5. Analisi dei dati di output

Corso di Simulazione

Anno accademico 2009/10

Analisi dei dati di Output

$Y_1, Y_2, \ldots, Y_m$: output della simulazione.
Le variabili casuali $Y_1, Y_2, \ldots, Y_m$ non sono in generale indipendenti
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 $Y_1, Y_2, \ldots, Y_m$:

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

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

Analisi del transitorio

Un problema di notevole importanza in una simulazione è quello della scelta delle condizioni iniziali.
Ad esempio, consideriamo il tempo medio di attesa, a regime, di fronte ad una data macchina di una linea di produzione, che indicheremo con d. Siano $Y_1, Y_2, \ldots, Y_m$, 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.

Una procedura per l'individuazione del transitorio

1. Si effettuano n repliche della simulazione, ciascuna di lunghezza m; sia y_{ij} l'osservazione *jesima* della *iesima* replica.
2. Costruiamo la sequenza $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_m$, 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 $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_m$, la nuova sequenza $\overline{Y}_1(k), \overline{Y}_2(k), \ldots,$ 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.
Scelta di m, n e k:

- m dovrà essere sufficientemente grande da risultare molto maggiore del valore atteso per l e tale da permettere nella simulazione un numero elevato di occorrenze di tutti gli eventi, anche di quelli poco probabili
- Per n, può essere opportuno di partire inizialmente con 5 o 10 repliche per poi aumentare tale valore se necessario
- 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.

Riduzione della varianza: *Variabili antitetiche*

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. 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

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

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):

$$X = h(U_1, U_2, \ldots, U_m).$$

Si può allora usare X come X_1 e porre

$$X_2 = h(1 - U_1, 1 - U_2, \ldots, 1 - U_m).$$

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: Tempo di attesa

D_i: tempo di attesa dell'*iesimo* cliente in una coda. Si vuole stimare $\theta = E[X]$, con X il tempo di attesa totale dei primi n clienti:

$$X = D_1 + \cdots + D_n.$$

X è una funzione monotona dei tempi di interarrivo e dei tempi di servizio:

$$X = h(T_1, \ldots, T_n, S_1, \ldots, S_n),$$

T e di S siano generati con il metodo dell'inversione a partire da $2n$ numeri casuali uniformi:

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

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

Esempio: Funzione di Affidabilità (a)

Consideriamo un sistema ad n componenti, ciascuna delle quali può essere funzionante o no.
$s = (s_1, s_2, \ldots, s_n)$ definisce lo *stato del sistema* con:

$$s_i = \begin{cases} 1, & \text{se } s_i \text{ funziona,} \\ 0, & \text{altrimenti.} \end{cases}$$

Funzione di struttura:

$$\Phi(s) = \begin{cases} 1, & \text{se il sistema funziona nello stato } s, \\ 0, & \text{altrimenti.} \end{cases}$$

Esempio: Funzione di Affidabilità (b)

Siano conosciute le probabilità di funzionamento delle diverse componenti:

$$P[s_i = 1] = p_i = 1 - P[s_i = 0]$$

Funzione di affidabilità:

$$R(p_1, p_2, \ldots, p_n) = P[\Phi(s) = 1] = E[\Phi(s)]$$

Esempio: Funzione di Affidabilità (c)

Per il calcolo di R si può operare con la simulazione, generando le s_i a partire da numeri casuali fra 0 ed 1, U_i:

$$s_i = \begin{cases} 1, & \text{se } U_i < p_i, \\ 0, & \text{altrimenti.} \end{cases}$$

Quindi è $\Phi(s_1, s_2, \ldots, s_n) = h(U_1, U_2, \ldots, U_n)$, con h funzione monotona. Allora con un solo vettore ad n componenti, U, si possono generare due valori della funzione h, $h(U)$ ed $h(1-U)$, ottenendo una riduzione della varianza della stima della funzione di affidabilità R.

Riduzione della varianza: *Condizionamento*

Sia X una *v.c.* di cui si voglia stimare la media $\theta = E[X]$, e sia Y un'altra *v.c.*. Assumiamo per semplicità che si tratti di variabili discrete.
Definiamo ora la nuova variabile casuale Z funzione di Y:

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

La media di Z è proprio il valore θ cercato.

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

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

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

e sommando membro a membro si ottiene

$$
E[Var[X|Y=y] + Var[Z] = E[X^2] - (E[X])^2 = Var[X]
$$

da cui

$$
Var[Z] = Var[X] - E[Var[X|Y=y]] \leq Var[X]
$$

Esempio: Somma dei tempi di attesa

Si voglia stimare la somma dei tempi di attesa in una coda

$$\theta = E[\sum_i W_i],$$

(W_i è il tempo di attesa dell'*iesimo* cliente)
N_i: numero dei clienti presenti nel sistema all'istante di arrivo del cliente *iesimo*. I tempi di servizio siano esponenziali con media μ.
Introduciamo la nuova *v.c.* $Z = \sum_i E[W_i|N_i]$.
Essendo

$$E[W_i|N_i] = N_i\mu,$$

si ha quindi

$$\theta = E[Z] = E[\sum_i N_i\mu].$$

Esempio: Funzione di affidabilità

Assumiamo che $S_1, S_2, \ldots, S_n$ siano variabili casuali indipendenti con probabilità:

$$P[S_i = 1] = p_i = 1 - P[S_i = 0]$$

e che si voglia stimare la funzione di affidabilità $E[\Phi(S_1, S_2, \ldots, S_n)]$. Si può usare la simulazione generando tutti i valori S_i tranne uno, k, e si stima la

$$E[\Phi(S_1, S_2, \ldots, S_n)|S_1, \ldots, S_{k-1}, S_{k+1}, \ldots, S_n]$$

Per ogni run della simulazione la stima può assumere tre valori, 1, se il sistema funzione indipendentemente dal valore di S_k, 0 se il sistema non funziona indipendentemente dal valore di S_k, e p_k altrimenti.