La generazione di campioni pseudocasuali da densità continue univariate (*)

Agostino Tarsitano
Università degli studi della Calabria
Dipartimento di Economia e Statistica
87030 Arcavacata di Rende (Cs)
agotar@unical.it

Riassunto.
In questo lavoro si discutono le tecniche correntemente in uso per la simulazione di densità continue, incluse quelle miste. In particolare, si parlerà della tecnica di inversione e della tecnica di trasformazione. Si commenta inoltre l'uso del computer (pregi e difetti) per la generazione dei numeri pseudo- casuali e si approfondirà il ruolo degli schemi congruenziali lineari con varie applicazioni dell'algoritmo di Wichmann e Hill.

keywords: numeri pseudocasuali, algoritmo di Wichmann-Hill

(*) Rapporto tecnico. Dipartimento di Economia politica, 1995.

L’autore si è limitato a riportare in forma digitale il testo e; laddove possibile, le tabelle e le formule del lavoro originale. Quando il risultato dello scanner si è rivelato inadeguato, formule e grafici sono stati rifatti. Solo la veste grafica è cambiata Solo qualche evidente errore ortografico od imperfezione nelle espressioni analitiche è stato modificato.

1. Introduzione.

I numeri pseudocasuali sono utilizzati in una ampia gamma di applicazioni tra cui:

1) Metodi Montecarlo;
2) Studio campionario ed asintotico;
3) Organizzazione di basi di dati;
4) Simulazione di modelli matematici;
5) Ricerca operativa.

Il secondo punto rappresenta l'interesse maggiore del presente lavoro il cui obiettivo è l'esposizione e l'approfondimento di alcuni algoritmi adoperati nei computer per la simulazione di campioni casuali da densità continue univariate.

Le finalità di questa simulazione sono ottimamente illustrate già nel lavoro di Teichroew (1965) che delinea in modo nitido e preciso l'evoluzione della *Distribution Sampling* iniziata da Student (1908a e 1908b).

Il problema, a grandi linee è il seguente: sia $S= (X_1, X_2,...,X_n)$ un campione casuale da una di densità $f(x)$ e sia $h(S)$ una statistica cioè una funzione delle osservazioni campionarie. Capita talvolta che la funzione di distribuzione di h(.)

$$H(z) = Pr(h(S) \leq z) \tag{1.1}$$

non possa essere espressa in forma esplicita oppure che la formula risultante si riveli poco trattabile. Si può allora costruire una tabella numerica che fornisca i valori di $H(z)$ per una opportuna selezione di valori della z.

Una situazione analoga si ha quando $h(S)$ è uno stimatore di una qualche caratteristica di una data distribuzione. In questo caso occorre stabilire quale sia la distribuzione campionaria ovvero quella asintotica dello stimatore, oppure da quale ampiezza campionaria in poi un'altra distribuzione costituisca una "buona approssimazione" di quella, solitamente incognita, di $h(S)$.

Per risolvere questo tipo di problemi vengono simulati L campioni casuali indipendenti di ampiezza n

$$S_i= (X_{1i}, X_{2i},...,X_{ni}); \quad i=1, 2 , ..., L \tag{1.2}$$

Per ogni campione si calcola la statistica $h(S)$ e per poi costruire con questi la distribuzione empirica della statistica o dello stimatore specifico

$$H(z) = \frac{\#\left[h(S) \leq z\right]}{L} \tag{1.3}$$

se i campioni sono in effetti casuali e indipendenti e provenienti dalla $f(x)$ allora $H(z)$ converge uniformemente per ogni z ad $H(z)$ (Teorema di Glivenko-Cantelli). Ne consegue che per L sufficientemente grande la (1.3) possa sostituire la (1.1) a tutti gli effetti.

La casualità dei campioni e l'accuratezza dei valori in essi inclusi è essenziale per lo studio campionario ed asintotico. Esistono diverse tecniche per realizzare la simulazione anche se in questo lavoro daremo esclusivo spazio alla generazione di numeri pseudocasuali con il metodo congruenziale lineare (paragrafo 2). Nel paragrafo 3 si discuterà uno specifico algoritmo per simulare numeri pseudo-causali dalla distribuzione uniforme. Nel

paragrafo 4 si faranno considerazioni sul problema della discretizzazione. Successivamente, si discuteranno le tecniche per generare campioni da densità continue univariate. In particolare, nel quinto paragrafo si approfondirà la tecnica dell'inversione e nel sesto la tecnica delle trasformazioni. Infine, nel paragrafo 7, si parlerà della generazione di campioni casuali da distribuzioni miste.

2. Generazione di numeri pseudocasuali

La simulazione di numeri casuali con il computer si basa su sequenze di numeri che scaturiscono da ben definite relazioni matematiche e per questo si parla di numeri pseudocasuali: infatti, pur conservando un precipuo carattere deterministico, il loro comportamento è, per molti versi, assimilabile a quello di una successione di variabili casuali. La presenza di legami funzionali tra di essi è considerata ininfluente poiché non affiora a livello statistico.

