Somma di due variabili casuali indipendenti

Siano X_1 e X_2 due variabili casuali indipendenti con valore atteso μ_1 e μ_2 e con varianza σ_1^2 e σ_2^2 rispettivamente. Sia $X = X_1 + X_2$ la somma delle due variabili casualicon valore atteso $\mu_1 + \mu_2$ e con varianza $\sigma_1^2 + \sigma_2^2$. Ci proponiamo di trovare (o, quanto meno, disegnare) la funzione di densità di probabilità di X. Per fare questo esiste, tra gli altri, un metodo generale detto *convoluzione*.[1] Seguendo quanto scrivono Devore, Berk e Carlton,[2] la funzione di densità di probabilità della variabile casuale $X = X_1 + X_2$ è data da

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X_1}(x_1) f_{X_2}(x - x_1) dx_1, \tag{1}$$

dove $f_{X_1}(x_1)$ e $f_{X_2}(x_2)$ sono, rispettivamente la funzione di densità di probabilità *marginale* di X_1 e X_2. I "veri" limiti di integrazione vanno determinati in modo che entrambi siano coerenti con l'intervallo di definizione sia di X_1, sia di X_2 (come vedremo negli esempi che seguono).

Somma di due variabili casuali esponenziali i.i.d.

Siano X_1 e X_2 due variabili casuali esponenziali indipendenti e identicamente distribuite (i.i.d.) con parametro λ e sia $X = X_1 + X_2$ la variabile casuale "somma" di X_1 e X_2. È davvero facile far vedere che X non può essere a usa volta una esponenziale. Infatti, il valore atteso di X è $\frac{2}{\lambda}$ e la varianza di X è $\frac{2}{\lambda^2}$, che non è il quadrato del valore atteso.

Qual è allora la funzione di densità di probabilità della variabile casuale X? Possiamo trovarla impiegando la convoluzione.

Partendo da $f_{X_1}(x_1) = \lambda e^{-\lambda x_1}$ e $f_{X_2}(x_2) = \lambda e^{-\lambda x_2}$, e ponendo $x_2 = x - x_1$, la (1) diventa

$$f_X(x) = \int_0^x \lambda e^{-\lambda x_1} \lambda e^{-\lambda(x - x_1)} dx_1.$$

[1] La convoluzione è un'operazione matematica che consiste nell'integrare il prodotto tra una funzione e una seconda funzione traslata di un certo valore. Trova applicazione in numerosi campi, quali l'elaborazione digitale delle immagini o dei segnali. Qui vedremo soltanto la sua applicazione in ambito probabilistico, per trovare la funzione di densità di probabilità della somma di due variabili casuali indipendenti.

[2] Devore J.L., Berk K.N., Carlton M.A. (2021). *Modern Mathematical Statistics with Applications. Third Edition.* Springer, pag. 307.

Gli estremi dell'intervallo di integrazione sono stati modificati rispetto a quelli riportati nella (1). L'estremo inferiore di integrazione è diventato 0 e questo è ovvio (dal momento che una variabile casuale esponenziale non può assumere valori negativi). L'estremo superiore è diventato x perché la condizione $x_2 \geq 0$ equivale a $x - x_1 \geq 0$ e, quindi, $x_1 \leq x$. Pertanto, integrando in dx_1, bisogna garantire di non andare oltre a x.

Semplificando si ottiene

$$f_X(x) = \int_0^x \lambda^2 e^{-\lambda x} dx_1$$

e, dal momento che $e^{-\lambda x}$ può essere considerato "costante", la funzione di densità di probabilità di X è

$$f_X(x) = \lambda^2 e^{-\lambda x} \int_0^x dx_1 = \lambda^2 e^{-\lambda x} x.$$

Questa funzione ha un massimo quando $x = \frac{1}{\lambda}$ e il massimo vale $\frac{\lambda}{e}$.

Vediamo con R un caso particolare ($\lambda = 5$) impiegando una simulazione che sfrutta la funzione `rexp`.

```
set.seed(123456)
n <- 100000
l <- 5
x <- rexp(2*n,l)
X <- matrix(x, ncol=2)
y <- rowSums(X)
hist(y, breaks = 50, xlim = c(0,1.5), prob=TRUE)
box()
curve(l^2*exp(-l*x)*x, add=TRUE)
```

La figura 1 riporta nel pannello di sinistra i risultati della simulazione sotto forma di istogramma, sul quale è sovraimposta la funzione di densità di probabilità teorica che si adatta molto bene ai dati osservati. Anche il *Q-Q plot* conferma questa impressione (figura 1, pannello di destra).

```
> pfun <- function(u) 25*(1/25-((5*u+1)*exp(-5*u))/25)
> x <- seq(0.1,2,0.1)
> p <- pfun(x)
> q <- quantile(y,p)
> plot(q,x)
```

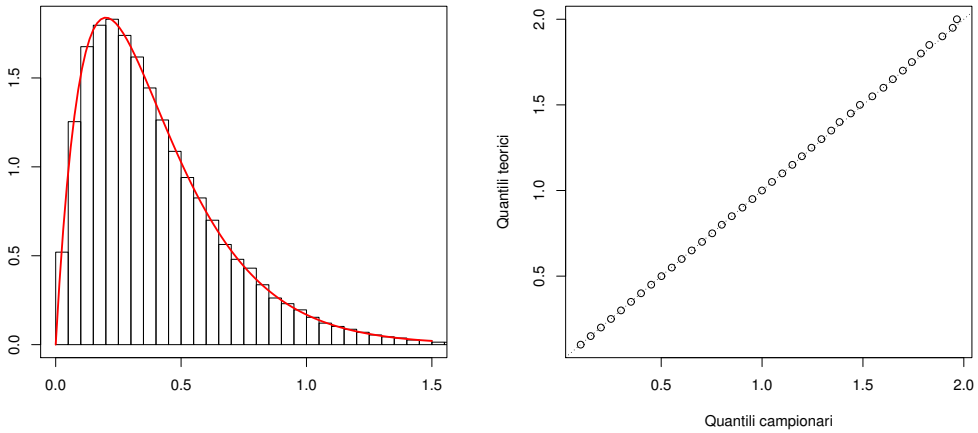


Figure 1: A sinistra: Distribuzione di probabilità della somma di due variabili casuali esponenziali i.i.d. ($\lambda = 5$). A destra: *Q-Q plot* per valutare l'adattamento dei dati simulati alla somma di due variabili casuali esponenziali i.i.d. ($\lambda = 5$).

