = 1$ la (5.11) fornisce il numero medio di utenti presenti nel sistema $M/M/1$, ossia $E(N) = \varrho/(1-\varrho)$, la (5.12) fornisce il tempo medio di attesa nel sistema per il modello $M/M/1$, ossia $E(W) = 1/[\mu\,(1-\varrho)]$, la (5.13) fornisce il tempo di permanenza nella fila di attesa per il sistema $M/M/1$, ossia $E(Q) = \varrho/[\mu\,(1-\varrho)]$, e la (5.14) permette di ottenere il numero medio di utenti presenti nella fila di attesa del sistema $M/M/1$, ossia $E(N_q) = \varrho^2/(1-\varrho)$.

Il confronto tra i sistemi $M/M/1$ e $M/M/2$ può anche essere esteso al confronto tra i sistemi $M/M/1$ e $M/M/s$. Infatti, si può dimostrare che *a parità del processo degli arrivi è meno conveniente scegliere un sistema con s servitori ognuno dei quali ha tempo medio di servizio pari a* $1/\mu$ *rispetto ad un sistema* $M/M/1$ *avente tempo medio di servizio pari a* $1/(s\,\mu)$*, ossia un sistema con un servitore s volte più veloce rispetto a ciascuno dei singoli servitori del sistema* $M/M/s$.

Esempio 5.1 Consideriamo un bar in cui gli utenti arrivano secondo un processo di Poisson con frequenza di 2 utenti al minuto. Supponiamo che i tempi di servizio di ogni servitore siano distribuiti esponenzialmente con tempo medio di servizio di 40 secondi. Se un sistema di servizio $M/M/s$ modella gli utenti del

$$\lambda_n = \lambda \qquad (n = 0, 1, \ldots) \qquad \mu_n = \begin{cases} n\,\mu & n = 1, 2, \ldots, s \\ s\,\mu, & n = s+1, s+2, \ldots \end{cases}$$

$$\varrho_s = \frac{\lambda}{s\,\mu} < 1 \qquad \text{(condizione di equilibrio statistico)}$$

$$q_0 = P(N = 0) = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!\,(1 - \varrho_s)} \right]^{-1},$$

$$q_n = P(N = n) = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \begin{cases} q_0\, \dfrac{(\lambda/\mu)^n}{n!}, & n = 1, 2, \ldots, s \\ q_0\, \dfrac{(\lambda/\mu)^n}{s^{n-s}\, s!}, & n = s+1, s+2, \ldots \end{cases}$$

$$\lambda^* = \lambda, \qquad \mu^* = \mu, \qquad a = \frac{\lambda^*}{\mu^*} = \frac{\lambda}{\mu}, \qquad \varrho^* = \frac{a}{s} = \frac{\lambda}{s\,\mu} = \varrho_s$$

$$C[s, \lambda/\mu] = P(N \geq s) = q_0\, \frac{(\lambda/\mu)^s}{s!\,(1 - \varrho_s)} \qquad \text{(formula C di Erlang)}$$

$$E(N) = \frac{\lambda}{\mu} + \frac{\varrho_s}{1 - \varrho_s}\, C[s, \lambda/\mu], \qquad E(W) = \frac{1}{\mu} + \frac{C[s, \lambda/\mu]}{s\,\mu\,(1 - \varrho_s)}$$

$$E(N_q) = \frac{\varrho_s}{1 - \varrho_s}\, C[s, \lambda/\mu], \qquad E(Q) = \frac{1}{s\,\mu\,(1 - \varrho_s)}\, C[s, \lambda/\mu]$$

$$E(N_s) = \frac{\lambda}{\mu}, \qquad E(S) = \frac{1}{\mu}$$

Tabella 5.4: Parametri prestazionali del sistema di servizio $M/M/s$

bar, determinare il numero di servitori necessari e sufficienti affinché il sistema non si congestioni e calcolare i principali parametri prestazionali del sistema.

In questo caso risulta $1/\mu = 40$ secondi $= 40/60$ minuti e

$$\lambda = 2 \text{ utenti al minuto}, \qquad \mu = \frac{60}{40} = \frac{3}{2} \text{ utenti al minuto},$$

e quindi l'intensità di traffico è $a = \lambda/\mu = 4/3$. Dalla (5.9), ossia $s-1 \leq \lambda/\mu < s$, segue che il numero di servitori necessari e sufficienti per evitare la congestione del bar è $s = 2$. Valutiamo ora i principali parametri prestazionali del sistema $M/M/2$ con $\lambda = 2$ e $\mu = 3/2$ utilizzando la Tabella 5.2. Il fattore di utilizzazione del sistema è $\varrho^* = \lambda/(2\mu) = 2/3 = \varrho_2$. Nella situazione di equilibrio statistico statistico, la distribuzione di equilibrio è

$$q_0 = \frac{1 - \varrho_2}{1 + \varrho_2} = \frac{1}{5} = 0.2, \qquad q_n = 2\, \frac{1 - \varrho_2}{1 + \varrho_2}\, \varrho_2^n = \frac{2}{5} \left(\frac{2}{3} \right)^n \quad (n = 1, 2, \ldots).$$

La probabilità che un utente in arrivo debba attendere, ossia che entrambi i

servitori siano occupati, è fornita dalla formula C di Erlang, ossia:

$$C[2, \lambda/\mu] = P(N \geq 2) = \frac{2\,\varrho_2^2}{1+\varrho_2} = \frac{8}{15} = 0.533.$$

Il numero medio di utenti nel sistema e nella fila di attesa sono:

$$E(N) = \frac{2\,\varrho_2}{1-\varrho_2^2} = \frac{12}{5} = 2.4 \text{ utenti}, \quad E(N_q) = \frac{2\,\varrho_2^3}{1-\varrho_2^2} = \frac{16}{15} = 1.07 \text{ utenti}.$$

Utilizzando le leggi di Little si ottengono i tempi medi di attesa nel sistema e in fila di attesa:

$$E(W) = \frac{E(N)}{\lambda} = \frac{6}{5} = 1.2 \text{ minuti} \qquad E(Q) = \frac{E(N_q)}{\lambda} = \frac{8}{15} = 0.533 \text{ minuti}.$$

Ovviamente, dalla terza legge di Little si può ottenere il numero medio di utenti in servizio:

$$E(N_s) = E(N) - E(N_q) = \frac{\lambda}{\mu} = \frac{4}{3} = 1.333 \text{ utenti},$$

e il tempo medio di servizio per servitore è:

$$E(S) = E(W) - E(Q) = \frac{2}{3} = 0.667 \text{ minuti}.$$

♢

5.5 Sistema di servizio $M/M/s/s$

Consideriamo un sistema di servizio con un'unica fila di attesa, s servitori, capacità finita s, rappresentato in Figura 5.5.

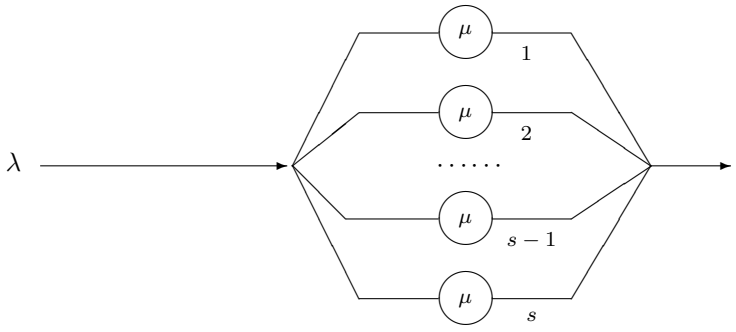

Figura 5.5: Sistema di servizio $M/M/s/s$.

La prima M significa che gli arrivi si verificano secondo un processo di Poisson di parametro λ. La seconda M significa che i tempi di servizio per ognuno dei servitori sono indipendenti ed esponenzialmente distribuiti con valore medio $1/\mu$. Se si verifica un arrivo quando tutti gli s servitori sono occupati la richiesta di servizio è rifiutata e quindi non ha effetto sul sistema. Se invece un utente in arrivo trova dei servitori liberi sceglie a caso uno di essi per essere servito.

Questo sistema di servizio è stato inizialmente proposto dal matematico e statistico danese *Agner Krarup Erlang* (1878–1929) come modello per analizzare il comportamento di un centralino telefonico caratterizzato da s linee disponibili. Le chiamate che arrivano e trovano tutte le s linee occupate sono rifiutate e quindi vengono perdute.

Denotiamo con $N(t)$ il numero di utenti presenti nel sistema $M/M/s/s$ al tempo t. Il processo stocastico $\{N(t), t \geq 0\}$ è descrivibile mediante un processo di nascita–morte caratterizzato da parametri

$$\begin{aligned} \lambda_n &= \begin{cases} \lambda, & n = 0, 1, \ldots, s-1 \\ 0, & n = s, s+1, \ldots \end{cases} \\ \mu_n &= \begin{cases} n\,\mu, & n = 1, 2, \ldots, s \\ 0, & n = s+1, s+2, \ldots \end{cases} \end{aligned} \tag{5.15}$$

Poiché il sistema di servizio $M/M/s/s$ è a capacità finita, raggiunge sempre una situazione di equilibrio statistico. Vogliamo ora determinare tale distribuzione. Facendo uso di (5.15) in (3.17) e ponendo $\varrho_s = \lambda/(s\mu)$ si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \sum_{n=1}^{s} \frac{\lambda^n}{\mu^n\, n!} = \sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!}\cdot$$

Proposizione 5.3 *In condizioni di equilibrio statistico per il sistema $M/M/s/s$ si ha:*

$$\begin{aligned} q_0 &= P(N=0) = \Big[\sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!}\Big]^{-1} \\ q_n &= P(N=n) = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \frac{(\lambda/\mu)^n}{n!}\, q_0\,, \qquad (n = 1, 2, \ldots, s). \end{aligned} \tag{5.16}$$

Tale distribuzione di probabilità è detta *distribuzione di Poisson troncata* di parametro λ/μ.

La *probabilità che un utente in arrivo sia rifiutato* può essere così ottenuta:

$$B[s, \lambda/\mu] = P(N=s) = q_s = \frac{\dfrac{(\lambda/\mu)^s}{s!}}{\displaystyle\sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!}} \tag{5.17}$$

ed è detta *formula B di Erlang*. Come vedremo nel seguito i principali parametri prestazionali del sistema $M/M/s/s$ coinvolgono la formula B di Erlang. In primo luogo, si nota che

$$q_n = P(N = n) = \frac{s!}{n!}\,(s\,\varrho_s)^{n-s}\,B[s, \lambda/\mu] \qquad (n = 0, 1, \ldots, s).$$

La frequenza media di arrivo e la frequenza media di partenza sono:

$$\lambda^* = \sum_{n=0}^{s-1} \lambda_n\, q_n = \lambda \sum_{n=0}^{s-1} q_n = \lambda\,(1 - q_s) = \lambda\,\{1 - B[s, \lambda/\mu]\},$$
$$\mu^* = \mu.$$

Si nota che λ^* è il prodotto della frequenza di arrivo λ e della probabilità che l'utente in arrivo non sia rifiutato. Pertanto l'intensità di traffico relativa al centro di servizio è

$$a = \frac{\lambda^*}{\mu^*} = \frac{\lambda\,(1 - q_s)}{\mu} = \frac{\lambda}{\mu}\,\big\{1 - B[s, \lambda/\mu]\big\}$$

e quindi il fattore di utilizzazione del sistema è

$$\varrho^* = \frac{\lambda^*}{s\,\mu^*} = \frac{\lambda\,(1 - q_s)}{s\mu} = \frac{\lambda}{s\,\mu}\,\{1 - B[s, \lambda/\mu]\}.$$

Nella situazione di equilibrio statistico il numero medio di utenti presenti nella fila di attesa ed il tempo medio di permanenza in coda sono nulli, ossia $E(N_q) = 0$ e $E(Q) = 0$. Poiché il tempo di attesa nel sistema coincide con il tempo di servizio segue che la variabile aleatoria W è distribuita esponenzialmente con valore medio $E(W) = 1/\mu$ e $E(N) = E(N_s)$.

Quindi il numero medio di utenti nel sistema, che coincide con il numero medio di clienti in servizio (numero medio di servitori occupati) è

$$E(N) = E(N_s) = \lambda^*\,E(W) = \frac{\lambda\,\big\{1 - B[s, \lambda/\mu]\big\}}{\mu}.$$

Si nota che il numero medio di clienti nel sistema (ossia il numero medio di servitori occupati) coincide con l'intensità di traffico del centro di servizio.

I principali risultati ottenuti per il modello $M/M/s/s$ sono elencati in Tabella 7.

Esempio 5.2 Si consideri un centralino telefonico con due linee disponibili che non consenta l'attesa di nessuna chiamata. Si supponga che arrivino con una frequenza di 2 chiamate al minuto e che la durata media di una telefonata è di 40 secondi. Se il sistema di servizio $M/M/2/2$ modella il centralino considerato, calcolare i principali parametri prestazionali del sistema.

In questo caso risulta $1/\mu = 40\,\text{secondi} = 40/60\,\text{minuti}$ e

$$\lambda = 2 \text{ telefonate al minuto}, \qquad \mu = \frac{60}{40} = \frac{3}{2} \text{ telefonate al minuto},$$

$$\lambda_n = \begin{cases} \lambda, & n = 0, 1, \ldots, s-1 \\ 0, & n = s, s+1, \ldots \end{cases} \qquad \mu_n = \begin{cases} n\,\mu, & n = 1, 2, \ldots, s \\ 0, & n = s+1, s+2, \ldots \end{cases}$$

$$q_0 = \left[\sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!}\right]^{-1} \qquad q_n = \frac{(\lambda/\mu)^n}{n!}\, q_0\,, \qquad (n = 1, 2, \ldots, s)$$

$$\lambda^* = \lambda\,\{1 - B[s, \lambda/\mu]\}, \qquad \mu^* = \mu,$$

$$a = \frac{\lambda}{\mu}\,\{1 - B[s, \lambda/\mu]\}, \qquad \varrho^* = \frac{\lambda}{s\,\mu}\,\{1 - B[s, \lambda/\mu]\}$$

$$B[s, \lambda/\mu] = q_s = \frac{(\lambda/\mu)^s/s!}{\sum_{n=0}^{s} (\lambda/\mu)^n/n!} \qquad \text{(formula B di Erlang)}$$

$$E(N_q) = 0, \qquad E(Q) = 0$$

$$E(N) = E(N_s) = \frac{\lambda}{\mu}\,\{1 - B[s, \lambda/\mu]\}$$

$$E(W) = E(S) = \frac{1}{\mu}$$

Tabella 5.5: Parametri prestazionali del sistema di servizio $M/M/s/s$.

e quindi l'intensità di traffico è $a = \lambda/\mu = 4/3$. La probabilità che in condizioni di equilibrio si abbiano k telefonate ($k = 0, 1, 2$) è:

$$q_0 = P(N=0) = \left[1 + \frac{(\lambda/\mu)}{1!} + \frac{(\lambda/\mu)^2}{2!}\right]^{-1} = \left[1 + \frac{4}{3} + \frac{8}{9}\right]^{-1} = \frac{9}{29},$$
$$q_1 = P(N=1) = \frac{\lambda}{\mu}\, q_0 = \frac{4}{3} \cdot \frac{9}{29} = \frac{12}{29},$$
$$q_2 = P(N=2) = \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 q_0 = \frac{16}{18} \cdot \frac{9}{29} = \frac{8}{29}\,.$$

La formula B di Erlang fornisce la probabilità che un utente in arrivo sia rifiutato:

$$B[s, \lambda/\mu] = B[2, 4/3] = q_2 = \frac{8}{29} = 0.2759.$$

Il coefficiente di utilizzazione del sistema è quindi:

$$\varrho^* = \varrho_s\{1 - B[s, \lambda/\mu]\} = \frac{2}{3} \cdot \frac{21}{29} = \frac{14}{29} = 0.4828.$$

In condizioni di equilibrio, il numero medio di linee occupate e la durata di ogni telefonata sono:

$$E(N) = \frac{\lambda}{\mu}\,\{1 - B[s, \lambda/\mu]\} = \frac{4}{3}\left(1 - \frac{8}{29}\right) = \frac{4}{3} \cdot \frac{21}{29} = \frac{28}{29} = 0.9656 \text{ telefonate},$$

$$E(W) = \frac{1}{\mu} = \frac{2}{3} = 0.66666 \text{ minuti.}$$

Si nota che in media è occupata una sola linea telefonica. ♢

5.6 Sistema di servizio $M/M/\infty$

Consideriamo un sistema di servizio a capacità infinita con un'unica fila di attesa e infiniti servitori. Supponiamo che i tempi di interarrivo siano indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$ e che i tempi di servizio per ognuno dei servitori siano indipendenti e distribuiti esponenzialmente con valore medio $1/\mu$. Poiché esistono infiniti servitori un utente che arriva può essere immediatamente servito, ossia il suo tempo di attesa nel sistema è uguale al suo tempo di servizio. Tale sistema è noto in letteratura come sistema $M/M/\infty$.

Il sistema $M/M/\infty$ è descrivibile mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri

$$\lambda_n = \lambda \quad (n = 0, 1, \ldots), \qquad \mu_n = n\,\mu \qquad (n = 1, 2, \ldots). \tag{5.18}$$

Tale sistema costituisce una buona approssimazione per molti sistemi reali del tipo self–service, quali grandi parcheggi, cinema, supermarket, ...; in tali casi si può ipotizzare che un utente in arrivo sia immediatamente servito.

Facendo uso di (5.18) in (3.17) si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \sum_{n=1}^{+\infty} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n = \exp\Big\{\frac{\lambda}{\mu}\Big\}.$$

Essendo la serie sempre convergente, il sistema $M/M/\infty$ raggiunge sempre una situazione di equilibrio statistico e si ha

$$q_0 = P(N = 0) = \exp\Big\{-\frac{\lambda}{\mu}\Big\},$$
$$q_n = P(N = n) = q_0\, \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n \exp\Big\{-\frac{\lambda}{\mu}\Big\}, \qquad (n = 1, 2, \ldots).$$

Proposizione 5.4 *Per il sistema $M/M/\infty$, in condizioni di equilibrio si ha:*

$$q_n = P(N = n) = \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n \exp\Big\{-\frac{\lambda}{\mu}\Big\}, \qquad (n = 0, 1, \ldots), \tag{5.19}$$

ossia una funzione di probabilità di Poisson di parametro λ/μ.

Il valore medio e la varianza del numero di utenti nel sistema sono:

$$E(N) = \frac{\lambda}{\mu}\,, \qquad \mathrm{Var}(N) = \frac{\lambda}{\mu}\cdot$$

Poiché nel sistema $M/M/\infty$ il tempo medio di interarrivo è $E(T) = 1/\lambda$ e il tempo medio di servizio per ognuno dei servitori è $E(S) = 1/\mu$ si ha

$$\lambda^* = \frac{1}{E(T)} = \lambda, \qquad \mu^* = \frac{1}{E(S)} = \mu.$$

L'intensità di traffico relativa al centro di servizio è quindi:

$$a = \frac{\lambda^*}{\mu^*} = \frac{\lambda}{\mu} < +\infty$$

e il fattore di utilizzazione del sistema è $\varrho^* = 0$.

Nella situazione di equilibrio statistico il numero medio di utenti presenti nella fila di attesa e il tempo medio di permanenza in coda sono nulli, ossia $E(N_q) = 0$ e $E(Q) = 0$. Poiché il tempo di attesa nel sistema coincide con il tempo di servizio, la variabile aleatoria W è distribuita esponenzialmente con valore medio $E(W) = 1/\mu$. Pertanto la densità di probabilità di W è:

$$f_W(t) = \begin{cases} \mu\, e^{-\mu\, t}, & t > 0 \\ 0, & \text{altrimenti.} \end{cases}$$

Il numero medio di utenti nel sistema, coincidente con il numero medio di utenti in servizio (numero medio di servitori occupati), è ottenibile dalle leggi di Little:

$$E(N) = E(N_s) = \lambda^*\, E(W) = \lambda^*\, E(S) = \frac{\lambda}{\mu}\,.$$

Si nota nuovamente che il numero medio di utenti nel sistema (ossia il numero medio di servitori occupati) coincide con l'intensità di traffico del centro di servizio.

Nel sistema $M/M/\infty$ i periodi di ozio del centro di servizio, che possono essere riguardati come tempi residui dei tempi di interarrivo, sono indipendenti e distribuiti esponenzialmente con valore medio $1/\lambda$. Pertanto la densità di probabilità della variabile aleatoria I descrivente un periodo di ozio è:

$$f_I(t) = \begin{cases} \lambda\, e^{-\lambda\, t}, & t > 0 \\ 0, & \text{altrimenti} \end{cases}$$

e quindi

$$E(B) = \frac{(1 - q_0)\, E(I)}{q_0} = \frac{(1 - q_0)}{\lambda\, q_0} = \frac{1 - e^{-\lambda/\mu}}{\lambda\, e^{-\lambda/\mu}} = \frac{1}{\lambda}\left(e^{\lambda/\mu} - 1\right).$$

Si nota che $E(B) > E(I)$, ossia il tempo medio di occupazione è maggiore del tempo medio di ozio del servitore.

I principali parametri prestazionali del sistema $M/M/\infty$ sono indicati in Tabella 5.6.

$$\lambda_n = \lambda \qquad (n = 0, 1, \ldots) \qquad \mu_n = n\,\mu \qquad (n = 1, 2, \ldots)$$

$$q_n = P(N = n) = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left\{-\frac{\lambda}{\mu}\right\}, \qquad (n = 0, 1, \ldots)$$

$$\lambda^* = \lambda, \qquad \mu^* = \mu, \qquad a = \frac{\lambda}{\mu}, \qquad \varrho^* = 0$$

$$E(N) = \frac{\lambda}{\mu}, \qquad E(W) = \frac{1}{\mu}$$

$$E(N_q) = 0, \qquad E(Q) = 0$$

$$E(N_s) = E(N) = \frac{\lambda}{\mu}, \qquad E(S) = \frac{1}{\mu}$$

$$E(I) = \frac{1}{\lambda}, \qquad E(B) = \frac{1}{\lambda}\left[\exp\left\{\frac{\lambda}{\mu}\right\} - 1\right]$$

Tabella 5.6: Parametri prestazionali del sistema di servizio $M/M/\infty$.

Esempio 5.3 Si desidera modellare il numero di utenti connessi simultaneamente ad una rete con un sistema $M/M/\infty$. Gli utenti accedono alla rete secondo un processo di Poisson con una frequenza media di 500 utenti all'ora e restano connessi alla rete in media per 20 minuti. Determinare i parametri prestazionali del sistema.
In questo caso $1/\mu = 20\,\text{minuti} = 20/60\,\text{ore}$ e

$$\lambda = 500 \text{ utenti all'ora}, \qquad \mu = \frac{60}{20} = 3 \text{ utenti all'ora},$$

e quindi l'intensità di traffico è $a = \lambda/\mu = 500/3 = 166.667$. Dalla (5.9), ossia $s - 1 \leq \lambda/\mu < s$, segue che il numero di servitori necessari e sufficienti in un sistema $M/M/s$ per evitare la congestione è $s = 167$, ossia occorrono un numero molto elevato di servitori. Per descrivere la rete si può quindi utilizzare un sistema $M/M/\infty$, con $\lambda = 500$ e $\mu = 3$. Dalla Tabella 5.5 segue che la distribuzione di equilibrio è

$$q_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left\{-\frac{\lambda}{\mu}\right\} = \frac{1}{n!}\left(\frac{500}{3}\right)^n \exp\left\{-\frac{500}{3}\right\} \qquad (n = 0, 1, \ldots).$$

Il numero medio di connessioni è $E(N) = E(N_s) = \lambda/\mu = 500/3 = 166.667$ e il tempo medio di connessione è $E(W) = E(S) = 1/\mu = 1/3$ di ora, ossia 20 minuti. ♢

5.7 Sistema con accelerazione del servizio

Il processo di nascita morte caratterizzato da parametri (5.18), ossia da parametri $\lambda_n = \lambda$ $(n = 0, 1, \ldots)$ e $\mu_n = n\,\mu$ $(n = 1, 2, \ldots)$, si può anche interpretare

come descrivente un sistema di servizio a capacità infinita con un'unica fila di attesa e un unico servitore che accelera il suo servizio all'aumentare della lunghezza della coda in maniera tale da soddisfare tutte le richieste degli utenti.

Il sistema di servizio con accelerazione del servizio ammette la stessa distribuzione di equilibrio del sistema $M/M/\infty$, ossia una distribuzione di Poisson di parametro λ/μ, fornita in (5.19). Il valore medio e la varianza del numero di utenti nel sistema sono quindi:

$$E(N) = \frac{\lambda}{\mu}\,, \qquad \mathrm{Var}(N) = \frac{\lambda}{\mu}\,\cdot$$

Anche se il sistema $M/M/\infty$ e quello con accelerazione del servizio possiedono la stessa distribuzione di equilibrio, hanno alcuni parametri prestazionali diversi. Infatti, per il sistema con accelerazione del servizio, la frequenza media di arrivo per unità di tempo è uguale a quella del sistema $M/M/\infty$, ossia $\lambda^* = \lambda$, mentre la frequenza media di partenza per unità di tempo è:

$$\mu^* = \frac{1}{1-q_0}\sum_{n=1}^{+\infty}\mu_n\, q_n = \frac{\mu}{1-q_0}\sum_{n=1}^{+\infty} n\, q_n = \frac{\mu}{1-q_0}\,\frac{\lambda}{\mu} = \frac{\lambda}{1-\exp\{-\lambda/\mu\}}\cdot$$

L'intensità di traffico, che coincide con il fattore di utilizzazione del sistema, è quindi:

$$a = \varrho^* = 1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\} < 1.$$

Dalla prima legge di Little segue che il tempo medio di attesa nel sistema è

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{1}{\mu}\cdot$$

Il numero medio di utenti in fila di attesa è

$$E(N_q) = \sum_{n=1}^{+\infty}(n-1)\, q_n = E(N) - (1-q_0) = \frac{\lambda}{\mu} - 1 + \exp\Big\{-\frac{\lambda}{\mu}\Big\}$$

e quindi dalla seconda legge di Little si ottiene il tempo medio di permanenza nella fila di attesa:

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{1}{\mu} - \frac{1}{\lambda}\Big[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\Big].$$

Il numero medio di utenti in servizio è

$$E(N_s) = E(N) - E(N_q) = 1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}$$

e coincide con l'intensità di traffico. Pertanto, dalla terza legge di Little otteniamo:

$$E(S) = \frac{E(N_s)}{\lambda^*} = \frac{1}{\lambda}\Big[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\Big].$$

$$\lambda_n = \lambda \qquad (n = 0, 1, \ldots) \qquad \mu_n = n\,\mu \qquad (n = 1, 2, \ldots)$$

$$q_n = P(N = n) = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left\{-\frac{\lambda}{\mu}\right\}, \qquad (n = 0, 1, \ldots)$$

$$\lambda^* = \lambda, \qquad \mu^* = \frac{\lambda}{1 - \exp\{-\lambda/\mu\}}, \qquad \varrho^* = 1 - \exp\left\{-\frac{\lambda}{\mu}\right\} < 1$$

$$E(N) = \frac{\lambda}{\mu}, \qquad E(W) = \frac{1}{\mu},$$

$$E(N_q) = \frac{\lambda}{\mu} - 1 + \exp\left\{-\frac{\lambda}{\mu}\right\}, \qquad E(Q) = \frac{1}{\mu} - \frac{1}{\lambda}\left[1 - \exp\left\{-\frac{\lambda}{\mu}\right\}\right]$$

$$E(N_s) = 1 - \exp\left\{-\frac{\lambda}{\mu}\right\} = \varrho^*, \qquad E(S) = \frac{E(N_s)}{\lambda^*} = \frac{1}{\lambda}\left[1 - \exp\left\{-\frac{\lambda}{\mu}\right\}\right]$$

Tabella 5.7: Parametri prestazionali del sistema con unico servitore che velocizza il suo servizio.

I principali parametri prestazionali del sistema con accelerazione del servizio sono indicati in Tabella 5.7.

Si nota che alcuni parametri prestazionali del sistema $M/M/\infty$ e di quello con accelerazione del servizio sono differenti, anche se entrambi i modelli sono caratterizzati dalla stessa distribuzione di equilibrio; ciò dipende dal fatto che nel modello con accelerazione del servizio è previsto un unico servitore, mentre nel modello $M/M/\infty$ si considerano infiniti servitori. Quindi, prima di determinare i parametri prestazionali di un sistema di servizio occorre sempre precisare il numero di servitori.

5.8 Sistema di servizio con scoraggiamento

Consideriamo un sistema di servizio a capacità infinita con unica fila di attesa. Supponiamo che gli utenti siano scoraggiati da una lunga coda e che il tempo di servizio del generico utente sia distribuito esponenzialmente con valore medio $1/\mu$.

Tale sistema può essere descritto mediante un processo di nascita–morte $\{N(t),\, t \geq 0\}$ caratterizzato da parametri

$$\lambda_n = \frac{\lambda}{n+1} \quad (n = 0, 1, \ldots), \qquad \mu_n = \mu \qquad (n = 1, 2, \ldots). \tag{5.20}$$

Vogliamo ora vedere in quali condizioni tale sistema raggiunge una situazione di equilibrio statistico. Facendo uso di (5.20) in (3.17) si ha:

$$1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} = 1 + \sum_{n=1}^{+\infty} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n = \exp\left\{\frac{\lambda}{\mu}\right\}.$$

Poiché tale serie è sempre convergente, il sistema considerato raggiunge sempre una situazione di equilibrio statistico e risulta

$$q_0 = \exp\Big\{-\frac{\lambda}{\mu}\Big\},$$

$$q_n = q_0\,\frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n} = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\Big\{-\frac{\lambda}{\mu}\Big\} \qquad (n = 1, 2, \ldots).$$

Si è ottenuta la stessa distribuzione di equilibrio (5.19) del sistema $M/M/\infty$, ossia una distribuzione di Poisson di parametro λ/μ. Pertanto, nella situazione di equilibrio statistico, il valore medio e la varianza e il coefficiente di variazione del numero di utenti presenti nel sistema sono

$$E(N) = \frac{\lambda}{\mu}\,, \qquad \mathrm{Var}(N) = \frac{\lambda}{\mu}\cdot$$

$$\lambda_n = \frac{\lambda}{n+1} \qquad (n = 0, 1, \ldots) \qquad \mu_n = \mu \qquad (n = 1, 2, \ldots)$$

$$q_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\Big\{-\frac{\lambda}{\mu}\Big\}, \qquad (n = 0, 1, \ldots)$$

$$\lambda^* = \mu\left[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\right], \qquad \mu^* = \mu, \qquad \varrho^* = 1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}$$

$$E(N) = \frac{\lambda}{\mu}\,, \qquad E(W) = \frac{\lambda}{\mu^2}\left[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\right]^{-1}$$

$$E(N_q) = \frac{\lambda}{\mu} - 1 + \exp\Big\{-\frac{\lambda}{\mu}\Big\}, \qquad E(Q) = \frac{\lambda}{\mu^2}\left[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\right]^{-1} - \frac{1}{\mu}$$

$$E(N_s) = 1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\} = \varrho^*, \qquad E(S) = \frac{1}{\mu}\cdot$$

Tabella 5.8: Parametri prestazionali del sistema di servizio con scoraggiamento degli utenti e unico servitore.

Anche se il sistema $M/M/\infty$ e quello con scoraggiamento possiedono la stessa distribuzione di equilibrio, essi sono fondamentalmente diversi sia nell'evoluzione transiente sia relativamente ai parametri prestazionali. Infatti, si ha:

$$\lambda^* = \sum_{n=0}^{+\infty} \lambda_n\, q_n = \sum_{n=0}^{+\infty} \frac{\lambda}{n+1}\,\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\Big\{-\frac{\lambda}{\mu}\Big\}$$

$$= \mu\,\exp\Big\{-\frac{\lambda}{\mu}\Big\} \sum_{n=0}^{+\infty} \frac{1}{(n+1)!}\left(\frac{\lambda}{\mu}\right)^{n+1} = \mu\,\exp\Big\{-\frac{\lambda}{\mu}\Big\} \sum_{k=1}^{+\infty} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k$$

$$= \mu \exp\Big\{-\frac{\lambda}{\mu}\Big\}\Big[\exp\Big\{\frac{\lambda}{\mu}\Big\} - 1\Big] = \mu\left[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\right],$$

$$\mu^* = \mu,$$

da cui l'intensità di traffico, che coincide con il fattore di utilizzazione del sistema, è

$$\varrho^* = \frac{\lambda^*}{\mu^*} = 1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}.$$

Dalla prima legge di Little si ricava:

$$E(W) = \frac{E(N)}{\lambda^*} = \frac{\lambda}{\mu^2}\left[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\right]^{-1}.$$

Osserviamo inoltre che Il numero medio di utenti in fila di attesa è

$$E(N_q) = \sum_{n=1}^{+\infty}(n-1)\,q_n = E(N) - (1 - q_0) = \frac{\lambda}{\mu} - 1 + \exp\Big\{-\frac{\lambda}{\mu}\Big\},$$

da cui, facendo ricorso alla seconda legge di Little, si ottiene il tempo medio di permanenza in fila di attesa:

$$E(Q) = \frac{E(N_q)}{\lambda^*} = \frac{\lambda}{\mu^2}\left[1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}\right]^{-1} - \frac{1}{\mu}$$

Infine, il numero medio di utenti in servizio è dato da

$$E(N_s) = E(N) - E(N_q) = 1 - \exp\Big\{-\frac{\lambda}{\mu}\Big\}$$

che coincide con l'intensità di traffico. Dalla terza legge di Little segue infine che

$$E(S) = \frac{E(N_s)}{\lambda^*} = \frac{1}{\mu}.$$

I principali parametri prestazionali del sistema di servizio con scoraggiamento degli utenti e unico servitore sono forniti in Tabella 5.8.

Si nota che il modello con accelerazione del servizio e quello con scoraggiamento hanno la stessa intensità di traffico, anche se le frequenze medie λ^* e μ^* sono differenti. Poiché le leggi di Little coinvolgono le frequenze medie di arrivo, i due sistemi hanno differenti parametri prestazionali

In conclusione, la distribuzione di equilibrio non è sufficiente per comprendere tutte le caratteristiche di un sistema di servizio. Dalle Tabelle 5.5, 5.6 e 5.8 si nota che il sistema $M/M/\infty$, il sistema con accelerazione del servizio e quello con scoraggiamento, anche se caratterizzati dalla stessa distribuzione di equilibrio, hanno parametri prestazionali fondamentalmente diversi. Ciò è dovuto al numero di servitori (infiniti nel modello $M/M/\infty$ e unico nei modelli con accelerazione del servizio e con scoraggiamento degli utenti) e a differenti frequenze medie di arrivo e di partenza per unità di tempo.

Capitolo 6

Simulazione

6.1 Introduzione alla simulazione

La teoria delle file di attesa ricorre a *modelli probabilistici* per ottenere i parametri prestazionali del sistema di servizio. Spesso si rivela difficile studiare tali modelli sia per le caratteristiche delle distribuzioni degli intervalli di interarrivo e dei tempi di servizio e sia per natura della disciplina di servizio. Inoltre, con i modelli probabilistici si riesce ad analizzare il sistema in condizioni di equilibrio statistico, mentre spesso si è interessati al comportamento del sistema di servizio anche nella sua fase transiente.

Per superare tali difficoltà si ricorre spesso a modelli di sistemi di servizio che utilizzano *tecniche di simulazione*. La simulazione consiste nel *riprodurre al computer il comportamento del sistema in esame*. Si basa sulla definizione di un modello, detto *modello di simulazione*, che descriva l'evoluzione del sistema nel tempo. Mediante la simulazione è possibile osservare il *comportamento dinamico del sistema* fornendo informazioni sulle sue prestazioni. Infatti, la simulazione permette di ottenere *stime* (ad esempio, *medie e varianze campionarie*) del tempo di permanenza nella fila di attesa, del tempo di attesa nel sistema, del numero di utenti in fila di attesa e nel sistema, del tempo di ozio e di occupazione del centro di servizio mediante l'analisi dell'evoluzione temporale del sistema.

Entrambi gli approcci (probabilistico e di simulazione) richiedono l'utilizzazione di un modello che permetta, in base ai parametri di input, di ottenere gli indici di prestazione del sistema.

Con un *modello probabilistico* di un sistema di servizio ci si prefigge di ottenere *soluzioni analitiche* e pertanto le ipotesi di base del modello sono semplificate in maniera tale da superare le difficoltà di natura matematica. Nel modello probabilistico si ottengono delle formule che permettono di esprimere gli indici di prestazione del sistema in funzione dei parametri di input (tempi medi di interarrivo, tempi medi di servizio, numero di servitori, ...) e delle caratteristiche

del sistema. Tali formule sono utilizzabili in maniera veloce ed efficiente per un ampio intervallo di valori dei parametri di input e permettono di interpretare qualitativamente il comportamento del sistema e di individuare le condizioni sui parametri che garantiscono il raggiungimento dell'equilibrio statistico.

In un *modello di simulazione*, invece, si possono includere un maggiore numero di caratteristiche significative, spesso rispecchiando meglio il comportamento del sistema reale. La simulazione comunque necessita di *lunghi periodi di esecuzione* per ottenere stime degli indici di prestazione e richiede anche indagini approfondite per l'individuazione delle condizioni che garantiscono il raggiungimento dell'equilibrio statistico.

Non è sempre semplice decidere se utilizzare modelli probabilistici o di simulazione per analizzare un sistema di servizio. I modelli di simulazione si rivelano particolarmente utili nello studio di sistemi di servizio complessi. Il loro utilizzo deve basarsi su due elementi essenziali: "*adeguatezza*" e "*semplicità d'uso*". Uno stesso sistema può essere descritto utilizzando diversi tipi di modelli; pertanto, un modello è tanto più adeguato alla descrizione del sistema di servizio reale quanto meglio rappresenta gli aspetti del sistema che sono di interesse per chi sta effettuando lo studio. La simulazione è quindi un metodo alternativo per la descrizione di un sistema di servizio e può essere utile se riesce a fornire soluzioni significative e di facile interpretazione da parte di coloro che saranno addetti all'utilizzazione del modello ad un costo competitivo rispetto ad altre tecniche di natura probabilistica.

Quando si decide di adottare tecniche di simulazione occorre pianificare l'esperimento in più fasi successive:

(i) formulazione del problema e del modello di simulazione;

(ii) acquisizione dei dati del sistema reale;

(iii) stima e verifica dei parametri e delle caratteristiche operative del sistema reale;

(iv) formulazione del programma di simulazione;

(v) progettazione degli esperimenti e analisi dei risultati.

(i) **Formulazione del problema e del modello di simulazione**
Formulare il problema significa *fissare gli obiettivi di studio* e *stabilire dei criteri per esaminare le soluzioni al problema.* Gli *obiettivi* di un esperimento di simulazione sono essenzialmente di due tipi:

(a) studio del problema di dimensionamento;

(b) studio degli effetti del cambiamento dei parametri o delle caratteristiche funzionali sul comportamento del sistema.

Relativamente al punto *(a)* ci si pone il problema di *stabilire il numero di centri di servizio in parallelo necessari per ottenere migliori prestazioni* oppure di

individuare come organizzare in modo efficiente il servizio in più fasi successive in una catena di produzione.

Con riferimento al punto *(b)* si desidera stabilire gli effetti dovuti al cambiamento dei parametri, quali i tempi medi di interarrivo degli utenti o i tempi medi di servizio per ognuno dei servitori, oppure se l'introduzione di una nuova apparecchiatura in una catena di produzione può permettere di migliorare i tempi di produzione.

Occorre infine individuare un modello di simulazione atto a descrivere in modo astratto il sistema reale

- precisando i *parametri di input*;

- individuando le *grandezze ritenute significative per la descrizione del sistema*;

- specificando *gli indici di prestazione di interesse.*

(ii) **Acquisizione dei dati del sistema reale**
L'acquisizione dei dati consiste nell'effettuare un'*analisi preliminare sul sistema reale per rilevare i dati* su cui si baserà la scelta del modello, la stima dei parametri e delle caratteristiche operative e anche la decisione sui componenti e sulle variabili da introdurre nel modello di simulazione. Occorrerà, ad esempio, effettuare opportuni campionamenti sui tempi di interarrivo degli autoveicoli ai caselli autostradali in varie ore della giornata, oppure sui tempi di produzione di particolari articoli in un'industria.

(iii) **Stima e verifica dei parametri e delle caratteristiche operative del sistema reale**
Questo passo è inteso a trasferire nel modello i parametri e le caratteristiche operative del sistema reale, stimate sulla base dei dati raccolti nel punto *(ii)*. Essa si traduce nell'applicare delle *tecniche statistiche per la stima del valore medio e della varianza* di una distribuzione di probabilità di tipo noto e talora di distribuzioni di probabilità non note. Spesso interessa anche stabilire quale sia il *tipo di distribuzione di probabilità più adeguata* da utilizzare.

Una volta che, sulla base dei dati raccolti, si è formulato un modello e si sono scelti i parametri e le caratteristiche funzionali è necessaria una valutazione della sua adeguatezza per descrivere il sistema reale prima di procedere alla costruzione del simulatore. Quindi, costruito il modello è opportuno effettuare degli opportuni *test di verifica di ipotesi statistiche sia sui parametri sia sulle distribuzioni di probabilità.* Ad esempio, se i tempi di interarrivo sono simulati mediante un opportuno generatore pseudocasuale, la valutazione richiederà l'analisi delle *proprietà statistiche del generatore prescelto* per stabilire se esso ha prodotto sequenze pseudocasuali adeguate al tipo di distribuzione di probabilità e ai suoi parametri.

