Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Simulazione e studio del modello Broken stick per l'analisi di sequenze geniche

Relatore:
Prof. Daniel Remondini

Correlatore:
Dott.ssa Alessandra Merlotti

Presentata da:
Andrea Edera

Anno Accademico 2019/2020

Abstract

Alcuni risultati ottenuti studiando le proprietà statistiche dei dinucleotidi all'interno del DNA umano, mostrano che l'andamento delle distribuzioni delle interdistanze dei dinucleotidi TA è ben descritto da una legge di potenza. É stato ideato un modello in grado di generare un andamento di questo tipo e che potrebbe rendere conto del meccanismo generativo delle distribuzioni osservate all'interno della sequenza del DNA umano. Questo risulta essere una variante del modello Broken stick.
Scopo di questa tesi è confrontare il modello Broken stick con la sua variante, che chiameremo Broken stick con memoria, in modo da valutarne analogie e differenze.
Si sono implementati i due modelli ed è stata condotta un'analisi variando il numero di tagli iniziali e il numero di iterazioni, tenendo fisso il valore della probabilità di taglio.
L'implementazione del modello ha messo in luce i limiti computazionali del programma utilizzato, Matlab, e mostrato che non è semplice fittare le distribuzioni ottenute.
Tagliando casualmente i segmenti si raggiunge un limite superato il quale il calcolatore non riesce più a distinguere gli estremi del segmento, generando così segmenti di lunghezza nulla. In questo modo non è possibile realizzare il modello così come è stato pensato, cioè in un dominio continuo in cui è possibile tagliare il segmento infinite volte. Di conseguenza la scelta delle condizioni iniziali non può essere arbitraria. Si sono studiate generazioni che hanno prodotto circa 1'000'000 di segmenti e si è visto che l'andamento descritto dai due modelli, a parità di condizioni iniziali, risulta differente. Il tipo di taglio iniziale determina una traslazione della distribuzione in scala log-log.
In conclusione il modello Broken stick con memoria genera un andamento riconducibile ad una legge di potenza lungo tutto il range in cui viene rappresentata la distribuzione, a differenza di quello semplice che presenta effetti di cutoff sia in intervalli piccoli che grandi.

Indice

Introduzione

Lo studio fatto da Paci et al. nell'articolo *Characterization of DNA methylation as a function of biological complexity via dinucleotide inter-distances* [6], mostra le differenze tra le distribuzioni delle interdistanze dei dinucleotidi CG e quelli non-CG. In particolare viene mostrato che nel DNA umano l'andamento dei CG è di tipo esponenziale decrescente, mentre quello non-CG è governato da una legge di potenza. Questa differenza riflette il ruolo peculiare dei dinucleotidi CG all'interno dei mammiferi come siti preferenziali in cui avviene la metilazione.
Uno studio successivo [5], analizzando le distribuzioni di 18 organismi, 9 mammiferi e 9 non-mammiferi, mostra che le distribuzioni dei CG nei mammiferi differiscono da quelle non-CG. Questa è una caratteristica propria di questi organismi infatti classi differenti non mostrano questi andamenti. Tutti gli organismi analizzati sono accomunati dal fatto che le distribuzioni non-CG nelle code hanno un andamento descritto da una legge di potenza [5, 6]. In particolare sembra che solo il dinucleotide TA mostri un andamento completamente descritto da una legge di potenza [4].

Partendo dai risultati precedenti, è stato proposto un modello in grado di generare lo stesso andamento e che risulta essere una variante del modello Broken stick.
Scopo di questa tesi è confrontare il modello Broken stick con la variante, che chiameremo Broken stick con memoria, per valutarne analogie e differenze. Si sono implementati i due modelli e si sono confrontate le generazioni ottenute a patire da condizioni iniziali differenti. Entrambi dipendono da 3 paremetri: numero di tagli iniziali m, probabilità con cui ogni segmento verrà tagliato p e numero di iterazioni a cui è soggetto il segmento iniziale n. Si è tenuta costante la probabilità di inserzione p e si sono fatti variare gli altri parametri in modo da ottenere circa 1'000'000 di segmenti a generazione.
É stato testato come varia la distribuzione al variare dei segmenti prodotti, al variare dei tagli fatti inizialmente e a seconda di come è stato fatto il taglio. I tagli sono stati fatti in 2 modi nel primo caso il segmento iniziale è stato tagliato in modo da ottenere m segmenti con la stessa lunghezza, mentre il secondo tipo di taglio è stato fatto in punti casuali in modo tale da avere una distribuzione normale delle lunghezze. Si sono comparate le distribuzioni ottenute con lo stesso modello partendo da condizioni iniziali differenti e quelle ottenute con modelli diversi a partire dalle stesse condizioni iniziali.

Capitolo 1

DNA e distribuzioni delle distanze fra dinucleotidi

In questo capitolo si forniscono le nozioni base di biologia per comprendere il contesto in cui può essere utile il modello analizzato.

1.1 Struttura e funzionamento del DNA

L'acido desossiribonucleioco, DNA, l'acido ribonucleico, RNA, e le proteine costituiscono la base degli esseri viventi. Nel DNA è contenuto tutto il materiale genetico che conserva l'informazione ereditaria, permette di trasmetterla e di esprimerla.
Il DNA e l'RNA sono costituiti da acidi nucleici, i quali sono polimeri a loro volta costituiti da monomeri chiamati nucleotidi.
Ogni nucleotide è composto da un gruppo fosfato, uno zucchero pentoso (desossiribosio nel DNA, ribosio nell'RNA) e una base azotata. Ci sono due famiglie di basi azotate: pirimidine e purine. La prima include citosina (C), timina (T) e uracile (U) (che è presente nell'RNA al posto di T) e sono caratterizzate da un anello composto da sei atomi di carbonio e azoto. Le purine sono molecole più grandi formate da 2 anelli di 5 o 6 elementi, ne fanno parte adenina (A) e guanina (G). Il polimero è composto da nucleotidi legati tra loro attraverso legami fosfodiesterici. Questo schema si ripete ed ha una direzionalità intrinseca; da una parte il gruppo fosfato è legato al carbonio 5 $'$, mentre al carbonio 3 $'$ è legato il gruppo idrossilico.

L'RNA solitamente è composto da una sola catena di polinucleotidi mentre il DNA da 2. I filamenti che compongono quest'ultimo formano una spirale attorno ad un asse immaginario componendo una struttura 3-dimensionale a doppia elica. I filamenti sono antiparalleli, con le colonne zucchero-fosfato disposte all'esterno della doppia elica, e corrono in direzione 5 $'$- 3 $'$ una opposta all'altra. Le basi azotate sono accoppiate

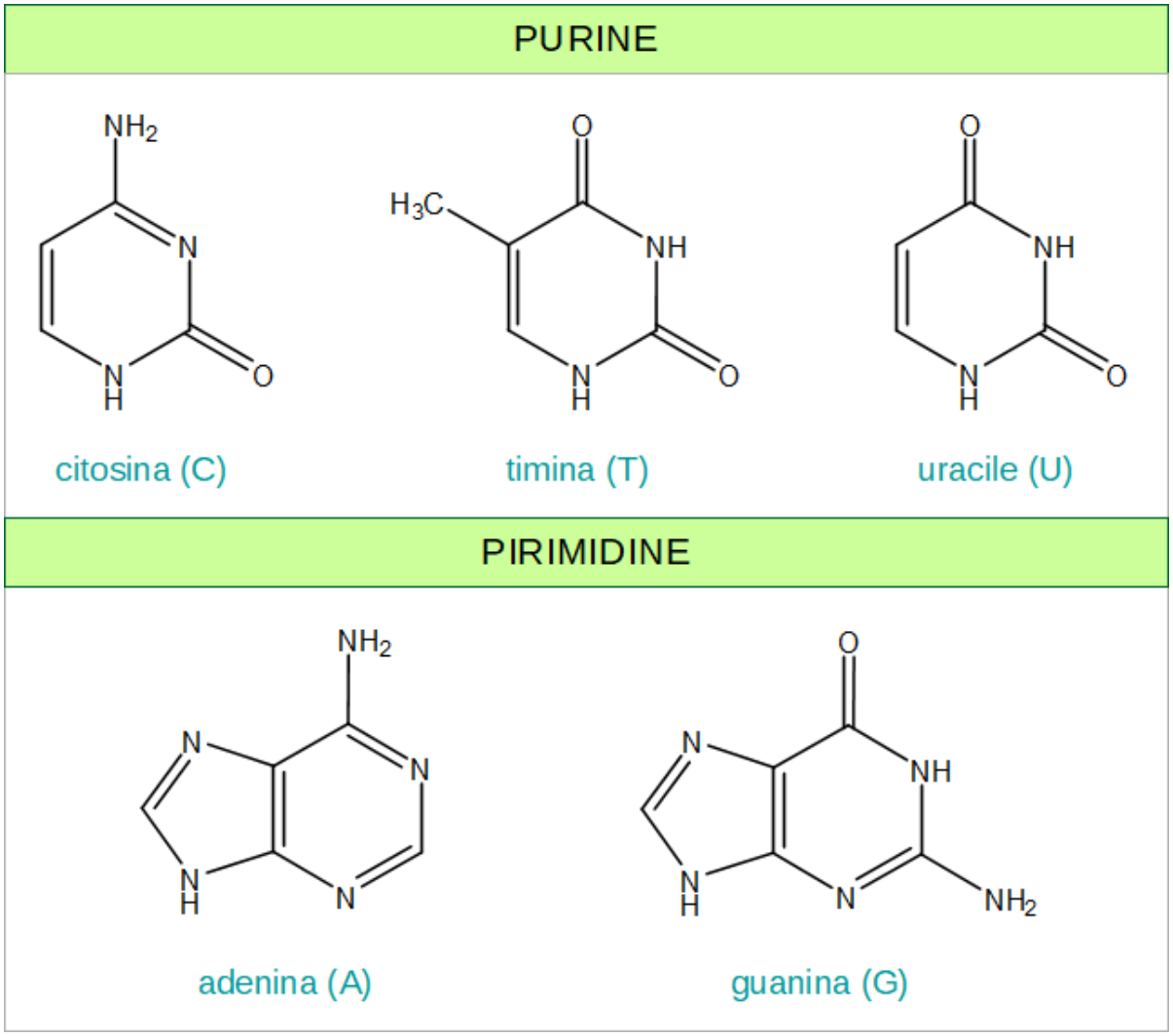


Figura 1.1: Struttura delle basi azotate. Immagine presa da [2].

all'interno dell'elica attraverso legami ad idrogeno che tengono insieme i due filamenti. Solo certe basi della doppia elica sono compatibili con altre: l'adenina si accoppia sempre con la timina attraverso 2 legami ad idrogeno e la citosina si accoppia sempre con la guanina attraverso 3 legami ad idrogeno. I due filamenti sono complementari: se leggiamo la sequenza delle basi lungo una catena della doppia elica, sappiamo la sequenza lungo l'altra catena. Questa unicità del DNA permette la creazione di 2 copie identiche di ogni molecola di DNA in una cellula che si prepara a dividersi, generando due cellule identiche alla madre.

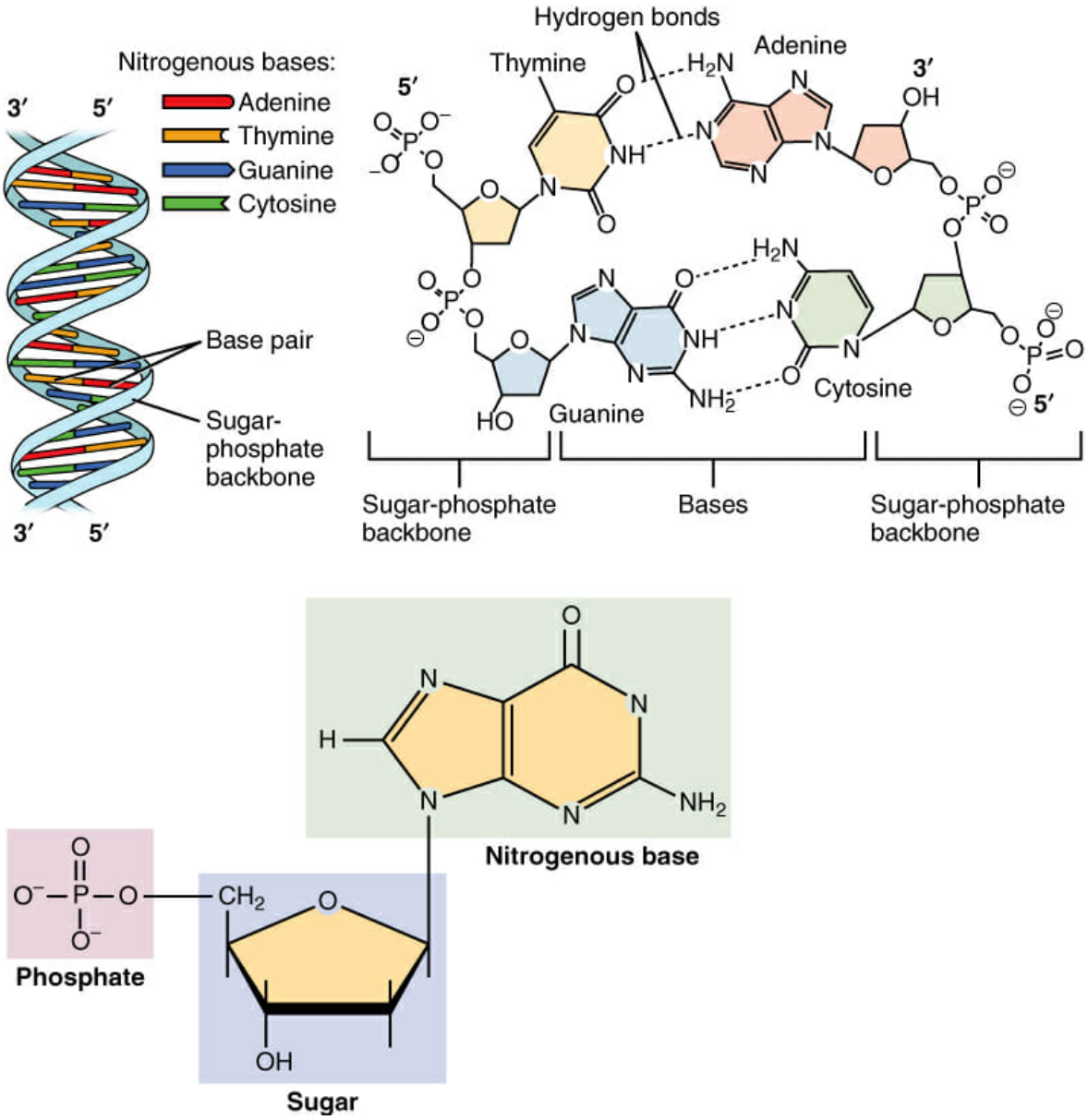


Figura 1.2: Struttura chimica del DNA e dei nucleotidi. Immagine presa da [1].

1.2 Distribuzioni delle distanze fra dinucleotidi

Negli ultimi 40 anni per studiare le sequenze di DNA si è sviluppato un approccio statistico. Questo è utile perchè può rivelare proprietà funzionali e strutturali biologicamente molto rilevanti. Un metodo di studio si basa sulla misura delle interdistanze tra i dinucleotidi, ovvero si considera un dinucleotide e si misura quante basi ci sono tra una

ripetizione e l'altra. Si considerano dincleotidi perchè svolgono un ruolo essenziale nella biologia del genoma e quindi possono essere la chiave per comprendere il DNA più a fondo. Le distribuzioni che si ottengono risultano caratteristiche per organismi differenti in particolare lo studio del genoma umano [7] mostra una forte caratterizzazione della distribuzione dei dinucletidi CG. Questi hanno un andamento molto diverso rispetto a quello delle altre coppie come si può vedere in Figura 1.3.

