RICHIAMI SU SIMULAZIONE E METODI MONTE CARLO

AGENDA

- introduzione della simulazione e dei metodi Monte Carlo
- tecniche simulative
  - ▷ generatore congruenziale lineare, variabili uniformi
  - ▷ trasformazione mediante inversa generalizzata
  - ▷ metodo di accettazione-rifiuto (rejection)
  - ▷ metodi ad hoc
  - ▷ simulazione di v.a. multidimensionali
- analisi dei risultati

SIMULAZIONE E MONTE CARLO

- si consideri una variabile aleatoria X, o un evento E, difficili da osservare/non ripetibili
  - ▷ ESEMPIO 1. X rappresenta il rendimento di un investimento che darà i suoi frutti fra 10 anni
    $\rightsquigarrow$ per osservare X bisogna dunque aspettare 10 anni
  - ▷ ESEMPIO 2. X rappresenta il danno provocato da un sinistro
    $\rightsquigarrow$ per osservare X bisogna che il sinistro si verifichi
  - ▷ ESEMPIO 3. E rappresenta l'evento {se cado dalle scale rimango illesa}
    $\rightsquigarrow$ per osservare E dovrei cadere dalle scale
- simulare la v.a. X, o l'evento E, significa considerare un'altra v.a., X', o un altro evento E', facili da osservare, osservarli e far finta (simulare) che le loro osservazioni siano quelle di X (o di E)
- affinché il procedimento sia corretto bisogna che le nuove variabili riproducano il medesimo stato di incertezza originario
  $\rightsquigarrow$ ciò equivale a richiedere che X' abbia la stessa distribuzione di probabilità di X (congiunta, se multidimensionale) o, rispettivamente, gli eventi E ad E' siano equiprobabili

SIMULAZIONE E MONTE CARLO

- la simulazione di variabili aleatorie è estremamente efficace per
  - ▷ generare scenari
  - ▷ calcolare speranze matematiche tramite il metodo Monte Carlo quando non sono disponibili formule analitiche
  - ▷ stimare la funzione di ripartizione di variabili aleatorie complesse (ad es. trasformate di v.a. con cdf nota)
- applicazioni:
  - ▷ generare traiettorie future di variabili rilevanti per un assicuratore (rendimenti degli investimenti, esborsi per sinistri, corresponsione di rendite vitalizie, ...) $\rightsquigarrow$ economic scenario generators
  - ▷ calcolare probabilità di rovina $\rightsquigarrow$ solvency
  - ▷ calcolare il prezzo di strumenti finanziari derivati (finanza) o il valore delle liabilities di un portafoglio assicurativo vita
  - ▷ ...
- simulazione Monte Carlo
  - ▷ funziona bene in ambito multidimensionale (con molte variabili)
  - ▷ generalmente richiede un elevato tempo di calcolo

SIMULAZIONE E MONTE CARLO

- il metodo Monte Carlo è basato sulla legge dei grandi numeri: se X è una variabile aleatoria con media finita, e $X_1, \ldots, X_n, \ldots$ sono variabili iid con la stessa distribuzione di X, allora, con prob. 1,

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = E[X]$$

- il metodo Monte Carlo esegue i seguenti passi:
  - ▷ simula $X_1, \ldots, X_n$ (con n grande!) e siano $x_1, \ldots, x_n$ le corrispondenti determinazioni simulate
  - ▷ approssima $\theta = E[X]$ con

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  - ▷ per stimare la cdf di X si usa la funzione di ripartizione empirica:

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \leq x\}}$$

SIMULAZIONE E MONTE CARLO

- typicamente, X è una variabile aleatoria molto complicata, es.

$$X = g(Z)$$

con Z vettore di variabili aleatorie (di elevata dimensione)
$\rightsquigarrow$ prima si simula Z e poi si calcola il corrispondente valore simulato di X applicando la funzione g al vettore simulato di Z

- ESEMPIO. si consideri un portafoglio/fondo composto da diverse classi di attività
  - ▷ il rendimento del portafoglio, X, dipende dai rendimenti delle singole attività, Z (azioni, proprietà immobiliari, obbligazioni, derivati, merci, ...)
  - ▷ si simula il rendimento di ciascuna attività, poi si combinano i risultati in modo da costruire il rendimento simulato del fondo

SIMULAZIONE E MONTE CARLO

- come illustrato finora, a noi interessano le applicazioni della simulazione in ambito stocastico, anche se in realtà le primissime applicazioni del metodo Monte Carlo si ritrovano in ambito deterministico, per il calcolo di integrali multipli di elevata dimensione:

$$\int_{a_1}^{b_1} \cdots \int_{a_m}^{b_m} f(y_1, \ldots, y_m) \mathrm{d}y_1 \ldots \mathrm{d}y_m = \text{[sostituzione di variabili]}$$

$$= \int_0^1 \cdots \int_0^1 g(u_1, \ldots, u_m) \mathrm{d}u_1 \ldots \mathrm{d}u_m = E[g(U_1, \ldots, U_m)]$$

con $U_1, \ldots, U_m$ v.a. iid con distribuzione uniforme in $[0, 1]$

SIMULAZIONE E MONTE CARLO

- come detto precedentemente, per la correttezza del procedimento è necessario che le nuove variabili, quelle simulanti, abbiano la stessa distribuzione di probabilità delle variabili originarie
- nella pratica, si simulano dapprima v.a. con distribuzione uniforme nell'intervallo $[0,1]$ e poi si trasformano in modo da ottenere una v.a. con distribuzione assegnata
- ci sono due alternative per generare v.a. $U(0,1)$:
  - ▷ si usa un procedimento fisico (ad es. si sorteggiano da un'urna, con reimbussolamento, le cifre decimali, teoricamente infinite, dei numeri da simulare) producendo così veri numeri casuali
  - ▷ si utilizza un calcolatore per generare una sequenza deterministica che mimi la sequenza casuale $\rightsquigarrow$ numeri pseudocasuali
- qualunque sia il metodo prescelto, deve essere statisticamente testato per
  - ▷ l'indipendenza
  - ▷ la distribuzione $U(0,1)$

SIMULAZIONE E MONTE CARLO

- l'utilizzo di un computer e di sequenze di numeri pseudocasuali è ovviamente preferibile (e necessario), per la loro
  - ▷ velocità
  - ▷ riproducibilità
  - ▷ facile portabilità
- riproducibilità - l'utilizzo della stessa sequenza di numeri simulati - è una proprietà molto importante nelle analisi di:
  - ▷ sensitività dei risultati rispetto ai parametri del problema
    $\rightsquigarrow$ la differenza nei risultati è attribuibile a diversi parametri, e non a diverse simulazioni
  - ▷ statica comparata: in particolare nei problemi decisionali bisogna utilizzare le stesse distribuzioni simulate delle variabili d'interesse