Si potrebbe addirittura chiedere a R di eseguire direttamente l'operazione di convoluzione, impiegando la funzione `integrate`, nel modo seguente

```
> f1 <- function(x) 5*exp(-5*x)
> f  <- function(z)
+ integrate(function(x,z) f1(x) * f1(z-x),0,z,z)$value
> f  <- Vectorize(f)
```

Un confronto fra i valori ottenuti eseguendo la convoluzione "per via analitica" e quelli ottenuti numericamente impiegando `integrate` mostra che sono del tutto equivalenti:

```
> z <- seq(0,2,0.001)
> y <- f(z)
> yy <- l^2*exp(-l*z)*z
> max(abs(y-yy))
[1] 4.440892e-16
```

La somma di due variabili casuali esponenziali i.i.d. (con parametro λ) è un caso particolare della variabile casuale *Gamma* con parametri λ e 2. Più

in generale, la somma di k variabili casuali esponenziali i.i.d. (con parametro λ) è una variabile casuale *Gamma* con parametri λ e k.

Somma di due variabili casuali esponenziali indipendenti

Siano X_1 e X_2 due variabili casuali esponenziali indipendenti, con parametri λ_1 e λ_2 rispettivamente. Vogliamo trovare la funzione di densità di probabilità della variabile casuale $X = X_1 + X_2$ impiegando la convoluzione.

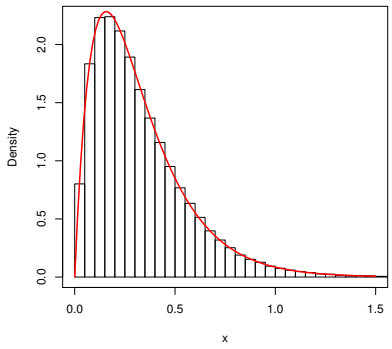
In questo caso abbiamo $f_{X_1}(x_1) = \lambda e^{-\lambda_1 x_1}$ e $f_{X_2}(x_2) = \lambda e^{-\lambda_2 x_2}$. Quindi la (1) diventa

$$f_X(x) = \int_0^x \lambda_1 e^{-\lambda_1 x_1} \lambda_2 e^{-\lambda_2 (x - x_1)} dx_1.$$

Semplificando, dopo una serie di passaggi, otteniamo la funzione di densità di probabilità di X:

$$f_X(x) = \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 x} - e^{-\lambda_2 x} \right).$$

Vediamo con R un caso particolare ($\lambda_1 = 5, \lambda_2 = 8$) impiegando una simulazione, i cui risultati sono presentati nella figura riportata qui di seguito.




```
> set.seed(654321)
> n <- 100000
> l1 <- 5; l2 <- 8
> x1 <- rexp(n,l1); x2 <- rexp(n,l2)
> hist(x1+x2, breaks = 50,
+ xlim=c(0,1.5), prob=TRUE)
> box()
> k <- l1*l2/(l2-l1)
> curve(k*(exp(-l1*x)-
+ exp(-l2*x)), add=TRUE,col="red")
```

Anche in questo caso la curva sovraimposta si adatta molto bene ai dati osservati.

Analogamente a quanto fatto in precedenza, chiediamo a R di eseguire direttamente la convoluzione:

```
> f1 <- function(x) 5*exp(-5*x)
> f2 <- function(x) 8*exp(-8*x)
```

```
> f  <- function(z)
+ integrate(function(x,z) f1(x) * f2(z-x),0,z,z)$value
> f  <- Vectorize(f)
```

Un confronto fra i valori ottenuti eseguendo la convoluzione "per via analitica" e quelli ottenuti impiegando `integrate` mostra che sono del tutto equivalenti:

```
> z <- seq(0,2,0.01)
> y <- f(z)
> k <- l1*l2/(l2-l1)
> yy <- k*(exp(-l1*z)-exp(-l2*z))
> max(abs(y-yy))
[1] 8.881784e-16
```

Somma di due variabili casuali uniformi i.i.d.

Siano X_1 e X_2 due variabili casuali uniformi i.i.d. fra a e b e sia $X = X_1 + X_2$ la variabile casuale "somma" di X_1 e X_2. Sappiamo che in questo caso la funzione risultante è una variabile casuale triangolare distribuita fra $2a$ e $2b$ con il massimo in $(a + b)$. La funzione di densità di probabilità di questa variabile casuale si può trovare facilmente per via "geometrica". Infatti il "triangolo" ha una base la cui lunghezza è $2(b - a)$, mentre l'altezza deve essere necessariamente $\frac{1}{b-a}$ in modo da garantire che l'area sia uguale a 1. Essendo la distribuzione simmetrica, la media, la moda e la mediana sono tutte uguali a $(a + b)$ e, quindi, la funzione di densità di probabilità è data da due segmenti di retta di pendenza uguali (in valore assoluto) a $\frac{1}{(b-a)^2}$, ma di segno opposto: positivo quando $2a \leq x \leq a + b$) e negativo quando $a + b \leq x \leq 2b$).

$$f(x) = \begin{cases} 0, & x < 2a, \\ \dfrac{x-2a}{(b-a)^2}, & 2a \leq x < (a+b), \\ \dfrac{1}{b-a}, & x = (a+b), \\ \dfrac{2b-x}{(b-a)^2}, & (a+b) < x \leq 2b, \\ 0, & x > 2b. \end{cases}$$

Questo risultato è immediato se si conosce già la forma della distribuzione risultante (triangolare, nel nostro caso). La forma deve però emergere anche se si esegue la convoluzione delle due funzioni di densità di probabilità $f_1(x_1)$ e $f_2(x_2)$ prestando però attenzione a definire in modo appropriato gli estremi dell'intervallo di integrazione. Infatti, la variabile casuale risultante è definita soltanto quando $2a \leq x \leq 2b$ (ovvero, all'esterno di questo intervallo la funzione di densità di probabilità della somma sarà sempre uguale a zero). Inoltre, essendo X_1 e X_2 identicamente distribuite (oltre che indipendenti) ciascuna di esse o varrà $\frac{1}{b-a}$ o varrà zero. Questo vuol dire che il prodotto $f_1(x_1)f_2(x_2)$ (prodotto che entra nell'integrale di convoluzione) o sarà $\frac{1}{(b-a)^2}$ o sarà zero.