(iv) **Formulazione del programma di simulazione**
Questo passo consiste nella traduzione del modello di simulazione in un modello interpretabile dall'elaboratore. Esso comprende la *stesura di un programma di simulazione* che descriva la successione logica delle operazioni necessarie per

analizzare il modello nella fase transiente (per descrivere l'evoluzione temporale del sistema), la scelta del linguaggio di programmazione e la scelta dello stato iniziale, ossia dei valori da assegnare inizialmente alle variabili del programma.

Occorre anche verificare se il simulatore si comporta come previsto e rifletta la situazione reale. Infatti, nella simulazione possono essere intervenuti errori di programmazione, errori di arrotondamento, problemi di convergenza, ...

(v) **Progettazione degli esperimenti e analisi dei risultati**
In tale passo occorre *progettare gli esperimenti da effettuare*. In particolare, occorre stabilire *quante esecuzioni sono necessarie per ogni esperimento* e *come effettuare le misure al termine della simulazione*. Aumentando il numero di esecuzioni aumenta la significatività statistica dei risultati, ma aumenta anche considerevolmente il costo computazionale stesso della simulazione. Infine, occorre effettuare un'analisi delle misure ottenute mediante gli esperimenti di simulazione in maniera da comprendere il comportamento del sistema reale.

Nella simulazione gioca un ruolo essenziale *l'inferenza statistica*, con particolare riguardo alla *stima dei parametri* e alla *verifica delle ipotesi*.

6.2 Classificazione dei simulatori

Esistono vari modi di classificare i simulatori.
• Una *prima classificazione* distingue i simulatori in *statici* e *dinamici*.
→ Nei *simulatori statici* la variabile temporale non gioca alcun ruolo e lo scopo fondamentale è quello di stimare alcune caratteristiche utilizzando prove ripetute indipendenti. Tipici esempi sono i simulatori che utilizzano il metodo di Monte Carlo.
→ Nei *simulatori dinamici* si descrive l'evoluzione temporale del modello e quindi il tempo diventa la variabile principale. Lo scopo della simulazione è raccogliere dati statistici su processi che evolvono nel tempo. A differenza delle prove ripetute viene a mancare l'indipendenza e occorre prendere in esame la correlazione delle osservazioni. Inoltre, il processo osservato può raggiungere o meno una situazione di equilibrio statistico. Poiché a priori non si hanno informazioni sull'evoluzione del sistema, occorre analizzare la fase transiente per decidere se si raggiungerà una situazione di stabilità.
• Una *seconda classificazione* distingue i simulatori in *deterministici* e *casuali*.
→ I *simulatori deterministici* sono basati su modelli la cui evoluzione è univocamente determinata una volta fissati i parametri di input. Un esempio tipico è un sistema di servizio $D/D/1$ con tempi di interarrivo e di servizio deterministici; infatti, assegnati i tempi di interarrivo e di servizio, di durata costante, l'evoluzione del sistema è completamente specificata.
→ I *simulatori casuali* sono basati su modelli che includono variabili aleatorie o processi stocastici e necessitano della generazione di variabili aleatorie; l'evoluzione del modello dipende dai parametri di input e dalla generazione di una o più variabili aleatorie. Ad esempio, nel sistema di servizio $D/M/1$ i tempi di interarrivo sono di durata costante, mentre i tempi di servizio debbono essere generati tramite la simulazione di una variabile aleatoria esponenziale.

• Una *terza classificazione* distingue i simulatori in *sincroni* e *asincroni* in base alle modalità con cui viene descritto il trascorrere del tempo.
→ Nei *simulatori sincroni* il tempo di simulazione viene suddiviso in tanti intervalli di uguale ampiezza. All'inizio (o alla fine) di ciascuno di questi intervalli

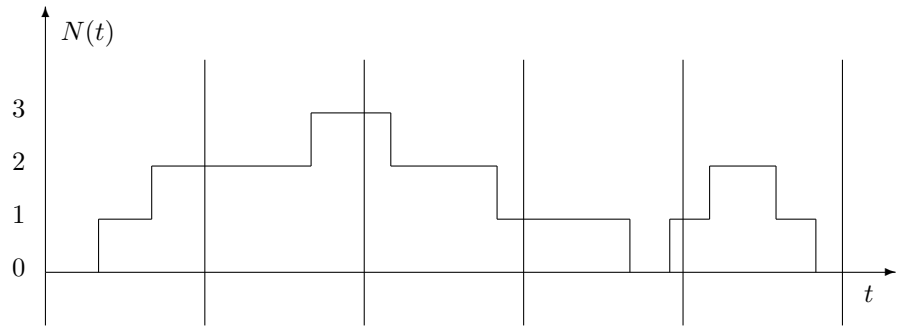

Figura 6.1: Effetto della discretizzazione con passo grande in un simulatore sincrono.

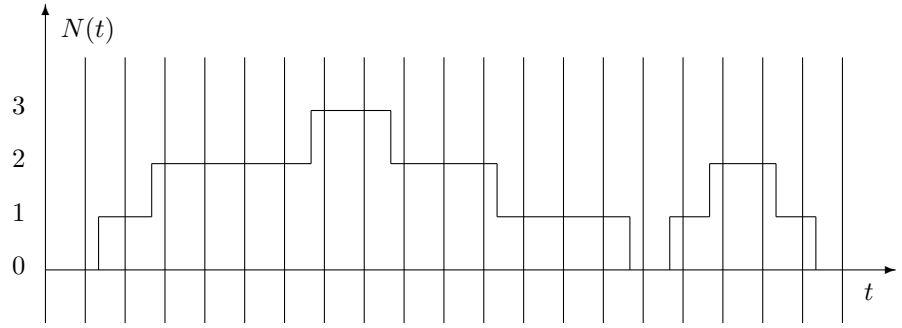

Figura 6.2: Effetto della discretizzazione con passo piccolo in un simulatore sincrono.

si determinano gli eventuali cambiamenti di stato del sistema. In tali simulatori è importante *scegliere accuratamente il passo di discretizzazione* dell'asse temporale. Infatti, se si sceglie un *passo molto piccolo* si rischia di aggiornare troppo spesso e inutilmente le variabili, la durata della simulazione diventa elevata e aumenta il costo computazionale. Invece, un *passo di discretizzazione molto grande* può fornire una rappresentazione imprecisa (grossolana) del comportamento del sistema. Se gli eventi si verificano su una scala temporale molto diversa dalla suddivisione temporale in intervalli di uguale ampiezza effettuata con un simulatore sincrono, si potrebbe verificare una *perdita di precisione* poiché si rischia di trascurare comportamenti particolari del sistema. Inoltre in un simulatore sincrono si potrebbe avere una *perdita di efficienza* in presenza di dinamiche temporali poco uniformi, come ad esempio quando si hanno periodi

in cui si verificano un gran numero di variazioni dello stato del sistema alternati a periodi in cui si verificano poche variazioni. La perdita di precisione e di efficienza di un simulatore sincrono comportano che le misure prestazionali del sistema potrebbero risultare imprecise. In Figura 6.1 e in Figura 6.2 è rappresentata una realizzazione descrivente l'evoluzione temporale del numero di utenti $N(t)$ in un sistema di servizio con singolo servitore e sono visualizzate due differenti scelte del passo di discretizzazione. Nella Figura 6.1 il passo di discretizzazione è grande e, osservando il sistema solo all'inizio o alla fine di ciascuno intervallo della suddivisione, si nota che alcuni cambiamenti di stato del processo $N(t)$ non possono essere osservati. Invece nella Figura 6.2 il passo di discretizzazione è piccolo, la simulazione è più precisa ma inefficace, poiché si osserva lo stato del sistema anche in istanti di tempo in cui non si sono verificati cambiamenti di stato.

→ Nei *simulatori asincroni* il tempo di simulazione viene aggiornato in modo irregolare in base ai cambiamenti di stato del sistema. Nella Figura 6.3 questi cambiamenti sono indicati con le frecce riportate sull'asse dei tempi. I simulatori asincroni si basano sul principio che lo stato del sistema rimane invariato tra gli istanti successivi di cambiamento di stato del processo e quindi non occorre osservare il sistema in questi periodi di tempo.

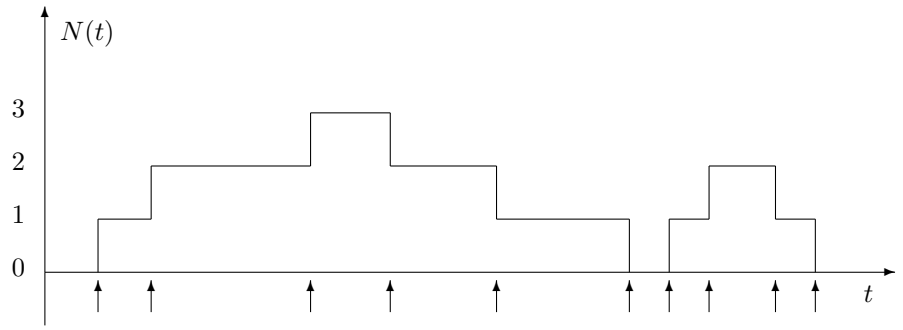

Figura 6.3: Effetto della discretizzazione in un simulatore asincrono.

• Una *quarta classificazione* distingue i simulatori in *orientati agli eventi* (*ad eventi discreti*) e *orientati ai processi* in base alle modalità di funzionamento.

→ Nei *simulatori ad eventi discreti* (*orientati agli eventi*) lo scorrere del tempo è strettamente legato al verificarsi di cambiamenti di stato (detti *eventi*) del sistema. Si definisce *evento* ogni possibile situazione che porta ad un cambiamento del valore delle variabili di stato che descrivono il comportamento del sistema. Gli eventi non avvengono in modo continuo ma soltanto negli istanti temporali in cui si verificano cambiamento di stato del sistema. Il simulatore ad eventi discreti quindi salta i periodi in cui non si verificano cambiamenti di stato poiché tali periodi non sono significativi per la descrizione dell'evoluzione del sistema. Nei simulatori ad eventi discreti i tempi in cui si verificano gli

eventi debbono essere mantenuti in ordine crescente e gli eventi debbono essere realizzati in sequenza.
→ Nei *simulatori orientati ai processi* il sistema è descritto in termini di processi (piuttosto che di eventi) che sono eseguiti in parallelo e che interagiscono tra loro scambiandosi informazioni.

In un simulatore ad eventi discreti occorre:

(i) definire i tipi di eventi che si possono verificare;

(ii) definire per ogni evento le modifiche da apportare allo stato che descrive il comportamento del sistema;

(iii) definire una struttura dati (*calendario degli eventi*) che permetta di ordinare gli eventi sulla base del loro istante di occorrenza e che raccolga le informazioni relative a tali eventi;

(iv) definire la fase di inizializzazione delle variabili;

(v) scorrere il calendario e ogni volta che si incontra un evento eseguire le modifiche delle variabili di stato corrispondenti a quell'evento;

(vi) valutare i parametri prestazionali del sistema.

Si nota che per *simulare un sistema di servizio* occorre scegliere un *simulatore dinamico, asincrono e ad eventi discreti*. In esso si possono identificare due tipi di eventi che causano cambiamenti di stato, ossia gli arrivi e le partenze degli utenti dal sistema.

6.3 Metodo di Monte Carlo

Nei *simulatori statici* si ricorre spesso al *metodo di Monte Carlo*. Il metodo di Monte Carlo è *un procedimento statistico basato sull'utilizzazione di numeri casuali*; il nome si riferisce alla capitale del Principato di Monaco e, più precisamente, alle roulette presenti nei casinò considerate come semplici congegni per la generazione di numeri casuali.

Il nome "metodo di Monte Carlo" venne usato per la prima volta verso la metà degli anni 40 nell'ambito del *progetto Manhattan* (Los Alamos, New Mexico) in connessione al trasporto dei neutroni. Il progetto Manhattan era un programma di ricerca e sviluppo in ambito militare che condusse anche alla realizzazione delle prime bombe atomiche durante la seconda guerra mondiale. Contribuirono alle ricerche il matematico, fisico e informatico ungherese *John von Neumann* (1903–1957), il matematico polacco *Stanislaw Marcin Ulam* (1909-1984), il matematico statunitense *Nicholas Constantine Metropolis* (1915-1999) e il fisico italiano *Enrico Fermi* (1901-1954). Nel 1949 Metropolis e Ulam, con il supporto di Fermi e von Neumann, pubblicarono il primo articolo ufficiale sul metodo Monte Carlo[1].

[1] N. Metropolis and S. Ulam: The Monte Carlo Method, Journal of the American Statistical Association 44, Num. 247, 335–341 (1949)

In generale, con il termine *metodo di Monte Carlo* si intende *rappresentare la soluzione di un problema come un parametro non noto (di una ipotetica popolazione) e si cerca di stimare tale parametro utilizzando un campione (estratto dalla popolazione) ottenuto mediante sequenze di numeri casuali.*

Spesso, invece di servirsi di un campione di numeri effettivamente estratti a caso, si ricorre a una sequenza di numeri ottenuti con un processo iterativo deterministico; tali numeri vengono detti *pseudo-casuali* poiché hanno proprietà statistiche analoghe a quelle dei veri numeri casuali.

A partire dalla fine degli anni 1950, sono apparsi in letteratura numerosi articoli scientifici che descrivono il metodo e le sue applicazioni in diversi campi scientifici: ingegneria, informatica, economia, finanza, fisica, biologia e matematica applicata. Il metodo di Monte Carlo si rivela spesso utile per risolvere vari tipi di problemi matematici: valutazione di integrali unidimensionali e multidimensionali, sistemi di equazioni lineari, inversioni di matrici. Inoltre, il metodo di Monte Carlo gioca un ruolo fondamentale in molti metodi di simulazione, che mirano a stimare parametri non noti di fenomeni complessi di carattere aleatorio.

Allo sviluppo dei metodi di Monte Carlo hanno contribuito i notevoli progressi in ambito informatico connessi sviluppo tecnologico e all'enorme crescita delle potenza di calcolo degli elaboratori.

6.3.1 Calcolo dell'area sottesa ad una curva

Applichiamo ora il metodo di Monte Carlo per *calcolare l'area sottesa da una curva*, che rappresenta il parametro non noto da stimare di una ipotetica popolazione. Sia $f(x)$ una funzione positiva definita nell'intervallo (a, b) e sia

$$J = \int_a^b f(x)\, dx, \tag{6.1}$$

ossia J è l'area sottesa dalla curva $f(x)$ nell'intervallo (a, b).

Considereremo prima una procedura numerica per calcolare l'integrale definito unidimensionale (6.1) e successivamente utilizzeremo due differenti metodi di Monte Carlo per calcolare lo stesso integrale.

► **Metodo numerico**

Suddividiamo l'intervallo (a, b) in N sottointervalli di ampiezza Δ. Quindi, $b - a = N\,\Delta$. Poniamo inoltre:

$$x_0 = a, \qquad x_i = a + i\,\Delta = a + i\,\frac{b-a}{N} \qquad (i = 1, 2, \ldots, N) \tag{6.2}$$

Si nota che $f(x_i)\,\Delta$ rappresenta l'area del rettangolo di altezza $f(x_i)$ e di base Δ. Un'approssimazione numerica per l'integrale definito J è:

$$\widehat{J} = \sum_{i=1}^{N} f(x_i)\,\Delta = \sum_{i=1}^{N} f(x_i)\,\frac{b-a}{N} = (b-a)\,\Big[\frac{1}{N}\,\sum_{i=1}^{N} f(x_i)\Big]. \tag{6.3}$$

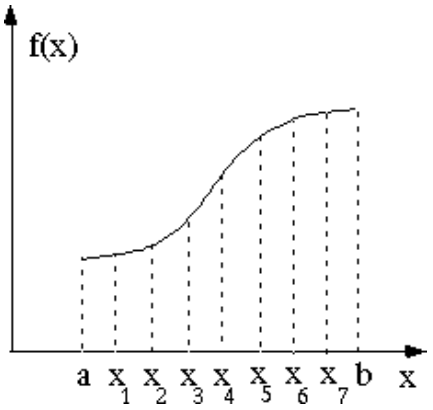

Figura 6.4: Metodo numerico per il calcolo di un integrale in (a, b).

Tale approssimazione numerica corrisponde al calcolo della media aritmetica di $f(x_1), f(x_2), \ldots, f(x_N)$ moltiplicata per l'ampiezza dell'intervallo di integrazione $b - a$.

► **Primo metodo di Monte Carlo**

Tale metodo consiste nel generare N variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$ e nel trasformarle in variabili aleatorie ancora indipendenti e uniformemente distribuite nell'intervallo di integrazione (a, b) mediante la trasformazione

$$X_i = a + (b - a)\, U_i \qquad (i = 1, 2, \ldots, N). \tag{6.4}$$

A differenza del metodo numerico ora gli N punti non vengono più scelti deterministicamente e equidistanziati di Δ nell'intervallo (a, b), ma *scelti in maniera probabilistica*, ossia si assume che siano *indipendenti e uniformemente distribuiti nell'intervallo* (a, b). Uno stimatore che utilizza il metodo di Monte Carlo per calcolare J è

$$\widehat{J} = (b - a)\left[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right]. \tag{6.5}$$

Vogliamo ora dimostrare che tale stimatore gode di importanti proprietà statistiche, ossia:

$$E(\widehat{J}) = J, \qquad \lim_{N\to+\infty} \mathrm{Var}(\widehat{J}) = 0. \tag{6.6}$$

Infatti, essendo $X_1, X_2, \ldots, X_N$ variabili aleatorie iuniformemente distribuite in (a, b) con densità $1/(b - a)$, risulta:

$$E[f(X_i)] = \int_a^b f(x)\,\frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b f(x)\,dx = \frac{J}{b-a}\,,$$
$$E[f^2(X_i)] = \int_a^b f^2(x)\,\frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b f^2(x)\,dx,$$

e per la linearità del valore medio si ha:

$$E(\widehat{J}) = E\Big\{(b-a)\Big[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\Big]\Big\} = \frac{b-a}{N}\sum_{i=1}^{N} E[f(X_i)] = J,$$

e per l'indipendenza nella scelta degli N punti, ossia delle variabili aleatorie $X_1, X_2, \ldots, X_N$, si ottiene:

$$\mathrm{Var}(\widehat{J}) = \mathrm{Var}\Big\{(b-a)\Big[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\Big]\Big\} = \frac{(b-a)^2}{N^2}\sum_{i=1}^{N} \mathrm{Var}[f(X_i)],$$

dimostrando così le (6.6).

I passi dell'algoritmo del primo metodo di Monte Carlo per valutare J sono quindi i seguenti:

Algoritmo

STEP 1: Generare una sequenza di N variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;

STEP 2: Generare una sequenza di N variabili aleatorie $X_1, X_2, \ldots, X_N$ indipendenti e uniformemente distribuite nell'intervallo (a, b) tramite la trasformazione

$$X_i = a + (b-a)\,U_i \qquad (i = 1, 2, \ldots, N)$$

STEP 3: Stimare J utilizzando lo stimatore:

$$\widehat{J} = (b-a)\Big[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\Big].$$

▶ Secondo metodo di Monte Carlo: successo e insuccesso

Sia $f(x)$ una funzione definita nell'intervallo (a, b) tale che $0 \le f(x) \le c$. Il metodo consiste nel generare indipendentemente N coppie di variabili aleatorie (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N) (rappresentanti N punti) uniformemente distribuiti nel rettangolo $R = \{(x, y) : \ a < x < b,\ 0 < y < c\}$. La funzione densità di probabilità di ogni singola coppia (X, Y) è uniforme nel rettangolo, ossia

$$g_{XY}(x, y) = \begin{cases} \dfrac{1}{c\,(b-a)}, & (x, y) \in R \\ 0, & \text{altrimenti.} \end{cases} \tag{6.7}$$

Le variabili aleatorie X e Y sono tra loro indipendenti e caratterizzate rispettivamente da densità marginali

$$g_X(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{altrimenti,} \end{cases} \qquad g_Y(y) = \begin{cases} \dfrac{1}{c}, & 0 < y < c \\ 0, & \text{altrimenti.} \end{cases}$$

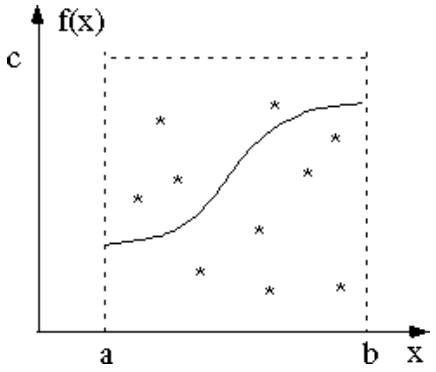

Figura 6.5: Secondo metodo di Monte Carlo per il calcolo di un integrale in (a, b).

Geometricamente è ovvio che se N_S di questi punti sono sotto la curva $y = f(x)$, sussiste l'approssimazione:

$$\frac{N_S}{N} \simeq \frac{\int_a^b f(x)\, dx}{c\,(b-a)},$$

che ci conduce a considerare come stimatore per l'integrale (6.1):

$$\widehat{J} = c\,(b-a)\,\frac{N_S}{N}\,. \tag{6.8}$$

Vogliamo ora dimostrare che tale stimatore gode delle proprietà (6.6). Osserviamo in primo luogo che ciascuna delle N prove costituisce una prova di Bernoulli in cui le probabilità di successo e insuccesso sono rispettivamente

$$p = \frac{\int_a^b f(x)\, dx}{c\,(b-a)}, \qquad\qquad q = 1 - p. \tag{6.9}$$

Pertanto, N_s rappresenta il numero di successi in N prove indipendenti di Bernoulli ed è quindi distribuita in modo *binomiale con probabilità di successo p in ogni singola prova*. Pertanto,

$$E(N_S) = N\,p, \qquad \mathrm{Var}(N_S) = N\,p\,q.$$

Dalla (6.8) si ha:

$$E(\widehat{J}) = E\Big[c\,(b-a)\,\frac{N_S}{N}\Big] = \frac{c\,(b-a)}{N}\,E(N_S) = \int_a^b f(x)\,dx = J$$
$$\mathrm{Var}(\widehat{J}) = \mathrm{Var}\Big[c(b-a)\,\frac{N_S}{N}\Big] = \frac{c^2(b-a)^2}{N^2}\,\mathrm{Var}(N_S) = \frac{c^2(b-a)^2}{N}\,p\,q.$$

Si nota nuovamente che sussistono le (6.6), ossia $E(\widehat{J}) = J$ e $\lim_{N\to+\infty} \mathrm{Var}(\widehat{J}) = 0$.

I passi dell'algoritmo del secondo metodo di Monte Carlo per valutare J sono quindi i seguenti:

Algoritmo
STEP 1: Generare indipendentemente due sequenze $U_1, U_2, \ldots, U_N$ e $V_1, V_2, \ldots, V_N$, ognuna costituita da N variabili aleatorie indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;
STEP 2: Generare una sequenza di N variabili aleatorie $X_1, X_2, \ldots, X_N$ uniformemente distribuite nell'intervallo (a, b) e inoltre generare una sequenza di N variabili aleatorie $Y_1, Y_2, \ldots, Y_N$ uniformemente distribuite nell'intervallo $(0, c)$ mediante le trasformazioni

$$X_i = a + (b - a)\, U_i, \qquad Y_i = c\, V_i \qquad (i = 1, 2, \ldots, N) \tag{6.10}$$

STEP 3: Porre inizialmente la variabile aleatoria contatore $N_S = 0$. Per ognuna delle coppie (X_i, Y_i) $(i = 1, 2, \ldots, N)$ determinate con il generatore uniforme controllare se $Y_i < f(X_i)$; se tale condizione è soddisfatta allora la variabile aleatoria contatore N_S è incrementata di un'unità, mentre se tale condizione non è soddisfatta N_S non viene incrementata;
STEP 4: Stimare J utilizzando lo stimatore:

$$\widehat{J} = c\,(b - a)\,\frac{N_S}{N}\,.$$

▶ Calcolo dell'area sottesa ad una curva su domini infiniti con il metodo di Monte Carlo
Finora ci siamo occupati della valutazione di integrali definiti su un dominio finito. Desideriamo ora calcolare con il metodo di Monte Carlo il seguente integrale:

$$J = \int_0^{+\infty} f(x)\, dx, \tag{6.11}$$

Effettuiamo in (6.11) il cambiamento di variabile $y = (x + 1)^{-1}$, ossia $x = (1 - y)/y$. Allora si ha:

$$J = \int_0^{+\infty} f(x)\, dx = \int_0^1 f\Big(\frac{1-y}{y}\Big)\,\frac{1}{y^2}\, dy = \int_0^1 h(y)\, dy, \tag{6.12}$$

dove si è posto

$$h(y) = f\Big(\frac{1-y}{y}\Big)\,\frac{1}{y^2} \tag{6.13}$$

Abbiamo ricondotto la valutazione di J alla valutazione di un integrale nell'intervallo $(0, 1)$ del tipo (6.1). Applicando quindi il primo metodo di Monte Carlo si ha:

$$\widehat{J} = \frac{1}{N}\sum_{i=1}^{N} h(U_i) = \frac{1}{N}\sum_{i=1}^{N} f\Big(\frac{1-U_i}{U_i}\Big)\,\frac{1}{U_i^2}. \tag{6.14}$$

Vogliamo dimostrare che sussistono ancora le (6.6). Infatti, ricordando la (6.13) si ha

$$E[h(U_i)] = \int_0^1 h(y)\, dy = \int_0^{+\infty} f(x)\, dx = J\,,$$
$$E[h^2(U_i)] = \int_0^1 h^2(y)\, dy = \int_0^{+\infty} f^2(x)\, dx,$$

da cui si ottiene:

$$E(\widehat{J}) = E\Big[\frac{1}{N}\sum_{i=1}^{N} h(U_i)\Big] = \frac{1}{N}\sum_{i=1}^{N} E[h(U_i)] = J,$$
$$\mathrm{Var}(\widehat{J}) = \mathrm{Var}\Big[\frac{1}{N}\sum_{i=1}^{N} h(U_i)\Big] = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}[h(U_i)],$$

da cui si ottengono le (6.6). I passi dell'algoritmo del primo metodo di Monte Carlo per calcolare J in (6.11) sono quindi i seguenti:

Algoritmo
STEP 1: Generare una sequenza di N variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;
STEP 2: Stimare J utilizzando lo stimatore:

$$\widehat{J} = \frac{1}{N}\sum_{i=1}^{N} f\Big(\frac{1-U_i}{U_i}\Big)\,\frac{1}{U_i^2}.$$

Nel caso di integrali multidimensionali il metodo di Monte Carlo si rivela spesso competitivo e più efficace rispetto ad altri metodi puramente numerici.

6.3.2 Calcolo di π con il metodo di Monte Carlo

Ci proponiamo ora di stimare π utilizzando il metodo di Monte Carlo. Osserviamo che l'area di un cerchio di raggio unitario è $A = \pi\, r^2 = \pi$ e quindi il settore circolare disegnato in Figura 6.6 ha area $A = \pi/4$. Un punto di coordinate (x, y) è interno al settore circolare se risulta $x^2 + y^2 < 1$, con $0 < x < 1$ e $0 < y < 1$.

Se si scelgono a caso N punti nel quadrato di lato unitario e se si denota con N_s il numero di punti che cadono nel settore circolare, si deve avere

$$\frac{N_s}{N} \simeq \frac{\pi/4}{1},$$

ossia N_s/N è approssimativamente uguale al rapporto tra l'area del settore circolare e l'area del quadrato circoscritto al settore. Pertanto, una stima di π è

$$\widehat{\Pi} = 4N_s/N.$$

L'algoritmo per il calcolo di π può essere così formalizzato:

Algoritmo

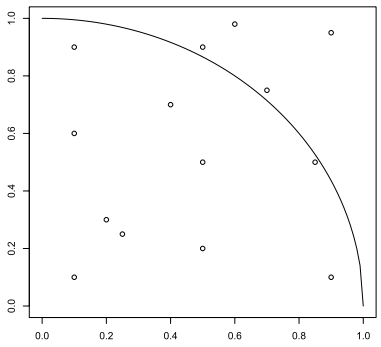

Figura 6.6: Calcolo di π con il metodo di Monte Carlo.

STEP 1: Generare indipendentemente due sequenze $U_1, U_2, \ldots, U_N$ e $V_1, V_2, \ldots, V_N$, ognuna costituita da N variabili aleatorie indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;
STEP 2: Porre inizialmente la variabile aleatoria contatore $N_S = 0$. Per ognuna delle coppie $(U_i, V_i) \quad (i = 1, 2, \ldots, N)$ determinate con il generatore uniforme controllare se $U_i^2 + V_i^2 < 1$; se tale condizione è soddisfatta allora la variabile aleatoria contatore N_S è incrementata di un'unità, mentre se tale condizione non è soddisfatta N_S non viene incrementata;
STEP 3: Stimare π utilizzando lo stimatore $\widehat{\Pi} = 4N_s/N$.

Tale stimatore gode di importanti proprietà statistiche, ossia:

$$E(\widehat{\Pi}) = \pi, \qquad \lim_{N \to +\infty} \mathrm{Var}(\widehat{\Pi}) = 0.$$

6.3.3 Somma nel lancio di dadi con il metodo di Monte Carlo

Consideriamo un esperimento consistente nel lanciare k dadi non truccati e nel registrare la somma dei risultati ottenuti. Denotiamo con X_i la variabile aleatoria che descrive il risultato ottenuto nel lanciare il dado i-esimo ($i = 1, 2, \ldots, k$) che assume i valori $1, 2, 3, 4, 5, 6$ ognuno con probabilità pari a $1/6$. Denotiamo inoltre con $S = X_1 + X_2 + \ldots + X_k$ la variabile aleatoria che descrive la somma dei risultati ottenuti nel lancio dei k dadi. Ovviamente S assume valori da k (tutti i k dadi hanno fornito come risultato 1) fino a $6\,k$ (tutti i k dadi hanno fornito come risultato 6). Utilizzando il metodo di Monte Carlo, si desidera stimare la probabilità $p = P(S = s)$ che la somma dei risultati ottenuti sia s, con $s \in [k, 6k]$.

L'algoritmo per stimare la probabilità p può essere così formalizzato:

Algoritmo

STEP 1: Simulare il lancio di ognuno dei k dadi per N volte ottenendo i vettori:

$$\begin{array}{ll} (x_{1,1}, \ldots x_{1,i}, \ldots, x_{1,N}) & (\textit{primo dado}) \\ (x_{2,1}, \ldots x_{2,i}, \ldots, x_{2,N}) & (\textit{secondo dado}) \\ \ldots\ldots\ldots\ldots & \\ (x_{k,1}, \ldots x_{k,i}, \ldots, x_{k,N}) & (k\ \textit{dado}) \end{array} \tag{6.15}$$

STEP 2: Creare il vettore $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_N)$ ottenuto sommando i risultati dei lanci dei k dadi, dove $y_i = x_{1,i} + x_{2,i} + \ldots + x_{k,i}$ $(k = 1, 2, \ldots, N)$.
STEP 3: Contare il numero N_s di volte in cui nel vettore $\mathbf{y}$ compare come somma s e stimare la probabilità p richiesta con lo stimatore

$$\widehat{P} = \frac{N_s}{N}.$$

Osserviamo che N_s rappresenta il numero di successi in N prove indipendenti di Bernoulli ed è quindi distribuita in modo binomiale con probabilità di successo p in ogni singola prova. Pertanto

$$E(\widehat{P}) = \frac{1}{N} E(N_s) = \frac{1}{N} N\, p = p,$$
$$\lim_{N \to +\infty} \mathrm{Var}(\widehat{P}) = \lim_{N \to +\infty} \Big[\frac{1}{N^2} N\, p\, (1-p) \Big] = 0,$$

che mostrano che lo stimatore $\widehat{P}$ della probabilità $p = P(S = s)$ gode di importanti proprietà statistiche.

Per realizzare al computer tale esperimento occorre simulare una variabile aleatoria discreta X descrivente il risultato del lancio di un dado. Come vedremo nel seguito ciò può essere realizzato simulando una variabile aleatoria uniforme U nell'intervallo $(0, 1)$, suddividendo l'intervallo in sei sottointervalli di uguale ampiezza $1/6$ e ponendo

$$X = \begin{cases} 1, & 0 \le U < 1/6 \\ 2, & 1/6 \le U < 2/6 \\ 3, & 2/6 \le U < 3/6 \\ 4, & 3/6 \le U < 4/6 \\ 5, & 4/6 \le U < 5/6 \\ 6, & 5/6 \le U < 1 \end{cases}$$

A partire dalla generazione di una singola variabile discreta X descrivente il risultato del lancio di un dado, occorre simulare il lancio di ognuno dei k dadi per N volte come descritto nello Step 1.

Nelle procedure del metodo di Monte Carlo precedentemente utilizzate intervengono due importanti punti:

(i) generazione di variabili aleatorie uniformemente distribuite nell'intervallo $(0, 1)$;

(ii) generazione, a partire dal punto *(i)*, di altre variabili aleatorie con opportune distribuzioni di probabilità.

In entrambi i punti l'affidabilità dei risultati che si ottengono si basa principalmente sulla qualità della sorgente dei numeri pseudocasuali e sulla scelta di un algoritmo computazionale efficiente.

Le problematiche indicate nei punti *(i)* e *(ii)* intervengono anche nella *simulazione dei sistemi di servizio*, come vedremo nei prossimi capitoli.

Capitolo 7

Generatori uniformi

7.1 Linguaggio R

In questo capitolo e nei successivi utilizzeremo il linguaggio R per creare sequenze di numeri pseudocasuali, per generare variabili aleatorie discrete e continue e per simulare alcuni sistemi di servizio.

R è contemporaneamente un linguaggio di programmazione ed un software, scaricabile gratuitamente da Internet sotto la licenza GPL (General Public License). Sul sito del progetto "The R Project for Statistical Computing", la cui home page è

http://www.r-project.org/

è possibile trovare ogni tipo di supporto per l'utilizzo di R. Inoltre dal sito di "The Comprehensive R Archive Network (CRAN)", il cui indirizzo Internet è

http://cran.r-project.org/

è possibile accedere ai numerosi mirror del CRAN per effettuare il download del software e di tutta la documentazione per diversi sistemi operativi: Linux, Windows, MacOS. Da tale sito è possibile scaricare, oltre al programma base, anche una vasta gamma di package aggiuntivi utilizzabili per la risoluzione di specifici problemi o per particolari analisi statistiche.

La versione iniziale di R è stata realizzata nel 1996 da Ross Ihaka e Robert Gentleman[1] del Dipartimento di Statistica dell'Università di Auckland in Nuova Zelanda. Successivamente, un nutrito gruppo di ricercatori operanti in ambito informatico e statistico hanno iniziato a fornire il loro contributo, dando così vita al "R Development Core Team", che dal 1997 si occupa dei codici sorgenti di R. Nel 2003 è stata costituita l'organizzazione non profit "R Foundation for Statistical Computing" avente come obiettivo quello di promuovere lo sviluppo

[1] Ihaka R., Gentleman R. *R: A Language for Data Analysis and Graphics*, Journal of Computational and Graphical Statistics, **5(3)**, 299-314, 1996.

e la diffusione del software, di gestire e tutelare il copyright di R e della relativa documentazione.

Le caratteristiche principali di R possono essere così sintetizzate:

1. è un software *open source*;
2. è un linguaggio di programmazione *object-oriented* e utilizza un *interprete*;
3. è un software *multipiattaforma* e può essere installato su Linux, Windows e MacOS;
4. dispone di un insieme di strumenti per il calcolo su vettori, matrici, data frame e per altre operazioni anche complesse;
5. è semplice da utilizzare nella gestione e nella manipolazione di dati;
6. è dotato di notevoli e flessibili potenzialità grafiche 2D e 3D consentendo la rappresentazione grafica di dati statistici;
7. fornisce la possibilità di programmare, creando funzioni e programmi ad hoc definiti dall'utente.

Nel linguaggio R tutto viene rappresentato mediante oggetti. Ogni oggetto (vettore, dataset, tabella, grafico, ...) è trattato dalle funzioni di R con uno specifico metodo ed è possibile implementare nuovi metodi per ampliare le possibilità delle stesse funzioni.

Il cuore di R è rappresentato dal modulo base che offre gli strumenti fondamentali per effettuare le usuali operazioni di lettura e scrittura dei dati da e su file, le operazioni su matrici e vettori e le elaborazioni statistiche connesse all'analisi esplorativa dei dati, e alla produzione di grafici. Alcune librerie sono già comprese nel modulo base, mentre altre librerie possono essere aggiunte e installate successivamente in base alle necessità. Tali package sono disponibili sul sito del CRAN (CRAN Task Views).

L'attivazione di una sessione di lavoro viene effettuata agendo con un doppio click del mouse sull'icona che identifica l'applicazione R presente sul desktop. Dopo qualche istante sullo schermo compare la tipica console di R che presenta un'immagine simile a quella illustrata qui di seguito:

```
R version 4.3.3 (2024-02-29)
Copyright (C) 2024 The R Foundation for Statistical Computing
R è un software libero ed è rilasciato SENZA ALCUNA GARANZIA.
Siamo ben lieti se potrai redistribuirlo, ma sotto certe condizioni.
Scrivi 'license()' o 'licence()' per dettagli su come distribuirlo.
R è un progetto di collaborazione con molti contributi esterni.
Scrivi 'contributors()' per maggiori informazioni e 'citation()'
per sapere come citare R o i pacchetti di R nelle pubblicazioni.

Scrivi 'demo()' per una dimostrazione, 'help()' per la guida in linea, o
'help.start()' per l'help navigabile con browser HTML.
Scrivi 'q()' per uscire da R.
>
```

Tutti i comandi debbono essere inseriti dopo il segnale di prompt ">", che indica la disponibilità del sistema ad accettare nuovi comandi. Scopo della console è quello di fornire all'utente un ambiente dove scrivere i comandi che desidera eseguire. Dopo aver premuto il tasto Invio, il corrispondente comando scritto nella riga, viene immediatamente verificato per controllarne la correttezza ed eseguito se l'esito della verifica è positivo.

Anche se il linguaggio è fornito con un'interfaccia a linea di comando, sono disponibili diverse interfacce grafiche che consentono di integrare R, tra cui *RStudio* il cui download può essere effettuato da *https://www.rstudio.com/*.

ll sistema R consente la progettazione e la realizzazione di nuove funzioni definite dall'utente. La definizione di una nuova funzione in R deve rispettare la seguente sintassi:

$> \mathsf{nomeFunzione} <- \mathsf{function}(\mathsf{arg1}, \mathsf{arg2}, \ldots, \mathsf{argN})\ \mathsf{corpo\ della\ funzione}$

Il nome della funzione è indicato alla sinistra dell'operatore di assegnazione $<-$. Gli argomenti, se presenti, sono inseriti separati da virgole nella coppia di parentesi tonde e sono usati per passare i valori iniziali alla funzione. Alla fine si definisce il corpo della funzione, ossia le istruzioni necessarie per ottenere il risultato atteso. Se il corpo è formato da più di un'istruzione occorre racchiuderlo tra una coppia di parentesi graffe. Una funzione ritorna il valore calcolato nell'ultima espressione oppure ritorna i valori $\mathsf{r1}, \mathsf{r2}, \ldots, \mathsf{rk}$ utilizzando

$$> \mathsf{return}(\mathsf{c}(\mathsf{r1}, \mathsf{r2}, \ldots, \mathsf{rk}))$$

dove $c()$ denota l'operatore di concatenazione, utilizzato per definire il vettore $(\mathsf{r1}, \mathsf{r2}, \ldots, \mathsf{rk})$. Una chiamata alla funzione si ottiene con l'istruzione

$> \mathsf{nomeFunzione}(\mathsf{arg1}, \mathsf{arg2}, \ldots, \mathsf{argN}).$

In questo capitolo utilizzeremo il linguaggio R per definire alcune funzioni utili a generare sequenze di numeri pseudocasuali utilizzando diversi tipi di algoritmi.

7.2 Numeri casuali e pseudocasuali

Nella simulazione giocano un ruolo fondamentale i *numeri casuali* uniformemente distribuiti in un fissato intervallo. In passato sono state proposte varie tecniche per generare numeri casuali. Una consisteva nel dotare l'elaboratore di una speciale apparecchiatura capace di generare numeri casuali sfruttando qualche particolare fenomeno fisico allo scopo di *produrre tabelle di numeri casuali* che gli studiosi potessero poi utilizzare nelle loro simulazioni. Furono così costruite delle *macchine* sia di *tipo meccanico* sia di *tipo elettronico* che presentavano però l'inconveniente di imporre una *laboriosa manutenzione* per garantire l'efficienza di apparecchiature, spesso delicate, atte a generare sequenze casuali, che risultavano talvolta poco adatte alle applicazioni desiderate.

Il metodo più classico per ottenere sequenze casuali uniformemente distribuite è il *metodo dell'urna* in cui dischetti numerati sono messi in un'urna e mescolati prima di ogni estrazione; ogni dischetto estratto, dopo essere stato letto, è rimesso nell'urna. Questo metodo ha discrete caratteristiche random

ma le sequenze ottenute, come tutte quelle effettivamente basate su un processo di successive estrazioni con rimpiazzamento, presentano l'inconveniente di essere *non ripetibili*, a meno che non si proceda ad una laboriosa ed ingombrante *registrazione di tutti i valori estratti in una tabella*, che può essere inserita in un elaboratore e successivamente utilizzata per ottenere sequenze casuali. Ogni metodo che porta alla costruzione di tabelle di numeri casuali, anche se permette di riottenere gli stessi risultati utilizzando gli stessi numeri, è *costoso* sia per lo *spazio di memoria occupato dalla tabella*, sia per il *tempo richiesto per accedervi* e sia poiché per analizzare determinati problemi potrebbero essere necessari molti più numeri di quelli presenti nelle tabelle.