GENERATORE CONGRUENZIALE LINEARE

- è utilizzato per generare variabili uniformi, numeri pseudocasuali, attraverso algoritmi deterministici
- la gran parte di questi algoritmi produce dapprima una sequenza ricorsiva $s_1, \ldots, s_n, \ldots$ tramite
  - $s_0 = s \rightsquigarrow$ seme (assegnato)
  - $s_{n+1} = f_1(s_n)$, $n = 0, 1, 2, \ldots$

  dopo di che ricava la sequenza $u_1, \ldots, u_n, \ldots$ di numeri pseudocasuali che mimano v.a. indipendenti $U(0,1)$ tramite

$$u_n = f_2(s_n), \; n \geq 1$$

  dove f_1 e f_2 sono funzioni reali, in particolare f_2 prende valori nell'intervallo $[0,1]$

GENERATORE CONGRUENZIALE LINEARE

- esistono sempre due numeri interi d ed ℓ tali che $s_d = s_{d+\ell}$, cosicché la sequenza si ripeterà dopo ℓ iterazioni
  $\rightsquigarrow$ il minimo ℓ per cui questo si verifica è il periodo della sequenza, che deve essere tanto più grande possibile
- se si conoscono il seme s e le due funzioni f_1, f_2, allora si può riprodurre l'intera sequenza (u_n)
- un esempio tipico è il generatore congruenziale lineare:

$$s_{n+1} = f_1(s_n) = (A\,s_n + C) \bmod M, \quad u_n = f_2(s_n) = \frac{s_n}{M}$$

  dove A, C, M sono interi tali che $A > 0$, $C \geq 0$, $M > 0$ e si parte da un seme intero non negativo
- quindi s_{n+1} è il resto della divisione intera di $A\,s_n + C$ per M, cosicché, per ogni n, s_n è un intero ≥ 0 e $\leq M-1$, e $u_n \in [0,1)$
- più precisamente, i numeri della sequenza appartengono all'insieme $\{0, \frac{1}{M}, \ldots, \frac{M-1}{M}\}$ e il periodo non può superare M

GENERATORE CONGRUENZIALE LINEARE

- ESEMPIO. per $s = 85$, $A = 11$, $C = 37$, $M = 100$, i primi 4 numeri generati dall'algoritmo sono

n	$A\,s_{n-1} + C$	s_n	u_n
0	—	85	—
1	972	72	0.72
2	829	29	0.29
3	356	56	0.56
4	653	53	0.53

- andando avanti con la sequenza si può verificare che $s_{50} = s_0$, $\rightsquigarrow \ell = 50$
- una scelta di parametri molto diffusa nelle prime generazioni di computer era $A = 7^5 = 16807, C = 0, M = 2^{31} - 1 = 2147483647$
- i moderni calcolatori usano algoritmi più complessi ma basati su simili idee (es. Mersenne Twister, con periodo $2^{199937} - 1$)
- vediamo ora come si simulano v.a. unidimensionali di distribuzione assegnata utilizzando numeri pseudocasuli come determinazioni di v.a. $U(0,1)$

TRASFORMAZIONE MEDIANTE INVERSA

- d'ora innanzi assumiamo che U sia una v.a. con distribuzione $U(0,1)$
- si consideri l'inversa generalizzata (o pseudoinversa, o quantile, $\cdots$) della funzione di ripartizione, F_X, di una v.a. X:

$$F_X^{-1}(p) = \inf\{x \in \mathbb{R} : \ F_X(x) \geq p\} \qquad \text{per} \quad 0 < p < 1$$

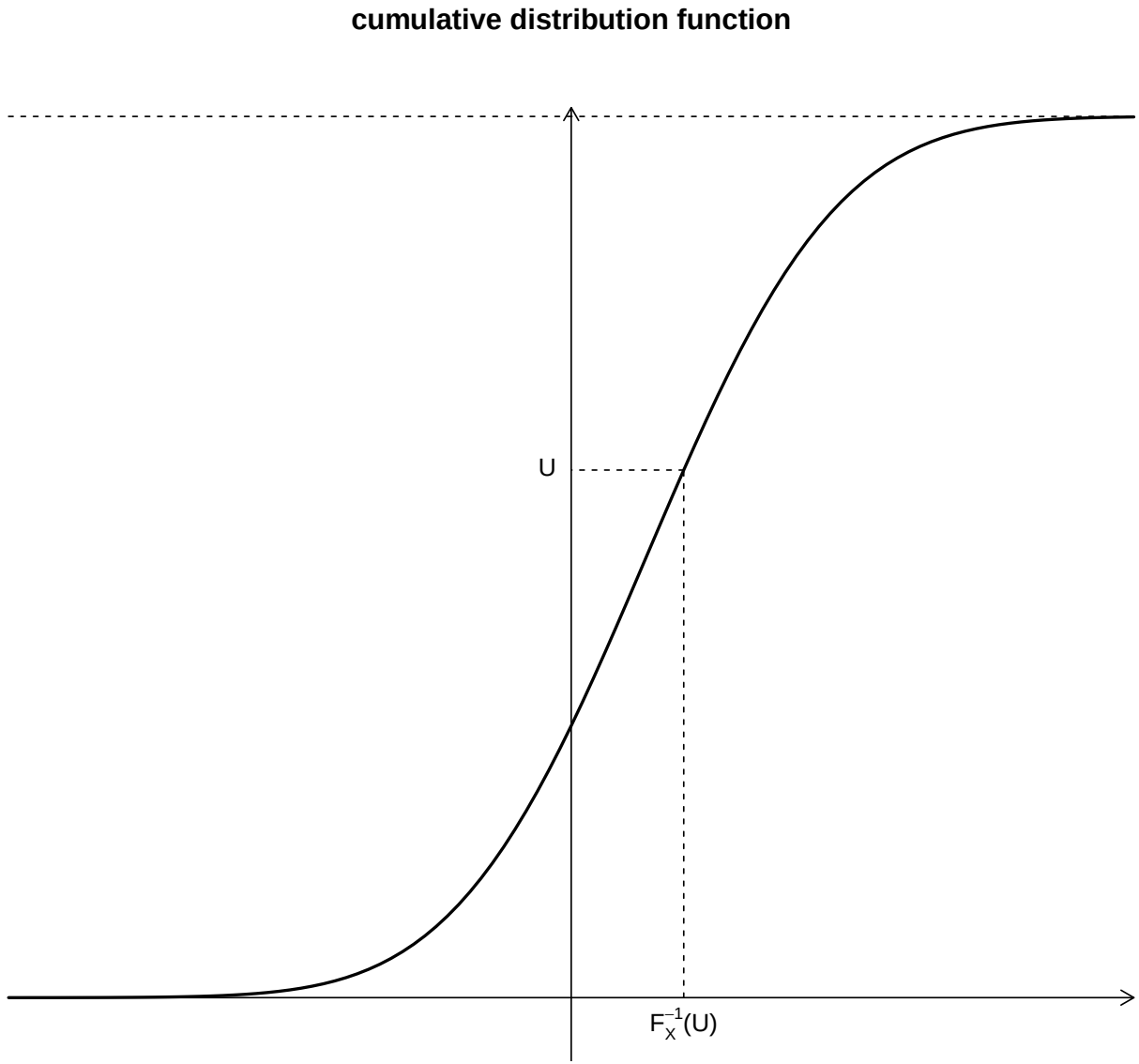
- in particolare, se F_X è invertible (dove non vale 0 o 1), allora F_X^{-1} coincide con l'usuale inversa
- si può usare il seguente risultato per simulare v.a. con funzione di ripartizione F_X:

