Negli *script* presentati nel testo, il numero delle repliche di ciascun "esperimento" è stato fissato generalmente pari a 10000 (qualche volta a 100000). In queste brevi note vorremmo proporre qualche riflessione sul modo in cui si possa cercare di identificare un numero "adeguato" di repliche per una simulazione.

Se indichiamo con $p$ la probabilità di un evento, analiticamente determinabile (almeno in linea di principio) impiegando le regole del calcolo delle probabilità e con $\hat{p}$ una sua approssimazione (*stima*) ottenuta attraverso una simulazione, allora, a patto che $p$ non sia troppo vicina a 0 oppure a 1, la distribuzione di campionamento di $\hat{p}$ (cioè, in buona sostanza, i diversi valori che possiamo aspettarci replicando un grandissimo numero di volte l'intera simulazione) potrà essere ragionevolmente approssimata impiegando la distribuzione gaussiana con media $p$ e varianza $\frac{p\times(1-p)}{n}$, dove $n$ è il numero di repliche fissato per la simulazione. In simboli: $\hat{p} \sim N\left(p, \frac{p\times(1-p)}{n}\right)$. Ricordando le proprietà della distribuzione normale, allora il 95% circa dei risultati (cioè dei valori $\hat{p}$ che le simulazioni potranno produrci) saranno compresi in un intervallo centrato sul valore analitico $p$ e di ampiezza pari a $2\times 1.96\times\sqrt{\frac{p\times(1-p)}{n}}$.

Prendiamo, a titolo di esempio, l'esercizio 3.36 a pag. 92 (il problema del Cavalier de Méré). Abbiamo visto che la probabilità di ottenere almeno un 1 in quattro lanci di un dado è pari a $1-\left(\frac{5}{6}\right)^4 = \frac{671}{1296}$, cioè (con 6 cifre decimali) 0.517747. Fissando allora un numero di repliche pari a 10000, la varianza della distribuzione di campionamento di $\hat{p}$ (impiegando l'approssimazione normale) sarà pari a circa 0.0000249685 e l'errore standard a 0.00499685. L'ampiezza dell'intervallo che comprende il 95% dei possibili risultati della simulazione è circa 0.0195876. In altre parole, con una simulazione che impiega 10000 repliche, dobbiamo considerare che tutti i valori di $\hat{p}$ compresi fra 0.498159 e 0.537335 sono delle valide approssimazioni numeriche del valore vero 0.517747. Notate che nell'esercizio, che proponeva il problema del Cavalier de Méré, era necessario che la simulazione riuscisse a distinguere fra 0.517747 (la soluzione corretta) e 0.491404 (la probabilità di vincere al secondo gioco). In effetti, con 10000 repliche, il risultato più estremo (ad un livello del 95%) è comunque superiore a tale valore. In altre parole, anche se la nostra simulazione fosse piuttosto "sfortunata", comunque, con una probabilità del 95%, non dovrebbe darci un valore inferiore a 0.498159.

L'ampiezza dell'intervallo sopra calcolato (facciamo notare che **non** si tratta di un intervallo di confidenza) è di poco inferiore a 2 punti percentuali. Supponiamo di volere un intervallo di ampiezza non superiore a 1 punto percentuale. È evidente che per raggiungere questo risultato dovremo aumentare il numero delle repliche; ma **quante** repliche dobbiamo fare?
Una prima risposta "intuitiva" è: circa 40000 repliche. Infatti per dimezzare l'ampiezza dell'intervallo dobbiamo quadruplicare il numero delle repliche (raddoppiarle non basta, perché il numero $n$ di repliche figura sotto il simbolo di radice quadrata).
Possiamo però dare anche una risposta "analitica". Si tratta infatti di risolvere la seguente disequazione: $2\times 1.96\times\sqrt{\frac{p\times(1-p)}{n}} \le 0.01$, dove a $p$ sostituiamo il valore $\frac{671}{1296}$ e $n$ è l'incognita. La soluzione ci dice che per ottenere il risultato cercato, dobbiamo eseguire almeno 38368 repliche. Possiamo verificare la correttezza del risultato ricorrendo a **R**, come già fatto altre volte nel testo:

```
> p <- 671/1296
> n <- c(38301:38399)
> amp <- 2*1.96*sqrt(p*(1-p)/n)
> amp[61:69]
[1] 0.010000861 0.010000730 0.010000600 0.010000470
[5] 0.010000339 0.010000209 0.010000079 0.009999948
[9] 0.009999818
```

Nel testo abbiamo impiegato 100000 repliche. In questo caso, il 95% dei possibili risultati della simulazione saranno compresi fra 0.511553 e 0.523941 (quello riportato nel testo, 0.5169, è uno di questi). L'ampiezza dell'intervallo con 100000 repliche è di poco superiore a 0.006, quasi a metà strada fra 0.01 e 0.001.