Per superare gli inconvenienti finora discussi sono stati sviluppati metodi che permettono di ottenere dagli elaboratori sequenze di numeri casuali attraverso il ripetuto uso di un *meccanismo algebrico deterministico* (ossia mediante opportune formule di ricorrenza) che presentano il vantaggio, rispetto alle altre tecniche già citate, di essere facilmente implementabili con algoritmi estremamente veloci computazionalmente, di richiedere poco spazio di memoria, di permettere di riprodurre sequenze identiche a quelle già utilizzate in modo da riottenere gli stessi risultati.

Storicamente i primi a proporre una tecnica di questo tipo furono von Neumann e Metropolis nel 1946 con il *metodo del centro del quadrato* (*midsquare method*). La procedura utilizzata per generare una sequenza $\alpha_0, \alpha_1, \ldots$ è la seguente.

Un arbitrario numero positivo α_0, detto *seme* (o *valore iniziale*), è scelto come input per generare il processo ricorrente. Tale numero α_0, rappresentato con $2k$ digits, è elevato al quadrato. Si produce così un numero α_0^2 di $4k$ digits (inserendo, se necessario, degli 0 alla sinistra del numero per formare esattamente $4k$ digits) dal quale viene estratto un numero α_1 costituito dai $2k$ digits centrali di α_0^2 (includendo quindi i bits da $k+1$ a $3k$), che a sua volta viene elevato al quadrato per generare α_2. Si prosegue poi in questo modo per generare gli altri numeri della sequenza.

Esempio 7.1 Sia $\alpha_0 = 8234$. Con il metodo del centro del quadrato si ottiene la sequenza

$$\begin{aligned}
&\alpha_0 = 8234 \\
&\alpha_0^2 = 67798756 \Longrightarrow \alpha_1 = 7987 \\
&\alpha_1^2 = 63792169 \Longrightarrow \alpha_2 = 7921 \\
&\alpha_2^2 = 62742241 \Longrightarrow \alpha_3 = 7422 \\
&\alpha_3^2 = 55086084 \Longrightarrow \alpha_4 = 0860 \\
&\alpha_4^2 = 00739600 \Longrightarrow \alpha_5 = 7396 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

♢

Il metodo del centro del quadrato ha scarse qualità statistiche; inoltre, il seme (valore iniziale) deve essere scelto accuratamente, come mostrato nei due esempi seguenti.

Esempio 7.2 Sia $\alpha_0 = 7182$. Con il metodo del centro del quadrato si ottiene la sequenza

$$\begin{aligned}
&\alpha_0 = 7182\\
&\alpha_0^2 = 51581124 \Longrightarrow \alpha_1 = 5811\\
&\alpha_1^2 = 33767721 \Longrightarrow \alpha_2 = 7677\\
&\alpha_2^2 = 58936329 \Longrightarrow \alpha_3 = 9363\\
&\alpha_3^2 = 87665769 \Longrightarrow \alpha_4 = 6657\\
&\alpha_4^2 = 44315649 \Longrightarrow \alpha_5 = 3156\\
&\alpha_5^2 = 09960336 \Longrightarrow \alpha_6 = 9603\\
&\alpha_6^2 = 92217609 \Longrightarrow \alpha_7 = 2176\\
&\alpha_7^2 = 04734976 \Longrightarrow \alpha_8 = 7349\\
&\alpha_8^2 = 54007801 \Longrightarrow \alpha_9 = 0078\\
&\alpha_9^2 = 00006084 \Longrightarrow \alpha_{10} = 0060\\
&\alpha_{10}^2 = 00003600 \Longrightarrow \alpha_{11} = 0036\\
&\alpha_{11}^2 = 00001296 \Longrightarrow \alpha_{12} = 0012\\
&\alpha_{12}^2 = 00000144 \Longrightarrow \alpha_{13} = 0001\\
&\alpha_{13}^2 = 00000001 \Longrightarrow \alpha_{14} = 0000\\
&\alpha_{14}^2 = 00000000 \Longrightarrow \alpha_{15} = 0000
\end{aligned}$$

Si nota che $\alpha_{14} = \alpha_{15} = \ldots = 0$. ♢

Esempio 7.3 Sia $\alpha_0 = 99$. Con il metodo del centro del quadrato si ottiene la sequenza

$$\begin{aligned}
&\alpha_0 = 99\\
&\alpha_0^2 = 9801 \Longrightarrow \alpha_1 = 80\\
&\alpha_1^2 = 6400 \Longrightarrow \alpha_2 = 40\\
&\alpha_2^2 = 1600 \Longrightarrow \alpha_3 = 60\\
&\alpha_3^2 = 3600 \Longrightarrow \alpha_4 = 60
\end{aligned}$$

Si nota che $\alpha_3 = \alpha_4 = \ldots = 60$. ♢

Il metodo del centro del quadrato agli inizi conseguì un discreto successo ma ben presto non venne più utilizzato principalmente per tre motivi:

(a) Relativa lentezza

Il metodo si rivelava lento computazionalmente nella generazione di sequenze di numeri sia a causa dell'operazione di esponenziazione che per la successiva operazione di selezione delle cifre centrali del numero ottenuto al passo precedente.

(b) Difficoltà analitiche

Queste consistono soprattutto nel determinare la lunghezza del ciclo della sequenza casuale ottenuta a partire da un generico valore iniziale fino al valore per il quale la sequenza inizia a ripetersi.

(c) Insoddisfacente comportamento statistico delle sequenze ottenute.

I numeri generati con questo metodo non sono uniformemente distribuiti ed inoltre viene meno l'indipendenza statistica tra gli elementi della sequenza.

Il metodo del centro del quadrato fu, quindi, abbandonato e ad esso seguirono i *metodi congruenti*. Attualmente proprio questi metodi e loro varianti sono spesso utilizzati nelle applicazioni.

In generale, un processo è veramente casuale se le predizioni sul suo comportamento futuro non possono essere migliorate dalla conoscenza del comportamento passato. Nell'adottare meccanismi algebrici deterministici per la generazione di sequenze casuali, si può incorrere nell'inconveniente di venire meno all'indipendenza statistica tra gli elementi della sequenza stessa. In processi di simulazione il termine casuale è quindi generalmente sostituito con il termine *pseudocasuale*.

Un metodo per la generazione di sequenze di numeri pseudocasuali con distribuzione uniforme è accettabile se soddisfa ai seguenti requisiti:

(i) i numeri debbono essere statisticamente uniformemente distribuiti nella sequenza;

(ii) i numeri debbono essere statisticamente indipendenti nella sequenza;

(iii) la sequenza deve essere riproducibile;

(iv) la sequenza deve poter avere un ciclo di lunghezza abbastanza grande;

(v) il metodo deve poter essere eseguito dall'elaboratore con rapidità e deve occupare poco spazio di memoria.

7.3 Metodo congruenziale moltiplicativo

I generatori lineari congruenziali moltiplicativi (GLCM) producono sequenze $\{x_n,\ n = 0, 1, 2, \ldots\}$ come segue:

(i) fissare un intero positivo m detto *modulo* del generatore;

(ii) scegliere degli interi positivi a e x_0 minori del modulo m; x_0 ($x_0 \neq 0$) è detto *valore iniziale* o *seme* e la costante a ($a \neq 0$) è detta *costante moltiplicativa* oppure *moltiplicatore*;

(iii) generare x_n mediante la relazione di congruenza lineare

$$x_{n+1} \equiv a\, x_n \pmod{m} \tag{7.1}$$

che si legge x_{n+1} *è congruente ad* $a\, x_n$ *modulo* m.

La procedura inizia con un valore iniziale x_0 che deve essere diverso da zero. Per determinare gli elementi della sequenza $\{x_n,\ n = 1, 2, \ldots\}$ occorre assegnare a x_{n+1} il resto r (con $0 \leq r \leq m-1$) della divisione di $a\, x_n$ per il modulo m.

La relazione di ricorrenza (7.1) è analoga all'equazione alle differenze del primo ordine $x_{n+1} = a\, x_n$ che ammette come soluzione

$$x_n = x_0\, a^n.$$

La relazione di congruenza lineare (7.1) può quindi essere così riscritta

$$x_n \equiv x_0\, a^n \pmod{m}. \tag{7.2}$$

Esempio 7.4 Sia $m = 32 = 2^5$, $x_0 = 1$ e $a = 3$. La relazione (7.1) diventa

$$x_{n+1} \equiv 3\, x_n \pmod{2^5}$$

e conduce alla sequenza

$$\begin{aligned}
&x_0 = 1\\
&x_1 \equiv 3\, x_0 = 3 \pmod{32} \Longrightarrow x_1 = 3\\
&x_2 \equiv 3\, x_1 = 9 \pmod{32} \Longrightarrow x_2 = 9\\
&x_3 \equiv 3\, x_2 = 27 \pmod{32} \Longrightarrow x_3 = 27\\
&x_4 \equiv 3\, x_3 = 81 \pmod{32} \Longrightarrow x_4 = 17\\
&x_5 \equiv 3\, x_4 = 51 \pmod{32} \Longrightarrow x_5 = 19\\
&x_6 \equiv 3\, x_5 = 57 \pmod{32} \Longrightarrow x_6 = 25\\
&x_7 \equiv 3\, x_6 = 75 \pmod{32} \Longrightarrow x_7 = 11\\
&x_8 \equiv 3\, x_7 = 33 \pmod{32} \Longrightarrow x_8 = 1.
\end{aligned}$$

Si nota che a partire da x_8 la sequenza 1, 3, 9, 27, 17, 19, 25, 11 comincia a ripetersi. ♢

Il più piccolo intero p tale che

$$x_0 = x_p \tag{7.3}$$

è detto *periodo fondamentale della sequenza*, ossia il periodo fondamentale rappresenta la lunghezza del ciclo della sequenza a partire da una generico valore iniziale fino al valore per il quale la sequenza inizia a ripetersi. Poiché i valori generati dalla (7.1) sono sempre minori del modulo m, è chiaro che tra due valori identici non possono presentarsi più di m valori diversi; quindi *la lunghezza del periodo fondamentale non può essere superiore a* m.

Utilizzando il linguaggio R è possibile creare una funzione `gcm()` che permette di ottenere sequenze di numeri pseudocasuali applicando il metodo congruenziale moltiplicativo.

```
> gcm<-function(N,x0,a,m){
+ n<-N
+ y<-numeric(n+1)
+ y[1]<-x0
+ for(i in 2:(n+1)) y[i]<-(a*y[i-1])%%m
+ return(c(y))
+ }
>
> gcm(8,1,3,2^5)
[1]  1  3  9 27 17 19 25 11  1
```

La funzione gcm ha come parametri la cardinalità N dei numeri da generare, il seme iniziale x0, la costante moltiplicativa a e il modulo m del generatore. Alla variabile n è associato N ed è definito un vettore numerico y avente $n+1$ elementi. Il primo elemento del vettore è il seme iniziale x0. Nel ciclo for si associa a y[i] il resto della divisione intera di $(a * y[i-1])$ e m. L'operatore di R %% restituisce il resto della divisione intera. La funzione gcm ritorna un vettore, mediante la funzione di concatenazione c(), contenente la sequenza di numeri pseudocasuali generati. Dopo aver definito tale funzione, otteniamo la sequenza pseudocasuale relativa all'Esempio 7.4.

Il passaggio dalla sequenza pseudocasuale $x_0, x_1, \ldots$ (con $0 \leq x_n < m$) ad una sequenza pseudocasuale di numeri $u_0, u_1, \ldots$ (con $0 \leq u_n < 1$) può essere effettuata ponendo

$$u_n = \frac{x_n}{m} \qquad (n = 0, 1, \ldots). \tag{7.4}$$

Analogamente, se si desidera passare dalla sequenza pseudocasuale $x_0, x_1, \ldots$ (con $0 \leq x_n \leq m-1$) ad una sequenza pseudo–casuale di numeri $u_0, u_1, \ldots$ (con $0 \leq u_n \leq 1$) basterà invece porre

$$u_n = \frac{x_n}{m-1} \qquad (n = 0, 1, \ldots). \tag{7.5}$$

Le (7.4) e (7.5) possono quindi essere utilizzate per generare numeri uniformemente distribuiti nell'intervallo $[0, 1)$ e $[0, 1]$, rispettivamente.

Per ottenere sequenze di numeri $u_0, u_1, \ldots$ appartenenti all'intervallo $[0, 1)$ utilizzando il metodo congruenziale moltiplicativo definiamo la funzione Ugcm(), ottenuta dividendo tutti i numeri della sequenza per il modulo del generatore. Utilizzando poi tale funzione otteniamo la sequenza di numeri uniformi in $[0, 1)$ relativa all'Esempio 7.4.

```
> Ugcm<-function(N,x0,a,m){
+ n<-N
+ y<-numeric(n+1)
+ y[1]<-x0
+ for(i in 2:(n+1)) y[i]<-(a*y[i-1])%%m
+ y<-y/m
+ return(c(y))
+ }
>
> Ugcm(8,1,3,2^5)
[1] 0.03125 0.09375 0.28125 0.84375 0.53125 0.59375 0.78125 0.34375
[9] 0.03125
```

Il *vantaggio del generatore congruente moltiplicativo* è l'estrema velocità di generazione dei numeri. Invece, *alcuni svantaggi* sono:

(a) la sequenza generata è periodica di periodo al più uguale a m;

(b) ogni valore della sequenza è completamente determinato dai tre parametri a, x_0 e m;

(c) esiste una correlazione tra i valori successivi della sequenza.

Come si vedrà nel seguito di questo capitolo occorre effettuare una scelta accurata dei parametri a, x_0 e m del generatore congruente moltiplicativo.

7.3.1 Scelta del modulo come potenza di 2

Sebbene non esistano restrizioni sulla scelta del modulo m, ai fini dell'implementazione su elaboratori binari una possibile scelta del modulo è $m = 2^b$ per rendere più veloce la generazione di ogni numero x_{n+1} della sequenza a partire dal numero x_n usando la relazione congruente (7.1). Con questa scelta la (7.2) diventa

$$x_{n+1} \equiv a\, x_n \pmod{2^b}. \tag{7.6}$$

Vogliamo ora mostrare che nella relazione di ricorrenza (7.6) è conveniente scegliere i parametri a e x_0 nel seguente modo:

(a) $a = 3, 5, \ldots, 2^b - 1$

(b) $x_0 = 1, 3, 5, \ldots, 2^b - 1$

(a) **Scelta della costante moltiplicativa** a

Occorre scegliere $a \neq 1$; infatti se $a = 1$ si ripete sempre lo stesso valore iniziale della sequenza generata da (7.6).

Esempio 7.5 Sia $m = 2^5$, $x_0 = 1$ e $a = 1$. La relazione (7.1) diventa

$$x_{n+1} \equiv x_n \pmod{2^5}$$

e conduce alla sequenza $x_0 = x_1 = x_2 = \ldots = 1$. ♢

Se a è pari, ossia $a = 2k$, l'equazione (7.6) diventa

$$x_{n+1} \equiv 2\, k\, x_n \pmod{2^b},$$

ossia per la (7.2):

$$x_n \equiv x_0\, (2\, k)^n \pmod{2^b}.$$

Quando $n = b$ tale relazione di congruenza conduce a

$$x_b \equiv x_0\, 2^b\, k^b \pmod{2^b},$$

da cui si deduce che $x_b = 0$; tutti i numeri successivamente generati sono quindi nulli, ossia $x_{b+1} = x_{b+2} = \ldots = 0$. Quindi se a è pari, a partire da $n = b$ viene meno l'indipendenza statistica dei valori generati.

Esempio 7.6 Sia $m = 2^5$, $x_0 = 1$ e $a = 2$. La relazione (7.1) diventa

$$x_{n+1} \equiv 2\, x_n \pmod{2^5}$$

e conduce alla sequenza

$$x_0 = 1,\ x_1 = 2,\ x_2 = 4,\ x_3 = 8,\ x_4 = 16,\ x_5 = 0,\ x_6 = 0, \ldots$$

♢

Occorre quindi scegliere a intero positivo dispari diverso da 1 e minore del modulo $m = 2^b$.

(b) **Scelta del valore iniziale** x_0

Se x_0 è pari, ossia $x_0 = 2\,k$, l'equazione (7.2) diventa

$$x_n \equiv 2\, k\, a^n \pmod{2^b}$$

L'operazione di determinare il resto della divisione di $2\,k\,a^n$ per 2^b conduce allo stesso risultato di dividere $k\,a^n$ per 2^{b-1} ossia

$$\frac{2\,k\,a^n}{2^b} = \frac{k\,a^n}{2^{b-1}}.$$

L'equazione di congruenza $x_n \equiv 2\,k\,a^n \pmod{2^b}$ è equivalente a calcolare $x_n = 2\,y_n$, dove $y_n \equiv k\,a^n \pmod{2^{b-1}}$. Scegliere quindi x_0 pari, ossia $x_0 = 2\,k$, è equivalente a ridurre il modulo m da 2^b ad almeno 2^{b-1}. Ovviamente se il resto della divisione è pari si può iterare il ragionamento ottenendo un modulo ancora più piccolo. Poiché il modulo stabilisce il limite superiore della lunghezza del periodo, essere riusciti a ridurre il modulo corrisponde ad aver considerato sequenze il cui periodo si discosta di molto dal limite superiore $m = 2^b$.

Esempio 7.7 Consideriamo le relazioni di congruenza

$$x_{n+1} \equiv 3\, x_n \pmod{2^5}, \qquad y_{n+1} \equiv 3\, y_n \pmod{2^4}$$

e sia $x_0 = 2$ e $y_0 = 1$. Quindi si ha

$$x_n \equiv 2\ 3^n \pmod{2^5}, \qquad y_n \equiv 3^n \pmod{2^4},$$

da cui si ottengono rispettivamente le seguenti sequenze:

$$x_0 = 2,\ x_1 = 6,\ x_2 = 18,\ x_3 = 22,\ x_4 = 2, \ldots$$
$$y_0 = 1,\ y_1 = 3,\ y_2 = 9,\ \ y_3 = 11,\ y_4 = 1, \ldots$$

Essendo x_0 pari, si ha $x_n = 2\,y_n$ $(n = 0, 1, \ldots)$. In entrambe le sequenze il periodo fondamentale è 4; tale periodo si discosta molto da $m = 2^5$. ♢

Occorre quindi scegliere x_0 dispari.

(c) **Altre considerazioni sulla costante moltiplicativa** a

Il modulo m stabilisce soltanto il limite superiore della lunghezza del periodo, che è invece fortemente influenzato dal valore del moltiplicatore a. Tale valore deve essere scelto in maniera tale che il periodo non si discosti molto dal limite superiore 2^b, ma anche in modo tale da avere una sequenza con i requisiti di casualità richiesti.

Esempio 7.8 Sia $m = 2^5$, $x_0 = 1$ con $a = 3, 5, 7, \ldots, 31$. La relazione (7.1) diventa

$$x_{n+1} \equiv a\, x_n \pmod{2^5}$$

e conduce alle sequenze indicate nella Tabella 7.1:

a	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31
x_0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
x_1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31
x_2	9	25	17	17	25	9	1	1	9	25	17	17	25	9	1
x_3	27	29	23	25	19	21			11	13	7	9	3	5	
x_4	17	17	1	1	17	17			17	17	1	1	17	17	
x_5	19	21			27	29			3	5			11	13	
x_6	25	9			9	25			25	9			9	25	
x_7	11	13			3	5			27	29			19	21	
x_8	1	1			1	1			1	1			1	1	
p	8	8	4	4	8	8	2	2	8	8	4	4	8	8	2

Tabella 7.1: Sequenze prodotte con il generatore congruente moltiplicativo per $m = 2^5$, $x_0 = 1$ e $a = 3, 5, 7, \ldots, 31$.

Osservando la Tabella 7.1 si può notare che

(a) il massimo periodo è 8 e si ottiene in otto casi su 15; in quattro casi il periodo è 4 e in tre casi il periodo è 2.

(b) la sequenza ottenuta per $a = 11$ è l'immagine speculare di quella ottenuta per $a = 3$; analogamente $a = 5$ e $a = 13$, $a = 19$ e $a = 27$, $a = 21$ e $a = 29$ sono immagini speculari.

(c) le sequenze aventi periodo maggiore che rispecchiano meglio la distribuzione uniforme sono quelle ottenute scegliendo $a = 5, 13, 21, 29$, ossia $1, 5, 9, 13, 17, 21, 25, 29$ (tutti i numeri sono distanziati di 4). Inoltre le sequenze ottenute scegliendo $a = 3, 11, 19, 27$ (aventi anch'esse periodo 8), ossia $1, 3, 9, 11, 17, 19, 25, 27$ sono più raggruppate e quindi rispecchiano meno la distribuzione uniforme.

$\diamondsuit$

Se si utilizza il generatore congruente moltiplicativo (7.1), per determinare una costante moltiplicativa a^* diversa da a che genera la *stessa sequenza in ordine speculare* (inverso) occorre scegliere

$$a^* \equiv a^{p-1} \pmod{m},$$

dove p è il periodo fondamentale della sequenza (lunghezza del ciclo). Infatti, scegliendo $m = 2^5$, dalla Tabella 7.1 si nota che se $a = 3$ allora $a^* = 11$, se $a = 5$ allora $a^* = 13$, se $a = 19$ allora $a^* = 27$ e se $a = 21$ allora $a^* = 29$.

Inoltre, nel generatore congruente moltiplicativo (7.1) per determinare una *sequenza antitetica* $x_1^*, x_2^*, \ldots$, ossia tale che $x_n^* = m - x_n$ $(n = 1, 2, \ldots)$, basta scegliere come seme della nuova sequenza $x_0^* = m - x_0$. Ad esempio, scegliendo $a = 3$, $m = 2^5$, $x_0 = 1$ e $x_0^* = m - x_0 = 31$, i due generatori moltiplicativi seguenti

$$x_{n+1} \equiv 3\, x_n \pmod{2^5}, \qquad x_{n+1}^* \equiv 3\, x_n^* \pmod{2^5}$$

producono rispettivamente le sequenze:

$$x_0 = 1,\ x_1 = 3,\ x_2 = 9,\ x_3 = 27,\ x_4 = 17,\ x_5 = 19,\ x_6 = 25,\ x_7 = 11$$
$$x_0^* = 31,\ x_1^* = 29,\ x_2^* = 23,\ x_3^* = 5,\ x_4^* = 15,\ x_5^* = 13,\ x_6^* = 7,\ x_7^* = 21$$

e si nota che $x_n + x_n^* = 2^5$ $(n = 0, 1, \ldots)$.

Per scegliere opportunamente a in maniera tale da ottenere un periodo massimo viene in aiuto la teoria dei numeri e si può dimostrare il seguente risultato.

Proposizione 7.1
(a) Sia $m = 2^b$ con $b \geq 4$. Se si scelgono i parametri interi a e x_0 del generatore congruente moltiplicativo minori del modulo e tali che

(i) x_0 positivo dispari,

(ii) $a = 8\,n + 3$ oppure $a = 8\,n + 5$, dove n è un qualsiasi intero non negativo,

si ottiene il periodo massimo 2^{b-2}.

(b) Sia $m = 10^b$ con $b \geq 5$. Se si scelgono i parametri del generatore congruente moltiplicativo minori del modulo e tali che che

(i) x_0 positivo dispari e non divisibile per 5,

(ii) $a = 200\,n \pm z$ dove z può assumere uno dei seguenti 32 valori 3, 11, 13, 19, 21, 27, 29, 37, 53, 59, 61, 67, 69, 77, 83, 91, 109, 117, 123, 131, 133, 139, 141, 147, 163, 171, 173, 179, 181, 187, 189, 197,

si ottiene il periodo massimo $5 \cdot 10^{b-2}$.

Esempio 7.9 Riconsideriamo la Tabella 7.1 in cui $m = 2^5$, $x_0 = 1$, $a = 3, 5, 7, \ldots, 31$. Le condizioni della Proposizione 7.1 sono soddisfatte. Infatti $x_0 = 1$ è dispari e se si sceglie $a = 8n+3$ oppure $a = 8n+5$, ossia $a = 3, 11, 19, 27$ oppure $a = 5, 13, 21, 29$, si ottiene il periodo è massimo $2^3 = 8$. ♢

Essendo il periodo del moltiplicatore sempre minore del modulo m, con il metodo congruenziale moltiplicativo non tutti i numeri sono presenti nel periodo. Il metodo presenta nell'intervallo $[0, m-1)$ delle zone vuote, ossia zone in cui non si presentano numeri. La distribuzione e la distanza tra le zone vuote varia a seconda del seme e del moltiplicatore.

Per elaboratori binari i più comuni valori del modulo m sono 2^{31} e 2^{35}. Un algoritmo introdotto nel 1960 dall'IBM (detto RANDU) era caratterizzato da $m = 2^{31} = 2.147.483.648$ e $a = 65539$; in tal caso poiché $a = 8 \cdot 8192 + 3 = 2^{16} + 3$ il periodo è 2^{29}. Knuth[2] nel 1981 ha dimostrato che tale generatore non ha buone proprietà di casualità. L'algoritmo SIMULA invece utilizza $m = 2^{35}$, $a = 5^{13} = 1.220.703.125$; in tal caso poiché $a = 8 \cdot 152587890 + 5$ il periodo è 2^{33}. Park e Miller nel 1988 hanno mostrato che anche tale generatore non ha buone proprietà statistiche.[3]

I passi dell'algoritmo per generare una sequenza pseudocasuale di numeri $u_0, u_1, \ldots$ (con $0 \leq u_n < 1$) utilizzando il metodo congruenziale moltiplicativo con $m = 2^b$ sono quindi i seguenti:

Algoritmo
STEP 1: fornire in input x_0, a e b tali da soddisfare le ipotesi della Proposizione 7.1
STEP 2: per ogni $n = 1, 2, \ldots, 2^{b-2} - 1$ calcolare

$$x_n \equiv a\, x_{n-1} \pmod{2^b}$$

STEP 3: per ogni $n = 0, 1, \ldots, 2^{b-2} - 1$ calcolare

$$u_n = x_n \cdot 2^{-b}.$$

7.3.2 Scelta del modulo come numero primo

Nel generatore congruente moltiplicativo (7.1) si può scegliere m come numero primo. Si può dimostrare il seguente risultato.

Proposizione 7.2 *Se si scelgono i parametri a e m del generatore congruente moltiplicativo (7.1) tali che*

(i) m è un numero primo (ad esempio, $m = 2^{31} - 1$);

(ii) a (minore del modulo) è un elemento primitivo modulo m

si ottiene il periodo massimo $m - 1$.

[2]Donald E. Knuth è un informatico statunitense (nato nel 1938), professore emerito presso la Stanford University. È autore di una serie di libri sulla programmazione degli algoritmi e la relativa analisi. È inoltre creatore del sistema tipografico TEX.

[3]Stephen K. Park and Keith W. Miller (1988) Random number generators: Good ones are hard to find. Communications of the ACM 31(10), 1192–1201.

Il numero a *è un elemento primitivo modulo* m *se il più piccolo numero intero* s *per il quale* $(a^s - 1)$ *è divisibile per* m *è proprio* $s = m - 1$. Quindi, a è un elemento primitivo modulo m se $a^s - 1$ è un multiplo di m per $s = m - 1$ ma non lo è per valori interi s più piccoli di $m - 1$.

Il generatore congruente moltiplicativo utilizzato nella Proposizione 7.2, genera nel periodo ogni intero compreso nell'intervallo $[1, m - 1)$ prima che la sequenza cominci nuovamente a ripetersi.

Esempio 7.10 Consideriamo il generatore congruente moltiplicativo

$$x_{n+1} \equiv 5\, x_n \pmod 7$$

Tale generatore soddisfa le ipotesi della Proposizione 7.2. Infatti, $m = 7$ è un numero primo ed inoltre $a = 5$ è un elemento primitivo modulo 7, poiché il più piccolo intero s tale che $a^s - 1 = 5^s - 1$ è divisibile per 7 è proprio $s = 6$. Infatti, si nota che

$$a^s - 1 = 5^s - 1 = \begin{cases} 4, & s = 1 \\ 24, & s = 2 \\ 124, & s = 3 \\ 624, & s = 4 \\ 3124, & s = 5 \\ 15624, & s = 6 \end{cases}$$

e risulta $15624 = 2232 \cdot 7$. Il generatore ha quindi un periodo massimo pari a $m - 1 = 6$.

x_0	x_1	x_2	x_3	x_4	x_5	x_6	p
1	5	4	6	2	3	1	6
2	3	1	5	4	6	2	6
3	1	5	4	6	2	3	6
4	6	2	3	1	5	4	6
5	4	6	2	3	1	5	6
6	2	3	1	5	4	6	6

Tabella 7.2: Sequenze prodotte con il generatore congruente moltiplicativo per $m = 7$, $a = 5$ e $x_0 = 1, 2, 3, 4, 5, 6$.

Nella Tabella 7.2 sono visualizzate tutte le sequenze generate per $m = 7$, $a = 5$, $x_0 = 1, 2, 3, 4, 5, 6$ e il loro rispettivo periodo $p = m - 1 = 6$. ♢

Esempio 7.11 Consideriamo il generatore congruente moltiplicativo

$$x_{n+1} \equiv 3\, x_n \pmod{31}$$

Tale generatore soddisfa le ipotesi della Proposizione 7.2. Infatti, $m = 31 = 2^5 - 1$ è un numero primo ed inoltre $a = 3$ è un elemento primitivo modulo 31, poiché il più piccolo intero s tale che $a^s - 1 = 3^s - 1$ è divisibile per 31 è proprio $s = 30$. Infatti, le radici primitive di 31 sono $a = 3, 11, 12, 13, 17, 21, 22, 24$.

Partendo con un seme $x_0 = 1$ si ottiene la sequenza in Tabella 7.3 il cui periodo è 30. ♢

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	3	9	27	19	26	16	17	20	29	25
	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
	13	8	24	10	30	28	22	4	12	5
	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	x_{30}
	15	14	11	2	6	18	23	7	21	1

Tabella 7.3: Sequenza prodotta con il generatore congruente moltiplicativo $x_{n+1} \equiv 3\,x_n \pmod{31}$ con $x_0 = 1$.

Utlizzando R si riottiene la sequenza precedentemente elencata:

```
> gcm(30,1,3,31)
 [1]  1  3  9 27 19 26 16 17 20 29 25 13  8 24 10 30 28 22  4 12  5
[22] 15 14 11  2  6 18 23  7 21  1
```

I passi dell'algoritmo per generare una sequenza pseudocasuale di numeri $u_0, u_1, \ldots$ (con $0 \le u_n < 1$) utilizzando il metodo congruenziale moltiplicativo con $m = 2^b - 1$ sono quindi i seguenti:

Algoritmo
STEP 1: fornire in input x_0, a e m tali da soddisfare le ipotesi della Proposizione 7.2
STEP 2: per ogni $n = 1, 2, \ldots, m-2$ calcolare

$$x_n \equiv a\,x_{n-1} \pmod{m}$$

STEP 3: per ogni $n = 0, 1, \ldots, m-2$ calcolare

$$u_n = x_n/m.$$

Una scelta spesso utilizzata dei parametri del generatore congruente moltiplicativo con m numero primo è $m = 2^{31} - 1 = 2.147.483.647$, $a = 7^5 = 16807$ oppure $a = 630360016$ oppure $a = 742938285$ oppure $a = 397204094$. Nella Tabella 7.4 sono riportati alcuni semi che conducono, almeno nella fase iniziale, a differenti sequenze di numeri pseudocasuali per il generatore congruente moltiplicativo

$$x_{n+1} \equiv 7^5\,x_n \pmod{2^{31} - 1}. \tag{7.7}$$

Tale generatore ha periodo massimo $2^{31} - 2$.

Il generatore congruente moltiplicativo (7.7) è stato estensivamente analizzato nella letteratura e fornisce uno standard minimo qualitativo che deve possedere un generatore di numeri pseudocasuali. Alcuni difetti di questo generatore sono che spesso numeri piccoli tendono ad essere seguiti da numeri grandi e che esiste una correlazione tra numeri successivi, come è possibile evidenziare disponendo i numeri presi a coppie in un grafico bidimensionale.

	Seme Iniziale		Seme Iniziale
1	748932582	16	1651217741
2	1985072130	17	909094944
3	1631331038	18	2095891343
4	67377721	19	203905359
5	366304404	20	2001697019
6	1094585182	21	431442774
7	1767585417	22	1659181395
8	1980520317	23	400219676
9	392682216	24	1904711401
10	64298628	25	263704907
11	250756106	26	350425820
12	1025663860	27	873344587
13	186056398	28	1416387147
14	522237216	29	1881263549
15	213453332	30	1456845529

Tabella 7.4: Semi da utilizzare per il generatore congruente moltiplicativo in (7.7).

A partire dal generatore congruente moltiplicativo (7.7) sono stati introdotti in letteratura generatori sempre più complessi in grado da soddisfare vari tipi di test statistici. Molti di questi generatori sono ottenuti combinando due o più generatori congruenziali moltiplicativi in modo tale da ottenere periodi più lunghi.

Dalla sequenza prodotta con il generatore congruente moltiplicativo otteniamo tramite la (7.4) una sequenza di numeri $u_0, u_1, \ldots$ in $[0, 1)$. Suddividiamo tale sequenza in sottosequenze consecutive di n numeri che sono utilizzate come coordinate di punti di un cubo n–dimensionale di lato unitario. Se i numeri fossero non correlati tenderebbero a coprire tutto il cubo n–dimensionale. In realtà, Marsaglia nel 1968 ha dimostrato che tali punti cadono al più in $(n!\, m)^{1/n}$ iperpiani paralleli. Se, ad esempio, si considera una sequenza suddivisa in sottosequenze di tre numeri consecutivi che sono utilizzati come punti di un cubo di lato unitario, si nota che tali punti cadono al più in $(3!\, m)^{1/3}$ piani paralleli; se si sceglie $m = 2^{15}$, tali punti cadono al più in 58 piani paralleli.

Nella Tabella 7.5 per $m = 2^{31}$ è indicato il valore approssimato $(n!\, m)^{1/n}$, ossia la maggiorazione di Marsaglia per il numero di distinti iperpiani paralleli per alcune scelte di n.

In alcune applicazioni, come ad esempio nel caso del metodo di Monte Carlo per il calcolo di integrali unidimensionali e multidimensionali, un piccolo numero di iperpiani paralleli può fornire risultati della simulazione non accettabili poiché il generatore non si comporterebbe in maniera simile ad un generatore perfettamente casuale.

Occorre sottolineare che il numero effettivo di iperpiani paralleli è, alcune volte, di gran lunga inferiore alla maggiorazione di Marsaglia $(n!\, m)^{1/n}$. Un tipico esempio è fornito dall'algoritmo RANDU dell'IBM basato sul generatore

n	$(n!\,m)^{1/n}$
1	2^{31}
2	2^{16}
3	2344
4	476
5	191
6	107

Tabella 7.5: Per $m = 2^{31}$ è indicata la maggiorazione di Marsaglia per il numero di distinti iperpiani paralleli per alcune scelte di n.

congruenziale moltiplicativo con $m = 2^{31}$ e $a = 65539$; se infatti si scelgono sottosequenze consecutive di $n = 3$ numeri si può dimostrare che tutti i punti cadono in soltanto 15 piani paralleli di un cubo unitario. Tale algoritmo (alcune volte ancora oggi utilizzato) è stato definito da Knuth "really horrible" per le sue insufficienti proprietà statistiche.[4]

7.4 Altri tipi di generatori congruenti

Esistono altri tipi di generatori, ossia *congruenti moltiplicativi misti*, *congruenti additivi*, *di Fibonacci*,...

I *generatori congruenti moltiplicativi misti,* introdotti da Lehmer nel 1951, producono sequenze $\{x_n,\ n = 0, 1, 2, \ldots\}$ come segue:

(i) fissare un intero positivo m detto *modulo* del generatore;

(ii) scegliere degli interi positivi a, c e x_0 minori del modulo m; x_0 è detto *valore iniziale* o *seme*, la costante a ($a \neq 0$) è detta *moltiplicatore* e la costante c è detta *incremento*;

(iii) generare x_n dalla relazione di congruenza lineare

$$x_{n+1} \equiv a\,x_n + c \pmod{m} \tag{7.8}$$

che si legge x_{n+1} *è congruente ad* $a\,x_n + c$ *modulo* m.

Se $c = 0$ si ottiene il *generatore congruente moltiplicativo* descritto nel paragrafo precedente, mentre se $c > 0$ si ottiene il *generatore congruente moltiplicativo misto.*

La procedura inizia con un valore iniziale x_0; se $c > 0$ il seme x_0 può anche essere scelto nullo. Per determinare gli elementi della sequenza $\{x_n,\ n = 1, 2, \ldots\}$ occorre assegnare a x_{n+1} il resto r (con $0 \leq r \leq m-1$) della divisione di $a\,x_n + c$ per il modulo m.

[4] Donald E. Knuth (1981) The Art of Computer Programming. Addison-Wesley.

La relazione di ricorrenza (7.8) è analoga all'equazione alle differenze del primo ordine $x_{n+1} = a\,x_n + c$ che ammette come soluzione

$$x_n = \begin{cases} x_0\,a^n + c\,\dfrac{a^n - 1}{a - 1}, & a \neq 1 \\ x_0 + n\,c, & a = 1. \end{cases}$$

Se $a \neq 1$ la (7.8) può essere scritta come

$$x_n \equiv x_0\,a^n + \frac{a^n - 1}{a - 1}\,c \pmod{m},$$

mentre se $a = 1$ si ha:

$$x_n \equiv x_0 + n\,c \pmod{m}.$$

La scelta dei parametri a, c, x_0 e m è importante per determinare la bontà del generatore considerato. Anche per il generatore congruente moltiplicativo misto si riesce a dimostrare un risultato analogo a quello espresso nella Proposizione 7.1 per il generatore congruente moltiplicativo.

Proposizione 7.3 *Sia $m = 2^b$ con $b \geq 2$. Se si scelgono i parametri a e c del generatore congruente moltiplicativo misto (7.8) minori del modulo e tali che*

(i) c intero positivo dispari;

(ii) $a = 8n + 1$ oppure $a = 8n + 5$ dove n è un qualsiasi intero non negativo (o equivalentemente $a = 4n + 1$);

si ottiene il periodo massimo 2^b.

Esempio 7.12 Sia $m = 2^4$, $x_0 = 7$ e $c = 3$. La relazione (7.8) diventa

$$x_n \equiv a\,x_{n-1} + 3 \pmod{2^4}.$$

Dalla Tabella 7.6 si nota che se $a = 1, 5, 9, 13$ il periodo p è $m = 2^4 = 16$. ♢

Utilizzando il linguaggio R è possibile creare una funzione `gcmm()` che permette di ottenere sequenze di numeri pseudocasuali applicando il metodo congruenziale moltiplicativo misto. I parametri presenti nella funzione sono la lunghezza `N` della sequenza generata, il seme `x0`, la costante moltiplicativa `a`, la costante additiva `c` e il modulo del generatore `m`. Utilizzando poi tale funzione otteniamo la sequenza di numeri pseudocasuali relativa alla Tabella 7.6.

```
> gcmm<-function(N,x0,a,c,m){
+  n<-N
+  y<-numeric(n+1)
+  y[1]<-x0
+ for(i in 2:(n+1)) y[i]<-(a*y[i-1]+c)%%m
+ return(c(y))
+ }
>
```

a	1	5	9	13
x_0	7	7	7	7
x_1	10	6	2	14
x_2	13	1	5	9
x_3	0	8	0	8
x_4	3	11	3	11
x_5	6	10	14	2
x_6	9	5	1	13
x_7	12	12	12	12
x_8	15	15	15	15
x_9	2	14	10	6
x_{10}	5	9	13	1
x_{11}	8	0	8	0
x_{12}	11	3	11	3
x_{13}	14	2	6	10
x_{14}	1	13	9	5
x_{15}	4	4	4	4
x_{16}	7	7	7	7
p	16	16	16	16

Tabella 7.6: Sequenze prodotte con il generatore congruente moltiplicativo misto per $m = 2^4$, $x_0 = 7$, $c = 3$ e $a = 1, 5, 9, 13$.

```
> gcmm(16,7,1,3,2^4)
 [1]  7 10 13  0  3  6  9 12 15  2  5  8 11 14  1  4  7
>
> gcmm(16,7,5,3,2^4)
 [1]  7  6  1  8 11 10  5 12 15 14  9  0  3  2 13  4  7
>
> gcmm(16,7,9,3,2^4)
 [1]  7  2  5  0  3 14  1 12 15 10 13  8 11  6  9  4  7
>
> gcmm(16,7,13,3,2^4)
 [1]  7 14  9  8 11  2 13 12 15  6  1  0  3 10  5  4  7
```

Per ottenere sequenze di numeri $u_0, u_1, \ldots$ appartenenti all'intervallo $[0, 1)$ utilizzando il metodo congruenziale moltiplicativo basta dividere tutti i numeri della sequenza per il modulo del generatore. Utilizzando poi tale funzione otteniamo la sequenza di numeri uniformi in $[0, 1)$ relativa alla Tabella 7.6

```
> Ugcmm<-function(N,x0,a,b,m){
+ n<-N
+ y<-numeric(n+1)
+ y[1]<-x0
+ for(i in 2:(n+1)) y[i]<-(a*y[i-1]+b)%%m
+ y<-y/m
+ return(c(y))
+ }
> Ugcmm(16,7,1,3,2^4)
 [1] 0.4375 0.6250 0.8125 0.0000 0.1875 0.3750 0.5625 0.7500 0.9375
```

```
[10] 0.1250 0.3125 0.5000 0.6875 0.8750 0.0625 0.2500 0.4375
>
> Ugcmm(16,7,5,3,2^4)
 [1] 0.4375 0.3750 0.0625 0.5000 0.6875 0.6250 0.3125 0.7500 0.9375
[10] 0.8750 0.5625 0.0000 0.1875 0.1250 0.8125 0.2500 0.4375
>
> Ugcmm(16,7,9,3,2^4)
 [1] 0.4375 0.1250 0.3125 0.0000 0.1875 0.8750 0.0625 0.7500 0.9375
[10] 0.6250 0.8125 0.5000 0.6875 0.3750 0.5625 0.2500 0.4375
>
> Ugcmm(16,7,13,3,2^4)
 [1] 0.4375 0.8750 0.5625 0.5000 0.6875 0.1250 0.8125 0.7500 0.9375
[10] 0.3750 0.0625 0.0000 0.1875 0.6250 0.3125 0.2500 0.4375
```

I passi dell'algoritmo per generare una sequenza pseudocasuale di numeri $u_0, u_1, \ldots$ (con $0 \leq u_n < 1$) utilizzando il metodo congruenziale moltiplicativo misto con $m = 2^b$ sono quindi i seguenti:

Algoritmo
STEP 1: fornire in input x_0, a, c e b tali da soddisfare le ipotesi del Teorema 7.3
STEP 2: per ogni $n = 1, 2, \ldots, 2^b - 1$ calcolare

$$x_n \equiv a\, x_{n-1} + c \pmod{m},$$

STEP 3: per ogni $n = 0, 1, \ldots, 2^b - 1$ calcolare

$$u_n = x_n \cdot 2^{-b}.$$

Dalle Proposizioni 7.1 e 7.3 emerge che se $m = 2^b$ il generatore congruente moltiplicativo ha un periodo massimo 2^{b-2} che è pari ad $1/4$ del periodo massimo 2^b del generatore congruente moltiplicativo misto. Occorre però sottolineare che il generatore congruente moltiplicativo è solitamente preferito a quello misto per la maggiore casualità con cui spesso si presentano i numeri della sequenza.

