# Analisi e scelta dei dati di input

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

Francesca Mazzia

Dipartimento di Matematica
Università di Bari

24 Aprile 2006

# 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.
- 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.

# Dati di input

- 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.

# Dati di input

- 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.

# Dati di input

- 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.

# Dati di input

- 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 rappresenta 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.

# Raccolta dei dati di input

- Uno dei problemi più difficili e più importanti è la raccolta dei dati di imput.
- Anche se la struttura del modello è valida, se i dati di input sono stati raccolti in modo inaccurato o non rappresentativo dell'ambiente, il risultato della simulazione sarà sbagliato.

# Identificare la distribuzione

- Una distribuzione delle frequenze o un istogramma è utile nell'identificare la forma della distribuzione
- Un istogramma si costruisce dividendo il range dei dati in intervalli uguali, determinando la frequenza delle occorrenze in ogni sottointervallo, e facendo il grafico delle frequenze
- Scegliere il numero degli intervalli come la radice quadrata del numero delle osservazioni di solito funziona bene.
- L'istogramma per dati continui corrisponde alla funzione densità di probabilità di una distribuzione teorica.
- Se l'istogramma è associato a dati discreti dovrebbe somigliare a una funzione di massa di probabilità.

# Esempio

- Il numero di veicoli che arrivano all'angolo nord-est di un incrocio in un periodo di 5 minuti fra le 7.00 a.m. e le 7.05 a.m. è stato monitorato per 5 giorni lavorativi in un periodo di 20 settimane.

| Arrivi | frequenze | arrivi | frequenze |
|---|---|---|---|
| 0 | 12 | 6 | 7 |
| 1 | 10 | 7 | 5 |
| 2 | 19 | 8 | 5 |
| 3 | 17 | 9 | 3 |
| 4 | 10 | 10 | 3 |
| 5 | 8 | 11 | 1 |

# Grafici Quantili-Quantili

- Riconosciuta la distribuzione che potrebbe rappresentare i dati possiamo utilizzare i grafici q-q come strumento per decidere quanto i dati si avvicinano alla distribuzione data.
- Se $X$ è una variabile aleatoria il $q$ quantile di $X$ è il valore $\gamma$ tale che $F(\gamma) = P(X \leq \gamma) = q$ per $0 < q < 1$.
- Siano $x_i, i = 1, n$ dei dati campione di una distribuzione $X$. Ordina le osservazioni dalla più piccola alla più grande e siano $y_j, j = 1, n$ i valori ordinati.
- I grafici q-q sono basati sul fatto che $y_j$ è una stima del $(j - 1/2)/n$ quantile di $X$, cioè $y_j \approx F^{-1}((j - 1/2)/n)$
- Se i dati appartengono alla distribuzione data il grafico q-q sarà circa una linea retta con pendenza 1.

# Stima dei parametri

- Una volta individuata la distribuzione è necessario determinarne i parametri.
- Ad esempio se la distribuzione è una esponenziale, allora bisogna determinare il valore di $\lambda$.
- 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 il test $\chi^2$ o il test di Kolmogorov-Smirnov.
- Statistiche preliminari sono determinare la media campione e la varianza campione.

# Statistiche preliminari

- Media campione:

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

- Varianza Campione

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}$$

# Stima dei parametri

- Distribuzione di Poisson: Parametro $\alpha$ stimatore $\hat{\alpha} = \overline{X}$
- Distribuzione Esponenziale: Parametro $\lambda$, stimatore $\hat{\lambda} = 1/\overline{X}$
- Distribuzione Normale: parametri $\mu$, $\sigma^2$, stimatori: $\hat{\mu} = \overline{X}$, $\hat{\sigma}^2 = S^2$

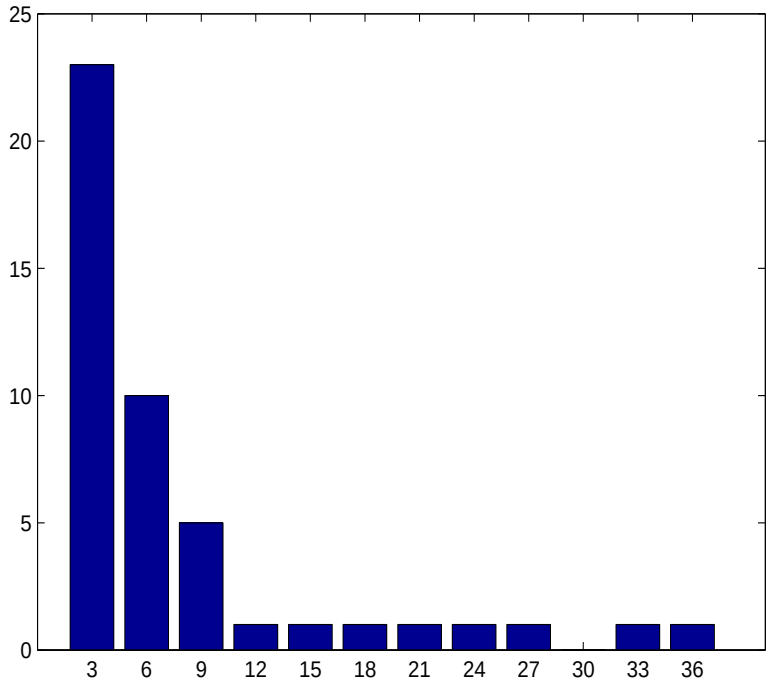
# Test Chi-quadro per la distribuzione esponenziale

