Μαρία Ι. Διαμαντοπούλου

Δασική Στατιστική

Θεωρία και Δασοβιομετρικές Εφαρμογές με χρήση IBM-SPSS και R

ΕΚΔΟΣΕΙΣ ΖΗΤΗ

Για επικοινωνία με τη συγγραφέα:

https://mdiamantopoulou.gr/

https://www.for.auth.gr/προσωπικο/μελη/δεπ/διαμαντοπουλου-μαρια

ISBN 978-960-456-621-1

Φωτοστοιχειοθεσία **Π. ΖΗΤΗ & Σια ΙΚΕ**
Εκτύπωση 18ο χλμ Θεσ/νίκης-Περαίας
Βιβλιοδεσία Τ.Θ. 4171 • Περαία Θεσσαλονίκης • Τ.Κ. 570 19
Τηλ.: 2392.072.222 - Fax: 2392.072.229 • e-mail: info@ziti.gr

ΕΚΔΟΣΕΙΣ ΖΗΤΗ ziti.gr

ΒΙΒΛΙΟΠΩΛΕΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ - ΚΕΝΤΡΙΚΗ ΔΙΑΘΕΣΗ:
Αρμενοπούλου 27, 546 35 Θεσσαλονίκη
Τηλ.: 2310.203.720, Fax: 2310.211.305 • e-mail: sales@ziti.gr

ΒΙΒΛΙΟΠΩΛΕΙΟ ΑΘΗΝΩΝ - ΠΩΛΗΣΗ ΛΙΑΝΙΚΗ-ΧΟΝΔΡΙΚΗ:
Χαριλάου Τρικούπη 22, 106 79 Αθήνα
Τηλ.-Fax: 210.3816.650 • e-mail: athina@ziti.gr

ΗΛΕΚΤΡΟΝΙΚΟ ΒΙΒΛΙΟΠΩΛΕΙΟ: www.ziti.gr

Πρόλογος

Ο σκοπός συγγραφής αυτού του βιβλίου ήταν η κάλυψη των αναγκών της διδασκαλίας του μαθήματος της «Εφαρμοσμένης Στατιστικής», του Τμήματος Δασολογίας και Φυσικού Περιβάλλοντος, του Α.Π.Θ. Η κάλυψη των αναγκών αυτών είχε διττό ρόλο. Αφενός την θεωρητική ανάπτυξη και κατανόηση βασικών μεθοδολογιών της εφαρμοσμένης στατιστικής σε βιολογικά και μάλιστα δασικά δεδομένα, αφετέρου την υλοποίηση των μεθοδολογιών αυτών, με χρήση των αυτοματοποιημένων ηλεκτρονικών διαδικασιών, οι οποίες προσφέρονται σήμερα μέσω στατιστικών προγραμμάτων και γλωσσών προγραμματισμού. Τέλος, δόθηκε ιδιαίτερη βαρύτητα στην κατανόηση των μεθοδολογιών οι οποίες αναπτύσσονται, κάνοντας χρήση του απαραίτητου βασικού μαθηματικού υποβάθρου αλλά και στην πρακτική κατανόηση της τελικής στατιστικής ερμηνείας και συμπερασματολογίας.

Κατ' αυτό τον τρόπο, πιστεύεται ότι το βιβλίο θα συμβάλλει έτσι ώστε η Δασική Στατιστική να αποτελέσει βάση και κατανοητό αρωγό στις ερευνητικές προσπάθειες αλλά και στην καθημερινή πρακτική των φοιτητριών/-ων του Τμήματος Δασολογίας και Φυσικού Περιβάλλοντος του Α.Π.Θ., αλλά και άλλων σχετικών Τμημάτων. Επιπλέον, φιλοδοξεί να αποτελέσει ένα κατανοητό βοήθημα για οποιονδήποτε επιστήμονα αντιμετωπίζει ζητήματα τα οποία άπτονται του πεδίου της Εφαρμοσμένης Στατιστικής.

Η ανάπτυξη των βιομετρικών μεθόδων από την εποχή του K. Pearson και του R. Fisher στις αρχές του 20^{ου} αιώνα, συνετέλεσε στη διαμόρφωση και ανάπτυξη της Δασικής Βιομετρίας ως τη βασική επιστήμη στη δασική έρευνα και πράξη, η οποία συνδυάζει μετρήσεις πεδίου με προηγμένες στατιστικές και υπολογιστικές τεχνικές, με σκοπό να παρέχει πληροφορίες για τη δυναμική των δασών και να καθοδηγεί τις διαδικασίες λήψης αποφάσεων. Ως αναπόσπαστο και βασικό μέρος της Δασικής Βιομετρίας, η Δασική Στατιστική δίνει έμφαση στην προσαρμογή των μεθόδων της εφαρμοσμένης στατιστικής, στη δασική έρευνα και πράξη. Για το λόγο αυτό, αποφασίστηκε η συγκεκριμένη επιλογή του τίτλου του βιβλίου ως: «*Δασική Στατιστική – Θεωρία και Δασοβιομετρικές εφαρμογές με χρήση IBM-SPSS και R*».

Η ύλη του βιβλίου βασίζεται στο ερευνητικό και συγγραφικό έργο της συγγραφέα και ενσωματώνει τη σύγχρονη επιστημονική γνώση σε αντικείμενα σχετικά με την Εφαρμοσμένη Δασική Στατιστική. Δόθηκε έμφαση, ώστε το βιβλίο να έχει έναν εφαρμοσμένο προσανατολισμό, αποφεύγοντας τις θεωρητικές αποδείξεις, παραπέμποντας τον ενδιαφερόμενο σε εξειδικευμένα βιβλία και επιστημονικά άρθρα, εφόσον επιθυμεί βαθύτερη μαθηματική αναζήτηση.

Επίσης, δόθηκε έμφαση, ώστε σε όλο το βιβλίο, η παρουσίαση της θεωρίας να συνοδεύεται από πληθώρα παραδειγμάτων, κυρίως δασοβιομετρικών μεταβλητών, για να είναι ευκολότερη η κατανόησή της. Όπου επίσης θεωρείται απαραίτητο, η επίλυση των παραδειγμάτων γίνεται με χρήση και αναλυτική περιγραφή της διαδικασίας, τόσο του στατιστικού πακέτου IBM-SPSS, όσο και της γλώσσας προγραμματισμού R.

Το βιβλίο αποτελείται από ένδεκα Κεφάλαια. Σε όλα τα κεφάλαια, δίνεται η επίλυση δασοβιομετρικών κυρίως παραδειγμάτων με χρήση και αναλυτική περιγραφή της διαδικασίας, τόσο του πακέτου IBM-SPSS, όσο και της γλώσσας προγραμματισμού R.

Στο πρώτο κεφάλαιο της Εισαγωγής περιγράφεται η έννοια και η διαχρονική διαδρομή της Στατιστικής, της Δασικής Βιομετρίας ως κλάδο της εφαρμοσμένης Στατιστικής, δίνονται εισαγωγικοί στατιστικοί ορισμοί και περιγράφονται οι εισαγωγικές βασικές έννοιες χρήσης της στατιστικής γλώσσας προγραμματισμού R και του στατιστικού πακέτου IBM® SPSS® Statistics (SPSS).

Στο δεύτερο κεφάλαιο περιγράφονται οι μέθοδοι και τα μέσα συλλογής στατιστικών στοιχείων, η προ-επεξεργασία των δεδομένων, οι διαδικασίες ελέγχου των πρωτογενών στοιχείων μέσω της διερευνητικής ανάλυσης δεδομένων, οι μέθοδοι παρουσίασης στατιστικών δεδομένων και αποτελεσμάτων στατιστικής ανάλυσης.

Στο τρίτο κεφάλαιο περιγράφονται εμπειρικές κατανομές συχνοτήτων και συγκεκριμένα η εμπειρική κατανομή συχνοτήτων ασυνεχούς μεταβλητής και η εμπειρική κατανομή συχνοτήτων συνεχούς μεταβλητής.

Στο τέταρτο κεφάλαιο περιγράφονται τα μέτρα κεντρικής τάσης, τα οποία περιλαμβάνουν τον αριθμητικό μέσο όρο και τις ιδιότητές του, τον σταθμισμένο αριθμητικό μέσο όρο, τον αποκομμένο αριθμητικό μέσο όρο, τον γεωμετρικό μέσο, τον αρμονικό μέσο, τον τετραγωνικό μέσο, τη σχέση μεταξύ αριθμητικού, γεωμετρικού, αρμονικού και τετραγωνικού μέσου, τους M-εκτιμητές, τη διάμεσο, τον υπολογισμό της διαμέσου σε μη ομαδοποιημένα στοιχεία και σε ομαδοποιημένα στοιχεία, την επικρατούσα τιμή, τον υπολογισμό της επικρατούσας τιμής σε μη ομαδοποιημένα στοιχεία και σε ομαδοποιημένα στοιχεία, τη σχέση μεταξύ αριθμητικού μέσου, διαμέσου και επικρατούσας τιμής, τα ποσοστημόρια και τα τεταρτημόρια.

Στο πέμπτο κεφάλαιο περιγράφονται τα μέτρα διασποράς, τα οποία περιλαμβάνουν το εύρος, το ενδοτεταρτημοριακό εύρος, τη μέση απόλυτη απόκλιση, τη διακύμανση, τις βασικές ιδιότητες της διακύμανσης, την τυπική απόκλιση, τις ιδιότητες της τυπικής απόκλισης, το συντελεστή μεταβλητότητας, το συντελεστή κύμανσης τεταρτημορίου και την καμπύλη του Lorenz.

Στο έκτο κεφάλαιο περιγράφονται τα μέτρα ασυμμετρίας και κύρτωσης, τα οποία περιλαμβάνουν τους συντελεστές ασυμμετρίας και κύρτωσης.

Στο έβδομο κεφάλαιο περιγράφονται τα στοιχεία πιθανοτήτων και οι θεωρητικές κατανομές πιθανοτήτων. Περιλαμβάνονται οι βασικοί ορισμοί και έννοιες της θεωρίας των πιθανοτήτων και οι αντίστοιχοι συμβολισμοί, η οπτική αναπαράσταση των συνόλων και των πράξεων αυτών, η έννοια και ο ορισμός της πιθανότητας, η στοχαστική ή στατιστική ανεξαρτησία, η δεσμευμένη πιθανότητα, το θεώρημα του Bayes, τα βασικά στοιχεία συνδυαστικής ανάλυσης, οι βασικές μέθοδοι προσδιορισμού του δειγματοχώρου σε σύνθετα πειράματα, η τυχαία ή στοχαστική μεταβλητή, η διωνυμική κατανομή, η κατανομή Poisson, η προσαρμογή εμπειρικής κατανομής δεδομένων δείγματος στις θεωρητικές συχνότητες της διωνυμικής κατανομής και της κατανομής Poisson, η κανονική και τυπική κανονική κατανομή, η προσαρμογή εμπειρικής κατανομής δεδομένων δείγματος σε κανονική κατανομή, η X^2-κατανομή, η *t*-κατανομή και η *F*-κατανομή.

Στο όγδοο κεφάλαιο περιγράφονται στοιχεία επαγωγικής στατιστικής και ο έλεγχος υποθέσεων. Περιλαμβάνονται οι ιδιότητες εκτιμητή, το διάστημα εμπιστοσύνης εκτιμητή, το διάστημα εμπιστοσύνης αριθμητικού μέσου όρου, το διάστημα εμπιστοσύνης της διαφοράς δύο αριθμητικών μέσων όρων, ο έλεγχος υποθέσεων αριθμητικού μέσου όρου, ο έλεγχος υποθέσεων περισσότερων των δύο πληθυσμών (ANOVA), ο έλεγχος υπόθεσης κατά έναν παράγοντα (one-way ANOVA), ο έλεγχος υπόθεσης κατά δύο παράγοντες (two-way ANOVA) και ο παραγοντικός έλεγχος υπόθεσης (factorial ANOVA).

