UNIVERSITA' DEGLI STUDI DI TRIESTE

Statistica e Informatica per l'azienda, la finanza e l'assicurazione

# ESERCIZIO LABORATORIO R


a cura di Giulio Fantuzzi

anno accademico 2021/2022

**Esercizio:** *sia* $(y_1, y_2, ..., y_n)$ *una sequenza di numeri pseudo casuali generati da* $Y \sim U(0,1)$

1. *Per n arbitrario, scrivere una funzione che restituisca la probabilità approssimata mediante simulazione della* $Pr(\max\{Y_1, Y_2, ..., Y_n\} > 0.5)$.

   Di seguito lo script del codice:

```
probabilita_approssimata<- function(n){
    nsim<- 10000
    count <- vector(mode = "numeric", length = nsim)
    for(i in 1:nsim){
        valori_generati<-runif(n, min = 0, max = 1)
        massimo=max(valori_generati)
        count[i]<- (massimo>0.5)
    }
    probabilita_empirica<- sum(count)/nsim
    return (probabilita_empirica)
}
```

**Figure 1:** Implementazione della funzione

***Commento del codice:*** come da specifiche, la funzione prende in input la numerosità campionaria ($n$) e restituisce in output la probabilità approssimata che il massimo campionario ($Y_{(n)}$) sia maggiore di 0.5. La funzione si basa sul principio del campionamento ripetuto, dunque sulla simulazione dello stesso processo in condizioni analoghe un gran numero di volte (*"nsim"*), rilevando ogni volta lo stato del sistema.

Tale struttura è stata realizzata mediante un ciclo for, in cui ad ogni iterazione:

- vengono generati $n$ dati campionari $(y_1, y_2, ..., y_n)$ da $Y \sim U(0,1)$;
- viene calcolato il massimo tra i valori generati (*massimo*);
- viene inserito nel vettore *count* il valore di verità (come 0 o 1) della proposizione logica $'massimo > 0.5'$. Dunque in *count* viene inserito 1 se $massimo > 0.5$, 0 altrimenti.

Terminato il ciclo, sommando i valori di *count* si ottiene il numero di volte che il massimo campionario è stato > 0.5 (casi favorevoli). Rapportando tale valore al numero di simulazioni effettuate, si ottiene una misura empirica della probabilità richiesta (*probabilita_empirica*), la quale verrà mandata in output. Inoltre, la chiamata della funzione sarà del tipo:

```
#Scelta della numerosità campionaria
n<- 5
#Chiamata della funzione
risultato_empirico= probabilita_approssimata(n)
```

**Figure 2:** Chiamata della funzione

*NOTA:* *nsim* è stato scelto in maniera arbitraria (10000 è sufficientemente alto).

Infine, per verificare l'attendibilità del risultato empirico, si è scelto di confrontarlo con il risultato esatto che ci si aspetta dalla teoria. Più precisamente, sapendo che il massimo campionario $Y_{(n)}$ si distribuisce come segue:

$$F_{Y_{(n)}}(y) = [F_Y(y)]^n, \quad \text{con} \quad F_Y(y) = y \quad [\text{essendo } Y \sim U(0,1)]$$

Risulta semplice ricavare il valore di $Pr(\max\{Y_1, Y_2, ..., Y_n\} > 0.5)$. In particolare:

$$Pr(\max\{Y_1, Y_2, ..., Y_n\} > 0.5) = Pr(Y_{(n)} > 0.5) = 1 - [F_Y(0.5)]^n$$

Ad esempio, chiamando la funzione con $n = 5$ si è ottenuto:

| Values | |
|---|---|
| n | 5 |
| risultato_empirico | 0.9685 |
| risultato_teorico | 0.96875 |

**Figure 3:** Risultati

*Commento:* il valore calcolato dalla funzione è pressoché uguale a quello teorico

2. *I valori* $\left(-\frac{\log(1-y_1)}{\theta}, -\frac{\log(1-y_2)}{\theta}, ..., -\frac{\log(1-y_n)}{\theta}\right)$ *sono determinazioni i.i.d da* $X \sim Esp(\theta)$ *?*