se U è $U(0,1)$ allora $F_X^{-1}(U)$ ha funzione di ripartizione F_X

- la verifica è facile: se F_X è invertibile poniamo $X' = F_X^{-1}(U)$

$$F_{X'}(x) = \Pr[X' \leq x] = \Pr[F_X^{-1}(U) \leq x] = \Pr[U \leq F_X(x)] = F_X(x)$$

TRASFORMAZIONE MEDIANTE INVERSA



TRASFORMAZIONE MEDIANTE INVERSA

- questo suggerisce il seguente algoritmo, se F_X^{-1} è nota (analiticamente):
  1. genera U da $U(0,1)$
  2. calcola $X' = F_X^{-1}(U)$
- anche se F_X^{-1} non è nota si può risolvere numericamente l'equazione in x

$$F_X(x) = U$$

- molti software (ad es. Excel) hanno già delle routine predefinite per invertire numericamente funzioni di ripartizione di cui non è nota la forma analitica dell'inversa

TRASFORMAZIONE MEDIANTE INVERSA

- ESEMPIO 1. genera una v.a. con distribuzione Weibull(c,γ)
  - la funzione di ripartizione della Weibull(c,γ) è

$$F_X(x) = 1 - \exp(-cx^\gamma), x > 0, \qquad \text{dove } c > 0, \gamma > 0$$

  - per trovare la sua inversa bisogna quindi risolvere l'equazione $F_X(x) = p$ con $0 < p < 1$
  - da cui si ottiene $x = F_X^{-1}(p) = \left(-\frac{1}{c}\log(1-p)\right)^{1/\gamma}$
  - quindi per generare una Weibull(c,γ) si genera prima una v.a. uniforme U e poi si calcola

$$X' = \left(-\frac{1}{c}\log U\right)^{1/\gamma}$$

  - abbiamo preso U, anziché $1-U$, tanto hanno la stessa distribuzione $U(0,1)$
  - se vogliamo un'esponenziale Exp(c) basta prendere $\gamma = 1$

TRASFORMAZIONE MEDIANTE INVERSA

- ESEMPIO 2. sia T_y la durata residua di vita di un individuo di età y
  - ▷ la sua funzione di ripartizione è

$$F_{T_y}(t) = \Pr[T_y \leq t] = {}_tq_y = 1 - {}_tp_y = 1 - \exp\left(-\int_0^t \mu(y+v)\mathrm{d}v\right),$$

    dove μ è l'intensità di mortalità
  - ▷ se μ è una Gompertz, $\mu(z) = \alpha\exp(\beta z)$ con $\alpha, \beta > 0$, si può prendere, come sostituta di T_y

$$T'_y = \frac{1}{\beta}\log\left(1 - \frac{\beta}{\alpha}\mathrm{e}^{-\beta y}\log(U)\right),$$

    dove abbiamo sostituito U a $1-U$. Infatti:

$$F_{T_y}(t) = 1 - \exp\left(-\frac{\alpha}{\beta}\exp(\beta y)(\exp(\beta t) - 1)\right)$$

$$\rightsquigarrow\ F_{T_y}(t) = p \ \Leftrightarrow\ t = F^{-1}_{T_y}(p) = \frac{1}{\beta}\log\left(1 - \frac{\beta}{\alpha}\mathrm{e}^{-\beta y}\log(1-p)\right)$$

TRASFORMAZIONE MEDIANTE INVERSA

- il metodo di trasformazione tramite inversa (generalizzata) si applica anche alle v.a. discrete, per cui la funzione di ripartizione non è invertibile bensì costante a tratti
- se la variabile assume come possibili valori $x_1, x_2, \ldots, x_n, \ldots$ con probabilità $\Pr[X = x_i] = p_i$ (tali che $p_i > 0$ e $\sum_i p_i = 1$),
  - ▷ la funzione di ripartizione di X è quindi

$$F_X(x) = \sum_{i:x_i \leq x} p_i$$

    e p_i è il salto di tale funzione in x_i
  - ▷ l'inversa generalizzata diventa

$$F_X^{-1}(p) = \min_i \left\{ x_i : F_X(x_i) = \sum_{j=1}^{i} p_j \geq p \right\}$$

