GENERAZIONE DI NUMERI PSEUDOCASUALI

Corso di Tecniche di Simulazione, a.a. 2005/2006

Francesca Mazzia

Dipartimento di Matematica
Università di Bari

24 Aprile 2006

Generazione di Numeri Casuali

- Generare NUMERI CASUALI con un computer convenzionale è un esercizio di SIMULAZIONE,
- La SIMULAZIONE DEL CASO ! quindi un Esercizio IMPOSSIBILE !
- Cosa significa infatti SIMULARE? significa, in primo luogo MODELLARE.
- Ma Modellare significa riprodurre la copia di qualcosa che si conosce.
- Abbiamo già serie difficoltà con i sistemi deterministici!
- Come facciamo a modellare il CASO?

Generazione di Numeri Casuali

- La soluzione con un computer non potrà che essere approssimata. Si dovrà infatti utilizzare un qualche algoritmo che, in quanto tale, strumento casuale non può essere.
- Può però far finta bene e quindi generare almeno numeri PSEUDO-CASUALI.
- Ma come facciamo a dire che una sequenza di numeri è abbastanza casuale ?
- E come facciamo a fidarci a priori della sufficiente casualità di un generatore?

Generazione di Numeri Casuali

- un generatore di numeri pseudo-casuali è buono se le sue prestazioni (ripetute tantissime volte) riproducono una Distribuzione Uniforme (vale a dire si collocano con la casualità tipica degli errori di misura nel corso di una normale esperienza).

Generazione di Numeri casuali

- R_1, R_2,.. devono essere uniformi e indipendenti.
- $E(R) = \int_0^1 x dx = 1/2$
- $V(R) = \int_0^1 x^2 dx - E(R)^2 = 1/12$

Distribuzione Uniforme

Nella attività di simulazione, al Generatore di Numeri (Pseudo-)Casuali si richiedono alcune proprietà in parte anche contraddittorie rispetto alla loro natura. In particolare si richiede di

- generare sequenze infinite di numeri statisticamente indipendenti all'interno di un dato intervallo: tipicamente l'intervallo (0,1) o l'intervallo (-1,1);
- essere molto efficiente (i.e. molto veloce); in quanto un programma di simulazione può avere la necessità di attivarlo decine di milioni di volte;
- essere ripetibile (!!!!!????); per ragioni di debug o di semplice ri-esecuzione di un dato run.

Metodi Congruenziali

- Una classe di generatori di numeri casuali a Distribuzione Uniforme è basata sui cosiddetti Metodi Congruenziali.
- Questo significa semplicemente che sono basati sulla operazione (di Congruenza): $x = y \mod m$, che nel concreto consiste in
  calcolare $q = x/m$
  calcolare $p = m * int(q)$
  porre $y = x - p$
- Esempi:
  $74 \mod 7 = 4, 74/7 = 10.571, 7 * 10 = 70, 74 - 70 = 4$
  $57 \mod 57 = 0, 57/57 = 1, 57 * 1 = 57, 57 - 57 = 0$

Metodi Congruenziali

I prototipi di questi generatori sono quello detto

- Additivo

$$X_{i+1} = X_i + X_{i-1} \mod m$$

- Moltiplicativo

$$X_{i+1} = aX_i \mod m$$

- Misto

$$X_{i+1} = (aX_i + c) \mod m$$

Metodi Congruenziali

- Nelle formule precedenti a, c, m e X_i sono da considerare interi non negativi.
- X_i è compreso tra 0 ed m.
- Se si desiderano numeri razionali compresi tra 0 e 1 e sufficiente calcolare il numero

$$R_i = X_i/m$$

- La nomenclatura è:
  a = Moltiplicatore
  c = Incremento
  m = Modulo
  $X_{iniziale}$ = Seme

Metodi congruenziali

- Il generatore di Lehmer è del tipo Moltiplicativo e lo si realizza assumendo
  - m intero, primo e grande
  - a intero nell'intervallo $1 < a < m$
  - y_0 intero nell'intervallo $1 \leq y_0 \leq m - 1$
- Con una scelta opportuna di a e m è possibile ottenere una sequenza che certamente non è casuale ma non è neppure facilmente distinguibile da una sequenza casuale.