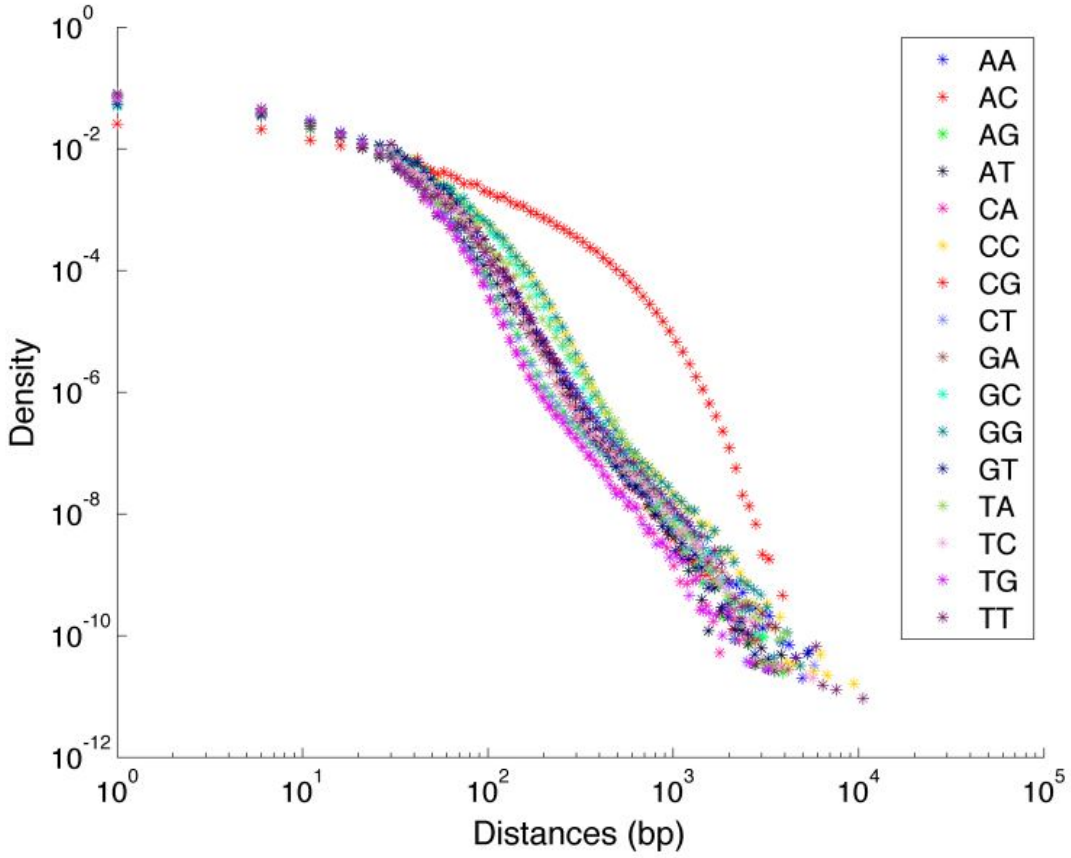


Figura 1.3: Rappresentazione in scala log-log delle distribuzioni delle interdistanze dei dinucleotidi del genoma umano. Si utilizza come unità misura il bp (numero di basi azotate). Figura presa da [7].

Dall'analisi dei dati risulta che le code delle distribuzioni delle interdistanze dei dinucleotidi CG hanno un decadimento esponenziale, mentre quelle non-CG mostrano un decadimento simile ad una legge di potenza. Questa differenza è stata motivata considerando il ruolo essenziale ricoperto dai dinucleotidi CG all'interno del genoma dei mammiferi. Questi sono siti preferenziali in cui avviene la metilazione [6], risulta infatti che la percentuale di CG metilati all'interno del DNA umano si aggiri tra il 70% e l'80%. Il lavoro [5] studiando il genoma di 18 organismi, 9 mammiferi e 9 non-mammiferi, ha

mostrato che la maggior parte delle distribuzioni delle interdistanze CG all'interno dei mammiferi seguono una distribuzione gamma. Inoltre la distribuzione delle interdistanze CG differisce da tutte le altre non-CG, mentre questo non accade per i non-mammiferi. Tutte le distribuzioni non-CG degli organismi considerati nelle code hanno un andamento descritto da una legge di potenza $p(\tau) \propto \tau^{-\alpha}$. Si nota anche che le distribuzioni CG e non-CG differiscono in quanto il numero dei dinucleotidi CG è minore di quelli non-CG. Per questo si sono domandati se, partendo da una distribuzione specifica non-CG, fosse possibile ottenere una distribuzione simile a quella CG cambiando semplicemente un certo numero d di dinucleotidi da un tipo a un altro in maniera causuale. Hanno osservato che solo le distribuzioni AT e TA diventano simili a quelle CG. É importante specificare che solo queste nell'*Homo sapiens* differiscono dalle altre in quanto sono le uniche a presentare un profilo simile a quello ottenuto con una legge di potenza pura. Dal lavoro [4], che studia la distribuzione della coppia CG e TA all'interno del genoma umano, emerge che la distribuzione della coppia TA è descritta interamente da una legge di potenza, a differenza delle altre che hanno un andamento di questo tipo solo nelle code.

Capitolo 2

Modello Broken stick

In questo capitolo si introduce il modello Broken stick con memoria, pensato come possibile modello generativo delle distribuzioni osservate per le coppie non-CG del genoma. Si descrive come sono stati implementati il programma che simula il modello Broken stick semplice e quello con memoria. Si riportano le analisi fatte e i risultati ottenuti.

2.1 Modello Broken stick semplice

Come si legge nell'articolo [3], il modello Broken stick è stato proposto da MacArthur, in ambito ecologico, per spiegare come le specie si dividono le risorse disponibili. Vi sono diversi approcci con cui considerare lo stesso modello, uno consiste nel misurare la sovrapposizione delle nicchie[1], mentre un approccio alternativo ne studia le distanze. Data una risorsa continua e una funzione di utilizzo della risorsa da parte delle specie associate, si possono calcolare le differenze tra i valori acquisiti da queste funzioni e la distribuzione delle differenze che ci si attende a partire da alcune assunzioni. Ad esempio, considerando l'abbondanza delle specie rispetto al modello Broken stick, si possono scegliere $n-1$ punti, in modo casuale, lungo un segmento di lunghezza c (dove n rappresenta il numero di specie e c il numero totale di individui dell'insieme) in corrispondenza dei quali rompere il bastoncino, e successivamente calcolare la distribuzione delle lunghezze dei bastoncini prodotti. Analogamente, é possibile pensare lo stesso modello applicato al DNA. In questo caso il bastoncino rappresenta una sequenza di DNA e i punti in cui si spezza il segmento corrispondono all'inserimento del dinucletide considerato all'interno del filamento di DNA.

[1] Con nicchia in ecologia si intende il ruolo e lo spazio che una specie occupa all'interno dell'ecosistema in cui vive.

2.2 Modello generativo proposto

Si considera una x_j distribuzione indipendente di variabili casuali tra $[0, 1]$ con distribuzione

$$p(x) = (b+1)x^b \quad b \in [-1, 0] \tag{2.1}$$

Considerando la sequenza di variabili casuali

$$y_k = \prod_{j=1}^{k} x_j \tag{2.2}$$

dove x_j è una variabile casuale indipendente distribuita con probabilità uniforme nell'intervallo $[0, 1]$. Si calcola la distribuzione di probabilità delle variabili y_k procedendo per induzione: $y_1 = x_1$, per y_2 si utilizza il cambio di variabili

$$u = x_2 \quad v = x_2 x_1$$

così

$$dx_1 dx_2 = \frac{dudv}{u}$$

con il dominio

$$u \in (0, 1] \quad v \in (0, u]$$

Andando a valutare

$$p(y_2) = -\int_0^1 \frac{du}{u} \int_0^u dv \delta(v - y_2) v^b (b+1)^2 = \int_0^1 dv \int_v^1 \frac{du}{u} \delta(v - y_2) v^b (b+1)^2$$

così

$$p(y_2) = -(b+1)^2 \int_0^1 dv \delta(v - y_2) v^b \ln v = -(b+1)^2 y_2^b \ln(y_2) \tag{2.3}$$

Introducendo

$$p(y_k) = (b+1)^k \frac{y_k^b (-\ln y_k)^{k-1}}{(k-1)!} \tag{2.4}$$

che produce per induzione

$$p(y_{k+1}) = -(b+1)^{k+1} \int_0^1 dv \int_v^1 \frac{du}{u} \delta(v - y_{k+1}) \frac{\ln(v/u)^{k-1}(v/u)^b}{(k-1)!} u^b$$

dove si utilizzano le variabili

$$u = x_{k+1} \quad v = x_{k+1} y_k$$

Un calcolo esplicito da

$$p(y_{k+1}) = -(b+1)^{k+1} \int_0^1 dv \int_{\ln v}^0 dz \delta(v - y_{k+1}) \frac{(-\ln(v) + z)^{k-1} v^b}{(k-1)!}$$

dove si definisce

$$z = \ln(u)$$

Così si ottiene

$$p(y_{k+1}) = (b+1)^{k+1} \frac{y_{k+1}^b (-\ln y_{k+1})^k}{k!} \tag{2.5}$$

che è consistente con l'assunzione fatta eq. (2.4). Considerando la variabile y che prende valori da alcuni y_k, la distribuzione risulta (alcune scelte di k sono possibili ma si deve introdurre il peso w^k, che definisce la relativa numerosità della realizzazione della variabile y_k: se $w > 1$ il peso della variabile y_k aumenta esponenzialmente con k)

$$p(y) \propto y^b \sum_{k \geq 1} (b+1)^{k-1} w^{k-1} \frac{(-\ln y_k)^{k-1}}{(k-1)!} \propto y^{b - w(b+1)} \qquad y \in (0, 1]$$

La distribuzione che ne risulta è una legge di potenza

$$p(y) \propto y^{-\alpha} \tag{2.6}$$

dove

$$\alpha = w(b+1) - b \quad b \in (-1, 0]$$

Se si considera la distribuzione uniforme $b = 0$, il modello può essere interpretato con un modello broken stick locale. Si divide un lungo segmento in N sottosegmenti realizzati dalla uscita della varibile x_1 N-volte. Si assume che ogni segmento sopravviva con una probabilità $(1-p)$ e che venga spezzato con una probabilità p attraverso la realizzazione della variabile $y_2 = x_1 x_2$ (x_1 è associato al segmento considerato). Il processo è iterato n-volte e ad ogni iterazione ogni segmento può avere una rottura con una probabilità p (ad ogni rottura si realizza una variabile y_k dove k è il numero totale di rotture). Si stima la numerosità relativa dei segmenti considerandone la probabilità: il numero di segmenti che realizzano la variabile y_k, sono soggetti a $k-1$ rotture con probabilità $P = p^{k-1}(1-p)^{n-k+1}$. Dove l'abbondanza relativa dei segmenti $y_{k+1}(n_{k+1})$ rispetto al segmento y_k è

$$\frac{n_{k+1}}{n_k} = \frac{2p}{(1-p)} \tag{2.7}$$

É presente un fattore 2 perchè ad ogni rottura di un segmento si generano 2 sottosegmenti. Dopo n step, la lunghezza delle distribuzioni è data da

$$p(y) \propto \sum_{k\geq 1}^{n} \frac{(-2^{k-1}p^{k-1}\ln y)^{k-1}}{(k-1)!(1-p)^{k-1}} \simeq y^{-\alpha} \tag{2.8}$$

Dove

$$\alpha = \frac{2p}{1-p} \tag{2.9}$$

Questa formula permette di ricavare il valore dell'esponente α conoscendo p e viceversa conoscendo α si può ricavare p.

2.3 Implementazione dei modelli

Si sono implementati e poi testati due modelli, il Broken stick semplice e quello con memoria.

2.3.1 Broken stick semplice

Il modello Broken Stick semplice, partendo da un segmento iniziale, genera un certo numero di segmenti finali utilizzando il seguente schema:

1. si taglia un segmento m volte;
2. ogni segmento ha una certa probabilità p di essere soggetto ad un taglio ulteriore in un punto casuale.

Il secondo punto si ripete per un numero n fissato. Ne risulta che la lunghezza dei segmenti generati dipende dai 3 parametri: tagli iniziali m, probabilità di taglio p e numero di iterazioni n.

Processo

É possibile rappresentare il modello attraverso due generazioni di numeri random. La prima determina quali segmenti saranno soggetti ad inserzione. Se si estrae un numero $x \in [0, 1]$ e vi è una certa probabilità $p \in [0, 1]$ che il segmento venga tagliato; allora verranno tagliati i segmenti che soddisfano la condizione $x \leq p$. Facendo una seconda estrazione di numeri random si possono generare i punti in cui avverrà il taglio per ogni segmento.

Implementazione

L'implementazione del processo è stata realizzata nel seguente modo. Il segmento iniziale $[0, 1]$ viene suddiviso in m segmenti. Questi sono rappresentati da un vettore contenente $m + 1$ elementi avente il primo elemento $= 0$ e l'ultimo $= 1$.

Questo vettore è soggetto al seguente processo n volte. Si genera un vettore di numeri random con dimensione pari al numero di segmenti. Attraverso la funzione `find` si trovano i segmenti che saranno soggetti ad inserzione, cioè quelli con $x \leq p$. Si genera quindi un altro vettore contenente tutti i numeri random $y \in]0, 1[$ che rappresentano i punti in cui viene applicato il taglio. Questi valori devono essere adattati al segmento considerato, per questo si crea il vettore $v = \{0, y, 1\}$. Utilizzando la funzione `rescale(v, a, b)` è possibile riscalare il vettore passatogli come argomento, v, nell'intervallo specificato

$[a, b]$. Dove a e b in questo caso corrispondono agli estremi del segmento da tagliare. Così si ottiene il punto in cui rompere il segmento e questo valore viene inserito in fondo al vettore contenente tutti gli estremi dei segmenti. Dopo aver analizzato tutti i segmenti si ordina il vettore risultante.

2.3.2 Broken stick con memoria

Questo modello è come il Broken Stick semplice, però ha una condizione aggiuntiva: se un segmento non è tagliato una volta, non sarà più tagliato.

Implementazione

Per implementare questo modello al posto dei vettori si sono utilizzate le matrici. In particolare dopo il primo taglio da cui si ottengono m segmenti, si costruisce una matrice $2 \times m$. In questo modo ogni colonna è costituita da 2 elementi che rappresentano gli estremi del segmento.

Come nell'implementazione del Broken Stick si genera un vettore di numeri casuali per capire se avverrà il taglio. Si applica la funzione `find` che determina quali segmenti saranno soggetti ad inserzione e per questi viene generato un vettore che ne contiene il punto di inserzione. Si applica un ciclo su tutti gli elementi: quelli da tagliare vengono tagliati e memorizzati in una matrice, mentre quelli che non sono soggetti ad inserzione vengono memorizzati in una matrice differente. Si utilizzano 2 matrici in modo da distinguere quali saranno soggetti ad ulteriori tagli e quali no. La funzione utilizzata per tagliare il segmento in questo caso è `rescale`.
Dopo il taglio si controlla che entrambi i segmenti generati non siano nulli. Se uno dei due segmenti prodotti risulta nullo non viene applicato il taglio e si memorizza il segmento iniziale.

