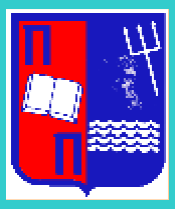
Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής



Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΚΙΝΔΥΝΩΝ**

ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΚΑΤΑΝΟΜΗΣ PARETO ΣΕ ΔΕΔΟΜΕΝΑ ΑΝΑΛΟΓΙΣΜΟΥ

ΜΑΡΙΑ-ΕΛΕΝΗ Ι.ΧΑΛΚΙΑ

Δ ι π λ ω μ α τ ι κ ή Ε ρ γ α σ ί α

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Αναλογιστική Επιστήμη και Διαχείριση Κινδύνων*

Πειραιάς
Ιούλιος 2023

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΚΙΝΔΥΝΩΝ**

APPLICATIONS OF PARETO DISTRIBUTION IN ACTUARIAL DATA

ΜΑΡΙΑ-ΕΛΕΝΗ Ι.ΧΑΛΚΙΑ

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 *Αναλογιστική Επιστήμη και Διαχείριση Κινδύνων*

Πειραιάς
Ιούλιος 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη Συνέλευση του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……..
συνεδρίασή της, σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Αναλογιστική Επιστήμη και Διαχείριση Κινδύνων.

Τα μέλη της Επιτροπής ήταν:

- Τζαβελάς Γεώργιος (Επιβλέπων)
- Βερροπούλου Γεωργία
- Πολίτης Κωνσταντίνος

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
ACTUARIAL SCIENCE AND RISK MANAGEMENT

APPLICATIONS OF PARETO DISTRIBUTION IN ACTUARIAL DATA

By

MARIA ELENI I. CHALKIA

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 Actuarial Science and Risk Management

Piraeus, Greece
July 2023

Στον σύζυγο και στην κόρη μας

Ευχαριστίες

Πρωτίστως θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης κ. Τζαβελά Γεώργιο για την λεπτομερή καθοδήγηση και πολύτιμη βοήθειά του, τόσο με την παροχή πλούσιου υλικού για την εκπόνηση της διπλωματικής μου, όσο και για τις σαφείς οδηγίες του, καθώς επίσης και τα μέλη της τριμελούς επιτροπής, κα. Βερροπούλου Γεωργία και κ. Πολίτη Κωνσταντίνο. Δε θα μπορούσα επίσης να παραλείψω να ευχαριστήσω τον σύζυγό μου για τη συμπαράσταση και τη στήριξή του όλο αυτό τον καιρό, αλλά και τους γονείς μου και τον αδελφό μου για τη βοήθειά τους

Περίληψη

Η παρούσα διπλωματική εργασία επικεντρώνεται στην κατανομή Pareto, τις μεθόδους εκτίμησης των παραμέτρων της, καθώς και στην αξιολόγηση των εκτιμητών αυτών. Ο κύριος σκοπός της εργασίας είναι η ανάλυση και η εξέταση της κατανομής Pareto με βάση πραγματικά δεδομένα, καθώς και η προσαρμογή της κατανομής σε αυτά τα δεδομένα.

Αρχικά, γίνεται μια εισαγωγή στην κατανομή και η σημασία της σε διάφορους τομείς. Εξετάζεται το θεωρητικό υπόβαθρο της κατανομής, συμπεριλαμβανομένου του ιστορικού, της μαθηματικής μορφής, των μέτρων θέσης, διασποράς και ροπής, καθώς και της λοξότητας και κύρτωσης. Επίσης, αναλύονται οι βασικές ιδιότητες της κατανομής Pareto και παρουσιάζονται οι συναρτήσεις επιβίωσης και μέσης υπολειπόμενης ζωής.

Στο τρίτο κεφάλαιο, αναλύονται μια σειρά μεθόδων εκτίμησης των παραμέτρων της κατανομής Pareto. Αρχικά, παρουσιάζεται η παραμετροποίηση της κατανομής, που περιλαμβάνει τη μορφή της κατανομής και τις παραμέτρους της. Με τη χρήση αυτών των μεθόδων εκτίμησης, είναι δυνατόν να εκτιμηθούν οι παράμετροι και να προσαρμοστεί η κατανομή στα παρατηρούμενα δεδομένα.

Στο τέταρτο κεφάλαιο περιγράφονται λεπτομερώς τα μέτρα καλής προσαρμογής που χρησιμοποιούνται για την αξιολόγηση των εκτιμητών και παρουσιάζονται οι στατιστικές συναρτήσεις που χρησιμοποιούνται για τον υπολογισμό τους. Αυτά τα μέτρα περιλαμβάνουν τον δείκτη Kolmogorov-Smirnov (KS), τον δείκτη Cramér-von Mises (CvM) και τον δείκτη Anderson-Darling (AD). Κάθε μέτρο έχει τη δική του στατιστική συνάρτηση που χρησιμοποιείται για τον υπολογισμό του και αποτελεί μια ποσοτική μέτρηση της προσαρμογής των εκτιμητών στα δεδομένα Pareto. Η παρουσίαση αυτών των στατιστικών συναρτήσεων επιτρέπει την ακριβή και αντικειμενική αξιολόγηση της απόδοσης των εκτιμητών.

Στο πέμπτο κεφάλαιο αναλύεται η δυσκολία των κατανομών να περιγράψουν πλήρως ένα σύνολο δεδομένων σε όλο το εύρος τους. Υποστηρίζεται ότι οι κατανομές μπορούν να περιγράψουν ικανοποιητικά τις μικρές ή μεσαίες τιμές, αλλά όχι τις μεγάλες. Παρουσιάζονται οι γενικές μορφές των σύνθετων (composite) συναρτήσεων, ειδικότερα της εκθετικής και της Pareto, καθώς και η σχετική μορφή της hazard.

Το έκτο κεφάλαιο επικεντρώνεται στους ευνοϊκούς εκτιμητές (favorable estimators) της κατανομής Pareto. Αρχικά, γίνεται μια εισαγωγή στο θέμα και αναλύονται τα μέτρα καλής προσαρμογής, τα οποία χρησιμοποιούνται για να αξιολογηθεί η προσαρμογή του μοντέλου στα δεδομένα. Ένα σημαντικό μέτρο καλής προσαρμογής είναι το σημείο απώλειας, το οποίο υπολογίζει το ποσοστό των παρατηρήσεων που πέφτουν έξω από την προσαρμοσμένη κατανομή. Στη συνέχεια, εξετάζεται το κριτήριο ευρωστίας, το οποίο χρησιμοποιεί τη διακύμανση των εκτιμητών ως ένα μέτρο της αστάθειας της εκτίμησης και παρουσιάζονται διάφοροι ευνοϊκοί εκτιμητές της κατανομής. Αυτοί περιλαμβάνουν εκτιμητές ποσοστημορίων, εκτιμητές κομμένου μέσου όρου και γενικευμένους εκτιμητές διαμέσων. Κάθε εκτιμητής έχει τις δικές του ιδιότητες και περιορισμούς και χρησιμοποιείται ανάλογα με την εφαρμογή και τα δεδομένα που διαθέτουμε.

Στο έβδομο κεφάλαιο της διπλωματικής αναλύεται η εφαρμογή όλων των εκτιμητών σε πραγματικά δεδομένα και παρουσιάζονται σημαντικά συμπεράσματα σχετικά με την απόδοση και την ακρίβεια των εκτιμητών. Αρχικά, περιγράφεται η επιλογή του δείγματος και η μεθοδολογία που ακολουθήθηκε για την αξιολόγηση των εκτιμητών. Εξηγείται ο τρόπος υπολογισμού των εκτιμητών και παρουσιάζονται γραφήματα που απεικονίζουν την αρχική εκτίμηση προσαρμογής των δεδομένων. Στη συνέχεια, παρουσιάζονται οι στατιστικοί έλεγχοι που πραγματοποιήθηκαν και τα αποτελέσματα τους. Τέλος, γίνεται η κατάταξη των εκτιμητών βάσει της ανθεκτικότητάς τους και της απόδοσής τους, προκειμένου να προκύψουν συγκεκριμένες συστάσεις για την επιλογή του βέλτιστου εκτιμητή. Οι αναλύσεις και τα αποτελέσματα αυτά συνοψίζουν την αξιοπιστία και την απόδοση των εκτιμητών και αποτελούν σημαντική συνεισφορά στην κατανόηση του μοντέλου. Τα σχήματα, ο προσδιορισμός των εκτιμητών και οι αντίστοιχοι έλεγχοι πραγματοποιήθηκαν με χρήση της Mathematica.

Abstract

The present thesis focuses on the Pareto distribution, the estimating methods of its parameters, and the evaluation of them. The main objective of the thesis is to analyze and examine the Pareto distribution based on real data, as well as to fit the distribution to this data.

Initially, an introduction to the Pareto distribution is provided, along with its significance in various fields. The theoretical background of the distribution is examined, including its history, mathematical form, measures of location, dispersion, and moments, as well as skewness and kurtosis. Furthermore, the basic properties of the Pareto distribution are analyzed, and the survival and mean residual life functions are presented.

In the third chapter, a series of methods for estimating the parameters of the Pareto distribution are discussed. Firstly, the parameterization of the distribution is presented, which includes the form of the distribution and its parameters. Using these estimation methods, it is possible to estimate the parameters of the Pareto distribution and fit the distribution to observed data.

The fourth chapter provides a detailed description of goodness-of-fit measures used for evaluating the estimators, and the statistical functions used to compute them. These measures include the Kolmogorov-Smirnov (KS) statistic, the Cramér-von Mises (CvM) statistic, and the Anderson-Darling (AD) statistic. Each measure has its own statistical function used for its computation and represents a quantitative measure of the fit of the estimators to the Pareto data. The presentation of these statistical functions allows for accurate and objective evaluation of the performance of the estimators.

In the fifth chapter, the difficulty of distributions to fully describe a dataset across its entire range is analyzed. It is argued that distributions can adequately describe small or medium values but not the large ones. The general forms of composite functions, specifically the exponential and Pareto distributions, are presented, along with their respective hazard functions.

The sixth chapter focuses on the favorable estimators of the Pareto distribution. Initially, an introduction to the topic is provided, and the goodness-of-fit measures used to assess the fit of the Pareto model to the data are analyzed. An important measure of goodness-of-fit is the loss function, which calculates the percentage of observations falling outside the fitted distribution. Furthermore, the criterion of robustness is examined, which uses the variance of the estimators as a measure of estimation instability, and various favorable estimators of the Pareto distribution are presented. These include percentile estimators, truncated mean estimators, and generalized median estimators. Each estimator has its own properties and limitations and is used according to the application and available data.

The seventh chapter analyzes the application of all estimators to real data and presents significant conclusions regarding the performance and accuracy of the estimators. Initially, the selection of the sample and the methodology followed for the evaluation of the estimators are described. The computation of the estimators is explained, and graphs depicting the initial fitting of the data are presented. Subsequently, the conducted statistical tests and their results are presented. Finally, a ranking of the estimators based on their robustness and performance is performed to derive specific recommendations for selecting the optimal estimator. These analyses and results summarize the reliability and performance of the estimators and contributes to the better understanding of the model. The graphs, determination of estimators, and corresponding tests were conducted using Mathematica.

Περιεχόμενα

Κατάλογος Πινάκων xvii

Κατάλογος Σχημάτων xix

1. Εισαγωγή 21

1.1 Εισαγωγή στο θέμα της κατανομής Pareto 21

1.2 Σημασία και σκοπός της εργασίας 22

2. Θεωρητικό υπόβαθρο 24

2.1 Ιστορικό της κατανομής Pareto 24

2.2 Μαθηματική μορφή 24

2.3 Μέτρα θέσης, διασποράς και ροπή 27

2.4 Λοξότητα και κύρτωση 28

2.5 Βασικές ιδιότητες 30

2.6 Ρυθμός αποτυχίας, συναρτήσεις επιβίωσης και μέσης υπολειπόμενης ζωής 31

3. Μέθοδοι εκτίμησης παραμέτρων 34

3.1 Παραμετροποίηση 34

3.2 Μέθοδος μέγιστης πιθανοφάνειας 35

3.3 Μέθοδος των ροπών 38

3.4 Μέθοδος των ποσοστιαίων σημείων 40

3.5 Μέθοδος των ελαχίστων τετραγώνων 41

4. Μέτρα καλής προσαρμογής 45

4.1 Εισαγωγή 45

4.2 KS στατιστική Dn 45

4.3 CvM στατιστική W_n^2 47

4.4 AD στατιστική A_n^2 48

5. Ο ρόλος της κατανομής Pareto στις σύνθετες (composite) συναρτήσεις 50

5.1 Εισαγωγή 50

5.2 Σύνθετες (composite) συναρτήσεις 51

6. Ευνοϊκοί εκτιμητές της κατανομής Pareto 55

6.1 Εισαγωγή 55

6.2 Κριτήριο ευρωστίας: Σημείο απώλειας 57

6.3 Κριτήριο αποτελεσματικότητας: Διακύμανση 58
6.4 Εκτιμητές ποσοστημορίων 59
6.5 Εκτιμητές κομμένου μέσου όρου 61
6.6 Γενικευμένοι εκτιμητές διαμέσων 62

7. Εφαρμογή σε πραγματικά δεδομένα 64
7.1 Επιλογή δείγματος 64
7.2 Μεθοδολογία 65
7.3 Αρχικές εκτιμήσεις 66
7.4 Αναλυτική εφαρμογή 71
7.5 Αποτελέσματα 77
8. Συμπεράσματα 82
8.1 Ανασκόπηση των βασικών αποτελεσμάτων της εφαρμογής 82
8.2 Συνολικά συμπεράσματα της εργασίας 83

Παραρτήματα 86
Π1. Κώδικας Σχημάτων με χρήση της Mathematica 87
Π2. Κώδικας Προγράμματος 90
Περίληψη 11
Abstract 13
Βιβλιογραφία 94

1 Κατάλογος Πινάκων

2-1	Είδη κατανομών Pareto	26
6-1	Τιμές C_k, για k από 2 μέχρι 10	62
7-1	Απώλειες λόγω καταστροφών από ανέμους	64
7-2	Τιμές του $\hat{a}$, στατιστικές και βαθμοί για τα δεδομένα ανέμου.	78
7-3	Τυπικές αποκλίσεις εκτιμητών	79
7-4	Συσχέτιση τιμών κατάταξης	80
7-5	Μεταβλητότητα μέτρων καλής προσαρμογής	80
7-6	Τελική κατάταξη εκτιμητών	81

2 Κατάλογος Σχημάτων

2-1	Διάγραμμα πυκνότητας Pareto για τιμές παραμέτρων σ=2 και α = 1, 2, 3	25
2-2	Διάγραμμα αθροιστικής συνάρτησης Pareto για τιμές παραμέτρων σ=2 και α = 1, 2, 3	26
2-3	Διάγραμμα επιβίωσης Pareto για τιμές παραμέτρων σ=2 και α = 1, 2, 3	27
2-4	Διάγραμμα συνάρτησης λοξότητας Pareto	29
2-5	Διάγραμμα συνάρτησης κύρτωσης Pareto	30
2-6	Διάγραμμα συνάρτησης λ(x) κατανομής Pareto για τιμές παραμέτρων α=1 και α=4.	32
5-1	Διάγραμμα Εκθετικής, Pareto και σύνθετης κατανομής για τιμές παραμέτρων α=0.35 και θ=1.	53
7-1	Ιστόγραμμα συχνοτήτων δεδομένων ανέμου	67
7-2	Διάγραμμα QQ κατανομής Pareto για τα δεδομένα μας	69
7-3	Διάγραμμα QQ εκθετικής κατανομής για τα δεδομένα μας	70
7-4	Διάγραμμα εμπειρικής – θεωρητικής CDF για α̂=0,795341 και σ̂=1,58	72
7-5	Διάγραμμα εμπειρικής – θεωρητικής CDF για α̂=1,2015 και σ̂=1,54712	74
7-6	Διάγραμμα εμπειρικής – θεωρητικής CDF για α̂=0.768658 και σ̂=1.54349	76

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

1.1 Εισαγωγή στο θέμα της κατανομής Pareto

Η κατανομή Pareto είναι μια στατιστική κατανομή που χρησιμοποιείται στη μελέτη των φαινομένων που εμφανίζουν μια μη γραμμική σχέση της μορφής $y = \alpha x^{\beta}$, όπου α και β είναι σταθερές. Τα φαινόμενα αυτά παρουσιάζουν μια σημαντική αυξητική συμπεριφορά για μικρές τιμές του x, αλλά αποκτούν ακόμα μεγαλύτερη αυξητική συμπεριφορά για μεγαλύτερες τιμές του x, δημιουργώντας μια "ουρά" στην κατανομή των τιμών.

Η περιγραφή της κατανομής δίνεται από μια παραμετροποιημένη συνάρτηση πυκνότητας πιθανότητας (PDF), η οποία έχει μια παράμετρο σχήματος και μια κατώτατη οριακή τιμή (ή κλίμακα). Η παράμετρος σχήματος αναφέρεται στο πώς αλλάζει η πιθανότητα των μεγαλύτερων τιμών καθώς αυξάνεται η τιμή, ενώ η κλίμακα καθορίζει την ελάχιστη τιμή που μπορεί να λάβει η μεταβλητή.

Η κατανομή Pareto διακρίνεται από μια “βαριά” ουρά, όπου οι μεγαλύτερες τιμές έχουν πολύ μικρή πιθανότητα να συμβούν, ενώ οι μικρότερες τιμές είναι σχετικά πιθανότερες να συμβούν. Αυτό το φαινόμενο αντικατοπτρίζει την ύπαρξη εξαιρετικά μεγάλων τιμών που μπορούν να εμφανιστούν σε ορισμένα γεγονότα, όπως για παράδειγμα στα κέρδη των επιχειρήσεων ή στο μέγεθος των πόλεων.

Η κατανομή είναι στενά συνδεδεμένη με τον νόμο των λίγων (the law of the few). Αυτός ο νόμος αναφέρεται στο γεγονός ότι μια μικρή μερίδα ατόμων, επιχειρήσεων ή πόλεων ελέγχει το μεγαλύτερο μέρος των πόρων ή των πλούτων και περιγράφει τη συμπεριφορά αυτών των λίγων ατόμων, εξηγώντας την ανισότητα στην κατανομή των πόρων στην κοινωνία.

Η κατανομή Pareto έχει εφαρμογές σε πολλούς τομείς, όπως η οικονομία, η χρηματοοικονομία, η κοινωνιολογία και η μηχανική. Η χρήση της μπορεί να βοηθήσει στην

κατανόηση των ανισοτήτων στην κατανομή των πόρων, στην ανάλυση των κερδών των επιχειρήσεων, στον σχεδιασμό συστημάτων ασφάλειας και στη μελέτη των δικτύων και της κατανομής της κίνησης στο διαδίκτυο.

Στον οικονομικό τομέα, χρησιμοποιείται για την ανάλυση της κατανομής του πλούτου και των εισοδημάτων και στην χρηματοοικονομία για τον υπολογισμό των αποδόσεων των επενδύσεων και των αναλύσεων της κίνησης των μετοχών.

Στην κοινωνιολογία, η κατανομή χρησιμοποιείται για την ανάλυση της κατανομής της πλούσιας και φτωχής τάξης σε μια κοινωνία, καθώς και της κατανομής της εξουσίας και της επιρροής στον κοινωνικό χώρο.

Στη μηχανική, χρησιμοποιείται για τη μοντελοποίηση της αντοχής των υλικών και των δομών, καθώς και για τον υπολογισμό των αποδόσεων και της αξιοπιστίας των συστημάτων.

Η κατανομή Pareto έχει επίσης εφαρμογές στην ανάλυση των δικτύων και της κίνησης στο διαδίκτυο. Για παράδειγμα, χρησιμοποιείται για την ανάλυση της κατανομής του μεγέθους των δικτυακών αρχείων, όπως εικόνες, βίντεο και ήχου. Βοηθάει στην κατανόηση του πόσο συχνά λαμβάνονται ή μεταφέρονται αρχεία με διαφορετικό μέγεθος, καθώς και πόσο σημαντικό είναι να διαθέτουν οι διακομιστές αποθήκευσης αρκετό χώρο για τα μεγάλα αρχεία. Μπορεί να συμβάλει στην πρόβλεψη της κίνησης και της επιβάρυνσης του δικτύου σε κάθε σημείο του χρόνου. Αυτό είναι σημαντικό για τους παρόχους διαδικτυακών υπηρεσιών, καθώς μπορούν να αντιμετωπίσουν καλύτερα τη ζήτηση και να βελτιώσουν την απόδοση του δικτύου τους.

Τέλος, η κατανομή Pareto χρησιμοποιείται σε πολλά προβλήματα στατιστικής ανάλυσης και πιθανοτήτων, όπως η πρόβλεψη της εμφάνισης των επιθέσεων κακόβουλου λογισμικού στα συστήματα ασφαλείας, η ανάλυση της διάγνωσης καρκίνου και η εκτίμηση του κόστους των ασφαλίστρων υγείας.

1.2 Σημασία και σκοπός της εργασίας

Σκοπός της εργασίας είναι η εξέταση της κατανομής Pareto και η κατανόηση της κατανομής η οποία περιγράφεται από μία συγκεκριμένη συνάρτηση πυκνότητας πιθανότητας και

χρησιμοποιείται για να περιγράψει τη συμπεριφορά των δεδομένων σε μια διανομή με βαριά ουρά. Ο σκοπός αυτός είναι σημαντικός, διότι μπορεί να βοηθήσει στην καλύτερη κατανόηση των φαινομένων που παρατηρούνται σε αυτούς τους τομείς και στην εξέλιξη αντίστοιχων πολιτικών ή οικονομικών αποφάσεων.

Η εργασία αποδίδει σημασία στην κατανόηση της κατανομής λόγω του γεγονότος ότι η συγκεκριμένη κατανομή συναντάται συχνά σε πραγματικά δεδομένα και επομένως είναι απαραίτητο να γνωρίζουμε πώς να τα αναλύουμε και να ερμηνεύουμε σωστά τα αποτελέσματα.

Επιπλέον, η σημασία της εργασίας είναι η περιγραφή γνωστών μεθόδων από την βιβλιογραφία, χρήσιμων στην επίλυση προβλημάτων σε διάφορους τομείς, όπως στην οικονομία, στη βιομηχανία, στην κοινωνιολογία και στην πληροφορική, οι οποίες κάνουν χρήση της κατανομής Pareto. Για παράδειγμα, μπορεί να χρησιμοποιηθεί για την ανάλυση του μεγέθους των δεδομένων, όπως στην ανάλυση δεδομένων που προκύπτουν από το διαδίκτυο. Η κατανομή Pareto μπορεί να χρησιμοποιηθεί επίσης για την ανάλυση της κατανομής του μεγέθους των αρχείων, των εικόνων ή των βίντεο σε μια ιστοσελίδα.

Κ Ε Φ Α Λ Α Ι Ο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό, οι πληροφορίες αντλήθηκαν από την ακόλουθη βιβλιογραφία και επιστημονικές πηγές: Βλαχάκης και Δαγκλής (2010), Κωνσταντάτος και Παπαγεωργίου (2006), Μπαλταδώρος (2011), Clauset, Shalizi και Newman (2009), Costa, Machado και Ferreira (2011), Jain και Khan (2018) και Weisstein (2003).

2.1 Ιστορικό της κατανομής Pareto

Η κατανομή Pareto έλαβε το όνομά της από τον Ιταλό οικονομολόγο Vilfredo Pareto (1843-1923) η οποία προτάθηκε από τον ίδιο το 1896. Η ιστορία της κατανομής ξεκίνησε όταν ο Pareto άρχισε να ερευνά την οικονομία της Ιταλίας την συγκεκριμένη εποχή. Ο Pareto μελετώντας την κατανομή των πλουσίων στην Ιταλία, ανακάλυψε ότι υπήρχε μια κατά πολύ ανισότερη κατανομή της περιουσίας από ό,τι θα περίμενε κανείς. Συγκεκριμένα, παρατήρησε ότι οι πλούσιοι αποτελούσαν μια μικρή μειονότητα που ελέγχαν το μεγαλύτερο μέρος της περιουσίας. Αργότερα, οι έρευνες του Pareto έδειξαν ότι η κατανομή ισχύει για πολλά διαφορετικά φαινόμενα στη φύση και στην κοινωνία, όπως την κατανομή των πλούσιων στην Αμερική, τη συχνότητα εμφάνισης των λέξεων σε μια γλώσσα και το μέγεθος των πόλεων.