Alcune scelte dei parametri m, a e c del generatore congruente moltiplicativo misto sono $m = 2^{31}$, $a = 314159269$, $c = 453806245$ oppure $m = 2^{35}$, $a = 5^{15} = 30517578125$, $c = 1$.

I generatori congruenti moltiplicativi e quelli congruenti moltiplicativi misti precedentemente discussi sono casi particolari del *generatore congruente additivo*

$$x_{n+1} \equiv a_0\, x_n + a_1\, x_{n-1} + \ldots + a_r\, x_{n-r} + c \pmod{m} \qquad (n = r, r+1, \ldots), \tag{7.9}$$

che richiede la conoscenza di $r+1$ valori iniziali $x_0, x_1, \ldots, x_r$, di $r+1$ costanti moltiplicative $a_0, a_1, \ldots, a_r$ e di una costante additiva c.

Un particolare generatore congruente additivo è il *generatore di Fibonacci*:

$$x_{n+1} \equiv x_n + x_{n-1} \pmod{m} \qquad (n = 1, 2, \ldots), \tag{7.10}$$

che richiede la conoscenza di soltanto due valori iniziali. Tale generatore, *di interesse storico*, è detto di Fibonacci per la sua similarità con la successione dei numeri di Fibonacci.

x_0	1	x_7	21	x_{14}	610
x_1	1	x_8	34	x_{15}	987
x_2	2	x_9	55	x_{16}	597
x_3	3	x_{10}	89	x_{17}	584
x_4	5	x_{11}	144	x_{18}	181
x_5	8	x_{12}	233	x_{19}	765
x_6	13	x_{13}	377	x_{20}	946

Tabella 7.7: Sequenza prodotta con il generatore di Fibonacci per $m = 1000$ e $x_0 = x_1 = 1$.

Se, ad esempio, si sceglie $m = 1000$ e $x_0 = x_1 = 1$, i primi valori della sequenza generata con il metodo di Fibonacci sono riportati in Tabella 7.7.

Spesso le sequenze prodotte con il generatore di Fibonacci non sono dotate di buone qualità statistiche poiché presentano una forte correlazione seriale tra i numeri della sequenza.

Utilizzando il linguaggio R è possibile creare una funzione fib() che permette di ottenere sequenze di numeri pseudocasuali applicando il metodo di Fibonacci. I parametri presenti nella funzione sono la lunghezza N della sequenza generata, i valori iniziali x0 e x1 e il modulo del generatore m. Utilizzando poi tale funzione otteniamo la sequenza di numeri pseudocasuali relativa alla Tabella 7.7.

```
> fib<-function(N,x0,x1,m){
+ n<-N
+ y<-numeric(n+1)
+ y[1]<-x0
+ y[2]<-x1
+ for(i in 3:(n+1)) y[i]<-(y[i-1]+y[i-2])%%m
+ return(c(y))
+ }
>
> fib(20,1,1,1000)
[1]   1   1   2   3   5   8  13  21  34  55  89 144 233 377 610 987
[17] 597 584 181 765 946
```

Un altro tipo di generatore congruente additivo è il seguente

$$x_{n+1} \equiv x_n + x_{n-r} \pmod{m} \qquad (n = r, r+1, \ldots), \tag{7.11}$$

che richiede la conoscenza di $r+1$ valori iniziali; tale generatore fornisce sequenze tanto più soddisfacenti dal punto di vista statistico quanto più grande si sceglie il parametro r.

Occorre sottolineare che i generatori congruenziali sono comunemente utilizzati in esperimenti di simulazione e in algoritmi probabilistici; essi sono predicibili e quindi completamente insicuri per scopi crittografici.

In conclusione, le caratteristiche che deve avere un buon generatore di sequenze pseudocasuali sono:

- *ripetibilità*, che garantisce la possibilità di ripetere più volte lo stesso esperimento di simulazione;

- *soddisfacimento di test statistici* di uniformità e di indipendenza, in maniera da verificare che il generatore sia abbastanza simile ad un generatore di numeri perfettamente casuali
- *semplicità e rapidità di utilizzazzione*, in maniera da risultare efficiente computazionalmente
- *periodo lungo*, in maniera tale da poter disporre di sequenze lunghe di numeri pseudocasuali;
- *portabilità*, in maniera tale da rendere l'implementazione del generatore indipendente dalla piattaforma.

7.5 Algoritmi per numeri pseudocasuali in R

Il linguaggio R mette a disposizione dell'utente diversi algoritmi predefiniti in grado di generare numeri pseudocasuali distribuiti uniformemente nell'intervallo $[0, 1)$. Per visualizzare gli algoritmi disponibili basta digitare all'interno della finestra di lavoro help(RNGkind), dove RNG indica *Random Number Generator*. Per sapere quale è il metodo impiegato da R basta digitare il nome della stessa funzione senza alcun argomento tra le parentesi tonde

```
> RNGkind()
[1] "Mersenne-Twister" "Inversion"     "Rejection"
```

che mostra che il metodo utilizzato di default da R per generare numeri pseudocasuali uniformemente distribuiti è quello di *Mersenne-Twister*. L'algoritmo di Mersenne-Twister è un algoritmo per la generazione di numeri pseudocasuali di tipo lineare congruenziale, sviluppato nel 1998 da Makoto Matsumoto e Takuji Nishimura.[5] Tale generatore ha i seguenti vantaggi:

- ha un periodo $2^{19937} - 1$, che è un *numero primo di Mersenne*;

- permette di generare punti equidistribuiti in spazi fino a 623 dimensioni;

- supera numerosi test statistici di casualità.

Ricordiamo che un numero di Mersenne è un *numero primo* esprimibile come $2^p - 1$, con p *intero positivo primo*. I numeri primi di Mersenne prendono il nome dal teologo, filosofo e matematico francese Marin Mersenne (1588-1648) che determinò alcuni di tali numeri. I primi sei numeri primi di Mersenne sono $3, 7, 31, 127, 8191, 131071$. Attualmente il progetto GIMPS (*Great Internet Mersenne Prime Search*), fondato nel 1996 da Gorge Woltman, effettua una ricerca sfruttando migliaia di computer in tutto il mondo, per trovare grandi numeri primi di Mersenne e fornisce una lista dei numeri finora ottenuti dai ricercatori di tutto il mondo.

[5] M. Matsumoto and T. Nishimura, "Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator", ACM Trans. on Modeling and Computer Simulation Vol. 8, No. 1, January pp.3-30 (1998) DOI:10.1145/272991.272995.

In R la funzione di default per generare una sequenza di lunghezza n di numeri uniformemente distribuiti nell'intervallo (a, b) è runif(), che è la funzione di base per la generazione di numeri pseudocasuali:

```
> set.seed(seme)
>runif(n,min=a,max=b)
```

La funzione set.seed(seme) permette di stabilire un intero che definisce il seme del generatore, mentre la funzione runif(n, min = a, max = b) genera una sequenza di numeri uniformi nell'intervallo (a, b). Se si sceglie $a = 0$ e $b = 1$ si genera una sequenza di numeri uniformemente distribuiti in $(0, 1)$. In particolare, se i parametri min e max sono omessi, allora di default sono assunti i valori $a = 0$ e $b = 1$, rispettivamente. Inoltre, se non si precisa il seme iniziale, ossia si omette l'istruzione set.seed(seme), il seme è scelto casualmente, si modifica automaticamente dopo ogni generazione della sequenza e si ottengono sequenze di numeri pseudocasuali differenti ripetendo la generazione della sequenza.

Ad esempio, generiamo una sequenza di numeri uniformi in $(0, 1)$ per due volte scegliendo casualmente il seme:

```
> runif(10)
 [1] 0.79546064 0.39851858 0.54863669 0.02115862 0.96387626
 [6] 0.54110397 0.37855422 0.83335040 0.60994644 0.50043137
>
> runif(10)
 [1] 0.5776252 0.1342261 0.3564915 0.1466240 0.3855348 0.4642717
 [7] 0.5387468 0.9136307 0.7487815 0.5364271
```

Si nota che le due sequenze generate sono differenti.

Generiamo infine una sequenza di numeri uniformi in $(0, 1)$ per due volte consecutive fissando lo stesso seme iniziale:

```
> set.seed(1)
> runif(10)
 [1] 0.26550866 0.37212390 0.57285336 0.90820779 0.20168193
 [6] 0.89838968 0.94467527 0.66079779 0.62911404 0.06178627
>
> set.seed(1)
> runif(10)
 [1] 0.26550866 0.37212390 0.57285336 0.90820779 0.20168193
 [6] 0.89838968 0.94467527 0.66079779 0.62911404 0.06178627
```

In questo ultimo caso le sequenze generate sono le stesse.

In un esperimento di simulazione occorre spesso generare più sequenze statisticamente indipendenti per rappresentare differenti variabili aleatorie. Ad esempio, nel caso di lanci successivi di dadi o di monete occorrono sequenze statisticamente indipendenti per ogni lancio. Invece, nella simulazione di un sistema di servizio occorre generare sequenze (esponenziali, di Erlang, iperesponenziali, ...) per rappresentare i tempi di interarrivo e di servizio e le sequenze generate debbono essere statisticamente indipendenti.

Per ottenere sequenze uniformi in $(0, 1)$ statisticamente indipendenti esistono le seguenti possibilità:

1) utilizzare *semi* (valori iniziali) *differenti*;

2) utilizzare una sola sequenza (generata con un unico seme iniziale) per ottenere istanze di una variabile aleatoria uniforme in $(0, 1)$ e successivamente *partizionare la sequenza generata in distinte sottosequenze* da utilizzare per generare le differenti variabili aleatorie indipendenti;

3) utilizzare *costanti moltiplicative* (moltiplicatori) *differenti* nel metodo congruenziale moltiplicativo.

Nel seguito utilizzeremo valori iniziali (semi) differenti per generare più sequenze statisticamente indipendenti con la funzione `set.seed()` di R.

Capitolo 8

Simulazione di variabili aleatorie discrete

8.1 Introduzione

Nel Capitolo 7 abbiamo considerato alcuni metodi per costruire generatori uniformi nell'intervallo $(0, 1)$. Vogliamo ora introdurre delle tecniche che permettano di *simulare variabili aleatorie discrete e continue* a partire da variabili aleatorie con distribuzione uniforme nell'intervallo $(0, 1)$.

In questo capitolo considereremo la simulazione di variabili aleatorie discrete fornendo metodi generali e metodi specifici per alcune variabili aleatorie. Questi metodi saranno utili per simulare il numero di utenti presenti nel sistema $M/M/1$, $M/M/2$ e $M/M/\infty$ in equilibrio statistico.

8.2 Variabili aleatorie discrete

Sia X una variabile aleatoria discreta che assume valori in un insieme finito o al più numerabile $S = \{x_1, x_2, \ldots\}$ e sia

$$p_j = P(X = x_j) \qquad (j = 1, 2, \ldots) \tag{8.1}$$

la sua funzione di probabilità. Ovviamente si deve avere che

$$p_j \geq 0 \qquad (j = 1, 2, \ldots), \qquad \sum_{j:\, x_j \in S} p_j = 1. \tag{8.2}$$

Per simulare la variabile aleatoria X suddividiamo l'intervallo (0,1) in tanti sottointervalli di ampiezze $p_1, p_2, p_3, \ldots$ in maniera tale che $p_1 + p_2 + \ldots = 1$, come mostrato in Figura 8.1

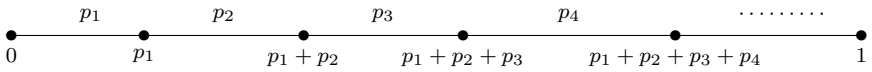

Figura 8.1: Suddivisione dell'intervallo (0,1) in sottointervalli.

Un metodo generale per simulare la variabile aleatoria discreta X che assume valori $0, 1, \ldots$ con probabilità $q_0, q_1, \ldots$ è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Porre

$$X = \begin{cases} x_1, & 0 \le U < p_1 \\ x_2, & p_1 \le U < p_1 + p_2 \\ \cdots\cdots\cdots & \\ x_j, & \sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i \\ \cdots\cdots\cdots & \end{cases}$$

Proposizione 8.1 *La variabile aleatoria discreta X generata con tale algoritmo ha funzione di probabilità (8.1).*

Dimostrazione Per ogni $j = 1, 2, \ldots$ si ha

$$P(X = x_j) = \sum_{k:x_k \in S} P\Big(X = x_j, \sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i\Big)$$
$$= \sum_{k:x_k \in S} P\Big(\sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i\Big)\, P\Big(X = x_j \,\Big|\, \sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i\Big)$$

dove si è posto $\sum_{i=1}^{k-1} p_i = 0$ se $k = 1$. Poiché

$$P\Big(X = x_j \,\Big|\, \sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i\Big) = \begin{cases} 1, & k = j \\ 0, & \text{altrimenti,} \end{cases}$$

segue che

$$P(X = x_j) = P\Big(\sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i\Big) = \sum_{i=1}^{j} p_i - \sum_{i=1}^{j-1} p_i = p_j.$$

La variabile aleatoria X generata con il precedente algoritmo ha quindi funzione di probabilità (8.1). □

L'algoritmo per simulare una variabile aleatoria discreta con funzione di probabilità (8.1) è quindi il seguente:

Algoritmo

STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Se $0 \leq U < p_1$ porre $X = x_1$ e terminare;
Se $p_1 \leq U < p_1 + p_2$ porre $X = x_2$ e terminare;
Se $p_1 + p_2 \leq U < p_1 + p_2 + p_3$ porre $X = x_3$ e terminare;
. . . . . .
. . . . . .

Vogliamo ora mostrare come generare alcune particolari variabili aleatorie discrete.

Problema 8.1 Simulazione di una variabile aleatoria di Bernoulli
Sia X una variabile aleatoria di Bernoulli caratterizzata da funzione di probabilità

$$P(X = 0) = 1 - p, \qquad P(X = 1) = p,$$

con $0 < p < 1$, dove p denota la probablità di successo e $1 - p$ la probabilità di insuccesso. La variabile aleatoria X può descrivere l'esperimento consistente nel lancio di una moneta con probabilità di successo p. Suddividiamo l'intervallo (0,1) in due sottointervalli come mostrato in Figura 8.2

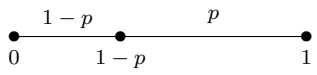

Figura 8.2: Suddivisione dell'intervallo (0,1) in due sottointervalli.

Un metodo generale per simulare X è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Porre

$$X = \begin{cases} 0, & 0 \leq U < 1 - p \\ 1, & 1 - p \leq U < 1. \end{cases}$$

La funzione bernsim permette di generare una sequenza di n numeri con distribuzione Bernoulli con probabilità di successo p. La sequenza che si ottiene può simulare n lanci indipendenti di una moneta.

```
> bernsim<-function(n,p){
+ u<-runif(n)
+ u[which(u<1-p)]<-0
+ u[which(u>=1-p)]<-1
+ return(u)
+ }
>
> set.seed(1)
> bernsim(30,0.5)
 [1] 0 0 1 1 0 1 1 1 1 0 0 0 1 0 1 0 1 1 0 1 1 0 1 0 0 0 0 0 1 0
```

□

Nella funzione bernsim riceve come parametri la lunghezza n della sequenza da generare e la probabilità di successo p. Si genera una sequenza di n numeri uniformi in $(0, 1)$ e si assegnano al vettore u. La funzione which(u $< 1 -$ p) restituisce gli indici del vettore u i cui elementi sono inferiori a $1 -$ p e successivamente agli elementi del vettore u corrispondenti a tali indici è assegnato il valore 0; analogamente, la funzione which(u $>= 1 -$ p) restituisce gli indici del vettore u i cui elementi sono maggiori o uguali a $1 -$ p e successivamente agli elementi del vettore u corrispondenti a tali indici è assegnato il valore 1. La funzione bernsim restituisce la simulazione di una sequenza di n numeri pseudocasuali con distribuzione di Bernoulli con probabilità di successo p.

Possiamo alternativamente utilizzare la funzione predefinita sample() di R nel seguente modo:

```
sample(x, size, replace = FALSE, prob = NULL)
```

dove x è un vettore di valori interi distinti assunti dalla variabile aleatoria discreta X a cui è associato un vettore di probabilità prob; size è la lunghezza della sequenza di numeri pseudocasuali che simulano X, replace indica se le estrazioni sono effettuate con reinserimento (TRUE) oppure senza reinserimento (FALSE). Se si omette di specificare il vettore prob la distribuzione di probabilità di X sarà di default quella equiprobabile.

Ad esempio, per simulare i risultati di 30 prove indipendenti di Bernoulli in cui la probabilità di successo è $p = 1/2$ basta considerare l'istruzione

```
> sample(c(0,1),30,replace=TRUE,prob=c(1/2,1/2))
 [1] 0 1 0 1 1 1 0 1 1 1 1 1 1 1 0 1 0 1 0 0 1 1 0 1 0 0 0 1 0 0
```

o equivalentemente:

```
> sample(0:1,30,replace=TRUE)
 [1] 1 0 0 0 1 0 0 1 0 1 0 1 0 0 0 1 1 0 1 1 0 1 0 0 1 0 1 0 0 0
```

Invece, per simulare i risultati di 30 prove indipendenti di Bernoulli in cui la probabilità di successo è $p = 3/4$ basta considerare l'istruzione

```
> sample(c(0,1),30,replace=TRUE,prob=c(1/4,3/4))
 [1] 1 1 1 1 0 1 0 1 1 1 0 1 0 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1
```

In ogni simulazione, la sequenza generata può essere differente dalla precedente a meno di non inserire l'istruzione set.seed(seme) (fissando il seme iniziale) prima di sample().

Problema 8.2 Simulazione del lancio di un dado

Sia X una variabile aleatoria che descrive il lancio di un dado. Pertanto $P(X = i) = 1/6 \quad (i = 1, 2, \ldots, 6)$. Suddividiamo l'intervallo (0,1) in sei sottointervalli di uguale ampiezza come mostrato in Figura 8.3

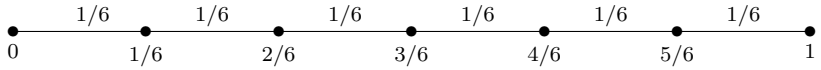

Figura 8.3: Suddivisione dell'intervallo (0,1) in 6 sottointervalli di uguale ampiezza.

Per generare una sequenza di numeri che simulano il lancio del dado possiamo utilizzare il seguente algoritmo:

$$X = \begin{cases} 1, & 0 \leq U < 1/6 \\ 2, & 1/6 \leq U < 2/6 \\ 3, & 2/6 \leq U < 3/6 \\ 4, & 3/6 \leq U < 4/6 \\ 5, & 4/6 \leq U < 5/6 \\ 6, & 5/6 \leq U < 1 \end{cases}$$

Definiamo ora una funzione dadosim che simula una sequenza di lanci di un dado non truccato.

```
> dadosim<-function(n){
+ u<-runif(n)
+ u[which(u<1/6)]<-1
+ u[which(u>=1/6&u<2/6)]<-2
+ u[which(u>=2/6&u<3/6)]<-3
+ u[which(u>=3/6&u<4/6)]<-4
+ u[which(u>=4/6&u<5/6)]<-5
+ u[which(u>=5/6&u<1)]<-6
+return(u)
+}
>
> set.seed(1)
> dadosim(30)
 [1] 2 3 4 6 2 6 6 4 4 1 2 2 5 3 5 3 5 6 3 5 6 2 4 1 2 3 1 3 6 3
```

Possiamo anche utilizzare la funzione predefinita sample() di R per simulare una sequenza di 30 lanci di un dado non truccato nel seguente modo:

```
> sample(1:6,30,replace=TRUE)
 [1] 2 1 4 6 5 5 3 3 5 4 4 3 2 6 4 2 1 3 6 4 6 5 3 3 1 1 5 1 3 4
```

□

Utilizziamo ora la funzione sample() per simulare un esperimento in cui le estrazioni sono eseguite senza reinserimento.

Problema 8.3 Estrazione di biglie da un'urna senza reinserimento
Consideriamo un'urna contenente 20 biglie numerate $1, 2, \ldots, 20$ ed estraiamo 10 biglie in sequenza senza reinserirle nell'urna. Supponiamo che la distribuzione di probabilità iniziale sia quella equiprobabile. Utilizziamo la funzione sample() con replace = FALSE:

```
> sample(1:20,10,replace=FALSE)
 [1]  7 12 15 10 20 13 14 19 18 11
```

□

Esempio 8.1 (Simulazione della somma del lancio di due dadi regolari) Consideriamo l'esperimento consistente nel lancio di due dadi regolari e registriamo la somma dei risultati ottenuti. Denotiamo con X_1 la variabile aleatoria che descrive il risultato ottenuto lanciando il primo dado e con X_2 la variabile aleatoria che descrive il risultato ottenuto lanciando il secondo dado e sia $S = X_1 + X_2$ la variabile aleatoria che descrive la somma dei risultati ottenuti nel lancio dei due dadi. In Tabella 8.1 sono visualizzate le probabilità della variabile aleatoria S ottenute come rapporto tra i casi favorevoli e i casi possibili.

Tabella 8.1: Probabilità $P(S = k)$ relative alla somma del risultati del lancio dei due dadi

S	$P(S = k)$	Risultati dei lanci
2	$1/36 = 0.0278$	$(1,1)$
3	$2/36 = 0.0556$	$(1,2)$ $(2,1)$
4	$3/36 = 0.0833$	$(1,3)$ $(2,2)$ $(3,1)$
5	$4/36 = 0.1111$	$(1,4)$ $(2,3)$ $(3,2)$ $(4,1)$
6	$5/36 = 0.1389$	$(1,5)$ $(2,4)$ $(3,3)$ $(4,2)$ $(5,1)$
7	$6/36 = 0.1667$	$(1,6)$ $(2,5)$ $(3,4)$ $(4,3)$ $(5,2)$ $(6,1)$
8	$5/36 = 0.1389$	$(2,6)$ $(3,5)$ $(4,4)$ $(5,3)$ $(6,2)$
9	$4/36 = 0.1111$	$(3,6)$ $(4,5)$ $(5,4)$ $(6,3)$
10	$3/36 = 0.0833$	$(4,6)$ $(5,5)$ $(6,4)$
11	$2/36 = 0.0556$	$(5,6)$ $(6,5)$
12	$1/36 = 0.0278$	$(6,6)$

Poiché $E(X_i) = 21/6 = 3.5$ e $\text{Var}(X_i) = 105/36 = 2.916667$ per $i = 1, 2$, si ha:

$$E(S) = E(X_1) + E(X_2) = 7, \qquad \text{Var}(S) = \text{Var}(X_1) + \text{Var}(X_2) = 5.833333.$$

Simuliamo ora il lancio dei due dadi e calcoliamo la somma dei risultati.

```
> set.seed(1)
> x1<- sample(1:6,50000,replace=TRUE)
>
> set.seed(2)
> x2<- sample(1:6,50000,replace=TRUE)
>
> s<-x1+x2
>
> table(s)
s
```

```
   2    3    4    5    6    7    8    9   10   11   12
1379 2710 4112 5596 6958 8303 6898 5612 4272 2747 1413
>
> table(s)/length(s)
s
      2       3       4       5       6       7       8       9
0.02758 0.05420 0.08224 0.11192 0.13916 0.16606 0.13796 0.11224
           10      11      12
       0.08544 0.05494 0.02826
>
> mean(s)
[1] 7.0154
> var(s)
[1] 5.824999
```

Abbiamo effettuato 50000 lanci dei due dadi e abbiamo inserito i risultati nei vettori x1 e x2. Il vettore s contiene la somma dei elementi dei due vettori. La funzione table(s) calcola le frequenze assolute ossia quante volte si presenta una certa somma nel vettore e table(s)/length(s) calcola le frequenze relative. Tali frequenze relative sono delle stime, ottenute tramite la simulazione, delle probabilità in Tabella 8.1. Le funzioni mean(s) e var(s) calcolano la media campionaria e la varianza campionaria, così definite:

$$\overline{s} = \frac{1}{n}\sum_{i=1}^{n} s_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(s_i - \overline{s})^2,$$

essendo n la lunghezza del campione simulato. Si nota che la media campionaria e la varianza campionaria stimano accuratamente la media e la varianza della variabile aleatoria S.

Problema 8.4 Simulazione di una variabile aleatoria binomiale di parametri (k, p)

Sia X una variabile aleatoria binomiale di parametri (k, p) caratterizzata da funzione di probabilità

$$P(X = j) = \binom{k}{j} p^j \, (1-p)^{k-j} \qquad (j = 0, 1, \ldots, k). \tag{8.3}$$

La variabile aleatoria binomiale X di parametri (k, p) permette di descrivere il numero di successi ottenuti in k prove indipendenti di Bernoulli in cui la probabilità di successo è p e la probabilità di insuccesso $1 - p$.

Una variabile aleatoria binomiale X di parametri (k, p) può essere facilmente simulata ricordando che essa si può esprimere come la somma di k variabili aleatorie indipendenti $X_1, X_2, \ldots, X_k$ distribuite secondo Bernoulli, ossia

$$X = X_1 + X_2 + \ldots + X_k,$$

dove le X_i sono tali che

$$P(X_i = 0) = 1 - p, \qquad P(X_i = 1) = p \qquad (i = 1, 2, \ldots, k).$$

Per simulare una variabile aleatoria binomiale X di parametri (k, p) possiamo quindi considerare il seguente algoritmo:

Algoritmo
STEP 1: Generare k variabili aleatorie indipendenti $U_1, U_2, \ldots, U_k$ uniformemente distribuite in $(0, 1)$;
STEP 2: Porre

$$X_i = \begin{cases} 0, & 0 \leq U_i < 1 - p \\ 1, & 1 - p \leq U_i < 1 \end{cases} \qquad (i = 1, 2, \ldots, k)$$

STEP 3: Valutare

$$X = \sum_{i=1}^{k} X_i.$$

In R esiste una funzione predefinita

rbinom(n, size, prob)

dove n è lunghezza della sequenza da generare, size è il numero complessivo delle prove (ossia k) e prob è la probabilità di successo in ciascuna prova. Ad esempio, desideriamo simulare una variabile aleatoria binomiale X di parametri $(20, 0.2)$ generando una sequenza lunga 50 contenente i numeri di successi in 20 prove indipendenti di Bernoulli):

```
> sim<-rbinom(50,size=20,prob=0.2)
> sim
[1] 5 4 1 3 2 7 4 1 3 3 2 4 3 4 3 3 6 7 3 5 1 2 5 8 5 5 4 6 1 7 5 3
[33] 3 7 1 4 4 8 2 6 5 1 3 1 4 3 6 4 1 3
```

Per ottenere sempre la stessa sequenza occorre fissare prima di rbinom(n, size, prob) il seme iniziale del generatore tramite la funzione set.seed(seme). □

Problema 8.5 Simulazione di una variabile aleatoria geometrica modificata
Sia X una variabile aleatoria con funzione di probabilità geometrica modificata

$$p_j = P(X = j) = (1 - p)^{j-1} p \qquad (j = 1, 2, \ldots). \tag{8.4}$$

La variabile aleatoria geometrica modificata X permette di descrivere il tempo di attesa per ottenere il primo successo in una successione di prove indipendenti di Bernoulli in cui la probabilità di successo è p e la probabilità di insuccesso è $1 - p$.

Un metodo generale per simulare X è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;

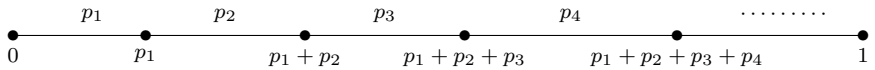

Figura 8.4: Suddivisione dell'intervallo (0,1) in sottointervalli per la geometrica modificata.

STEP 2: Porre

$$X = \begin{cases} 1, & 0 \le U < p_1 \\ 2, & p_1 \le U < p_1 + p_2 \\ \cdots\cdots\cdots & \\ j, & \sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i \\ \cdots\cdots\cdots & \end{cases}$$

dove $\sum_{i=1}^{j-1} p_i = 0$ se $j = 1$. Dalla (8.4) risulta

$$\sum_{i=1}^{j} p_i = \sum_{i=1}^{j} (1-p)^{i-1}\, p = p \sum_{k=0}^{j-1} (1-p)^k = 1 - (1-p)^j .$$

e quindi occorre porre $X = j$ se e solo se:

$$\begin{aligned} \sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i &\iff 1 - (1-p)^{j-1} \le U < 1 - (1-p)^j \\ &\iff (1-p)^j < 1 - U \le (1-p)^{j-1} \\ &\iff j \log(1-p) < \log(1-U) \le (j-1)\log(1-p) \\ &\iff \frac{\log(1-U)}{\log(1-p)} < j \le 1 + \frac{\log(1-U)}{\log(1-p)} . \end{aligned}$$

Per simulare una variabile aleatoria geometrica modificata possiamo quindi utilizzare il seguente algoritmo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Porre

$$X = \left\lfloor \frac{\log(1-U)}{\log(1-p)} + 1 \right\rfloor .$$

dove $\lfloor x \rfloor$ denota il più grande intero minore o uguale di x.

Vogliamo ora generare una sequenza di numeri con distribuzione geometrica modificata e confrontare la media campionaria ottenuta mediante la simulazione e la media teorica $E(X) = 1/p$. Definiamo la funzione geomsim e successivamente generiamo una sequenza di lunghezza 1000 di numeri distribuiti geometricamente con $p = 0.6$.

```
> geomsim<-function(n,p){
+  u<-runif(n)
+ w<-log(1-u)/log(1-p)+1
+ y<-floor(w)
+ return(y)
+ }
>
> set.seed(5)
> x<-geomsim(1000,0.6)
> mean(x)
[1] 1.669
```

dove floor(w) fornisce il vettore dei più grandi interi minori o uguali degli elementi di w.

La media teorica della distribuzione geometrica modificata è $E(X) = 1/p = 5/3 = 1.666$ mentre quella campionaria ottenuta mediante la simulazione è 1.669.

In R esiste una funzione predefinita

$$\mathsf{rgeom}(\mathsf{n}, \mathsf{prob}) + 1$$

dove n è lunghezza della sequenza da generare e prob è la probabilità di successo in ciascuna prova. Ad esempio, se desideriamo simulare 20 tempi di attesa per ottenere il primo successo in lanci successivi di una moneta con probabilità di successo $p = 0.2$ si ha:

```
> sim<-rgeom(20,prob=0.2)+1
> sim
 [1] 28  1  2  1  9  2  2  1  5 12  9  1  2  7  2  4  3  1 21  2
```

□

Problema 8.6 Simulazione di una variabile aleatoria di Poisson
Sia X una variabile aleatoria di Poisson di parametro λ caratterizzata da funzione di probabilità

$$p_j = P(X = j) = \frac{\lambda^j}{j!}\, e^{-\lambda} \qquad (j = 0, 1, \ldots) \tag{8.5}$$

Si noti che le probabilità (8.5) possono essere calcolate tramite la procedura iterativa:

$$\begin{aligned} p_0 &= e^{-\lambda} \\ p_{j+1} &= \frac{\lambda^{j+1}}{(j+1)!}\, e^{-\lambda} = \frac{\lambda}{j+1}\, p_j \qquad (j = 0, 1, \ldots). \end{aligned} \tag{8.6}$$

Un metodo per simulare la variabile aleatoria di Poisson X è quindi il seguente:

Algoritmo

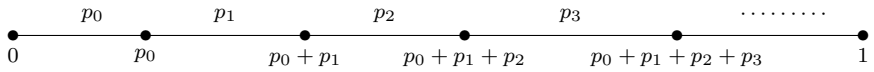

Figura 8.5: Suddivisione dell'intervallo (0,1) in sottointervalli per Poisson.

STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0,1)$;
STEP 2: Porre

$$X = \begin{cases} 0, & 0 \le U < p_0 \\ 1, & p_0 \le U < p_0 + p_1 \\ \cdots\cdots\cdots & \\ n, & \sum_{i=0}^{n-1} p_i \le U < \sum_{i=0}^{n} p_i \\ \cdots\cdots\cdots & \end{cases}$$

dove le p_i sono valutate tramite la procedura iterativa (8.6).

In R esiste la funzione predefinita

rpois(n, lambda)

dove n è lunghezza della sequenza da generare e lambda è il valore medio. Ad esempio, se desideriamo generare una sequenza di 50 numeri pseudocasuali simulando una variabile aleatoria di Poisson di valor medio $\lambda = 3$ si ha:

```
> set.seed(7)
> sim<-rpois(100,lambda=3)
> sim
  [1] 8 2 1 1 2 4 2 7 1 3 1 2 4 1 3 1 3 0 7 2 3 2 9 5 8 1 3 3 7 2 4
 [32] 2 1 1 2 5 3 4 5 3 4 2 4 3 5 2 1 5 5 7 3 4 4 3 4 2 1 1 2 2 1 3
 [63] 4 4 3 4 3 3 2 3 2 3 5 3 1 3 3 3 3 6 5 5 2 4 2 1 1 4 2 2 2 0 5
 [94] 1 4 3 3 2 5 1
> mean(sim)
[1] 3.08
> var(sim)
[1] 3.326869
```

La media teorica della distribuzione di Poisson è $E(X) = \lambda = 3$ mentre la media campionaria ottenuta mediante la simulazione è 3.08. La varianza teorica della distribuzione di Poisson coincide con il valore medio, ossia $\mathrm{Var}(X) = \lambda = 3$ mentre la varianza campionaria ottenuta mediante la simulazione è 3.33. A differenza della distribuzione teorica, si nota che la media campionaria e la varianza campionaria ottenute tramite la simulazione sono differenti.

Per avere sempre la stessa sequenza è stato fissato prima di rpois(n, lambda) il seme iniziale del generatore tramite la funzione set.seed(seme). □

8.3 Simulazione sistemi di servizio in equilibrio statistico

In questo paragrafo utilizzeremo i generatori uniformi e la generazione di variabili aleatorie discrete per simulare

- il numero di utenti nel sistema $M/M/1$ in equilibrio statistico;
- il numero di utenti nel sistema $M/M/2$ in equilibrio statistico;
- il numero di utenti nel sistema $M/M/\infty$ in equilibrio statistico.

8.3.1 Simulazione $M/M/1$ in equilibrio statistico

Consideriamo un sistema di servizio $M/M/1$ caratterizzato da tempi di interarrivo distribuiti esponenzialmente con valore medio $1/\lambda$, tempi di servizio distribuiti esponenzialmente con valore medio $1/\mu$, con unico servitore e capacità del sistema infinita.

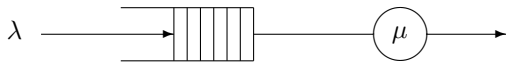

Figura 8.6: Sistema di servizio $M/M/1$.

Il sistema $M/M/1$ non si congestiona se

$$\varrho = \frac{\lambda}{\mu} < 1.$$

Se si denota con N la variabile aleatoria che descrive il numero di utenti presenti nel sistema di servizio $M/M/1$ in condizioni di equilibrio statistico, si ha:

$$q_j = P(N = j) = (1 - \varrho)\, \varrho^j, \qquad j = 0, 1, \ldots \tag{8.7}$$

ossia una funzione di probabilità geometrica. Ricordiamo che una *variabile geometrica* X descrive il *numero di fallimenti che precedono il primo successo in lanci successivi di una moneta* ed è caratterizzata da probabilità

$$P(X = j) = (1 - p)^j\, p, \qquad j = 0, 1, \ldots$$

essendo p la probabilità di successo. Si nota che per il sistema $M/M/1$, il numero N di utenti in condizioni di equilibrio è descritto da una distribuzione geometrica con $p = 1 - \varrho$.

Se $\varrho < 1$, il numero medio di utenti presenti nel sistema $M/M/1$ è

$$E(N) = \frac{\varrho}{1 - \varrho}.$$

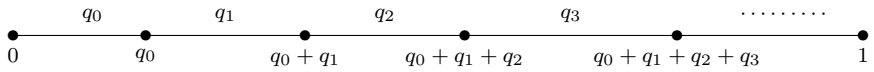

Figura 8.7: Suddivisione dell'intervallo (0,1) in sottointervalli.

Per simulare la variabile aleatoria discreta N occorre suddividere l'intervallo $(0,1)$ in sottointervalli di ampiezza $q_0, q_1, \ldots$ in maniera tale che $q_0+q_1+\ldots = 1$.

Ricordiamo che il metodo generale per simulare il numero di utenti N in un sistema di servizio in equilibrio statistico è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0,1)$;
STEP 2: Porre

$$N = \begin{cases} 0, & 0 \le U < q_0 \\ 1, & q_0 \le U < q_0 + q_1 \\ \vdots\vdots\vdots & \\ j, & \sum_{i=0}^{j-1} q_i \le U < \sum_{i=0}^{j} q_i \\ \vdots\vdots\vdots & \end{cases}$$

dove le q_i sono definite in (8.7).

Seguendo la stessa procedura adottata per la variabile geometrica modificata, determiniamo una procedura più semplice per simulare la variabile aleatoria geometrica N, descrivente il numero di utenti nel sistema $M/M/1$ in condizioni di equilibrio. Dalla (8.7) si ha

$$\sum_{i=0}^{j-1} q_i = (1-\varrho)\sum_{i=0}^{j-1} \varrho^i = 1-\varrho^j.$$

Occorre quindi porre $N = j$ se risulta

$$\begin{aligned} \sum_{i=0}^{j-1} q_i \le U < \sum_{i=0}^{j} q_i &\iff 1-\varrho^j \le U < 1-\varrho^{j+1} \\ &\iff \varrho^{j+1} < 1-U \le \varrho^j \iff (j+1)\log\varrho < \log(1-U) \le j\log\varrho \\ &\iff \frac{\log(1-U)}{\log\ \varrho} - 1 < j \le \frac{\log(1-U)}{\log\ \varrho}, \end{aligned}$$

dove l'ultima disuguaglianza segue poiché, nell'effettuare la divisione per $\log\varrho < 0$, occorre invertire i segni delle disuguaglianze essendo $\varrho < 1$.

Quindi, l'algoritmo per simulare la variabile aleatoria N, che descrive il numero di utenti presenti nel sistema $M/M/1$ in condizioni di equilibrio statistico, è il seguente:

Algoritmo

STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Porre

$$N = \Big\lfloor \frac{\log(1-U)}{\log \varrho} \Big\rfloor,$$

dove $\lfloor x \rfloor$ denota il più grande intero minore o uguale di x.
Vogliamo ora generare una sequenza che descriva il numero di utenti presenti nel sistema $M/M/1$ in condizioni di equilibrio e confrontare la media campionaria ottenuta mediante la simulazione con il valore medio teorico $E(N) = \varrho/(1-\varrho)$. Definiamo la funzione MM1queue e successivamente generiamo una sequenza di lunghezza 1000 relativa al numero di utenti presenti nel sistema $M/M/1$ in condizioni di equilibrio.

```
> MM1queue<-function(n,rho,seme){
+ set.seed(seme)
+ u<-runif(n)
+ w<-log(1-u)/log(rho)
+ N<-floor (w)
+ return(N)
+ }
>
> utenti<-MM1queue(1000,0.6,7)
> mean(utenti)
[1] 1.525
```

dove floor(w) fornisce il vettore dei più grandi interi minori o uguali degli elementi di w.

