ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΜΟΝΤΕΛΑ ΠΙΘΑΝΟΤΗΤΩΝ ΓΙΑ ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΟΥ ΟΓΚΟΥ

Αναστασία Τζογάνη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούνιος 2018

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΜΟΝΤΕΛΑ ΠΙΘΑΝΟΤΗΤΩΝ ΓΙΑ ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΟΥ ΟΓΚΟΥ

Αναστασία Τζογάνη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούνιος 2018

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:
-Καθηγητής: Κούτρας Μάρκος(Επιβλέπων)
-Καθηγητής: Θεοδωρίδης Ιωάννης
- Επίκουρος Καθηγητής: Πελέκης Νικόλαος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

PROBABILITY MODELS FOR FITTING BIG DATA

By

Anastasia Tzogani

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus. Greece
June 2018

Στους γονείς μου
Δημήτρη και Ιωάννα

Περίληψη

Τα τελευταία χρόνια πολλοί ερευνητές επικεντρώθηκαν στην στοχαστική μοντελοποίηση σε φυσικών και κοινωνικών φαινομένων και έφτασαν στο συμπέρασμα ότι τα περισσότερα από αυτά ακολουθούν την Κατανομή Νόμου Δύναμης (Power Law Distribution. PLD). Στη παρούσα εργασία δίνουμε αρχικά τον ορισμό της συνεχούς PLD και τις ιδιότητές της. Στην συνέχεια παραθέτουμε διάφορες στατιστικές μεθόδους εκτίμησης των παραμέτρων της μαζί με τον έλεγχο καλής προσαρμογής Kolmogorov-Smirnov. Ακολούθως, προχωρούμε στην παρουσία τεχνικών σύγκρισης της PLD έναντι άλλων κατανομών βαριάς ουράς μέσω του ελέγχου του λόγου πιθανοφανειών και αναλύουμε κάποιες από αυτές. Επιπρόσθετα, χρησιμοποιώντας προσομοιωμένα δεδομένα από την PLD εξετάζουμε ποια από τις μεθόδους εκτίμησης δίνει αξιόπιστα αποτελέσματα και επιπλέον ποια εναλλακτική κατανομή προσαρμόζεται καλύτερα σε δεδομένα τα οποία προέρχονται από την PLD. Τέλος, δίνουμε κάποιες εφαρμογές της PLD και παρουσιάζουμε πως μπορούμε να εφαρμόσουμε τις τεχνικές εκτίμησης των παραμέτρων της μέσα από διάφορες γλώσσες προγραμματισμού.

Abstract

Over the last few decades, many scientists focused on the stochastic modeling of natural and social phenomena and arrived at the conclusion that many of them follow Power Law Distribution. This dissertation contains some brief definitions of the continuous PLD and presents in detail its properties. Furthermore, we present statistical methods for estimating the parameters of PLD along with Kolmogorov-Smirnov goodness-of-fit. Moreover, we compare PLD with other alternative fat-tailed distributions using likelihood ratios and also we describe some of them. Additionally, we use simulated data from PLD in order to find out which of the estimation methods gives best results and also which alternative distribution can describe better data from PLD. Finally, we present some applications of PLD and provide information on how one can apply estimating methods for the PLD parameters via several programming languages.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

1.1 ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ

Ο όρος μεγάλα δεδομένα έχει να κάνει με σύνολα δεδομένων τα οποία είναι τόσο μεγάλα που δεν μπορούν να αντιμετωπιστούν όπως οι παραδοσιακές βάσεις δεδομένων για την συλλογή, διαχείριση και επεξεργασία των δεδομένων. Τα μεγάλα δεδομένα δεν είναι ένα καινούργιο φαινόμενο, αλλά αποτελεί μέρος μιας μακράς εξέλιξης της συλλογής και χρήσης των δεδομένων. Το Ίντερνετ, η τεχνητή νοημοσύνη, το κινητό και τα μέσα μαζικής δικτύωσης είναι οι κύριες πηγές που οδηγούν στην δημιουργία τεράστιου όγκου δεδομένων τα οποία χαρακτηρίζονται από την πολυπλοκότητά τους.

Υπάρχει μεγάλο χάσμα μεταξύ της παραγωγής δεδομένων και της δυνατότητας να εξάγουμε γνώση από αυτά. Αυτό απαιτεί κάποια εργαλεία που να αντιμετωπίζουν τις τέσσερις προκλήσεις που διέπουν τα μεγάλα δεδομένα ή *4 Vs* όπως τα αποκαλούν. Πιο συγκεκριμένα αυτές οι προκλήσεις είναι ο όγκος (*volume*), η ταχύτητα (*velocity*), η ποικιλία (*variety*) και η εγκυρότητα (*veracity*) (Camacho et al 2014).

Ο όγκος έχει να κάνει με την ποσότητα των δεδομένων, δηλαδή τονίζει την αποθήκευση, τον υπολογισμό της χωρητικότητας και την μνήμη. Η ταχύτητα αναφέρεται στην ταχύτητα εισαγωγής των δεδομένων αλλά και την εξόρυξη γνώσης από αυτά. Η ποικιλία τονίζει το εύρος τύπων των δεδομένων και τις πηγές του. Υπάρχουν διαφορετικές πηγές δεδομένων, που έχουν δομημένη και μη δομημένη πληροφορία, οι οποίες θα πρέπει να συνδυαστούν σωστά έτσι ώστε να γίνει η σωστή ανάλυση των δεδομένων. Στα μεγάλα δεδομένα εντοπίζεται συχνά το πρόβλημα του θορύβου και της αναληθής συσχέτισης (*spurious correlation*) συνεπώς η τελευταία πρόκληση είναι η εγκυρότητα δηλαδή η αναζήτηση αξιόπιστης πληροφορίας από μεγάλα σύνολα δεδομένων (Camacho 2014).

Μέθοδοι μέτρησης όπως το μέγιστο, το ελάχιστο, οι μέσες τιμές δεν μπορούν να περιγράψουν αυτά τα δεδομένα διότι χάνουν μεγάλο μέρος της πληροφορίας. Επιπρόσθετα, συνηθισμένες κατανομές όπως η Κανονική Κατανομή δεν μπορούν να περιγράψουν τα μεγάλα δεδομένα λόγω της πολυπλοκότητάς τους καθώς εμφανίζουν αρκετά συχνά ασσυμετρία

πράγμα που σημαίνει ότι η μέση τιμή δεν είναι αντιπροσωπευτική για όλες τις παρατηρήσεις. Μια δημοφιλής κατανομή που χρησιμοποιείται για να περιγράψει αυτά τα δεδομένα είναι η Κατανομή του Νόμου Δύναμης ή *Power Law Distribution* (PLD) (Gadepally και Kepner 2015).

1.2 ΚΑΤΑΝΟΜΗ ΝΟΜΟΥ ΔΥΝΑΜΗΣ

Το 1890 ο Vilfredo Pareto μελέτησε τον φόρο εισοδήματος από πολλές χώρες και για πολλές χρονικές περιόδους. Μελέτησε το γράφημα του αριθμού των ατόμων που κατέχουν εισόδημα πάνω από ένα όριο έναντι του αντίστοιχου ορίου σε διπλά λογαριθμησμένους άξονες. Με το παραπάνω γράφημα ανακάλυψε ότι υπάρχει μια γραμμική σχέση καθώς και ότι η κατανομή του εισοδήματος είναι εξαιρετικά ασύμμετρη και με πιο βαριά ουρά σε σχέση με την Κανονική Κατανομή (Barabasi 2015).

Ο Vilfredo Pareto, θεώρησε ότι ανακάλυψε έναν καινούργιο τύπο «παγκόσμιου νόμου» καθώς κατάφερε να μετρήσει την ανισότητα στην κατανομή του εισοδήματος για πολλές χώρες (Geerolf 2016). Αυτό που ισχυρίστηκε είναι ότι λίγοι άνθρωποι κατέχουν ένα μεγάλο μέρος των χρημάτων σε σχέση με το μεγαλύτερο μέρος του πληθυσμού και για αυτό υπάρχει και η βαριά ουρά στην κατανομή του εισοδήματος. Αυτή η ανομοιότητα στις παρατηρήσεις είχε ως αποτέλεσμα να συνδέσει την κατανομή του εισοδήματος με την PLD, όπου για πρώτη φορά αναφέρεται (Barabasi 2015).

Όλα τα παραπάνω συνετέλεσαν στην διατύπωση του νόμου του Pareto 80/20, σύμφωνα με τον οποίο το 80% των χρημάτων βρίσκονται στο 20% του πληθυσμού. Ουσιαστικά, αυτός ο νόμος μας δηλώνει ότι το 80% των αποτελεσμάτων μας οφείλεται στο 20% των προσπαθειών μας.

Ο νόμος του Pareto εμφανίζεται σε πάρα πολλούς τομείς. Για παράδειγμα, το 80% των κερδών παράγεται από το 20% των εργαζομένων, το 80% των αποφάσεων παίρνεται από το 20% της συνάντησης. Το 80% των συνδέσεων στο Ίντερνετ αναφέρονται στο 15% των ιστοσελίδων, το 10% των πόλεων στις ΗΠΑ φιλοξενεί το 60% του συνολικού πληθυσμού της χώρας.

Ο μαθηματικός τύπος της PLD κάνει την εμφάνισή του σε πολλές εφαρμογές, παρόλα αυτά σε κάποιες περιπτώσεις αν και μιλάμε για την PLD εκείνη αναφέρεται με διαφορετικά ονόματα (Milojevic 2010). Αρχικά ο νόμος του Lotka ή Lotka’s Law είναι το πιο γνωστό παράδειγμα

της PLD. Ο Lotka το 1926 ανακάλυψε ότι ένας μικρός αριθμός συγγραφέων παράγει ένα μεγάλο αριθμό δημοσιεύσεων. Πιο συγκεκριμένα διαπίστωσε ότι το 60% των συγγραφέων κάνει μία δημοσίευση το 15% κάνει δύο και το 7% των συγγραφέων κάνει 3 δημοσιεύσεις (Friedman 2015).

Ακολούθως. υπάρχει και ο νόμος του Zipf ή Zipf's Law ο οποίος προέρχεται από την γλωσσολογία. Ο Zipf το 1949 μελέτησε πόσες φορές εμφανίζονται οι λέξεις σε ένα κείμενο από την πιο συχνή λέξη με την πιο σπάνια λέξη. Αυτό που ανακάλυψε είναι ότι η συχνότητα μιας οποιαδήποτε λέξης είναι αντιστρόφως ανάλογη της κατάταξης της στον πίνακα των συχνοτήτων. Η κατανομή αυτή είναι PLD καθώς υπάρχουν πολλές λέξεις οι οποίες εμφανίζονται σπάνια μέσα σε ένα κείμενο και υπάρχουν λίγες λέξεις τις οποίες τις συναντάμε πιο συχνά (Milojevic 2010).

Τέλος, υπάρχει και ο νόμος του Bradford ή Bradford's Law ο οποίος αναφέρεται στην αθροιστική συνάρτηση των περιοδικών που αναφέρονται σε ένα συγκεκριμένο θέμα. Όλες οι παραπάνω κατανομές είναι PLD και ανάλογα με την μορφή που έχουν ονομάζονται Κατανομή του Lotka ή Κατανομή του Zipf ή Κατανομή Pareto ή Κατανομή Bradford (Milojevic 2010).

1.3 ΠΕΡΙΓΡΑΦΗ ΚΕΦΑΛΑΙΩΝ

Στο δεύτερο κεφάλαιο, με ένα παράδειγμα που αναφέρεται στο μέγεθος των πόλεων στις ΗΠΑ προσπαθήσαμε να δείξουμε γιατί κατανομές όπως η Κανονική Κατανομή η οποία χρησιμοποιείται αρκετά συχνά για την περιγραφή δεδομένων, δεν μπορεί να περιγράψει δεξιά ασύμμετρες κατανομές με βαριά ουρά. Δίνουμε ακολούθως τον ορισμό και τον τύπο της PLD και στην συνέχεια αναφερόμαστε στη γραμμική σχέση που υπάρχει όταν λογαριθμίσουμε και τους δύο άξονες της συνάρτησης πυκνότητας πιθανότητας αλλά και της συνάρτησης επιβίωσης. Τέλος, αναφερόμαστε στις ροπές της αλλά και στο γεγονός ότι η PLD αναφέρεται και ως κατανομή χωρίς κλίμακα.

Στο τρίτο κεφάλαιο αρχικά, εξετάζουμε τις μεθόδους εκτιμήσεις των παραμέτρων της PLD και στην συνέχεια τον έλεγχο καλής προσαρμογής μεταξύ των δεδομένων μας και των δεδομένων που έχουμε παράγει από την PLD σύμφωνα με τις εκτιμήσεις των παραμέτρων που έχουμε βρει. Ακολούθως, προχωρούμε στις συγκρίσεις της PLD με εναλλακτικές κατανομές που έχουν παρόμοια συμπεριφορά με εκείνη.

Στο Τέταρτο κεφάλαιο, αρχικά ορίζουμε τη διακριτή PLD και παρουσιάζουμε διάφορες μεθόδους εκτίμησης των παραμέτρων της, οι οποίες δεν διαφέρουν και πολύ με αυτές της συνεχούς PLD. Ακολούθως, ορίζουμε εναλλακτικές κατανομές με παρόμοια συμπεριφορά με την συνεχή PLD, δηλαδή την Λογαριθμοκανονική Κατανομή, την Εκθετική Κατανομή, την Επεκτεινόμενη Εκθετική Κατανομή (*Stretched Exponential Distribution*) και τέλος την PLD με εκθετική αποκοπή (*PLD with exponential cutoff*).

Στο Πέμπτο κεφάλαιο, αρχικά προβαίνουμε σε μια σύγκριση των μεθόδων εκτίμησης της εκθετικής παραμέτρου της PLD και αναφερόμαστε στην δημιουργία τυχαίων αριθμών από την κατανομή αυτή. Ακολούθως, δημιουργούμε τρία σύνολα δεδομένων από την PLD και προσπαθούμε να εντοπίσουμε ποια εναλλακτική κατανομή μπορεί να προσαρμοστεί καλύτερα σε δεδομένα τα οποία προέρχονται από την PLD.

Στο Έκτο Κεφάλαιο, αναφερόμαστε αρχικά σε εφαρμογές που έχει χρησιμοποιηθεί η PLD. Το 1999 οι Albert. Jeong και Barabasi μελέτησαν πραγματικά δίκτυα και διαπίστωσαν ότι η εμφάνιση δύο χαρακτηριστικών, δηλαδή η αύξηση και προτίμηση σύνδεσης, είχαν σαν αποτέλεσμα η κατανομή των βαθμών στην ουρά της να έχει συμπεριφορά PLD. Οι εφαρμογές που αναπτύσσουμε έχουν να κάνουν με τον παγκόσμιο ιστό. το Ίντερνετ, το Δίκτυο αναφορών, το Δίκτυο τηλεφωνικών κλήσεων. το Δίκτυο ηλεκτρονικού ταχυδρομείου και τέλος το Δίκτυο των ηθοποιών. Στην συνέχεια γίνεται αναφορά στις γλώσσες προγραμματισμού R, Matlab και Python οι οποίες έχουν χρησιμοποιηθεί για την εκτίμηση των παραμέτρων της PLD.

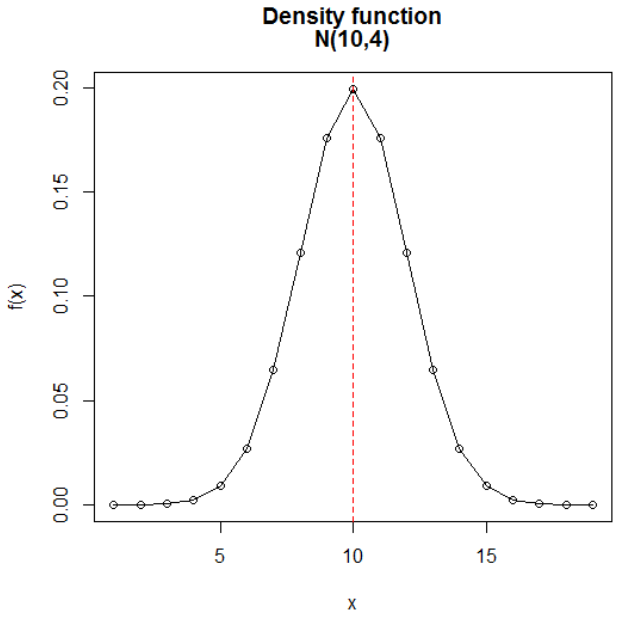
ΚΕΦΑΛΑΙΟ 2

ΟΡΙΣΜΟΙ-ΙΔΙΟΤΗΤΕΣ

2.1 ΕΙΣΑΓΩΓΗ

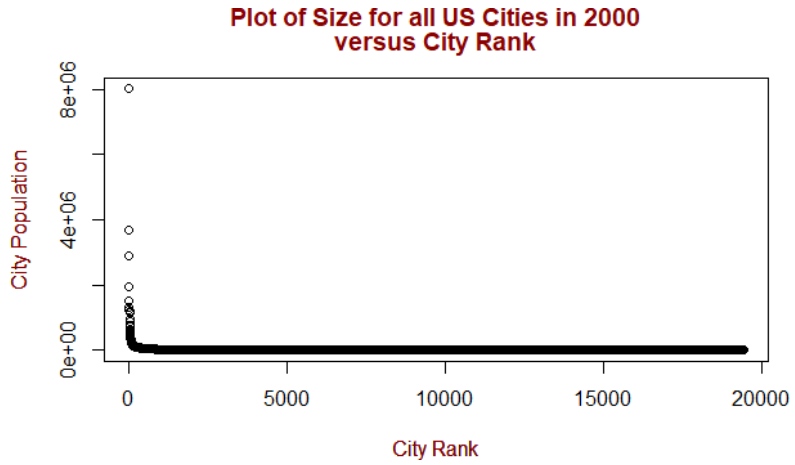
Πολλές εμπειρικές ποσότητες ομαδοποιούνται γύρω από μια τιμή η οποία είναι αντιπροσωπευτική για τις περισσότερες παρατηρήσεις. Έτσι ακόμα και αν υπάρχουν μεγάλες αποκλίσεις από αυτή την τιμή, το οποίο είναι σπάνιο φαινόμενο, οι κατανομές αυτές χαρακτηρίζονται καλά έχοντας δηλαδή μέση τιμή και τυπική απόκλιση (οι Clauset et al. (2009)*)*. Ένα απλό παράδειγμα είναι η Κανονική κατανομή, που περιγράφει τυχαίες μεταβλητές πραγματικών τιμών οι οποίες συγκεντρώνονται γύρω από μια μέση τιμή. Αυτό φαίνεται και στο παρακάτω γράφημα όπου δίνουμε την γραφική παράσταση της συνάρτησης πυκνότητας της Κανονικής κατανομής με μέση τιμή 10 και τυπική απόκλιση 2. Η καμπύλη είναι συμμετρική με το υψηλότερο σημείο να είναι στην μέση τιμή.

Σχήμα 2-1: Συνάρτηση Πυκνότητας της Κανονικής Κατανομής



Έστω τώρα, θέλουμε να εξετάσουμε το μέγεθος των πόλεων (Newman (2005)) των ΗΠΑ που έχουν πληθυσμούς 100,000 και μεγαλύτερους το έτος 2000 (US Census Bureau in 2000). Παίρνοντας τα δεδομένα για τις πόλεις αρχικά κατατάσσουμε τις παρατηρήσεις μας, N=242. με φθίνουσα σειρά. Ακολούθως ταξινομούμε τις πόλεις ανάλογα με τον πληθυσμό δηλαδή, το 1 θα αντιστοιχεί στην πόλη New York, το 2 στην πόλη Los Angeles κ.τ.λ. (Gabaix (2016)). Στην συνέχεια θα δημιουργήσουμε το γράφημα, με την βοήθεια της R, όπου στον άξονα των x θα τοποθετήσουμε την ταξινόμησης (ranks) και στον άξονα των y τους πληθυσμούς της αντίστοιχης πόλης και προκύπτει το παρακάτω γράφημα.

Σχήμα 2-2: Γράφημα του πληθυσμού των πόλεων το έτος 2000 έναντι των βαθμών των πόλεων



Το παραπάνω γράφημα, αρχικά, φαίνεται να είναι δεξιά ασύμμετρο (*right-skewed*) πράγμα που σημαίνει ότι το μεγαλύτερο μέρος της κατανομής προκύπτει από πολλά αλλά μικρά γεγονότα. Αυτό πρακτικά σημαίνει ότι, η μέση τιμή δεν είναι αντιπροσωπευτική για όλες τις παρατηρήσεις και πιο συγκεκριμένα υπάρχουν τιμές οι οποίες είναι πολύ πιο μεγάλες από την μέση τιμή (Shatnawi and Althebyan(2013)*)*. Αυτή η δεξιά ασυμμετρία είναι ποιοτικά διαφορετική από το ιστόγραμμα της Κανονικής Κατανομή. Επιπλέον, παρατηρούμε ότι η ουρά της εμπειρικής κατανομής φθίνει πιο αργά από ότι η Κανονική Κατανομή η οποία φθίνει εκθετικά.

Πολλές εμπειρικές κατανομές δεν ακολουθούν αυτό το μοτίβο, δηλαδή δεν χαρακτηρίζονται από μια μέση τιμή και μια τυπική απόκλιση αλλά οι παρατηρήσεις τους κυμαίνονται σε ένα μεγάλο δυναμικό εύρος (Newman(2005)). Στο παραπάνω γράφημα παρατηρείται ότι 243 πόλεις από τις 19,447 έχουν πληθυσμό μεγαλύτερο από την μέση τιμή. Για παράδειγμα ο πληθυσμός της Νέας Υόρκης είναι 8.008.654 ενώ η δειγματική μέση τιμή των δεδομένων μας είναι $\bar{x}$=9002.051. Συνεπώς, η δειγματική μέση τιμή μια πόλης στις ΗΠΑ δεν δίνει καμία ένδειξη ότι ένα σημαντικό μέρος του πληθυσμού μένει στην Νέα Υόρκη και στο Λος Άντζελες (Virkar and Clauset (2014)). Η ένδειξη ότι αυτή η κατανομή δεν εξηγείται καλά από την Κανονική Κατανομή είναι το γεγονός ότι η δειγματική τυπική απόκλιση είναι περίπου οχτώ φορές μεγαλύτερη από την μέση τιμή. Έτσι, εάν προσαρμόζαμε στα δεδομένα μας την Κανονική Κατανομή δεν θα αναμέναμε να δούμε πόλη τόσο μεγάλη όσο είναι η Νέα Υόρκη.