2.2 Μαθηματική μορφή

Μια συνεχής τυχαία μεταβλητή X λέμε ότι ακολουθεί την κατανομή Pareto με παραμέτρους α (σχήματος) και σ (κλίμακας), το οποίο συμβολίζουμε με $X \sim P(\alpha, \sigma)$, αν η **συνάρτηση πυκνότητας πιθανότητας** δίνεται από τον τύπο:

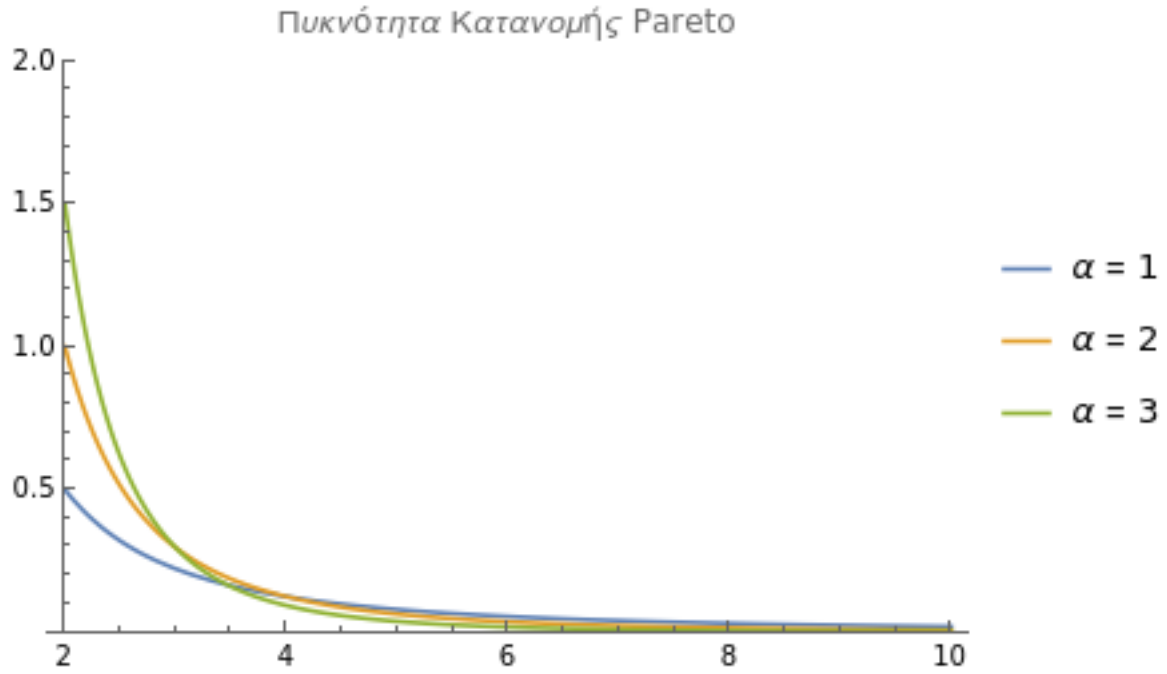
$$f_X(x) = \begin{cases} \dfrac{\alpha\sigma^{\alpha}}{x^{\alpha+1}} & x \geq \sigma, \alpha > 0 \\ 0 & x < \sigma \end{cases}$$

Στο παρακάτω Σχήμα 2-1 δείχνουμε την επίδραση των παραμέτρων στην συνάρτηση

πυκνότητας πιθανότητας Pareto.

ΣΧΗΜΑ 2-1

Διάγραμμα πυκνότητας Pareto για τιμές παραμέτρων σ = 2 και α = 1, 2, 3



Η **αθροιστική συνάρτηση κατανομής** (cumulative distribution function - CDF) δίνεται από τον τύπο

$$F_X(x) = 1 - \left(\frac{\sigma}{x}\right)^{\alpha}$$

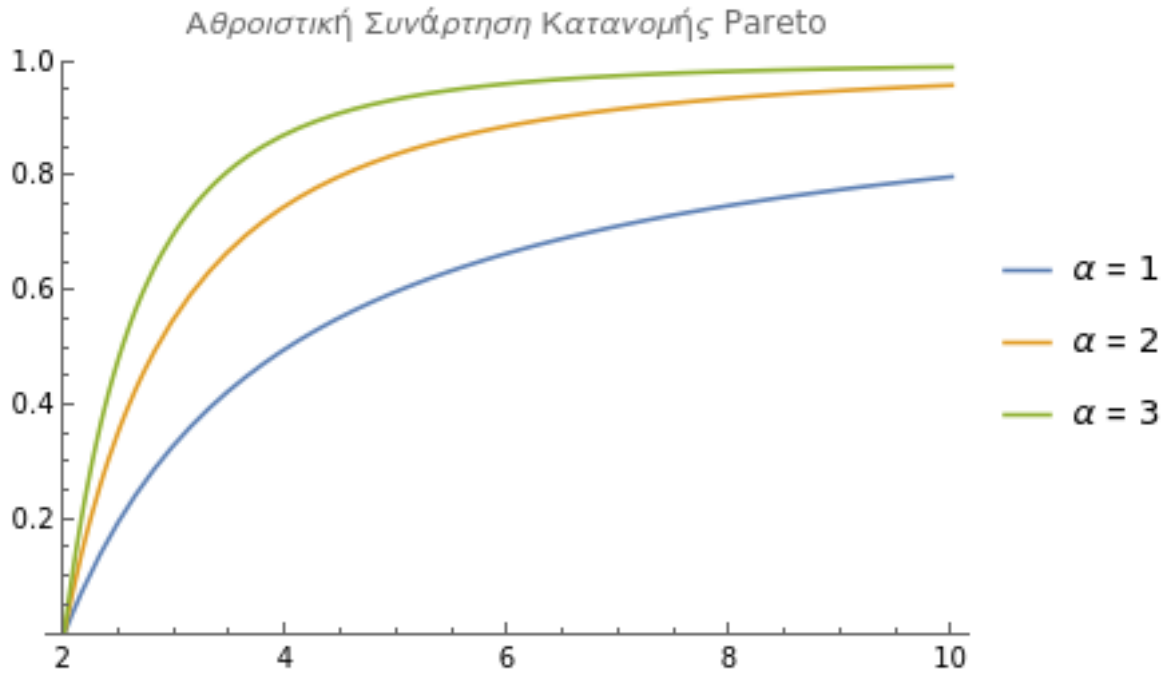
όπου $x \geq \sigma,\ \alpha > 0$. Όταν $x < \sigma$ τότε $F_X(x) = 0$.

Μπορεί εύκολα να φανεί ότι η αθροιστική συνάρτηση κατανομής είναι μια φθίνουσα συνάρτηση ως προς την ανεξάρτητη μεταβλητή x, το οποίο αποτελεί μια ακόμη ένδειξη ότι η κατανομή έχει βαριά ουρά.

Στο παρακάτω Σχήμα 2-2 δείχνουμε την επίδραση των παραμέτρων στην αθροιστική συνάρτηση κατανομής Pareto.

ΣΧΗΜΑ 2-2

Διάγραμμα αθροιστικής συνάρτησης Pareto για τιμές παραμέτρων σ=2 και α = 1, 2, 3.



Η **συνάρτηση επιβίωσης** (survival function) της κατανομής ορίζεται ως η πιθανότητα να επιβιώσει η μεταβλητή X πάνω από μια συγκεκριμένη τιμή x και δίνεται από την σχέση

$$S(x) = P(X > x) = 1 - F(x) = \left(\frac{\sigma}{x}\right)^{\alpha}$$

Η κατανομή Pareto μπορεί να γενικευθεί με διάφορους τρόπους. Ο παρακάτω πίνακας παρουσιάζει κάποιες γενικεύσεις της.

ΠΙΝΑΚΑΣ 2.1
Είδη κατανομών Pareto

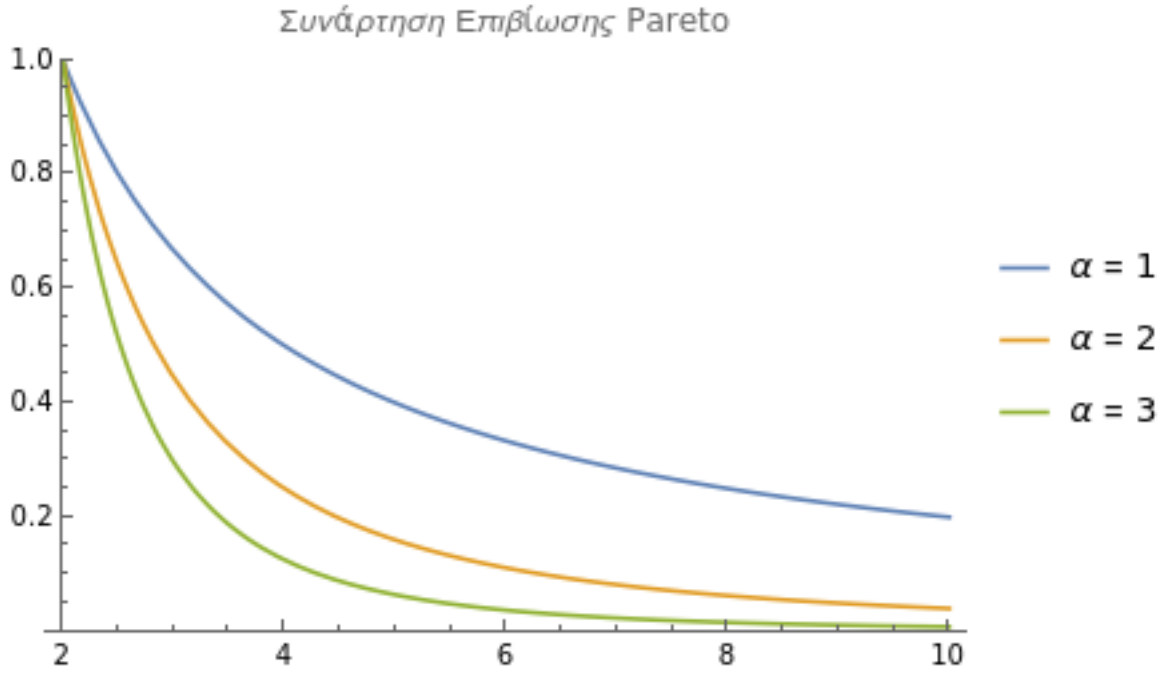
Τύπος	Συνάρτηση επιβίωσης	Πεδίο ορισμού	Παράμετροι
Τύπος I	$\left(\frac{\sigma}{x}\right)^{\alpha}$	$x \geq \sigma$	$\sigma > 0, \alpha > 0$
Τύπος II	$\left(1 + \frac{x-\mu}{\sigma}\right)^{-\alpha}$	$x \geq \mu$	$\mu \in \mathbb{R}, \sigma > 0, \alpha > 0$
Lomax	$\left(1 + \frac{x}{\sigma}\right)^{-\alpha}$	$x \geq 0$	$\mu = 0, \sigma > 0, \alpha > 0$
Τύπος III	$\left[1 + \left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}\right]^{-1}$	$x \geq \mu$	$\mu \in \mathbb{R}, \sigma > 0, \gamma > 0$
Τύπος IV	$\left[1 + \left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}\right]^{-\alpha}$	$x \geq \mu$	$\mu \in \mathbb{R}, \sigma > 0, \gamma > 0, \alpha > 0$

Από τους παραπάνω τύπους στην παρούσα διπλωματική θα ασχοληθούμε με τον τύπο I και ο οποίος συνήθως αναφέρεται σαν κατανομή Pareto.

Στο παρακάτω Σχήμα 2-3 δείχνουμε την επίδραση των παραμέτρων στην συνάρτηση επιβίωσης Pareto.

ΣΧΗΜΑ 2-3

Διάγραμμα επιβίωσης Pareto για τιμές παραμέτρων $\sigma = 2$ και $\alpha = 1, 2, 3$



2.3 Μέτρα θέσης, διασποράς και ροπή

Η **μέση τιμή** και η **διακύμανση** δίνονται από τους τύπους

$$\mathrm{E(X)} = \frac{\alpha\sigma}{(\alpha - 1)} \quad \text{για } \alpha > 1$$

και

$$\mathrm{Var(X)} = \frac{\alpha\sigma^2}{(\alpha - 1)^2(\alpha - 2)} \quad \text{για } \alpha > 2$$

αντίστοιχα.

Η **k-η ροπή** της κατανεμημένης τυχαίας μεταβλητής Pareto υπάρχει, αν και μόνο εάν, $\alpha > k$. Αν k είναι θετικός ακέραιος τότε

$$E(X^k) = \frac{k!\,\sigma^k}{(\alpha-1)\cdots(\alpha-k)} \quad \alpha > k$$

Η συνθήκη α > k επιβάλλεται για να εξασφαλιστεί ότι ο παρονομαστής της εξίσωσης δεν γίνεται μηδέν ή αρνητικός. Όταν αυτό συμβαίνει, η ροπή δεν μπορεί να οριστεί μαθηματικά.

Στη συνθήκη α > k, η παράμετρος α αντιπροσωπεύει τη σχετική σπανιότητα των μεγάλων τιμών, ενώ η τάξη k αντιπροσωπεύει τον βαθμό της ροπής που εξετάζουμε. Όταν η τιμή του α είναι μεγαλύτερη από την τάξη k, αυτό υποδηλώνει ότι η κατανομή έχει αρκετά μεγάλες τιμές και η ροπή αυτής της τάξης μπορεί να υπολογιστεί με ακρίβεια. Από στατιστικής άποψης, η κατανόηση της συνθήκης α > k μας βοηθά να εξασφαλίσουμε πότε μπορούμε να χρησιμοποιήσουμε τον τύπο για τον υπολογισμό της k-ης ροπής της κατανεμημένης τυχαίας μεταβλητής Pareto.

Από την άλλη πλευρά, όταν ικανοποιείται η συνθήκη α ≤ k, η κατανομή Pareto δεν έχει ορισμένη την k-η ροπή. Αυτό συμβαίνει όταν η τιμή της α είναι μικρότερη ή ίση με την τάξη k. Σε αυτήν την περίπτωση, η κατανομή έχει ακραίες μεγάλες τιμές που δεν συγκλίνουν και η ροπή δεν μπορεί να οριστεί με ακρίβεια.

2.4 Λοξότητα και κύρτωση

Η κατανομή Pareto έχει δύο κύρια χαρακτηριστικά, τη λοξότητα και την κύρτωση, που περιγράφουν το σχήμα της κατανομής. Οι δύο βασικοί συντελεστές για την μέτρηση των χαρακτηριστικών αυτών είναι: ο συντελεστής λοξότητας και ο συντελεστής κύρτωσης.

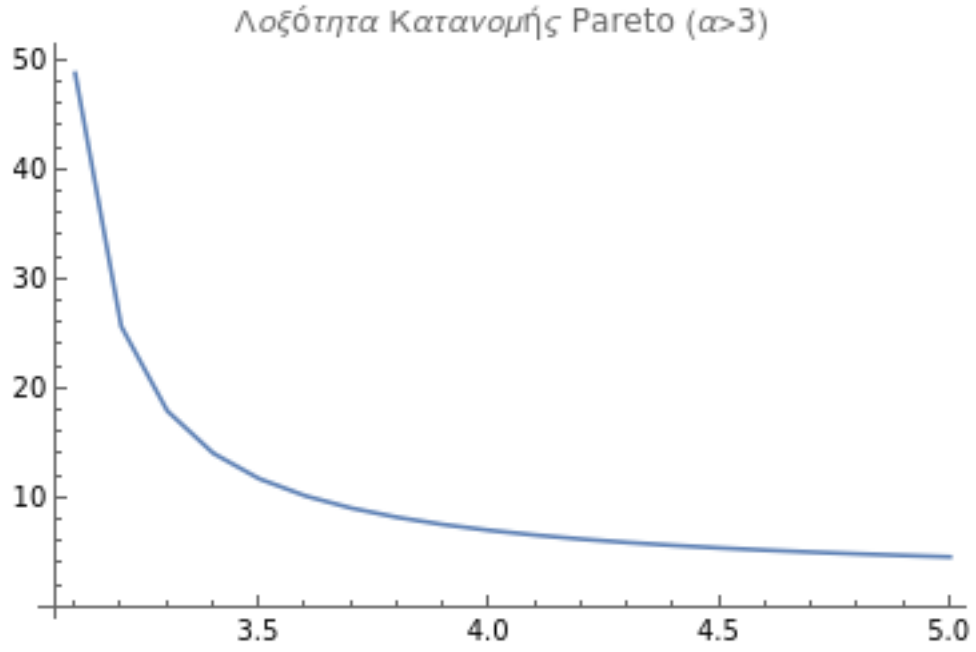
Ο συντελεστής λοξότητας (skewness coefficient) καθορίζει πόσο ασύμμετρη είναι η κατανομή. Αν ο συντελεστής λοξότητας είναι μικρότερος από μηδέν, η κατανομή είναι αριστερά λοξή (αριστερή ουρά είναι μεγαλύτερη από τη δεξιά), ενώ αν είναι μεγαλύτερος από μηδέν, η κατανομή είναι δεξιά λοξή (δεξιά ουρά είναι μεγαλύτερη από την αριστερή). Η κατανομή Pareto είναι εξίσου πιθανή να είναι αριστερά ή δεξιά λοξή, ανάλογα με τον συντελεστή λοξότητας.

$$\text{Λοξότητα} = \frac{2(1+\alpha)}{\alpha-3}\sqrt{\frac{\alpha-2}{\alpha}}\ ,\ \alpha > 3$$

Στο παρακάτω Σχήμα 2-4 δείχνουμε την γραφική παράσταση της συνάρτησης λοξότητας.

ΣΧΗΜΑ 2-4

Διάγραμμα συνάρτησης λοξότητας Pareto



Ο συντελεστής κύρτωσης (kurtosis coefficient) καθορίζει πόσο κεντραρισμένη ή "στενή" είναι η κατανομή σε σχέση με την κανονική κατανομή. Αν ο συντελεστής κύρτωσης είναι μικρότερος από μηδέν, η κατανομή είναι "πλατύτερη" από την κανονική κατανομή, με περισσότερα ακραία σημεία. Αντίθετα, αν ο συντελεστής κύρτωσης είναι μεγαλύτερος από μηδέν, η κατανομή είναι "στενότερη" από την κανονική κατανομή, με λιγότερα ακραία σημεία.

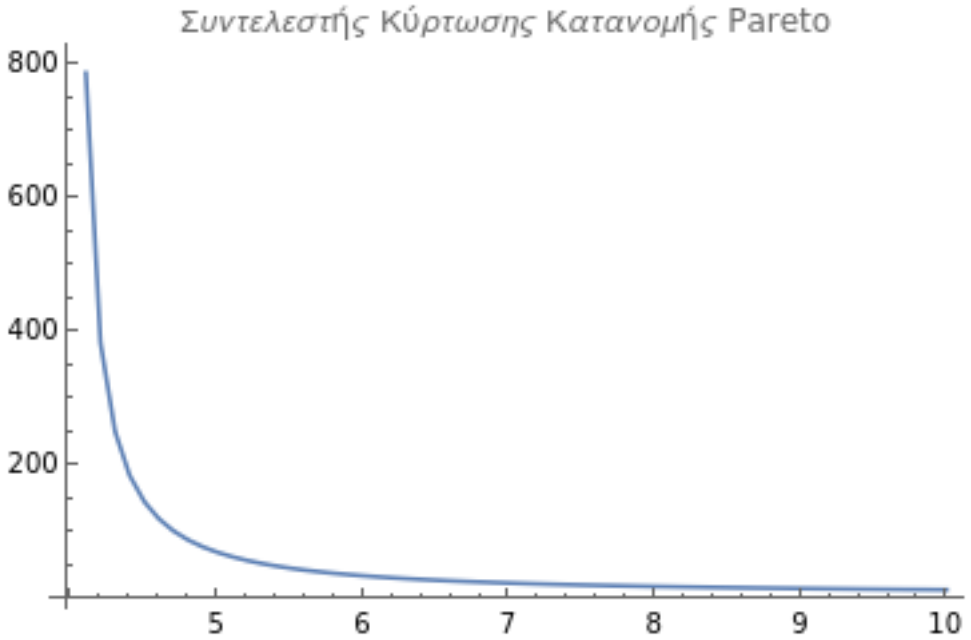
Μια κατανομή με μεγάλη συγκέντρωση τιμών γύρω από τον μέσο της όρο ονομάζεται λεπτόκυρτη (leptokurtic), ενώ μια κατανομή με μικρή συγκέντρωση τιμών κοντά στον μέσο όρο της λέγεται πλατύκυρτη (platykurtic). Στην περίπτωση της λεπτόκυρτης κατανομής, η κύρτωση είναι μεγαλύτερη του 3 (kurtosis > 3), ενώ στην περίπτωση της πλατύκυρτης κατανομής είναι μικρότερη του 3 (kurtosis < 3). Οι κατανομές που προσεγγίζονται από την κανονική κατανομή ονομάζονται μεσόκυρτες (mesokurtic) και η τιμή του συντελεστή κύρτωσης είναι ίση με 3.

$$\text{Κύρτωση} = \frac{6(\alpha^3+\alpha^2-6\alpha-2)}{\alpha(\alpha-3)(\alpha-4)} \ , \ \alpha > 4$$

Στο παρακάτω Σχήμα 2-5 δείχνουμε την γραφική παράσταση της συνάρτησης κύρτωσης.

ΣΧΗΜΑ 2-5

Διάγραμμα συνάρτησης κύρτωσης Pareto



Οι συντελεστές λοξότητας και κύρτωσης βοηθούν να κατανοήσουμε καλύτερα τη συμπεριφορά των δεδομένων που ακολουθούν αυτή την κατανομή. Επίσης, οι συντελεστές αυτοί μπορούν να χρησιμοποιηθούν για να συγκρίνουμε την κατανομή Pareto με άλλες κατανομές και να αξιολογήσουμε πόσο καλά προσαρμόζεται στα δεδομένα μας.

2.5 Βασικές ιδιότητες

Η κατανομή Pareto έχει τις παρακάτω βασικές ιδιότητες:

● *Βαριά ουρά*: Η κατανομή έχει μια βαριά ουρά, που σημαίνει ότι οι ακραίες τιμές είναι πολύ πιθανό να εμφανίζονται συχνότερα από ό,τι σε άλλες κατανομές. Είναι μια ιδιότητα που αναφέρεται στο γεγονός ότι η πιθανότητα να παρατηρηθεί μια ακραία τιμή στην κατανομή είναι αρκετά υψηλή σε σχέση με τις πιθανότητες των μετρήσεων στην κορυφή της κατανομής. Αυτό σημαίνει ότι η κατανομή Pareto παρουσιάζει μια πτώση της πυκνότητας των παρατηρήσεων στην ουρά της κατανομής, αλλά αυτή η πτώση είναι πολύ αργή σε σχέση με άλλες κατανομές. Η ιδιότητα αυτή έχει σημαντικές επιπτώσεις στα αποτελέσματα που προκύπτουν από την ανάλυση των δεδομένων. Συγκεκριμένα, καθιστά δυσκολότερη την εκτίμηση της μέσης τιμής και της

διασποράς των δεδομένων. Επιπλέον, η βαριά ουρά επιτρέπει την ύπαρξη ακραίων τιμών, οι οποίες μπορεί να είναι σημαντικές για την κατανόηση των δεδομένων και των φαινομένων που εξετάζονται.

● *Δύναμη κλιμάκωσης (power law):* Η κατανομή Pareto είναι μια κατανομή με δύναμη κλιμάκωσης, που σημαίνει ότι η συμπεριφορά της στην ουρά δεν είναι παρόμοια με τη συμπεριφορά της στη βάση. Πιο συγκεκριμένα, η ιδιότητα της δύναμης κλιμάκωσης δείχνει ότι η συχνότητα εμφάνισης των τιμών στην ουρά της κατανομής, ακολουθεί μια κλιμάκωση στις ακραίες τιμές, δηλαδή ότι η πιθανότητα εμφάνισης μιας μεγάλης τιμής μειώνεται ραγδαία καθώς αυξάνονται οι τιμές αυτές.

● *Σχέση μεταξύ μέσης τιμής και διακύμανσης*: Η κατανομή δεν έχει ορισμένη διακύμανση αν ο δείκτης σχήματος είναι μικρότερος από 2, αλλά έχει μια σχέση μεταξύ μέσης τιμής και διακύμανσης. Αυτό σημαίνει ότι όσο αυξάνεται η μέση τιμή της κατανομής, τόσο ο αριθμός των ακραίων τιμών αυξάνεται και έτσι αυξάνεται και η ασυμμετρία της κατανομής. Όσο αυξάνεται ο δείκτης σχήματος α, τόσο μειώνεται η μέση τιμή της κατανομής, ενώ αυξάνεται η ασυμμετρία της κατανομής και η ανομοιομορφία της. Επίσης, παρατηρούμε ότι όσο μεγαλώνει ο δείκτης σχήματος α, τόσο αυξάνεται η διακύμανση της κατανομής, αλλά αυτό γίνεται με έναν ρυθμό που μειώνεται με τον αριθμό α, ενώ για α=2 δεν υπάρχει διακύμανση.

2.6 Ρυθμός αποτυχίας και συνάρτηση μέσης υπολειπόμενης ζωής

Η κατανομή Pareto έχει συχνά χρησιμοποιηθεί για να περιγράψει συστήματα όπου συμβαίνουν σπάνιες αλλά σημαντικές παρεμβάσεις ή αποτυχίες. Ένας τρόπος για να περιγράψουμε την απόδοση ενός τέτοιου συστήματος είναι μέσω του ρυθμού αποτυχίας και της συνάρτησης μέσης υπολειπόμενης ζωής οι οποίες αποτελούν δύο σημαντικές μετρικές που μας βοηθούν να κατανοήσουμε καλύτερα τις ιδιότητες αυτής της κατανομής.

Ο **ρυθμός αποτυχίας** είναι ο αριθμός των αποτυχιών που συμβαίνουν ανά μονάδα χρόνου και εκφράζεται ως ο λόγος των αποτυχιών προς τον χρόνο. Η **συνάρτηση μέσης υπολειπόμενης ζωής** είναι η μέση διάρκεια ζωής ενός συστήματος μετά την αποτυχία του.