L'evoluzione dei generatori di numeri ha una sua pietra miliare nei generatori congruenziali lineari introdotti da Lehmer (1949). La loro formula è

$$X_n = (aX_{n-1} + c) \; Mod \; m; \qquad n=1,2, ..., m \tag{2.1}$$

dove $\{X_i\}$ è una successione di interi, $[.]$ è la funzione parte intera e $\equiv$ è il simbolo di congruenza cioè $p \equiv q \; Mod \; m$ significa che

$$p = q - \left[\frac{q}{m}\right] m \tag{2.2}$$

ovvero p è il resto intero della divisione di q per m. Ad esempio $5 \equiv 17$ Mod 12 risulta da: 17 - [17/12]* 12 = 17 - 12 = 5. La formula ricorsiva (2.1) è molto semplice, di agevole traduzione nel linguaggio interno di ogni computer e può quindi risultare di rapidissima esecuzione. In essa compaiono quattro parametri (tutti numeri interi)

X_o Valore di partenza $X_0 \geq 0$ *sec* $>O$ *e* $X_0 > 0$ *se* $c = 0$;
a Moltiplicatore $1 < a < m$;
c Incremento $0 \leq c < m$;
m Modulo $m > 2$;

se $c=0$ i generatori sono detti puri in contrasto alla denominazione di misti attribuita agli schemi con $c>0$. L'uso della (2.1) di rado coinvolge direttamente gli interi che da essa scaturiscono. Piuttosto, poiché $0 \leq X_n < m$ si opera con le frazioni

$$U_n = \frac{X_n}{m}, \quad n = 0,1,2,\ldots,n-1 \tag{2.3}$$

Le sequenze ottenute dalla (2.1) sono deterministiche. Dato un certo termine è possibile calcolarne qualsiasi altro che interverrà in successione, ad esempio il k-esimo termine è

$$X_k = \left[aX_0 + \frac{\left(a^k - 1\right)c}{a-1} \right] Mod \;\; m \tag{2.4}$$

La conseguenza è che ogni successione ottenuta dalla (2.1) è riproducibile: per ripetere integralmente una stessa successione di numeri pseudo casuali è sufficiente conservarne il valore iniziale. Un limite dei generatori congruenziali lineari è che possono produrre al massimo *m* numeri pseudocasuali diversi; giunto all'm-esimo la (2.1) entra in ciclo riprendendo la successione dal punto iniziale X_o. Il numero di valori diversi prodotto prima di ripetersi costituisce il periodo del generatore. E' evidente che un periodo elevato è uno dei desiderata richiesti ai generatori.

L'estrema essenzialità dello schema ed i risultati con esso ottenuti ha indotto una ampia ricerca sulla definizione di criteri che portino alla scelta dei parametri *(X_o, a, c, m)* tali da assicurare il periodo pieno del generatore per valori elevati del modulo e la "casualità" delle serie da essi generate. L'argomento è troppo vasto anche solo per un breve accenno in questa sede. Il rinvio d'obbligo è al magnifico testo di Knuth (1981).

3. L'algoritmo di Wichmann-Hlll

Una variante dei generatori congruenziali che si è ormai consolidata in letteratura prevede l'uso congiunto di più generatori, preferibilmente di tipo puro (1a cui esecuzione è più rapida comportando meno operazioni), Fra i molti programmi comparsi su varie riviste scientifiche merita attenzione quello suggerito da Wichmann e Hill (1982). L'algoritmo, noto con la sigla AS183, si basa sulla combinazione di tre generatori congruenziali puri

$$\begin{cases} X_n \equiv \left(171 X_{n-1}\right) Mod \;\; 30269 \\ Y_n \equiv \left(170 Y_{n-1}\right) Mod \;\; 30307 \\ Z_n \equiv \left(172 Z_{n-1}\right) Mod \;\; 30323 \end{cases} \tag{3.1}$$

Fissati i ,tre valori di partenza: X_O, Y_O e Z_o l'algoritmo procede generando ogni volta tre numeri pseudocasuali dalle (3.1) sommandoli poi in base alla formula

$$I_n = \frac{X_n}{30269} + \frac{Y_n}{30307} + \frac{Z_n}{30323} \tag{3.2}$$

Le tre frazioni componenti di In hanno distribuzione uniforme sull'intervallo unitario (estremi esclusi). Distribuzione analoga avrà pure la parte frazionaria di I_n

$$U_n \equiv I_n Mod \;\; 1 = I_n - \left[I_n \right] \tag{3.3}$$

che è poi il numero pseudo-casuale adoperato nelle applicazioni.

La scelta dei parametri, come provano Wichmann e Hill (1982, 1984), assicura che il periodo dei singoli generatori sia pieno (vi compaiono tutti i numeri minori del modulo tranne lo zero) mentre, per il generatore composto, il periodo è circa 6.95×10^{12}: se i numeri fossero generati alla velocità della luce, la successione comincerebbe a ripetersi non prima di otto mesi. Di conseguenza solo alcuni segmenti dell'algoritmo di Wichmann-Hill possono essere sottoposti a verifica.