2.3.3 Parametri

Entambi i modelli hanno 3 parametri indipendenti:

- m numero di tagli iniziali;
- n numero di iterazioni a cui sono soggetti i segmenti;
- p probabilità che il segmento venga tagliato.

Si è fissato il parametro $p = 0.72$ e sono state fatte diverse generazioni facendo variare i parametri m ed n. Questi ultimi sono stati scelti in modo da generare 1'000, 10'000, 1'000'000 di segmenti.

Generazioni effettuate			
Broken stick	Tagli inziali m	iterazioni n	segmenti prodotti tagli uguali-random
Semplice	200	3	1'038 - 1'012
Memoria	200	5	1'890 - 1'865
Semplice	200	7	9'206 - 9'023
Memoria	200	9	8'515 - 8'841
Semplice	500	14	975'873 - 991'074
Memoria	500	20	1'217'237 - 1'155'017
Semplice	5'000	10	1'135'373 - 1'134'668
Memoria	5'000	13	968'230 - 931'258
Semplice	50'000	6	1'291'398 - 1'297'998
Memoria	50'000	8	1'473'775 - 1'481'410

Tabella 2.1: Parametri con cui sono state fatte le generazioni e segmenti ottenuti per ogni generazione.

Queste generazioni sono state fatte per testare l'andamento delle distribuzioni e si è deciso di analizzare le generazioni aventi il campione più numeroso, circa 1'000'000 di dati.

2.3.4 Problema dei segmenti nulli

Generando un grande numero di segmenti emerge un problema, la produzione di segmenti di lunghezza nulla. Come riportato nelle sezioni 2.3.1 e 2.3.2 il punto in cui avviene il taglio è generato casualemente attraverso la funzione `rand` nell'intervallo $]0, 1[$ e attraverso la funzione `reascale` si trova il punto corrispondente del segmento considerato. In questo modo, riducendosi ad ogni iterazione la lunghezza del segmento, diventa sempre minore la distanza tra l'estremo e il punto in cui si è fatta l'inserzione. Quello che accade è che il punto generato si può trovare tra gli estremi considerati ma il calcolatore non ne riconosce la differenza. Occorre quindi fare i conti con un limite computazione che non permette di distinguere due numeri la cui differenza si trova oltre la 16-esima cifra significativa. Non avendo senso la produzione di segmenti nulli il problema è stato ovviato introducendo un controllo post-taglio. Dopo aver fatto il taglio si controlla che i segmenti prodotti abbiano lunghezza $\neq 0$; in caso contrario non si opera l'inserzione e si prende nota del taglio non avvenuto. In questo modo si evita la produzione di segmenti con lunghezza nulla e si può vedere quante inserzioni non sono state fatte per ogni ciclo. In tabella 2.2 sono riportati i dati relativi all'ultima iterazione.

Segmenti non tagliati durante l'ultima iterazione			
Broken stick	Tagli inziali m	iterazione n	uguali-random
Semplice	500	14	19 - 34
Memoria	500	20	3'327 - 4'106
Semplice	5'000	10	5 - 2
Memoria	5'000	13	93 - 150
Semplice	50'000	6	1 - 1
Memoria	50'000	8	8 - 16

Tabella 2.2: Parametri con cui sono state fatte le generazioni e segmenti non tagliati durante l'ultima iterazione.

2.4 Risultati delle generazioni

Le implementazioni dei due modelli, spiegate nelle sezioni precedenti, producono dei vettori o delle matrici contenenti gli estremi dei segmenti generati. Per ricavare le lunghezze dei segmenti si è utilizzata la funzione `diff` con la quale è possibile ottenere la differenza tra gli elementi consecutivi del vettore passatogli per argomento. Passando per argomento alla funzione `hist` il vettore ottenuto da `diff` e specificando il numero di bin in cui si vogliono suddividere i dati (100'000), si sono ottenuti 2 vettori: il primo contenente le occorrenze per ogni bin e il secondo contenente il centro di ogni bin. In questo modo, dopo aver normalizzato le occorrenze ottenute, è stato possibile rappresentare in un grafico la distribuzione delle lunghezze utilizzando la funzione `plot`.
Le lunghezze che si ottengono attraverso questi processi differiscono tra loro di diversi ordini di grandezza; per apprezzarne la distribuzione occorre quindi utilizzare un scala diversa da quella lineare, per questo si è utilizzata la scala log-log. Quello che ci si aspetta di vedere in questa scala, nel caso in cui si rappresenti una distribuzione avente legge di potenza $p(x) = a \cdot x^b$, è una retta del tipo $y = p1 \cdot x + p2$. Per questo motivo si è fatto il log dei dati ottenuti, dopo aver rimosso i bin nulli, e si sono fittati con una retta attraverso la funzione `fit`.
Nelle sezioni successive sono riportate le distribuzioni ottenute dalle generazioni più numerose riportate in Tabella 2.1.

2.4.1 Tagli iniziali $m = 500$

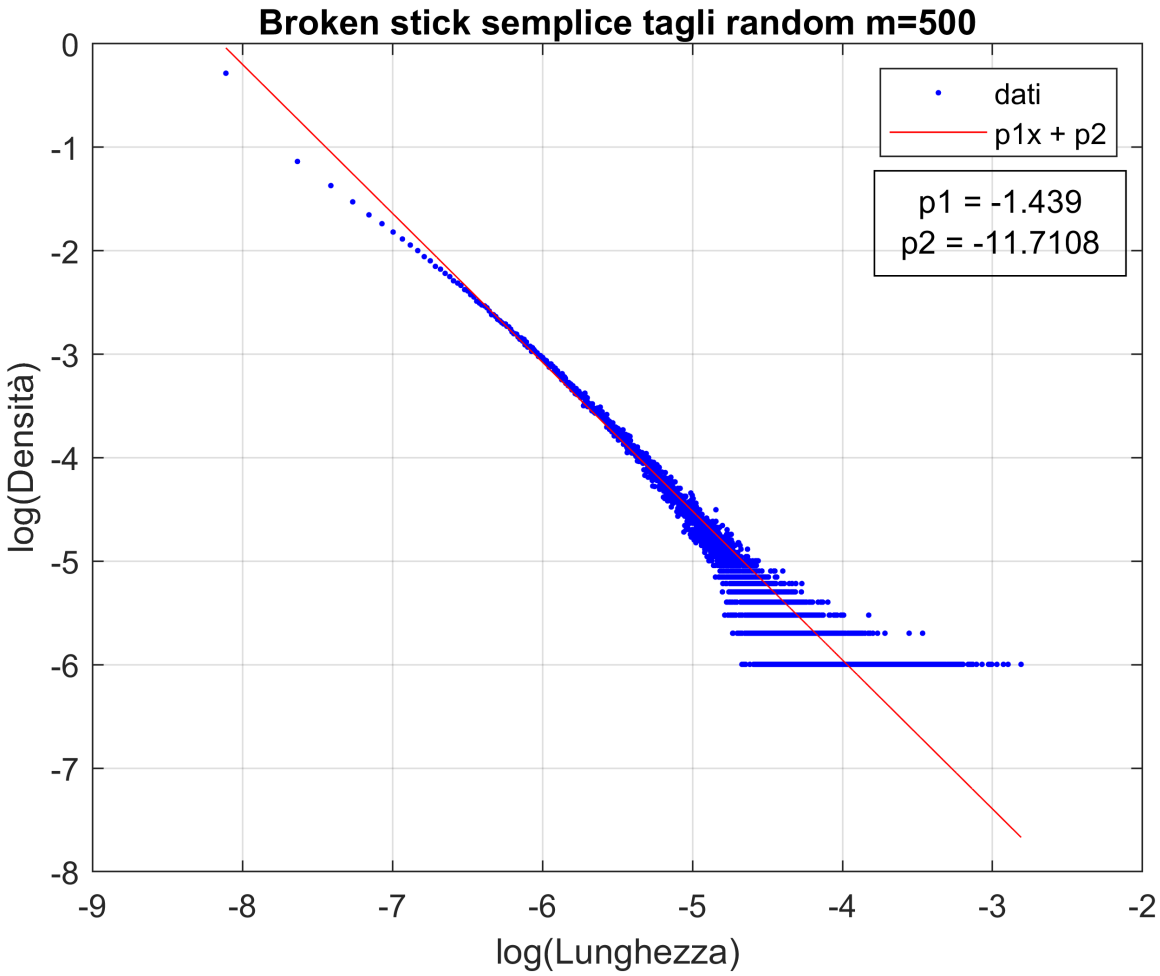


Figura 2.1: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-8, -5]$.

Broken stick semplice tagli random m=500 p=0.72 n=14	
Par	Val
p1	-1.44 ± 0.01
p2	-11.71 ± 0.07
rsquare	0.9868

Tabella 2.3: Parametri del fit lineare

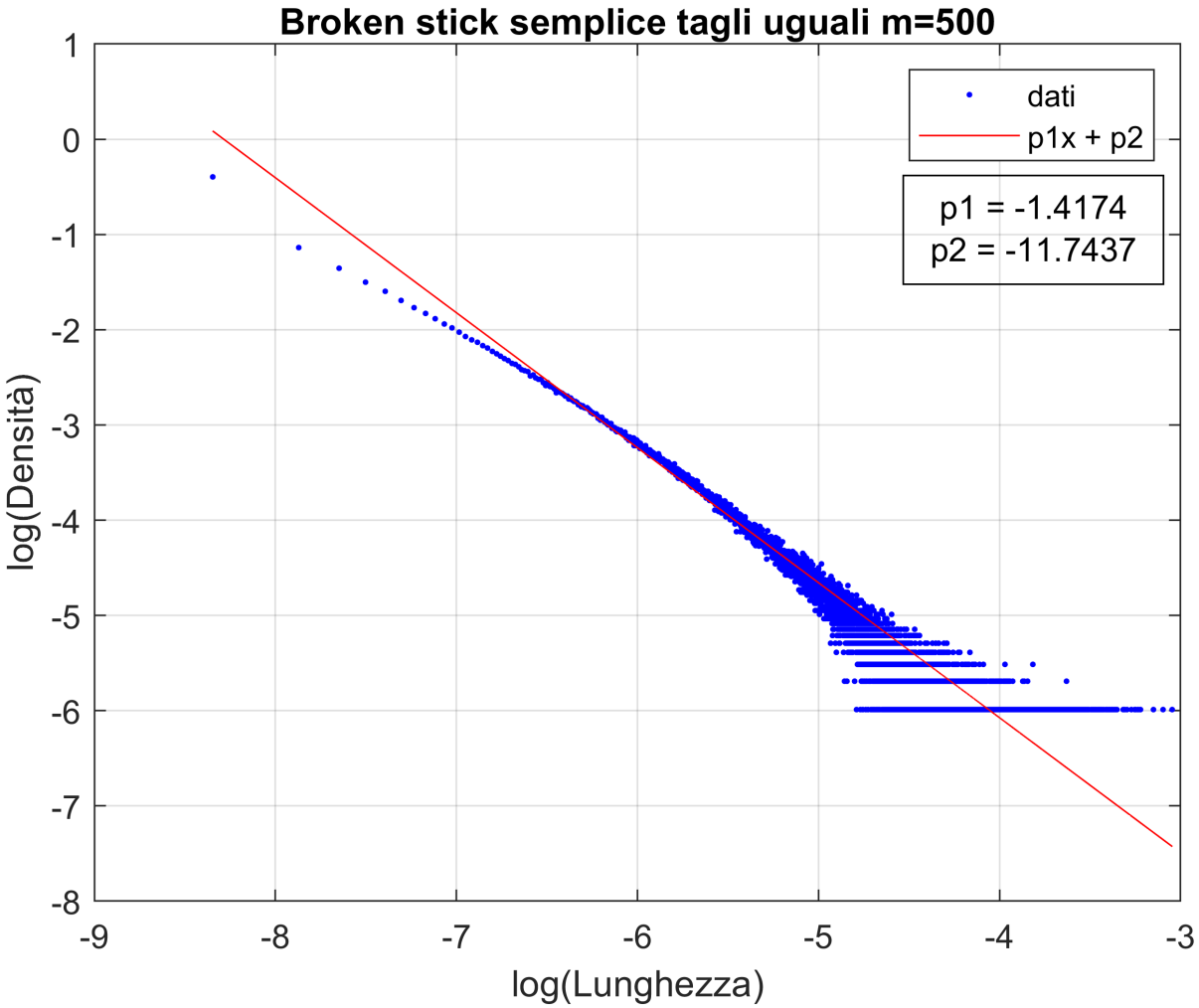


Figura 2.2: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-8, -5]$.

Broken stick semplice tagli uguali m=500 p=0.72 n=14	
Par	Val
p1	-1.42 ± 0.01
p2	-11.74 ± 0.06
rsquare	0.9825

Tabella 2.4: Parametri del fit lineare

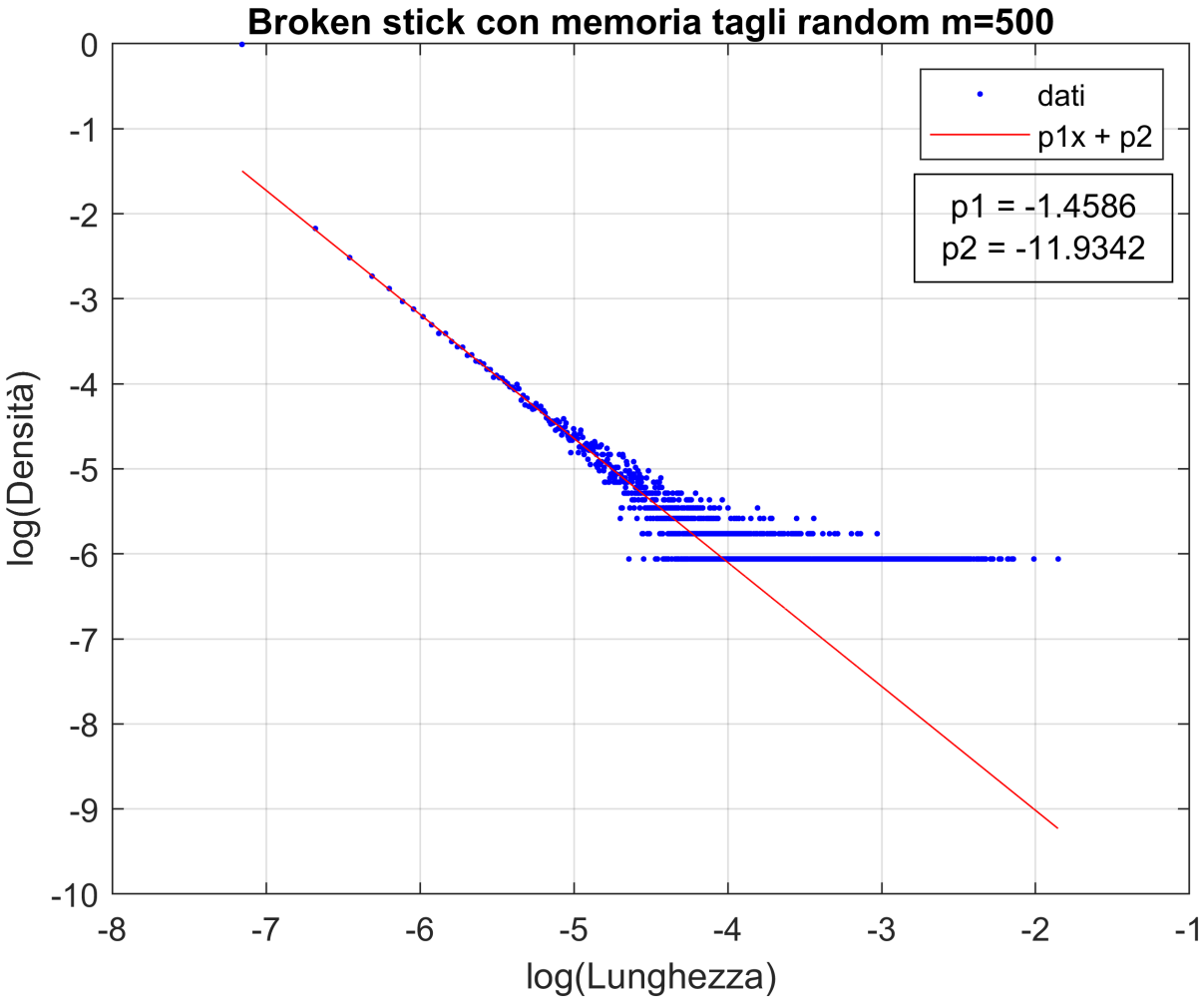


Figura 2.3: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-7, -5]$.