Ο ρυθμός αποτυχίας ή ρυθμός κινδύνου ή ένταση θνησιμότητας μιας θετικής συνεχούς τυχαίας μεταβλητής X, που έχει συνάρτηση πυκνότητας πιθανότητας f(x) και αθροιστική

συνάρτηση κατανομής F(x), ορίζεται από την σχέση:

$$\lim_{\Delta x \to 0} \frac{P(X < x + \Delta x \mid X > x)}{\Delta x} = \frac{f(x)}{1 - F(x)}$$

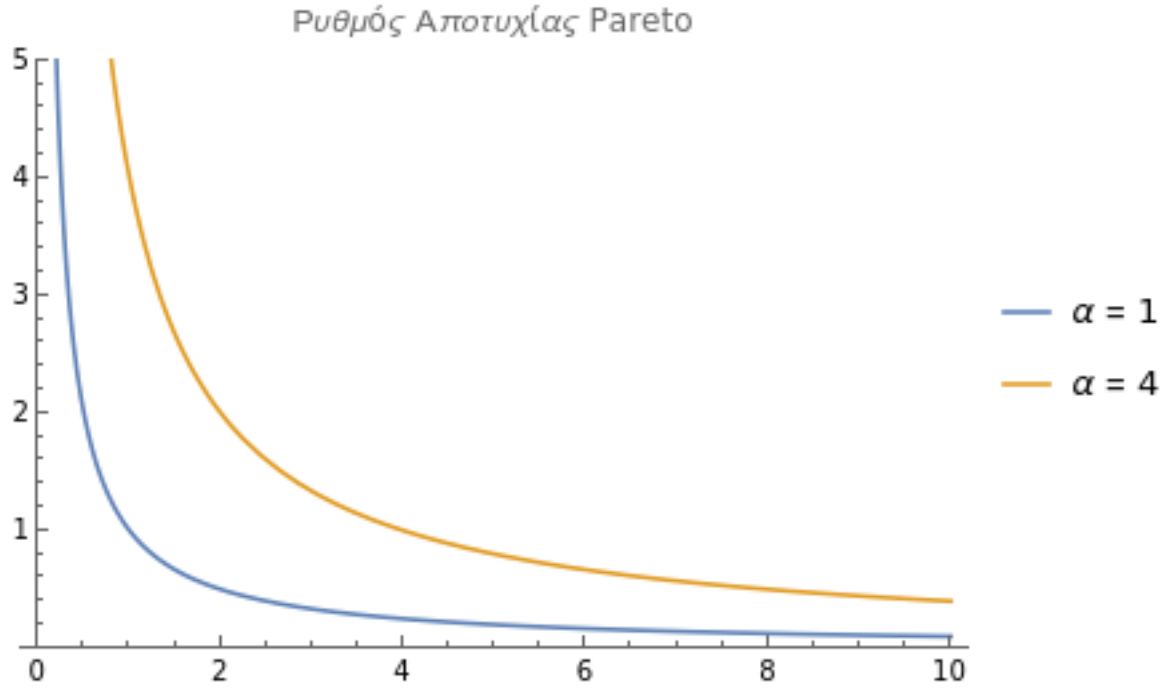
Αντικαθιστώντας τους τύπους για την κατανομή Pareto έχουμε

$$\lambda(x) = \frac{\alpha}{x}$$

Στο παρακάτω Σχήμα 2-6 δείχνουμε την επίδραση των παραμέτρων στον ρυθμό αποτυχίας της κατανομής Pareto.

ΣΧΗΜΑ 2-6

Διάγραμμα συνάρτησης λ(x) κατανομής Pareto για τιμές παραμέτρων α=1 και α=4.



Η συνάρτηση μέσης υπολειπόμενης ζωής (mean residual life function) μιας κατανομής δίνεται από τον τύπο

$$m(t) = \frac{1}{S(t)} \int_t^{+\infty} S(x)\, dx$$

όπου S(x) είναι η συνάρτηση επιβίωσης.

Αντικαθιστώντας τους τύπους της κατανομής Pareto η συνάρτηση μέσης υπολειπόμενης ζωής είναι

$$m(t) = \frac{1}{1 - \left(\frac{\theta}{t+\theta}\right)^{\alpha}} \int_{t}^{+\infty} \left(1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha}\right) dx$$

Το ολοκλήρωμα $\int_{t}^{+\infty} \left(1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha}\right) dx$ αποκλίνει, επομένως η συνάρτηση m(t) δεν είναι πεπερασμένη.

Κ Ε Φ Α Λ Α Ι Ο 3

Μέθοδοι εκτίμησης παραμέτρων

Στο κεφάλαιο αυτό, οι πληροφορίες αντλήθηκαν από την ακόλουθη βιβλιογραφία και επιστημονικές πηγές: Παπαδογιάννης και Γαλετάκης (2008), Τσάτσος, Παναγιωτόπουλος και Κατσιφαράκης (2009), Chatfield (2013) και Zar (2010).

3.1 Παραμετροποίηση

Όπως έχουμε πει, η κατανομή Pareto χαρακτηρίζεται από δύο παραμέτρους: την παράμετρο **μορφής ή δείκτη κλίσης** (shape parameter) και την **ελάχιστη τιμή** (minimum value parameter).

Ο δείκτης κλίσης, είναι η παράμετρος α, και αναπαριστά την κλίση της κατανομής. Όταν η τιμή του δείκτη κλίσης είναι μικρότερη του 1, η κατανομή δεν έχει ορισμένη μέση τιμή και διακύμανση. Η τιμή του δείκτη κλίσης επηρεάζει το σχήμα της κατανομής, καθώς καθορίζει το πόσο απότομα αυξάνεται ή μειώνεται η πιθανότητα των ακραίων τιμών ενώ η ελάχιστη τιμή έχει τη θέση παραμέτρου κλίμακας.

Γενικά, η εκτίμηση των παραμέτρων της κατανομής Pareto είναι ένα σημαντικό ζήτημα στη στατιστική και στην οικονομετρία, καθώς επιτρέπει τη μοντελοποίηση των καταναλωτικών και οικονομικών δεδομένων και την πρόβλεψη των μελλοντικών τους στοιχείων.

Στο κεφάλαιο της παραμετροποίησης και εκτίμησης παραμέτρων της κατανομής Pareto, εξετάζονται τρόποι για την εύρεση των παραμέτρων α και σ που περιγράφουν τη συγκεκριμένη κατανομή στο πλαίσιο μιας συγκεκριμένης εφαρμογής.

Η πρώτη διαδικασία είναι η εκτίμηση των παραμέτρων από δεδομένα παρατηρήσεων. Συγκεκριμένα, η μέθοδος των ελαχίστων τετραγώνων μπορεί να χρησιμοποιηθεί για να εκτιμήσει την παράμετρο α από ένα σύνολο δεδομένων. Επίσης, η παράμετρος σ μπορεί να εκτιμηθεί από το ελάχιστο δείγμα παρατήρησης.

Στη συνέχεια, εξετάζονται μέθοδοι παραμετροποίησης της κατανομής Pareto για να περιγράψουν συγκεκριμένα προβλήματα. Για παράδειγμα, στο πρόβλημα του υπερπληθυσμού της πόλης, η κατανομή Pareto μπορεί να παραμετροποιηθεί έτσι ώστε να περιγράφει το μέγεθος των πόλεων. Τέλος, αναλύονται οι διάφορες μέθοδοι για την εκτίμηση των παραμέτρων της κατανομής Pareto, οι οποίες είναι αυτή της μέγιστης πιθανοφάνειας, των ροπών, των ποσοστιαίων σημείων και των ελαχίστων τετραγώνων.

3.2 Μέθοδος Μέγιστης Πιθανοφάνειας

Η **Μέθοδος Μέγιστης Πιθανοφάνειας** (Maximum Likelihood Estimation - MLE): Αυτή η μέθοδος επιλέγει τις τιμές των παραμέτρων που μεγιστοποιούν τη συνάρτηση πιθανοφάνειας, δηλαδή την πιθανότητα να προκύψουν τα δεδομένα που παρατηρούμε, με βάση μια συγκεκριμένη τιμή των παραμέτρων. Η MLE είναι η πιο δημοφιλής μέθοδος εκτίμησης παραμέτρων στη στατιστική. Στην περίπτωση της κατανομής Pareto, η MLE χρησιμοποιείται για την εκτίμηση των δύο παραμέτρων της κατανομής, του συντελεστή κλίσης α και της ελάχιστης τιμής σ.

Αν καταφέρουμε να βρούμε τις παραμέτρους που μεγιστοποιούν την πιθανοφάνεια, τότε έχουμε εκτιμήσει τις παραμέτρους της κατανομής Pareto από τα δεδομένα.

Για την εκτίμηση των παραμέτρων της κατανομής Pareto με τη χρήση της MLE, η πιθανοφάνεια είναι η πιθανότητα να παρατηρηθούν οι συγκεκριμένες τιμές των δεδομένων μας στην κατανομή Pareto με δεδομένες τις παραμέτρους α και σ. Η πιθανοφάνεια αυτή είναι μια συνάρτηση των παραμέτρων α και σ που δίνεται από την εξίσωση:

$$L(\alpha, \sigma|x) = \prod_{i=1}^{n} f(x_i; \alpha, \sigma)$$

όπου n είναι ο αριθμός των δεδομένων μας, x_i είναι η i-οστή παρατήρηση και $f(x_i|\alpha, \sigma)$ είναι η συνάρτηση πυκνότητας πιθανότητας της κατανομής Pareto με παραμέτρους α και σ, επομένως στόχος μας είναι να βρούμε τις τιμές αυτών των παραμέτρων που μεγιστοποιούν την πιθανοφάνεια $L(\alpha, \sigma|x)$.

Το επόμενο θεώρημα δίνει τις βασικές ιδιότητες του ΕΜΠ όταν το n λαμβάνει πολύ μεγάλες τιμές.

ΘΕΩΡΗΜΑ 3.1 Έστω ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$ από μια κατανομή που εξαρτάται από μια $s-$διάστατη παράμετρο $\theta = (\theta_1, \theta_2, \dots, \theta_s)$. Κάτω από τις ανάλογες συνθήκες για την πολυδιάστατη περίπτωση ισχύουν τα εξής:

(α) Με πιθανότητα που τείνει στο 1, υπάρχει μια λύση $\hat{\theta}_n$ του συστήματος εξισώσεων πιθανοφάνειας:

$$\frac{\partial \log \ell(\theta|X)}{\partial \theta_1} = 0$$

$$\vdots$$

$$\frac{\partial \log \ell(\theta|X)}{\partial \theta_s} = 0$$

η οποία αποτελεί μια συνεπή εκτίμηση της παραμέτρου σ.

β) Η $\sqrt{n}(\hat{\theta}_n - \theta)$ συγκλίνει κατά κατανομή, στην $s-$διάστατη κανονική κατανομή με μέση τιμή $0 = (0,0, \dots, 0)^T$, και πίνακα συνδιακύμανσης $I^{-1}(\theta)$.

Ο πίνακας $I(\theta)$ ονομάζεται πίνακας πληροφορίας του Fisher και ορίζεται ως εξής:

$$I(\theta) = \left[E_\theta \left(\frac{\partial \ln f(X, \theta)}{\partial \theta_i} \cdot \frac{\partial \ln f(X, \theta)}{\partial \theta_j} \right) \right]_{sxs}$$

όπου $[.]_{sxs}$ συμβολίζει έναν πίνακα διαστάσεων $s \times s$. Κατά παρόμοιο τρόπο με την μονοδιάστατη περίπτωση, αν η παράγωγος ως προς θ_i, $1 \le i \le k$, μπορεί να εναλλάσσεται με το ολοκλήρωμα ως προς x τότε ισχύει

$$I(\theta) = \left[E_\sigma \left(\frac{\partial \ln f(X, \theta)}{\partial \theta_i} \cdot \frac{\partial \ln f(X, \theta)}{\partial \theta_j} \right) \right] = - \left[E_\sigma \left(\frac{\partial^2 \ln f(X, \theta)}{\partial \theta_i \, \partial \theta_j} \right) \right].$$

Άλλοι τρόποι εύρεσης της μέγιστης πιθανοφάνειας επιτυγχάνονται μέσω της εφαρμογής αλγορίθμων βελτιστοποίησης, όπως ο αλγόριθμος Newton-Raphson ή ο αλγόριθμος BFGS, που αναζητούν τις τιμές των παραμέτρων που μεγιστοποιούν την πιθανοφάνεια.

Συγκεκριμένα, ο αλγόριθμος Newton-Raphson χρησιμοποιεί τις πρώτες και δεύτερης

τάξης παραγώγους της πιθανοφάνειας για να βρει τις τιμές των παραμέτρων που μεγιστοποιούν την πιθανοφάνεια. Στη συνέχεια, εκτελείται μια επαναληπτική διαδικασία ώστε να βρεθεί η βέλτιστη τιμή των παραμέτρων. Ο αλγόριθμος BFGS είναι μια γενικότερη μέθοδος που χρησιμοποιεί την επίλυση προβλημάτων βελτιστοποίησης με περιορισμούς και χωρίς περιορισμούς.

Ο αλγόριθμος BFGS χρησιμοποιεί μια σειρά από επαναλήψεις για να ενημερώσει την εκτίμηση των παραμέτρων σε κάθε βήμα. Κατά τη διάρκεια κάθε επανάληψης, ο αλγόριθμος χρησιμοποιεί την πληροφορία από τα προηγούμενα βήματα για να υπολογίσει μια νέα εκτίμηση των παραμέτρων.

Γενικά, ο αλγόριθμος BFGS είναι πιο αποδοτικός από τον αλγόριθμο Newton-Raphson, καθώς απαιτεί λιγότερους υπολογιστικούς πόρους και μπορεί να χρησιμοποιηθεί για πιο γενικά προβλήματα βελτιστοποίησης. Ωστόσο, ο αλγόριθμος Newton-Raphson είναι ακόμα ένας χρήσιμος αλγόριθμος για την εκτίμηση των παραμέτρων της κατανομής Pareto, καθώς παρέχει μια ακριβή λύση στο πρόβλημα.

Είναι σημαντικό να επισημανθεί ότι η συνάρτηση πιθανότητας αποτελεί ένα μέτρο που μας δίνει πληροφορίες για την πιθανότητα να παρατηρήσουμε τα συγκεκριμένα δεδομένα που έχουν ήδη παρατηρηθεί. Μέσω της μέγιστης συνάρτησης πιθανότητας, προσπαθούμε να εκτιμήσουμε τις τιμές των παραμέτρων α και σ που μεγιστοποιούν την πιθανότητα εμφάνισης των δεδομένων που διαθέτουμε. Αξίζει να σημειωθεί ότι η τιμή του σ δεν μπορεί να υπερβαίνει την μικρότερη τιμή των δεδομένων x. Με αυτόν τον τρόπο, μπορούμε να μεγιστοποιήσουμε την πιθανότητα εντοπίζοντας τον λογάριθμο της προαναφερθείσας πιθανότητας. Έχουμε

$$L(\alpha, \sigma|x) = \prod_{i=1}^{n} \frac{\alpha\sigma^{\alpha}}{x_i^{\alpha+1}} \quad \text{με} \quad \alpha > 0 \text{ και } 0 < \sigma \leq \min\{x_i\}$$

Αρχικά επειδή η συνάρτηση L είναι αύξουσα ως προς σ και η μεγαλύτερη τιμή που μπορεί να λάβει το σ είναι $\min\{x_i\}$, συνάγουμε ότι $\hat{\sigma} = \min\{x_i\}$.

Λογαριθμίζοντας την συνάρτηση L παίρνουμε

$$\ln L(\alpha,\sigma|x) = \ln\prod_{i=1}^{n}\frac{\alpha\sigma^{\alpha}}{x_i{}^{\alpha+1}} = \sum_{i=1}^{n}\ln\left(\frac{\alpha\sigma^{\alpha}}{x_i{}^{\alpha+1}}\right) = n\ln\alpha + n\alpha\ln\sigma - (\alpha+1)\sum_{i=1}^{n}\ln x_i$$

Θεωρώντας γνωστό το σ η παραπάνω σχέση γίνεται

$$\frac{d\ln L}{d\alpha} = \frac{n}{\alpha} + n\ln\hat{\sigma} - \sum_{i=1}^{n}\ln x_i$$

Ακολούθως μηδενίζουμε την παραπάνω σχέση, λύνουμε ως προς α και παίρνουμε

$$\hat{\alpha}_{MLE} = \frac{n}{\sum_{i=1}^{n}\ln(x_i/\hat{\sigma})}$$

3.3 Μέθοδος των Ροπών

Ο ΕΜΠ παρουσιάζει αξιόλογα χαρακτηριστικά, αλλά η χρήση του δεν είναι πάντα εφικτή. Ακόμα και όταν είναι εφικτή, υφίσταται η δυσκολία της πολυπλοκότητας επίλυσης του συστήματος εξισώσεων. Για αυτόν τον λόγο, αναζητούμε διάφορες εναλλακτικές μεθόδους εκτίμησης.

Μια διαδεδομένη μέθοδος είναι η μέθοδος των ροπών.

Η βασική ιδέα είναι να βρούμε τις τιμές της παραμέτρου σ, για τις οποίες οι θεωρητικές ροπές συμπίπτουν με τις δειγματικές ροπές. Έτσι, αν η διάσταση του σ είναι s, εξισώνουμε τις πρώτες s θεωρητικές ροπές με τις αντίστοιχες δειγματικές ροπές και λύνουμε το σύστημα των s εξισώσεων ως προς σ.

Έστω $\mu_k(\sigma) = E_\sigma[X^\kappa]$ και $\overline{X^\kappa} = \frac{\sum_{i=1}^{n} x_i^\kappa}{n}$ η $k-$τάξεως θεωρητική και δειγματική ροπή αντίστοιχα.

Αν η διάσταση του σ είναι 1 λύνουμε την εξίσωση

$$\mu_1(\sigma) = \bar{X}.$$

Αν η διάστασή του σ είναι s, τότε λύνουμε το σύστημα

$$\mu_1(\sigma) = \bar{X}$$
$$\vdots$$
$$\mu_s(\sigma) = \overline{X^s}\,.$$

Πρόταση 3.1 Έστω ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$ που ακολουθεί την κατανομή Pareto με παραμέτρους α και σ. Τότε η εκτίμηση των παραμέτρων με την μέθοδο των ροπών είναι:

$$\hat{\alpha} = \frac{n\bar{X} - X_m}{n(\bar{X} - X_m)} \text{ και } \hat{\sigma} = \frac{(\hat{\alpha}n - 1)X_m}{\hat{\alpha}n}$$

όπου $X_m = \min\{x_i\}$

Απόδειξη.

Έστω ότι α > 1.

Από την θεωρία ισχύουν οι σχέσεις $E(X) = \frac{\alpha\sigma}{(\alpha-1)}$ και $E(X^2) = \frac{\alpha}{(\alpha-2)}\sigma^2$ επομένως θα έχουμε $E(X) = \frac{\alpha\sigma}{(\alpha-1)} = \bar{X}$ από όπου προκύπτει:

$$\hat{\alpha} = \frac{\bar{X}}{\bar{X} - \hat{\sigma}}$$

Δεδομένου ότι $P(X > x) = 1 - F_X(x) = \left(\frac{\sigma}{x}\right)^{\alpha} \quad x > \sigma, \quad \alpha > 0, \quad \sigma > 0$ η πιθανότητα ταυτόχρονα όλα τα X_i να είναι μεγαλύτερα από την τιμή του x είναι

$$P(X_1, \dots, X_n > x) = \prod_{i=1}^{n} P(X_i > x) = \left(\frac{\sigma}{x}\right)^{\alpha n} \qquad (3.1)$$

Ωστόσο λόγο του ορισμού του X_m θα ισχύει επίσης $P(X_m > x) = \prod_{i=1}^{n} P(x_i > x)$ δηλαδή

$$P(X_m > x) = \left(\frac{\sigma}{x}\right)^{\alpha n} \text{ ή αλλιώς } P(X_m \le x) = 1 - \left(\frac{\sigma}{x}\right)^{\alpha n}$$

που σημαίνει ότι η συνάρτηση πυκνότητας πιθανότητας της σ θα είναι

$$h(x) = \frac{\alpha n \sigma^{\alpha n}}{x^{1+\alpha n}}$$

επομένως $E(X_m) = \int_{\sigma}^{+\infty} \frac{\alpha n \sigma^{\alpha n}}{x^{\alpha n}} = \frac{\alpha n \sigma}{-1+\alpha n}$

Για να βρούμε την εκτίμηση του σ αρκεί να λύσουμε την σχέση: $E(X_m) = X_m = \frac{\alpha n \sigma}{-1+\alpha n}$ η οποία οδηγεί στην $\hat{\sigma} = \frac{(\hat{\alpha} n - 1) X_m}{\hat{\alpha} n}$.

Εν τέλη από την σχέση (3.1) καταλήγουμε ότι $\hat{\alpha} = \frac{\bar{X}}{\bar{X} - \frac{(\hat{\alpha} n - 1) X_m}{\hat{\alpha} n}} = \frac{n\bar{X} - X_m}{n(\bar{X} - X_m)}$.

3.4 Μέθοδος των Ποσοστιαίων Σημείων

Η αξιολόγηση με τη μέθοδο των ροπών αναγνωρίζεται ως μια εναλλακτική προσέγγιση στην περίπτωση που ο ΕΜΠ δεν είναι διαθέσιμος ή δυσκολεύει την εύρεσή του. Αυτή η μέθοδος εφαρμόζεται όταν υπάρχουν κατάλληλες ροπές που μπορούν να χρησιμοποιηθούν. Ωστόσο, πρέπει να σημειωθεί ότι οι κατάλληλες ροπές δεν είναι πάντα διαθέσιμες, ιδίως όταν αντιμετωπίζουμε κατανομές με βαριά ουρά. Αυτή η κατάσταση δημιουργεί προβλήματα σε επιστημονικούς τομείς όπως ο αναλογισμός που βασίζονται σε τέτοιου είδους κατανομές.

Η εκτίμηση με τη μέθοδο των ποσοστιαίων σημείων (ΕΠΣ) αποτελεί μια εναλλακτική προσέγγιση στον ΕΜΡ. Η μεθοδολογία που ακολουθείται είναι η ίδια. Τα δειγματικά ποσοστιαία σημεία εξισώνονται με τα αντίστοιχα θεωρητικά σημεία και επιλύεται η αντίστοιχη εξίσωση ή σύστημα εξισώσεων. Συνήθως, όταν ο αριθμός των αγνώστων παραμέτρων είναι 1, εξισώνουμε το δειγματικό ποσοστό $\hat{\pi}_{0.5}$ με το αντίστοιχο θεωρητικό ποσοστό που συμβολίζεται με $\pi_{0.5}(\sigma)$. Άρα, η ΕΠΣ αντιπροσωπεύει τη λύση της αντίστοιχης εξίσωσης:

$$\pi_{0.5}(\sigma) = \hat{\pi}_{0.5}$$

Στην περίπτωση όπου η διάσταση είναι 2, λύνουμε το σύστημα εξισώσεων:

$$\pi_{0.25}(\sigma) = \hat{\pi}_{0.25}$$
$$\pi_{0.75}(\sigma) = \hat{\pi}_{0.75}$$

Σε περιπτώσεις όπου η κατανομή f είναι συμμετρική γύρω από το σ, δηλαδή f(x;σ) = f(x-σ), εξισώνουμε τη δειγματική διάμεσο $\hat{\pi}_{0.5}$, που αντιστοιχεί στο ΕΠΣ, με τη θεωρητική $\pi_{0.5}(\sigma)$.

Για παράδειγμα, στην περίπτωση της κανονικής κατανομής, η $\hat{\sigma}$ είναι ο δειγματικός μέσος (Χ), ενώ ο ΕΜΠ και ο ΕΜΡ αντιστοιχούν στον δειγματικό μέσο (Μ).

Σε περίπτωση δείγματος από την κατανομή Cauchy, δεν είναι δυνατό να υπολογιστεί ο ΕΜΡ, καθώς δεν υπάρχει μέση τιμή. Έτσι ο ΕΜΠ της σ αντιστοιχεί στο δειγματικό μέσο. Επίσης, ο ΕΠΣ είναι ισοδύναμος με τη διάμεσο, καθώς $\pi_{0.5} = \theta$.

Πρόταση 3.2 Θεωρούμε τυχαία μεταβλητή Χ η οποία ακολουθεί την κατανομή Pareto με παραμέτρους α και σ. Έστω $P_1=P(X_1 \leq x_1)$ και $P_2=P(X_2 \leq x_2)$ για κάποιες τιμές x_1, x_2 αντίστοιχα. Τότε χρησιμοποιώντας την μέθοδο των ποσοστιαίων σημείων οι εκτιμητές των παραμέτρων είναι:

$$\hat{\alpha}_q = \frac{\log\frac{1-P_1}{1-P_2}}{\log\frac{x_2}{x_1}} \quad \text{και} \quad \hat{\theta}_q = x_1(1-P_1)^{1/\hat{\alpha}_q}$$

Απόδειξη.

Έχουμε $P_1 = 1 - \left(\frac{\theta}{X_1}\right)^{\alpha}$ επομένως $\log(1-P_1) = \alpha \log\frac{\theta}{x_1}$ (3.2)

Ομοίως $P_2 = 1 - \left(\frac{\theta}{X_2}\right)^{\alpha}$ επομένως $\log(1-P_2) = \alpha \log\frac{\theta}{x_2}$ (3.3)

Από τις σχέσεις (3.2) και (3.3) προκύπτει ότι $\log\frac{(1-P_1)}{(1-P_2)} = \alpha \log\frac{X_2}{X_1}$ επομένως $\hat{\alpha}_q = \frac{\log\frac{1-P_1}{1-P_2}}{\log\frac{x_2}{x_1}}$.

Με αντικατάσταση του $\widehat{\alpha_q}$ στην σχέση (1) παίρνουμε το $\widehat{\theta_q}$, συγκεκριμένα θα έχουμε

$$1 - P_1 = \left(\frac{\theta}{X_1}\right)^{\alpha} \Rightarrow \hat{\theta}_q = x_1(1-P_1)^{1/\hat{\alpha}_q}$$

3.5 Μέθοδος των Ελαχίστων Τετραγώνων

Η **Μέθοδος των Ελαχίστων Τετραγώνων** (Least Squares Estimation - LSE): Η μέθοδος των ελαχίστων τετραγώνων είναι μια στατιστική τεχνική που χρησιμοποιείται για την εκτίμηση των παραμέτρων ενός μαθηματικού μοντέλου από παρατηρήσεις δεδομένων. Σκοπός της μεθόδου είναι να βρεθούν οι τιμές των παραμέτρων που ελαχιστοποιούν την απόκλιση μεταξύ των πραγματικών παρατηρήσεων και των προβλέψεων που προκύπτουν από το μαθηματικό μοντέλο.

