Tor Vergata
Laurea Magistrale in Informatica

Inferenza Statistica e Teoria dell'Informazione

Stick-Breaking Problem

Leonardo Tamiano

August 15, 2020

Abstract

Il seguente documento contiene la trattazione di un problema probabilistico da noi chiamato *stick-breaking problem*. Il documento è strutturato come segue: nella sezione (1) è presente una breve descrizione del problema analizzato; le sezioni (2) e (3) invece contengono l'analisi nel problema, prima per una istanza specifica e successivamente per quella generale. Infine, nella sezione (4) è presente una breve discussione su un possibile utilizzo delle variabili aleatorie utilizzate per la trattazione del problema.

Contents

1 Descrizione del Problema

Il problema analizzato in questa sede può essere descritto informalmente come segue: si prende un segmento di una certa lunghezza, lo si spezza in un certo numero di pezzettini utilizzando n variabili aleatorie, e si studia la v.a. che rappresenta la lunghezza del pezzettino più corto. Per andare a costruire un modello effettivo che potrà poi essere analizzato bisogna quindi definire la lunghezza del segmento e le distribuzioni delle variabili aleatorie che vengono utilizzate per rompere il segmento, oltre ad eventuali ipotesi di indipendenza.

Nel nostro caso particolare abbiamo che il segmento preso in considerazione è l'intervallo unitario $[0, 1]$, mentre le variabili aleatorie utilizzate per spezzare il segmento sono i.i.d. uniformi in $[0, 1]$. In [1] sono riportati dei risultati teorici anche nei casi di v.a. i.i.d. con distribuzione esponenziale $Exp(\lambda)$ o con una generica funzione di ripartizione $F_X(t)$. Procediamo quindi con la formalizzazione del nostro modello.

Siano $X_1, X_2, \ldots, X_n$ i.i.d. con $X_i \sim U[0, 1]$ e consideriamo la loro statistica d'ordine $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ per definire le seguenti v.a.

$$\begin{aligned} L_1 &:= X_{(1)} - X_{(0)} \\ L_2 &:= X_{(2)} - X_{(1)} \\ &\vdots \\ L_i &:= X_{(i)} - X_{(i-1)} \\ &\vdots \\ L_n &:= X_{(n)} - X_{(n-1)} \\ L_{n+1} &:= X_{(n+1)} - X_{(n)} \end{aligned}$$

con $X_{(0)} := 0$ e $X_{(n+1)} := 1$ in modo da avere $L_1 = X_{(1)}$ e $L_{n+1} = 1 - X_{(n)}$. Consideriamo infine la statistica d'ordine

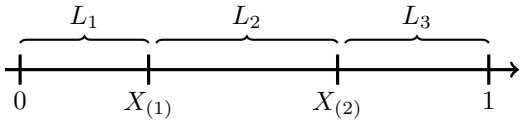
$$L_{(1)} \leq L_{(2)} \leq \ldots \leq L_{(n)} \leq L_{(n+1)}$$

e concentriamoci sulla v.a. $L_{(1)}$, che rappresenta la lunghezza del pezzettino più corto. Tra le varie domande che ci possiamo porre su $L_{(1)}$ siamo interessati in particolare a calcolare la sua distribuzione di probabilità, ovvero il valore di $P(L_{(1)} \leq t)$, oltre al suo valore atteso $\mathbb{E}[L_{(1)}]$.

L'analisi del problema è stata effettuata prima considerando il caso specifico in cui $n = 2$, e poi considerando come le risposte ottenute nel caso già studiato cambiano al crescere di n.

2 Trattazione Caso Specifico ($n = 2$)

Andiamo adesso ad analizzare il caso in cui $n = 2$. In questo caso abbiamo due uniformi $X_1, X_2 \sim U[0, 1]$ che ci spezzano l'intervallo unitario in tre pezzi, come mostrato in figura