Broken stick com memoria tagli random m=500 p=0.72 n=20	
Par	Val
p1	-1.46 ± 0.03
p2	-11.9 ± 0.2
rsquare	0.9923

Tabella 2.5: Parametri del fit lineare

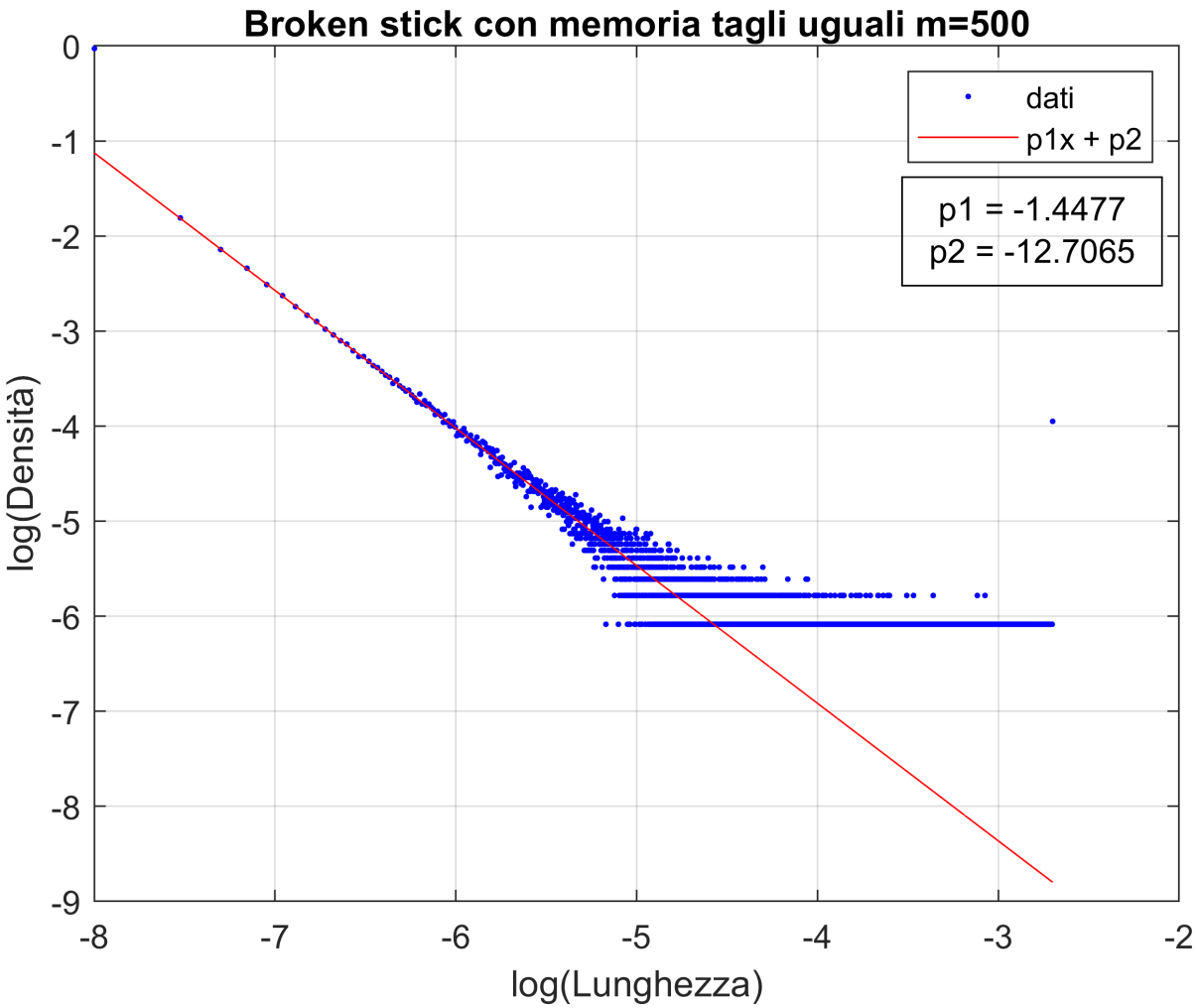
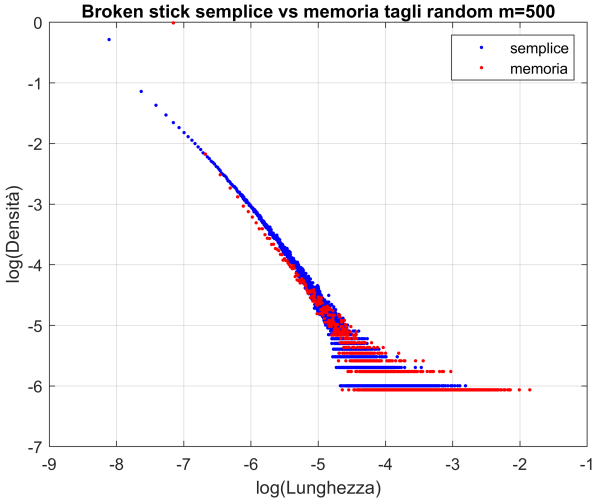


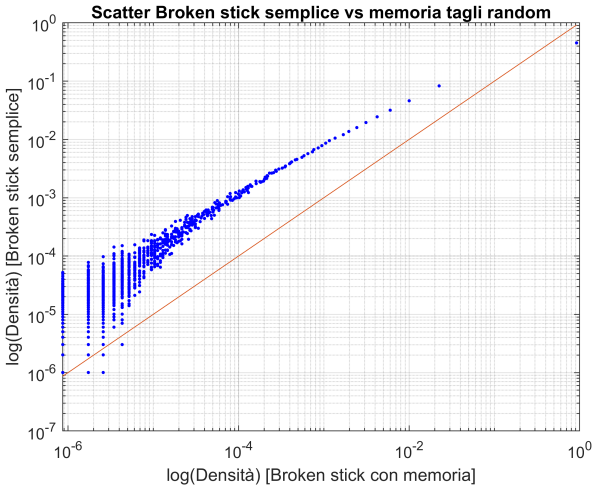
Figura 2.4: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-7.9, -5.5]$.

Broken stick con memoria tagli uguali m=500 p=0.72 n=20	
Par	Val
p1	-1.45 ± 0.02
p2	-12.7 ± 0.1
rsquare	0.9900

Tabella 2.6: Parametri del fit lineare

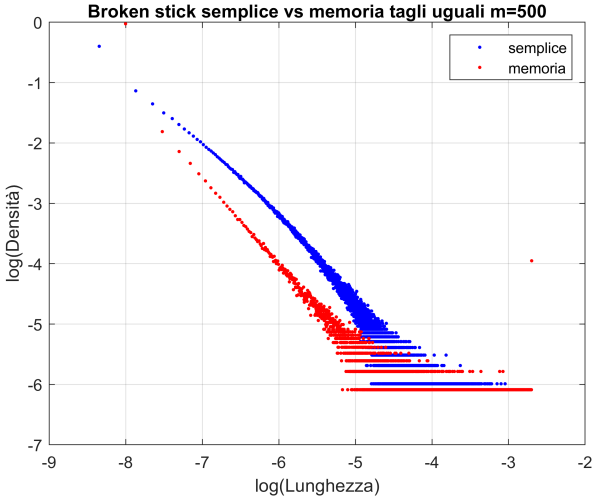


(a) Broken stick semplice vs memoria tagli random.

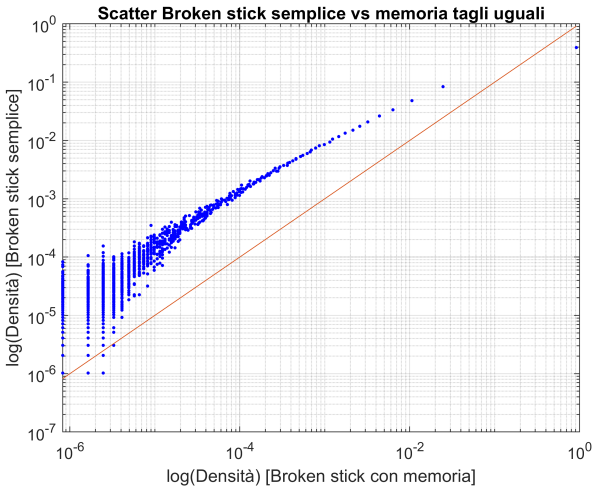


(b) Scatter Broken stick semplice vs memoria tagli random.

Figura 2.5: Grafico Broken stick semplice vs memoria tagli random e relativo scatter.

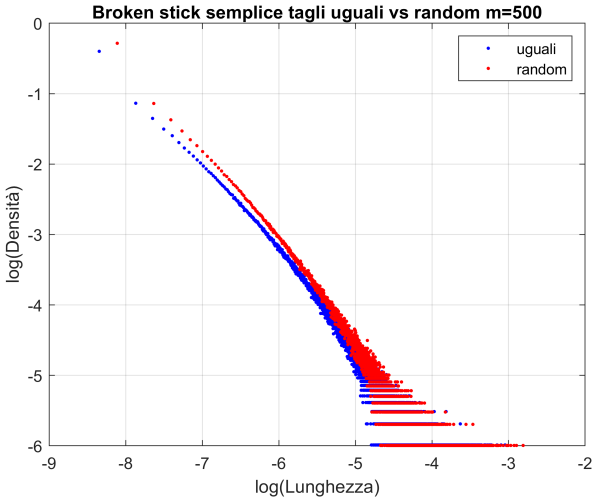


(a) Broken stick semplice vs memoria tagli uguali.

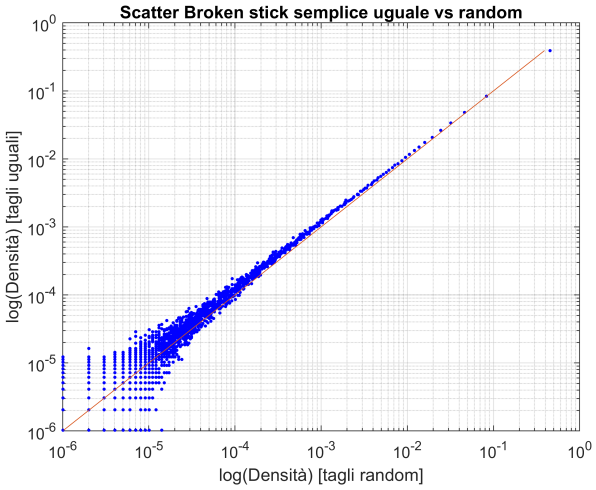


(b) Scatter Broken stick semplice vs memoria tagli uguali.

Figura 2.6: Grafico Broken stick semplice vs memoria tagli uguali e relativo scatter.

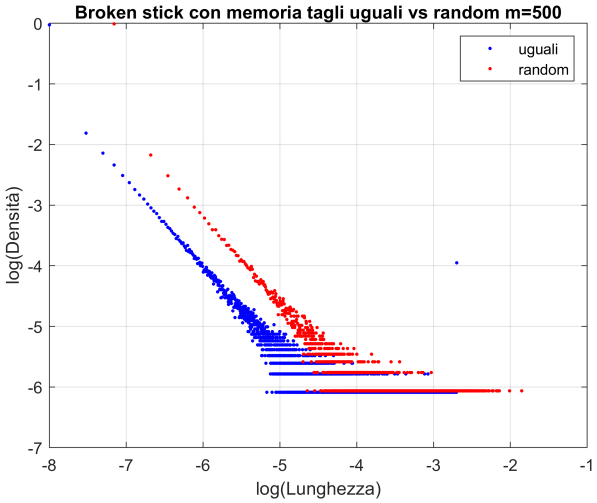


(a) Broken stick semplice tagli uguali vs random.

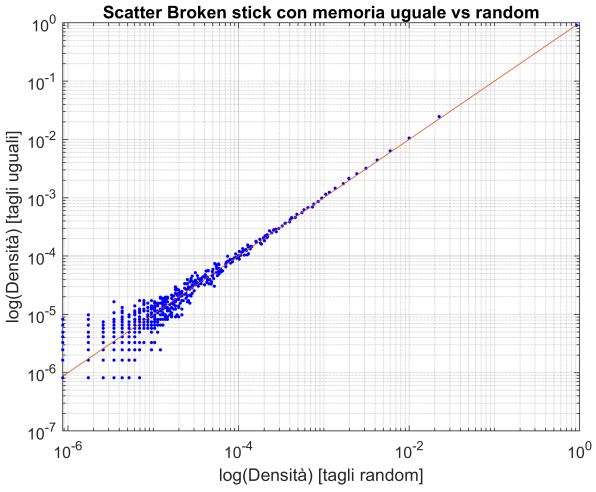


(b) Scatter Broken stick semplice tagli uguali vs random.

Figura 2.7: Grafico Broken stick semplice tagli uguali vs random e relativo scatter.



(a) Broken stick con memoria tagli uguali vs random.



(b) Scatter Broken stick con memoria tagli uguali vs random.

Figura 2.8: Grafico Broken stick con memoria tagli uguali vs random e relativo scatter.