Η μέθοδος εφαρμόζεται συνήθως σε προβλήματα παλινδρόμησης, όπου θέλουμε να προβλέψουμε μια συνεχή μεταβλητή από μια ή περισσότερες ανεξάρτητες μεταβλητές. Ας

υποθέσουμε ότι έχουμε ένα μοντέλο που περιγράφεται από μια εξίσωση της μορφής:

$$Y = \sum_{i=1}^{n} \beta_i X_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

όπου Y είναι η εξαρτημένη μεταβλητή που θέλουμε να προβλέψουμε, $X_1, X_2, \ldots, X_n$ είναι οι ανεξάρτητες μεταβλητές, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ είναι οι παράμετροι που θέλουμε να εκτιμήσουμε, και ε είναι οι τυχαίοι όροι (σφάλματα) που περιγράφουν τη διακύμανση που δεν εξηγείται από τις ανεξάρτητες μεταβλητές.

Στόχος μας είναι να εκτιμήσουμε τις τιμές των παραμέτρων β_0, β_1, β_2, ..., β_n έτσι ώστε η διαφορά ανάμεσα στις πραγματικές παρατηρήσεις και τις προβλέψεις που προκύπτουν από το μοντέλο να ελαχιστοποιείται.

Η μέθοδος λαμβάνει το όνομά της από το γεγονός ότι ελαχιστοποιεί το άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των πραγματικών παρατηρήσεων και των προβλέψεων. Δηλαδή, ελαχιστοποιείται η έκφραση:

$$S = \Sigma(y_i - \hat{y}_i)^2$$

όπου y_i είναι η i-οστή πραγματική παρατήρηση, $\hat{y}_i$ είναι η i-οστή πρόβλεψη που προκύπτει από το μοντέλο, και Σ συμβολίζει το άθροισμα για όλες τις παρατηρήσεις.

Για να ελαχιστοποιηθεί αυτή η έκφραση, χρησιμοποιούμε τη μέθοδο των μερικών παραγώγων. Υπολογίζουμε τις μερικές παραγώγους της έκφρασης S με βάση τις παραμέτρους β_0, β_1, β_2, ..., β_n και θέτουμε τις παραγώγους ίσες με μηδέν για να βρούμε τις τιμές των παραμέτρων που ελαχιστοποιούν την έκφραση S. Αυτό οδηγεί στην επίλυση ενός συστήματος εξισώσεων, γνωστού ως "συνθήκη ελαχίστων τετραγώνων". Οι λύσεις αυτού του συστήματος δίνουν τις εκτιμήσεις για τις παραμέτρους του μοντέλου.

Μετά τον υπολογισμό των τιμών των παραμέτρων, μπορούμε να εκτιμήσουμε την εξαρτημένη μεταβλητή Y για νέες τιμές των ανεξάρτητων μεταβλητών X_1, X_2, ..., X_n χρησιμοποιώντας το μοντέλο

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Η μέθοδος των ελαχίστων τετραγώνων είναι μια ευρέως αποδεκτή μέθοδος για την εκτίμηση παραμέτρων, καθώς παρέχει τις βέλτιστες εκτιμήσεις υπό την υπόθεση ότι τα σφάλματα είναι κανονικά κατανεμημένα και ανεξάρτητα μεταξύ τους.

Είναι σημαντικό να σημειωθεί ότι η μέθοδος των ελαχίστων τετραγώνων προσπαθεί να ελαχιστοποιήσει την άθροιση των τετραγώνων των αποκλίσεων, αλλά δεν εξετάζει την αιτιολογία των αποκλίσεων ή την ανάλυση της διάσπασης της διακύμανσης. Επιπλέον, πρέπει να λάβουμε υπόψη την ερμηνεία των αποτελεσμάτων και τις προϋποθέσεις της μεθόδου για τη σωστή εκτίμηση των παραμέτρων. Είναι μια ισχυρή και ευέλικτη τεχνική για την εκτίμηση παραμέτρων σε προβλήματα παλινδρόμησης, παρέχοντας μια βέλτιστη λύση που ελαχιστοποιεί τις αποκλίσεις μεταξύ των παρατηρήσεων και των προβλέψεων.

Πρόταση 3.3 Έστω ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$ που ακολουθούν την κατανομή Pareto με παραμέτρους α και σ. Τότε χρησιμοποιώντας την μέθοδο των ελαχίστων τετραγώνων οι εκτιμητές των παραμέτρων θα είναι:

$$\hat{\alpha} = -\frac{\sum_{i=1}^{n}\left(\log x_i - \frac{\log\prod_{i=1}^{n} x_i}{n}\right)\left(\log\left(1-F(x_i)\right) - \frac{\log\prod_{i=1}^{n}(1-F(x_i))}{n}\right)}{\sum_{i=1}^{n}\left(\log x_i - \frac{\log\prod_{i=1}^{n} x_i}{n}\right)^2}$$

και

$$\hat{\sigma} = \exp\left\{\frac{\log\prod_{i=1}^{n}(1-F(x_i))}{n\hat{\alpha}} + \frac{\log\prod_{i=1}^{n} x_i}{n}\right\}$$

Απόδειξη. Αρχικά ισχύει ότι $\left(\frac{\sigma}{x_i}\right)^{\alpha} = 1 - F(x_i)$ επομένως $\alpha\log\sigma - \alpha\log x_i = \log(1-F(x_i))$. Στη συνέχεια θέτουμε $Y_i = \log(1-F(x_i))$, $Z_i = \log x_i$, $\beta = \alpha\log\sigma$, $\gamma = -\alpha$ και πέρνουμε το γραμμικό μοντέλο:

$$Y_i = \beta + \gamma Z_i$$

από το οποίο προκύπτει

$$\varepsilon_i = Y_i - \beta - \gamma Z_i \Rightarrow$$

$$\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(Y_i - \beta - \gamma Z_i)^2$$

και συνεπώς αρκεί να ελαχιστοποιήσουμε αυτή την σχέση ως προς τους συντελεστές β και γ.

Έχουμε

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^{\,2}}{\partial \beta} = -2 \sum_{i=1}^{n} (Y_i - \beta - \gamma Z_i)$$

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^{\,2}}{\partial \gamma} = -2 \sum_{i=1}^{n} (Y_i - \beta - \gamma Z_i) Z_i$$

Θέτοντας τις παραπάνω παραγώγους ίσες με μηδέν τελικά έχουμε:

$$\hat{\beta} = \bar{Y} - \hat{\gamma}\bar{Z}$$

και

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} (\bar{Z} - Z_i)(\bar{Y} - Y_i)}{\sum_{i=1}^{n} (\bar{Z} - Z_i)^2}$$

αντίστοιχα, όπου

$$\bar{Z} = \frac{\log \prod_{i=1}^{n} x_i}{n}$$

$$\bar{Y} = \frac{\log \prod_{i=1}^{n} (1 - F(x_i))}{n}$$

και συνεπώς επειδή

$$\hat{\alpha} = -\gamma$$

και

$$\hat{\sigma} = \exp\{(\bar{Y} + \hat{\alpha}\bar{Z})/\hat{\alpha}\}$$

προκύπτουν οι ζητούμενοι εκτιμητές.

Κ Ε Φ Α Λ Α Ι Ο 4

Μέτρα καλής προσαρμογής

Στο κεφάλαιο αυτό, οι πληροφορίες αντλήθηκαν από την ακόλουθη βιβλιογραφία και επιστημονικές πηγές: Βλαχογιάννης, Μπούγιας και Φωτίου (2016), Brazauskas και Serfling (2003), Klugman, Panjer και Willmot (2018) και Stephens (2001).

4.1 Εισαγωγή

Τα μέτρα καλής προσαρμογής στη στατιστική χρησιμοποιούνται για να αξιολογήσουν πόσο καλά μια κατανομή προσαρμόζεται σε ένα σύνολο δεδομένων. Αυτά τα μέτρα μας δίνουν μια ποσοτική εκτίμηση του βαθμού προσαρμογής και μας επιτρέπουν να συγκρίνουμε διάφορες κατανομές μεταξύ τους.

Θεωρούμε ένα τυχαίο δείγμα $X_1,\ldots,X_n$ για το οποίο ισχύει η διάταξη $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ και την εμπειρική αθροιστική συνάρτηση κατανομής:

$$F_X(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq x\}, \quad -\infty < x < \infty.$$

Επίσης, για έναν εκτιμητή $\hat{\alpha}$ έστω $\hat{F}(X_{(j)})$ η πιθανότητα που αντιστοιχεί στο $X_{(j)}$ από το μοντέλο $P(\sigma, \hat{\alpha})$, για $j = 1,\ldots,n$. Σημειώνουμε ότι $F_n(X_{(j)}) = j/n$, για $j = 1,\ldots,n$. Οι στατιστικές καλής προσαρμογής ορίζονται στις επόμενες ενότητες.

4.2 KS στατιστική D_n

Η KS (Kolmogorov-Smirnov) στατιστική D_n είναι ένα μέτρο που χρησιμοποιείται για να αξιολογήσει την απόκλιση ενός δείγματος από μια θεωρητική κατανομή πιθανοτήτων. Αποτελεί

έναν μη-παραμετρικό τρόπο για να εξετάσουμε αν τα δεδομένα μας προέρχονται από μια συγκεκριμένη κατανομή πιθανοτήτων. Η στατιστική D_n υπολογίζεται από τη μέγιστη απόλυτη απόκλιση μεταξύ της εμπειρικής συνάρτησης κατανομής (ECDF) του δείγματος και της θεωρητικής συνάρτησης κατανομής (CDF) της προκαθορισμένης κατανομής. Απλώς είναι η μέγιστη απόλυτη διαφορά μεταξύ των δύο συναρτήσεων.

Ο τιμή τής D_n χρησιμοποιείται για να αξιολογήσει πόσο καλά το δείγμα συμφωνεί με την υποθετική θεωρητική κατανομή. Όσο μικρότερη είναι η τιμή του D_n, τόσο πιο κοντά είναι η ECDF στην CDF και τόσο καλύτερα το δείγμα προσαρμόζεται στην θεωρητική κατανομή. Αντίθετα, μια μεγαλύτερη τιμή του D_n υποδηλώνει ότι το δείγμα αποκλίνει περισσότερο από την θεωρητική κατανομή.

Για να χρησιμοποιήσουμε την KS στατιστική D_n, αρχικά προσδιορίζουμε την υποθετική κατανομή που θέλουμε να ελέγξουμε. Στη συνέχεια, υπολογίζουμε την ECDF του δείγματός μας και την CDF της υποθετικής κατανομής. Στην συνέχεια, υπολογίζουμε τη μέγιστη απόλυτη απόκλιση (D_n) μεταξύ των δύο συναρτήσεων.

Αφού υπολογιστεί η τιμή του D_n, μπορούμε να τη συγκρίνουμε με την κρίσιμη τιμή του κριτηρίου Kolmogorov-Smirnov από τους πίνακες κριτικών τιμών. Αν η τιμή του D_n είναι μικρότερη από την κρίσιμη τιμή, τότε απορρίπτουμε τη μηδενική υπόθεση και συμπεραίνουμε ότι το δείγμα προέρχεται από την υποθετική κατανομή που εξετάσαμε. Αυτό υποδηλώνει ότι δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ του δείγματος και της θεωρητικής κατανομής.

Αντίθετα, αν η τιμή του D_n είναι μεγαλύτερη από την κρίσιμη τιμή, τότε δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση και δεν έχουμε αποδείξει ότι το δείγμα προέρχεται από την υποθετική κατανομή. Αυτό υποδηλώνει ότι υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ του δείγματος και της θεωρητικής κατανομής.

Γενικά η KS στατιστική D_n είναι ένα μέτρο που χρησιμοποιείται για να εκτιμήσει την καλή προσαρμογή ενός δείγματος προς μια υποθετική κατανομή πιθανοτήτων. Αν η τιμή του D_n είναι μικρή, συμπεραίνουμε ότι το δείγμα προσαρμόζεται καλά στην υποθετική κατανομή. Αντίθετα, μια μεγαλύτερη τιμή του D_n υποδηλώνει απόκλιση του δείγματος από την υποθετική κατανομή.

Η KS στατιστική D_n είναι ευρέως χρησιμοποιούμενη για την αξιολόγηση της προσαρμογής των δεδομένων μας σε μια θεωρητική κατανομή, όπως η κανονική κατανομή ή η εκθετική κατανομή. Επιπλέον, μπορεί επίσης να χρησιμοποιηθεί για σύγκριση δύο δειγμάτων για την εκτίμηση αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ τους.

Η KS στατιστική D_n έχει τύπο

$$D_n^+ = \max_{1 \le j \le n} \left(\frac{j}{n} - \hat{F}(X_{(j)}) \right), \quad D_n^- = \max_{1 \le j \le n} \left(\hat{F}(X_{(j)}) - \frac{j-1}{n} \right)$$

$$D_n = \max(D_n^+, D_n^-).$$

4.3 CvM στατιστική W_n^2

Η CvM (Cramér-von Mises) στατιστική W_n^2 είναι ένα μέτρο καλής προσαρμογής που χρησιμοποιείται για να εκτιμήσει την απόκλιση ενός δείγματος από μια υποθετική κατανομή πιθανοτήτων. Η στατιστική CvM υπολογίζεται από την αθροιστική συνάρτηση κατανομής (CDF) της υποθετικής κατανομής και την CDF του δείγματος.

Η τιμή W_n^2 κινείται μεταξύ 0 και άπειρο. Όσο μικρότερη είναι η τιμή της CvM στατιστικής, τόσο πιο κοντά είναι το δείγμα στην υποθετική κατανομή. Αν η τιμή της στατιστικής είναι πολύ μικρή, συμπεραίνουμε ότι το δείγμα προσαρμόζεται καλά στην υποθετική κατανομή.

Για να πάρουμε απόφαση σχετικά με την απόρριψη ή αποδοχή της μηδενικής υπόθεσης, συγκρίνουμε την τιμή της CvM στατιστικής με την κρίσιμη τιμή από τον πίνακα κρίσιμων τιμών. Αν η τιμή της CvM στατιστικής είναι μικρότερη από την κρίσιμη τιμή, τότε απορρίπτουμε τη μηδενική υπόθεση και συμπεραίνουμε ότι το δείγμα δεν προέρχεται από την υποθετική κατανομή. Αντίθετα, αν η τιμή της CvM στατιστικής είναι μεγαλύτερη από την κρίσιμη τιμή, δεν έχουμε αρκετά στοιχεία για να απορρίψουμε τη μηδενική υπόθεση και δεν μπορούμε να συμπεράνουμε ότι το δείγμα δεν προέρχεται από την υποθετική κατανομή.

Η CvM στατιστική είναι ευρέως χρησιμοποιούμενη στην ανάλυση καλής προσαρμογής δεδομένων σε υποθετικές κατανομές, όπως η κανονική κατανομή. Επιπλέον, μπορεί επίσης να χρησιμοποιηθεί για τη σύγκριση δύο δειγμάτων και την εκτίμηση της διαφοράς μεταξύ τους.

● Σε σχέση με την κατανομή KS στατιστική, η CvM στατιστική δεν είναι γενική και δεν μπορεί να εφαρμοστεί σε οποιαδήποτε υποθετική κατανομή.
● Επίσης η KS στατιστική είναι πιο ευαίσθητη στην μηδενική υπόθεση, που αναφέρεται στην απόρριψη της υπόθεσης ότι το δείγμα προέρχεται από μια συγκεκριμένη υποθετική κατανομή.
● Η KS στατιστική λαμβάνει υπόψη τόσο το εύρος των αποκλίσεων όσο και τη συχνότητα των αποκλίσεων από την υποθετική κατανομή.

Αυτή η ευαισθησία στην μηδενική υπόθεση καθιστά την KS στατιστική πιο ισχυρή σε σχέση με τη CvM στατιστική για τον έλεγχο καλής προσαρμογής. Ωστόσο, είναι σημαντικό να σημειωθεί ότι και οι δύο στατιστικές (CvM και KS) έχουν τα πλεονεκτήματά τους και χρησιμοποιούνται ανάλογα με το πλαίσιο και τις απαιτήσεις της εκάστοτε ανάλυσης.

Η CvM στατιστική W_n^2 έχει τύπο

$$W_n^2 = \sum_{j=1}^{n} \left(\hat{F}(X_{(j)}) - \frac{2j-1}{2n} \right)^2 + \frac{1}{12n}$$

4.4 AD στατιστική A_n^2

Η στατιστική AD όπως και η CvM χρησιμοποιείται για τον έλεγχο καλής προσαρμογής ενός δείγματος σε μια υποθετική κατανομή πιθανοτήτων, όπως η κανονική κατανομή. Η διαφορά μεταξύ της AD (Anderson-Darling) στατιστικής και της CvM (Cramér-von Mises) στατιστικής είναι ότι η AD στατιστική υπολογίζεται με βάση την εναπόθεση (επικαλυπτόμενη περιοχή) μεταξύ της CDF (συνάρτηση κατανομής) του δείγματος και της θεωρητικής CDF της υποθετικής κατανομής. Αναλυτικά, η στατιστική υπολογίζεται ως ένας συνδυασμός των αποκλίσεων και των συχνοτήτων των αποκλίσεων από την υποθετική κατανομή.

Όσον αφορά την ευαισθησία στην μηδενική υπόθεση η AD στατιστική είναι πιο ευαίσθητη στην ανίχνευση απόκλισης από την υποθετική κατανομή στις ουρές της κατανομής. Αυτό σημαίνει ότι επιδεικνύει μεγαλύτερη ευαισθησία σε κατανομές με αραιές ουρές από ότι η CvM.

Η AD στατιστική A_n^2 έχει τύπο

$$A_n^2 = n - \frac{1}{n}\sum_{j=1}^{n}\{(2j-1)\log\hat{F}(X_{(j)}) + (2n+1-2j)\log(1-\hat{F}(X_{(j)}))\}$$

Όταν η παράμετρος α εκτιμάται από την $\hat{\alpha}_{ML}$, οι κρίσιμες τιμές και οι τύποι στα επίπεδα σημαντικότητας για τις στατιστικές Dn, W_n^2 και A_n^2 είναι διαθέσιμες στους D'Agostino και Stephens (1986).

Κ Ε Φ Α Λ Α Ι Ο 5

Ο ρόλος της κατανομής Pareto στις σύνθετες (composite) συναρτήσεις

Στο κεφάλαιο αυτό, οι πληροφορίες αντλήθηκαν από την ακόλουθη βιβλιογραφία και επιστημονικές πηγές: Cooray & Ananda (2005), Johnson, Kotz & Balakrishnan (1994), Kaas, Goovaerts, Denuit & Dhaene (2001), Klugman, Panjer & Willmot (2004) και Teodorescu, Vernic (2006).

5.1 Εισαγωγή

Ο ρόλος της κατανομής Pareto στις σύνθετες (composite) συναρτήσεις είναι ένα θέμα που αξίζει να διερευνηθεί. Συχνά, μια κατανομή δεν μπορεί να παράσχει μια ικανοποιητική περιγραφή ενός συνόλου δεδομένων σε ολόκληρο το εύρος τους. Αυτό σημαίνει ότι μια κατανομή μπορεί να είναι καλή στην περιγραφή των μικρών ή μεσαίων τιμών ενός συνόλου δεδομένων, αλλά αντιμετωπίζει δυσκολία στην αναπαράσταση των μεγάλων τιμών και αντίστροφα.

Αυτή η προβληματική κατάσταση είναι εξαιρετικά συχνή στην ασφαλιστική επιστήμη. Στον τομέα αυτό, οι αναλυτές αρκετές φορές εργάζονται με στατιστικές μεθόδους για να προβλέψουν τις μελλοντικές ασφαλιστικές απαιτήσεις ή αποζημιώσεις. Η κατανομή που χρησιμοποιείται μπορεί να περιγράψει αποτελεσματικά τα συνηθισμένα αιτήματα, αλλά μπορεί να αποτυγχάνει στην ακριβή πρόβλεψη των ακραίων αιτημάτων που είναι σπάνια, αλλά μπορεί να είναι ιδιαίτερα δαπανηρά.

Αν επικεντρωθούμε σε αυτό το πρόβλημα, μπορούμε να εξετάσουμε πιθανές προσεγγίσεις ή μεθόδους που μπορούν να βοηθήσουν στην αντιμετώπισή του, όπως η χρήση εναλλακτικών κατανομών ή μοντέλων. Επιπλέον, μπορούμε να εξετάσουμε πρακτικά παραδείγματα ή να προσφέρουμε προτάσεις για βελτιώσεις στις ασφαλιστικές μεθόδους που βασίζονται στην κατανομή Pareto.

Συνολικά, στην εργασία αυτή, θα εμβαθύνουμε την κατανόηση της προκλητικής φύσης της κατανομής Pareto και θα προτείνουμε πιθανές λύσεις για τη βελτίωση της ακρίβειας και αξιοπιστίας των προβλέψεων.

5.2 Σύνθετες (composite) συναρτήσεις

Κατά τη μελέτη των απαιτήσεων που επηρεάζουν ένα συγκεκριμένο ασφαλιστικό χαρτοφυλάκιο, μια συχνή κατάσταση είναι ότι υπάρχουν πολλές μικρές αξιώσεις, αλλά επίσης λίγες μεγάλες αξιώσεις που δημιουργούν μια βαριάς ουράς κατανομή. Τέτοιες καταστάσεις συναντώνται συχνά στην ασφάλεια περιουσιακών στοιχείων, αυτοκινήτων, κλπ.

Στη συνέχεια, μια κατανομή αξιώσεων μπορεί να μοντελοποιηθεί ως συνδυασμός δύο πυκνοτήτων, που αποτελούνται από μια λιγότερο βαριάς ουράς κατανομή μέχρι ένα συγκεκριμένο σημείο και από μια κατανομή βαριάς ουράς πέρα από αυτό το σημείο. Τέτοιου είδους κατανομές είναι οι λεγόμενες σύνθετες, όπως προτείνουν οι Cooray και Ananda (2005). Οι ίδιοι κατασκεύασαν ένα σύνθετο μοντέλο ως εξής:

$$f(x) = \begin{cases} cf_1(x) & 0 < x \leq \theta \\ cf_2(x) & \theta < x \leq \infty \end{cases} \tag{5.1}$$

όπου f_1 και f_2 είναι συναρτήσεις πυκνότητας πιθανότητας (pdf), ενώ το c είναι μια σταθερά.

Προκειμένου να έχουμε μια ομαλή πυκνότητα πιθανότητας (pdf), επιβάλλεται η ύπαρξη συνθηκών συνέχειας και διαφορισιμότητας στο σημείο θ, από όπου προκύπτει η σταθερά c, καθώς και μια συνθήκη που μειώνει κατά ένα τον αριθμό των άγνωστων παραμέτρων των f_1 και f_2. Όπως αναφέρθηκε προηγουμένως, συνήθως το f_1 θεωρείται μια κατανομή με μικρή ουρά, ενώ το f_2 μια κατανομή με βαριά ουρά.

Στη θέση της f_2 χρησιμοποιείται σχεδόν αποκλειστικά κάποια μορφή της Pareto. Ενδεικτικά αναφέρεται η εξής βιβλιογραφία:

● Ο Cooray K. (2009) and Calderin-Ojeda(2018) προτείνουν το Weibull-Pareto μοντέλο για την μελέτη μονοκόρυφων δεδομένων αποτυχίας.

● Οι Majid and Ibrahim (2021) μελετούν τα σύνθετα μοντέλα Pareto από την πλευρά της Μπευσιανής στατιστικής.

● Οι Benatmare et. Al (2020) εισάγουν το σύνθετο Rayleigh-Pareto μοντέλο με το οποίο μελετούν ασφαλιστικές αποζημιώσεις λόγω φωτιάς

● Οι Aminzadeh and Deng (2019) and Lui and Ananda(2022) μελετούν το Inverse Gamma-Pareto composite model και

● Οι Cooray & Cheng (2015) & Scollnik (2007) μελετούν το Lognormal-Pareto μοντέλο.