Scegliendo $\varrho = 0.6$, la media campionaria ottenuta mediante la simulazione è 1.525, mentre la media teorica è

$$E(N) = \frac{\varrho}{1-\varrho} = \frac{0.6}{0.4} = 1.5.$$

La funzione table() permette di determinare la frequenza assoluta del numero di utenti nella sequenza. Il rapporto table(utenti)/length(utenti) fornisce la frequenza relativa del numero di utenti nella sequenza. La frequenza relativa, ottenuta tramite la simulazione, fornisce una stima della distribuzione di probabilità del numero di utenti presenti nel sistema $M/M/1$ in condizioni di equilibrio:

```
> table(utenti)
utenti
  0   1   2   3   4   5   6   7   8   9  10  11  13
394 240 136 103  48  34  15   9  11   5   1   3   1
>
> round(table(utenti)/length(utenti),3)
utenti
    0     1     2     3     4     5     6     7     8
0.394 0.240 0.136 0.103 0.048 0.034 0.015 0.009 0.011
    9    10    11    13
0.005 0.001 0.003 0.001
```

Ad esempio, se $\varrho = 0.6$ la probabilità simulata di avere 0 utenti è 0.394, mentre la probabilità teorica è

$$q_0 = P(N = 0) = 1 - \varrho = 0.4.$$

Per simulare una variabile geometrica esiste una funzione predefinita di R

$$\mathsf{rgeom}(\mathsf{n}, \mathsf{prob})$$

dove n è lunghezza della sequenza da generare e prob è la *probabilità di successo in ciascuna prova*. Pertanto, per simulare il numero di utenti N nel sistema $M/M/1$ in condizioni di equilibrio tramite la funzione predefinita rgeom(n, prob) occorre scegliere come probabilità di successo $p = 1 - \varrho$. Quindi, un metodo alternativo per simulare il numero di utenti presenti nel sistema $M/M/1$ in condizioni di equilibrio è:

```
> MM1queueBis<-function(n,rho,seme){
+ set.seed(seme)
+ N<-rgeom(n, 1-rho)
+ return(N)
+ }
>
utentiBis<-MM1queueBis(1000,0.6,7)
> mean(utentiBis)
[1] 1.523
> table(utentiBis)
utentiBis
  0   1   2   3   4   5   6   7   8   9  10  12
399 237 141  86  52  34  24  10   8   2   6   1
>
> round(table(utentiBis)/length(utentiBis),3)
utentiBis
    0     1     2     3     4     5     6     7     8     9    10
0.399 0.237 0.141 0.086 0.052 0.034 0.024 0.010 0.008 0.002 0.006
   12
0.001
```

Scegliendo $\varrho = 0.6$, la media campionaria ottenuta mediante la simulazione è 1.523, mentre la media teorica è $E(N) = 0.5$. Inoltre, se $\varrho = 0.6$ la probabilità simulata di avere 0 utenti è 0.399, mentre la probabilità teorica è $q_0 = P(N = 0) = 1 - \varrho = 0.4$.

8.3.2 Simulazione $M/M/2$ in equilibrio statistico

Consideriamo un sistema di servizio $M/M/2$ caratterizzato da tempi di interarrivo distribuiti esponenzialmente con valore medio $1/\lambda$, tempi di servizio per servitore distribuiti esponenziali con valore medio $1/\mu$, due servitori identici che lavorano in parallelo e capacità del sistema infinita. Il sistema $M/M/2$ non si congestiona se

$$\varrho_2 = \frac{\lambda}{2\mu} < 1$$

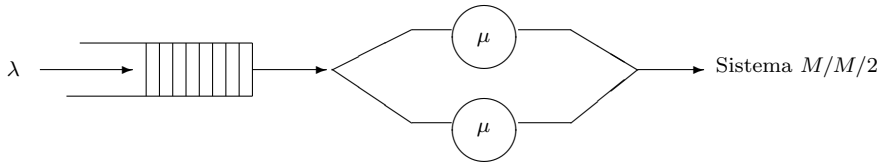

Figura 8.8: Il sistema di servizio $M/M/2$.

Se si denota con N il numero di utenti presenti nel sistema di servizio $M/M/2$ in condizioni di equilibrio statistico, si ha:

$$q_0 = P(N=0) = \frac{1-\varrho_2}{1+\varrho_2}, \qquad q_n = P(N=n) = 2\, q_0\, \varrho_2^n, \qquad n = 1, 2, \ldots \tag{8.8}$$

Se $\varrho_2 < 1$, il numero medio di utenti nel sistema $M/M/2$ è

$$E(N) = \frac{2\varrho_2}{1-\varrho_2^2}.$$

Determiniamo una procedura più semplice, rispetto al metodo generale, per simulare la variabile aleatoria N, descrivente il numero di utenti nel sistema $M/M/2$ in condizioni di equilibrio statistico. Dalla (8.8) segue che

$$\begin{aligned}\sum_{i=0}^{j-1} q_i &= q_0 + 2\, q_0 \sum_{i=1}^{j-1} \rho_2^i = q_0 + 2\, q_0 \Big(\frac{1-\varrho_2^j}{1-\varrho_2} - 1\Big) \\ &= q_0\, \frac{1+\varrho_2-2\varrho_2^j}{1-\varrho_2} = \frac{1+\varrho_2-2\varrho_2^j}{1+\varrho_2} \qquad (j = 1, 2, \ldots).\end{aligned}$$

Pertanto occorre porre $N = j$ se risulta

$$\begin{aligned}&\sum_{i=0}^{j-1} q_i \le U < \sum_{i=0}^{j} q_i \Longleftrightarrow 1 - \frac{2\varrho_2^j}{1+\varrho_2} \le U < 1 - \frac{2\varrho_2^{j+1}}{1+\varrho_2} \\ &\quad\Longleftrightarrow \frac{2\varrho_2^{j+1}}{1+\varrho_2} < 1-U \le \frac{2\varrho_2^j}{1+\varrho_2} \\ &\quad\Longleftrightarrow \log\Big(\frac{2\varrho_2^{j+1}}{1+\varrho_2}\Big) < \log(1-U) \le \log\Big(\frac{2\varrho_2^j}{1+\varrho_2}\Big) \\ &\quad\Longleftrightarrow \log\Big(\frac{2}{1+\varrho_2}\Big) + (j+1)\log\varrho_2 < \log(1-U) \le \log\Big(\frac{2}{1+\varrho_2}\Big) + j\log\varrho_2 \\ &\quad\Longleftrightarrow (j+1)\log\varrho_2 < \log\Big(\frac{(1-U)(1+\varrho_2)}{2}\Big) \le j\log\varrho_2 \\ &\quad\Longleftrightarrow \frac{1}{\log\varrho_2}\log\Big(\frac{(1-U)(1+\varrho_2)}{2}\Big) - 1 < j \le \frac{1}{\log\varrho_2}\log\Big(\frac{(1-U)(1+\varrho_2)}{2}\Big)\end{aligned}$$

dove l'ultima disuguaglianza segue poiché, nell'effettuare la divisione per $\log \varrho_2 < 0$, occorre invertire il segno delle disuguaglianze essendo $\varrho_2 < 1$.

Quindi, l'algoritmo per simulare la variabile aleatoria N, che descrive il numero di utenti presenti nel sistema $M/M/2$ in condizioni di equilibrio statistico, è il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Porre

$$N = \Big\lfloor \frac{1}{\log \varrho_2} \log\Big(\frac{(1-U)(1+\varrho_2)}{2}\Big) \Big\rfloor,$$

dove $\lfloor x \rfloor$ denota il più grande intero minore o uguale di x.
Vogliamo ora generare una sequenza che descriva il numero di utenti presenti nel sistema $M/M/2$ in condizioni di equilibrio e confrontare la media campionaria ottenuta con la simulazione con il valore medio teorico $E(N) = 2\varrho_2/(1-\varrho_2^2)$.

Definiamo la funzione MM2queue e successivamente generiamo una sequenza di lunghezza 1000 relativa al numero di utenti presenti nel sistema $M/M/2$ in condizioni di equilibrio.

```
> MM2queue<-function(n,rho2,seme){
+ set.seed(seme)
+ u<-runif(n)
+ w<-log((1-u)*(1+rho2)/2)/log(rho2)
+ N<-floor (w)
+ return(N)
+ }
>
> utenti<-MM2queue(1000,0.8,3)
>
> mean(utenti)
[1] 4.619
```

La media campionaria ottenuta mediante la simulazione è 4.619, mentre la media teorica è

$$E(N) = \frac{2\varrho_2}{1-\varrho_2^2} = 4.444.$$

```
> table(utenti)
utenti
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
106 170 153  99 106  67  56  43  46  36  21  14  13  14  11   6   4

17  18  19  21  22  23  24  25  26  33  38
5   5   7   3   5   3   1   1   3   1   1
>
> round(table(utenti)/length(utenti),3)
utenti
    0     1     2     3     4     5     6     7     8     9    10
0.106 0.170 0.153 0.099 0.106 0.067 0.056 0.043 0.046 0.036 0.021
   11    12    13    14    15    16    17    18    19    21    22
0.014 0.013 0.014 0.011 0.006 0.004 0.005 0.005 0.007 0.003 0.005
   23    24    25    26    33    38
0.003 0.001 0.001 0.003 0.001 0.001
```

La funzione table() permette di determinare la frequenza assoluta del numero di utenti nella sequenza. Il rapporto table(utenti)/length(utenti) fornisce la frequenza relativa del numero di utenti nella sequenza e fornisce una stima della distribuzione di probabilità, ottenuta con la simulazione, del numero di utenti presenti nel sistema $M/M/2$ in condizioni di equilibrio. Ad esempio, se $\varrho_2 = 0.8$ la probabilità simulata di avere 0 utenti è 0.106, mentre la probabilità teorica è

$$q_0 = P(N = 0) = \frac{1 - \varrho_2}{1 + \varrho_2} = 0.111.$$

8.3.3 Simulazione $M/M/\infty$ in equilibrio statistico

Consideriamo un sistema di servizio $M/M/\infty$ caratterizzato da tempi di interarrivo distribuiti esponenzialmente con valore medio $1/\lambda$, tempi di servizio per servitore distribuiti esponenziali con valore medio $1/\mu$, infiniti servitori identici che lavorano in parallelo e capacità del sistema infinita. Il sistema $M/M/\infty$ raggiunge sempre una situazione di equilibrio statistico e si ha:

$$q_n = P(N = n) = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\Big\{-\frac{\lambda}{\mu}\Big\}, \qquad (n = 0, 1, \ldots), \tag{8.9}$$

ossia una funzione di probabilità di Poisson di parametro $\varrho = \lambda/\mu > 0$. Il valore medio e la varianza del numero di utenti nel sistema sono:

$$E(N) = \varrho = \frac{\lambda}{\mu}, \qquad \mathrm{Var}(N) = \varrho = \frac{\lambda}{\mu}.$$

Per simulare il numero di utenti possiamo utilizzare la funzione predefinita di R rpois(n, ϱ), dove n è la lunghezza della sequenza da generare e $\varrho = \lambda/\mu$ il valore medio:

```
> MMinfty<-function(n,rho,seme){
+ set.seed(seme)
+ N<-rpois(n, rho)
+ return(N)
+ }
>
> utenti<-MMinfty(1000,0.6,7)
> mean(utenti)
[1] 0.601
> var(utenti)
[1] 0.6104094
>
> table(utenti)
utenti
  0   1   2   3   4
549 332  92  23   4
> table(utenti)/length(utenti)
utenti
    0     1     2     3     4
0.549 0.332 0.092 0.023 0.004
```

Scegliendo $\varrho = \lambda/\mu = 0.6$, la media campionaria ottenuta mediante la simulazione è 0.601, mentre la media teorica è $E(N) = 0.6$. Inoltre, se $\varrho = 0.6$ la probabilità simulata di avere 0 utenti è 0.549, mentre la probabilità teorica è $q_0 = P(N = 0) = e^{-\varrho} = 0.5488116$. Si nota che nella simulazione la media campionaria e la varianza campionaria sono prossime ma non uguali come nel modello teorico.

Cambiamo ora il valore del parametro ϱ.

```
> utenti<-MMinfty(1000,2,7)
> mean(utenti)
[1] 2.027
> var(utenti)
[1] 2.002273
>
> table(utenti)
utenti
  0   1   2   3   4   5   6   7
130 269 265 193  92  30  16   5
> table(utenti)/length(utenti)
utenti
    0     1     2     3     4     5     6     7
0.130 0.269 0.265 0.193 0.092 0.030 0.016 0.005
```

Scegliendo $\varrho = \lambda/\mu = 2$, la media campionaria ottenuta mediante la simulazione è 2.027, mentre la media teorica è $E(N) = 2$. Inoltre, se $\varrho = 2$ la probabilità simulata di avere 0 utenti è 0.130, mentre la probabilità teorica è $q_0 = P(N = 0) = e^{-\varrho} = 0.1353353$.

Nei prossimi capitoli cercheremo di simulare tali sistemi di servizio nel transiente. In questo caso sarà necessario simulare i tempi di interarrivo e di servizio. Pertanto nel prossimo capitolo considereremo la generazione di variabili aleatorie continue.

Capitolo 9

Simulazione di variabili aleatorie continue

9.1 Introduzione

In questo capitolo, introduciamo delle tecniche per *simulare variabili aleatorie continue* a partire da variabili aleatorie con distribuzione uniforme nell'intervallo $(0, 1)$. Per la simulazione di variabili aleatorie continue forniremo metodi generali e metodi specifici per alcune variabili aleatorie.

I metodi generali più utilizzati per simulare variabili aleatorie continue sono due:

- *metodo di inversione della funzione di distribuzione*;

- *metodo di reiezione*.

Nel seguito denoteremo con U una variabile aleatoria uniformemente distribuita nell'intervallo $(0, 1)$ e con

$$f_U(u) = \begin{cases} 1, & 0 < u < 1 \\ 0, & \text{altrimenti,} \end{cases} \qquad F_U(u) = P(U < u) = \begin{cases} 0, & u \leq 0 \\ u, & 0 < u \leq 1 \\ 1, & u > 1, \end{cases} \tag{9.1}$$

la sua funzione densità e la sua funzione di distribuzione, rispettivamente.

9.2 Metodo di inversione della funzione di distribuzione

Desideriamo simulare una variabile aleatoria continua X con funzione di distribuzione $F_X(x)$. Il metodo di inversione della funzione di distribuzione si basa sulla seguente proposizione:

Proposizione 9.1 *Sia U una variabile aleatoria uniformemente distribuita nell'intervallo $(0,1)$. Definiamo una variabile aleatoria X continua tramite la trasformazione*

$$X = F^{-1}(U), \quad \text{o equivalentemente} \quad U = F(X) \tag{9.2}$$

essendo $F(x)$ una funzione di distribuzione invertibile. La funzione di distribuzione della variabile aleatoria X è:

$$F_X(x) = \begin{cases} 0, & x \le F^{-1}(0) \\ F(x), & F^{-1}(0) < x \le F^{-1}(1) \\ 1, & x > F^{-1}(1). \end{cases} \tag{9.3}$$

Dimostrazione Poiché $F(x)$ è una funzione di distribuzione, essa è non decrescente. Ricordando la (9.2) e facendo uso di (9.1) si ha

$$\begin{aligned} F_X(x) = P(X < x) &= P\big[F^{-1}(U) < x\big] = P\big[U < F(x)\big] \\ &= \begin{cases} 0, & F(x) \le 0 \\ F(x), & 0 < F(x) \le 1 \\ 1, & F(x) > 1, \end{cases} \end{aligned}$$

da cui segue direttamente la (9.3). □

La Proposizione 9.1 mostra che è possibile simulare una variabile aleatoria X continua caratterizzata da funzione di distribuzione $F(x)$ invertibile simulando una variabile aleatoria U uniformemente distribuita in $(0,1)$ e ponendo $X = F^{-1}(U)$ o equivalentemente $U = F(X)$.

L'algoritmo per simulare la variabile aleatoria continua X con il metodo di inversione della funzione di distribuzione è quindi il seguente:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0,1)$;
STEP 2: Porre

$$X = F^{-1}(U) \qquad \text{o equivalentemente} \quad U = F(X).$$

Per generare una sequenza di valori reali di una data variabile aleatoria X con funzione di distribuzione $F(x)$ si procede quindi nel seguente modo:

- si genera un reale u_i uniformemente distribuito nell'intervallo $(0,1)$, si pone $u_i = F(x_i)$;

- si ricava tramite l'applicazione della funzione inversa il valore $x_i = F^{-1}(u_i)$ corrispondente al numero u_i generato;

- si itera questo metodo ai vari elementi della sequenza $u_1, u_2, \ldots$ ottenendo la sequenza $x_1, x_2, \ldots$ di numeri reali che costituiscono osservazioni di una variabile aleatoria X con funzione di distribuzione $F(x)$.

Problema 9.1 Simulazione di una variabile aleatoria distribuita uniformemente in (a, b)
Consideriamo una variabile aleatoria X uniformemente distribuita nell'intervallo (a, b) e sia

$$F_X(x) = P(X < x) = \begin{cases} 0, & x \le a, \\ \dfrac{x-a}{b-a} & a < x \le b, \\ 1, & x > b \end{cases}$$

la sua funzione di distribuzione. Applicando il metodo di inversione della funzione di distribuzione possiamo scrivere

$$U = F(X) = \frac{X-a}{b-a},$$

da cui segue che

$$X = a + (b-a)\,U,$$

dove U è una variabile aleatoria uniformemente distribuita nell'intervallo $(0, 1)$.

Il *metodo di inversione della funzione di distribuzione* permette di simulare la variabile aleatoria X *uniformemente distribuita in* (a, b) nel seguente modo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
STEP 2: Porre

$$X = a + (b-a)\,U.$$

A partire dalla sequenza $u_1, u_2, \ldots$ di numeri uniformemente distribuiti in $(0, 1)$ possiamo quindi ottenere la sequenza $x_1, x_2, \ldots$ di numeri uniformemente distribuiti in (a, b) tramite la relazione

$$x_i = a + (b-a)\,u_i \qquad (i = 1, 2, \ldots).$$

Tale procedura è stata applicata nel Capitolo 6 per calcolare l'area sottesa da una curva con il metodo di Monte Carlo.

Definiamo in R la funzione unifsim che permette di generare una sequenza di n numeri pseudocasuali distribuiti uniformemente nell'intervallo (a, b). Successivamente utilizziamo tale funzione per generare una sequenza di 30 numeri uniformi nell'intervallo $(0, 5)$ a partire dalla sequenza di numeri uniformi in $[0, 1)$:

```
> unifsim<-function(n,a,b){
+ x<-a+(b-a)*runif(n)
+ return(x)
+ }
```

```
>
> set.seed(3)
> round(unifsim(30,0,5),3)
 [1] 0.840 4.038 1.925 1.639 3.011 3.022 0.623 1.473 2.888 3.155
[11] 2.560 2.525 2.670 2.786 4.340 4.149 0.557 3.518 4.487 1.399
[21] 1.141 0.077 0.645 0.467 1.184 3.956 2.999 4.551 2.802 3.779
```

La funzione round(x, k) permette di arrotondare gli elementi di un vettore x contenente numeri reali considerando solo k cifre decimali.

Per generare una sequenza di numeri uniformi nell'intervallo (a, b) si può anche utilizzare la funzione predefinita di R

$$\mathsf{runif}(\mathsf{n}, \mathsf{min} = \mathsf{a}, \mathsf{max} = \mathsf{b})$$

che utilizza lo stesso algoritmo. Infatti, con lo stesso seme iniziale, si ottiene la stessa sequenza precedente di 30 numeri uniformi nell'intervallo $(0, 5)$.

```
> set.seed(3)
> round(runif(30,min=0,max=5),3)
 [1] 0.840 4.038 1.925 1.639 3.011 3.022 0.623 1.473 2.888 3.155
[11] 2.560 2.525 2.670 2.786 4.340 4.149 0.557 3.518 4.487 1.399
[21] 1.141 0.077 0.645 0.467 1.184 3.956 2.999 4.551 2.802 3.779
```

□

Problema 9.2 Simulazione di una variabile aleatoria esponenziale
Consideriamo una variabile aleatoria X esponenzialmente distribuita con valore medio $1/\lambda$ e sia

$$F_X(x) = P(X < x) = \begin{cases} 0, & x \leq 0, \\ 1 - e^{-\lambda x}, & x > 0 \end{cases}$$

la sua funzione di distribuzione. Applicando il metodo di inversione della funzione di distribuzione si ha:

$$U = F(X) = 1 - e^{-\lambda X}$$

ossia

$$e^{-\lambda X} = 1 - U \quad \Longrightarrow \quad -\lambda X = \log(1 - U),$$

da cui segue

$$X = -\frac{1}{\lambda} \log(1 - U),$$

dove U è una variabile aleatoria uniformemente distribuita nell'intervallo $(0, 1)$.

Il *metodo di inversione della funzione di distribuzione* permette di simulare la variabile aleatoria X esponenzialmente distribuita con valore medio $1/\lambda$ nel seguente modo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;

STEP 2: Porre

$$X = -\frac{1}{\lambda}\log(1-U).$$

A partire dalla sequenza $u_1, u_2, \ldots$ di numeri uniformemente distribuiti in $(0,1)$, possiamo ottenere la sequenza $x_1, x_2, \ldots$ di numeri esponenzialmente distribuiti con valore medio $1/\lambda$ tramite la relazione

$$x_i = -\frac{1}{\lambda}\log(1-u_i) \qquad (i = 1, 2, \ldots).$$

Se U è uniformemente distribuita in $(0,1)$, anche la variabile aleatoria $1-U$ è uniformemente distribuita in $(0,1)$. Quindi X può essere anche espressa tramite la relazione $X = -(1/\lambda)\log U$.

Definiamo ora in R la funzione expsim che permetta di generare una sequenza di n numeri distribuiti esponenzialmente di valore medio $1/\lambda$. Successivamente utilizziamo tale funzione per generare una sequenza di 30 numeri distribuiti esponenzialmente con $\lambda = 2$ e seme iniziale 3.

```
> expsim<-function(n,lambda){
+ x<- -log(1-runif(n))/lambda
+ return(x)
+ }
>
> set.seed(3)
> round(expsim(30,2),3)
 [1] 0.092 0.824 0.243 0.199 0.461 0.464 0.067 0.174 0.431 0.498
[11] 0.359 0.352 0.382 0.407 1.012 0.885 0.059 0.608 1.139 0.164
[21] 0.130 0.008 0.069 0.049 0.135 0.783 0.458 1.205 0.411 0.705
```

□

In R esiste una funzione predefinita che simula la variabile aleatoria esponenziale

```
rexp(n, rate=lambda)
```

dove n è lunghezza della sequenza da generare e *rate* è la frequenza λ della densità esponenziale.

⋆ 9.1 Simulazione tempi di interarrivo e di servizio del sistema $M/M/1$
Consideriamo un sistema di servizio $M/M/1$. Desideriamo simulare i tempi di interarrivo esponenzialmente distribuiti con valore medio $E(T) = 1/\lambda$, e i tempi di servizio esponenzialmente distribuiti con valore medio $E(S) = 1/\mu$.

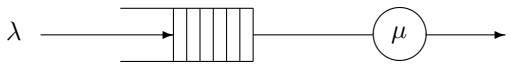

Figura 9.1: Sistema di servizio $M/M/1$.

In questo caso, occorre generare due sequenze indipendenti di numeri uniformi in $(0,1)$, ossia

$$u_1, u_2, \ldots, \qquad v_1, v_2, \ldots.$$

I *tempi di interarrivo* t_i possono essere generati con

$$t_i = -\frac{1}{\lambda}\log(1-u_i), \quad i = 1, 2, \ldots$$

e i *tempi di servizio* possono essere generati con

$$s_i = -\frac{1}{\mu}\log(1-v_i), \quad i = 1, 2, \ldots$$

Il codice seguente permette di creare una sequenza di 1000 tempi di interarrivo tint distribuiti esponenzialmente con valore medio $E(T) = 1/\lambda = 1$ e una sequenza di 1000 tempi di servizio tserv distribuiti esponenzialmente con valore medio $E(S) = 1/\mu = 1/2$. Si nota che

$$\varrho = \frac{\lambda}{\mu} = 0.5 < 1$$

e quindi il sistema di servizio $M/M/1$ non si congestiona. Le funzioni mean(tint) e mean(tserv) permettono di ottenere una stima del tempo medio di interarrivo e una stima del tempo medio di servizio tramite la simulazione. Si nota che i valori medi simulati dei tempi di interarrivo e di servizio sono vicini ai valori medi teorici $E(T) = 1/\lambda = 1$ e $E(S) = 1/\mu = 1/2$.

```
> set.seed(3)
> tint<-rexp(1000,1)
> mean(tint)
[1] 0.9837254
>
> set.seed(5)
> tserv<-rexp(1000,2)
> mean(tserv)
[1] 0.5013411
```

♢

⋆ 9.2 Simulazione tempi di interarrivo e di servizio del sistema $M/U/1$
Consideriamo il sistema di servizio $M/U/1$. Desideriamo simulare i tempi di interarrivo esponenzialmente distribuiti con valore medio $E(T) = 1/\lambda$ e i tempi di servizio uniformemente distribuiti nell'intervallo $(0, 2/\mu)$, con valore medio $E(S) = 1/\mu$.

In questo caso occorre generare due sequenze indipendenti di numeri uniformi in $(0, 1)$, ossia

$$u_1, u_2, \ldots \qquad\qquad v_1, v_2, \ldots.$$

I *tempi di interarrivo* t_i possono essere generati con

$$t_i = -\frac{1}{\lambda}\log(1-u_i), \quad i = 1, 2, \ldots$$

e i tempi di servizio possono essere calcolati ponendo $a = 0$ e $b = 2/\mu$, ossia generati con

$$s_i = \frac{2}{\mu}\, v_i \quad i = 1, 2, \ldots$$

Il codice seguente permette di creare una sequenza di 1000 tempi di interarrivo tint distribuiti esponenzialmente con valore medio $E(T) = 1/\lambda = 1$ e una sequenza di 1000 tempi di servizio tserv distribuiti uniformemente nell'intervallo $(0, 1/2)$ con valore medio $E(S) = 1/\mu = 1/4 = 0.25$. Si nota che

$$\varrho = \frac{\lambda}{\mu} = 0.25 < 1$$

e quindi il sistema di servizio $M/U/1$ non si congestiona. La funzione mean(tint) e mean(tserv) permette di ottenere una stima del tempo medio di interarrivo e una stima del tempo medio di servizio ottenuti tramite simulazione. Si nota che i valori medi simulati dei tempi di interarrivo e di servizio sono vicini ai valori medi teorici $E(T) = 1/\lambda = 1$ e $E(S) = 1/\mu = 1/4 = 0.25$.

```
> set.seed(3)
> tint<-rexp(1000,1)
> mean(tint)
[1] 0.9837254
>
> set.seed(5)
> tserv<-runif(1000,0,0.5)
> mean(tserv)
[1] 0.2507395
```

♢

Problema 9.3 Simulazione di una variabile aleatoria distribuita secondo Rayleigh
Consideriamo una variabile aleatoria X distribuita secondo Rayleigh e siano

$$f_X(x) = \begin{cases} x\, e^{-x^2/2}, & x > 0 \\ 0, & \text{altrimenti,} \end{cases} \qquad F_X(x) = P(X < x) = \begin{cases} 0, & x \le 0 \\ 1 - e^{-x^2/2}, & x > 0 \end{cases}$$

la sua densità di probabilità e la sua funzione di distribuzione, rispettivamente.

La distribuzione di Rayleigh svolge un importante ruolo nella modellizzazione della durata di vita di componenti elettroniche e nella teoria dell'affidabilità dei sistemi. È caratterizzata da valore medio $E(X) = \sqrt{\pi/2} = 1.25331$ e varianza $\mathrm{Var}(X) = 2 - \pi/2 = 0.429204$.

Applicando il metodo di inversione della funzione di distribuzione possiamo scrivere

$$U = F(X) = 1 - e^{-X^2/2},$$

ossia

$$e^{-X^2/2} = 1 - U \quad \Longrightarrow \quad -\frac{X^2}{2} = \log(1-U) \quad \Longrightarrow \quad X^2 = -2\log(1-U),$$

da cui, essendo X una variabile positiva, segue

$$X = \sqrt{-2\,\log(1-U)},$$

dove U è una variabile aleatoria uniformemente distribuita nell'intervallo $(0, 1)$.

Il *metodo di inversione della funzione di distribuzione* quindi permette di simulare la variabile aleatoria X distribuita secondo Rayleigh nel seguente modo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
STEP 2: Porre

$$X = \sqrt{-2 \log(1 - U)},$$

A partire dalla sequenza $u_1, u_2, \ldots$ di numeri uniformemente distribuiti in $(0, 1)$, possiamo quindi ottenere la sequenza $x_1, x_2, \ldots$ di numeri distribuiti secondo Rayleigh tramite la relazione

$$x_i = \sqrt{-2 \log(1 - u_i)}, \qquad i = 1, 2, \ldots$$

Definiamo ora in R la funzione raysim che permette di generare una sequenza di n numeri distribuiti secondo Rayleigh. Successivamente utilizziamo tale funzione per generare una sequenza di 30 numeri pseudocasuali con seme iniziale 3.

```
> raysim<-function(n){
+ y<-sqrt(-2*log(1-runif(n)))
+ return(y)
+ }
>
> set.seed(3)
>  round(raysim(30),3)
 [1] 0.607 1.815 0.986 0.891 1.358 1.362 0.516 0.835 1.313 1.412
[11] 1.198 1.186 1.236 1.277 2.012 1.882 0.486 1.560 2.134 0.810
[21] 0.720 0.176 0.526 0.443 0.735 1.770 1.353 2.195 1.282 1.679
>
> mean(raysim(5000))
[1] 1.257916
> var(raysim(5000))
[1] 0.4285866
```

□

Esempio 9.1 Desideriamo generare N punti distribuiti uniformemente in un cerchio di raggio R centrato nell'origine.
Possiamo procedere generando un angolo θ uniforme in $[0, 2\pi)$ e un raggio r uniforme in $[0, R)$. Un generico punto ha coordinate (x, y), dove $x = r \cos\theta$ e $y = r \sin\theta$. Si nota che $x^2 + y^2 = r^2$, da cui $r = \sqrt{x^2 + y^2}$.

```
> gencerchio<-function(N, R){
+ set.seed(5)
+ u<-runif(N)
+ set.seed(7)
+ v<-runif(N)
+ r<-numeric(N)
+ teta<-numeric(N)
+ ascissa<-numeric(N)
+ ordinata<-numeric(N)
+ df<-data.frame(X=1:N,Y=1:N)
```

```
+ row.names(df)<-1:N
+ for(i in 1:N){
+ r[i]<-R*sqrt(u[i])
+ teta[i]<-2*pi*v[i]
+ ascissa[i]<-r[i]*cos(teta[i])
+ ordinata[i]<-r[i]*sin(teta[i])
+ df[i,1]<-ascissa[i]
+ df[i,2]<-ordinata[i]
+ }
+ return(df)
+ }
```

Utilizziamo due vettori u e v di lunghezza N uniformi in $(0,1)$ e definiamo quattro vettori numerici r, teta, ascissa e ordinata di lunghezza N. Consideriamo poi un data frame df con due colonne di lunghezza N che contiene le ascisse e le ordinate degli N punti. Nell'iterazione i-esima del ciclo for generiamo il raggio r(i) e l'angolo teta[i] che serviranno per determinare l'ascissa ascissa[i] e l'ordinata ordinata[i] del punto i-esimo nel cerchio. Ad esempio,

```
> gencerchio(10,2)
              X           Y
1   0.89273464 -0.06231105
2  -1.32545629  0.99198791
3   1.43095087  1.27274613
4   0.96578844  0.45260415
5   0.02540332  0.64649453
6   0.43690678 -1.61658353
7  -0.77915330  1.22668662
8   1.77007898 -0.31394462
9   0.98662929  1.68895327
10 -0.64286586  0.16892472
```

genera 10 punti in un cerchio di raggio 2 centrato nell'origine. Con la funzione

```
plot(gencerchio(1000,2), col="blue")
```

possiamo visualizzare la generazione di 1000 punti nel cerchio di raggio 2 centrato nell'origine ♢

9.3 Metodo di reiezione

Il metodo di inversione della funzione di distribuzione trova un limite preciso nel suo campo di applicazione in presenza di variabili aleatorie X per le quali risulti molto difficile, o addirittura impossibile, esprimere analiticamente la funzione di distribuzione inversa. Un tipico esempio di funzione di distribuzione difficile da invertire è quella della variabile aleatoria normale standard la cui funzione di distribuzione è

$$F(x) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{x} e^{-z^2}\, dz, \qquad x \in \mathbb{R}.$$

Spesso in questi casi è conveniente utilizzare il metodo di reiezione.

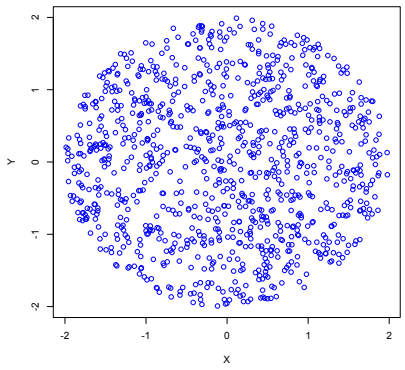

Figura 9.2: Generazione di 1000 punti in un cerchio di raggio 2 centrato nell'origine

Sia X la variabile aleatoria che si desidera simulare caratterizzata da densità di probabilità $f(x)$. Supponiamo di disporre di un metodo che permetta di simulare un'altra variabile aleatoria Y caratterizzata da densità di probabilità $g(y)$. Il *metodo di reiezione* simula la variabile aleatoria X nel seguente modo:

Algoritmo
STEP 1: Generare indipendentemente una variabile aleatoria Y avente densità $g(y)$ e una variabile aleatoria U uniformemente distribuita nell'intervallo $(0, 1)$;
STEP 2: Sia c una costante reale positiva scelta in modo tale che

$$\frac{f(y)}{c\, g(y)} \leq 1 \tag{9.4}$$

per ogni y tale che $f(y) > 0$ e $g(y) > 0$. Se risulta

$$U < \frac{f(Y)}{c\, g(Y)} \tag{9.5}$$

porre $X = Y$, altrimenti ritornare al passo *1.*

Proposizione 9.2 *La variabile aleatoria continua X generata con il metodo di reiezione è caratterizzata da densità di probabilità $f(x)$.*

Dimostrazione Vogliamo dimostrare che la variabile aleatoria X che si desidera simulare con il metodo di reiezione è caratterizzata da densità di probabilità $f(x)$. Dal passo 2 dell'algoritmo segue che

$$P(X < x) = P\left\{ Y < x \,\middle|\, U < \frac{f(Y)}{c\, g(Y)} \right\} = \frac{P\left\{ Y < x,\, U < \frac{f(Y)}{c\, g(Y)} \right\}}{P\left\{ U < \frac{f(Y)}{c\, g(Y)} \right\}}$$

$$= \frac{\displaystyle\int_{-\infty}^{\infty} P\Big\{Y < x,\, U < \frac{f(Y)}{c\, g(Y)} \,\Big|\, Y = y\Big\}\, g(y)\, dy}{\displaystyle P\Big\{U < \frac{f(Y)}{c\, g(Y)}\Big\}}$$

$$= \frac{\displaystyle\int_{-\infty}^{x} P\Big\{U < \frac{f(y)}{c\, g(y)}\Big\}\, g(y)\, dy}{\displaystyle P\Big\{U < \frac{f(Y)}{c\, g(Y)}\Big\}}. \tag{9.6}$$

Poiché U è uniformemente distribuita in $(0, 1)$, facendo uso di (9.1) in (9.6) si ricava:

$$P(X < x) = \frac{\displaystyle\int_{-\infty}^{x} \frac{f(y)}{c\, g(y)}\, g(y)\, dy}{\displaystyle P\Big\{U < \frac{f(Y)}{c\, g(Y)}\Big\}} = \frac{\displaystyle\int_{-\infty}^{x} f(y)\, dy}{\displaystyle c\, P\Big\{U < \frac{f(Y)}{c\, g(Y)}\Big\}}. \tag{9.7}$$

Procedendo al limite quando x tende all'infinito nella (9.7), si ottiene:

$$1 = \frac{\displaystyle\int_{-\infty}^{\infty} f(y)\, dy}{\displaystyle c\, P\Big\{U < \frac{f(Y)}{c\, g(Y)}\Big\}} = \frac{1}{\displaystyle c\, P\Big\{U < \frac{f(Y)}{c\, g(Y)}\Big\}},$$

ossia

$$P\Big\{U < \frac{f(Y)}{c\, g(Y)}\Big\} = \frac{1}{c}. \tag{9.8}$$

Sostituendo (9.8) in (9.7) si ha:

$$P(X < x) = \frac{\displaystyle\int_{-\infty}^{x} f(y)\, dy}{\displaystyle c\, P\Big\{U < \frac{f(Y)}{c\, g(Y)}\Big\}} = \int_{-\infty}^{x} f(y)\, dy,$$

ossia $P(X < x)$ coincide con la funzione di distribuzione di una variabile aleatoria continua di densità $f(x)$. La variabile aleatoria X generata con il metodo di reiezione è quindi caratterizzata da densità di probabilità $f(x)$. □

Problema 9.4 Simulazione del valore assoluto di una variabile aleatoria normale standard
Sia Z una variabile aleatoria di densità normale standard

$$f_Z(z) = \frac{1}{\sqrt{2\,\pi}}\, \exp\Big\{-\frac{z^2}{2}\Big\}, \qquad z \in \mathbb{R}, \tag{9.9}$$

La densità normale standard ha una curva a campana, simmetrica intorno allo zero e caratterizzata da valore medio nullo e varianza unitaria, ossia $E(Z) = 0$ e $\mathrm{Var}(Z) = 1$. Consideriamo la variabile aleatoria

$$X = |Z|,$$

con funzione di distribuzione

$$P(X < x) = P(|Z| < x) = P(-x < Z < x) = P(Z < x) - P(Z < -x)$$

e densità di probabilità:

$$f(x) = \begin{cases} \dfrac{2}{\sqrt{2\,\pi}} \exp\Big\{-\dfrac{x^2}{2}\Big\}, & x > 0 \\ 0, & \text{altrimenti.} \end{cases}$$

Si può mostrare che la variabile aleatoria X è caratterizzata da valore medio, momento del secondo ordine e varianza:

$$E(X) = \sqrt{\frac{2}{\pi}} = 0.797885, \qquad E(X^2) = 1, \qquad \mathrm{Var}(X) = 1 - 2/\pi = 0.36338.$$

Desideriamo simulare la variabile aleatoria X utilizzando il metodo di reiezione. Nell'algoritmo di reiezione scegliamo una variabile aleatoria Y esponenzialmente distribuita di valore medio unitario, ossia di densità di probabilità:

$$g(y) = \begin{cases} e^{-y}, & y > 0. \\ 0, & \text{altrimenti.} \end{cases}$$

Per applicare il metodo di reiezione occorre determinare una costante c che soddisfi la condizione (9.4), ossia tale che per ogni $y > 0$ risulti $f(y)/g(y) \le c$. Se $y > 0$ si ha

$$\frac{f(y)}{g(y)} = \frac{2}{\sqrt{2\,\pi}} \exp\Big\{-\frac{y^2}{2} + y\Big\} = \sqrt{\frac{2e}{\pi}} \exp\Big\{-\frac{(y-1)^2}{2}\Big\}.$$

Poiché

$$\frac{f(y)}{g(y)} \le \sqrt{\frac{2e}{\pi}},$$

affinché la (9.4) sia soddisfatta per ogni $y > 0$ basterà scegliere nel metodo di reiezione

$$c = \sqrt{\frac{2e}{\pi}}\,.$$

Pertanto, se $y > 0$ si ha

$$\frac{f(y)}{c\,g(y)} = \exp\Big\{-\frac{(y-1)^2}{2}\Big\}\,.$$

L'algoritmo per simulare la variabile aleatoria $X = |Z|$, con Z avente distribuzione normale standard, mediante il metodo di reiezione è quindi il seguente:

Algoritmo
STEP 1: Generare le variabili aleatorie indipendenti Y esponenzialmente distribuita di valore medio unitario e U uniformemente distribuita in $(0, 1)$;

STEP 2: Se risulta

$$U < \exp\Big\{-\frac{(Y-1)^2}{2}\Big\}$$

porre $X = Y$, altrimenti ritornare al passo *1.*

Il codice seguente permette di applicare il metodo di reiezione per la simulazione del valore assoluto di una variabile aleatoria normale standard, ossia di $X = |Z|$.

```
> set.seed(3)
> u<-runif(3000) # genera un vettore di  numeri uniformi
> set.seed(5)
> y<--log(1-runif(3000)) # genera un vettore di  numeri esponenziali
> w<-exp(-(y-1)^2/2)
> z<-which(u<w)
> x<-y[z]
> length(x)
[1] 2270
>
> mean(x)
[1] 0.7875984
> var(x)
[1] 0.3436235
```

Si genera un vettore u di lunghezza 3000 con distribuzione uniforme nell'intervallo $(0, 1)$ e un vettore y con distribuzione esponenziale di valore medio unitario. Se si eliminano le istruzioni set.seed(3) e set.seed(5) i semi vengono scelti automaticamente da R e quindi una successiva esecuzione del codice conduce a sequenze differenti. Si calcola il vettore w contenente i valori $\exp\{-(y-1)^2/2\}$ e si esegue il test di reiezione. Si determinano con la funzione which(u < w) gli indici dei vettori che soddisfano la condizione u < w e si associano al vettore z. Il vettore z contiene gli indici di y che corrispondono ai valori casuali accettati. Si costruisce così il vettore x che conterrà i valori simulati della variabile aleatoria $X = |Z|$. La lunghezza del vettore x ottenuto con il metodo di reiezione è 2270 e si può calcolare la media campionaria e la varianza campionaria utilizzando rispettivamente mean(x) e var(x). Si nota che la media campionaria e la varianza campionaria ottenuti sono vicini alla media $E(X) = 0.797885$ e alla varianza $\mathrm{Var}(X) = 0.36338$. □

Problema 9.5 Simulazione di una variabile aleatoria normale standard

La simulazione della variabile aleatoria $X = |Z|$ discussa nel Problema 9.4 permette di simulare una variabile aleatoria Z di densità normale standard (9.9). Essendo la densità normale standard simmetrica intorno allo zero, si può porre $Z = X$ oppure $Z = -X$ con uguale probabilità, ossia

$$Z = \begin{cases} -X, & 0 \le V < 1/2 \\ X, & 1/2 \le V < 1, \end{cases}$$

essendo V una variabile aleatoria uniformemente distribuita in $(0, 1)$. Infatti risulta

$$P(Z < z) = P\Big(Z < z, 0 \le V < \frac{1}{2}\Big) + P\Big(Z < z, \frac{1}{2} \le V < 1\Big)$$

$$= \frac{1}{2} P\Big(Z < z \mid 0 \le V < \frac{1}{2}\Big) + \frac{1}{2} P\Big(Z < z \mid \frac{1}{2} \le V < 1\Big)$$
$$= \frac{1}{2} P(-X < z) + \frac{1}{2} P(X < z) = \frac{1}{2} P(X > -z) + \frac{1}{2} P(X < z),$$

da cui derivando rispetto a z si ottiene la densità normale standard (9.9).