L'approccio adottato può essere riassunto nei seguenti punti:

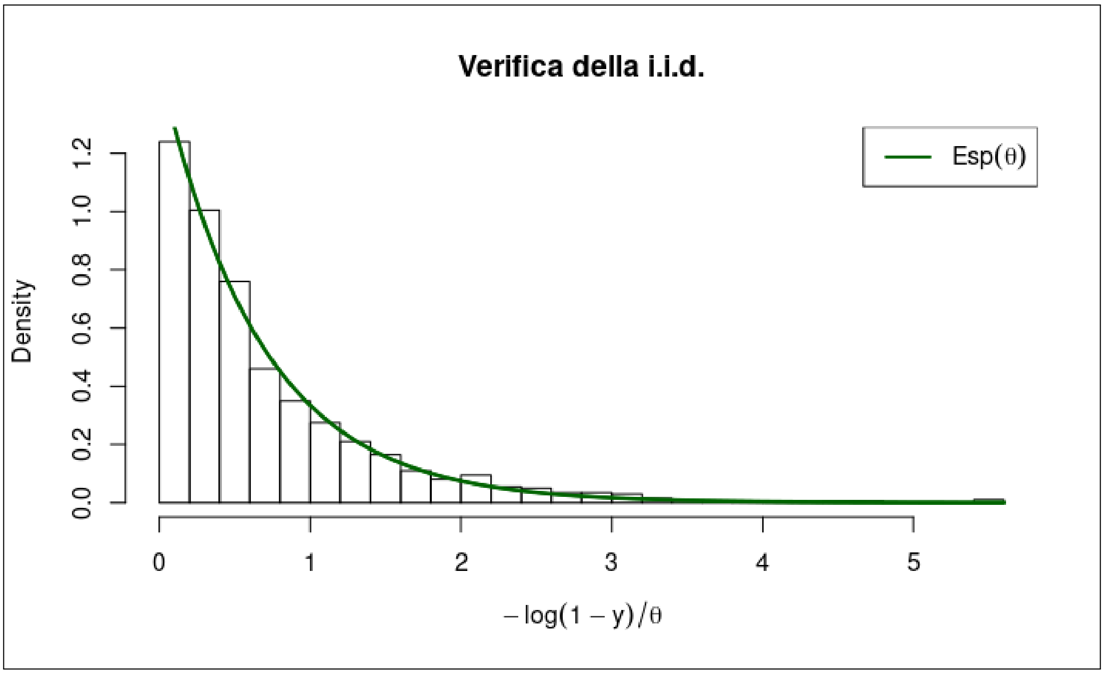
- è stato fissato in maniera arbitraria il parametro $\theta$ dell'Esponenziale;
- si sono generati $n$ valori da $Y \sim U(0,1)$, con $n$ arbitrario;
- per ogni $y$ generata si è calcolato il corrispondente $-\frac{\log(1-y)}{\theta}$;

```
#Fisso arbitrariamente il parametro dell'esponenziale
theta=1.5

#genero i valori da una U(0,1)
n<- 1000
values<-runif(n,min=0,max=1)

#calcolo i vari -log(1-Y) / theta
log_values<- vector(mode = "numeric", length = n)
for(i in 1:n){
    log_values[i]= -log(1-values[i])/theta
}
```

E' stato poi realizzato un istogramma per rappresentare la distribuzione empirica dei dati ottenuti, e ad esso è stata sovrapposta la densità di $X \sim Esp(\theta)$.

```
#istogramma rappresentativo dei risultati empirici
hist(log_values, prob = T, breaks = 25, plot=T,
     main='Verifica della i.i.d.', xlab=expression(- log(1-y)/theta))
#Sovrapponiamo la curva dell'esponenziale
curve(dexp(x,theta),from=0, add=TRUE, col='dark green', lwd=2)
legend("topright", legend= expression(Esp(theta)), lwd=2,
       lty=1,col="dark green")
```