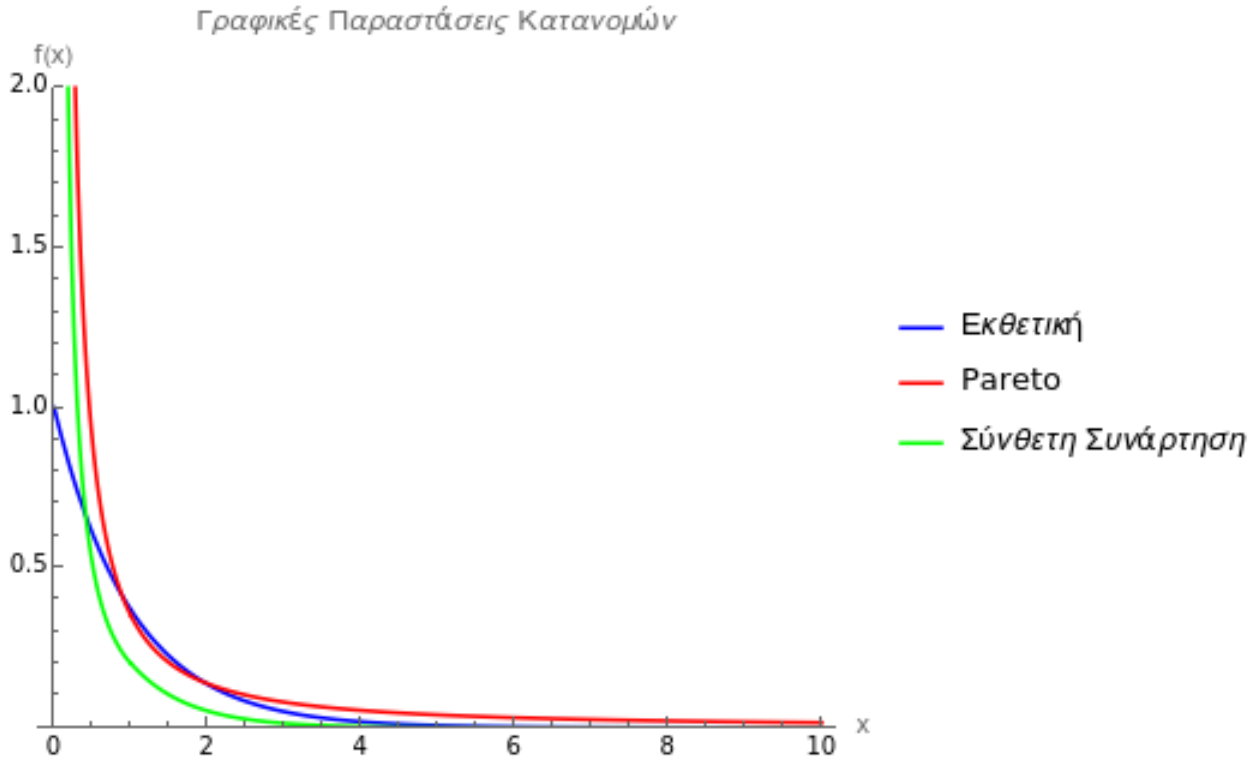
Εδώ θα περιγράψουμε το Εκθετικό - Pareto μοντέλο των Theodoresku and Vernic (2009). Συγκεκριμένα, επιλέγουμε την Εκθετική πυκνότητα για την f_1 και μια Pareto πυκνότητα για την f_2. Αυτή η επιλογή είναι εμπνευσμένη από το γεγονός ότι ανάμεσα στις κατανομές με βαριές ουρές, η κατανομή Pareto προτιμάται όταν πρόκειται να μοντελοποιηθούν μεγαλύτερες αξιώσεις ή πληρωμές επανασφάλισης, ενώ η Εκθετική κατανομή είναι εύκολη στον χειρισμό. Επίσης, η Εκθετική κατανομή εμφανίζεται συχνά σε ασφαλιστικά μοντέλα για αξιώσεις και, ως εκ τούτου, παίζει σημαντικό ρόλο σε μοντέλα που επιτρέπουν αναλυτικό υπολογισμό πιθανοτήτων καταστροφής (Kaas, Goovaerts, Denuit & Dhaene (2001)). Για λεπτομέρειες σχετικά με την Εκθετική και την Pareto κατανομή, δείτε το Johnson, Kotz & Balakrishnan (1994).

Ένα σύνθετο μοντέλο Εκθετικής – Pareto της μορφής (5.1) προτάθηκε ήδη από τους Teodorescu και Vernic (2006). Το κύριο χαρακτηριστικό της πυκνότητάς του είναι ότι, ακόμα κι αν η μορφή του μοιάζει με την Εκθετική, έχει μια μεγαλύτερη ουρά από την αντίστοιχη Εκθετική και μια ελαφρύτερη ουρά από την αντίστοιχη Pareto. Η σύνθετη Εκθετική - Pareto πυκνότητα είναι

$$f(x) = \begin{cases} \frac{0.775}{\theta} \exp\left\{-\frac{1.35x}{\theta}\right\} & 0 < x \leq \theta \\ \\ 0.2\frac{\theta^{0.35}}{x^{1.35}} & \theta < x \leq \infty \end{cases} \qquad (5.2)$$

ΣΧΗΜΑ 5-1

Διάγραμμα Εκθετικής, Pareto και σύνθετης κατανομής για τιμές παραμέτρων α=0.35 και θ=1.



Με απλή ολοκλήρωση βρίσκουμε ότι η hazard συνάρτηση της (5.2) είναι

$$h(x) = \frac{f_X(x)}{\bar{F}_X(x)} = \begin{cases} 0.422605 + 0.574074e^{-1.35x} & 0 < x < 1 \\ \dfrac{0.35}{x} & x \geq 1 \end{cases}$$

Παρατηρούμε ότι για $x \geq 1$ η hazard rate της σύνθετης συνάρτησης είναι η ίδια με την hazard rate της Pareto. Δηλαδή η hazard rate της σύνθετης συνάρτησης για μεγάλα x ταυτίζεται με την hazard rate της Pareto.

Κ Ε Φ Α Λ Α Ι Ο 6

Ευνοϊκοί εκτιμητές της κατανομής Pareto

Στο κεφάλαιο αυτό, οι πληροφορίες αντλήθηκαν από την ακόλουθη βιβλιογραφία και επιστημονικές πηγές: Βλαχογιάννης, Μπούγιας και Φωτίου (2016), Brazauskas και Serfling (2003), Klugman, Panjer και Willmot (2018) και Stephens (2001).

6.1 Εισαγωγή

Μεγάλο μέρος της βιβλιογραφίας αναφέρεται στην ανθεκτική και αποδοτική εκτίμηση παραμέτρων βαριάς ουράς για τα (ισοδύναμα) μοντέλα Pareto και εκθετικής κατανομής, για μεγάλα και μικρά δείγματα.

Πέρα από αυτούς τους εκτιμητές, έχουν εισαχθεί και νέοι ανθεκτικότεροι, όπως της "γενικευμένης διαμέσου" (GM) και του "κομμένου μέσου όρου" (T) οι οποίοι δείχνουν να παρέχουν εξίσου καλές ισορροπίες μεταξύ αποδοτικότητας και ανθεκτικότητας όπως αρκετούς ήδη γνωστούς εκτιμητές, συμπεριλαμβανομένων εκείνων που αντιστοιχούν στις μεθόδους μέγιστης πιθανοφάνειας, ποσοστιαίων σημείων κ.α.

Η απόδοση των προαναφερθέντων εκτιμητών σε πραγματικά δεδομένα μπορούν να καθιερωθούν μέσω της χρήσης μέτρων καλής προσαρμογής και έτσι οι ευνοϊκές θεωρητικές ιδιότητες των εκτιμητών του τύπου GM και T οδηγούν σε μια εξαιρετική πρακτική απόδοση. Οι εκτιμητές αυτοί κατατάσσονται και συγκρίνονται με βάση τα στατιστικά τεστ Kolmogorov-Smirnov, Cramér-von Mises και Anderson-Darling.

Μια κατανομή Pareto διαδραματίζει πολύ σημαντικό ρόλο στην αναλογιστική μοντελοποίηση λόγω της εννοιολογικής απλότητας και της ευκολίας εφαρμογής της στην πράξη. Η αθροιστική συνάρτηση κατανομής του μοντέλου Pareto P(α, σ) δίνεται από τον τύπο

$$F(x) = 1 - (\sigma/x)^{\alpha} \quad \text{για } x > \sigma \tag{6.1}$$

όπου $\alpha > 0$ είναι η παράμετρος σχήματος που χαρακτηρίζει την ουρά της κατανομής και $\sigma > 0$ είναι η παράμετρος κλίμακας.

Η υπόθεση της γνώσης του σ είναι αρκετά τυπική στην αναλογιστική βιβλιογραφία γιατί, όπως χαρακτηριστικά αναφέρει ο Philbrick (1985), «Αν και μπορεί να υπάρχουν περιπτώσεις όπου αυτή η τιμή πρέπει να εκτιμηθεί, ουσιαστικά όλες οι αιτήσεις ασφάλισης αυτής της αξίας θα επιλεγούν εκ των προτέρων.»

Σε αρκετές πρόσφατες εργασίες οι Brazauskas and Serfling ανέπτυξαν μια ισχυρή και αποτελεσματική εκτίμηση του δείκτη ουράς για διάφορες περιπτώσεις: για μεγάλα και μικρά δείγματα και για μοντέλα μιας και δύο παραμέτρων (με γνωστό σ ή άγνωστο). Για την ανάπτυξη της μεθοδολογίας τους ανέπτυξαν μια γνωστή σχέση ισοδυναμίας μεταξύ του μοντέλου (6.1) και της εκθετικής κατανομής E(μ,θ) η οποία έχει αθροιστική συνάρτηση κατανομής

$$G(z) = 1 - e^{-\frac{z-\mu}{\theta}}, \quad z > \mu \tag{6.2}$$

για $\theta > 0$ και $-\infty < \mu < \infty$. Συγκεκριμένα, εάν η τυχαία μεταβλητή X έχει αθροιστική συνάρτηση κατανομής F που δίνεται από τον τύπο (6.1) τότε η μεταβλητή $Z = \log X$ έχει αθροιστική συνάρτηση κατανομής G που δίνεται από το τύπο (6.2), με $\mu = \log \sigma$ και $\theta = \alpha^{-1}$

Σε μελέτες μεγάλου δείγματος, για παράδειγμα, εισήχθησαν νέοι ισχυροί εκτιμητές του τύπου GM και του τύπου T και προσαρμόστηκαν από τη βιβλιογραφία του μοντέλου E(μ,θ). Αυτοί οι εκτιμητές ήταν τότε συγκρίσιμοι με τη μέγιστη πιθανοφάνεια, την μέθοδο του ποσοστιαίου σημείου και άλλους εκτιμητές. Χρησιμοποιώντας ως κριτήριο απόδοσης την ασυμπτωτική σχετική αποτελεσματικότητα (ARE) σε σχέση με τον εκτιμητή MLE και ως κριτήριο ευρωστίας το σημείο απώλειας (BP), ο τύπος της «γενικευμένης διαμέσου» φάνηκε να κυριαρχεί σε όλους τους ανταγωνιστές, με τον τύπο T ως δεύτερο καλύτερο. Από πρακτική άποψη, το ARE είναι ισοδύναμο με την ακρίβεια του εκτιμητή και μπορεί να ερμηνευθεί ως προς το μήκος του διαστήματος εμπιστοσύνης.

Σε αυτό το κεφάλαιο αναλύουμε την απόδοση των προαναφερθέντων εκτιμητών και καθορίζουμε την καταλληλότητά τους μέσω της χρήσης κατάλληλων μετρικών. Αποδεικνύουμε ότι οι εκτιμητές των τύπων GM και T, λόγω των ευνοϊκών θεωρητικών τους ιδιοτήτων, επιδεικνύουν εξαιρετική πρακτική απόδοση.

Τα μέτρα καλής προσαρμογής, χρησιμοποιούνται εδώ για δύο σκοπούς: (i) για (τυπική) δοκιμή της καταλληλότητας του εκτιμώμενου μοντέλου Pareto για ένα συγκεκριμένο σύνολο δεδομένων όταν αυτό εκτιμάται από το MLE και (ii) για την αξιολόγηση και σύγκριση της ομοιότητας του μοντέλου Pareto όταν χρησιμοποιούνται διάφοροι εκτιμητές (όχι μόνο του MLE) της παραμέτρου α.

Στην αναλογιστική βιβλιογραφία το ζήτημα της καλής προσαρμογής αντιμετωπίζεται μέσω ενός συνδυασμού άτυπων μεθόδων και επίσημων στατιστικών δοκιμών. Οι πιο ανεπίσημες τεχνικές βασίζονται στη διαφορά (απόλυτη ή σχετική) μεταξύ των εφαρμοσμένων και εμπειρικών τιμών των σχετικών ποσοτήτων, όπως ο αριθμός των αξιώσεων ή αναμενόμενη τιμή αξιώσεων για διαφορετικά επίπεδα αξίωσης.

Όπως είναι γνωστό στη στατιστική βιβλιογραφία (π.χ., D'Agostino και Stephens (1986)), η x^2 δοκιμή είναι λιγότερο ισχυρή από τις δοκιμές που βασίζονται στην εμπειρική αθροιστική συνάρτηση κατανομής. Επομένως, εδώ χρησιμοποιούμε τρία ευρέως δημοφιλή μέτρα καλής προσαρμογής που βασίζονται στην εμπειρική αθροιστική συνάρτηση κατανομής — την προαναφερθείσα στατιστική KS, την στατιστική Cramérvon Mises (CvM) και την στατιστική Anderson-Darling (AD).

Όλα αυτά τα στατιστικά στοιχεία μετρούν την απόσταση κατά κάποιο τρόπο μεταξύ του προσαρμοσμένου μοντέλου $\hat{F}$ και του εμπειρικού F_n. Επομένως, είναι προτιμότεροι οι εκτιμητές που οδηγούν σε μικρότερες τιμές αυτών των στατιστικών.

6.2 Κριτήριο ευρωστίας: Σημείο απώλειας

Ένα δημοφιλές και αποτελεσματικό κριτήριο για την ανθεκτικότητα ενός εκτιμητή είναι το **σημείο απώλειας** (breakdown point - BP), το οποίο χαρακτηρίζεται ελλιπώς ως ο μεγαλύτερος

λόγος των στρεβλωμένων δειγματικών παρατηρήσεων που μπορεί να αντιμετωπίσει ο εκτιμητής. Δηλαδή, το σημείο απώλειας ενός εκτιμητή μετρά τον βαθμό ανθεκτικότητας του εκτιμητή στην επίδραση ακραίων παρατηρήσεων που μπορεί να αποτελέσουν δυσμενή επιρροή στα συνολικά δεδομένα, και όχι απαραίτητα ασυνήθιστες παρατηρήσεις που προέρχονται από το παραμετρικό μοντέλο που μελετάμε.

Οι Brazauskas και Serfling (2000) εξέτασαν δύο τύπους επιρροής - άνω και κάτω επιρροή - και κατά συνέπεια προσδιόρισαν ξεχωριστές εκδόσεις του σημείου απώλειας (breakdown point, BP):

Σημείο απώλειας Κάτω (άνω) Ορίου (Lower (upper) Breakdown Point, LBP/UBP) η μεγαλύτερη αναλογία των κάτω (άνω) παρατηρήσεων δείγματος που μπορεί να ληφθεί σε ένα κατώτατο (άνω) όριο, χωρίς να οδηγηθεί ο εκτιμητής σε ένα όριο που δεν εξαρτάται από την εκτιμώμενη παράμετρο.

Για τον μοντελισμό των απωλειών από ασφαλιστικά συμβόλαια, η επιρροή των χαμηλότερων τύπων είναι λιγότερο ανησυχητική, διότι το κατώτατο όριο των απωλειών ορίζεται συνήθως προκαταβολικά από το συμβόλαιο. (Για παράδειγμα, το χαμηλότερο όριο μπορεί να αναπαρασταθεί ως αυτοσυμμετοχή.) Επομένως, στην παρούσα αντιμετώπιση προτιμούμε εκτιμητές που έχουν μη μηδενική UBP (Μέγιστη Πιθανή Υποψία).

6.3 Κριτήριο αποτελεσματικότητας: Διακύμανση

Αν τα δείγματα παρατηρήσεων ακολουθούν το υποθετικό παραμετρικό μοντέλο, τότε είναι γνωστό ότι, για μεγάλα σύνολα δεδομένων, ο Εκτιμητής Μέγιστης Πιθανοφάνειας (MLE) επιτυγχάνει (στην προσεγγιστική του κανονική κατανομή) την ελάχιστη δυνατή διακύμανση ανάμεσα σε ένα μεγάλο σύνολο ανταγωνιστικών εκτιμητών. Επομένως, μπορεί να θεωρηθεί ως ένας ποσοτικός δείκτης αποτελεσματικότητας. Ειδικότερα, για το μοντέλο P(σ,θ) με το σ γνωστό, ο MLE του α υπολογίζεται εύκολα στο άρθρο Arnold (1983), και δίνεται από την εξίσωση:

$$\hat{\alpha}_{ML} = \frac{1}{\sum_{i=1}^{n} \log(X_i/\sigma)}$$

Μπορεί να αποδειχθεί (Lehmann and Casella 1998) ότι η $\frac{2n\alpha}{\hat{\alpha}_{ML}}$ έχει αθροιστική συνάρτηση κατανομής χ^2_{2n} όπου χ^2_v αναφέρεται στην χι-τετράγωνο κατανομή με n βαθμούς ελευθερίας. Αυτό συνεπάγεται ότι η εκτιμήτρια $\hat{\alpha}_{ML}$ είναι μια μεροληπτική εκτιμήτρια του α, αλλά ο πολλαπλασιασμός της με τον παράγοντα (n – 1)/n παρέχει έναν αμερόληπτο εκτιμητή.

$$\hat{\alpha}_{MLU} = \frac{n-1}{n}\hat{\alpha}_{ML} = \frac{n-1}{\sum_{i=1}^{n} \log(X_i/\sigma)}$$

Για περαιτέρω λεπτομέρειες σχετικά με τη θεωρία της ακριβούς κατανομής της εκτιμήτριας $\hat{\alpha}_{ML}$ υπάρχουν στο Rytgaard (1990). Ακολουθώντας τεχνικές από τους Brazauskas και Serfling (2000a,b), μπορεί να δειχθεί ότι για μεγάλο μέγεθος δείγματος n, η $\hat{\alpha}_{MLU}$ είναι περίπου κανονικά κατανεμημένη με μέση τιμή α και διακύμανση α^2/n. Επιπλέον, άλλες ανταγωνιστικές εκτιμήτριες $\hat{\alpha}$ για την παράμετρο α που εξετάζονται εδώ είναι περίπου κανονικά κατανεμημένες με μέση τιμή α και διακύμανση $c\alpha^2/n$, όπου $c>1$ είναι μια σταθερά και n μεγάλος αριθμός. Αυτό σημαίνει ότι τα διαστήματα εμπιστοσύνης για την παράμετρο α που βασίζονται στις ανταγωνιστικές εκτιμήτριες θα είναι c φορές πιο ευρύτερα από αυτές που βασίζονται στην MLU. Ωστόσο, αυτή η βέλτιστη ακρίβεια της MLU επιτυγχάνεται με το κόστος της ανθεκτικότητας, που γίνεται κρίσιμη όταν τα πραγματικά δεδομένα αποκλίνουν από το υποθετικό παραμετρικό μοντέλο. Συνεπώς, η MLU είναι η πιο αποδοτική εκτιμήτρια αλλά δεν είναι ανθεκτική, με UBP = 0.

Στη συνέχεια θα παρουσιάσουμε τις μεθόδους για την εκτίμηση των παραμέτρων. Συγκεκριμένα, παρουσιάζουμε τους εκτιμητές των τύπων του ποσοστού, του κομμένου μέσου όρου και της γενικευμένης διαμέσου. Για περαιτέρω λεπτομέρειες και συζήτηση, ο αναγνώστης μπορεί να ανατρέξει στο έργο των Brazauskas και Serfling (2000a,b).

6.4 Εκτιμητές ποσοστημορίων

Οι εκτιμητές ποσοστημορίων της παραμέτρου α δεν επηρεάζονται καθόλου από επιπλέον πληροφορίες σχετικά με το σ. Για αυτό το λόγο και για τη συμβατότητα με την υπάρχουσα βιβλιογραφία, περιγράφουμε αυτήν την προσέγγιση εδώ για την περίπτωση που το σ θεωρείται

ως άγνωστη παράμετρος.

Οι εκτιμητές ποσοστημορίων για $k \geq 2$ και πιθανότητες $0 < p_i < \ldots < p_k < 1$ ορίζονται ως εξής:

$$\hat{\alpha}_Q = \left(\sum_{i=1}^{n} b_i \log X_{(\lceil np_i \rceil)}\right)^{-1},$$

$$\hat{\sigma}_Q = \exp\left\{\log X_{(\lceil np_i \rceil)} - u_i/\hat{\alpha}_Q\right\},$$

με

$$b_1 = \frac{1}{L}\frac{u_2 - u_1}{e^{u_2} - e^{u_1}}$$

$$b_i = \frac{1}{L}\left[\frac{u_i - u_{i-1}}{e^{u_i} - e^{u_{i-1}}} - \frac{u_{i+1} - u_i}{e^{u_{i+1}} - e^{u_i}}\right]$$

όπου $2 \leq i \leq k - 1$

$$b_k = \frac{1}{L}\frac{u_k - u_{k-1}}{e^{u_k} - e^{u_{k-1}}}$$

και

$$L = \sum_{i=2}^{k} \frac{(u_i - u_{i-1})^2}{e^{u_i} - e^{u_{i-1}}}$$

όπου $u_i = -\log(1 - p_i)$, $1 \leq i \leq k$, και $\lceil X \rceil$ υποδηλώνει τον μικρότερο ακέραιο αριθμό που είναι μεγαλύτερος ή ίσος του x.

Τέτοιες εκτιμήσεις εισήχθησαν και μελετήθηκαν για το πρόβλημα Pareto από τον Quandt (1966) για $k = 2$ και από τον Koutrouveli (1981) για $k \geq 2$.

Επιλέγοντας το ελάχιστο της ορίζουσας του ασυμπτωτικού πίνακα συνδιακύμανσης των εκτιμητών των παραμέτρων σ και α ως κριτήριο βέλτιστης επιλογής, ο Koutrouvelis (1981) βρήκε ότι η βέλτιστη επιλογή του p_1 είναι πάντα:

$$p_1^0 = \frac{1}{n + 0{,}5}$$

και οι υπόλοιπες βέλτιστες πιθανότητες $p_1^0, \dots, p_k^0$

Για k=2 είναι $p_2^0 = 1 - (1 - p_1^0)e^{-1.5936} \approx 0.80$

Για k=5 είναι $p_2^0 = 1 - (1 - p_1^0)e^{-0.6003} \approx 0.45$, $p_3^0 = 1 - (1 - p_1^0)e^{-1.3544} \approx 0.74$, $p_4^0 = 1 - (1 - p_1^0)e^{-2.3721} \approx 0.91$ και $p_5^0 = 1 - (1 - p_1^0)e^{-3.9657} \approx 0.98$.

Συμβολίζουμε τους βέλτιστους εκτιμητές της θετικής παραμέτρου α με $\hat{\alpha}_Q^{opt,k}$. Εξετάζουμε επίσης μια μη βέλτιστη περίπτωση (συμβολίζεται με $\hat{\alpha}_Q^*$):

• Για k = 5, έχουμε τις τιμές $p_1 = 0.13$, $p_2 = 0.315$, $p_3 = 0.50$, $p_4 = .685$, και $p_5 = 0.87$.

Παρατήρηση. Όταν ο αριθμός k των ποσοστιαίων σημείων επιλεγεί ίσος με τον αριθμό των άγνωστων παραμέτρων του μοντέλου, η μέθοδος αντιστοιχεί σε αυτό που ονομάζεται *αντιστοίχιση ποσοστών* από τον Klugman, Panjer, και Willmot (1998).

6.5 Εκτιμητές κομμένου μέσου όρου

Για τις καθορισμένες τιμές των β_1 και β_2 που ικανοποιούν τη σχέση $0 \le \beta_1, \beta_2 < 1/2$, δημιουργείται ένας κομμένος μέσος όρος απορρίπτοντας το ποσοστό β_1 των παρατηρήσεων με τις χαμηλότερες τιμές και το ποσοστό β_2 των παρατηρήσεων με τις υψηλότερες τιμές, δημιουργώντας τον μέσο όρο των υπόλοιπων παρατηρήσεων. Ειδικότερα, για το α εισάγουμε τον εκτιμητή του κομμένου μέσου όρου

$$\hat{\alpha}_T = \left(\sum_{i=1}^{n} c_{ni} \log\left(X_{(i)}/\sigma\right)\right)^{-1},$$

με $c_{ni} = 0$ για $1 \le i \le [n\beta_1]$, $c_{ni} = 0$ για $n - [n\beta_2] + 1 \le i \le n$, και $c_{ni} = 1/d(\beta_1, \beta_2, n)$ για $[n\beta_1] + 1 \le i \le n - [n\beta_2]$, όπου $[\cdot]$ υποδηλώνει το "ακέραιο μέρος" και

$$d(\beta_1, \beta_2, n) = \sum_{j=[n\beta_1]+1}^{n-[n\beta_2]} \sum_{i=0}^{j-1} (n - i)^{-1}.$$

Αυτοί οι εκτιμητές αντιστοιχούν στους εκτιμητές κομμένου μέσου που εισήγαγε και

μελέτησε ο Kimber (1983a, b) για το αντίστοιχο πρόβλημα εκτίμησης του $\theta = \alpha^{-1}$ στο μοντέλο $E(\mu,\theta)$ με γνωστή την παράμετρο μ. Οι παραπάνω τιμές των c_{ni} είναι μια επιλογή που καθιστά το $\hat{\theta}_T = \hat{\alpha}_T^{-1}$ αμερόληπτο για $\theta = \alpha^{-1}$.

6.6 Γενικευμένοι εκτιμητές διαμέσων

Οι στατιστικές των Γενικευμένων Διάμεσων (GM) ορίζονται λαμβάνοντας τη διάμεσο των $\binom{n}{k}$ τιμών ενός δοθέντος πυρήνα $h(x_1,...,x_k)$ πάνω από όλα τα k σύνολα των δεδομένων. Δείτε το έργο των Serfling (1984, 2000) για γενική συζήτηση. Στην εργασία των Brazauskas και Serfling (2000a), τέτοιοι εκτιμητές εξετάστηκαν για την παράμετρο α στην περίπτωση όπου το σ είναι γνωστό

$$\hat{\alpha}_{GM} = \text{Διάμεσος}\{h(X_{i_1}, \dots, X_{i_k})\}$$

με συγκεκριμένο πυρήνα $h(x_1, \dots, x_k)\}$:

$$h(x_1, \dots, x_k; \sigma) = \frac{1}{C_k} \frac{k}{\sum_{j=1}^{k} \log(X_j/\sigma)}$$

όπου το C_k είναι ένας αμερόληπτος πολλαπλασιαστικός παράγοντας της διαμέσου, δηλαδή επιλέγεται έτσι ώστε η κατανομή του $h(x_1, \dots, x_k; \sigma)$ να έχει διάμεσο α. Οι τιμές του C_k, για k = 2:10, παρέχονται στον ακόλουθο πίνακα. (Για k > 10, ο C_k προσεγγίζεται από τον τύπο $C_k \approx k/(k - 1/3)$.