Στο ένατο κεφάλαιο περιγράφονται η συσχέτιση και η παλινδρόμηση. Περιλαμβάνει την συνδιακύμανση, την συσχέτιση, την μερική συσχέτιση, την παλινδρόμηση, την απλή γραμμική παλινδρόμηση, τον έλεγχο υπόθεσης των συντελεστών παλινδρόμησης, τον έλεγχο ανάλυσης διακύμανσης για τη διαπίστωση της γραμμικότητας της εξίσωσης, τα μέτρα αξιολόγησης της προσαρμοσμένης εξίσωσης, την πολλαπλή γραμμική παλινδρόμηση, τη μη-γραμμική παλινδρόμηση, την παλινδρόμηση Ridge και την παλινδρόμηση Lasso.

Στο δέκατο κεφάλαιο περιγράφεται η διαχείριση ποιοτικών δεδομένων και περιλαμβάνονται τα κατάλληλα περιγραφικά στατιστικά για κατηγορικές μεταβλητές, ο έλεγχος καλής προσαρμογής (goodness of fit test), ο έλεγχος ανεξαρτησίας (chi-square test for independence/of association), ο έλεγχος ομοιογένειας

(test of homogeneity), η διωνυμική λογιστική παλινδρόμηση (binomial logistic regression), η ανάλυση συστάδων (cluster analysis), η τεχνική *K*-modes (*K*-modes), η ιεραρχική ταξινόμηση (hierarchical clustering), η παραγοντική ανάλυση μικτών μεταβλητών (Factorial Analysis of Mixed Data, FAMD), η εγκυρότητα (validity) και η αξιοπιστία (reliability) ερωτηματολογίου.

Στο ενδέκατο κεφάλαιο περιγράφονται οι μη-παραμετρικοί έλεγχοι και συγκεκριμένα, η δοκιμασία Wilcoxon (Wilcoxon Signed-Rank Test), η δοκιμασία Mann-Whitney *U* (Wilcoxon Rank-Sum Test), η δοκιμασία Kolmogorov-Smirnov (*K*-*S*) για ένα δείγμα, η δοκιμασία Kruskal-Wallis *H* (Kruskal-Wallis H Test) και η δοκιμασία Friedman (Friedman Test).

Για τη συγγραφή αυτού του βιβλίου, η συγγραφέας άντλησε γνώση, εκτός των άλλων βιβλιογραφικών πηγών, και από τα διδακτικά συγγράμματα τα οποία και η ίδια διδάχθηκε, σε αντικείμενα Δασικής Βιομετρίας, του συναδέλφου Ομότιμου Καθηγητή του Εργαστηρίου Δασικής Βιομετρίας, του Τμήματος Δασολογίας και Φυσικού Περιβάλλοντος, ΑΠΘ, κ. Μάτη Κωνσταντίνου, τον οποίο ως καθηγητή μου, θερμά ευχαριστώ.

Επίσης, η συγγραφέας άντλησε γνώση και εμπειρία από την πολυετή διδασκαλία που κλήθηκε να δώσει σε αντικείμενα Δασικής Βιομετρίας και Εφαρμοσμένης Στατιστικής, σε διάφορα Τμήματα της Τριτοβάθμιας εκπαίδευσης, συνειδητοποιώντας κατ' αυτό τον τρόπο τις υφιστάμενες ανάγκες σε παροχή έκτασης και βάθους γνώσης. Επίσης, γνώση αντλήθηκε από τη διαθέσιμη εκτενή βιβλιογραφία καθώς και από τις επιστημονικές δημοσιεύσεις και διδακτικές σημειώσεις που κατά καιρούς συνέγραψε η συγγραφέας, οι οποίες και αναφέρονται στην Βιβλιογραφία του παρόντος βιβλίου.

Στην ολοκλήρωση του βιβλίου αυτού θεωρώ χρέος και χαρά μου να ευχαριστήσω θερμά την οικογένειά μου η οποία μου συμπαραστάθηκε με κατανόηση, υπομονή και εμψύχωση, προκειμένου να ολοκληρωθεί η προσπάθεια.

Το βιβλίο αυτό αποτελεί πρώτη έκδοση. Η προσπάθεια για τη συγγραφή του είναι πιθανό να περιέχει λάθη ή ελλείψεις, οι οποίες προφανώς δεν κατέστη δυνατό να διαπιστωθούν στην παρούσα φάση. Για το λόγο αυτό η παροχή σχετικών υποδείξεων είναι ιδιαιτέρως ευπρόσδεκτη.

Θεσσαλονίκη, 2024
Μαρία Ι. Διαμαντοπούλου,
Επίκουρη Καθηγήτρια Α.Π.Θ.

Περιεχόμενα

Κεφάλαιο 4

Κεφάλαιο 5

Κεφάλαιο 6

Κεφάλαιο 7

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

«Τύχην νόμιζε» (να λαμβάνουμε υπ' όψη «το τυχαίο»)
ΔΕΛΦΙΚΑ ΠΑΡΑΓΓΕΛΜΑΤΑ

«Νουν ηγεμόνα ποιεί» (Το μυαλό να κυριαρχεί στις αποφάσεις)
ΣΟΛΩΝ ο ΑΘΗΝΑΙΟΣ

1.1 Η Στατιστική ως έννοια και επιστήμη διαχρονικά

Ο όρος στατιστική εικάζεται ότι προέρχεται είτε από την αρχαία ελληνική λέξη ίστημι και του εξ αυτού παραγώγου ρήματος στατίζω, που σημαίνει τοποθετώ, ταξινομώ, συμπεραίνω, είτε από την λατινική λέξη status, που σημαίνει πολιτεία (κράτος). Από αρχαιοτάτων χρόνων, ήταν γνωστή η στατιστική, ως τρόπος σκέψης και ενέργειας, χωρίς αυτό να γίνεται συνειδητά, αρκετές προ Χριστού χιλιετίες. Υπάρχουν δε αναφορές συλλογής στοιχείων που παραπέμπουν σε στατιστική από πολλούς φιλοσόφους και συγγραφείς της αρχαιότητας, όπως στον Κομφούκιο (551 - 479 π.Χ.), στον Ξενοφώντα (430 - 355 π.Χ.), Ηρόδοτο (484 - 425 π.Χ.), Αριστοτέλη (384 - 322 π.Χ.). Στη β' ραψωδία της Ιλιάδας του Ομήρου κατά τον $8^{ο}$ - $9^{ο}$ π.Χ. αιώνα διασώζεται ο κατάλογος των πλοίων των Αχαιών που εκστράτευσαν κατά της Τροίας (Τρωικός πόλεμος, $13^{ος}$ - $12^{ος}$ αιώνας π.Χ.). Επίσης, είναι γνωστό το γεγονός της απογραφής του Ιουδαϊκού πληθυσμού την εποχή της γέννησης του Χριστού.

Ως πατέρας της Στατιστικής, αναφέρεται ο Γερμανός Gottfrient Anchenwall (1719-1772), ο οποίος εργαζόμενος ως οικονομολόγος και αργότερα ως διδάσκαλος, για πρώτη φορά εισήγαγε τη λέξη στατιστική (Statistik), προκειμένου να αναφερθεί και να περιγράψει αριθμητικά δεδομένα τα οποία συγκέντρωνε, αλλά ήταν και ο πρώτος που χρησιμοποίησε πίνακες και γραφήματα γι' αυτόν το σκο-

πό. Υπάρχουν θεωρίες αμφισβητήσεων επ' αυτού, οι οποίες θεωρούν ότι αντίστοιχη δουλειά η οποία έγινε από τον Άγγλο William Petty (1623-1687), δεν έτυχε της ίδιας αναγνώρισης, γι' αυτό και δεν αναφέρεται.

Το 1654, ο Γάλλος Blaise Pascal (1623-1662) έθεσε τις βάσεις για τη Συνδυαστική Ανάλυση και το Λογισμό των Πιθανοτήτων. Ο Ελβετός Jacob Bernoulli, (1655-1705) ήταν εκείνος ο οποίος με δημοσίευμά του το οποίο κυκλοφόρησε μετά το θάνατό του από συγγενή του, περιέγραψε την εφαρμογή της θεωρίας πιθανοτήτων σε τυχερά παιχνίδια και πρότεινε το θεώρημα του ισχυρού νόμου των μεγάλων αριθμών.

Στη συνέχεια, άρχισε ο συσχετισμός κάποιων στοιχείων, όπως η σχέση μεταξύ ηλικίας και θανάτου, όπου παρατηρήθηκε η κανονικότητα της θνησιμότητας των διαφόρων ηλικιών. Πρώτος ο Γάλλος Abraham de Moivre, το 1773, έθεσε τη βάση στην ανακάλυψη της κανονικής κατανομής, που αποτελεί τη βάση του μεγαλύτερου μέρους της στατιστικής, η οποία ονομάζεται και κατανομή του Gauss, προς τιμή του Γερμανού Karl Friedrich Gauss (1776-1855), που την ίδια σχεδόν εποχή κατέληξε στα ίδια συμπεράσματα. Λίγο αργότερα, ο Βέλγος A. Quetelet (1796-1874), επηρεαζόμενος από τους Laplace και Fourier, προσπάθησε να ορίσει τον "μέσο άνθρωπο" με βάση σωματομετρικά μεγέθη, δηλ. ουσιαστικά τη μέση τιμή, η οποία σαν έννοια είχε χρησιμοποιηθεί και από τον Αρχιμήδη ως "κέντρο βάρους". Επίσης, ήταν ο πρώτος που αντιμετώπισε την κανονική κατανομή, ως το νόμο των σφαλμάτων και παράλληλα το 1853 οργάνωσε το πρώτο διεθνές στατιστικό συνέδριο, του οποίου η συμβολή τόσο στη στατιστική θεωρία όσο και στην εφαρμογή ήταν εξαιρετικά σημαντική.

Αργότερα, η ωριμότητα της σκέψης των ανθρώπων που ασχολήθηκαν με την επιστήμη της στατιστικής, έδωσε φοβερή ώθηση σ' αυτό τον κλάδο της επιστήμης. Ενδεικτικά, ο Άγγλος Francis Galton (1822-1911) ήταν αυτός που έθεσε τις βάσεις της εφαρμογής της στατιστικής επιστήμης στα έμβια όντα. Λίγο αργότερα, ο Άγγλος Karl (Carl) Pearson (1857-1936), επηρεασμένος από τις μελέτες του Francis Galton, ήταν αυτός ο οποίος έθεσε τις βάσεις των στατιστικών μεθόδων που εδράζονται στη μέθοδο των ελαχίστων τετραγώνων, όπως είναι η θεωρία της παλινδρόμησης. Τέλος, ο Βρετανός Sir Ronald Aylmer Fisher (1890-1962), έθεσε τις βάσεις της θεωρίας του σχεδιασμού πειραμάτων σε αγροτικές καλλιέργειες.

Όσον αφορά την κατάρτιση σε θέματα Εφαρμοσμένης Στατιστικής των φοιτητριών/φοιτητών του Τμήματος Δασολογίας και Φυσικού Περιβάλλοντος, σύμφωνα με όσα αναφέρει ο Μάτης (2003), στο Τμήμα Δασολογίας και Φυσικού Περι-

βάλλοντος του Α.Π.Θ. πριν το ακαδημαϊκό έτος 1971-72 δεν διδασκόταν το μάθημα της Στατιστικής ως ξεχωριστό μάθημα. Στοιχεία Εφαρμοσμένης Στατιστικής διδασκόταν εντός της ύλης του μαθήματος της Δεντρομετρίας. Στη συνέχεια, για τα επόμενα δύο χρόνια η Στατιστική διδασκόταν ως μάθημα επιλογής στο δεύτερο έτος σπουδών. Από το ακαδημαϊκό έτος 1974-75 συμπεριλήφθηκε στο πρόγραμμα σπουδών του Τμήματος ως υποχρεωτικό μάθημα, διδασκόμενο στο πρώτο εξάμηνο του δεύτερου έτους σπουδών, ως Δασική Βιομετρία Ι, μέχρι και την τελευταία αναθεώρηση προπτυχιακού προγράμματος σπουδών.