In questo modo si è ottenuto il seguente grafico:




Il grafico suggerisce che le $-\frac{\log(1-y_1)}{\theta}, -\frac{\log(1-y_2)}{\theta}, \ldots, -\frac{\log(1-y_n)}{\theta}$ seguono la distribuzione di $X \sim Esp(\theta)$, ed effettivamente è così. Infatti, l' *i.i.d.* può essere dimostrata analiticamente:

- **INDIPENDENZA**$\rightarrow$ le $y_i$ sono state generate da $Y \sim U(0,1)$ in maniera casuale, dunque è ragionevole considerarle indipendenti. Dall'altro lato, le $-\frac{\log(1-y_i)}{\theta}$ sono funzioni monotone delle $y_i$ indipendenti (e di un parametro fissato, che però non influisce), dunque potranno essere considerate anch'esse indipendenti

- **IDENTICA DISTRIBUZIONE**$\rightarrow$ posta $X = -\frac{\log(1-Y)}{\theta}$ si ha:

$$F_X(x) = P(X \leq x) = P(-\frac{\log(1-Y)}{\theta} \leq x) = P(\log(1-Y) \geq -\theta x) =$$
$$= P(1-Y \geq e^{-\theta x}) = P(Y \leq 1-e^{-\theta x}) = F_Y(1-e^{-\theta x}) = 1-e^{-\theta x}$$

  ed $1-e^{-\theta x}$ è proprio la *f.d.r* di una $Esp(\theta)$

3. *Posto $n = 5$, come si può valutare con un piccolo studio di simulazione il comportamento di $T = -\sum_{i=1}^{5} \frac{\log(1-Y_i)}{\theta}$, dove $(Y_1, ..., Y_5)$ è un c.c. da $Y \sim U(0,1)$? Si confronti la distribuzione simulata con quella esatta. Cosa si osserva se n ha un valore molto più grande di 5?*

Dopo aver fissato in maniera arbitraria $\theta$, il comportamento di $T$ è stato analizzato mediante uno studio di simulazione Monte Carlo. Di seguito lo script del codice:

```
theta=1.5 #arbitrario
n<- 5 #come da consegna
nsim<- 10000 #arbitrario (purché sufficientemente alto)

#vettore che conterrà le determinazioni (empiriche) di T
T_empirica1 <- vector(mode = "numeric", length = nsim)
for(i in 1:nsim){
    #1)generiamo n valori da Y~U(01)
    values<-runif(n, min = 0, max = 1)
    #2)calcoliamone la funzione -log(1-Y) / theta
    log_values<- vector(mode = "numeric", length = n)
    #3)somma dei valori ottenuti
    for(j in 1:n){
        log_values[j]= -log(1-values[j])/theta
    }
    t<- sum(log_values)
    T_empirica1[i]<- t
}
```

***Commento del codice:*** innanzitutto si sono scelti in maniera arbitraria i valori del parametro $\theta$ e del numero di simulazioni da effettuare (*nsim*). L'idea di base è stata quella di implementare un ciclo for, in cui ad ogni iterazione:

- vengono generati 5 ($n$) valori casuali da $Y \sim U(0,1)$: $y_1, ..., y_5$;
- vengono calcolati i corrispettivi $-\frac{\log(1-y_1)}{\theta}, -\frac{\log(1-y_2)}{\theta}, ..., -\frac{\log(1-y_n)}{\theta}$;
- vengono sommati questi ultimi

Così facendo, ad ogni ciclo si otterrà una determinazione $t$ della variabile aleatoria $T$, la quale verrà inserita nel vettore *T_empirica1*. Al termine del ciclo, il vettore conterrà in tutto *nsim* determinazioni empiriche di $T$, grazie alle quali è possibile realizzare un istogramma rappresentativo della distribuzione empirica di $T$. Dall'altro lato, la distribuzione teorica si è potuta determinare facilmente a partire dalla dimostrazione precedente. In particolare:

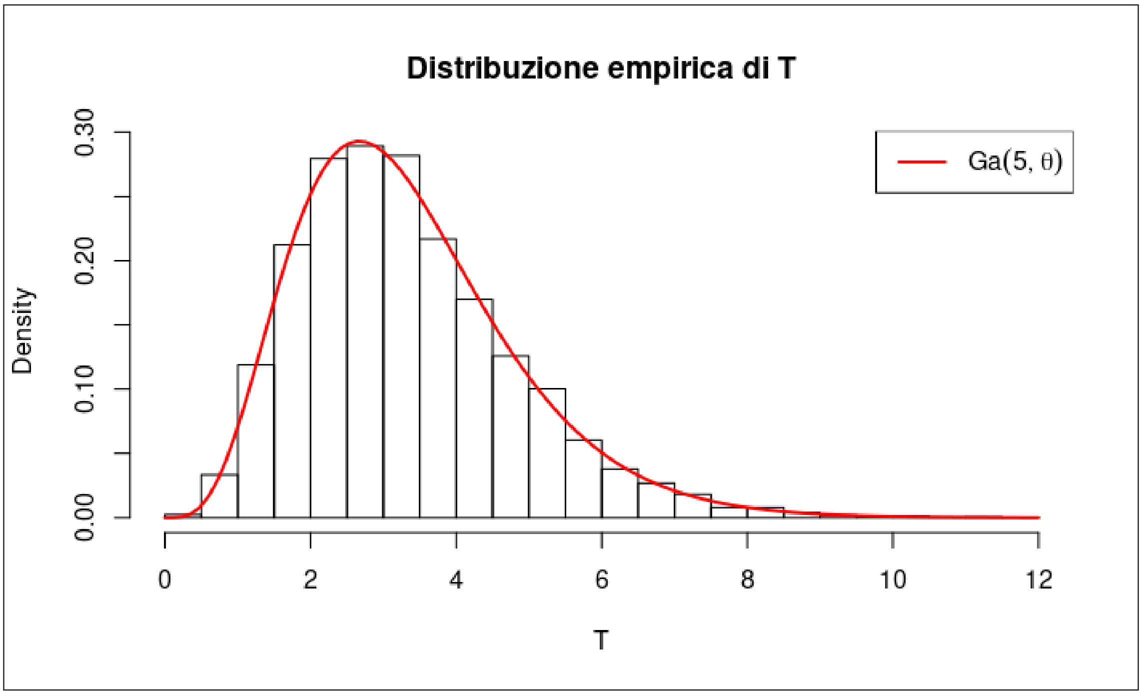
$$-\frac{\log(1-y)}{\theta} \sim Esp(\theta) \implies T = -\sum_{i=1}^{5} \frac{\log(1-Y_i)}{\theta} \sim Ga(5,\theta)$$

*NOTA:* sfruttando il fatto che la somma di Esponenziali è una Gamma

Come anticipato, è stato realizzato un istogramma rappresentativo della distribuzione empirica di $T$, e ad esso è stata sovrapposta la densità teorica di $T \sim Ga(5,\theta)$

```
#istogramma rappresentativo dei risultati empirici
hist(T_empirica1, prob = T, breaks = 25, plot=T,
     xlab='T', main='Distribuzione empirica di T')
#Controlliamo se la Gamma teorica approssima i dati empirici
curve(dgamma(x, n, theta), add=T, col=2, lwd=2)
legend("topright", legend= expression(Ga(5,theta)), lwd=2, lty=1,col="red")
```

In questo modo si è ottenuto il seguente grafico:



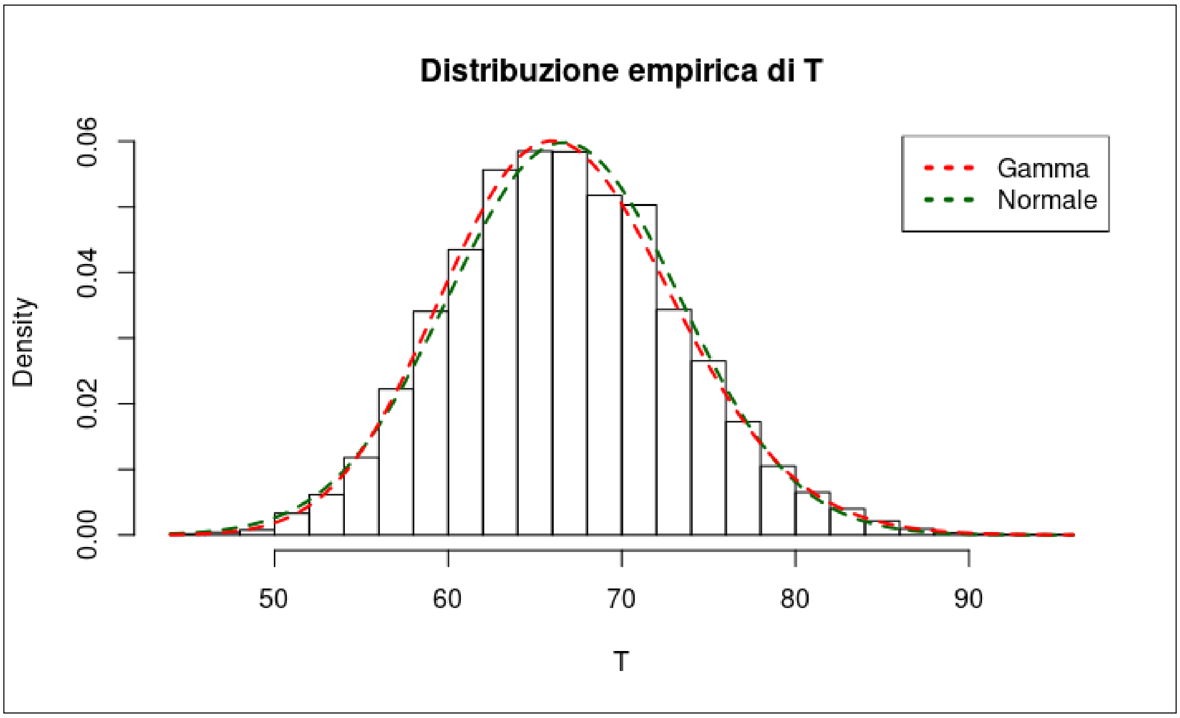

Infine è stato analizzato il caso con $n$ molto maggiore di 5. Lo studio di simulazione è stato realizzato in maniera analoga a prima, ed anche in questo caso è stata rappresentata la distribuzione empirica di $T$ a confronto con quella teorica ($T \sim Ga(n,\theta)$).

Tuttavia, rispetto al caso precedente è stato possibile fare delle ulteriori considerazioni basate sul teorema del limite centrale. In particolare, interpretando $T$ come somma di Esponenziali:

$$T \sim Ga(n,\theta) \stackrel{TLC}{\Longrightarrow} T \dot{\sim} N\left(\frac{n}{\theta}, \frac{n}{\theta^2}\right)$$

```
#istogramma rappresentativo dei risultati empirici:
hist(T_empirica2, prob = T, breaks = 25, plot=T, xlab='T',
     main='Distribuzione empirica di T')

#Distribuzioni teoriche:
mu= n/theta
sigma= sqrt(n/(theta^2))
curve(dnorm(x, mu, sigma), add=T, col='dark green', lwd=2, lty=2)
curve(dgamma(x, n, theta), add=T, col='red', lwd=2, lty=2)
legend("topright", legend= c("Gamma", "Normale"), lwd=3,
       lty=3,col=c("red", "dark green"))
```

Combinando tutto in un unico grafico:




Notiamo che la distribuzione empirica è approssimata con grandissima precisione sia dalla Gamma fornita dalla teoria, sia dalla Normale fornita dal teorema del limite centrale.

Il grafico mostra anche come la curva della Gamma riesca a descrivere i dati con una precisione leggermente maggiore rispetto alla Normale. Questo risultato è sensato: la Gamma è una distribuzione teorica esatta; la Normale è una distribuzione approssimata.