Note di Simulazione

G. Gallo

2006/2007

Indice

Introduzione

Con il termine *simulazione* si intende l'attività del replicare per mezzo di opportuni modelli una realtà già esistente o da progettare, al fine di studiare, nel primo caso, gli effetti di possibili interventi o eventi in qualche modo prevedibili, o, nel secondo, di valutare diverse possibili scelte progettuali alternative.

L'uso di modelli come strumento di aiuto nei processi decisionali è antico e diffusissimo. Un tipico esempio è quello dei *modelli a scala*, usati soprattutto in fase di progettazione. Si tratta di modelli che replicano fedelmente, anche se a scala ridotta, la realtà che si vuole rappresentare. Tipici modelli di questo tipo sono i plastici che vengono utilizzati nella progettazione architettonica, o i modelli di strutture che vengono utilizzati per studiare gli effetti di sollecitazioni, ad esempio di tipo sismico. Questi strumenti sono caratterizzati da un notevole costo di realizzazione e da una grande rigidità di uso, e pertanto sempre meno utilizzati in pratica.

Altri importantissimi modelli molto usati come strumenti decisionali, soprattutto con lo sviluppo e la diffusione della Ricerca Operativa, sono i *modelli analitici*. Si tratta di modelli in cui la realtà sotto esame viene rappresentata per mezzo di variabili e relazioni di tipo logico/matematico. A questa classe appartengono, fra gli altri, i modelli di *programmazione lineare* (più in generale di *programmazione matematica*) o i modelli di *file d'attesa*. Si tratta di modelli di notevole potenza, che consentono in molti casi di determinare, con un costo contenuto, una o più soluzioni ottime (o comunque soluzioni molto buone) per il problema considerato. Tuttavia al crescere della complessità e della dimensione dei problemi tali modelli diventano di uso sempre più difficile e costoso, ed in molti casi, per le loro stesse caratteristiche, inapplicabili.

La complessità di un processo decisionale ha diverse dimensioni: il numero delle variabili, il tipo di relazioni che legano fra loro le variabili, il numero

di obiettivi, il numero di attori, cioè di persone che hanno la possibilità di prendere decisioni o di influire su esse, ed infine il grado di incertezza con cui le grandezze in gioco e le relazioni fra le variabili sono conosciute. In generale situazioni in cui il numero di obiettivi ed il numero di decisori è sufficientemente limitato (idealmente un obiettivo ed un unico decisore) ed in cui le relazioni fra le variabili e le grandezze in gioco sono conosciute con sufficiente approssimazione si prestano abbastanza bene all'uso di modelli di tipo analitico. Ovviamente purché il numero di variabili necessarie per descrivere la realtà sotto esame non sia eccessivamente grande, dove i limiti nelle dimensioni dei problemi trattabili dipendono grandemente dalla complessità delle relazioni che legano fra loro le variabili.

Qui faremo riferimento a processi decisionali caratterizzati da elevati livelli di complessità, in cui l'uso di modelli analitici sia poco proponibile. I modelli che presenteremo si differenziano da quelli di tipo analitico per l'uso del calcolatore come strumento non solo di calcolo, come ad esempio nei modelli di programmazione matematica, ma anche di rappresentazione degli elementi che costituiscono la realtà in studio e delle relazioni fra di essi. La corrispondenza tra realtà e modello non è basata su una riduzione proporzionale delle dimensioni, ma è di tipo funzionale: ad ogni elemento del sistema reale corrisponde un oggetto informatico (una sottoprogramma, una struttura di dati, ...) che ne svolge la funzione nel modello.

Rispetto alla sperimentazione diretta, costosissima e spesso praticamente impossibile, o a quella realizzata per mezzo di modelli a scala, la simulazione ha il vantaggio della grande versatilità, della velocità di realizzazione e del (relativamente) basso costo. È possibile attraverso la simulazione provare rapidamente politiche e scelte progettuali alternative e modellare sistemi anche di grandissima complessità studiandone il comportamento e l'evoluzione nel tempo.

Capitolo 1

Problemi e Modelli

1.1 Processi decisionali e modelli di simulazione

Il *processo decisionale*, cioè il processo attraverso cui, a partire dall'emergere di una situazione che richiede una scelta o una azione, si arriva alla scelta dell'azione da intraprendere e poi alla sua realizzazione, è oggetto di studio in settori notevolmente diversi che vanno dall'ingegneria all'informatica, dalla sociologia alla teoria della politica, dall'economia alle scienze gestionali.

Lo studio dei processi decisionali, la capacità di analizzarne e scomporne i meccanismi, e soprattutto la messa a punto di strumenti sia metodologici che tecnici di supporto è essenziale per pervenire a 'buone' decisioni. Spesso è il processo decisionale in se stesso che produce risultati significativi al di là delle decisioni ed azioni alle quali esso porta; questo per la sua caratteristica di essere un processo di apprendimento che in qualche modo cambia gli attori stessi in esso coinvolti.

In generale in un processo decisionale il punto di partenza è l'individuazione di una *realtà problematica* che richiede un cambiamento e quindi una decisione. La realtà così individuata viene analizzata in modo da evidenziare al suo interno il *sistema* da studiare ai fini della o delle decisioni da prendere; vengono cioè scelti quegli elementi che ci sembrano più rilevanti, evidenziate le relazioni che li collegano, e definiti gli obiettivi da raggiungere. A questo punto si costruisce un *modello* formale che permetta di riprodurre (*simulare*) il sistema individuato, allo scopo di comprenderne il comportamento e di arrivare ad individuare le decisioni da prendere.

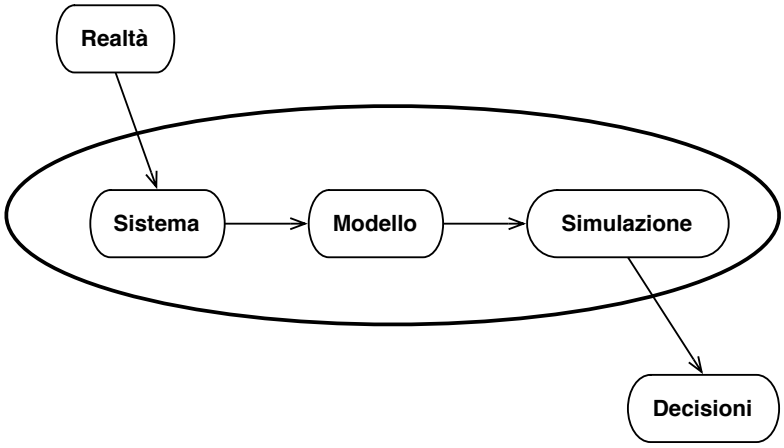

Figura 1.1. *Processo decisionale*

Il processo che abbiamo delineato parte quindi dalla *realtà* ed arriva alla o alle *decisioni* finali attraverso tre passi: l'individuazione del *sistema* da studiare, la costruzione del *modello* e la *simulazione*. Questo processo è sinteticamente rappresentato in figura 1.1, dove è stata evidenziata la parte che qui ci interessa, quella che va dalla definizione del sistema fino alla simulazione.

Esamineremo nel seguito più in dettaglio i passi del processo decisionale.

Il sistema. La realtà oggetto di indagine viene rappresentata attraverso un sistema, cioè un insieme di elementi interagenti fra loro. Quella di rappresentare la realtà come un sistema è una scelta, ed il risultato di tale rappresentazione è la conseguenza di una successione di scelte specifiche, tutte caratterizzate da un certo grado di arbitrarietà e quindi suscettibili di revisione nel corso del processo decisionale. La principale e più critica scelta riguarda i *confini* del sistema, cioè quali elementi della realtà debbano essere inseriti nel sistema che la rappresenta e quali invece lasciati fuori. Ad esempio, nello studio del traffico privato in una area urbana dovrò scegliere quale porzione della rete viaria considerare e dove disegnare i confini dell'area da studiare. Si tratta di scelte che riguardano aspetti fisici della realtà che sembra non pongano rilevanti problemi: alcune strade secondarie sono chiaramente poco rilevanti ai fini dei flussi principali di traffico, e appare naturale

limitarsi a considerare l'area nella quale ci interessa conoscere la distribuzione del traffico. In realtà le scelte fatte anche su aspetti di questo tipo non sono neutre, e possono falsare i risultati ottenuti. La possibilità di usare alcune strade apparentemente secondarie, ad esempio, può avere effetti imprevisti sulla distribuzione del traffico. Inoltre può accadere che una parte dei veicoli in ingresso nell'area possa utilizzare più punti di accesso, e la scelta dipende dalla distribuzione del traffico dentro l'area stessa; questo porta ad effetti sul traffico totale che possono non essere valutabili senza ampliare l'area sotto esame. Altre scelte riguardano quali variabili considerare e la definizione delle relazioni fra le variabili o elementi del sistema. Naturalmente in questo lavoro bisogna essere guidati dagli obiettivi che il nostro processo decisionale ha. Diversi obiettivi portano a rappresentazioni diverse della stessa realtà.

È necessario tenere sempre presente lo scarto che esiste tra il sistema e la realtà che esso rappresenta. Questo scarto può essere maggiore o minore, ma è comunque ineludibile. La realtà non è direttamente conoscibile se non attraverso una 'concettualizzazione' da parte dell'osservatore, e l'ottica sistemica è proprio lo strumento usiamo a questo scopo. Noi conosciamo la realtà attraverso il sistema con cui la rappresentiamo. Si tratta di una rappresentazione che dobbiamo essere sempre disponibili a rimettere in discussioni. Situazioni nuove ed impreviste ci potranno portare a rivedere il sistema che abbiamo definito, arricchendolo e modificandolo.

Il modello. Il modello costituisce il modo con cui noi formalizziamo il sistema che rappresenta la realtà in esame. I modelli possono essere di tipo diverso, e possono essere sia quantitativi che qualitativi. Ad esempio, modelli classici ed un tempo molto usati sono i cosiddetti *modelli a scala*. Si tratta di modelli fisici che rappresentano a scala ridotta un sistema. Tipici esempi sono il plastico di un quartiere o di una intera città, utilizzato per scelte di tipo architettonico o urbanistico, oppure il modello a scala della struttura di un edificio che viene utilizzato per valutare la risposta della struttura a sollecitazioni ad esempio di tipo sismico. Questi sono modelli di tipo analogico; la riduzione della scala viene fatta in modo che le caratteristiche di interesse si mantengano, eventualmente riducendosi nella loro intensità secondo scala scelta. Questi modelli con la diffusione dei calcolatori elettronici sono sempre meno usati.

Una altra classe di modelli è costituita dai *modelli analitici*. Si tratta di modelli in cui il sistema viene formalizzato attraverso un insieme di variabili

e un insieme di relazioni matematiche che limitano e definiscono i valori che tali variabili possono assumere. Ad esempio una rete elettrica può essere rappresentata per mezzo di un opportuno sistema di equazioni, la cui soluzione fornisce i valori che variabili quali intensità di corrente e differenze di potenziale possono assumere. Molto spesso in questo tipo di modelli viene anche definito una funzione obiettivo da minimizzare o massimizzare. Si ricorre in questo caso ad algoritmi di ottimizzazione. Modelli analitici sono quelli studiati nell'ambito della Programmazione Matematica (Bigi et al., 2003) oppure nell'ambito della teoria delle code.

I modelli a scala e quelli analitici sono, sia pure in forma diversa, modelli abbastanza 'rigidi' e caratterizzati da una limitata ricchezza espressiva. Possono essere usati per modellare sistemi 'relativamente' semplici[1]. Molto poco rigidi e capaci di rappresentare una grande varietà di diverse situazioni sono modelli qualitativi, quali ad esempio i diversi tipi di *mappe cognitive* che vengono utilizzate per rappresentare realtà problematiche nell'ambito della cosiddetta "Soft Operations Research"(Checkland, 1989; Rosenhead, 1989).

I *modelli di simulazione*, che sono quelli che più ampiamente tratteremo nel seguito, sono modelli che si differenziano da quelli di tipo analitico per l'uso del calcolatore come strumento non solo di calcolo, come ad esempio nei modelli di programmazione matematica, ma anche di rappresentazione degli elementi che costituiscono la realtà in studio e delle relazioni fra di essi. La corrispondenza tra realtà e modello non è basata su una riduzione proporzionale delle dimensioni, ma è di tipo funzionale: ad ogni elemento del sistema reale corrisponde un oggetto informatico (un sottoprogramma, una struttura di dati, ...) che ne svolge la funzione nel modello. Questi modelli sono particolarmente flessibili consentendo di rappresentare e di studiare sistemi molto complessi, e dei quali conosciamo alcune caratteristiche solo attraverso analisi di tipo statistico.

I modelli di simulazione sono in genere modelli dinamici, cioè includono la dimensione temporale, e hanno lo scopo di studiare l'andamento nel tempo di un sistema. Al contrario i modelli analitici sono spesso (anche se non sempre) statici: forniscono la soluzione al problema studiato a partire da dati che descrivono il sistema in un dato istante o intervallo temporale[2].

[1] In realtà i modelli di Programmazione Matematica consentono di affrontare problemi anche con milioni di variabili. Le difficoltà nascono soprattutto quando si è in presenza di forti nonlinearità e di situazioni caratterizzate da incertezza.

[2] È bene avvertire il lettore che ogni tentativo di classificazione è necessariamente problematica. La realtà sfugge sempre ai nostri tentativi di racchiuderla in categorie predefi-

La simulazione. La fase della simulazione vera e propria è quella conclusiva, che porterà poi alle decisioni finali. Innanzitutto il modello viene tradotto in un programma su calcolatore che viene fatto girare. In questo modo, analizzando il comportamento del modello e confrontandolo con i dati in nostro possesso sulla da cui eravamo partiti, è possibile verificare quanto il modello costruito rappresenti, correttamente rispetto ai nostri obiettivi[3], la realtà sotto studio. La correttezza può essere vista da due punti di vista diversi:

- *Correttezza d'insieme* (black box validity): gli output che il modello produce riflettono accuratamente quelli del sistema reale.
- *Correttezza delle singole componenti* del sistema (white box validity): le componenti del sistema sono consistenti con la realtà e/o la teoria esistente.

L'implementazione del modello può essere realizzata con diversi strumenti. È possibile usare linguaggi *general purpose* quali Pascal, C, C++, per i quali esistono delle librerie di routines orientate alla simulazione. Esistono anche diversi linguaggi specializzati, quali ad esempio SIMSCRIPT, MODSIM e GPSS. Una interessante alternativa è quella di ricorrere ad applicazioni di tipo interattivo per la simulazione quali, fra gli altri, *Arena*, *Witness*, *Extend* e *Micro Saint*. Tali applicazioni sono di facile uso e quindi molto adatte a costruire rapidamente modelli anche sofisticati, ma sono meno versatili e potenti dei linguaggi specializzati o di quelli *general purpose*. Per problemi di piccole dimensioni è anche possibile usare strumenti informatici di uso comune quali le *spreadsheet*. Tali strumenti possono essere utili quando si vuole rapidamente avere un'idea del funzionamento di una singola componente o di un sottosistema di un sistema complesso.

Una volta assicuratici della correttezza del modello inizia la parte conclusiva e fondamentale del lavoro, quella della sperimentazione che porterà alle decisioni finali. Questa fase richiede l'uso di strumenti statistici, sia per l'analisi dei dati di partenza sia per la valutazione dei risultati della simulazione. A questi strumenti verrà dedicato un ampio spazio nel seguito. La

nite, e questo vale anche per i modelli: esistono modelli analitici che tengono conto delle dinamiche temporali, e d'altra parte alcuni modelli di simulazione sono essenzialmente statici.

[3]Va sempre ricordato che un modello non è la realtà, e che di conseguenza non ha senso parlare di correttezza in sé del modello; la correttezza è relativa agli scopi per cui abbiamo costruito il modello.

sperimentazione deve essere realizzata in modo che l'influenza dei diversi fattori sui risultati ottenuti sia chiaramente evidenziata. Bisogna sempre tenere presente che, in ultima analisi, la simulazione è uno strumento conoscitivo.

1.2 Classi di modelli di simulazione

Un modello di simulazione può essere deterministico o stocastico, discreto o continuo. Si parla di *simulazione deterministica* quando l'evoluzione nel tempo del modello costruito è univocamente determinata dalle sue caratteristiche e dalle condizioni iniziali. Quando nel modello sono presenti grandezze aleatorie che a seconda del valore che assumono possono portare a diversi comportamenti si parla invece di *simulazione stocastica*. Con *simulazione continua* si intende una simulazione in cui il valore delle variabili coinvolte varia in modo continuo nel tempo (anche se poi esse saranno in pratica valutate in istanti discreti). Si ha invece una *simulazione discreta* quando lo stato del sistema studiato, e quindi il valore delle variabili relative, cambia in ben definiti istanti di tempo. Illustreremo meglio questi concetti nel seguito attraverso alcuni esempi.

1.2.1 Modello Preda-Predatore

In una isola sono presenti due popolazioni di animali, conigli e linci. La vegetazione dell'isola fornisce ai conigli nutrimento in quantità che possiamo assumere illimitata, mentre sono i conigli l'unico alimento disponibile per le linci. Possiamo considerare costante nel tempo il tasso di natalità dei conigli; questo significa che in assenza di predatori i conigli crescerebbero con legge esponenziale. Il loro tasso di mortalità invece dipende dalla probabilità che essi hanno di divenire preda di una lince e quindi dal numero di linci presenti per unità di superficie. Per quel che riguarda le linci, il tasso di mortalità è costante, mentre il loro tasso di crescita dipende dalla disponibilità di cibo e quindi dal numero di conigli per unità di superficie presenti nell'isola. Si vuole studiare l'andamento della dimensione delle due popolazioni nel tempo a partire da una situazione iniziale (numero di conigli e di linci) data.

Per effettuare una simulazione di questo sistema biologico possiamo modellarlo per mezzo del seguente sistema di equazioni alle differenze finite, dove $x(t)$ e $y(t)$ sono rispettivamente il numero dei conigli e delle linci al tempo t:

$$x(t+1) - x(t) = N_c x(t) - F(y(t))x(t),$$
$$y(t+1) - y(t) = -M_l y(t) + G(x(t))y(t).$$

In questo modello sono sottese le seguenti ipotesi: (*i*) in assenza di predatori il numero dei conigli cresce secondo una legge esponenziale, cioè con tasso costante; (*ii*) analogamente, in assenza di prede, il numero delle linci decresce con tasso costante.

Nella figura 1.2 viene indicato l'andamento delle due popolazioni nel corso di 50 anni, avendo scelto l'anno come unità di tempo ed avendo posto N_c=0.35, M_l=0.25, $F(y) = 1.25(1 - e^{\frac{-35y}{S}})$ e $G(x) = 0.5(1 - e^{\frac{-0.8x}{S}})$. Con S si è indicata la superficie dell'isola, che si è assunta essere pari a 6000 ettari. I valori iniziali delle due popolazioni sono stati posti a 6000 per i conigli e a 70 per le linci.

Questo modello deriva dal classico *modello preda/predatore* proposto dal matematico italiano Vito Volterra[4] nel 1926 per studiare le variazioni di alcune popolazioni di pesci nell'Adriatico, che, nella sua forma più semplice si presenta così:

$$\frac{dx(t)}{dt} = Ax(t) - Bx(t)y(t),$$
$$\frac{dy(t)}{dt} = -Cy(t) + Dx(t)y(t).$$

Chiaramente si tratta di un modello deterministico e di simulazione continua. Infatti lo stato del sistema (dimensione delle due popolazioni) in ogni istante di tempo è univocamente determinato dato lo stato iniziale ed i parametri del modello. Inoltre, almeno in principio, le variabili, cioè le dimensioni delle popolazioni, variano con continuità nel tempo. Il fatto che esse vengano valutate in un insieme discreto di istanti di tempo non è in contraddizione con questo fatto. Chiaramente questa variabilità continua nel tempo è propria del modello usato, non della realtà che esso rappresenta. Avremmo potuto

[4]Il modello che noi abbiamo qui presentato può essere considerato la discretizzazione di una variante dell'originale modello di Volterra. In essa si è usato un intervallo di discretizzazione pari ad 1. I risultati ottenuti dipendono da questa scelta; torneremo in seguito sugli effetti della scelta dell'intervallo di discretizzazione nel caso di modelli di simulazione basati su sistemi di equazioni alle differenze finite.

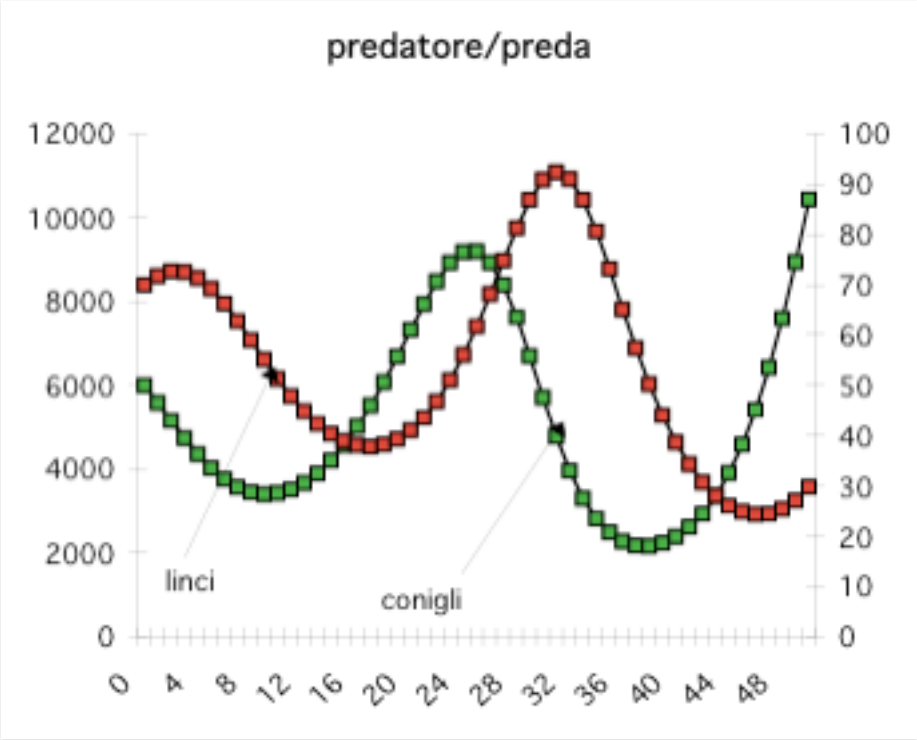

Figura 1.2. *Andamento delle popolazioni nel tempo*

	M_1	M_2
L_1	4	6
L_2	0	5
L_3	3	8
L_4	4	0
L_5	6	3

Tabella 1.1. *Tempi di lavorazione*

scegliere un modello diverso in cui fossero evidenziati gli eventi delle catture, delle morti e delle nascite. Un tale modello sarebbe però di difficile realizzazione in considerazione della dimensione del problema e non ci darebbe, ai fini dello studio dell'andamento della dimensione delle popolazioni, una informazione più accurata di quella fornita dal modello scelto.

1.2.2 Modello dell'officina

In un'officina ci siano 2 macchine M_1 e M_2, e all'inizio della giornata siano in attesa di essere eseguiti 5 lavori, L_1, L_2, L_3, L_4 e L_5. Nella tabella 1.1 sono indicati i tempi richiesti dai lavori sulle macchine (in decine di minuti). Uno zero indica che un lavoro non richiede quella macchina. I lavori che richiedono due macchine devono passare prima per M_1 e poi per M_2.

Supponiamo che si decida di eseguire i lavori assegnando a ciascuna macchina, quando essa si rende disponibile, il primo lavoro eseguibile, nell'ordine da 1 a 5. Se allo stesso istante più lavori sono eseguibili sulla stessa macchina si eseguirà quello con indice minore. Ci si chiede quale sia il tempo minimo necessario per completare tutti i lavori.

Gli eventi in cui si possono avere cambiamenti di stato nel sistema sono:

1. un lavoro si rende disponibile per una macchina;

2. una macchina inizia un lavoro;

3. una macchina termina un lavoro

La sequenza dei cambiamenti di stato nella simulazione è riportata nella tabella 1.2.

Tempo	M_1		M_2	
	Finisce	*Inizia*	*Finisce*	*Inizia*
0		L_1		L_2
4	L_1	L_3		
5			L_2	L_1
7	L_3	L_4		
11	L_4	L_5	L_1	L_3
17	L_5			
19			L_3	L_5
22			L_5	

Tabella 1.2. *Cambiamenti di stato del sistema*

Con la politica scelta sono quindi necessari 220 minuti per l'esecuzione di tutti i lavori.

Anche questo, come il precedente è un esempio di simulazione deterministica, ma al contrario di esso qui la simulazione è di tipo discreto; infatti i cambiamenti di stato del sistema avvengono solamente in alcuni istanti, e sono determinati dalle durate delle lavorazioni sulle macchine. Osserviamo che in questo esempio si è usata una tempificazione diversa da quella usata nell'esempio precedente. Invece di valutare lo stato del sistema ad intervalli regolari, il che avrebbe comportato un gran numero di valutazioni inutili, lo si è fatto solamente negli istanti in cui avviene effettivamente un cambiamento di stato (*tempificazione per eventi*).

1.2.3 Un problema di manutenzione

Un *server* ha due unità disco. Un guasto ad uno dei dischi comporta la perdita delle informazioni in esso contenute e la necessità di ricorrere alle copie di back-up con relativa interruzione del servizio e costo di intervento. La riparazione di un disco costa 150 *Euro*, mentre la sua revisione, se ancora funzionante, costa 50 *Euro*. Nella tabella 1.3 viene riportata la probabilità p_i di un guasto nell'i^{esimo} mese dopo una riparazione o una revisione. Abbiamo assunto che la probabilità che il disco funzioni più di 6 mesi senza guasti sia trascurabile. Con P_i sono state indicate le probabilità cumulate.

Attualmente i dischi vengono riparati ogni volta che si guastano, l'uno

i	p_i	P_i
1	0.05	0.05
2	0.15	0.20
3	0.20	0.40
4	0.30	0.70
5	0.20	0.90
6	0.10	1

Tabella 1.3. *Probabilità delle durata di funzionamento dei dischi*

indipendentemente dall'altro. Si vuole valutare la convenienza di effettuare una politica congiunta: ogni qualvolta un disco si guasta, esso viene riparato e l'altro, se ancora funzionante, viene revisionato.

Per ricorrere ad una simulazione dobbiamo essere in grado di generare gli "eventi guasto" in modo che rispettino le probabilità date. A questo scopo basta generare numeri (pseudo)casuali nell'intervallo [0, 1) e generare l'evento "guasto nel mese i" se il numero generato appartiene all'intervallo $[P_{i-1},\ P_i)$, avendo posto $P_0 = 0$. Nella Tabella 1.4 sono indicati i risultati della simulazione per le due politiche in esame. Con t e T vengono indicati rispettivamente l'intervallo tra due interventi successivi ed il tempo simulato.

Num. Casuali		**Riparaz. separate**				**Ripar. congiunte**	
		t		**T**		**t**	**T**
Disco A	Disco B	Disco A	Disco B	Disco A	Disco B		
0.71	0.37	5	3	5	3	3	3
0.58	0.34	4	3	9	6	3	6
0.21	0.89	3	5	12	11	3	9
0.81	0.08	5	2	17	13	2	11
0.94	0.67	6	4	23	17	4	15
0.58	0.19	4	2	27	19	2	17
0.19	0.22	2	3	29	22	2	19
0.37	0.33	3	3	32	25	3	22
0.88	0.56	5	4	37	29	4	26
0.67	0.77	4	5	41	34	4	30
0.36	0.27	3	3	44	37	3	33
0.67	0.90	4	5	48	42	4	37
0.30	0.08	3	2	51	44	2	39
0.50	0.84	4	5		49	4	43
0.56	0.27	4	3		52	3	46
0.21	0.35	3	3			3	49
0.11	0.16	2	2			2	51

Tabella 1.4. *Riparazione dischi: risultato della simulazione*

Capitolo 2

Simulazione discreta

In questo capitolo presentiamo i concetti base per la costruzione di modelli di simulazione e i principali approcci alla simulazione discreta. La trattazione è in parte basata sul testo di Pidd (1998), al quale rimandiamo per approfondimenti.

2.1 Il sistema da modellare

Come abbiamo già visto nel precedente capitolo, in generale, in un progetto di simulazione il primo passo consiste nell'analizzare la realtà nella quale il problema che si vuole affrontare e risolvere. Questa realtà viene rappresentata come un *sistema*. Un sistema è un insieme di elementi e di relazioni far essi. L'aspetto più rilevante di un sistema è che esso è qualcosa di più dell'insieme delle sue parti. Cioè il suo comportamento non può essere appieno conosciuto solamente analizzando separatamente il comportamento delle parti. È la struttura e la natura delle relazioni fra le parti che determina, secondo modalità a volte 'apparentemente' imprevedibili, il comportamento complessivo del sistema.

Come abbiamo evidenziato, il sistema non è la realtà ma ne è una rappresentazione, in ultima analisi è una costruzione mentale. Sta a noi scegliere il livello di dettaglio del sistema che vogliamo definire. Questo livello dipenderà dagli obiettivi del nostro studio, cioè dal tipo di problema che vogliamo risolvere. Ad esempio, se ci proponiamo di studiare il traffico di veicoli privati allo scopo di definire politiche per la riduzione della congestione e dei tempi di trasporto, non è rilevante né il colore né l'anno di produzione delle

auto. Quest'ultimo dato potrebbe essere invece interessante se fossimo interessati all'inquinamento prodotto dal traffico; in questo caso infatti l'anno di produzione ci può fornire indicazioni sulle emissioni del motore.

In generale una buona norma è che il livello di dettaglio sia il minimo, compatibilmente con le risposte che noi vogliamo ottenere dalla simulazione. Usualmente si parte da un modello semplice, con pochi elementi, e poi lo si raffina, aumentando man mano il grado di dettaglio, fino ad arrivare ad un modello che sia soddisfacente rispetto ai nostri obiettivi.

Un elemento molto importante è il tipo di notazioni formali usato per rappresentare e descrivere il sistema. Innanzitutto le notazioni devono essere sufficientemente semplici e chiare da potere essere comprese facilmente da tutti coloro che sono coinvolti sia esplicitamente che implicitamente nel processo di modellazione, cioè da quelli che spesso vengono chiamati *portatori di interessi*. Si tratta non solo di chi avrà il compito di prendere le decisioni finali e di farle eseguire, ma anche di dovrà poi concretamente metterle in pratica, e di chi, operando nella realtà sotto esame, detiene sia conoscenze essenziali per la definizione e modellazione del sistema che i dati necessari per la sperimentazione. Un modello facilmente comprensibile permette una verifica della sua correttezza da parte di tutti e quindi anche una maggiore efficienza ed efficacia del processo di modellazione. Oltre che garantire chiarezza e semplicità, le notazioni devono essere anche sufficientemente rigorose da permettere una veloce ed efficiente implementazione del modello per mezzo del software scelto.

In questo paragrafo descriveremo gli elementi formali che utilizzeremo per caratterizzare un sistema, introducendo anche le notazioni che verranno più frequentemente usate. L'attenzione sarà diretta al sistema fisico che si vuole modellare, agli elementi che in esso compaiono, alle relazioni tra tali elementi e alle attività che vi si svolgono. Successivamente riesamineremo il processo di modellazione avendo come obiettivo l'implementazione di un simulatore del sistema in esame.

I principali elementi che utilizzeremo per rappresentare un sistema sono le *entità* e le *operazioni*. Le prime descrivono ciò che caratterizza un sistema da un punto di vista statico, mentre le seconde permettono di descriverne l'evoluzione nel tempo e di evidenziare le relazioni fra le sue parti.

2.1.1 Entità

Gli oggetti base del sistema da modellare sono le *entità*. Si tratta di elementi del sistema che vengono considerati individualmente, e del cui stato si mantiene informazione nel corso della simulazione. Tipiche entità sono il paziente che si presenta all'accettazione di un ospedale, il pezzo che viene lavorato in una catena di montaggio, oppure l'aereo in attesa di atterrare a un aereoporto.

Nel seguito distingueremo fra tipi di entità, o *classi*, e entità individuali o *oggetti*[1].

Attraverso una classe viene definito un tipo astratto di entità con le sue proprietà. Le proprietà sono costituite da un insieme di *attributi* che caratterizzano tutte le entità che di quella classe sono istanziazioni. Ad esempio in un sistema costituito da uno sportello postale a cui si rivolgono clienti per richiedere alcuni tipi di servizi, possiamo definire la classe *cliente*, caratterizzata da attributi quali: tipo e quantità di servizio richiesto, ora di arrivo allo sportello, ... Un oggetto della classe sarà invece un particolare cliente, cioè ad esempio il cliente che si presenta allo sportello alle 11:32, con due bollettini di conto corrente da pagare. Pertanto ad un dato tipo di entità corrisponderà una sola classe ed un numero variabile, limitato o illimitato, di oggetti. Naturalmente la realtà può essere più complessa, ad esempio con più sottoclassi della stessa classe.