Από το πανεπιστημιακό έτος 2021-2022 για τις/τους φοιτήτριες/φοιτητές με έτος εισαγωγής το έτος 2020, ο τίτλος του μαθήματος τροποποιήθηκε και έγινε Εφαρμοσμένη Στατιστική και διδάσκεται ως υποχρεωτικό μάθημα στο 1ο εξάμηνο του πρώτου έτους σπουδών του Τμήματος.

1.2 Η Δασική Βιομετρία ως κλάδος της εφαρμοσμένης Στατιστικής σήμερα

Σε γενικές γραμμές μπορεί να λεχθεί ότι **εφαρμοσμένη στατιστική** είναι εκείνος ο κλάδος της Στατιστικής επιστήμης, που ασχολείται με εμπειρικά δεδομένα τα οποία μπορεί να προέρχονται είτε από μετρήσεις/παρατηρήσεις γεγονότων που συμβαίνουν στον περιβάλλοντα χώρο είτε από καταγραφή αποτελεσμάτων πειραμάτων, με σκοπό να εξάγει βάσιμα επιστημονικά συμπεράσματα, λαμβάνοντας υπόψη το στοιχείο αβεβαιότητας που μπορεί να περικλείουν αυτά τα δεδομενα. Η ιδιαιτερότητά της ως κλάδος της γενικότερης μαθηματικής επιστήμης, έγκειται στο γεγονός ότι ενώ η γλώσσα των αριθμών εμπεριέχει έναν χαρακτήρα απόλυτης ακρίβειας, η εφαρμοσμένη στατιστική περιλαμβάνει ανακρίβειες, οι οποίες οφείλονται στα ίδια αυτά τα μεγέθη που μελετά, υπό περιορισμούς και οριοθετήσεις μέσω στατιστικών και μαθηματικών νόμων. Αποτελεί ως επιστήμη, τον καλύτερο δυνατό τρόπο με τον οποίο μπορούν να αναλυθούν και να αντιμετωπιστούν όλες εκείνες οι καταστάσεις όπου τα «καθαρά» μαθηματικά, εξαιτίας του απόλυτου των αριθμητικών μεγεθών τους, δεν είναι ευέλικτα. Έτσι λοιπόν η εφαρμοσμένη στατιστική, ως απαραίτητο και μοναδικό εργαλείο, έχει πάρα πολλά πεδία εφαρμογής και ανάλογα με την γνωστική περιοχή στην οποία εφαρμόζεται, παίρνει και το αντίστοιχο χαρακτηριστικό όνομα καθώς και τις αντίστοιχες θεωρητικές και πρακτικές εξειδικεύσεις που οφείλονται στη φύση του εκάστοτε γνωστικού αντικειμένου.

Η **Δασική Στατιστική**, ως κλάδος της εφαρμοσμένης Στατιστικής είναι η επιστήμη της μέτρησης (ποσοτικοποίησης) των χαρακτηριστικών (μεταβλητών) του δασικού οικοσυστήματος. Ο Prodan (1968) έδωσε τον ορισμό της Δασικής Βιομετρίας, ως το πεδίο της δασικής επιστήμης το οποίο περιλαμβάνει τις μεθόδους μαθηματικής στατιστικής και βιομετρίας που είναι σημαντικές για το δασικό οικοσύστημα. Δηλαδή, η Δασική Βιομετρία περιλαμβάνει την ποσοτικοποίηση, μέσα σε κάποιο εύλογο επίπεδο ακρίβειας και συνέπειας, των βιολογικών και φυσικών χαρακτηριστικών ενός δασικού οικοσυστήματος, μέσω της συγκέντρωσης, της παρουσίασης, της επεξεργασίας, της ανάλυσης και ερμηνείας αυτών των δεδομένων, προκειμένου να βρεθούν λύσεις ή απαντήσεις σε αναζητήσεις της δασικής επιστήμης. Ο ορισμός ο οποίος δόθηκε από τον M. Prodan υπογραμμίζει ότι η δασική Βιομετρία ασχολείται θεμελιωδώς με την εφαρμογή ποσοτικών μεθόδων για την κατανόηση και την περιγραφή των δασικών οικοσυστημάτων. Επίσης, σύμφωνα με τον ορισμό που δίνεται από το Ερευνητικό Ινστιτούτο Δασικής Στατιστικής του Portland, USA, η Δασική Στατιστική/Βιομετρία είναι η επιστήμη της μέτρησης (-μετρία) των δασών (βιο-). Περιλαμβάνει τον ποσοτικό προσδιορισμό των βιολογικών και φυσικών χαρακτηριστικών των δέντρων και της σχετικής βλάστησης, των εντόμων, των ασθενειών, της άγριας ζωής, της τοπογραφίας, των εδαφών και του κλίματος, μεμονωμένα και συλλογικά. Τα χαρακτηριστικά αυτά περιλαμβάνουν όλα τα ποσοτικώς προσδιορίσιμα χαρακτηριστικά εντός του δάσους, τόσο χρονικά όσο και χωρικά.

Η Δασική Στατιστική γενικά αποτελεί απαραίτητο εργαλείο για να αναλυθεί και να κριθεί μια κατάσταση, καθώς και να προβλεφθεί η πορεία κάποιου φαινομένου στο μέλλον. Πολλές φορές, κάποιες καταστάσεις δεν οδηγούν σε αυτονόητα ή αυταπόδεικτα συμπεράσματα, πχ. ποιός είναι ο συνολικός ξυλώδης όγκος που μπορεί ένα δασικό οικοσύστημα να παράγει, έτσι ώστε να γίνει προγραμματισμός ανάλογων υλοτομιών ή ποιό είναι το ποσοστό φυτρωτικότητας κάποιων σπόρων, κλπ. Σ' αυτή την περίπτωση, αλλά και σε πάρα πολλές άλλες, η χρήση της Στατιστικής, μπορεί να κρίνει την κατάσταση και να οδηγήσει ασφαλώς σε συμπεράσματα και αποφάσεις. Επίσης, τις περισσότερες φορές καλούμαστε να περιγράψουμε και να αναλύσουμε πληθυσμούς των οποίων τα μέλη είναι τόσα πολλά πχ. όλα τα δέντρα που απαρτίζουν ένα δασικό οικοσύστημα, ώστε να είναι αδύνατο να πάρουμε τις απαραίτητες πληροφορίες απ' όλα. Και σ' αυτές τις περιπτώσεις, η εφαρμογή των μεθόδων της Στατιστικής, δίνει τη δυνατότητα λύσεων. Κατ' αυτό τον τρόπο, η Δασική Στατιστική, εφαρμόζει τις αρχές και μεθόδους της εφαρμοσμένης στατιστικής χρησιμοποιώντας στοιχεία που αφορούν το δασικό οικοσύστημα. Περιλαμβάνει δε τρία βασικά μεθοδολογικά κεφάλαια. Το πρώ-

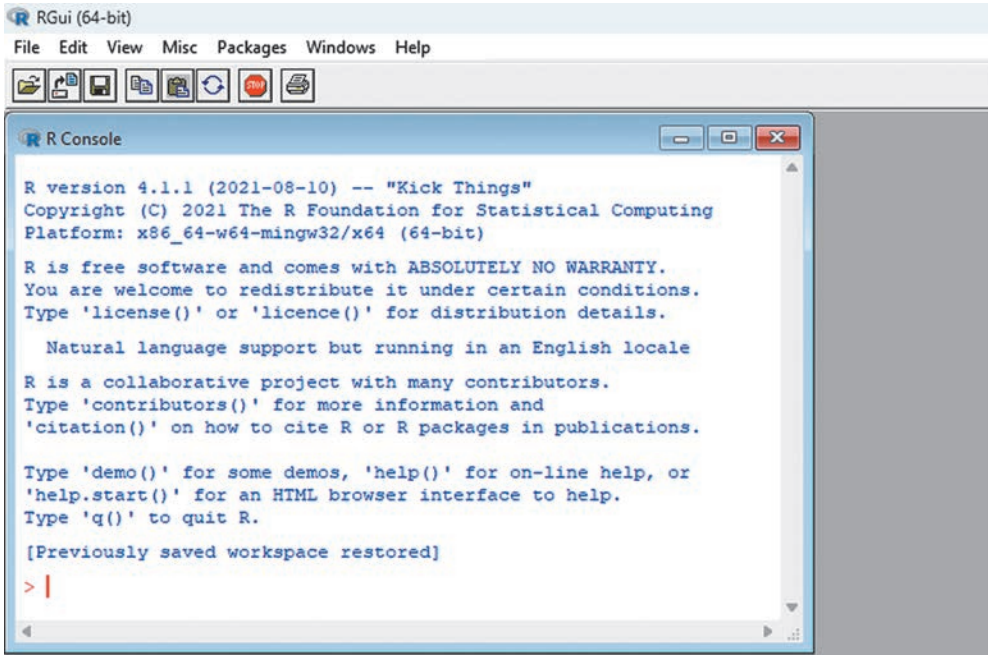
το, αφορά το σχεδιασμό και την εφαρμογή της διαδικασίας της συλλογής των πρωτογενών στοιχείων, είτε στο δασικό οικοσύστημα (πεδίο), είτε στο εργαστήριο μέσω εφαρμογής των αρχών του κλάδου του σχεδιασμού πειραμάτων. Η λήψη αυτή των πρωτογενών στοιχείων, όταν πρόκειται για λήψη στοιχείων δείγματος, ακολουθεί επιπλέον τις μεθοδολογίες και τις αρχές που αναπτύσσονται στον κλάδο της δειγματοληψίας. Το δεύτερο, αφορά την περιγραφή και διερευνητική ανάλυση των δεδομένων, μέσω κυρίως περιγραφικών στατιστικών μεθόδων, ενώ το τρίτο μεθοδολογικό κεφάλαιο αφορά την ανάλυση των δεδομένων μέσω εφαρμογής στατιστικών τεχνικών και διαδικασιών, οι οποίες οδηγούν τέλος στην στατιστική συμπερασματολογία.

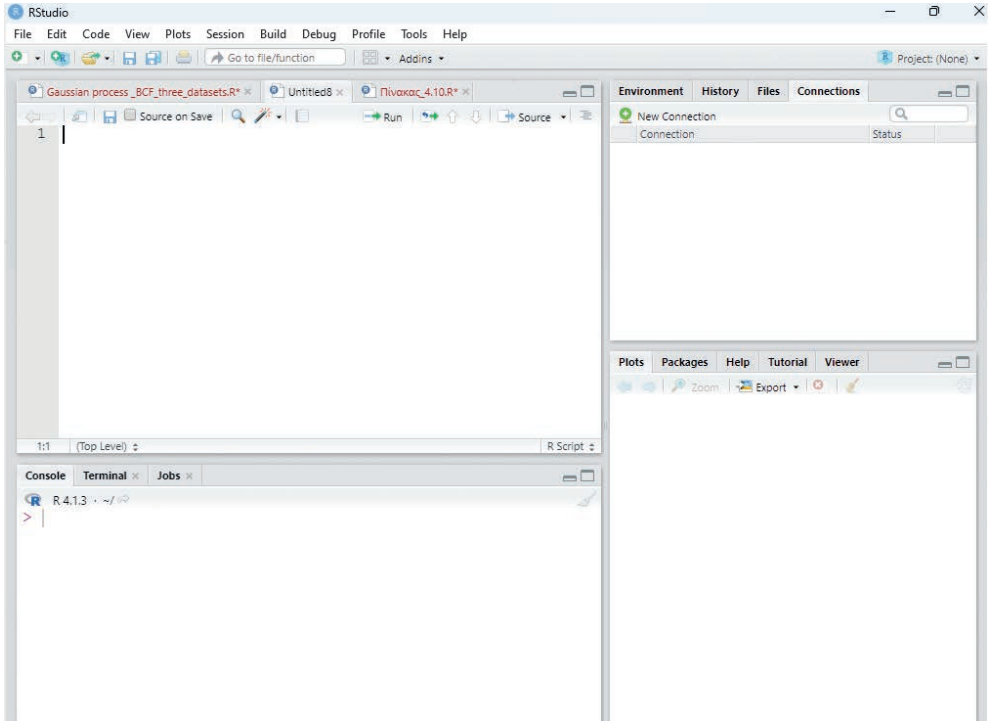
Η γνώση της εφαρμογής των μεθόδων της εφαρμοσμένης στατιστικής στη δασική έρευνα είναι απαραίτητο να συνοδεύεται από τουλάχιστον τη βασική γνώση προγραμματισμού και γνώση χρήσης στατιστικών πακέτων (statistical softwares), προκειμένου να είναι δυνατή η επεξεργασία και ανάλυση πολλών αριθμητικά δεδομένων, αλλά και να υπάρχει η δυνατότητα εφαρμογής προηγμένων στατιστικών μεθόδων και τεχνικών, οι οποίες μπορούν να οδηγήσουν σε τελικά ασφαλή συμπεράσματα. Στα πλαίσια αυτού του βιβλίου, θα δοθούν στοιχεία χρήσης της γλώσσας προγραμματισμού R και η εφαρμογή της στην επίλυση στατιστικών διαδικασιών, οι οποίες θα αναπτυχθούν στα επόμενα κεφάλαια, καθώς και η αντίστοιχη χρήση του στατιστικού πακέτου IBM-SPSS (από εδώ και στο εξής η αναφορά στο πακέτο θα γίνεται ως SPSS) για την επίλυση στατιστικών προβλημάτων, με χρήση δασικών δεδομένων.