2.1 Simulazioni

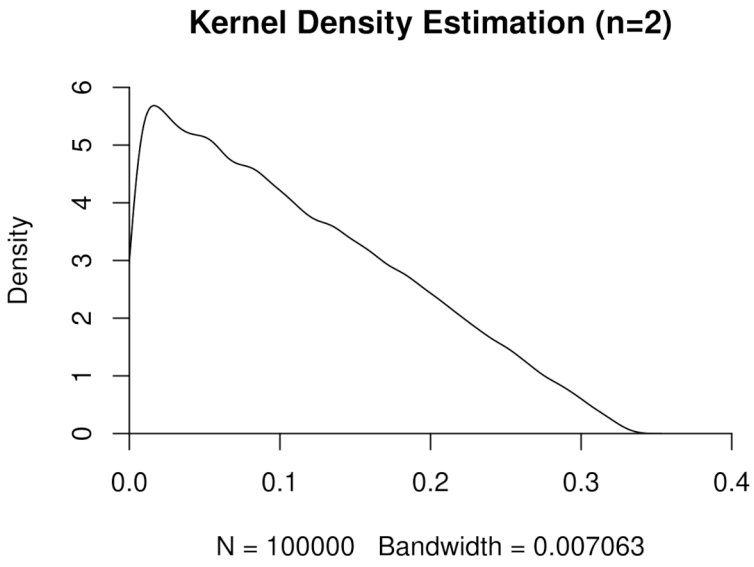
Per cercare di capire la forma della densità di $L_{(1)}$ si è simulato il procedimento di generare due uniformi e calcolare la lunghezza del pezzettino più piccolo utilizzando il seguente codice R

```r
m <- 100000
vec <- vector("double", length=m)

for (i in 1:m) {
  vars <- runif(2, min=0, max=1)

  p1 <- min(vars)
  p2 <- max(vars)

  l1 <- p1
  l2 <- p2 - p1
  l3 <- 1 - p2

  vec[i] <- min(l1,l2,l3)
}
```

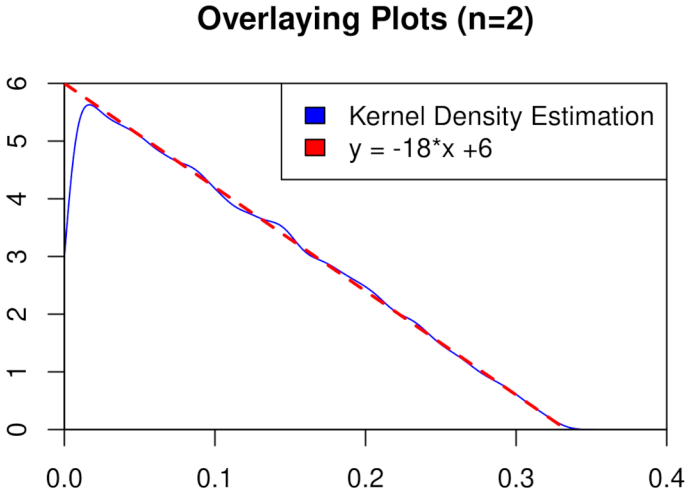
Il seguente codice R invece utilizza la funzione **density()**, che calcola la kernel density estimation del sample generato dalla simulazione precedente.

```r
plot(density(vec),
     main="Kernel Density Estimation (n=2)",
     xlim = c(0, 0.4),
     ylim = c(0, 6),
     yaxs="i",
     xaxs="i"
)
```

Una volta eseguito, il codice genera il seguente grafico



come è possibile vedere, la stima della densità sembra assomigliare ad una retta della forma $y = m \cdot x + h$, con $m < 0$. Dopo vari tentativi nel cercare di capire il valore corretto per i cofficienti m e h si è trovata la seguente retta



2.2 Teoria

Procediamo adesso calcolando l'espressione teorica della distribuzione di $L_{(1)}$ nel caso in cui $n = 2$. Piuttosto che calcolare $P(L_{(1)} \leq t)$, il nostro piano sarà quello di calcolare $P(L_{(1)} > t)$ per poi utilizzare una nota proprietà che ci dice che $P(L_{(1)} \leq t) = 1 - P(L_{(1)} > t)$.

Iniziamo notando che non può essere $L_{(1)} > 1/3$, in quanto altrimenti la somma dei tre pezzettini farebbe $L_1 + L_2 + L_3 > 3 \cdot 1/3 > 1$. Dunque per $t > 1/3$ si ha $P(L_{(1)} > t) = 0$.

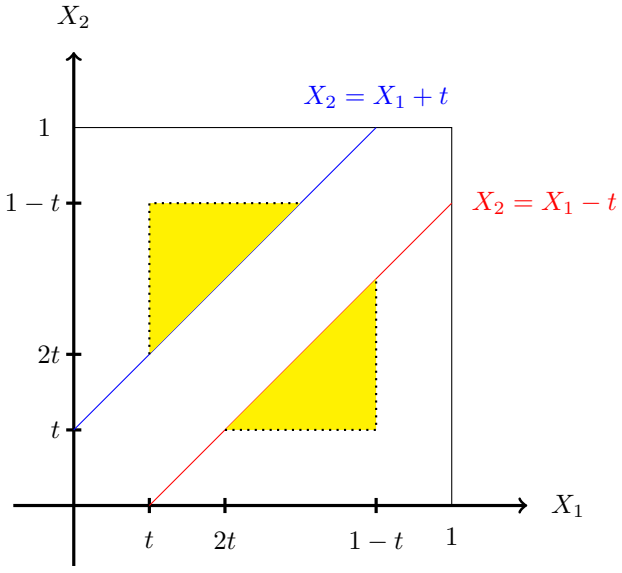
Consideriamo adesso il caso $t \in [0, 1/3]$. Dire che il pezzo più piccolo è maggiore di t equivale a dire che tutti i pezzi sono più grandi di t. In altre parole, troviamo la seguente equivalenza di eventi