2.4.2 Tagli iniziali $m = 5'000$

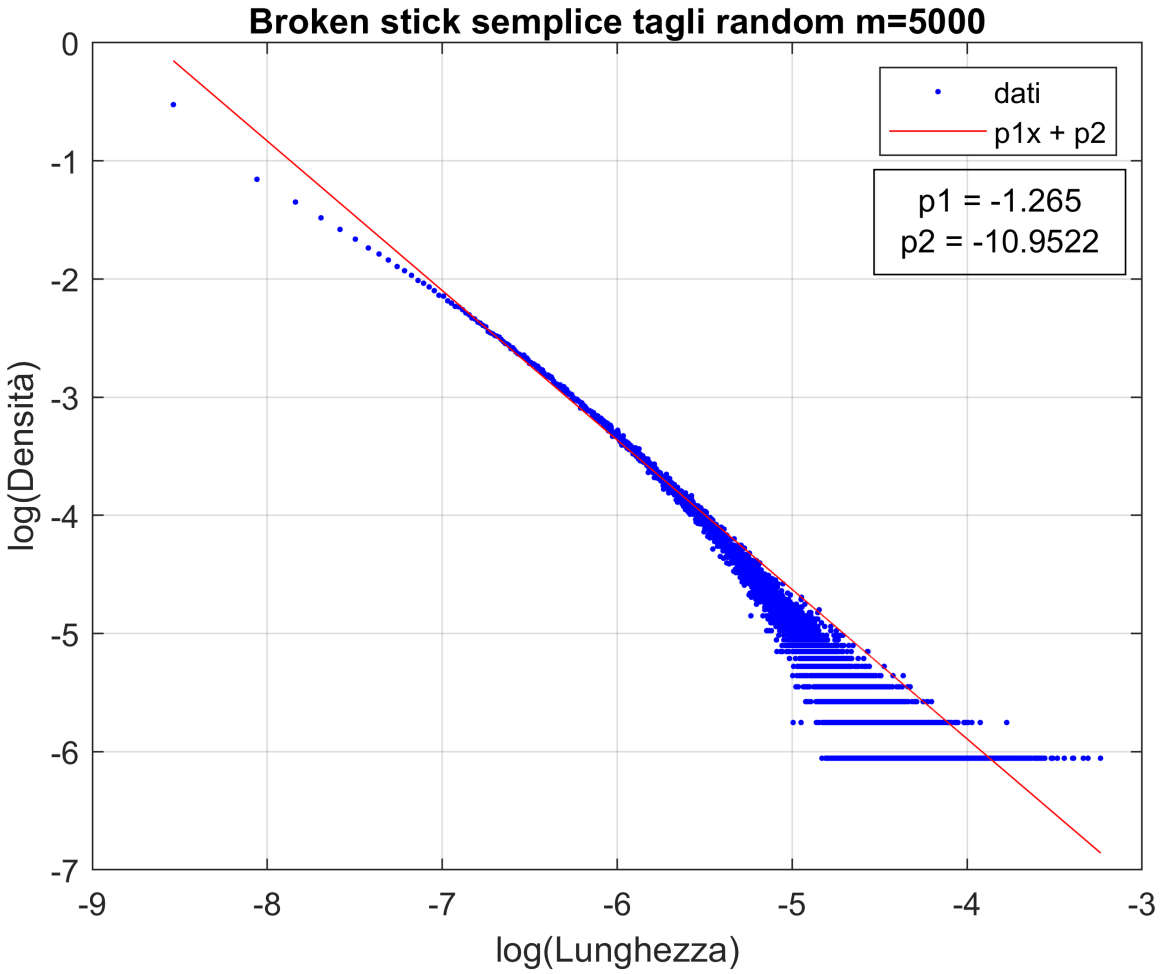


Figura 2.9: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-8, -5.5]$.

Broken stick semplice tagli random m=5000 p=0.72 n=10	
Par	Val
p1	-1.27 ± 0.01
p2	-10.95 ± 0.06
rsquare	0.9914

Tabella 2.7: Parametri del fit lineare

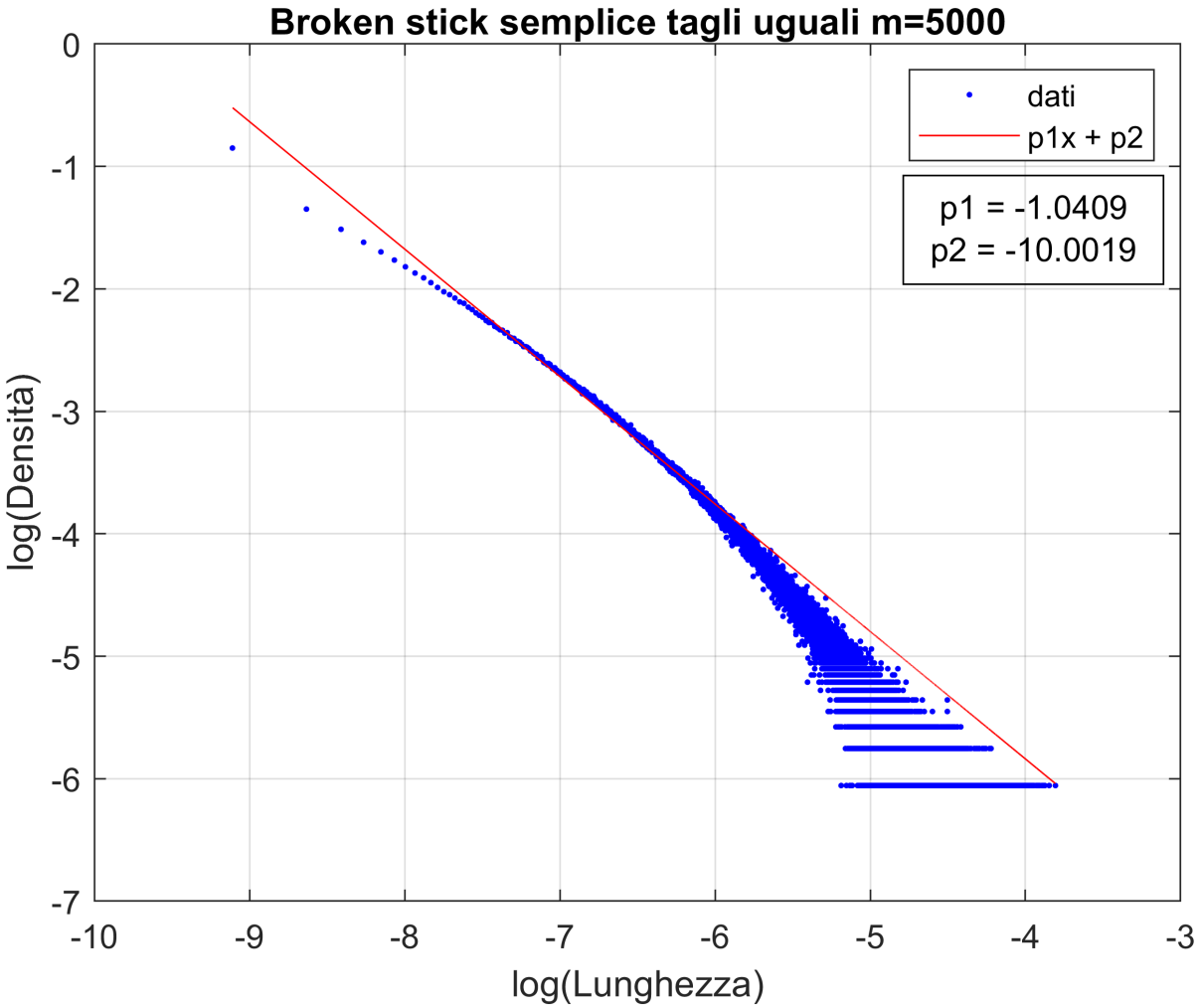


Figura 2.10: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-9, -6]$.

Broken stick semplice tagli uguali m=5000 p=0.72 n=10	
Par	Val
p1	-1.04 ± 0.01
p2	-10.00 ± 0.05
rsquare	0.9910

Tabella 2.8: Parametri del fit lineare

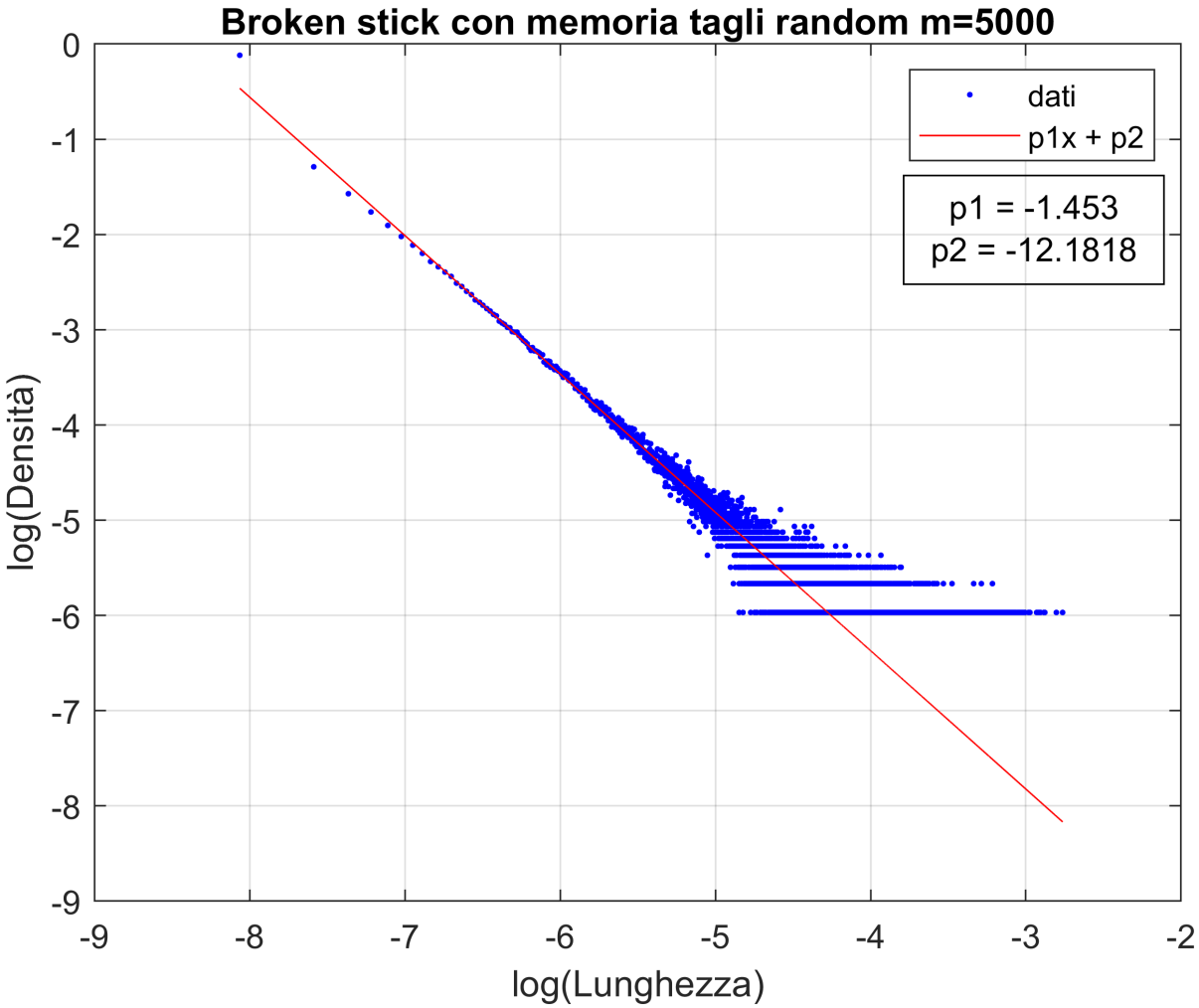


Figura 2.11: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-8, -5]$.

Broken stick com memoria tagli random m=5000 p=0.72 n=13	
Par	Val
p1	-1.45 ± 0.02
p2	-12.2 ± 0.1
rsquare	0.9811

Tabella 2.9: Parametri del fit lineare

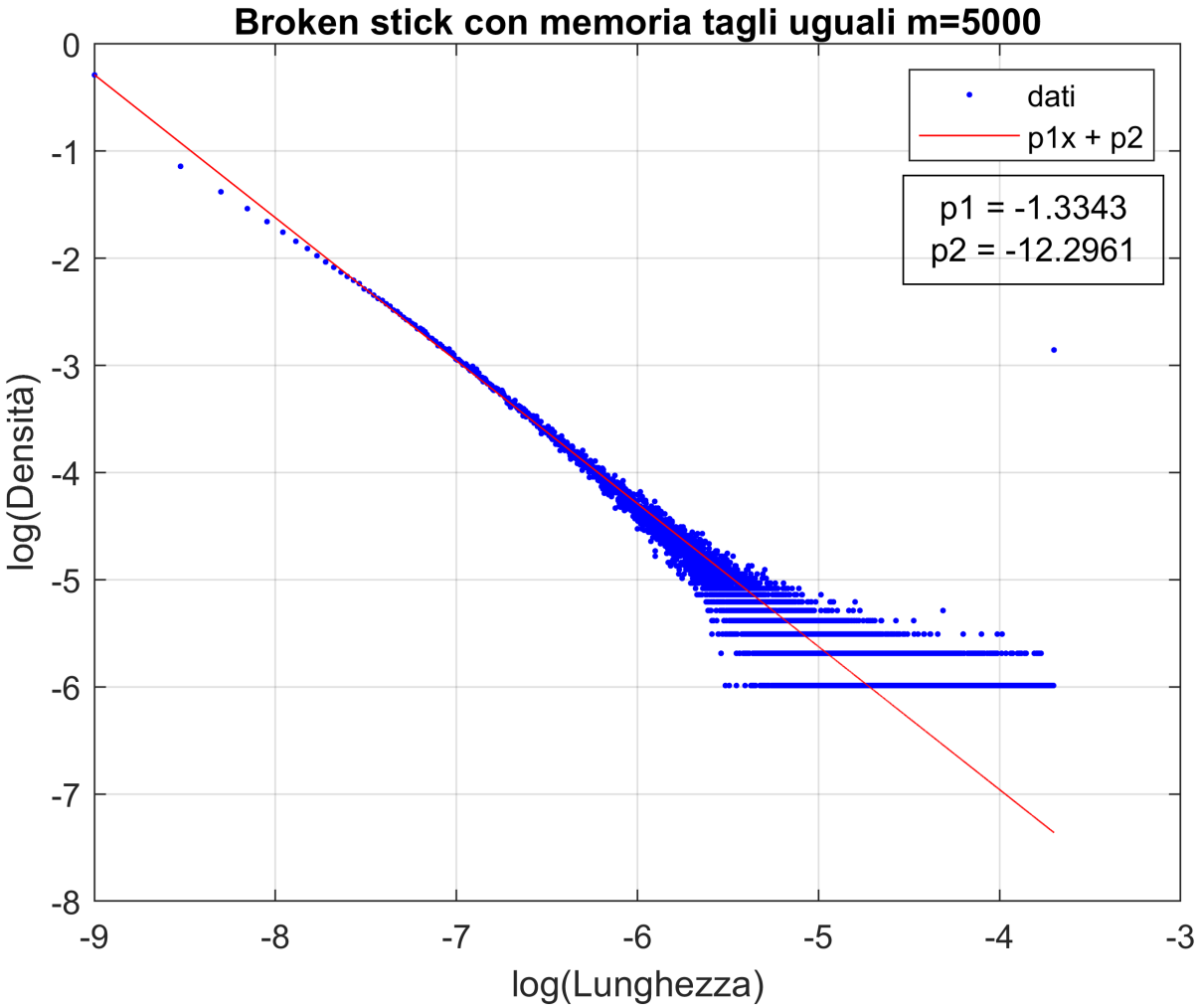
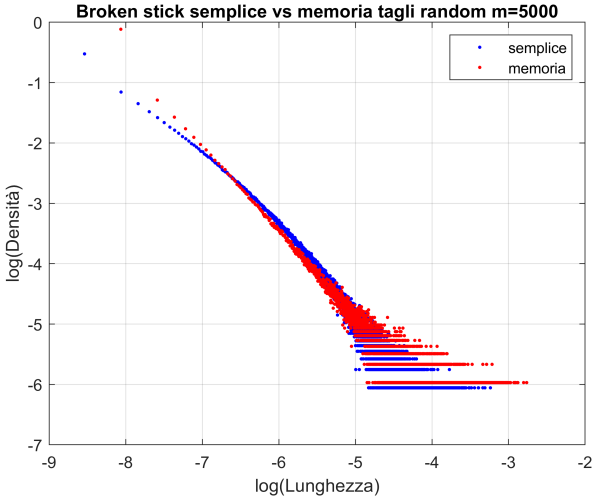


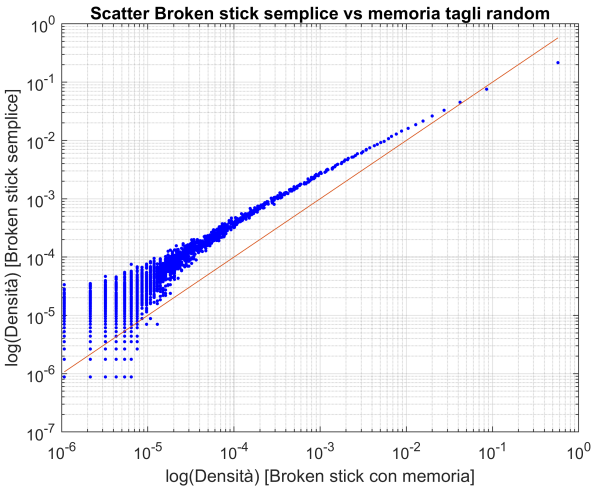
Figura 2.12: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-8.9, -6.15]$.