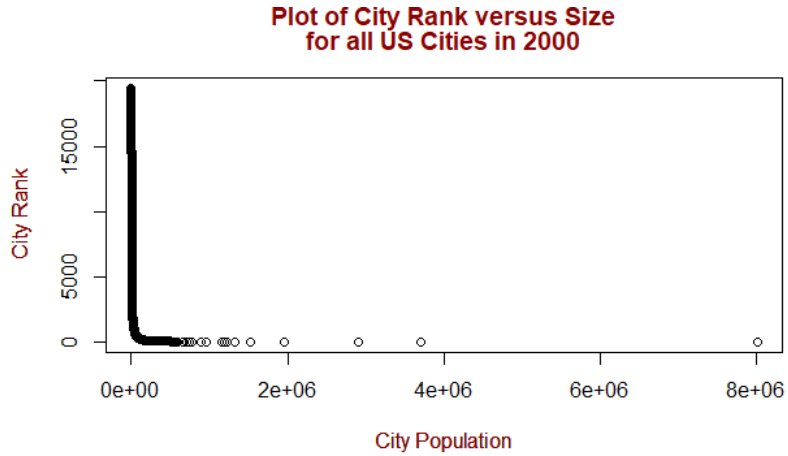
Πίνακας 2-1: Περιγραφικά στοιχεία των πληθυσμών των πόλεων

Number of Observations	Number of Observations greater than 99.216	Max	Min	Mean	Standard Deviation
19.447	**243**	**8.008.654**	**1**	**9002.051**	**77825.05**

Μια κατανομή η οποία μπορεί να περιγράψει τέτοια δεδομένα, είναι η Κατανομή Νόμου Δύναμης δηλαδή *Power-Law Distribution(PLD)*, η οποία περιγράφει ένα μεγάλο φάσμα από φαινόμενα στα οποία συμπεριλαμβάνονται οι σεισμοί (Bak and Tang(1989)), οι πυρκαγιές στο δάσος (Newman(2006)), ο αριθμός των αναφορών σε επιστημονικά άρθρα(Price (1965)) και ο αριθμός των επισκέψεων σε ιστοσελίδες (Adamic and Huberman (2000)).

Η PLD συχνά αποκαλείται και Κατανομή Pareto ή Νόμος του Zipf (Zipf's law), διότι αυτοί οι δύο ερευνητές την ανακάλυψαν σε κάποια από τα δεδομένα τους. Η Κατανομή Pareto και ο Zipf's law διαφέρουν μεταξύ τους στον τρόπο που σχεδιάζουμε το γράφημα. Το γράφημα 2.2 έχει σχεδιαστεί σύμφωνα με τον Pareto ενώ εάν γίνει αντιστροφή των αξόνων προκύπτει το γράφημα 2.3 σύμφωνα με τον Zipf (Newman 2005).

Σχήμα 2-3: Γράφημα των βαθμών των πόλεων των ΗΠΑ έναντι του μεγέθους των πόλεων το έτος 2000.



2.2 ΒΑΣΙΚΟΙ ΟΡΙΣΜΟΙ ΤΗΣ PLD

Στην Στατιστική, ο όρος Νόμος Δύναμης (*Power Law*) αποτελεί μια συγκεκριμένη συναρτησιακή σχέση μεταξύ δύο ποσοτήτων. Πιο συγκεκριμένα, μια αλλαγή στην μια ποσότητα αντιστοιχεί σε αναλογική αλλαγή στην άλλη ποσότητα. Μαθηματικά, μια ποσότητα ακολουθεί την PLD εάν προέρχεται από μια κατανομή πυκνότητας:

$$p(x) \propto x^{-a}. \tag{2.1}$$

όπου a είναι μια σταθερή παράμετρος της κατανομής η οποία ονομάζεται εκθετική παράμετρος ή παράμετρος κλίμακας (*scaling parameter*) και το $\propto$ σημαίνει ανάλογο. Σε πραγματικά δεδομένα συνήθως η εκθετική παράμετρος συνήθως βρίσκεται στην περιοχή $2 < a < 3$, παρόλα αυτά υπάρχουν και εξαιρέσεις (Clauset et al. (2009)). Στην πράξη λίγες εμπειρικές ποσότητες ακολουθούν την PLD για όλες τις τιμές του x. Πιο συχνά ισχύει να ακολουθούν την PLD από ένα κατώτερο όριο x_{min} και έπειτα. Σε αυτές τις περιπτώσεις λέμε ότι η ουρά της εμπειρικής κατανομής ακολουθεί την PLD.

Μια συνεχής *Power Law* κατανομή περιγράφεται από μια συνάρτηση πυκνότητας πιθανότητας τέτοια ώστε:

$$p(x)dx = \Pr(x \leq X < x + dx) = C \cdot x^{-a}dx \tag{2.2}$$

όπου X είναι η παρατηρούμενη τιμή και C είναι μια σταθερά (*normalization constant*). Όσο το $x \to 0$, η συνάρτηση αυτή αποκλίνει και αυτό έχει σαν αποτέλεσμα να υπάρχει ένα κατώτερο όριο x_{min} για το οποίο να ισχύει η συμπεριφορά *Power Law*.

Η σταθερά C ομαλοποίησης, ορίζεται εφόσον βρεθεί η εκθετική παράμετρος, και καθορίζεται από την απαίτηση ο λογάριθμος της συνάρτησης πυκνότητας πιθανότητας να ισούται με ένα:

$$1 = \int_{x_{min}}^{\infty} p(x)dx = C \cdot \int_{x_{min}}^{\infty} x^{-a}dx = \frac{C}{1-a}[x^{-a+1}]_{x_{min}}^{\infty} \Leftrightarrow$$

$$C = (a-1) \cdot x^{a-1} \tag{2.3}$$

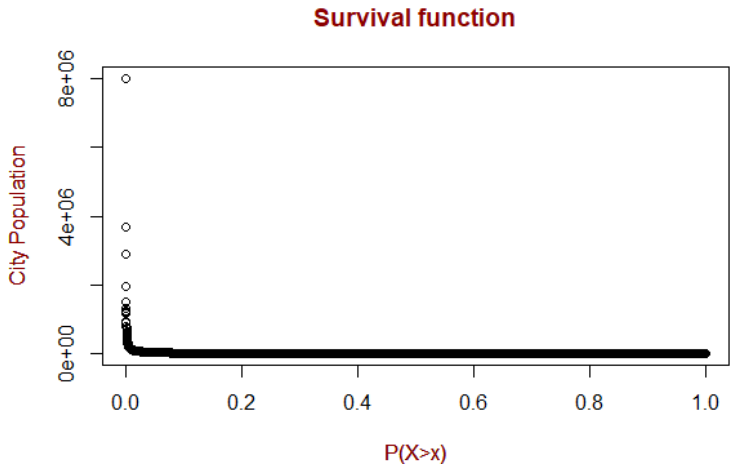
Αυτό που παρατηρούμε είναι ότι ο παραπάνω υπολογισμός είναι εφικτός για $\alpha > 1$, καθώς διαφορετικά, η δεξιά πλευρά της εξίσωσης αποκλίνει. Έτσι, για $\alpha > 1$, αντικαθιστώντας τον τύπο (2.3) στη συνάρτηση πυκνότητας πιθανότητας p(x), μπορούμε να πάρουμε την επόμενη έκφραση της συνεχούς PLD:

$$p(x) = \frac{a-1}{x_{min}} \cdot \left(\frac{x}{x_{min}}\right)^{-a}, \quad x \geq x_{min} \tag{2.4}$$

Σε πολλές περιπτώσεις χρησιμοποιείται η συνάρτηση επιβίωσης (survival function) της PLD. Η συνάρτηση επιβίωσης είναι η πιθανότητα μια τυχαία μεταβλητή X να έχει τιμή μεγαλύτερη από x, δηλαδή είναι το ποσοστό των παρατηρήσεων που έχουν τιμή μεγαλύτερη από x (Siganos et al. (2003)). Η παραπάνω συνάρτηση ορίζεται ως εξής:

$$P(x) = \Pr(X > x) = \int_x^{\infty} p(x)dx = \left(\frac{x}{x_{min}}\right)^{-a+1} \tag{2.5}$$

Σχήμα 2-4: Γράφημα της συνάρτησης επιβίωσης της PLD



Η *P(x)* είναι καλά ορισμένη για κάθε τιμή του *X* και προτιμάται για την σχεδίαση των δεδομένων καθώς δεν χάνουμε πληροφορία. Επιπλέον, ακολουθεί και αυτή τον νόμο της δύναμης και αυτό που αλλάζει είναι η εκθετική παράμετρος η οποία είναι α-1, δηλαδή εάν από την αρχική εκθετική παράμετρο αφαιρέσουμε την μονάδα. Είναι ένας εναλλακτικός τρόπος για να εξάγουμε πληροφορίες από την ουρά της κατανομής και όπως θα δούμε και στο τρίτο κεφάλαιο επιτρέπει την πιο ακριβή εκτίμηση του εκθέτη (Barabasi (2015)). Τέλος, αξίζει να σημειωθεί σε αυτό το σημείο ότι η μέθοδος που χρησιμοποιήσαμε για να δημιουργήσουμε τα γραφήματα 2.2 και 2.3 είναι κοντά στην κατασκευή της συνάρτησης επιβίωσης (Clauset et al. (2009)).

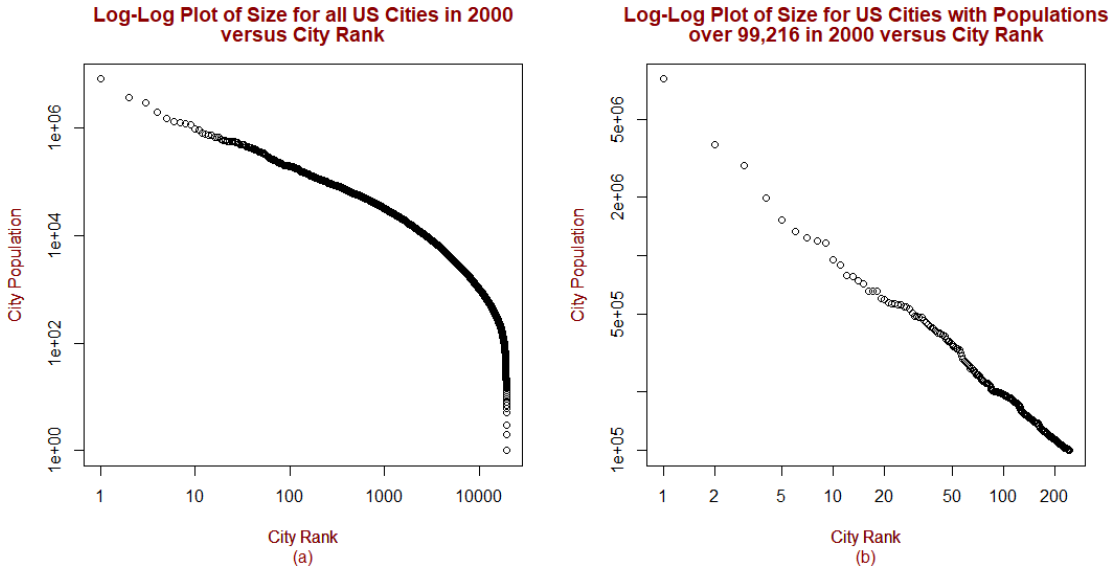
Πίνακας 2-2: Υπολογισμός της συνάρτησης επιβίωσης της PLD

X_i	f(x_i)	$p(x_i)$	$P(x_i)$
1	2	0.0001028436	1
2	2	0.0001028436	0.9999486
3	1	5.142181e-05	0.9998972
5	1	5.142181e-05	0.9998457
6	7	0.0003599527	0.9997943
…..			
1,517,550	1	5.142181e-05	2.571091e-04
1,953,633	1	5.142181e-05	2.056873e-04
2,896,047	1	5.142181e-05	1.542654e-04
3,694,742	1	5.142181e-05	1.028436e-04
8,008,654	1	5.142181e-05	5.142181e-05

2.3 ΓΡΑΜΜΙΚΗ ΣΧΕΣΗ

Εάν στο γράφημα 2.2 λογαριθμίσουμε και τους δύο άξονες δηλαδή φτιάξουμε το *log-log plot*, τότε όπως διαπιστώνεται και παρακάτω, η κατανομή φαίνεται να προσεγγίζεται από μία ευθεία γραμμή για τα $x \geq x_{min}$. Λογαριθμίζοντας, την εξίσωση (2.2) και στις δύο πλευρές βλέπουμε ότι η PLD ακολουθεί την ευθεία :

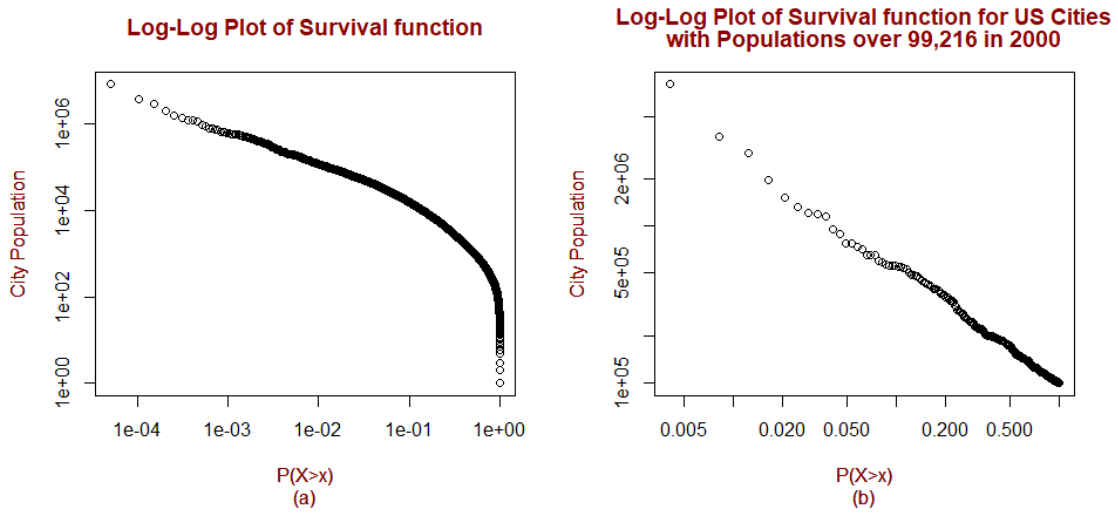
$$\ln\big(p(x)\big) = -a \cdot \ln(x) + c\,. \tag{2.5}$$



Σχήμα 2-5: (α) Log-log plot του γραφήματος 2.2 (b) Log-log plot των δεδομένων για τους πληθυσμούς μεγαλύτερους από το $x_{min} = 99{,}216$.

Από την παραπάνω σχέση διαπιστώνεται ότι υπάρχει μια γραμμική σχέση μεταξύ του λογαρίθμου της συνάρτησης πυκνότητας πιθανότητας και του λογάριθμου του μεγέθους των παρατηρήσεων. Η κλίση της ευθείας είναι η εκθετική παράμετρος α και το μείον που υπάρχει δηλώνει ότι η ευθεία έχει αρνητική κλίση. Η παραπάνω μέθοδος χρησιμοποιείται για να ανιχνεύσουμε συμπεριφορά *Power Law* στα δεδομένα μας, παρόλα αυτά πολλές φορές είναι παραπλανητική και δεν προτιμάται (Mitzenmacher(2004)). Την παραπάνω γραμμική σχέση την παρατηρούμε και εάν κατασκευάσουμε το γράφημα της συνάρτησης επιβίωσης καθώς ακολουθεί τον νόμο δύναμης (Σχήμα 2.6).

Η PLD μπορεί οπτικά να ανιχνευτεί μέσω της παραπάνω γραμμικής σχέσης σε *log-log plot*. Παρόλα αυτά, σπάνια μπορούμε να αναγνωρίσουμε με σιγουριά ότι μια PLD και για αυτό τον λόγο, στο τρίτο κεφάλαιο, υπάρχει λεπτομερής αναφορά στις στατιστικές τεχνικές που μας επιτρέπου να ισχυριστούμε την υπόθεση ότι οι παρατηρήσεις μας είναι συνεπείς στην PLD .



Σχήμα 2-6: (a) Log-log plot του γραφήματος 2.4 (b) Log-log plot της συνάρτησης επιβίωσης για πληθυσμούς πάνω από 99,216.

2.4 ΙΔΙΟΤΗΤΕΣ ΤΗΣ PLD

Υπάρχουν αρκετοί μηχανισμοί με τις οποίες μπορούμε να παράγουμε κατανομές βαριάς ουράς. Τα τελευταία χρόνια, πολλοί έχουν επικεντρωθεί στην PLD χάρη στις ενδιαφέρουσες μαθηματικές της ιδιότητες. Σε αυτή την ενότητα, αναφερόμαστε στην κατανομή βαριάς ουράς αλλά και στις ροπές της PLD οι οποίες λόγω της δεξιάς ασυμμετρίας δεν είναι όλες πεπερασμένες.

2.4.1 ΚΑΤΑΝΟΜΗ ΒΑΡΙΑΣ ΟΥΡΑΣ ΚΑΙ ΡΟΠΕΣ

Ένα χαρακτηριστικό γνώρισμα των συνηθισμένων κατανομών είναι ότι οι ουρές τους φθίνουν πιο γρήγορα, σε αντίθεση με κατανομές οι οποίες για μεγάλες τιμές του x φθίνουν πολύ πιο αργά και συναντάμε γεγονότα που χαρακτηρίζονται από μεγάλες τιμές (Barabasi(2015)). Αυτές οι κατανομές ονομάζονται κατανομές βαριάς ουράς (*heavy-tailed*

distribution) και έχουν τα λεγόμενα "κακά δεδομένα" λόγω των ακραίων σημείων (*outliers*). Με τον όρο ακραίες τιμές εννοούμε εκείνες που είναι πιο απομακρυσμένες από τις υπόλοιπες παρατηρήσεις. Για αυτό το λόγο στις κατανομές βαριάς ουράς η διακύμανση είναι πολύ πιο μεγάλη από ότι η μέση τιμή καθώς υπάρχει και μεγάλη διαφορά μεταξύ μέγιστης και ελάχιστης τιμής. Έτσι μέθοδοι που βασίζονται στο ελάχιστο, μέγιστο και στην μέση τιμή δεν περιγράφουν καλά τις ασύμμετρες κατανομές (Siganos (2003)).

Η PLD αντιπροσωπεύει το καλύτερο παράδειγμα από κατανομές βαριάς ουράς, διότι το $\frac{1}{x^a}$ μειώνεται πολύ πιο αργά όσο το x αυξάνεται με αποτέλεσμα ακραία γεγονότα να συμβαίνουν συχνότερα από ότι στην Κανονική Κατανομή (Easley (2010)). Σε γενικές γραμμές, σε μια PLD οι ουρές πέφτουν ασυμπτωτικά σύμφωνα με μια δύναμη α. Έτσι, το αποτέλεσμα είναι να έχουμε πιο βαριές ουρές από ότι άλλες κατανομές (Mitzenmacher (2004)). Σίγουρα στατιστικές τεχνικές στα δεδομένα μας εάν χρησιμοποιηθούν με προσοχή μπορούν να μας βοηθήσουν στην ερμηνεία των δεδομένων και να εκτιμήσουμε πόσο συχνά ακραία γεγονότα συμβαίνουν στο σύστημα (Stumpf and Porter (2012)).

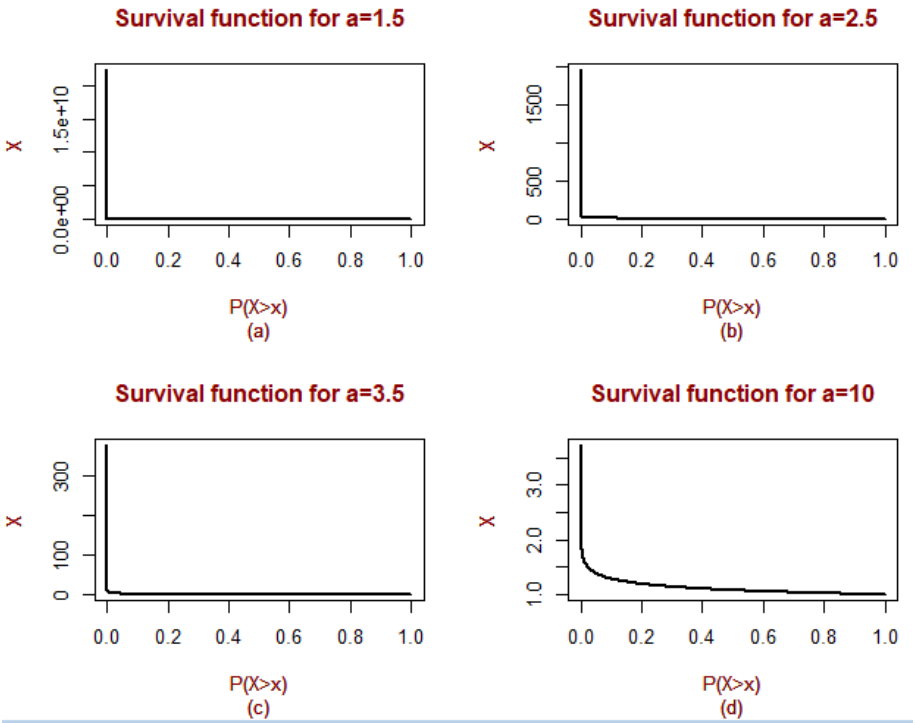
2.4.2 ΡΟΠΕΣ

Η PLD έχει έναν αριθμό από ιδιότητες οι οποίες είναι διαφορετικές από τις συνηθισμένες κατανομές. Αυτό γίνεται διότι η PLD δεν έχει όλες της τις ροπές πεπερασμένες και για αυτό τον λόγο χρησιμοποιείται για να μοντελοποιήσει το ακραίο κομμάτι της εμπειρικής κατανομής (Chatterjee (2014)). Δηλαδή, η εμπειρική ποσότητα X δεν παρουσιάζει μικρή μεταβλητότητα και δεν συγκεντρώνεται γύρω από τον μέσο όρο (Willinger et al. (2004)). Ένα στάδιο πριν τον υπολογισμό των ροπών είναι η εκτίμηση της εκθετικής παραμέτρου αλλά και του κατώτερου σημείου.

Να τονίσουμε σε αυτό το σημείο ότι η εκτίμηση της εκθετικής παραμέτρου είναι σημαντική καθώς μας δείχνει τον βαθμό ανισότητας της κατανομής. Μια μικρή τιμή του εκθέτη *α* σημαίνει ότι υπάρχει μεγάλος βαθμός ανισότητας στα δεδομένα (Gabaix (2016)). Πιο συγκεκριμένα, αυτό σημαίνει ότι υπάρχει μεγάλη πιθανότητα να βρούμε πολύ μεγάλα γεγονότα (παρατηρήσεις). Επιπλέον, όσο πιο μικρός είναι ο εκθέτης τόσο πιο μεγάλες είναι και οι ουρές (Gabaix (2008)). Παρακάτω έχουμε δημιουργήσει γραφήματα της συνάρτησης επιβίωσης για

διάφορες τιμές του a και παρατηρούμε ότι για όσο πιο μικρές είναι οι τιμές της εκθετικής παραμέτρου ισχύουν τα παραπάνω.