Dobbiamo a questo proposito considerare due casi, a seconda che x sia minore o maggiore di $a+b$ (il valore centrale dell'intervallo di definizione di X); infatti, la condizione $a \leq x_2 \leq b$ equivale alla condizione $a \leq x - x_1 \leq b$ e, quindi, per un x fissato, x_1 deve essere compreso fra $x-b$ e $x-a$. L'integrale "generale" di convoluzione (1) si spezza quindi in due parti. La prima vale quando $2a \leq x \leq (a+b)$:

$$f_X(x) = \int_a^{x-a} \frac{1}{(b-a)^2} dx_1 = \frac{x-2a}{(b-a)^2}.$$

La seconda parte vale quando $(a+b) \leq x \leq 2b$:

$$f_X(x) = \int_{x-b}^{b} \frac{1}{(b-a)^2} dx_1 = \frac{2b-x}{(b-a)^2}.$$

E questo ci (ri)porta alla funzione di densità di probabilità che avevamo trovato per via "geometrica".

Somma di due variabili casuali uniformi indipendenti

Siano X_1 e X_2 due variabili casuali indipendenti distribuite in modo uniforme fra a_1 e b_1 e fra a_2 e b_2 rispettivamente. Sia poi $X = X_1 + X_2$ la somma di X_1 e X_2. Questo caso più generale è più complesso del precedente, dal momento che la distribuzione di X non è più necessariamente triangolare. Dobbiamo infatti distinguere due casi principali:

- $a_1 + b_2 = a_2 + b_1$: in questo caso la variabile casuale X segue una distribuzione triangolare;
- $a_1 + b_2 \neq a_2 + b_1$: in questo caso la variabile casuale X non segue una distribuzione triangolare, bensì una distribuzione *trapezoidale*.

Iniziamo prendendo in esame il primo caso. La condizione $a_1+b_2 = a_2+b_1$ implica che $b_1-a_1 = b_2-a_2$. Cosa questo comporti lo possiamo vedere molto bene con R considerando il caso di due variabili casuali uniformi discrete:

```
> x1 <- c(10:18)
> x2 <- c(32:40)
> x <- outer(x1,x2,"+")
> x
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,]   42   43   44   45   46   47   48   49   50
[2,]   43   44   45   46   47   48   49   50   51
[3,]   44   45   46   47   48   49   50   51   52
[4,]   45   46   47   48   49   50   51   52   53
[5,]   46   47   48   49   50   51   52   53   54
[6,]   47   48   49   50   51   52   53   54   55
[7,]   48   49   50   51   52   53   54   55   56
[8,]   49   50   51   52   53   54   55   56   57
[9,]   50   51   52   53   54   55   56   57   58
```

Ci sono 9 risultati possibili per X_1 e 9 risultati possibili per X_2. In questo modo X può assumere tutti i valori interi da 42 a 58, crescendo "regolarmente" da 42 fino a 50 (dove la distribuzione ha un massimo) e decrescendo in modo speculare fino ad arrivare a 58 (vedi il pannello di sinistra della figura 2).

Il valore più frequente è, appunto, 50, che si può ottenere in 9 possibili modi diversi, disposti lungo una diagonale del "quadrato" individuato dai 9 possibili valori di X_1 e dai 9 possibili valori di X_2.

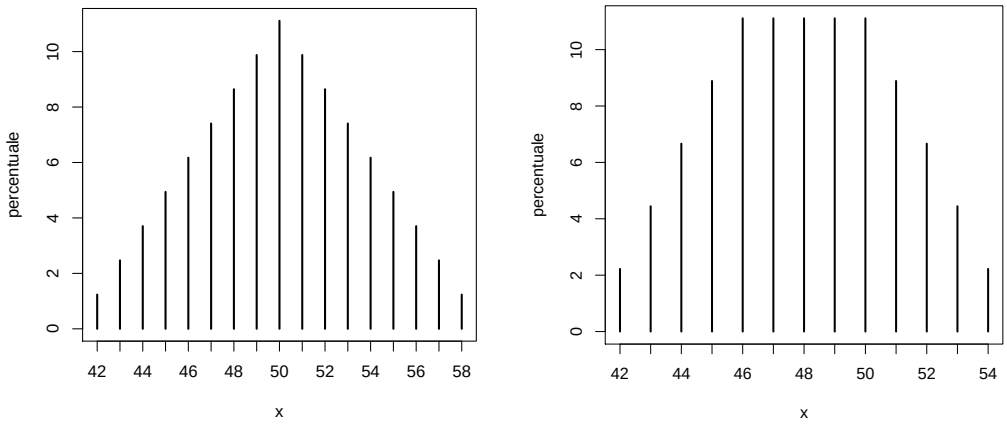


Figure 2: Distribuzione di probabilità della somma X di due variabili casuali uniformi discrete X_1, X_2 indipendenti. Pannello di sinistra: $a_1 = 10, b_1 = 18$ e $a_2 = 32, b_2 = 40$. La distribuzione risultante è *triangolare* (discreta), dal momento che $a_1 + b_2 = a_2 + b_1 = 50$. Pannello di destra: $a_1 = 32, b_1 = 40$ e $a_2 = 10, b_2 = 14$. La distribuzione risultante è *trapezoidale* (discreta), dal momento che $a_1 + b_2 = 46 \neq a_2 + b_1 = 50$.

```
> table(x)
x
42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58
1  2  3  4  5  6  7  8  9  8  7  6  5  4  3  2  1
```

Una cosa analoga accade, seppure con una "continuità" di valori, quando X_1 e X_2 sono due variabili casuali uniformi continue. La distribuzione risultante della somma X è quindi triangolare; il valore più piccolo possibile è a_1+a_2, il più grande è b_1+b_2 e il valore intermedio è $\frac{a_1+a_2+b_1+b_2}{2}$; dal momento che $a_1 + b_2 = a_2 + b_1$, questo valore (che rappresenta la media, la moda e la mediana di X) può anche essere scritto come $a_1 + b_2$ (o, indifferentemente, come $a_2 + b_1$).