$$\begin{aligned}\{L_{(1)} > t\} &= \{L_1 > t,\ L_2 > t,\ L_3 > t\} \\ &= \{X_{(1)} > t,\ X_{(2)} - X_{(1)} > t,\ 1 - X_{(2)} > t\}\end{aligned}$$

Dato che non sappiamo quale delle due variabili sarà quella più grande, rispetto alle variabili originali X_1 e X_2 l'evento $\{L_{(1)} > t\}$ è verificato se e solo se vengono rispettate le seguenti tre condizioni

- $t < X_1 < 1 - t$
- $t < X_2 < 1 - t$
- $X_2 - X_1 > t$ se $X_2 > X_1$, altrimenti $X_1 - X_2 > t$.

Riportando la zona del piano cartesiano che soddisfa tutte e tre le condizioni troviamo la seguente figura



Calcolare la nostra probabilità di interesse si riduce quindi al calcolo dei seguenti integrali

$$P(L_{(1)} > t) = \int_{x_1=2t}^{1-t} \int_{x_2=t}^{x_1-t} f(x_1, x_2)\ dx_2\, dx_1 \ + \int_{x_1=t}^{1-2t} \int_{x_2=x_1+t}^{1-t} f(x_1, x_2)\ dx_2\, dx_1$$

Ricordiamo a questo punto il fatto che X_1 e X_2 sono variabili indipendenti, e quindi per $t \in [0, 1]$ si ha che

$$f(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = 1$$

Da questo, e dal fatto che le due zone gialle del grafico rappresentano due triangoli che hanno la stessa area, e che se uniti formano un quadrato di lato $(1-3t)$, ci possiamo semplificare il calcolo degli integrali. Possiamo infatti calcolare direttamente il volume di interesse, che rappresenta proprio il valore di probabilità dell'evento $\{L_{(1)} > t\}$ e che è pari a $(1-3t)^2$, in quanto l'area di un quadrato è il quadrato di uno dei suoi lati e l'altezza associata ad ogni punto dell'aria da $f(x_1, x_2)$ è pari a 1.

Riassumendo, abbiamo trovato che la funzione di ripartizione di $L_{(1)}$ è data da

$$F_{L_{(1)}}(t) = P(L_{(1)} \leq t) = 1 - P(L_{(1)} > t) = 1 - \big(1-3t\big)^2 = 6t - 9t^2$$

andandola a derivare troviamo quindi la densità di $L_{(1)}$

$$f_{L_{(1)}}(t) = \frac{dF_{L_{(1)}}(t)}{dt} = -18t + 6$$

che è proprio la retta $m \cdot x + h$ con $m = -18$ e $h = 6$ che avevamo stimato tramite la simulazione del processo.

Da questo inoltre segue che, nel caso $n = 2$ il valore atteso della lunghezza del pezzo più corto è uguale a $1/9$. Infatti,

$$\begin{aligned}
\mathbb{E}[L_{(1)}] = \int_0^{1/3} t \cdot (-18t + 6)\ dt = \int_0^{1/3} -18t^2 + 6t\ dt &= \left.\frac{-18}{3}t^3 + \frac{6}{2}t^2\right|_{t=0}^{t=1/3} \\
&= \frac{18}{3} \cdot \frac{1}{27} + \frac{6}{2} \cdot \frac{1}{9} \\
&= \frac{1}{9}
\end{aligned}$$

3 Trattazione Caso Generale ($n > 2$)

Andiamo adesso a vedere come variano i risultati ottenuti prima quando $n > 2$. Il caso generale verrà trattato come si è trattato il caso specifico: prima si effettueranno delle simulazioni e successivamente si proverà a trovare dei risultati teorici.