D'acchito i generatori componenti ispirano una certa diffidenza a causa del moltiplicatore piuttosto piccolo rispetto al modulo. Secondo i suggerimenti di Knuth il modulo dovrebbe essere, in tutti i tre casi, almeno il doppio (soprattutto ai fini della correlazione lineare tra termini successivi). Gli autori affermano comunque che il loro algoritmo ha superato in maniera soddisfacente diversi test di casualità. A maggiore conforto si sono calcolate le autocorrelazioni

$$r_k(n) = \frac{12}{n} \sum_{i=k+1}^{n} (U_i - 0.5)(U_{i-k} - 0.5); \quad k = 1,2,3,4 \tag{3.4}$$

per vari valori di n e con valori iniziali: $X_o = 4649$, $Y_o = 13367$; $Z_o = 26317$. I risultati sono esposti nella tabella 1. In questa, ogni entrata è la media aritmetica calcolata sulle $L = 1000$ repliche di ogni situazione sperimentale.

Tabella_l: autocorrelazioni nel generatore di Wichmann-Hill

n	$r_1(n)$	$r_2(n)$	$r_3(n)$	$r_4(n)$
30	0.00651	-0.00373	0.00353	0.00577
60	0.01411	-0.00271	0.00331	-0.00144
120	-0.00010	0.00022	-0.00600	0.00116
200	-0.00297	0.00468	0.00000	0.00187
500	-0.00069	0.00213	-0.00041	0.00142
1000	0.00050	0.00109	0.00081	-0.00049
2000	-0.00118	0.00084	-0.00002	-0.00046
5000	0.00051	-0.00037	0.00027	-0.00009

come è facile constatare, solo in un'occasione la correlazione lineare è stata, in media, superiore all'l % e cioè per $n = 60$.

La routine di Wichmann e Hill è stata utilizzata in tutte le simulazioni necessarie per questo ed altri lavori. Alla luce delle esperienze fatte emergono due suggerimenti:
1) utilizzare sempre la doppia precisione aritmetica per la realizzazione della (3.3;
2) inserire un controllo sul valore di u_n se questo è l'argomento di un logaritmo o denominatore di frazione (così facendo si evitano sgradite interruzioni di lunghi e costosi programmi di simulazione).
Per il resto l'algoritmo è risultato del tutto affidabile.

4. Problemi di rappresentazione numerica

Si consideri una variabile casuale x con densità $dF(x) = f\{x\}dx$. La X viene simulata (o forse è meglio dire approssimata) dalla variabile pseudo-casuale X^* generata con il computer e che ha funzione di distribuzione F^*. In che relazione stanno X e X^*?

Una prima riflessione si impone a proposito del dominio delle due variabili: anche se lo spazio campionario della *X* è l'asse reale, quello della *X** finisce ai limiti imposti dalla capacità di rappresentazione della macchina. Per fare un esempio gli estremi di rappresentazione sul supermini V AX 11/780 sono

Tabella_2: estremi ra resentabili a vari livelli di precision

D- e F-Floating	[-2.9*10- ,1.7*1]
G-Floating	[-0.56*10-308,0.9*10308]
H-Floating	[-0.84*10-4932,0.59*104932]

La contrazione del campo di variazione, pure ovviabile con degli artifici di programmazione, non è, come si vede, molto severa dato che le distribuzioni statistiche più diffuse non hanno code così cospicue da impensierire a questi valori estremi. E' importante però ricordare che tutte le variabili casuali simulate con il computer sono in realtà delle variabili tronche che variano cioè all'interno dell'intervallo *[P, G]* dove *P* è il numero più piccolo e *G* il numero più grande accettabile dal computer.

Un'altra necessaria considerazione riguarda la precisione con cui il computer rappresenta i numeri. E' chiaro che in una architettura finita non si potranno mai rappresentare correttamente i numeri irrazionali o periodici che quindi rimangono esclusi dall'insieme dei valori rappresentabili. Con riferimento alla tabella 2 le cifre esatte sono

Tabella_3: cifre esatte a vari livelli di rappresentazione

F-Floating	7 cifre
D-Floating	16 cifre
G-Floating	15 cifre
H-Floating	33 cifre

Sia *b* il numero di cifre significative ed X_i un numero rappresentabile correttamente dal computer; allora lo stesso X_i rappresenterà tutti i numeri reali compresi nell'intervallo

$$]X_i - 0.5*10^{-b}, X_i + 0.5*10^{-b}[$$

in pratica X_i è il valore centrale della classe di ampiezza 10^{-b} chiamato in causa nelle elaborazioni al posto di tutti i membri della classe. Indichiamo con *I* l'insieme dei valori diversi che si possono trattare con il computer

$$I = \left\{ X_0^*, X_1^*, X_2^*, \ldots, X_k^* \right\} \quad con \quad X_0^* = P, \quad X_k^* = G \tag{4.1}$$

con *k* finito. L'insieme *I* è ordinato, ma i valori non sono equispaziati. Il dominio della *X** oltre ad essere limitato è anche discreto e con dei vuoti che diventano sempre più larghi man mano che ci si avvicina agli estremi (perché si riduce frazione di cifre esatte nella rappresentazione del numero). In questo senso è essenziale disporre di generatori che coinvolgano tutti i numeri rappresentabili nel computer e che questo arco sia il più ampio possibile. Ad esempio, nei generatori congruenziali, gli U_n sono tutti multipli di *(1/m)* dove *m* è il modulo. In pratica è la frazione *1/m* che stabilisce il vuoto minimo tra due numeri effettivamente generati: maggiore è il modulo più piccolo è l'intervallo in cui il generatore non produce numeri.

