Laboratorio di R

Simulazioni Monte Carlo: teoremi limite

Valore atteso di una combinazione lineare di variabili casuali. Siano $X_1, \ldots, X_n$, n variabili casuali definite tutte sullo stesso spazio di probabilità $(\Omega, \mathcal{A}, \mathsf{P})$ e siano $c_1, \ldots, c_n$ costanti reali. Allora

$$\mathrm{E}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i \, \mathrm{E}(X_i).$$

Varianza di una combinazione lineare di variabili casuali indipendenti. Siano $X_1, \ldots, X_n$, n variabili casuali indipendenti definite tutte sullo stesso spazio di probabilità $(\Omega, \mathcal{A}, \mathsf{P})$ e siano $c_1, \ldots, c_n$ costanti reali. Allora

$$\mathrm{Var}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 \, \mathrm{Var}(X_i).$$

Media campionaria. Siano $X_1, \ldots, X_n$, n variabili casuali indipendenti tutte con lo stesso valore atteso $\mathrm{E}(X_i) = \mu$, $i = 1, \ldots, n$, e tutte con la stessa varianza $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, \ldots, n$, e sia $\overline{X}_n = (1/n) \sum_{i=1}^{n} X_i$, la loro *media aritmetica*. Allora

$$\begin{aligned} \mathrm{E}(\overline{X}_n) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mathrm{E}\left(\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\right) \\ &= \frac{1}{n}\,\mathrm{E}(X_1) + \cdots + \frac{1}{n}\,\mathrm{E}(X_n) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\, n\mu = \mu, \end{aligned}$$

e

$$\begin{aligned} \mathrm{Var}(\overline{X}_n) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mathrm{Var}\left(\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\right) \\ &= \frac{1}{n^2}\,\mathrm{Var}(X_1) + \cdots + \frac{1}{n^2}\,\mathrm{Var}(X_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Si noti che per derivare che $\mathrm{E}(\overline{X}_n) = \mu$ e $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$, non abbiamo assunto che le variabili casuali $X_1, \ldots, X_n$ abbiano tutte la stessa identica distribuzione, ma solo che queste siano indipendenti e che abbiano tutte la stessa media e la stessa varianza.

Legge (debole) dei grandi numeri. Fondamentalmente, esistono due leggi dei grandi numeri, quella debole e quella forte. Matematicamente queste due leggi esprimono due risultati completamente diversi, anche se nel linguaggio comune di tutti i giorni (lotterie, giochi di sorte ecc.) hanno entrambe a che vedere con l'evidenza empirica che al crescere delle ripetizioni di un esperimento sotto identiche condizioni, le frequenze relative e, più in generale, le medie aritmetiche, tendono a stabilizzarsi. Nello specifico, la *legge debole dei grandi numeri* è un teorema che riguarda la distribuzione della media aritmetica fatta su un grande numero di osservazioni. In particolare, sia $X_1, \ldots, X_n, \ldots$ una successione infinita di variabili casuali indipendenti ed identicamente distribuite (i.i.d.) con valore atteso $\mathrm{E}(X_i) = \mu$, $i = 1, \ldots, n, \ldots$, e varianza $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, \ldots, n, \ldots$, e siano $\overline{X}_n = (1/n) \sum_{i=1}^{n} X_i$, per $n = 1, 2, \ldots$, le medie aritmetiche fatte sulle prime n variabili casuali. Allora, per ogni $\varepsilon > 0$, piccolo a piacere, si ha che

$$\lim_{n \to \infty} P(|\overline{X}_n - \mu| \geq \varepsilon) = 0,$$

o, equivalentemente, che

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| < \varepsilon) = 1.$$

Teorema del limite centrale. Il *teorema del limite centrale* (o *teorema centrale del limite*) prende il suo nome dal fatto che per lungo tempo nei secoli scorsi l'evidenza empirica della sua esistenza fu considerata come il problema centrale di tutta la teoria della probabilità. Questo teorema è strettamente collegato alla distribuzione normale (chiamata anche gaussiana) e ne fa di questa la più importante tra tutte le distribuzioni. Date n variabili casuali $X_1, \ldots, X_n$, indipendenti con valore atteso $\mathrm{E}(X_i) = \mu$, $i = 1, \ldots, n$, e varianza $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, \ldots, n$, sia $S_n = X_1 + \cdots + X_n$ la loro somma, sia $\overline{X}_n = (1/n)\sum_{i=1}^n X_i = S_n/n$ la loro media aritmetica, e sia quindi