Gli oggetti possono a volte essere raggruppati in *insiemi*. Ad esempio in un pronto soccorso ospedaliero possiamo considerare l'insieme degli infermieri. Nel corso delle operazioni del sistema, ciascuna attività potrà richiedere uno o più infermieri; se il numero degli infermieri liberi non è sufficiente per lo svolgimento dell'attività, allora essa non potrà iniziare finché non si rendano liberi tanti infermieri quanti ne servono.

Alcune entità sono di tipo collettivo. Ad esempio un parcheggio con 100 posti macchina può essere rappresentato mediante la classe "posto macchina" con 100 istanze, cioè oggetti del tipo "posto macchina". Ma, a meno di situazioni molto particolari, non ci interesserà sapere se il singolo posto macchina è libero o occupato; ci basta invece sapere se ci sono posti liberi nel parcheggio e quanti ce ne sono. In questo caso il parcheggio viene considerato

[1]Così facendo prendiamo a prestito termini propri dell'informatica ed in particolare della programmazione ad oggetti. Questa scelta è stata fatta per accentuare il fatto che il tipo di modellazione che stiamo presentando è orientata alla implementazione del modello e quindi alla simulazione.

come una unica entità collettiva: si dice allora che è una *risorsa*. Ovviamente il decidere se una data entità sia o non sia da considerare una risorsa è una scelta che dipende da noi: al limite anche una entità individuale può essere considerata come una risorsa di cardinalità 1. In generale si tratta di una scelta che deve essere guidata di considerazioni legate all'economia globale del modello che si sta costruendo[2].

Le *entità* possono essere *permanenti* o *temporanee*, *attive* o *passive*. Ad esempio, nel caso di una coda a un botteghino teatrale, i clienti che arrivano si mettono in coda ed escono dal sistema una volta serviti, possono essere considerati come entità temporanee e passive, mentre il botteghino svolge il ruolo di entità permanente ed attiva. Comunque la distinzione è a volte arbitraria, dipendendo dalla percezione che ha del sistema colui che costruisce il modello e dalle sue scelte.

Infine una entità può essere in uno stato di attesa oppure può essere occupata nello svolgimento di una qualche attività. Uno *stato* del sistema è definito come l'insieme degli stati delle entità che lo costituiscono.

2.1.2 Operazioni

Con il termine *operazioni* indichiamo tutto ciò che attiene alla dinamica del sistema e che fa sì che esso evolva passando da uno stato all'altro. In particolare parleremo di *eventi*, *attività* e *processi*.

Eventi. Nei testi di Simulazione molto spesso con il termine *evento* si intende l'istante di tempo in cui avviene un cambiamento di stato del sistema. Qui lo intenderemo piuttosto come un fatto che produce un cambiamento di stato nel sistema. Ad esempio la fine del servizio di un cliente ad uno sportello fa sì che il cliente esca dal sistema e che l'impiegato allo sportello passi dallo stato di occupato a quello di libero.

Attività. Ognuna delle entità/oggetti presenti nel nostro sistema in ogni istante di tempo svolge delle attività, avendo considerato anche l'attesa come una particolare attività. Una attività è qualcosa che si svolge fra due eventi

[2]Nella costruzione di un modello di simulazione è bene farsi guidare dal principio enunciato dal filosofo medievale Guglielmo da Ockham, meglio noto come il *rasoio di Ockham*: "Entia non sunt multiplicanda praeter necessitatem" (gli enti non vanno moltiplicati al di là di ciò che è strettamente necessario).

e corrisponde ad uno stato di una o più entità. Ad esempio l'attività di servizio ad uno sportello si svolge tra l'evento di 'inizio servizio' e quello di 'fine servizio', e coinvolge sia l'impiegato che opera allo sportello che il cliente che viene servito. Il primo svolge un ruolo attivo in questa attività, mentre il secondo svolge un ruolo passivo.

Processi. I processi sono delle sequenze o cicli predefiniti di attività (e quindi di eventi). Ad esempio il passeggero di un aereo passa attraverso la seguente sequenza di attività: arrivo al *check-in*, attesa in coda, *check-in*, arrivo al controllo di sicurezza, attesa in coda, controllo di sicurezza, attesa per la chiamata di imbarco, controllo alla porta d'imbarco, imbarco sull'aereo.

2.1.3 Cicli delle attività

Un metodo molto usato per descrivere le transizioni da uno stato all'altro in un sistema è costituito dai cosiddetti *cicli di attività*. In un ciclo di attività gli stati vengono rappresentati come nodi in un grafo ed i passaggi da uno stato all'altro come archi orientati. Si distingue fra stati attivi e stati passivi utilizzando due tipi diversi di rappresentazione grafica dei nodi: quadrati per rappresentare gli stati attivi e cerchi per rappresentare quelli passivi (figura 2.1). Spesso i due tipi di stati si alternano nello stesso ciclo: ad uno stato attivo segue uno passivo e viceversa.

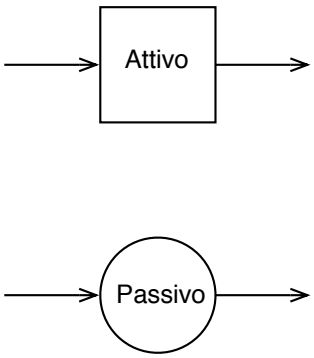

Figura 2.1. *Rappresentazione degli stati*

Riprendiamo l'esempio del servizio allo sportello già utilizzato precedentemente. In questo esempio, che consideriamo nella sua versione più semplice, si hanno due classi, la classe 'sportello' e la classe 'cliente'. La prima

supponiamo abbia una sola istanziazione, avendo supposto l'esistenza di un solo sportello, mentre la seconda ha un numero a priori illimitato di istanziazioni. Gli oggetti della classe cliente sono entità temporanee nel nostro sistema; infatti entrano nel sistema quando arrivano nell'ufficio e scompaiono dal sistema dopo essere state servite. L'entità sportello è invece una entità permanente.

I cicli delle attività per le due classi sono quelli indicati nelle figure 2.2 e 2.3. Osserviamo come il ciclo delle attività ci permetta di evidenziare bene non solo gli stati veri e propri, ma anche le attività ad essi associate e gli eventi che segnano il passaggio dall'uno all'altro.

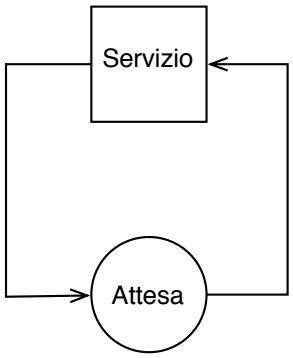

Figura 2.2. *Il ciclo degli stati dell'addetto allo sportello*

Combinando i cicli delle attività delle diverse classi/entità presenti possiamo costruire il ciclo delle attività del sistema, che riportiamo nella figura 2.4.

In questo esempio abbiamo un solo tipo di cliente e questo rende particolarmente semplice il ciclo delle attività. Un esempio un po' più complesso è quello del 'botteghino del teatro' utilizzato da Pidd (1998). In questo esempio il sistema è costituito dal botteghino di un teatro in cui vengono venduti i biglietti. L'addetto al botteghino (*servente*), oltre a vendere i biglietti ai clienti deve anche rispondere alle telefonate fornendo le informazioni richieste. Si hanno pertanto due code, una fisica di *clienti*, davanti allo sportello, ed una, virtuale, formata da *chiamate* in attesa (il sistema telefonico si suppone abbastanza sofisticato da consentirlo). Entrambe le code vengono processate con una politica di tipo FIFO e i clienti hanno sempre la precedenza sulle chiamate (mai rischiare di perdere un cliente pagante!). In questo caso abbiamo tre classi di entità: cliente (un numero illimitato di entità temporanee

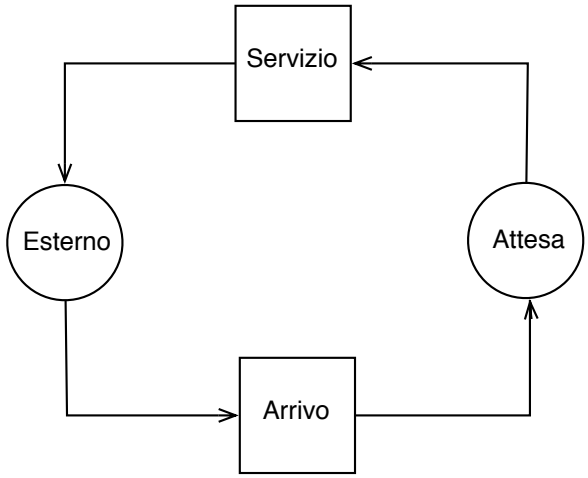

Figura 2.3. *Il ciclo degli stati del cliente*

e passive), chiamata (un numero illimitato di entità temporanee e passive), botteghino (una sola entità, permanente e attiva). I cicli delle attività per le classi 'cliente' e 'chiamata' sono identici a quello di figura 2.3, mentre il ciclo delle attività della classe 'botteghino' è quello riportato nella figura 2.5. Riprenderemo nel seguito questo esempio più in dettaglio.

2.2 UML: un linguaggio di modellazione

Nel seguito, per rappresentare graficamente i modelli di simulazione useremo la grafica e la sintassi del linguaggio per modellazione UML. Si tratta di un linguaggio che pur essendo nato in altro contesto, quello della modellazione di sistemi software, tuttavia per la sua versatilità si presta molto bene anche alla costruzione di modelli di simulazione. Ed in effetti già da alcuni anni ha fatto la sua comparsa anche in questo contesto. Per una introduzione al linguaggio UML rimandiamo al testo di Fowler (2000).

Le *classi* e gli *oggetti* che le istanziano vengono rappresentati per mezzo di rettangoli divisi in tre sezioni che contengono il *nome* della classe/entità, gli *attributi* e le *operazioni*. Per gli attributi viene spesso anche indicato il tipo di variabile che li definisce (intero, booleano, ...), mentre per le operazioni viene indicata la variabile il cui valore l'operazione calcola e il tipo di tale variabile. Una o entrambe le sezioni relative agli attributi ed alle ope-

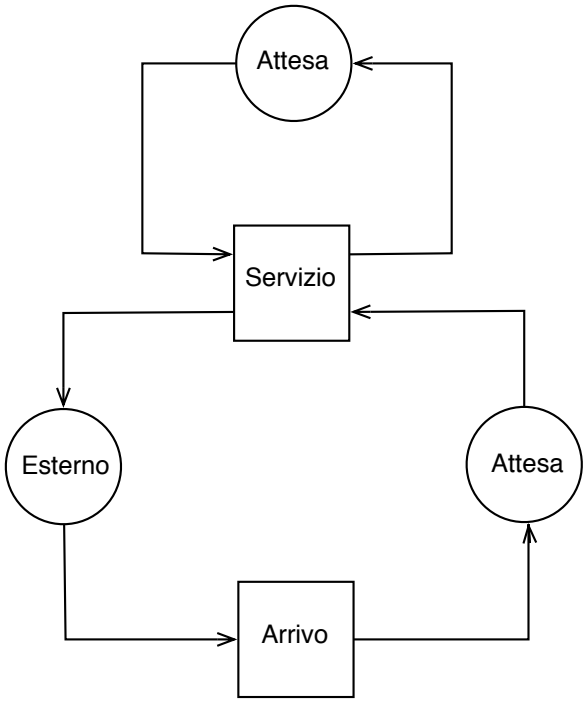

Figura 2.4. *Il ciclo degli stati del sistema*

razioni possono anche mancare. Nella figura 2.6 sono rappresentate le classi dell'esempio dello sportello. Rispetto a quanto visto precedentemente abbiamo aggiunto la classe 'coda' che ha una ben precisa relazione con la classe cliente: gli oggetti della prima sono aggregazioni di oggetti della seconda. Questo fatto è graficamente rappresentato dal particolare simbolo usato per connettere le due classi. Nella classe *Impiegato* abbiamo indicato come attributo una variabile booleana che è vera se l'impiegato è libero (in attesa) e falsa altrimenti[3], ed una operazione che è quella di servire il cliente e che programma la fine del servizio al tempo $T_C + t$, dove T_C è il tempo corrente nella simulazione e t è il tempo di servizio del cliente che viene servito. Nella classe coda sono stati indicati un attributo, la lunghezza, variabile di tipo intero, e due operazioni, $Estrai()$ e $Inserisci()$. L'operazione $Estrai()$ forni-

[3]Seguendo un uso diffuso nell'area della programmazione, abbiamo indicato la variabile con la notazione *nome.tipo*, dove il primo è il nome della variabile, mentre il secondo è il tipo della variabile (un tempo nel nostro caso).

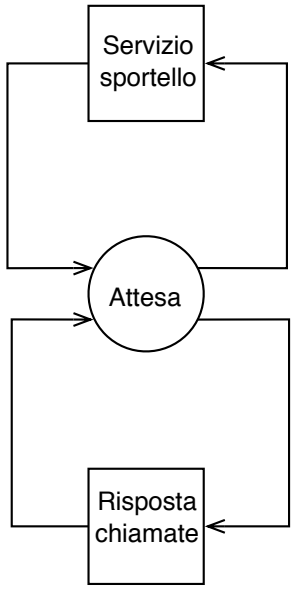

Figura 2.5. *Il ciclo degli stati del botteghino del teatro*

sce il suo primo elemento, che nel nostro caso sarà un oggetto di tipo Cliente, e lo cancella dalla coda; analogamente l'operazione $Inserisci(Q)$ inserisce un oggetto di tipo Cliente nella coda. Un altro attributo importante potrebbe essere il tipo di gestione degli oggetti nella coda: *FIFO*, *LIFO*, ... La classe Cliente ha come attributo il tempo di servizio, t.

Gli oggetti sono rappresentati con simboli simili a quelli usati per rappresentare le classi di cui sono istanziazioni. Naturalmente qui potremo avere più copie dello stesso oggetto in dipendenza del numero di istanze presenti nel sistema. Ad esempio se nel sistema avessimo due sportelli, nel diagramma avremo due copie dell'oggetto sportello. Nella figura 2.7 è rappresentato il diagramma degli oggetti presenti nel sistema costituito da un solo sportello. Ogni oggetto è caratterizzato dal suo nome e dalla classe alla quale appartiene, con la notazione "*Nome*: *Classe*"[4]. Come per le classi il rettangolo che li rappresenta può contenere altre informazioni, quali gli attributi e le operazioni che gli oggetti possono compiere.

[4]Per gli oggetti della classe *Cliente* abbiamo utilizzato i nomi $C1, C2, \ldots$, e ne abbiamo indicati alcuni, l'uno sovrapposto all'altro, per indicare simbolicamente che il loro numero non è a priori limitato.

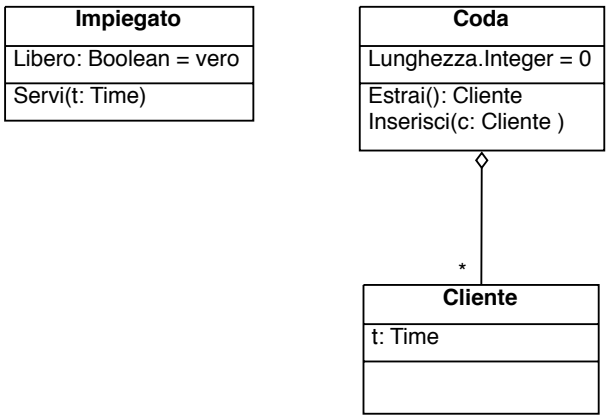

Figura 2.6. *Diagramma delle classi*

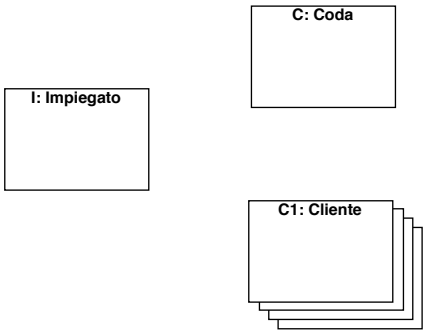

Figura 2.7. *Diagramma degli oggetti*

Gli stati vengono rappresentati con i simboli di figura 2.8, dove sono indicati anche due stati speciali, quello iniziale e quello finale, ed il simbolo usato per rappresentare le transizioni. Le transizioni hanno una etichetta costituita da tre sezioni (non tutte necessariamente presenti): *evento*, *condizione* ed *azione*. L'evento è il fatto che produce la transizione, la condizione, se verificata, è quella che fa sì che al verificarsi dell'evento effettivamente si abbia la transizione, mentre l'azione è l'oggetto della transizione stessa. Ad esempio l'evento 'fine del servizio', se la condizione 'coda non vuota' è verificata produce l'azione di chiamare il primo cliente in coda e di servirlo. Nel rettangolo arrotondato che rappresenta uno stato, nella zona sotto il nome dello stato possono essere indicate le diverse azioni che vengono effettuate

nello stato.

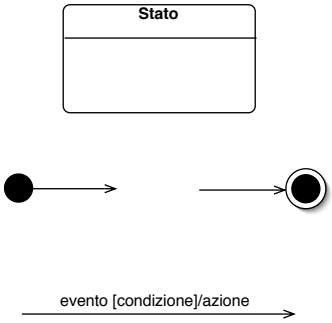

Figura 2.8. *Stati e transizioni*

Illustreremo ora l'uso del linguaggio UML attraverso la modellazione del problema del botteghino teatrale presentato precedentemente. Il diagramma delle entità è quello indicato in figura 2.9, dove sono indicate le classi e gli oggetti che le istanziano. Attributo della classe cliente è il tipo, che è una variabile booleana: può assumere i valori p (cliente che si presenta di persona allo sportello) e t (cliente che chiama al telefono)[5]. Osserviamo la presenza di due code, una per ciascun tipo di cliente: Qp è la coda dei clienti che arrivano di persona al botteghino, mentre Qt è la coda dei clienti che telefonano.

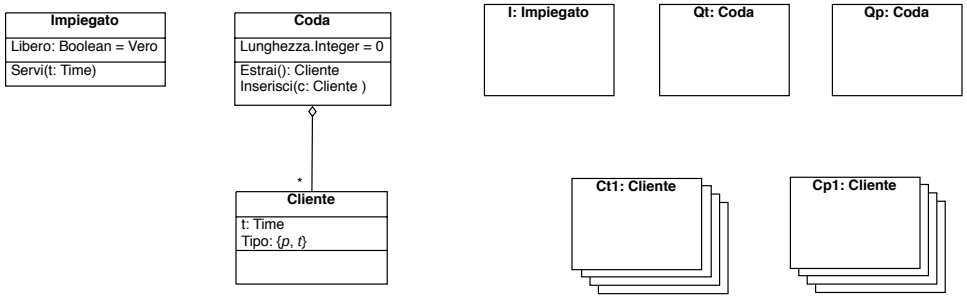

Figura 2.9. *Classi ed oggetti nell'esempio del botteghino teatrale*

In figura 2.10 è riportato il diagramma degli stati per quel che riguarda l'impiegato addetto allo sportello. Quando entra in servizio si trova nello

[5]Qui abbiamo scelto di avere una sola classe di clienti, distinguendo per mezzo di un attributo i due tipi di clienti. Ovviamente avremmo potuto in alternativa rappresentare i clienti per mezzo di una classe con due sottoclassi.

stato di attesa, dal quale esce per iniziare un servizio non appena arrivi il primo cliente fisico oppure la prima telefonata. In questo caso l'evento che fa partire la transizione è l'arrivo del cliente e l'azione è l'inizio dell'atto di servire il cliente. Osserviamo che se arriva un cliente fisico, esso deve essere servito subito (i clienti successivi andranno in coda), mentre se arriva una chiamata, essa verrà servita solo se allo stesso tempo non è arrivato un cliente fisico[6]. Nel caso che l'addetto stia servendo un cliente, non appena il servizio finisce (evento $FineServizio$) verrà controllata la lunghezza della coda Qp, e se non vuota (condizione) si inizierà a servire il primo cliente in coda $Qp.Estrai()$. L'operazione $InizioServizio(Qp.Estrai())$ esegue anche la cancellazione di $Qp.Estrai()$ da Qp. Se poi Qp risulta vuota, si controlla la coda Qt e, se non è vuota, si risponde alla prima chiamata in coda. Anche qui il cliente di cui si inizia il servizio viene cancellato dalla coda.

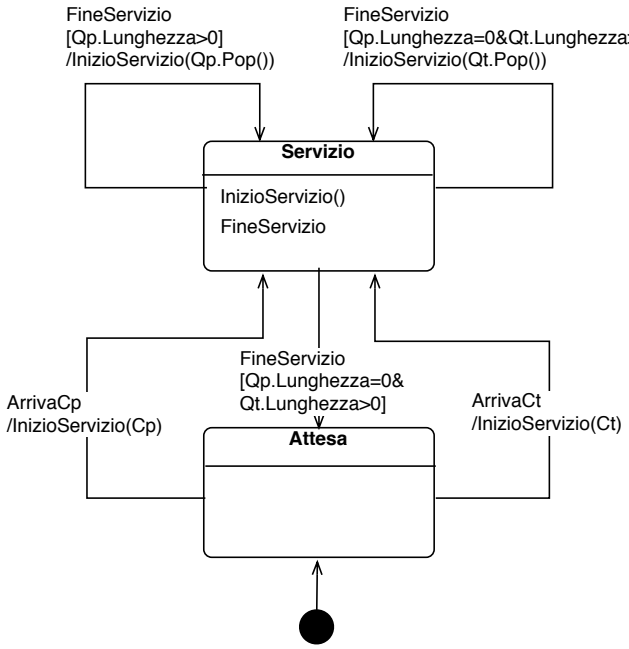

Figura 2.10. *Il Diagramma degli stati dell'impiegato nell'esempio del botteghino teatrale*

Il diagramma di figura 2.10 non contiene una descrizione completa di ciò

[6]Questa relazione di precedenza non appare esplicitamente nel diagramma di figura 2.10

che accade nel nostro sistema. È infatti necessario anche descrivere gli stati dei clienti, cosa che viene fatta in figura 2.11, con riferimento ai clienti fisici; per gli altri il diagramma è nella sua struttura identico. Nel successivo paragrafo vengono riportati alcuni esempi svolti con un certo livello di dettagliodettaglio.

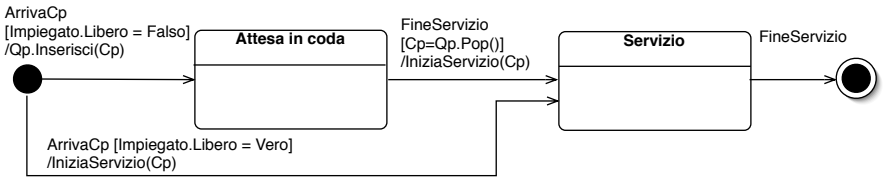

Figura 2.11. *Diagramma degli stati del cliente nell'esempio del botteghino teatrale*

In realtà si sarebbe potuto evitare di costruire un diagramma degli stati anche per i clienti, anche se forse si sarebbe perso un po' in leggibilità del modello. I clienti infatti in questo esempio sono entità passive ed una ragionevole scelta può essere quella di definire gli stati solamente per le entità attive. Se avessimo deciso di fare così naturalmente avremmo dovuto arricchire il diagramma degli stati per l'impiegato, ad esempio aggiungendo due ulteriori transizioni come in figura 2.12. Queste transizioni corrispondono all'arrivo dei clienti ed al loro inserimento in coda. Dal punto di vista formale è come se si desse all'impiegato anche il compito di gestire le code non solo per quel che riguarda le estrazioni ma anche gli inserimenti. In pratica si tratta di una operazione fittizia che richiede tempo nullo. La nostra scelta qui, anche se a prezzo di una certa ridondanza, è stata guidata da esigenze di chiarezza e leggibilità del modello.

Nel seguito saranno descritti alcuni esempi per i quali verranno presentati i diagrammi degli stati.

2.2.1 Esempi

Servizio piccoli prestiti

Si consideri una banca in cui il servizio dei piccoli prestiti funziona secondo le seguenti modalità. Il cliente si rivolge all'ufficio apposito, dove un impiegato esamina la richiesta. Se questa rispetta un certo numero di criteri prefissati, viene approvata; altrimenti il cliente viene rinviato al funzionario responsabile

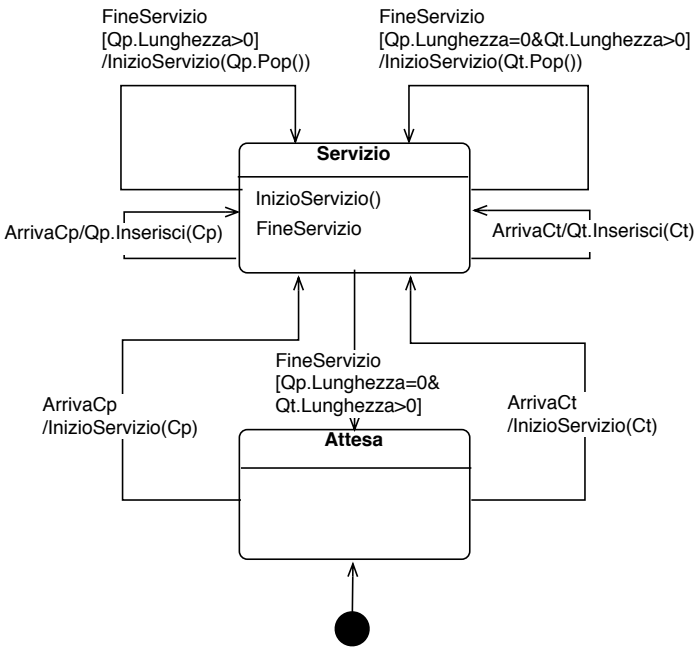

Figura 2.12. *Inclusione dell'operazione di inserimento in coda dei clienti nel Diagramma degli stati nell'esempio del botteghino teatrale*

del servizio crediti. Il funzionario riesamina la pratica, rivede col cliente l'importo del prestito e le eventuali condizioni, e quindi o approva la richiesta oppure la respinge definitivamente.

Ci sono in questo sistema tre classi di entità: *cliente*, *impiegato* e *funzionario*. Le entità cliente sono temporanee e passive: entrano nel sistema secondo una data legge di probabilità, passano attraverso una o due code e vengono servite dai relativi serventi, quindi escono dal sistema. Invece, le entità impiegato e funzionario sono permanenti ed attive. Nella figura 2.13 sono indicate le tre entità. Abbiamo indicato qui un solo attributo per ciascuna classe. Il cliente ha come attributo 'Prestito approvato', di tipo Booleano, che viene inizializzata a *Falso*. L'impiegato ed il funzionario hanno l'attributo, 'Disponibile' di tipo Booleano, che viene inizializzato al valore *Vero*, e che indica la disponibilità a servire un nuovo cliente; se il funzionario e l'impiegato stanno servendo un cliente, allora è 'Disponibile'=*Falso*.

Nella figura 2.14 è stato indicato il diagramma degli stati dei clienti. Una volta arrivato, il cliente entra in coda e vi aspetta. La transizione avviene

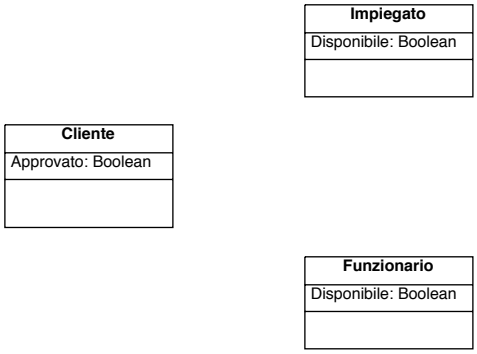

Figura 2.13. *Le entità*

quando si verifica l'evento che l'impiegato si è reso libero ed è soddisfatta la condizione che il cliente è il primo della coda. A questo punto il cliente cambia di stato ed inizia il servizio (primo esame). In questo stato vengono effettuate le seguenti azioni: appena il servizio inizia l'attributo 'Disponibile' dell'impiegato viene posto a $Falso$[7], viene poi svolto il servizio ed infine si riporta al valore $Vero$ l'attributo 'Disponibile'. Poi, se il prestito è stato approvato il cliente esce dal sistema, altrimenti passa alla seconda coda. I passaggi dalla seconda coda al secondo esame e poi l'uscita del cliente hanno un andamento analogo a quello relativo al primo esame.

Analogamente si possono costruire i diagrammi degli stati dell'impiegato e del funzionario, che sono fra loro identici. In figura 2.15 è riportato quello del funzionario[8].

Il centro prelievi

Il centro prelievi di un ospedale è aperto nei giorni feriali dalle 7,30 alle 10. Anche qui abbiamo una classe clienti. Un cliente appena arrivato ritira un numero da un'apposita macchinetta distributrice ed attende di essere chiamato allo sportello per l'accettazione, dove presenterà la richiesta di analisi effettuata dal suo medico curante. I clienti sono chiamati all'accettazione in

[7]In alcuni tipi di servizio questo è ciò che avviene quando la luce sullo sportello cambia di colore passando da verde a rossa.

[8]Qui abbiamo scelto di fare passare sempre nello stato Libero il funzionario dopo ogni fine di servizio. Avremmo potuto invece, analogamente a quanto fatto nell'esempio del botteghino teatrale, farlo rimanere nello stato di servizio fino allo svuotamento della coda.

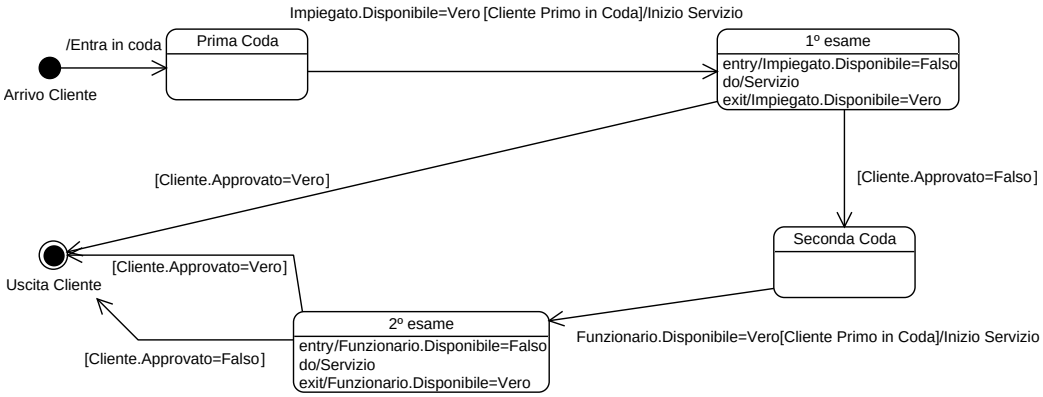

Figura 2.14. *Servizio Piccoli Prestiti: Diagramma degli stati del cliente*

ordine di numero crescente. Dopo l'accettazione il cliente si reca allo sportello per il pagamento del ticket per poi mettersi in fila davanti all'ambulatorio in cui si effettuano i prelievi; se è esente dal ticket, si recherà direttamente all'ambulatorio per il prelievo.

Le entità sono: i clienti, lo sportello per l'accettazione, lo sportello per il pagamento del ticket e l'ambulatorio per i prelievi. I clienti sono entità temporanee e passive, in numero illimitato, caratterizzate da due attributi, *Esente* e *Numero*. Il primo indica se il cliente è esente o no dal pagamento del ticket; si tratta di un attributo di tipo *Boolean*. Il secondo attributo è il numero che viene assegnato al cliente all'arrivo; questo numero servirà per la coda dell'accettazione e per quella dei prelievi. Le altre entità sono invece entità attive e permanenti; possono essere presenti in più istanze: ad esempio nell'ambulatorio possono esserci più infermiere/i che effettuano i prelievi. Il diagramma degli stati relativo ai clienti è riportato nella figura 2.16.

Il cliente, appena entrato nel sistema, riceve un numero, che corrisponde al suo ordine di arrivo, ed aspetta che il numero venga chiamato. A questo punto accede ad uno degli sportelli dell'accettazione, e interagisce con l'addetto, che, alla fine del servizio, gli dà un foglio con tutti i dati degli esami richiesti. Poi, se non è esente dal ticket, il cliente si mette in fila alla cassa e quando si libera una cassa (numero di casse libere, NP, maggiore di zero)

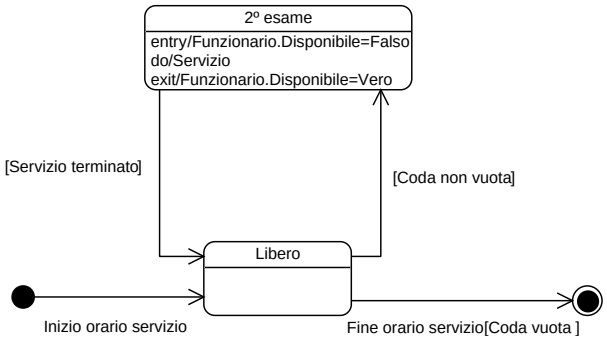

Figura 2.15. *Diagramma degli stati del funzionario*

accede al pagamento; dopo di che passa alla coda 3. Qui vale il suo numero iniziale, per cui quando il numero viene chiamato il cliente può accedere al servizio prelievo.