Metodi congruenziali

- Fissati a ed m, risulta anche fissata la lunghezza p ($\leq m$) del periodo con il quale la sequenza si ripete.
- Esistono valori di a ed m che generano sequenze di Periodo Completo, vale a dire sequenze che consistono in una permutazione completa dei numeri $1, 2, 3, ..m-1$ e quindi dei numeri normalizzati $1/m, 2/m, .., (m-1)/m$ (a media 0.5 e varianza 1/12).
- Questo significa che per avere una sequenza dall'aspetto verosimilmente casuale occorre trovare una coppia di a ed m che fornisca un periodo sufficientemente lungo.

Esempio

- Siano $X_0 = 27$, $a = 17$,$c = 43$,e $m = 100$.
- Generiamo dei numeri interi fra 0 e 99, quindi $R_i = X_i/99$
- $X_0 = 27$
- $X_1 = (17 \cdot 27 + 43) \mod 100 = 502 \mod 100 = 2$
  $R_1 = 2/100 = 0.02$
- $X_2 = (17 \cdot 2 + 43) \mod 100 = 77 mod 100 = 77$
  $R_2 = 77/100 = 0.77$
- $X_3 = (77 \cdot 17 + 43) \mod 100 = 1352 mod 100 = 52$
  $R_3 = 52/100 = 0.52$

Generazione di numeri casuali

- Ogni R_i è discreto, questa osservazione ha scarse conseguenze se m è molto grande (sono usati valori come $2^{31} - 1$).
- Con densità massima si intende che i valori assunti da R_i non lasciano grandi vuoti in $[0, 1]$.
- Per avere la massima densità ed evitare i cicli in applicazioni pratiche un generatore dovrebbe avere il periodo più grande possibile.
- Il periodo massimo può essere raggiunto con scelte particolari di a, c, m e X_0.

Generazione di numeri casuali

- Se m è una potenza di 2, $m = 2^b$, e $c \neq 0$, il periodo più lungo è $P = m = 2^b$. Tale periodo si ottiene se c è primo rispetto a m e $a = 1 + 4k$, con k intero.
- Se m è una potenza di 2, $m = 2^b$, e $c = 0$, il periodo più lungo è $P = m/4 = 2^{b-2}$. Tale periodo si ottiene se X_0 è dispari, $a = 3 + 8k$ oppure $a = 5 + 8k$, con k intero positivo.
- Se m è un numero primo e $c = 0$, il periodo più lungo è $P = m - 1$. Tale periodo si ottiene se a ha la seguente proprietà: il più piccolo intero tale che $a^k - 1$ è divisibile per m è $k = m - 1$.

Test per i Numeri Casuali

- Per assicurarci che i numeri generati sono difficilmente distinguibili da numeri casuali veri possiamo eseguire alcuni test.
- Test di frequenza: si usa il test di Kolmogorov-Smirnov oppure il test del χ^2 per confrontare la distribuzione dell'insieme dei numeri generati con la distribuzione uniforme.
- Test dei run: vengono studiate le sequenze crescenti o decrescenti, oppure le sequenze dei valori sopra o sotto la media, confrontando i valori ottenuti con quelli che ci aspettiamo. Viene usato il test del χ^2.

Test per i Numeri Casuali

- Test di autocorrelazione: confronta la correlazione dei numeri generati con quella attesa (cioè zero).
- Test dei gap: conta il numero delle cifre che appaiono fra le ripetizioni di una particolare cifra, e utilizza il test di Kolmogorov-Smirnov per confrontare con il valore atteso.
- Poker test: Tratta i gruppi di cifre come una mano di poker. Le mani ottenute sono confrontate con quello che ci aspettiamo usando il test del χ^2.