Σχήμα 2-7: Η συνάρτηση επιβίωσης της PLD για διάφορες τιμές του α



Η μέση τιμή της PLD είναι υπολογίζεται από τον παρακάτω τύπο:

$$E(X) = \int_{x_{min}}^{\infty} x \cdot p(x)dx = C\int_{x_{min}}^{\infty} x^{-a+1}dx = \frac{C}{2-a} \cdot [x^{-a+2}]_{x_{min}}^{\infty} \Leftrightarrow$$

$$\Leftrightarrow E(X) = \frac{a-1}{a-2} \cdot x_{min} \quad (2.6)$$

Η παραπάνω σχέση ξεδιαλύνει το γεγονός της απόκλισης της μέση τιμής για $\alpha \leq 2$. Εάν το σύνολο δεδομένων μας είναι πεπερασμένο, η μέση τιμή θα είναι πεπερασμένη. Παρόλα αυτά, όσο πηγαίνουμε σε πιο μεγάλα σύνολα δεδομένων η μέγιστη τιμή θα αυξάνεται και έτσι δεν θα υπάρχει κάποιο ανώτερο όριο (Newman (2005)).

Η ροπή δεύτερης τάξης. δηλαδή η μέση τιμή του x^2, δίνεται από τον παρακάτω τύπο:

$$E(X^2) = \int_{x_{min}}^{\infty} x^2 \cdot p(x)dx = \int_{x_{min}}^{\infty} x^{-a+2}dx = C \cdot [x^{-a+3}]_{x_{min}}^{\infty} \Leftrightarrow$$

$$\Leftrightarrow E(X^2) = \frac{a-1}{a-3} \cdot x_{min}^2. \quad (2.7)$$

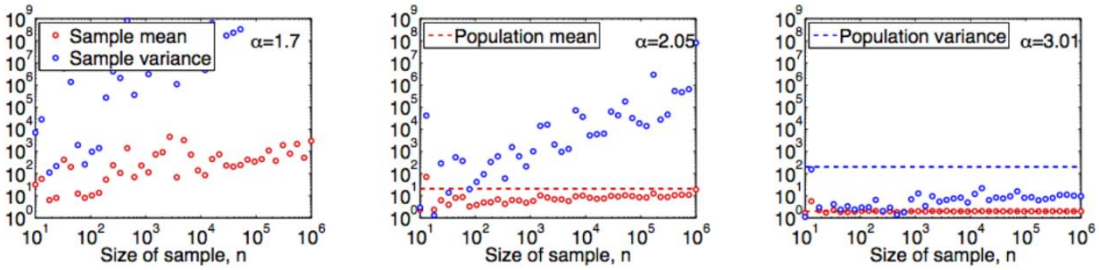
Αυτό που παρατηρείται και εδώ είναι ότι εάν η εκθετική παράμετρος πάρει τιμές μικρότερες ή ίσες από 3 ο παραπάνω τύπος θα αποκλίνει με αποτέλεσμα η ροπή δεύτερης τάξης να είναι μη πεπερασμένη. Σε αυτή την περίπτωση η $\langle x_{max} \rangle$ θα εξακολουθεί να αποκλίνει, παρόλα αυτά, δεν θα κυριαρχεί στην μέση τιμή (Newman(2005)).

Η συνέπεια της άπειρης μέσης τιμής και τυπικής απόκλισης είναι ότι δεν μπορεί να εφαρμοσθεί το Κεντρικό Οριακό Θεώρημα για αυτές τις κατανομές. Έτσι, συμπεράσματα βασιζόμενα στην δειγματική μέση τιμή και τυπική απόκλιση είναι υπό αμφισβήτηση (Baxter et al. (2006)).

Τα παραπάνω αποτελέσματα μπορούν να γενικευτούν και να ισχυριστούμε ότι γενικά όλες οι ροπές $E(x^m)$ υπάρχουν για $m < a-1$ και όλες οι ροπές που είναι μεγαλύτερες αποκλίνουν. Οι ροπές οι οποίες δεν αποκλίνουν θα είναι της παρακάτω μορφής:

$$E(x^m) = \frac{a-1}{a-1-m} \cdot x_{min}^m \tag{2.8}$$

Ο βαθμός ανισότητας στα δεδομένα απεικονίζεται στα παρακάτω γραφήματα. Στο πρώτο γράφημα η εκθετική παράμετρος έχει την τιμή 1.7 και όπως αναφέραμε και παραπάνω οι ροπές πρώτης και δεύτερης τάξης θα είναι μη πεπερασμένες και αυτό απεικονίζεται με διασπαρμένα σημεία στο γράφημα. Στο δεύτερο γράφημα η εκθετική παράμετρος έχει τιμή ίση με 2.05, πράγμα που σημαίνει πως η ροπή πρώτης τάξης είναι πεπερασμένη και τα σημεία των δειγματικών μέσων τιμών σχηματίζουν μια ευθεία. Αυτό δεν ισχύει για την δειγματική διακύμανση καθώς η ροπή δεύτερης τάξης είναι μη πεπερασμένη. Τέλος. όταν η εκθετική παράμετρος είναι μεγαλύτερη του 3 και οι ροπές αυτές είναι πεπερασμένες θα σχηματίζουν ευθεία και οι δειγματικές μέσες τιμές αλλά και οι δειγματικές διακυμάνσεις.



mael

Σχήμα 2-8: Δειγματικές μέσες τιμές και δειγματικές διακυμάνσεις για διάφορες τιμές της εκθετικής παραμέτρου και πλήθος παρατηρήσεων.

2.5 ΚΑΤΑΝΟΜΕΣ ΧΩΡΙΣ ΚΛΙΜΑΚΑ

Η PLD ονομάζεται πολλές φορές και κατανομή χωρίς κλίμακα (*scale-free distribution*). Αυτό πρακτικά σημαίνει ότι η κατανομή είναι η ίδια ανεξάρτητα από την κλίμακα που την βλέπουμε (Newman(2005)). Η πηγή της ονομασίας *scale-free* προκύπτει από το γεγονός ότι σε κάποιο κρίσιμο σημείο (*critical point or continuous phase transition*) κάποιο χαρακτηριστικό της κλίμακας όπως η μέση τιμή και η διακύμανση αποκλίνει. Αυτό έχει ως αποτέλεσμα, το σύστημα να μένει χωρίς καθορισμένο μέγεθος-κλίμακα.

Η παραπάνω ιδιότητα είναι ευαίσθητη στην τιμή που θα πάρει ο εκθέτης καθώς όπως διαπιστώθηκε και παραπάνω για α < 2 έχουμε μη-πεπερασμένη μέση τιμή ενώ για $2 \leq \alpha < 3$ έχουμε μη-πεπερασμένη διακύμανση. Αυτό σημαίνει ότι, όταν τυχαία επιλέξουμε μια παρατήρηση δεν ξέρουμε αν η τιμή της θα είναι μεγάλη ή μικρή καθώς δεν υπάρχει κάποια εσωτερική κλίμακα (Barabasi 2015)).

Εάν κάποιος αλλάξει την κλίμακα από x σε cx, δηλαδή η κλίμακα πολλαπλασιαστεί με κάποιον παράγοντα, τότε έχουμε τα εξής:

$$p(c \cdot x) = C \cdot (c \cdot x)^{-a} p(c \cdot x) = c^{-a} \cdot p(x) \iff$$

$$p(c \cdot x) \propto p(x) \tag{2.9}$$

Συμπεραίνουμε ότι οι παραπάνω συναρτήσεις πυκνότητας πιθανότητας είναι ανάλογες και η αρχική PLD πολλαπλασιάζεται με τον παράγοντα c^{-a}. Έτσι οι σχετικές πιθανότητες μεταξύ σπάνιων και μη-σπάνιων γεγονότων παραμένουν ίδιες ανεξάρτητα της κλίμακας (Chatterje(2014)). Αυτό που παρατηρείται είναι ότι, λίγα γεγονότα με υψηλές τιμές συνυπάρχουν με πολλά γεγονότα με χαμηλές τιμές (Barabasi(2015)). Τα μεγάλα γεγονότα δεν έχουν ποιοτικά διαφορετική ερμηνεία από ότι τα μικρά γεγονότα (Clauset et al. (2007)). Επιπρόσθετα. η *scale-free* συμπεριφορά σημαίνει ότι η αναλογία των μικρών και των μεγάλων γεγονότων διατηρείται κάθε φορά που το σύστημα εξελίσσεται (Shatnawi and Althebyan (2013)).

Κ Ε Φ Α Λ Α Ι Ο 3

ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ

Οι PLD έχουν προσελκύσει το ευρύ επιστημονικό ενδιαφέρον για τις μαθηματικές τους ιδιότητες αλλά και το γεγονός ότι εμφανίζονται σε πολλά φυσικά ή και ανθρώπινα συστήματα. Παρόλα αυτά, ο εντοπισμός τους μπορεί να αποβεί δύσκολος λόγω των μεγάλων διακυμάνσεων (fluctuations) που λαμβάνουν χώρα στην ουρά της εμπειρικής κατανομής. Σε αυτό το κεφάλαιο, θα παρουσιάσουμε ένα σύνολο στατιστικών μεθόδων όπως πρότειναν οι Clauset et al. (2009) για την διερεύνηση της υπόθεσης PL. Αρχικά, θα βρούμε το καλύτερο μοντέλο PL κάνοντας εκτίμηση της εκθετικής παραμέτρους αλλά και του κάτω ορίου. Ακολούθως, με τον έλεγχο καλής προσαρμογής Kolmogorov-Smirnov θα κάνουμε έλεγχο της αξιοπιστίας του μοντέλου PL αλλά και σύγκριση της PLD έναντι άλλων κατανομών βαριάς ουράς μέσω του ελέγχου του λόγου πιθανοφανειών.

3.1 ΕΚΤΙΜΗΣΗ ΤΟΥ ΚΑΤΩΤΕΡΟΥ ΟΡΙΟΥ

Τα εμπειρικά δεδομένα συνήθως ακολουθούν την PLD από ένα σημείο και πάνω, δηλαδή η κεφαλή της κατανομής αποκλίνει από την PLD και το μέρος της κατανομής που εξετάζουμε είναι η ουρά της (Goldstein et al. (2004)). Γι' αυτόν τον λόγο θα εστιάσουμε σε εκείνη την περιοχή. Επιπρόσθετα, όπως θα διαπιστωθεί και παρακάτω, για την εκτίμηση της εκθετικής παραμέτρου είναι απαραίτητος ο προσδιορισμός του κάτω ορίου x_{min}.

Η εύρεση του κατάλληλου κάτω σημείου είναι πολύ σημαντική. Εάν επιλεγεί μια χαμηλή τιμή τότε θα έχουμε μια μεροληπτική παράμετρο καθώς το μοντέλο θα προσαρμόζεται σε δεδομένα που δεν έχουν την PL συμπεριφορά. Από την άλλη πλευρά, εάν επιλέξουμε μεγάλη τιμή για το κάτω όριο υπάρχει κίνδυνος να παραλείψουμε αρκετά δεδομένα με αποτέλεσμα να υπάρξει απώλεια πληροφορίας. Συνεπώς αυτό που χρειαζόμαστε είναι μια μέθοδο που να περιλαμβάνει και τους δύο περιορισμούς. Η τιμή του x_{min} που θα επιλεγεί θα πρέπει να

ελαχιστοποιεί τις διαφορές μεταξύ της εμπειρικής συνάρτησης κατανομής των δεδομένων και του μοντέλου που έχει την καλύτερη προσαρμογή.

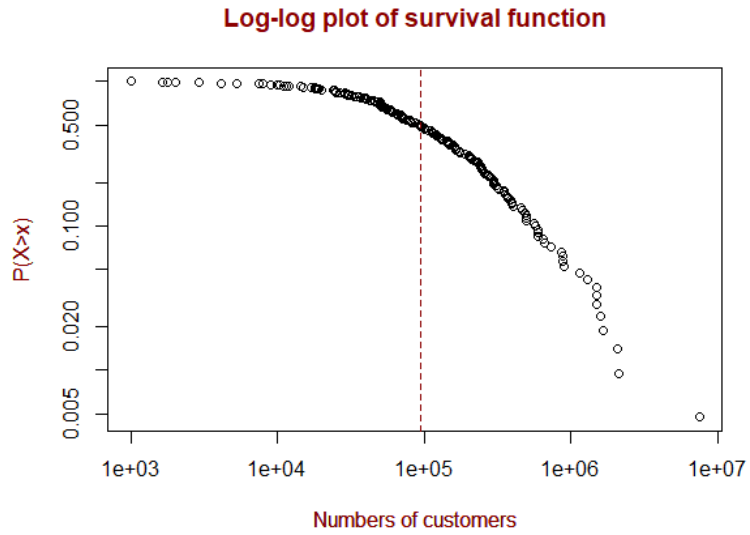
Για την εύρεση του κάτω ορίου έχουν προταθεί κάποιες μέθοδοι εκτίμησης. Η πιο συνηθισμένη και γρήγορη μέθοδος είναι στο γράφημα της συνάρτησης επιβίωσης με λογαριθμισμένους άξονες να εντοπίσουμε πότε η εμπειρική κατανομή γίνεται ευθεία. Επιπλέον, μπορούμε να κατασκευάσουμε το γράφημα της εκθετικής παραμέτρου συναρτήσει του x_{min} και να διερευνήσουμε πάνω από ποια τιμή η συνάρτηση γίνεται σταθερή. Παρόλα αυτά, το σύνηθες γράφημα με τα ποσοστημόρια της κατανομής (*QQ-plot*) ενδέχεται να δώσει καλύτερη εικόνα του πραγματικού μοντέλου. Σε κάθε περίπτωση οι παραπάνω μέθοδοι είναι υποκειμενικές και δεν προτιμώνται στην πράξη.

Μία λογική επιλογή μεθόδου εκτίμησης είναι η μέθοδος της μέγιστης πιθανοφάνειας. Δυστυχώς, δεν ενδείκνυται η συγκεκριμένη μέθοδος καθώς όσο αυξάνουμε την τιμή του κάτω ορίου τόσο μειώνεται ο αριθμός των παρατηρήσεων στο δείγμα μας (Clauset(2011)). Η μέθοδος η οποία συνήθως χρησιμοποιείται βασίζεται σε απόσταση κατανομών. Υπάρχει μια πληθώρα μέτρων για τον υπολογισμό απόστασης δύο συναρτήσεων πυκνότητας, αλλά για μη κανονικά δεδομένα η συνηθέστερη είναι η K-S μέθοδος (*Kolmogorov-Smirnov statistic*), η οποία δίνεται από τον τύπο:

$$D = \sup_x |S(x) - P(x)| \tag{3.1}$$

όπου, για $x_i \geq x_{min}$, η $S(x)$ είναι η συνάρτηση επιβίωσης των δεδομένων μας και $P(x)$ είναι η συνάρτηση επιβίωσης του υπό εξέταση μοντέλου.

Εάν το x_{min} έχει μικρή τιμή η οποία αντιστοιχεί στο κομμάτι της εμπειρικής κατανομής που δεν έχει την PL συμπεριφορά τότε η ποσότητα D θα είναι μεγάλη καθώς το μοντέλο της PL δεν περιγράφει καλά τα δεδομένα. Ομοίως, εάν η τιμή του x_{min} είναι μεγάλη, το μέγεθος του δείγματος θα είναι μικρό με αποτέλεσμα η ποσότητα D να είναι μεγάλη. Όταν η τιμή του x_{min} συμπίπτει με την αρχή της PL συμπεριφοράς τότε θα είναι μικρή. Η εκτίμηση του x_{min} είναι η τιμή εκείνη η οποία ελαχιστοποιεί την ποσότητα D (Virkar και Clauset (2014)). Το παρακάτω γράφημα αντιστοιχεί στην συνάρτηση επιβίωσης του συνόλου δεδομένων των ηλεκτρικών διακοπών. Με αυτό το παράδειγμα, θα ερευνήσουμε τους αριθμούς των πελατών που επλήγησαν από ηλεκτρικές διακοπές στις Η.Π.Α μεταξύ 1984 και 2002 (Newman(2006)). Με την βοήθεια της R, διαπιστώθηκε ότι για D(94,285)= 0.02227983 έχουμε την μικρότερη απόσταση μεταξύ εμπειρικής και PLD με α= 1.891775 και x_{min}=94,285.



Σχήμα 3-1: Log-Log plot της συνάρτησης επιβίωσης των αριθμών των πελατών που επλήγησαν από ηλεκτρικές διακοπές.

Να τονίσουμε σε αυτό το σημείο ότι η KS μέθοδος, είναι λίγο συντηρητική καθώς μπορεί να δώσει σχετικά μεγαλύτερες εκτιμήσεις για το κάτω όριο. Στη βιβλιογραφία υπάρχουν και άλλες προτεινόμενες μέθοδοι, όπως παραλλαγές της KS μεθόδου. το τεστ καλής προσαρμογής του Kuiper και των Anderson-Darling. Τα αποτελέσματα από το τεστ καλής προσαρμογής του Kuiper δίνουν παρόμοια αποτελέσματα με το KS ενώ το στατιστικό τεστ των Anderson-Darling δίνει πολύ μεγάλες τιμές για το x_{min}.

3.2 ΕΚΤΙΜΗΣΗ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ ΤΟΥ ΕΚΘΕΤΗ

Πολλές έρευνες οι οποίες αναφέρονται σε εμπειρικές κατανομές PL χρησιμοποιούν αδύναμες στατιστικές μεθόδους για την εκτίμηση της εκθετικής παραμέτρου α. Η σταθερά α μπορεί να εκτιμηθεί με πολλούς τρόπους όπως η μέθοδος ελαχίστων τετραγώνων και η μέθοδος μέγιστης πιθανοφάνειας (MLE). Η μέθοδος ελαχίστων τετραγώνων χρησιμοποιήθηκε από τον Pareto στις κατανομές του πλούτου(Arnold (1983)) καθώς και από τον Richardson στην ανάλυση του μεγέθους των πολέμων στον 20^{ο} αιώνα(Richardson (1960)).

Η μέθοδος μέγιστης πιθανοφάνειας σύμφωνα με τους Clauset et al. (2009) παράγει πιο ακριβείς εκτιμήσεις καθώς και οι Goldstein et al. (2004) έδειξαν σε έρευνά τους ότι η μέθοδος μέγιστης πιθανοφάνειας είναι πιο ισχυρή από ότι οι γραφικές μέθοδοι που βασίζονται στην γραμμική προσαρμογή αφού λογαριθμίσουμε τους άξονες. Στην συνέχεια θα δώσουμε

αναφορές για τις διάφορες γραφικές μεθόδους εκτίμησης της εκθετικής παραμέτρου χρησιμοποιώντας γραμμική παλινδρόμηση με την μέθοδο των ελαχίστων τετραγώνων και θα καταλήξουμε στην πιο αξιόπιστη και ισχυρή εκτίμηση με την μέθοδο της μέγιστης πιθανοφάνειας.

Μέχρι πρόσφατα, για το αν η εμπειρική κατανομή ακολουθεί PL συμπεριφορά καθορίζεται από την οπτική εξέταση σε διπλά λογαρθμισμένους άξονες. Παρόλα αυτά, η παραπάνω προσέγγιση είναι αδύναμη στο να αναγνωρίσει σημαντικές διαφορές της PLD και των άλλων εναλλακτικών κατανομών και πόσο μάλλον όταν το μέγεθος του δείγματος είναι αρκετά μικρό (Klaus et al. (2011)). Ο πιο απλός και συνηθέστερος τρόπος ταυτοποίησης της PL συμπεριφοράς στις εμπειρικές κατανομές είναι, όπως αναφέρθηκε και στο δεύτερο κεφάλαιο, η ευθεία γραμμή που σχηματίζεται αφού λογαριθμίσουμε τους δύο άξονες.

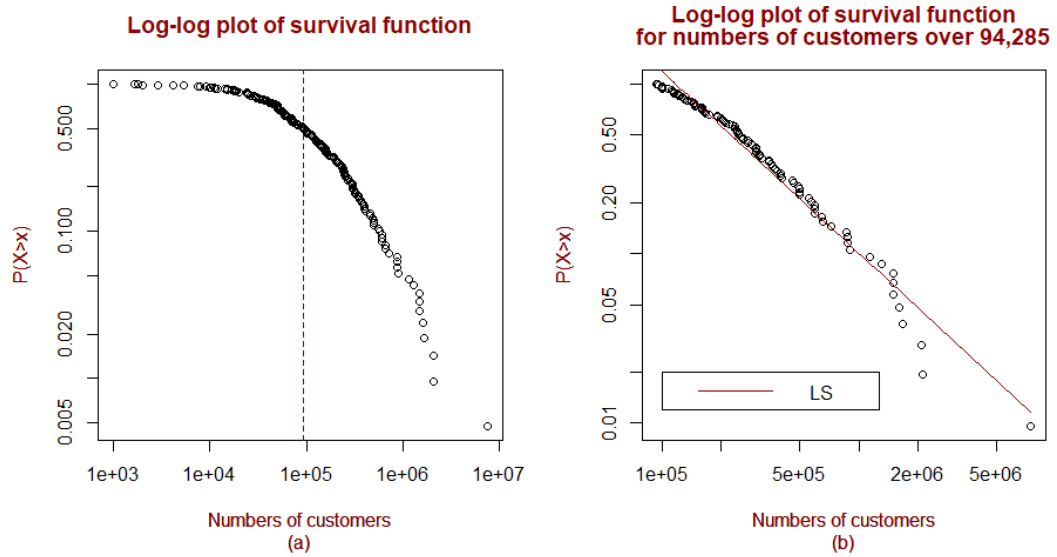
Όπως αναφέρθηκε και παραπάνω, μια μέθοδος εκτίμησης της εκθετικής παραμέτρου είναι η μέθοδος των ελαχίστων τετραγώνων. Πολλοί ερευνητές χρησιμοποιούν τέσσερις γραφικές μεθόδους για την εκτίμηση της εκθετικής παραμέτρου με την παραπάνω μέθοδο. Η πρώτη αναφέρεται στην εφαρμογή της μεθόδου ελαχίστων τετραγώνων στο γράφημα των δεδομένων με διπλούς λογαριθμικούς άξονες .

Η παραπάνω μέθοδος, παρόλα αυτά, δεν λαμβάνει υπόψη ότι σχεδόν το μεγαλύτερο μέρος των δεδομένων παρατηρείται στα πρώτα σημεία της κατανομής και ότι στην ουρά της κατανομής υπάρχει θόρυβος ως αποτέλεσμα του μικρού αριθμού των παρατηρήσεων. Σύμφωνα με τους Bauke (200,7) Goldstein et al. (2004), Pueyo and Jovani (2006) η μέθοδος των ελαχίστων τετραγώνων στο γράφημα των δεδομένων δεν είναι αξιόπιστη. Σαν συνέπεια αυτών, προτάθηκαν δύο παραλλαγές στην απευθείας γραμμική παλινδρόμηση. Πιο συγκεκριμένα, η μία παραλλαγή είναι η χρησιμοποίηση των πέντε πρώτων σημείων για απευθείας παλινδρόμηση και η δεύτερη η λογαριθμική ομαδοποίηση των δεδομένων (logarithmically binned data) (Goldstein et al. (2004)).