- Vengono eseguiti dei test sul tempo di rottura di chip elettronici, i risultati in giorni sono:

| | | | | |
|---|---|---|---|---|
| 79.919 | 3.081 | 0.062 | 1.961 | 5.845 |
| 3.027 | 6.505 | 0.0021 | 0.0013 | 0.123 |
| 6.769 | 59.899 | 1.192 | 34.760 | 5.009 |
| 18.387 | 0.141 | 43.565 | 24.420 | 0.433 |
| 144.695 | 3.663 | 17.967 | 0.091 | 9.003 |
| 0.941 | 0.878 | 3.371 | 2.157 | 7.579 |
| 0.624 | 5.380 | 3.148 | 7.078 | 23.960 |
| 0.590 | 1.928 | 0.3 | 0.002 | 0.543 |
| 7.004 | 31.764 | 1.005 | 1.147 | 0.219 |
| 3.217 | 14.382 | 1.008 | 2.336 | 4.562 |

# Istogramma

Nell'intervallo $[0, 3)$ ci sono 23 dati, nell'intervallo $[3, 6)$ ci sono 10 dati ..., possiamo tracciare l'Istogramma:



# Test Chi-quadro per la distribuzione esponenziale

- Assumiamo che i dati dell'esempio precedente appartengano a una distribuzione esponenziali
- Il parametro $\hat{\lambda} = 1/\hat{X} = 0.084$ al giorno
- L'ipotesi nulla è $H_0$ è distribuita esponenzialmente.
- Eseguiamo il test con intervalli di uguale probabilità.
- Il numero di intervalli si sceglie $\leq n/5 = 10$.
- Poniamo $k = 8$, ogni intervallo ha probabilità p=0.125.
- Gli estremi degli intervalli si calcolano come segue: $F(a_i) = 1 - e^{-\lambda a_i}$, probabilità cumulativa da zero ad $a_i$.
- Poniamo $F(a_i) = ip$ e risolviamo per $a_i = -(1/\lambda)ln(1 - ip)$
- $a_0 = 0$ e $a_k = \infty$

# Test Chi-quadro per la distribuzione esponenziale

| Intervallo | $O_i$ | $E_i$ | $(O_i - E_i)^2/E_i$ |
|---|---|---|---|
| [0,1.590) | 19 | 6.25 | 26.01 |
| [1.590, 3.425) | 10 | 6.25 | 2.25 |
| [3.425, 5.595) | 3 | 6.25 | 0.81 |
| [5.595, 8.252) | 6 | 6.25 | 0.01 |
| [8.252, 11.677) | 1 | 6.25 | 4.41 |
| [11.677,16.503 | 1 | 6.25 | 4.41 |
| [16.503, 24.755) | 4 | 6.25 | 0.81 |
| [24.755, $\infty$) | 6 | 6.25 | 0.01 |
| | — — — | — — — | — — — |
| | 50 | 50 | 39.6 |

Il valore di $\chi_0^2 = 39.6$. I gradi di libertà sono dati da $k - s - 1$ con $s$ il numero di parametri della distribuzione. In questo caso $s = 1$ e con $\alpha = 0.05$ il valore da usare per il confronto è 12.6 e quindi l'ipotesi nulla è rigettata.

# Test di Kolmogorov-Smirnov per la distribuzione esponenziale

Supponiamo di avere calcolato 50 tempi di interarrivo (in minuti) per un intervallo di T=100 minuti:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.44 | 0.53 | 2.04 | 2.74 | 2.00 | 0.30 | 2.54 | 0.52 | 2.02 | 1.89 | 1.53 | 0.21 |
| 2.80 | 0.04 | 1.35 | 8.32 | 2.34 | 1.95 | 0.10 | 1.42 | 0.46 | 0.07 | 1.09 | 0.76 |
| 5.55 | 3.93 | 1.07 | 2.26 | 2.88 | 0.67 | 1.12 | 0.26 | 4.57 | 5.37 | 0.12 | 3.19 |
| 1.63 | 1.46 | 1.08 | 2.06 | 0.85 | 0.83 | 2.44 | 2.11 | 3.15 | 2.90 | 6.58 | 0.64 |

L'ipotesi nulla è $H_0$ : i tempi di interarrivo sono distribuiti esponenzialmente. Si può mostrare che se i tempi di interarrivo sono distribuiti esponenzialmente, allora i tempi di arrivo $T_1, T_1 + T_2, T_1 + T_2 + T_3, \ldots$ sono distribuiti uniformemente in $(0, T)$. I tempi di arrivo normalizzati $T_1/T, (T_1 + T_2)/T, \ldots$ sono distribuiti uniformemente in $(0, 1)$

Possiamo applicare lo stesso test utilizzato per la distribuzione uniforme, calcolando i valori di $D^+ = 0.1054$, $D_- = 0.0080$ e $D = max(D^+, D_-) = 0.1054$. Il valore critico è $D_{0.05} = 1.36/\sqrt{n} = 0.1923$. L'ipotesi nulla non può essere rigettata.

# Modelli di input senza dati

- A volte è necessario sviluppare un modello di simulazione senza avere a disposizione dati.
- In questo caso bisogna affidarsi alle opinioni di esperti oppure bisogna anlizzare limiti fisici o convenzionali, oppure studiare la natura del processo, oppure affidarsi a dati di tipo ingegneristico (tempo medio di rottura di dischi ...)