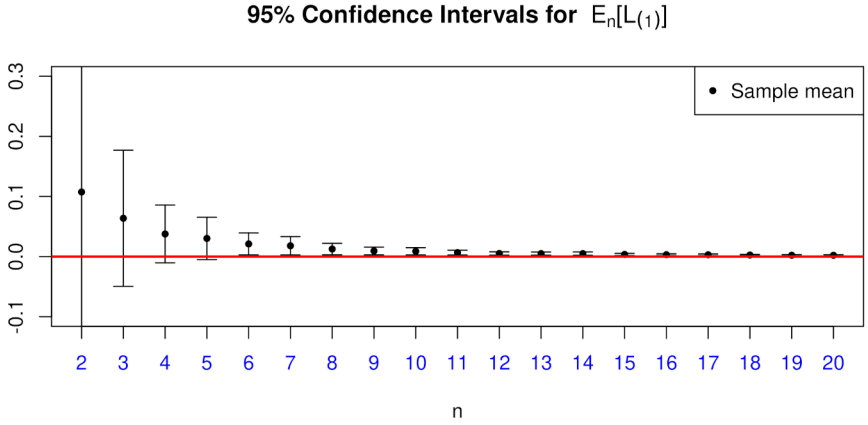
3.1 Simulazioni

Per simulare i casi in cui $n > 2$ è stato modificato e generalizzato il codice visto prima, ottenendo il seguente risultato

```
1  simulate_shortest_stick  <- function(n, m) {
2      shortest_intervals <- vector("double", length=m)
3      intervals <- vector("double", length=n+1)
4
5      for (i in 1:m) {
6          vars <- runif(n, min=0, max=1)
7
8          ## sort the vector vars
9          sorted_vars  <- sort(vars, decreasing = FALSE)
10
11         ## compute the lengths for all intervals
12         intervals[1] = sorted_vars[1]
13         intervals[n+1] = 1 - sorted_vars[n]
14
15         for (j in 2:n){
16             intervals[j] = sorted_vars[j] - sorted_vars[j-1]
17         }
18
19         ## sort the length of the intervals
20         sorted_intervals  <- sort(intervals, decreasing = FALSE)
21
22         ## memorize shortest interval
23         shortest_intervals[i]  <- sorted_intervals[1]
24     }
25
26     return (shortest_intervals)
27 }
```

Andiamo adesso ad analizzare la relazione tra il size dell'istanza n e il valore atteso della variabile aleatoria $L_{(1)}$. A tale fine indichiamo con $\mathbb{E}_n[L_{(1)}]$ il valore atteso di $L_{(1)}$ quando si utilizzano n v.a. per spezzare l'intervallo unitario in $n+1$ pezzettini.

Osserviamo che all'aumentare del numero di pezzi con cui spezziamo l'intevallo unitario $[0,1]$ ci aspettiamo che la lunghezza attesa del pezzo più piccolo diventi sempre più piccola. Infatti, andando a plottare degli intervalli di confidenza del 95% per il valore atteso $\mathbb{E}_n[L_{(1)}]$, al variare del size dell'istanza $n \in \{1, 2, \ldots, 20\}$ otteniamo il seguente risultato



dove gli intervalli di confidenza sono stati calcolati utilizzando il seguente codice

```
compute_confidence_interval  <- function(n) {
    iterations <- 100
    vec <- simulate_shortest_stick(n, iterations)

    # compute sample mean, standard deviation, and standard error.
    s_mn  <- mean(vec)
    s_stdev <- sd(vec)
    s_standard_error  <- s_stdev / sqrt(n)

    # compute distance to move from the mean of a t-student
    # distribution from both sides to have 95% probability.
    z  <- -qt(0.025, df=n-1)

    l_limit  <- s_mn - z * s_standard_error
    u_limit  <- s_mn + z * s_standard_error

    return(c(s_mn, l_limit, u_limit))
}
```

Notiamo dal grafico appena visto che $\mathbb{E}_n[L_{(1)}]$ e n sembrano relazionati tra loro da una *power law*, ovvero da una relazione della forma $\mathbb{E}_n[L_{(1)}] = a \cdot n^k$. Al fine di verificare l'esistenza effettiva di questa relazione sono stati calcolati i seguenti log-log plots