ΠΙΝΑΚΑΣ 6-1

Τιμές C_k, για k από 2 μέχρι 10

k	2	3	4	5	6	7	8	9	10
C_k	1.1916	1.1219	1.0893	1.0705	1.0582	1.0495	1.0431	1.0382	1.0343

Κ Ε Φ Α Λ Α Ι Ο 7

Εφαρμογή σε πραγματικά δεδομένα

7.1 Επιλογή δείγματος

Το σύνολο δεδομένων Καταστροφών λόγω ανέμων (1977) προέρχεται από τον Hogg και τον Klugman (1984) Αντιπροσωπεύει 40 απώλειες που συνέβησαν το 1977 λόγω καταστροφών από ανέμους. Τα δεδομένα καταγράφηκαν στα πλησιέστερα εκατομμύρια δολάρια και περιλαμβάνουν μόνο αυτές τις απώλειες που ήταν $2,000,000 ή περισσότερο. Στον επόμενο πίνακα παρουσιάζονται οι απώλειες (σε εκατομμύρια δολάρια):

ΠΙΝΑΚΑΣ 7-1

Απώλειες λόγω καταστροφών από ανέμους

2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	4	4	4	5
5	5	5	6	6	6	6	8	8	9	15	17	22	23	24	24	25	27	32	43

Οι Hogg και Klugman (1984) χρησιμοποίησαν δύο παραμετρικά μοντέλα για να ταιριάξουν τα δεδομένα των ανέμων: την διακεκομμένη εκθετική (με το σημείο διακοπής 1,5) και την Pareto. Οι Derrig, Ostaszewski και Rempala (2000) μελέτησαν επίσης αυτό το σύνολο δεδομένων και, εκτός από τα παραμετρικά μοντέλα που αναφέρθηκαν παραπάνω, χρησιμοποίησαν μη παραμετρικές εμπειρικές προσεγγίσεις για να εκτιμήσουν την πιθανότητα ότι οι απώλειες από τους ανέμους θα υπερβούν τα 29,5 εκατομμύρια δολάρια. Επιπλέον, ο Philbrick (1985), ανάμεσα σε πολλές εφαρμογές της κατανομής Pareto σε πραγματικά δεδομένα, διερεύνησε την P(σ, α) προσαρμογή δεδομένων των ανέμων με το σημείο διακοπής σ = 2 και πρότεινε τη χρήση του MLE για την εκτίμηση του, αλλά φαίνεται ότι δεν γνώριζε ότι αυτός ο εκτιμητής είναι μεροληπτικός.

Ξεκινάμε με ένα παράδειγμα βασισμένο στα δεδομένα ανέμου. Οι απώλειες που καταγράφονται στρογγυλοποιούνται στο πλησιέστερο εκατομμύριο, πράγμα που υποδηλώνει ότι οι πραγματικές απώλειες που αντιστοιχούν στο 2, δεν ήταν ακριβώς 2 αλλά κάπου ανάμεσα στο

1,5 και το 2,5. Για να αποφευχθεί η συγκέντρωση δεδομένων και οι ακατάλληλες ομαδοποιήσεις απαιτούμενες από αυτήν τη στρογγυλοποίηση, θα εφαρμόσουμε έναν απλό τρόπο για τον αναπροσδιορισμό των δεδομένων.

Συνεχίζοντας με τα δεδομένα του ανέμου και, ειδικότερα, τις απώλειες μεγέθους 2, είναι λογικό να υποθέσουμε ότι οι πραγματικές παρατηρήσεις που αντιστοιχούν στο 2 είναι εξίσου απομακρυσμένες (ή, αντίστοιχα, ομοιόμορφα κατανεμημένες) στο διάστημα (1,5, 2,5). Έτσι, για τα δεδομένα του ανέμου, αντί για τις 12 παρατηρήσεις με τιμή "2" θα χρησιμοποιήσουμε ως πραγματικά δεδομένα τα 1,58, 1,65, 1,73, 1,81, 1,88, 1,96, 2,04, 2,12, 2,19, 2,27, 2,35, 2,42. Πιο αυστηρά:

Εάν (A, B) είναι ένα διάστημα απωλειών και m είναι ο αριθμός των απωλειών μέσα σε αυτό το διάστημα, τότε m ομοιόμορφα κατανεμημένες απώλειες $x_1, ..., x_m$ σε αυτό το διάστημα βρίσκονται σύμφωνα με τον τύπο

$$x_k = \left(1 - \frac{k}{m+1}\right)A + \frac{k}{m+1}B\,, \quad k = 1, \dots, m$$

Επισημαίνουμε ότι μια τέτοια προσέγγιση δεν παραμορφώνει την αρχική ομαδοποίηση ή δεν αλλάζει το συνολικό ποσό απωλειών στην ομάδα. Είναι εύκολο να εφαρμοστεί στην πράξη και, πιο σημαντικό ακόμη, καθιστά τα δεδομένα συνεχή, επιτρέποντας έτσι την άμεση εφαρμογή μεθόδων εκτίμησης και ελέγχου καλής προσαρμογής. Τέλος, μπορεί κανείς να εξετάσει και πιο περίπλοκες διαδικασίες απο-ομαδοποίησης δεδομένων χρησιμοποιώντας, για παράδειγμα, την οικογένεια κατανομών beta αντί της ομοιόμορφης κατανομής. Σε αυτήν την περίπτωση, όμως, απαιτείται μια επιπλέον πληροφορία, όπως η μέση τιμή και η διακύμανση των απωλειών στο διάστημα.

7.2 Μεθοδολογία

Για να αξιολογήσουμε εάν το σύνολο δεδομένων μας έχει την μορφή κατανομής βαριάς ουράς, όπως η κατανομή Pareto, θα ακολουθήσουμε την παρακάτω διαδικασία:

1. Εκτίμηση των παραμέτρων της θεωρητικής κατανομής: Χρησιμοποιούμε της μεθόδους εκτίμησης παραμέτρων που περιγράψαμε στο κεφάλαιο 3.

2. Έλεγχος της εμπειρικής κατανομής: Συγκρίνουμε τις τιμές που προκύπτουν από την παραπάνω θεωρητική κατανομή με την εμπειρική κατανομή. Μπορούμε να χρησιμοποιήσουμε γραφήματα, όπως το γράφημα της εμπειρικής συνάρτησης επιβίωσης ή το γράφημα QQ-plot, για να συγκρίνουμε την εμπειρική κατανομή με τη θεωρητική κατανομή Pareto.

3. Στατιστική αξιολόγηση: Χρησιμοποιούμε τις στατιστικές μετρικές για να αξιολογήσουμε την αντιστοιχία μεταξύ της θεωρητικής και της εμπειρικής κατανομής.

Μπορούν να υπάρχουν και άλλα στάδια ή στοιχεία τα οποία είναι χρήσιμα να περιληφθούν στη διαδικασία αξιολόγησης της αντιστοιχίας μεταξύ της θεωρητικής και της εμπειρικής κατανομής. Ορισμένα από αυτά μπορούν να περιλαμβάνουν:

4. Υποθέσεις και προϋποθέσεις: Πριν από την εκτέλεση της ανάλυσης, είναι σημαντικό να γνωρίζουμε και να ελέγχουμε τις υποθέσεις και τις προϋποθέσεις που απαιτούνται για τη χρήση της κατανομής Pareto και των σχετικών μεθόδων εκτίμησης.

5. Αξιολόγηση ποιότητας εφαρμογής: Μπορεί να αξιολογηθεί η ποιότητα της εφαρμογής της θεωρητικής κατανομής στα δεδομένα. Αυτό μπορεί να γίνει μέσω γραφημάτων, όπως το ιστόγραμμα και το QQ-plot, καθώς και μέσω της επισκόπησης των αποτελεσμάτων της ανάλυσης.

6. Προσαρμογή μοντέλου: Σε ορισμένες περιπτώσεις, μπορεί να είναι απαραίτητο να προσαρμοστεί ένα πιο περίπλοκο μοντέλο στα δεδομένα, αν η απλή κατανομή Pareto δεν αρκεί.

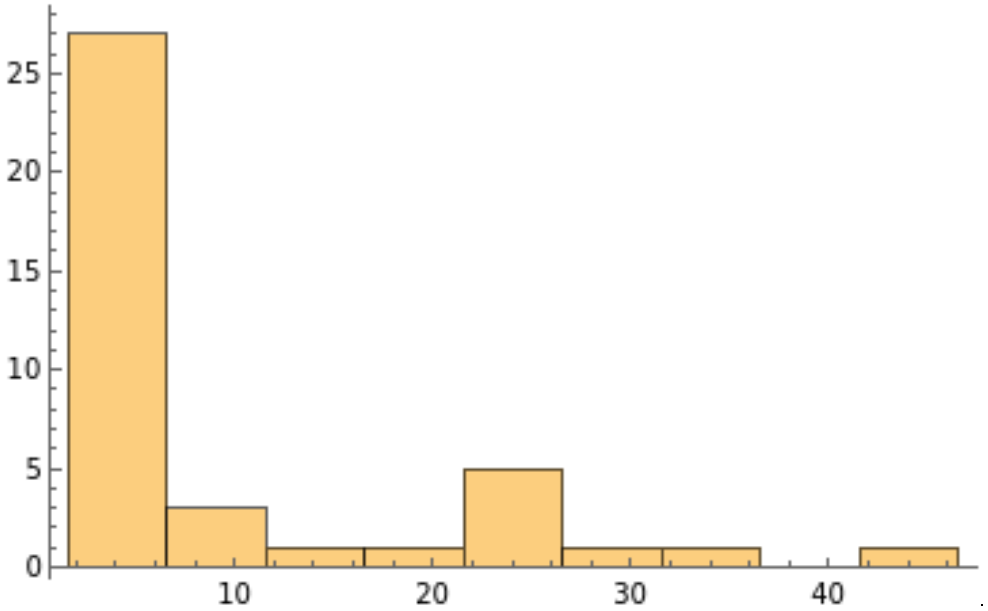
7.3 Αρχικές εκτιμήσεις

Αρχικά παρουσιάζουμε τα δεδομένα μας σε μορφή ιστογράμματος συχνοτήτων, το οποίο

αντιπροσωπεύει τις αναφερόμενες απώλειες από τους ανέμους.

ΣΧΗΜΑ 7-1

Ιστόγραμμα συχνοτήτων δεδομένων ανέμου



Η μέση τιμή των απωλειών και η τυπική απόκλιση αντίστοιχα είναι

$$\mu = 9.23 \text{ και } s = 10.24$$

Παρατηρώντας το ιστόγραμμα, εκτιμάμε ότι τα δεδομένα μας κατανέμονται με τέτοιο τρόπο, που πιθανόν να προδίδουν την ύπαρξη μιας κατανομής με βαριά ουρά, καθώς εμφανίζονται ακραίες τιμές και μεγάλη διασπορά στις υψηλές τιμές του δείγματος. Αυτό σημαίνει ότι ορισμένες τιμές είναι πολύ μεγαλύτερες από την μέση τιμή του δείγματος και την τυπική απόκλιση.

Συγκεκριμένα, αντιλαμβανόμαστε ότι οι ακραίες τιμές επηρεάζουν την τελική τιμή του μέσου όρου, καθώς αυξάνουν την συνεισφορά τους στον υπολογισμό του. Έτσι, ο μέσος όρος μπορεί να μην αποτελεί αντιπροσωπευτικό μέτρο της κεντρικής τάσης των δεδομένων μας σε αυτή την περίπτωση.

Όσον αφορά τη διακύμανση, η παρουσία ακραίων τιμών έχει σημαντική επίδραση, καθώς αυξάνει τη διασπορά των δεδομένων. Αυτό οδηγεί σε μεγαλύτερη τυπική απόκλιση και δυσκολεύει την εκτίμηση της πραγματικής διακύμανσης της κατανομής.

Η σχέση μεταξύ της τυπικής απόκλισης και της μέσης τιμής δεν μαρτυρά απαραίτητα την ύπαρξη μιας κατανομής με βαριά ουρά. Η τυπική απόκλιση είναι μεγαλύτερη από τη μέση τιμή και αυτό υποδεικνύει την υψηλή διασπορά των δεδομένων, αλλά δεν σημαίνει απαραιτήτως ότι η κατανομή έχει βαριά ουρά.

Η βαριά ουρά αναφέρεται στο γεγονός ότι ορισμένες τιμές του δείγματος είναι ασυνήθιστα μεγάλες συγκριτικά με τη μέση τιμή και τις υπόλοιπες τιμές. Η τυπική απόκλιση δεν αποτυπώνει απευθείας αυτήν την πτυχή της κατανομής.

Για να εξακριβώσουμε μια ισχυρότερη εκτίμηση για την ύπαρξη κατανομής με βαριά ουρά, θα εξετάσουμε τη συμπεριφορά των ακραίων τιμών σε σχέση με τη μέση τιμή και τη διασπορά των δεδομένων, ως προς τα επόμενα χαρακτηριστικά:

1. Ύπαρξη ασυνήθιστα μεγάλων τιμών που βρίσκονται μακριά από τη μέση τιμή και τις υπόλοιπες τιμές του δείγματος.

2. Ύπαρξη σημαντικού πλήθους ακραίων τιμών σε σχέση με τον συνολικό αριθμό των παρατηρήσεων.

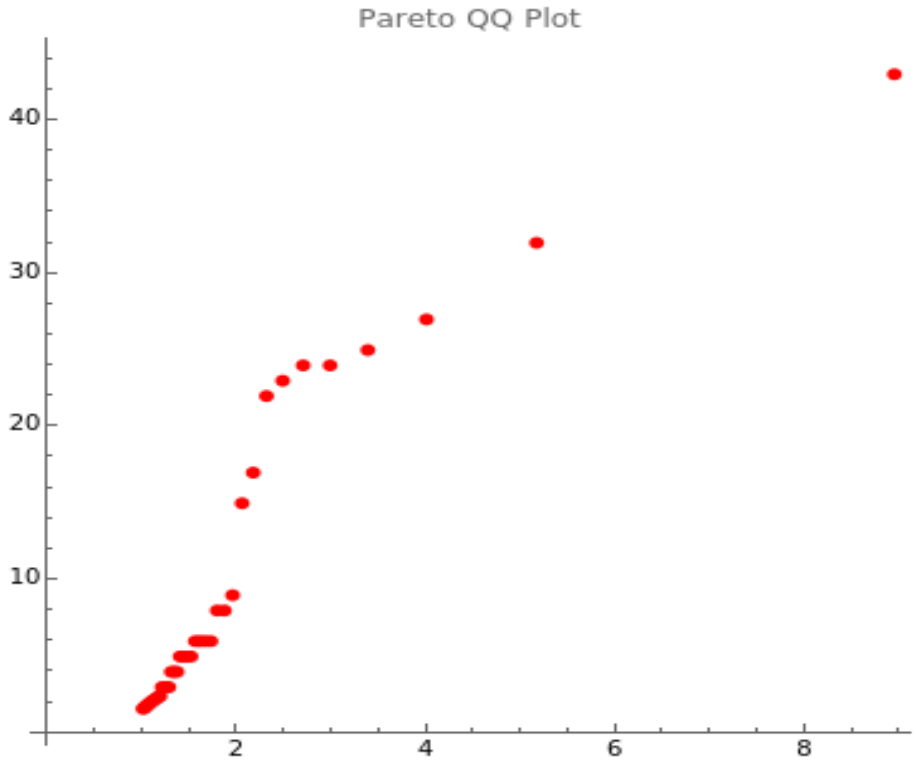
3. Ασυνήθιστα υψηλή πιθανότητα εμφάνισης ακραίων τιμών σε σχέση με την κανονική κατανομή.

Διαπιστώνουμε ότι τα παραπάνω χαρακτηριστικά εμφανίζονται στο υπόψη ιστόγραμμα καθώς τιμές μεγαλύτερες από το 20 μπορούν να θεωρηθούν ακραίες. Οι τιμές αυτές αποτελούν το 20% των δεδομένων μας και επομένως και η πιθανότητα εμφάνισής τους είναι πολύ μεγαλύτερη απ ότι στην κανονική κατανομή όπου η αντίστοιχη πιθανότητα συνήθως είναι μικρότερη από 5%.

Ένας άλλος τρόπος για να εκτιμήσουμε ότι τα δεδομένα μας κατανέμονται με τέτοιο τρόπο, που πιθανόν να προδίδουν την ύπαρξη μιας κατανομής με βαριά ουρά είναι το QQ διάγραμμα το οποίο μας παρέχει μια οπτική αναπαράσταση της προσαρμογής της Pareto κατανομής στα δεδομένα μας και μας βοηθά να αξιολογήσουμε την καταλληλότητα της κατανομής για την περιγραφή των δεδομένων μας.

ΣΧΗΜΑ 7-2

Διάγραμμα QQ κατανομής Pareto για τα δεδομένα μας



Βάσει του παραπάνω διαγράμματος μπορούμε να εξάγουμε ορισμένα συμπεράσματα σχετικά με την προσαρμογή των δεδομένων στην κατανομή Pareto:

- Το διάγραμμα QQ παρουσιάζει μια σχετική ευθυγράμμιση των σημείων με μια γραμμή, υποδεικνύοντας ότι η κατανομή Pareto μπορεί να προσαρμοστεί σχετικά καλά στα δεδομένα.

- Οι παράμετροι της κατανομής Pareto (α=1, β=2) που χρησιμοποιούνται φαίνεται να δημιουργούν μια καλή ευθεία προσαρμογής στα δεδομένα.

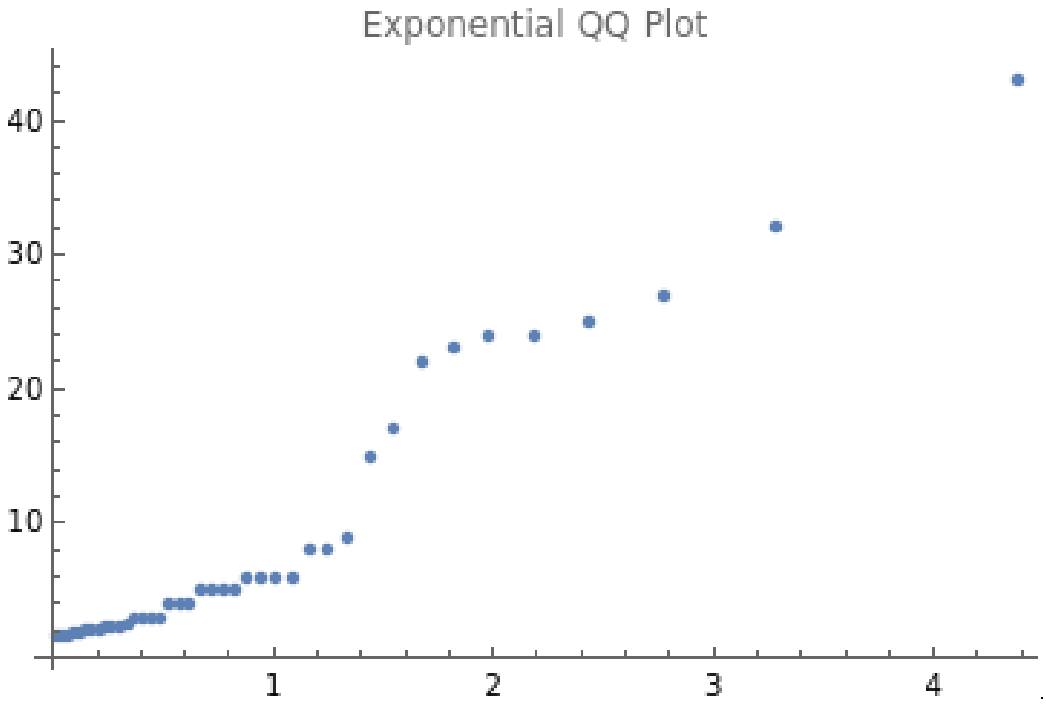
Τα δεδομένα προσαρμόζονται σχετικά καλά στην κατανομή Pareto με τις συγκεκριμένες παραμέτρους α=1 και β=2. Ωστόσο, είναι σημαντικό να σημειωθεί ότι η εκτίμηση αυτή βασίζεται στις επιλεγμένες παραμέτρους της κατανομής Pareto. Παρότι τα δεδομένα φαίνεται να προσαρμόζονται σχετικά καλά στην κατανομή Pareto με τις επιλεγμένες παραμέτρους, είναι σημαντικό να εξετάσουμε και άλλες πιθανές κατανομές που μπορεί να ταιριάζουν επίσης στα δεδομένα μας. Η επιλογή της κατάλληλης κατανομής για τη μοντελοποίηση των δεδομένων

εξαρτάται από τη φύση του προβλήματος και την κατανόηση του πεδίου εφαρμογής.

Για λόγους σύγκρισης θα χρησιμοποιήσουμε την εκθετική κατανομή για να δούμε κατά πόσον τα δεδομένα μας προσαρμόζονται και σε αυτή την κατανομή.

ΣΧΗΜΑ 7-3

Διάγραμμα QQ εκθετικής κατανομής για τα δεδομένα μας



Βάσει των αποτελεσμάτων του παραπάνω διαγράμματος, μπορούμε να εξάγουμε ορισμένα συμπεράσματα σχετικά με την προσαρμογή των δεδομένων στην εκθετική κατανομή:

- Το διάγραμμα QQ παρουσιάζει μια απόκλιση των σημείων από την ευθεία, υποδεικνύοντας ότι η εκθετική κατανομή δεν προσαρμόζεται ιδανικά στα δεδομένα.

- Τα σημεία δεν ακολουθούν στενά τις ταξινομημένες τιμές των δεδομένων.

Συναφώς, συμπεραίνουμε ότι η εκθετική κατανομή δεν είναι η κατάλληλη επιλογή για να μοντελοποιήσουμε τα δεδομένα μας. Οι τιμές των δεδομένων φαίνεται να αποκλίνουν από την εκθετική κατανομή και είναι σημαντικό να σημειωθεί ότι η εκτίμηση αυτή βασίζεται στην επιλογή της εκθετικής κατανομής με παράμετρο λ=1.

Η εκτίμηση της προσαρμογής της κατανομής στα δεδομένα αποτελεί μια αρχική αξιολόγηση που μπορεί να μας κατευθύνει για περαιτέρω ανάλυση. Η ανίχνευση της βαριάς ουράς τόσο από το ιστόγραμμα όσο και από το διάγραμμα QQ είναι μια εκτίμηση που δεν παρέχει απόλυτη απόδειξη και για πιο ακριβείς και στατιστικά επιβεβαιωμένες εκτιμήσεις, απαιτείται να χρησιμοποιήσουμε τις προχωρημένες μεθόδους ανάλυσης δεδομένων.

7.4 Αναλυτική Εφαρμογή

Για να εκτιμήσουμε τις παραμέτρους της κατανομής Pareto, θα χρησιμοποιήσουμε αρχικά και αναλυτικά την μέθοδο μέγιστης πιθανοφάνειας. Οι εκτιμήτριες για τις παραμέτρους της κατανομής Pareto με αυτή την μέθοδο είναι:

Για την ελάχιστη τιμή (σ):

$$\hat{\sigma} = \min\{x_i\} = 1{,}58$$

Για τον δείκτη κλίσης (α):

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln(x_i/\hat{\sigma})} = 0{,}795341$$

Για να συνεχίσουμε, χρειαζόμαστε την αθροιστική συνάρτηση κατανομής (CDF) της θεωρητικής κατανομής Pareto με τις εκτιμηθείσες παραμέτρους. Η CDF της κατανομής Pareto με παραμέτρους $\hat{\alpha}$ και $\hat{\sigma}$ ορίζεται ως εξής:

$$F_X(x) = 1 - \left(\frac{1{,}58}{x}\right)^{0{,}795341}$$

όπου $x \geq 1{,}58$.

Από τα δεδομένα που μας δόθηκαν, μπορούμε να υπολογίσουμε τις τιμές της CDF της κατανομής Pareto για κάθε παρατήρηση. Στη συνέχεια, συγκρίνουμε αυτές τις τιμές με την αντίστοιχη εμπειρική CDF των δεδομένων.

Τιμές Δεδομένων: 1.58, 1.65, 1.73, 1.81, 1.88, 1.96, 2.04, 2.12, 2.19, 2.27, 2.35, 2.42, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 8, 8, 9, 15, 17, 22, 23, 24, 24, 25, 27, 32, 43.