Η δεύτερη παραλλαγή βασίζεται στην προσθήκη όλων των τιμών που εμπίπτουν σε ομάδες(bins) οι οποίες έχουν λογαριθμική απόσταση. Το πλεονέκτημα αυτής της μεθόδου είναι ότι ομαδοποιώντας τα δεδομένα κατά αυτόν τον τρόπο μειώνεται και ο θόρυβος. Αυτό οφείλεται από το γεγονός ότι, οι ομάδες που βρίσκονται στην ουρά της κατανομής έχουν περισσότερα δείγματα και μειώνονται τα στατιστικά σφάλματα(Newman(2006)).

Τέλος, μια άλλη μέθοδος, η οποία είναι ανώτερη από πολλές απόψεις, είναι η γραμμική προσαρμογή στο γράφημα της συμπληρωματικής αθροιστικής συνάρτησης κατανομής. Όπως

αναφέρθηκε και στο δεύτερο κεφάλαιο, η συνάρτηση επιβίωσης ακολουθεί PLD με διαφορετικό εκθέτη δηλαδή, α-1. Με την παραπάνω συνάρτηση, δεν χρειάζεται να κάνουμε ομαδοποίηση στα δεδομένα καθώς εξ ορισμού είναι καλά ορισμένη και όλη η πληροφορία βρίσκεται στο γράφημα.



Σχήμα 3-2: (a) Log-log plot της συνάρτησης επιβίωσης των αριθμών των πελατών που επλήγησαν από τις ηλεκτρικές διακοπές με $x_{min} = 94.285$. (b) Log-log plot της συνάρτησης επιβίωσης όταν ο αριθμός των πελατών ξεπέρασε τους 94.285. Σε αυτό το γράφημα έχουμε προσαρμόσει και ευθεία ελαχίστων τετραγώνων.

Στο δεύτερο γράφημα έχουμε εφαρμόσει την μέθοδο των ελαχίστων τετραγώνων για $x_i \geq x_{min}$ με συντελεστή προσδιορισμού αρκετά υψηλό, δηλαδή $R^2 = 0.972$. Η ευθεία ελαχίστων τετραγώνων που εφαρμόζεται καλύτερα στα δεδομένα, σύμφωνα με τα αποτελέσματα που μας έδωσε η R, είναι η $\ln(P(x)) = 12.459 - 1.071 \ln(x)$. Συνεπώς η εκτίμηση της εκθετικής παραμέτρου α προκύπτει ως εξής α=1.071+1=2.071.

Στις παραπάνω μεθόδους που αναπτύξαμε δεν αντιμετωπίζεται πλήρως το πρόβλημα του θορύβου στην ουρά και σαν συνέπεια δεν έχουμε ακριβή αποτελέσματα εκτίμησης της παραμέτρου α. Επιπλέον, τα σφάλματα είναι δύσκολο να εκτιμηθούν καθώς δεν ακολουθούν τις συνθήκες που απαιτούν οι συνήθεις τύποι της παλινδρόμησης. Με τις μεθόδους γραμμικής παλινδρόμησης, οι εκτιμήσεις δεν ικανοποιούν βασικές απαιτήσεις των κατανομών πιθανοτήτων όπως η κανονικοποίηση (normalization). Ακόμα, όσον αφορά την συνάρτηση επιβίωσης, δεν ισχύει η υπόθεση της ανεξαρτησίας διότι οι γειτονικές τιμές είναι ισχυρά

συσχετισμένες. Τέλος, ακόμα και αν τα δεδομένα μας δεν ακολουθούν την PLD, υπάρχει ο κίνδυνος να λάβουμε μεγάλο R^2, συνεπώς μεγάλες τιμές στον συντελεστή προσδιορισμού δεν θα πρέπει να λαμβάνονται ως ένδειξη υπέρ της PL μορφής (Clauset et al. (2009)).

Μια εναλλακτική μέθοδος, η οποία χρησιμοποιείται για την προσαρμογή παραμετροποιημένων μοντέλων (parameterized models) σε εμπειρικά δεδομένα, είναι η εκτίμηση με την μέθοδο της μέγιστης πιθανοφάνειας η οποία δίνει πιο ακριβείς εκτιμήσεις παραμέτρων για μεγάλο μέγεθος δείγματος (Virkar και Clauset(2014)). Στην συγκεκριμένη μέθοδο συναντάμε επιθυμητά χαρακτηριστικά, δηλαδή, ότι ο εκτιμητής υπάρχει και είναι μοναδικός και επιπλέον είναι συνεπής που αυτό πρακτικά σημαίνει πώς όσο το $n \to \infty$ τόσο θα προσεγγίζουμε την πραγματική τιμή. Στην περίπτωση των συνεχών δεδομένων ο εκτιμητής μέγιστης πιθανοφάνειας της εκθετικής παραμέτρου έχει προέλθει από τον Muniruzzaman (1957). Έχοντας ορίσει την συνάρτηση πυκνότητας πιθανότητας (2.2) η πιθανοφάνεια των δεδομένων ορίζεται ως εξής:

$$p(x|a) = \prod_{i=1}^{n} \frac{a-1}{x_{min}} \cdot \left(\frac{x_i}{x_{min}}\right)^{-a} \tag{3.2}$$

Υποθέτοντας ότι, οι παρατηρήσεις μας προέρχονται από την PLD πάνω από την τιμή x_{min}, η συνάρτηση του λογαρίθμου της πιθανοφάνειας είναι:

$$L = \ln(p(x|a) = ln \prod_{i=1}^{n} \frac{a-1}{x_{min}} \cdot \left(\frac{x_i}{x_{min}}\right)^{-a}$$

$$= \sum_{i=1}^{n} \left[\ln(a-1) - lnx_{min} - a \cdot ln \frac{x_i}{x_{x\,min}}\right]$$

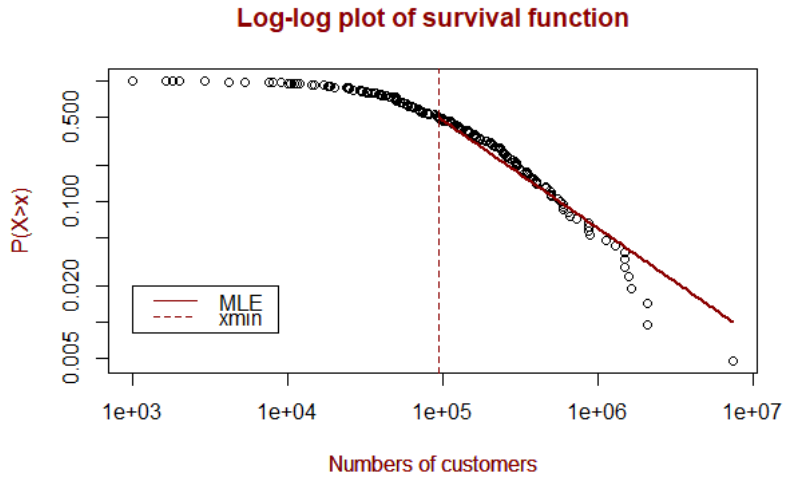
$$= n \cdot \ln(a-1) - n \cdot lnx_{min} - a \cdot \sum_{i=1}^{n} ln \frac{x_i}{x_{min}}. \tag{3.3}$$

Θέτοντας $\frac{\partial L}{\partial a} = 0$ και λύνοντας ως προς a θα λάβουμε την εκτίμηση μέγιστης πιθανοφάνειας. Ο λογάριθμος πιθανοφάνειας χρησιμοποιείται διότι απλοποιεί τον υπολογισμό, καθώς και ότι η συνάρτηση του λογαρίθμου είναι μια αύξουσα και μονότονη συνάρτηση και δεν διαταράσσει το σημείο που αποκτάται το μέγιστο (Goldstein et al. (2004)).

Ο εκτιμητής μέγιστης πιθανοφάνειας (*MLE*) για την συνεχή PLD είναι ο παρακάτω:

$$\hat{\alpha} = 1 + n\left[\sum_{i=1}^{n} ln\frac{x_i}{x_{min}}\right]^{-1} \quad (3.4)$$

όπου x_{min} είναι οι παρατηρούμενες τιμές της ποσότητας x τέτοιες ώστε $x_i \geq x_{min}$. Το παρακάτω γράφημα απεικονίζει την συνάρτηση επιβίωσης σε λογαριθμισμένους άξονες καθώς και την μέγιστη πιθανοφάνεια της PLD. Όπως προκύπτει η εκτίμηση για την παράμετρο του εκθέτη α είναι ίση με 1.89 συνεπώς το σφάλμα μεροληψίας είναι 8.7% για την μέθοδο των ελαχίστων τετραγώνων.



Σχήμα 3-3: Log-log plot της συνάρτησης επιβίωσης του αριθμών των πελατών που επλήγησαν από ηλεκτρικές διακοπές με την ευθεία μέγιστης πιθανοφάνειας για $x_i \geq x_{min}$

3.3 ΕΛΕΓΧΟΣ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ

Οι παραπάνω μέθοδοι εκτίμησης μας δίνουν την δυνατότητα να προσαρμόσουμε την PLD σε οποιοδήποτε σύνολο δεδομένων. Παρόλα αυτά δεν μας δίνουν καμία προειδοποίηση για το αν η κατανομή είναι όντως αξιόπιστη για τα δεδομένα μας. Επιπλέον, υπάρχουν πολλές κατανομές βαριάς ουράς από τις οποίες παράγονται δείγματα τα οποία μοιάζουν με την κατανομή του νόμου δύναμης(Virkar και Clauset(2014)). Συνεπώς, είναι σημαντική και η πραγματοποίηση ενός ελέγχου καλής προσαρμογής το οποίο να παράγει ένα p_value που ποσοτικοποιεί την αξιοπιστία της PLD(623).

Ένας από τους πιο γνωστούς ελέγχους είναι ο χ^2, παρόλα αυτά με το χ^2 του Pearson αντιμετωπίζουμε σοβαρά προβλήματα όσον αφορά την επιλογή του αριθμού των κλάσεων που

θα πρέπει να χρησιμοποιήσουμε (Goldstein et al. (2004)). Ένας άλλος έλεγχος είναι η απευθείας σύγκριση με κατανομές που μοιάζουν με την PLD όπως η Εκθετική Κατανομή, η Λογαριθμοκανονική Κατανομή και η Κατανομή Weibull μέσω του λόγου πιθανοφανειών. Παρόλα αυτά, ένας τέτοιος έλεγχος θα ήταν καλό να πραγματοποιηθεί εφόσον έχουμε διαπιστώσει ότι η PLD προσαρμόζεται καλά στα δεδομένα μας, τον οποίο θα περιγράψουμε στην συνέχεια. Για αυτό τον λόγο, χρησιμοποιούμε έναν έλεγχο που βασίζεται στην απόσταση, και συγκεκριμένα τον έλεγχο *Kolmogorov Smirnov(KS)* όπως περιγράφεται από τους Clauset et al. (2009). Στους υπολογισμούς θα χρησιμοποιήσουμε τον έλεγχο KS που είδαμε και στην ενότητα 3.2 (3.4).

Ο έλεγχος που γίνεται έχει την παρακάτω μορφή:

H_0: τα δεδομένα μας προέρχονται από την κατανομή νόμου δύναμης

$\big(P(x) \equiv S(x)\big)$ έναντι της

H_1: τα δεδομένα μας δεν προέρχονται από την κατανομή νόμου δύναμης.

Με την παραπάνω μέθοδο θα μπορέσουμε να διακρίνουμε εάν η διαφορά μεταξύ των δεδομένων μας και του μοντέλου μπορεί να αποδοθεί ή όχι σε στατιστικές διακυμάνσεις (statistical fluctuations) (Virkar και Clauset(2014)). Η διαδικασία του ελέγχου της υπόθεσης PL ξεκινάει με την παραγωγή τυχαίων αριθμών από την PLD ίδιου μεγέθους με τα δεδομένα μας χρησιμοποιώντας την μέθοδο της αντιστροφής (White et al (2008)), με παραμέτρους που έχουν υπολογιστεί με βάση τις ενότητες 3.1 και 3.2. Χρησιμοποιώντας την παραπάνω μέθοδο μπορούμε να παράγουμε δεδομένα που προέρχονται από την PLD, χρησιμοποιώντας μόνο μια τυχαία μεταβλητή από την Ομοιόμορφη κατανομή. Η διαδικασία είναι η ακόλουθη:

$$U = 1 - \left(\frac{X}{x_{min}}\right)^{-a+1} \Rightarrow \ln(1-U) = (1-a)\cdot[\ln(X) - \ln(x_{min})]$$

$$\Rightarrow -\frac{1}{a-1}\cdot\ln(1-U) + \ln(x_{min}) = \ln(X)$$

$$\Rightarrow X = x_{min}\cdot(1-U)^{-\frac{1}{a-1}}. \qquad (3.5)$$

Στην συνέχεια πραγματοποιούμε τον έλεγχο KS και λαμβάνουμε ένα p-value που ποσοτικοποιεί την υπόθεση της PLD (Hilbert (2013)). Παρόλα αυτά, επειδή θέλουμε η τιμή του p-value να είναι ακριβής με 2 δεκαδικά ψηφία την παραπάνω διαδικασία την εκτελούμε 2500 . Το p-value ορίζεται ως το κλάσμα, των φορών που το p-value είναι μεγαλύτερο του

επιπέδου σημαντικότητας 0.05 δια του αριθμού των επαναλήψεων. Στο παράδειγμά μας, το p-value του ελέγχου είναι 0.8794 με το οποίο δεχόμαστε ότι τα δεδομένα μας προέρχονται από την PLD.

3.4 ΕΝΑΛΛΑΚΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Η παραπάνω μέθοδος αποτελεί μια αξιόπιστη μέθοδο για το αν τα δεδομένα μας προέρχονται πραγματικά από την PLD. Παρόλα αυτά, μια μεγάλη τιμή του p-value σημαίνει ότι μπορεί να υπάρχουν εναλλακτικές κατανομές οι οποίες προσαρμόζονται καλύτερα στα δεδομένα μας από ότι η PLD. Για να έχουμε μια πιο ισχυρή ένδειξη ότι η PL είναι αυτή που περιγράφει καλά τα δεδομένα της εμπειρικής κατανομής θα πρέπει να την συγκρίνουμε με κατανομές όπως η Εκθετική Κατανομή, η Λογαριθμοκανονική Κατανομή και η Κατανομή Weibull.

Η μέθοδος που θα εφαρμοστεί είναι αυτή που ακολούθησαν οι Clauset et al. (2009) , δηλαδή έναν έλεγχο λόγου πιθανοφάνειας που προτάθηκε από τον Vuong (1989). Η βασική ιδέα του λόγου πιθανοφανειών είναι να κάνουμε τον έλεγχο του λόγου πιθανοφάνειας των δεδομένων κάτω από δύο «ανταγωνιστικές» κατανομές. Έτσι, η κατανομή με τη μεγαλύτερη πιθανοφάνεια προσαρμόζεται καλύτερα στα δεδομένα. Εναλλακτικά μπορούμε να κατασκευάσουμε τον λογάριθμο του λόγου των πιθανοφανειών, όπου επιθυμούμε να έχει θετική ή αρνητική τιμή ώστε να αποφανθούμε ποια κατανομή είναι καλύτερη.

Όταν η τιμή είναι κοντά στο μηδέν τότε δεν μπορεί να προσδιοριστεί με σιγουριά ποιο από τα δύο μοντέλα είναι το καλύτερο και συνεπώς δεν καταλήγουμε σε κάποιο συμπέρασμα. Το παραπάνω συμβαίνει καθώς τα εμπειρικά δεδομένα υπόκεινται σε στατιστικές διακυμάνσεις(Virkar και Clauset(2014)). Για να θεωρηθεί αξιόπιστη η συγκεκριμένη στατιστική συνάρτηση πρέπει θετική ή αρνητική. Ο τύπος δίνεται από τη σχέση:

$$LR = log\left(\frac{L_A\left(H|\hat{\theta}_A\right)}{L_B\left(H|\hat{\theta}_B\right)}\right). \tag{3.6}$$

όπου L_A είναι η πιθανοφάνεια του μοντέλου που ακολουθεί κατανομή νόμου δύναμης και L_B είναι η πιθανοφάνεια της εναλλακτικής κατανομής.

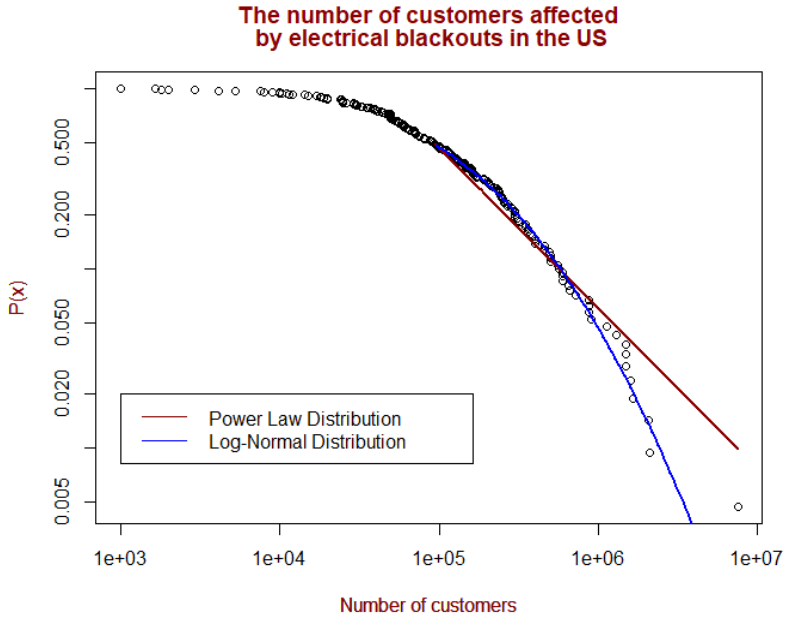
Εφόσον έχουμε υπολογίσει τον λογάριθμο του λόγου πιθανοφάνειας, για να διαπιστώσουμε ότι είναι αρκετά μακριά από το μηδέν και η ένδειξη αυτή είναι στατιστικά σημαντική, θα χρησιμοποιήσουμε την μέθοδο που πρότεινε ο Q.H. Vuong (1989). (*Likelihood Ratio Statistic (LR)*). Θα πρέπει αρχικά να χρησιμοποιήσουμε κάποιο μέτρο μεταβλητότητας. Το μέτρο μεταβλητότητας που θα χρησιμοποιηθεί στην παρούσα εργασία είναι η τυπική απόκλιση του LR. Ο τύπος της διακύμανσης δίνεται από την σχέση:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(l_i^{(A)} - l_i^{(B)}\right) - \left(\bar{l}^{(A)} - \bar{l}^{(B)}\right)\right]^2. \qquad (3.7)$$

όπου $l_i^{(A)}$ είναι ο λογάριθμος της πιθανοφάνειας της κατανομής νόμου δύναμης, $l_i^{(B)}$ είναι ο λογάριθμος της πιθανοφάνειας της εναλλακτικής κατανομής και τα $\bar{l}^{(A)}$, $\bar{l}^{(B)}$ είναι οι αντίστοιχες μέσες τιμές.

Αυτή η μέθοδος μας δίνει ένα *p-value*, το οποίο αν είναι μεγαλύτερο από το επίπεδο σημαντικότητας $a = 0.05$ τότε δεν έχουμε σαφή εικόνα για το ποιο μοντέλο είναι κατάλληλο για τα δεδομένα μας. Αντίθετα αν είναι μικρότερο, τότε ο λογάριθμος του λόγου πιθανοφανειών δεν είναι αποτέλεσμα διακυμάνσεων και μπορούμε να βασιστούμε σε αυτόν για τα συμπεράσματά μας.

Στην συνέχεια θα προβούμε σε απευθείας σύγκριση της PLD με την Λογαριθμοκανονική Κατανομή. Να τονίσουμε σε αυτό το σημείο ότι, για να συγκρίνουμε τις δύο κατανομές, θα πρέπει κάθε κατανομή να έχει το ίδιο κάτω όριο. Το παρακάτω γράφημα απεικονίζει την συνάρτηση επιβίωσης του συνόλου δεδομένων των ηλεκτρικών διακοπών μαζί με τις ευθείες καλής προσαρμογής των δύο κατανομών. Από το γράφημα προκύπτει ότι η Λογαριθμοκανονική Κατανομή προσαρμόζεται καλύτερα στα δεδομένα μας από ότι η PLD.



Σχήμα 3-4: Log-log plot της συνάρτησης επιβίωσης του αριθμών των πελατών που επλήγησαν από ηλεκτρικές διακοπές. Η προσαρμογή του PL μοντέλου είναι η κόκκινη γραμμή ενώ η μπλε γραμμή αντιστοιχεί στην Λογαριθμοκανονική Κατανομή.

Εάν εφαρμόσουμε και το στατιστικό τεστ του *Vuong* για να συγκρίνουμε αυτές τις δύο κατανομές, το αποτέλεσμα θα συμπίπτει με αυτό που φάνηκε στο παραπάνω γράφημα.

3-1 Αποτελέσματα του στατιστικό τεστ Vuong για τον αριθμό των πελατών που επλήγησαν από ηλεκτρικές διακοπές.

Vuong's Test Statistic	*p-value one sided*	*p-value two sided*
-1.654581	0.9509952	0.09800961

Από τον παραπάνω πίνακα παρατηρούμε ότι η τιμή του στατιστικού τεστ είναι αρνητική, το οποίο μας αποδεικνύει ότι η Λογαριθμοκανονική κατανομή είναι καλύτερη. Το *p-value one-sided* είναι αυτό που θα χρησιμοποιήσουμε, δηλαδή το 0.9509952, άρα η Λογαριθμοκανονική Κατανομή είναι καλύτερο μοντέλο για την περιγραφή των δεδομένων μας.

Κ Ε Φ Α Λ Α Ι Ο 4

Ε Ν Α Λ Λ Α Κ Τ Ι Κ Α Μ Ο Ν Τ Ε Λ Α

4.1 ΔΙΑΚΡΙΤΗ PLD

Στα προηγούμενα κεφάλαια εστιάσαμε στην PLD για συνεχή δεδομένα. Παρόλα αυτά, πολλές εμπειρικές ποσότητες είναι στην πραγματικότητα διακριτές και συνήθως λαμβάνουν ακέραιες τιμές. Στο κεφάλαιο αυτό θα εξετάσουμε την διακριτή κατανομή που είναι αντίστοιχη της συνεχούς PLD. Σύμφωνα με τους Clauset et.al (2007) και τα δύο είδη PLD έχουν πολλά κοινά χαρακτηριστικά, διαφέρουν ωστόσο στις στατιστικές μεθόδους που εφαρμόζονται σε αυτές.