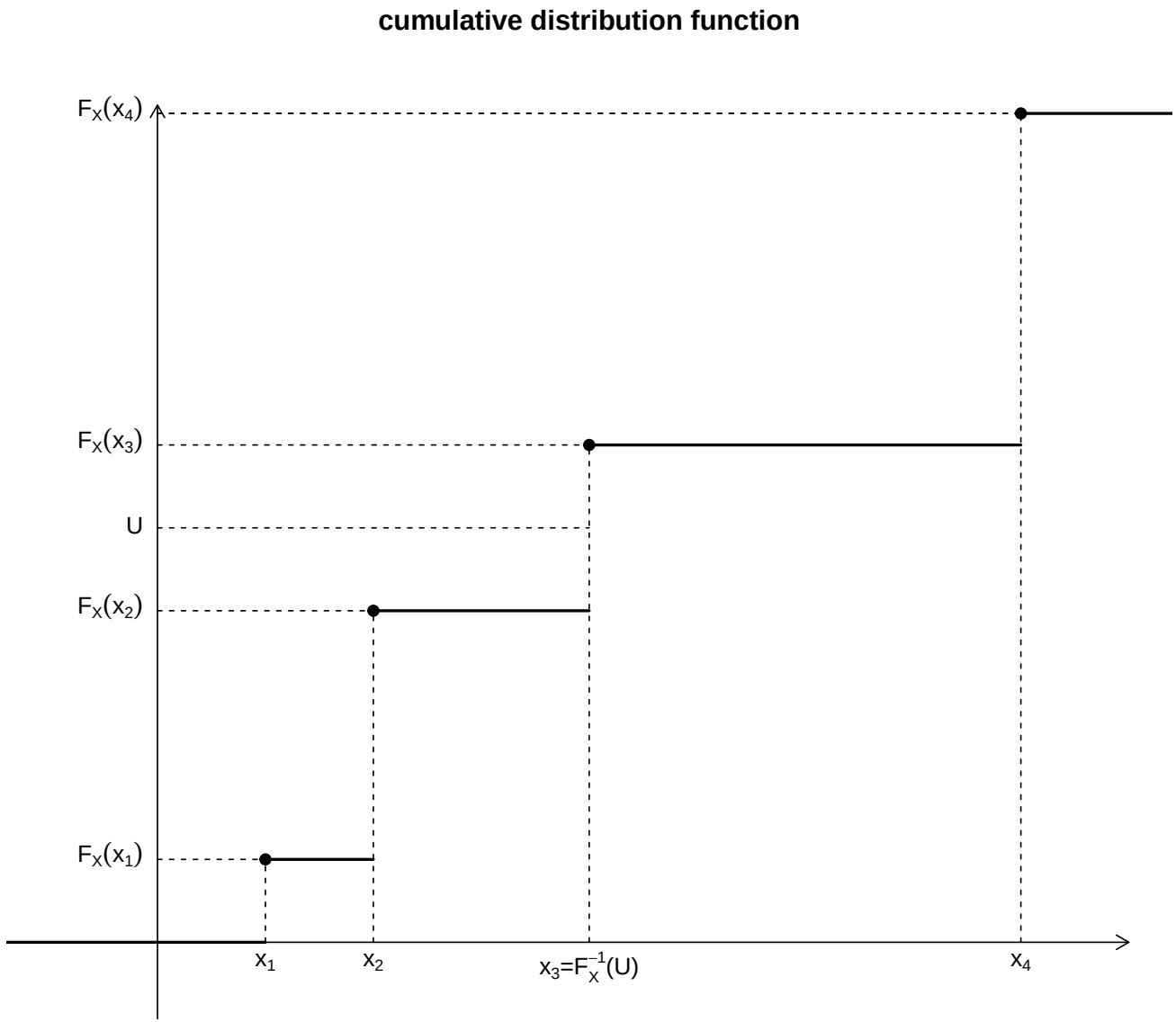
TRASFORMAZIONE MEDIANTE INVERSA

- l'algoritmo di simulazione esegue i seguenti passi:
  1. genera U da $U(0,1)$
  2. trova l'intero i tale che $F_X(x_{i-1}) < U \leq F_X(x_i)$
     (dove si è posto, per convenzione, $F_X(x_0) = 0$)
  3. restituisce x_i come valore simulato di X (cioè $X' = x_i$)
- si può facilmente verificare che X' così ottenuto ha la stessa distribuzione di X:

$$\Pr[X' = x_i] = \Pr[F_X(x_{i-1}) < U \leq F_X(x_i)] = F_X(x_i) - F_X(x_{i-1}) = p_i$$

- dal punto di vista pratico, per calcolare x_i si può eseguire
  - ▷ una ricerca sequenziale
  - ▷ una ricerca binaria data la monotonia di F_X (e quindi anche della sua inversa generalizzata)
  - ▷ approssimare l'inversa generalizzata tramite una funzione analitica (ad es. un polinomio di terzo grado), calcolarla in U in modo da porsi vicino alla soluzione cercata, e poi proseguire la ricerca (in avanti o all'indietro) sequenzialmente

TRASFORMAZIONE MEDIANTE INVERSA



TRASFORMAZIONE MEDIANTE INVERSA

- ESEMPIO 1. si consideri una v.a. $X \sim$Poisson(μ):

$$\Pr[X = i] = p_i = \mathrm{e}^{-\mu}\frac{\mu^i}{i!}$$

se ne calcolino tre valori simulati nel caso $\mu = 5$ corrispondenti a $u_1 = 0.47327$, $u_2 = 0.28951$, $u_3 = 0.20806$

  - ▷ tabuliamo la funzione di ripartizione di X:

i	$\Pr[X = i]$	$\Pr[X \leq i]$
0	0.006738	0.006738
1	0.033690	0.040428
2	0.084224	0.124652
3	0.140374	0.265026
4	0.175467	0.440493
5	0.175467	0.615961
⋮	⋮	⋮

  - ▷ i tre valori simulati di X sono $t_1 = 5$, $t_2 = 4$, $t_3 = 3$

TRASFORMAZIONE MEDIANTE INVERSA

- ESEMPIO 2. sia $\bar{T}_y$ la durata residua di vita di un individuo di età y arrotondata per eccesso all'intero più vicino:

$$\bar{T}_y = i \quad \Leftrightarrow \quad i-1 < T_y \leq i, \quad , \quad i = 1,2,\ldots,$$

  - ▷ se si estraggono da una tavola di sopravvivenza ℓ le probabilità di morte annuali (salti della funzione di ripartizione),

$$\frac{\ell_{y+i-1} - \ell_{y+i}}{\ell_y}, \qquad i = 1,2,\ldots,\omega - y,$$

  - ▷ il valore simulato $\bar{T}'_y = i \quad \Leftrightarrow \quad 1 - \frac{\ell_{y+i-1}}{\ell_y} < U \leq 1 - \frac{\ell_{y+i}}{\ell_y}$

TRASFORMAZIONE MEDIANTE INVERSA

- ESEMPIO 3. come caso particolare della simulazione di una v.a. discreta, possiamo simulare un evento A di probabilità $\Pr[A] = p$
  - ▷ basta infatti considerare la variabile indicatore

$$I_A = \begin{cases} 1 & \text{se } A \text{ è vero} \\ 0 & \text{se } A \text{ è falso} \end{cases}$$

  - ▷ l'algoritmo di simulazione diventa dunque:
    1. simula U da $U(0,1)$
    2. se $U > 1-p$ allora A' (evento sostituto di A) è VERO, altrimenti A' è FALSO
  - ▷ poiché $\Pr[U > 1-p] = \Pr[U < p] = p$, è equivalente rimpiazzare il passo 2. con