In generale ci saranno disponibili più sportelli per l'accettazione, più cassieri per il pagamento e più ambulatori per il prelievo. Nel diagramma non abbiamo evidenziato questo fatto. Lo avremmo potuto fare espandendo ciascuno degli stati relativi a servizi. Ad esempio, in figura 2.17 è indicata l'espansione dello stato Accettazione nell'ipotesi di due sportelli di servizio. Quando il numero viene chiamato viene anche indicato lo sportello relativo ed il cliente chiamato accederà a quello sportello.

Possiamo immaginare che ci sia una classe *Accettazione*, che ha come attributo la variabile LN, ultimo numero chiamato, e come istanze le entità *Sportello 1* e *Sportello 2*. Ciascuna delle due entità ha come attributo la variabile booleana *Disponibile*. Quando una di queste entità si rende disponibile chiama il numero successivo; la chiamata, che immaginiamo avvenga attraverso un display che porta oltre al numero chiamato anche quello dello sportello, viene letta dal cliente che si reca allo sportello.

Nella figura 2.18 abbiamo indicato il diagramma degli stati relativo all'accettazione. Gli sportelli dell'accettazione aprono alle 7:30 e chiudono dopo le 10:00 non appena è stato servito l'ultimo cliente arrivato. Durante l'apertura, non appena ci sia uno sportello disponibile, allora viene chiamato il numero successivo. Con LN, variabile che all'inizio ha valore 0, abbiamo indicato il numero dell'ultimo cliente chiamato al servizio, e con CMAX abbiamo indicato il numero dell'ultimo cliente arrivato. La condizione $LN \leq CMAX$ indica che il prossimo numero viene chiamato solo se ad esso corrisponde un

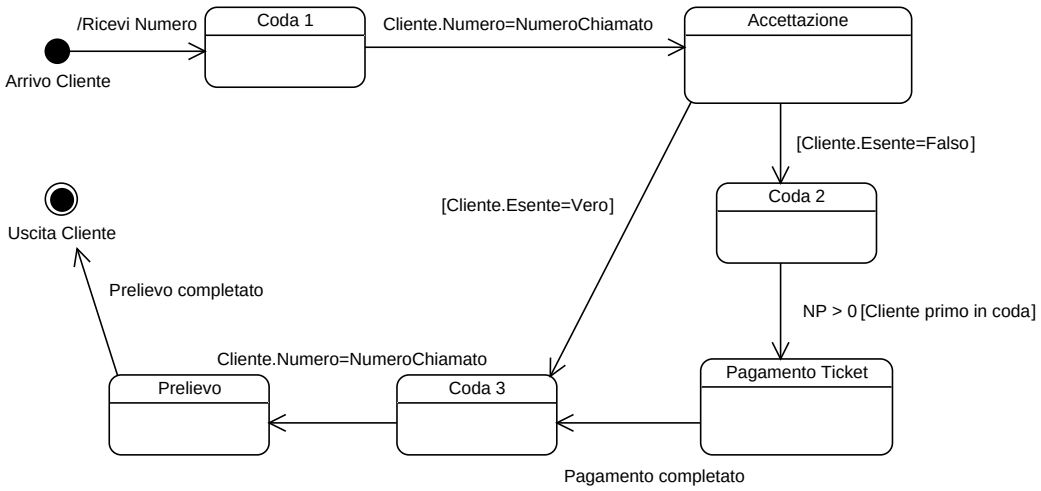

Figura 2.16. *Ciclo delle attività dei clienti del Centro Prelievi*

cliente in attesa. Dopo la chiamata inizia il servizio del cliente chiamato. Prima di iniziare le operazioni si rende non disponibile lo sportello[9], ed alla fine lo si rende di nuovo disponibile.

Un deposito per la distribuzione di merci

Una cooperativa di distribuzione ha un solo deposito, dove arrivano camion provenienti dai produttori con le merci richieste, e da cui partono i furgoni con le merci destinate ai diversi supermercati appartenenti alla cooperativa. Ci sono 2 banchine per lo scarico dai camion e 4 per il carico dei furgoni. Ci sono 5 squadre di 2 addetti ciascuna, che provvedono a scaricare, mettere in stock e caricare la merce sui furgoni. Esiste poi solo una via di accesso e di uscita, che consente il passaggio o di un camion (indipendentemente dal senso di marcia) o di due furgoni (purché in senso opposto di marcia); non c'è lo spazio perché passino contemporaneamente un camion ed un furgone.

Se l'obiettivo della simulazione è di studiare le attese dei camion e dei furgoni, possiamo pensare di considerare camion e furgoni come delle entità,

[9]Questa operazione potrebbe in pratica corrispondere al fatto che il display relativo allo sportello su cui era comparso il numero chiamato smetta di lampeggiare.

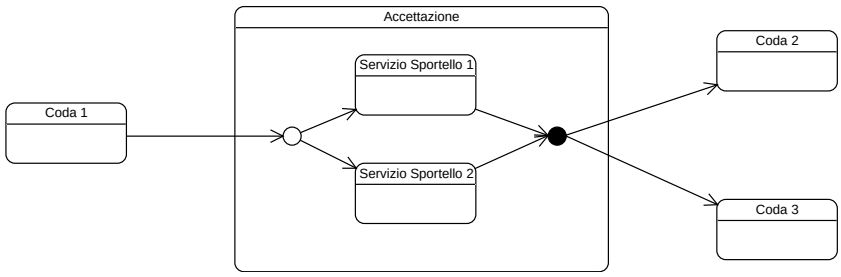

Figura 2.17. *Espansione dello stato Accettazione*

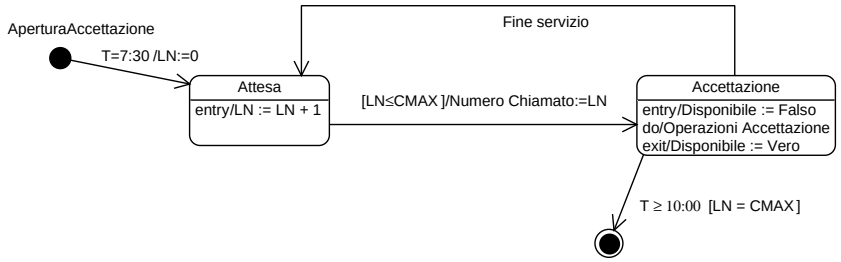

Figura 2.18. *Diagramma degli stati dei servizio Accettazione*

mentre le banchine di carico e scarico, la via di accesso e le squadre come risorse.

I cicli delle attività per i camion è riportato nella figura 2.19, dove abbiamo indicato con NBc e NS rispettivamente il numero delle banchine disponibili per i camion e quello delle squadre disponibili; all'inizio è NB = 2 e NS = 5. Abbiamo assunto che il camion che ha finito di scaricare rimane in attesa alla banchina fino a che l'ingresso non risulti libero e quindi possa uscire.

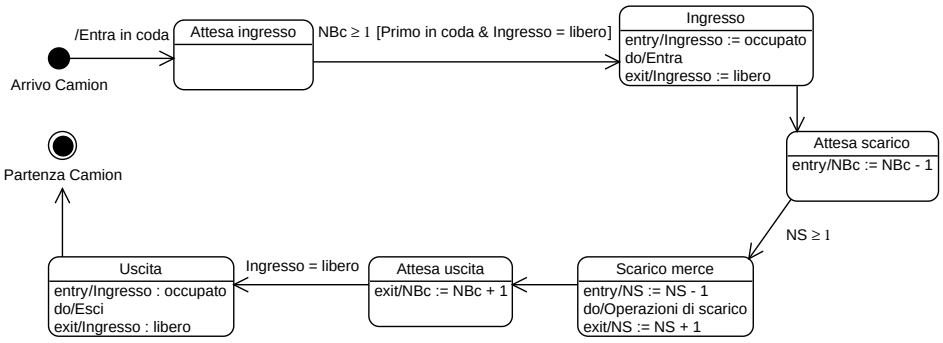

Figura 2.19. *Diagramma degli stati dei camion*

2.3 Approcci alla simulazione

In generale in un programma di simulazione sono sempre presenti i seguenti elementi:

- Controllore
- Tempo di simulazione
- Generatore dei dati di input
- Entità
- Eventi
- Attività
- Stati
- Processi

Il controllore è la componente del sistema che gestisce la sequenza degli eventi e l'evolversi dello stato del sistema nel tempo. In particolare il controllore ha il compito di fare avanzare il tempo di simulazione. Per quanto questa funzione sia presente in ogni modello, tuttavia il modo in cui essa viene svolta

può essere diverso. A diversi approcci alla simulazione corrispondono diversi modi di implementare il controllore. Un processo è una sequenza di stati e di transizioni da uno stato al successivo.

I diversi approcci corrispondono alla diversa enfasi che può essere data agli elementi. In generale una simulazione può essere realizzata a partire dagli eventi, dalle attività, oppure dai processi. Ciascuno di questi tre approcci ha svantaggi e vantaggi. Nel seguito li presenteremo più in dettaglio, servendoci di alcuni degli esempi visti precedentemente, e discuteremo i relativi meriti.

2.3.1 Simulazione per eventi

L'approccio basato sugli eventi ha una notevole importanza storica: i principali linguaggi orientati alla simulazione lo inglobano, anche se in genere come una delle opzioni. Permette di costruire programmi di simulazione molto compatti ed efficienti, ma anche proprio per questo più costosi dal punto di vista della manutenzione e degli aggiornamenti.

Per modellare un simulatore per eventi abbiamo bisogno di introdurre delle nuove classi, che sono indicate nella figura 2.20.

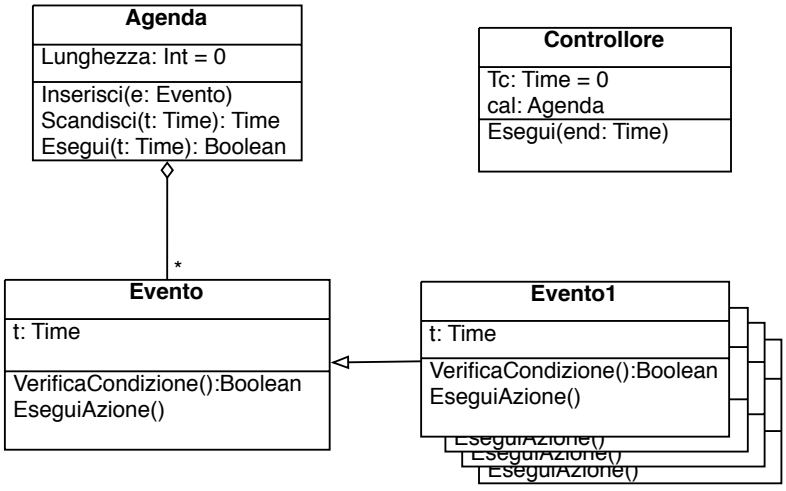

Figura 2.20. *Diagramma delle classi in un simulatore per eventi*

Oltre al controllore abbiamo la classe degli eventi (*Evento*), le diverse entità della classe (*Evento*1, *Evento*2, ...) ed il calendario degli eventi che devono avvenire (*Agenda*). La classe *Agenda* è caratterizzata da un

attributo, *Lunghezza*, che fornisce il numero di eventi in essa contenuti, e da tre operazioni: *inserisci*, che inserisce un nuovo evento; *scandisci*, che opera una scansione della lista al fine di determinare il tempo del prossimo evento; *esegui*, che, per ognuno degli eventi che avvengono al tempo corrente, ne controlla le condizioni e, se verificate, ne esegue le azioni.

Nel diagramma di figura 2.21 riportiamo la sequenza delle operazioni che effettua il controllore.

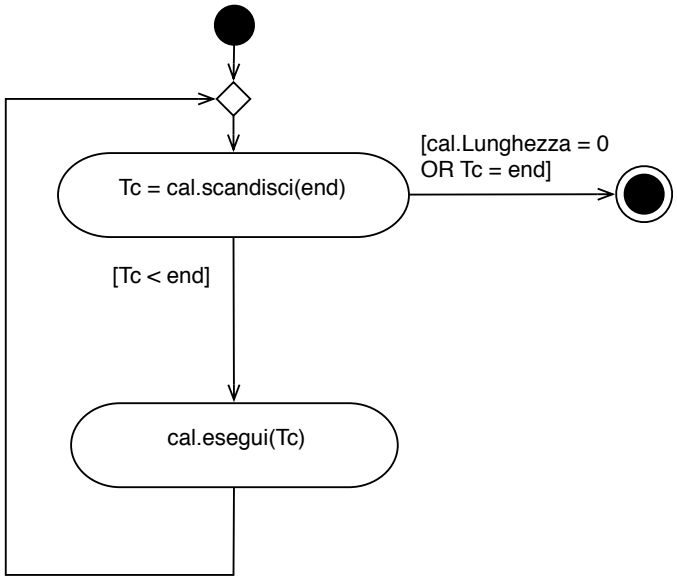

Figura 2.21. *Diagramma degli stati in un simulatore per eventi*

Innanzitutto il controllore fa una scansione del calendario degli eventi *cal*in modo da determinare il nuovo tempo corrente, T_C, che viene posto al valore del tempo in cui è previsto il primo degli eventi. Se questo tempo risulta maggiore o uguale al tempo massimo, allora a T_C viene attribuito il valore *end* e la simulazione termina. Invece se il tempo corrente è inferiore al tempo massimo della simulazione, allora vengono eseguite le azioni relative agli eventi di *cal* le cui condizioni sono verificate. Il calcolo del tempo corrente è ciò che fa avanzare il tempo di simulazione.

Nel caso del botteghino di teatro abbiamo tre tipi di eventi: *ArrivaCp*, *ArrivaCt*, *FineServizio*. Tutti e tre questi eventi agiscono, producendo azioni, sia sull'entità *sportello* che sulle entità di tipo *cliente*, anche se lo fanno seguendo un ordine prestabilito. In particolare *ArrivaCp* agisce prima

di *ArrivaCt*, nel caso i due eventi si verifichino nello stesso istante di tempo. Questo garantisce che un cliente fisico che arriva allo stesso tempo di una telefonata venga servito prima.

Esaminiamo ora più in dettaglio gli effetti dei tre eventi.

- *ArrivaCp*. Se L'impiegato è libero, allora l'evento fa iniziare il servizio del cliente arrivato: l'impiegato passa allo stato di servizio, e lo stesso accade per il cliente che non passa attraverso la coda.

- *ArrivaCt*. Se L'impiegato è libero, e non è arrivato contemporaneamente un cliente fisico, allora l'evento fa iniziare il servizio del cliente arrivato: l'impiegato passa allo stato di servizio, e lo stesso accade per il cliente che non passa attraverso la coda.

- *FineServizio*. Se l'impiegato non è libero, rimane nello stato di servizio ed inizia a servire il primo cliente della coda fisica, se ce ne sono, altrimenti risponde alla prima telefonata in coda. Se infine entrambe le due code sono vuote, l'impiegato si mette in attesa, cambiando così di stato. Contemporaneamente il cliente appena servito esce dal sistema, e il primo cliente in coda (nell'ordine fisica e telefonica), se esiste, passa dallo stato di attesa a quello di servizio.

2.3.2 Simulazione per attività

In questo approccio, a partire dal diagramma degli stati, si decompongono le attività svolte nel sistema in attività elementari. Queste attività corrispondono agli eventi, cioè a quei fatti o azioni che portano ai cambiamenti di stato del sistema e che quindi danno origine alle transizioni. Il ruolo del controllore sarà quello di gestire la lista delle attività individuate in modo che vengano eseguite. Ogni attività è caratterizzata dalla condizione che la fa avvenire e dalle azioni che vengono corrispondentemente fatte, e viene rappresentato per mezzo di una tabella in cui sono riportate le informazioni rilevanti.

Utilizzando anche qui la notazione UML, in figura 2.22 abbiamo indicato le principali classi ed entità presenti in un simulatore per attività.

Oltre al controllore abbiamo la classe delle attività (*Attività*), le diverse entità della classe (*Attività*1, *Attività*2, ...) e la lista delle attività (*ListaAttività*). La classe delle attività è caratterizzata da due attributi: il tempo al quale l'attività deve essere realizzata e la condizione che la fa realizzare; uno solo dei due attributi sarà sarà presente in una data attività. La lista

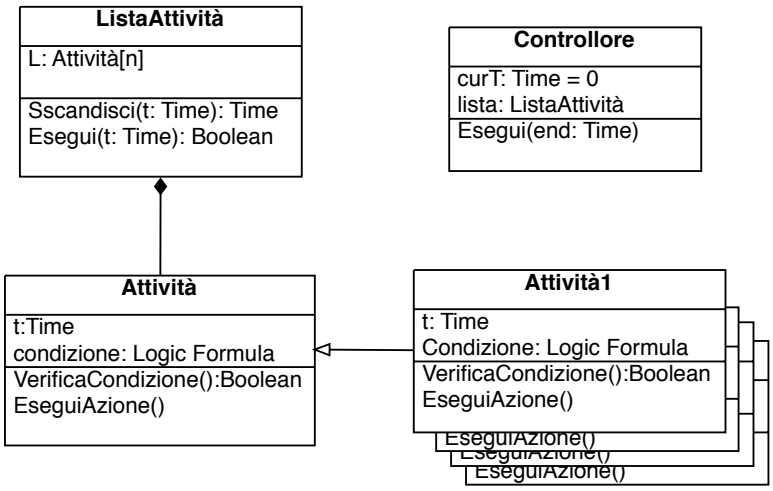

Figura 2.22. *Diagramma delle classi in un simulatore per attività*

può essere realizzata in diversi modi: il modo più semplice è costituito da un vettore. Alla lista, che può essere vista come un calendario delle attività, sono associate due operazioni: *scandisci*, che opera una scansione della lista al fine di determinare il tempo della prossima o delle prossime attività e*esegui*, che esegue le attività che vanno eseguite al tempo t e fornisce in output una variabile booleana che è vera se tutte le attività sono state eseguite.

Nel diagramma di figura 2.23 riportiamo la sequenza delle operazioni che effettua il controllore.

Innanzitutto il controllore fa una scansione della lista di attività in modo da determinare il nuovo tempo corrente, T_C, che viene posto al valore del tempo in cui è prevista la prima delle attività da eseguire. Se questo tempo risulta maggiore o uguale al tempo massimo, allora a T_C viene attribuito il valore *end* e la simulazione termina. Invece se il tempo corrente è inferiore al tempo massimo della simulazione, allora vengono eseguite tutte le attività della lista le cui condizioni sono verificate (ciclo interno). Per fare questa operazione è necessaria una nuova scansione della lista, che però va eseguita più volte: infatti l'esecuzione di una attività può rendere soddisfatte le condizioni per un'altra fra quelle già esaminate. L'operazione è completata quando una intera scansione della lista non porta alla esecuzione di nessuna attività. Quando tutte le attività da eseguire sono state completate il controllore va a calcolare il nuovo tempo corrente ed inizia un nuovo ciclo. Il

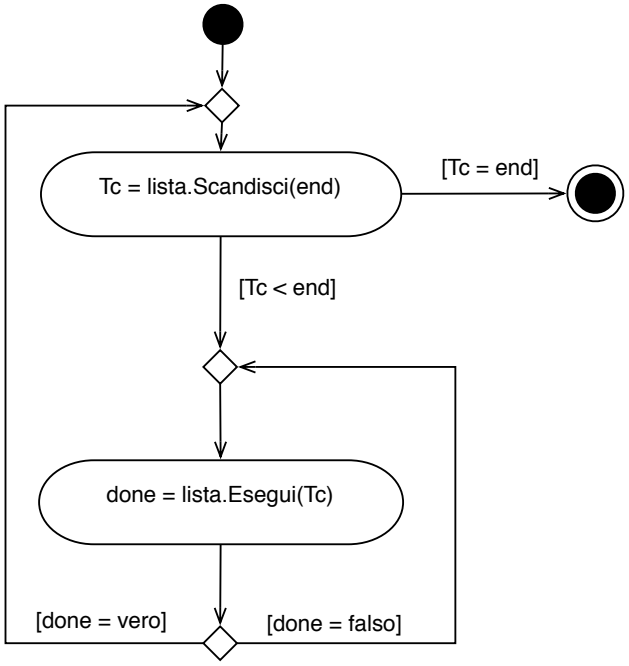

Figura 2.23. *Diagramma delle operazioni effettuate dal controllore in un simulatore per attività*

calcolo del tempo corrente è ciò che fa avanzare il tempo di simulazione.

Consideriamo l'esempio del botteghino di teatro e cerchiamo di individuare quali sono le attività. Possiamo definire 6 attività:

- $a[1]$ Inizio servizio allo sportello
- $a[2]$ Inizio servizio al telefono
- $a[3]$ Fine servizio allo sportello
- $a[4]$ Fine servizio al telefono
- $a[5]$ Arrivo cliente allo sportello
- $a[6]$ Arrivo cliente al telefono

Osserviamo come queste attività siano attività elementari e che non coincidano con le operazioni descritte nelle figure 2.10 e 2.11. Esse fanno piuttosto riferimento all'inizio ed alla fine degli stati attivi. Le descriviamo più

in dettaglio di seguito, indicando per ogni attività gli attributi (il tempo e condizione).

La prima attività ha la struttura seguente:

Attività	$a[1]$: Inizio servizio allo sportello
Attributi	- $I.Libero = vero$ & $Qp.lunghezza > 0$
Azioni	$a[3].t := T_C + Qp.Estrai().t$; $I.Libero := falso$

In questo caso l'attributo tempo è vuoto, mentre la condizione è che allo stesso tempo abbia valore vero l'attributo Libero dell'Impiegato I ($I.Libero = vero$) e risulti non vuota la coda ($Qp.lunghezza > 0$). Se la condizione è verificata, si programma la fine del servizio al tempo corrente, T_C, più il tempo di servizio del cliente, cioè il suo attributo t, e si pone a falso il valore dell'attributo Libero dell'impiegato. Osserviamo che il cliente è ottenuto effettuando l'operazione Estrai sulla coda Qp. Simile è la seconda attività:

Attività	$a[2]$: Inizio servizio al telefono
Attributi	- $I.Libero = vero$ & $Qp.lunghezza = 0$ & $Qt.lunghezza > 0$
Azioni	$a[4].t := T_C + Qt.Estrai().t$; $I.Libero := falso$

La terza e la quarta attività hanno le seguente struttura comune:

Attività	$a[3]/a[4]$: Fine servizio
Attributi	t vero
Azioni	$I.Libero := vero$

Qui esiste un tempo al quale le attività vanno realizzate, quello di fine servizio calcolato dalle attività a[1] o a[2], mentre la condizione è sempre verificata. L'unica azione che viene fatta è quella di rendere libero l'impiegato.

l'attività a[5] è riportata di seguito. Non viene invece riportata la a[6] che è sostanzialmente identica.

Attività	$a[5]$: Arrivo cliente fisico
Attributi	t vero
Azioni	$Qp.Inserisci(nuovoCliente(rndp1()))$ $a[5].t := T_C + rndp2()$

Qui le azioni sono due. Innanzitutto viene chiamato un generatore di numeri casuali ($rndp1$) che genera il cliente che arriva, con le sue caratteristiche; il cliente così generato viene inserito nella coda. Poi viene programmato il prossimo arrivo, modificando il tempo della attività. Anche qui si usa un generatore di numeri casuali ($rndp2$) per determinare l'intervallo di tempo fra l'arrivo attuale ed il prossimo[10].

In questo approccio abbiamo fatto la scelta di considerare attività elementari comportanti una singola operazione di cambiamento di stato. Ad esempio, nel caso delle attività "Arrivo cliente", anche se la coda è vuota e l'impiegato disponibile, si mette il cliente in coda; ci sarà poi un'attività di tipo "Inizio servizio" che si verificherà subito dopo (però sempre nello stesso istante del tempo di simulazione) e che provvederà a fare iniziare il servizio estraendo il cliente dalla coda; per cui il tempo effettivo di permanenza in coda è nullo. Ciascuna di queste attività elementari è indipendente dalle altre. Questo ha il vantaggio di rendere abbastanza semplice l'aggiornamento o la manutenzione di programmi di simulazione basati su questo approccio. Ha però un prezzo, l'effettuazione da parte del controllore di ripetute scansioni di una lista che può essere notevolmente lunga.

Analizzando le attività elementari si vede come esse possano essere considerate di due tipi, *attività condizionate* e *attività programmate*. Le prime sono attività che si verificano, indipendentemente dal valore del tempo di simulazione, ogni qualvolta siano verificate determinate condizioni logiche. Appartengono a questa classe nell'esempio del botteghino teatrale le attività $a[1]$ ed $a[2]$. Infatti nella descrizione di queste attività l'attributo tempo è vuoto: lo svolgersi di esse dipende dallo stato complessivo del sistema. Le seconde invece sono destinate a svolgersi in tempi prefissati, indipendentemente dallo stato del sistema. Sono di questo tipo nell'esempio le attività $a[3]$, $a[4]$, $a[5]$, ed $a[6]$.

Da questa osservazione segue che le scansioni ripetute di tutta la lista

[10]I generatori di numeri casuali, dei quali parleremo più in dettaglio nel seguito servono per realizzare nella simulazione dei tempi di arrivo e di servizio che abbiano le stesse proprietà statistiche di quelli reali.

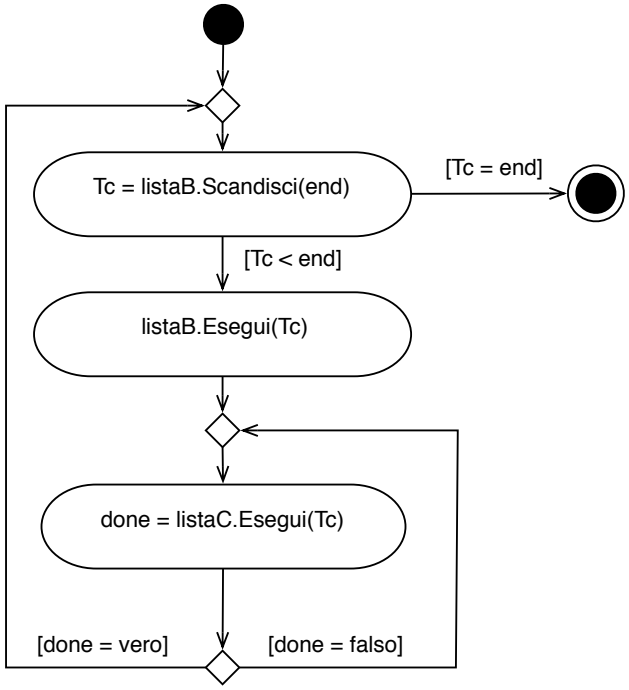

Figura 2.24. *Diagramma delle operazioni effettuate dal controllore nel metodo delle tre fasi*

delle attività da parte del controllore portano a dei controlli non necessari. In effetti le attività programmate possono essere scandite una sola volta, mentre le scansioni ripetute possono essere limitate alla parte della lista che contiene le attività condizionate. Questo permette di rendere più efficienti le operazioni del controllore. Un approccio di questo tipo viene chiamato *metodo delle tre fasi* (Pidd, 1998). Il diagramma degli stati di un simulatore che utilizzi il metodo delle tre fasi è indicato in figura 2.24, dove abbiamo indicato *listaB* e *listaC* rispettivamente la lista delle attività programmate e quella delle attività condizionate.

2.3.3 Simulazione per processi

In questo approccio tutti gli eventi del ciclo degli stati di una entità, con le relative operazioni, vengono vengono accorpati in una sequenza detta *proces-*

so. Un processo è sostanzialmente la sequenza delle operazioni descritte dal diagramma degli stati.

Nel caso del botteghino teatrale, ogni processo è relativo ad un cliente. Ad esempio il processo relativo ad un cliente fisico C può essere pensato come costituito dalle seguenti operazioni, in cui abbiamo usato le notazioni già introdotte a proposito della simulazione per attività ed abbiamo indicato con $P(C)$il processo che è esso stesso una entità del modello.

1. Arrivo del cliente C quando il tempo di simulazione è uguale al suo tempo di arrivo, cioè $T_C = P(C).t$.

2. Calcolo del tempo di arrivo , $P(C').t$ del cliente successivo, C';

3. Generazione dell'entità C' e creazione del processo $P(C')$;

4. Il processo $P(C')$ viene posto in stato di attesa fino a che non risulti $T_C = P(C').t$;

5. Il cliente C viene inserito nella prima coda, e viene posto in stato di attesa fino a quando non si trovi in testa alla coda e risulti Impiegato.Libero = *vero*;

6. Si pone Impiegato.Libero := $falso$, si pone $P(C)$.Tempo_Prossima_Attività $=T_C + C.t$, e si pone il processo in stato di attesa fino a quando $T_C = P(C)$.Tempo_Prossima_Attività;

7. Si pone Impiegato.Libero := $Vero$ e si interrompe il processo.

Un processo può essere *attivo* oppure *in attesa*. In quest'ultimo caso, si può trattare di una attesa condizionata, quando la ripresa dell'attività del processo dipende dal realizzarsi di condizioni esterne, oppure di un'attesa programmata, quando il tempo in cui il processo verrà riattivato è predefinito. Nell'esempio c'è una attesa condizionata (operazione n.5) e due attese programmate (operazioni n.4 e n.6).

Nell'esempio considerato, l'impiegato può essere considerato come una risorsa che viene utilizzata dai processi (clienti). In casi più complessi si possono avere diverse classi di entità e quindi tipi di processi che competono per l'uso di risorse comuni. In questo caso si hanno processi che interagiscono, ciascuno condizionato dallo stato degli altri.

In questo tipo di approccio, il programma di controllo deve mantenere una lista contenente, per ciascuna entità/processo due informazioni: il tempo di riattivazione (se conosciuto) ed il punto nel processo in cui la prossima riattivazione deve avvenire. Questa lista può essere suddivisa in due sottoliste: quella degli *eventi futuri*, che contiene i processi in attesa non condizionata ed il cui tempo di riattivazione è maggiore del tempo corrente; quella degli *eventi correnti*, che contiene i processi in attesa non condizionata che devono al tempo corrente essere riattivati, e tutti quelli in attesa condizionata. Questi ultimi stanno in questa lista, anche se non necessariamente verranno riattivati, perché il riattivarsi di uno degli altri può creare le condizioni per la loro riattivazione.

L'operazione tipica che il controllore effettua in questo tipo di approccio è:

(*i*) esegue una scansione della sottolista degli eventi futuri, in modo da determinare il nuovo tempo di simulazione che viene quindi aggiornato;

(*ii*) sposta dalla sottolista degli eventi futuri a quella degli eventi correnti le entità il cui tempo di riattivazione coincide con il nuovo tempo di simulazione;

(*iii*) esegue ripetute scansioni della sottolista degli eventi correnti, cercando di spingere ciascuna entità il più avanti possibile nel suo processo (quando un'entità viene posta in attesa non condizionata, essa viene spostata alla sottolista degli eventi futuri).

Naturalmente all'inizio il controllore deve inizializzare il sistema dando ad esempio nel caso del botteghino valori *vero* all'attributo *Libero* dell'impiegato e generando il primo cliente.

L'approccio per processi e porta a programmi di simulazione efficienti computazionalmente. Esso richiede però, soprattutto per modelli complessi, grande attenzione nella gestione delle interazioni tra processi: c'è ad esempio il rischio che si creino situazioni di stallo. Questo fatto è particolarmente critico quando si debba intervenire su un modello già esistente, per aggiornarlo o modificarlo.

Capitolo 3

Funzioni di distribuzione e test statistici

Presentiamo in questo capitolo i concetti e gli strumenti del Calcolo delle Probabilità e della Statistica indispensabili per la costruzione e l'uso di modelli di simulazione stocastica. La trattazione, che sarà necessariamente molto sintetica, è basata sul testo di Mood, Graybill e Boes **?** e su quello di Ross **?**, a cui rimandiamo per approfondimenti.

3.1 Variabili casuali

Uno spazio di probabilità è una tripla $(\Omega, \mathcal{F}, P)$, dove:

· Ω , lo *spazio campione*, è un insieme di elementi (tipicamente l'insieme dei possibili esiti di un esperimento);

· $\mathcal{F}$, lo *spazio degli eventi*, è una famiglia di sottoinsiemi di Ω, caratterizzata delle seguenti proprietà :

$$\begin{array}{ll} i) & \Omega \in \mathcal{F}, \\ ii) & A \in \mathcal{F} \Longrightarrow \Omega \setminus A \in \mathcal{F}, \\ iii) & A, B \in \mathcal{F} \Longrightarrow A \cup B \in \mathcal{F}; \end{array}$$

· P: $\mathcal{F} \rightarrow$[0, 1], la *funzione di probabilità*, è una funzione reale avente le seguenti proprietà :

$i)$ $P(A) \geq 0, \forall A \in \mathcal{F}$,
$ii)$ $P(\Omega) = 1$,
$iii)$ $(A, B \in \mathcal{F}) \wedge (A \cap B = \emptyset) \Longrightarrow P(A \cup B) = P(A) + P(B)$;

Dalle proprietà sopra definite si possono derivare facilmente le seguenti:

$$A, B \in \mathcal{F} \Longrightarrow A \cap B \in \mathcal{F}, \text{ e } B \setminus A \in \mathcal{F},$$
$$P(\Omega \setminus A) = 1 - P(A),$$
$$A \subseteq B \Longrightarrow P(B \setminus A) = P(B) - P(A),$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Dato uno spazio di probabilità, $(\Omega, \mathcal{F}, P)$, una *variabile casuale* è una funzione X: $\Omega \to \Re$, avente la proprietà che, per ogni reale r, $\{\omega \in \Omega$: $X(\omega) \leq r\} \in \mathcal{F}$. L'uso dell'espressione "variabile casuale" non ha convincenti giustificazioni ed è causa di ambiguità; è comunque un'espressione universalmente accettata e pertanto verrà usata anche qui.