Dobbiamo a questo punto considerare due casi, a seconda che x sia minore o maggiore di $a_1 + b_2$. La condizione $a_2 \leq x_2 \leq b_2$ equivale alla condizione $a_2 \leq x - x_1 \leq b_2$ e, quindi, per un x fissato, x_1 deve essere compreso fra $x - b_2$ e $x - a_2$. L'integrale "generale" di convoluzione (1) si spezza anche in

questo caso in due parti. La prima vale quando $a_1 + a_2 \leq x \leq a_1 + b_2$:

$$f_X(x) = \int_{a_1}^{x-a_2} \frac{1}{(b_1 - a_1)(b_2 - a_2)} dx_1 = \frac{x - (a_1 + a_2)}{(b_1 - a_1)(b_2 - a_2)}.$$

La seconda parte vale quando $a_1 + b_2 \leq x \leq b_1 + b_2$:

$$f_X(x) = \int_{x-b_2}^{b_1} \frac{1}{(b_1 - a_1)(b_2 - a_2)} dx_1 = \frac{(b_1 + b_2) - x}{(b_1 - a_1)(b_2 - a_2)}.$$

Lasciamo al lettore la verifica di questo risultato impiegando la via "geometrica".

Il caso più generale si ha quando $a_1 + b_2 \neq a_2 + b_1$. Dovremmo distinguere il caso $a_1 + b_2 < a_2 + b_1$ da quello $a_1 + b_2 > a_2 + b_1$; tuttavia, senza perdere di generalità, prenderemo in esame solo il primo.

Vediamo cosa comporta la condizione $a_1 + b_2 < a_2 + b_1$ prendendo in esame il caso discreto .

```
> x1 <- c(32:40)
> x2 <- c(10:14)
> x <- outer(x2,x1,"+")
> x
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,]   42   43   44   45   46   47   48   49   50
[2,]   43   44   45   46   47   48   49   50   51
[3,]   44   45   46   47   48   49   50   51   52
[4,]   45   46   47   48   49   50   51   52   53
[5,]   46   47   48   49   50   51   52   53   54
```

Questa volta ci sono 9 risultati possibili per X_1 e 5 risultati possibili per X_2. In questo modo X può assumere tutti i valori interi da 42 a 54. Osserviamo una prima "fase" di crescita "regolare" da 42 fino a 46, valore che si può realizzare in 5 modi diversi. A questo valore corrisponde il massimo della distribuzione, ma questo massimo non è unico. Anche 47, 48, 49 e 50 si possono realizzare in 5 modi diversi. Infine osserviamo una "fase speculare" di decrescita che porta al valore massimo possibile per la somma (vale a dire 54).

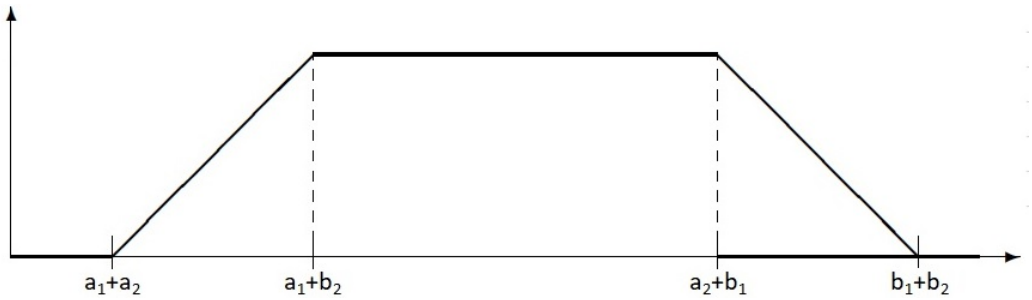


Figure 3: Grafico della funzione di densità di probabilità della somma X di due variabili casuali uniformi continue X_1, X_2 indipendenti, distribuite, rispettivamente, tra a_1 e b_1 e tra a_2 e b_2 ($a_1 + b_2 < a_2 + b_1$).

```
> table(x)
x
 42 43 44 45 46 47 48 49 50 51 52 53 54
  1  2  3  4  5  5  5  5  5  4  3  2  1
```

Questo accade perché il "quadrato" osservato nel caso precedente (quando $a_1 + b_2 = a_2 + b_1$) è stato sostituito da un "rettangolo". La distribuzione risultante ha quindi una forma trapezoidale (discreta), con una probabilità uniforme fra $x = 46$ e $x = 50$, vale a dire fra $a_1 + b_2$ e a $a_2 + b_1$, come si può vedere nel pannello di destra della figura 2.

Lo stesso accade, ma con una "continuità" di valori, quando X_1 e X_2 sono due variabili casuali uniformi continue. La distribuzione risultante della somma X ha la forma di un trapezio isoscele. Il valore più piccolo possibile è $a_1 + a_2$, il più grande è $b_1 + b_2$. Quando $a_1 + b_2 \leq x \leq a_2 + b_1$ la densità di probabilità è uniforme (figura 3).

Il trapezio è isoscele, perché $(a_1 + b_2) - (a_1 + a_2) = (b_1 + b_2) - (a_2 + b_1) = (b_2 - a_2)$. Il "rettangolo centrale" ha quindi lo stesso intervallo di definizione di X_2. Tuttavia l'altezza di questo "rettangolo" non può per ovvi motivi essere $f_{X_2}(x_2)$ (cioè $\frac{1}{b_2 - a_2}$). Vedremo che l'altezza è invece $f_{X_1}(x_1)$ (cioè $\frac{1}{b_1 - a_1}$).

Sono quindi tre i casi da prendere in esame:

- $a_1 + a_2 \leq x \leq a_1 + b_2$;
- $a_1 + b_2 \leq x \leq a_2 + b_1$;
- $a_2 + b_1 \leq x \leq b_1 + b_2$.