4.1.1 Ορισμοί

Στις περισσότερες περιπτώσεις η PL συμπεριφορά ισχύει στην ουρά της κατανομής, όπου οι παρατηρούμενες τιμές είναι τόσο μεγάλες, που τα διακριτά σύνολα δεδομένων μπορούν με ασφάλεια να αντιμετωπιστούν ως συνεχή (Newman (2006). Clauset et.al (2007)). Ωστόσο, θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στο αν τα δεδομένα μας πρέπει να αντιμετωπιστούν ως διακριτά ή συνεχή, για να μην οδηγηθούμε σε μεροληπτικά αποτελέσματα όσο αφορά τις εκτιμήσεις των παραμέτρων και τις στατιστικές αναλύσεις.

Η συνάρτηση μάζας πιθανότητας (*probability mass function*) της διακριτής PLD έχει την παρακάτω μορφή:

$$p(x) = \Pr(X = x) = C \cdot x^{-a}, \quad x = 1.2. \ldots. \tag{4.1}$$

όπου α είναι η εκθετική παράμετρος και C είναι μια σταθερά όπως ακριβώς και στην συνεχή περίπτωση.

Για μια διακριτή τυχαία μεταβλητή που ακολουθεί την PLD, η σταθερά C καθορίζεται από την παρακάτω απαίτηση *:*

$$\sum_{x=1}^{\infty} p(x) = 1 \Rightarrow C \cdot \sum_{x=1}^{\infty} x^{-a} = 1 \Rightarrow C = \frac{1}{\zeta(a)}. \quad (4.2)$$

όπου $\zeta(\alpha) = (\sum_{x=1}^{\infty} x^{-a})^{-1}$ είναι η συνάρτηση ζήτα (*the Riemann ζ-function*). Αντικαθιστώντας τον τύπο (3.2) στην συνάρτηση μάζας πιθανότητας $p(x)$, μπορούμε να πάρουμε την επόμενη έκφραση της διακριτής PLD:

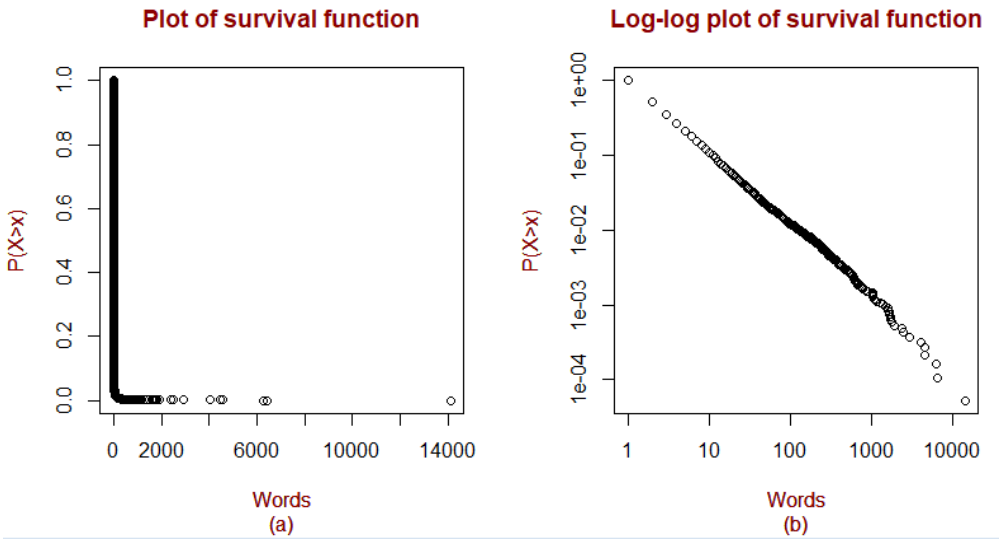
$$p(x) = \frac{x^{-a}}{\zeta(\alpha)}, \quad x = 1.2. \ldots \quad (4.3)$$

Εάν, η PL συμπεριφορά ισχύει στην ουρά της κατανομής, δηλαδή για $x \geq x_{min}$, τότε η ισοδύναμη έκφραση της διακριτής PLD είναι η παρακάτω:

$$p(x) = \frac{x^{-a}}{\zeta(\alpha, x_{min})}, \qquad x \geq x_{min} \quad (4.4)$$

όπου $\zeta(\alpha, x_{min}) = \sum_{x=x_{min}}^{\infty} (n + x_{min})^{-a}$ είναι η γενικευμένη συνάρτηση ζήτα (*Hurwitz zeta function*). Η συνάρτηση επιβίωσης στην διακριτή περίπτωση έχει την παρακάτω μορφή:

$$P(x) = \frac{\zeta(\alpha, x)}{\zeta(\alpha, x_{min})}. \quad (4.5)$$



Σχήμα 4-1: (a) Γράφημα της συνάρτησης επιβίωσης της συχνότητας εμφάνισης των λέξεων στο βιβλίο του Moby Dick. (b) Log-log plot της συνάρτησης επιβίωσης των δεδομένων.

Ένα παράδειγμα διακριτής PLD, είναι η συχνότητα εμφάνισης των λέξεων στο μυθιστόρημα του Moby Dick από τον Herman Melville (Newman (2006), Clauset et.al(2009), Gillespie (2015)). Στο συγκεκριμένο παράδειγμα έχουμε την συχνότητα 18,855 λέξεων, και όπως στην συνεχή PLD, σύμφωνα με τον Πίνακα 4-1 παρατηρείται ότι η δειγματική τυπική απόκλιση είναι περίπου 6 φορές μεγαλύτερη από την δειγματική μέση τιμή.

Επιπρόσθετα, υπολογίζοντας την συχνότητα εμφάνισης των λέξεων μέσω της γλώσσας προγραμματισμού R, παρατηρούμε ότι η συχνότητα εμφάνισης κάποιων λέξεων είναι μεγαλύτερη από την δειγματική μέση τιμή, με αποτέλεσμα να δημιουργείται και η βαριά ουρά στο Σχήμα 4-1(a) της συνάρτησης επιβίωσης των δεδομένων. Τέλος, λογαριθμίζοντας τις ποσότητες που υπάρχουν στους άξονες, προκύπτει και εδώ σύμφωνα με το Σχήμα 4-1(b). γραμμική σχέση μεταξύ του λογαρίθμου της συνάρτησης πιθανότητας και του λογάριθμου της συχνότητας των λέξεων. Επομένως, υπάρχουν ενδείξεις ότι τα δεδομένα μας ακολουθούν την PLD.

4-1 Περιγραφικά στοιχεία της συχνότητας των λέξεων στο βιβλίο του Moby Dick.

Number of Observations	Number of Observations greater than 7	Max	Min	Mean	Standard Deviation
18.855	**2.958**	**14.086**	**7**	**60.89**	**370.5983**

4.1.2 Στατιστική συμπερασματολογία

Όσο αφορά την εκτίμηση των παραμέτρων ακολουθούμε την ίδια διαδικασία όπως και στην συνεχή PLD, παρόλα αυτά θα επισημάνουμε παρακάτω τις διαφορές που παρατηρούνται. Η εκτίμηση του κατώτερου ορίου γίνεται με βάση την μέθοδο που έχει προταθεί από τους Clauset et.al (2009), και είναι ο στατιστικός έλεγχος *Kolmogorov-Smirnov*. Η εκτίμηση του x_{min} είναι η τιμή εκείνη η οποία ελαχιστοποιεί την ποσότητα:

$$D = \max_{x \geq x_{min}} |S(x) - P(x)| \tag{4.6}$$

όπου P(x) και S(x) είναι οι συναρτήσεις επιβίωσης του μοντέλου και των δεδομένων αντίστοιχα για $x \geq x_{min}$.

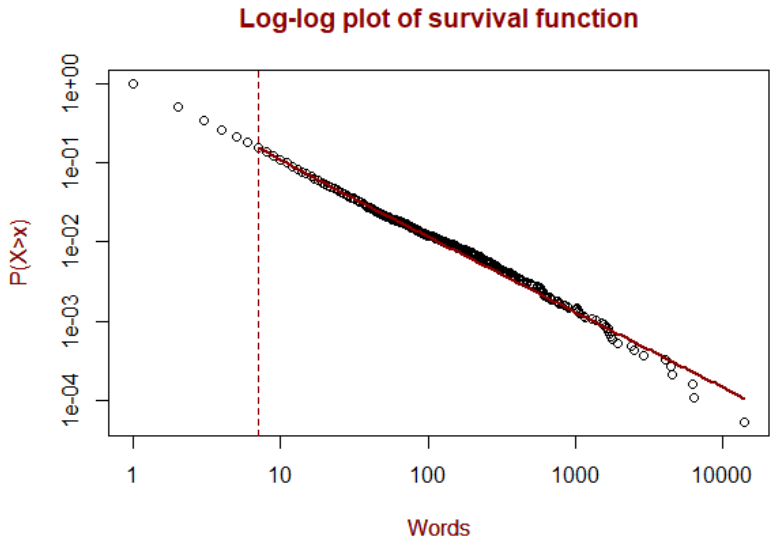
Η εκτίμηση της εκθετικής παραμέτρου γίνεται και στην διακριτή περίπτωση με την μέθοδο της μέγιστης πιθανοφάνειας, καθώς δίνει πιο ακριβή αποτελέσματα σε αντίθεση με άλλες τεχνικές. Έτσι, εάν η ουρά της εμπειρικής κατανομής ξεκινάει από κάποιο κατώτερο όριο $x_{min} > 1$, η συνάρτηση του λογαρίθμου πιθανοφάνειας είναι:

$$\mathcal{L} = ln \prod_{i=1}^{n} \frac{x_i^{-a}}{\zeta(a.x_{min})} = -nln\zeta(a.x_{min}) - a\sum_{i=1}^{n} lnx_i \tag{4.7}$$

Θέτοντας $\frac{\partial \mathcal{L}}{\partial a} = 0$ και λύνοντας ως προς a φτάνουμε στο παρακάτω αποτέλεσμα:

$$\frac{-n}{\zeta(a.x_{\min})}\frac{\partial}{\partial a}\zeta(a.x_{min}) - \sum_{i=1}^{n} lnx_i = 0$$

$$\Rightarrow \frac{\zeta'(\hat{a}.x_{min})}{\zeta(\hat{a}.x_{min})} = -\frac{1}{n}\cdot\sum_{i=1}^{n}\ln x_i . \tag{4.8}$$



Σχήμα 4-2 : Log-log plot της συνάρτησης επιβίωσης της συχνότητας εμφάνισης των λέξεων στο βιβλίο του Moby Dick μαζί με την ευθεία μέγιστης πιθανοφάνειας για $x_i \geq x_{min}$.

Παρόλα αυτά, επειδή η εξίσωση (4.8) είναι δύσκολο να λυθεί αριθμητικά, χρησιμοποιείται η παρακάτω προσέγγιση (4.9) όπως έχει δοθεί από τους Clauset et.al (2007), στην οποία οι ακέραιοι αριθμοί που έχουν παραχθεί από την PLD προσεγγίζονται ως συνεχείς πραγματικοί αριθμοί στρογγυλοποιημένοι στον πλησιέστερο ακέραιο αριθμό:

$$\hat{\alpha} \simeq 1 + n \cdot \left[\sum_{i=1}^{n} ln \frac{x_i}{x_{min} - \frac{1}{2}} \right]^{-1}. \tag{4.9}$$

Η διακριτή προσέγγιση του εκτιμητή μέγιστης πιθανοφάνειας είναι παρόμοια με αυτή της συνεχούς PLD με την διαφορά ότι έχουμε τον όρο $\frac{1}{2}$ στον παρονομαστή. Για το παραπάνω σύνολο δεδομένων, με την βοήθεια της γλώσσας προγραμματισμού R έχουμε βρει την εκτίμηση του κατώτερου ορίου, δηλαδή $x_{min} = 7$, και τον εκτιμητή μέγιστης πιθανοφάνειας $\hat{\alpha} = 1.95$.

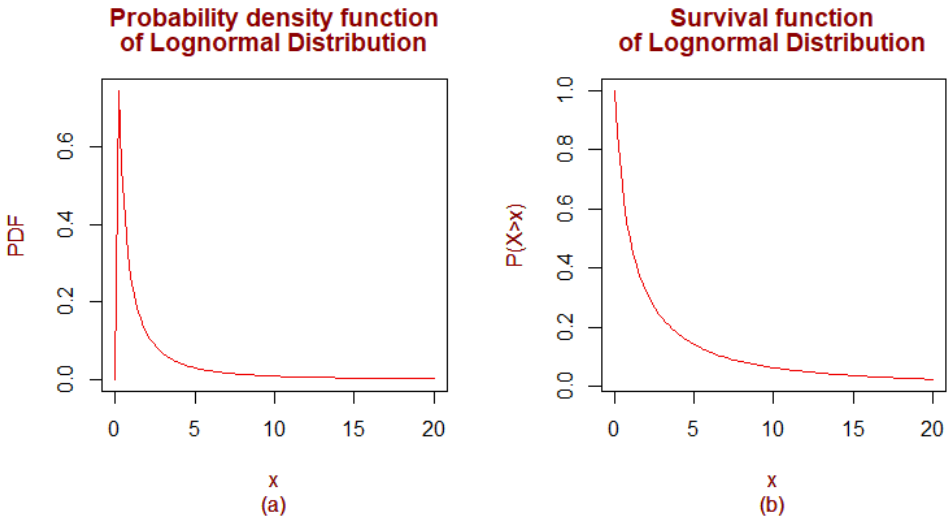
4.2 ΛΟΓΑΡΙΘΜΟΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Οι ασύμμετρες κατανομές έχουν συνήθως μικρές τιμές για τη μέση τιμή και μεγάλες τιμές για την διακύμανση. Πέρα από την PLD, μια ασύμμετρη κατανομή είναι και η Λογαριθμοκανονική, δηλαδή η *Lognormal Distribution (LD)*. Αρκετοί είναι οι ερευνητές που υποστηρίζουν, ότι η LD, δίνει την καλύτερη περιγραφή σε αντίθεση με την PLD. Η LD δίνει μεγάλη πιθανότητα σε μικρά μεγέθη και μικρή, αλλά σημαντική, πιθανότητα σε μεγάλα μεγέθη (Albert et.al (1999)). Μια LD αποτελείται από δύο παραμέτρους και περιγράφεται από μια συνάρτηση πυκνότητας πιθανότητας τέτοια ώστε:

$$p(x) = \frac{1}{x} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(lnx - \mu)^2}{2\sigma^2} \right). \tag{4.9}$$

Πίνακας 4-2: Μέτρα θέσης της LD

Mean	Median	Mode
$\boldsymbol{\exp\left(\mu + {\sigma^2}/{2}\right)}$	$\exp(\mu)$	$\exp(\mu - \sigma^2)$

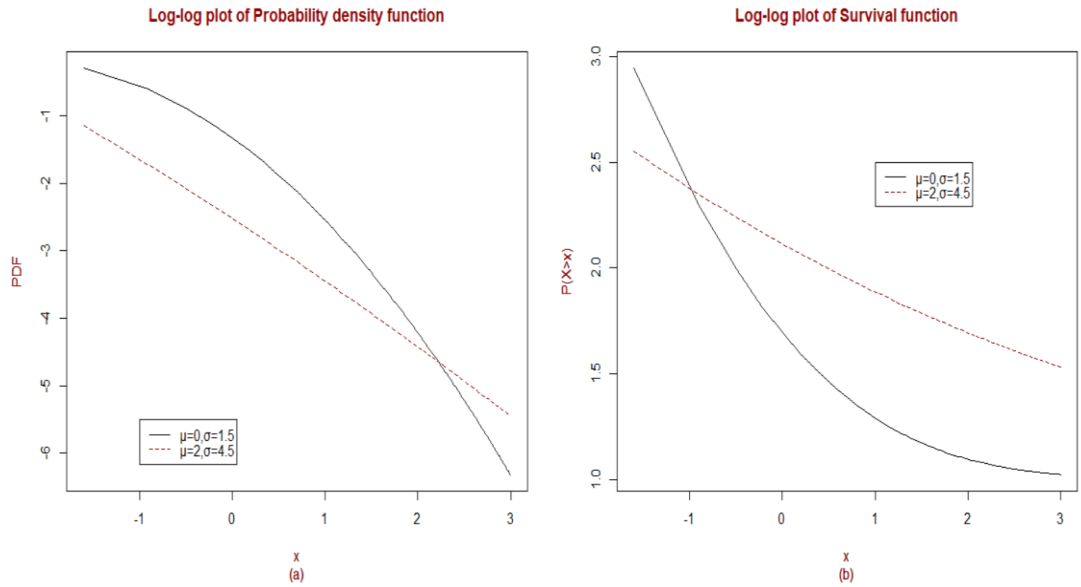


Σχήμα 4-3: (a) Γράφημα της συνάρτησης πυκνότητας πιθανότητας επιβίωσης της LD μ=4 και σ=20. (b)Log-log plot της συνάρτησης επιβίωσης της LD.

Σύμφωνα με τους Gong et al. (2001) σε ένα γράφημα της συνάρτησης επιβίωσης με λογαριθμισμένους άξονες η PLD φθίνει με σταθερή κλίση ενώ η LD φθίνει με αυξανόμενη κλίση. Παρόλα αυτά, οι Mitzenmacher (2001) και Gong et al. (2001) διαπίστωσαν ότι όσο πιο μεγάλη είναι η διακύμανση, τόσο η συνάρτηση επιβίωσης αλλά και η συνάρτηση πυκνότητας της LD θα προσεγγίζεται από μια ευθεία γραμμή.

Για να διαπιστώσουμε το παραπάνω έχουμε λάβει αρχικά δύο LD με παραμέτρους μ=0, σ=1.5 και μ=2, σ=4.5. Ακολούθως, όπως φαίνεται στο Σχήμα 4-4 (α) και στο Σχήμα 4-4 (b), έχουμε σχεδιάσει τις συναρτήσεις πυκνότητας πιθανότητας και τις συναρτήσεις επιβίωσης έχοντας λογαριθμήσει τους δύο άξονες. Σύμφωνα με τα παρακάτω σχήματα, στην περίπτωση που η διακύμανση έχει μεγαλύτερη τιμή, η κατανομή προσεγγίζεται από μια ευθεία γραμμή και το σχήμα αυτό είναι εξαιρετικά παρόμοιο με αυτό της PLD.

Σύμφωνα με τους Aitchison και Brown (1954) η LD είναι ένα στατιστικό εργαλείο ειδικά για την ανάλυση των δεδομένων εισοδήματος. Οι ιδιότητές της είναι απλές λόγω της στενής της σχέσης με την Κανονική Κατανομή, και αυτό της παρέχει εύκολη πρόσβαση στις μεθόδους εκτίμησης αλλά και σε στατιστικούς ελέγχους για σύγκριση. Όπως στην Κανονική Κατανομή το άθροισμα δύο ανεξάρτητων κανονικών μεταβλητών είναι κανονική τυχαία μεταβλητή, έτσι και στην LD ισχύει ότι το γινόμενο δύο τυχαίων μεταβλητών από την LD είναι επίσης LD (Mitzenmacher (2001), Limpert et.al(2001)).



Σχήμα 4-4: (a) Γράφημα της συνάρτησης πυκνότητας πιθανότητας σε log-log άξονες για μ=0, σ=1.5 και μ=2, σ=4.5 (b) Γράφημα της συνάρτησης επιβίωσης σε log-log άξονες για μ=0, σ=1.5 και μ=2, σ=4.5.

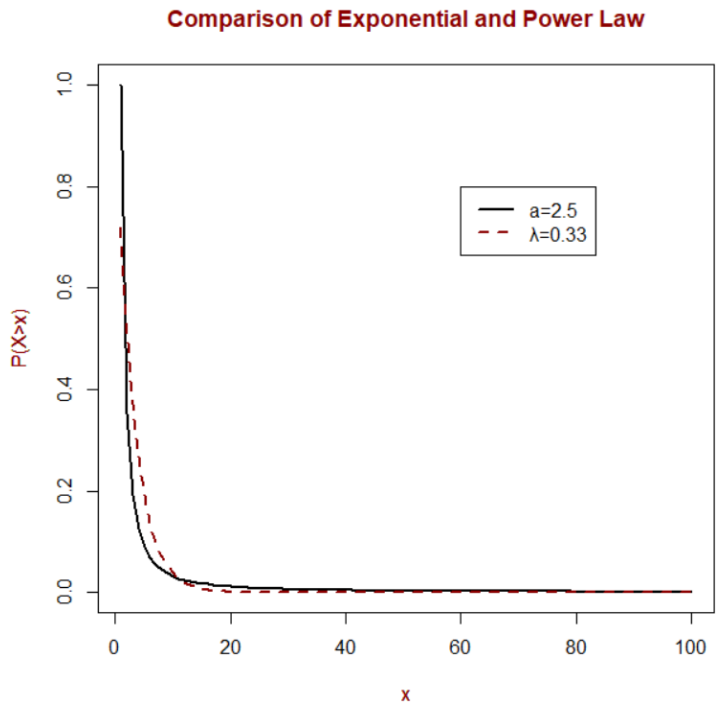
Μερικές φορές η προσαρμογή της LD σε ένα σύνολο δεδομένων είναι καλύτερη της PLD, καθώς αποτελείται από μια ακόμη παράμετρο και με αυτόν τον τρόπο βελτιώνεται και η προσαρμογή στα δεδομένα (Gabaix (2009)). Παρόλα αυτά, υπάρχει κίνδυνος να μην εντοπίσουμε την ουσία του φαινομένου. Επιπρόσθετα, είναι δύσκολο να χαρακτηρίσουμε την ουρά μιας κατανομής η οποία βασίζεται σε πεπερασμένο σύνολο δεδομένων. Για αυτό τον λόγο θα πρέπει να είμαστε προσεκτικοί στο ποιο από τα δύο μοντέλα θα επιλέξουμε και σίγουρα ο στατιστικός έλεγχος Kolmogorov-Smirnof καθώς και η απευθείας σύγκριση των δύο μοντέλων με τον έλεγχο του λόγου πιθανοφάνειας του Vuong, όπως είδαμε και στο Κεφάλαιο 3, θα μας βοηθήσουν για να κάνουμε την καλύτερη επιλογή.

4.3 ΕΚΘΕΤΙΚΗ ΚΑΤΑΝΟΜΗ

Μια συνεχής κατανομή, η οποία έχει ευρεία χρήση είναι η Εκθετική Κατανομή ή *Exponential Distribution (ED)*. Η πιο συνηθισμένη περίπτωση εμφάνισης της παραπάνω κατανομής είναι όταν μελετάμε τον χρόνο αναμονής μέχρι την πραγματοποίηση ενός σπανίου

γεγονότος. Η ED έχει μια παράμετρο λ και συνάρτηση πυκνότητας πιθανότητας ίση με $p(x) = \lambda \cdot e^{-\lambda x}$.