L'algoritmo per simulare la variabile aleatoria Z con distribuzione normale standard è quindi il seguente:

Algoritmo
STEP 1: Generare le variabili aleatorie indipendenti Y esponenzialmente distribuita di valore medio unitario e U uniformemente distribuita in $(0, 1)$;
STEP 2: Se risulta

$$U < \exp\Big\{-\frac{(Y-1)^2}{2}\Big\}$$

porre $X = Y$, altrimenti ritornare al passo *1.*
STEP 3: Generare una variabile aleatoria V uniformemente distribuita in $(0, 1)$ e porre

$$Z = \begin{cases} -X, & 0 \le V < 1/2 \\ X, & 1/2 \le V < 1. \end{cases}$$

Il codice seguente permette di applicare il metodo di reiezione per la simulazione una variabile aleatoria normale standard. Abbiamo aggiunto al codice dell'algoritmo del Problema 9.4 le istruzioni per generare Z.

```
> set.seed(3)
> u<-runif(3000)
> set.seed(5)
> y<--log(1-runif(3000))
> w<-exp(-(y-1)^2/2)
> z<-which(u<w)
> x<-y[z]
> length(x)
[1] 2270
>
> set.seed(7)
> v<-runif (length (x))
> length(v)
[1] 2270
> v[ which (v <0.5)] <- -1
> v[ which (v >=0.5)] <-1
> z<-x*v
> length(z)
[1] 2270
> mean(z)
[1] -0.04189519
> var (z)
[1] 0.9624521
>
> plot(density(z),main="Densita' normale standard simulata",xlab="x")
> curve(dnorm(x,mean=0,sd=1),add=TRUE,col="red")
```

Si nota che abbiamo generato un vettore v uniformemente distribuito nell'intervallo $(0, 1)$ della stessa lunghezza del vettore x. Successivamente in corrispon-

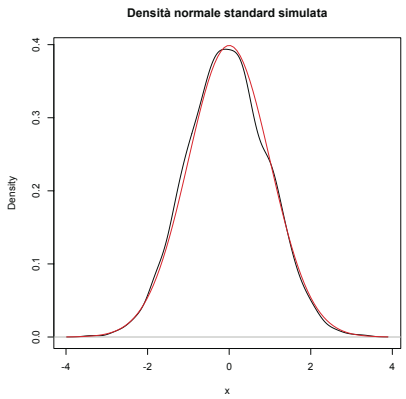

Figura 9.3: Densità normale standard teorica (colore rosso) e simulata (colore nero).

denza degli indici per i quali l'elemento del vettore è minore di $1/2$ assegnamo il valore -1, mentre in corrispondenza degli indici per i quali l'elemento del vettore è maggiore o uguale di $1/2$ assegniamo il valore 1. Il vettore v così creato avrà lunghezza 2270 con elementi -1 e 1. Infine, il vettore z, descrivente la sequenza di numeri distribuiti secondo una normale standard, è ottenuto effettuando il prodotto dei due vettori x e v di uguale lunghezza elemento per elemento. Si nota che la media campionaria e la varianza campionaria sono vicini alla media $E(Z) = 0$ e alla varianza $\mathrm{Var}(Z) = 1$ della variabile aleatoria normale standard.

La funzione plot() permette di effettuare il grafico della densità simulata ottenuta dal vettore z mediante la funzione density(z). La funzione curve() permette di graficare una funzione di qualsiasi tipo; nel nostro caso si è rappresentata la densità normale standard teorica mediante il comando dnorm(x, mean = 0, sd = 1), dove mean e sd denotano la media (zero) e la deviazione standard (unitaria) di tale densità normale standard. Il parametro add = TRUE serve per rappresentare nella stessa finestra grafica le due densità teorica e simulata, mentre col = "red") serve per specificare il colore rosso. In Fig. 9.3 è visualizzata la funzione densità normale standard (curva rossa) e quella simulata (curva nera). □

Problema 9.6 Simulazione di una variabile aleatoria normale
Sia T una variabile aleatoria distribuita normalmente con valore medio μ e varianza σ^2, caratterizzata da densità di probabilità

$$f_T(t) = \frac{1}{\sigma\sqrt{2\,\pi}}\,\exp\Big\{-\frac{(t-\mu)^2}{2\,\sigma^2}\Big\} \qquad (t \in \mathbb{R}).$$

La densità normale ha una curva a campana, simmetrica intorno al valore medio $E(T) = \mu$, con varianza $\mathrm{Var}(T) = \sigma^2$. Se si considera la trasformazione

$$Z = \frac{T-\mu}{\sigma}\,,$$

si ottiene una variabile aleatoria Z con distribuzione normale standard (di valore medio nullo e varianza unitaria).

Ricordando i Problemi 9.4 e 9.5, l'algoritmo per simulare una variabile aleatoria T distribuita normalmente con valore medio $E(T) = \mu$ e varianza $\mathrm{Var}(T) = \sigma^2$, è il seguente:

Algoritmo
STEP 1: Generare le variabili aleatorie indipendenti Y esponenzialmente distribuita di valore medio unitario e U uniformemente distribuita in $(0, 1)$;
STEP 2: Se risulta

$$U < \exp\Big\{-\frac{(Y-1)^2}{2}\Big\}$$

porre $X = Y$, altrimenti ritornare al passo *1.*
STEP 3: Generare una variabile aleatoria V uniformemente distribuita in $(0, 1)$ e porre

$$Z = \begin{cases} -X, & 0 \le V < 1/2 \\ X, & 1/2 \le V < 1. \end{cases}$$

STEP 4: Porre

$$T = \mu + \sigma\, Z.$$

Le seguenti linee di codice permettono di ottenere una sequenza di numeri distribuiti normalmente con valore medio $\mu = 2$ e deviazione standard $\sigma = 0.1$. Abbiamo aggiunto al codice del Problema 9.5 il codice per simulare una variabile normale di valore medio $\mu = 2$ e deviazione standard $\sigma = 0.1$.

```
> set.seed(3)
> u<-runif(3000)
> set.seed(5)
> y<--log(1-runif(3000))
> w<-exp(-(y-1)^2/2)
> z<-which(u<w)
> x<-y[z]
> length(x)
[1] 2270
>
> set.seed(7)
> v<-runif (length (x))
> length(v)
[1] 2270
> v[ which (v <0.5)] <- -1
> v[ which (v >=0.5)] <-1
> z<-x*v
> length(z)
[1] 2270
>
> t<-2+0.1*z
> length(t)
[1] 2270
> mean(t)
[1] 1.99581
> var(t)
[1] 0.009624521
>
```

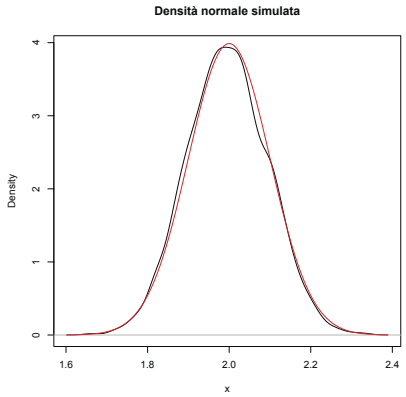

Figura 9.4: Densità normale teorica (colore rosso) e simulata (colore nero) per $\mu = 2$ e per $\sigma = 0.1$.

```
>plot(density(t),main="Densita' normale simulata",xlab="x")
> curve(dnorm(x,mean=2,sd=0.1),add=TRUE,col="red")
```

Si nota che la media campionaria e la varianza campionaria sono prossimi ai valori teorici della media $E(T) = \mu = 2$ e della deviazione standard $\sqrt{\text{Var}(T)} = \sigma = 0.1$ della variabile aleatoria normale considerata.

In Fig. 9.4 è rappresentata la funzione densità normale (curva rossa) e quella simulata (curva nera) per $\mu = 2$ e per $\sigma = 0.1$. □

9.4 Particolari variabili aleatorie continue: normale e di Erlang

Per alcuni tipi di variabili aleatorie continue esistono idonei algoritmi che risultano *più efficienti computazionalmente* di quelli ottenibili con il metodo di inversione della funzione di distribuzione o con il metodo di reiezione.

Vogliamo ora analizzare metodi specifici per simulare alcuni tipi di variabili aleatorie continue, ossia *normale* e *di Erlang*.

Problema 9.7 Simulazione di una variabile aleatoria normale standard
Esistono vari metodi per generare una variabile aleatoria con distribuzione normale standard oltre a quello discusso nel Problema 9.5. Uno di questi si basa sul teorema centrale di convergenza. Tale teorema afferma che se $X_1, X_2, \ldots$ è una successione di variabili aleatorie indipendenti e identicamente distribuite, con valore medio $E(X_i) = \mu$ finito e varianza $\text{Var}(X_i) = \sigma^2$ finita, allora

$$\lim_{N \to +\infty} P\left(\frac{X_1 + X_2 + \ldots + X_N - N\,\mu}{\sigma\,\sqrt{N}} < x \right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\,\pi}} \exp\left\{ -\frac{z^2}{2} \right\} dz. \tag{9.10}$$

La (9.10) mostra che la funzione di distribuzione della variabile aleatoria

$$\frac{X_1 + X_2 + \ldots + X_N - N\,\mu}{\sigma\sqrt{N}}$$

è approssimabile, per N sufficientemente grande, con la funzione di distribuzione di una variabile aleatoria normale standard.

Per simulare la variabile normale standard possiamo scegliere le variabili $X_1, X_2, \ldots$ uniformi nell'intervallo $(0,1)$. Quindi, se si considera una sequenza di variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0,1)$, ognuna caratterizzata da valore medio $E(U_i) = 1/2$ e varianza $\mathrm{Var}(U_i) = 1/12$, dal teorema centrale di convergenza segue che per N abbastanza grande la funzione di distribuzione della variabile aleatoria

$$Z = \frac{U_1 + U_2 + \ldots + U_N - N/2}{\sqrt{N/12}} \tag{9.11}$$

è approssimativamente quella di una variabile aleatoria normale standard. Un metodo per simulare una variabile aleatoria con distribuzione normale standard è quindi il seguente:

Algoritmo
STEP 1: Generare N variabili aleatorie $U_1, U_2, \ldots, U_N$ indipendenti e uniformemente distribuite nell'intervallo $(0,1)$;
STEP 2: Valutare Z tramite la relazione

$$Z = \frac{U_1 + U_2 + \ldots + U_N - N/2}{\sqrt{N/12}}.$$

Solitamente si sceglie $N \geq 12$ poiché sperimentalmente si è visto che già 12 variabili aleatorie con distribuzione uniforme sono sufficienti per simulare una variabile aleatoria con distribuzione normale standard. In particolare, scegliendo $N = 12$ l'algoritmo di simulazione della variabile aleatoria Z con distribuzione normale standard diventa il seguente.

Algoritmo
STEP 1: Generare 12 variabili aleatorie $U_1, U_2, \ldots, U_{12}$ indipendenti e uniformemente distribuite nell'intervallo $(0,1)$;
STEP 2: Valutare Z tramite la relazione

$$Z = U_1 + U_2 + \ldots + U_{12} - 6.$$

La variabile aleatoria $U_1 + U_2 + \ldots + U_{12} - 6$ assume valori nell'intervallo $(-6, 6)$. Questo intervallo è sufficiente per la generazione di una variabile aleatoria con distribuzione normale standard; infatti, ricordando che

$$P(-3 < Z < 3) = P(Z < 3) - P(Z < -3),$$

utilizzando R si ottiene

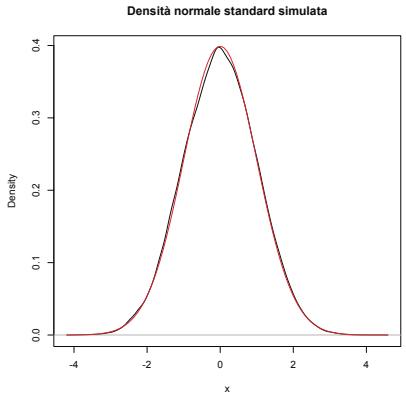

Figura 9.5: Densità normale teorica (colore rosso) e simulata (colore nero) per $\mu = 0$ e per $\sigma = 1$.

```
> pnorm(3,mean=0,sd=1)-pnorm(-3,mean=0,sd=1)
[1] 0.9973002
```

Quindi, $P(-3 < Z < 3) \simeq 0.9973$, ossia la probabilità che la variabile aleatoria Z assuma valori nell'intervallo $(-3, 3)$ è prossima all'unità.

Generiamo una sequenza di 50000 numeri distribuiti secondo una normale standard attraverso il precedente algoritmo facendo scegliere i semi delle 12 sequenze uniformi automaticamente da R.

```
> u1<-runif(50000)
> u2<-runif(50000)
> u3<-runif(50000)
> u4<-runif(50000)
> u5<-runif(50000)
> u6<-runif(50000)
> u7<-runif(50000)
> u8<-runif(50000)
> u9<-runif(50000)
> u10<-runif(50000)
> u11<-runif(50000)
> u12<-runif(50000)
> z<-u1+u2+u3+u4+u5+u6+u7+u8+u9+u10+u11+u12-6
> mean(z)
[1] -0.0003385987
> var(z)
[1] 1.006875
> plot(density(z),main="Densita' normale standard simulata",xlab="x")
> curve(dnorm(x,mean=0,sd=1),add=TRUE,col="red")
```

In Fig. 9.5 è visualizzata la funzione densità normale standard (curva rossa), con $\mu = 0$ e per $\sigma = 1$, e quella simulata (curva nera) ottenuta sommando 12 uniformi in $(0, 1)$.

Un altro metodo per simulare una variabile aleatoria con distribuzione normale standard è il *metodo di Box–Muller* che permette di ottenere una coppia di valori della variabile aleatoria normale standard sfruttando ogni volta la simulazione di due variabili aleatorie uniformi in $(0, 1)$. Il metodo di Box-Muller per simulare una variabile aleatoria con distribuzione normale standard può essere sintetizzato nel seguente algoritmo:

Algoritmo
STEP 1: Generare due variabili aleatorie U_1 e U_2 indipendenti e uniformemente distribuite nell'intervallo $(0, 1)$;
STEP 2: Porre

$$X = \sqrt{-2\,\log(1-U_1)}\,\cos(2\,\pi\,U_2)$$
$$Y = \sqrt{-2\,\log(1-U_1)}\,\sin(2\,\pi\,U_2).$$

Come mostrato nel Problema 9.3 la variabile aleatoria $\sqrt{-2\,\log(1-U_1)}$ utilizzata nell'algoritmo di Box-Muller è caratterizzata da una funzione di distribuzione di Rayleigh. □

In R esiste una funzione predefinita che simula la variabile aleatoria normale

```
rnorm(n, mean=mu, sd=sigma)
```

dove `n` è lunghezza della sequenza da generare, μ è il valore medio e σ la deviazione standard della densità normale.

Desideriamo concludere questo capitolo con la simulazione di variabili aleatorie di Erlang di ordine k e iperesponenziali di ordine k.

Problema 9.8 Simulazione di una variabile aleatoria di Erlang
Siano $X_1, X_2, \ldots, X_k$ variabili aleatorie indipendenti e esponenzialmente distribuite con valore medio $1/(k\varrho)$, con $\varrho > 0$. La variabile aleatoria

$$Y = X_1 + X_2 + \ldots + X_k$$

è caratterizzata da una densità di Erlang di ordine k con valore medio $E(Y) = 1/\varrho$ e varianza $\mathrm{Var}(Y) = 1/(k\varrho^2)$. La densità di Erlang di ordine k della variabile aleatoria Y è:

$$f_Y(x) = \begin{cases} \dfrac{(k\varrho)^k}{(k-1)!}\, e^{-k\varrho\,x}\; x^{k-1}, & x > 0 \\ 0, & x \le 0, \end{cases} \tag{9.12}$$

Desideriamo simulare la variabile aleatoria Y. Notiamo che metodo di inversione della funzione di distribuzione non è utilizzabile poiché la funzione di distribuzione della variabile aleatoria Y è difficile da invertire. Inoltre, il metodo di reiezione, anche se applicabile (scegliendo come variabile aleatoria di confronto che si riesce a simulare una variabile esponenziale di valore medio coincidente con quello della variabile aleatoria di Erlang Y) è di difficile utilizzazione pratica. Occorre quindi considerare il metodo descritto nel seguente algoritmo.

Algoritmo

STEP 1: Generare k variabili aleatorie indipendenti $X_1, X_2, \ldots, X_k$ esponenzialmente distribuite di valore medio $1/(k\varrho)$;
STEP 2: Calcolare

$$Y = \sum_{i=1}^{k} X_i.$$

Ricordando il Problema 9.2, nel precedente algoritmo si può porre

$$X_i = -\frac{1}{k\,\varrho}\,\log(1-U_i), \qquad i = 1, 2, \ldots, k$$

con $U_1, U_2, \ldots, U_k$ uniformemente distribuite in $(0, 1)$.

L'algoritmo per simulare una variabile aleatoria di Erlang di ordine k può quindi anche essere così formulato:

Algoritmo
STEP 1: Generare k variabili aleatorie indipendenti $U_1, U_2, \ldots, U_k$ uniformemente distribuite in $(0, 1)$;
STEP 2: Porre

$$X_i = -\frac{1}{k\,\varrho}\,\log(1-U_i), \qquad i = 1, 2, \ldots, k.$$

STEP 3: Calcolare

$$Y = \sum_{i=1}^{k} X_i = -\frac{1}{k\,\varrho}\sum_{i=1}^{k}\log(1-U_i) = -\frac{1}{k\,\varrho}\,\log\Big[\prod_{i=1}^{k}(1-U_i)\Big].$$

□

La variabile aleatoria di Erlang di ordine k gioca un ruolo fondamentale nelle file di attesa nella descrizione di alcuni meccanismi di arrivo e di servizio denotati con E_k. Nel *meccanismo degli arrivi* si suppone che i tempi di interarrivo degli utenti che accedono al sistema siano indipendenti ed esponenzialmente distribuiti con valore medio $1/(k\lambda)$ e che esiste un distributore che assegna ordinatamente a ciascuno delle k file di attesa gli arrivi. In una generica delle k file di attesa tra un arrivo ed il successivo intercorrono k intervalli di interarrivo esponenziali indipendenti ed identicamente distribuiti.

Denotando con T un tempo di interarrivo in una generica fila si ha $T = T_1 + T_2 + \ldots + T_k$ e il valore medio e la varianza sono:

$$E(T) = E(T_1) + E(T_2) + \ldots + E(T_k) = k\,\frac{1}{k\lambda} = \frac{1}{\lambda},$$
$$\mathrm{Var}(T) = \mathrm{Var}(T_1) + \mathrm{Var}(T_2) + \ldots + \mathrm{Var}(T_k) = k\,\frac{1}{(k\lambda)^2} = \frac{1}{k\lambda^2}.$$

Nel *meccanismo di servizio* abbiamo supposto che il centro di servizio è organizzato in k identiche ed indipendenti fasi poste in sequenza. Il tempo di servizio di una generica fase j $(j = 1, 2, ..., k)$ è descritto da una variabile aleatoria esponenziale di valore medio $1/(k\mu)$.

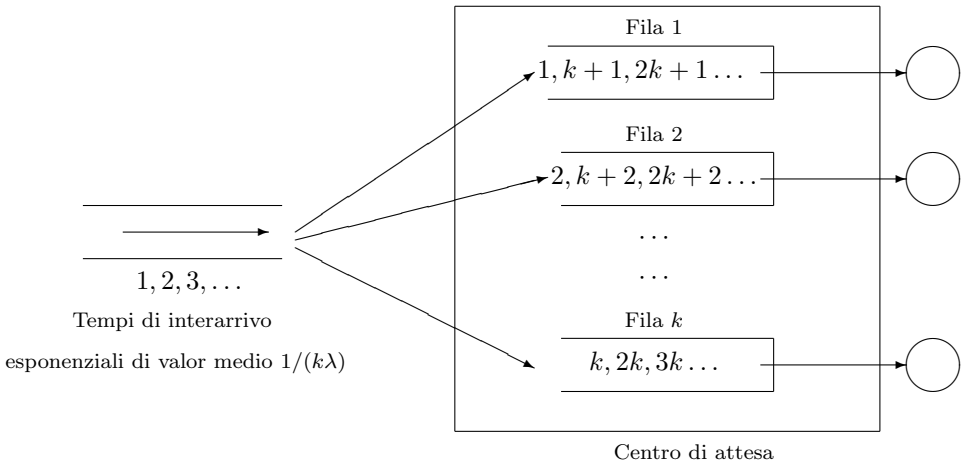

Figura 9.6: Tempi di interarrivo di tipo E_k in ognuna delle k file di attesa.

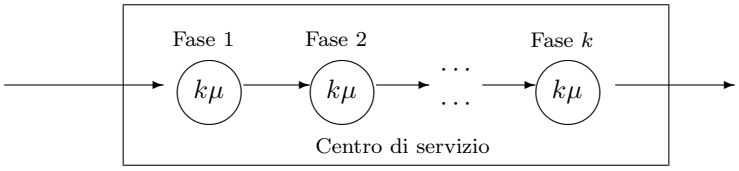

Figura 9.7: Tempi di servizio di tipo E_k nel caso in cui il centro di servizio prevede k successive fasi esponenziali di valore medio $1/(k\mu)$.

Denotando con S il tempo complessivo di servizio si ha che $S = S_1 + S_2 + \ldots + S_k$ e il valore medio e la varianza sono:

$$E(S) = E(S_1) + E(S_2) + \ldots + E(S_k) = k\,\frac{1}{k\mu} = \frac{1}{\mu},$$
$$\mathrm{Var}(S) = \mathrm{Var}(S_1) + \mathrm{Var}(S_2) + \ldots + \mathrm{Var}(S_k) = k\,\frac{1}{(k\mu)^2} = \frac{1}{k\mu^2}.$$

⋆ 9.3 Simulazione tempi di interarrivo e di servizio del sistema $M/E_2/1$ Consideriamo un sistema di servizio $M/E_2/1$, con tempi di interarrivo esponenzialmente distribuiti con valore medio $E(T) = 1/\lambda$; il servizio è organizzato in due fasi successive indipendenti, ognuna distribuita esponenzialmente con valore medio $1/(2\mu)$. Pertanto, i tempi di servizio sono distribuiti secondo Erlang di ordine 2 con valore medio $E(S) = 1/\mu$ e varianza $\mathrm{Var}(S) = 1/(2\mu^2)$.

Desideriamo simulare i tempi di interarrivo esponenzialmente distribuiti con valore medio $E(T) = 1/\lambda$ e i tempi di servizio distribuiti secondo Erlang di ordine 2 con valore medio $E(S) = 1/\mu$ e varianza $\mathrm{Var}(S) = 1/(2\mu^2)$. Ponendo

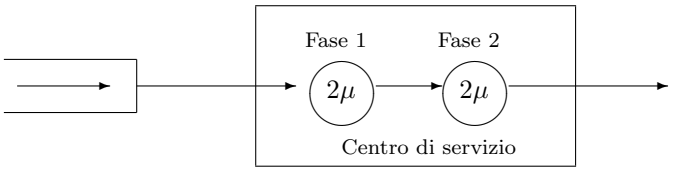

Figura 9.8: Tempi di servizio di tipo E_2 con centro di servizio che prevede 2 fasi successive esponenziali di valore medio $1/(2\mu)$.

$k = 2$ nella (9.12), la densità dei tempi di servizio è

$$b(t) = \begin{cases} (2\mu)^2\, e^{-2\mu t}\, t, & t > 0 \\ \\ 0, & t \leq 0. \end{cases}$$

Occorre generare tre sequenze indipendenti di numeri uniformi in $(0, 1)$, ossia $u_1, u_2, \ldots, v_1, v_2, \ldots$ e $z_1, z_2, \ldots$. I tempi di interarrivo t_i possono essere generati con

$$t_i = -\frac{1}{\lambda} \log(1 - z_i), \qquad i = 1, 2, \ldots$$

e i tempi di servizio con

$$s_i = -\frac{1}{2\mu}\, \log(1 - u_i) - \frac{1}{2\mu}\, \log(1 - v_i), \qquad i = 1, 2, \ldots$$

Il codice seguente permette di creare una sequenza di 5000 tempi di interarrivo tint distribuiti esponenzialmente con valore medio $1/\lambda = 1/2$ e una sequenza di 5000 tempi di servizio tserv distribuiti secondo Erlang con valore medio complessivo relativo alle due fasi $E(S) = 1/\mu = 1/4$. Si nota che

$$\varrho = \frac{\lambda}{\mu} = 0.5 < 1$$

e quindi il sistema di servizio $M/E_2/1$ non si congestiona. La densità di probabità teorica dei tempi di servizio è

$$b(t) = \begin{cases} 64\, t\, e^{-8\, t}, & t > 0 \\ \\ 0, & t \leq 0, \end{cases}$$

da cui

$$E(S) = \frac{1}{\mu} = 0.25, \qquad \mathrm{Var}(S) = \frac{1}{2\mu^2} = \frac{1}{32} = 0.03125.$$

```
> set.seed(7)
> z<-runif(5000)
> tint<--log(1-z)/2 # tempi di interarrivo
> mean(tint)
```

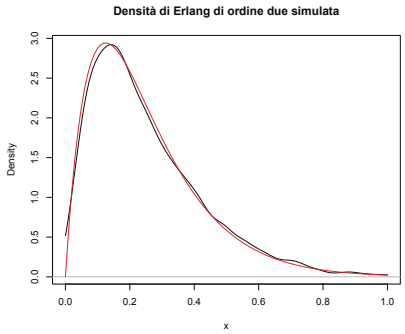

Figura 9.9: Densità di Erlang teorica (colore rosso) e simulata per $\mu = 4$.

```
[1] 0.4996218
> var(tint)
[1] 0.2580656
>
> set.seed(3)
> u<-runif(5000)
> set.seed(5)
> v<-runif(5000)
> tserv<-log(1-u)/8-log(1-v)/8 # tempi di servizio
> mean(tserv)
[1] 0.2529793
> var(tserv)
[1] 0.03192714
>
> plot(density(tserv,from=0,to=1),
+ main="Densita' di Erlang di ordine due simulata",xlab="x")
> curve(64*x*exp(-8*x),add=TRUE,col="red")
```

Si nota che la media campionaria e la varianza campionaria ottenuti mediante la simulazione sono prossimi ai rispettivi valori teorici. In Fig. 9.9 è visualizzata la funzione densità di Erlang teorica e quella simulata. ♢

Siamo infine interessati a simulare la variabile aleatoria iperesponenziale che è utilizzata in alcuni meccanismi di interarrivo e di servizio. Vedremo che per la sua simulazione gioca un ruolo fondamentale il metodo composto.

9.4.1 Metodo composto e variabile iperesponenziale

Assumiamo che la densità di probabilità di una variabile aleatoria continua X si possa porre nella forma

$$f_X(x) = \sum_{j=1}^{k} p_j \, g_j(x) \tag{9.13}$$

dove $p_1, p_2, \ldots, p_k$ soddisfano le condizioni

$$p_j \geq 0 \qquad (j = 1, 2, \ldots, k), \qquad \sum_{j=1}^{k} p_j = 1 \tag{9.14}$$

e dove $g_1(x), g_2(x), \ldots, g_k(x)$ sono le densità di probabilità delle variabili aleatorie continue $Z_1, Z_2, \ldots, Z_k$. Supponiamo di disporre di un metodo che permetta di simulare le k variabili aleatorie $Z_1, Z_2, \ldots, Z_k$.

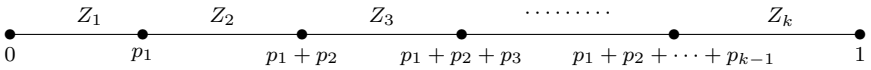

Figura 9.10: Suddivisione dell'intervallo (0,1) in k sottointervalli.

Per simulare la variabile aleatoria X utilizzando il metodo composto suddividiamo l'intervallo (0,1) in k sottointervalli di ampiezze $p_1, p_2, \ldots, p_k$ in maniera tale che $p_1 + p_2 + \ldots + p_k = 1$, come mostrato in Figura 9.10.

Il *metodo composto* simula la variabile aleatoria X con densità di probabilità (9.13) nel seguente modo:

Algoritmo
STEP 1: Generare una variabile aleatoria U uniformemente distribuita in $(0, 1)$;
STEP 2: Porre

$$X = \begin{cases} Z_1, & 0 \leq U < p_1 \\ Z_2, & p_1 \leq U < p_1 + p_2 \\ \cdots\cdots\cdots & \\ Z_j, & \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \\ \cdots\cdots\cdots & \\ Z_k, & \sum_{i=1}^{k-1} p_i \leq U < \sum_{i=1}^{k} p_i = 1. \end{cases}$$

Proposizione 9.3 *La variabile aleatoria continua X generata con il metodo composto ha densità di probabilità (9.13).*

Dimostrazione Osserviamo che

$$\begin{aligned} P(X < x) &= \sum_{j=1}^{k} P\Big(X < x, \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i\Big) \\ &= \sum_{j=1}^{k} P\Big(\sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i\Big)\, P\Big(X < x \,\Big|\, \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i\Big), \end{aligned}$$

dove si è posto $\sum_{i=1}^{j-1} p_i = 0$ per $j = 1$. Poiché risulta

$$P\Big(X < x \,\Big|\, \sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i\Big) = P(Z_j < x) \qquad (j = 1, 2, \ldots, k),$$

segue che

$$P(X < x) = \sum_{j=1}^{k} p_j \, P(Z_j < x).$$

Derivando ambo i membri rispetto a x, segue la (9.13). La variabile aleatoria X è quindi caratterizzata da densità di probabilità (9.13). □

Problema 9.9 Simulazione di una variabile aleatoria iperesponenziale
Il metodo composto gioca un ruolo importante nella simulazione di una variabile iperesponenziale. Sia X una variabile aleatoria caratterizzata da una densità iperesponenziale di ordine k, ossia

$$f_X(x) = \begin{cases} \displaystyle\sum_{j=1}^{k} p_j \, \mu_j \, e^{-\mu_j x}, & x > 0 \\ 0, & \text{altrimenti,} \end{cases} \tag{9.15}$$

dove $p_1, p_2, \ldots, p_k$ soddisfano le (9.14).

Osserviamo che la densità iperesponenziale si presenta come una *combinazione lineare di funzioni densità esponenziali*. Per simulare X si può quindi utilizzare il *metodo composto*. In questo caso la variabile aleatoria Z_j del metodo composto è caratterizzata da densità esponenziale di valore medio $1/\mu_j$ $(j = 1, 2, \ldots, k)$.

Per simulare la variabile aleatoria iperesponenziale X suddividiamo l'intervallo (0,1) in k sottointervalli di ampiezze $p_1, p_2, \ldots, p_k$ in maniera tale che $p_1 + p_2 + \ldots + p_k = 1$, come mostrato in Figura 9.11.

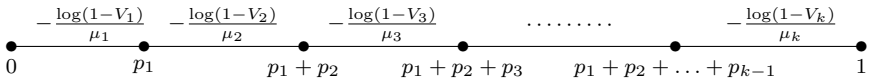

Figura 9.11: Suddivisione dell'intervallo (0,1) in sottointervalli.

Ricordando la simulazione di una variabile aleatoria esponenziale fornita nel Problema 9.2, il metodo composto simula la variabile aleatoria X iperesponenziale nel seguente modo:

Algoritmo
STEP 1: Generare $k + 1$ variabili aleatorie $V_1, V_2, \ldots, V_k, U$ indipendenti e uniformemente distribuite in $(0, 1)$;

STEP 2: Porre

$$X = \begin{cases} -\log(1-V_1)/\mu_1, & 0 \le U < p_1 \\ -\log(1-V_2)/\mu_2, & p_1 \le U < p_1 + p_2 \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ -\log(1-V_j)/\mu_j, & \displaystyle\sum_{i=1}^{j-1} p_i \le U < \sum_{i=1}^{j} p_i \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ -\log(1-V_k)/\mu_k, & \displaystyle\sum_{i=1}^{k-1} p_i \le U < \sum_{i=1}^{k} p_i. \end{cases}$$

La variabile aleatoria iperesponenziale si rivela utile per descrivere alcuni meccanismi di arrivo e di servizio.

Nel meccanismo di arrivo i potenziali utenti sono suddivisi in k diverse sorgenti a causa di differenti livelli di priorità loro assegnati oppure a causa di loro diverse provenienze geografiche. I tempi di interarrivo degli utenti che accedono alla sorgente j-esima sono descritti da variabili aleatorie indipendenti e distribuite esponenzialmente con valore medio $1/\lambda_j$ $(j = 1, 2, \ldots, k)$. Il centro di attesa è provvisto di un ingresso unico che provvede a scegliere con probabilità p_j la sorgente j-esima $(j = 1, 2, ..., k)$ ed ad avviare al centro di attesa la prima delle richieste di servizio relative alla sorgente selezionata.

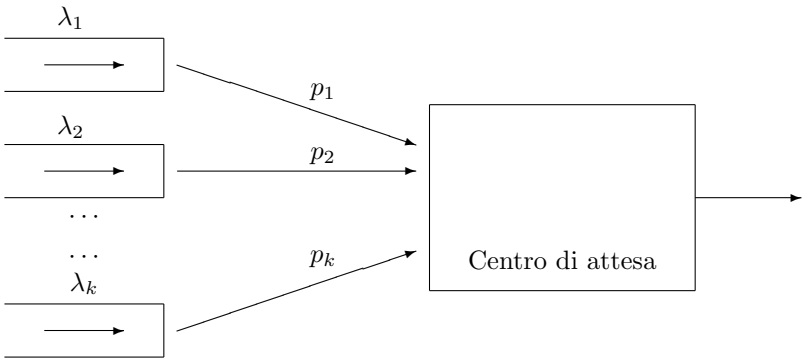

Figura 9.12: Tempi di interarrivo di tipo H_k nella fila di attesa.

La variabile aleatoria T, che descrive la lunghezza dell'intervallo di tempo tra due arrivi successivi al centro di attesa del sistema, ha densità ipereresponenziale di ordine k.

Nel meccanismo di servizio si considera un centro di servizio costituito da un unico servitore che provvede a fornire k tipi di differenti servizi. Si suppone che la probabilità che l'utente richieda un servizio di tipo j sia p_j per ogni

$j = 1, 2, \ldots, k$ e che la durata del servizio di tipo j sia esponenziale con valore medio $1/\mu_j$ $(j = 1, 2, \ldots, k)$.

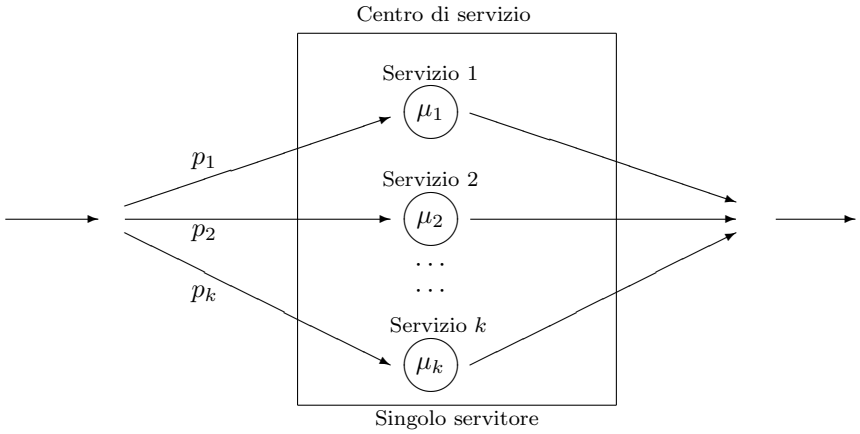

Figura 9.13: Tempi di servizio di tipo H_k per un singolo servitore che offre k differenti servizi.

La variabile aleatoria S, che descrive il tempo di servizio necessario per soddisfare un tipo qualsiasi di richiesta fatta dall'utente, è iperesponenziale di ordine k.

□

⋆ 9.4 Simulazione tempi di interarrivo e di servizio del sistema $U/H_2/1$
Consideriamo un sistema di servizio $U/H_2/1$ con i tempi di interarrivo distribuiti uniformementi in $(0, 2/\lambda)$ e i tempi di servizio distribuiti con densità iperesponenziale di ordine 2, ossia

$$b(t) = \begin{cases} p\,\mu_1 e^{-\mu_1 t} + (1-p)\,\mu_2 e^{-\mu_2 t}, & t > 0 \\ 0, & t \le 0, \end{cases}$$

da cui

$$E(S) = \frac{p}{\mu_1} + \frac{1-p}{\mu_2}, \qquad E(S^2) = \frac{2p}{\mu_1^2} + \frac{2(1-p)}{\mu_2^2}.$$

Esiste un unico servitore che fornisce due differenti servizi distribuiti esponenzialmente con valori medi $1/\mu_1$ e $1/\mu_2$; l'utente sceglie il primo servizio con probabilità p ed il secondo servizio con probabilità $1 - p$.

Occorre generare quattro sequenze indipendenti uniformemente distribuite in $(0, 1)$, ossia $u_1, u_2, \ldots$, $v_1, v_2, \ldots$ e $h_1, h_2, \ldots$, $k_1, k_2, \ldots$

Essendo i tempi di interarrivo distribuiti uniformementi in $(0, 2/\lambda)$, sono generati con

$$t_i = \frac{2}{\lambda}\, u_i, \qquad i = 1, 2, \ldots$$

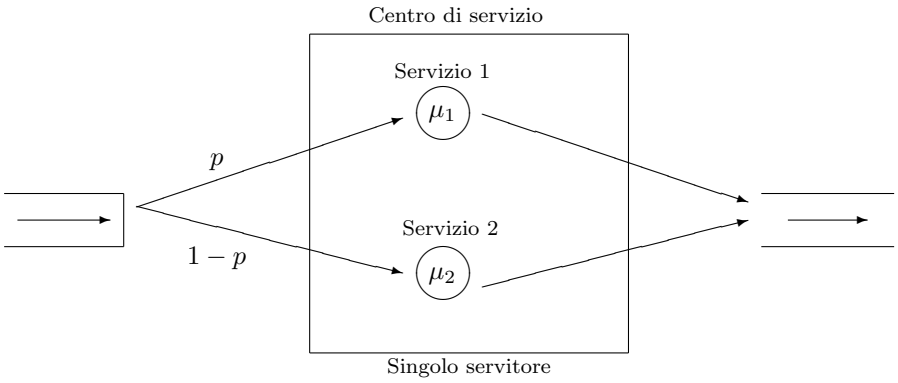

Figura 9.14: Tempi di servizio di tipo H_2 per un singolo servitore che offre 2 differenti servizi.

Essendo i tempi di servizio distribuiti con densità iperesponenziale di ordine 2, sono generati con

$$s_i = \begin{cases} -\dfrac{1}{\mu_1}\,\log(1-h_i), & 0 \le v_i < p \\ -\dfrac{1}{\mu_2}\,\log(1-k_i), & p \le v_i < 1, \end{cases} \qquad i = 1, 2, \ldots$$

Il codice seguente permette di creare una sequenza di 10000 tempi di interarrivo tint distribuiti uniformemente nell'intervallo $(0, 1)$ (con valore medio $E(T) = 1/\lambda = 1/2 = 0.5$) e una sequenza di 10000 tempi di servizio tserv distribuiti con densità iperesponenziale di ordine 2. Scegliamo $p = 1/4$, $\mu_1 = 2$ e $\mu_2 = 6$, in modo tale che il valore medio della variabile iperesponenziale sia

$$E(S) = \frac{p}{\mu_1} + \frac{1-p}{\mu_2} = \frac{1}{8} + \frac{3}{24} = \frac{1}{4}.$$

In questo caso

$$\varrho = \frac{E(S)}{E(T)} = \frac{1}{2} = 0.5 < 1$$

e quindi il sistema di servizio $U/H_2/1$ non si congestiona. Si nota che

$$E(S^2) = \frac{2p}{\mu_1^2} + \frac{2(1-p)}{\mu_2^2} = \frac{1}{6}, \qquad \mathrm{Var}(S) = E(S^2) - [E(S)]^2 = \frac{5}{48} = 0.1041667.$$

La densità iperesponenziale dei tempi di servizio è quindi

$$b(t) = \begin{cases} p\mu_1\, e^{-\mu_1 t} + (1-p)\mu_2\, e^{-\mu_2 t} = \dfrac{1}{2}\, e^{-2t} + \dfrac{9}{2}\, e^{-6t}, & t > 0 \\ 0, & t \le 0, \end{cases}$$

```
> set.seed(3)
> u<-runif(10000)
> tint<-u # tempi di interarrivo uniformi in (0,1)
> mean(tint)
[1] 0.4973273
> var(tint)
[1] 0.08394002
>
> set.seed(5)
> v<-runif(10000)
> set.seed(7)
> h<-runif(10000)
> set.seed(9)
> k<-runif(10000)
> y1<-which(v<1/4)
> y2<-which(v>=1/4)
> tserv<-numeric(10000)
> tserv[y1]<--log(1-h[y1])/2
> tserv[y2]<--log(1-k[y2])/6
> length(tserv)
[1] 10000
> mean(tserv)
[1] 0.2515293
> var(tserv)
[1] 0.1069151
>
> hist(tserv,freq=FALSE,breaks=100,xlim=c(0,3),
+ main="Istogramma della densita' iperesponenziale simulata")
> curve(exp(-2*x)/2+9*exp(-6*x)/2,add=TRUE,col="red",xlim=c(0,3) )
```

Si nota che la media campionaria e la varianza campionaria sono vicine alla media teorica $E(S) = 0.25$ e alla varianza teorica $\mathrm{Var}(S) = 0.1041667$.