L'integrale "generale" di convoluzione (1) si spezza in tre parti. La prima vale quando $a_1 + a_2 \leq x \leq a_1 + b_2$:

$$f_X(x) = \int_{a_1}^{x-a_2} \frac{1}{(b_1 - a_1)(b_2 - a_2)} dx_1 = \frac{x - (a_1 + a_2)}{(b_1 - a_1)(b_2 - a_2)}.$$

La seconda parte vale quando $a_1 + b_2 \leq x \leq a_2 + b_1$:

$$f_X(x) = \int_{x-b_2}^{x-a_2} \frac{1}{(b_1 - a_1)(b_2 - a_2)} dx_1 = \frac{b_2 - a2}{(b_1 - a_1)(b_2 - a_2)} = \frac{1}{b_1 - a_1}.$$

La terza parte vale quando $a_2 + b_1 \leq x \leq b_1 + b_2$:

$$f_X(x) = \int_{x-b_2}^{b_1} \frac{1}{(b_1 - a_1)(b_2 - a_2)} dx_1 = \frac{(b_1 + b_2) - x}{(b_1 - a_1)(b_2 - a_2)}.$$

Somma di due variabili casuali "qualsiasi" indipendenti

Se X_1 e X_2 sono due variabili casuali indipendenti, in generale non si riesce ad ottenere la funzione di densità di probabilità della loro somma X in forma "chiusa" usando la convoluzione, ma è necessario ricorrere a metodi numerici. Vediamo come sia possibile arrivare a questo risultato impiegando R in alcuni esempi particolari.

0.0.1 Somma di di due variabili casuali triangolari indipendenti

Per il primo esempio di questa sezione sarà necessario impiegare il *package* `triangle` che "provide information about the triangle distribution on the interval from a to b with a maximum at c" (a, b e c rappresentano i parametri della distribuzione triangolare). In particolare impiegheremo la funzione `rtriangle` per simulare dati estratti dalle due distribuzioni triangolari rappresentate nella figura 4. Entrambe le distribuzioni si estendono da 20 a 30 e hanno un massimo, rispettivamente, in 20 (per la variabile casuale X_1) e in 30 (per la variabile casuale X_2). Ricordando la formula per calcolare l'area di un triangolo e che l'area deve essere uguale a 1, è semplice trovare la funzione di densità di probabilità di ciascuna delle due variabili casuali. Abbiamo quindi che $f_{X_1}(x_1) = 0.6 - 0.2x_1$ e che $f_{X_2}(x_2) = -0.4 + 0.2x_2$. Naturalmente al di fuori dell'intervallo $[20, 30]$ le densità sono sempre nulle.

Estraiamo 100000 osservazioni da ciascuna delle due distribuzioni e facciamone la somma.

Figure 4: Grafico della funzione di densità di probabilità di due variabili casuali triangolari X_1 e X_2, entrambe distribuite fra 20 e 30. X_1 ha il massimo in $x_1 = 30$. X_2 ha il massimo in $x_2 = 20$. Per entrambe il massimo vale $1/5$.

```
> library(triangle)
> set.seed(123456)
> n <- 100000
> x1 <- rtriangle(n,20,30,20) # massimo su 20
> x2 <- rtriangle(n,20,30,30) # massimo su 30
> x <- x1 + x2
```

L'istogramma relativo alla somma delle due triangolari è riportato nella figura a pagina 16. La forma, pur essendo simmetrica, non è a forma di campana e non è nemmeno triangolare. Si osserva infatti una "curvatura", con la concavità che sembra rivolta sempre verso l'esterno.

Vogliamo trovare, impiegando R, la convoluzione delle due funzioni per disegnare la funzione di densità di probabilità della variabile casuale $X = X_1 + X_2$. Naturalmente la formula generale rimane quella riportata all'inizio, ma vedremo subito che occorre prestare molta attenzione nel definire gli estremi di integrazione. Abbiamo quindi

$$f_X(x) = \int_{-\infty}^{+\infty} (0.6 - 0.2x_1)(-0.4 + 0.02(50 - x_1))dx_1$$

e vogliamo calcolare, ad esempio, $f_X(50)$ per via analitica (senza impiegare R per il momento).

È del tutto evidente che, se $x_1 \in [20, 30]$ e $x_2 \in [20, 30]$, sarà, necessariamente $x \in [40, 60]$. Il valore che stiamo considerando (50) è quello centrale. Se $x = 50$, x_1 può assumere qualsiasi valore nell'intervallo di definizione e,

contestualmente, x_2 assumerà il valore $50 - x_1$, che sarà sempre all'interno dell'intervallo di definizione. Avremo quindi:

$$f_X(50) = \int_{20}^{30} (0.6 - 0.2x_1)(-0.4 + 0.02(50 - x_1))dx_1 = \frac{2}{15}.$$

Cosa accade, però, se vogliamo calcolare, ad esempio, $f_X(45)$? In questo caso, x_1 può assumere qualsiasi valore nell'intervallo $[20, 25]$ e, contestualmente, x_2 assumerà il valore $45 - x_1$ (quindi, da 25 a 20). Se, però, assegnassimo a x_1 il valore 26, x_2 dovrebbe valere 19, che è al di fuori dell'intervallo di definizione di x_2 (lo stesso varrebbe per qualsiasi valore maggiore di 25). Ecco quindi che, per calcolare $f_X(45)$, gli estremi di integrazione non saranno più gli stessi del caso precedente o, più precisamente, l'estremo superiore non potrà più essere 30 ma 25:

$$f_X(45) = \int_{20}^{25} (0.6 - 0.2x_1)(-0.4 + 0.02(45 - x_1))dx_1 = \frac{1}{24}.$$

Prendiamo in esame il caso $f_X(55)$, che può essere considerato "speculare" rispetto al precedente. Non possiamo certo assegnare a x_1 il valore 20, perché allora sarebbe $x_2 = 35$, al di fuori del suo campo di definizione. Il più piccolo valore che possiamo assegnare a x_1 è 25, perché in questo caso $x_2 = 30$; il più grande è $x_1 = 30$, perché il valore corrispondente di x_2 è 25 (all'interno del suo campo di definizione). Ancora una volta, quindi, dovremo modificare gli estremi di integrazione, che andranno da 25 a 30 (il minimo e il massimo ammissibili per x_1 quando $x = 55$):

$$f_X(55) = \int_{25}^{30} (0.6 - 0.2x_1)(-0.4 + 0.02(55 - x_1))dx_1 = \frac{1}{24}$$

(naturalmente $f_X(55) = f_X(45)$ perché la distribuzione di X è simmetrica rispetto al valore centrale 50).

Potremmo ripetere tutte queste considerazioni, calcolando per via analitica $f_X(40), f_X(41), f_X(42), \ldots, f_X(58), f_X(59), f_X(60)$ e disegnando "a mano" la funzione di densità di probabilità risultante. Se vogliamo che sia R a farlo per noi, dobbiamo "spiegare" in modo molto preciso quello che vogliamo che faccia.