La funzione di distribuzione della variabile casuale simulata x* può essere formalizzata nel modo che segue

$$F*(x)=\begin{cases} x_i^*, \ i=0,1,2,\ldots,k \\ p_0=0, \quad p_k=1 \\ p_i=F\left(x_i^*\right)-F\left(x_{i-1}^*\right), \quad i=1,2,\ldots,k-1 \end{cases} \tag{4.2}$$

Affinché l'approssimazione di *X* a mezzo della *X** sia adeguata, il numero di cifre significative deve essere elevato ed inoltre la distribuzione da simulare deve avere code molto sottili per i valori vicini agli estremi rappresentabili dal computer. Holland (1975) ha studiato l'effetto del processo di discretizzazione sulla media e varianza di alcune variabili casuali continue concludendo che l'ordine di grandezza delle distorsioni è in funzione inversa del numero di cifre esatte e che ai livelli di precisione correntemente in uso (si vedano le tabelle 2 e 3) la distorsione dovuta agli arrotondamenti è trascurabile.

5. Simulazione per inversione

La simulazione di un gran numero di variabili casuali parte dalla simulazione dalla distribuzione uniforme nell'intervallo unitario. Il motivo della primarietà di questa distribuzione nel contesto della simulazione di variabili casuali è riassunto dal seguente teorema (Mood-Graybill-Boes,1974)

Teorema
Se la variabile casuale X ha funzione di densitˆ f(x) allora la variabile casuale

$$u=\int_{-\infty}^{x} f(t)dt \tag{5.1}$$

ha densitˆ uniforme sull'intervallo [0,1].

Per ottenere una osservazione X_i dalla densità *f(x)* bisognerà risolvere rispetto ad X_i l'equazione

$$\int_{-\infty}^{X_i} f(t)dt=U_i \tag{5.2}$$

dove U_i è stato generato dalla uniforme [0,1]. Questo però richiede che la *f(x)* non diventi mai identicamente nulla ovvero la *X* nella (5.2) deve essere una funzione monotona strettamente crescente della *U*, diciamo *X=h(u)*. La funzione inversa di *h(.)* è proprio la (5.1).

Vediamo in che modo sia possibile sfruttare il teorema per determinare la variabile casuale *X*. La sua funzione di ripartizione: *F(x)* consiste nella probabilità che *X* sia inferiore ad un valore prefissato di volta in volta: *F(x)=Pr(x<t)*. Se la X è sostituita la sua espressione in termini della *U* avremo

$$F(x)=Pr(x<t)=Pr\left(h(u)<t\right) \tag{5.3}$$

Poiché la *h(.)* è strettamente crescente, la disuguaglianza $h(u)<t$ è soddisfatta da tutti quei valori che verificano anche $u<h^{-1}(t)$. Ne consegue che

$$F(x) = Pr\left(u < h^{-1}(x)\right) \tag{5.4}$$

Il calcolo della probabilità in (5.4) non crea alcuna difficoltà in quanto la densità di probabilità della variabile casuale *U* è, nel nostro quadro di ragionamento, perfettamente nota e corrispondente alla uniforme *[0,1]*. Quindi

$$F(x) = \int_0^{h^{-1}(x)} f_U(t)dt = \int_0^{h^{-1}(x)} dt = h^{-1}(x) \tag{5.5}$$

dove $f_U(t)=1$ *per* $0<t<1$ è la densità della variabile casuale uniforme. Infatti, sostituendo alla inversa $h^{-1}(x)$ la sua espressione si otterrà:

$$h^{-1}(x) = u = \int_{-\infty}^{x} f(t)dt \tag{5.6}$$

In molte occasioni la *h(u)* può essere direttamente utilizzata ai fini della simulazione dalla densità *f(x)*. Si abbia ad esempio una variabile casuale con funzione di ripartizione *F(x)* del tipo Gompertz-Werhulst

$$F(x) = 1 - ae^{-\{g(x)\}} \tag{5.7}$$

dove *g(x)* è positiva, continua e strettamente crescente nel dominio della *X*. Se *U* ha densità uniforme, lo stesso sarà per *(1-U)*. Perciò:

$$1 - u = 1 - F(x) = 1 - ae^{-\{g(x)\}} \tag{5.8}$$

che ci porta alla relazione inversa: $x = g^{-1}\left[Ln\left(u/a\right)\right]$.