In Figura 9.15 la densità iperesponenziale teorica è sovrapposta all'istogramma delle frequenze relative ottenuto tramite il campione simulato tserv. Il parametro freq = FALSE permette di ottenere l'istogramma delle frequenze relative e il parametro breaks = 100 suddivide i dati del vettore tserv in 100 classi di uguale ampiezza. L'area di ogni rettangolo dell'istogramma fornisce la frequenza relativa della classe e la somma delle aree di tutti i rettangoli è unitaria. ♢

La generazione dei tempi di interarrivo e di servizio ci permetterà di simulare ed analizzare il comportamento di vari sistemi di servizio nel transiente.

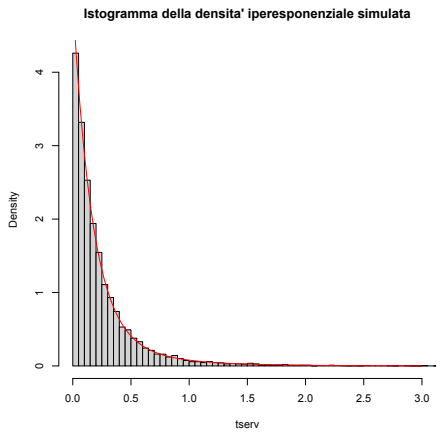

Figura 9.15: Densità iperesponenziale teorica (colore rosso) e istogramma ottenuto dai dati simulati sui tempi di servizio.

Capitolo 10

Simulazione di sistemi con singolo servitore

10.1 Introduzione

In questo capitolo desideriamo descrivere alcune procedure per simulare sistemi di servizio a capacità infinita con singolo servitore. In particolare, considereremo

- il sistema $M/M/1$ nel transiente stimando alcuni parametri prestazionali;
- il sistema $M/E_2/1$ nel transiente stimando alcuni parametri prestazionali.

Per simulare un sistema di servizio occorre utilizzare un simulatore dinamico, asincrono e ad eventi discreti. Gli eventi che causano cambiamenti di stato sono gli arrivi e le partenze degli utenti.

Consideriamo un sistema di servizio singolo servitore, singola fila di attesa, a capacità infinita, con disciplina di servizio FIFO in cui i tempi di interarrivo e i tempi di servizio hanno una distribuzione di probabilità di tipo generale (deterministica, uniforme, esponenziale, di Erlang, iperesponenziale, ...). In Figura 10.1 è mostrata una tipica realizzazione del numero di utenti presenti in tale sistema di servizio.

Supponiamo che i cambiamenti di stato siano istantanei ed avvengano in corrispondenza di eventi. Nel caso di un sistema di servizio con unico servitore si possono identificare due tipi di eventi:

- l'arrivo di un utente nel sistema che causa l'incremento del numero di utenti nel sistema;
- la fine del servizio di un utente con la conseguente uscita dal sistema e l'accesso al servizio di un altro utente (se ne esistono altri in fila di attesa); ciò causa un decremento del numero di utenti nel sistema.

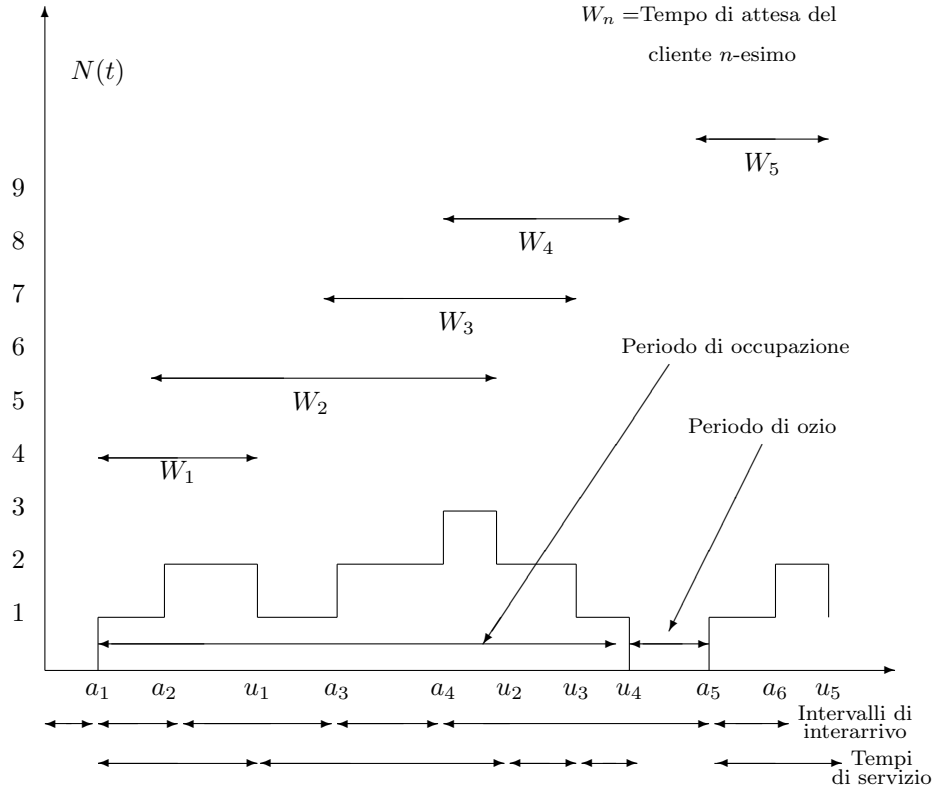

Figura 10.1: Una tipica realizzazione di un sistema di servizio.

Scopo della simulazione è quello di ricostruire l'evoluzione nel tempo del comportamento del sistema. Le componenti fondamentali della simulazione del sistema di servizio sono:

- la simulazione dei tempi di interarrivo con specificata distribuzione;

- la simulazione dei tempi di servizio con specificata distribuzione;

- il meccanismo con cui la simulazione termina.

La scelta della distribuzione teorica dei tempi di interarrivo e di servizio che meglio approssima il comportamento del sistema reale avviene calcolando le medie campionarie e le deviazioni standard campionarie ricavate da un'analisi storica del sistema reale ed effettuando un test di verifica di ipotesi sulla distribuzione teorica ipotizzata.

Nella procedura di simulazione che descriveremo desideriamo stimare soltanto alcuni parametri prestazionali: tempo medio di interarrivo, tempo medio di servizio, tempo medio di attesa, tempo medio di permanenza in fila di attesa. Non prestiamo attenzione al numero di utenti in fila di attesa e nel sistema.

Riferendoci alla Figura 10.1, utilizziamo le seguenti notazioni:

- $T(k)$ tempo di interarrivo k–esimo, ossia la lunghezza dell'intervallo di tempo intercorrente tra il $(k-1)$-esimo arrivo e il k-esimo arrivo al sistema di servizio ($k = 1, 2, \ldots$).

- $S(k)$ tempo di servizio del k-esimo utente ($k = 1, 2, \ldots$).

- $A(k)$ istante di arrivo del k-esimo utente nel sistema, ossia l'istante di tempo in cui effettivamente arriva il k-esimo utente. Sussiste la relazione ricorsiva:

$$\begin{aligned} &A(1) = T(1) \\ &A(k) = T(1) + T(2) + \ldots + T(k-1) + T(k) = A(k-1) + T(k) \quad (10.1) \\ &\qquad\qquad (k = 2, 3, \ldots). \end{aligned}$$

- $U(k)$ istante di uscita del k-esimo utente dal sistema, ossia l'istante di tempo in cui il k-esimo utente completa il servizio e lascia il sistema di servizio ($k = 1, 2, \ldots$). In un sistema di servizio con unico servitore e disciplina di servizio FIFO, l'istante di partenza del k-esimo utente può essere ricavato utilizzando la relazione ricorsiva:

$$\begin{aligned} &U(1) = A(1) + S(1) \\ &U(k) = \max\big\{A(k), U(k-1)\big\} + S(k) \qquad (k = 2, 3, \ldots). \qquad (10.2) \end{aligned}$$

Tale formula ricorsiva è detta "*legge di Lindley*". Si nota che se $k = 2, 3, \ldots$

$$U(k) = \begin{cases} A(k) + S(k), & A(k) \geq U(k-1) \\ U(k-1) + S(k), & A(k) < U(k-1) \end{cases}$$

Ciò significa che se l'utente $(k-1)$-esimo esce dal sistema prima o nello stesso istante dell'entrata dell'utente k-esimo, allora questo ultimo utente entra immediatamente in servizio e il suo tempo di uscita è uguale al suo tempo di arrivo più il suo tempo di servizio. Se, invece, il $(k-1)$-esimo utente esce dal sistema dopo l'entrata dell'utente k-esimo, allora questo ultimo utente non può entrare immediatamente in servizio (deve attendere nella fila di attesa) e quindi il suo tempo di uscita è uguale al tempo di uscita dell'utente $(k-1)$-esimo più il tempo di servizio dell'utente k-esimo.

- $W(k)$ tempo di attesa nel sistema di servizio del k-esimo utente ($k = 1, 2, \ldots$). Si deve avere

$$\begin{aligned} &W(1) = S(1) \\ &W(k) = U(k) - A(k) \qquad (k = 2, 3, \ldots). \end{aligned}$$

- $Q(k)$ tempo di permanenza nella fila di attesa del k-esimo utente ($k = 1, 2, \ldots$). Si deve avere

$$\begin{aligned} &Q(1) = 0 \\ &Q(k) = W(k) - S(k) \qquad (k = 2, 3, \ldots). \end{aligned}$$

- $O(k)$ tempo di ozio del centro di servizio fino all'istante di arrivo del k-esimo utente ($k = 1, 2, \ldots$). Si deve avere

$$O(k) = \begin{cases} 0, & U(k-1) \geq A(k) \\ A(k) - U(k-1), & U(k-1) < A(k). \end{cases} \tag{10.3}$$

La sequenza $T(1), T(2), \ldots$ dei tempi di interarrivo è ottenuta simulando una variabile aleatoria con la distribuzione di probabilità desiderata (uniforme, esponenziale, di Erlang, iperesponenziale). Analogamente, la sequenza $S(1), S(2), \ldots$ dei tempi di servizio è anch'essa ottenuta simulando una variabile aleatoria con la desiderata distribuzione di probabilità (uniforme, esponenziale, di Erlang, iperesponenziale). I tempi di interarrivo e di servizio possono anche essere ricavati osservando il funzionamento di un sistema di servizio reale, ottenendo così delle sequenze campionarie delle osservazioni effettuate.

10.2 Simulazione $M/M/1$ nel transiente

Consideriamo un sistema di servizio $M/M/1$ con tempi di interarrivo distribuiti esponenzialmente con valore medio $1/\lambda$, ossia il processo degli arrivi è descritto da un processo di Poisson di parametro λ, e tempi di servizio distribuiti esponenzialmente con valore medio $1/\mu$. Ci proponiamo ora di simulare i tempi di arrivo e di partenza degli utenti nel sistema $M/M/1$ e di calcolare alcuni parametri prestazionali del sistema tramite la simulazione, ossia i tempi medi di attesa nel sistema e i tempi di permanenza in fila di attesa. Fissati i parametri λ e μ del sistema $M/M/1$, nella simulazione procederemo nel seguente modo:

- generiamo i tempi di interarrivo esponenziali e calcoliamo i tempi di arrivo degli utenti nel sistema e il numero cumulativo di arrivi;
- generiamo i tempi di servizio esponenziali e calcoliamo i tempi di partenza degli utenti nel sistema;
- calcoliamo i tempi di permanenza in fila di attesa e di attesa nel sistema;
- ricaviamo alcuni parametri prestazionali del sistema.

10.2.1 Tempi di arrivo e numero cumulativo di arrivi

Desideriamo simulare i tempi di arrivo degli utenti t_A e in corrispondenza di tali tempi individuare il numero cumulativo di arrivi N_A fino a quell'istante di tempo. Denotando con $A(k)$ l'istante di arrivo del k-esimo utente nel sistema, sussiste la dalla (10.1) risulta che

$$A(1) = T(1), \qquad A(k) = A(k-1) + T(k) \qquad (k = 2, 3, \ldots),$$

dove $T(k)$ denota l'intervallo di tempo tra l'arrivo $k-1$ e l'arrivo k (tempo di interarrivo), distribuiti esponenzialmente nel sistema $M/M/1$. Per determinare

i tempi di arrivo degli utenti occorre simulare i tempi di interarrivo, distribuiti esponenzialmente con valore medio $1/\lambda$ utilizzando il metodo di inversione della funzione di distribuzione descritto nel precedente capitolo, ossia

$$t_k = -\frac{1}{\lambda}\log(1-u_k), \qquad k = 1, 2, \ldots, n,$$

dove $u_1, u_2, \ldots, u_n$ è una sequenza uniforme in $(0, 1)$.

Definiamo una funzione arrivi() che ha come parametri la lunghezza della sequenza da generare n, il parametro λ e un seme iniziale e deve restituire un data frame contenente nella prima colonna i tempi di arrivo degli utenti t_A e nella seconda colonna il numero cumulativo degli arrivi N_A fino a quell'istante di tempo.

```
> arrivi<-function(n,lambda,seme){
+ set.seed(seme)
+ u<-runif(n)
+ a<-numeric(n) # vettore dei tempi di arrivo
+ a[1]<--log(1-u[1])/lambda
+ df<-data.frame(t_A=1:n,N_A=1:n)
+ row.names(df)<-1:n
+ df[1,1]<-a[1]
+ df[1,2]<-1
+ for(i in 2:n){
+ a[i]<-a[i-1]-log(1-u[i])/lambda
+ df[i,1]<-a[i]
+ df[i,2]<-i}
+ return(df)
+ }
```

Il vettore a contiene i tempi di arrivo degli utenti. Nel data frame df sono presenti n righe e due colonne. La prima colonna contiene i tempi di arrivo t_A degli utenti e la seconda colonna il numero cumulativo di arrivi N_A fino a quell'istante di tempo.

Ad esempio, scegliendo $\lambda = 2$ e seme iniziale 3, generiamo $n = 10$ tempi di arrivo e il numero cumulativo di arrivi fino all'istante di tempo considerato:

```
> arrivi(10,2,3)
          t_A N_A
1  0.09198638   1
2  0.91585853   2
3  1.15887817   3
4  1.35742900   4
5  1.81820713   5
6  2.28187545   6
7  2.34843173   7
8  2.52292752   8
9  2.95384053   9
10 3.45229177  10
```

10.2.2 Tempi di arrivo e di partenza

Denotiamo con $A(k)$ e con $U(k)$ l'istante di arrivo e l'istante di partenza del k-esimo utente, rispettivamente. Essendo la disciplina di servizio FIFO, dalla

(10.2) ricaviamo:

$$U(1) = A(1) + S(1), \qquad U(k) = \max\big\{A(k), U(k-1)\big\} + S(k) \quad (k = 2, 3, \ldots),$$

dove $S(k)$ denota il tempo necessario per servire il k-esimo utente. Per determinare i tempi di partenza degli utenti occorre simulare i tempi di interarrivo, distribuiti esponenzialmente con valore medio $1/\lambda$, ed i tempi di servizio, distribuiti esponenzialmente con valore medio $1/\mu$, utilizzando il metodo di inversione della funzione di distribuzione, ossia

$$t_k = -\frac{1}{\lambda}\log(1-u_k), \qquad s_k = -\frac{1}{\mu}\log(1-v_k), \qquad k = 1, 2, \ldots, n,$$

dove $u_1, u_2, \ldots, u_n$ e $v_1, v_2, \ldots, v_n$ sono sequenze indipendenti uniformi in $(0, 1)$.

Definiamo una funzione arrivipartenze() che ha come parametri la lunghezza della sequenza da generare n, il parametro λ, il parametro μ e un seme iniziale e restituisce un data frame contenente nella prima colonna i tempi di arrivo t_A e nella seconda colonna i tempi di partenza t_P dei vari utenti del sistema di servizio $M/M/1$:

```
> arrivipartenze<-function(n,lambda,mu,seme){
+ set.seed(seme) # scelta di un  seme
+ u<-runif(n)
+ set.seed(seme+2) # scelta di un differente seme
+ v<-runif(n)
+ a<-numeric(n) # vettore dei tempi di arrivo
+ p<-numeric(n) # vettore dei tempi di partenza
+ a[1]<--log(1-u[1])/lambda
+ p[1]<--log(1-u[1])/lambda-log(1-v[1])/mu
+ df<-data.frame(t_A=1:n,t_P=1:n)
+ row.names(df)<-1:n
+ df[1,1]<-a[1]
+ df[1,2]<-p[1]
+ for(i in 2:n){
+ a[i]<-a[i-1]-log(1-u[i])/lambda
+ p[i]<-max(a[i],p[i-1])-log(1-v[i])/mu
+ df[i,1]<-a[i]
+ df[i,2]<-p[i]
+ }
+ return(df)
+ }
```

Il vettore a contiene i tempi di arrivo degli utenti e il vettore p contiene i tempi di partenza (uscita dal sistema dei vari utenti) del sistema $M/M/1$. Nel data frame df sono presenti n righe e due colonne. La prima colonna contiene i tempi di arrivo t_A degli utenti e la seconda colonna i tempi di uscita t_P degli utenti.

Ad esempio, scegliendo $\lambda = 2$, $\mu = 4$ e seme iniziale 3, generiamo i tempi di arrivo e tempi di partenza dei primi 10 utenti:

```
> arrivipartenze(10,2,4,3)
          t_A       t_P
1  0.09198638 0.1478393
2  0.91585853 1.2048277
```

```
3  1.15887817 1.8266825
4  1.35742900 1.9103408
5  1.81820713 1.9379760
6  2.28187545 2.5837514
7  2.34843173 2.7714243
8  2.52292752 3.1839049
9  2.95384053 3.9676542
10 3.45229177 3.9969150
```

Avendo scelto $\lambda = 2$ e $\mu = 4$, il sistema $M/M/1$ non si congestiona al crescere del tempo poiché $\varrho = \lambda/\mu = 1/2 = 0.5$.

10.2.3 Tempi di permanenza in fila e di attesa nel sistema

Possiamo ora calcolare i tempi di permanenza in coda e di attesa nel sistema dei vari utenti utilizzando le formule:

$$W(1) = S(1), \qquad W(k) = U(k) - A(k) \quad (k = 2, 3, \ldots).$$

e

$$Q(1) = 0, \qquad Q(k) = W(k) - S(k) \quad (k = 2, 3, \ldots).$$

Pertanto, definiamo una funzione attesa() che ha come parametri la lunghezza n della sequenza da generare, il parametro λ, il parametro μ e un seme iniziale e restituisce un data frame contenente i tempi di servizio S, i tempi di permanenza in fila di attesa Q e i tempi di attesa W dei vari utenti.

```
> attesa<-function(n,lambda,mu,seme){
+ set.seed(seme)
+ u<-runif(n)
+ set.seed(seme+2)
+ v<-runif(n)
+ a<-numeric(n) # vettore dei tempi di arrivo
+ p<-numeric(n) # vettore dei tempi di partenza
+ a[1]<--log(1-u[1])/lambda
+ p[1]<--log(1-u[1])/lambda-log(1-v[1])/mu
+ df<-data.frame(S=1:n,Q=1:n,W=1:n)
+ row.names(df)<-1:n
+ df[1,1]<--log(1-v[1])/mu
+ df[1,2]<-0
+ df[1,3]<--log(1-v[1])/mu
+ for(i in 2:n){
+ a[i]<-a[i-1]-log(1-u[i])/lambda
+ p[i]<-max(a[i],p[i-1])-log(1-v[i])/mu
+ df[i,1]<--log(1-v[i])/mu
+ df[i,2]<-p[i]-a[i]+log(1-v[i])/mu
+ df[i,3]<-p[i]-a[i]
+ }
+ return(df)
+ }
```

Nel data frame sono presenti n righe e tre colonne. La prima colonna contiene i tempi di servizio degli utenti, la seconda colonna i tempi di permanenza in fila di attesa e la terza colonna i tempi di attesa nel sistema.

Ad esempio, scegliendo $\lambda = 2$, $\mu = 4$ e seme iniziale 3, generiamo i tempi di servizio, tempi di permanenza in fila di attesa e i tempi di attesa nel sistema $M/M/1$.

```
> attesa(10,2,4,3)
            S             Q           W
1   0.05585291 0.000000e+00 0.05585291
2   0.28896921 0.000000e+00 0.28896921
3   0.62185477 4.594957e-02 0.66780434
4   0.08365829 4.692535e-01 0.55291181
5   0.02763518 9.213368e-02 0.11976886
6   0.30187597 1.110223e-16 0.30187597
7   0.18767288 2.353197e-01 0.42299257
8   0.41248062 2.484968e-01 0.66097740
9   0.78374930 2.300644e-01 1.01381369
10  0.02926074 5.153625e-01 0.54462320
```

Per un sistema $M/M/1$ in condizioni di equilibrio statistico ($\varrho = \lambda/\mu < 1$) abbiamo mostrato che

$$E(S) = \frac{1}{\mu}, \qquad E(W) = \frac{1}{\mu - \lambda}, \qquad E(Q) = E(W) - E(S) = \frac{\varrho}{\mu - \lambda}, \qquad .$$

Desideriamo avere informazioni sui parametri prestazionali del sistema $M/M/1$ attraverso la simulazione. Inoltre, se $\varrho < 1$ desideriamo confrontare i risultati ottenuti con la simulazione con quelli teorici in condizioni di equilibrio. Nel seguito analizziamo tre differenti casi.

Caso 1

Se scegliamo $\lambda = 2$ e $\mu = 4$, il sistema $M/M/1$ non si congestiona essendo $\varrho = \lambda/\mu = 0.5$. Il tempo medio di servizio è $E(S) = 0.25$, il tempo medio di permanenza in fila di attesa è $E(Q) = 0.25$ e $E(W) = 0.5$.

La seguente linea di codice

```
> colMeans(attesa(5000,2,4,3))
        S         Q         W
0.2545113 0.2779548 0.5324661
```

permette di determinare una stima del tempo medio di servizio, del tempo medio di permanenza in fila di attesa e del tempo medio di attesa nel sistema $M/M/1$ considerando un campione di 5000 utenti arrivati, avendo supposto che $\lambda = 2$, $\mu = 4$. Le medie campionarie sono effettuate sugli elementi delle colonne del data frame, ottenendo un vettore di lunghezza tre. Si nota che i risultati teorici sono in accordo con le stime ottenute mediante la simulazione.

Caso 2

Aumentiamo la frequenza di servizio degli utenti nel sistema $M/M/1$ scegliendo $\lambda = 2$ e $\mu = 6$. Il sistema $M/M/1$ non si congestiona, essendo $\varrho = \lambda/\mu = 0.333$. Il tempo medio di servizio è $E(S) = 0.166$, il tempo medio di permanenza in fila di attesa è $E(Q) = 0.083$ e $E(W) = 0.25$.

Consideriamo un campione di 5000 utenti arrivati nel sistema a cui prestare servizio scegliendo $\lambda = 2$ e $\mu = 6$. Stime del tempo medio di servizio, del tempo medio di permanenza in fila di attesa e del tempo medio di attesa nel sistema sono:

```
> colMeans(attesa(5000,2,6,3))
         S          Q          W
0.16967419 0.09157458 0.26124877
```

Si nota nuovamente che i risultati teorici sono in accordo con le stime ottenute mediante la simulazione. Essendo l'intensità di traffico $\varrho = 0.333$ inferiore a quella del Caso 1 ($\varrho = 0.5$), il tempo medio di attesa degli utenti nel sistema è minore rispetto al Caso 1.

Caso 3

Diminuiamo infine la frequenza del servizio degli utenti nel sistema $M/M/1$ scegliendo $\lambda = 2$ e $\mu = 1$. In questo ultimo caso, poiché $\lambda = 2$ e $\mu = 1$, il sistema $M/M/1$ è destinato a congestionarsi, essendo $\varrho = \lambda/\mu = 2$.

Consideriamo un campione di 5000 utenti arrivati a cui fornire il servizio. Stime del tempo medio di servizio, del tempo medio di permanenza in fila di attesa e del tempo medio di attesa nel sistema sono:

```
> colMeans(attesa(5000,2,1,3))
          S           Q           W
   1.018045 1294.133710 1295.151755
```

Si nota che il tempo medio di attesa e di permanenza in fila di attesa dei 5000 utenti arrivati cresce notevolmente rispetto ai casi precedentemente analizzati, sintomo del fatto che un unico servitore non riesce a soddisfare le richieste degli utenti. Ricordiamo che il numero necessario e sufficiente di servitori s identici da disporre in parallelo in un sistema $M/M/s$ per evitare la congestione deve soddisfare la disuguaglianza:

$$s - 1 \leq \frac{\lambda}{\mu} < s.$$

Nel caso esaminato in cui $\lambda = 2$ e $\mu = 1$ sono necessari 3 servitori in parallelo per evitare la congestione del sistema di servizio.

10.3 Simulazione del sistema di servizio $M/E_2/1$ nel transiente

Consideriamo un sistema di servizio $M/E_2/1$, con tempi di interarrivo esponenzialmente distribuiti con valore medio $E(T) = 1/\lambda$; il servizio è organizzato in due fasi successive indipendenti, ognuna distribuita esponenzialmente con valore medio $1/(2\mu)$. Pertanto, i tempi di servizio sono distribuiti secondo Erlang di ordine 2 con valore medio e varianza

$$E(S) = \frac{1}{\mu}, \qquad \mathrm{Var}(S) = \frac{1}{2\mu^2}$$

Il sistema $M/E_2/1$ è illustrato in Figura 10.2.

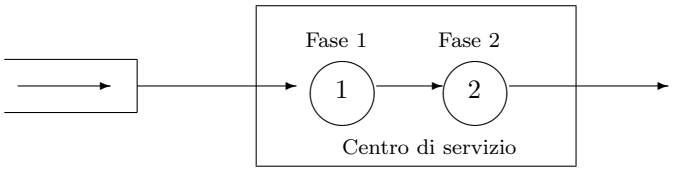

Figura 10.2: Sistema di servizio $M/E_2/1$.

Nel Capitolo 4, abbiamo mostrato che un sistema $M/G/1$ con $E(S) = 1/\mu$ raggiunge una situazione di equilibrio se $\varrho = \lambda/\mu < 1$ e sussiste la formula di Pollaczek–Khintchine (4.13):

$$E(N) = \varrho + \frac{\varrho^2\,[1 + C^2(S)]}{2\,(1-\varrho)} \qquad (\varrho < 1), \tag{10.4}$$

dove $C(S) = \sqrt{\mathrm{Var}(S)}/E(S)$ denota il coefficiente di variazione. Applicando la prima legge di Little risulta

$$E(W) = \frac{E(N)}{\lambda} = \frac{1}{\mu} + \frac{\varrho\,[1 + C^2(S)]}{2\,\mu\,(1-\varrho)}, \qquad E(Q) = \frac{\varrho\,[1 + C^2(S)]}{2\,\mu\,(1-\varrho)}.$$

In particolare, per il sistema $M/E_2/1$ il coefficiente di variazione è

$$C(S) = \frac{\sqrt{\mathrm{Var}(S)}}{E(S)} = \frac{1}{\sqrt{2}},$$

e quindi:

$$E(S) = \frac{1}{\mu}, \qquad E(W) = \frac{1}{\mu} + \frac{3\varrho}{4\mu\,(1-\varrho)}, \qquad E(Q) = \frac{3\varrho}{4\mu\,(1-\varrho)}. \tag{10.5}$$

Nel sistema $M/E_2/1$, essendo le singole fasi esponenziali di valore medio $1/(2\mu)$, i tempi di servizio s_i possono essere generati con

$$s_i = -\frac{1}{2\mu}\,\log(1 - v_{1i}) - \frac{1}{2\mu}\,\log(1 - v_{2i}), \qquad i = 1, 2, \ldots,$$

dove $v_{11}, v_{12}, \ldots, v_{1n}$ e $v_{21}, v_{22}, \ldots, v_{2n}$ sono due sequenze indipendenti e uniformi in $(0, 1)$.

Definiamo una funzione attesaME21() che ha come parametri la lunghezza n della sequenza da generare, il parametro λ, il parametro μ e un seme iniziale e restituisce un data frame contenente i tempi di servizio S, i tempi di permanenza in fila di attesa Q e i tempi di attesa W dei vari utenti.

```
> attesaME21<-function(n,lambda,mu,seme){
+ set.seed(seme)
+ u<-runif(n)
+ set.seed(seme+2)
```

```
+ v1<-runif(n)
+ set.seed(seme+3)
+ v2<-runif(n)
+ a<-numeric(n) # vettore dei tempi di arrivo
+ p<-numeric(n) # vettore dei tempi di partenza
+ a[1]<--log(1-u[1])/lambda
+ p[1]<--log(1-u[1])/lambda-log(1-v1[1])/(2*mu)-log(1-v2[1])/(2*mu)
+ df<-data.frame(S=1:n,Q=1:n,W=1:n)
+ row.names(df)<-1:n
+ df[1,1]<--log(1-v1[1])/(2*mu)-log(1-v2[1])/(2*mu)
+ df[1,2]<-0
+ df[1,3]<--log(1-v1[1])/(2*mu)-log(1-v2[1])/(2*mu)
+ for(i in 2:n){
+ a[i]<-a[i-1]-log(1-u[i])/lambda
+ p[i]<-max(a[i],p[i-1])-log(1-v1[i])/(2*mu)-log(1-v2[i])/(2*mu)
+ df[i,1]<--log(1-v1[i])/(2*mu)-log(1-v2[i])/(2*mu)
+ df[i,2]<-p[i]-a[i]+log(1-v1[i])/(2*mu)+log(1-v2[i])/(2*mu)
+ df[i,3]<-p[i]-a[i]
+  }
+ return(df)
+ }
```

La prima colonna del data frame contiene i tempi di servizio degli utenti, la seconda colonna i tempi di permanenza in fila di attesa e la terza colonna i tempi di attesa nel sistema.

Ad esempio, se si sceglie $\lambda = 2$, $\mu = 4$ e seme iniziale 3, generiamo i tempi di servizio, tempi di permanenza in fila di attesa e i tempi di attesa nel sistema $M/E_2/1$.

```
> attesaME21(10,2,4,3)
            S             Q         W
1   0.14443715 0.000000e+00 0.1444372
2   0.49134246 0.000000e+00 0.4913425
3   0.34930284 2.483228e-01 0.5976257
4   0.10160256 3.990748e-01 0.5006774
5   0.21976419 3.989927e-02 0.2596635
6   0.62845809 2.775558e-16 0.6284581
7   0.48990003 5.619018e-01 1.0518018
8   0.38606085 8.773061e-01 1.2633669
9   0.48095376 8.324539e-01 1.3134077
10  0.02296154 8.149564e-01 0.8379180
```

Si nota che i tempi di attesa sono inferiori a quelli del sistema $M/M/1$ con lo stesso tempo medio di interarrivo $E(T) = 1/\lambda$ e lo stesso tempo medio di servizio $E(S) = 1/\mu$.

Desideriamo avere informazioni sui parametri prestazionali del sistema di servizio $M/E_2/1$ attraverso la simulazione. Inoltre, se $\varrho < 1$ desideriamo confrontare i risultati con quelli teorici in condizioni di equilibrio. Nel seguito consideriamo tre differenti casi.

Caso 1

Se scegliamo $\lambda = 2$ e $\mu = 4$, il sistema $M/E_2/1$ non si congestiona essendo $\varrho = \lambda/\mu = 0.5$. Ricordando (10.5), il tempo medio di servizio è $E(S) = 0.25$, il tempo medio di permanenza in fila di attesa è $E(Q) = 0.1875$ e $E(W) = 0.4375$.

La seguente linea di codice

```
> colMeans(attesaME21(5000,2,4,3))
        S         Q         W
0.2543332 0.2122030 0.4665362
```

permette di determinare una stima del tempo medio di servizio, del tempo medio di permanenza in fila di attesa e del tempo medio di attesa nel sistema $M/E_2/1$ considerando un campione di 5000 utenti arrivati, avendo supposto che $\lambda = 2$ e $\mu = 4$. Si nota che i risultati teorici sono in accordo con le stime ottenute mediante la simulazione. Inoltre, il tempo medio di attesa è inferiore a quello del sistema $M/M/1$ con gli stessi parametri.

Caso 2

Aumentiamo la frequenza di servizio degli utenti nel sistema $M/E_2/1$ scegliendo $\lambda = 2$ e $\mu = 6$. Il sistema $M/E_2/1$ non si congestiona, essendo $\varrho = \lambda/\mu = 0.333$. Facendo uso della (10.5), il tempo medio di servizio è $E(S) = 0.166$, il tempo medio di permanenza in fila di attesa è $E(Q) = 0.0625$ e $E(W) = 0.2291667$.

Consideriamo un campione di 5000 utenti arrivati nel sistema a cui prestare servizio scegliendo $\lambda = 2$ e $\mu = 6$. Stime del tempo medio di servizio, del tempo medio di permanenza in fila di attesa e del tempo medio di attesa nel sistema sono:

```
> colMeans(attesaME21(5000,2,6,3))
         S          Q          W
0.16955549 0.06872395 0.23827944
```

Si nota nuovamente che i risultati teorici sono in accordo con le stime ottenute mediante la simulazione. Inoltre, il tempo medio di attesa è inferiore a quello del sistema $M/M/1$ con gli stessi parametri.

Caso 3

Diminuiamo infine la frequenza del servizio degli utenti nel sistema $M/E_2/1$ scegliendo $\lambda = 2$ e $\mu = 1$. In questo ultimo caso, poiché $\lambda = 2$ e $\mu = 1$, il sistema $M/E_2/1$ è destinato a congestionarsi, essendo $\varrho = \lambda/\mu = 2$.

Consideriamo un campione di 5000 utenti arrivati a cui fornire il servizio. Forniamo delle stime del tempo medio di servizio, del tempo medio di permanenza in fila di attesa e del tempo medio di attesa nel sistema $M/E_2/1$ sono:

```
> colMeans(attesaME21(5000,2,1,3))
          S           Q           W
   1.017333 1285.515335 1286.532668
```

Per entrambi i sistemi $M/M/1$ e $M/E_2/1$ si riscontra un aumento considerevole del tempo di attesa degli utenti, sintomo del fatto che entrambi i sistemi si congestionano al crescere del tempo quando $\varrho \geq 1$.

È possibile estendere la procedura di simulazione ad altri sistemi di servizio con differenti distribuzioni dei tempi di interarrivo o di servizio.

Nel prossimo capitolo desideriamo descrivere alcune procedure più generali per simulare sistemi di servizio includendo orari di chiusura al pubblico e servitori in parallelo con differenti velocità di servizio, prestando attenzione sia ai tempi di attesa che al numero di utenti nel sistema.

Capitolo 11

Procedure generali di simulazione in sistemi di servizio

11.1 Introduzione

In questo capitolo desideriamo descrivere alcune procedure generali per simulare sistemi di servizio a capacità infinita. In particolare, considereremo

- un sistema di servizio con singolo servitore nel transiente con orario di chiusura al pubblico;
- un sistema con due servitori aventi differenti velocità di servizio nel transiente.

11.2 Sistema di servizio con orario di chiusura e singolo servitore

Consideriamo un sistema di servizio singolo servitore, singola fila di attesa, a capacità infinita, con disciplina di servizio FIFO in cui i tempi di interarrivo e i tempi di servizio hanno una distribuzione di probabilità di tipo generale (deterministica, uniforme, esponenziale, di Erlang, iperesponenziale, ...) che preveda un orario di chiusura al pubblico.

Denotiamo con t_c l'istante dopo il quale non sia più consentito agli utenti in arrivo di accedere al sistema, sebbene dopo tale istante il servitore dovrà completare il servizio di tutti gli utenti presenti nel sistema al tempo t_c. Questa situazione si verifica frequentemente nei sistemi di servizio (banche, uffici postali,

centri ambulatoriali, ...) in cui è previsto un orario giornaliero di apertura e di chiusura al pubblico. Inoltre, sia t_f l'istante finale in cui tutti gli utenti che hanno avuto accesso al sistema sono stati serviti.

Per simulare il sistema di servizio occorre una variabile che rappresenti il tempo e una struttura che raccolga le informazioni relative agli eventi. Questa struttura è detta *calendario degli eventi*. Il calendario degli eventi è realizzato tramite una lista ordinata in base al campo che contiene l'istante di tempo in cui gli eventi si verificano. Tale lista deve contenere per ogni *tempo di osservazione* t (dove t è l'istante di tempo in cui si verifica un arrivo o una partenza), il *numero di utenti presenti nel sistema*, il *numero cumulativo di arrivi*, il *numero cumulativo di partenze*, il *tempo del prossimo arrivo dopo* t e il *tempo di completamento del servizio dell'utente attualmente in servizio*.

Occorrono inoltre delle variabili che consentano di trarre informazioni complessive sulla simulazione, ossia le variabili di input, di stato e di output.

Variabili di input

- i parametri necessari per generare i tempi di interarrivo degli utenti con assegnata distribuzione di probabilità;
- i parametri necessari per generare i tempi di servizio con assegnata distribuzione di probabilità;
- il valore che permette di definire la fine del processo di simulazione (numero massimo di utenti da servire, tempo massimo di simulazione, ...)

Le sequenze dei tempi di interarrivo e di servizio di probabilità permettono di ottenere:

- istanti di arrivo di ogni utente;
- istanti di partenza di ogni utente;
- istante t_f di fine della simulazione, ossia l'istante di tempo in cui l'ultimo utente lascia il sistema (nel sistema non sono più presenti utenti).

Variabili temporale

- t (tempo di osservazione, relativo agli istanti di tempo in cui si verificano gli eventi)

Variabili contatore

- N_A (numero cumulativo di arrivi fino al tempo t);
- N_U (numero cumulativo di partenze fino al tempo t).

Variabili di stato

- n (numero di utenti nel sistema al tempo di osservazione t);
- n_q (numero di utenti in fila di attesa al tempo di osservazione t).

Consideriamo inoltre i seguenti eventi futuri:

- t_A il tempo del nuovo arrivo (successivo al tempo di osservazione t);

- t_U tempo di completamento del servizio dell'utente attualmente in servizio. Se nessun utente è presente in servizio allora t_U verrà posto uguale a ∞.

Variabili di output
Occorre ricavare delle stime delle seguenti grandezze:

- l'istante di tempo t_f in cui la simulazione ha termine;

- numero medio di utenti in fila di attesa;

- numero medio di utenti nel sistema;

- fattore di utilizzazione del sistema;

- tempo medio di permanenza nella fila di attesa;

- tempo medio di attesa di un utente nel sistema.

Nella formulazione dell'algoritmo generale per la simulazione di un sistema di servizio occorre distinguere i seguenti casi:

- l'arrivo di un utente quando il sistema di servizio è vuoto (inizio immediato del servizio);

- l'arrivo di un utente quando il sistema di servizio non è vuoto (l'utente è inserito in fila di attesa);

- la fine del servizio con la fila di attesa vuota (non esistono utenti che immediatamente entrano in servizio);

- la fine del servizio con la fila di attesa non vuota (inizio immediato del servizio del successivo utente).

Algoritmo per simulare il sistema di servizio con unico servitore

Inizializzazione:
Al tempo $t = 0$ si inizializzano le seguenti variabili:

- $N_A \leftarrow 0$ (il numero cumulativo di arrivi fino al tempo $t = 0$ è nullo);

- $N_U \leftarrow 0$ (il numero cumulativo di partenze fino al tempo $t = 0$ è nullo);

- $n \leftarrow 0$ (il numero di utenti presenti nel sistema al tempo $t = 0$ è nullo);

- $n_q \leftarrow 0$ (il numero di utenti nella fila di attesa al tempo $t = 0$ è nullo).

Questi stati debbono essere aggiornati in modo appropriato al verificarsi di ogni evento muovendosi lungo l'asse temporale fino ad incontrare il prossimo evento.
Nella fase di inizializzazione, occorre

- *generare un tempo di interarrivo T con la distribuzione di probabilità desiderata* in modo tale da far partire il processo di simulazione generando il primo evento, ossia un arrivo, e innescando la procedura arrivo;

- porre il tempo del nuovo arrivo $t_A = T$ e il tempo di completamento del servizio $t_U = \infty$ (non essendoci utenti in servizio al tempo $t = 0$).