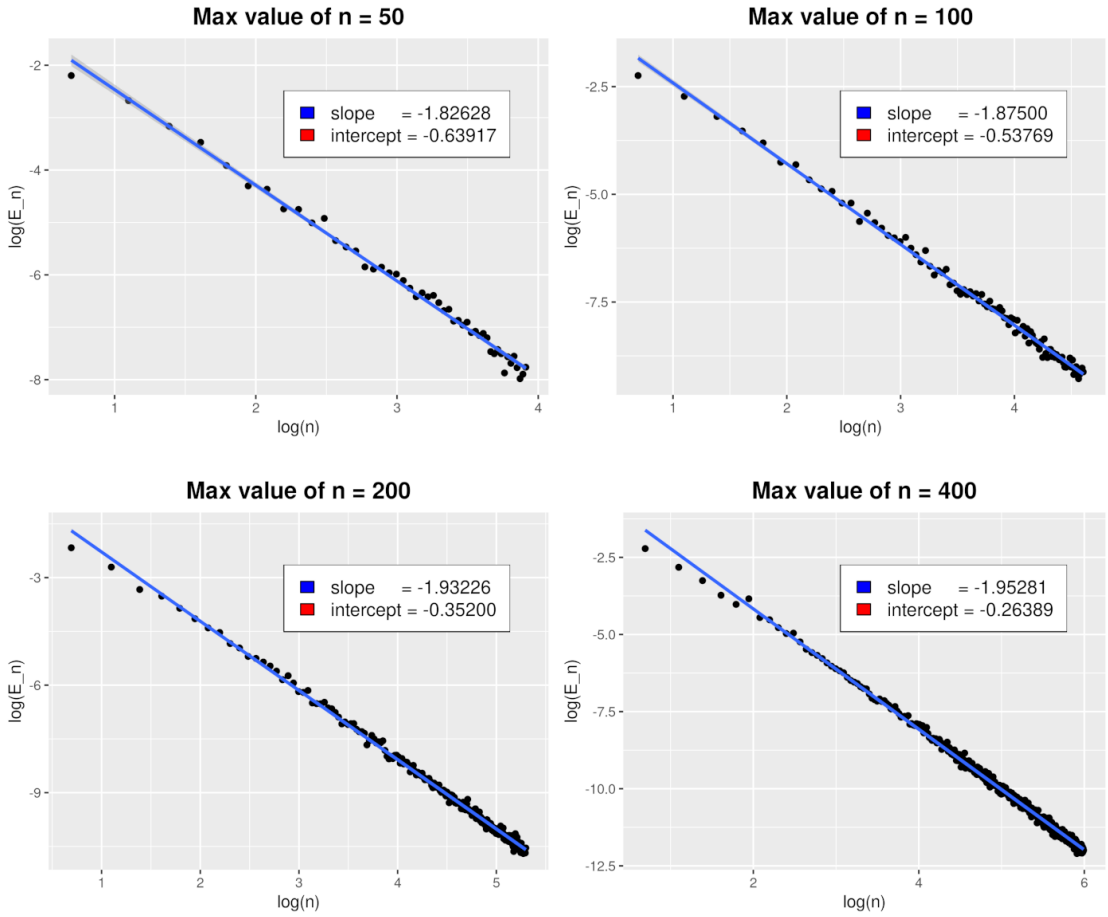
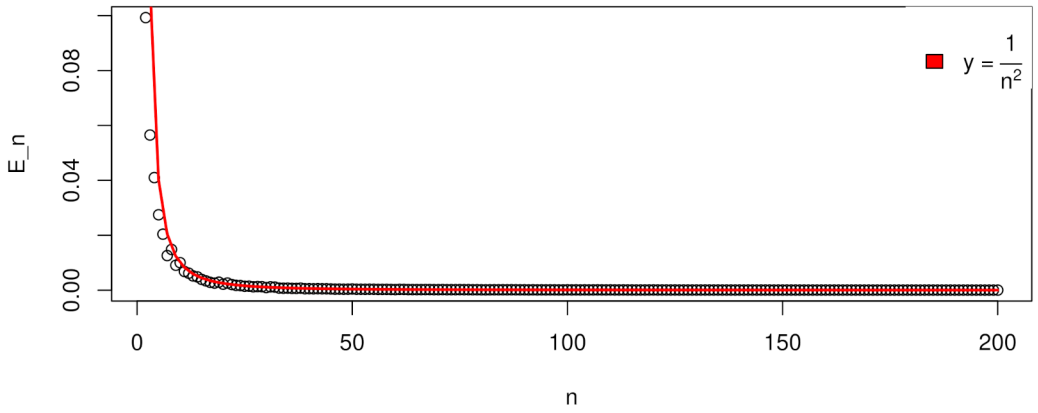


Figure 1: Log-Log plots tra $\mathbb{E}_n[L_{(1)}]$ e n

Dove la retta è stata calcolata utilizzando la funzione **lm**() offerta dal linguaggio R. È possibile vedere come il coefficiente angolare della retta tende al valore -2 all'aumentare del valore di n. Questo risultato suggerisce che $k = -2$, ovvero che la decrescita di $\mathbb{E}_n[L_{(1)}]$ sembra essere dell'ordine di $O\left(\frac{1}{n^2}\right)$.

Andando a imporre la curva $f(n) = 1/n^2$ ai valori di media ottenuti simulando il processo troviamo il seguente grafico



che sembra coincidere con i risultati ottenuti.