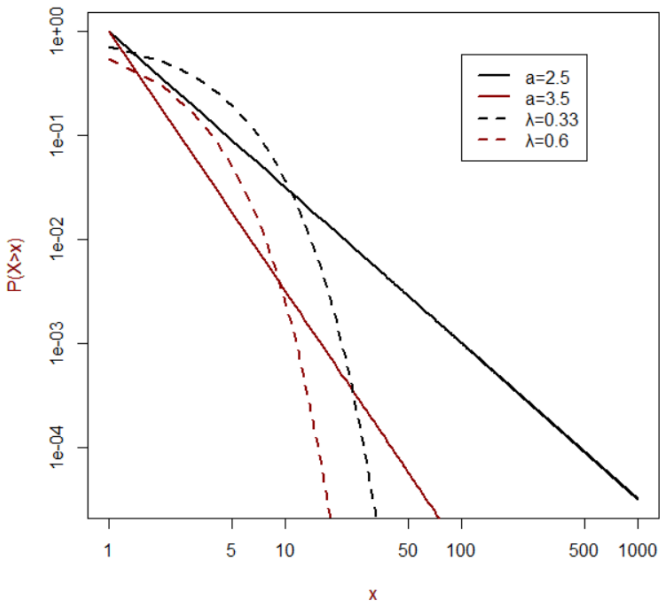
Η PLD διαφέρει από την ED, πέρα από τις μαθηματικές ιδιότητες, κυρίως στην ουρά. Σε μια PLD, η ουρά της εμπειρικής κατανομής, ασυμπτωτικά, φθίνει σύμφωνα με μια δύναμη α. Όπως αναφέραμε και στον Κεφάλαιο 2, η κατανομή αυτή οδηγεί σε πιο βαριές ουρές σε αντίθεση με την ED που έχει πιο λεπτές ουρές (Mitzenmacher, Barari και Mitra (2008)). Συνεπώς, ακραίες τιμές που αναμένονται στην PLD δεν αναμένονται στην ED, δηλαδή έχουμε λιγότερο ακραίες διακυμάνσεις των τιμών (Chu-Shore et.al (2010)). Στο Σχήμα 4-6 συγκρίνουμε τις συναρτήσεις επιβίωσης των δύο κατανομών. Να επισημάνουμε σε αυτό το σημείο ότι, η επιλογή της παραμέτρου της Εκθετικής Κατανομής έχει γίνει εξισώνοντας την ροπή πρώτης τάξης της PLD, για α=2.5 με x_{min}=1, με την ροπή πρώτης τάξης της Εκθετικής Κατανομής, δηλαδή $E(x) = \frac{1}{\lambda}$.



Σχήμα 4-5: Log-log plot των συναρτήσεων επιβίωσης της PLD και ED για διάφορες τιμές των παραμέτρων

Άλλη μια διαφορά αυτών των δύο κατανομών παρατηρείται όταν λογαριθμήσουμε τους άξονες των συναρτήσεων επιβίωσης τους. Στο Σχήμα 4-7, έχουμε κατασκευάσει τις συναρτήσεις επιβίωσης, για τις δύο κατανομές σε *log-log* άξονες, για διάφορες επιλογές των παραμέτρων τους. Η διαδικασία επιλογής των παραμέτρων της Εκθετικής Κατανομής έχει

προκύψει με την διαδικασία που αναφέρθηκε παραπάνω. Στην ED φαίνεται να υπάρχει έντονη καμπυλότητα σε αντίθεση με την PLD που φαίνεται να προσεγγίζεται από μια ευθεία γραμμή.



Σχήμα 4-6: Log-log plot των συναρτήσεων επιβίωσης της PLD και ED για διάφορες τιμές των παραμέτρων τους.

Σύμφωνα με τους Chu-Shore et.al(2010) μπορούμε να ξεχωρίσουμε την PLD από την ED, παρόλα αυτά το άθροισμα δύο ή τριών εκθετικών συναρτήσεων, δηλαδή η Κατανομή Γάμμα. ενδέχεται να μοιάζει στην PLD καθώς προσεγγίζονται από μια ευθεία γραμμή όταν λογαριθμίσουμε και τους δύο άξονες. Τέλος, σύμφωνα με τον Newman(2009) ο συνδυασμός δύο εκθετικών κατανομών αποτελεί έναν μηχανισμό παραγωγής PLD. Ας υποθέσουμε ότι μια ποσότητα Y προέρχεται από την ED ώστε $p(y) \sim e^{ay}$. Εάν, μια ποσότητα X σχετίζεται εκθετικά με την Y δηλαδή $Y = e^{bX}$ τότε η κατανομή πιθανότητας της X είναι:

$$p(x) = p(y)\frac{dy}{dx} \sim \frac{e^{ay}}{b \cdot e^{by}} = \frac{x^{-1+a/b}}{b} \qquad (4.10)$$

η οποία αποτελεί μια PLD με εκθετική παράμετρο $\alpha = 1 - \alpha/b$.

4.4 ΕΠΕΚΤΕΙΝΟΜΕΝΗ ΕΚΘΕΤΙΚΗ ΚΑΤΑΝΟΜΗ ΚΑΙ ΚΑΤΑΝΟΜΗ PLD ΜΕ ΕΚΘΕΤΙΚΗ ΑΠΟΚΟΠΗ

Σύμφωνα με τον Barabasi (2015), για την περιγραφή δεδομένων αρκετά συχνά χρησιμοποιούνται κατανομές που αποτελούν συνδυασμό της PLD και της ED. Αυτές οι κατανομές μπορεί να είναι είτε εκθετικά φραγμένες όπως η PLD με εκθετική αποκοπή (*PLD with exponential cutoff*) ή φραγμένες αλλά που να φθίνουν πολύ πιο γρήγορα από την PLD όπως είναι η επεκτεινόμενη εκθετική Κατανομή (*Stretched Exponential Distribution (SED)*).

Η PLD με εκθετική αποκοπή είναι απλά ένας PL πολλαπλασιασμένος με την εκθετική συνάρτηση. Συνεπώς συνδυάζει την PLD, η οποία οδηγεί σε πιο βαριές ουρές, μαζί με την εκθετική συνάρτηση που είναι φραγμένη. Η συνάρτηση πυκνότητας πιθανότητας έχει την παρακάτω μορφή:

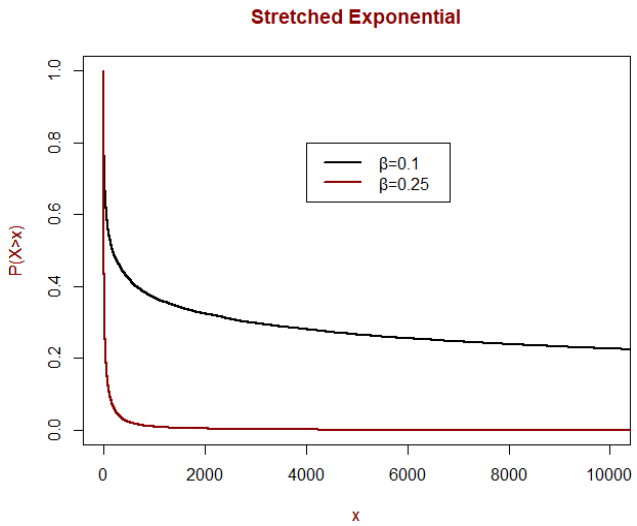
$$p(x) = C \cdot x^{-a} \cdot e^{-\lambda x}. \tag{4.11}$$

όπου $C = \frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha.\lambda x_{min})}$.

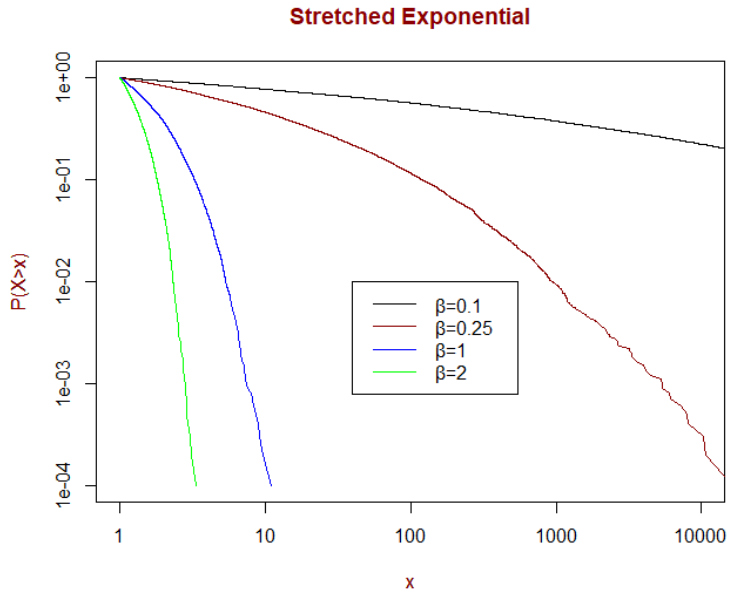
Εάν πάρουμε τον λογάριθμο της εξίσωσης (4.11) τότε έχουμε ότι:

$$\ln\big(p(x)\big) = lnC - a \cdot lnx - \lambda \cdot x. \tag{4.12}$$

Με την παραπάνω εξίσωση διαπιστώνουμε ότι εάν $x \ll \frac{1}{\lambda}$ τότε η κατανομή ακολουθεί την PLD με εκθετική παράμετρο ίση με *α*. Αντίθετα, εάν $x \gg \frac{1}{\lambda}$ τότε η κατανομή ακολουθεί την PLD with exponential cutoff για μεγάλες τιμές του *x* (Barabasi (2015)).



Σχήμα 4-7: Συνάρτηση επιβίωσης της SED για β=0.1 και β=0.25.



Σχήμα 4-8: Log-log plot της συνάρτησης επιβίωσης για διάφορες τιμές του β.

Όσον αφορά την SED η μορφή της είναι παρόμοια με της (4.11) με την διαφορά ότι ο PL βρίσκεται πάνω στην εκθετική συνάρτηση. Η SED είναι γνωστή και ως η συνάρτηση επιβίωσης της κατανομής Weibull (Cardona et.al (2007)). Η συνάρτηση πυκνότητας πιθανότητας της SED έχει την παρακάτω μορφή:

$$p(x) = C \cdot x^{\beta-1} \cdot e^{-\lambda \cdot x^{\beta}}. \tag{4.13}$$

όπου $C = \beta \cdot \lambda \cdot e^{\lambda \cdot x_{min}^{\beta}}$.

Η SED είναι μια κατανομή βαριάς ουράς που παράγει δείγματα τα οποία μοιάζουν στην PLD. Όταν η εκθετική παράμετρος κινείται στην περιοχή $0 < \beta < 1$, δύσκολα μπορούμε να διακρίνουμε την SED από την PLD και ειδικότερα όσο το $\beta \simeq 0$. Συνεπώς, η εκθετική παράμετρος δηλώνει το πόσο βαριά είναι η ουρά και όσο πιο μικρή είναι τόσο πιο βαριά ουρά θα έχουμε (Laherrere και Sornette (1998)). Όταν $\beta = 1$ προκύπτει η εκθετική συνάρτηση. Η SED έχει ουρά πιο βαριά από την εκθετική κατανομή αλλά λιγότερο βαριά από ότι η PLD. Ένα ακόμη πλεονέκτημα που έχει είναι ότι οι ροπές της είναι πεπερασμένες και συνεπώς η διακύμανση, η ασυμμετρία και η κύρτωση μπορούν εύκολα να εκτιμηθούν.

Κ Ε Φ Α Λ Α Ι Ο 5

ΣΥΓΚΡΙΣΗ ΜΕΘΟΔΩΝ ΚΑΙ ΕΛΕΓΧΟΙ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ

5.1 ΣΥΓΚΡΙΣΗ ΜΕΘΟΔΩΝ

Σε αυτήν την ενότητα θα συγκρίνουμε τρεις μεθόδους εκτίμησης της εκθετικής παραμέτρου της PLD, και επιπλέον θα εξετάσουμε για κάθε μια ξεχωριστά πότε μας δίνουν πιο αξιόπιστα αποτελέσματα. Για την εύρεση των εκτιμητών αυτών των μεθόδων, δημιουργήσαμε τυχαίους αριθμούς από την PLD σύμφωνα με τον τύπο (3.5) που είδαμε στο Κεφάλαιο 3. Ένας δεύτερος τρόπος δημιουργίας τυχαίων αριθμών από την PLD είναι χρησιμοποιώντας το διαθέσιμο CRAN πακέτο PoweRlaw της R και την εντολή *rplcon(n. xmin. alpha)*. Να τονίσουμε σε αυτό το σημείο, ότι έχουμε χρησιμοποιήσει στην αρχή του αλγορίθμου και την συνάρτηση *set.seed(10)*, έτσι ώστε να διασφαλίσουμε ότι όλα τα αποτελέσματα μπορούν να αναπαραχθούν.

Πριν δημιουργήσουμε τους τυχαίους αριθμούς έχουμε ορίσει την εκθετική παράμετρο και το κατώτερο όριο της PLD, καθώς και το πλήθος των παρατηρήσεων. Πιο αναλυτικά, ως κατώτερο όριο έχουμε επιλέξει την τιμή $x_{min} = 1$, ως εκθετική παράμετρο έχουμε επιλέξει τις τιμές $\alpha = 1.5,\ \alpha = 2, \alpha = 2.5,\ \alpha = 3.5$ και τέλος ως πλήθος παρατηρήσεων έχουμε επιλέξει τις τιμές $N = 10^2, N = 10^3, N = 10^4, N = 10^5$ *και* $N = 10^6$.

Αρχικά, για την Εκτίμηση Μέγιστης Πιθανοφάνειας (MLE). χρησιμοποιήσαμε τον τύπο (3.3) που είδαμε στο Κεφάλαιο 3. Όπως αναφέραμε και στο Κεφάλαιο 2, η συνάρτηση επιβίωσης της PLD ακολουθεί και αυτή PLD αλλά με εκθέτη $\alpha - 1$. Συνεπώς, για να βρούμε τον εκτιμητή με την μέθοδο ελαχίστων τετραγώνων, αρχικά δημιουργήσαμε την συνάρτηση επιβίωσης της PLD. Ακολούθως εφαρμόσαμε την μέθοδο ελαχίστων τετραγώνων, και αφού αυξήσαμε την κλίση της ευθείας κατά μία μονάδα βρήκαμε την εκτίμηση της εκθετικής παραμέτρου. Τέλος, για την μέθοδο των ροπών αφού προσδιορίσαμε αρχικά την ροπή πρώτης

τάξης, καθώς έχουμε μια άγνωστη παράμετρο. στη συνέχεια την εξισώσαμε με την δειγματική μέση τιμή. Την παραπάνω μέθοδο την εφαρμόσαμε για $\alpha = 2.5, \alpha = 3.5$ καθώς για μικρότερες τιμές η ροπή πρώτης τάξης είναι μη πεπερασμένη.

***Πίνακας 5-1**: Αποτελέσματα της σύγκρισης των μεθόδων εκτίμησης της εκθετικής παραμέτρου.*

Number	MLE	LS	Methods of moments	Error (Methods of moments)	Error (MLE)	Error (LS)
$\alpha = 1.5 \quad x_{min} = 1$						
$N = 10^2$	1.66155	1.68715	-	-	10.77%	12.48%
$N = 10^3$	1.47614	1.43916	-	-	-1.6%	-4.06%
$N = 10^4$	1.49566	1.48901	-	-	-0.29%	-0.73%
$N = 10^5$	1.502988	1.501323	-	-	0.2%	0.09%
$N = 10^6$	1.500707	1.5002948	-	-	0.05%	0.02%
$\alpha = 2 \quad x_{min} = 1$						
$N = 10^2$	2.323101	2.37430	-	-	16.16%	18.72%
$N = 10^3$	1.952288	1.87833	-	-	-2.39%	-6.08%
$N = 10^4$	1.991333	1.97802	-	-	-0.43%	-1.1%
$N = 10^5$	2.005975	2.002646	-	-	0.3%	0.13%
$N = 10^6$	2.001414	2.0005897	-	-	0.07%	0.03%
$\alpha = 2.5 \quad x_{min} = 1$						
$N = 10^2$	2.984651	3.06145	1.876818	-37.44%	19.37%	22.46%
$N = 10^3$	2.428432	2.31749	3.788415	25.96%	-2.86%	-7.30%
$N = 10^4$	2.487	2.46703	3.072443	2.41%	-0.52%	-1.32%
$N = 10^5$	2.508963	2.503969	2.944191	-1.86%	0.36%	0.16%
$N = 10^6$	2.502121	2.5008845	3.029895	1%	0.08%	0.04%
$\alpha = 3.5 \quad x_{min} = 1$						
$N = 10^2$	4.307751	4.43575	1.409892	-15.41%	23.08%	26.74%
$N = 10^3$	3.38072	3.19582	1.776377	6.58%	-3.41%	-8.7%
$N = 10^4$	3.478333	3.44505	1.683085	0.98%	-0.62%	-1.57%
$N = 10^5$	3.514938	3.506615	1.661637	0.301%	0.43%	0.189%
$N = 10^6$	3.503535	3.5014741	1.665933	-0.04%	0.101%	0.04%

Τα αποτελέσματα των τριών αυτών μεθόδων βρίσκονται στον παραπάνω Πίνακα 5-1 όπου έχουμε υπολογίσει και το ποσοστό σφάλματος που δίνεται από τον παρακάτω τύπο:

$$\%_{σφάλματος} = \frac{εκτιμώμενη\ τιμή - πραγματική\ τιμή}{πραγματική\ τιμή} \cdot 100$$

Ο παραπάνω τύπος λαμβάνει θετικές τιμές και αυτό δηλώνει ότι η μέθοδος υπερεκτιμά την τιμή της παραμέτρου ή αρνητικές τιμές και αυτό δηλώνει ότι η μέθοδος υποεκτιμά την τιμή της παραμέτρου.

Από τα παραπάνω αποτελέσματα είναι φανερό ότι όσο πιο μεγάλο είναι το πλήθος των παρατηρήσεών μας τόσο περισσότερο οι εκτιμήσεις και των τριών μεθόδων προσεγγίζουν την πραγματική τιμή της παραμέτρου. Αν και η μέθοδος των ροπών έχει όλο και πιο μικρό ποσοστό σφάλματος όσο αυξάνονται οι παρατηρήσεις, παρόλα αυτά δεν μπορεί να μας δώσει εκτιμήσεις για την εκθετική παράμετρο όταν αυτή είναι μικρότερη του δύο, καθώς όπως αναφέραμε και παραπάνω η ροπή πρώτης τάξης είναι μη πεπερασμένη.

Όσον αφορά τις άλλες δύο μεθόδους, σε πολλά επιστημονικά άρθρα υπάρχει προτίμηση της μεθόδου μέγιστης πιθανοφάνειας έναντι της μεθόδου ελαχίστων τετραγώνων όπως αναλύσαμε και στο Κεφάλαιο 3. Τα αποτελέσματα που προέκυψαν συμφωνούν με το παραπάνω συμπέρασμα, καθώς με μια πρώτη ματιά διαπιστώσαμε ότι η εκτίμηση μέγιστης πιθανοφάνειας δίνει καλύτερα αποτελέσματα έναντι των άλλων μεθόδων καθώς έχουμε τα μικρότερα ποσοστά σφάλματος.

Επιπρόσθετα, διαπιστώσαμε ότι σημαντικό ρόλο παίζει αν η εκθετική παράμετρος λαμβάνει μεγάλες ή όχι τιμές. Σύμφωνα με τον Πίνακα 5-1, όταν η εκθετική παράμετρος ισούται με 1.5 και ο αριθμός των παρατηρήσεων είναι $N = 10^2$ το ποσοστό σφάλματος και στις δύο μεθόδους είναι μικρό, δηλαδή 10.77% για την μέθοδο μέγιστης πιθανοφάνειας και 12.48% για την μέθοδο ελαχίστων τετραγώνων. Παρόλα αυτά, το παραπάνω δεν συμβαίνει όταν η εκθετική παράμετρος ισούται με 3.5 για τον ίδιο αριθμό παρατηρήσεων καθώς το ποσοστό σφάλματος έχει αυξηθεί και στις δύο μεθόδους. Συνεπώς, όσο πιο μεγάλες τιμές λαμβάνει η εκθετική παράμετρος και όσο πιο μικρός είναι ο αριθμός των παρατηρήσεων, και οι δύο μέθοδοι αποτυγχάνουν να προσεγγίσουν την πραγματική τιμή της εκθετικής παραμέτρου.

5.2 Έλεγχοι καλής προσαρμογής

Σε αυτή την ενότητα σκοπός μας είναι να διαπιστώσουμε μέσω κάποιων κριτηρίων, ποια κατανομή μπορεί να προσαρμοστεί καλύτερα σε δεδομένα τα οποία προέρχονται από την PLD. Οι κατανομές που θα εξετάσουμε είναι η Εκθετική Κατανομή, η Κανονική Κατανομή, η Λογαριθμοκανονική Κατανομή και τέλος η Κατανομή Weibull. Έχουμε δημιουργήσει τρία σύνολα δεδομένων τυχαίων αριθμών από την PLD και πιο συγκεκριμένα, το πρώτο αποτελείται από 100 παρατηρήσεις με x_{min}=1 και α=2.5, το δεύτερο και αυτό με 100 παρατηρήσεις με x_{min}=1 και α=1.5, και τέλος το τρίτο αποτελείται από 1,000 παρατηρήσεις με x_{min}=1 και α=3.

Πίνακας 5-2: Εκτιμήσεις παραμέτρων των κατανομών.

	Normal Distribution		**Exponential**	**Lognormal Distribution**		**Weibull**	
	Mean	SD	Rate	Mean	SD	Shape	Scale
N=10^2 x_{min}=1 α=2.5	1.876818	1.255625	0.5328166	0.503867	0.4464478	1.704551	2.125958
N=10^2 x_{min}=1 α=1.5	22.08272	84.93563	0.04528427	1.511601	1.339344	0.5681838	9.608637
N=10^3 x_{min}=1 α=3	2.231736	4.332943	0.4480816	0.5250511	0.5580172	1.093976	2.340859

Η γλώσσα προγραμματισμού *R* διαθέτει στην βιβλιοθήκη της ένα πακέτο το οποίο ονομάζεται *MASS* το οποίο περιέχει την συνάρτηση *fitdistr*. Με την βοήθεια αυτής της συνάρτησης θα βρούμε τις εκτιμήσεις των παραμέτρων των παραπάνω κατανομών για τα τρία σύνολα δεδομένων που έχουμε δημιουργήσει. Να τονίσουμε σε αυτό το σημείο, ότι η παραπάνω συνάρτηση χρησιμοποιεί την μέθοδος μέγιστης πιθανοφάνειας για τις εκτιμήσεις των παραμέτρων. Τα αποτελέσματα των εκτιμήσεων βρίσκονται στον Πίνακα 5-2.