La funzione $F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$, definita sull'insieme dei reali, è detta *funzione di distribuzione*.

L'uso di variabili casuali è fondamentale nella simulazione stocastica. Nei sistemi da simulare si presentano usualmente fenomeni non (facilmente) prevedibili apriori (arrivo di clienti ad uno sportello, quantità di pioggia in una data stagione, guasti in un'apparecchiatura, ...). Tali fenomeni vengono rappresentati per mezzo di variabili casuali, delle quali, per mezzo di serie storiche o di indagini campionarie, viene poi studiata la funzione di distribuzione. Vengono quindi costruiti generatori di numeri casuali, aventi le stesse distribuzioni, che verranno usati nella simulazione per modellare i fenomeni stessi.

Esempio 1 Ad uno sportello di banca assumiamo che si possano fare solamente tre operazioni, incasso di un assegno (operazione a), bonifico (operazione b) e versamento (operazione v), e che il singolo cliente faccia una sola di esse. Consideriamo come esperimento l'arrivo del prossimo cliente, come esito dell'esperimento la richiesta di una delle operazioni, a, b e v, e come evento il fatto che il cliente chieda una in un sottoinsieme delle operazioni (ad esempio la a o la v).

Poniamo allora $\Omega = \{a, b, v\}$, $\mathcal{F} = 2^{\Omega}$. Sia poi X la funzione così definita:

$$\begin{aligned} X(a) &= 0, \\ X(b) &= 1, \\ X(v) &= 2. \end{aligned}$$

La funzione X è una variabile casuale, infatti si ha:

$$\begin{aligned} r < 0 \quad \{\omega : X(\omega) \leq r\} &= \emptyset, \\ 0 \leq r < 1 \quad \{\omega : X(\omega) \leq r\} &= \{a\}, \\ 1 \leq r < 2 \quad \{\omega : X(\omega) \leq r\} &= \{a, b\}, \\ 2 \leq r \quad \{\omega : X(\omega) \leq r\} &= \Omega. \end{aligned}$$

Esempio 2 Si consideri il numero di pazienti che si presentano ad un ambulatorio tra le 9 e le 10 di mattina, e poniamo $\Omega = \{0, 1, 2, \ldots\}$, $\mathcal{F} = 2^{\Omega}$, e $X(\omega) = \omega$ (la funzione identità). La funzione X così definita è una variabile casuale, infatti, per ogni reale r è

$$\{\omega : X(\omega) \leq r\} = \{0, \ldots, \lfloor r \rfloor\} \in \mathcal{F}.$$

3.1.1 Distribuzioni discrete

Una variabile casuale è detta *discreta*, se l'insieme dei valori che può assumere è numerabile. Sia $(\Omega, \mathcal{F}, P)$ uno spazio di probabilità, e X una variabile casuale discreta che possa assumere i valori x_1, x_2, ..., x_k, ... Definiamo la *funzione di densità discreta*:

$$f_X(x) = \begin{cases} P(X = x), & \text{se } x = x_i, \text{per qualche } i = 1, 2, \ldots, \\ 0, & \text{altrimenti.} \end{cases} \tag{3.1}$$

Le funzioni di densità e di distribuzione di X sono legate dalle seguenti relazioni:

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i), \tag{3.2}$$

$$f_X(x_i) = F_X(x_i) - lim_{h \to 0^+} F_X(x_i - h). \tag{3.3}$$

La *media* di X, che sarà indicata con μ_X, è definita dalla

$$E[X] = \sum_i x_i f_X(x_i). \tag{3.4}$$

La media della variabile casuale X^r viene detta *momento* r^{esimo} di X e viene denotata come μ_X^r.

La *varianza* X, indicata con σ_X^2, è la media degli scarti quadratici rispetto alla media μ_X, e rappresenta una misura di dispersione di X. La sua radice quadrata, σ_X, è detta *deviazione standard*. La varianza è definita dalla

$$Var[X] = \sum_i (x_i - \mu_X)^2 f_X(x_i), \tag{3.5}$$

da cui è immediato derivare la

$$Var[X] = E[X^2] - (E[X])^2. \tag{3.6}$$

Se X e Y sono variabili casuali, e $\alpha \in \Re$, allora valgono le seguenti proprietà:

$$\begin{aligned} E[\alpha] &= \alpha, \\ E[\alpha X] &= \alpha E[X], \\ E[X+Y] &= E[X] + E[Y], \\ Var[\alpha] &= 0, \\ Var[\alpha X] &= \alpha^2 Var[X]. \end{aligned}$$

Se X e Y sono variabili casuali, la varianza della loro somma è data dalla:

$$\begin{aligned} Var[X+Y] &= E[(X+Y-E[X+Y])^2] \\ &= E[(X+Y-\mu_X-\mu_Y)^2] \\ &= E[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)] \\ &= \sigma_X^2 + \sigma_Y^2 + 2Cov[X,Y], \end{aligned}$$

dove $Cov[X,Y]$ è la covarianza di X ed Y:

$$Cov[X,Y] = E[(X-E[X])(Y-E[Y])].$$

Se X ed Y sono indipendenti, allora la loro covarianza è nulla e si ha:

$$Var[X+Y] = Var[X] + Var[Y].$$

Un ruolo rilevante nel legare fra loro media è varianza di una distribuzione hanno le *Disuguaglianze di Chebyshef*:

$$\begin{aligned} P(|X-\mu_X| > r\sigma_X) &\leq 1/r^2, & (3.7)\\ P(|X-\mu_X| < r\sigma_X) &\geq 1-1/r^2, & (3.8) \end{aligned}$$

dove r è un reale positivo.

Introduciamo ora una funzione particolarmente importante ai fini della determinazione di media e varianza di distribuzioni, la *funzione generatrice dei momenti*:

$$\begin{aligned} m_X(t) &= E[e^{tX}] = E[1 + Xt + \frac{1}{2!}(Xt)^2 + \frac{1}{3!}(Xt)^3 + \ldots] & (3.9)\\ &= 1 + \mu_{1X}t + \frac{1}{2!}\mu_{2X}t^2 + \ldots = \sum_{i=0}^{\infty}\frac{1}{i!}\mu_{iX}t^i. & (3.10) \end{aligned}$$

dove la seconda uguaglianza deriva dall'espansione in della funzione e^{tX}. Abbiamo assunto l'esistenza di un intervallo di ampiezza positiva, $[-h, h]$, tale che per ogni t in esso contenuto la funzione $m_x(t)$ è definita..

È immediato verificare che risulta:

$$\frac{d^r}{dt^r}m_X(0) = \mu_{rX}. \quad (3.11)$$

Nel seguito descriveremo brevemente alcune delle più comuni funzioni di distribuzione. Dato un insieme S, con $I_S(x)$ indicheremo una funzione che vale 1 se $x \in S$ e 0 altrimenti.

Distribuzione uniforme

Sia X una variabile casuale che assume i valori $1, 2, ..., n$. Essa viene detta avere una distribuzione uniforme se risulta:

$$f_X(x) = f_X(x;n) = \begin{cases} \frac{1}{n}, & x = 1,2,...,n \\ 0, & altrimenti \end{cases} = \frac{1}{n}I_{\{1,2,\ldots,n\}}(x), \quad (3.12)$$

Si ha

$$\begin{aligned} E[X] &= \frac{1}{n}\sum_{i=1}^{n} i = \frac{(n+1)n}{2n} = \frac{n+1}{2}; \\ Var[X] &= E[X^2] - (E[X])^2 = \frac{1}{n}\sum_{i=1}^{n} i^2 - \frac{(n+1)^2}{4} \\ &= \frac{n(n+1)(2n+1)}{6n} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}; \\ m_X(t) &= \frac{1}{n}\sum_{j=1}^{n} e^{jt}. \end{aligned}$$

Esempio La variabile casuale che rappresenta l'esito del lancio di un dado ha distribuzione uniforme con n=6 (naturalmente nell'ipotesi che il dado non sia truccato).

Distribuzione binomiale

Consideriamo un esperimento che abbia due possibili esiti, che possiamo chiamare S (successo) e F (fallimento), l'uno con probabilità p e l'altro con probabilità $q = 1 - p$. Assumiamo ora di eseguire n volte l'esperimento in modo che ciascun esito sia indipendente dagli altri, e consideriamo come variabile casuale X il numero di volte in cui si ha un successo cioè in cui l'esito dell'esperimento è S.

Chiaramente è:

$$f_X(x) = f_X(x; n, p) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0, 1, 2, ..., n \\ 0, & altrimenti \end{cases} . \tag{3.13}$$

La variabile casuale X viene detta avere una *distribuzione binomiale.*

$$m_X(t) = \sum_{i=0}^{n} e^{ti} \binom{n}{i} p^i q^{n-i} = \sum_{i=0}^{n} \binom{n}{i} (pe^t)^i q^{n-i} = (q + pe^t)^n. \tag{3.14}$$

Possiamo pertanto calcolare la media e la varianza. Essendo

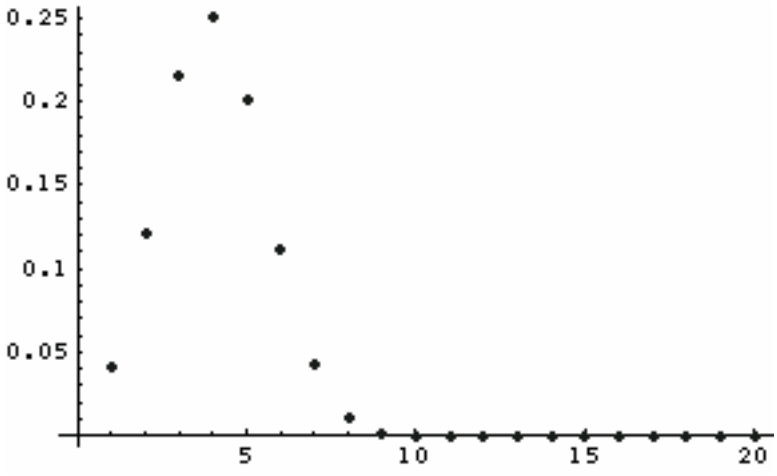

Figura 3.1. *Distribuzione binomiale con* $n = 10$ *e* $p = 0.6$

$$\frac{dm_X(t)}{dt} = pe^t n(pe^t + q)^{n-1}, \tag{3.15}$$

e

$$\frac{d^2 m_X(t)}{dt^2} = pe^t n(pe^t + q)^{n-2}(npe^t + q), \tag{3.16}$$

e ricordando che $p + q = 1$, si ha

$$E[X] = \frac{dm_X(0)}{dt} = pn, \tag{3.17}$$

$$E[X^2] = \frac{d^2 m_X(0)}{dt^2} = np(np + q). \tag{3.18}$$

È poi

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 \\ &= np(np + q) - (np)^2 = npq. \end{aligned}$$

Un esempio di distribuzione binomiale con $n = 10$ e $p = 0.6$ è riportato in figura 3.1.

Nel caso particolare in cui $n = 1$ si parla di *distribuzione di Bernoulli*. Supponiamo ora di avere n realizzazioni di una variabile casuale con distribuzione di Bernoulli. La probabilità di avere x di tali realizzazioni uguali ad 1 è $\binom{n}{x} p^x q^{n-x}$. Abbiamo così ottenuto una *v.c.* con distribuzione binomiale.

È possibile dimostrare che se $X_1, X_2, \ldots, X_m$ sono variabili casuali indipendenti, con distribuzione binomiale con parametri $n_1, n_2, \ldots, n_m$ e p, allora la variabile casuale $Y = \sum_{i=1}^{m} X_i$ ha una distribuzione binomiale con parametri $n = \sum_{i=1}^{m} n_i$.

Oserviamo infine che la distribuzione binomiale è simmetrica se e solo se $p = 0.5$

Una applicazione interessante della distribuzione binomiale si ha nel controllo di qualità: il numero di pezzi difettosi in un lotto di dimensione n, assumendo che sia p la probabilità che un pezzo abbia dei difetti, ha una distribuzione binomiale con parametri n e p.

Esempio Nel volo Roma-Milano delle 16 della compagnia aerea AirPadania ci sono disponibili 80 posti. La probabilità che un viaggiatore prenotato non si presenti alla partenza sia indicata con p. Assumiamo che il valore di p dipenda dalla fascia oraria e dal tipo di volo, ma che per dato volo sia lo stesso per ogni passeggero. Avendo già 80 prenotazioni, la AirPadania, per decidere che politica di "overbooking" seguire, vuol sapere quale è la distribuzione di probabilità della *v.c.* X = numero di viaggiatori che non si presentano. Il presentarsi o non presentarsi di un singolo viaggiatore può essere visto come il realizzarsi di una *v.c.* con distribuzione di Bernoulli. Pertanto la X ha una distribuzione binomiale con $n = 80$.

Distribuzione geometrica

Supponiamo di effettuare una sequenza di esperimenti identici ed indipendenti, ciascuno dei quali ha come esito S (successo) con probabilita p e F (fallimento) con probabilità $q = 1 - p$. Consideriamo come variabile casuale X il numero di fallimenti prima di ottenere un successo. Abbiamo allora che è

$$P[X = k] = p(1 - p)^k, \qquad k = 0, 1, 2, \ldots$$

La funzione generatrice dei momenti è data da:

$$m_X(t) = \sum_{i=0}^{\infty} e^{it} p(1-p)^i = p \sum_{i=0}^{\infty} [e^t(1-p)]^i.$$

e se assumiamo che essa sia definita in un intorno sufficientemente piccolo dello 0 per cui risulti $e^t(1-p) < 1$, allora è[1]

$$m_X(t) = \frac{p}{1 - e^t(1-p)}.$$

Possiamo ora calcolare la media e la varianza:

$$E[X] = m'_X(0) = \frac{1-p}{p};$$

$$Var[X] = E[X^2] - E[X]^2 = m''_X(0) - E[X]^2 = \frac{(1-p)(2-p)}{p^2} - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}.$$

Distribuzione di Poisson

Consideriamo eventi che accadono nel tempo, quali l'arrivo di clienti ad uno sportello (di telefonate ad un centralino, ...); sia ν il numero medio di occorrenze dell'evento nell'unità di tempo, e supponiamo che valgano le seguenti proprietà.

- La probabilità di avere esattamente una occorrenza in un intervallo di tempo di ampiezza h opportunamente piccola ($h \ll 1$) è $\nu h + o(h)$, dove con $o(h)$ viene indicato un infinitesimo di ordine superiore rispetto ad h.
- La probabilità di più di un'occorrenza in un intervallo di ampiezza h è un $o(h)$.
- I numeri di occorrenze in intervalli disgiunti sono indipendenti.

Una sequenza temporale di eventi che abbiano le proprietà indicate sopra viene anche detta un *Processo di Poisson*.

Dato un processo di Poisson, consideriamo la variabile casuale X uguale al numero di eventi che si verificano in un dato intervallo $(0, t)$. Dividiamo

[1] Ricordiamo che, per le proprietà della serie geometrica, se $a < 1$, allora è $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$.

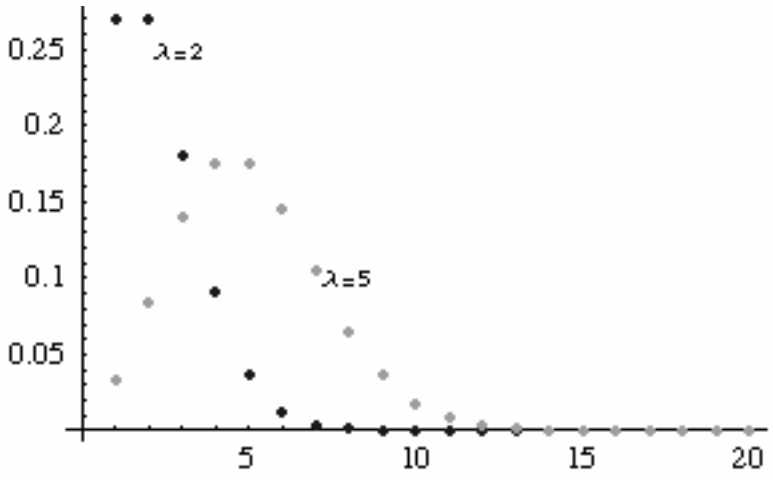

Figura 3.2. *Distribuzione di Poisson*

l'intervallo in n intervallini di ampiezza $\frac{t}{n}$. La probabilità di avere esattamente una occorrenza in un dato intervallino, a meno di un infinitesimo di ordine superiore rispetto a $\frac{t}{n}$, è $\nu\frac{t}{n}$, e per la proprietà dell'indipendenza, abbiamo che la probabilità di k occorrenze è, a meno di un infinitesimo di ordine superiore data dalla

$$\begin{aligned} P[X=k] &= \binom{n}{k}\left(\frac{\nu t}{n}\right)^k\left(1-\frac{\nu t}{n}\right)^{n-k} \\ &= \frac{n(n-1)...(n-k+1)}{k!n^k}(\nu t)^k\left(1-\frac{\nu t}{n}\right)^n\left(1-\frac{\nu t}{n}\right)^{-k} \\ &\xrightarrow[n\to\infty]{} \frac{(\nu t)^k e^{-\nu t}}{k!}. \end{aligned}$$

Abbiamo così ricavato una distribuzione molto usata, nota come *distribuzione di Poisson.*

$$f_X(x) = f_X(x;\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}, & x = 0,1,2,... \\ 0, & altrimenti \end{cases}, \tag{3.19}$$

Esempi di distribuzioni di Poisson con $\lambda = 2$ e $\lambda = 5$ sono riportati in figura 3.2.

Calcoliamo la funzione generatrice dei momenti

$$m_X(t) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{(e^t\lambda)^i}{i!} = e^{\lambda(e^t-1)}, \tag{3.20}$$

dove l'uguaglianza deriva dal fatto che è $\sum_{i=0}^{\infty} \frac{a^i}{i!} = e^a$.

Si ha allora

$$\frac{dm_X(t)}{dt} = \lambda e^{\lambda(e^t-1)+t}, \tag{3.21}$$

$$\frac{d^2m_X(t)}{dt^2} = \lambda e^{\lambda(e^t-1)+t}(\lambda e^t + 1), \tag{3.22}$$

da cui

$$E[X] = \frac{dm_X(0)}{dt} = \lambda, \tag{3.23}$$

$$E[X^2] = \frac{d^2m_X(0)}{dt^2} = \lambda(\lambda + 1). \tag{3.24}$$

Possiamo quindi calcolare la varianza:

$$Var[X] = E[X^2] - E[X]^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda. \tag{3.25}$$

Questa distribuzione fornisce un ragionevole modello per molti fenomeni casuali in cui si vuole descrivere il numero di volte che un dato evento avviene nell'unità di tempo, ad esempio numero di arrivi ad uno sportello nell'unità di tempo.

3.1.2 Distribuzioni continue

Una variabile casuale, X, è detta continua se esiste una funzione reale f_X tale che per ogni x reale :

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du,$$

dove f_X è la funzione di densità di probabilità, o più semplicemente la *funzione di densità*.

Per i punti x in cui la $F_X(x)$ è differenziabile, vale la:

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

È quindi possibile data l'una delle due funzioni, densità o distribuzione, trovare l'altra.

Nel caso di variabili continue, la media della variabile X viene definita come segue

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Di conseguenza si definiscono la varianza, i momenti e la funzione generatrice dei momenti, per cui valgono le proprietà già viste a proposito delle distribuzioni discrete:

$$\begin{aligned} Var[X] &= E[(X-\mu_X)^2] = E[X^2] - (E[X])^2, \\ \mu_X^r &= E[X^r], \\ m_X(t) &= E[e^{tX}] \end{aligned}$$

Le proprietà della funzione generatrice dei momenti sono simili a quelle della funzione generatrice dei momenti per il caso discreto.

Distribuzione uniforme

Una variabile casuale, X, è uniformemente distribuita nell'intervallo reale $[a, b]$ se è caratterizzata dalle seguenti funzioni di densità e distribuzione

$$f_X(x) = f_X(x; a, b) = \frac{1}{b-a} I_{[a,b]}(x),$$

$$F_X(x) = \left(\frac{x-a}{b-a}\right) I_{[a,b]}(x) + I_{(b,\infty)}(x),$$

Si ha allora

$$\begin{aligned} E[X] &= \int_a^b \frac{x}{b-a}dx = \frac{b+a}{2}, \\ Var[X] &= E[X^2] - (E[X])^2 = \int_a^b \frac{x^2}{b-a}dx - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}, \\ m_X(t) &= \int_a^b e^{tx}\frac{1}{b-a}dx = \frac{e^{bt}-e^{at}}{(b-a)t} \end{aligned}$$

La distribuzione uniforme gioca un ruolo particolarmente importante nella simulazione. Usualmente infatti si parte da generatori di variabili casuali uniformi per derivare le diverse distribuzioni che servono. Osserviamo che in questo caso la funzione generatrice dei momenti non è definita nello 0.

Distribuzione normale

Una distribuzione di particolare importanza sia dal punto di vista della teoria che da quello delle applicazioni pratiche è la distribuzione normale:

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

dove i parametri μ e σ sono rispettivamente la media e la deviazione standard; infatti è

$$\begin{aligned} m_X(t) &= E[e^{tX}] = e^{t\mu}E[e^{t(X-\mu)}] \\ &= e^{t\mu}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{t(x-\mu)}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\ &= e^{t\mu}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2-2\sigma^2t(x-\mu)}{2\sigma^2}}dx \\ &= e^{\mu t+\frac{\sigma^2t^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu-\sigma^2t)^2}{2\sigma^2}}dx. \end{aligned}$$

Osserviamo che l'integrale fornisce l'area sotto la curva che definisce la densità di una variabile casuale normale con media $\mu - \sigma^2 t$ e varianza σ^2, e pertanto vale 1. Si ha allora

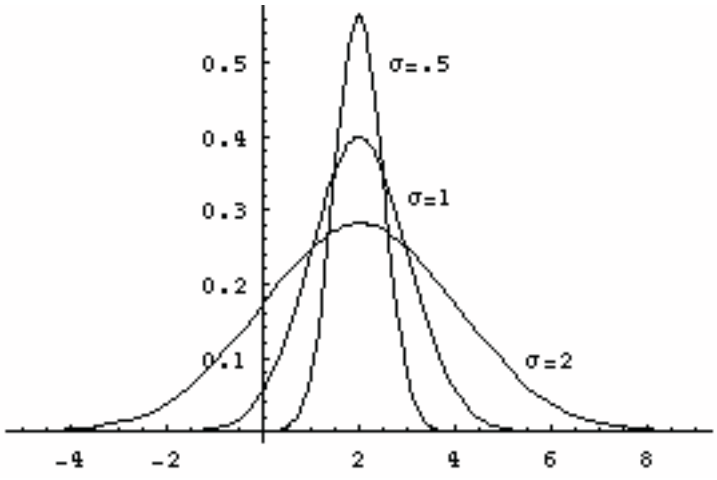

Figura 3.3. *Distribuzione normale*

$$m_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

Possiamo ora verificare che effettivamente μ e σ^2 sono la media e la varianza. Infatti si ha

$$\begin{aligned} E[X] &= \frac{d}{dt} m_X(0) = \mu \\ Var[X] &= E[X^2] - (E[X])^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2 \end{aligned}$$

Una variabile casuale con distribuzione normale è detta *standard* se ha media 0 e varianza 1, e viene denotata con $N(0,1)$

La distribuzione normale è la distribuzione limite di molte altre distribuzioni di probabilità. Essa si presta bene alla modellazione di variabili casuali rappresentanti lo scarto in più o in meno rispetto ad un qualche prefissato obiettivo.

Esempi di distribuzioni normali con $\mu = 2$ e diversi valori di σ sono riportati in figura 3.3.

Distribuzione esponenziale

E' una variabile casuale definita nello spazio dei reali non negativi con distribuzione

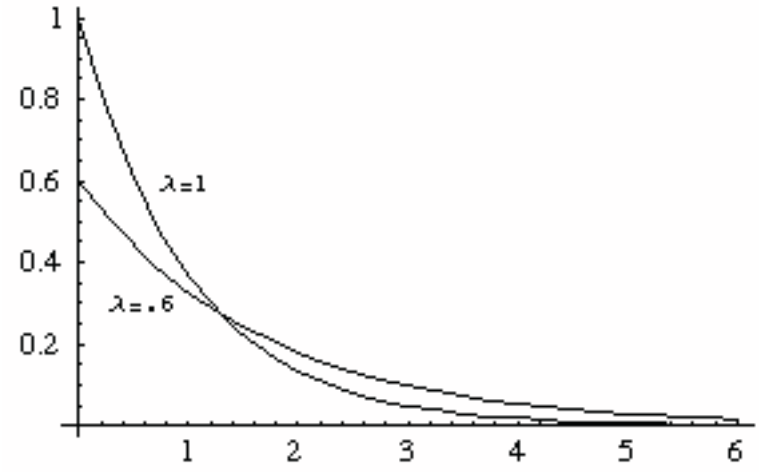

Figura 3.4. *Distribuzione esponenziale*

$$f_X(x;\lambda) = \lambda e^{-\lambda x},$$

$$F_X(x) = 1 - e^{-\lambda x},$$

con λ un parametro positivo.
La funzione generatrice dei momenti, per $t < \lambda$, è

$$\begin{aligned} m_X(t) &= E[e^{tX}] = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx \\ &= \lambda\int_0^\infty e^{-(\lambda-t)x}dx = \frac{\lambda}{\lambda - t}. \end{aligned}$$

Da essa è immediato derivare la media e la varianza:

$$E[X] = \frac{1}{\lambda}, Var[X] = \frac{1}{\lambda^2}.$$

Esempi di distribuzione esponenziale sono riportati in figura 3.4.

La distribuzione esponenziale si presta bene a modellare le distanze temporali tra un evento ed il successivo, quando il numero di eventi in un fissato intervallo di tempo ha una distribuzione di Poisson.

Consideriamo un evento le cui occorrenze nel tempo hanno una distribuzione di Poisson. Supponendo che si sia appena verificata un'occorrenza, chiamiamo con X la variabile casuale "tempo da attendere prima della occorrenza successiva". È allora

$$P[X > t] = P[\text{nessuna occorrenza fino al tempo } t] = e^{-\nu t},$$

e di conseguenza

$$F_X(t) = P[X \leq t] = 1 - e^{-\nu t}, t \geq 0.$$

Una caratteristica importante della distribuzione esponenziale è che, per ogni coppia (s, t) di reali positivi, vale la

$$P[X > s + t | X > s] = P[X > t]. \tag{3.26}$$

Infatti è

$$P[X > s + t | X > s] = \frac{P[X > s + t]}{P[X > s]} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P[X > t].$$

Si parla in questo caso di proprietà di *assenza di memoria*. Per capirne il senso assumiamo di avere un processo di Poisson e che sia passato il tempo s dall'ultimo evento verificatosi. Ci si chiede quale sia la probabilità che passi ancora almeno un tempo t prima che si verifichi il prossimo evento. La proprietà di assenza di memoria per la distribuzione esponenziale ci dice che la probabilità cercata è indipendente dal valore di s: possiamo cioè fare come se il processo iniziasse nel momento in cui ci troviamo.

Vale inoltre la seguente proprietà:

$$P[cX \leq x] = P[X \leq \frac{x}{c}] = 1 - e^{\frac{\lambda}{c}x}$$

Cioè se X è una variabile casuale esponenziale con parametro λ, allora cX è una variabile casuale esponenziale con parametro $\frac{\lambda}{c}$.

Distribuzione Gamma

Siano $X_1, X_2, \ldots, X_n$ variabili casuali indipendenti con distribuzione esponenziale e parametro λ. Consideriamo la nuova variabile casuale

$$Y_n = \sum_{i=1}^{n} X_i.$$

Per studiare la funzione di distribuzione di Y_n, osserviamo che ciascuna X_i può essere pensata come il tempo intercorso fra due eventi successivi in un processo poissoniano con λ occorrenze in media nell'unità di tempo. Allora la probabilità che Y_n sia minore o uguale a t è pari alla probabilità che nel tempo t si verifichino almeno n eventi, cioè:

$$F_{Y_n}(t) = P[Y_n \leq t] = \sum_{j=n}^{\infty} \frac{(\lambda t)^j e^{-\lambda t}}{j!}.$$

Possiamo calcolare ora la funzione di densità:

$$f_{Y_n}(t) = F'_{Y_n}(t) = \sum_{j=n}^{\infty} \frac{\lambda j (\lambda t)^{j-1} e^{-\lambda t} - \lambda (\lambda t)^j e^{-\lambda t}}{j!}$$

$$= \lambda e^{-\lambda t} \left(\sum_{j=n}^{\infty} \frac{(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=n}^{\infty} \frac{(\lambda t)^j}{j!} \right) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!}.$$

Questa distribuzione viene chiamata *distribuzione Gamma* con parametri n e λ.

Esempi di distribuzione Gamma sono riportati in figura 3.5.

La media e la varianza sono date dalle:

$$E[Y_n] = \frac{n}{\lambda}, \qquad Var[Y_n] = \frac{n}{\lambda^2}.$$

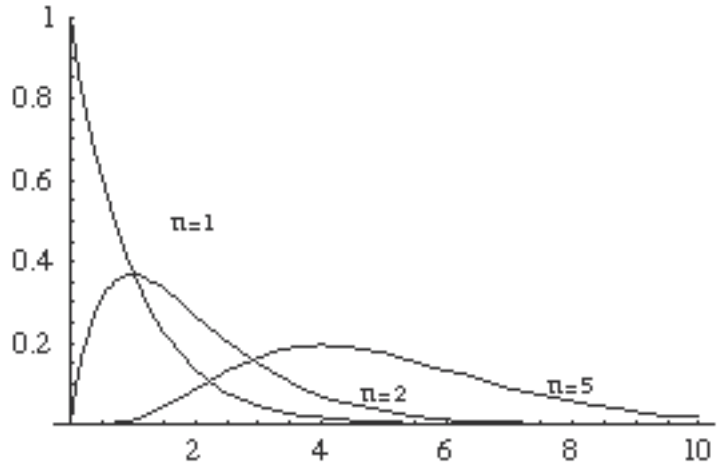

Figura 3.5. *Distribuzione Gamma con* $\lambda = 1$.

3.2 Stima di parametri

3.2.1 Media e varianza del campione

Siano $X_1, X_2, ..., X_n$ *v.c.* indipendenti con una data distribuzione F, e con $E[X_i] = \mu$ e $Var[X_i] = \sigma^2$, $i = 1, \ldots, n$.

la quantità

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

è detta *media campionaria*, ed è uno stimatore di μ; può quindi essere usata per stimare questo parametro, quando esso non sia noto. Osserviamo che si tratta di uno stimatore corretto, infatti è

$$E[\overline{X}_n] = \mu.$$

Per valutare la bontà di $\overline{X}_n$ come stimatore osserviamo che risulta

$$Var[\overline{X}_n] = \frac{1}{n}\sigma^2,$$

e pertanto lo stimatore è tanto più accurato quanto più grande è n.

Uno stimatore corretto della varianza σ^2 è

$$S_n^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}.$$

Infatti si ha:

$$\begin{aligned}(n-1)E[S_n^2] &= E[\sum_{i=1}^{n}(X_i - \overline{X}_n)^2] \\ &= \sum_{i=1}^{n} E[X_i^2] - nE[\overline{X}_n^2].\end{aligned}$$

Dalla (3.6) si ha che

$$E[X_i^2] = Var[X_i] + E[X_i]^2,$$

e pertanto

$$\begin{aligned}(n-1)E[S_n^2] &= \sum_{i=1}^{n}(Var[X_i] + E[X_i]^2) - n(Var[\overline{X}_n] + E[\overline{X}_n]^2) \\ &= n\sigma^2 + n\mu^2 - n\frac{\sigma^2}{n} - n\mu^2 \\ &= (n-1)\sigma^2.\end{aligned}$$

Abbiamo così mostrato che è $E[S_n^2] = \sigma^2$.