3.2 Teoria

Andiamo adesso a calcolare la distribuzione teorica di $L_{(1)}$ nel caso in cui utilizziamo n v.a. $X_1, X_2, ..., X_n$ i.i.d. con $X_i \sim U[0,1]$ per spezzare l'intervallo $[0,1]$.

Iniziamo notando che in [2] si dimostra che se $Y_1, Y_2, ..., Y_n$ è un sample i.i.d. con pdf $f(y)$, allora la densità congiunta della statistica d'ordine $\underline{Y} = (Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)})$ è data da

$$f_{\underline{Y}}(y_1, y_2, \ldots, y_n) = \begin{cases} n! \cdot f(y_1)f(y_2)\ldots f(y_n) & , \ y_1 \leq y_2 \leq \ldots \leq y_n \\ 0 & , \ \text{altrimenti} \end{cases}$$

nel nostro caso, dato che $X_i \sim U[0,1]$, abbiamo che $f(x_i) = 1$ per $i = 1, \ldots, n$, e quindi

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n) = \begin{cases} n! & , \ x_1 \leq x_2 \leq \ldots \leq x_n \\ 0 & , \ \text{altrimenti} \end{cases}$$

Continuando, ricordiamo che avevamo definito gli intervalli di interesse $L_1, L_2, \ldots, L_n, L_{n+1}$ come segue

$$L_i := X_{(i)} - X_{(i-1)} \quad , \quad \forall i = 1, \ldots, n+1$$

con $X_{(0)} := 0$ e $X_{(n+1)} := 1$. Notando che la lunghezza dell'ultimo intervallo L_{n+1} dipende da quella dei precedenti n intervalli, ci interessa calcolare la densità congiunta di $\underline{L} = (L_1, L_2, \ldots, L_n)$.

Al fine di calcolare la densità congiunta di $\underline{L}$ utilizziamo la seguente formula di trasformazione

$$f_{\underline{L}}(L_1, L_2, \ldots, L_n) = f_{\underline{X}}(X_{(1)}, X_{(2)}, ..., X_{(n)}) \cdot \frac{1}{\det\left(\dfrac{d\underline{L}}{d\underline{X}}\right)}$$

e notiamo che il dominio in cui è definita $f_{\underline{L}}$ è l'insieme degli $(L_1, L_2, ..., L_n) \in \mathbb{R}^n$ tali che

(i) $0 \leq L_i \leq 1$, per $i = 1, ..., n$

(ii) $\sum_{i=1}^{n} L_i \leq 1$

Procedendo con i calcoli troviamo

$$\det\left(\frac{d\underline{L}}{d\underline{X}}\right) = \det\begin{pmatrix} \frac{dL_1}{dX_{(1)}} & \frac{dL_1}{dX_{(2)}} & \frac{dL_1}{dX_{(3)}} & \cdots & \frac{dL_1}{dX_{(n)}} \\ \frac{dL_2}{dX_{(1)}} & \frac{dL_2}{dX_{(2)}} & \frac{dL_2}{dX_{(3)}} & \cdots & \frac{dL_2}{dX_{(n)}} \\ \vdots & \vdots & & \ddots & \\ \frac{dL_n}{dX_{(1)}} & \frac{dL_n}{dX_{(2)}} & \frac{dL_n}{dX_{(3)}} & \cdots & \frac{dL_n}{dX_{(n)}} \end{pmatrix} = \det\begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & & \ddots & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix} = 1$$

in quanto il determinante di una matrice triangolare è il prodotto degli elementi sulla diagonale. Mettendo tutto assieme otteniamo la seguente espressione

$$f_{\underline{L}}(L_1, L_2, \ldots, L_n) = \begin{cases} n! & , \quad 0 \leq L_i \leq 1, \ \sum_{i=1}^{n} L_i \leq 1 \\ 0 & , \quad \text{altrimenti} \end{cases}$$

3.2.1 Distribuzione di $L_{(1)}$

Procediamo ora con il calcolo della distribuzione di $L_{(1)}$. Iniziamo notando la seguente equivalenza di eventi

$$(L_{(1)} > t) = (L_1 > t, \ L_2 > t, \ \ldots, \ L_n > t, \ L_{n+1} > t)$$

Un possibile approccio è dunque il calcolo del seguente integrale multiplo

$$\int \cdots \int_D n! \ \ dL_1 \ dL_2 \ dL_3 \ \ \ldots \ dL_n$$

dove $D \subseteq \mathbb{R}^n$ è il dominio formato dalle n -uple $(L_1, L_2, ..., L_n)$ tali che

1. $L_i > t$, per $i = 1, ..., n$
2. $\sum_{i=1}^{n} L_i \leq 1 - x$

Per semplificare i calcoli però presentiamo una argomentazione ripresa da [3].