1.3 Εισαγωγή στη στατιστική γλώσσα προγραμματισμού R

Η γλώσσα προγραμματισμού R, αποτελεί ένα ελεύθερο λογισμικό, το οποίο διατίθεται για διάφορα λειτουργικά συστήματα (Windows, Linux, κλπ) από την ηλεκτρονική διεύθυνση http://www.r-project.org/. Στο συγκεκριμένο site δίνονται όλες οι απαραίτητες οδηγίες εγκατάστασης. Επίσης, η R μπορεί να αποκτηθεί και από ένα άλλο δίκτυο διανομής, το οποίο είναι ο πρότυπος καθρέφτης (mirror) του CRAN (Comprehensive R Archive), όπου η διαδικτυακή του διεύθυνση είναι http://cran.r-project.org.

Είναι δυνατό να “δουλέψει” στη βασική της έκδοση, αλλά και μέσα από το περιβάλλον RStudio (Εικόνα 1.1), το οποίο διανέμεται ελεύθερα από τη διεύθυνση http://www.rstudio.org.





Εικόνα 1.1. Περιβάλλον εργασίας της R α) σε βασική έκδοση και β) σε περιβάλλον RStudio.

Λεπτομερής ανάπτυξη της χρήσης της R, της σύνταξης των εντολών και της δημιουργίας γραφημάτων μπορεί ο αναγνώστης να βρει στην ιστοσελίδα της R και συγκεκριμένα στη διεύθυνση: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf. Η εκμάθηση της γλώσσας R δεν αποτελεί αντικείμενο και σκοπό του βιβλίου αυτού. Στο πλαίσιο των στατιστικών διαδικασιών οι οποίες αναπτύσσονται και αποτελούν το βασικό αντικείμενο του βιβλίου, θα δίνονται τμήματα κώδικα της R, έτσι ώστε να είναι εύκολη η υλοποίηση των στατιστικών διαδικασιών και η εξαγωγή των σχετικών αποτελεσμάτων. Παρ' όλα αυτά θα αναφερθούν κάποια πολύ βασικά στοιχεία της χρήσης της R, προκειμένου να είναι σε θέση ο αναγνώστης, ακόμη και αν δεν γνωρίζει προγραμματισμό στην R, να χρησιμοποιήσει τον κώδικα και να πάρει τα σχετικά αποτελέσματα.

Όπως φαίνεται στην Εικόνα 1.1, το βασικό περιβάλλον εργασίας (Εικόνα 1,α) δίνει ως βασικό μενού επιλογών (πρώτη μπάρα επιλογών) τα: File, Edit, View, Misc, Packages, Windows και Help, ενώ το περιβάλλον εργασίας RStudio (Εικόνα 1,β) δίνει ως βασικό μενού επιλογών (πρώτη μπάρα επιλογών) τα: File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools και Help. Επιπλέον, υπάρχει η δυνατότητα διαίρεσης της κεντρικής οθόνης σε 4 τμήματα, γεγονός το οποίο διευκολύνει την εγγραφή εντολών κώδικα, την παρακολούθηση της «εκτέλεσής» τους και την εμφάνιση των αποτελεσμάτων της «εκτέλεσης» του κώδικα. Γι' αυτό το λόγο, προτείνεται η χρήση της R μέσα από το περιβάλλον του Rstudio, ως ευκολότερη και σε μεγαλύτερο βαθμό κατανοητή η χρήση της, όσον αφορά τουλάχιστον τους μη έμπειρους χρήστες. Στη συνέχεια, η περιγραφή και ανάπτυξη του κώδικα και των διαδικασιών που θα υλοποιούνται με τη γλώσσα προγραμματισμού R, θα αφορούν το περιβάλλον RStudio. Τα αρχεία εντολών προγραμματισμού τα οποία δημιουργούνται στην R (scripts), είναι της μορφής όνομα_αρχείου.R

Όπως φαίνεται στο κάτω αριστερό (όπως το βλέπει κάποιος στην οθόνη) ορθογώνιο του RStudio υπάρχει η κονσόλα στην οποία εγγράφονται εντολές μετά από το «>». Εφόσον ολοκληρωθεί η εγγραφή τους, πατώντας «enter», εκτελείται η εντολή και φαίνεται το αποτέλεσμά της. Οι βασικές αριθμητικές πράξεις εγγράφονται σύμφωνα με τον Πίνακα 1.1 με το πάτημα του πλήκτρου «enter». Επίσης κατά τον ίδιο τρόπο (Πίνακας 1.1) εγγράφονται και οι βασικές λογικές εκφράσεις οι οποίες έχουν ως αποτέλεσμα έκφραση αλήθειας (TRUE) ή λάθους (FALSE).

Πίνακας 2.3. Στάδιο 3ο: *Τμήμα περιεχομένου αρχείου αποτελεσμάτων.*

Ένδειξη οθόνης αποτελεσμάτων

File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help

utput
- Log
- Explore
  - Title
  - Notes
  - Case Processing Summary
  - Descriptives
  - M-Estimators
  - Percentiles
  - Extreme Values
  - Tests of Normality
  - diameter at breast height
    - Title
    - Histogram
    - Stem-and-Leaf Plot
    - Normal Q-Q Plot
    - Detrended Normal Q-Q
    - Boxplot

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
diameter at breast height	532	100.0%	0	0.0%	532	100.0%

Descriptives

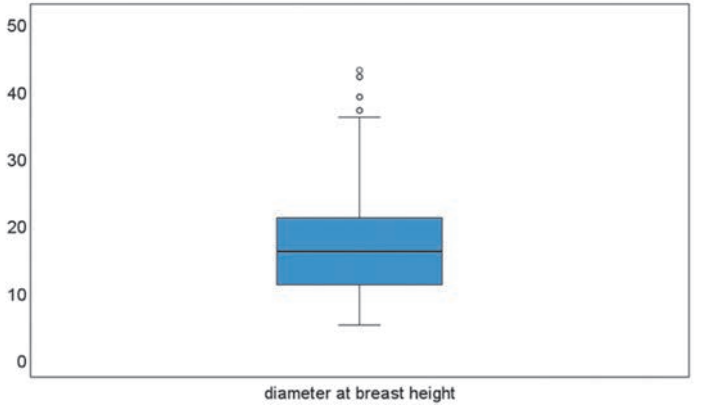
			Statistic	Std. Error
diameter at breast height	Mean		16.79511	.325198
	95% Confidence Interval for Mean	Lower Bound	16.15628	
		Upper Bound	17.43395	
	5% Trimmed Mean		16.41228	
	Median		16.00000	
	Variance		56.261	
	Std. Deviation		7.500743	
	Minimum		5.000	
	Maximum		43.000	
	Range		38.000	
	Interquartile Range		10.000	
	Skewness		.726	.106
	Kurtosis		.222	.211

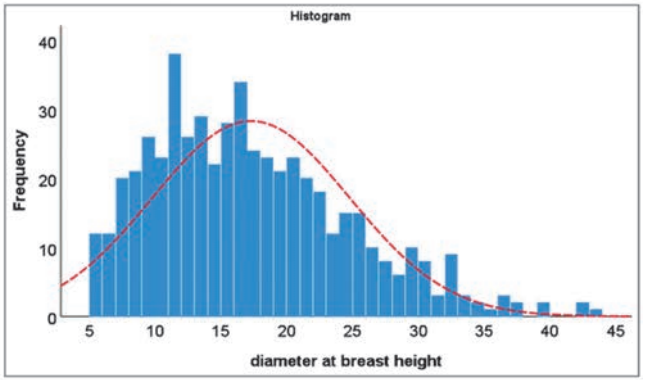
```
diameter at breast height Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    12.00        0 .  555555555555
    32.00        0 .  66666666666677777777777777777777
    47.00        0 .  88888888888888888888899999999999999999999999999
    61.00        1 .  00000000000000000000000011111111111111111111111111111111111
    55.00        1 .  22222222222222222222222233333333333333333333333333333
    50.00        1 .  44444444444444444444455555555555555555555555555555
    58.00        1 .  66666666666666666666666666666666667777777777777777777777777
    44.00        1 .  88888888888888888888888999999999999999999999
    43.00        2 .  00000000000000000000000111111111111111111111
    30.00        2 .  22222222222222222233333333333
    30.00        2 .  44444444444444455555555555555
    18.00        2 .  66666666667777777
    16.00        2 .  8888889999999999
    11.00        3 .  00000000111
    12.00        3 .  222222222333
     3.00        3 .  445
     3.00        3 .  666
     7.00 Extremes    (>=37)

 Stem width:     10.000
 Each leaf:       1 case(s)
```





5. Ενσωματώνεται στο γράφημα η γραμμή των 45°, η οποία αντιπροσωπεύει την πλήρη εξίσωση των τιμών των ζευγών (x,y), η οποία αποτελεί την ιδανική κατάσταση.

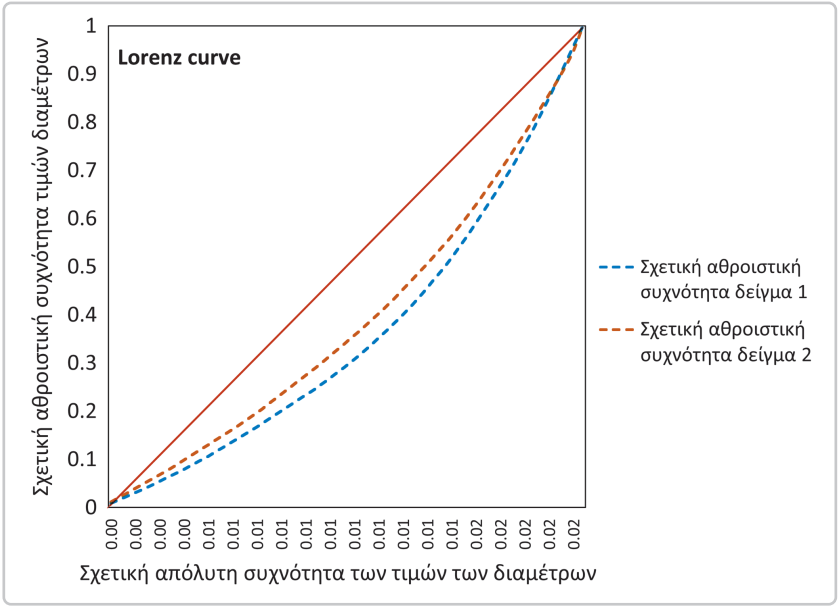
Όσο πιο μακριά βρίσκεται η καμπύλη του Lorenz των δεδομένων από την ιδανική κατάσταση, τόσο οι ανισότητες είναι μεγαλύτερες και η διασπορά μεγαλύτερη. Το συγκεκριμένο γράφημα είναι ιδανικό προκειμένου να συγκριθούν δύο ή και περισσότερες κατανομές συχνοτήτων μεταξύ τους. Η καμπύλη του Lorenz η οποία απέχει περισσότερο από την ιδανική κατάσταση, έχει τις μεγαλύτερες ανισότητες και τη μεγαλύτερη διακύμανση τιμών.