Test per i Numeri Casuali

- Per testare l'uniformità l'ipotesi nulla è $H_0 : R_i \sim U(0,1)$. Dice che i numeri sono distribuiti uniformemente in $(0,1)$.
- L'ipotesi alternativa è $H_1 : R_i \nsim U(0,1)$.
- Un fallimento nel rigettare l'ipotesi nulla indica che non è stata evidenziata una nonuniformità.
- Questo non implica che ulteriori test non siano necessari.

Test per i Numeri Casuali

- Per testare l'indipendenza l'ipotesi nulla è $H_0 : R_i \sim \textit{indipendenti}$. Dice che i numeri sono distribuiti uniformemente in $(0, 1)$.
- L'ipotesi alternativa è $H_1 : R_i \nsim \textit{indipendente}$.
- Un fallimento nel rigettare l'ipotesi nulla indica che non è stata evidenziata una non indipendenza.
- Questo non implica che ulteriori test non siano necessari.

Generazione di Numeri Casuali

- Per ogni test bisogna indicare un livello di significatività α.
- α è la probabilità di rigettare l'ipotesi nulla quando essa è vera.
- Di solito α è 0.01 o 0.05.

Test di frequenza

- Un test di base che dovrebbe essere eseguito su nuovi generatori è il test dell'uniformità.
- Sono disponibili due test diversi, il test di Kolmogorov-Smirnov e il test del χ^2.

Test di Kolmogorov-Smirnov

- Questo test confronta la cdf continua $F(x)$ della distribuzione uniforme, con la cdf empirica $S_N(x)$ degli N campioni.
- Sappiamo che $F(x) = x, \quad 0 \leq x \leq 1$.
- Se i numeri che abbiamo generato sono $R_1, R_2, ..$ allora la cdf empirica $S_N(x)$ è definita da:

$$S_N(x) = \frac{\text{numero di } R_1, R_2, ... \leq x}{N}$$

- all'aumentare di N, $S_N(x)$ dovrebbe diventare una approssimazione migliore di $F(x)$, se l'ipotesi nulla è vera.

Test di Kolmogorov-Smirnov

- Il test di Kolmogorov-Smirnov si basa su una misura di quanto $S_N(x)$ si discosta da $F(x)$, cioè su:

$$D = max_x |F(x) - S_N(x)|$$

- La distribuzione di D è nota ed è stata tabulata in funzione si N.
- Passo 1: Ordina i dati dal più piccolo al più grande, generando $R_{(i)}$ l'insieme ordinato dei dati.
- Passo 2: calcola $D^+ = max_{1 \leq i \leq n}\{i/N - R_{(i)}\}$ e $D^- = max_{1 \leq i \leq n}\{R_{(i)} - (i-1)/N\}$
- Passo 3: calcola $D = max(D^+, D^-)$

Test di Kolmogorov-Smirnov

- Determina il valore critico D_α dalla tabella.
- Se $D > D_\alpha$, l'ipotesi nulla è rigettata, altrimenti nessuna differenza è stata registrata fra la distribuzione vera dei dati e quella uniforme.

Test di Kolmogorov-Smirnov

- Supponiamo di avere generato 5 numeri $0.44, 0.81, 0.14, 0.05, 0.93$, e sia $\alpha = 0.05$.
- Al primo passo ordiniamo i numeri ottenendo la nuova sequenza $0.05, 0.14, 0.44, 0.81, 0.93$
- Calcoliamo $D^+ = max\{(1/5 - 0.05), (2/5 - 0.14), (3/5 - 0.44), (4/5 - 0.81), (5/5 - 0.93)\} = 0.26$.
- Calcoliamo $D^- = max\{(0.05 - 0/5), (0.14 - 1/5), (0.44 - 2/5), (0.81 - 3/5), (093 - 4/5)\} = 0.21$.
- Calcoliamo D $= 0.26$.
- Determiniamo $D_{0.05} = 0.567$
- Il valore calcolato è inferiore al valore critico, l'ipotesi non è rigettata.

Test chi-quadro

- questo test usa la grandezza

$$\chi_0^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

  con O_i i valori osservati e E_i i valori attesi nella i-sima classe, n è il numero delle classi.
- Per la distribuzione uniforme, se suddividiamo l'intervallo [0,1] in n sottointervalli uguali n corrisponderà al numero di suddivisione. Ci aspettiamo che ogni intervallo contenga N/n valori, dove N è il numero totale delle osservazioni.