Procedura di arrivo ($t_A \leq t_U, t_A = \min(t_A, t_U) \leq t_c$)

In generale, la procedura arrivo si innesca quando il tempo del prossimo arrivo dopo il tempo di osservazione t è minore o uguale al tempo di completamento del servizio dell'utente attualmente in servizio ed inoltre il tempo del prossimo arrivo dopo t (minimo tra i due tempi) è minore o uguale a t_c.

1. $t \leftarrow t_A$ (aggiornare il tempo di osservazione t all'attuale istante di arrivo);

2. $N_A \leftarrow N_A + 1$ (incrementare di uno il numero cumulativo degli arrivi);

3. $n \leftarrow n + 1$ (incrementare di uno il numero di utenti nel sistema);

4. se $n = 1$ (l'utente che arriva trova il servitore libero, ossia l'utente arrivato è l'unico nel sistema):

   - generare un tempo di servizio S con la distribuzione di probabilità desiderata e calcolare l'istante di completamento del servizio (registrandolo come evento futuro). La nuova partenza avverrà al tempo $t_U = t + S$, dove S è il tempo di servizio (generato mediante simulazione).

   altrimenti se $n > 1$ (l'utente che arriva non trova il servitore libero)

   - $n_q \leftarrow n_q + 1$ (inserire l'utente in coda secondo la disciplina di servizio e incrementare di uno il numero di utenti in coda).

5. generare un tempo di interarrivo T con la distribuzione di probabilità desiderata e calcolare l'istante del prossimo arrivo (registrandolo come evento futuro). Il nuovo arrivo avverrà al tempo $t_A = t + T$, dove T è un tempo di interarrivo (generato mediante la simulazione).

Si nota che se il servitore è libero, l'utente appena arrivato non aspetta in coda e uscirà dal sistema dopo un tempo pari al suo tempo di servizio. Quando viene scandito un evento arrivo e il sistema è vuoto viene inserito un nuovo evento partenza nel calendario ad un tempo pari al precedente tempo più il valore del tempo di servizio.

Procedura di partenza ($t_U < t_A, t_U = \min(t_A, t_U) \leq t_c$)

In generale, la procedura di partenza si innesca quando il tempo del prossimo arrivo dopo il tempo di osservazione t è maggiore del tempo di completamento del servizio dell'utente attualmente in servizio ed, inoltre, questo tempo di completamento del servizio (minimo tra i due tempi) è minore o uguale a t_c

1. $t \leftarrow t_U$ (aggiornare il tempo t di osservazione all'attuale istante di partenza);
2. $N_U \leftarrow N_U + 1$ (incrementare di uno il numero cumulativo di utenti serviti);
3. $n \leftarrow n - 1$ decrementare di uno il numero di utenti nel sistema;
4. se $n > 0$ (l'utente in partenza lascia altri utenti in fila di attesa, ossia il sistema non è vuoto)
   - prelevare un utente dalla fila di attesa secondo la disciplina di servizio e porre $n_q \leftarrow n_q - 1$;
   - generare il tempo di servizio S con la distribuzione di probabilità desiderata e calcolare l'istante di partenza (registrandolo come evento futuro). La nuova partenza avverrà al tempo $t_U = t + S$, dove S è il tempo di servizio (generato mediante la simulazione).

   altrimenti se $n = 0$ (non ci sono utenti nel sistema, il sistema è vuoto)
   - porre $t_U = \infty$ (porre il prossimo istante di partenza uguale ad infinito).

Quando viene scandito un evento fine servizio con coda non vuota viene inserito un nuovo evento fine del servizio nel calendario degli eventi ad un tempo pari al precedente tempo più il valore del tempo di servizio.

Procedura di terminazione $(\min(t_A, t_U) > t_c)$

La procedura di terminazione si innesca quando il minimo tra il tempo del prossimo arrivo dopo il tempo di osservazione t ed il tempo di completamento del servizio dell'utente attualmente in servizio dopo il tempo di osservazione t sono maggiori di t_c. In questo caso non possono più accedere nuovi utenti nel sistema ed occorre fornire il servizio agli utenti già entrati nel sistema.

- innescare ripetutamente la procedura di partenza finché nel sistema non sono più presenti utenti, nel qual caso terminare la simulazione e registrare nella variabile t_f l'istante di tempo in cui l'ultimo utente lascia il sistema.

Procedura statistica

Al termine della simulazione abbiamo ottenuto N_A (numero totale degli arrivi) che sarà anche uguale a N_U (numero totale delle partenze) ed inoltre anche t_f, ossia il tempo finale di simulazione corrispondente all'istante in cui l'ultimo utente lascia il sistema. Dalle sequenze dei tempi di interarrivo $T(1), T(2), \ldots, T(N_A)$ e di servizio $S(1), S(2), \ldots, S(N_A)$ si possono immediatamente ricavare i tempi di interarrivo, i tempi di servizio, i tempi di permanenza in coda e nel sistema, gli istanti di arrivo e di partenza e alcune caratteristiche statistiche quali medie campionarie e varianze campionarie. Le stime della media dei tempi di interarrivo e di servizio sono:

$$\overline{T} = \frac{1}{N_A} \sum_{k=1}^{N_A} T(k), \qquad \overline{S} = \frac{1}{N_A} \sum_{k=1}^{N_A} S(k). \tag{11.1}$$

Inoltre, avendo ottenuto tramite le (10.1) i tempi di arrivo $A(1), A(2), \ldots, A(N_A)$ degli utenti nel sistema e tramite la (10.2) i tempi di uscita $U(1), U(2), \ldots, U(N_A)$ degli utenti dal centro di servizio, allora le differenze $W(k) = U(k) - A(k)$ $(k = 1, 2 \ldots, N_A)$ rappresentano i tempi di attesa dei vari utenti nel sistema e le differenze $Q(k) = W(k) - S(k)$ $(k = 1, 2 \ldots, N_A)$ forniscono i tempi di permanenza dei vari utenti nella fila di attesa. Le stime della media dei tempi di attesa nel sistema e nella fila di attesa sono:

$$\overline{W} = \frac{1}{N_A} \sum_{k=1}^{N_A} W(k), \qquad \overline{Q} = \frac{1}{N_A} \sum_{k=1}^{N_A} Q(k) \tag{11.2}$$

I risultati della simulazione permettono anche di ottenere il numero di utenti $N(t)$ presenti nel sistema al tempo t a partire dalle coppie (n, t), dove t denota l'istante di tempo in cui si è verificato un evento (arrivo o partenza) e dove n denota il numero di utenti presenti nel sistema al tempo t. Una stima della probabilità di avere k utenti nel sistema nell'intervallo $(0, t_f)$ può essere così ottenuta:

$$\widehat{q}_k = \frac{\text{tempo trascorso nello stato } k}{t_f} \qquad (k = 0, 1, \ldots, N_A), \tag{11.3}$$

da cui è possibile ricavare una stima della media e della varianza del numero di utenti presenti nel sistema nell'intervallo $(0, t_f)$:

$$\overline{N} = \sum_{i=1}^{N_A} i\ \widehat{q}_i, \qquad \overline{N}_q = \sum_{i=2}^{N_A} (i-1)\ \widehat{q}_i \tag{11.4}$$

Esempio 11.1 Consideriamo un sistema di servizio singolo servitore, singola fila di attesa, a capacità infinita, con disciplina di servizio FIFO.

Supponiamo di aver osservato (oppure simulato) le due sequenze di tempi di interarrivo e di servizio, misurati in minuti, riportate in Tabella 11.1.

k	1	2	3	4	5	6	7	8	9	10	11
$T(k)$	1.73	1.35	0.71	0.62	14.28	0.70	15.52	3.15	0.76	1.00	0.50
$S(k)$	2.90	1.76	3.39	4.52	4.46	4.36	2.09	3.36	2.37	5.38	0.50

Tabella 11.1: Sequenze dei tempi di interarrivo e di servizio.

Scegliamo $t_c = 40$ minuti come istante di tempo oltre il quale non è più concesso agli utenti di accedere al sistema.

Costruiamo un data frame contenente i tempi di interarrivo, i tempi di servizio, i tempi di arrivo, i tempi di partenza, i tempi di permanenza in fila di attesa e i tempi di attesa nel sistema.

```
> tabella<-function(){
+ tint<-c(1.73,1.35,0.71,0.62,14.28,0.70,15.52,3.15,0.76,1.00,0.50)
+ tserv<-c(2.90,1.76,3.39,4.52,4.46,4.36,2.09,3.36,2.37,5.38,0.50)
+ n<-length(tint)
+ a<-numeric(n) # vettore dei tempi di arrivo
```

```
+ p<-numeric(n) # vettore dei tempi di partenza
+ a[1]<-tint[1]
+ p[1]<-a[1]+tserv[1]
+ df<-data.frame(T=1:n,S=1:n,A=1:n,U=1:n,Q=1:n,W=1:n)
+ row.names(df)<-1:n
+ df[1,1]<-tint[1]
+ df[1,2]<-tserv[1]
+ df[1,3]<-a[1]
+ df[1,4]<-p[1]
+ df[1,5]<-0
+ df[1,6]<-tserv[1]
+ for(i in 2:n){
+ a[i]<-a[i-1]+tint[i]
+ p[i]<-max(a[i],p[i-1])+tserv[i]
+ df[i,1]<-tint[i]
+ df[i,2]<-tserv[i]
+ df[i,3]<-a[i]
+ df[i,4]<-p[i]
+ df[i,5]<-round(p[i]-a[i]-tserv[i],4)
+ df[i,6]<-round(p[i]-a[i],4)
+ }
+ return(df)
+ }
```

```
> tabella()
       T    S      A      U    Q    W
1   1.73 2.90   1.73   4.63 0.00 2.90
2   1.35 1.76   3.08   6.39 1.55 3.31
3   0.71 3.39   3.79   9.78 2.60 5.99
4   0.62 4.52   4.41  14.30 5.37 9.89
5  14.28 4.46  18.69  23.15 0.00 4.46
6   0.70 4.36  19.39  27.51 3.76 8.12
7  15.52 2.09  34.91  37.00 0.00 2.09
8   3.15 3.36  38.06  41.42 0.00 3.36
9   0.76 2.37  38.82  43.79 2.60 4.97
10  1.00 5.38  39.82  49.17 3.97 9.35
11  0.50 0.50  40.32  49.67 8.85 9.35
```

Osservando i tempi di arrivo emerge che il tempo di arrivo 40.32 (presente nella terza colonna) è maggiore del tempo $t_c = 40$ di chiusura al pubblico. Selezioniamo la terza colonna e determiniamo il numero di utenti arrivati prima dell'orario di chiusura.

```
> attesa<-tabella()[,3]
> attesa
[1]  1.73   3.08   3.79   4.41  18.69  19.39  34.91  38.06  38.82  39.82  40.32
 >
> tc<-40
> Narrivi<-max(which(attesa<tc))
> Narrivi
[1] 10
```

Consideriamo il data frame ridotto contenente soltanto i risultati per gli utenti che hanno avuto accesso al sistema.

```
> Ridottatabella<-tabella()[1:Narrivi,]
```

```
> Ridottatabella
       T    S     A     U    Q    W
1   1.73 2.90  1.73  4.63 0.00 2.90
2   1.35 1.76  3.08  6.39 1.55 3.31
3   0.71 3.39  3.79  9.78 2.60 5.99
4   0.62 4.52  4.41 14.30 5.37 9.89
5  14.28 4.46 18.69 23.15 0.00 4.46
6   0.70 4.36 19.39 27.51 3.76 8.12
7  15.52 2.09 34.91 37.00 0.00 2.09
8   3.15 3.36 38.06 41.42 0.00 3.36
9   0.76 2.37 38.82 43.79 2.60 4.97
10  1.00 5.38 39.82 49.17 3.97 9.35
```

Le funzioni $\mathsf{apply}(\mathsf{X}, 2, \mathsf{mean})$, $\mathsf{apply}(\mathsf{X}, 2, \mathsf{var})$ e $\mathsf{apply}(\mathsf{X}, 2, \mathsf{sd})$ permettono di calcolare la media campionaria, la varianza campionaria e la deviazionne standard campionaria delle colonne di un data frame X. Per il data frame precedente si ha:

```
> apply(Ridottatabella,2,mean)
     T      S      A      U      Q      W
 3.982  3.459 20.270 25.714  1.985  5.444
>
> apply(Ridottatabella,2,var)
         T           S           A           U           Q           W
 33.768840   1.433721 269.194133 273.655716   3.917294   7.810449
>
> apply(Ridottatabella,2,sd)
         T          S          A          U          Q          W
 5.811096  1.197381 16.407137 16.542543  1.979216  2.794718
```

Le medie campionarie del tempi di interarrivo e dei tempi di servizio di questi dieci utenti sono:

$$\overline{T} = \frac{1}{10}\sum_{k=1}^{10} T(k) = 3.982 \text{ minuti}, \quad \overline{S} = \frac{1}{10}\sum_{k=1}^{10} S(k) = 3.459 \text{ minuti}.$$

Le medie campionarie dei tempi di attesa nel sistema e di permanenza in fila di attesa sono:

$$\overline{Q} = \frac{1}{10}\sum_{k=1}^{10} Q(k) = 1.985 \text{ minuti}, \quad \overline{W} = \frac{1}{10}\sum_{k=1}^{10} W(k) = 5.444 \text{ minuti}. \quad (11.5)$$

Per ottenere tutti i parametri prestazionali del sistema occorre ordinare in ordine crescente gli istanti di arrivo e di partenza ed utilizzare l'algoritmo generale per la simulazione del sistema di servizio con unico servitore. I risultati della simulazione sono elencati in Tabella 11.2, in cui t indica il tempo di osservazione, N_A il numero cumulativo di arrivi, N_U il numero di completamenti di servizio, n il numero di utenti nel sistema, n_q il numero di utenti in fila di attesa, $S(k)$ il tempo di servizio dell'utente k-esimo, t_U l'istante di completamento del servizio (evento futuro), $T(k)$ il tempo di interarrivo intercorrente tra il $(k-1)$-esimo e il k-esimo utente e t_A l'istante del prossimo arrivo (evento futuro).

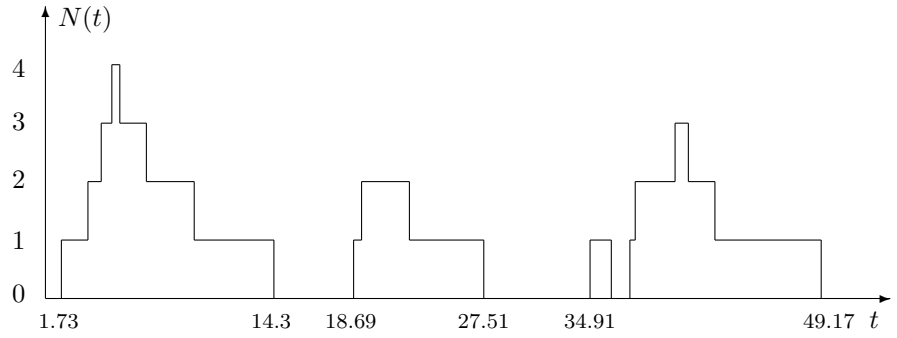

Figura 11.1: Realizzazione del numero di utenti presenti nel sistema di servizio con singolo servitore utilizzando i tempi di arrivo e di partenza.

Nell'ultima colonna Tabella 11.2, Iniz denota la fase di inizializzazione, PA la procedura di arrivo, PU la procedura di partenza e PT la procedura di terminazione. Osservando il tempo di osservazione $t = 39.82$ della Tabella 11.2 emerge che il tempo del prossimo arrivo $t_A = 40.32$ è maggiore del tempo $t_c = 40$ di chiusura dell'accesso degli utenti al sistema. In questo istante di tempo sono arrivati 10 utenti, di cui soltanto 7 sono già stati serviti. Pertanto a partire da questo tempo di osservazione occorre servire gli utenti ancora presenti nel sistema. Inoltre, dalla Tabella 11.2 emerge anche che l'istante di tempo in cui l'ultimo utente lascia il sistema è $t_f = 49.17$ minuti.

Dalla Tabella 11.2 si ricava:

$$N(t) = \begin{cases} 0, & 0 \le t < 1.73 \\ 1, & 1.73 \le t < 3.08 \\ 2, & 3.08 \le t < 3.79 \\ 3, & 3.79 \le t < 4.41 \\ 4, & 4.41 \le t < 4.63 \\ 3, & 4.63 \le t < 6.39 \\ 2, & 6.39 \le t < 9.78 \\ 1, & 9.78 \le t < 14.3 \\ 0, & 14.3 \le t < 18.69 \\ 1, & 18.69 \le t < 19.39 \\ 2, & 19.39 \le t < 23.15 \\ 1, & 23.15 \le t < 27.51 \\ 0, & 27.51 \le t < 34.91 \\ 1, & 34.91 \le t < 37.00 \\ 0, & 37.00 \le t < 38.06 \\ 1, & 38.06 \le t < 38.82 \\ 2, & 38.82 \le t < 39.82 \\ 3, & 39.82 \le t < 41.42 \\ 2, & 41.42 \le t < 43.79 \\ 1, & 43.79 \le t < 49.17 \\ 0, & t = 49.17, \end{cases} \qquad N_q(t) = \begin{cases} 0, & 0 \le t < 3.08 \\ 1, & 3.08 \le t < 3.79 \\ 2, & 3.79 \le t < 4.41 \\ 3, & 4.41 \le t < 4.63 \\ 2, & 4.63 \le t < 6.39 \\ 1, & 6.39 \le t < 9.78 \\ 0, & 9.78 \le t < 19.39 \\ 1, & 19.39 \le t < 23.15 \\ 0, & 23.15 \le t < 38.82 \\ 1, & 38.82 \le t < 39.82 \\ 2, & 39.82 \le t < 41.42 \\ 1, & 41.42 \le t < 43.79 \\ 0, & 43.79 \le t < 49.17 \\ 0, & t = 49.17, \end{cases} \tag{11.6}$$

da cui è possibile ricostruire la realizzazione, illustrata in Figura 11.1, del numero di utenti presenti nel sistema di servizio con singolo servitore. Il grafico di

t	N_A	N_U	n	n_q	$S(k)$	t_U	$T(k)$	t_A	Casi
0	0	0	0	0	-	∞	1.73	1.73	Iniz
1.73	1	0	1	0	2.90	4.63	1.35	3.08	PA
3.08	2	0	2	1	-	-	0.71	3.79	PA
3.79	3	0	3	2	-	-	0.62	4.41	PA
4.41	4	0	4	3	-	-	14.28	18.69	PA
4.63	4	1	3	2	1.76	6.39	-	-	PP
6.39	4	2	2	1	3.39	9.78	-	-	PP
9.78	4	3	1	0	4.52	14.3	-	-	PP
14.3	4	4	0	0	-	∞	-	-	PP
18.69	5	4	1	0	4.46	23.15	0.70	19.39	PA
19.39	6	4	2	1	-	-	15.52	34.91	PA
23.15	6	5	1	0	4.36	27.51	-	-	PP
27.51	6	6	0	0	-	∞	-	-	PP
34.91	7	6	1	0	2.09	37.00	3.15	38.06	PA
37.00	7	7	0	0	-	∞	-	-	PP
38.06	8	7	1	0	3.36	41.42	0.76	38.82	PA
38.82	9	7	2	1	-	-	1.00	39.82	PA
39.82	10	7	3	2	-	-	0.5	$40.32 > t_c$	PA
41.42	10	8	2	1	2.37	43.79	-	-	PT
43.79	10	9	1	0	5.38	49.17	-	-	PT
49.17	10	10	0	0	-	∞	-	-	PT

Tabella 11.2: Esempio di funzionamento del simulatore del sistema di servizio con singolo servitore con soltanto dieci arrivi.

Figura 11.1 può anche essere realizzato in R nel seguente modo:

```
> tempiarrivi<-Ridottatabella[,3]
> tempipartenze<-Ridottatabella[,4]
> tempi<-sort(c(0,tempiarrivi,tempipartenze))
> n<-c(0,1,2,3,4,3,2,1,0,1,2,1,0,1,0,1,2,3,2,1,0)
> plot(tempi,n,type="s",col="red")
```

La funzione c(0, tempiarrivi, tempipartenze) crea un vettore contenente 0, i tempi di arrivo e i tempi dii partenza degli utenti e la funzione sort() ordina in ordine crescente gli elementi del vettore. L'opzione type = "s" presente nella funzione plot() permette di unire i punti avente per ascissa i tempi e per ordinate il numero di utenti nel sistema ai vari tempi utilizzando una funzione a gradini.

Inoltre, a partire dalla Tabella 11.6 è anche possibile ricostruire una realizzazione, illustrata in Figura 11.2, del numero di utenti presenti in fila di attesa. Il grafico di Figure 11.2 può anche essere così realizzato in R:

```
> tempiarrivi<-Ridottatabella[,3]
> tempipartenze<-Ridottatabella[,4]
> tempi<-sort(c(0,tempiarrivi,tempipartenze))
> nq<-c(0,0,1,2,3,2,1,0,0,0,1,0,0,0,0,0,1,2,1,0,0)
> plot(tempi,nq,type="s",col="blue")
```

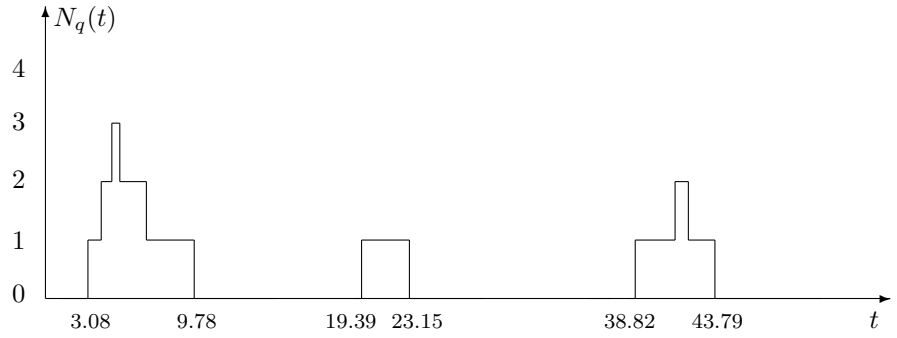

Figura 11.2: Una realizzazione del numero di utenti presenti in fila di attesa.

Facendo uso di (11.6) in (11.3) si ottiene poi una stima delle probabilità di avere k utenti nel sistema nell'intervallo $(0, 49.17)$:

$$\begin{aligned}
\widehat{q}_0 &= \frac{1.73 + 4.39 + 7.4 + 1.06}{49.17} = \frac{14.58}{49.17} = 0.2965,\\
\widehat{q}_1 &= \frac{1.35 + 4.52 + 0.7 + 4.36 + 2.09 + 0.76 + 5.38}{49.17} = \frac{19.16}{49.17} = 0.3897,\\
\widehat{q}_2 &= \frac{0.71 + 3.39 + 3.76 + 1.00 + 2.37}{49.17} = \frac{11.23}{49.17} = 0.2284,\\
\widehat{q}_3 &= \frac{0.62 + 1.76 + 1.6}{49.17} = \frac{3.98}{49.17} = 0.0809,\\
\widehat{q}_4 &= \frac{0.22}{49.17} = 0.0045.
\end{aligned}$$

Utilizzando poi la (11.4) con $N_A = 10$ si ricava una stima del numero medio di utenti nel sistema nell'intervallo $(0, t_f) = (0, 49.17)$:

$$\overline{N} = 0 \cdot \widehat{q}_0 + 1 \cdot \widehat{q}_1 + 2 \cdot \widehat{q}_2 + 3 \cdot \widehat{q}_3 + 4 \cdot \widehat{q}_4 = \frac{54.44}{49.17} = 1.1072.$$

Ricordando la (11.6) si ottiene anche una stima delle probabilità di avere k utenti nella fila di attesa nell'intervallo $(0, t_f)$:

$$\begin{aligned}
\widetilde{q}_0 &= \frac{3.08 + 9.61 + 15.67 + 5.38}{49.17} = \frac{33.74}{49.17} = 0.6862,\\
\widetilde{q}_1 &= \frac{0.71 + 3.39 + 3.76 + 1.00 + 2.37}{49.17} = \frac{11.23}{49.17} = 0.2284,\\
\widetilde{q}_2 &= \frac{0.62 + 1.76 + 1.6}{49.17} = \frac{3.98}{49.17} = 0.0809,\\
\widetilde{q}_3 &= \frac{0.22}{49.17} = 0.0045,
\end{aligned}$$

da cui utilizzando la (11.4) con $N_A = 10$ si ricava una stima del numero medio di utenti in fila di attesa nell'intervallo $(0, t_f)$:

$$\overline{N}_q = 0 \cdot \widetilde{q}_0 + 1 \cdot \widetilde{q}_1 + 2 \cdot \widetilde{q}_2 + 3 \cdot \widetilde{q}_3 = \frac{19.85}{49.17} = 0.4037,$$

o equivalentemente:

$$\overline{N}_q = 1 \cdot \widehat{q}_2 + 2 \cdot \widehat{q}_3 + 3 \cdot \widehat{q}_4 = \frac{19.85}{49.17} = 0.4037.$$

Una stima del fattore di utilizzazione (intensità di traffico) del sistema nell'intervallo di tempo $(0, t_f)$ è rappresentata dalla media campionaria degli utenti in servizio:

$$\overline{N}_s = \overline{N} - \overline{N}_q = 0.7035.$$

Si nota che sussiste l'identità:

$$\overline{N}_s = 1 - \widehat{q}_0,$$

ossia la media campionaria degli utenti in servizio coincide con la probabilità stimata di avere almeno un utente nel sistema.

Nella simulazione dei sistemi di servizio nel transiente non sussistono le leggi di Little (che sono applicabili soltanto quando il sistema ha raggiunto una situazione di equilibrio statistico). ♢

11.3 Sistema con due differenti servitori che lavorano in parallelo

Consideriamo un sistema di servizio con due servitori che lavorano in parallelo, singola fila di attesa a capacità infinita con disciplina di servizio FIFO, in cui i tempi di servizio del servitore i-esimo ($i = 1, 2$) hanno una distribuzione di probabilità G_i di tipo generale (deterministica, esponenziale, ...). Tale sistema di servizio è illustrato in Figura 11.3.

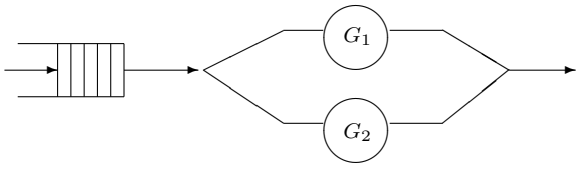

Figura 11.3: Sistema di servizio con due servitori che lavorano in parallelo.

Il sistema di servizio schematizzato in Figura 11.3 può prevedere servitori con differenti velocità di servizio, ad esempio uno molto veloce e l'altro estremamente lento. In particolare, se la distribuzione dei tempi di interarrivo è esponenziale e se i due servitori sono identici, con distribuzione di servizio esponenziale, si ottiene il sistema di servizio $M/M/2$.

Dopo l'arrivo l'utente si metterà in fila di attesa se entrambi i servitori sono occupati, entrerà in servizio dal servitore 1 se il servitore 2 è occupato, entrerà in servizio dal servitore 2 se il primo servitore è occupato. Assumiamo che *se entrambi i servitori sono liberi l'utente sceglie in maniera casuale uno dei due servitori*. Quando l'utente completa il servizio da uno dei due servitori, l'utente uscirà dal sistema e l'utente arrivato per primo (se ne sono presenti più di uno) entrerà nel centro di servizio.

Scegliamo come modalità di terminazione della simulazione il *numero massimo di utenti a cui il sistema deve fornire servizio* e denotiamo con t_f il *tempo in cui l'ultimo utente lascia il sistema*.

Supponiamo di voler simulare il precedente sistema registrando il

- *tempo di attesa nel sistema di ogni utente*,

- *numero di completamenti di servizio effettuati da ognuno dei due servitori*.

Poiché sono presenti due servitori, *gli utenti non partiranno necessariamente nell'ordine in cui sono arrivati*. Quindi, per conoscere quale utente ha lasciato il sistema dopo il completamento del servizio occorre registrare gli utenti presenti nel sistema. Così *numeriamo gli utenti che arrivano nel sistema*: il primo arrivo è l'utente numero 1, il successivo è l'utente numero 2 e così via. Poiché la disciplina di servizio è quella FIFO (entrano in servizio secondo il loro arrivo), segue che la conoscenza di *quali utenti* (ossia del numero ad essi associato) *sono in servizio* e di *quanti utenti sono nel sistema*, ci permette di identificare gli utenti in fila di attesa. Ad esempio, supponiamo che gli utenti numerati i e j sono in servizio con $i < j$, e che ne esistano $n > 2$ nel sistema, ossia due in servizio e $n-2$ in fila di attesa. Poiché tutti gli utenti con numero minore di j sono già entrati in servizio prima dell'utente j-esimo e inoltre nessun utente con numero più alto di j ha completato il servizio, segue che gli utenti numerati con $j+1, \ldots, j+n-2$ sono in fila di attesa.

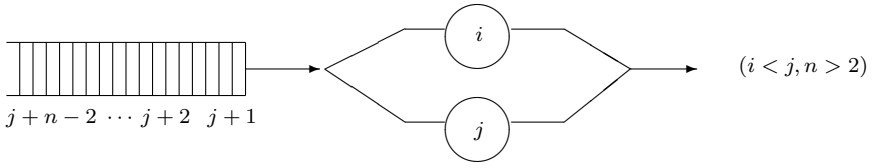

Figura 11.4: Sistema di servizio con due servitori con n utenti nel sistema.

Procedura generale di simulazione

Per effettuare la simulazione del sistema di servizio illustrato in Figura 11.3, con due servitori che lavorano in parallelo, usiamo le seguenti variabili:

Variabile temporale: t (tempo di osservazione)

Variabili contatore:

N_A: numero cumulativo di arrivi fino al tempo t

C_1: numero cumulativo di partenze dal primo servitore fino al tempo t;

C_2: numero cumulativo di partenze dal secondo servitore fino al tempo t.

Variabile di stato del sistema:

SS $= (n, i_1, i_2)$ (sono presenti n utenti nel sistema al tempo t; i_1 è il numero associato all'utente attualmente in servizio dal primo servitore e i_2 è il numero associato all'utente attualmente in servizio dal secondo servitore);

SS $= (0)$ quando il sistema è vuoto al tempo t;

SS $= (1, j, 0)$ quando al tempo t l'unico utente nel sistema è il j-esimo ed è servito dal primo servitore;

SS $= (1, 0, j)$ quando al tempo t l'unico utente è il j-esimo ed è servito dal secondo servitore.

Poiché le precedenti variabili cambiano valore quando si verifica un arrivo oppure una partenza di un utente dal primo servitore oppure una partenza dal secondo servitore, scegliamo *gli arrivi dall'esterno e le partenze da ognuno dei due servitori come eventi.* La lista degli eventi contiene il tempo del nuovo arrivo e il tempo della partenza dell'utente attualmente in servizio dal primo e dal secondo servitore del sistema di servizio:

- t_A è il tempo del nuovo arrivo (successivo a t),

- t_i è il tempo di completamento del servizio dell'utente attualmente in servizio presso il servitore i-esimo ($i = 1, 2$). Se non ci sono utenti in servizio presso il servitore i-esimo allora t_i ($i = 1, 2$) verrà posto uguale a ∞.

Le **variabili di output** da calcolare sono:

- l'istante di tempo t_f in cui la simulazione termina;

- $A(n)$ il tempo di arrivo dell'utente n-esimo nel sistema ($n = 1, 2, \ldots$),

- $U(n)$ il tempo di partenza dell'utente n-esimo dal sistema ($n = 1, 2, \ldots$),

- la quadrupla (n, i_1, i_2, t), ossia il numero n di utenti nel sistema, il numero i_1 associato all'utente in servizio dal primo servitore e il numero i_2 associato all'utente in servizio dal secondo servitore in ogni istante di tempo $t = t_A, t_1, t_2$ in cui si verifica un evento.

Per iniziare la simulazione occorre inizializzare le variabili e i tempi degli eventi come segue:

Inizializzazione:

Al tempo $t = 0$ si inizializzano le seguenti variabili:

- $N_A \leftarrow 0$, $C_1 \leftarrow 0$, $C_2 \leftarrow 0$ (il numero cumulativo di arrivi e il numero cumulativo di partenze da ognuno dei due servitori fino al tempo $t = 0$ sono nulli);

- **SS** $= (0)$ (il sistema è vuoto al tempo $t = 0$);

- Generare T (il tempo del prossimo arrivo nel sistema dopo $t = 0$) e porre $t_A = T$ e $t_1 = t_2 = \infty$ (si pone il tempo di servizio infinito per entrambi i servitori non essendoci utenti in servizio al tempo $t = 0$).

Per procedere nella simulazione occorre muoversi lungo l'asse temporale fino ad incontrare il prossimo evento. Al passo successivo occorre considerare casi differenti a seconda di quale evento si verifica per primo.

Procedura di arrivo: **SS** $= (n, i_1, i_2), t_A = \min(t_A, t_1, t_2)$
In questo caso il tempo del prossimo arrivo dopo t è inferiore al tempo di completamento del servizio per entrambi i servitori. I passi da compiere sono:

- $t \leftarrow t_A$ (ci muoviamo fino al tempo t_A);

- $N_A \leftarrow N_A + 1$ (poiché esiste un nuovo arrivo al tempo t_A);

- Generare un tempo di interarrivo T e porre $t_A = t + T$ (questo è il tempo del nuovo arrivo dopo t);

- In output registrare $A(N_A) = t$ (il tempo di arrivo dell'utente numerato N_A è t).

Inoltre poiché si ha un arrivo:

- Se lo stato del sistema era **SS** $= (0)$ (entrambi i servitori liberi)

  - generare una variabile aleatoria discreta che assume i valori 1 e 2 (numeri associati ai servitori) con distribuzione equiprobabile.

  - Se la generazione fornisce come risultato 1, l'utente si dirigerà verso il primo servitore, lo stato del sistema diventa **SS** $= (1, N_A, 0)$ ed occorre generare un tempo di servizio S_1 e porre $t_1 = t + S_1$.

  - Se la generazione fornisce come risultato 2, l'utente si dirigerà verso il secondo servitore, lo stato del sistema diventa **SS** $= (1, 0, N_A)$ ed occorre generare un tempo di servizio S_2 e porre $t_2 = t + S_2$.

- Se lo stato del sistema era **SS** $= (1, j, 0)$, lo stato del sistema diventa

  - **SS** $= (2, j, N_A)$ (nel sistema è presente un solo utente che è in servizio dal primo servitore e arriva un nuovo utente che accede al servizio offerto dal secondo servitore; il numero associato all'utente in servizio dal secondo servitore è N_A)

  - Generare S_2 (tempo di servizio dell'utente attualmente in servizio dal secondo servitore) e porre $t_2 = t + S_2$ (tempo di completamento del servizio dell'utente in servizio dal secondo servitore)

- Se lo stato del sistema era $\mathbf{SS} = (1, 0, j)$ lo stato del sistema diventa

  $\mathbf{SS} = (2, N_A, j)$ (nel sistema è presente un solo utente che è in servizio dal secondo servitore e arriva un nuovo utente che accede al servizio offerto dal primo servitore; il numero associato all'utente in servizio dal primo servitore è N_A)

  - Generare S_1 e porre $t_1 = t + S_1$

- Se il numero di utenti presenti nel sistema era $n > 1$ e $SS = (n, i_1, i_2)$, lo stato del sistema diventa

  - $\mathbf{SS} = (n+1, i_1, i_2)$ (entrambi i servitori sono occupati e il nuovo utente che arriva si metterà in fila di attesa).

Procedura di partenza dal primo servitore: $\mathbf{SS} = (n, i_1, i_2)$, $t_1 < t_A$, $t_1 \leq t_2$
In questo caso il tempo del prossimo arrivo dopo t è maggiore del tempo di completamento del servizio dell'utente attualmente in servizio dal primo servitore ed inoltre tale tempo t_1 è inferiore al tempo di completamento del servizio dell'utente attualmente in servizio dal secondo servitore. Quindi l'uscita dal sistema di un utente si deve verificare dal primo servitore. I passi da compiere sono:

- $t \leftarrow t_1$ (ci muoviamo fino al tempo t_1);

- $C_1 \leftarrow C_1 + 1$ (esiste una partenza dal primo servitore);

- In output registrare il tempo $U(i_1) = t$ (il tempo di partenza dell'utente i_1 è t).

Inoltre, poiché si ha una partenza dal primo servitore:

- Se il numero di utenti presenti nel sistema era $n = 1$

  - $\mathbf{SS} = (0)$ (è presente un utente dal primo servitore e tale utente lascia il sistema)

  - $t_1 = \infty$

- Se il numero di utenti presenti nel sistema era $n = 2$

  - $\mathbf{SS} = (1, 0, i_2)$ (sono presenti due utenti in servizio e si verifica una partenza dal primo servitore)

  - $t_1 = \infty$

- Se il numero di utenti presenti nel sistema era $n > 2$ e $\mathbf{SS} = (n, i_1, i_2)$, posto $m = \max(i_1, i_2)$

  - $\mathbf{SS} = (n-1, m+1, i_2)$ (esistono utenti in coda e una partenza dal primo servitore consente ad un utente in fila di attesa di accedere al servizio del primo servitore; il numero associato a tale utente è $m + 1$)

  - Generare S_1 e porre $t_1 = t+S_1$ (poiché un nuovo utente accede al servizio del primo servitore essendo non vuota la fila di attesa)

Procedura di partenza dal secondo servitore: **SS** $= (n, i_1, i_2)$, $t_2 < t_A$, $t_2 \leq t_1$
In questo caso il tempo del prossimo arrivo dopo t è maggiore del tempo di completamento del servizio dell'utente attualmente in servizio dal secondo servitore ed inoltre tale tempo t_2 è inferiore al tempo di completamento del servizio dell'utente attualmente in servizio dal primo servitore. Quindi l'uscita dal sistema di un utente si deve verificare dal secondo servitore. I passi da compiere sono:

- $t \leftarrow t_2$ (ci muoviamo fino al tempo t_2);

- $C_2 \leftarrow C_2 + 1$ (esiste una partenza dal secondo servitore);

- In output registrare il tempo $U(i_2) = t$ (il tempo di partenza dell'utente i_2 è t).

Inoltre poiché si ha una partenza dal secondo servitore:

- Se il numero di utenti presenti nel sistema era $n = 1$

  - **SS** $= (0)$ (è presente un utente dal secondo servitore e tale utente lascia il sistema)

  - $t_2 = \infty$

- Se il numero di utenti presenti nel sistema era $n = 2$

  - **SS** $= (1, i_1, 0)$ (sono presenti due utenti e si verifica una partenza dal secondo servitore)

  - $t_2 = \infty$

- Se il numero di utenti presenti nel sistema era $n > 2$ e **SS** $= (n, i_1, i_2)$, posto $m = \max(i_1, i_2)$

  - **SS** $= (n - 1, i_1, m + 1)$ (esistono utenti in coda e una partenza dal secondo servitore consente ad un utente in fila di attesa di accedere al servizio del secondo servitore; il numero associato a tale utente è $m + 1$)

  - Generare S_2 e porre $t_2 = t + S_2$ (poiché un nuovo utente accede al servizio del secondo servitore essendo non vuota la fila di attesa)

Utilizzando questo algoritmo è possibile ottenere i *tempi di arrivo e di partenza dei vari utenti*, così come il *numero di completamenti di servizio effettuati da ognuno dei due servitori*. Al termine della simulazione abbiamo ottenuto N_A (numero complessivo di arrivi), C_1 (numero complessivo di partenze dal primo servitore) e C_2 (numero complessivo di partenze dal secondo servitore), da cui è possibile ricavare il numero complessivo di utenti a cui è stato fornito il servizio, ossia $N_U = C_1 + C_2$. La simulazione termina quando abbiamo fornito il servizio a tutti gli utenti previsti.

Per ogni $i = 1, 2, \ldots, N_A$ abbiamo registrato $A(i)$ (il tempo di arrivo dell'utente i-esimo) e $U(i)$ il tempo di uscita dal sistema dell'utente i-esimo). Quindi

$U(i) - A(i)$ rappresenta il tempo che l'utente i-esimo spende nel sistema. La media campionaria del tempo di attesa nel sistema è:

$$\overline{W} = \frac{1}{N_A}\sum_{i=1}^{N_A}[U(i) - A(i)] \tag{11.7}$$

I risultati della simulazione permettono inoltre di ottenere il numero di utenti presenti nel sistema al tempo t a partire dalle coppie (n, t), dove t denota l'istante di tempo in cui si è verificato un evento e n il numero di utenti presenti nel sistema al tempo t. Una stima delle probabilità di avere i utenti nel sistema nell'intervallo $(0, t_f)$ può essere così ottenuta:

$$\widehat{q}_i = \frac{\text{tempo trascorso nello stato i}}{t_f} \qquad (i = 0, 1, \ldots, N_A), \tag{11.8}$$

da cui è possibile ricavare una stima della media del numero di utenti presenti nel sistema nell'intervallo $(0, t_f)$:

$$\overline{N} = \sum_{i=1}^{N_A} i\ \widehat{q}_i. \tag{11.9}$$

È anche possibile studiare le potenzialità dei due servitori analizzando i numeri medi complessivi di partenze $\overline{C}_1$ e $\overline{C}_2$ dal primo e dal secondo servitore fino al tempo t_f.

Spero che questi appunti possano costituire un valido aiuto nella preparazione dell'esame di Simulazione. Vi auguro di proseguire e completare con successo la vostra carriera universitaria.

Amelia G. Nobile

Indice