Broken stick con memoria tagli uguali m=5000 p=0.72 n=13	
Par	Val
p1	-1.33 ± 0.01
p2	-12.30 ± 0.06
rsquare	0.9945

Tabella 2.10: Parametri del fit lineare

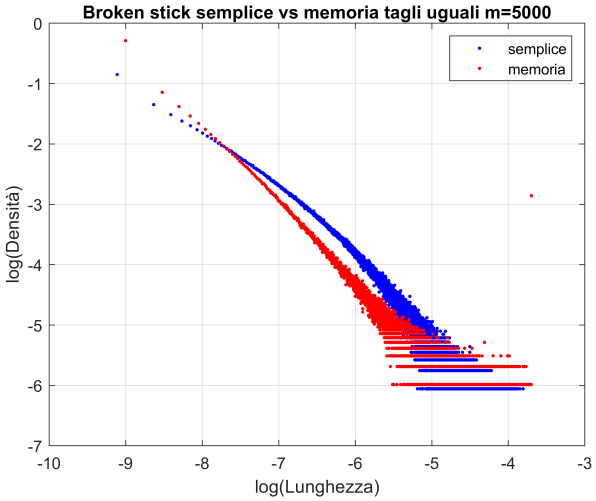


(a) Broken stick semplice vs memoria tagli random.

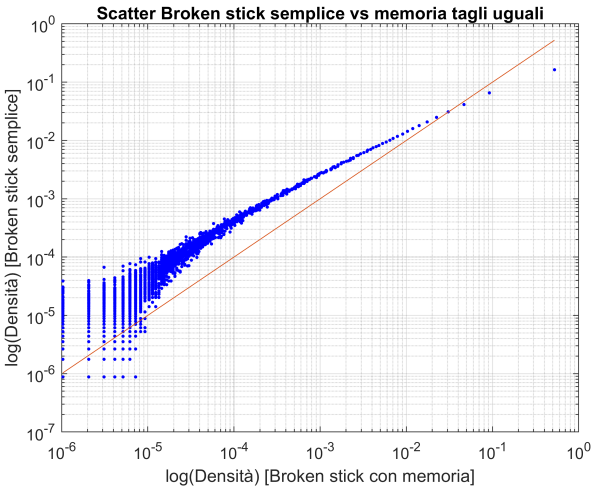


(b) Scatter Broken stick semplice vs memoria tagli random.

Figura 2.13: Grafico Broken stick semplice vs memoria tagli random e relativo scatter.

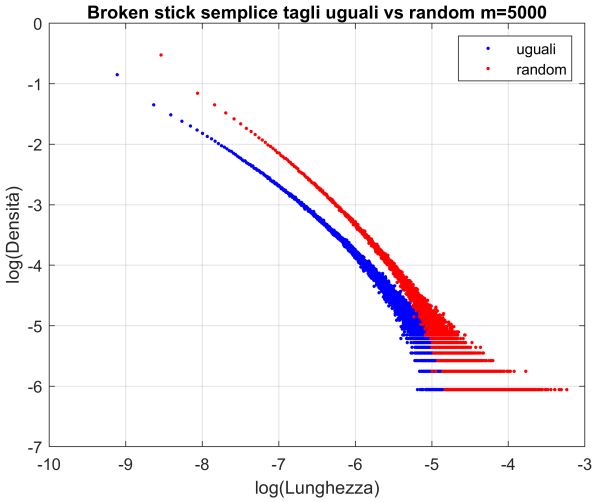


(a) Broken stick semplice vs memoria tagli uguali.

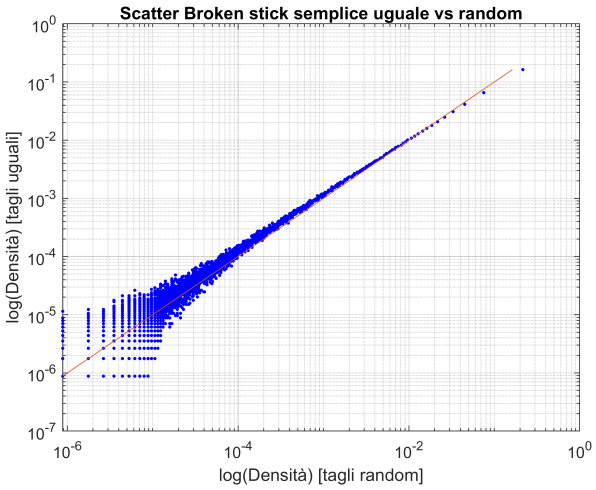


(b) Scatter Broken stick semplice vs memoria tagli uguali.

Figura 2.14: Grafico Broken stick semplice vs memoria tagli uguali e relativo scatter.

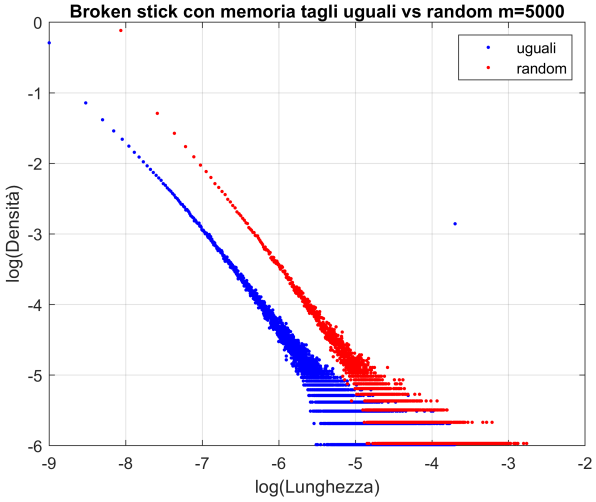


(a) Broken stick semplice tagli uguali vs random.

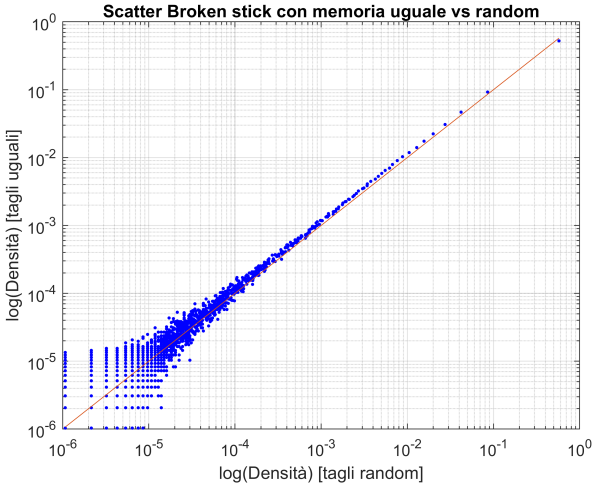


(b) Scatter Broken stick semplice tagli uguali vs random.

Figura 2.15: Grafico Broken stick semplice tagli uguali vs random e relativo scatter.



(a) Broken stick con memoria tagli uguali vs random.



(b) Scatter Broken stick con memoria tagli uguali vs random.

Figura 2.16: Grafico Broken stick con memoria tagli uguali vs random e relativo scatter.

2.4.3 Tagli iniziali $m = 50'000$

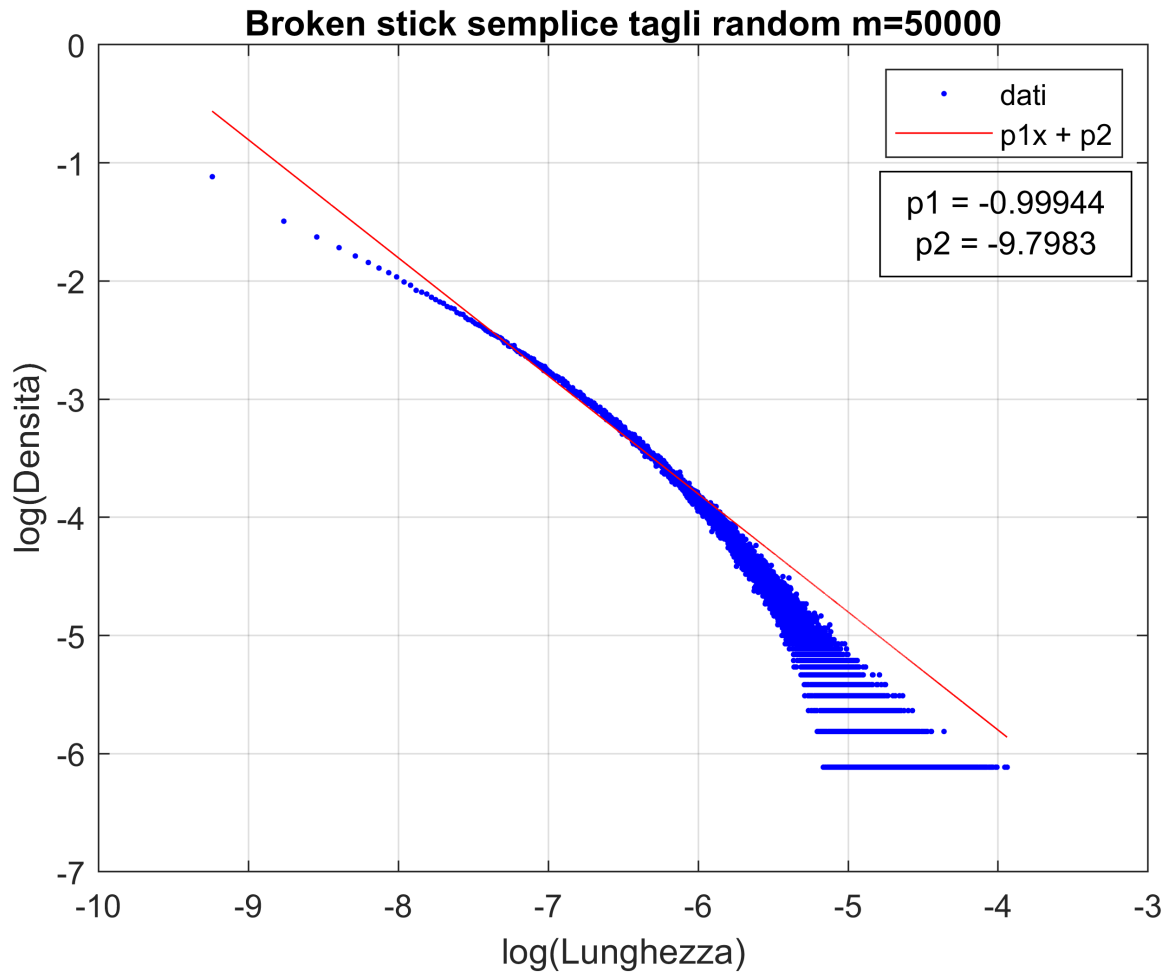


Figura 2.17: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-9, -6]$.

Broken stick semplice tagli random m=50000 p=0.72 n=6	
Par	Val
p1	-0.999 ± 0.009
p2	-9.80 ± 0.05
rsquare	0.9847

Tabella 2.11: Parametri del fit lineare

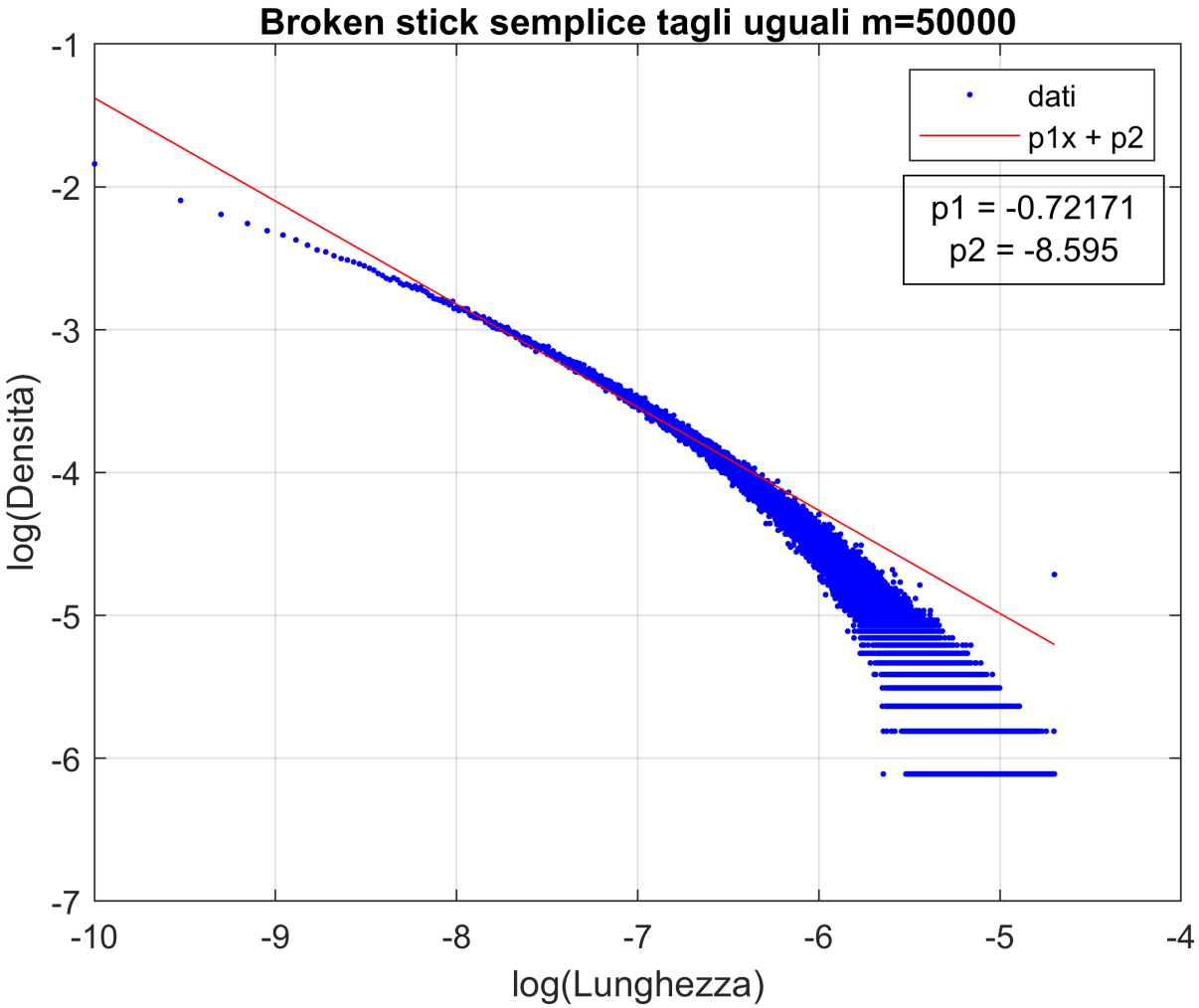


Figura 2.18: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-10, -6.5]$.