Test chi-quadro

- La distribuzione di χ_0^2 è circa la distribuzione chi-quadro con $n-1$ gradi di libertà.
- Il test del chi-quadro è valido solo per campioni grandi.
- il test di Kolmogorov-Smirnov è più potente ed è valido anche per piccoli campioni.

Test chi-quadro

- Supponiamo di avere generato 100 numeri e di volere usare il test del chi-quadro per verificare l'uniformità.
- Nel test usiamo $n = 10$ intervalli di uguale ampiezza $[0, 0.1), [0.1, 0.2), ..., , [0.9, 1.0)$.
- Il valore critico di $\chi^2_{0.05,9} = 16.9$.
- Se il risultato χ^2_0 è minore del valore critico l'ipotesi nulla non è rigettata.

$$\chi^2_0 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

con O_i i valori osservati e E_i i valori attesi nella i-sima classe, n è il numero delle classi.

Test per l'indipendenza

- Supponiamo che un insieme di 100 numeri siano testati per l'uniformità eche i primi 10 valori siano nell'intervallo 0.01,0.10, gli altri 10 in 0.11,020 e così via.
- questo insieme passa il test di frequenza, ma l'ordine in cui vengono prodotti i numeri non è casuale
- Bisogna quindi fare dei test per controllare l'indipendenza dei numeri generati.

Test per l'indipendenza

- Un esempio di sequenza di numeri che non sono indipendenti sono numeri casuali che però presentano particolari caratteristiche.
- Per esempio possono essere sempre crescenti o decrescenti per blocchi di lunghezza prefissata, oppure possono essere sempre sopra la media o sotto la media per blocchi di lunghezza prefissata.
- Un test che permette di controllare questo è il run test.

Test per l'indipendenza

- Esempio: 0.87^{-} 0.15^{+} 0.23^{+} 0.45^{+} 0.69^{-} 0.32^{-} 0.30^{-} 0.19^{+} 0.24^{-} 0.18^{+} 0.65^{+} 0.83^{+} 0.22^{+} 0.81 in questa sequenza tutti i numeri sono segnati con + se sono seguiti da un numero più grande, con - se sono seguiti da un numero più piccolo. L'ultimo numero non viene segnato.
- La sequenza dei più e dei meno nell'esempio è la seguente:

$$- + + + - - - + - + + + - +$$

ogni successione di più o di meno forma un run. Ci sono 8 successioni e quindi 8 run, ognuno con una sua lunghezza.

Test per l'indipendenza

- Se a è il numero totale di run in una sequenza puramente casuale, allora la media e la varianza di a sono dati da:

$$\mu_a = \frac{2N-1}{3}$$

$$\sigma_a^2 = \frac{16N-29}{90}$$

- Per $N > 20$ la distribuzione di a può essere approssimata con una distribuzione normale $N(\mu_a, \sigma_a^2)$.
- Questa distribuzione può essere usata per verificare l'indipendenza.

Test per l'indipendenza

- Il test statistico è
$$Z_0 = \frac{a - \mu_a}{\sigma_a}$$
- l'ipotesi nulla non viene rigettata se $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$

Test per l'indipendenza

- Un test analogo è verificare la sequenza dei run sopra e sotto la media
- Se b è il numero totale di run, n_1 il numero di run sopra la media n_2 il numero di run sotto la media, la media e la varianza di b per un sequenza indipendente sono:

$$\mu_b = \frac{2n_1 n_2}{N} + \frac{1}{2}$$

$$\sigma_b^2 = \frac{2n_1 n_2(2n_1 n_2 - N)}{N^2(N-1)}$$

Test per l'indipendenza

- Se $n_1 > 20$ o $n_2 > 20$ la sequenza è distribuita normalmente.
- Il test statistico è
$$Z_0 = \frac{a - \mu_b}{\sigma_b}$$
- l'ipotesi nulla non viene rigettata se $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$