    2'. se $U < p$ allora A' è VERO, altrimenti A' è FALSO

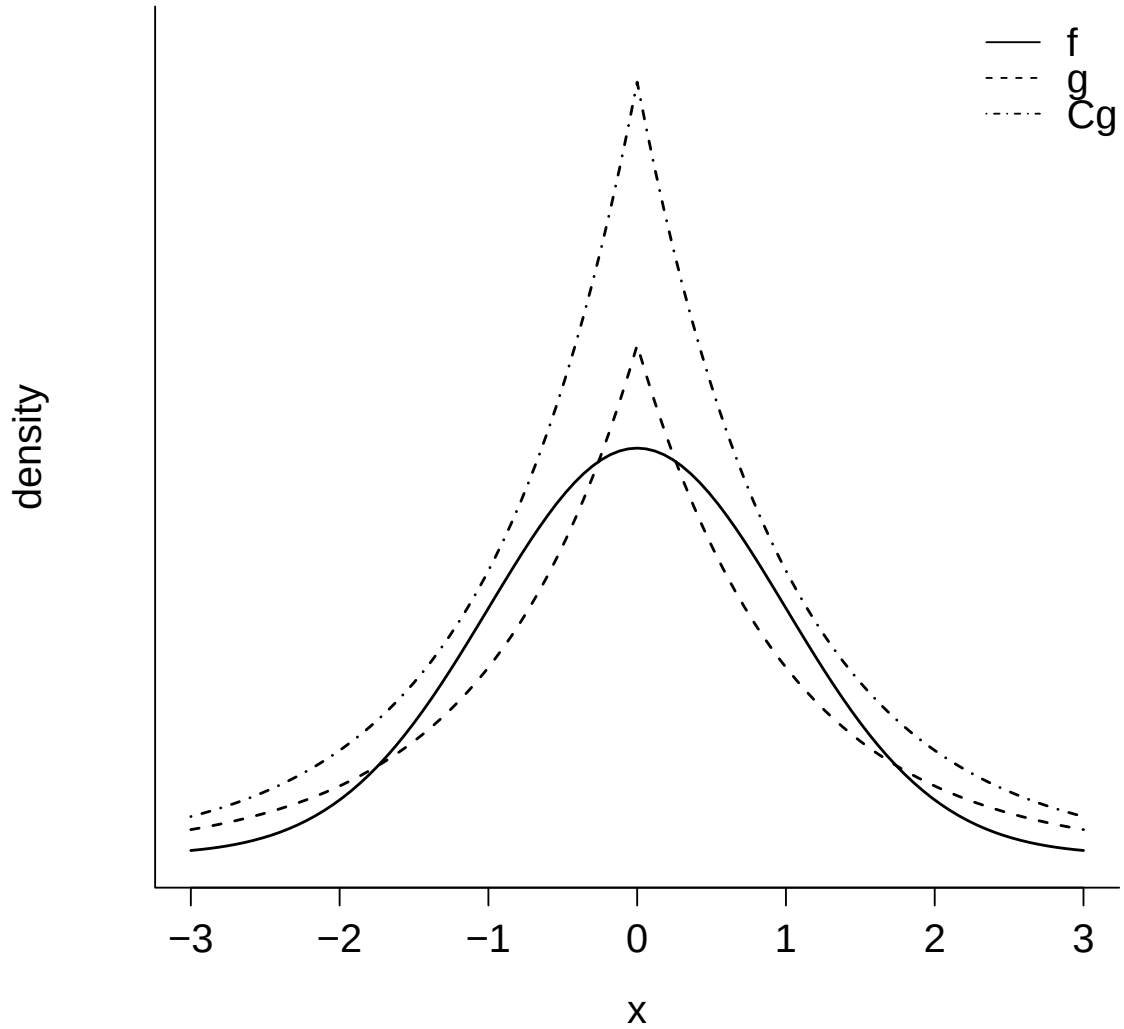
METODO DI ACCETTAZIONE-RIFIUTO

- si supponga di voler simulare una v.a. X con funzione di densità f, e di saper simulare v.a. con densità g
- si supponga che esista un numero $C>1$ tale che

$$f(x) \leq C \cdot g(x) \text{ per ogni } x \in \mathbb{R}$$

  - ▷ in altre parole, il grafico di g può essere 'gonfiato' in modo da dominare quello di f
  - ▷ ovviamente la scelta di g non è libera ma è vincolata da f
- si scelga a caso un punto (Z_1, Z_2) sotto la curva $C \cdot g$; se il punto cade anche sotto la curva f allora l'ascissa dello stesso, Z_1, viene presa (accettata) come valore simulato di X, altrimenti viene rifiutata e si ritenta con un nuovo punto
- si può dimostrare che tale procedimento genera v.a. con densità f

METODO DI ACCETTAZIONE-RIFIUTO



METODO DI ACCETTAZIONE-RIFIUTO

- l'algoritmo di simulazione prevede i seguenti passi:
  1. simula U da $U(0,1)$ e Z_1 da g in condizioni di indipendenza
  2. poni $Z_2 = C \cdot g(Z_1) \cdot U \rightsquigarrow 0 \leq Z_2 \leq C \cdot g(Z_1) \rightsquigarrow (Z_1, Z_2)$ sta sotto la curva $C \cdot g$
  3. se $Z_2 \leq f(Z_1)$ allora accetta Z_1 come determinazione simulata di X, altrimenti rifiuta e torna al passo 1.
- la probabilità di accettazione è pari a

$$\Pr\left[C \cdot g(Z_1) \cdot U \leq f(Z_1)\right] = \Pr\left[U \leq \frac{f(Z_1)}{C \cdot g(Z_1)}\right]$$

$$= E\left[\Pr\left[U \leq \frac{f(Z_1)}{C \cdot g(Z_1)} \middle| Z_1\right]\right] = E\left[\frac{f(Z_1)}{C \cdot g(Z_1)}\right]$$

$$= \int_{-\infty}^{+\infty} \frac{f(y)}{C \cdot g(y)} g(y)\mathrm{d}y = \frac{1}{C}$$

METODO DI ACCETTAZIONE-RIFIUTO

$\rightsquigarrow$ tale risultato è d'altronde immediato se si pensa alla probabilità come quoziente tra 'casi favorevoli' e 'casi possibili' misurati, rispettivamente, dall'area sottostante il grafico di f e quella sottostante il grafico di $C \cdot g$:

$$\frac{\int_{-\infty}^{+\infty} f(y)\mathrm{d}y}{C \int_{-\infty}^{+\infty} g(y)\mathrm{d}y} = \frac{1}{C}$$

$\rightsquigarrow$ il numero aleatorio di tentativi da fare prima di ottenere un'accettazione ha distribuzione geometrica di parametro $1/C$ e quindi il suo valore atteso è pari a C

$\rightsquigarrow$ la costante C deve essere tanto più piccola possibile

- il valore ottimo di C viene determinato risolvendo il problema

$$C^* = \max_{x \in \mathbb{R}} \frac{f(x)}{g(x)}$$

$\rightsquigarrow$ $C \cdot g$ dev'essere quanto più possibile vicino a f

METODO DI ACCETTAZIONE-RIFIUTO

- Il C^* ottenuto risolvendo il problema di massimo precedente è effettivamente il

$$\min\{C : f(x) \leq C \cdot g(x) \text{ per ogni } x \in \mathbb{R}\}$$

  - ▷ Infatti, indicando con x^* un punto di massimo della funzione f/g, si ha che

$$C^* = \frac{f(x^*)}{g(x^*)} \geq \frac{f(x)}{g(x)} \; \forall\, x \in \mathbb{R} \Rightarrow C^* \cdot g(x) \geq \frac{f(x)}{g(x)} \cdot g(x) = f(x) \; \forall\, x \in \mathbb{R}$$