Παράδειγμα 5.10

Για τις διαμέτρους του **Παραδείγματος 5.9** να γίνει γραφικός έλεγχος με την καμπύλη του Lorenz. Ποια από τις δύο εμπειρικές κατανομές εμφανίζει μεγαλύτερες ανισότητες και τελικά ποιο από τα δύο δείγματα είναι πιο ομοιογενές;

Απάντηση

Σύμφωνα με όσα αναπτύχθηκαν, δημιουργούνται οι καμπύλες του Lorenz για τα δύο δείγματα στο ίδιο γράφημα, προκειμένου να είναι δυνατή η σύγκρισή τους (Σχήμα 5.4).



Σχήμα 5.4. Η καμπύλη του Lorenz για τη σύγκριση των δύο εμπειρικών κατανομών συχνοτήτων των δειγμάτων 1 και 2.

Όπως φαίνεται στο Σχήμα 5.4, η εμπειρική κατανομή του δείγματος 1 εμφανίζει τη μεγαλύτερη απόκλιση από την ευθεία 45° γραμμή, σε σχέση με την αντίστοιχη κατανομή του δείγματος 2. Άρα επιβεβαιώνεται το συμπέρασμα το οποίο έχει εξαχθεί μέσω του υπολογισμού των συντελεστών κύμανσης, ότι δηλαδή το δείγμα 1 εμφανίζει τη μεγαλύτερη μεταβλητότητα.

Το γράφημα της καμπύλης προσφέρεται και από τις βιβλιοθήκες της R. Το σχετικό script σε γλώσσα προγραμματισμού R και το παραγόμενο αποτέλεσμα για το πρώτο δείγμα των τιμών διαμέτρων (δείγμα 1), χρησιμοποιώντας μόνο τις σχετικές αθροιστικές συχνότητες δίνεται στον Πίνακα 5.14. Όπως μπορεί κανείς να παρατηρήσει από το Σχήμα 5.4 και το παραγόμενο αποτέλεσμα του Πίνακα 5.14, είτε χρησιμοποιηθούν στον άξονα των x'x οι τιμές της σχετικής απόλυτης συχνότητας είτε οι τιμές της σχετικής αθροιστικής συχνότητας, το οπτικό αποτέλεσμα παραμένει παρόμοιο.

Πίνακας 5.14. *Γράφημα της καμπύλης Lorenz στην R.*

Κώδικας R (R-script)

```
install.packages("ineq")
library(ineq)
# Create data,
x<-c(12,10,812,10,8,9,10,8,8,9,12,8,12,11,12,11,8,12,9,12,11,8,……,
    12,14,21,21,10,9)
# Lorenz Curve steps
# Sort the data in ascending order
sorted_data <- sort(x)
# Calculate the cumulative proportions
cumulative_proportions <- cumsum(sorted_data) / sum(sorted_data)
# Create a vector of cumulative proportions
cumulative_proportions_vector <- c(0, cumulative_proportions)
# Create a vector of the perfect equality line
perfect_equality_line <- seq(0, 1, length.out =
    length(cumulative_proportions_vector))
# Plot the Lorenz curve
plot(perfect_equality_line, cumulative_proportions_vector, type = "l",
    xlab = "Cumulative Proportion", ylab = "Cumulative Proportion",
    main = "Lorenz Curve")
# Add a diagonal reference line (perfect equality)
abline(0, 1, col = "red", lty = 2)
```

```
#Παράδειγμα 7.20                    > #Παράδειγμα 7.20
dbinom(2, size=10, prob=0.4)        > dbinom(2, size=10, prob=0.4)
                                    [1] 0.1209324
dbinom(7, size=10, prob=0.4)        > dbinom(7, size=10, prob=0.4)
                                    [1] 0.04246733
                                    >
#Παράδειγμα 7.21                    > #Παράδειγμα 7.21
dbinom(40, size=90, prob=0.33)      > dbinom(40, size=90, prob=0.33)
                                    [1] 0.0066300010
```

7.11 Κατανομή Poisson (Poisson distribution)

Η **κατανομή Poisson** (Poisson distribution), αποτελεί μια πολύ σημαντική διακριτή κατανομή, η οποία βρίσκει εφαρμογή σε πολλά φυσικά φαινόμενα. Ονομάζεται αλλιώς και ως κατανομή των σπάνιων γεγονότων, επειδή η εμφάνιση των γεγονότων είναι πιθανότερη στις μικρές τιμές της μεταβλητής *X*, ενώ οι μεγαλύτερες τιμές της μεταβλητής είναι σπάνιο να εμφανιστούν. Γενικά, φαινόμενα των οποίων η πιθανότητα εμφάνισής τους περιγράφεται από την κατανομή αυτή, είναι εκείνα τα οποία συμβαίνουν (εμφάνιση) σε ένα σταθερό διάστημα χρόνου ή/και χώρου, συμβαίνουν με ένα γνωστό μέσο ρυθμό, ενώ η εμφάνιση του επόμενου γεγονότος, είναι ανεξάρτητη από το διάστημα εμφάνισης του προηγούμενου. Παράδειγμα τέτοιων γεγονότων μπορεί να είναι ο αριθμός δασικών πυρκαγιών σε μια περιοχή, μέσα σε ορισμένο χρονικό διάστημα, ή ο αριθμός των τηλεφωνικών κλήσεων που φθάνουν σε ένα δασαρχείο, κλπ.

Η κατανομή Poisson, αποτελεί μια ειδική περίπτωση διωνυμικής κατανομής, η οποία εμφανίζεται όταν συμβαίνουν ταυτόχρονα οι εξής προϋποθέσεις: α) η πιθανότητα επιτυχίας του πειράματος είναι μικρότερη από 10% και β) ο μέσος όρος της κατανομής βρίσκεται μεταξύ του 0 και του 10. Τότε η πιθανότητα εμφάνισης του γεγονότος ακολουθεί την κατανομή Poisson, με μέσο όρο *λ*. Όταν μια μεταβλητή *X* ακολουθεί την κατανομή Poisson, με μέσο όρο *λ*, τότε συμβολίζεται ως $X \sim P(\lambda)$ και η ποσότητα *λ* είναι η παράμετρος της κατανομής. Ο αριθμητικός μέσος της κατανομής είναι ίσος με

$$\lambda = n \cdot p,$$

ενώ η διακύμανσή της ισούται με

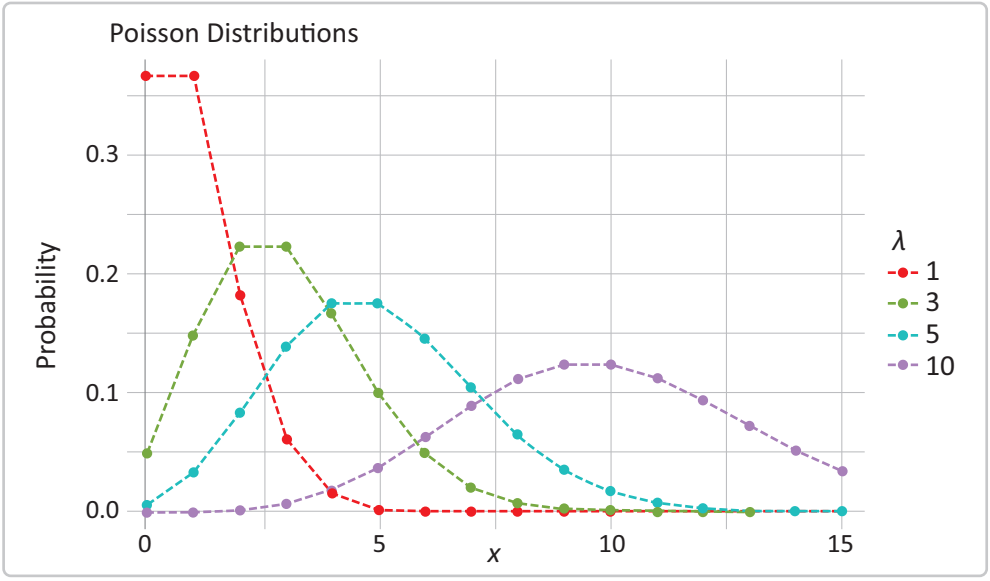
$$\sigma^2 = \lambda\ .$$

Δηλαδή, ο μέσος όρος της κατανομής είναι ίσος με την τιμή της διασποράς της.

Επειδή η τιμή της διακύμανσης μπορεί να διαφοροποιηθεί εξαιτίας της διακύμανσης της δειγματοληψίας, θεωρούμε ότι αν $\sigma^2 = \lambda \pm 0{,}2 \cdot \lambda$, τότε ισχύει κατά προσέγγιση η ισότητα διακύμανσης με το μέσο όρο της κατανομής. Η συνάρτηση πιθανότητας, η οποία περιγράφει την κατανομή Poisson για τη μεταβλητή $X \sim P(\lambda)$, η οποία παίρνει την τιμή *x*, είναι:

$$P(X = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!} \tag{7.30}$$

Η γραφική παράσταση διαφόρων κατανομών Poisson, με διαφορετικούς μέσους όρους (*λ*), δίνεται στο Σχήμα 7.4, όπου η ένωση των σημείων με διακεκκομένη γραμμή έχει γίνει μόνο για λόγους καλύτερης κατανόησης των διαφορετικών μορφών που παίρνει η κατανομή, για διαφορετικές τιμές του μέσου όρου της και για τιμές της μεταβλητής *X* από 0 έως 15. Πρόκειται για μία διακριτή κατανομή και οι τιμές μεταξύ των σημείων δεν έχουν νόημα.



Σχήμα 7.4. Γραφική παράσταση κατανομής Poisson, για διαφορετικές τιμές λ.

Για τον υπολογισμό των πιθανοτήτων της κατανομής Poisson, όπως και στην περίπτωση της διωνυμικής κατανομής, μπορούν να δημιουργηθούν πίνακες είτε απλοί, είτε αθροιστικής κατανομής. Προκειμένου να δημιουργηθεί πίνακας απλών πιθανοτήτων της κατανομής Poisson, έτσι ώστε η τυχαία μεταβλητή *X* να πάρει κάποια συγκεκριμένη τιμή, η οποία εκφράζει την πιθανότητα επιτυχίας, αρκεί να χρησιμοποιηθεί η σχέση (7.30). Η εύρεση της αθροιστικής πιθανότητας,

7.13 Κανονική και τυπική κανονική κατανομή (normal and standard normal distribution)

Η **κανονική κατανομή** θεωρείται ως η σπουδαιότερη συνεχής κατανομή πιθανοτήτων, γιατί βρίσκει πάρα πολλές εφαρμογές σε προβλήματα της πράξης, προσεγγίζοντας πολλά φυσικά φαινόμενα.