Proposizione 3.1. Sia $X_1, ..., X_n$ un campione i.i.d. con $X_i \sim U[0,1]$, e consideriamo la statistica ordine $X_{(1)}, X_{(2)}, ..., X_{(n)}$ per definire $L_i := X_{(i)} - X_{(i-1)}$, per $i = 1, ..., n+1$, con $X_{(0)} := 0$, $X_{(n+1)} := 1$. Allora si ha che

$$P(L_1 > t,\ L_2 > t,\ \ldots,\ L_n > t\ , L_{n+1} > t) = \begin{cases} (1-(n+1)\cdot t)^n & ,\ \ t \in [0, \frac{1}{n+1}] \\ 0 & ,\ \ t > \frac{1}{n+1} \end{cases}$$

Proof. Iniziamo notando che se $t > 1/(n+1)$ allora non è possibile avere che tutti gli L_i siano più grandi di t, in quanto altrimenti la loro somma sarebbe più grande di 1. Sia quindi $t \in [0, 1/(n+1)]$.

L'evento $\{L_1 > t\} = \{X_{(1)} > t\}$ è equivalente a dire che tutte le v.a. $X_1, X_2, ..., X_n$ sono $> t$, e quindi che assumono valore nell'intervallo $(t, 1)$.
Per simmetria poi, dato vero l'evento $\{L_1 > t\}$, abbiamo che la probabilità che la distanza tra il primo e il secondo punto sia maggiore di t, equivale alla probabilità che la distanza dal punto t al primo punto sia maggiore di t. In formula,

$$P(L_2 > t \mid L_1 > t) = P(L_1 > 2t \mid L_1 > t)$$

Ma allora troviamo

$$\begin{aligned} P(L_1 > t\ ,\ L_2 > t) &= P(L_2 > t \mid L_1 > t) \cdot P(L_1 > t) \\ &= P(L_1 > 2t \mid L_1 > t) \cdot P(L_1 > t) \\ &= \frac{P(L_1 > 2t\ ,\ L_1 > t)}{P(L_1 > t)} \cdot P(L_1 > t) \\ &= P(L_1 > 2t) \end{aligned}$$

Utilizzando le ipotesi di i.i.d. del campione e il fatto che $L_1 = X_{(1)}$, troviamo

$$P(X_{(1)} > 2t) = P(X_1 > 2t, X_2 > 2t, \ldots, X_n > 2t) = \prod_{i=1}^{n} P(X_i > 2t) = \prod_{i=1}^{n} (1-2t) = (1-2t)^n$$

Abbiamo quindi dimostrato che $P(L_2 > t, L_1 > t) = (1-2t)^n$. Iterando il ragionamento appena fatto è possibile far vedere che

$$\forall r = 1, ..., n+1: \quad P(L_1 > t, L_2 > t, ..., L_r > t) = (1-rt)^n$$

in particolare quindi per $r = n+1$ si ottiene

$$P(L_1 > t, L_2 > t, ..., L_{n+1} > t) = (1-(n+1)t)^n$$

□

Utilizzando la proposizione appena dimostrata possiamo facilmente calcolare la funzione di ripartizione di $L_{(1)}$ come segue: se $t > 1/(n+1)$ allora $F_{L_{(1)}}(t) = 1$, se invece $t \in [0, 1/(n+1)]$, allora

$$\begin{aligned} F_{L_{(1)}}(t) &= P(L_{(1)} \leq t) \\ &= 1 - P(L_{(1)} > t) \\ &= 1 - P(L_1 > t, L_2 > t, ..., L_{n+1} > t) \\ &= 1 - \Big(1 - (n+1)t\Big)^n \end{aligned}$$

la densità di $L_{(1)}$ per $t \in [0, 1/(n+1)]$ è quindi data da

$$\begin{aligned} f_{L_{(1)}}(t) &= \frac{d}{dt} F_{L_{(1)}}(t) \\ &= \frac{d}{dt}\Big[1 - \Big(1 - (n+1)t\Big)^n\Big] \\ &= (n+1) \cdot n \cdot (1 - (n+1)t)^{n-1} \end{aligned}$$

3.2.2 Media di $L_{(1)}$

Utilizzando l'espressione per la densità appena trovata possiamo calcolare la media di $L_{(1)}$ risolvendo il seguente integrale

$$\mathbb{E}[L_{(1)}] = \int_0^{1/(n+1)} t \cdot f_{L_{(1)}}(t)\ dt = \int_0^{1/(n+1)} t \cdot (n+1) \cdot n \cdot (1 - (n+1)t)^{n-1}\ dt$$