Ma potremmo seguire un altro approccio, in particolare in quei casi in cui non è disponibile una soluzione analitica al problema proposto. Si pensi, ad esempio, al problema dei cerini di Banach (vedi pag. 184 del testo). In questo caso, possiamo trattare il risultato ottenuto $\hat{p}$ come una *stima puntuale* del valore vero e sconosciuto $p$ e costruire un *intervallo di confidenza* la cui ampiezza è una misura della *precisione* con cui stimiamo il valore esatto $p$. La formula dell'intervallo di confidenza in parola è la seguente (vedi pag. 236):

$$\left[ \hat{p} - z_{\alpha/2} \times \sqrt{\frac{p \times (1-p)}{n}} \; ; \hat{p} + z_{\alpha/2} \times \sqrt{\frac{p \times (1-p)}{n}} \right]$$

Purtroppo, non conoscendo il vero valore $p$, dovremo sostituire, nella formula precedente, a questo la sua stima $\hat{p}$, rendendo la risposta ancora più approssimata. L'ampiezza dell'intervallo di confidenza sarà quindi data da $2 \times z_{\alpha/2} \times \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}}$ e non è calcolabile se non dopo avere eseguito la simulazione ed avere osservato il risultato $\hat{p}$. Ricordando però quando detto a pag. 254 a proposito della varianza di una variabile casuale di Bernoulli, qualora si disponga di una formula approssimata per la soluzione ovvero si abbia una ragionevole idea di quanto possa valere $p$, risulta possibile usare "conservativamente" tale informazione, approssimandola con un valore più vicino a 0.5 (dove la varianza è massima).

Riprendiamo in esame il problema del Cavalier de Méré e supponiamo di non conoscere la soluzione analitica, ma di essere ragionevolmente sicuri che essa sia "vicina" a 0.5 (che è anche il valore per cui la varianza di una Bernoulli è massima). Allora l'ampiezza di un intervallo di confidenza al 95% non potrà, **in nessuna caso**, essere superiore a $2 \times 1.96 \times \sqrt{\frac{0.5 \times (1-0.5)}{n}} = \frac{1.96}{\sqrt{n}}$.
Per ottenere una ampiezza di un punto percentuale sono necessarie almeno 38416 repliche. Vediamo di ripetere l'*esperimento* proposto a pag. 92 con questo numero di repliche.

```
> dado <- c(1:6)
> prove <- 38416
> l <- sample(dado,size=4*prove,replace=TRUE)
> m <- l==1; rm(l)
> n <- matrix(m,ncol=4,byrow=TRUE); rm(m)
> ris <- apply(n,1,sum); rm(n)
> ris <- ris>0
```

```
> num <- sum(ris)
> ph <- num/prove
> se <- sqrt(ph*(1-ph)/prove)
> cil <- ph-1.96*se
> cih <- ph+1.96*se
> cat(ph,cil,cih,(cih-cil),"\n")
0.5205383 0.5155425 0.5255341 0.00999156
```

Lo *script* è un po' diverso da quello proposto nel testo, ma dovreste essere oramai in grado di comprenderne il significato; la differenza più sostanziale consiste nell'avere evitato il ciclo `for` memorizzando i risultati sotto forma di matrice (i cui elementi valgono `TRUE` o `FALSE`) alle cui righe viene applicata la funzione `sum`. Possiamo osservare che l'intervallo di confidenza va da circa 0.516 a circa 0.526 e che la sua ampiezza, come richiesto, è appena inferiore a 0.01.

Supponiamo di volerci spingere più in avanti con la richiesta di precisione per arrivare ad una ampiezza non superiore a 0.001; saranno necessarie 3841600 repliche (ampiezza **10** volte più piccola, numero di repliche **100** volte più grande). È facile rendersi conto di quanto "costi" aumentare la precisione con cui si stima il risultato. La simulazione prevederà l'estrazione di $3841600 \times 4 = 15366400$ (oltre quindici milioni!) di numeri pseudo-casuali, il che può comportare alcuni problemi di memoria se il *computer* su cui si esegue la simulazione non ha RAM a sufficienza (e, dopo tutto, anche **R** non può gestire una quantità di memoria illimitata). Ecco comunque lo *script* e, soprattutto, il risultato ottenuto. Facciamo notare che con la funzione `rm` liberiamo spazio in memoria non appena un oggetto "corposo" non ci serve più. Se nello *script* precedente questa poteva apparire una inutile "raffinatezza", in questo caso diventa indispensabile per non avere da **R** un *out of memory error*.

```
 > dado <- c(1:6)
> prove <- 3841600
> l <- sample(dado,size=4*prove,replace=TRUE)
> m <- l==1; rm(l)
> n <- matrix(m,ncol=4,byrow=TRUE); rm(m)
> ris <- apply(n,1,sum); rm(n)
> ris <- ris>0
> num <- sum(ris)
> ph <- num/prove
> se <- sqrt(ph*(1-ph)/prove)
> cil <- ph-1.96*se
> cih <- ph+1.96*se
> cat(ph,cil,cih,(cih-cil),"\n")
0.5174716 0.5169719 0.5179713 0.0009993893
```

Il risultato analitico dovrebbe essere compreso (con una confidenza del 95%) fra circa 0.517 e 0.518, con una incertezza sul terzo decimale. Spingersi oltre con la richiesta di precisione (ampiezza 0.0001) potrebbe rivelarsi davvero proibitivo (oltre 384 milioni di repliche!).

Queste poche righe fanno capire come l'approccio seguito nel testo sia solo il primo passo per affrontare il problema della simulazione. È effettivamente possibile (oltre che necessario) spingersi in avanti con la richiesta di precisione di una stima, ma, allora, diventa necessario fare ricorso a tecniche *Monte Carlo* molto più raffinate di quella proposta nel nostro testo didattico, che è stata volutamente *naive*. Per questi ed altri approfondimenti si invita il lettore interessato alla consultazione di testi specifici. Un semplice esempio introduttivo, che propone la stima del numero

*e* mediante simulazione è proposto nel capitolo 1 *on line* su questa stessa pagina *web*. Buone simulazioni!