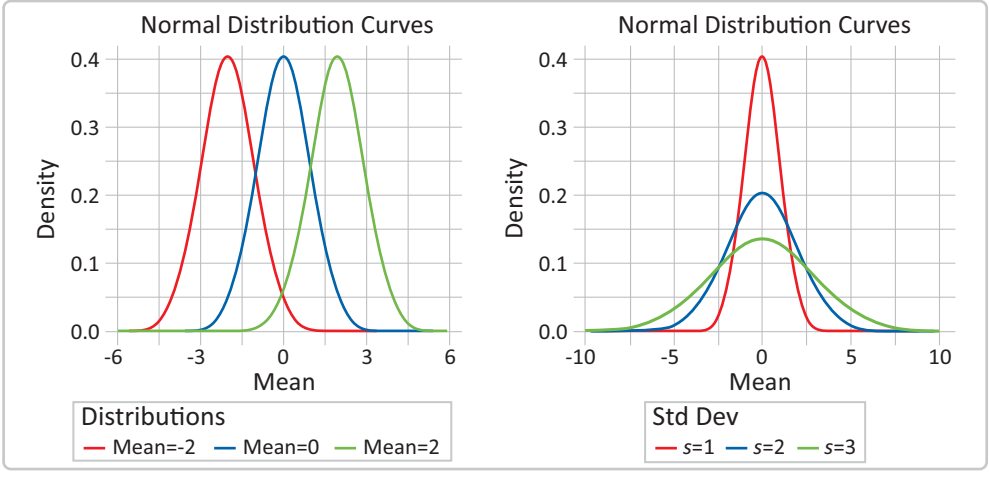
Πολλοί ερευνητές ασχολήθηκαν με την κανονική κατανομή. Πρώτος τη διατύπωσε ο μαθηματικός De Moivre γύρω στο 1733, ο οποίος διαπίστωσε ότι στη διωνυμική κατανομή, καθώς το μέγεθος του δείγματος τείνει στο άπειρο, το διωνυμικό ανάπτυγμα $(p+q)^n$ τείνει να προσεγγίσει την κανονική κατανομή, ενώ αργότερα ξαναμελετήθηκε από τους Gauss και Laplace γύρω στο 1812, ως κατανομή των σφαλμάτων.

Πρόκειται για μια συμμετρική κατανομή, κωδωνοειδούς μορφής, με ασύμπτωτο τον άξονα των τετμημένων, η οποία έχει δύο παραμέτρους και συμβολίζεται με: $X \sim N(\mu, \sigma^2)$. Δηλαδή, η τυχαία μεταβλητή X η οποία είναι συνεχής, ακολουθεί την κανονική κατανομή με μέσο όρο μ και διασπορά σ^2.

Η συνάρτηση πυκνότητας-πιθανότητας της κατανομής δίνεται από την σχέση:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-0{,}5 \cdot \left(\frac{x-\mu}{\sigma}\right)^2} \tag{7.32}$$

Στο Σχήμα 7.7, δίνεται η γραφική παράσταση διαφορετικών κανονικών κατανομών, οι οποίες έχουν διαφορετικούς μέσους όρους και ίδια τιμή διακύμανσης (Σχήμα 7.7α) και διαφορετικές τιμές διακύμανσης και ίσους μέσους όρους (Σχήμα 7.7β).

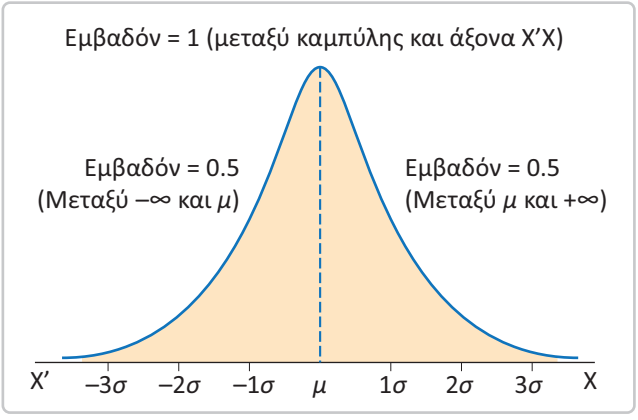


Σχήμα 7.7. Διαφορετικές μορφές κανονικών κατανομών.

Όπως διαπιστώνεται, η κανονική κατανομή ανεξάρτητα από την τιμή του μέσου όρου της και της διακύμανσής της είναι κωδωνοειδούς μορφής, συμμετρική γύρω από το μέσο όρο της. Η διαφορετική τιμή μέσου όρου χαρακτηρίζει τη θέση της κατανομής ως προς τον άξονα των *X'X*, ενώ η διαφορετική τιμή της διακύμανσης της κατανομής, άρα και της τυπικής απόκλισης, καθορίζει τη μορφή της κατανομής.

Θεωρείται ότι το εμβαδόν μεταξύ της συμμετρικής κωδωνοειδούς καμπύλης της κανονικής κατανομής και του άξονα *X'X*, είναι ίσο με ένα (Σχήμα 7.8), το οποίο μοιράζεται εξίσου εκατέρωθεν του κάθετου άξονα στον *X'X*, ο οποίος περνάει από το μέσο όρο της κατανομής. Αυτό μαθηματικά εκφράζεται με το αόριστο ολοκλήρωμα της συνάρτησης πυκνότητας πιθανότητας:

$$\int_{-\infty}^{+\infty} f(x)d(x) = \int_{-\infty}^{+\infty} \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-0,5 \cdot \left(\frac{x-\mu}{\sigma}\right)^2} \, \mathrm{d}(x) = 1 \qquad (7.33)$$



Σχήμα 7.8. Εμβαδόν μεταξύ κανονικής κατανομής και άξονα x'x.

Οι αθροιστικές πιθανότητες της κατανομής *F(x)*, δηλαδή η πιθανότητα η τυχαία μεταβλητή $X \sim N(\mu, \sigma^2)$ να πάρει τιμές σε εύρος μικρότερο από μία συγκεκριμένη τιμή x, δηλ. $P(X \le x)$, αποδίδεται από τη γραφική παράσταση της αθροιστικής συνάρτησης της κανονικής κατανομής, η οποία φαίνεται στο Σχήμα 7.9. Στο Σχήμα 7.9α φαίνεται η γραφική παράσταση αθροιστικών κανονικών κατανομών με διαφορετικούς μέσους όρους, αλλά ίση διακύμανση, ενώ στο Σχήμα 7.9β φαίνεται η γραφική παράσταση αθροιστικών κανονικών κατανομών με ίσους μέσους όρους, αλλά διαφορετικές διακυμάνσεις.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**1.892500**	**0.092774**	**20.40**	9.02e-07 ***
X	0.075167	0.003674	20.46	8.87e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1191 on 6 degrees of freedom
Multiple R-squared: 0.9859, Adjusted R-squared: 0.9835
F-statistic: 418.5 on 1 and 6 DF, p-value: 8.873e-07

Από τη γραφική παράσταση του παραγόμενου αποτελέσματος του Πίνακα 9.9, φαίνεται η άριστη προσαρμογή της εξίσωσης ευθείας γραμμής στα μετρημένα δεδομένα, τα οποία αναπαρίστανται με στικτό διάγραμμα.

9.3.2 Πολλαπλή γραμμική παλινδρόμηση (multiple linear regression, MLR)

Η πολλαπλή γραμμική παλινδρόμηση αποτελεί μια προέκταση της απλής γραμμικής παλινδρόμησης. Δηλαδή επιλύει τα γραμμικά εκείνα μοντέλα, τα οποία έχουν περισσότερες από μία ανεξάρτητες μεταβλητές. Και σ' αυτή την περίπτωση, όπως και στην απλή γραμμική παλινδρόμηση, υφίστανται οι περιορισμοί της ύψωσης σε πρώτη δύναμη των ανεξάρτητων μεταβλητών όπως επίσης και ο τρόπος σύνδεσης των συντελεστών. Εδώ μπορούν να συμπεριληφθούν και τα μοντέλα εκείνα τα οποία δεν είναι γραμμικά, όπως πχ. εκθετικά μοντέλα, τα οποία με διάφορες διαδικασίες μπορούν να μετασχηματιστούν σε γραμμικά.

Η γενική μορφή του τυπικού μοντέλου που επιλύεται με την εφαρμογή της πολλαπλής παλινδρόμησης, είναι:

$$y_i = \hat{b}_0 + \sum_{k=1}^{p-1} \hat{b}_k \cdot x_{ik} + \varepsilon_i \qquad (9.33)$$

όπου *p* είναι ο αριθμός των συντελεστών παλινδρόμησης του μοντέλου.

Η τιμή του συντελεστή $\hat{b}_0$, δίνει η τιμή που μπορεί να πάρει η μεταβλητή *Y*, όταν γίνει ταυτόχρονη μηδένιση των τιμών όλων των ανεξάρτητων μεταβλητών.

Ο συντελεστής $\hat{b}_k$ δείχνει τη μέση μεταβολή (αύξηση ή μείωση) της εξαρτημένης μεταβλητής, όταν η μεταβλητή x_{ik} μεταβληθεί κατά μία μονάδα, με την προϋπόθεση ότι όλες οι υπόλοιπες $(p-2)$ ανεξάρτητες μεταβλητές παραμένουν στα-

θερές, ανεξάρτητα από την τιμή που έχουν. Κατ' επέκταση, το ίδιο ισχύει και για τους υπόλοιπους συντελεστές παλινδρόμησης.

Οι ίδιες βασικές προϋποθέσεις που αναφέρθηκαν στην περίπτωση της απλής γραμμικής παλινδρόμησης επεκτείνονται και ισχύουν και για την περίπτωση της εφαρμογής της πολλαπλής γραμμικής παλινδρόμησης. Επιπρόσθετα, επειδή στην εξίσωση υπάρχουν περισσότερες από μία ανεξάρτητες μεταβλητές, αυτές δεν θα πρέπει να εμφανίζουν υψηλή συσχέτιση μεταξύ τους, δηλαδή δεν θα πρέπει να εμφανίζεται **πολυσυγγραμικότητα** (multicollinearity).

Το αντιπροσωπευτικό μοντέλο της πολλαπλής παλινδρόμησης μπορεί να δοθεί και με μορφή πινάκων ως εξής:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix} \cdot \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \Leftrightarrow Y = X \cdot \hat{b} + \varepsilon \qquad (9.34)$$

Όπως φαίνεται στη σχέση (9.34), η εξαρτημένη μεταβλητή Y είναι ένας πίνακας διαστάσεων *n* γραμμών επί μία στήλη: (*n*×1), οι ανεξάρτητες μεταβλητές είναι ένας πίνακας διαστάσεων (*n*×*p*), οι συντελεστές παλινδρόμησης αποτελούν έναν πίνακα διαστάσεων (*p*×1) και τέλος τα σφάλματα είναι ένας πίνακας διαστάσεων (*n*×1).

Στην πολλαπλή γραμμική παλινδρόμηση, οι παράμετροι του πληθυσμού εκτιμώνται μέσω της μεθόδου των ελαχίστων τετραγώνων κατά τρόπο ανάλογο μ'αυτόν που προαναφέρθηκε. Η μόνη διαφορά έγκειται στο γεγονός ότι εδώ το σύστημα εξισώσεων που προκύπτει δεν είναι 2×2 όπως στην περίπτωση της απλής γραμμικής παλινδρόμησης, αλλά *p*×*p*, οπότε γίνεται ευρεία χρήση της θεωρίας των πινάκων βάσει της οποίας λύνονται οι περισσότερες σχέσεις που ακολουθούν. Δηλαδή, επιδιώκεται η ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων της παραγόμενης εξίσωσης παλινδρόμησης μέσω του μηδενισμού των μερικών παραγώγων του αθροίσματος αυτού ως προς τον καθένα από τους συντελεστές παλινδρόμησης. Οι εκτιμητές που προκύπτουν κατά αυτόν τον τρόπο, ενσωματώνουν όλες τις επιθυμητές ιδιότητες των εκτιμητών, δηλαδή είναι αμερόληπτοι, αποτελεσματικοί, επαρκείς και συνεπείς. Υπολογίζονται με βάση τη σχέση:

$$\hat{b} = (X'X)^{-1} \cdot (X'Y) \qquad (9.35)$$

όπου X' είναι ο ανάστροφος πίνακας του *X* και έχει διαστάσεις (*p*×*n*) και $(X'X)^{-1}$

Παράδειγμα 9.7

Σε δείγμα 94 δέντρων, μετρήθηκαν οι διάμετροι των κορμών τους, ανά ένα μέτρο και το ολικό τους ύψος. Με εφαρμογή μεθόδου τμηματικής ογκομέτρησης υπολογίστηκε ο ολικός έμφλοιος κορμικός όγκος τους.