3.2.2 Intervalli di confidenza

Ci proponiamo ora di ottenere una valutazione della bontà della stima di μ fornita dalla media campionaria $\overline{X}_n$. La possibilità di valutare la bontà della stima è importante al fine di determinare il valore di n che consente di stimare μ con la voluta accuratezza.

Per il teorema del limite centrale, per n opportunamente grande, è

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \dot{\sim} N(0,1),$$

dove $\dot{\sim} N(0,1)$ significa: "è approssimativamente distribuito come una normale standard".

La stessa cosa vale se sostituiamo σ, che non conosciamo, con la sua stima S_n.

Sia Z una *v.c.* normale standard; per ogni $\alpha \in (0,1)$ sia z_α il valore per cui è

$$P(Z > z_\alpha) = \alpha.$$

Abbiamo allora

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

e quindi

$$P(-z_{\alpha/2} < \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} < z_{\alpha/2}) \approx 1 - \alpha,$$

o equivalentemente, moltiplicando per -1,

$$P(-z_{\alpha/2} < \sqrt{n}\frac{\mu - \overline{X}_n}{S_n} < z_{\alpha/2}) \approx 1 - \alpha,$$

da cui

$$P(\overline{X}_n - z_{\alpha/2}\frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2}\frac{S_n}{\sqrt{n}}) \approx 1 - \alpha,$$

Abbiamo così trovato che, con probabilità $1 - \alpha$, il valore μ incognito si trova nell'intervallo $\overline{X}_n \pm z_{\alpha/2}\frac{S_n}{\sqrt{n}}$. Si dirà allora che abbiamo una stima di μ con un intervallo di confidenza del $100(1 - \alpha)\%$.

Ad esempio, essendo $P(Z < 1.96) = 0.975$, si ha che la probabilità che la media campionaria $\overline{X}_n$ differisca da μ di più di $1.96\frac{S_n}{\sqrt{n}}$ è circa 0.05.

Supponiamo ora di volere stimare la media di una variabile casuale X, la cui distribuzione non è nota, in modo che la probabilità che di fare un errore maggiore di d sia pari ad α, con d ed α valori prefissati.

Se c è il reale per cui risulta $P(Z < c) = 1 - \alpha/2$, si generano successive realizzazioni di X fino ad averne un numero k tale che risulti $c\frac{S_k}{\sqrt{k}} < d$. È comunque opportuno che tale valore non sia inferiore a 30.

Per realizzare in modo efficiente il calcolo è opportuno disporre di formule ricorsive per il calcolo di $\overline{X}_k$ e di S_k^2. Tale formula è facilmente derivabile per la media:

$$\begin{aligned}\overline{X}_{k+1} &= \frac{1}{k+1}\sum_{j=1}^{k+1} X_j \\ &= \overline{X}_k - \overline{X}_k + \frac{k\overline{X}_k + X_{k+1}}{k+1} \\ &= \overline{X}_k + \frac{X_{k+1} - \overline{X}_k}{k+1}.\end{aligned}$$

Per quel che riguarda la varianza, possiamo scrivere:

$$\begin{aligned}S^2_{k+1} &= \sum_{j=1}^{k+1} \frac{(X_j - \overline{X}_{k+1})^2}{k} \\ &= \sum_{j=1}^{k} \frac{(X_j - \overline{X}_k + \overline{X}_k - \overline{X}_{k+1})^2}{k} + \frac{(X_{k+1} - \overline{X}_{k+1})^2}{k} \\ &= \sum_{j=1}^{k} \frac{(X_j - \overline{X}_k)^2 + (\overline{X}_k - \overline{X}_{k+1})^2 + 2(X_j - \overline{X}_k)(\overline{X}_k - \overline{X}_{k+1})}{k} \\ &\quad + \frac{(X_{k+1} - \overline{X}_{k+1})^2}{k} \\ &= (1 - \frac{1}{k})S^2_k + (\overline{X}_k - \overline{X}_{k+1})^2 + \frac{(X_{k+1} - \overline{X}_{k+1})^2}{k},\end{aligned}$$

dove l'ultima uguaglianza deriva dal fatto che è $\sum_{j=1}^{k}(X_j - \overline{X}_k) = 0$.

Essendo

$$\begin{aligned}X_{k+1} - \overline{X}_{k+1} &= \frac{(k+1)X_{k+1} - \sum_{j=1}^{k+1} X_j}{k+1} \\ &= \frac{kX_{k+1} - \sum_{j=1}^{k} X_j}{k+1} \\ &= \frac{kX_{k+1} - \sum_{j=1}^{k}[(k+1)X_j - kX_j]}{k+1} \\ &= k(\overline{X}_{k+1} - \overline{X}_k),\end{aligned}$$

si ha

$$S^2_{k+1} = (1 - \frac{1}{k})S^2_k + (1 + k)(\overline{X}_{k+1} - \overline{X}_k)^2.$$

3.2.3 Massima verosimiglianza

Sia $X_1, X_2, \ldots, X_n$ il campione che assumiamo provenga da una distribuzione con funzione di densità $f_\theta(x)$, dove con θ si è indicato il parametro che caratterizza la distribuzione (o il vettore dei parametri, nel caso ve ne siano più di uno). Allora, nell'ipotesi che le osservazioni siano indipendenti, una misura della probabilità di avere ottenuto proprio quel campione da una popolazione con la distribuzione data è fornita dalla funzione

$$L(\theta) = f_\theta(X_1) f_\theta(X_2) \ldots f_\theta(X_n)$$

che è detta *funzione di verosimiglianza*.

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

Osserviamo che, nel caso di distribuzioni discrete, $L(\theta)$ è proprio la probabilità di avere ottenuto il campione $X_1, X_2, \ldots, X_n$. È diverso invece il caso di distribuzioni continue, per le quali la probabilità di un particolare insieme finito di valori è comunque nulla. In questo caso possiamo però affermare che la probabilità che l'estrazione casuale di un elemento da una popolazione con la distribuzione data sia un valore compreso in un'intorno di raggio $\varepsilon/2$ di X_i è approssimativamente $\varepsilon f_\theta(X_i)$, con un'approssimazione tanto più accurata quanto più piccolo è ε. Pertanto $L(\theta)$ è approssimativamente proporzionale alla probabilità dell'estrazione di un campione di n elementi, $Y_1, Y_2, \ldots, Y_n$, con $Y_i \in [X_i - \varepsilon, X_i + \varepsilon]$, $i = 1, 2, \ldots, n$, e con ε opportunamente piccolo.

Esempi

Stima del parametro λ in una distribuzione esponenziale. Supponiamo di volere stimare con il metodo della massima verosimiglianza il parametro λ di una esponenziale. Si ha

$$\begin{aligned} L(\lambda) &= (\lambda e^{-\lambda X_1})(\lambda e^{-\lambda X_2}) \ldots (\lambda e^{-\lambda X_n}) \\ &= \lambda^n e^{-\lambda \sum_{i=1}^n X_i} \\ &= \lambda^n e^{-\lambda n \overline{X}_n} \end{aligned}$$

La derivata di $L(\lambda)$ è

$$\frac{dL}{d\lambda} = n\lambda^{n-1}e^{-\lambda n\overline{X}_n} - \lambda^n n\overline{X}_n e^{-\lambda n\overline{X}_n}$$

e, uguagliando a 0, si ottiene

$$\lambda = \frac{1}{\overline{X}_n}$$

che era quello che ci si aspettava essendo la media media campionaria uno stimatore corretto di $1/\lambda$, la media della distribuzione.

Stima dei parametri di una distribuzione uniforme. Assumiamo che $x_1 \leq x_2 \leq \cdots \leq x_n$ siano realizzazioni di una *v.c.* X uniforme, della quale non conosciamo né gli estremi né l'ampiezza dell'intervallo.

Dalle proprietà della distribuzione uniforme sappiamo che

$$\sigma^2 = \frac{(b-a)^2}{12} \Rightarrow b - a = 2\sqrt{3}\sigma,$$

e quindi possiamo scrivere la funzione di densità come segue

$$f(x;\mu,\sigma) = \frac{1}{2\sqrt{3}\sigma} I_{[\mu-\sqrt{3}\sigma,\mu+\sqrt{3}\sigma]}(x).$$

La funzione di massima verosimiglianza è allora

$$\begin{aligned} L(\mu,\sigma;x_1,x_2,\ldots,x_n) &= \prod_{i=1}^{n} \frac{1}{2\sqrt{3}\sigma} I_{[\mu-\sqrt{3}\sigma,\mu+\sqrt{3}\sigma]}(x_i) \\ &= \left(\frac{1}{2\sqrt{3}\sigma}\right)^n I_{[\mu-\sqrt{3}\sigma,x_n]}(x_1) I_{[x_1,\mu+\sqrt{3}\sigma]}(x_n) \\ &= \left(\frac{1}{2\sqrt{3}\sigma}\right)^n I_{[\frac{\mu-x_1}{\sqrt{3}},+\infty]}(\sigma) I_{[\frac{x_n-\mu}{\sqrt{3}},+\infty]}(\sigma). \end{aligned}$$

La funzione $L(\mu,\sigma)$ vale $\left(\frac{1}{2\sqrt{3}\sigma}\right)^n$ nell'area che si trova, in figura 3.6, si al di sopra delle due rette e 0 altrove. Il massimo si ha allora quando σ è minimo, cioè in corrispondenza dell'incrocio fra le rette:

$$\hat{\mu} = \frac{x_n + x_1}{2} \qquad \hat{\sigma} = \frac{x_n - x_1}{2\sqrt{3}}.$$

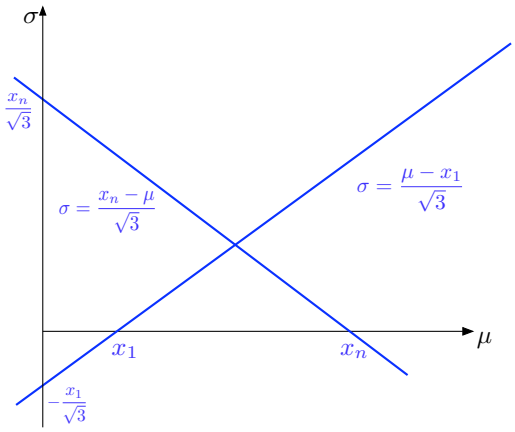

Figura 3.6. *Le rette al di sopra delle quali la funzione di verosimiglianza ha valore positivo*

3.2.4 Stima dell'errore quadratico medio

Siano $X_1, X_2, ..., X_n$ *v.c.* indipendenti con distribuzione F. Indichiamo con $\theta(F)$ un parametro della distribuzione F che si vuole stimare (media, varianza, ...) e con $g(X_1, X_2, ..., X_n)$ lo stimatore che si vuole utilizzare. Definiamo l'*Errore Quadratico Medio*:

$$EQM(F, g) = E_F[(g(X_1, X_2, ..., X_n) - \theta(F))^2].$$

Il problema che ci poniamo è la stima di $EQM(F, g)$. Osserviamo che F non è nota e quindi $E_F[(g(X_1, X_2, ..., X_n) - \theta(F))^2]$ non può essere determinata per via analitica.

Supponiamo di disporre di una realizzazione $(x_1, x_2, ..., x_n)$ delle *v.c.* X_1, X_2, ..., X_n, e definiamo la variabile casuale discreta X_e che assume i valori $x_1, x_2, ..., x_n$ con funzione di distribuzione:

$$F_e(x) = \frac{|\{i : x_i \leq x\}|}{n}.$$

In pratica ordiniamo le x_i

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)},$$

dove $x_{(i)}$ indica l'iesima osservazione in ordine crescente di valore.

Si ha allora

$$F_e(x) = \begin{cases} 0, & \text{se } x < x_{(1)} \\ \frac{j}{n}, & \text{se } x_{(j)} \leq x < x_{(j+1)} \\ 1, & \text{se } x_{(n)} \leq x \end{cases}$$

La distribuzione F_e può essere considerata una *stima empirica* della F; infatti, per la legge dei grandi numeri, è

$$F_e(x) \xrightarrow[n\to\infty]{} F(x).$$

Allora $\theta(F_e)$ è una approssimazione di $\theta(F)$, e

$$EQM(F_e, g) = E_{F_e}[(g(X_1, X_2, ..., X_n) - \theta(F_e))^2]$$

è una approssimazione di $EQM(F, g)$. Poiché F_e è nota, sia $\theta(F_e)$ che $EQM(F_e, g)$ sono calcolabili e quindi è possibile avere una approssimazione di $EQM(F, g)$ tanto più buona quanto più è grande n.

In pratica però il calcolo di $EQM(F_e)$ può risultare notevolmente oneroso. Infatti è

$$EQM(F_e) = \sum_{y \in \{x_1, x_2, ..., x_n\}^n} \frac{(g(y) - \theta(F_e))^2}{n^n},$$

dove con $\{x_1, x_2, ..., x_n\}^n$ abbiamo indicato l'insieme di tutti i vettori ad n componenti i cui elementi possono assumere valori nell'insieme $\{x_1, x_2, ..., x_n\}$.

In pratica vengono generati k vettori $y \in \{x_1, x_2, ..., x_n\}^n$, $y_1, y_2, ..., y_k$, e si pone

$$EQM(F_e) \simeq \sum_{i=1}^{k} \frac{(g(y_i) - \theta(F_e))^2}{k}.$$

Questo modo di procedere si giustifica col fatto che le $(g(y_i) - \theta(F_e))^2$ possono essere considerate come valori assunti da variabili casuali indipendenti con media $EQM(F_e)$, e quindi la loro media è una stima corretta di $EQM(F_e)$.

3.3 Test di ipotesi

3.3.1 Test Chi-Quadro

Siano date n variabili casuali discrete, $X_1, X_2, \ldots, X_n$, assumenti valori 1, 2, $\ldots$, k. Assumiamo tali variabili identicamente distribuite e sia X una *v.c.* che rappresenti ciascuna di esse.

Vogliamo validare la correttezza della seguente ipotesi H_0 (ipotesi nulla)

$$H_0 : P[X = i] = p_i, i = 1, ..., k$$

dove $p_1, p_2, ..., p_k$ sono valori dati con somma 1.

Definiamo le nuove *v.c.* $N_i = |\{j : X_j = i\}|, i = 1, ..., k$. Sotto l'ipotesi H_0, N_i ha distribuzione binomiale con parametri n e p_i, per ogni i, e quindi ha media np_i. Consideriamo ora la grandezza

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

Chiaramente più grande è T meno è probabile che l'ipotesi H_0 sia corretta. Per valori di n grandi T ha approssimativamente una distribuzione *Chi-Quadro* con $k - 1$ gradi di libertà. È quindi

$$P_{H_0}[T \geq t] \cong P[\chi^2_{k-1} \geq t]$$

Se è t il valore assunto da T, $P[\chi^2_{k-1} \geq t]$ fornisce la probabilità di errore nel caso che si decida di scartare l'ipotesi H_0. Valori che vengono considerati ragionevoli per rigettare l'ipotesi sono $P[\chi^2_{k-1} \geq t] = 0.05$ (oppure più conservativamente 0.01).

Una più accurata approssimazione del valore $P_{H_0}[T \geq t]$ può essere ottenuta per mezzo di una simulazione. Si generano a questo scopo le variabili casuali $T^{(1)}, T^{(2)}, ..., T^{(r)}$, ciascuna con la distribuzione di T sotto l'ipotesi H_0, e si pone

$$P_{H_0}[T \geq t] \cong \frac{\left|\{j : T^{(j)} \geq t\}\right|}{r}$$

Al crescere di r migliora la bontà dell'approssimazione.

Consideriamo ora il caso in cui $X_1, X_2, \ldots, X_n$ siano variabili indipendenti identicamente distribuite e l'ipotesi H_0 sia che abbiano una comune distribuzione continua F data.

Possiamo ricondurci al caso precedente suddividendo l'insieme dei possibili valori assunti dalle X_i in k intervalli distinti

$$(-\infty, x_1), (x_1, x_2), ..., (x_{k-2}, x_{k-1}), (x_{k-1}, +\infty).$$

Si considerano quindi le *v.c.* discrete X_i^d con $X_i^d = i$ se X_i si trova nell'intervallo (x_{i-1}, x_i), e l'ipotesi H_0 diviene $P[X_i^d = i] = F(x_i) - F(x_{i-1})$, $i = 1, \ldots, k$.

Si pone il problema di come scegliere gli intervalli in modo da garantire la validità del test. Una ragionevole scelta è quella di intervalli equiprobabili, cioè tali che risulti $F(x_1) - F(x_0) = F(x_2) - F(x_1) = \ldots = F(x_k) - F(x_{k-1})$, e tali che il numero di osservazioni che ricade in ciascuno di essi non sia troppo piccolo (ad esempio non inferiore a 5)[2].

3.3.2 Test di Kolmogorov-Smirnov per distribuzioni continue

Il caso in cui $X_1, X_2, \ldots, X_n$ sono variabili indipendenti identicamente distribuite e l'ipotesi H_0 che vogliamo valutare è che abbiano una comune distribuzione continua F data, può essere trattato in modo diretto e più efficiente. A questo scopo utilizziamo l'approssimazione empirica F_e della F, costruita come già visto nel paragrafo 3.2.4:

$$F_e(x) = \frac{|\{i : X_i \leq x\}|}{n}$$

Se l'ipotesi H_0 è corretta allora $F_e(x)$ è una buona approssimazione di $F(x)$. Una misura dello scostamento è

$$D = max_x \left| F_e(x) - F(x) \right|.$$

Dati i valori osservati $x_1, \ldots, x_n$ di $X_1, \ldots, X_n$, ricaviamo il valore osservato d di D. Essendo

$$\begin{aligned} Max_x\{F_e(x) - F(x)\} &= Max\{\frac{j}{n} - F(x_{(j)}) : j = 1 \ldots n\}, \\ Max_x\{F(x) - F_e(x)\} &= Max\{F(x_{(j)}) - \frac{j-1}{n} : j = 1 \ldots n\}, \end{aligned}$$

si ha

[2]Per una trattazione più approfondita di questo punto rinviamo a **?**.

$$d = max\left\{\frac{j}{n} - F(x_{(j)}), F(x_{(j)}) - \frac{j-1}{n} : j = 1, \dots n\right\}.$$

Se conoscessimo la probabilità $P_F(D \geq d)$ nell'ipotesi che la distribuzione vera sia F, avremmo la probabilità di fare un errore se decidessimo di rigettare H_0.

Osserviamo che

$$\begin{aligned} P_F[D \geq d\,] &= P\left[Max_x \left|\frac{|\{i : X_i \leq x\}|}{n} - F(x)\right| \geq d\right] \\ &= P\left[Max_x \left|\frac{|\{i : F(X_i) \leq F(x)\}|}{n} - F(x)\right| \geq d\right] \\ &= P\left[Max_x \left|\frac{|\{i : U_i \leq F(x)\}|}{n} - F(x)\right| \geq d\right] \end{aligned}$$

dove $U_1, U_2, \dots, U_n$ sono variabili casuali indipendenti uniformi in (0.1). La prima uguaglianza deriva dal fatto che la funzione F è monotona crescente; la seconda dal fatto che se X è una *v.c.* con distribuzione continua F, allora $F(X)$ è una *v.c.* uniforme in $(0, 1)$. Infatti, ponendo $Y = F(X)$ si ha $P[Y \leq y] = P[X \leq F^{-1}(y)] = y$.

La distribuzione di D sotto H_0 non dipende quindi da F ed è

$$P[D \geq d\,] = P\left[Max_{0 \leq y \leq 1} \left|\frac{|\{i : U_i \leq y\}|}{n} - y\right| \geq d\right].$$

Possiamo allora stimare $P[D \geq d\,]$ iterando il seguente procedimento:

1. Si generano $u_1, u_2, \dots, u_n$, uniformi in $(0, 1)$,
2. Si calcola $Max_{0 \leq y \leq 1} \left|\frac{|\{i:u_i \leq y\}|}{n} - y\right| = Max\left\{\frac{j}{n} - u_{(j)}, u_{(j)} - \frac{j-1}{n} : j = 1, ..., n\right\}$.

Si ripete più volte e si prende come valore per $P[D \geq d\,]$ la proporzione di volte in cui il valore trovato risulta $\geq d$.

Se $P[D \geq d]$ è sufficientemente basso (es. 0.05) l'ipotesi viene rigettata, altrimenti viene accettata.

3.3.3 Il test della somma dei ranghi

Consideriamo una grandezza rilevante del sistema che si vuole modellare, e siano date m osservazioni, $Y_1, Y_2, \ldots, Y_m$, di questa grandezza (ad esempio i tempi totali di attesa in m giorni). Sotto opportune ipotesi le Y_i possono essere considerate come *v.c.* identiche e indipendenti.

Siano $X_1, X_2, \ldots, X_n$ i valori forniti dalla simulazione per la stessa grandezza in n esecuzioni del modello. Anche le X_i saranno *v.c.* identicamente distribuite e indipendenti, con distribuzione F (in generale non nota). L'ipotesi H_0 da verificare è che anche le Y_i abbiano la stessa distribuzione, cioè che

$$X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$$

siano *v.c.* identicamente distribuite e indipendenti.

Operiamo come segue: ordiniamo le $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ in ordine crescente di valore e per $i = 1, \ldots, n$ sia R_i il rango di X_i, cioè la sua posizione nella lista ordinata.

Ad esempio se le sequenze sono:
$X : 20, 15, 38, 40, 35, 31$

$Y : 25, 30, 29, 34.$
Si ha
$R_1 = 2$, $R_2 = 1$, $R_3 = 9$, $R_4 = 10$, $R_5 = 8$ e $R_6 = 6$.
Consideriamo ora la quantità

$$R = \sum_{i=1}^{n} R_i$$

($R = 36$ nell'esempio precedente)

Chiaramente un valore troppo piccolo o troppo grande di R falsificherebbe con alta probabilità l'ipotesi H_0. Supponendo di ritenere accettabile una probabilità α (ad es. 0.05) di sbagliare nel rigettare l'ipotesi, rigetteremo H_0 se risulta

$$2 \min \left\{ P_{H_0} \left[R \leq r \right], P_{H_0} \left[R \geq r \right] \right\} \leq \alpha.$$

Si pone allora il problema di determinare la distribuzione di R. Ponendo $F_{n,m}(r) = P_{H_0}[R \leq r]$, vale la seguente equazione ricorsiva

$$F_{n,m}(r) = \frac{n}{n+m} F_{n-1,m}(r - n - m) + \frac{m}{n+m} F_{n,m-1}(r),$$

con

$$F_{1,0}(r) = \begin{cases} 0, & \text{se } r < 1 \\ 1, & \text{se } r \geq 1 \end{cases}$$
$$F_{0,1}(r) = \begin{cases} 0, & \text{se } r < 0 \\ 1, & \text{se } r \geq 0 \end{cases}$$

Si ha allora un sistema di equazioni ricorsive che consente di calcolare $F_{n,m}(r)$ e quindi la distribuzione di R.

In pratica il calcolo di $F_{n,m}(r)$ utilizzando la formula ricorsiva risulta molto oneroso. Si ricorre allora ad una approssimazione di $F_{n,m}(r)$.

È possibile dimostrare che

$$\frac{R - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

è, approssimativamente, per n ed m grandi, una normale standard, $N(0,1)$. Pertanto è

$$P[R \leq r] \cong P[Z \leq r*]$$

con Z una *v.c.* $N(0,1)$ e

$$r* = \frac{r - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}.$$

3.4 Modelli di processi di arrivo

In molti casi ci si trova ad analizzare sistemi caratterizzati da arrivi di entità, come ad esempio nel caso dei clienti che arrivano ad un ufficio postale. Abbiamo già visto come la distribuzione di Poisson svolga un ruolo rilevante in questi casi. Si parla in situazioni di questo tipo di *processi di Poisson* e si distingue tra due tipi di processi, quelli stazionari e quelli non stazionari.

Nel seguito descriveremo brevemente i due tipi di processi, indicando con λ il numero medio di arrivi nell'unità di tempo, e con $N(t)$ il numero di arrivi nell'intervallo temporale [0, t].

Un processo di Poisson si dice *stazionario* quando valgono le seguenti proprietà:

1. arriva un individuo alla volta, cioè non ci sono arrivi di gruppi di individui;

2. il numero di arrivi nell'intervallo $(t, t+s]$, $N(t+s)-N(t)$, è indipendente da $N(u)$, per ogni $u \in [0, t]$;

3. la distribuzione di $N(t+s)-N(t)$ è indipendente da t per ogni $(t, s) \geq 0$.

Secondo quanto abbiamo già visto nel paragrafo 3.1.1, gli arrivi in queste ipotesi sono rappresentabili per mezzo di una variabile casuale con distribuzione di Poisson, cioè:

$$P[N(t+s) - N(t) = k] = \frac{e^{\lambda s}(\lambda s)^k}{k!}, \qquad k = 0, 1, 2, \ldots, \quad t, s \geq 0.$$

Ricordiamo che, come abbiamo già visto, i tempi di interarrivo sono *v.c.* con distribuzione esponenziale e media λ.

In molti casi reali il numero medio di arrivi nell'unità di tempo non è indipendente dal tempo: il numero di clienti che si presenta ad uno sportello non è in media lo stesso in ogni intervallo temporale. Possiamo allora sostituire alla costante λ una funzione del tempo $\lambda(t)$. In questo caso si parla di *processo di Poisson non stazionario*. Un tale processo è definito dalle seguenti condizioni:

1. arriva un individuo alla volta, cioè non ci sono arrivi di gruppi di individui;

2. il numero di arrivi nell'intervallo $(t, t+s]$, $N(t+s)-N(t)$, è indipendente da $N(u)$, per ogni $u \in [0, t]$.

La distribuzione degli arrivi è sempre poissoniana, ma con parametro variabile. Si ha allora:

$$P[N(t+s) - N(t) = k] = \frac{e^{b(t,s)} b(t,s)^k}{k!}, \qquad k = 0, 1, 2, \ldots, \quad t, s \geq 0.$$

Indicando con $\Lambda(t)$ la media di $N(t)$, cioè

$$\Lambda(t) = E[N(t)],$$

possiamo scrivere

$$b(t, s) = \Lambda(t + s) - \Lambda(t).$$

Nel caso in cui $\Lambda(t)$ è differenziabile si ha

$$\lambda(t) = \frac{d\Lambda(t)}{dt},$$

e quindi

$$b(t,s) = \Lambda(t+s) - \Lambda(t) = \int_t^{t+s} \lambda(y)dy.$$

Per potere stimare la funzione $\lambda(t)$ è necessario disporre delle serie temporali corrispondenti a più giorni. Si supponga di conoscere gli arrivi in un intervallo di tempo T per n giorni. Assumiamo che non ci sia motivo di ritenere che il comportamento degli arrivi sia diverso da un giorno all'altro. Dividiamo l'intervallo T in p intervallini di uguale ampiezza Δ, $[t_1, t_2], [t_2, t_3], \ldots, [t_p, t_{p+1}]$.

Sia x_{ij} il numero di arrivi nell'intervallo i del giorno j. Possiamo allora calcolare la media del numero di arrivi in ciascuno degli intervallini:

$$\bar{x}_i = \frac{\sum_j x_{ij}}{n},$$

e di conseguenza costruire una approssimazione della funzione $\lambda(t)$:

$$\bar{\lambda}(t) = \frac{\bar{x}_i}{\Delta}, \qquad t \in [t_i, t_{i+1}], \;\; i = 1, 2, \ldots, p.$$

Naturalmente Δ non dovrà essere troppo piccolo, altrimenti risulterebbe priva di senso la media $\bar{x}_i$, né troppo grande, altrimenti non si riuscirebbe a catturare la variabilità di $\lambda(t)$.

In diversi casi i clienti arrivano a gruppi. Tipico è il caso degli arrivi ad un ristorante. In questo caso possiamo pensare di operare a due livelli. Ad un primo livello consideriamo i gruppi come gli individui che arrivano. L'arrivo dei gruppi può essere considerato come un processo di tipo poissoniano, e $N(t)$ è il numero dei gruppi che arrivano entro il tempo t.

Definiamo poi, per ogni gruppo i, la variabile casuale discreta, B_i, che può assumere valori 1, 2, ... Tale variabile definisce la cardinalità del gruppo. Il numero di arrivi individuali entro il tempo t è allora dato dalla:

$$X(t) = \sum_{i=1}^{N(t)} B_i, \quad t \geq 0.$$

Capitolo 4

Analisi e scelta dei dati di input

4.1 Introduzione

Per l'esecuzione di una simulazione è necessario disporre di dati di input che siano una adeguata rappresentazione di ciò che accadrà in realtà nel sistema oggetto di studio. Ad esempio se stiamo simulando il funzionamento di un ambulatorio per un periodo di un anno, avremo bisogno di generare per ogni giorno un flusso di clienti, che per caratteristiche (tipo di trattamento richiesto) e per distribuzione temporale, sia il più realistico possibile.

In generale le caratteristiche dell'input possono essere rappresentate per mezzo di opportune variabili casuali (ad esempio la *v.c.* tempo di interarrivo fra due clienti successivi), e possiamo ragionevolmente supporre che su queste variabili siano disponibili dei dati sperimentali; dati raccolti durante il funzionamento del sistema da simulare, se già esistente, oppure dati relativi a sistemi simili nel caso che la simulazione riguardi un sistema da realizzare.

Possiamo pensare, in prima istanza, a tre approcci alternativi di uso dei dati disponibili per la preparazione dell'input della simulazione.

1. I dati disponibili vengono utilizzati direttamente nella simulazione.

2. I dati disponibili vengono usati per costruire una funzione di distribuzione empirica che verrà poi usata per generare l'input della simulazione.

3. Si utilizzano tecniche statistiche per derivare dai dati una funzione di distribuzione *teorica* che rappresenti bene il loro andamento e per

stimarne i parametri; questa distribuzione sarà poi usata nella simulazione.

Il primo approccio ha senso nel caso in cui sia facile raccogliere grandi quantità di dati rappresentativi delle effettive condizioni di funzionamento del sistema sotto esame. Se, ad esempio, stiamo studiando la politica di gestione degli accessi ad una memoria a dischi in un sistema di calcolo, possiamo facilmente disporre di lunghe sequenze di *queries* rilevate nel funzionamento di sistemi di calcolo esistenti, sotto diverse condizioni di uso. Tuttavia con questo approccio c'è sempre il rischio di riprodurre solamente ciò che è avvenuto nel passato, perdendo la possibilità di valutare il funzionamento del sistema in condizioni diverse e non previste.

Il secondo approccio è in genere preferibile, essendo meno condizionato dalla abbondanza dei dati disponibili. Questo approccio è particolarmente utile in fase di validazione del modello, quando si vogliono confrontare gli output del modello e quelli del sistema reale.

Se si riesce a stimare una distribuzione teorica che rappresenti bene i dati osservati, allora il terzo approccio è preferibile per le seguenti ragioni:

- Una distribuzione empirica può mostrare irregolarità (dovute ad esempio al numero limitato di dati), mentre una distribuzione teorica tende a "regolarizzare" i dati.

- Al contrario di una distribuzione empirica, una distribuzione teorica consente di generare valori delle variabili casuali che siano al di fuori dell'intervallo dei valori osservati.

- Una distribuzione teorica costituisce un modo molto compatto per rappresentare i valori dei dati di input, mentre l'uso di distribuzioni empiriche richiede il mantenimento in memoria di grandi quantità di dati.

Sia che si usi una distribuzione empirica oppure una distribuzione teorica (ad esempio una di quelle viste nel paragrafo 3.1), è necessario disporre della capacità di generare numeri casuali con la voluta distribuzione. Ad esempio nel problema della riparazione e manutenzione dei dischi (paragrafo 1.2.3) abbiamo assunto la disponibilità di una sequenza di numeri casuali uniformemente distribuiti tra 0 ed 1 che ci è servito per generare una sequenza di numeri tra 1 e 6 aventi la funzione di probabilità richiesta. Nel prossimo paragrafo verrà mostrato come si può derivare una distribuzione empirica dai

dati disponibili. Successivamente affronteremo il problema della individuazione della distribuzione che corisponde ai dati e della generazione di variabili casuali con la distribuzione scelta.

4.2 Distribuzioni empiriche

Supponiamo di disporre di un insieme di osservazioni di una data variabile casuale, $X_1, X_2, \ldots, X_n$, e di volere costruire a partire da esse una distribuzione continua. Supponiamo che i valori siano tutti distinti, ed indichiamo con $X_{(i)}$ la *iesima* osservazione in ordine crescente di valore, cioè sia $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$.[1]

Consideriamo gli $n-1$ intervalli del tipo $[X_{(i)}, X_{(i+1)})$, ed assumiamo che la distribuzione all'interno dell' intervallo i sia uniforme con densità $\frac{1}{(n-1)(X_{(i+1)}-X(i))}$, con $i = 1, 2, \ldots, n-1$.

Allora la distribuzione empirica continua F è data dalla:

$$F(x) = \begin{cases} 0 & , x < X_{(1)}, \\ \frac{i-1}{n-1} + \frac{(x-X_{(i)})}{(n-1)(X_{(i+1)}-X_{(i)})} & , X_{(i)} \leq x < X_{(i+1)}, i = 1, \ldots, n-1, \\ 1 & , X_{(n)} \leq x. \end{cases}$$

Ad esempio le osservazioni 0.6, 1, 3.4, 2, 2.5, 4, 3.2 danno origine alla distribuzione empirica di figura 4.1.

