5. Analisi dei dati di input

Corso di Simulazione

Anno accademico 2009/10

Analisi e scelta dei dati di input

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.
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.

Uso dei dati disponibili

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.

Distribuzione teorica

Se si riesce a stimare una distribuzione teorica che rappresenti bene i dati osservati, allora è ragionevole utilizzarla nella simulazione

- 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.

Distribuzioni empiriche

$X_1, X_2, \ldots, X_n$, osservazioni distinte
$X_{(i)}$ la *iesima* osservazione in ordine crescente di valore,

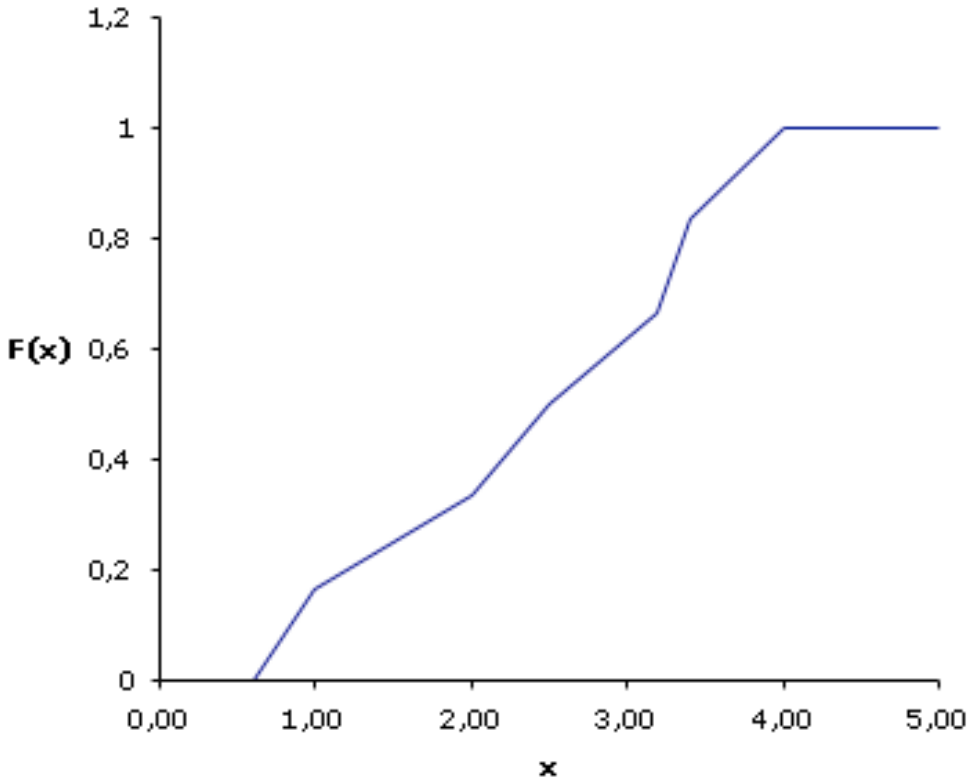
$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}$$

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 $i = 1, 2, \ldots, n-1$.
Distribuzione empirica continua F:

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

Distribuzioni empiriche

Ad esempio le osservazioni 0.6, 1, 3.4, 2, 2.5, 4, 3.2 danno origine alla distribuzione empirica:



Se le osservazioni non sono tutte distinte si possono usare tecniche di perturbazione per renderle distinte

Distribuzioni empiriche

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.

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. Siano $X_1, X_2, \ldots, X_n$ le osservazioni. Un'idea della loro indipendenza la possiamo ottenere analizzando la correlazione fra le diverse osservazioni. Indichiamo con $\overline{\rho}_j$ la stima del coefficiente di correlazione fra osservazioni distanti j posizioni nella sequenza:

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

Il coefficiente di correlazione delle variabili casuali X e Y è

$$\rho_{XY} = \frac{Cov[X, Y]}{\sigma_X \sigma_Y}$$

dove $Cov[X, Y] = E[(X - \mu_X)(Y - \mu_Y)]$ è la covarianza di X e Y

Individuazione della distribuzione

Che tipo di distribuzione da scegliere per la variabile di input sotto esame?
A volte la risposta risulta dalla conoscenza *a priori* del tipo di fenomeno da cui la variabile casuale deriva.
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.

	Esponenziale	Gamma ($n > 1$)	Poisson	Binomiale
$\frac{\sigma}{\mu}$	1	< 1		
$\frac{\sigma^2}{\mu}$			1	< 1

Istogramma

Funzioni continue

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.

Funzioni discrete

Si tracciano 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.

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*.
Una volta stimati i parametri, una verifica di quanto la distribuzione scelta approssima la distribuzione dei dati nel campione può essere effettuata con un test statistico.

Numeri pseudocasuali

Come generare sequenze di numeri casuali?
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.

Numeri pseudocasuali con distribuzione uniforme

I metodi più frequentemente usati sono i *metodi congruenziali*:

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

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 $c = 0$), 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).