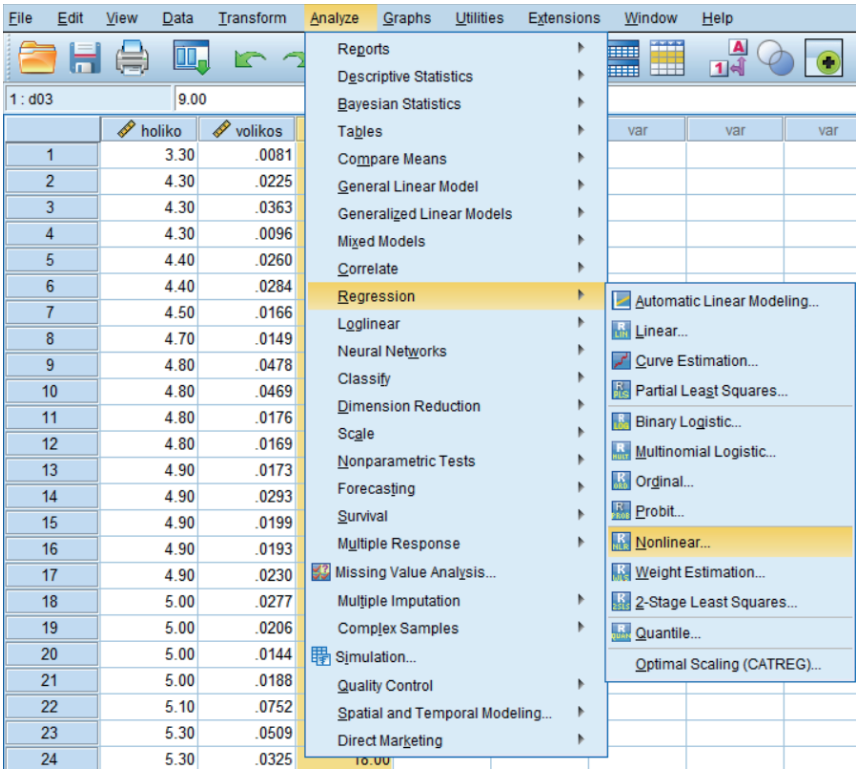
Να καταρτιστεί εξίσωση μη-γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή τον όγκο των δέντρων και ανεξάρτητη μεταβλητή το ολικό ύψος τους.

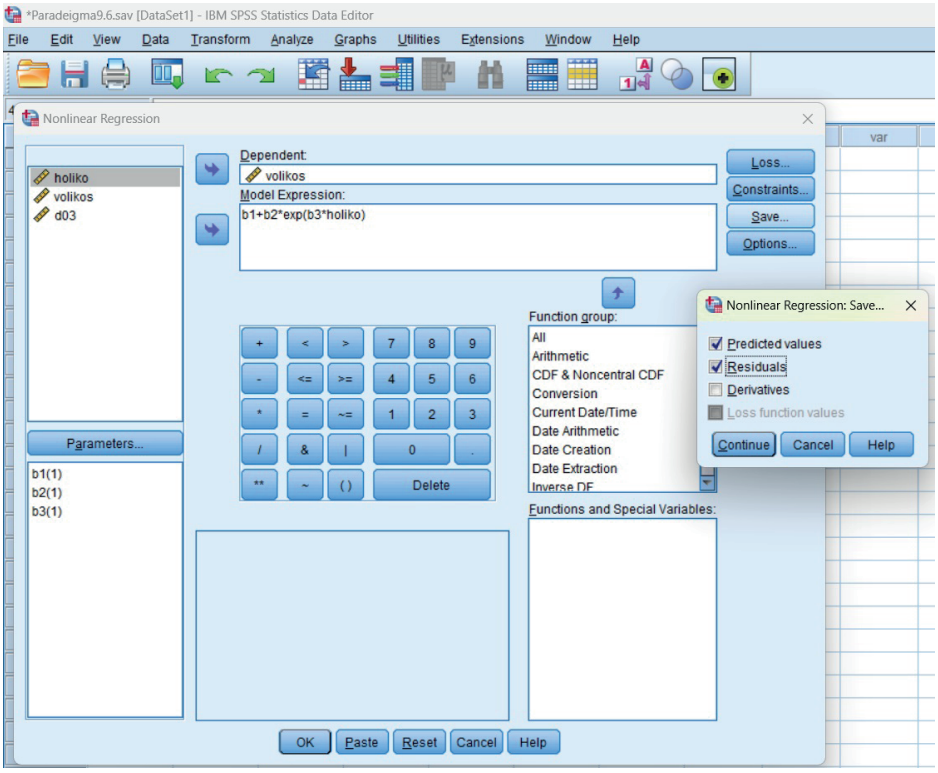
Απάντηση

Πριν την εφαρμογή μη-γραμμικής παλινδρόμησης, θα πρέπει να εξαντληθεί κάθε πιθανότητα, ένα γραμμικό μοντέλο να έχει τη δυνατότητα να περιγράψει με ικανοποιητικό τρόπο, τα δεδομένα. Μετά τη διερεύνηση αυτή και εφόσον η απλή γραμμική παλινδρόμηση και η πολλαπλή γραμμική παλινδρόμηση απέτυχαν να παράγουν ικανοποιητικά αποτελέσματα, τότε εφαρμόζεται η μη-γραμμική παλινδρόμηση. Η εφαρμογή της στο SPSS περιγράφεται στον Πίνακα 9.13.

Πίνακας 9.13. *Επίλυση προβλήματος μη-γραμμικής παλινδρόμησης, με το SPSS.*

Ένδειξη οθόνης





Περιγραφή

1. Εισάγουμε τα δεδομένα
2. Επιλέγουμε Regression → Nonlinear
3. Στη θυρίδα διαλόγου που αναδύεται τοποθετούμε την εξαρτημένη μεταβλητή στο «Dependent» γράφουμε με κωδικοποίηση του SPSS την εξίσωση την οποία επιθυμούμε να προσαρμόσει το SPSS στα δεδομένα
4. Από τις διαθέσιμες εντολές επιλέγουμε «Save» και από το παράθυρο που αναδύεται επιλέγουμε «Predicted values», «Residuals»
5. Στη συνέχεια επιλέγουμε «Continue» και OK
6. Τέλος, μέσω της επιλογής «Graphs» από το κύριο menu των επιλογών επιλέγουμε «Regression Variable Plots» και δημιουργούμε το στικτόδιάγραμμα των δεδομένων με τα αντίστοιχα θηκογράμματα και την γραμμή προσαρμογής του μοντέλου

Εντολές σε γλώσσα SPSS:

```
DATASET ACTIVATE DataSet1.
* NonLinear Regression.
MODEL PROGRAM b1=1 b2=1 b3=1.
COMPUTE PRED_=b1+b2*exp(b3*holiko).
```

```
NLR volikos
 /OUTFILE='C:\Users\maria\spss22540\SPSSFNLR.TMP'
 /PRED PRED_
 /SAVE PRED RESID
 /CRITERIA SSCONVERGENCE 1E-8 PCON 1E-8.

STATS REGRESS PLOT YVARS=volikos XVARS=holiko
/OPTIONS CATEGORICAL=BARS GROUP=1 BOXPLOTS
INDENT=15 YSCALE=75
/FITLINES APPLYTO=TOTAL.
```

Παραγόμενο αποτέλεσμα:

Iteration History[b]

Iteration Number[a]	Residual Sum of Squares	Parameter b1	b2	b3
1.0	3.955E+10	1.000	1.000	1.000
1.1	39.737	.031	3.433E-5	1.000
2.0	39.737	.031	3.433E-5	1.000
2.1	4.867	.031	3.430E-5	.921
3.0	4.867	.031	3.430E-5	.921
3.1	.198	.029	7.206E-5	.759
4.0	.198	.029	7.206E-5	.759
14.1	.075	.006	.003	.407
15.0	.075	.006	.003	.407
15.1	.075	.006	.003	.407

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval Lower Bound	Upper Bound
b1	.006	.010	-.014	.026
b2	.003	.002	.000	.006
b3	.407	.046	.315	.499

Correlations of Parameter Estimates

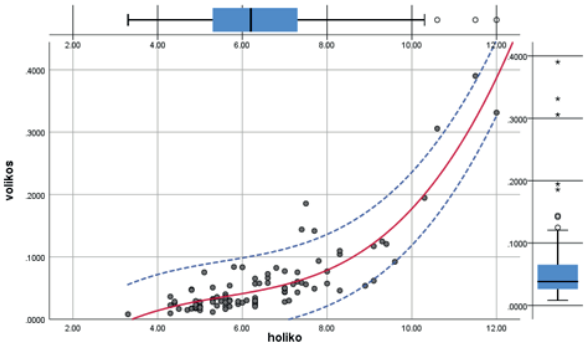
	b1	b2	b3
b1	1.000	-.936	.913
b2	-.936	1.000	-.995
b3	.913	-.995	1.000

ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	.629	3	.210
Residual	.075	91	.001
Uncorrected Total	.703	94	
Corrected Total	.375	93	

Dependent variable: volikos

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .801.



Όπως φαίνεται στο συγκεκριμένο Παράδειγμα 9.7, ζητήθηκε η προσαρμογή του ασυμπτωτικού μη-γραμμικού μοντέλου: $\hat{y}_i = b_1 + b_2 \cdot e^{(b_3 \cdot x_i)}$, το οποίο είναι εσωτερικά μη-γραμμικό, δηλαδή είναι αδύνατο να μετασχηματιστεί σε γραμμική μορφή, εκτός αν τεθεί η προϋπόθεση ότι $b_1 = 0$. Το μοντέλο επιλύθηκε με την εφαρμογή του αλγορίθμου βελτιστοποίησης των Levenberg-Marquardt, με εφαρμογή επαναληπτικής διαδικασίας, ενώ χρησιμοποιήθηκε το μέσο τετραγωνικό σφάλμα, ως το μέτρο αξιολόγησης του μοντέλου.

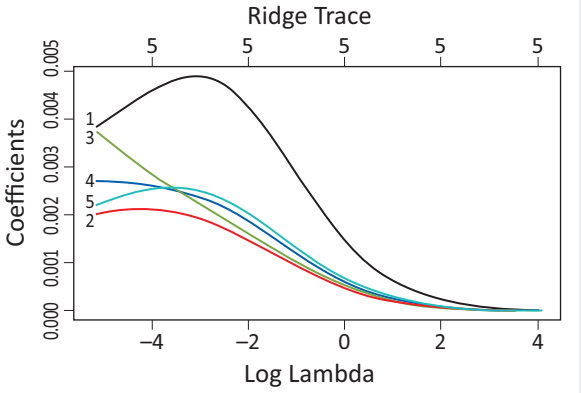
Πίνακας 9.16. *Εφαρμογή της παλινδρόμησης Ridge, στην R.*

Κώδικας R (R-script)

```
library(glmnet)
library(car)
# Load the data
data <- read.csv("C:/Users/MD/ Ridge.csv")
colnames(data) <- make.names(colnames(data))
colnames(data)
colnames(data) <- c("h", "d1", "d2", "d3", "d4", "y")
# Prepare the predictor and response variables
x <- as.matrix(data[, c('h', 'd1', 'd2', 'd3', 'd4')])
y <- data$y
# Fit the ridge regression model with a sequence of lambda values
ridge_model <- glmnet(x, y, alpha = 0)
# Plot the Ridge trace
plot(ridge_model, xvar = "lambda", label = TRUE)
title(main = "Ridge Trace")
# Perform cross-validation to find the best lambda
cv_ridge <- cv.glmnet(x, y, alpha = 0)
best_lambda <- cv_ridge$lambda.min
print(best_lambda)
# Fit the final model with the best lambda
final_ridge_model <- glmnet(x, y, alpha = 0, lambda = best_lambda)
ridge_coefficients <- coef(final_ridge_model)
print(ridge_coefficients)
```

Παραγόμενο αποτέλεσμα

```
> print(best_lambda)
[1] 0.00602468
> ridge_coefficients <-
coef(final_ridge_model)
> print(ridge_coefficients)
(Intercept)   −0.101289505
h              0.003833729
d1             0.002030172
d2             0.003749633
d3             0.002708144
d4             0.002194844
```



Σε περίπτωση που πρέπει να αναλυθούν ταυτόχρονα ποσοτικές και κατηγορικές μεταβλητές, τότε η κοινή ανάλυσή τους θα πρέπει να χρησιμοποιεί τη μετρική απόστασης Gower, η οποία λειτουργεί με αποδεκτό τρόπο για συνδυασμό ποσοτικών και ποιοτικών μεταβλητών.