Al fine di risolvere l'integrale iniziamo integrando per parti

$$\begin{aligned} \int_0^{1/(n+1)} \underbrace{t}_{f(t)} \cdot \underbrace{(n+1) \cdot n \cdot (1 - (n+1)t)^{n-1}}_{g'(t)}\ dt &= f(t) \cdot g(t) \Big|_{t=0}^{t=1/(n+1)} - \int_0^{1/(n+1)} f'(t) \cdot g(t)\ dt \\ &= t \cdot -(1 - (n+1)t)^n \Big|_{t=0}^{t=1/(n+1)} - \int_0^{1/(n+1)} 1 \cdot -(1 - (n+1)t)^n\ dt \\ &= 0 + \int_0^{1/(n+1)} (1 - (n+1)t)^n\ dt \end{aligned}$$

andando ad effettuare la sostituzione $u = (1 - (n+1)t)$ abbiamo che

$$\begin{cases} \dfrac{du}{dt} = -(n+1) \cdot 1 \iff dt = -\dfrac{du}{n+1} \\ t = 0 \implies u = 1 \\ t = \dfrac{1}{n+1} \implies u = 0 \end{cases}$$

sostituendo all'integrale otteniamo quindi

$$\begin{aligned}\int_1^0 u^n \cdot -\frac{1}{n+1}\, du &= -\frac{1}{n+1} \cdot \Big[\frac{u^{n+1}}{n+1}\Big] \Big|_{t=1}^{t=0} \\ &= -\frac{1}{n+1} \cdot \Big[\frac{0^{n+1}}{n+1} - \frac{1^{n+1}}{n+1}\Big] \\ &= \frac{1}{(n+1)^2}\end{aligned}$$

abbiamo quindi dimostrato che

$$\mathbb{E}[L_{(1)}] = \frac{1}{(n+1)^2}$$

il che conferma i dati teorici ottenuti nel caso $n = 2$ e i dati ottenuti dalla simulazione nella sezione 3.1.

4 Applicazione alla Teoria delle Decisioni

Le variabili aleatorie $L_1, L_2, ..., L_{n+1}$ trattate in precedenza possono essere utilizzate per definire dei test statistici di tipo *goodness of fit* in cui abbiamo un campione i.i.d. $X_1, X_2, ..., X_n$ con funzione di ripartizione $F_X(\cdot)$ e vogliamo capire se $F_X(\cdot) = F(\cdot)$ per una data $F(\cdot)$. Utilizzando il linguaggio statistico, siamo interessati al seguente test di ipotesi

$$H_0 : \ F_x = F \quad , \quad H_A : \ F_x \neq F$$

Un approccio standard per trattare questo problema impiega l'utilizzo della trasformazione $U = F(X)$. È infatti possibile dimostrare che sotto H_0 si ha $F(X) \sim U[0, 1]$. Dunque, se H_0 è vero, la trasformazione $U = F(X)$ trasforma il sample $X_1, \ldots, X_n$ in un sample di v.a. uniformi in $[0, 1]$.
A partire da queste variabili uniformi $U_1, U_2, ..., U_n$, con $U_i = F(X_i)$, possiamo considerare le variabili aleatorie $L_i := U_{(i)} - U_{(i-1)}$, per $i = 1, \ldots, n + 1$, con $U_{(0)} := 0$, $U_{(n+1)} := 1$ per definire delle particolari statistiche che possono poi essere utilizzate all'interno di test di ipotesi.
Un esempio di test che impiegano l'utilizzo delle v.a. L_i può essere trovato in [4], dove Greenwood ha definito la seguente statistica come possibile strumento per capire meglio le dinamiche di un contagio

$$G_n := \sum_{i=1}^{n+1} (L_i)^2$$

Più alto è il valore di G_n e più gli eventi presi in considerazione sono raggruppati tra loro nel tempo; viceversa, più basso è il valore di G_n e più gli eventi sono distribuiti in modo uniforme nel tempo.
È possibile trovare in [1] altre statistiche che fanno utilizzo delle v.a. L_i.

References

[1] R. Pyke, "Spacings," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 27, no. 3, pp. 395–449, 1965.

[2] *Order Statistics*. Available at `https://www.colorado.edu/amath/sites/default/files/attached-files/order_stats.pdf`.

[3] L. Holst, "On the lengths of the pieces of a stick broken at random," *Journal of Applied Probability*, vol. 17, no. 3, p. 623–634, 1980.

[4] M. Greenwood, "The statistical study of infectious diseases," *Journal of the Royal Statistical Society*, vol. 109, no. 2, pp. 85–110, 1946.