Il primo passo, il più semplice, è definire le due funzioni di "partenza", vale a dire $f_{X_1}(x_1)$ e $f_{X_2}(x_2)$:

```
> f.x1 <- function(x)  0.6 - 0.02*x
> f.x2 <- function(x) -0.4 + 0.02*x
```

Il secondo passo è quello più "delicato": dobbiamo definire la funzione[3] che impiega `integrate` (cioè la funzione che calcola, numericamente, il valore della funzione di densità di probabilità della somma). Come abbiamo già visto in precedenza l'integrale "generale" si dovrà spezzare in due parti, a seconda che $x < 50$ o che $x > 50$ (il caso $x = 50$ può ricadere indifferentemente in uno dei due casi precedenti). Dobbiamo allora trovare un modo generale per definire gli estremi di integrazione nei due casi.

Quando $x < 50$ abbiamo visto che x_1 può assumere il valore minimo del suo campo di definizione (20), ma non può assumere il valore massimo (30). Se rivediamo quanto scritto in precedenza, vediamo che, per un dato valore della somma x, $x - x_1$ (vale a dire x_2) potrà essere al massimo $x - 20$ (quando $x = 45$, x_1 poteva andare da 20 a 25, mentre x_2 poteva andare da 25 a 20). Quindi, quando $x < 50$, il primo integrale sarà:

$$f_X(x) = \int_{20}^{x-20} (0.6 - 0.2x_1)(-0.4 + 0.02(x - x_1))dx_1.$$

Quando $x > 50$, x_1 non può più assumere il valore minimo del suo campo di definizione (cioè 20), mentre può assumere il valore massimo; per un dato valore della somma x, x_1 non può essere inferiore a $x - 30$. Quindi, quando $x > 50$, il secondo integrale sarà:

$$f_X(x) = \int_{x-30}^{30} (0.6 - 0.2x_1)(-0.4 + 0.02(x - x_1))dx_1$$

(osserviamo che, quando $x = 50$, gli estremi di integrazione di entrambi gli integrali considerati vanno da 20 a 30, come deve essere).

Siamo ora pronti per "spiegare" a R come calcolare numericamente i due integrali:

```
> f.xa  <- function(z) integrate(function(x,z)
+ f.x1(x)*f.x2(z-x),20,(z-20),z)$value
> f.xa  <- Vectorize(f.xa)
> f.xb  <- function(z) integrate(function(x,z)
+ f.x1(x)*f.x2(z-x),(z-30),30,z)$value
> f.xb  <- Vectorize(f.xb)
```

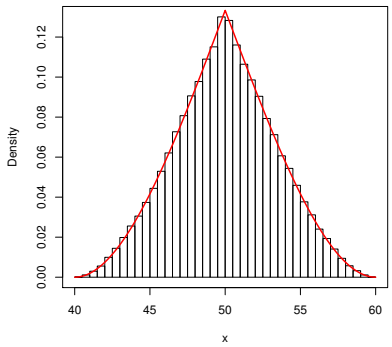
[3] In realtà, come vedremo, saranno due.

(naturalmente la responsabilità di chiamare `f.xa` o `f.xb` a seconda del valore della somma x sarà nostra e soltanto nostra).

Per verificare di aver definito tutto in modo corretto, facciamo calcolare a R i valori della funzione di densità di probabilità della somma per alcuni valori di x, controllando che coincidano con quelli calcolati analiticamente:

```
> xa <- seq(40,50,1); fa <- f.xa(xa)
> xb <- seq(51,60,1); fb <- f.xb(xb)
> data.frame("x"=c(xa,xb),"f(x)"=round(c(fa,fb),3))
    x  f(x)
1  40 0.000
2  41 0.002
3  42 0.007
4  43 0.016
5  44 0.028
6  45 0.042
7  46 0.058
8  47 0.075
9  48 0.094
10 49 0.113
11 50 0.133
12 51 0.113
13 52 0.094
14 53 0.075
15 54 0.058
16 55 0.042
17 56 0.028
18 57 0.016
19 58 0.007
20 59 0.002
21 60 0.000
```

Siamo adesso pronti per disegnare l'istogramma dei valori ottenuti mediante la simulazione e sovrapporre ad esso la funzione di densità di probabilità della somma X calcolata numericamente. Come si può vedere dalla figura riportata di seguito, l'adattamento è ottimo.




```
> hist(x, breaks=50, prob=TRUE,
+ main="", xlab="x"); box()
> xa <- seq(40,50,0.1)
> fa <- f.xa(xa)
> xb <- seq(50.1,60,0.1)
> fb <- f.xb(xb)
> xx <- c(xa,xb)
> f <- c(fa,fb)
> lines(xx,f,lwd=2,col="red")
```

Possiamo anche calcolare agevolmente la funzione di densità di probabilità della somma X ricorrendo ad uno "stratagemma". Infatti, la funzione da integrare (nell'integrale di convoluzione) è una parabola. Pertanto qualsiasi primitiva è una cubica e la funzione di densità di probabilità della somma X è rappresentata dai rami di due cubiche che si incrociano in $x = 50$. Possiamo calcolare i coefficienti delle due cubiche a partire dai valori analitici della $f_X(x)$ (ricordando che questi valori sono "simmetrici" rispetto a 50). Sappiamo già che $f_X(40) = 0$, $f_X(50) = 2/15$, $f_X(60) = 0$; abbiamo poi già calcolato $f_X(45) = f_X(55) = 1/24$. Basterebbe a questo punto calcolare il valore di $f_X(x)$ in corrispondenza di un solo altro valore di x per poter trovare i coefficienti. Noi abbiamo calcolato analiticamente tutti i valori di $f_X(x)$ per $x = 40, 41, 42, \ldots, 60$ e possiamo trovare i coefficienti impiegando la regressione multipla (e il metodo dei minimi quadrati).

```
> x <- c(40,41,42,43,44,45,46,47,48,49,50)
> y <- c(0,29/15000,14/1875,81/5000,52/1875,1/24,
+ 36/625,1127/15000,176/1875,567/5000,2/15)
> x2 <- x^2
> x3 <- x^3
> fit.a <- lm(y ~ x + x2 + x3)
> sum(fit.a$residuals^2)
[1] 3.644766e-30
> #
> x <- c(60:50)
> x2 <- x^2
> x3 <- x^3
> fit.b <- lm(y ~ x + x2 + x3)
> sum(fit.b$residuals^2)
```

```
[1] 1.429817e-30
> cbind(fit.a$coef,fit.b$coef)
[,1]          [,2]
(Intercept)  7.466667e+00 -7.200000e+00
x           -4.800000e-01  4.800000e-01
x2           1.000000e-02 -1.000000e-02
x3          -6.666667e-05  6.666667e-05
```