Πίνακας 10.15. *Επιλογή κατάλληλων τεχνικών στην Ιεραρχική ταξινόμηση, ανάλογα με τον τύπο των μεταβλητών που αναλύονται.*

Βασικές διαθέσιμες τεχνικές στην Ιεραρχική ταξινόμηση				
Μεταβλητές	**Μετρική απόστασης**	**Μετασχηματισμός**	**Μέθοδος σύνδεσης**	**Δενδρόγραμμα**
Ποσοτικές	Ευκλείδεια, L1 (Manhattan), Cosine, Chebyshev, Block, and Minkowski, Pearson correlation	Μπορεί να χρησιμοποιηθεί κανονικοποίηση	Simple (Nearest-neighbor), Complete (Furthest-Neighbor), Ward's, Centroid clustering, Median clustering	Η συγχώνευση βασίζεται σε αριθμητικές αποστάσεις
Κατηγορικές	Gower, συντελεστής απλός ομοιότητας, δείκτης του Jaccard	Απαραίτητη η κωδικοποίηση	Simple (Nearest-neighbor), Complete (Furthest-neighbor), Ward's	Η συγχώνευση βασίζεται σε κατηγορική ομοιότητα

Παράδειγμα 10.10

Ερωτήθηκαν εννιά άτομα τυχαία από μια γειτονιά, στην οποία πρόκειται να κατασκευαστεί πάρκο τσέπης, σχετικά με τις προτιμήσεις τους σε είδος δέντρου (μεταξύ τριών διαφορετικών ειδών), σχήμα (μεταξύ τριών διαφορετικών σχημάτων) και μέγεθος πάρκου (μεταξύ τριών διαφορετικών μεγεθών). Οι απαντήσεις που δόθηκαν καταγράφονται στον παρακάτω πίνακα:

Ερωτώμενος	**Απαντήσεις σχετικά με**		
	Είδος δέντρου	**Σχήμα**	**Μέγεθος**
1	είδος1	κυκλικό	μικρό
2	είδος2	τετραγωνικό	μεγάλο
3	είδος3	τριγωνικό	μέσο
4	είδος2	κυκλικό	μεγάλο
5	είδος1	τετραγωνικό	μικρό

Ερωτώμενος	Απαντήσεις σχετικά με		
	Είδος δέντρου	Σχήμα	Μέγεθος
6	είδος2	τετραγωνικό	μεγάλο
7	είδος3	τριγωνικό	μέσο
8	είδος2	κυκλικό	μεγάλο
9	είδος1	τετραγωνικό	μικρό

Μπορεί να ομαδοποιηθεί η συμπεριφορά των ερωτώμενων ως προς τις προτιμήσεις τους;

Απάντηση

Πρόκειται για τη διερεύνηση των προτιμήσεων 9 ατόμων, των οποίων οι απαντήσεις δημιούργησαν τρεις κατηγορικές μεταβλητές, τριών επιπέδων η καθεμία. Θα χρησιμοποιηθεί η R, για την εφαρμογή της ιεραρχικής ταξινόμησης στα κατηγορικά δεδομένα. Στον Πίνακα 10.16, δίνεται ο βασικός κώδικας και τα παραγόμενα αποτελέσματα.

Πίνακας 10.16. *Εφαρμογή της τεχνικής της ιεραρχικής ταξινόμησης, με την R.*

Κώδικας R (R-script)

```
install.packages("cluster")
library(cluster)
# Sample data with categorical variables
data <- data.frame(
  Color = as.factor(c("1", "2", "3", "2", "1", "2", "3", "2", "1")),
  Shape = as.factor(c("Circle", "Square", "Triangle", "Circle", "Square", "Square",
"Triangle", "Circle", "Square")),
  Size = as.factor(c("Small", "Large", "Medium", "Large", "Small", "Large", "Me-
dium", "Large", "Small"))
)
print(data)
# Compute Ward distance matrix
Wards_dist <- daisy(data, metric = "Ward")
# Print distance matrix
print(as.matrix(Wards_dist))
# Perform hierarchical clustering
hc <- agnes(Wards_dist, method = "average")
```

```
# Print clustering result
print(hc)
# Plot dendrogram
plot(hc, which.plots = 2, main = "Dendrogram of Hierarchical Clustering")
```

Παραγόμενο αποτέλεσμα

```
# Print distance matrix
> print(as.matrix(Wards_dist))
```

	1	2	3	4	5	6	7	8	9
1	0.0000000	1.0000000	1	0.6666667	0.3333333	1.0000000	1	0.6666667	0.3333333
2	1.0000000	0.0000000	1	0.3333333	0.6666667	0.0000000	1	0.3333333	0.6666667
3	1.0000000	1.0000000	0	1.0000000	1.0000000	1.0000000	0	1.0000000	1.0000000
4	0.6666667	0.3333333	1	0.0000000	1.0000000	0.3333333	1	0.0000000	1.0000000
5	0.3333333	0.6666667	1	1.0000000	0.0000000	0.6666667	1	1.0000000	0.0000000
6	1.0000000	0.0000000	1	0.3333333	0.6666667	0.0000000	1	0.3333333	0.6666667
7	1.0000000	1.0000000	0	1.0000000	1.0000000	1.0000000	0	1.0000000	1.0000000
8	0.6666667	0.3333333	1	0.0000000	1.0000000	0.3333333	1	0.0000000	1.0000000
9	0.3333333	0.6666667	1	1.0000000	0.0000000	0.6666667	1	1.0000000	0.0000000

```
>
```

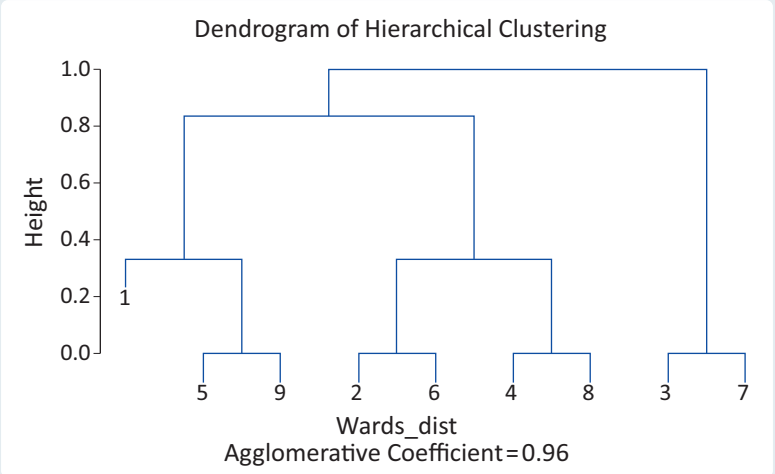
Agglomerative coefficient: **0.962963**

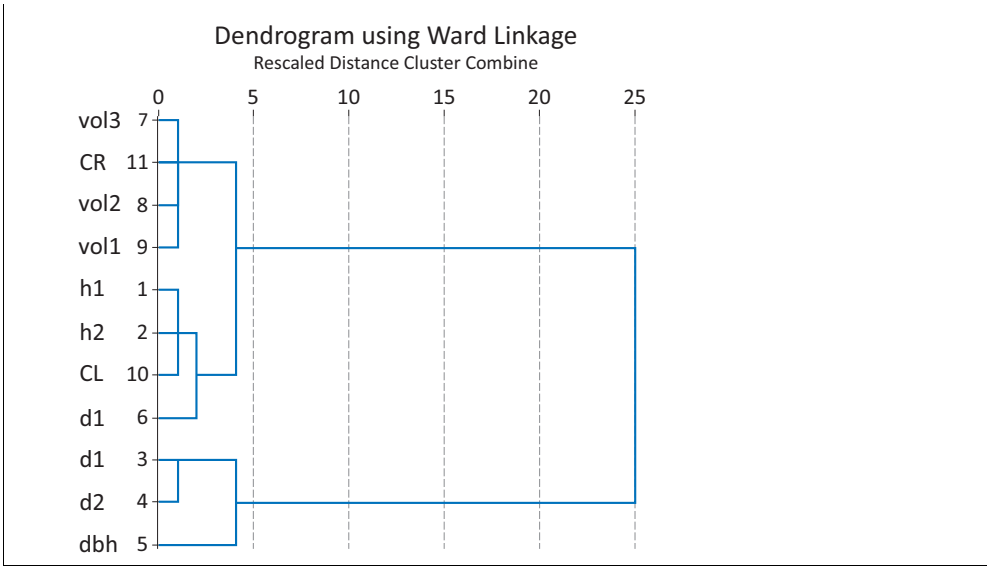
Order of objects:

[1] 1 5 9 2 6 4 8 3 7

Height (summary):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0000	0.1667	0.3125	0.4583	1.0000





Όπως φαίνεται στον Πίνακα 10.17, ο πρώτος πίνακας δίνει τον αριθμό των στοιχείων ανά στήλη τα οποία αναλύθηκαν και την πληροφορία της ύπαρξης ελλειπουσών τιμών η οποία είναι πολύ σημαντική, γιατί μεγάλο ποσοστό ελλειπουσών τιμών μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα.

Ο επόμενος πίνακας (Proximity Matrix) δίνει όπως αναφέρει και το όνομά του πληροφορίες σχετικά με την εγγύτητα μεταξύ των μεταβλητών. Δηλαδή οι μεταβλητές που έχουν μικρότερη απόσταση μεταξύ τους είναι περισσότερο όμοιες. Όπως μπορεί να παρατηρήσει κανείς, η διαγώνιος του πίνακα είναι μηδέν, γιατί η απόσταση της κάθε μεταβλητής από τον εαυτό της είναι μηδενική. Οι αποστάσεις αυτές που δίνονται στον πίνακα εγγύτητας, είναι οι τετραγωνικές Ευκλείδειες αποστάσεις γιατί αυτή η επιλογή έγινε κατά την κατάστρωση της διαδικασίας.

Ο πίνακας της συσσώρευσης (Agglomeration Schedule) δίνει πληροφόρηση σχετικά με τη διαδικασία με την οποία εξελίσσεται η ομαδοποίηση. Στον πίνακα που έχει παραχθεί αξίζει να προσεχθούν οι ακόλουθες πληροφορίες. Φαίνονται 10 βήματα (Stage), σε καθένα από τα οποία μια μεταβλητή (δίνεται ο αύξοντας αριθμός της) μπορεί να αποτελέσει ομάδα με κάποια άλλη κοντινή, από άποψη ομοιότητας, μεταβλητή. Στη συνέχεια, οι συντελεστές ποσοτικοποιούν την διαφορετικότητα μεταξύ των μεταβλητών. Μεγάλη τιμή του συντελεστή αντιπροσωπεύει μεγαλύτερη διαφορετικότητα μεταξύ των μεταβλητών. Αυτή η στήλη των τιμών των συντελεστών μπορεί να αξιοποιηθεί για τη βέλτιστη επιλογή του αριθμού των ομάδων. Δηλαδή, στο σημείο που εμφανίζεται μια απότομη και μεγάλη σε μέγεθος αλλαγή τιμών, σηματοδοτείται η πιθανή βέλτιστη τιμή του αριθμού των ομάδων.

Ευρετήριο όρων