Lezione n. 6

ESEMPIO 6.1 *Distribuzione esponenziale*
Sia $\underline{x} = (x_1, \ldots, x_n)$ un campione estratto da una popolazione esponenziale di parametro λ: $X_i \sim$ i.i.d. Exp(λ).

Verificare che lo stimatore $T_n = 1/\bar{X}$ è uno stimatore distorto di λ e proporre uno stimatore non distorto. Valutare inoltre la distorsione di T_n al variare di n.
SOLUZIONE. *Per calcolare il valore atteso dello stimatore* $T_n = 1/\bar{X} = n/\sum_{i=1}^n X_i$, *utilizziamo la conoscenza della distribuzione della variabile aleatoria somma posta al denomiatore. Essendo la popolazione esponenziale,* $Y_n = \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$*: usando infatti la f.g.m. abbiamo, per l'indipendenza e per la somiglianza delle v.a. coinvolte e ricordando che la f.g.m. di una v.a. esponenziale di parametro* λ *è* $\frac{1}{1-\lambda t}$*:*

$$g_Y(t) = \prod_{i=1}^n g_{X_i}(t) = \left(\frac{1}{1-\lambda t}\right)^n,$$

che è appunto la f.g.m. di una Gamma di parametri n *e* λ*. Per calcolare il valore atteso di* $1/Y$ *sapendo che* Y *ha densità* $f_y(\cdot)$ *in generale possiamo utilizzare l' espressione* $E(1/Y) = \int 1/y f(y) dy$ *("formula dello statistico incosciente"). Nel nostro caso conosciamo la densità di* $Y_n = \sum_{i=1}^n X_i$ *(Gamma con parametri* n, λ*), per cui, essendo* $\text{E}(T_n) = \text{E}(n/Y_n)$*, scriviamo*

$$\text{E}(\frac{n}{\sum_{i=1}^n X_i}) = n \int_0^\infty \frac{1}{s} \frac{\lambda^n s^{n-1} e^{-\lambda s}}{(n-1)!} ds = \frac{n\lambda^n}{(n-1)!} \int_0^\infty s^{n-2} e^{-\lambda s} ds.$$

Riconosciamo nell'ultimo integrale il nucleo di una Gamma$(n-1, \lambda)$*; di conseguenza otteniamo*

$$E(\frac{n}{\sum_{i=1}^n X_i}) = n \frac{\lambda^n}{(n-1)!} \frac{(n-2)!}{\lambda^{n-1}} = \lambda \frac{n}{n-1}$$

Quindi $T_n = \frac{n}{\sum X_i}$ *ha valore atteso pari a* $\frac{n}{n-1}\lambda$*. Se ne deduce che* T_n *è uno stimatore* distorto *di* λ*. Tuttavia essendo* n *noto, si può facilmente pervenire a uno stimatore* non distorto*:* $\tilde{T}_n = \frac{n-1}{\sum_i X_i}$ *è infatti uno stimatore non distorto di* λ*.*

Si dimostra facilmente che la distorsione di T_n tende a zero al crescere di n: essa è infatti pari a: $B(T_n, \lambda) = E(T_n) - \lambda = \lambda/(n-1)$. *Lo stimatore è dunque* asintoticamente non distorto. *Come si vedrà in seguito, lo stimatore T_n è anche lo* stimatore di massima verosimiglianza *di* λ. ■

Analizziamo per simulazione quali sono le proprietà dei due stimatori. La simulazione ci permette di costruire un'approssimazione della distribuzione di uno stimatore, naturalmente per una prefissata scelta del parametro (che quindi di fatto è noto!). Infatti, sulla base di un numero elevato m di campioni simulati, si costruiscono m realizzazioni di ciascuno stimatore in analisi, e tali realizzazioni simulate consentono di valutare per approssimazione le caratteristiche della distribuzione di tali stimatori. Ad esempio si può calcolare la media sulle m simulazioni delle realizzazioni, ottenendo cosí un'approssimazione del valore atteso dello stimatore (in corrispondenza del valore del parametro che è stato scelto). Più in generale, l'istogramma costruito su questi m valori permette di approssimare la distribuzione campionaria dello stimatore che si analizza.

Per ottenere un'approssimazione della distribuzione campionaria di una statistica T_n possiamo dunque simulare l'operazione del campionamento ripetuto: estraiamo, nelle stesse condizioni, una serie (con un numero di termini m sufficientemente elevato) di campioni casuali di numerosità n da una distribuzione di probabilità prefissata f_θ, quindi su ciascun campione $\underline{x}^j (j = 1, \ldots, m)$ calcoliamo il valore osservato per la statistica, $T_n(\underline{x}^j) = t_n^j$ Ripetendo l'operazione per $j = 1, \ldots m$ otteniamo un vettore di risultati t_n^j la cui distribuzione empirica (visualizzabile tramite un istogramma) approssima, per m sufficientemente elevato, la distribuzione campionaria di T_n.