Πίνακας 5-3: Αποτελέσματα κριτηρίων

	Normal Distribution		
	KS	BIC	AIC
N=10^2. x_{min}=1 α=2.5	D=0.24492 *p-value*=1.232 ·10^{-5}	338.5247	333.3143
N=10^2. x_{min}=1 α=1.5	D= 0.40212 *p-value*= 1.799·10^{-14}	1181.377	1176.166
N=10^3. x_{min}=1. α=3	D=0.38816 *p-value* < 2.2 ·10^{-16}	5784.186	5774.371
	Exponential Distribution		
	KS	BIC	AIC
N=102. xmin=1 α=2.5	D=0.41609 p-value=1.887·10-15	330.5208	327.9156
N=102. xmin=1 α=1.5	D= 0.47779 p-value < 2.2·10-16	823.5643	820.9591
N=103. xmin=1. α=3	D= 0.36134 p-value < 2.2·10-16	3612.468	3607.56
	Lognormal Distribution		
	KS	BIC	AIC
N=102. xmin=1 α=2.5	D= 0.15002 p-value= 0.02219	232.4849	227.2746
N=102. xmin=1 α=1.5	D= 0.15002 p-value= 0.02219	653.7542	648.5438
N=103. xmin=1. α=3	D=0.17368 p-value < 2.2·10-16	2735.064	2725.248
	Weibull Distribution		
	KS	BIC	AIC
N=102. xmin=1 α=2.5	D= 0.24504 p-value= 1.218·10-5	288.6721	283.4618
N=102. xmin=1 α=1.5	D= 0.24504 p-value= 1.218·10-5	709.9414	704.731
N=103. xmin=1. α=3	D= 0.3261 p-value < 2.2·10-16	3596.388	3607.56

Το πρώτο κριτήριο που χρησιμοποιήσαμε είναι ο έλεγχος *Kolmogorov-Smirmof* , ο οποίος, όπως αναφέραμε και στο Κεφάλαιο 3, βασίζεται στην μέγιστη διαφορά μεταξύ των δεδομένων

που προέρχονται από την PLD και της αθροιστικής συνάρτησης κατανομής της Κανονικής Κατανομής, της Εκθετικής Κατανομής, της Λογαριθμοκανονικής Κατανομής ή της Κατανομής Weibull. Όλα τα *p-value* σύμφωνα με τον Πίνακα 5-3 είναι μικρότερα του επιπέδου σημαντικότητας α=0.05, συνεπώς δεν υπάρχουν στατιστικές ενδείξεις ότι τα δεδομένα μας προέρχονται από τις παραπάνω κατανομές.

Το παραπάνω αποτέλεσμα είναι αναμενόμενο καθώς τα δεδομένα μας έχουν παραχθεί από την PLD. Αξίζει να σημειωθεί ότι, η Λογαριθμοκανονική Κατανομή έχει την μικρότερη απόσταση D έναντι των άλλων κατανομών, αν και όπως αναφέραμε και παραπάνω δεν υπάρχουν στατιστικές ενδείξεις ότι τα δεδομένα προσαρμόζονται στην Λογαριθμοκανονική Κατανομή.

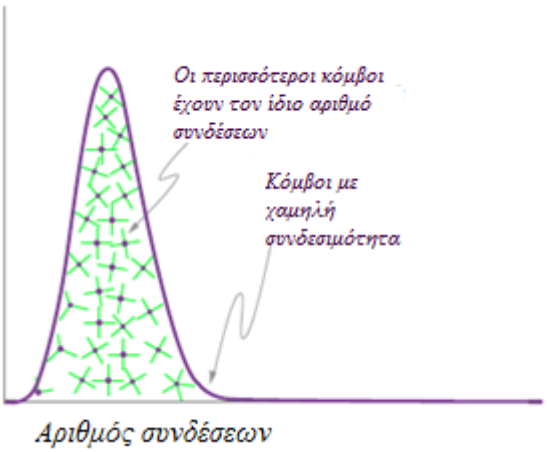
Για να βρούμε την κατανομή που εφαρμόζει καλύτερα στα δεδομένα θα εφαρμόσουμε και τα κριτήρια AIC (*Akaike Information Criteria*) και BIC (*Bayesian Information Criteria*). Η κατανομή που προσαρμόζεται καλύτερα στα δεδομένα είναι αυτή που έχει την μικρότερη τιμή AIC και BIC αντίστοιχα. Σύμφωνα με τα αποτελέσματα του Πίνακα 5-3 καταλήγουμε στο συμπέρασμα ότι η κατανομή που φαίνεται ως καλύτερη είναι η Λογαριθμοκανονική Κατανομή καθώς έχει και στα τρία σύνολα δεδομένων τις μικρότερες τιμές και για τα δύο κριτήρια. Το παραπάνω αποτέλεσμα συμφωνεί με τον έλεγχο καλής προσαρμογής *Kolmogorov-Smirnov*, διότι όπως αναφέρθηκε και παραπάνω για την Λογαριθμοκανονική Κατανομή η απόσταση ήταν η μικρότερη συγκριτικά με τις υπόλοιπες κατανομές. Η δεύτερη καλύτερη κατανομή είναι η Weibull ενώ οι υπόλοιπες δύο έχουν πάντα πιο υψηλές τιμές με την Κανονική Κατανομή να έχει πάντα τις υψηλότερες τιμές.

Κ Ε Φ Α Λ Α Ι Ο 6

ΕΦΑΡΜΟΓΕΣ ΤΗΣ PLD ΚΑΙ ΓΛΩΣΣΕΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ

6.1 ΕΦΑΡΜΟΓΕΣ ΤΗΣ PLD

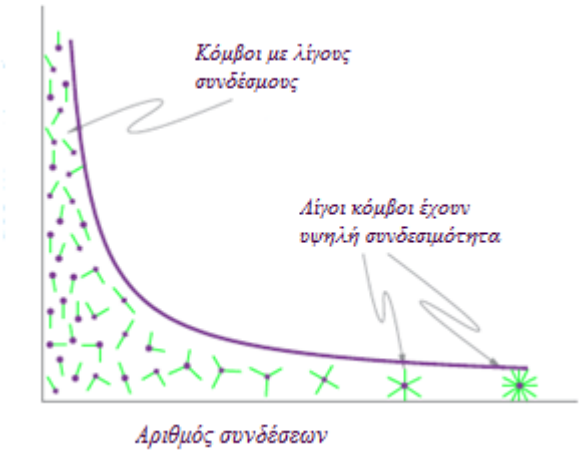
Υπάρχουν ποικίλα δίκτυα, όπως ο παγκόσμιος ιστός (*World Wide Web* , WWW), όπου η δυσκολία περιγραφής τους έγκειται στην πολύπλοκη τοπολογία τους. Πιο συγκεκριμένα, σε αυτά τα δίκτυα οι κόμβοι αντιπροσωπεύουν τα στοιχεία του δικτύου των οποίων οι ακμές αντιπροσωπεύουν τις αλληλεπιδράσεις μεταξύ τους (Albert και Barabasi (2002)). Σε αυτά τα πραγματικά δίκτυα, έχουν παρατηρηθεί δύο βασικά χαρακτηριστικά. Το πρώτο είναι η αύξηση ή growth και το άλλο είναι η προτίμηση σύνδεσης (*preferential attachment*).



Σχήμα 6-1: Γράφημα της κατανομή των βαθμών ενός τυχαίου δικτύου η οποία ακολουθεί κατανομή Poisson (Barabasi (2015)).

Το πρώτο χαρακτηριστικό υποδηλώνει ότι αυτά τα δίκτυα εξελίσσονται μέσω μιας αναπτυσσόμενης διαδικασίας. Πιο συγκεκριμένα, τα δίκτυα αυτά συνεχώς επεκτείνονται με την προσθήκη νέων κόμβων τα οποία συνδέονται με άλλους κόμβους. Η προτίμηση σύνδεσης υποδηλώνει ότι καινούργιοι κόμβοι θα συνδεθούν με μεγάλη πιθανότητα με κόμβους οι οποίοι έχουν μεγάλη συνδεσιμότητα (Barabasi και Albert (1999). Jeong et al. (2003)). Σύμφωνα με τους Adamic και Huberman (2000) αλλά και τους Easley και Kleinberg (2010) το παραπάνω χαρακτηριστικό ονομάζεται και ο πλούσιος γίνεται πλουσιότερος (*rich-get-richer*). Αυτό πρακτικά σημαίνει ότι, η *preferential attachment* οδηγεί στο φαινόμενο *rich-get*-richer καθώς οι παλιοί κόμβοι θα αυξήσουν την συνδεσιμότητά τους σε βάρος των καινούργιων κόμβων. Έτσι, με την πάροδο του χρόνου κάποιοι κόμβοι θα έχουν μεγάλη συνδεσιμότητα. *Οι* βαθμοί των κόμβων χαρακτηρίζονται από μια κατανομή $P(k)$, η οποία εκφράζει την πιθανότητα ένας κόμβος να αλληλοεπιδρά με *k* άλλους κόμβους (Barabasi και Albert(1999)).

Στο μοντέλο των Erdos και Renyi (1960) η κατανομή του βαθμού των κόμβων περιγράφεται από την κατανομή Poisson. Στο παραπάνω γράφημα παρατηρείται ότι η πιθανότητα να βρεθεί κόμβος με μεγάλη συνδεσιμότητα είναι πάρα πολύ μικρή. Παρόλα αυτά, στην κατανόηση των πολύπλοκων δικτύων παρατηρήθηκε ότι κάποιοι κόμβοι διατηρούν μεγάλο αριθμό συνδέσεων αναλογικά με το συνολικό μέγεθος του δικτύου (Adamic και Huberman (2002)).



Σχήμα 6-2: Γράφημα της κατανομής των βαθμών ενός τυχαίου δικτύου η οποία ακολουθεί PLD. (Barabasi (2015)).

Οι Albert et al. (1999) πρότειναν μια βελτιωμένη θεωρία για πραγματικά δίκτυα όπου η κατανομή των βαθμών στην ουρά της έχει συμπεριφορά PL. Συνεπώς, ανεξάρτητα από το σύστημα που μελετάμε, η πιθανότητα $P(k)$ ένας κόμβος να αλληλοεπιδρά με *k* κόμβους φθίνει

σύμφωνα με την PL, δηλαδή $p(k) \sim k^{-\gamma}$. Η εμφάνιση της PLD είναι το αποτέλεσμα της συνύπαρξης αύξησης και προτίμηση σύνδεσης και τα παραπάνω δίκτυα ονομάζονται *Barabasi and Albert networks (BA model)* ή *scale-free networks*.

Η βασική διαφορά μεταξύ των τυχαίων δικτύων και των *scale-free* δικτύων έγκειται στην ουρά της degree distribution. Η PL συμπεριφορά στην ουρά δείχνει η πιθανότητα να βρούμε κόμβους με μεγάλο αριθμό συνδέσεων είναι σημαντική (Σχήμα 6-2). Επιπλέον, πολλοί κόμβοι με μικρή συνδεσιμότητα συνυπάρχουν στο δίκτυο με λίγους κόμβους με υψηλή συνδεσιμότητα. Στα *scale-free* δίκτυα η εκθετική παράμετρος κυμαίνεται μεταξύ του 2 και του 3, και αυτό πρακτικά σημαίνει, όπως είδαμε και στο Κεφάλαιο 2, ότι η ροπή πρώτης τάξης θα είναι πεπερασμένη ενώ η ροπή δεύτερης τάξης θα τείνει στο άπειρο.

Θα δώσουμε στην συνέχεια περισσότερες λεπτομέρειες για τον WWW, το Ίντερνετ, το Δίκτυο αναφορών, το Δίκτυο τηλεφωνικών κλήσεων, το Δίκτυο ηλεκτρονικού ταχυδρομείου και τέλος το Δίκτυο των ηθοποιών.

α. WWW

Ο WWW αντιπροσωπεύει το μεγαλύτερο δίκτυο το μέγεθος του οποίου ξεπερνάει και τον ανθρώπινο εγκέφαλο (Barabasi (2015)). Σε αυτό το δίκτυο οι κόμβοι αποτελούν τις ιστοσελίδες και οι ακμές είναι οι υπερσύνδεσμοι οι οποίοι ενώνουν την μια ιστοσελίδα με την άλλη (Albert και Barasi (2002)). Επειδή ο WWW είναι ένα κατευθυνόμενο δίκτυο, θα έχουμε δύο κατανομές βαθμών $p_{in}(k)$ και $p_{out}(k)$ που αντιστοιχούν στους εισερχόμενους και εξερχόμενους υπερσυνδέσμους. Πιο αναλυτικά, η $p_{in}(k)$ εκφράζει την πιθανότητα k υπερσύνδεσμοι να οδηγούν σε μια συγκεκριμένη ιστοσελίδα, ενώ η $p_{out}(k)$ εκφράζει την πιθανότητα μια ιστοσελίδα να έχει k υπερσυνδέσμους. Οι παραπάνω πιθανότητες ακολουθούν PLD δηλαδή $p_{in}(k) \sim k^{-\gamma_{in}}$ και $p_{in}(k) \sim k^{-\gamma_{out}}$ όπου γ_{in} και γ_{out} είναι οι εκθετικές παράμετροι για τις δύο περιπτώσεις.

Οι Albert et al. (1999) μελέτησαν τους κόμβους του WWW (N=325,729) και υπολόγισαν ότι $\gamma_{in} = 2.1$ και $\gamma_{out} = 2.45$. Έτσι, κόμβοι με διαφορετική συνδεσιμότητα συνυπάρχουν στο ίδιο δίκτυο. Οι Kumar et al. (1999) μελέτησαν 40 εκατομμύρια ιστοσελίδες και έφτασαν στο συμπέρασμα ότι $\gamma_{in} = 2.1$ και $\gamma_{out} = 2.38$ ενώ οι Broder et al. (2000) χρησιμοποιώντας 200 εκατομμύρια ιστοσελίδες κατέληξαν ότι $\gamma_{in} = 2.1$ και $\gamma_{out} = 2.72$ (Albert και Barabasi (2002)). Το παρακάτω σχήμα είναι ένα *scale-free* δίκτυο με $\gamma = 2.1$ και $\bar{k} = 3$ στο οποίο

παρατηρείται η συνύπαρξη λίγων κόμβων με μεγάλη συνδεσιμότητα μαζί με έναν μεγάλο αριθμό κόμβων με μικρή συνδεσιμότητα (Barabasi (2015)).

Σχήμα 6-3: Scale-free δίκτυο με $\gamma = 2.1$ *και* $E(x) = 3$. *(Barabasi (2015)).*

Τέλος, οι Adamic και Huberman (2000) μελέτησαν την ανίχνευση 260,000 ιστοσελίδων (crawl of sites), οι οποίες αντιπροσωπεύουν έναν ξεχωριστό όνομα τομέα (domain name). Πιο αναλυτικά, μέτρησαν πόσους συνδέσμους δέχτηκαν οι ιστοσελίδες από άλλες ιστοσελίδες. Το αποτέλεσμα της έρευνας τους, ήταν η κατανομή των βαθμών των εισερχόμενων υπερσυνδέσμων να ακολουθεί PLD με εκθετική παράμετρο $\gamma_{in} = 1.94$.

β. Ίντερνετ

Το ίντερνετ αποτελεί ένα δίκτυο με φυσικές συνδέσεις καθώς υπάρχουν συνδέσεις μεταξύ των υπολογιστών αλλά και τηλεπικοινωνιακών συσκευών. Η τοπολογία του Ίντερνετ μπορεί να μελετηθεί σε δύο επίπεδα. Το πρώτο επίπεδο είναι αυτό του δρομολογητή ή *router* (router level) στο οποίο οι κόμβοι είναι οι routers και οι ακμές είναι οι φυσικές συνδέσεις μεταξύ τους. Το δεύτερο επίπεδο έχει να κάνει με τα αυτόνομα συστήματα (inter-domain level), όπου το κάθε σύστημα αποτελείται από χιλιάδες routers και υπολογιστές και αντιπροσωπεύει έναν κόμβο, ενώ η ακμή είναι η σύνδεση δύο αυτόνομων συστημάτων εάν υπάρχει τουλάχιστον μια διαδρομή που να τα συνδέει (Albert και Barabasi (2002)).

Η τοπολογία του Ίντερνετ έχει προσελκύσει έντονο ερευνητικό ενδιαφέρον καθώς έχουν προκύψει ερωτήματα όπως: "Γιατί δεν μπορούμε να προσομοιώσουμε το Ίντερνετ;", "Υπάρχουν ιδιότητες στην τοπολογία του Ίντερνετ οι οποίες δεν αλλάζουν με τον χρόνο;", Μελέτες που έχουν γίνει στο παρελθόν χρησιμοποιούν μέτρα που βασίζονται στην μέση τιμή, στο μέγιστο και στο ελάχιστο τα οποία οδηγούν σε απώλεια σημαντικής πληροφορίας καθώς

δεν μπορούν να περιγράψουν ικανοποιητικά τις ασύμμετρες κατανομές (Faloutsos et al. (1999)).

Οι Faloutsos et al. (1999) μελέτησαν την τοπολογία του Ίντερνετ και στα δύο επίπεδα και κατέληξαν, ότι η κατανομή βαθμών ακολουθεί PLD. Για την έρευνά τους, όσον αφορά το δεύτερο επίπεδο, χρησιμοποίησαν τρία σύνολα δεδομένων. Τα πρώτο σύνολο δεδομένων αντιστοιχεί στον Νοέμβριο του 1997 με 3,015 κόμβους και 5,126 ακμές, το δεύτερο στον Απρίλιο του 1998 με 3,530 κόμβους και 6,432 ακμές, και τέλος το τρίτο στον Δεκέμβριο του 1998 με 4,389 κόμβους και 8,256 ακμές. Στο επίπεδο του δρομολογητή τα δεδομένα αντιστοιχούν στο έτος 1995 και αποτελούνται από 3,888 κόμβους και 5,012 ακμές. Οι εκθετικές παράμετροι για το δεύτερο επίπεδο είναι $\gamma_I^{as} = 2.15$. $\gamma_I^{as} = 2.16$ και $\gamma_I^{as} = 2.2$ αντίστοιχα, ενώ για το πρώτο $\gamma_I^r = 2.48$. Να τονίσουμε σε αυτό το σημείο ότι η αύξηση του Ίντερνετ με τον χρόνο είναι 45%.

Σύμφωνα με τους Albert et al. (1999) μια σημαντική ποσότητα προς μελέτη είναι η κοντινότερη απόσταση μεταξύ δύο ιστοσελίδων, και πιο συγκεκριμένα ο μικρότερος αριθμός υπερσυνδέσμων που απαιτούνται για την μετάβαση από μια ιστοσελίδα σε μια άλλη. Η διάμετρος του Ίντερνετ ορίζεται ως η μέση τιμή των κοντινότερων αυτών διαδρομών και η PLD καταφέρνει να απαντήσει σε ερωτήματα όπως ποια θα ήταν η διάμετρος του Ίντερνετ εάν διπλασιαζόταν ο αριθμός των κόμβων αλλά και το πόσες ακμές θα αναμέναμε με αυτόν τον διπλασιασμό.

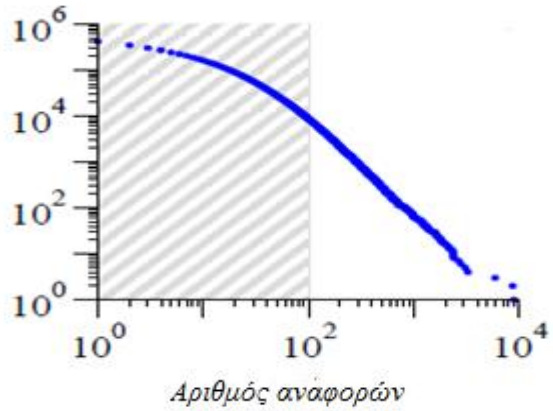
γ. Δίκτυο αναφορών (*Citation network)*

Ένα πολύπλοκο δίκτυο είναι αυτό των αναφορών στις επιστημονικές δημοσιεύσεις. Όσο περισσότερο αναφέρεται μια εργασία, τόσο πιο μεγάλη πιθανότητα έχει να την ακούσουμε και να την διαβάσουμε. Σε αυτό το δίκτυο οι κόμβοι αντιπροσωπεύουν τα δημοσιευμένα άρθρα, ενώ οι ακμές αντιπροσωπεύουν τις αναφορές τους σε προηγούμενα δημοσιευμένα άρθρα (Albert και Barabasi (2002)).

Ο Redner το 1998 μελέτησε την κατανομή της αναφοράς και κατέληξε στο συμπέρασμα ότι, ένα δημοσιευμένο άρθρο το οποίο αναφέρεται k φορές ακολουθεί την PLD με εκθετική παράμετρο $\gamma_{cite} = 3$. Συνεπώς, η πιθανότητα ένα επιστημονικό άρθρο να δεχτεί k αναφορές είναι ανάλογη του k^{-3}.

Πιο αναλυτικά, ο Redner (1998) μελέτησε δύο σύνολα δεδομένων. Στο πρώτο σύνολο δεδομένων οι αναφορές έγιναν το χρονικό διάστημα μεταξύ του 1981 και του 1997. Σε αυτά

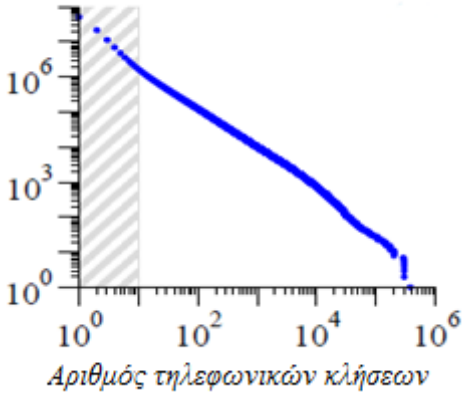
τα δεδομένα έχουμε 783,339 άρθρα που υπάρχουν στο Ινστιτούτο Επιστημονικών Πληροφοριών (*Institute for Scientific Information)* ενώ οι αναφορές τους είναι 6,716,198. Στο δεύτερο σύνολο δεδομένων έχουμε 24,296 έγγραφα στα οποία έχει γίνει αναφορά τουλάχιστον μια φορά, και το σύνολο των αναφορών είναι 351,872. Τα παραπάνω έγγραφα έχουν δημοσιευτεί στους τόμους 11 μέχρι 50 στο *Physical Review D*, μεταξύ του 1975 και του 1994. Αυτό που διαπιστώθηκε είναι ότι η εισερχόμενη κατανομή βαθμού για το δίκτυο αναφορών έχει συμπεριφορά PL με εκθετική παράμετρο ίση με 3.



Σχήμα 6-4:Συνάρτηση επιβίωσης του αριθμού των αναφορών σε δημοσιευμένα επιστημονικά έγγραφα (Newman(2009)).