  - ▷ Se $C < C^* = \frac{f(x^*)}{g(x^*)}$, allora $C \cdot g(x^*) < \frac{f(x^*)}{g(x^*)} \cdot g(x^*) = f(x^*)$

METODO DI ACCETTAZIONE-RIFIUTO

- ESEMPIO 1. si consideri la distribuzione normale standard:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \; x \in \mathbb{R}$$

- vogliamo applicare il metodo di accettazione-rifiuto usando una distribuzione di Laplace, che ha densità

$$g(x) = \frac{1}{2} \exp(-|x|)$$

METODO DI ACCETTAZIONE-RIFIUTO

- ESEMPIO 1 (continua) scegliamo questa distribuzione perché
  - ▷ è simmetrica come la normale e ha code più spesse (si veda grafico a p. 25)
  - ▷ può essere facilmente simulata usando il metodo di trasformazione tramite inversa. La sua funzione di ripartizione G e l'inversa G^{-1} sono rispettivamente date da

$$G(x) = \begin{cases} \frac{1}{2}\int_{-\infty}^{x} \mathrm{e}^{t}\mathrm{d}t = \frac{1}{2}\mathrm{e}^{x} & \text{se } x \leq 0 \\ \\ \frac{1}{2}\int_{-\infty}^{0} \mathrm{e}^{t}\mathrm{d}t + \frac{1}{2}\int_{0}^{x} \mathrm{e}^{-t}\mathrm{d}t = 1 - \frac{1}{2}\mathrm{e}^{-x} & \text{se } x \geq 0 \end{cases}$$

$$G^{-1}(p) = \begin{cases} \log(2p) & \text{se } 0 < p \leq \frac{1}{2} \\ \\ -\log\left(2(1-p)\right) & \text{se } \frac{1}{2} \leq p < 1 \end{cases}$$

METODO DI ACCETTAZIONE-RIFIUTO

- ESEMPIO 1 (continua)
  - ▷ troviamo ora il valore ottimo C^*, massimo della funzione:

$$\frac{f(x)}{g(x)} = \sqrt{\frac{2}{\pi}}\, \mathrm{e}^{-\frac{x^2}{2}+|x|}$$

  - ▷ tale funzione è pari (grafico simmetrico rispetto all'asse delle ordinate) con derivata, per $x \neq 0$

$$\sqrt{\frac{2}{\pi}}\, \mathrm{e}^{-\frac{x^2}{2}+|x|} \left(-x + \frac{|x|}{x}\right)$$

  - ▷ presenta due punti di massimo, $x^* = \pm 1$
  - ▷ il suo massimo è quindi

$$C^* = \frac{f(\pm 1)}{g(\pm 1)} = \sqrt{\frac{2\,\mathrm{e}}{\pi}} \approx 1.31$$

METODO DI ACCETTAZIONE-RIFIUTO

- ESEMPIO 1 (continua)
  - ▷ la condizione di accettazione

$$U \leq \frac{f(Z_1)}{C^* \cdot g(Z_1)} = \exp\left(-\frac{1}{2}\left(|Z_1| - 1\right)^2\right)$$

è equivalente a

$$-2\log U \geq \left(|Z_1| - 1\right)^2$$

- l'algoritmo di simulazione è quindi:
  1. simula U_1 e U_2 da $U(0,1)$ (in condizioni di indipendenza stocastica); siano u_1 e u_2 le corrispondenti determinazioni simulate
  2. calcola la determinazione simulata di Z_1: $z_1 = G^{-1}(u_1)$
  3. se $-2\log u_2 \geq \left(|z_1| - 1\right)^2$ restituisci z_1 come determinazione simulata di X, altrimenti torna al passo 1.
- la probabilità di accettazione è

$$\frac{1}{C^*} \approx 0.76$$

METODO DI ACCETTAZIONE-RIFIUTO

- ESEMPIO 2. sia T_y la durata residua di vita di un individuo di età y con intensità di mortalità di Makeham, $\mu(z) = \gamma + \alpha e^{\beta z}$, $\alpha, \beta, \gamma > 0$. Vogliamo simulare T_y usando come densità dominante quella di una Gompertz, che si ottiene dalla precedente ponendo $\gamma = 0$
  - ▷ la funzione di ripartizione di T_y è

$$F(x) = 1 - \exp\left(-\int_0^x \mu(y+v)\mathrm{d}v\right) = 1 - \exp\left(-\gamma x - \frac{\alpha}{\beta} e^{\beta y}(e^{\beta x} - 1)\right)$$

  - ▷ la densità di T_y è

$$f(x) = F'(x) = \left(\gamma + \alpha e^{\beta(y+x)}\right) \exp\left(-\gamma x - \frac{\alpha}{\beta} e^{\beta y}(e^{\beta x} - 1)\right)$$

  - ▷ la densità della Gompertz ($\rightsquigarrow \gamma = 0$) è

$$g(x) = \alpha \exp\left(\beta(y+x) - \frac{\alpha}{\beta} e^{\beta y}(e^{\beta x} - 1)\right)$$

METODO DI ACCETTAZIONE-RIFIUTO

- ESEMPIO 2 (continua)
  - ▷ troviamo ora il valore ottimo C^*, massimo della funzione:

$$\frac{f(x)}{g(x)} = \exp(-\gamma x) + \frac{\gamma}{\alpha}\exp\left(-\beta y - (\beta+\gamma)x\right)$$

  - ▷ tale funzione ha derivata

$$-\gamma\exp(-\gamma x) - (\beta+\gamma)\frac{\gamma}{\alpha}\exp\left(-\beta y - (\beta+\gamma)x\right) < 0$$

  - ▷ trattandosi di una funzione strettamente decrescente, il suo massimo è

$$C^* = \frac{f(0)}{g(0)} = 1 + \frac{\gamma}{\alpha}\exp\left(-\beta y\right)$$

METODO DI ACCETTAZIONE-RIFIUTO

- ESEMPIO 2 (continua)
  - ▷ la condizione di accettazione è

$$U \leq \frac{f(Z_1)}{C^* \cdot g(Z_1)} = \frac{\alpha e^{-\gamma Z_1} + \gamma e^{-\beta y - (\beta+\gamma) Z_1}}{\alpha + \gamma e^{-\beta y}}$$

- l'algoritmo di simulazione è quindi:
  1. simula U_1 e U_2 da $U(0,1)$ (in condizioni di indipendenza stocastica); siano u_1 e u_2 le corrispondenti determinazioni simulate
  2. calcola la determinazione simulata di Z_1:

$$z_1 = G^{-1}(u_1) = \frac{1}{\beta} \log\left(1 - \frac{\beta}{\alpha} e^{-\beta y} \log(u_1)\right)$$

     dove G è la funzione di ripartizione della Gompertz
  3. restituisci z_1 come determinazione simulata di T_y se

$$u_2 \leq \frac{\alpha e^{-\gamma z_1} + \gamma e^{-\beta y - (\beta+\gamma) z_1}}{\alpha + \gamma e^{-\beta y}},$$

     altrimenti torna al passo 1.