Un altro esempio lo si può vedere con la famiglia di distribuzioni di Burr

$$F(x) = \frac{1}{1 + e^{\{-g(x)\}}} \tag{5.9}$$

dove g(x) ha le medesime caratteristiche indicate per la (5.7). La (5.6) porta alla legge di inversione $x = g^{-1}\left[Ln\left(u/(1-u)\right)\right]$.

Nella tabella 4 sono riportate le trasformazioni della uniforme necessarie per generare alcune variabili casuali continue. L'efficacia del metodo di inversione più che sulla semplicità (che è solo apparente) dipende dalla qualità delle routine di sistema per operazioni che coinvolgono potenze, logaritmi, funzioni trigonometriche etc.

Tabella_4: formule di inversione per alcune variabili casuali continue

Denominazione	F(X)	Condizioni	Trasformazione
Arcoseno	$\frac{2}{\pi}\text{sen}^{-1}\left[\left(\frac{x}{2b}\right)^c\right]$	$0\le x\le 2b$	$x=2b\left[\text{sen}\left(\frac{\pi u}{2}\right)\right]^{1/c}$
Benini	$1-\left[\frac{x}{a}\right]^{\left\{-b\text{Ln}\left[\frac{x}{a}\right]\right\}}$	$x>a$	$x=e^{\left\{\text{Ln}(a)+\sqrt{-\frac{\text{Ln}(u)}{c}}\right\}}$
Burr III	$\left[\frac{1}{(1+x^{-a})}\right]^b$	$x>0$	$x=\left[\frac{1}{(u^{-1/a}-1)}\right]^{-1/b}$
Cauchy	$\frac{1}{2}+\frac{\text{Tan}^{-1}\left(\frac{x-a}{b}\right)}{\pi}$	$]-\infty;\infty[$	$x=a+b\text{Tan}\left(\pi\left[u-\frac{1}{2}\right]\right)$
Champernowne	$1-\frac{2^*\text{Tan}^{-1}\left(\frac{a}{x}\right)^b}{\pi}$	$0<a<\pi\ ;x>0$	$x=\frac{a}{\left(\text{Tan}\left(\frac{\pi}{2}-u\right)\right)^{1/b}}$
Dagum	$\alpha+\frac{(1-\alpha)}{[(1+x^{-a})]^b}$	$0\le\alpha\le 1;\ x>0$	$x=\left\{1-\left[\frac{(u-\alpha)}{(1-\alpha)}\right]^{-1/b}\right\}^{-1/a}$
Gumbel	$1-e^{\left\{-e^{\left\{\frac{x-a}{b}\right\}}\right\}}$	$]-\infty;\infty$	$x=a+b\text{Ln}[-\text{Ln}(u)]$
Gompertz	$1-\frac{1}{1+e^{\left\{\frac{x-a}{b}\right\}}}$	$]-\infty;\infty[$	$x=a+b\text{Ln}\left(\frac{1-u}{u}\right)$
Pareto	$1-\left[\frac{b-a}{x-a}\right]^c$	$x>a$	$x=a+\frac{b-a}{u^{1/c}}$
Perks	$\frac{1}{2}\left\{1+\text{Tanh}\left(\frac{x-a}{b}\right)\right\}$	$]-\infty;\infty[$	$x=a+b\text{Tanh}^{-1}(2u-1)$
Uniforme	$\left[\frac{x-a}{b-a}\right]^c$	$a\le x\le b$	$x=a+(b-a)u$
Weibull	$1-e^{\left\{-\left(\frac{x-a}{b}\right)^c\right\}}$	$x>a$	$x=a+b[-\text{Ln}(u)]^{1/c}$

6. Simulazione per trasformazione

Per molte variabili casuali la funzione di distribuzione inversa non può essere scritta in forma esplicita. Tale difficoltà può essere aggirate a mezzo di opportune trasformazioni applicate ad altre variabili casuali la cui funzione di distribuzione sia più trattabile. A questo proposito si riporta l'enunciato del teorema dato in (Mood-Graybill-Boes,1974)

Teorema

Siano X_1 ed X_2 due variabili casuali con densità congiunta $f_X(X_1,X_2)$ e sia inoltre $D_f=\{(X_1,X_2) \mid f_X(X_1,X_2)\geq 0\}$. Se valgono le seguenti ipotesi:

a) $Y_1=g_1(X_1,X_2)$ ed $Y_2=g_2(X_1,X_2)$ esprimono una relazione biunivoca tra D_f e D_g che è l'immagine di D_f costruita attraverso g_1 e g_2.

b) Le derivate di $X_1 = g_1^{-1}(X_1, X_2)$ e di $X_2 = g_2^{-1}(X_1, X_2)$ esistono e sono continue in D_g.

c) Lo jacobiano $|J|$ è diverso da zero in ogni punto di D_g.

Allora, la densità congiunta di Y_1 ed Y_2 sarà data da

$$f_y(Y_1, Y_2) = |J| f_x\left\{g_1^{-1}(X_1, X_2), g_2^{-1}(X_1, X_2)\right\}$$

Questo teorema è importante perché se f_x è semplice da simulare, una scelta accorta di g_1 e g_2 può rendere altrettanto semplice la simulazione della f_y.