Osserviamo che la distribuzione empirica così costruita è diversa da quella introdotta nel precedente capitolo. La differenza sta nel fatto che questa distribuzione, a differenza di quella, è continua, e presumibilmente approssima meglio la distribuzione cercata, permettendo di avere valori di $F(x)$ distinti e crescenti all'interno di ciascun intervallo $[x_j, x_{j+1})$.

In certi casi non si dispone di singole osservazioni, ma si conosce solamente quante osservazioni cadono in ciascuno di k intervalli contigui, $[a_0, a_1)$, $[a_1, a_2)$, $\ldots$, $[a_{k-1}, a_k)$. In questo caso possiamo ragionevolmente assumere la distribuzione in ciascun intervallo uniforme con densità $n_i/n(a_i - a_{i-1})$, dove n è il numero totale delle osservazioni e n_i è il numero di esse che cadono nell'*iesimo* intervallo.

[1] Nel caso in cui alcuni valori fossero coincidenti possiamo riportarci al caso di valori distinti con tecniche di perturbazione, cioè modificando alcuni di essi (in più o in meno) di quantità opportunamente piccole.

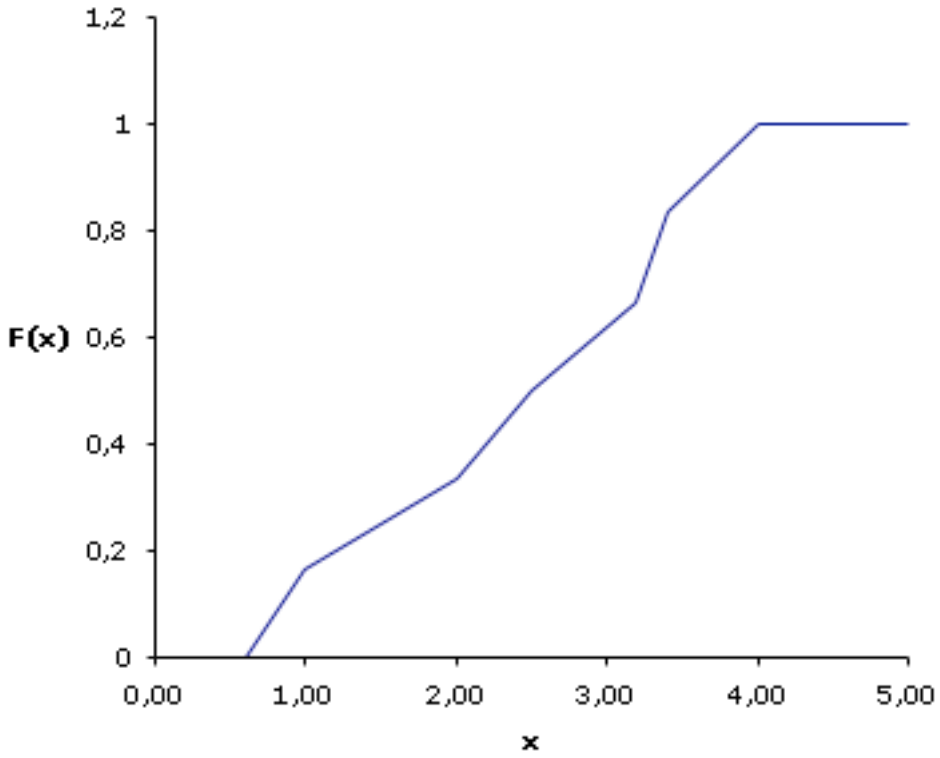

Figura 4.1. *Distribuzione empirica*

4.3 Analisi dei dati di input

4.3.1 Indipendenza delle osservazioni

Una ipotesi essenziale nelle stime di parametri discusse nel precedente capitolo è che le osservazioni siano indipendenti. Questa è un'ipotesi che nella realtà può non essere soddisfatta. Ad esempio, se rileviamo i tempi di attesa in una coda di clienti successivi abbiamo delle osservazioni strettamente correlate.

Siano $X_1, X_2, \ldots, X_n$ le nostre osservazioni. Un'idea della loro indipendenza la possiamo ottenere analizzando la correlazione fra le diverse osservazioni[2]. Indichiamo con $\overline{\rho}_j$ la stima del coefficiente di correlazione fra

[2]Ricordiamo che, date due variabili casuali X e Y, il loro coefficiente di correlazione è

osservazioni distanti j posizioni nella sequenza:

$$\overline{\rho_j} = \frac{\sum_{i=1}^{n-j}(X_i - \overline{X}_n)(X_{i+j} - \overline{X}_n)}{(n-j)S_n^2}$$

Chiaramente nel caso di osservazioni indipendenti ci si aspetta che $\overline{\rho}_j$ sia molto prossimo a zero. L'esame del grafico di $\overline{\rho}_j$ al variare di j dà una buona idea dell'indipendenza delle osservazioni. Una conferma può essere ottenuta esaminando il grafico che si ottiene riportando su un piano i punti (X_i, X_{i+1}), per $i = 1, 2, \ldots, n-1$. In caso di osservazioni indipendenti i punti dovrebbero risultare distribuiti casualmente sul piano; invece in presenza di correlazioni essi appariranno concentrati intorno a rette con pendenza positiva o negativa a seconda del tipo di correlazione.

4.3.2 Individuazione della distribuzione

Il passo successivo, una volta verificata l'indipendenza delle osservazioni, è quello della individuazione del tipo di distribuzione da scegliere per la variabile di input sotto esame. Tale individuazione può essere ottenuta in certi casi da una conoscenza *a priori* del tipo di fenomeno da cui la variabile casuale deriva. Un tipico esempio sono le considerazioni svolte a proposito della distribuzione di Poisson.

Più spesso si ricorre alla stima di opportuni parametri che ci forniscono un'idea delle caratteristiche della distribuzione ed all'esame dell'andamento delle osservazioni per mezzo di grafici.

Un confronto della *media* e della *mediana* può farci capire se è ragionevole o no considerare la distribuzione simmetrica (come ad esempio nel caso della normale): infatti nel caso di distribuzioni continue simmetriche media e mediana coincidono. Nel caso di distribuzioni discrete ciò non è vero quando il numero di valori distinti che la variabile può assumere è pari. Ricordiamo che la mediana di una variabile casuale X con distribuzione F_X è il più piccolo valore x per cui risulta $F_X(x) \geq 0.5$. Dobbiamo naturalmente sempre tenere presente che noi disponiamo solo di stime dei parametri, pertanto non possiamo aspettarci che anche nel caso di distribuzioni continue e simmetriche le stime della media e della varianza siano esattamente uguali.

Parametri che misurano la variazione di una *v.c.* sono il rapporto σ/μ ed il rapporto σ^2/μ. Il primo ha la caratteristica di essere uguale ad 1 per la

$\rho_{XY} = \frac{Cov[X,Y]}{\sigma_X \sigma_Y}$, dove $Cov[X,Y] = E[(X - \mu_X)(Y - \mu_Y)]$ è la covarianza di X e Y.

distribuzione esponenziale e < 1 per la gamma con $n > 1$, mentre il secondo è 1 per la distribuzione di Poisson e < 1 per la binomiale. Una stima di tali parametri può quindi dare utili indicazioni.

Infine è molto utile tracciare un istogramma dei valori assunti dalla variabile casuale nel campione di cui disponiamo. A questo scopo si suddivide l'intervallo tra il minimo ed il massimo dei valori assunti in intervalli disgiunti di uguale ampiezza, $[b_0, b_1), [b_1, b_2), \ldots, [b_{k-1} - b_k)$, con $\Delta = b_k - b_{k-1}$. Si definisce quindi la funzione $h(x)$:

$$h(x) = \begin{cases} 0, & \text{se} \quad x < b_0 \\ h_j, & \text{se} \quad b_{j-1} \leq x < b_j, \quad j = 1, 2, \ldots, k \\ 0, & \text{se} \quad b_k \leq x \end{cases}$$

dove h_j è il numero di osservazioni che cadono nel *jesimo* intervallo diviso il numero totale di osservazioni.

Il grafico della funzione $h(x)$ può dare una buona idea del tipo di distribuzione che ha la variabile casuale in esame.

Nel caso di variabili discrete si può ottenere lo stesso effetto tracciando su un grafico i punti $(n_j/n, x_j)$, dove x_j è il *jesimo* valore assunto nel campione dalla variabile casuale, n_j è il numero di occorrenze di tale valore, e n è la cardinalità del campione.

4.3.3 Stima dei parametri della distribuzione

Una volta individuata la distribuzione è necessario determinarne i parametri. Ad esempio se la distribuzione è una esponenziale, allora bisogna determinare il valore di λ. Uno degli approcci più usati per la determinazione dei parametri di una distribuzione è quello della *massima verosimiglianza* presentato capitolo 3.

Una volta stimati i parametri, una verifica di quanto la distribuzione scelta approssima la distribuzione dei dati nel campione può essere effettuata con uno dei test visti nel capitolo 3.

4.4 Numeri pseudocasuali

Il problema che affrontiamo qui è quello dei metodi che possono essere usati per generare sequenze di numeri casuali. Metodi meccanici caratterizzati da

un'intrinseca casualità, quali il lancio di un dado, possono portare alla produzione di numeri casuali con la voluta distribuzione. Tali metodi però sono impraticabili nella simulazione, dove si richiede la generazione in tempi molto piccoli di lunghe sequenze di numeri. Si ricorre pertanto alla generazione su calcolatore di numeri cosiddetti *pseudocasuali*. Si tratta di sequenze di numeri generati deterministicamente, e quindi "per nulla casuali", ma aventi proprietà statistiche che approssimano bene quelle di sequenze di numeri realmente casuali. Si tratta quindi di sequenze che all'analisi statistica risultano indistinguibili da sequenze di numeri casuali.

4.4.1 Numeri pseudocasuali con distribuzione uniforme

Esistono diverse tecniche per generare numeri pseudocasuali con distribuzione uniforme. I metodi più frequentemente usati sono i cosiddetti *metodi congruenziali*. In questi metodi una sequenza viene generata per mezzo della seguente formula:

$$X_{i+1} = aX_i + c \pmod{m}.$$

Se c è zero, il metodo viene detto *moltiplicativo*, altrimenti si parla di metodo *misto*. Un generatore di questo tipo genera al più m numeri distinti (m-1 se $c = 0$, poiché lo zero non può apparire nella sequenza generata), nell'intervallo $[0, m-1]$, e la sequenza generata è periodica. Il generatore ha periodo pieno se ha periodo m, e quindi genera tutti i numeri compresi tra 0 e m-1. Dividendo poi i numeri generati per m, si ottengono numeri compresi nell'intervallo [0,1).

Naturalmente ogni valore nell'intervallo deve avere la stessa probabilità di essere presente. Si vuole inoltre che tutti i numeri, le coppie e le triple abbiano la stessa probabilità di comparire in qualsiasi porzione della sequenza generata. Il primo numero della sequenza, X_0, è detto il *seme*. La scelta del seme è importante al fine di assicurare che la sequenza abbia un periodo sufficientemente lungo. Ad esempio nel caso di un generatore moltiplicativo ($c = 0$), X_0 ed m devono essere primi fra loro. Sempre nel caso di generatori moltiplicativi, delle scelte che garantiscono delle sequenze con buone proprietà statistiche sono, nel caso di macchine a 32 bit, $m = 2^{31} - 1$ e $a = 7^5 = 16,807$.

Esempio Usando un generatore moltiplicativo e $a = 3$, $X_0 = 3$ e $m = 7$ si ha:

$$\begin{aligned} X_1 &= 9 (\text{mod } 7) &= 2 \\ X_2 &= 6 (\text{mod } 7) &= 6 \\ X_3 &= 18 (\text{mod } 7) &= 4 \\ X_4 &= 12 (\text{mod } 7) &= 5 \\ X_5 &= 15 (\text{mod } 7) &= 1 \\ X_6 &= 3 (\text{mod } 7) &= 3 \\ X_7 &= 9 (\text{mod } 7) &= 2 \end{aligned}$$

La disponibilità nei linguaggi di programmazione più usati ed in particolare in quelli orientati alla simulazione di ottimi generatori di numeri casuali con distribuzione uniforme permette di considerare uno studio più approfondito dei generatori di numeri pseudocasuali e dei metodi statistici per analizzarne la qualità come al di là degli obiettivi del corso. Per un approfondimento dell'argomento rimandiamo ai testi di Law e Kelton **?** e di Knuth **?**.

Nel seguito vedremo come, a partire da una sequenza di numeri pseudocasuali distribuiti uniformemente nell'intervallo [0, 1), sia possibile generare numeri pseudocasuali con la distribuzione voluta.

4.4.2 Distribuzioni discrete

Inversione

Sia Y una variabile casuale discreta, che può assumere i valori $y_1 < y_2 < y_3 < \ldots$, e siano

$$\begin{aligned} f_Y(y_i) &= p_i, \\ F_Y(y_i) &= \sum_{j \leq i} p_j. \end{aligned}$$

le corrispondenti funzioni di densità e distribuzione.

Sia U una variabile casuale uniforme in [0, 1). A partire da essa definiamo la variabile casuale X:

$$X(U) = Max\{y_i : U \in [F_Y(y_{i-1}), F_Y(y_i)]\},$$

avendo posto $F_Y(y_0) = 0$.

Abbiamo così costruito una variabile casuale che ha la stessa distribuzione della variabile Y a partire da una distribuzione uniforme. Infatti è:

$$P[X = y_i] = P[F_Y(y_{i-1}) \leq U < F_Y(y_i)] = p_i,$$

dove l'ultima uguaglianza deriva dal fatto che U è uniforme in [0, 1).

È possibile allora dato una sequenza si numeri compresi tra 0 ed 1, con distribuzione uniforme, generare una corrispondente sequenza di numeri appartenenti ad un dato insieme discreto, con una prefissata distribuzione di probabilità.

In pratica dal punto di vista computazionale si sfrutta il fatto che per le distribuzione discrete esprimibili analiticamente come quelle viste nel paragrafo 3.1.1 è

$$\begin{aligned} P[X = i+1] &= a(i+1)P[X = i], \\ P[X \leq i+1] &= P[X \leq i] + P[X = i+1], \end{aligned}$$

dove $a(i)$ una opportuna funzione.

Ad esempio, nel caso di una distribuzione di Poisson,

$$a(i) = \lambda/i.$$

Un algoritmo generale di inversione itera le seguenti operazioni, che ad ogni iterazione forniscono un numero casuale con la voluta distribuzione.

$k := 0$,

$P := P[X = 0]$,

$S := P$,

Estrai $u \in [0, 1]$ (variabile pseudocasuale uniforme),

While$(u > S)$ **do**

$k := k + 1$,
$P := a(k)P$,
$S := S + P$,

Restituisci k.

Distribuzione di Poisson

Oltre al metodo di inversione del precedente paragrafo, si può usare la relazione fra distribuzione di Poisson e distribuzione esponenziale.

Ricordiamo che, se X una *v.c.* con distribuzione di Poisson e media λ, e Y una *v.c.* con distribuzione esponenziale e media $1/\lambda$, allora la prima fornisce il numero di eventi nell'unità di tempo e la seconda il tempo fra un evento ed il successivo.

Possiamo allora generare una sequenza di numeri pseudocasuali con distribuzione esponenziale, $y_1, y_2, ...$, e fermarci non appena risulti

$$\sum_{1}^{k+1} y_i > 1 \geq \sum_{1}^{k} y_i;$$

Si pone $x_1 = k$, e si ripete generando successivamente $x_2, x_3, ...$

Abbiamo così ottenuto una sequenza di numeri pseudocasuali con distribuzione di Poisson.

4.4.3 Distribuzioni continue

Inversione

Sia Y una variabile casuale continua, con funzioni di densità e distribuzione f_Y e F_Y, rispettivamente.

Sia U una variabile casuale uniforme in $[0, 1)$. A partire da essa, definiamo la variabile casuale $X = F_Y^{-1}(U)$. Cioè, per ogni valore u assunto da U, il corrispondente valore di X è $x = F_Y^{-1}(u)$.

Si ha allora

$$\begin{aligned} F_X(x) &= P[X \leq x] = P[F_Y^{-1}(U) \leq x] \\ &= P[U \leq F_Y(x)] = F_Y(x); \end{aligned}$$

La terza uguaglianza deriva dal fatto che la funzione di distribuzione è monotona e pertanto $[F_Y^{-1}(U) \leq x] \Rightarrow [F_Y(F_Y^{-1}(U)) \leq F_Y(x)]$. L'ultima uguaglianza deriva dal fatto che U è uniforme.

Abbiamo così costruito una variabile casuale che ha la stessa distribuzione della variabile Y a partire da una distribuzione di probabilità uniforme. È possibile allora data una sequenza si numeri compresi tra 0 ed 1, con

distribuzione uniforme, generare una corrispondente sequenza di numeri appartenenti ad un dato insieme discreto, con una prefissata distribuzione di probabilità.

Ad esempio supponiamo di volere costruire una sequenza di numeri pseudocasuali con distribuzione esponenziale.

Se Y è una variabile casuale con distribuzione esponenziale, è

$$F_Y(y) = 1 - e^{-\lambda y},$$

e quindi

$$F_Y^{-1}(x) = -\frac{1}{\lambda}\ln(1-x).$$

Pertanto se U è una *v.c.* con distribuzione uniforme in [0,1), allora

$$X = F_Y^{-1}(U) = -\frac{1}{\lambda}\ln(1-U)$$

è una *v.c.* con distribuzione esponenziale che assume valori $[0, \infty)$.

Data una sequenza di numeri pseudocasuali con distribuzione uniforme in $[0, 1)$, possiamo allora derivare una sequenza con distribuzione esponenziale.

Distribuzione normale

Siano $X_1, X_2, ..., X_n$, *v.c.* indipendenti, aventi la stessa distribuzione, con:

$$E[X_i] = \mu, Var[X_i] = \sigma^2, i = 1, ..., n.$$

Consideriamo la *v.c.*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

con

$$E[\overline{X}_n] = \mu, Var[\overline{X}_n] = \frac{1}{n}\sigma^2.$$

Introduciamo ora la variabile casuale

$$Z_n = \frac{\overline{X}_n - E[\overline{X}_n]}{\sqrt{Var[\overline{X}_n]}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

che, per il *Teorema del limite centrale*, al crescere di n, tende ad una *v.c.* con *distribuzione normale standard*, $N(0, 1)$.

Se le X_i sono distribuite in modo uniforme fra 0 e 1, si ha che $\mu = \frac{1}{2}$ e $\sigma^2 = \frac{1}{12}$, e quindi:

$$Z_n = \frac{\overline{X}_n - \frac{1}{2}}{1/\sqrt{12n}}. \tag{4.1}$$

Pertanto per ottenere una variabile normale standard basterà generare sequenze di n numeri casuali con distribuzione uniforme ed utilizzare poi la (4.1). In pratica un ragionevole valore per n è 12. Se poi si vuole ottenere una v.c. con media $\overline{\mu}$ e varianza $\overline{\sigma}^2$, basterà moltiplicare per $\overline{\sigma}$ il valore ottenuto e sommare $\overline{\mu}$.

Quest'approccio alla generazione di distribuzioni normali non fornisce buoni risultati né dal punto di vista della qualità delle sequenze di numeri generati né da quello della efficienza computazionale.

È preferibile il metodo che viene brevemente descritto nel seguito, detto *metodo polare* (per una trattazione più approfondita rimandiamo a **?** e **?**).

Definiamo le due variabili casuali

$$\begin{aligned} V_1 &= 2U_1 - 1, \\ V_2 &= 2U_2 - 1, \end{aligned}$$

dove U_1 e U_2 sono *v.c.* uniformi in [0, 1). Chiaramente V_1 e V_2 saranno uniformi in (-1, 1).

Data una sequenza di coppie (V_1, V_2), tratteniamo solo quelle per cui è $V_1^2 + V_2^2 \leq 1$. Avremo così costruito una *v.c.* a due dimensioni uniformemente distribuita nel cerchio unitario di raggio 1.

Ponendo $S = V_1^2 + V_2^2$, si ha che le variabili casuali X e Y così definite:

$$\begin{aligned} X &= V_1 \sqrt{\frac{-2 \lg S}{S}} \\ Y &= V_2 \sqrt{\frac{-2 \lg S}{S}} \end{aligned}$$

sono *v.c.* normali indipendenti con media 0 e varianza 1.

Metodo della reiezione

Si tratta di un metodo di semplice implementazione che permette di generare sequenze di numeri casuali, $x_1, x_2, ..., x_n, ...$, aventi una data funzione di densità, f_X. Si assume che tale funzione sia limitata e definita in un intervallo prefissato $[a, b]$:

$$0 \leq f_X(x) \leq M, \quad \text{per } a \leq x \leq b$$
$$f_X(x) = 0, \qquad \text{altrove.}$$

Vogliamo costruire una sequenza di numeri casuali, $x_1, x_2, ..., x_n, ...$, aventi questa funzione di densità.

Si può procedere iterando la seguente operazione, partendo con $i = 1$:

1. si genera una coppia di numeri pseudocasuali uniformi (r, s) con $r \in [a, b]$, e $s \in [0, M]$;
2. se $0 \leq s \leq f_X(r)$, allora si pone $x_i = r$.

Capitolo 5

Analisi dei dati di output

Una fase essenziale di ogni studio di simulazione è l'analisi dei risultati della simulazione stessa. Supponiamo di avere costruito il modello di un sistema e siano $Y_1, Y_2, \ldots, Y_m$ i dati di output che ci interessa studiare. Ad esempio Y_i rappresenti la lunghezza di una coda alla fine dell'*iesimo* intervallo di tempo in cui è stata suddivisa la giornata. Chiaramente Y_i può essere pensata come una variabile casuale. La difficoltà qui sta nel fatto che le variabili casuali $Y_1, Y_2, \ldots, Y_m$ non sono in generale indipendenti; quindi i metodi visti nei precedenti capitoli non possono essere direttamente utilizzati.

Se supponiamo però di avere effettuato n diversi *run* di simulazioni utilizzando diverse sequenze di numeri pseudocasuali, abbiamo diverse sequenze di realizzazioni delle variabili casuali $Y_1, Y_2, \ldots, Y_m$:

$$\begin{array}{ccccccc} y_{11}, & \ldots, & y_{1i}, & \ldots, & y_{1m} \\ y_{21}, & \ldots, & y_{2i}, & \ldots, & y_{2m} \\ \vdots & & & & \vdots \\ y_{n1}, & \ldots, & y_{ni}, & \ldots, & y_{nm} \end{array} \tag{5.1}$$

Una sequenza $y_{1i}, y_{2i}, \ldots, y_{ni}$ può essere vista come una sequenza di realizzazioni di n variabili casuali identicamente distribuite e indipendenti; ciò permette l'utilizzazione delle tecniche di analisi studiate.

5.1 Analisi del transitorio

Un problema di notevole importanza in una simulazione è quello della scelta delle condizioni iniziali. A questo proposito è essenziale distinguere tra si-

stemi con terminazione e sistemi di cui siamo interessati al comportamento a regime.

Ad esempio lo sportello di una banca che apre alle 8,30 e chiude alle 13,30 è un tipico esempio di sistema del primo tipo. Si tratta di un sistema in cui l'andamento della coda dei clienti ha un andamento intrinsecamente variabile: si parte all'inizio con nessun cliente in coda, e alla chiusura non si accettano più clienti e si esaurisce la coda. È proprio a questa variabilità che siamo interessati.

Diverso è il caso di un sistema di produzione continuo, in cui siamo interessati ad analizzare, ad esempio, il tempo richiesto a regime perché un pezzo venga prodotto, oppure l'intervallo di tempo (sempre a regime) tra l'arrivo di un ordine e la spedizione del prodotto ordinato. In questo caso il transitorio può falsare in modo rilevante i risultati del modello.

Ad esempio, supponiamo di essere interessati al tempo medio di attesa, a regime, di fronte ad una data macchina di una linea di produzione, che indicheremo con d. Siano Y_1, Y_2, ..., Y_m, i valori del parametro che si vuole stimare ottenuti tramite una simulazione. Se indichiamo con Y la variabile casuale 'tempo di attesa a regime', abbiamo

$$d = E[Y] = lim_{j\to\infty} E[Y_j].$$

Una stima di d possiamo ottenerla usando la media campionaria

$$\overline{Y}_m = \frac{\sum_{j=1}^{m} Y_j}{m},$$

dove m è il numero di osservazioni di cui disponiamo.

Nell'effettuare la simulazione dobbiamo scegliere le condizioni iniziali del sistema. Ad esempio, se decidiamo di iniziare la simulazione col sistema scarico, sarà $Y_1 = 0$. Ciò ovviamente si riflette sulla stima ottenuta falsandola. Si potrebbe pensare di partire da una situazione il più possibile simile a quella che si ha a regime, ma questo sposta solo il problema essendo proprio questa situazione quella che noi vogliamo stimare.

Una possibile soluzione è allora quella di non considerare nella stima le prime osservazioni, quelle che sono più influenzate dalle condizioni iniziali. La media viene allora stimata dalla

$$\overline{Y}_{ml} = \frac{\sum_{j=l+1}^{m} Y_j}{m-l},$$

dove l è il numero di osservazioni che vengono scartate, quelle cioè che corrispondono alla fase del transitorio. Il problema è quello di una corretta scelta di l; infatti, un valore troppo basso rischia di portare ad una stima in cui si risente delle condizioni iniziali, mentre un valore troppo alto porta a simulazioni eccessivamente costose.

Non esistono regole sicure per l'individuazione del transitorio. Una ragionevole procedura è riportata nel seguito.

1. Si effettuano n repliche della simulazione, ciascuna di lunghezza m; sia y_{ij} l'osservazione *jesima* della *iesima* replica.

2. Costruiamo la sequenza $\overline{Y}_1$, $\overline{Y}_2$, ..., $\overline{Y}_m$, con $\overline{Y}_j = \frac{\sum_{i=1}^{n} y_{ij}}{n}$. Risulta $E[\overline{Y}_j]$=$E[Y_j]$, e $Var[\overline{Y}_j]$=$Var[Y_j]/n$.

3. Sostituiamo ora alla sequenza $\overline{Y}_1$, $\overline{Y}_2$, ..., $\overline{Y}_m$, la nuova sequenza $\overline{Y}_1(k)$, $\overline{Y}_2(k)$, ..., con $k \leq \lfloor m/2 \rfloor$, e

$$\overline{Y}_j(k) = \begin{cases} \frac{\sum_{h=-(j-1)}^{j-1} \overline{Y}_{j+h}}{2j-1} & ,j = 1, \ldots, k \\ \frac{\sum_{h=-k}^{k} \overline{Y}_{j+h}}{2k+1} & j = k+1, \ldots, m-k. \end{cases}$$

4. Scegliere infine quel valore di l oltre il quale la sequenza $\{\overline{Y}_j(k)\}$ appare giunta a convergenza.

Si tratta, come si vede facilmente, di un approccio basato sulla ispezione da parte dell'esperto, ma dopo avere sottoposto i dati ad un trattamento che ha lo scopo fondamentale di ridurne la varianza. Tale trattamento consiste prima in una media fra dati corrispondenti nelle diverse repliche, poi nella sostituzione di ognuno dei dati così ottenuti con la media fra esso e i dati immediatamente precedenti e quelli immediatamente seguenti.

Grande cura bisognerà avere nella scelta dei parametri m, n e k. Il primo dovrà essere sufficientemente grande da risultare molto maggiore del valore aspettato per l e tale da permettere nella simulazione un numero elevato di occorrenze di tutti gli eventi, anche di quelli poco probabili. Per il secondo può essere opportuno di partire inizialmente con 5 o 10 repliche per poi aumentare tale valore se necessario. Infine k va scelto sufficientemente grande da rendere regolare il grafico della $\overline{Y}_j(k)$, ma non così grande da non permettere di distinguere bene il transitorio. Ovviamente nella scelta è fondamentale il ruolo dell'esperienza di chi effettua la simulazione.

Esercizio Una linea di produzione consiste di una cella di lavorazione ed una stazione di ispezione, in serie. I pezzi semilavorati arrivano alla cella con tempi di interarrivo esponenziali con media 1 minuto. I tempi di lavorazione nella cella sono uniformi nell'intervallo [0.65, 0.70] (in minuti). I tempi di ispezione sono uniformi in [0.75, 0.80]. Il 10% dei pezzi risultano difettosi e sono rimandati alla cella per essere lavorati di nuovo. La cella è soggetta ad interruzioni nella lavorazione a causa di guasti; l'intervallo di tempo fra un guasto ed il successivo ha legge esponenziale con media 6 ore, ed il tempo di riparazione è uniforme nell'intervallo [8, 12] (in minuti). Si voglia determinare la produzione oraria (numero di pezzi) a regime. Si usi l'approccio precedentemente descritto per l'individuazione del transitorio.

5.2 Tecniche per la riduzione della varianza

Tipico obiettivo di uno studio di simulazione è la stima di uno o più parametri. La bontà della stima che si ottiene sarà tanto migliore quanto minore sarà la varianza dello stimatore usato. Nel seguito presenteremo alcune tecniche per la riduzione della varianza.

5.2.1 Variabili antitetiche

Supponiamo di volere stimare $\theta = E[X]$, e supponiamo di avere generato due variabili casuali, X_1 e X_2, identicamente distribuite con media θ. È allora

$$Var\left[\frac{X_1 + X_2}{2}\right] = \frac{1}{4}\left(Var[X_1] + Var[X_2] + Cov[X_1, X_2]\right) \tag{5.2}$$

Se le due variabili casuali X_1 e X_2 fossero correlate negativamente, attraverso il loro uso potremmo ottenere una sostanziale riduzione della varianza.

Supponiamo che la variabile casuale X di cui vogliamo stimare la media sia una funzione di m numeri casuali, uniformi in [0,1):

$$X = h(U_1, U_2, \ldots, U_m). \tag{5.3}$$

Si può allora usare X come X_1 e porre

$$X_2 = h(1 - U_1, 1 - U_2, \ldots, 1 - U_m). \tag{5.4}$$

Essendo $1-U$ anch'essa una variabile casuale uniforme in [0,1), X_2 ha la stessa distribuzione di X, ed essendo $1-U$ negativamente correlata con U, si può provare che, se h è una funzione monotona, allora anche X_1 ed X_2 sono correlate negativamente.

Esempio Consideriamo una coda, e sia D_i il tempo di attesa in coda dell'*iesimo* cliente. Supponiamo di volere stimare $\theta = E[X]$, avendo indicato con X il tempo di attesa totale dei primi n clienti:

$$X = D_1 + \cdots + D_n. \tag{5.5}$$

Indicando con T_i l'*iesimo* tempo di interarrivo e con S_i l'*iesimo* tempo di servizio. Possiamo allora scrivere

$$X = h(T_1, \ldots, T_n, S_1, \ldots, S_n), \tag{5.6}$$

dove h può ragionevolmente essere assunta monotona.

Siano F e G le distribuzioni di T e di S rispettivamente, e supponiamo di usare il metodo dell'inversione per generare tali variabili casuali a partire da $2n$ numeri casuali uniformi:

$$T_i = F^{-1}(U_i),\ S_i = G^{-1}(U_{n+i}),\ i = 1, \ldots, n. \tag{5.7}$$

Una variabile casuale "antitetica" con la stessa distribuzione di X è ottenibile effettuando una seconda simulazione usando i numeri casuali $1-U_i$, $i = 1, \ldots, 2n$.

5.2.2 Condizionamento

Sia X una *v.c.* di cui si voglia stimare la media $\theta = E[X]$, e sia Y un'altra *v.c.*. Assumiano nel seguito che sia X che Y siano *v.c.* discrete; il caso di *v.c.* continue è analogo e viene lasciato per esercizio al lettore.

Definiamo ora la nuova variabile casuale Z funzione di Y:

$$\begin{aligned} Z = E[X|Y=y] &= \sum_x x P[X=x|Y=y] \\ &= \sum_x x \frac{P[X=x, Y=y]}{P[Y=y]}. \end{aligned}$$

Facciamo vedere che la media di Z è proprio il valore θ cercato:

$$E[Z] = \sum_y E[X|Y=y]P[Y=y] = \sum_y \sum_x xP[X=x,Y=y]$$
$$= \sum_x x \sum_y P[X=x,Y=y] = \sum_x xP[X=x] = E[X] = \theta.$$

Analizziamo la varianza di Z. Abbiamo che è

$$\begin{aligned} Var[X|Y=y] &= E[(X - E[X|Y=y])^2|Y=y] \\ &= E[X^2|Y=y] - (E[X|Y=y])^2. \end{aligned}$$

Si ha allora:

$$\begin{aligned} E[Var[X|Y=y]] &= E[E[X^2|Y=y] - (E[X|Y=y])^2] \\ &= E[X^2] - E[(E[X|Y=y])^2], \\ Var[Z] &= E[Z^2] - (E[Z])^2 \\ &= E[(E[X|Y=y])^2] - (E[X])^2, \end{aligned}$$

e sommando membro a membro si ottiene

$$E[Var[X|Y=y] + Var[Z] = E[X^2] - (E[X])^2 = Var[X], \tag{5.8}$$

da cui

$$Var[Z] = Var[X] - E[Var[X|Y=y]] \leq Var[X]. \tag{5.9}$$

Quindi la *v.c.* Z ha la stessa media di X con una varianza minore (o uguale): può allora essere conveniente usare Z per la stima di θ.