$$Z_n = \frac{\overline{X}_n - \mathrm{E}(\overline{X}_n)}{\sqrt{\mathrm{Var}(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

la loro media aritmetica standardizzata, che può anche essere scritta come

$$Z_n = \frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}.$$

Nello specifico, il teorema del limite centrale è un teorema che riguarda la distribuzione della media aritmetica standardizzata fatta su un grande numero di osservazioni. In particolare, sia $X_1, \ldots, X_n, \ldots$ una successione infinita di variabili casuali indipendenti ed identicamente distribuite (i.i.d.) con valore atteso $\mathrm{E}(X_i) = \mu$, $i = 1, \ldots, n, \ldots$, e varianza $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, \ldots, n, \ldots$, e siano $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$, per $n = 1, 2, \ldots$, le medie aritmetiche fatte sulle prime n variabili casuali. Allora, per ogni $z \in \mathbf{R}$,

$$\lim_{n\to\infty} P\left(\frac{\overline{X}_n - \mathrm{E}(\overline{X}_n)}{\sqrt{\mathrm{Var}(\overline{X}_n)}} \le z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,\mathrm{d}u = \Phi(z),$$

dove $\Phi(z)$ è la funzione di ripartizione della distribuzione normale standard $N(0,\ 1)$.

Simulazioni Monte Carlo. I metodi di simulazione Monte Carlo forniscono un potentissimo strumento per la risoluzione di problemi complessi. Fondamentalmente, si può sempre vedere una simulazione Monte Carlo come un metodo per approssimare un valore atteso attraverso una semplice media aritmetica. Tratteremo questo argomento implementando alcuni semplici esempi con il software R.

Esercizio 1 (legge debole dei grandi numeri)
Simulare 1000 realizzazioni delle variabili casuali $X_1, \ldots, X_n$ (per diversi valori di n) indipendenti e aventi distribuzione Bernoulliana con probabilità pari a $p = 0.5$.
a) Si ottenga l'istogramma di frequenze delle medie aritmetiche (sulle n variabili casuali) basato sulle 1000 realizzazioni.
b) Si ripeta l'esperimento assmendo che le variabili casuali abbiano distribuzione uniforme tra zero e uno.

Soluzione Monte Carlo

```
prove <- 1000; n <- 50; # n = 10, 50, 500, 5000
x <- 0; xmedia <- 0;

set.seed(1325)
```

```
t0 <- proc.time()
for (i in 1:prove) {
  x <- rbinom(n, size=1, prob=0.5)
  xmedia[i] <- mean(x)
}
t1 <- proc.time() - t0; t1

hist(xmedia,xlab="x",col="red",freq=FALSE,breaks=seq(0,1,0.01))

##### variabili casuali con distribuzione uniforme in (0;1))
x <- 0; xmedia <- 0;
mu <- 0.5; sigma2 <- 1/12;

set.seed(1325)
t0 <- proc.time()
for (i in 1:prove) {
  x <- runif(n, min=0, max=1)
  xmedia[i] <- mean(x)
}
t1 <- proc.time() - t0; t1

hist(xmedia,xlab="x",col="red",freq=FALSE,breaks=seq(0,1,0.01))
```

Esercizio 2 (passeggiata casuale simmetrica)
Simulare 10 realizzazioni delle variabili casuali dicotomiche $X_1, \ldots, X_n$ (per $n = 5000$) indipendenti con $P(X_i = -1) = P(X_i = 1) = 0.5$, per $i = 1, \ldots, n$.
a) Si ottengano, per ognuna delle 10 realizzazioni, le somme cumulate $S_m = X_1 + \cdots + X_m$, per $m = 1, \ldots, n$, e si rappresentino in un grafico.
b) Si ottengano, per ognuna delle 10 realizzazioni, le medie aritmetiche cumulate $\overline{X}_m = (1/m) \sum_{i=1}^{m} X_i$, per $m = 1, \ldots, n$, e si rappresentino in un grafico.