Broken stick semplice tagli uguali m=50000 p=0.72 n=6	
Par	Val
p1	-0.722 ± 0.005
p2	-8.60 ± 0.04
rsquare	0.9806

Tabella 2.12: Parametri del fit lineare

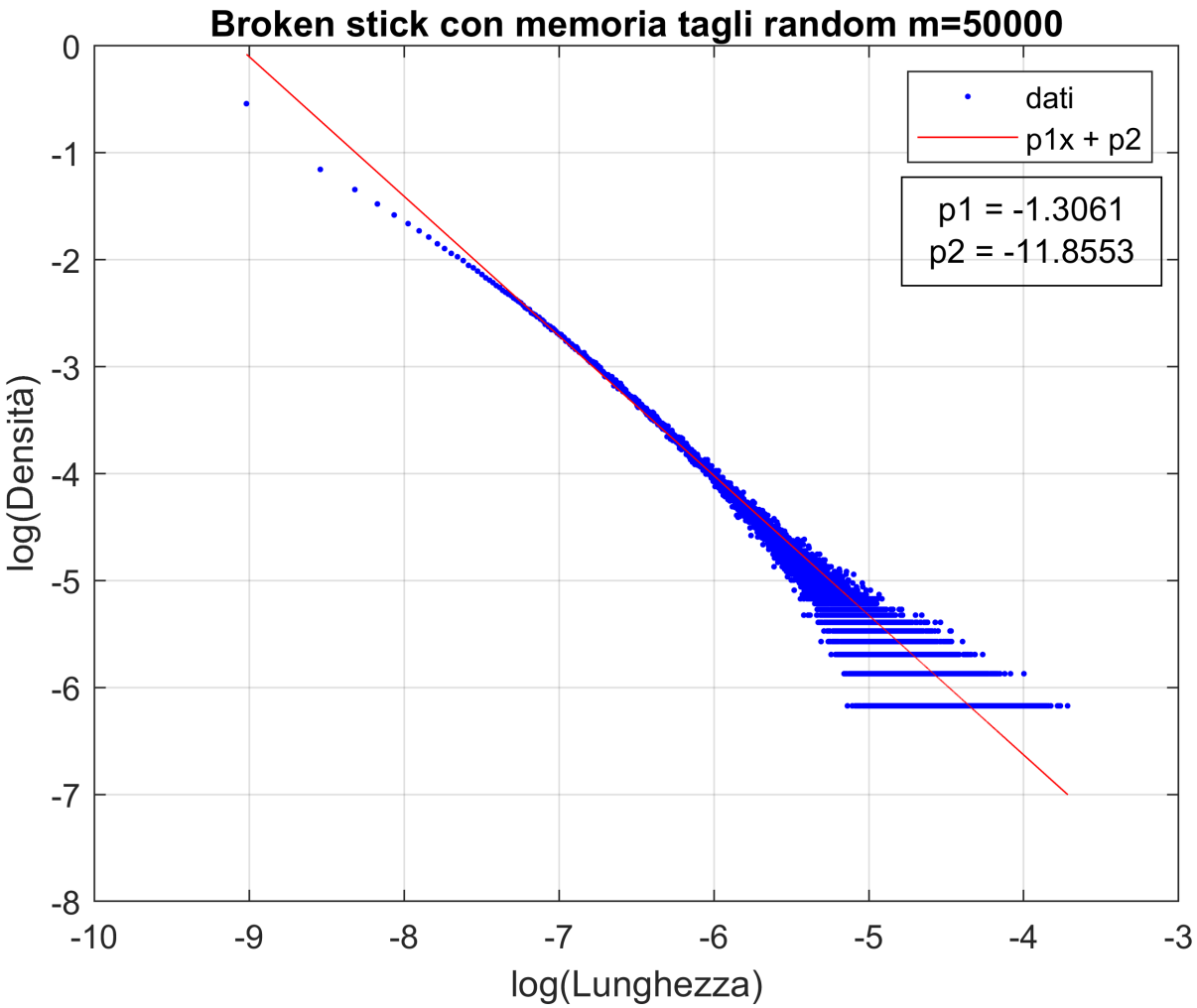


Figura 2.19: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-8.5, -5.8]$.

Broken stick com memoria tagli random m=50000 p=0.72 n=8	
Par	Val
p1	-1.306 ± 0.008
p2	-11.86 ± 0.04
rsquare	0.9922

Tabella 2.13: Parametri del fit lineare

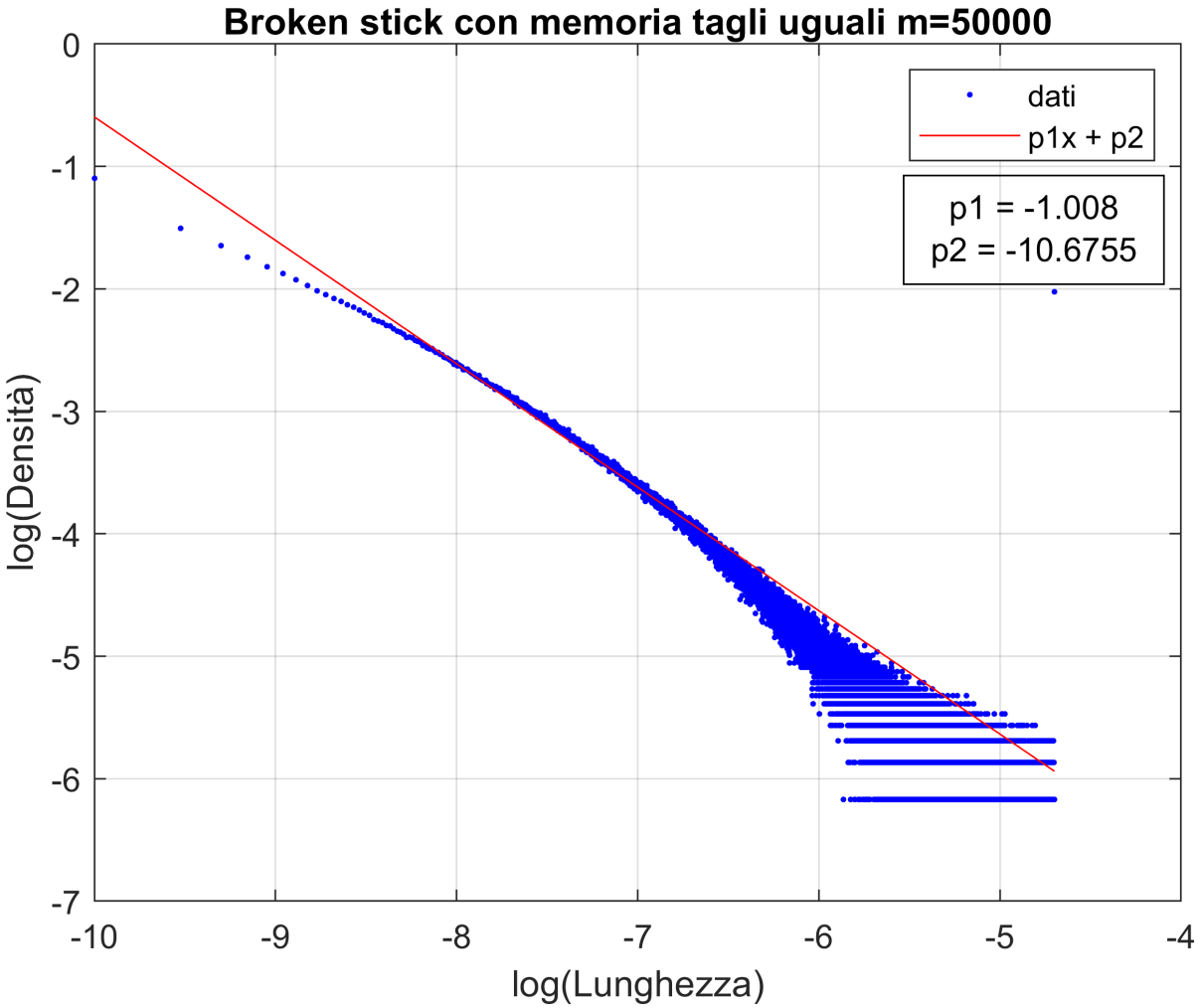
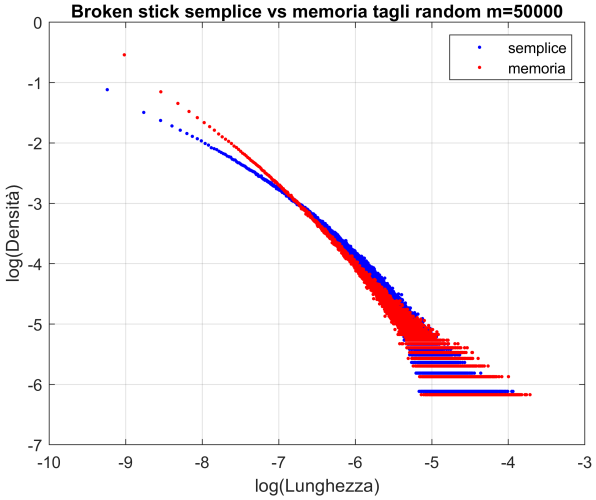


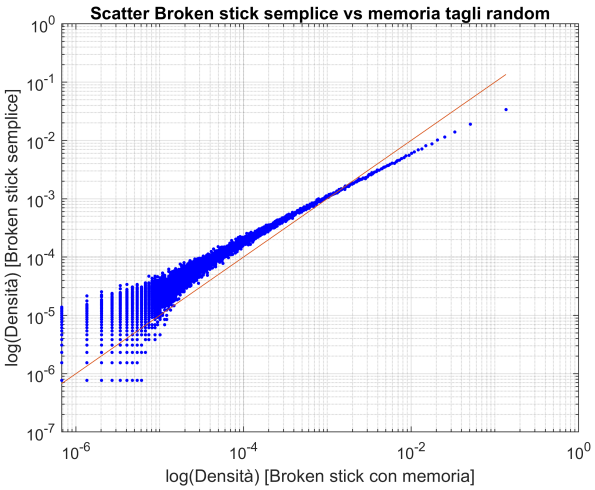
Figura 2.20: Grafico Densità-Lunghezza dei segmenti in scala log-log. I dati sono fittati con una retta nell'intervallo $[-9.5, -6.7]$.

Broken stick con memoria tagli uguali m=50000 p=0.72 n=8	
Par	Val
p1	-1.008 ± 0.006
p2	-10.68 ± 0.04
rsquare	0.9897

Tabella 2.14: Parametri del fit lineare

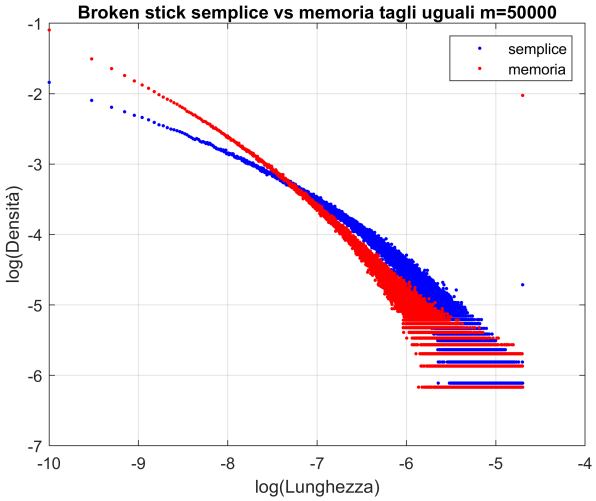


(a) Broken stick semplice vs memoria tagli random.

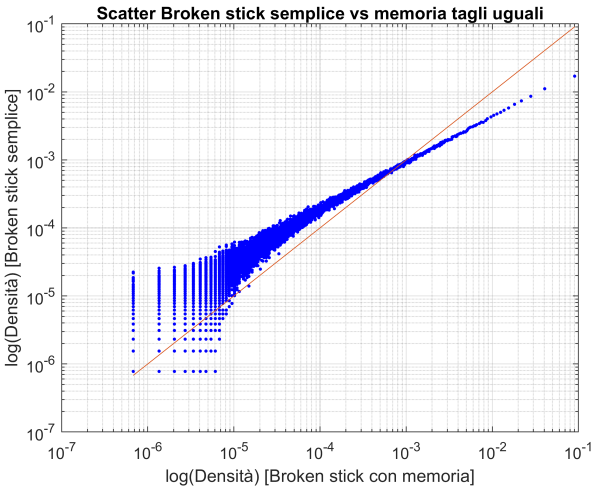


(b) Scatter Broken stick semplice vs memoria tagli random.

Figura 2.21: Broken stick semplice vs memoria tagli random.

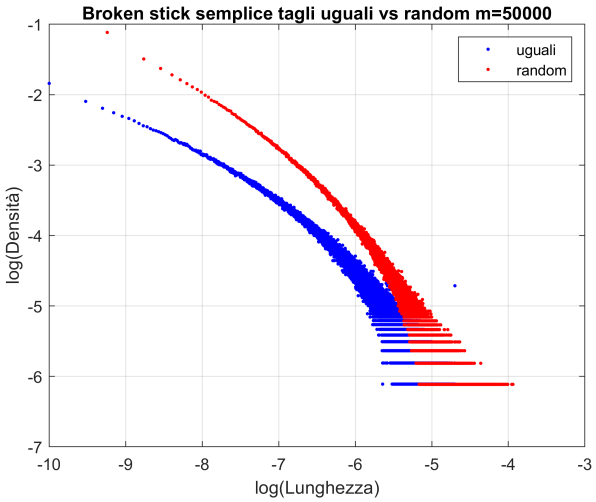


(a) Broken stick semplice vs memoria tagli uguali.

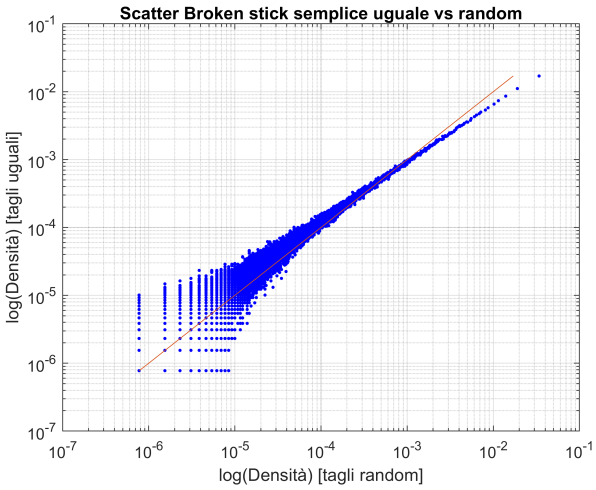


(b) Scatter Broken stick semplice vs memoria tagli uguali.

Figura 2.22: Grafico Broken stick semplice vs memoria tagli uguali e relativo scatter.

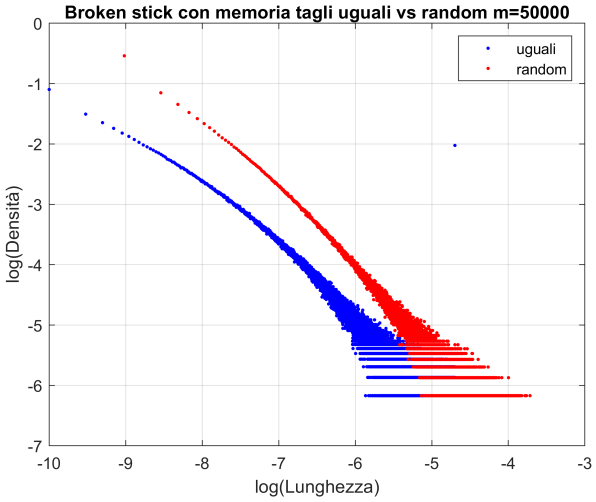


(a) Broken stick semplice tagli uguali vs random.