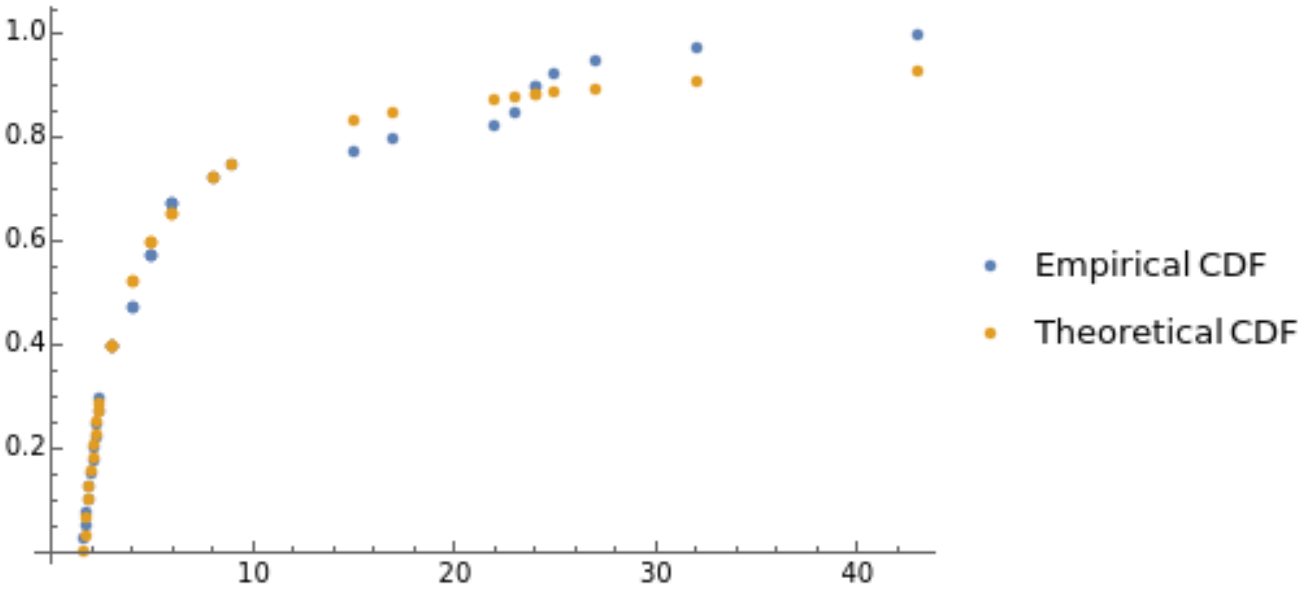
Τιμές Εμπειρικής $F_X(x)$: 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25, 0.275, 0.3, 0.4, 0.475, 0.575, 0.675, 0.725, 0.75, 0.775, 0.8, 0.825, 0.85, 0.9, 0.925, 0.95, 0.975, 1.

Τιμές Θεωρητικής $F_X(x)$: 0, 0.034, 0.070, 0.102, 0.129, 0.158, 0.184, 0.208, 0.229, 0.250, 0.271, 0.288, 0.399, 0.522, 0.6, 0.654, 0.725, 0.749, 0.833, 0.849, 0.877, 0.881, 0.885, 0.889, 0.895, 0.909, 0.928.

Στο επόμενο διάγραμμα 6-4 παρουσιάζεται η σύγκριση μεταξύ της παραπάνω θεωρητικής και εμπειρικής αθροιστικής κατανομής της Pareto.

ΣΧΗΜΑ 7-4

Διάγραμμα εμπειρικής – θεωρητικής CDF για $\hat{\alpha} = 0{,}795341$ και $\hat{\sigma} = 1{,}58$



Αναλύοντας το διάγραμμα, παρατηρούμε μια μικρή σχετικά απόσταση μεταξύ των δύο κατανομών. Αυτή η μικρή απόκλιση μας κάνει να υποπτευθούμε ότι η κατανομή Pareto πιθανόν προσαρμόζεται ιδανικά στα δεδομένα μας.

Για να επαληθεύσουμε με ακρίβεια την παρατήρηση αυτή, θα εκτελέσουμε τον επόμενο στατιστικό έλεγχο:

H_0 : Η κατανομή προσαρμόζεται στα δεδομένα μας

H_1 : Η κατανομή δεν προσαρμόζεται στα δεδομένα μας.

Εφαρμόζοντας τον KS στατιστικό έλεγχο βρίσκουμε την στατιστική D_n ίση με 0.0722496 και την τιμή p-value ίση με 0.997. Συνεπώς επειδή η τιμή αυτή είναι μεγαλύτερη από 0.05, που είναι το επιλεγμένο επίπεδο σημαντικότητας του ελέγχου, η μηδενική υπόθεση H_0 δεν απορρίπτεται.

Εφαρμόζοντας τον CvM στατιστικό έλεγχο βρίσκουμε την στατιστική W_n^2 ίση με 0.0298354 και την τιμή p-value ίση με 0.976706. Συνεπώς επειδή η τιμή αυτή είναι μεγαλύτερη από 0.05, που είναι το επιλεγμένο επίπεδο σημαντικότητας του ελέγχου, η μηδενική υπόθεση H_0 δεν απορρίπτεται.

Εφαρμόζοντας τον AD στατιστικό έλεγχο βρίσκουμε την στατιστική A_n^2 ίση με 0.334556 και την τιμή p-value ίση με 0.909716. Συνεπώς επειδή η τιμή αυτή είναι μεγαλύτερη από 0.05, που είναι το επιλεγμένο επίπεδο σημαντικότητας του ελέγχου, η μηδενική υπόθεση H_0 δεν απορρίπτεται

Συνεχίζοντας την ανάλυση χρησιμοποιούμε τώρα την μέθοδο των ροπών. Οι εκτιμήτριες για τις παραμέτρους της κατανομής Pareto με αυτή την μέθοδο είναι:

Για την ελάχιστη τιμή (σ):

$$\hat{\sigma} = \frac{(\hat{\alpha}n - 1)X_m}{\hat{\alpha}n} = 1{,}54712$$

Για τον δείκτη κλίσης (α):

$$\hat{\alpha} = \frac{n\bar{X} - X_m}{n(\bar{X} - X_m)} = 1{,}2015$$

Η CDF της κατανομής Pareto με παραμέτρους $\hat{\alpha}$ και $\hat{\sigma}$ ορίζεται ως εξής:

$$F_X(x) = 1 - \left(\frac{1{,}54712}{x}\right)^{1{,}2015}$$

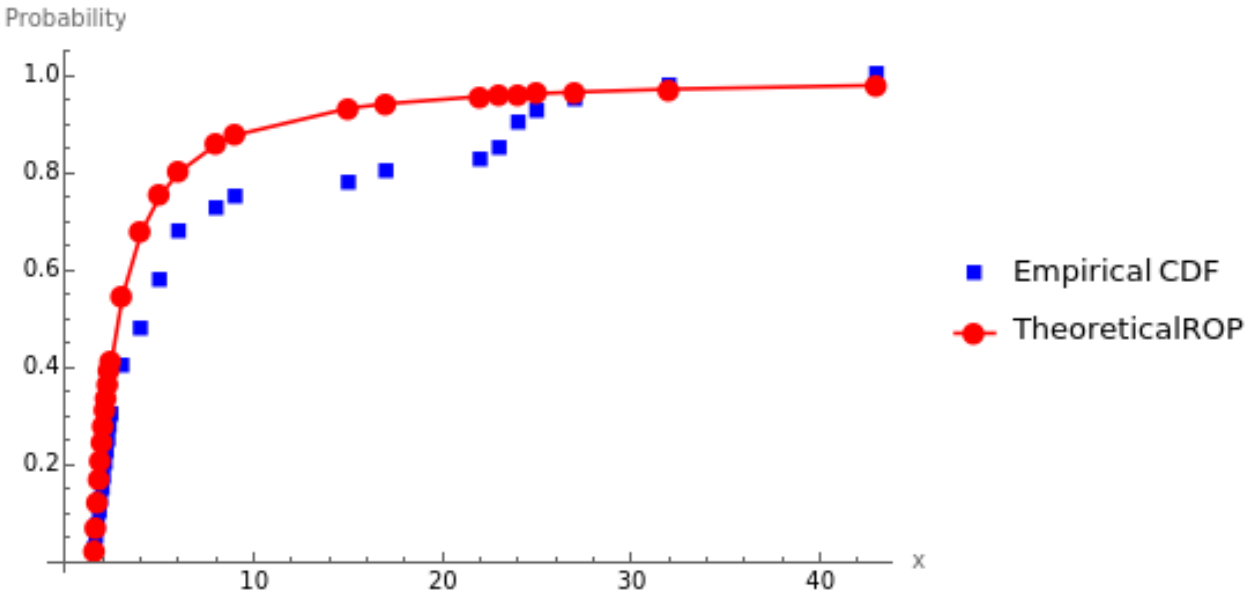
όπου x ≥ 1,54712.

Τιμές Θεωρητικής $F_X(x)$: 0.0249473, 0.0744337, 0.125616, 0.171841, 0.20875, 0.247393, 0.282712, 0.315109, 0.341326, 0.369117, 0.394832, 0.415803, 0.548714, 0.680598, 0.755713, 0.803771, 0.861117, 0.879444, 0.934742, 0.943853, 0.95881, 0.960952, 0.962899, 0.964675, 0.967795, 0.973741, 0.981588

Στο επόμενο διάγραμμα 6-5 παρουσιάζεται η σύγκριση μεταξύ της παραπάνω θεωρητικής και εμπειρικής αθροιστικής κατανομής της Pareto.

ΣΧΗΜΑ 7-5

Διάγραμμα εμπειρικής – θεωρητικής CDF για $\hat{\alpha} = 1{,}2015$ και $\hat{\sigma} = 1{,}54712$



Αναλύοντας το διάγραμμα, παρατηρούμε μια μικρή σχετικά απόσταση μεταξύ των δύο κατανομών στις τιμές κάτω από 8 και άνω του 22. Στο διάστημα 8 – 22 οι αποκλίσεις φαίνεται να είναι μεγαλύτερες.

Για να διαπιστώσουμε με ακρίβεια την υπόθεση ότι τα δεδομένα προσαρμόζονται στην κατανομή Pareto, θα εφαρμόσουμε, όπως και πριν, τον KS στατιστικό έλεγχο. Ο έλεγχος βρίσκει ότι D_n = 0.205598 και η τιμή p-value είναι ίση με 0.1766. Αν και η τιμή αυτή είναι πολύ μικρότερη σε σχέση με την προηγούμενη μέθοδο, είναι μεγαλύτερη από 0.05 που είναι το επιλεγμένο επίπεδο σημαντικότητας του ελέγχου και επομένως η μηδενική υπόθεση H_0 δεν απορρίπτεται.

Εφαρμόζοντας τον CvM στατιστικό έλεγχο βρίσκουμε ότι W_n^2 = 0.199931 και η τιμή p-value είναι ίση με 0.267603. Η τιμή αυτή είναι μεγαλύτερη από 0.05 που είναι το επιλεγμένο επίπεδο σημαντικότητας του ελέγχου και επομένως η μηδενική υπόθεση H_0 δεν απορρίπτεται.

Εφαρμόζοντας τον AD στατιστικό έλεγχο βρίσκουμε ότι A_n^2 = 0.95116 και η τιμή p-value είναι ίση με 0.286158. Η τιμή αυτή είναι μεγαλύτερη από 0.05 που είναι το επιλεγμένο επίπεδο σημαντικότητας του ελέγχου και επομένως η μηδενική υπόθεση H_0 δεν απορρίπτεται.

Τέλος, χρησιμοποιούμε την μέθοδο των ποσοστιαίων σημείων. Οι εκτιμήτριες για τις παραμέτρους της κατανομής Pareto με αυτή την μέθοδο είναι:

Για την ελάχιστη τιμή (σ):

$$\hat{\sigma} = x_1(1 - P_1)^{1/\hat{\alpha}} = 1.54349$$

Για τον δείκτη κλίσης (α):

$$\hat{\alpha} = \frac{\log\frac{1 - P_1}{1 - P_2}}{\log\frac{x_2}{x_1}} = 0.768658$$

όπου εδώ επιλέξαμε $x_1 = 1.65$, $x_2 = 3$, $P_1 = 0.05$ και $P_2 = 0.4$.

Η CDF της κατανομής Pareto με παραμέτρους $\hat{\alpha}$ και $\hat{\sigma}$ ορίζεται ως εξής:

$$F_X(x) = 1 - \left(\frac{1.54349}{x}\right)^{0.768658}$$

όπου $x \geq 1.54349$.

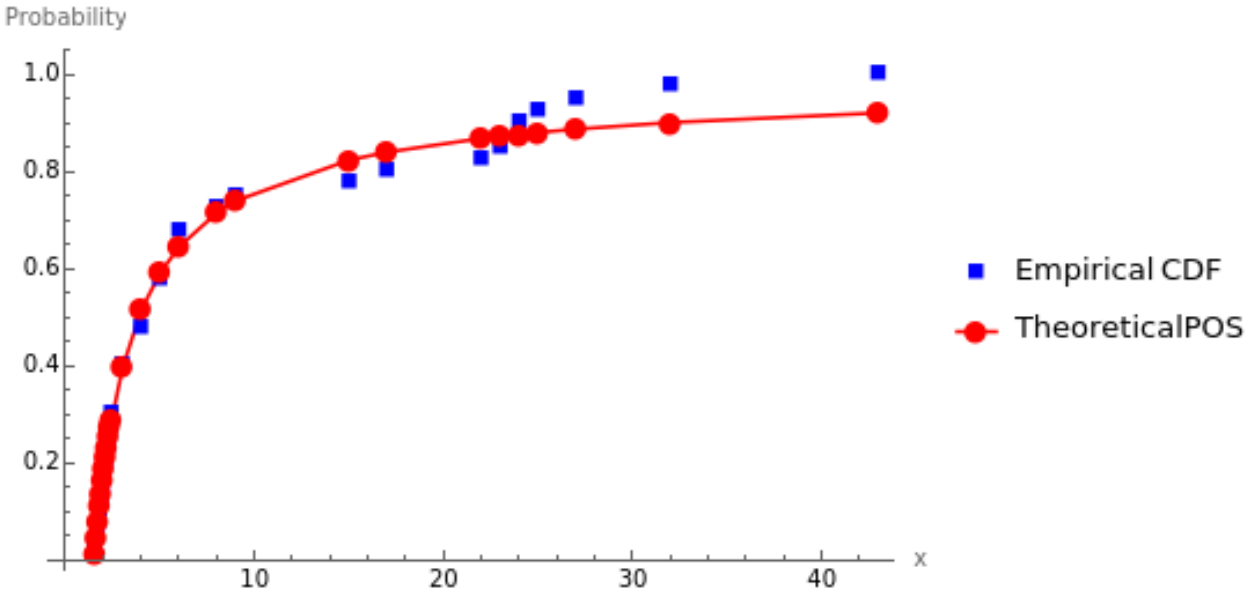
Τιμές Θεωρητική $F_X(x)$: 0.0178111, 0.05, 0.0839518, 0.115236, 0.140668, 0.167758, 0.192961, 0.216474, 0.235796, 0.256583, 0.276114, 0.292264, 0.4, 0.519032, 0.594841, 0.647822, 0.71769, 0.742126, 0.825867, 0.841839, 0.870273, 0.874631, 0.878666, 0.882414, 0.889168, 0.902737, 0.922498

Στο επόμενο διάγραμμα 6-6 παρουσιάζεται η σύγκριση μεταξύ της παραπάνω θεωρητικής

και εμπειρικής αθροιστικής κατανομής της Pareto.

ΣΧΗΜΑ 7-6

Διάγραμμα εμπειρικής – θεωρητικής CDF για $\hat{\alpha} = 0.768658$ και $\hat{\sigma} = 1.54349$



Αναλύοντας το διάγραμμα, παρατηρούμε μια μικρή σχετικά απόσταση μεταξύ των δύο κατανομών και μάλιστα τα δεδομένα φαίνονται να ταιριάζουν περισσότερο με το αρχικό διάγραμμα που απεικονίζει την προσαρμογή με την μέθοδο των μέγιστων πιθανοτήτων.

Για να διαπιστώσουμε με ακρίβεια την υπόθεση ότι τα δεδομένα προσαρμόζονται στην κατανομή Pareto, θα εφαρμόσουμε όπως και πριν, τον KS στατιστικό έλεγχο. Ο έλεγχος βρίσκει ότι D_n = 0.0775025 και η τιμή p-value είναι ίση με 0.9912. Η τιμή αυτή είναι μεγαλύτερη από 0.05, που είναι το επιλεγμένο επίπεδο σημαντικότητας του ελέγχου και επομένως η μηδενική υπόθεση H_0 δεν απορρίπτεται.

Εφαρμόζοντας όπως και πριν, τον CvM στατιστικό έλεγχο παίρνουμε W_n^2 = 0.0332647 και η τιμή p-value είναι ίση με 0.964251. Η τιμή αυτή είναι μεγαλύτερη από 0.05, που είναι το επιλεγμένο επίπεδο σημαντικότητας του ελέγχου και επομένως η μηδενική υπόθεση H_0 δεν απορρίπτεται.

Εφαρμόζοντας όπως και πριν, τον AD στατιστικό έλεγχο παίρνουμε A_n^2 = 0.400457 και η τιμή p-value είναι ίση με 0.847628. Η τιμή αυτή είναι μεγαλύτερη από 0.05, που είναι το

επιλεγμένο επίπεδο σημαντικότητας του ελέγχου και επομένως η μηδενική υπόθεση H_0 δεν απορρίπτεται.

7.5 Αποτελέσματα

Εδώ θα παρουσιάσουμε τον συνοπτικό πίνακα των δεδομένων μας, που δείχνει τις εκτιμήσεις της παραμέτρου α, τις τιμές των στατιστικών KS, CvM και AD, καθώς και τις θέσεις των εκτιμητών βάσει αυτών των μετρήσεων καλής εφαρμογής. Στην συνέχεια συνάγονται συμπεράσματα και παρέχονται συστάσεις. Οι επιδόσεις των εκτιμητών που εξετάζονται είναι αυτοί που παρουσιάστηκαν στο 5ο κεφάλαιο και στο 6ο κεφάλαιο. Οι τιμές των εκτιμητών του 5ου κεφαλαίου προσδιορίστηκαν από τους Brazauskas και Serfling (2003).

Παρατηρήσεις. Οι θέσεις των εκτιμητών ανατίθενται ως εξής: Ο εκτιμητής με τη χαμηλότερη τιμή για μια επιλεγμένη μέτρηση καλής εφαρμογής, λαμβάνει θέση 1, ο εκτιμητής με τη δεύτερη χαμηλότερη τιμή (για την ίδια μέτρηση) λαμβάνει θέση 2, κ.λπ. Η ιδέα της κατάταξης εκτιμητών ή μοντέλων βάσει κάποιου κριτηρίου δεν είναι καινούργια. Έχει προταθεί και συζητηθεί αρκετά από τους Klugman, Panjer και Willmot (1998).

Το προσαρμοσμένο μοντέλο για τους εκτιμητές του 5ου κεφαλαίου είναι το P(â, σ=1.5) με τιμές του $\hat{\alpha}$ να κυμαίνονται από 0,605 (για $\hat{\alpha}_Q\ \text{opt}, 2$) έως 0,791 (για $\hat{\alpha}_Q\ \text{opt}, 5$). Ο κύριος λόγος που υπάρχει αυτή η διαφορά είναι η επιλογή του σημείου διάσπασης. Στον Philbrick (1985), χρησιμοποιείται το σ = 2. Ωστόσο, με βάση τους Hogg και Klugman (1984), η επιλογή του 1,5 είναι πιο φυσική.

Επιπλέον, όλα τα τεστ καλής προσαρμογής υποστηρίζουν ισχυρά την καταλληλότητα του μοντέλου $P(\hat{\alpha}_{ML} = 0.764, \sigma = 1.5)$ με τις τιμές των δεικτών καλής προσαρμογής 0.1071 (KS), 0.1106 (CvM), 0.7329 (AD) και αντίστοιχα p-value: 0.51 (KS), 0.27 (CvM), 0.24 (AD). Ενώ τα αντίστοιχα p-value για το μοντέλο $P(\hat{\alpha}_{ML} = 0.945, \sigma = 2.0)$ είναι συγκρίσιμα για τα στατιστικά CvM και AD, το p-value για την στατιστική KS είναι σημαντικά χαμηλότερο: 0.33 (KS), 0.30 (CvM), 0.23 (AD). Έτσι, βάσει αυτής της συζήτησης, επιλέγεται το μοντέλο $P(\hat{\alpha}, \sigma = 1.5)$.

Ο πίνακας 6-2 υποδηλώνει ότι, αν και το μοντέλο $P(\sigma = 1.5, \hat{\alpha}_{ML})$ είναι αποδεκτό και από

τους τρεις ελέγχους καλής προσαρμογής, επιπλέον βελτιώσεις της προσαρμογής είναι δυνατές αν χρησιμοποιήσουμε την αξιόπιστη έκδοση MLU, η οποία μπορεί να βελτιωθεί ακόμα περισσότερο από ανθεκτικούς εκτιμητές. Για παράδειγμα, οι εκτιμητές $\hat{\alpha}_T$ (με $\beta_1 = 0$, $\beta_2 = .05$), $\hat{\alpha}_{GM}$ (με k = 4, k = 5, και k = 10) και $\hat{\alpha}_Q$ έχουν όλοι ομοιόμορφα μικρότερες τάξεις από τον $\hat{\alpha}_{MLU}$.

ΠΙΝΑΚΑΣ 7-2

Τιμές του $\hat{\alpha}$, στατιστικές και βαθμοί για τα δεδομένα ανέμου.

Εκτιμητής	$\hat{\alpha}$	KS	rank	CvM	rank	AD	rank	Overall Score
$\hat{\alpha}_{MLE}$	0.795	0.0722	1	0.0298	1	0.3346	1	3
$\hat{\alpha}_{ΕΠΣ}$	0.769	0.0775	2	0.0333	2	0.4005	2	6
T, β_1=0, β2=0,10	0.677	0.1031	10	0.0562	4	0.5335	4	18
T, β_1=0, β2=0,25	0.673	0.1045	11	0.0561	3	0.5368	5	19
GM, k=3	0.692	0.0981	9	0.0587	7	0.5316	3	19
T, β_1=0, β2=0,05	0.707	0.0932	6	0.0642	9	0.5457	7	22
T, β_1=0, β2=0,20	0.667	0.1066	12	0.0564	5	0.5441	6	23
GM, k=4	0.714	0.0912	5	0.0679	10	0.5576	9	24
GM, k=5	0.723	0.0884	3	0.0734	11	0.5777	11	25
T, β_1=0, β2=0,15	0.644	0.1077	13	0.0568	6	0.5487	8	27
Q^*, k=5	0.731	0.0911	4	0.0792	12	0.5999	12	28
GM, k=2	0.653	0.1118	14	0.0594	8	0.5720	10	32
GM, k=10	0.744	0.0975	7	0.0901	13	0.6445	13	33
MLU	0.745	0.0980	8	0.0911	14	0.6484	14	36
$Q^{opt,2}$	0.605	0.1320	16	0.0956	15	0.7939	15	46
$Q^{opt,5}$	0.791	0.1198	15	0.1445	16	0.8881	16	47
$\hat{\alpha}_{EMP}$	1.202	0.2056	17	0.1999	17	0.9512	17	51

Στον επόμενο πίνακα υπολογίζουμε τις τυπικές αποκλίσεις κατάταξης (rank) για κάθε εκτιμητή προκειμένου να αποκτήσουμε μια εικόνα για την ανθεκτικότητα των εκτιμητών στα διαφορετικά μέτρα καλής προσαρμογής.

ΠΙΝΑΚΑΣ 7-3

Τυπικές αποκλίσεις εκτιμητών

Εκτιμητής	KS Rank	CvM Rank	AD Rank	Standard Deviation
$\hat{\alpha}_{MLE}$	1	1	1	0.00
$\hat{\alpha}_{ΕΠΣ}$	2	2	2	0.00
T, β_1=0, β2=0,10	10	4	4	3.46
T, β_1=0, β2=0,25	11	3	5	4.16
GM, k=3	9	7	3	3.06
T, β_1=0, β2=0,05	6	9	7	1.53
T, β_1=0, β2=0,20	12	5	6	3.79
GM, k=4	5	10	9	2.65
GM, k=5	3	11	11	4.62
T, β_1=0, β2=0,15	13	6	8	3.61
Q^*, k=5	4	12	12	4.62
GM, k=2	14	8	10	3.06
GM, k=10	7	13	13	3.46
MLU	8	14	14	3.46
$Q^{opt,2}$	16	15	15	0.58
$Q^{opt,5}$	15	16	16	0.58
$\hat{\alpha}_{EMP}$	17	17	17	0.00

Η μέση τιμή των μη μηδενικών τυπικών αποκλίσεων του παραπάνω πίνακα ανέρχεται στο 3.04. Αυτή η προσέγγιση υποδηλώνει ότι τιμές πάνω από 4.00 μπορούν να θεωρηθούν ότι παρουσιάζουν ακραία μικρή ανθεκτικότητα ενώ αυτές κάτω από τον μέσο όρο ότι παρουσιάζουν μεγαλύτερη ανθεκτικότητα.

Παρατηρούμε επίσης ότι οι έλεγχοι CvM και AD δίνουν σχεδόν σε κάθε εκτιμητή την ίδια κατάταξη και επομένως κρίνεται σκόπιμο να εξετάσουμε την συσχέτιση μεταξύ των τιμών κατάταξης των τριών μέτρων η οποία φαίνεται στον επόμενο πίνακα που υπολογίστηκε με χρήση του λογισμικού spss.

ΠΙΝΑΚΑΣ 7-4

Συσχέτιση τιμών κατάταξης

Correlations

		KS_Rank	CvM_Rank	AD_Rank
KS_Rank	Pearson Correlation	1	,419	,505*
	Sig. (2-tailed)		,094	,039
	N	17	17	17
CvM_Rank	Pearson Correlation	,419	1	,958**
	Sig. (2-tailed)	,094		,000
	N	17	17	17
AD_Rank	Pearson Correlation	,505*	,958**	1
	Sig. (2-tailed)	,039	,000	
	N	17	17	17

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Από τον παραπάνω πίνακα συσχετίσεων προκύπτει ότι υπάρχει πολύ ισχυρή (σχεδόν τέλεια) θετική και στατιστικά σημαντική συσχέτιση στις κατατάξεις των CvM και AD σε επίπεδο σημαντικότητας 1%, όπου ο συντελεστής συσχέτισης ανέρχεται στο 0,958. Αντίστοιχα υπάρχει μέτρια θετική και στατιστικά σημαντική συσχέτιση στις κατατάξεις των KS και AD σε επίπεδο σημαντικότητας 5%, όπου ο συντελεστής συσχέτισης ανέρχεται στο 0,505.