Soluzione Monte Carlo

```
prove <- 10; n <- 5000; # n = 10, 50, 500, 5000
x <- 0; xmedia <- 0; somma <- 0; media <- 0;

# set.seed(1325)
plot(c(seq(0,n,1),0,0),c(rep(0,n+1),-sqrt(n)*2,sqrt(n)*2),xlab="",ylab="",pch=".")
t0 <- proc.time()
for (i in 1:prove) {
  x <- rbinom(n, size=1, prob=0.5)*2 - 1
  for (j in 1:n) {
    somma[j] <- sum(x[1:j])
  }
  points(seq(1,n,1),somma,type="l",pch=".",col=i)
}
t1 <- proc.time() - t0; t1

plot(c(seq(0,n,1),0,0),c(rep(0,n+1),-1,1),xlab="",ylab="",pch=".")
t0 <- proc.time()
for (i in 1:prove) {
  x <- rbinom(n, size=1, prob=0.5)*2 - 1
  for (j in 1:n) {
    media[j] <- mean(x[1:j])
  }
  points(seq(1,n,1),media,type="l",pch=".",col=i)
}
```

```
t1 <- proc.time() - t0; t1
```

Esercizio 3 (teorema del limite centrale)
Simulare 1000 realizzazioni delle variabili casuali $X_1, \ldots, X_n$ (per diversi valori di n) indipendenti e aventi distribuzione Bernoulliana con probabilità pari a $p = 0.5$. Si costruiscano quindi le medie aritmetiche $\overline{X}_n = (1/n) \sum_{i=1}^{n} X_i$ e le medie aritmetiche standardizzate

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

dove μ e σ^2 sono il valore atteso e la varianza delle variabili casuali $X_1, \ldots, X_n$.
a) Si ottenga l'istogramma di frequenze delle medie aritmetiche standardizzate (sulle n variabili casuali), basato sulle 1000 realizzazioni.
b) Si costruisca la funzione di ripartizione delle medie aritmetiche standardizzate (sulle n variabili casuali), basata sulle 1000 realizzazioni.
c) Paragonare la funzione di ripartizione delle medie aritmetiche standardizzate con la funzione di ripartizione della normale standard.
d) Si ripeta l'esperimento assmendo che le variabili casuali $X_1, \ldots, X_n$ abbiano distribuzione uniforme tra zero e uno, e quindi che abbiano distribuzione esponenziale di parametro $\lambda = 1$.

Soluzione Monte Carlo

```
prove <- 1000; n <- 50; # n = 10, 50, 500, 5000
x <- 0; xmedia <- 0; xzeta <- 0;
mu <- 0.5; sigma2 <- 1/4;

mediazeta <- function(x,n,mu,sigma2) {(x-mu)/(sqrt(sigma2/n))}

set.seed(1325)
t0 <- proc.time()
for (i in 1:prove) {
  x <- rbinom(n, size=1, prob=0.5)
  xmedia[i] <- mean(x)
  xzeta[i] <- mediazeta(xmedia[i],n,mu,sigma2)
}
t1 <- proc.time() - t0; t1

hist(xzeta,xlab="x",col="red",freq=FALSE, breaks=seq(-4,4,0.1))

plot(ecdf(xzeta), do.points=FALSE,verticals=TRUE)
mticks <- seq(-4, 4, 0.1)
lines(mticks, pnorm(mticks, mean=0, sd=1), lty=3, col="red")

##### distribuzione uniforme in (0;1)) ##########################
x <- 0; xmedia <- 0; xzeta <- 0;
mu <- 0.5; sigma2 <- 1/12;

set.seed(1325)
t0 <- proc.time()
for (i in 1:prove) {
  x <- runif(n, min=0, max=1)
  xmedia[i] <- mean(x)
  xzeta[i] <- mediazeta(xmedia[i],n,mu,sigma2)
}
t1 <- proc.time() - t0; t1
```

```
hist(xzeta,xlab="x",col="red",freq=FALSE, breaks=seq(-4,4,0.1))

plot(ecdf(xzeta), do.points=FALSE,verticals=TRUE)
mticks <- seq(-4, 4, 0.1)
lines(mticks, pnorm(mticks, mean=0, sd=1), lty=3, col="red")

##### distribuzione esponenziale ################################
x <- 0; xmedia <- 0; xzeta <- 0;
lambda <- 1; mu <- 1/lambda; sigma2 <- 1/lambda^2;

set.seed(1325)
t0 <- proc.time()
for (i in 1:prove) {
  x <- rexp(n, rate=lambda)
  xmedia[i] <- mean(x)
  xzeta[i] <- mediazeta(xmedia[i],n,mu,sigma2)
}
t1 <- proc.time() - t0; t1

hist(xzeta,xlab="x",col="red",freq=FALSE, breaks=seq(-6,6,0.1))

plot(ecdf(xzeta), do.points=FALSE,verticals=TRUE)
mticks <- seq(-6, 6, 0.1)
lines(mticks, pnorm(mticks, mean=0, sd=1), lty=3, col="red")
```