Simulazione dalla densità normale e lognormale

Fra le molte tecniche per ottenere variabili pseudocasuali con densità normale, quella più semplice è la cosiddetta tecnica polare dovuta a Marsaglia-Bray (1964). Siano r_1 ed r_2 due variabili casuali indipendenti con distribuzione uniforme sull'intervallo *[-1,1]*. Sia inoltre

$$s = r_1^2 + r_2^2 \tag{6.1}$$

Se risulta $s \leq 1$, le trasformazioni:

$$\begin{cases} X_1 = r_1\sqrt{\dfrac{-2Ln(s)}{s}} \\ X_2 = r_2\sqrt{\dfrac{-2Ln(s)}{s}} \end{cases} \tag{6.2}$$

sono indipendenti e con distribuzione normale standardizzata. Quindi, per generare variabili pseudocasuali normali standardizzate occorre generare coppie di uniformi sull'intervallo *[-1,1]* fino a che non si verifichi la condizione (6.1). A questo punto le (6.2) sono le variabili normali desiderate. Poi, per ottenere variabili da una distribuzione normale con media μ e varianza σ^2 basterà porre $Y=\mu+\sigma X$. Infine, la trasformazione esponenziale $z=exp\{y\}$ darà luogo ad una variabile casuale lognormale con parametri μ e σ.

Simulazione dalla densità Gamma

La variabile casuale *X* è di tipo Gamma *G(a,b)* se la sua funzione di densità è

$$dF(x) = \frac{a^b}{\Gamma(b)} x^{b-1} e^{-ax} dx, \quad x > 0, \;\; a, b > 0 \tag{6.3}$$

Richiamiamo subito alcune proprietà della variabile casuale Gamma che semplificano molto la sua simulazione. Innanzitutto, se *U* ha distribuzione uniforme, la sua trasformata *-Ln(U)* ha densità esponenziale ovvero densità Gamma *G(l,l)* come si verifica facilmente. Inoltre, la somma di due variabili casuali indipendenti e con rispettiva distribuzione $G(a,b_1)$ e $G(a,b_2)$, è ancora di tipo Gamma con densità $G(a,b_l + b_2)$. Definiamo $b=m + r$ dove *m* è la parte intera e *r* la parte frazionaria. Se le $\{U_i , i=1,2, ..., m\}$ sono delle variabili casuali uniformi, allora

$$Z = -Ln\left(\prod_{i=1}^{m} U_i \right) \tag{6.4}$$

avrà densità Gamma *G(1,m)*. Consideriamo ora tre ulteriori variabili uniformi, diciamo U_{m+l}; U_{m+2}; U_{m+3}; e costruiamo le trasformazioni

$$\begin{cases} V_1 = u_{m+1}^{1/r} \\ V_2 = u_{m+2}^{1/(1-r)} \end{cases} \tag{6.5}$$

se risulta $s= V_I + V_2 \leq 1$ si potrà porre

$$Y = -\frac{V_1}{s} Ln\left(U_{m+3}\right)$$

ed Y avrà densità Gamma *G(l,r)* (Atkinson-Pierce, 1976). Ne consegue che la somma W= Y + Z sarà *G(l,b)* e la trasformazione di scala X=aY porterà infine ad una osservazione dalla *G(a,b)*.

Particolari combinazioni dei parametri *a* e *b* definiscono speciali distribuzioni che hanno un loro ruolo autonomo (si è già vista la relazione con la esponenziale). Se *b* è un intero (il che elimina la parte di simulazione relativa alla parte frazionaria *r*) la distribuzione *G(a,b)* è detta di Erlang. Se si pongono *a=2* e *b=(g/2)* con *g* intero si ottiene la variabile casuale χ^2. Un ulteriore esempio è la variabile di Maxwell ottenibile dalla Gamma con la combinazione $a=(l/c)^2$ e *b=1.5*.

Simulazione dalla densità Beta
La variabile casuale Beta è associata alla densità

$$dF(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}dx, \quad 0 \le x \le 1, \;\; a,b > \tag{6.6}$$

Per simulare la variabile casuale Beta si possono utilizzare le relazioni che la legano alla variabile casuale Gamma. Siano g_l e g_2 due variabili casuali indipendenti e con distribuzione *G(l/a,1)* e *G(l/b,l)* rispettivamente. La trasformazione

$$X = \frac{g_1}{g_1 + g_2} \tag{6.7}$$

avrà densità Beta con parametri *a* e *b*. La simulazione della Beta può quindi sfruttare quanto già predisposto per la simulazione della Gamma, o se risulta più efficiente, ci si può basare sulla simulazione dalla esponenziale. Le stesse argomentazioni sono valide per la simulazione della densità Beta del secondo tipo

$$dF(x) = \frac{x^{a-1}(1+x)^{-b/a}}{B(a,b)}dx, \quad x \ge 0, \;\; a,b > 0 \tag{6.8}$$

La trasformazione da adottare in questo caso è $X=g_1/g_2$.