Esempio Si voglia stimare, come nell'esempio precedente, la somma dei tempi di attesa in una coda

$$\theta = E[\sum_i W_i], \tag{5.10}$$

dove W_i è il tempo di attesa dell'*iesimo* cliente. Si assume una politica di tipo *FIFO*.

Sia N_i il numero dei clienti presenti nel sistema all'istante di arrivo del cliente *iesimo*, cioè il numero dei clienti in attesa più l'eventuale cliente che viene servito. Supponiamo anche che i tempi di servizio siano esponenziali con media μ.

Introduciamo allora la nuova *v.c.* $Z = \sum_i E[W_i|N_i]$. Essendo

$$E[W_i|N_i] = N_i\mu, \tag{5.11}$$

si ha quindi

$$\theta = E[Z] = E[\sum_i N_i\mu]. \tag{5.12}$$

Pertanto l'uso delle N_i invece delle W_i consente di ottenere uno stimatore con minore varianza e quindi più accurato.

Osserviamo che abbiamo usato la proprietà della distribuzione esponenziale per cui se il tempo di servizio è esponenziale con media μ, anche il tempo necessario per completare, a partire da un tempo fissato (quello dell'arrivo del cliente i^{esimo}), il servizio è esponenziale con la stessa media (*assenza di memoria* della distribuzione esponenziale).

Capitolo 6

Dinamica dei sistemi

6.1 Introduzione

Riprendiamo il modello Preda-Predatore visto nel primo capitolo. Come abbiamo osservato, in questo modello ciò che ci interessa non è l'informazione riguardante le singole prede e i singoli predatori. Siamo piuttosto interessati a seguire l'evoluzione nel tempo delle due popolazioni a livello aggregato; le due variabili di stato principali sono pertanto il numero delle linci ed il numero dei conigli.

Un'altro aspetto interessante di questo modello è la presenza di effetti di retroazione per cui la variazione in un senso di una variabile può portare, attraverso catene di tipo *causa-effetto* ad ulteriori variazioni nella stessa direzione o in direzione opposta. Un tipico esempio è quello della figura 6.1 dove la freccia che va da una variabile ad un'altra indica che una variazione dell'una produce una variazione dell'altra variabile, ed il segno accanto alla freccia indica il tipo di variazione. Ad esempio un aumento del numero dei conigli produce un aumento della disponibilità di cibo per le linci, che a sua volta produce una diminuizione del tasso di mortalità.

Modelli caratterizzati da variabili di tipo aggregato e da cicli di retroazione possono essere trattati per mezzo delle tecniche della *Dinamica dei Sistemi*, oggetto di questo capitolo. Per un approfondimento delle tecniche della Dinamica dei Sistemi rimandiamo al testo di Sterman **?**.

Nel seguito riprenderemo il modello Preda-Predatore, descrivendone una versione più completa di quella semplificata vista nel primo capitolo, ed attraverso esso introdurremo i principali concetti della Dinamica dei Sistemi.

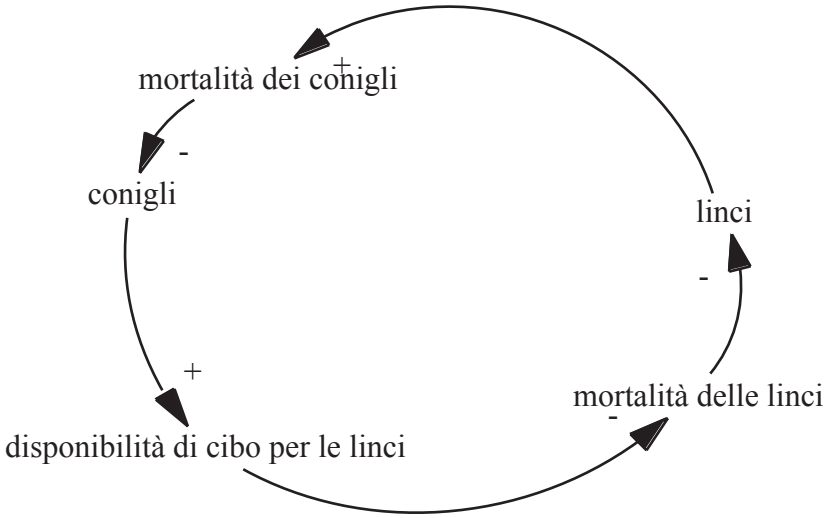

Figura 6.1.

6.2 Modello Preda-Predatore

Il modello Preda-Predatore è stato sviluppato dal matematico italiano Vito Volterra (1860-1940) per studiare un fenomeno che era stato evidenziato dallo zoologo Umberto D'Ancona.[1]

Analizzando le statistiche relative alla pesca nel nord dell'Adriatico, D'Ancona aveva osservato che durante gli ultimi anni della prima guerra mondiale e negli anni immediatamente seguenti si era verificato un sostanziale aumento della percentuale dei predatori (Selaci) pescati. L'unica circostanza che appariva collegabile a questo incremento era la diminuzione dell'attività di pesca causata dalle attività belliche.

Il modello proposto da Volterra aveva proprio lo scopo di studiare questo fenomeno. Qui presenteremo un modello di Dinamica dei Sistemi che è sostanzialmente equivalente a quello di Volterra e può esserne considerato una versione approssimata. Questo modello ci servirà per introdurre i principali concetti della Dinamica dei Sistemi.

[1]Le equazioni che stanno alla base del modello sono oggi note come equazioni di Volterra-Lotka, dallo statistico americano Alfred J. Lotka che le aveva derivate indipendentemente e contemporaneamente, anche se in una forma meno generale.

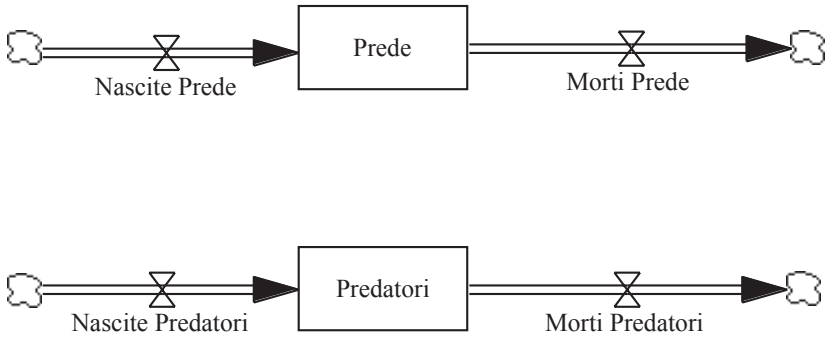

Figura 6.2. *Livelli e Flussi*

6.2.1 Livelli e flussi

Componenti fondamentali di un modello di Dinamica dei Sistemi sono le *variabili di livello* (o semplicemente *livelli*) e le *variabili di flusso* (o semplicemente *flussi*). Le prime rappresentano grandezze che mantengono il loro valore anche in condizione statiche, cioè nell'ipotesi che si blocchi lo scorrere del tempo. Le seconde sono invece variabili che rappresentano delle variazioni nell'unità di tempo, e che quindi assumerebbero valore nullo se bloccassimo lo scorrere del tempo. Nel nostro caso sono variabili di livello le cardinalità delle due popolazioni, *Prede* e *Predatori*. Sono invece variabili di flusso il numero di individui delle due popolazioni che nascono e muoiono nell'unità di tempo, *Nascite_Prede*, *Morti_Prede*, *Nascite_Predatori* e *Morti_Predatori*.

Dal punto di vista grafico i livelli vengono rappresentati per mezzo di rettangoli (che richiamano dei serbatoi), con canali di input e di output, mentre i flussi vengono rappresentati per mezzo di valvole su questi canali. Un esempio di rappresentazione grafica per il nostra problema è fornito in figura 6.2, dove le *nuvole* rappresentano la realtà esterna al sistema sotto considerazione.

Dal punto di vista analitico possiamo rappresentare le relazioni fra livelli e flussi attraverso equazioni del tipo

$$Livello(t + \Delta t) = Livello(t) + (FlussoInput(t) - FlussoOutput(t))\Delta t \tag{6.1}$$

dove con *FlussoInput* e *FlussoOutput* si sono indicate le variazioni per unità di tempo, in aumento ed in diminuzione rispettivamente, della variabile *Livello*, mentre con Δt si è indicato l'intervallo di tempo tra una valutazione delle variabili e la successiva nel processo di simulazione. L'equazione esprime

il fatto che il valore della variabile *Livello* al tempo $t + \Delta t$ è uguale al valore al tempo t più la variazione totale che si è verificata nell'intervallo di tempo $[t, t + \Delta t]$. Osserviamo che le variabili $FlussoInput(t)$ e $FlussoOutput(t)$ rappresentano le variazioni medie unitarie nell'intervallo $[t, t + \Delta t]$, pertanto l'equazione 6.1 descrive tanto più accuratamente l'andamento della variabile *Livello* quanto più è piccolo il valore di Δt. Sul problema della scelta dell'intervallo di tempo nei modelli di Dinamica dei Sistemi torneremo in seguito.

Nel problema in esame le equazioni che descrivono l'andamento nel tempo delle variabili di livello sono:

$$\begin{array}{lcl} Prede(t + \Delta t) & = & Prede(t) + \\ & & (Nascite\_Prede(t) - Morti\_Prede(t))\Delta t, \\ Predatori(t + \Delta t) & = & Predatori(t) + \\ & & (Nascite\_Predatori(t) - Morti\_Predatori(t))\Delta t. \end{array} \tag{6.2}$$

6.2.2 variabili ausiliarie e costanti

Le equazioni 6.2 consentono di descrivere l'andamento dei livelli, cioè nel nostro caso delle popolazioni delle prede e dei predatori, una volta note le condizioni iniziali (la dimensione delle popolazioni al tempo $t = 0$) e una volta che siano state esplicitate le natalità e mortalità delle due popolazioni in funzione del tempo. A questo scopo abbiamo bisogno di introdurre delle nuove grandezze: *variabili ausiliarie* e *costanti*.

Consideriamo innanzitutto le prede: assumendo che l'ambiente in cui vivono (il mare nel caso che stiamo esaminando) fornisca loro tutto il nutrimento necessario, possiamo assumere un tasso di natalità costante, e pertanto porre

$$Nascite\_Prede(t) = A \cdot Prede(t)$$

dove A è una costante che rappresenta il tasso di natalità delle prede.

La diminuzione delle prede è dovuta principalmente a due effetti, uno interno all'ecosistema, l'azione dei predatori, e l'altro esterno, la pesca.

Il numero di prede catturate da predatori nell'unità di tempo può essere assunto come proporzionale al numero dei possibili incontri, che è dato dal prodotto del numero delle prede per il numero dei predatori:

$$Prede\_Catturate(t) = B \cdot Incontri\_Prede\_Predatori,$$

dove è

$$Incontri\_Prede\_Predatori = Prede(t) \cdot Predatori(t).$$

Per quel che riguarda la pesca, possiamo assumere che nell'unità di tempo venga pescata una frazione ε della popolazione esistente:

$$Prede\_Pescate(t) = \varepsilon \cdot Prede(t).$$

Si ha allora:

$$Morti\_Prede = Prede\_Catturate(t) + Prede\_Pescate(t).$$

Consideriamo ora i predatori. Possiamo assumere che, in mancanza di prede e di azioni esterne, i predatori si estinguano con tasso di mortalità costante:

$$Mortalità\_Predatori(t) = C \cdot Predatori(t)$$

dove C è una costante che rappresenta il tasso di mortalità dei predatori.

Alle morti dovute a mancanza di cibo si aggiungono quelle dovute alla pesca. Nell'ipotesi che la proporzione di pesci pescati non dipenda dal tipo di pesce si ha:

$$Predatori\_Pescati(t) = \varepsilon \cdot Predatori(t).$$

Possiamo allora scrivere:

$$Morti\_Predatori = Mortalità\_Predatori(t) + Predatori\_Pescati(t).$$

Per quel che riguarda la crescita dei predatori, essa dipende dalla disponibilità di preda e quindi è proporzionale al numero di incontri tra prede e predatori; ciò viene rappresentato per mezzo della seguente equazione che lega le nascite dei predatori al numero di incontri che esse hanno con le prede:

$$Nascite\_Predatori(t) = D \cdot Prede(t) \cdot Predatori(t).$$

In generale le costanti B e D saranno diverse: la sparizione di una preda non comporta immediatamente la nascita di un predatore.

Dal punto di vista grafico il modello è rappresentato in figura 6.3, dove le frecce indicano le dipendenze funzionali tra le diverse variabili.

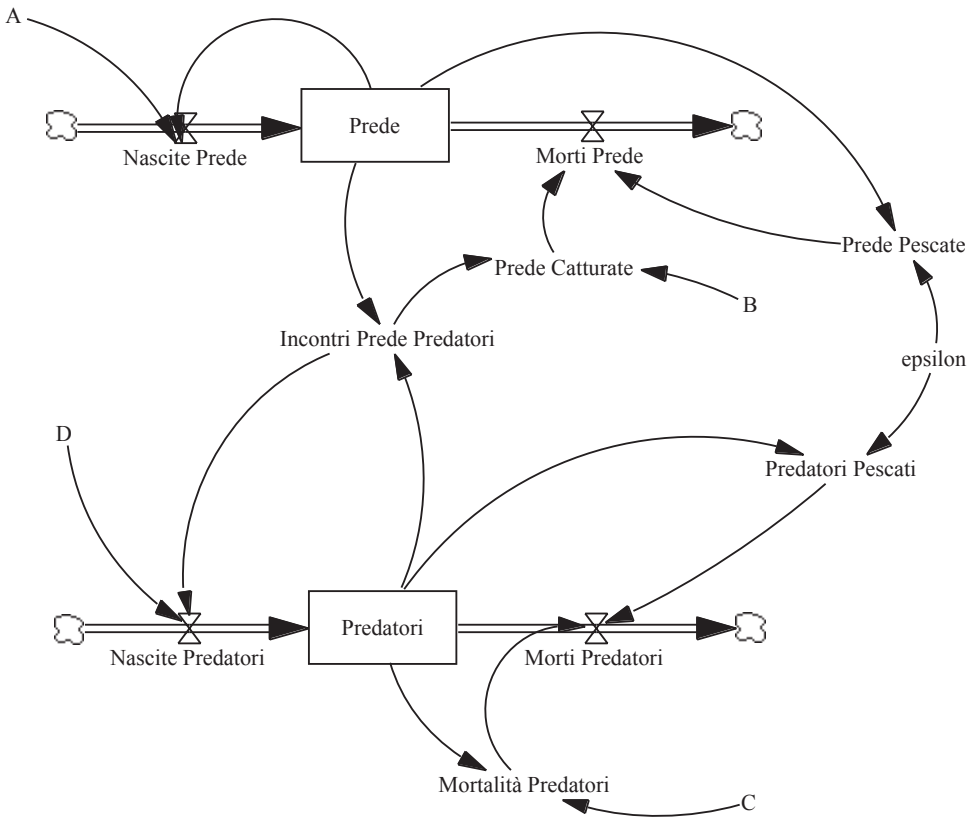

Figura 6.3. *Il modello Predatore-Preda*

6.2.3 Cicli causali

Le frecce nel modello che abbiamo costruito, sia che corrispondano a flussi fisici o a passaggi di informazioni, rappresentano delle relazioni cusa-effetto. Ad esempio la freccia che va dalla variabile di flusso *Nascite_Prede* alla variabile di livello *Prede* indica che il valore della prima ha un effetto sul valore della seconda; questo effetto è di segno positivo, nel senso che un aumento della prima comporta un aumento della seconda, anzi un rafforzamento della sua crescita. Anche la freccia che va dalla seconda alla prima, pur non rappresentando un flusso fisico, descrive una relazione causale di tipo positivo:

all'aumentare della popolazione, in assenza di altri elementi, corrisponde un aumento delle nascite per unità di tempo. Le due frecce insieme individuano un ciclo di *retroazione positiva*: ad un aumento della popolazione corrisponde un aumento delle nascite e quindi un ulteriore aumento della popolazione.

Ci possono essere relazioni con segno negativo, cioè rappresentanti una relazione tra due variabili in cui ad un aumento del valore della prima corrisponde una diminuizione del valore della seconda oppure una riduzione della sua velocità di crescita. Se in un ciclo appare un numero dispari di tali relazioni, allora tutto il ciclo è di segno negativo, e si parla di retroazione negativa. Ad esempio il Ciclo "Prede-Incontri_Prede_Predatori-Prede_Catturate-Morti_Prede-Prede" è un ciclo con segno negativo, infatti la relazione "Morti_Prede-[2] è di tipo negativo perché ad un aumento delle morti di una popolazione, la popolazione diminuisce di numero, oppure si rafforza una sua eventualmente già esistente decrescita, oppure infine si rallenta la sua eventuale crescita. Osserviamo che un ciclo causale, perché possa contribuire all'evoluzione del sistema nel tempo, dovrà contenere almeno una variabile di livello ed una variabile di flusso.

In un modello sono in genere presenti molti cicli causali ed una analisi del loro segno è essenziale per una piena comprensione del comportamento del sistema che si sta studiando.

Nella figura 6.4[3] è riportato l'andamento del numero delle prede e del numero dei predatori in una simulazione in cui si è assunto che l'attività di pesca terminasse alla 250.esima settimana, e che prima della terminazione il numero di pesci pescati nell'unità di tempo fosse pari all'1% della popolazione totale dei pesci ($\epsilon = 0.01$). In linea con quanto osservato si ha che il numero dei predatori aumenta, mentre decresce il numero delle prede. Il risultato è che la proporzione dei predatori cresce in modo significativo.

Questo andamento che non è immediatamente intuitivo si spiega con il fatto che un aumento dell'attività di pesca fa crescere la mortalità dei predatori attraverso due catene causa-effetto concordanti: da un lato c'è l'effetto diretto della pesca, dall'altro c'è quello indiretto dovuto al fatto che diminuiscono le prede. Invece l'effetto dell'aumento dell'attività della pesca sulle prede avviene attraverso due catene causa-effetto discordi: da un lato aumenta la mortalità a causa della pesca stessa, ma dall'altro la mortalità diminuisce

[2]A questa relazione causa-effetto non corrisponde nella figura 6.3 in una freccia che va dalla variabile di flusso "Morti_Prede" al livello "Prede"; questa relazione è invece rappresentata dalla freccia che va nella direzione opposta.

[3]Osservare che le scale sono diverse per le due popolazioni.

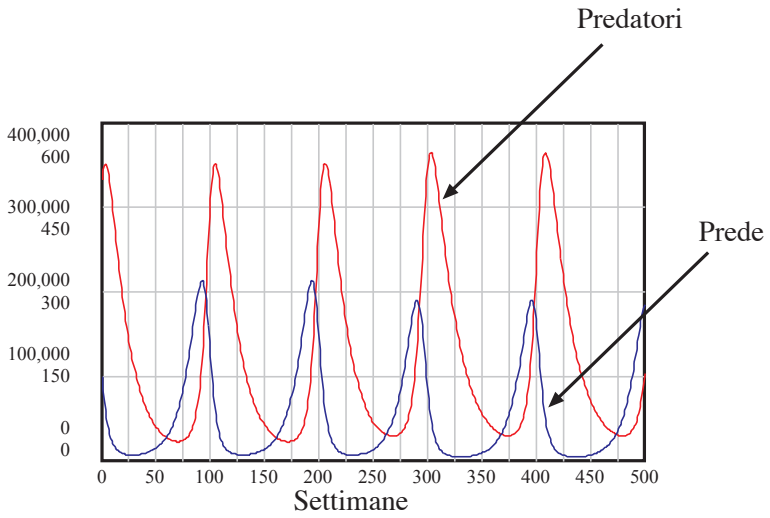

Figura 6.4. *Effetto della cessazione della pesca sull'andamento delle popolazioni delle prede e dei predatori.*

perchè diminuiscono i predatori. Questo spiega il fatto che nella curve della figura 6.4 appare, quando finisce l'attività di pesca, una diminuzione delle prede ed un aumento dei predatori.

6.3 Ritardi

Nei modelli di dinamica dei sistemi le catene ed i cicli causali, come abbiamo visto, giocano un ruolo molto importante. La loro presenza rende spesso difficile da prevedere il comportamento di un sistema. Ciò è tanto più vero in quanto gli effetti non si verificano immediatamente dopo le azioni che li causano; ci sono molto spesso ritardi nel manifestarsi degli effetti, e ciò può rendere particolarmente difficile da analizzare e spesso controintutitivo il comportamento di un sistema.

Ad esempio in una linea di produzione ci troviamo di fronte ad un flusso di pezzi che entrano in celle in cui vengono lavorati e ad un corrispondente flusso di pezzi finiti che escono: fra l'entrata di un pezzo nella cella e la sua uscita passa un tempo dato R. Quando viene lanciata una campagna pubblicitaria per la vendita di un nuovo prodotto passa in genere un certo

tempo prima che cominci ad avere effetto, e questi effetti si mantengono per un certo tempo anche dopo che la campagna è terminata.

In generale un ritardo può essere rappresentato per mezzo di un blocco (*black box*) in cui entra un segnale ed esce una risposta (vedi figura 6.5).

Figura 6.5. *Blocco di ritardo*

Presentiamo ora due tipi fondamentali di blocchi che consentono di rappresentare ed introdurre i ritardi in un modello di dinamica dei sistemi, il ritardo *pipeline*, ed il ritardo *esponenziale*.

Il primo corrisponde ad una situazione in cui ad un segnale di una data forma ed intensità corrisponde una risposta di uguale forma ed intensità, ma differita nel tempo. Ad esempio un ritardo *pipeline* di valore 3 è riportato nella figura 6.6: il segnale in ingresso, un impulso di ampiezza unitaria al tempo 0, produce un uguale impulso in uscita al tempo 3.

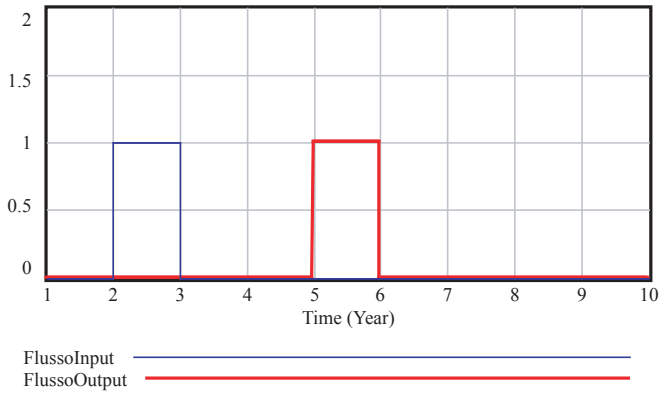

Figura 6.6. *Ritardo pipeline*

Un ritardo di questo tipo può essere realizzato concettualmente per mezzo di un blocco come quello in figura 6.7, in cui si pone:

$$FlussoOutput(t) = FlussoInput(t - R), \tag{6.3}$$

dove R è il valore del ritardo[4].

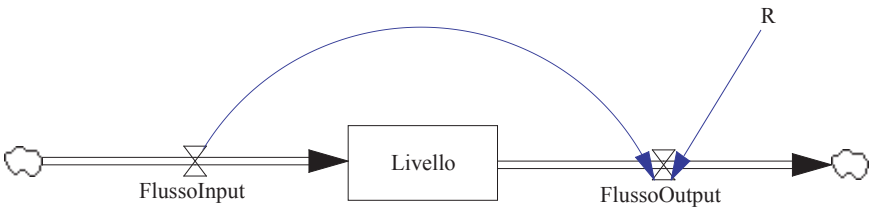

Figura 6.7. *Blocco di ritardo pipeline:* $FlussoOutput(t) = FlussoInput(t - R)$

Nel caso del ritardo esponenziale la situazione è diversa: il blocco funziona da livello, accumulando le quantità in ingresso (segnale) e fornendo in uscita (risposta) un flusso unitario pari al valore del livello diviso per un coefficiente di ritardo. La situazione è illustrata nella figura 6.8, dove un segnale in ingresso che vale 1 al tempo 0, e 0 dal tempo 1 in poi, produce in uscita una risposta smorzata che tende asintoticamente a zero. Nella figura sono presentate le risposte per due valori di ritardo, 3 e 6. Nel primo caso la risposta sale al valore 1/3 per poi decrescere asintoticamente a 0, mentre nel secondo caso sale al valore 1/6 per poi decrescere. Come si vede più grande è il ritardo più smorzata è la risposta e più lentamente il suo effetto si annulla nel tempo. Un ritardo di questo tipo può essere rappresentato per mezzo del blocco di figura 6.9, in cui si pone:

$$FlussoOutput(t) = Livello(t)/R, \tag{6.4}$$

dove R è il valore del ritardo.

Altri tipi di ritardi possono essere ottenuti mettendo in serie più blocchi di ritardo esponenziale. Particolarmente usato è il ritardo del terzo ordine

[4]In *Vensim* questo tipo di ritardo è realizzato per mezzo della funzione DELAY FIXED(Input, R, Output(0)).

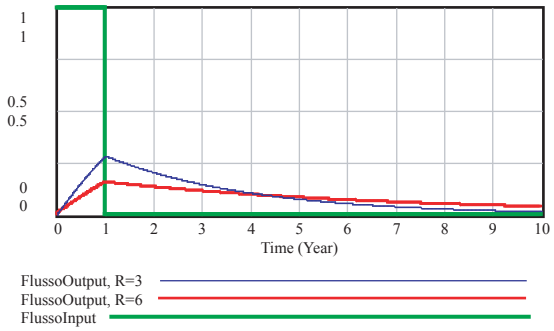

Figura 6.8. *Ritardo esponenziale*

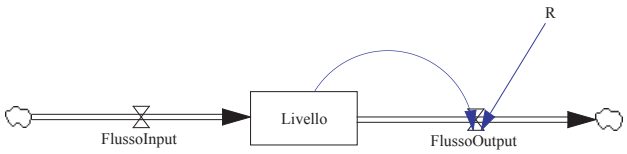

Figura 6.9. *Blocco di ritardo esponenziale del primo ordine*

ottenuto con tre blocchi in serie, ed utilizzando per ciascuno di essi un valore di ritardo pari ad un terzo del valore voluto[5]

Molto spesso il ruolo dei ritardi viene sottovalutato e ciò può portare a prendere decisioni sbagliate con risultati opposti a quelli che si vogliono ottenere. Riportiamo nel seguito alcuni esempi.

6.3.1 Un problema di magazzino

Consideriamo un semplice problema gestionale: la gestione del magazzino di una azienda per quel che riguarda un unico tipo di prodotto. Il livello del

[5]In *Vensim* esistono diverse funzioni che realizzano ritardi esponenziali. Ad esempio DELAY1(Input, R) realizza un ritardo esponenziale del 1^o ordine e DEALY3(Input, R) un ritardo esponenziale del 3^o ordine.

magazzino (numero di pezzi del prodotto in stock) può essere rappresentato da un livello, con le vendite come flusso in uscita, e il numero dei nuovi pezzi prodotti in ingresso. Livello e flussi sono rappresentati nella figura 6.10, che rappresenta il nucleo di partenza del modello che vogliamo costruire. In

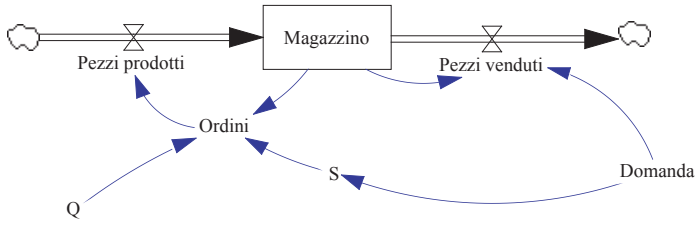

Figura 6.10. *Il magazzino con i flussi in ingresso e uscita*

questo modello abbiamo assunto la *Domanda* come una variabile esogena, ed abbiamo indicato con Q la quantità (il lotto) che viene ordinato ogni volta che si decide di rifornire il magazzino e con S il livello di stock raggiunto il quale si decide di fare un nuovo ordine. Immaginiamo che in questo modello semplificato un ordine venga fatto all'inizio di un periodo e che i nuovi pezzi arrivino all'inizio del successivo. Supponiamo poi di avere una domanda costante di 10 pezzi alla settimana (l'unità di tempo scelta), di decidere di riordinare quando il livello di magazzino è tale da potere soddisfare la domanda solamente per una settimana, che il valore scelto per Q sia 100, e che questo sia anche il valore del livello iniziale del magazzino pari a Q. La relazione fra gli ordini ed il livello del magazzino è data dalla:

$$Ordini = IF\_THEN\_ELSE(Magazzino \leq S, Q, 0).$$

Effettuando allora una simulazione su 50 settimane si ha l'andamento[6] indicato in figura 6.11.

[6]L'andamento riportato nel grafico è un po' falsato dall'effetto dell'interpolazione nel tracciamento del grafico. Ad esempio, alla fine della nona settimana il livello ha raggiunto il valore 10 che fa partire un nuovo ordine e nel corso della decima settimana il magazzino si svuota completamente e l'arrivo del nuovo lotto lo porta all'inizio dell'undicesima settimana al valore 100. A causa dell'interpolazione nel grafico il livello invece di arrivare a 0, cresce dal livello 10 al livello 100 nel corso della decima settimana.

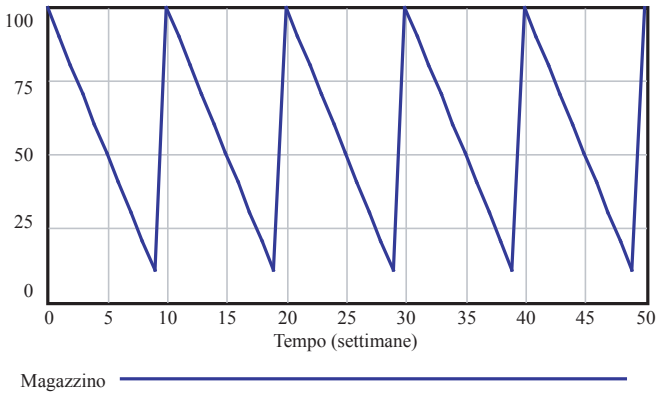

Figura 6.11. *Il livello del magazzino*

Osserviamo che in questo modello abbiamo assunto che la produzione dei nuovi pezzi ordinati avvenga in modo praticamente istantaneo; l'unico ritardo è dovuto all'effetto della discretizzazione. Infatti se avessimo decidere di porre $S = 0$, i nuovi ordini sarebbero partiti all'inizio della undicesima settimana e per tutta quella settimana il magazzino non sarebbe stato in grado di soddisfare la domanda. Il magazzino sarebbe poi ritornato al livello Q all'inizio della dodicesima settimana. Naturalmente avremmo potuto ridurre l'effetto della discretizzazione utilizzando una unità di tempo più piccola (ad esempio il giorno), ma non saremmo comunque riusciti ad eliminarlo del tutto.

In pratica l'ipotesi fatta non è realistica. La produzione richiede un tempo finito. Possiamo allora arricchire il modello inserendo un nuovo livello corrispondente ad i pezzi in fase di produzione. Il modello diventa allora quello della figura 6.12.

Abbiamo usato per la produzione un ritardo di tipo pipeline, ponendo

$$Pezziprodotti(t) = Pezziordinati(t - R),$$

con $R = 3$. Abbiamo poi incrementato S ponendolo al valore $R + 1$. Il risultato della simulazione è quello indicato nella figura 6.13.