Osserviamo che i residui di interpolazione sono tutti nulli, a riprova del fatto che, in questo caso, la cubica è il vero modello. Le funzioni di densità di probabilità di X_1 e X_2 sono, pertanto:

$$f_{X_1}(x_1) = \frac{112}{15} - \frac{12}{25}x + \frac{1}{100}x^2 - \frac{1}{15000}x^3$$
$$f_{X_2}(x_2) = -\frac{36}{5} + \frac{12}{25}x - \frac{1}{100}x^2 + \frac{1}{15000}x^3$$

Lasciamo al lettore verificare che il seguente *script* riproduce il grafico della funzione di densità di probabilità di X.

```
b0 <- 112/15
b1 <- -12/25
b2 <- 1/100
b3 <- -1/15000
curve(b0+b1*x+b2*x^2+b3*x^3,40,60,ylim=c(0,2/15))
#
b0 <- -36/5
b1 <- 12/25
b2 <- -1/100
b3 <- 1/15000
curve(b0+b1*x+b2*x^2+b3*x^3,add=TRUE,col="red")
```

Lasciamo infine al lettore come esercizio disegnare, con R:

- la funzione di densità di probabilità della somma di due variabili casuali triangolari indipendenti, entrambe definite fra 20 e 30 e con massimo in 20;
- la funzione di densità di probabilità della somma di due variabili casuali triangolari indipendenti, entrambe definite fra 20 e 30 e con massimo in 30.

Il risultato è riportato nella figura 5.

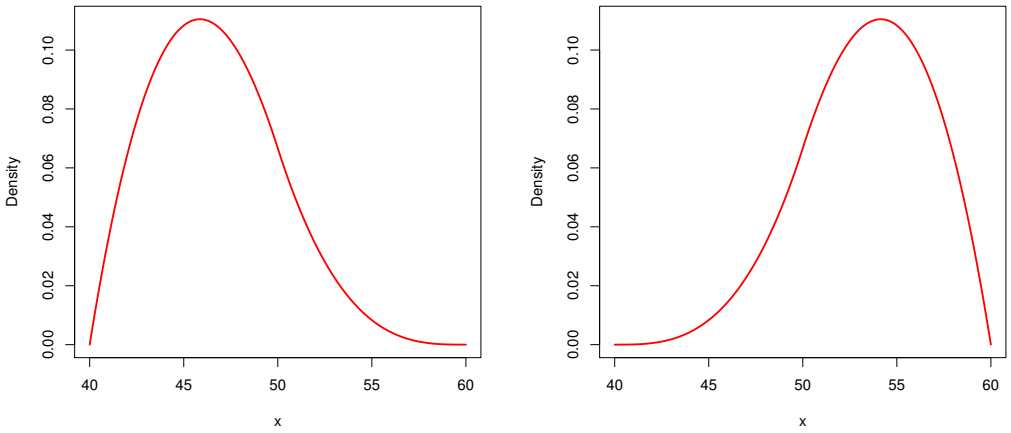


Figure 5: A sinistra: funzione di densità di probabilità della somma di due variabili casuali triangolari indipendenti, entrambe distribuite tra 20 e 30 e con massimo in 20. A destra: funzione di densità di probabilità della somma di due variabili casuali triangolari indipendenti, entrambe distribuite tra 20 e 30 e con massimo in 30.

0.0.2 Somma di una normale e di una lognormale indipendenti

Sia X una variabile casuale normale con media 2 e deviazione standard 0.25 e Y una variabile casuale log-normale con parametri 1 e 0.5.[4] Estraiamo 100000 osservazioni da ciascuna delle due distribuzioni e facciamone la somma.

```
> set.seed(123456)
> X <- rnorm(100000,2,0.25)
> Y <- rlnorm(100000,1,0.5)
> Z <- X + Y
```

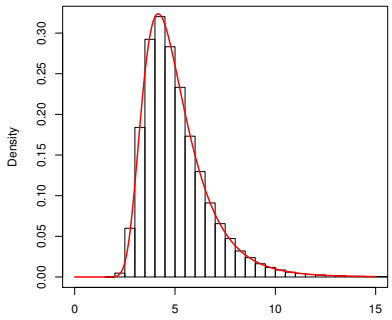
Prepariamo adesso la funzione che dovrà calcolare (numericamente, impiegando la funzione `integrate`) la funzione di densità di probabilità della somma $Z = X + Y$ mediante la convoluzione.

```
> f.X <- function(x) dnorm(x,2,0.25)
```

[4]Nella parametrizzazione che impiega R, la distribuzione log-normale che consideriamo è quella i cui logaritmi hanno media 1 e deviazione standard 0.5.

```
> f.Y <- function(y) dlnorm(y,1,0.5)
> f.Z <- function(z) integrate(function(x,z)
+ f.X(x)*f.Y(z-x),-Inf,Inf,z)$value
> f.Z <- Vectorize(f.Z)
```

Calcoliamo quindi i valori numerici della funzione di densità di probabilità di Z (chiamando la funzione `f.Z` appena definita) corrispondenti ad alcuni valori di Z, disegniamo l'istogramma dei valori ottenuti mediante la simulazione e sovrapponiamo la funzione di densità di probabilità di Z calcolata numericamente. Come si può vedere dalla figura riportata di seguito, l'adattamento è ottimo.




```
> z <- seq(0,15,0.01)
> f <- f.Z(z)
> hist(Z, breaks=50, xlim=c(0,15),
+ prob=TRUE)
> box()
> lines(z,f,lwd=2,col="red")
```