Simulazione della variabile casuale di Fisher
La variabile casuale di Fisher ha funzione di densità

$$dF(x) = \frac{1}{B(a,b)}\sqrt{\frac{a}{b}x^a}\left[1 + \frac{a}{b\sqrt{x^{a+b}}}\right]dx, \quad x \ge 0, \;\; a,b \;\; \textit{interi} \;\; \textit{positivi} \tag{6.9}$$

La simulazione di questa variabile casuale può basarsi su due variabili casuali Gamma: g_l con densità *G(0.5,a/2)* e g_2 con densità *G(0.5,b/2)*. Applicando la trasformazione

$$X = \frac{g_1/b}{g_2/a} \tag{6.10}$$

si darà origine ad una variabile casuale di Fisher con *a* e *b* gradi di libertà rispettivamente al denominatore ed al numeratore.

Simulazione della variabile casuale di Student
La densità dio una variabile casuale di Student è

$$dF(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}} dx, \quad x \geq 0, \; k \; intero \; positiv \tag{6.11}$$

La simulazione di questa variabile casuale può avvenire in vari modi. Ad esempio utilizzando la relazione che la lega alla variabile casuale Beta con *k=2a* e *b=1/2*. Se Y è una tale variabile, la trasformazione

$$X = \sqrt{k\left(\frac{1}{Y}-1\right)} \tag{6.12}$$

produrrà una variabile casuale di Student con *k* gradi di libertà.

Simulazione della variabile casuale di Laplace
La variabile casuale di Laplace ha funzione di densità

$$dF(x) = \frac{1}{2b} e^{-\left\{\left|\frac{x-a}{b}\right|\right\}} dx, \quad b > 0 \tag{6.13}$$

La funzione di distribuzione di questa variabile casuale è invertibile solo a tratti e questo impedisce un'applicazione diretta della tecnica di inversione. Infatti:

$$F(x) = \frac{1}{2b}\int_{-\infty}^{x} e^{-\left\{\left|\frac{t-a}{b}\right|\right\}} dt = 0.5\begin{cases}\left[1-e^{-\left\{\frac{x-a}{b}\right\}}\right] & se \;\; x \geq a \\ \left[e^{-\left\{\frac{x-a}{b}\right\}}\right] & se \;\; < a\end{cases} \tag{6.14}$$

Poiché la variabile casuale è simmetrica, la sua simulazione può basarsi sulla generazione di due osservazioni indipendenti dalla uniforme: U_1 ed U_2 e sulla relazione

$$X = a + \text{sgn}(0.5 - U_1) Ln(2U_2 - 1) \tag{6.15}$$

dove la funzione *sgn(.)* assume valori *-1, 0,1* secondo il segno del suo argomento.

7. SimuIazione di distribuzioni miste

Le tecniche di simulazione discusse nei due precedenti paragrafi possono essere estese a distribuzioni risultanti come mistura di altre distribuzioni. Se ad esempio si ha

$$F(x) = pF_1(x) + (1-p)F_2(x) \tag{7.1}$$

la simulazione dalla F(x) potrebbe articolarsi come segue:

1. Simulare due osservazioni indipendenti dalla uniforme sull'intervallo unitario, diciamo U_l ed U_2.
2. Se $U_l \leq p$ allora si pone $X = F_1^{-1}(U_2)$ altrimenti si pone $X = F_2^{-1}(U_2)$

Questa procedura, nota in letteratura come tecnica di composizione, diviene utilissima se p è vicino all'unità ed $F_l(X)$ ha una inversa "facile" da gestire. Anche se la inversa della $F_2(X)$ risultasse particolarmente ostica la tecnica di composizione ridurrebbe di molto le difficoltà dato che il suo utilizzo si avrebbe solo l'(l-p)% delle volte. Marsaglia (1984) è andato piuttosto avanti in questo senso ed anche se le sue proposte non hanno ancora avuto completa espressione, ha aperto un sentiero che può portare molto in avanti le tecniche di simulazione.

Quanto detto ha soprattutto scopo pratico: agevolare la simulazione della $F(x)$ quando questa possa essere espressa come nella (7.1), ma ha anche una interpretazione diversa, più impegnativa in un certo senso. La descrizione della casualità di un fenomeno che abbia radici profonde nella realtà non sempre può reggersi su di una singola distribuzione. Un modello unico, anche molto flessibile (cioè con molti parametri, spesso di interpretazione dubbia) potrebbe lasciar fuori alcuni aspetti della realtà e forse non quelli meno significativi.