Chiaramente il risultato ottenuto non è quello che ci aspettavamo. Il problema sta nel ritardo. Ad ogni settimana, nel nostro modello, viene fatto un controllo sul livello del magazzino e se lo si trova minore o uguale del

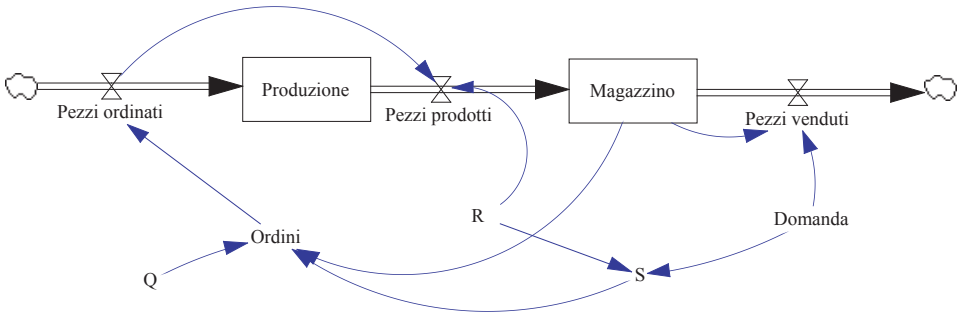

Figura 6.12. *Il modello ampliato*

valore della soglia S, si fa un ordine di Q unità. Quello che accade è che per il ritardo, fatto il primo ordine, la settimana successiva non sono ancora arrivati i pezzi richiesti e quindi si procede ad un altro ordine e così via fino ad accumulare 4 ordini di Q pezzi di seguito. Questo porta il livello del magazzino a crescere come indicato nella figura. Il modello va allora modificato; basta aggiungere la nuova condizione che l'ordine viene fatto solo se non ci sono pezzi correntemente in produzione[7]:

$$Ordini = IF\_THEN\_ELSE(Magazzino \leq S : AND : Produzione = 0, Q, 0).$$

Effettuando di nuovo la simulazione si riottiene, come ci aspettavamo, l'andamento riportato in figura 6.11.

Nelle simulazioni effettuate abbiamo considerato la domanda come costante e da questo valore abbiamo fatto dipendere il valore di soglia, scelto in modo tale da garantire il soddisfacimento della domanda fra il momento in cui viene effettuato l'ordine e quello in cui i pezzi ordinati arrivano in magazzino. Nel caso di domanda variabile può essere utile introdurre il concetto di *Domanda attesa*, che ci servirà per la determinazione di S, e che può essere anche usata per la determinazione del valore di Q che finora abbiamo considerato come variabile esogena. Se C è la copertura massima del magazzino, cioè il numero di settimane di vendite che il lotto da ordinare dovrà garantire, possiamo porre $S = (R+1)\times$Domanda attesa e $Q = C\times$Domanda attesa. Il

[7] Abbiamo qui usato la sintassi propria di *Vensim*.

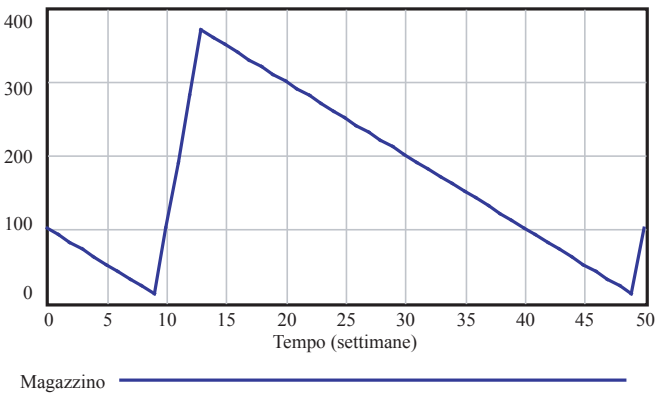

Figura 6.13. *Il livello del magazzino nel modello ampliato*

calcolo della domanda attesa può essere fatto per mezzo di una media delle domande passate attraverso la formula:

$$\bar{Y}_{i+1} = \sum_{k=0}^{n} a_k Y_{i-k}, \tag{6.5}$$

dove con Y_i e con $\bar{Y}_i$ abbiamo indicato rispettivamente la domanda e la domanda attesa al tempo i, mentre gli a_k sono dei coefficienti positivi con $\sum_{k=0}^{n} a_k = 1$. Chiaramente più grande è il valore di n scelto maggiore è il peso che si dà all'andamento passato della domanda. Un caso limite è quello in cui considero solo l'intervallo di tempo precedente ponendo $n = 0$ e $a_0 = 1$; in questo case si considera come domanda attesa al tempo i la domanda al tempo $i - 1$. È ragionevole che gli a_k siano decrescenti al crescere di k, cioè che si dia meno importanza alle osservazioni più lontane nel tempo.

Una tecnica che viene spesso usata per effettuare una stima di una grandezza a partire dai valori che essa ha assunto nel passato è quella del cosiddetto *Exponential Smoothing*. In questa tecnica si utilizza la formula 6.5, ma portando n ad infinito, cioè considerando idealmente un infinito numero di osservazioni, ed utilizzando i coefficienti $a_i = \alpha(1-\alpha)^i$, $i = 0, 1, \ldots,$ con

$0 < \alpha < 1$. Poiché si ha che[8]

$$\sum_{k=0}^{\infty}(1-\alpha)^k = \frac{1}{\alpha},$$

i coefficienti hanno somma unitaria.

Possiamo allora scrivere:

$$\begin{aligned}
\bar{Y}_{i+1} &= \alpha\sum_{k=0}^{\infty}(1-\alpha)^k Y_{i-k} \\
&= \alpha Y_i + \alpha\sum_{k=1}^{\infty}(1-\alpha)^k Y_{i-k} \\
&= \alpha Y_i + (1-\alpha)\alpha\sum_{k=0}^{\infty}(1-\alpha)^k Y_{i-k-1} \\
&= \alpha Y_i + (1-\alpha)\bar{Y}_i.
\end{aligned}$$

Si ha allora:

$$\bar{Y}_{i+1} - \bar{Y}_i - \alpha(\bar{Y}_i - Y_i). \tag{6.6}$$

La 6.6 può essere realizzata per mezzo di un livello, $\bar{Y}$, con $\alpha\bar{Y}$ come Onput e αY come Input. In pratica esistono delle funzioni predefinite che realizzano l'*Exponential Smoothing*.

6.3.2 Diffusioni di inquinanti

In Olanda, fra gli anni 1960 e 1990, fu abbondantemente usato, nelle coltivazioni di patate e di bulbi, un disinfettante del suolo, il DCPe (1-2 Dicloropropene) contenente un inquinante, il 1-2 Dicloropropano (DCPa), che ha una vita molto lunga e filtra nel terreno fino a raggiungere, dopo molto tempo (alcuni decenni) le falde acquifere, inquinandole. Pertanto anche se l'uso del DCPe è stato bandito nel 1990, ci si aspetta nei prossimi anni un inquinamento molto consistente (superiore ai livelli accettabili) delle falde[9].

Una situazione come questa può essere rappresentata per mezzo del modello di figura 6.14. In questo modello gli elementi principali sono due livelli

[8] Si tratta di una serie geometrica.

[9] Questo esempio è stato ripreso da Meadows et al. Meadows et al. (1992).

ed un blocco di ritardo *pipeline* fra di essi; i due livelli realizzano essi stessi dei ritardi di tipo esponenziale. L'inquinante entra nel sistema (flusso di ingresso a sinistra) e viene immagazzinato nel terreno, dove in parte viene assimilato o assorbito attraverso processi naturali ed in parte percola verso gli strati intermedi del terreno. Nel terreno una percentuale dell'inquinante si decompone e scompare (la decomposizione è un processo che avviene con un ritardo esponenziale), mentre un'altra parte scende nelle falde sottostanti attraverso un processo di percolazione (anche qui un ritardo esponenziale). Il processo di percolazione dura un certo tempo (ritardo pipeline), dopo di che l'inquinante raggiunge la falda acquifera, dove si accumula e man mano viene liberato attraverso processi di decadimento naturale: anche qui il flusso in uscita è proporzionale alla quantità immagazzinata in falda, e quindi si ha di nuovo un ritardo esponenziale.

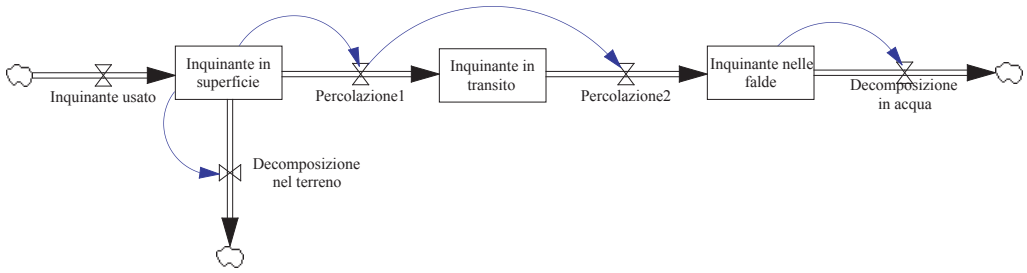

Figura 6.14. *Un modello di diffusione di inquinanti*

Nella figura 6.15 è stato riportato un tipico andamento nel tempo della concentrazione dell'inquinante nelle falde acquifere confrontato con le curve dell'inquinante usato e di quello rilasciato nel terreno. Qui abbiamo supposto l'uso di una quantità costante di inquinante dall'istante 0 per 10 anni, dopo di che l'uso dell'inquinante è stato interrotto. È interessante osservare come la presenza di inquinante in falda continui ad aumentare anche dopo l'interruzione del suo uso, e come solo dopo diversi anni il suo livello cominci a decrescere. Al di là del fatto che i numeri usati possano essere più o meno realistici e del fatto che il modello sia troppo generico per rappresentare bene un qualche particolare tipo di inquinante, esso ci fa comprendere

bene l'effetto dei ritardi, che in maggiore o minore quantità sono comunque sempre presenti in situazioni del tipo che abbiamo illustrato. Tra l'altro la figura presenta un andamento molto simile a quello riportato con riferimento al caso del DCPa dagli autori precedentemente citati.

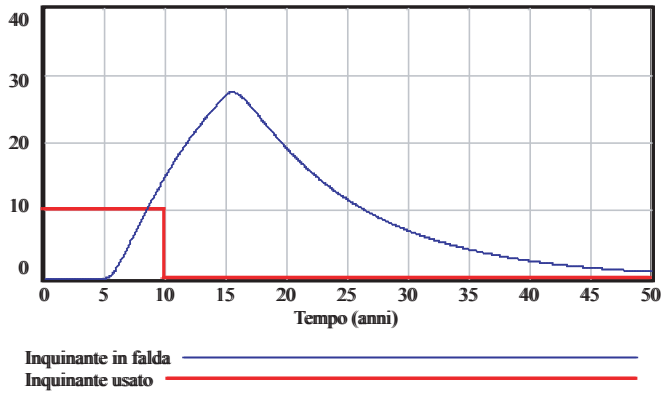

Figura 6.15. *Diffusione dell'inquinante*

Un caso simile, sempre riportato dagli stessi autori, è quello che riguarda i PCB (PolyChlorinated Biphenyls). Si tratta di materiali chimici, stabili, oleosi, non infiammabili, usati principalmente per il raffreddamento di componenti elettriche, capacità e trasformatori. Dal 1929 sono state prodotte circa 2 milioni di tonnellate di PCB, il quale è stato in genere eliminato dopo l'uso in discariche nel terreno, nelle fogne o anche in acqua. Nel 1966 uno studio sulla diffusione del DDT nell'ambiente portò a scoprire anche la presenza dei PCB in praticamente ogni componente dell'ecosistema, dall'atmosfera alla catena del cibo. La maggior parte dei PCB sono poco solubili in acqua, ma solubili nei grassi ed hanno una vita molto lunga. Si muovono lentamente nel terreno e in acqua, fino a che non si inseriscono in qualche forma di vita, dove si accumulano nei tessuti grassi ed aumentano in concentrazione man mano che si muovono in alto nella catena alimentare. Si trovano in pesci carnivori, uccelli e mammiferi marini, nei grassi umani ed anche nel latte umano. Interferiscono col sistema immunitario ed endocrino, in particolare con la riproduzione e lo sviluppo del feto. Negli anni '70 la sua produzione ed il suo uso sono stati proibiti in molti paesi. Tuttavia nel

1992 ancora circa il 70% del PCB prodotto era o in uso oppure contenuto in apparecchiature elettriche abbandonate. Del rimanente 30% solamente l'1% era già apparso nella catena alimentare. Anche qui i danni provocati dall'uso di un materiale inquinante saranno ancora evidenti molto tempo dopo la decisione di abbandonare l'uso di quel prodotto.

6.3.3 Inquinamento atmosferico ed effetto serra

Un caso che evidenzia bene come l'interagire di flussi e di livelli porti a ritardi è quello dell'effetto serra. Riportiamo nella figura 6.16 un modello semplificato dei rapporti fra ciclo del carbonio e riscaldamento globale.

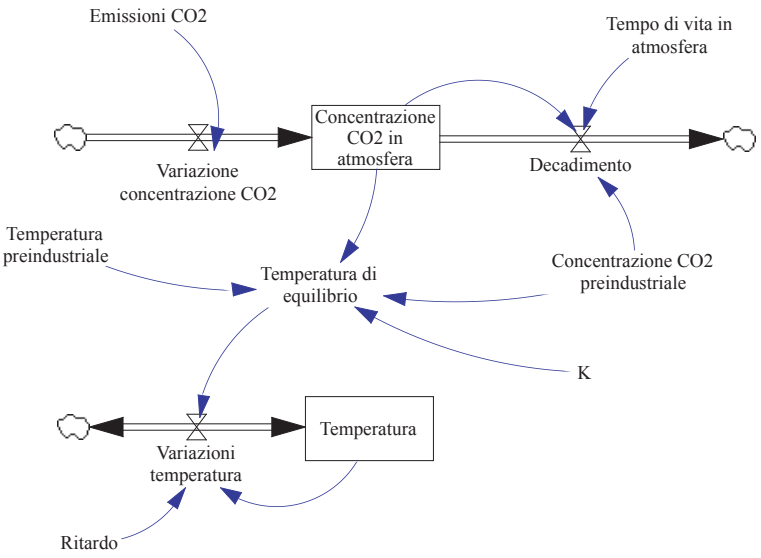

Figura 6.16. *Un modello semplificato del ciclo del carbonio e della temperatura globale*

I fenomeni connessi sono abbastanza complessi, ma possiamo cercare di delinearne gli elementi fondamentali in modo semplice. La temperatura sulla superficie terrestre - terra, strati inferiori dell'atmosfera e fascia superiore dei mari (la zona dei 50-100 metri in cui si concentra la vita marina) - è determinata pricipalmente dal bilancio fra le radiazioni solari (energia che

entra per radiazione) e l'energia che viene radiata indietro nello spazio. La terra è una massa calda circondata da uno spazio freddo ed emette radiazioni la cui distribuzione di frequenza ed intensità dipende dalla sua temperatura superficiale. Più calda è la terra maggiore è il flusso di energia che viene emessa per radiazione verso lo spazio: si crea un effetto di *feedback* negativo, per cui le radiazioni solari in arrivo scaldano la terra, facendone aumentare la temperatura superficiale, fino al punto in cui l'energia emessa per radiazione bilancia quella ricevuta; a questo punto la temperatura non cresce più.

La quantità di energia emessa verso lo spazio dipende pure dalla composizione dell'atmosfera. I cosiddetti *gas serra*, tra cui principalmente l'anidride carbonica (biossido di carbonio), assorbono una parte di questa energia. Quindi un aumento nella concentrazione dei gas serra fa aumentare la temperatura della terra fino a che essa non raggiunga un valore che permette di nuovo il bilanciamento tra energia in arrivo ed energia in uscita. Va osservato che i gas serra svolgono un ruolo fondamentale: sono essi che riducono le radiazioni di quel che serve per mantenere una temperatura media di circa 15 ^{0}C, necessaria per la vita sulla terra. In assenza di gas serra la temperatura media superficiale sarebbe di circa -17 ^{0}C.

Diversi processi naturali di natura biochimica e geotermica hanno causato nel tempo fluttuazioni di concentrazione di anidride carbonica nell'atmosfera. Oggi le attività umane hanno raggiunto una scala tale da avere effetti particolarmente rilevanti: le emissioni di gas serra sono andate crescendo in modo esponenziale dall'inizio della rivoluzione industriale. Conseguentemente la concentrazione di questi gas nell'atmosfera è anch'essa cresciuta esponenzialmente. La concentrazione nell'atmosfera di CO_2, che era prima dell'era industriale di circa 280 ppm (parti per milione), è ora di 370 ppm e tende a crescere[10].

Attualmente c'è uno sbilancio nelle radiazioni di circa 2.4 w/m^2, cioè la radiazione solare in arrivo supera di questa quantità la radiazione emessa dalla terra. Da qui il continuo aumento della temperatura: secondo l'IPCC la temperatura media è cresciuta nel ventesimo secolo fra 0.4 ed 0.8 ^{0}C. Il riscaldamento è stato accompagnato, fra gli altri fenomeni, dal ritirarsi dei ghiacciai, dalla dimuzione dello spessore dei ghiacci artici, e da un aumento dell'ordine di 10-20 cm del livello dei mari.

[10]L'IPCC (Intergovernmental Panel on Climate Change), un'agenzia promossa dall'Onu, afferma: *La presente concentrazione di CO_2 non è mai stata superata negli ultimi 420.000 anni e molto probabilmente neppure negli ultimi 20 milioni di anni. L'attuale tasso di crescita non ha pari almeno negli ultimi 20.000 anni* **?**

Nel modello della figura 6.16, abbiamo utilizzato due livelli, uno per rappresentare la concentrazione di CO_2 nell'atmosfera e l'altro per rappresentare la temperatura. Abbiamo poi supposto l'esistenza di un effetto di ritardo esponenziale per il decadimento dell'anidride carbonica atmosferica, per cui, da un lato ci sono le emissioni industriali di CO_2 e dall'altra il suo decadimento, smorzato sulla base di una stima della sua vita media nell'atmosfera. Abbiamo assunto come base il livello di concentrazione preindustriale[11] e come vita media 100 anni, per cui il decadimento avviene secondo la legge:

$$Decadimento = \frac{C_{CO_2} - C_{pi}}{100},$$

dove C_{CO_2} è la concentrazione di CO_2 e C_{pi} il suo livello preindustriale. Cioè, essendo di 100 anni il tempo medio di permanenza nell'atmosfera di una molecola di anidride carbonica, il numero medio di molecole che in media scompaiono ogni anno è approssimativamente un centesimo di quelle presenti in atmosfera[12]. Naturalmente qui abbiamo assunto l'anno come unità di tempo.

Abbiamo poi assunto che gli incrementi della temperatura di equilibrio rispetto alla temperatura media preindustriale siano crescenti con il rapporto fra la concentrazione di CO_2 ed il suo valore preindustriale secondo una legge del tipo:

$$T(C_{CO_2}) = T_{pi} + K \ln \frac{C_{CO_2}}{C_{pi}},$$

dove $T(C_{CO_2})$ è la temperatura di equilibrio in funzione della concentrazione di CO_2, T_{pi} è il suo livello preindustriale e K è un coefficiente di proporzionalità. Per temperatura di equilibrio intendiamo quella che, per dati livelli di concentrazione di CO_2, consente di realizzare l'equilibrio fra radiazioni in entrata e radiazioni in uscita.

L'effettiva temperatura, indicata con il secondo livello del modello, cresce quando essa è più bassa di quella di equilibrio e decresce quando è più alta; anche qui però c'è un effetto di smorzamento e quindi un ritardo.

[11] Il livello preindustriale viene qui considerato come un dato, anche se è esso stesso il risultato di un equilibrio fra bioemissioni e decadimento nell'atmosfera.

[12] Il fatto di considerare qui non tutta la quantità di anidride carbonica presente, ma solamente la frazione dovuta alle emissioni industriali, $C_{CO_2} - C_{pi}$, è dovuto al fatto che, per semplicità, abbiamo considerato la quantità di gas presente a regime in atmosfera a causa delle bioemissioni una costante, cioè una variabile esogena.

Si tratta di un modello che, come tutti i modelli, è solamente una approssimazione della realtà, in questo caso particolarmente semplificata. Tuttavia anche un modello così semplificato è in grado di darci delle informazioni, sia pure di tipo essenzialmente qualitativo. Ad esempio si è fatta una simulazione con il modello, mettendo in ingresso delle emissioni con un andamento simile a quello delle emissioni di CO_2 degli ultimi 100 anni, e poi, in 20 anni, si sono portate le emissioni al livello di 100 anni fa. I risultati sono indicati in figura 6.17.

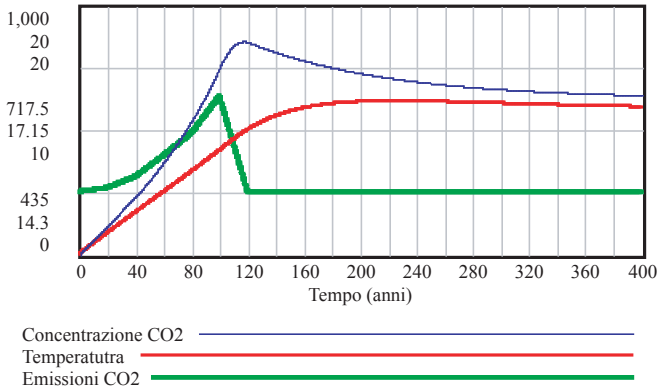

Figura 6.17. *Effetto dei ritardi sulla temperatura globale: risultati della simulazione*

Si vede chiaramente come, malgrado le emissioni si riducano enormemente in un breve lasso di tempo, la diminuzione della quantità totale di CO_2 nell'atmosfera avvenga lentamente, mentre la temperatura continua ad aumentare per oltre un secolo per poi cominciare a declinare molto lentamente, stabilizzandosi ad un livello notevolmente più alto di quello del momento in cui è stata presa la decisione di ridurre le emissioni.

Abbiamo poi provato a vedere cosa accade se, invece di ridurre al livello di 100 anni prima le emissioni, si riducono drasticamente a zero. I risultati sono quelli riportati in figura 6.18.

Si vede chiaramente come, malgrado le emissioni cessino in modo veloce, la diminuzione della quantità totale di CO_2 nell'atmosfera avvenga lentamente, mentre la temperatura continua ad aumentare per diversi decenni e solo

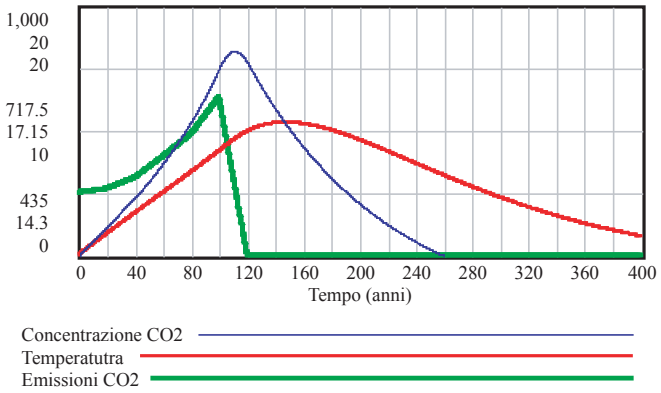

Figura 6.18. *Effetto dei ritardi sulla temperatura globale: risultati della simulazione*

dopo più di un secolo ritorna ai valori che aveva nel momento in cui si era presa la decisione di azzerare le emissioni.

Per quanto il modello usato sia molto semplificato, i risultati ottenuti sono consistenti con quelli forniti dai ben più sofisticati modelli usati dall'IPCC, che riportiamo in figura 6.19. In questo caso le emissioni sono state ridotte ad un livello sostenibile, in modo da garantire una stabilizzazione della temperatura. La stabilizzazione però, proprio per effetto dei ritardi, non avviene subito, ma solo dopo un rilevante lasso di tempo in cui la temperatura continua ad aumentare. Questo comporta una temperatura di equilibrio consistentemente più alta di quella esistente nel momento in cui è stata presa la decisione di ridurre le emissioni.

Il tenere in conto l'effetto dei ritardi nel modello è molto importante nel prendere decisioni: quando si deciderà di intervenire potrà essere troppo tardi! Questo è proprio quello che rischia di accadere se consideriamo che nella Conferenza di Kyoto del 1997, 38 paesi industrializzati si sono accordati a ridurre le emissioni a . . . circa il 95% dei livelli del 1990 entro il 2012. Anche se il trattato di Kyoto fosse realizzato completamente, le emissioni continuerebbero a superare la capacità di assorbimento e la concentrazione nell'atmosfera dei gas serra continuerebbe ad aumentare. E comunque la probabilità che esso venga effettivamente messo in pratica sono limitatissime dato che gli Usa

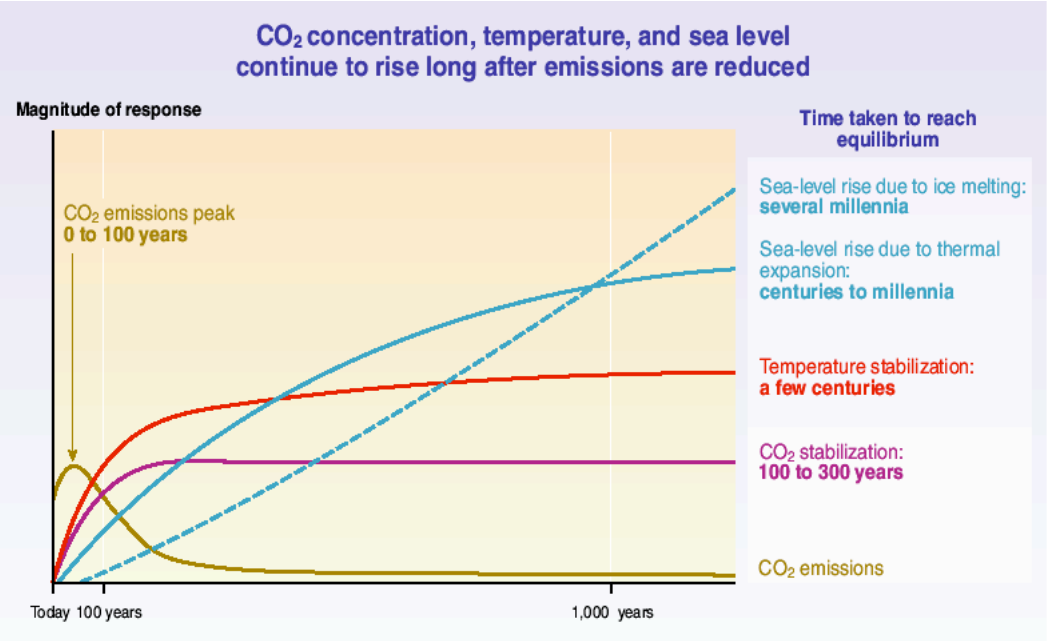

Figura 6.19. *Inerzia della temperatura globale rispetto alle variazioni delle emissioni di anidride carbonica (IPCC 2001)*

non lo hanno ratificato e che l'unica azione che intendono portare avanti sembra sia quella di ridurre le emissioni per unità di prodotto, senza porre però limiti al volume totale delle attività economiche che producono emissioni di gas serra.

6.3.4 La matematica dei ritardi

Riprendiamo il ritardo esponenziale del primo ordine e l'equazione 6.4. Questa equazione è la discretizzazione della seguente equazione differenziale lineare del primo ordine:

$$\frac{dL(t)}{dt} + \frac{L(t)}{R} = I(t). \tag{6.7}$$

Il primo membro della 6.7 ricorda la derivata di un prodotto di cui $L(t)$ sia uno dei fattori. Ed in effetti moltiplicando il primo ed il secondo membro per $e^{\frac{t}{R}}$ si ottiene

$$e^{\frac{t}{R}}\frac{dL(t)}{dt} + e^{\frac{t}{R}}\frac{L(t)}{R} = e^{\frac{t}{R}}I(t), \tag{6.8}$$

che può essere riscritta come segue

$$\frac{d[e^{\frac{t}{R}}L(t)]}{dt} = e^{\frac{t}{R}}I(t). \tag{6.9}$$

Assumiamo che la funzione $L(t)$ sia definita sull'insieme dei reali non negativi ($t \geq 0$). Possiamo allora integrare primo e secondo membro della 6.9 fra 0 e t, ottenendo

$$\int_0^t \frac{d[e^{\frac{s}{R}}L(s)]}{ds}ds = \int_0^t e^{\frac{s}{R}}I(s)ds, \tag{6.10}$$

da cui

$$e^{\frac{t}{R}}L(t) - L(0) = \int_0^t e^{\frac{s}{R}}I(s)ds. \tag{6.11}$$

Possiamo allora ricavare L(t)

$$L(t) = L(0)e^{-\frac{t}{R}} + e^{-\frac{t}{R}}\int_0^t e^{\frac{s}{R}}I(s)ds. \tag{6.12}$$

Consideriamo ora 2 casi, quello in cui si ha un input costante con valore unitario e quello in cui l'input è un impulso unitario di durata 1.

Caso 1 - $I(t) = 1, t \in [0, +\infty]$.

In questo caso si ha

$$\int_0^t e^{\frac{s}{R}}I(s)ds = R(e^{\frac{t}{R}} - 1),$$

e quindi sostituendo nella 6.12 si ottiene

$$L(t) = L(0)e^{-\frac{t}{R}} + R(1 - e^{-\frac{t}{R}}),$$

che nel caso in cui sia $L(0) = 0$, si riduce alla

$$L(t) = R(1 - e^{-\frac{t}{R}}), \tag{6.13}$$

il cui andamento è riportato nella figura 6.20 dove si è posto $R = 3$.

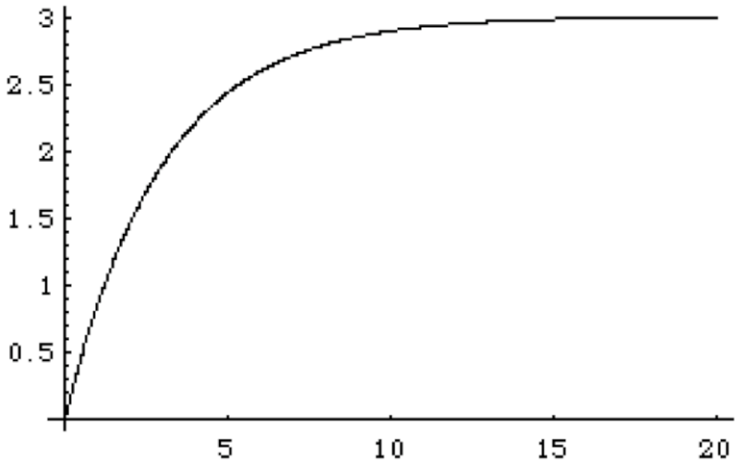

Figura 6.20. *Andamento di $L(t)$ nel caso di $R = 3$ ed $I(t) = 1$.*

Caso 2 - $I(t) = 1$ per $t \in [0, 1]$ e 0 per $t > 1$.
In questo caso si ha

$$\int_0^t e^{\frac{s}{R}} I(s)ds = \begin{cases} R(e^{\frac{t}{R}} - 1) & 0 \leq t \leq 1 \\ R(e^{\frac{1}{R}} - 1) & t > 1 \end{cases},$$

e quindi sostituendo nella 6.12, avendo posto $L(0) = 0$, si ottiene

$$L(t) = \begin{cases} R(1 - e^{-\frac{t}{R}}) & 0 \leq t \leq 1 \\ R(e^{-\frac{t-1}{R}} - e^{-\frac{t}{R}}) & t > 1 \end{cases},$$

il cui andamento, sempre nel caso di $R = 3$, è riportato nella figura 6.21.

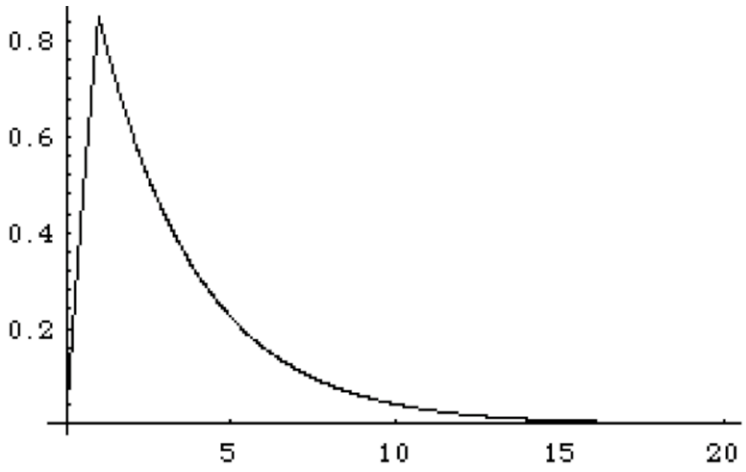

Figura 6.21. *Andamento di* $L(t)$ *nel caso di* $R = 3$ *ed* $I(t) = 1$ *per* $t \in [0, 1]$ *e 0 per* $t > 1$.

Bibliografia

Giancarlo Bigi, Antonio Frangioni, Giorgio Gallo, Stefano Pallottino, and Maria Grazia Scutellà. *Appunti di Ricerca Operativa*. SEU - Servizio Editoriale Universitario Pisano, 2003.

Peter Checkland. Soft systems methodology. In Jonathan Rosenhead, editor, *Rational analysis for a problematic world*. J. Wiley, 1989.

M. Fowler. *UML Distilled*. Addison-Wesley, 2000.

Donella H. Meadows, Dennis L. Meadows, and Jørgen Randers. *Beyond the Limits. Confronting global collapse. Envisioning a sustainable future*. Chelsea Green Publishing Company, 1992.

Michael Pidd. *Computer Simulation in Management Science*. McGraw-Hill, 1998.

Jonathan Rosenhead, editor. *Rational analysis for a problematic world*. J. Wiley, 1989.