1 Somma di v.a. indipendenti

Prendiamo 100 v.a. indipendenti $X_1, ..., X_{100}$ e studiamo la v.a.

$$S = X_1 + ... + X_{100}.$$

Ad esempio, in un gioco, se X_n rappresenta la vincita o perdita alla partita n-esima, S rappresenta la vincita o perdita globale dopo 100 partite. Che valori ci aspettiamo da S?

Un'altro esempio interessante in relazione alle esponenziali che useremo tra breve, è il seguente: se X_n rappresenta il tempo che intercorre tra l'arrivo di una richiesta di servizio e la successiva (in un sistema di servizio), cosa possiamo dire dell'istante in cui arriva la centesima richiesta? Supponiamo ad esempio che dopo 100 richieste si debba procedere a rifornimenti, sostituzioni, ecc...

Con R, generiamo 100 numeri casuali da una distribuzione a piacere. Ad esempio una esponenziale di parametro 1:

$$x < -r\exp(100, rate = 1).$$

Poi calcoliamo la somma

$$s < -sum(x).$$

Il valore che si trova non è lontano da 100, ma ad esempio (dipenden dalla simulazione), può essere 96.76397. Cosa possiamo immaginare? Eseguiamo qualche altra simulazione. Basta ripetere i due comandi visti sopra più volte; ad esempio si può trovare

$$\begin{array}{lll} 100.0858, & 91.43403, & 102.7003 \\ 97.46942, & 106.692, & 99.20531 \\ 114.9421, & 85.89443, & 82.23042. \end{array}$$

Abbastanza approssimativamente, questi numeri si trovano intorno al 100 (un po' meno), con una deviazione molto variabile, mediamente tra i 5 e 10. Questi numeri sperimentali non permetterebbero di identificare il risultato teorico, forse per il fatto che sono pochi, forse per la modesta aleatorietà del generatore di numeri casuali. Il risultato teorico è che S ha media 100 e deviazione standard 10. Infatti, per la linearità del valor medio

$$E[S] = E[X_1] + ... + E[X_{100}]$$

e per l'indipendenza

$$Var[S] = Var[X_1] + ... + Var[X_{100}]$$

ed inoltre tutte le v.a. X_i, in quanto esponenziali di parametro 1, hanno media 1 e varianza 1. Riassumendo, a meno di imperfezioni dovute o alla piccolezza del campione o alla poca bontà del simulatore, vediamo che i valori sperimentali di S rispecchiano abbastanza le previsioni teoriche.

Sorgono però due domande:

1) che distribuzione ha S?

2) se invece di effettuare la somma di 100 valori ne usassimo 1000, otterremmo un'approssimazione migliore della media?

2 Distribuzione della somma

Vorremmo fare un istogramma dei valori sperimentali di S. Quindi bisogna generare molte volte un campione di esponenziali di numerosità 100, registrare i relativi valori di S e tracciare un istogramma. Come si fa con R in modo automatizzato? Scriviamo un brevissimo programma iterativo.

- Creiamo un vettore di somme S: s<-1:100

- riempiamolo con le somme di 100 numeri esponenziali:

```
for(k in 1:100) {
s[k] <- sum(rexp(100, rate = 1))
}
```

- osserviamo media, deviazione, istogramma, q-q plot gaussiano:

```
s
mean(s)
sd(s)
hist(s)
hist(s,20)
qqnorm(s)
```

Tutto questo porta a pensare che S sia approssimativamente gaussiana, pur essendoci un certo margine di errore sperimentale. Non scordiamoci però che

$$E[S] = 100E[X_1]$$
$$Var[S] = 100Var[X_1]$$

per cui non si tratta di una gaussiana canonica.

A parole, si usa dire che la somma S_n di n v.a. indipendenti ugualmente distribuite $X_1, ..., X_n$, aventi media μ e deviazione σ, è approssimativamente una gaussiana di media $n \cdot \mu$ e deviazione $\sqrt{n} \cdot \sigma$. Questo enunciato non possiede l'usuale precisione dei teoremi matematici in quanto non stabilisce cosa significhi "approssimativamente" (termine che invece ha un certo significato pratico). Per questo si usa dare l'enunciato nella forma di teorema limite: la grandezza standardizzata

$$\frac{S_n - n \cdot \mu}{\sqrt{n} \cdot \sigma}$$

tende (in un senso opportuno) ad una gaussiana canonica $N(0,1)$.

3 Convergenza alla media

Il fatto che

$$\mu_{S_n} = n \cdot \mu$$
$$\sigma_{S_n} = \sqrt{n} \cdot \sigma$$

mostra che, anche se mediamente i valori di S_n sono pari a $n \cdot \mu$, la loro deviazione cresce all'infinito. Quindi i valori sperimentali di S_{1000} non sono più vicini a $1000 \cdot \mu$ di quanto i valori di S_{100} fossero vicini a $100 \cdot \mu$. La vicinanza si può cogliere dal punto di vista *relativo*:

$$\frac{S_n}{n}$$

è vicino a μ, ed il grado di approssimazione migliora al crescere di n. Verifichiamo con R quest'affermazione, in due modi completamente diversi.

Il primo modo consiste nell'osservare gli istogrammi di $\frac{S_{100}}{100}$ e $\frac{S_{1000}}{1000}$. Si vede chiaramente che i valori di $\frac{S_{1000}}{1000}$ sono molto più concentrati attorno a μ. Dal punto di vista teorico, questo si apprezza col calcolo:

$$Var\left[\frac{S_n}{n}\right] = \frac{1}{n^2} Var\left[S_n\right] = \frac{\sigma^2}{n}$$

per cui

$$\sigma_{\frac{S_{100}}{100}} = \frac{\sigma}{\sqrt{100}} = 0.1$$
$$\sigma_{\frac{S_{1000}}{1000}} = \frac{\sigma}{\sqrt{1000}} = 0.0316.$$

Si deve però anche riconoscere che il miglioramento non va proporzionalmente alla crescita di n, ma secondo la sua radice quadrata. Questo va sempre ricordato quando si vuole migliorare la stima di un valor medio (nel senso che, se gli esperimenti sono costosi, bisogna accontentarsi, in quanto servono troppi esperimenti per ottenere miglioramenti apprezzabili).

Un teorema che stabilisce un'affermazione in linea con la simulazione ora fatta è la *legge debole dei grandi numeri.* In realtà essa stabilisce solo che la probabilità di osservare $|S_n - \mu| > \delta$, con $\delta > 0$ fissato, tende a zero per $n \to \infty$.

Un secondo modo di cogliere l'avvicinamento di $\frac{S_n}{n}$ a μ per $n \to \infty$ è quello di osservare *una traiettoria, una storia*, *al crescere di* n, invece che fissare n e osservare con un istogramma tante simulazioni relative a quel valore di n. Questo è il semplice programma:

```
mu.n<-1:1000
for(k in 1:1000) {
mu.n[k] <- sum(rexp(k, rate = 1))/k
}
plot(mu.n)
```

Già che ci siamo, è interessante osservare invece il comportamento di $S_n - \mu \cdot n$ al crescere di n:

```
y.n<-1:1000
for(k in 1:1000) {
y.n[k] <- sum(rexp(k, rate = 1))-k
}
plot(y.n)
```

Si vede l'allargamento che va come $\sqrt{n}$. Qui si vede che, in un gioco di azzardo equo, non conviene usare la strategia di giocare ad oltranza sapendo che prima o poi la nostra vincita supererà il valore che ci prefissiamo inizialmente: infatti, può capitare che prima di quel momento la nostra perdita sia stata tale da mandarci in bancarotta.