Una ipotesi alternativa è che la casualità possa essere più compiutamente descritta da una distribuzione mista il cui effetto possa essere ricondotto alla formula

$$F(x) = \sum_{i=1}^{k} F_i(x) \tag{7.2}$$

dove i $\{p_i\}$ sono dei pesi non negativi che assommano ad uno. L'algoritmo che simula la (7.2) potrebbe incentrarsi sui passi seguenti:

1. Si definiscono le $P_i = \sum_{j=0}^{i} p_j; \quad p_0 = 0$ quantità;
2. Si ottengono due osservazioni dalla uniforme cioè U_I ed U_2;
3. *Se* $U_1 \in [P_{i-1}, P_i[$ *allora si pone* $X = F_i^{-1}(U_2)$

Lo schema è molto semplice e la sua efficacia dipende dalle tecniche di simulazione delle singole $F_i(x)$ e dai pesi $\{p_i\}$ A scopo illustrativo si è applicata questa tecnica ad una combinazione della Weibull con una Pareto, in particolare si sono adoperate le distribuzioni:

$$W(0,1,2.5) = 1 - e^{-x^{2.5}}, \;\; P(0,1,2.5) = 1 - \left(\frac{1}{x}\right)^{2.5} \tag{7.3}$$

nelle proporzioni

$$F(x) = 0.25W(0,1,2.5) + 0.75P(0,1,2.5) \tag{7.4}$$

Nella tabella 5 sono stati inseriti, per varie ampiezze campionarie, i valori del test di Kolmogorov-Smirnov

$$KS = \sqrt{L}\ \underset{i=1,2,\ldots,L}{Massimo}\left\{\frac{i}{L} - F(x)\right\} \tag{7.5}$$

applicato ai valori medi che della KS ha riportato sulle L=1000 repliche di ogni singola situazione. In particolare l'ultima colonna è stata ottenuta adoperando la distribuzione esatta

$$F(KS) = 1 - e^{-2(KS)^2} \tag{7.6}$$

che sarebbe valida solo per L tendente ad infinito.

Tabella_5: test di Kolmogorov-Smirnov.

n	Valore medio della KS	Pr(X>KS)
15	0.22126	9.32%
30	0.11076	2.42%
60	0.16053	5.02%
100	0.16348	5.20%
200	0.04906	0.48%
500	0.12496	3.07%
1000	0.05465	0.59%
2000	0.02274	0.10%
5000	0.00201	0.00%
10000	0.03648	0.26%

I valori ottenuti nella sperimentazione sono piuttosto soddisfacenti. Ad esempio, per *n=60* la statistica KS è stata superiore a 0.16053 intorno al 5.02% delle volte. Questo significa che su 1000 campioni di ampiezza 60, solo per una cinquanta di essi, si è dovuta rifiutare l'ipotesi che esso fosse stato generato dalla distribuzione mista (7.4). In questo senso va detto che il test di Kolmogorov-Smirnov è "conservatore" cioè il livello di significatività reale è inferiore o uguale a quello nominale (Massey, 1950) per cui, tra i campioni rigettati in base alla KS, qualcuno era forse accettabile. All'aumentare dell'ampiezza campionaria migliora (sebbene non in maniera progressiva) l'accostamento tra distribuzione simulata e distribuzione teorica, fino ad arrivare alla virtuale identità per n=5000. I risultati contraddittori per *n=10000* sono da attribuire agli errori di arrotondamento che per questo ordine di grandezza dei campioni hanno maggiormente fatto sentire la loro influenza.

Bibliografia

Atkinson AC., Pierce M.C. (1976): The Computer Generation of Beta, Gamma, and Normal Random Variable. *Journal of the Royal Statistical Society, A,* 139,431-448.

Holland S. B. (1975): Some Results on the Discretization of Continuous Probability Distributions. *Technometrics,* 17, 333-339.

Knuth D. E. (1981): The Art of Computer Programming. Vol. 2: Seminumerical AIgorithms. 2nd edition, Addison-Wesley, Reading, Mass.

Lehmer D. W.(1949): Mathematical Methods in Large Scale Computing Units. Proceedings of a second symposium on Large-Scale Digital Calculation Machinery. 564-567. Harvard University Press, Cambridge, Mass.

Marsaglia G. (1984): The Exact-Approximation Method for Generating Random Variables in a Computer. *Journal of the American Statistical Association,.* 79, 218-221.

Marsaglia G., Bray T. A. (1964): A Convenient Method for Generating Random Variable. *SIAM Review*, 6, 260-264.

Massey F .J. (1957): The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46, 68-77.

Mood A M. ,Graybill F. A., Boes D. C. (1974): lntroduction to the Theory of Statistics. McGraw-Hill. Kogakush Ltd, Tokyo.

"Student" (W.S. Gosset): (1908a): The Probable Error of a Mean. *Biometrika*. 6,1-25.

"Student" (W.S. Gosset): (1908b): Probable Error of a Correlation. *Biometrika*. 6,302-310.

Teichroew D; (1965): A History of Distribution Sampling Prior to the Era of the Computer and Its Relevance to Simulation. *Journal of the American Statistical Association.* 60,27-29.

Wichmann B. A, Hill I. D. (1982): Algorith AS183: an Efficient and Portable Pseudo-random Number Generator. *Applied Statistics*., 31, 188-190.

Wichmann B. A., Hill T. D. (1984): Algorithm AS183: Correction. *Applied Statistics,* 33,123.