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 (generatore moltiplicativo con $a = 3$, $X_0 = 3$ e $m = 7$):

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

Distribuzioni discrete: Inversione

Y *v.c.* discreta, con valori $y_1 < y_2 < y_3 < \ldots$

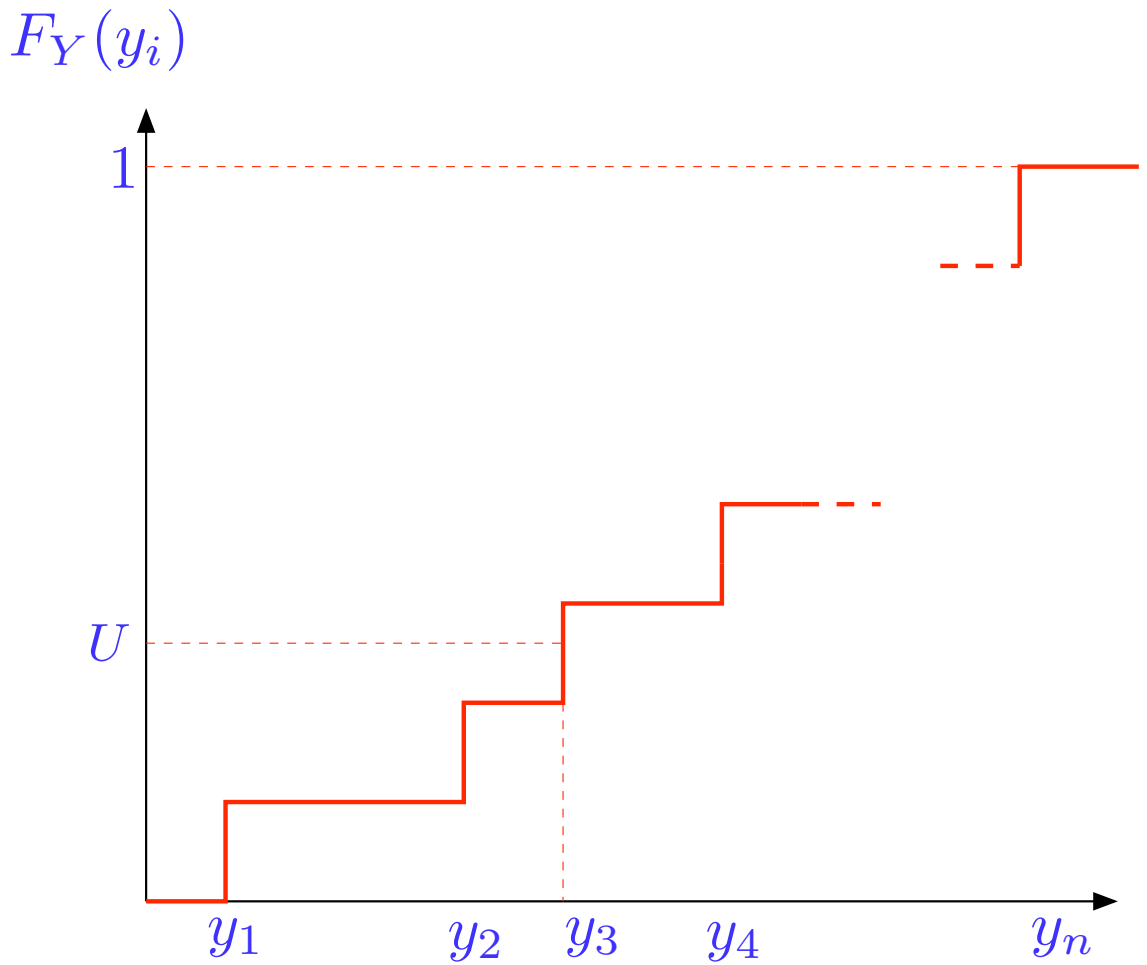
$$\begin{aligned} f_Y(y_i) &= p_i, \\ F_Y(y_i) &= \sum_{j \le i} p_j. \end{aligned}$$

U *v.c.*uniforme in $[0, 1)$.
Definiamo la variabile casuale X:

$$X(U) = Max\{y_i : U \in [F_Y(y_{i-1}), F_Y(y_i)]\}, \qquad (F_Y(y_0) = 0)$$

Inversione: esempio

$$X(U) = Max\{y_i : U \in [F_Y(y_{i-1}), F_Y(y_i)]\}, \qquad (F_Y(y_0) = 0)$$



In pratica dal punto di vista computazionale si sfrutta il fatto che per molte delle distribuzione discrete di uso frequente 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

Distribuzione	$P(i)$	$a(i)$
Binomiale	$\binom{n}{i} p^i (1-p)^{n-i}$	$\frac{(n-i+1)p}{iq}$
Geometrica	$p(1-p)^i$	$1-p$
Poisson	$\frac{\lambda^i e^{-\lambda}}{i!}$	$a(i) = \lambda/i$

Inversione: algoritmo di generazione

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

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 $x_2, x_3, ...$

Distribuzioni continue: inversione

Y *v.c.* continua, con funzioni di densità e distribuzione f_Y e F_Y
U *v.c.* uniforme in $[0, 1)$
Definiamo $X = F_Y^{-1}(U)$. 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.

Esempio: 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 $[0, \infty)$.

Distribuzione normale

$X_1, X_2, ..., X_n$, *v.c.* indipendenti, 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:

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

Pertanto per ottenere una variabile normale standard basterà generare sequenze di n numeri casuali con distribuzione uniforme ed utilizzare poi la v.c. Z_n. 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.

Distribuzione normale: metodo polare

Definiamo le *v.c.*

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

U_1 e U_2 uniformi in $[0, 1) \Rightarrow$ V_1 e V_2 uniformi in (-1, 1)
Generiamo coppie (V_1, V_2), trattenendo 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.
Le variabili casuali:

$$\begin{aligned} X &= V_1 \sqrt{\frac{-2 \lg(V_1^2 + V_2^2)}{V_1^2 + V_2^2}} \\ Y &= V_2 \sqrt{\frac{-2 \lg(V_1^2 + V_2^2)}{V_1^2 + V_2^2}} \end{aligned}$$

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

Metodo della reiezione

Si vuole generare una sequenza di numeri casuali, $x_1, x_2, ..., x_n, ...$, aventi funzione di densità, f_X con:

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

Si procede 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 $x_i = r$.

Metodo della reiezione