METODI AD HOC

- i metodi di accettazione-rifiuto e di trasformazione tramite inversa sono generali e possono essere applicati in pratica a molte situazioni
- alcuni metodi ad hoc possono essere usati per simulare v.a. in situazioni particolari, sfruttando la loro specifica struttura; vediamo alcuni semplici esempi:
  - ▷ per simulare una v.a. $Y \sim N(\mu, \sigma^2)$ basta simulare una normale standard X e poi usare la trasformazione $Y = \mu + \sigma X$
  - ▷ per simulare una v.a. Y con distribuzione lognormale basta simulare la normale $X = \log Y$ e poi usare la trasformazione $Y = \exp(X)$

METODI AD HOC

- per simulare una v.a. Y con distribuzione binomiale di parametri (n,p) basta simulare n eventi indipendenti A_i, $i = 1,\ldots,n$, di comune probabilità p, e poi usare la trasformazione

$$Y = I_{A_1} + \ldots + I_{A_n}$$

- per simulare una v.a. Y distribuita secondo una χ^2 con n (intero) gradi di libertà basta simulare n normali standard indipendenti $X_1,\ldots,X_n$ e poi usare la trasformazione $Y = X_1^2 + \ldots + X_n^2$
  In alternativa, se n è pari, basta simulare $n/2$ esponenziali i.i.d. $Z_1,\ldots,Z_{n/2}$ di parametro $1/2$, e poi usare la trasformazione $Y = Z_1 + \ldots + Z_{n/2}$

METODI AD HOC - DISTRIBUZIONE NORMALE

- per simulare v.a. normali standard esistono vari procedimenti ad hoc, oltre a quelli generali di inversione della cdf (lo standard in **R**) e di accettazione-rifiuto illustrato nell'Esempio 1 di p. 28
- il più semplice è basato sul Teorema del limite centrale:
  - ▷ se $X_1, \ldots, X_n$ sono v.a. iid con media e varianza finite θ e τ^2, allora la loro media aritmetica standardizzata converge in legge verso una $N(0,1)$

$$\boxed{\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \theta}{\tau/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\theta}{\tau\sqrt{n}} \approx N(0,1)}$$

- in particolare, se $X_i \sim U(0,1) \rightsquigarrow \theta = 1/2$ e $\tau^2 = 1/12$
  - ▷ il numero magico che si sceglie nelle applicazioni pratiche è $n = 12$, che suggerisce il seguente algoritmo:
    1. simula $U_1, \ldots, U_{12}$ variabili i.i.d da $U(0,1)$ e siano $u_1, \ldots, u_{12}$ le corrispondenti determinazioni simulate
    2. prendi $\sum_{i=1}^{12} u_i - 6$ come determinazione simulata di una $N(0,1)$
  - ▷ malgrado n sia molto piccolo, il procedimento supera bene i test di normalità ed è molto rapido da utilizzare

METODI AD HOC - DISTRIBUZIONE NORMALE

- il metodo di Box-Muller consente di generare coppie di normali standard indipendenti a partire da coppie $U(0,1)$ indipendenti
- 1. genera U_1, U_2 indipendenti da $U(0,1)$ e siano u_1, u_2 le corrispondenti determinazioni simulate
  2. restituisci

$$z_1 = \sqrt{-2\log u_1}\,\cos(2\pi u_2)$$
$$z_2 = \sqrt{-2\log u_1}\,\sin(2\pi u_2)$$

  come determinazioni simulate di una coppia (Z_1, Z_2) di $N(0,1)$ indipendenti

- svantaggio: bisogna calcolare funzioni 'speciali' come sin, cos

METODI AD HOC - DISTRIBUZIONE NORMALE

- idea: se il punto del piano (Z_1, Z_2) ha coordinate normali standard indipendenti, si può dimostrare che il quadrato della sua distanza dall'origine, $R^2 = Z_1^2 + Z_2^2$, e l'angolo $\beta = \arctan \frac{Z_2}{Z_1}$ che la retta passante per $(0,0)$ e (Z_1, Z_2) forma con il verso positivo dell'asse x, sono indipendenti con distribuzione Exp(1/2) e $U(0, 2\pi)$
- infatti abbiamo già visto (Esempio 1 a p. 16) che per simulare una Exp(1/2) basta usare la trasformazione $-2\log U(0,1)$, mentre per simulare una $U(0,2\pi)$ basta usare $2\pi U(0,1)$
  $\rightsquigarrow$ quindi bisogna risolvere il seguente sistema:

$$\begin{cases} z_1^2 + z_2^2 = -2\log u_1 \\ \arctan \frac{z_2}{z_1} = 2\pi u_2 \end{cases}$$

METODI AD HOC - DISTRIBUZIONE NORMALE

- partiamo dalla seconda equazione:

$$\frac{z_2}{z_1} = \tan(2\pi u_2) = \frac{\sin(2\pi u_2)}{\cos(2\pi u_2)} \Rightarrow z_2 = z_1 \frac{\sin(2\pi u_2)}{\cos(2\pi u_2)}$$

- sostituendo z_2 nella prima equazione si ottiene:

$$z_1^2 \left[1 + \frac{\sin^2(2\pi u_2)}{\cos^2(2\pi u_2)} \right] = \left(\frac{z_1}{\cos(2\pi u_2)} \right)^2 = -2 \log u_1$$

$$\Rightarrow z_1 = \sqrt{-2 \log u_1} \cos(2\pi u_2)$$

- sostituendo infine z_1 nella seconda equazione si perviene a

$$z_2 = \sqrt{-2 \log u_1} \cos(2\pi u_2) \frac{\sin(2\pi u_2)}{\cos(2\pi u_2)} = \sqrt{-2 \log u_1} \sin(2\pi u_2)$$