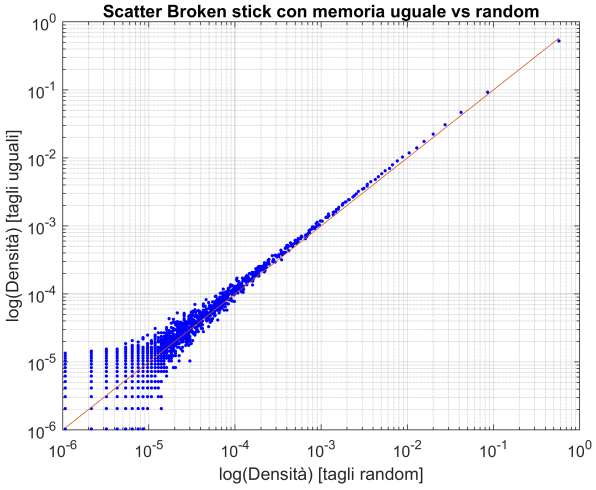


(b) Scatter Broken stick semplice tagli uguali vs random.

Figura 2.23: Grafico Broken stick semplice tagli uguali vs random e relativo scatter.



(a) Broken stick con memoria tagli uguali vs random.



(b) Scatter Broken stick con memoria tagli uguali vs random.

Figura 2.24: Grafico Broken stick con memoria tagli uguali vs random e relativo scatter.

2.5 Commenti

Dai grafici precedenti è possibile notare che

- il fit non è stato fatto sull'intera distribuzione ma solo in un intervallo. Infatti se si considera l'intera distribuzione il fit non ne riproduce l'andamento. Questo perchè la distribuzione è caratterizzata da code molto dense ed ampie che, utilizzando il metodo dei minimi quadrati, hanno un effetto dominante. Ne risulta quindi che il fit passi circa al centro della coda non rispecchiando l'andamento complessivo.

- I grafici contenenti gli andamenti prodotti dal modello Broken stick con memoria con tagli iniziali uguali, hanno tutti sulla destra un punto che si trova fuori dalla distribuzione complessiva. Questo rappresenta i segmenti iniziali, della stessa lunghezza, che non sono stati tagliati alla prima iterazione.

- In tutte le distribuzioni si nota che il primo punto a partire da sinistra si discosta dalla distribuzione. Questo è causato dal binnaggio utilizzato; avendo fatto un binning uniforme per ricoprire diversi ordini di grandezza, risulta che tutti i segmenti con lunghezza molto piccola vengano conteggiati in un unico bin. Si genera in questo modo un punto avente tante occorrenze che risulta essere non indicativo della distribuzione. Per riuscire a conteggiare queste occorrenze in bin diversi bisognerebbe diminuire l'ampiezza del bin, che equivale ad aumentare il numero di bin che ricoprono lo stesso intervallo.

- Confrontando a coppie gli andamenti ottenuti si evince chiaramente che le distribuzioni generate con i due modelli sono diverse. In particolare risulta che la distribuzione generata con il modello Broken stick con memoria abbia un andamento lineare in un range di valori più ampio rispetto a quella generata dal modello semplice. Quest'ultima presenta effetti di cutoff in piccoli e grandi intervalli.

- Le distribuzioni ottenute con lo stesso modello a partire da tipi di tagli differenti mostrano il medesimo andamento semplicemente traslato. Questo risultato è chiaramente conseguenza dei tagli iniziali; tagliando il segmento a random avremo una distribuzione normale attorno a valor medio identificato dalla lunghezza del segmento iniziale diviso il numero di tagli. Viceversa partendo da segmenti aventi la stessa lunghezza non abbiamo una distribuzione iniziale, ma solo un punto di partenza coincidente con il centro della distribuzione precedente. A partire da queste considerazioni, ci si attende di vedere la distribuzione raffigurante i tagli iniziali uguali a sinistra di quella con tagli iniziali a random. Ed effettivamente è quello che si osserva.

In seguito si comparano gli andamenti ottenuti con gli stessi modelli al variare del numero di tagli iniziali.

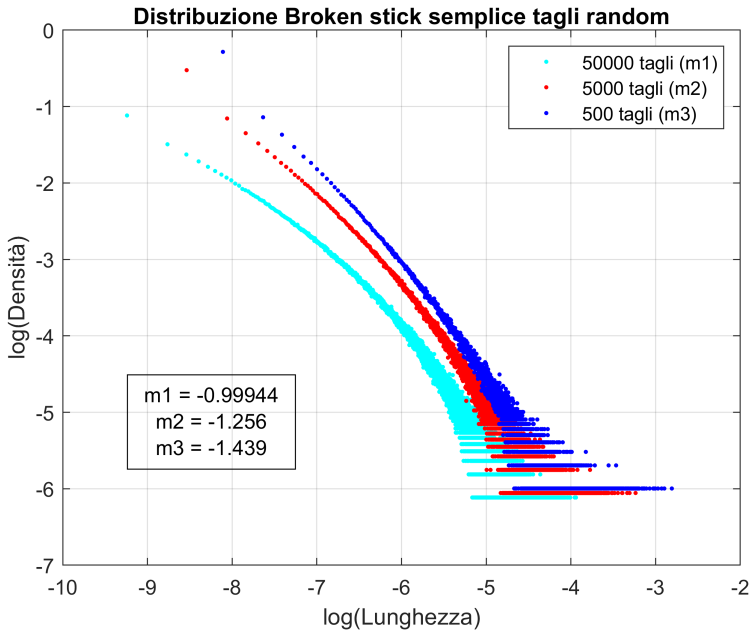


Figura 2.25: Grafico Densità-Lunghezza rappresentate le distribuzioni ottenute a partire da un diverso numero di segmenti inziali, applicando il modello Broken stick semplice con tagli inziali a random.

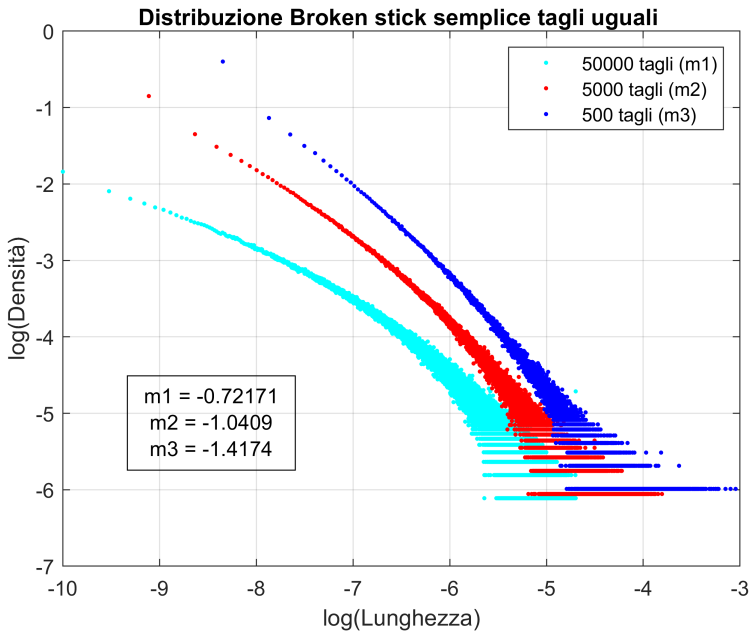


Figura 2.26: Grafico Densità-Lunghezza rappresentate le distribuzioni ottenute a partire da un diverso numero di segmenti inziali, applicando il modello Broken stick semplice con tagli inziali uguali.

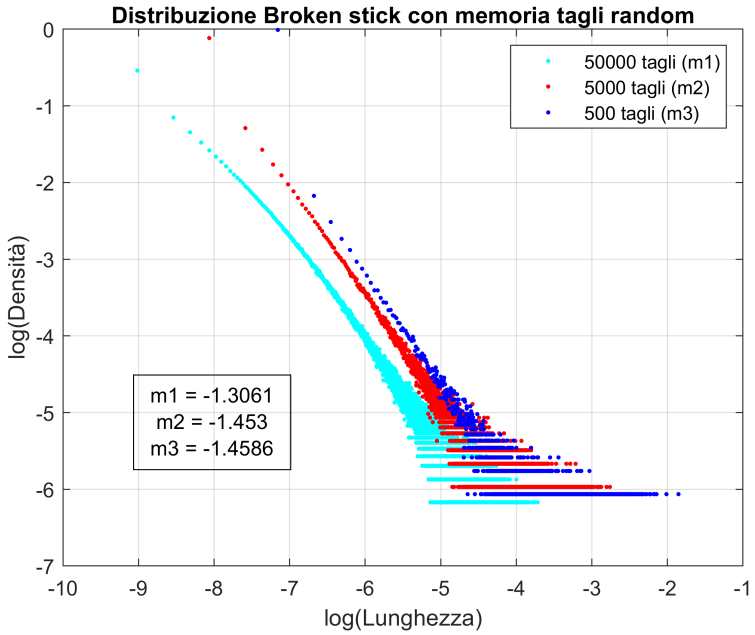


Figura 2.27: Grafico Densità-Lunghezza rappresentate le distribuzioni ottenute a partire da un diverso numero di segmenti inziali, applicando il modello Broken stick con memoria con tagli inziali a random.

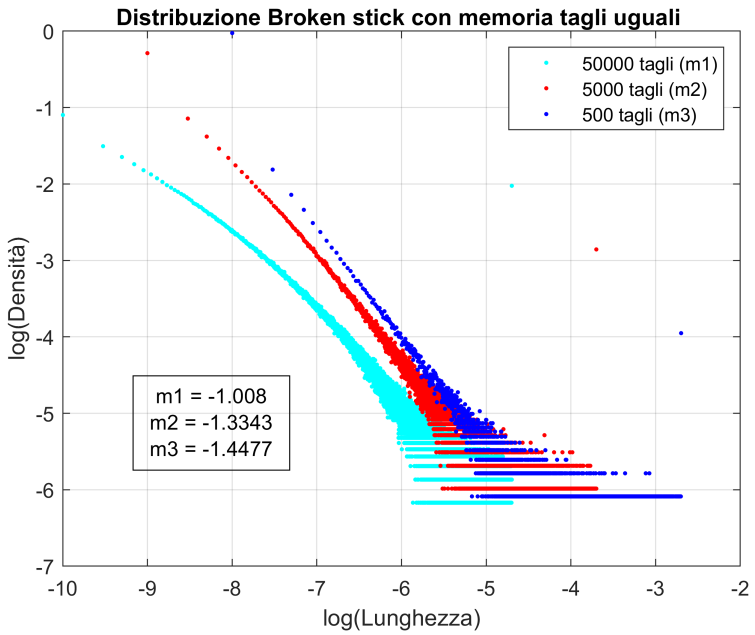


Figura 2.28: Grafico Densità-Lunghezza rappresentate le distribuzioni ottenute a partire da un diverso numero di segmenti inziali, applicando il modello Broken stick con memoria con tagli inziali uguali.

A partire dai grafici precedenti è possibile notare le seguenti caratteristiche:

- tutti gli andamenti si incurvano all'aumentare del numero di tagli iniziali. Occorre considerare che, partendo da un numero maggiore di segmenti, per ottenere generazioni aventi circa lo stesso numero di dati (1'000'000 di segmenti), bisogna diminuire il numero iterazioni. In questo modo è possibile che i modelli risentano maggiormente delle condizioni iniziali e che quindi per questo l'andamento risulti più curvo.
- Si osserva chiaramente che, in scala log-log, all'aumentare dei tagli iniziali ci sia una traslazione delle distribuzioni verso sinistra.

Capitolo 3

Conclusioni

Alla luce dei risultati ottenuti, si può concludere che l'implementazione dei modelli risulta non banale e che il modello teorico proposto non è pienamente realizzabile a causa dei limiti computazioni.

Le distribuzioni prodotte dai due modelli sono differenti. In particolare quella prodotta dal modello Broken stick con memoria risulta lineare, in scala log-log, lungo tutto il range di valori in cui è rappresentata; quindi l'intero andamento potrebbe essere ricondotto ad una legge di potenza. Mentre la distribuzione generata con il modello Broken stick semplice presenta effetti di cutoff sia in intervalli piccoli che grandi. L'analisi statistica condotta non permette di dire con certezza che tipo di distribuzioni siano quelle prodotte, però mette in luce la problematica del fit. Essendo generate distribuzioni molto dense ed ampie nelle code, il fit con il metodo dei minimi quadrati risulta inadeguato per descriverne l'andamento analizzando tutti i dati. Per questo si è fittata la curva in un intervallo non troppo denso in cui la distribuzione presentava un andamento rettilineo.

La distribuzione generata dal modello Broken stick con memoria a partire da 500 tagli iniziali risulta essere lineare in un range di valori più ampio rispetto a quella ottenuta con 50'000 tagli. Il fatto che le due distribuzioni abbiano 2 andamenti differenti potrebbe essere dovuto al numero diverso di iterazioni realizzate. Il modello con tagli minori è stato sottoposto a 20 iterazioni per produrre 1'000'000 di segmenti, mentre quello avente 50'000 segmenti di partenza ne ha subite 8. É possibile che il numero di iterazioni influisca sulla distribuzione facendo emergere aspetti propri delle condizioni iniziali piuttosto che del modello in sè. Da sottolineare che a causa dei limiti computazionali non è possibile scegliere in maniera arbitraria le condizioni iniziali. Ad esempio le 20 iterazioni applicate al caso di $m = 500$ non sono realizzabili a partire da $m = 50'000$. Si potrebbe invece diminuire il numero di tagli iniziali e aumentare il numero di iterazioni in modo da studiarne il comportamento più lontano dalle condizioni iniziali.

Bibliografia

[1] https://biologydictionary.net/nucleotide/

[2] https://www.gmpe.it/chimica/acidi-nucleici

[3] De Vita J., *Niche Separation and the Broken-Stick Model*, The American Naturalist, Aug., 1979, Vol. 114, No. 2 (Aug., 1979), pp. 171-178 http://www.jstor.com/stable/2460215

[4] D'Alberto J. (2019). *Studio delle interdistanze dei dinucleotidi CG e TA nei cromosomi umani*. Tesi di laurea triennale in Fisica, Università di Bologna, Relatore: Remondini D.; Correlatore: Merlotti A.

[5] Merlotti A. (2016), *DNA sequence analysis: a statistical characterization of dinucleotides interdistances across multiple organisms*. Tesi di laurea magistrale in Fisica, Università di Bologna, Relatore: Remondini D.; Correlatore: Faria do Valle Í.

[6] Paci G, Cristadoro G, Monti B, Lenci M, Esposti MD, Castellani G, et al. *Characterization of DNA methylation as a function of biological complexity via dinucleotide inter-distances*. Phil. Trans. R. Soc. A 374: 20150227. http://dx.doi.org/10.1098/rsta.2015.0227.

[7] Paci G. (2014). *Statistical methods for the analysis of DNA sequences: application to dinucleotide distribution in the human genome*. Tesi di laurea magistrale in Fisica, Università di Bologna, Relatore: Remondini D.; Correlatore: Cristadoro G.