Prima di tutto generiamo m campioni tra loro indipendenti di ampiezza n, tutti estratti dalla stessa popolazione, su cui poi calcoleremo t_n e $\tilde{t}_n$. Dobbiamo estrarre per m volte n realizzazioni indipendenti da una legge Exp(λ), con λ a nostra scelta. Data l'indipendenza é quindi sufficiente estrarre un lungo vettore di $n \cdot m$ relizzazioni da tale legge e poi disporre questo vettore in una matrice $n * m$. Scegliamo un valore di n, ad esempio 10 e un valore di m sufficientemente elevato, ad esempio 5000, e creiamo una matrice in cui ogni colonna rappresenta un campione di n elementi estratto da una distribuzione esponenziale di parametro λ, con λ fissato a propria scelta, ad esempio 1/2:

```
n <- 10
m <- 5000
lam <- 1/2
mat <- matrix(rexp(n*m,lam), nrow=n)
```

A questo punto su ciascuno dei 5000 campioni generati calcoliamo i valori dei due stimatori

$T_n = n/\sum_i X_i$ e $\tilde{T}_n = (n-1)/\sum_i X_i$. A tale scopo possiamo utilizzare la funzione `apply`, che consente di applicare la stessa funzione per riga o per colonna ad una matrice. Come già visto in precedenza, i suoi argomenti sono: la matrice a cui applicare la funzione, l'argomento (1,2) che indica se la funzione va applicata per riga o per colonna, e infine la funzione da applicare. Ad esempio per calcolare $\sum_{i=1}^{n} X_i^j$ per $j = 1, \ldots, 5000$ scriviamo `somme <- apply(mat, 2, sum)`. Il risultato è un vettore di $m = 5000$ elementi ciascuno dei quali rappresenta la somma dei 10 elementi di uno dei 5000 campioni. Calcoliamo dunque le realizzazioni di T_n e $\tilde{T}_n$:

```
stim1 <- n/somme
stim2 <- (n-1)/somme
# si noti che lo stesso risultato si poteva ottenere scrivendo ad esempio
# stim1 <- n/apply(mat, 2, sum)
# oppure
# stim1<-1/apply(mat, 2, mean)
```

N.B.: Volendo ripetere la simulazione in diverse condizioni (diversi valori di n, m o λ, è possibile inserire questi comandi in una funzione parametrizzata da `n,m,lam`:

```
simula <- function(n=10,m=5000, lam=0.5)
{
mat <- matrix(rexp(n*m,lam), nrow=n)
somme <- apply(mat, 2, sum)
stim1 <- n/somme
stim2 <- (n-1)/somme
return(cbind(stim1,stim2))
}
```

Si noti che la funzione restituisce una matrice con due vettori colonna.

Calcoliamo le medie aritmetiche degli m valori simulati dei due stimatori. Come abbiamo visto, queste medie forniscono un'approssimazione al valore atteso (rispetto alla distribuzione campionaria) di ciascuno degli stimatori considerati. Ricordando che λ è stato posto pari a 0.5, scriviamo:

```
> mean(stim1)
[1] 0.5574001
> mean(stim2)
[1] 0.5016601
```

```
#oppure, usando la funzione costruita sopra (con i parametri di default):
#a<-simula()
#apply(a,2,mean)
```

Notiamo che in ambedue i casi la media della distribuzione empirica di simulazione dello stimatore (che approssima la distribuzione campionaria sempre meglio al crescere di m) è prossima al valore prefissato per λ (0.5). Naturalmente, a meno di non fissare il seme della simulazione con il comando `set.seed`, ripetendo la procedura per l'estrazione si ottiene un risultato differente. Notiamo che, pur essendo il secondo stimatore corretto, la media dei valori simulati per $\tilde{T}_n$ non è esattamente 0.5; questo è ragionevole in quanto stiamo approssimando una distribuzione continua con un istogramma basato su 5000 valori simulati. Si può verificare che al crescere di m tale valore si avvicina sempre di piú a λ. In ogni caso, anche usando $m = 5000$ replicazioni dell'operazione di campionamento, si verifica facilmente che il primo stimatore fornisce in media una sovrastima del valore del parametro. Effettivamente, per quanto dimostrato in precedenza, T_n ha una distorsione sempre positiva; il valore atteso teorico di T_n per $n = 9$ risulta pari a $10/18 = 0.556$, che nel nostro esempio viene approssimato con 0.5574001.

Ripetendo la procedura, ad esempio, con $n = 30$, possiamo anche verificare che la distorsione tende a zero al crescere di n:

```
>a.30 <- simula(n=30)
>apply(a.30,2,mean)
    stim1     stim2
0.5187508 0.5014591
```

Come si vede, la differenza tra le due medie è minore, segno che i valori attesi di cui esse costituiscono un'approssimazione sono più simili; in effetti, la distorsione del primo stimatore passa da $\lambda/9 = 0.06$ a $\lambda/29 = 0.02$ e per il primo stimatore il valore atteso (teorico) diviene pari a 0.517.

Possiamo infine visualizzare l'istogramma dei vettori simulati per avere un'idea della distribuzione campionaria dei due stimatori considerati:

```
par(mfrow=c(2,1)
hist(a[,1],prob=T,nclass=30, main='stim1')
abline(v=c(0.5, 10/18),col=2:3)
hist(a[,2],prob=T,nclass=30, main='stim2')
abline(v=0.5,col=2)
```