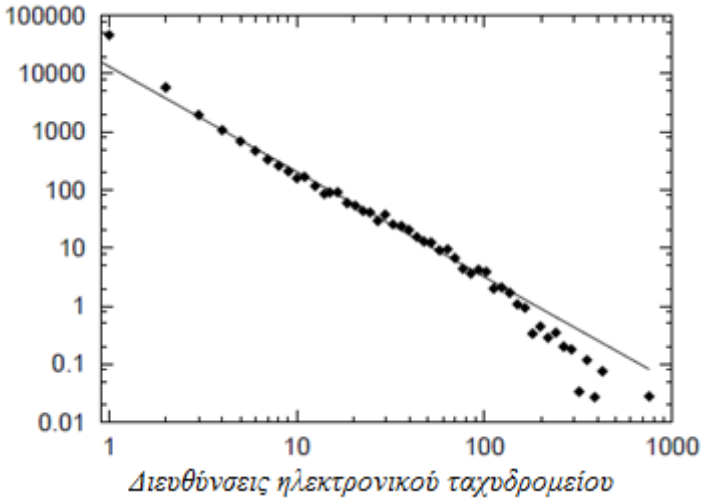
δ. Δίκτυο τηλεφωνικών κλήσεων και δίκτυο ηλεκτρονικού ταχυδρομείου

Σύμφωνα με τους Albert και Barabasi (2002), οι Abello et al. (1999), καθώς και οι Aiello et al. (2000) μελέτησαν το κατευθυνόμενο δίκτυο των τηλεφωνικών κλήσεων (*Phone call network*). Οι παραπάνω κατέληξαν στο συμπέρασμα ότι η κατανομή του βαθμού των εισερχόμενων και εξερχόμενων τηλεφωνικών κλήσεων ακολουθεί PLD με εκθετική παράμετρο $\gamma_{in} = \gamma_{out} = 2.1$ όπως φαίνεται και στο Σχήμα 6.5. Οι κόμβοι σε αυτό το δίκτυο είναι οι τηλεφωνικοί αριθμοί και κάθε ολοκληρωμένη τηλεφωνική κλήση αποτελεί μια ακμή. Σύμφωνα με τον Newman (2009) το σύνολο δεδομένων που μελέτησαν οι Aiello et al. (2000) αφορά τον αριθμό των τηλεφωνικών κλήσεων που δέχτηκαν σε μια μέρα οι 51 εκατομμύρια χρήστες της τηλεφωνικής υπηρεσίας μεγάλων τηλεφωνικών αποστάσεων *AT&T*. Αυτό που διαπίστωσαν ήταν ότι ο μεγαλύτερος αριθμός κλήσεων που δέχτηκε ένας πελάτης την συγκεκριμένη μέρα ήταν 375,746 κλήσεις ή 260 κλήσεις ανά λεπτό.



Σχήμα 6-5: Συνάρτηση επιβίωσης του αριθμού των τηλεφωνικών κλήσεων (Newman(2009)).

Μια παρόμοια κατανομή με τον αριθμό των τηλεφωνικών κλήσεων είναι και ο αριθμός των μηνυμάτων μέσω του ηλεκτρονικού ταχυδρομείου (email). Σύμφωνα με τους Ebel et al. (2002), οι κόμβοι αυτού του δικτύου είναι οι διευθύνσεις των emails οι οποίες συνδέονται μεταξύ τους κατά την ανταλλαγή αλληλογραφίας. Σε αυτήν την έρευνα, μελέτησαν 59,912 διευθύνσεις ηλεκτρονικού ταχυδρομείου και διαπίστωσαν ότι η κατανομή βαθμού ακολουθεί την PLD με εκθετική παράμετρο $\gamma = 1.81$ (βλέπε Σχήμα 6.6).



Σχήμα 6-6: Η κατανομή βαθμού του δικτύου ηλεκτρονικού ταχυδρομείου σε Log-log plot (Ebel et al. (2002)).

Πίνακας 6-1: Εφαρμογές της PLD

	Size	γ_{in}	γ_{out}
WWW	325,729	2.1	2.45
WWW	40,000,000	2.1	2.38
WWW	200,000,000	2.1	2.72
Internet, domain	3,015	2.15	2.15
Internet, domain	3,530	2.16	2.16
Internet, domain	4,389	2.2	2.2
Internet, router	3,888	2.48	2.48
E-mail	59,912	1.81	1.81
Phone call	51,000,000	2.1	2.1
Citation	783,339	3	-
Movie actors	212,250	2.3	2.3

ε. Δίκτυο ηθοποιών (movie actor network)

Σε αυτό το δίκτυο οι κόμβοι είναι οι ηθοποιοί, ενώ οι άκρες δείχνουν ότι δύο ηθοποιοί συνεργάστηκαν σε ταινία τουλάχιστον μια φορά. Σύμφωνα με τους Barabasi και Albert (1999) μελετήθηκε η συνεργασία 212,250 ηθοποιών και το συμπέρασμα ήταν ότι η πιθανότητα ένας ηθοποιός να έχει συνεργαστεί με *k* ηθοποιούς, δηλαδή να έχει *k* συνδέσεις. ακολουθεί την PLD με εκθετική παράμετρο ίση με $\gamma_{actor} = 2.3 \pm 0.1$.

6.2 ΓΛΩΣΣΕΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ

Υπάρχουν αρκετές γλώσσες προγραμματισμού που μπορούν να εφαρμόσουν τις τεχνικές που πρότειναν οι Clauset et al. (2009) για την εκτίμηση της εκθετικής παραμέτρου αλλά και του κατώτερου ορίου. Εμείς έχουμε ξεχωρίσει τρείς γλώσσες προγραμματισμού διότι παρέχουν βιβλιοθήκες ή εργαλεία που μπορούμε εύκολα να βρούμε αυτές τις τεχνικές αλλά και είναι εύκολες προς χρήση. Αυτές οι γλώσσες προγραμματισμού είναι η *R*, *Python* και *Matlab* και παρακάτω θα τις παρουσιάσουμε έχοντας βάλει και τον αντίστοιχο κώδικα που χρησιμοποιείται στην καθεμία.

- **R**

Η γλώσσα προγραμματισμού *R* δημιουργήθηκε από τους Ross Ihaka και Robert Gentleman και τέθηκε σε εφαρμογή το 1995. Αποτελεί ένα ισχυρό στατιστικό εργαλείο για την ανάλυση μεγάλων δεδομένων. Διαθέτει μια μεγάλη ποικιλία στατιστικών τεχνικών όπως Ανάλυση Παλινδρόμησης, Γενικευμένα Γραμμικά Μοντέλα καθώς και πολλές γραφικές τεχνικές όπως το πακέτο "ggplot2". Πέρα από τα παραπάνω η *R* μπορεί να επεκταθεί μέσω των πακέτων της καθώς διαθέτει μια τεράστια βιβλιοθήκη με επιστημονικούς αλγόριθμους οι οποίοι βοηθούν τους αναλυτές δεδομένων να δημιουργούν γρήγορα ευφυείς αναλυτικές εφαρμογές μεγάλων δεδομένων. Τέλος, σε αντίθεση με άλλα λογισμικά τα οποία χρησιμοποιούν πιο περίπλοκους κώδικες για την ανάλυση των δεδομένων. στην *R* μπορούν να γραφτούν μέσα σε λίγες γραμμές.

Η γλώσσα προγραμματισμού *R* διαθέτει στην βιβλιοθήκη της ένα πακέτο το οποίο ονομάζεται *poweRlaw*. Το συγκεκριμένο πακέτο περιέχει συναρτήσεις της *R* για προσαρμογή, σύγκριση και οπτικοποίηση των κατανομών βαριάς ουράς (Gillespie (2015)). Πέραν της συνεχούς και διακριτής PLD, το πακέτο αυτό μπορεί να εφαρμοστεί και σε έναν αριθμό κατανομών όπως η Εκθετική κατανομή και η Λογαριθμοκανονική κατανομή. Επιπρόσθετα, το πακέτο *poweRlaw*, μας παρέχει εύκολη χρήση των προτεινόμενων τεχνικών από τους Clauset et al. (2009) και με την βοήθειά του έχουμε βρει τις εκτιμήσεις για τα παραδείγματα των Κεφαλαίων 2.3 και 4.

Με τις εντολές που παρέχει το πακέτο *poweRlaw* μπορούμε με να βρούμε για την συνεχή PLD την συνάρτηση πυκνότητας πιθανότητας, την συνάρτηση κατανομής και ακόμα να παράγουμε εύκολα τυχαίους αριθμούς. Στην διακριτή περίπτωση οι εντολές είναι παρόμοιες, μόνο που εδώ αντί για *plcon* έχουμε *pldis*. Οι εντολές φαίνοται παρακάτω:

```
>dplcon(x, xmin, alpha, log = FALSE)
>pplcon(q, xmin, alpha, lower.tail = TRUE)
>rplcon(n, xmin, alpha)
```

Όπως αναφέρθηκε και παραπάνω με την βοήθεια του πακέτου *poweRlaw*, μπορεί να βρεθεί η εκτίμηση της εκθετικής παραμέτρου με την μέθοδο της μέγιστης πιθανοφάνειας. Παρακάτω έχουμε γράψει τις εντολές με τις οποίες μπορούμε να βρούμε την εκθετική παράμετρο αλλά

και το κατώτερο όριο για το σύνολο δεδομένων του μέγεθος των πόλεων που αναφέραμε στο Κεφάλαιο 2.

```
>library(poweRlaw)
>a<-read.table("F:/cities.txt");
>attach(a)
>anew=conpl$new(V1)
>est=estimate_xmin(anew);est
>anew$setXmin(99216)
>estimate_pars(anew)[1]
```

- **Python**

Η γλώσσα προγραμματισμού *Python* έχει ευρεία χρήση, δημιουργήθηκε από τον Guido van Rossum και η εφαρμογή της ξεκίνησε το 1989. Σε αντίθεση με την *R*, η παραπάνω γλώσσα προγραμματισμού είναι μια γλώσσα γενικού σκοπού αφού πρέπει να έχει ευρύτερη χρήση πέραν από την στατιστική ανάλυση των δεδομένων. Το κοινό τους χαρακτηριστικό, παρόλα αυτά, είναι ότι και οι δύο γλώσσες προγραμματισμού μπορούν να επεκταθούν μέσω των βιβλιοθηκών τους.

Η προσαρμογή κατανομών βαριάς ουράς περιλαμβάνει πολλούς και σύνθετους αλγορίθμους. Προκειμένου να μειωθούν σημαντικά τα εμπόδια στην εφαρμογή των στατιστικών τεχνικών στις παραπάνω κατανομές, υπάρχει το πακέτο *powerlaw* της *Python*. Αυτό το πακέτο είναι πιο βελτιωμένο σε σχέση με το προηγούμενο διαθέσιμο λογισμικό και έχει ως αποτέλεσμα ο χρήστης μπορεί να δημιουργήσει μικρότερο κώδικα για να εκτελεστεί η πλήρης ανάλυση((Alstott (2014)).

Με το πακέτο *powerlaw* μπορούμε να δημιουργήσουμε το γράφημα της συνάρτησης πυκνότητας αλλά και της συνάρτησης επιβίωσης της PLD. Αυτό το πακέτο έχει σχεδιαστεί για εύκολη πλοήγηση, συντήρηση και επεκτασιμότητα για αυτόν τον λόγο υπάρχουν αρκετές βελτιώσεις οι οποίες θα προστεθούν στις μελλοντικές εκδόσεις του *powerlaw* (Alstott (2014)). Παρακάτω παραθέτουμε τις εντολές της *Python* για την εύρεση της εκτίμησης της εκθετικής παραμέτρου μέσω της μεθόδου μέγιστης πιθανοφάνειας καθώς και την εύρεση του κατώτερου ορίου. Με την τελευταία εντολή μπορούμε να κάνουμε απευθείας σύγκριση της PLD με τις εναλλακτικές κατανομές.

```
>import powerlaw
>fit=powerlaw.Fit(data)
>fit.power_law.alpha
>fit.power_law.xmin
>fit.distribution_compare('powerlaw'. 'exponential')
```

- **Matlab**

Η γλώσσα προγραμματισμού *Matlab* είναι μια γλώσσα προγραμματισμού προσανατολισμένη στις μαθηματικές διαδικασίες. Παρέχει μαθηματικές συναρτήσεις καθώς και ένα εκτεταμένο σύνολο συναρτήσεων για την γραμμική άλγεβρα και πολλαπλασιασμό πινάκων. Τα εργαλεία της *Matlab* δηλαδή *Matlab toolboxes* παρέχουν στατιστικά εργαλεία, μηχανική εκμάθηση (*machine learning*), επεξεργασία σήματος (*signal processing*). βελτιστοποίηση (*optimization*) κ.α (https://www.mathworks.com/discovery/matlab-vs-r.html).

Επιπρόσθετα, μπορούμε να δημιουργήσουμε μια αποθήκη δεδομένων με την οποία μπορούμε να διαβάζουμε τα δεδομένα και με τις στατιστικές συναρτήσεις της *Matlab* να γίνεται ο καθαρισμός και η επεξεργασία τους για την εξόρυξη γνώσης. Αυτή η γλώσσα προγραμματισμού παρέχει υψηλή ταχύτητα για την επεξεργασία δεδομένων παρόλα αυτά για να χρησιμοποιηθεί πρέπει κάποιος να πληρώσει την άδεια που απαιτείται. Αυτό αποτελεί και ένα μειονέκτημά της σε σχέση με τις δύο παραπάνω γλώσσες προγραμματισμού οι οποίες είναι ελεύθερες προς χρήση.

Με την βοήθεια της γλώσσας προγραμματισμού *Matlab* μπορούμε να βρούμε την εκτίμηση της εκθετικής παραμέτρου καθώς και το κατώτερο όριο για την PLD με τις μεθόδους που περιέγραψαν οι Clauset et al. (2007). Παρακάτω παραθέτουμε την αντίστοιχη συνάρτηση η οποία αυτόματα ανιχνεύει εάν τα δεδομένα μας είναι συνεχή ή διακριτά. Επιπρόσθετα, εάν έχουμε διακριτό σύνολο δεδομένων, και η ελάχιστη τιμή των δεδομένων αυτών είναι μεγαλύτερη του 1000, τότε αυτά τα δεδομένα αντιμετωπίζονται ως συνεχή.

```
>function [alpha, xmin, L]=plfit(x, varargin)
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξένη

[1] Adamic L.A., Huberman B.A, "The nature of markets in the World Wide Web, *Quarterly Journal of Electronic Commerce,* Volume **1**, 1-10.

[2] Aitchison J., Brown J.A.C., (1954), "On Criteria for descriptions of income distribution", Metroeconomica, *International review of economics*, Volume **6**, Issue 3, 88-107.

[3] Albert R., Barabasi A.L, (2002), "Statistical Mechanics of complex networks", *Reviews of Physics*, Volume **74**, Issue 1, pages 47-97.

[4] Albert R., Jeong H., Barabasi A.L. (1999), "Internet: Diameter of the World Wide Web", *Nature*, Volume **401**, pages 130-131.

[5] Alstott J., Bullmore E., Dietmar P. (2014), "powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions, *PloS One* 9(1)

[6] Arnold B.C., (1983), "Pareto Distributions Statistical Distributions in Scientific Work", *International Co-operative Publishing House*, Burtonsville, MD, MR0751409.

[7] Bak P., Tang C. (1989), "Earthquakes as a Self-Organized Critical Phenomenon", *Journal Geographical Research,* Volume **94**, NO.B11, 15635-15637.

[8] Barabasi A.L(2015), "Network Science: The Scale-Free Property", *Cambridge University Press.*

[9] Barabasi A.L., Albert R. (1999), "Emergence of Scaling in Random Networks", *Science*, Volume **286**, Issue 5439, pages 509-512.

[10] Barari M., Saibal M., (2008), "Power Law Versus Exponential Law in Characterizing Stock Market Results", *International Atlantic Economic*, Volume **36**, 377-179.

[11] Bauke H., (2007), "Parameter estimation for power-law distributions by maximum likelihood methods", *The European Physical Journal B.* **58**, Issue 2, pre-print 167-173.

[12] Baxter G., Frean M., Noble J., Rickerby M., Smith H., Visser M., Melton H., Tempero E., (2006), "Understanding the shape of Java software", *Proceedings of the ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages, and Applications*, Preprint 397-412, New York.

[13] Camacho J., Fernandez G.M., Verdejo J.D., Teodoro P.G., (2014), "Tackling the Big Data 4vs for anomaly detection", *Computer Communications Workshops*, Preprint, Pages 500-505.

[14] Cardona M., Chamberlin R.V., MarxW., (2007), "The history of the stretched exponential function", *Annual Physics*, Volume 16, No 12, 842-845.

[15] Chatterjee R., (2014), "Practical Methods of Financial Engineering and Risk Management: Tools for Modern Financial Professionals", Chapter 8, *Berklly CA*, Prepr*int* 315-331.

[16] Chu-Shoer J., Westover M.B., Bianchi M.T., (2010) "Power Law Versus Exponential State Transition Dynamics: Application to Sleep-Wake Architecture", *Plos ONE,* Volume **5**, Issue 12, 1-11.

[17] Clauset A., Shalizi C.R., Newman M.E.J., (2009), "Power-law distributions in empirical data", *Society for Industrial and Applied Mathematics Review,* **51**(4), 661-703.

[18] Clauset A., Young M., Gleditsch K.S. (2007), "On the Frequency of Severe Terrorist Events.", *Journal of Conflict Resolution,* Volume **51**, Issue 1, 58-87.

[19] Clauset A., (2011), "Power-law distributions and working with empirical data", *Inference. Models and Simulation for Complex Systems,* Lecture 3.

[20] Easley D., Kleinberg J. (2010), "Power Laws and Rich-Get-Richer Phenomena", *Networks, Crowds and Markets: Reasoning about a Highly Connected World*, Cambridge University Press, 543-560.

[21] Ebel H., Mielsch L.I., Bornholdt S. (2002), "Scale-free topology of e-mail networks", *Physics Review*, Volume **66**, 035103.

[22] Faloutsos M., Faloutsos P., Faloutsos C. (1999), "On Power-Law Relationships of the Internet Topology", *Computer Communication Review*, Volume **29**, Issue 4, 251-262.

[23] Friedman A. (2015), "The power of Lotka's law through the eyes of R", *Romanian Statistical Review*, Volume **2**, Issue 63, Preprint, 69-77.

[24] Gabaix X., (2009), "Power Laws in Economics and Finance", *Annual Review of Economics,* Volume **1**, 255-293.

[25] Gabaix X., (2016), "Power Laws in Economics: An Introduction", *Journal of Economics Perspectives,* Volume **30**, Number 1, 185-206.

[26] Gadepally V., Kepner J., (2015), "Using a Power Law Distribution to describe Big Data", *High Performance Extreme Computing Conference*, Preprint, 1-5.

[27] Geerolf. F. (2016): "A Theory of Pareto Distributions." Tech, rep., UCLA.

[28] Gillespie C.S., (2015), "Fitting Heavy Tailed Distribution: The poweRlaw Package", Journal of Statistical Software, Volume **64**, Issue 2, 1-16.

[29] Goldstein M.L., Morris S.A., Yen G.G., (2004), "Problems with fitting to the power-law distribution", *The European Physical Journal B,* Volume **41**, 255-258.

[30] Gong W., Liu Y., Misra V., Towsley D., (2001), "On the tails of web file size distributions"

[31] Hilbert M., (2013), "Scale-Free Power Laws as Interaction between Progress and Diffusion", *Complexity,* Volume **19**, Issue 4, 56-65.

[32] Klaus A., Yu S., Plenz D., (2011), "Statistical Analyses Support Power Law Distributions found in Neuronal Avalanches", *Plos ONE* 6.

[33] Laherrere J., Sornette D., (1998), "Stretched Exponential distributions in nature and economy: "fat tails" with characteristic scales", *European Physics Journal*, Volume **2**, 525-539.

[34] Limpert E., Stahel W.A., Abbt M., (2001), "Log-normal Distribution across the Sciences: Key and Clues", *Bioscience,* Volume **51**, 341-352.

[35] Malevergne Y., Pisarenko V., Sornette D., (2005), "Empirical distributions of Stock returns: between the stretched exponential and power law?", *Journal Quantitative Finance*, Volume **5**, Isue 4, 379-401.

[36] Milojevic S. (2010), "Power Law Distributions in Information Science: Making the Case for Logarithmic Binning", *Journal of the Accosiation for Infromartion Science and Technology,* Volume **61**, Issue 12, 2417-2425.

[37] Mitzebmacher M., (2004), "A brief history of generative models for power law and lognormal distributions, *Internet Mathematics,* Volume **1**, No. 2, 226-251.

[38] Muniruzzaman A.N.M., (1957), "On Measures of Location and Dispersion and Tests of Hypotheses in a Pare to Population", *Calcutta Statistical Association Bulletin,* Volume **7**, Issue 3, 115.

[39] Newman M.E.J., (2006), "Power Laws, Pareto distributions and Zipf's law", *Contemporary Physics,* Volume **46**, 323-351.

[40] Pueyo S., Jovani R., (2006), "Comment on "A Keystone Mutualism Drives Pattern in a Power Function"", *Science,* Volume **313**, 1739.

[41] Redner S. (1998), "How Popular is Your Paper? An Empirical Study of the Citation Distribution", *The European Physical Journal B*, Volume **4**, 131-134.

[42] Richardson L.F., (1969), "Statistics of Deadly Quarrels", *The Boxwood Press, Pittsburg.*

[43] Shatnawi R., Althebyan Q., (2013), "An Empirical Study of the Effect of Power Law Distribution on the Interpretation of OO Metrics", *ISRN Software Engineering,* Volume 2013, 1-18.

[44] Siganos G., Faloutsos M., Faloutsos P., Faloutsos C., (2003), "Power laws and the AS-level Internet topology", *IEEE|ACM Transactions on Networking*, Volume **11**, Issue 4, 514-524.

[45] Stumpf M.P.H., Porter M.A., (2012), "Critical Truths About Power Laws", *Science,* Volume 135, Issue 6069, Preprint 665-666.

[46] Virkar Y., Clauset A., (2014), "Power-law distributions in binned empirical data", *The Annals of Applied Statistics,* Volume **8**, No. 1, 89-119.

[47] Vuong Q.H., (1989), "Likelihood ratio tests for model selection and non-nested hypotheses", *Econometrica*, Volume **57**, No. 2, 307-333.

[48] White E.P., Enquist B.J., Green J.L., (2008), "On estimating the exponent of power-law frequency distributions", *Ecology,* Volume **89**, No. 4., 905-912.

[49] Willinger W., Alderson D., Doyle J.C., Li L., (2004), "More "Normal" Than Normal: Scaling Distributions and Complex Systems", *Simulation Conference Washington DC, USA*, 130-141.

[50] Zhukov L.E., (2016), "Network Science", School of Data Analysis and Artificial Intelligence, Department of Computer Science*, National Research University Higher School of Economics.*