METODI AD HOC - DISTRIBUZIONE NORMALE

- l'algoritmo di Polar-Marsaglia è una variante di quello di Box-Muller che evita l'utilizzo di funzioni trigonometriche ma richiede un test di accettazione-rifiuto (ad hoc)
- 1. genera U_1, U_2 da $U(0,1)$ e siano u_1, u_2 le corrispondenti determinazioni simulate
  2. poni $v_1 = 2u_1 - 1$, $v_2 = 2u_2 - 1$ e $s = v_1^2 + v_2^2$
  3. se $s > 1$ torna al passo 1. (rifiuta), altrimenti accetta restituendo

$$z_1 = \sqrt{\frac{-2\log s}{s}}\, v_1$$

$$z_2 = \sqrt{\frac{-2\log s}{s}}\, v_2$$

  come determinazioni simulate della coppia (Z_1, Z_2) di normali standard indipendenti
- qui il punto (v_1, v_2) appartiene al quadrato di vertici $(\pm 1, \pm 1)$ ed è rigettato se non appartiene anche al cerchio unitario
- la probabilità di accettazione è $\frac{\pi}{4} \approx 0.78$

SIMULAZIONE DI V.A. MULTIDIMENSIONALI

- per simulare un vettore di v.a. $(X_1, X_2, \ldots, X_m)$ (che potrebbe rappresentare, ad es., la traiettoria di un processo stocastico) si procede come segue:
- 1. simula X_1 dalla distribuzione marginale F_{X_1}, e sia x_1 la sua determinazione simulata
  2. simula X_2 dalla distribuzione condizionata $F_{X_2|X_1=x_1}$, e sia x_2 la sua determinazione simulata

  ...

  m. simula X_m dalla distribuzione condizionata $F_{X_m|X_1=x_1, X_2=x_2, \ldots, X_{m-1}=x_{m-1}}$
- ovviamente, se $X_1, X_2, \ldots, X_m$ sono stocasticamente indipendenti, si usano le distribuzioni marginali per simulare tutte le variabili del vettore
- ci sono anche dei metodi ad hoc per particolari processi stocastici, in cui ci si riconduce al caso di variabili indipendenti (ad es. moto Browniano, processo di Poisson, ...)

ANALISI DEI RISULTATI

- abbiamo visto all'inizio che, se $X_1, X_2, \ldots, X_n$ sono iid con media finita θ, la loro media aritmetica

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  converge quasi certamente a θ (legge dei grandi numeri), per cui si può usare $\widehat{\theta}_n$ come stimatore della media θ, se incognita
- abbiamo inoltre visto che, se tali variabili hanno anche varianza finita τ^2, la media aritmetica standardizzata, $(\widehat{\theta}_n - \theta)/(\tau/\sqrt{n})$, converge in legge ad una $N(0,1)$ (teorema del limite centrale)
- ci si pone allora il problema dell'accuratezza della stima

ANALISI DEI RISULTATI

- approssimando dunque $(\widehat{\theta}_n - \theta)/(\tau/\sqrt{n}) \approx N(0,1)$ otteniamo

$$\Pr\left[-z_{1-\alpha/2} \leq \frac{\widehat{\theta}_n - \theta}{\tau/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1-\alpha$$

  dove $z_p = \Phi^{-1}(p)$ è il p-esimo quantile della distribuzione normale standard, con Φ funzione di ripartizione della stessa

- l'espressione precedente può essere anche riscritta come

$$\Pr\left[\widehat{\theta}_n - z_{1-\alpha/2}\frac{\tau}{\sqrt{n}} \leq \theta \leq \widehat{\theta}_n + z_{1-\alpha/2}\frac{\tau}{\sqrt{n}}\right] = 1-\alpha$$

  dove si evidenzia un intervallo di confidenza per θ, usualmente associato ad un valore di α 'piccolo', ad es. 1%, o 5%

ANALISI DEI RISULTATI

- poiché solitamente anche τ^2 non è noto, lo si può stimare tramite

$$\widehat{\tau}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\theta}_n)^2$$

- l'intervallo di confidenza diventa allora

$$\left(\widehat{\theta}_n - z_{1-\alpha/2}\frac{\widehat{\tau}_n}{\sqrt{n}}, \widehat{\theta}_n + z_{1-\alpha/2}\frac{\widehat{\tau}_n}{\sqrt{n}}\right)$$

ANALISI DEI RISULTATI

- ci chiediamo ora quante simulazioni dobbiamo effettuare per ottenere una fissata precisione
  - ▷ se ci concentriamo sull'errore assoluto $|\widehat{\theta}_n - \theta|$ e vogliamo maggiorarlo con δ si ha:

$$\Pr[|\widehat{\theta}_n - \theta| \leq \delta] \geq 1 - \alpha \rightsquigarrow \delta \geq z_{1-\alpha/2} \frac{\widehat{\tau}_n}{\sqrt{n}} \rightsquigarrow n \geq z^2_{1-\alpha/2} \frac{\widehat{\tau}^2_n}{\delta^2}$$

  - ▷ se ci concentriamo sull'errore relativo $|\widehat{\theta}_n - \theta|/|\theta|$ e vogliamo maggiorarlo con ε otteniamo:

$$\Pr\left[\frac{|\widehat{\theta}_n - \theta|}{|\theta|} \leq \varepsilon\right] \geq 1 - \alpha \rightsquigarrow n \geq z^2_{1-\alpha/2} \frac{\widehat{\tau}^2_n}{\theta^2 \varepsilon^2} \approx z^2_{1-\alpha/2} \frac{\widehat{\tau}^2_n}{\widehat{\theta}^2_n \varepsilon^2}$$

- poiché $\widehat{\tau}^2_n$ e $\widehat{\theta}_n$ non sono noti a priori, bisogna preliminarmente eseguire un certo numero di simulazioni per calcolarli, e poi verificare se tale numero è o meno sufficiente allo scopo, eventualmente adattando la procedura fino al raggiungimento dell'obiettivo (anche se in questo contesto di solito ci si accontenta di un campione ridotto di simulazioni)

ANALISI DEI RISULTATI

- ESEMPIO. si supponga che un assicuratore voglia stimare il numero atteso di sinistri θ in un dato portafoglio; l'analisi preliminare suggerisce che $\widehat{\theta}_n = 750$ e $\widehat{\tau}_n = 400$
  - ▷ quante simulazioni sono richieste per maggiorare l'errore assoluto con $\delta = 10$?
  - ▷ quante per maggiorare quello relativo con $\varepsilon = 0.5\%$?

  (intervallo di confidenza del 95% $\rightsquigarrow$ $\alpha = 5\%$ $\rightsquigarrow$ $z_{1-\alpha/2} = 1.96$)
  - ▷ per l'errore assoluto si ottiene

$$n \geq \frac{1.96^2 \cdot 400^2}{10^2} \approx 6147$$

  - ▷ per quello relativo si ha

$$n \geq \frac{1.96^2 \cdot 400^2}{750^2 \cdot 0.005^2} \approx 43709$$