0.0.3 Somma di una normale e di una esponenziale indipendenti

Sia X_1 una variabile casuale esponenziale con media $1/5$ e X_2 una variabile casuale normale con media 10 e deviazione standard 2. Estraiamo 100000 osservazioni da ciascuna delle due distribuzioni e facciamone la somma.

```
> m <- 10; s <- 2; l <- 5
> set.seed(123456)
> x1 <- rexp(100000,l)
> x2 <- rnorm(100000,m,s)
> x <- x1 + x2
```

Prepariamo adesso la funzione che dovrà calcolare numericamente la funzione di densità di probabilità della somma $X = X_1 + X_2$ mediante la convoluzione.

```
> f.x1 <- function(y) dexp(y,l)
> f.x2 <- function(x) dnorm(x,m,s)
> f.x <- function(z) integrate(function(x,z)
+ f.x1(x)*f.x2(z-x),0,Inf,z)$value
> f.x <- Vectorize(f.x)
```

Nel definire i limiti di integrazione (ricordando che stiamo integrando in dx_1) occorre tenere presente che, nel caso di un variabile casuale X_1 esponenziale, deve sempre essere $X_1 \geq 0$. Abbiamo quindi che per soddisfare contemporaneamente $X_1 \geq 0$ e $-\infty < X_2 < +\infty$ dovremo integrare a partire da zero.

Calcoliamo ora i valori numerici della funzione di densità di probabilità di X (chiamando la funzione `f.x` appena definita) corrispondenti ad alcuni valori di X, disegniamo l'istogramma dei valori ottenuti mediante la simulazione e sovrapponiamo la funzione di densità di probabilità di X calcolata numericamente. Come si può vedere dal grafico riportato nel pannello di sinistra della figura 6, l'adattamento è ottimo.

```
> xx <- seq(0,20,0.1)
> f <- f.x(xx)
> hist(x, breaks=50, xlim=c(0,20), prob=TRUE, main="",
+ xlab="x"); box()
> lines(xx,f,lwd=2,col="red")
```

Il grafico della funzione di densità di probabilità della variabile casuale X rappresentato nel pannello di sinistra della figura 6 sembra quello di una normale. Anche la media e la deviazione standard dei valori simulati sono molto vicini ai parametri della normale che entra nella convoluzione.

```
> c(mean(x),sd(x))
[1] 10.202163  2.014442
```

Il motivo principale è legato al fatto che la variabile casuale esponenziale X_1, avendo $\lambda = 5$, scende in modo molto "ripido". Fra 0 e 1 è compreso il 99.3% delle osservazioni (e fra 0 e 0.1 è compreso il 39.3% delle osservazioni).

Vediamo un secondo esempio in cui la funzione di densità di probabilità della somma è alquanto diversa dalla normale.

Sia X_1 una variabile casuale esponenziale con media $1/2$ e X_2 una variabile casuale normale con media 1 e deviazione standard 0.2. Estraiamo 100000 osservazioni da ciascuna delle due distribuzioni e facciamone la somma.

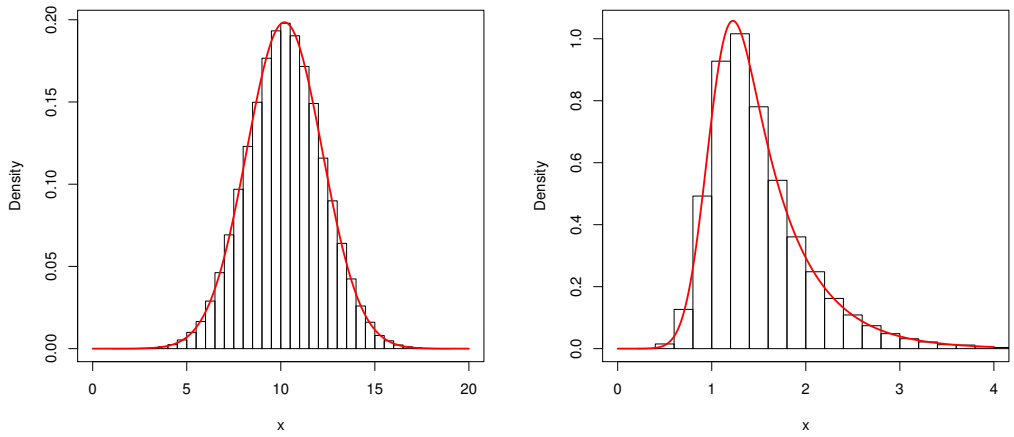


Figure 6: Distribuzione di probabilità della somma X di una variabile casuale esponenziale e di una variabile casuale normale indipendenti. Pannello di sinistra: $\lambda = 5, \mu = 10, \sigma = 2$. Pannello di destra: $\lambda = 2, \mu = 1, \sigma = 0.2$.

```
> m <- 1; s <- 0.2; l <- 2
> set.seed(123456)
> x1 <- rnorm(100000,m,s)
> x2 <- rexp(100000,l)
> x <- x1 + x2
```

La funzione che dovrà calcolare numericamente la funzione di densità di probabilità della somma $X = X_1 + X_2$ mediante la convoluzione è la stessa dell'esempio precedente.

```
> f.x1 <- function(y) dexp(y,l)
> f.x2 <- function(x) dnorm(x,m,s)
> f.x <- function(z) integrate(function(x,z) f.x1(x)*f.x2(z-x),
+ 0,Inf,z)$value
> f.x <- Vectorize(f.x)
```

Calcoliamo ora i valori numerici della funzione di densità di probabilità di X (chiamando la funzione `f.x`) corrispondenti ad alcuni valori di X, disegniamo l'istogramma dei valori ottenuti mediante la simulazione e sovrapponiamo la funzione di densità di probabilità di X calcolata numericamente.

Come si può vedere dal grafico riportato nel pannello di destra della figura 6, l'adattamento è ottimo.

```
> x <- seq(0,4,0.1)
> f <- f.x(x)
> hist(Z, breaks=50, xlim=c(0,4), prob=TRUE,
+ ylim=c(0,1.05), main="", xlab="x")
> box()
> lines(x,f,lwd=2,col="red")
```

In questo secondo esempio il grafico della funzione di densità di probabilità della variabile casuale X rappresentato nel pannello di destra della figura 6 è ben diverso quello di una normale. Anche la media e la deviazione standard dei valori simulati sono del tutto diversi da quelli della normale che entra nella convoluzione.

```
> c(mean(x),sd(x))
[1] 2.000000 1.197915
```