Στον επόμενο πίνακα παρουσιάζουμε την μεταβλητότητα των μέτρων καλής προσαρμογής υπολογίζοντας τον συντελεστή μεταβλητότητας CV (σ / μ).

ΠΙΝΑΚΑΣ 7-5

Μεταβλητότητα μέτρων καλής προσαρμογής

CV KS	CV CvM	CV AD
0,28	0,53	0,26

Παρατηρούμε ότι ο CvM παρουσιάζει την μεγαλύτερη μεταβλητότητα σε σχέση με τα άλλα δύο μέτρα τα οποία έχουν περίπου την ίδια μεταβλητότητα.

Σύμφωνα με τα παραπάνω αποτελέσματα, συνάγεται ότι το μέτρο CvM μπορεί να θεωρηθεί ότι δίνει ταυτόσημη κατάταξη με τον AD, διότι παρόλο που ο CvM έχει μεγαλύτερη μεταβλητότητα από όλα τα μέτρα, έχει σχεδόν τέλεια συσχέτιση με τον AD και συμπεριφέρεται όπως αυτόν.

Αντίθετα ο KS παρόλο που έχει μικρή μεταβλητότητα όπως ο AD, παρουσιάζει σημαντικές διαφορές με τον AD και κατά συνέπεια και με τον CvM χαρακτηρίζοντας τον αντίστοιχο εκτιμητή ως μη ανθεκτικό. Για τους λόγους αυτούς και χωρίς βλάβη της γενικότητας, θα θεωρήσουμε ότι αν ένας εκτιμητής παρουσιάζει τουλάχιστον 2 από τις 3 κατατάξεις των αντίστοιχων στατιστικών μικρότερες από έναν άλλο εκτιμητή, τότε η απόδοσή του μπορεί να θεωρείται καλύτερη.

Με βάση αυτό το κριτήριο, η τελική κατάταξη των εκτιμητών προσδιορίζεται όπως τον παρακάτω πίνακα:

ΠΙΝΑΚΑΣ 7-6

Τελική κατάταξη εκτιμητών

Εκτιμητής	KS Rank	CvM Rank	AD Rank	Standard Deviation	Overall Score
$\hat{\alpha}_{\mathrm{MLE}}$	1	1	1	0.00	3
$\hat{\alpha}_{\mathrm{ΕΠΣ}}$	2	2	2	0.00	6
GM, k=3	9	7	3	3.06	19
T, β_1=0, β2=0,10	10	4	4	3.46	18
T, β_1=0, β2=0,25	11	3	5	4.16	19
T, β_1=0, β2=0,20	12	5	6	3.79	23
T, β_1=0, β2=0,05	6	9	7	1.53	22
T, β_1=0, β2=0,15	13	6	8	3.61	27
GM, k=4	5	10	9	2.65	24
GM, k=2	14	8	10	3.06	32
GM, k=5	3	11	11	4.62	25
Q^*, k=5	4	12	12	4.62	28
GM, k=10	7	13	13	3.46	33
MLU	8	14	14	3.46	36
$Q^{opt,2}$	16	15	15	0.58	46
$Q^{opt,5}$	15	16	16	0.58	47
$\hat{\alpha}_{\mathrm{EMP}}$	17	17	17	0.00	51

Κ Ε Φ Α Λ Α Ι Ο 8

Συμπεράσματα

8.1 Ανασκόπηση των βασικών βημάτων της εφαρμογής

Κατά την εφαρμογή του μοντέλου Pareto στα δεδομένα απωλειών από καταστροφές ανέμων, υλοποιήθηκαν τα παρακάτω βήματα:

1. Τα δεδομένα αρχικά ομαδοποιήθηκαν χωρίς να αλλάξουν οι συνολικές τιμές των απωλειών, καθιστώντας τα έτσι συνεχή και επιτρέποντας την εφαρμογή σε αυτά των μεθόδων εκτίμησης και ελέγχου καλής προσαρμογής.

2. Χρησιμοποιήσαμε διαγνωστικά ιστογράμματα και γραφήματα QQ για να εκτιμήσουμε οπτικά εάν το μοντέλο Pareto είναι κατάλληλο για την περιγραφή των δεδομένων μας.

3. Υπολογίσαμε όλους τους εκτιμητές και εφαρμόσαμε για τον καθένα τους στατιστικούς ελέγχους KS, CvM και AD για να ελέγξουμε με αυστηρότητα εάν το μοντέλο Pareto παρέχει μια επαρκή εφαρμογή στα δεδομένα μας και παρουσιάσαμε συνολικά τα αποτελέσματα στον Πίνακα 6-2.

4. Υπολογίσαμε την τυπική απόκλιση κατάταξης των εκτιμητών στα τρία μέτρα καλής προσαρμογής για να εκτιμήσουμε τον βαθμό ανθεκτικότητάς τους και μετρήσαμε την μεταβλητότητα του κάθε μέτρου.

5. Τέλος αξιολογήσαμε συνολικά τα αποτελέσματα και παρουσιάσαμε την τελική κατάταξη των εκτιμητών.

8.2 Συνολικά συμπεράσματα εργασίας

Βασιζόμενοι στις συγκρίσεις των αποτελεσμάτων των Πινάκων 6-2, 6-3, 6-4, 6-5 και 6-6 προκύπτουν τα ακόλουθα συμπεράσματα:

• Αρχικά και τα τρία τεστ καλής προσαρμογής υποστηρίζουν το μοντέλο Pareto σε όλους τους εκτιμητές, με διαφορετική απόδοση ο καθένας.

• Οι συγκρίσεις δείχνουν ότι οι εκτιμητές $\hat{\alpha}_{MLE}$ και $\hat{\alpha}_{ΕΠΣ}$ κυριαρχούν στον ανταγωνισμό και ακολούθως οι ευνοϊκοί εκτιμητές GM (k=3), T (β_1=0, β2=0.10) και T (β_1=0, β2=0.25).

• Ο τελικός πίνακας ενδεικτικά εκφράζει ότι ο εκτιμητής $\hat{\boldsymbol{\alpha}}_{\mathbf{T}}$ ($\beta_1 = 0$, $\beta_2 = 0.10$) με κατατάξεις (10, 4, 4) και απόδοση 18, είναι καλύτερος από τον εκτιμητή $\hat{\boldsymbol{\alpha}}_{\mathbf{T}}$ ($\beta_1 = 0$, $\beta_2 = 0.050$) με κατατάξεις (6, 9, 7) αλλά χειρότερος από τον εκτιμητή $\hat{\boldsymbol{\alpha}}_{\mathbf{GM}}$ (k = 3) με κατατάξεις (9, 7, 3) και απόδοση 19.

• Οι εκτιμητές τύπου GM προσφέρουν τις καλύτερες ισορροπίες μεταξύ ανθεκτικότητας και αποδοτικότητας, το οποίο συνεπάγεται εξαιρετική απόδοση όσον αφορά την καλή προσαρμογή. Οι καλύτερες προσαρμογές παρέχονται από τους εκτιμητές $\hat{\alpha}_{GM}$ (k = 3 και k = 4).

• Οι εκτιμητές τύπου T είναι ελαφρώς λιγότερο ανταγωνιστικοί όσον αφορά τις συγκρίσεις "ανθεκτικότητας έναντι αποδοτικότητας". Ωστόσο, η απόδοσή τους στην καλή προσαρμογή είναι τουλάχιστον ισάξια με αυτή των εκτιμητών τύπου GM. Οι καλύτερες προσαρμογές παρέχονται από τους εκτιμητές $\hat{\alpha}_{T}$ ($\beta_1 = 0$, $\beta_2 = 0.05$ και $\beta_1 = 0$, β2 = 0.10).

• Οι εκτιμητές τύπου Q υπολείπονται και ως προς τα δύο κριτήρια, "ανθεκτικότητα έναντι αποδοτικότητας" και καλή προσαρμογή, από ότι τους εκτιμητές τύπου T και GM και επομένως, είναι λιγότερο ανταγωνιστικοί.

• Ο μη ανθεκτικός αλλά πιο αποδοτικός εκτιμητής MLU δεν μπορεί ούτε να βελτιωθεί ούτε να βελτιώσει κάποιον άλλο εκτιμητή όσον αφορά το κριτήριο "ανθεκτικότητας έναντι αποδοτικότητας". Ωστόσο, η απόδοσή του όσον αφορά την καλή προσαρμογή είναι συνεχώς

ανάμεσα στις χειρότερες, υποδηλώνοντας ότι για τις συγκρίσεις "ανθεκτικότητας έναντι αποδοτικότητας" η ανθεκτικότητα πρέπει να έχει υψηλότερη προτεραιότητα.

ΠΑΡΑΡΤΗΜΑΤΑ

Π1 Κώδικας Σχημάτων με χρήση της Mathematica

Π2 Κώδικας Προγράμματος

Π1 Κώδικες Σχημάτων με χρήση της Mathematica

Σχήμα 2-1

```
xM = 2; (* Ελάχιστη τιμή *)
α = {1, 2, 3}; (* Παράμετρος μορφής *)
Plot[Evaluate[(α xM^α)/x^(α + 1)], {x, xM, 10},
 PlotRange -> {0, 2},
 PlotLegends -> {"α = 1", "α = 2", "α = 3"},
 FrameLabel -> {"x", "f(x)"},
 PlotLabel -> "Πυκνότητα ΚατανομήςPareto"]
```

Σχήμα 2-2

```
xM = 2; (* Ελάχιστη τιμή *)
α = {1, 2, 3}; (* Παράμετρος μορφής *)
Plot[Evaluate[1 - (xM/x)^α], {x, xM, 10},
 PlotRange -> {0, 1},
 PlotLegends -> {"α = 1", "α = 2", "α = 3"},
 FrameLabel -> {"x", "F(x)"},
 PlotLabel -> "Αθροιστική Συνάρτηση Κατανομής Pareto"]
```

Σχήμα 2-3

```
σ = 2; (* Παράμετρος κλίμακας *)
α = {1, 2, 3}; (* Παράμετρος μορφής *)
Plot[Evaluate[(σ/x)^α], {x, σ, 10},
 PlotRange -> {0, 1},
 PlotLegends -> {"α = 1", "α = 2", "α = 3"},
 FrameLabel -> {"x", "S(x)"},
 PlotLabel -> "Συνάρτηση Επιβίωσης Pareto"]
```

Σχήμα 2-4

```
α = Range[3.1, 5, 0.1]; (* Περιοχή τιμών για τη παράμετρο α *)
Λοξότητα = (2(1+α))/(α-3) Sqrt[(α-2)/α];
ListPlot[Transpose[{α, Λοξότητα}],
 Joined -> True,
 FrameLabel -> {"α", "Λοξότητα"},
 PlotLabel -> "Λοξότητα Κατανομής Pareto (α>3)",
 PlotRange -> All]
```

Σχήμα 2-5

```
α = Range[4.1, 10, 0.1]; (* Περιοχή τιμών για τη παράμετρο α *)
Κύρτωση = (6(α^3+α^2-6α-2))/(α(α-3)(α-4));
ListPlot[Transpose[{α, Κύρτωση}],
 Joined -> True,
 FrameLabel -> {"α", "Κύρτωση"},
 PlotLabel -> "Συντελεστής Κύρτωσης Κατανομής Pareto",
 PlotRange -> All]
```

Σχήμα 2-6

```
α = {1, 4}; (* Παράμετρος μορφής *)
Plot[Evaluate[α/x], {x, 0.01, 10},
 PlotRange -> {0, 5},
 PlotLegends -> {"α = 1", "α = 4"},
 FrameLabel -> {"x", "λ(x)"},
 PlotLabel -> "Ρυθμός Αποτυχίας Pareto"]
```

Σχήμα 5-1

```
(* Ορισμός των συναρτήσεων *)
fExp[x_] := Exp[-x]
fPareto[x_] := (0.35*1^0.35)/x^(0.35 + 1)
fComposite[x_] := Piecewise[{{0.775*Exp[-1.35 x], x >= 1},
{0.2*1^0.35/x^1.35, x < 1}}]
(* Απεικόνιση των γραφικών παραστάσεων στο ίδιο σχήμα *)
Plot[{fExp[x], fPareto[x], fComposite[x]}, {x, 0, 10},
  PlotLegends -> {"Εκθετική", "Pareto", "Σύνθετη Συνάρτηση"},
  PlotStyle -> {Blue, Red, Green},
  AxesLabel -> {"x", "f(x)"},
  PlotRange -> {0, 2},
  PlotLabel -> "Γραφικές Παραστάσεις Κατανομών",
  AspectRatio -> 0.8]
```

Σχήμα 7-1

```
data = {1.58, 1.65, 1.73, 1.81, 1.88, 1.96, 2.04, 2.12, 2.19, 2.27,
2.35, 2.42, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 8, 8, 9, 15,
17, 22, 23, 24, 24, 25, 27, 32, 43};
binWidth = 5;
binRanges = Range[Min[data], Max[data] + binWidth, binWidth];
histogramData = HistogramList[data, {binRanges}];
Histogram[data, {binRanges}, "Count"]
```

Σχήμα 7-2

```
data = {1.58, 1.65, 1.73, 1.81, 1.88, 1.96, 2.04, 2.12, 2.19, 2.27,
2.35, 2.42, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 8, 8, 9, 15,
17, 22, 23, 24, 24, 25, 27, 32, 43};
sortedData = Sort[data];
n = Length[sortedData];
probabilities = (Range[n] - 0.5)/n;
α = 1;
β = 2;
paretoQuantiles = Quantile[ParetoDistribution[α, β], probabilities];
ListPlot[Transpose[{paretoQuantiles, sortedData}],
  PlotRange -> All,
  FrameLabel -> {"Pareto Quantiles", "Ordered Data"},
  AspectRatio -> 1,
  PlotLabel -> "Pareto QQ Plot",
  PlotMarkers -> {Automatic, 10},
  PlotStyle -> Directive[PointSize[0.02], Red]
]
```

Σχήμα 7-3

```
data = {1.58, 1.65, 1.73, 1.81, 1.88, 1.96, 2.04, 2.12, 2.19, 2.27,
2.35, 2.42, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 8, 8, 9, 15,
17, 22, 23, 24, 24, 25, 27, 32, 43};
sortedData = Sort[data];
n = Length[sortedData];
probabilities = Table[(i - 0.5)/n, {i, 1, n}];
exponentialQuantiles = Table[InverseCDF[ExponentialDistribution[1],
p], {p, probabilities}];
ListPlot[Transpose[{exponentialQuantiles, sortedData}],
  PlotRange -> All,
  FrameLabel -> {"Exponential Quantiles", "Data"},
  PlotLabel -> "Exponential QQ Plot"]
```

Τα σχήματα 7.4 έως 7.6 έγιναν με την χρήση του επόμενου προγράμματος

Π2 Κώδικας Προγράμματος

```
data = {1.58, 1.65, 1.73, 1.81, 1.88, 1.96, 2.04, 2.12, 2.19, 2.27,
2.35, 2.42, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 8, 8, 9, 15,
17, 22, 23, 24, 24, 25, 27, 32, 43};

σhat = Min[data];
αhat = Length[data]/Total[Log[#/σhat] & /@ data];

x̄ = Mean[data];
xₘ = Min[data];
n = Length[data];

αˆ= (n x̄ - xₘ)/(n (x̄ - xₘ));
σˆ= ((αˆn - 1) xₘ)/(αˆn);

aq=Log[(1-0.05)/(1-0.4)]/Log[3/1.65];
θq=1.65(1-0.05)^(1/(aq));

Print["Οι εκτιμητές με τη μέθοδο Μέγιστης Πιθανοφάνειας είναι:"]
σhat
αhat

Print["Οι εκτιμητές με την μέθοδο των Ροπών είναι:"]
αˆ
σˆ
Print["Οι εκτιμητές με την μέθοδο των Ποσοστιαίων Σημείων είναι:"]
aq
θq

empiricalFx = EmpiricalDistribution[data];
xValues = Sort[DeleteDuplicates[data]];
empiricalFxValues = N[CDF[empiricalFx, xValues]];

k = Length[xValues];

TheoreticalMLE= 1-( (σhat)/(xValues))^(αhat)

squares = Transpose[{xValues, empiricalFxValues}];
circles = Transpose[{xValues, TheoreticalMLE}];
ListPlot[{squares, circles}, PlotStyle -> {Directive[Blue,
PointSize[0.02]], Directive[Red, PointSize[0.02]]}, PlotMarkers -> {{
"■", 15}, {"●", 15}}, Joined -> {False, True}, PlotLegends ->
{"Empirical CDF", "Theoretical MLE"}, AxesLabel -> {"x",
"Probability"}, ImageSize -> 400]

DMLE = Max[Abs[empiricalFxValues - TheoreticalMLE]]
```

```
m = k; n = 10000; s = 0; d = DMLE;
Do[U = Table[Random[], {m}]; U = Sort[U]; d1 = Max[Table[i/m - U[[i]],
{i, 1, m}]]; d2 = Max[Table[U[[i]] - (i - 1)/m, {i, 1, m}]];
If[Max[{d1, d2}] >= d, s = s + 1], {j, 1, n}]; ep = N[s/n];
Print["p-value Simulation estimate : ", ep, " ", {ep-(ep(1-
ep)/n)^0.5*1.96, ep+(ep(1-ep)/n)^0.5*1.96}];

TheoreticalROP= 1-( (σ)/(xValues))^(α)

squares = Transpose[{xValues, empiricalFxValues}];
circles = Transpose[{xValues, TheoreticalROP}];
ListPlot[{squares, circles}, PlotStyle -> {Directive[Blue,
PointSize[0.02]], Directive[Red, PointSize[0.02]]}, PlotMarkers -> {{
"■", 15}, {"●", 15}}, Joined -> {False, True}, PlotLegends ->
{"Empirical CDF", "TheoreticalROP"}, AxesLabel -> {"x",
"Probability"}, ImageSize -> 400]

DROP = Max[Abs[empiricalFxValues - TheoreticalROP]]

 m = k; n = 10000; s = 0; d = DROP;
Do[U = Table[Random[], {m}]; U = Sort[U]; d1 = Max[Table[i/m - U[[i]],
{i, 1, m}]]; d2 = Max[Table[U[[i]] - (i - 1)/m, {i, 1, m}]];
If[Max[{d1, d2}] >= d, s = s + 1], {j, 1, n}]; ep = N[s/n];
Print["p-value Simulation estimate : ", ep, " ", {ep-(ep(1-
ep)/n)^0.5*1.96, ep+(ep(1-ep)/n)^0.5*1.96}];

TheoreticalPOS= 1-( (θq)/(xValues))^(aq)

squares = Transpose[{xValues, empiricalFxValues}];
circles = Transpose[{xValues, TheoreticalPOS}];
ListPlot[{squares, circles}, PlotStyle -> {Directive[Blue,
PointSize[0.02]], Directive[Red, PointSize[0.02]]}, PlotMarkers -> {{
"■", 15}, {"●", 15}}, Joined -> {False, True}, PlotLegends ->
{"Empirical CDF", "TheoreticalPOS"}, AxesLabel -> {"x",
"Probability"}, ImageSize -> 400]

 DPOS = Max[Abs[empiricalFxValues - TheoreticalPOS]]

 m = k; n = 10000; s = 0; d = DPOS;
Do[U = Table[Random[], {m}]; U = Sort[U]; d1 = Max[Table[i/m - U[[i]],
{i, 1, m}]]; d2 = Max[Table[U[[i]] - (i - 1)/m, {i, 1, m}]];
If[Max[{d1, d2}] >= d, s = s + 1], {j, 1, n}]; ep = N[s/n];
Print["p-value Simulation estimate : ", ep, " ", {ep-(ep(1-
ep)/n)^0.5*1.96, ep+(ep(1-ep)/n)^0.5*1.96}];

dataAndEmpiricalFx = Transpose[{xValues, empiricalFxValues}]

TableForm[dataAndEmpiricalFx, TableHeadings -> {None, {"x", "Empirical
F_X(x)"}}]
```

```
CvMStat = CramerVonMisesTest[empiricalFxValues, TheoreticalMLE,
"TestStatistic"]
CvMPvalue = CramerVonMisesTest [empiricalFxValues, TheoreticalMLE,
"PValue"]

CvMStat = CramerVonMisesTest[empiricalFxValues, TheoreticalROP,
"TestStatistic"]
CvMvalue = CramerVonMisesTest[empiricalFxValues, TheoreticalROP,
"PValue"]

CvMStat = CramerVonMisesTest[empiricalFxValues, TheoreticalPOS,
"TestStatistic"]
CvMPvalue = CramerVonMisesTest[empiricalFxValues, TheoreticalPOS,
"PValue"]

ADStat = AndersonDarlingTest[empiricalFxValues, TheoreticalMLE,
"TestStatistic"]
ADPvalue = AndersonDarlingTest[empiricalFxValues, TheoreticalMLE,
"PValue"]

ADStat = AndersonDarlingTest[empiricalFxValues, TheoreticalROP,
"TestStatistic"]
ADvalue = AndersonDarlingTest[TheoreticalROP, empiricalFxValues,
"PValue"]

ADStat = AndersonDarlingTest[empiricalFxValues, TheoreticalPOS,
"TestStatistic"]
ADPvalue = AndersonDarlingTest[empiricalFxValues, TheoreticalPOS,
"PValue"]
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Βλαχάκης Ν., Δαγκλής Μ., (2010). *Εισαγωγή στην Στατιστική*. Αθήνα: Εκδόσεις Κλειδάριθμος.

Βλαχογιάννης Α., Μπούγιας Ν., Φωτίου Ε., Κοτζανίκος Π., (2016*). Στατιστική Εκτίμηση. Θεσσαλονίκη*: Εκδόσεις Ζήτη

Κωνσταντάτος Μ., Παπαγεωργίου Ι., (2006). *Στοιχεία Στατιστικής με Χρήση του Mathematica*. Αθήνα: Εκδόσεις Συμμετρία. (Διαθέσιμο στο διαδίκτυο μέσω του ιστότοπου της Εκδόσεων Συμμετρία:

Μπαλταδώρος Π. Γ., (2011). *Στατιστική Ανάλυση με το R. Αθήνα*: Εκδόσεις Κριτική. (Διαθέσιμο στο διαδίκτυο μέσω του Google Books)

Παπαδογιάννης Α., Γαλετάκης Ν., (2008). *Στατιστική Εφαρμογή με το λογισμικό R*. Αθήνα: Εκδόσεις Σταμούλη.

Τσάτσος Κ., Παναγιωτόπουλος Π., Κατσιφαράκης Γ., (2009). *Εισαγωγή στην Στατιστική*. Αθήνα: Εκδόσεις Συμμετρία.

Ξένη

Brazauskas V., Serfling R., (2003). *Favorable estimators for fitting pareto models: a study using goodness-of-fit measures with actual data*, Cambridge University Press, Vol. 33, No. 2, pp. 365-381.

Chatfield C. (2013). *The Analysis of Time Series: An Introduction (6th Edition)*. Boca Raton, FL: CRC Press.

Clauset A., Shalizi C. R., Newman M. E. J.. (2009). *Power-law distributions in empirical data*. SIAM Review, 51(4), 661-703

Cooray, K, Ananda, M.A. (2005), "Modeling actuarial data with a composite Lognormal-Pareto model". *Scandinavian Actuarial Journal,* 5: 321-334.

Costa M. L.,Machado J. A. T., Ferreira O. P.. (2011). *The Pareto Distribution: Concepts, Models, and Applications*. Boca Raton, FL: CRC Press. (Διαθέσιμο στο διαδίκτυο μέσω του CRC Press)

Hogg, P.V. and Klugman, s.a. (1984). Loss Distributions. Wiley, New York.

Philbrick, S.W. (1985) A practical guide to the single parameter Pareto distribution. Proceedings of the Casualty Actuarial Society LXXII, 44-84.

Jain C., Khan S. Z. & Jain N. K.. (2018). *Pareto Distribution: Theory, Methods, and Applications*. Boca Raton, FL: CRC Press. (Διαθέσιμο στο διαδίκτυο μέσω του CRC Press)

Johnson, N. L., Kotz, S., Balakrishnan, N. (1994), *Continuous Univariate Distributions*. Wiley, New York.

Kaas, R., Goovaerts, M., Denuit, M., Dhaene, J. (2001), *Modern Actuarial Risk Theory*. Kluwer Academic Publishers, Boston.

Klugman A. S., Panjer H. H., Willmot G. E., (1998). *Loss models: from data to decisions, 3rd ed.*, Drake University, University of Waterloo, University of Waterloo.

Klugman A. S., Panjer H. H., Willmot G. E., (2018). *Loss models, from data to decisions, fifth edition*, U.S.A, Wiley.

Klugman, S.A., Panjer, H.H., Willmot, G.E. (2004), *Loss models: from data to decisions* (2nd edition). New York, John Wiley & Sons, Inc.

Stephens M. A. (2001*). EDF Statistics for Goodness of Fit and Some Comparisons*. Journal of the American Statistical Association, 69(347), 730-737.

Teodorescu, S., Vernic, R. (2006), "A composite Exponential–Pareto distribution". *The Annals of the "Ovidius" University of Constanta*, Mathematics Series, vol.XIV (1): 99-108.

Weisstein E. C. (2003). *CRC Concise Encyclopedia of Mathematics (3rd Edition)*. Boca Raton, FL: CRC Press.

Zar J. H. (2010). *Biostatistical Analysis (5th Edition).* Upper Saddle River, NJ: Pearson Education