Κεφάλαιο 1

Περιγραφή, παρουσίαση και σύνοψη δεδομένων

Μαθησιακοί στόχοι

Στο τέλος της ενότητας αυτής, ο αναγνώστης θα είναι ικανός να:

1) Διαχωρίσει ποιες είναι οι ποιοτικές και ποιες οι ποσοτικές μεταβλητές και οι διάφοροι τύποι τους
2) Παρουσιάσει και να συνοψίσει με τον κατάλληλο τρόπο τα δεδομένα
3) Διαμορφώσει και να ερμηνεύσει τα αντίστοιχα γραφήματα
4) Κατανοήσει τα χαρακτηριστικά της κανονικής κατανομής και τη διαφορά τους από τις ασύμμετρες κατανομές
5) Εφαρμόσει τον κατάλληλο μετασχηματισμό δεδομένων

Στην ιατρική, συνήθως μελετούνται είτε ασθενείς ή υγιή άτομα που αποτελούν στατιστικές μονάδες. Οι πληροφορίες που συλλέγονται από αυτά τα άτομα αποκαλούνται μεταβλητές. Με τον όρο μεταβλητή, αποκαλούμε κάθε τι το οποίο μεταβάλλεται, αλλάζει, παραλλάσσει ή ποικίλει. Οι μεταβλητές αποτελούν μετρήσιμα στοιχεία ή χαρακτηριστικά των ατόμων που μελετώνται. Στην ουσία, οι επιλεγμένες μεταβλητές χρησιμοποιούνται για να εκτιμηθούν τα χαρακτηριστικά του πραγματικού πληθυσμού. Ο τύπος της μεταβλητής θα πρέπει να προσδιοριστεί για να περιγραφούν σωστά τα δεδομένα και να χρησιμοποιηθούν οι κατάλληλες στατιστικές δοκιμασίες για τον έλεγχο μιας συγκεκριμένης υπόθεσης. Επομένως, περιγραφή είναι η μορφή ενός χαρακτηριστικού για μια συγκεκριμένη στατιστική μονάδα και αποδίδεται με τους όρους **τιμή** και **δεδομένο**.

1.1 Τύποι μεταβλητών

Αρχικά, πριν γίνει οποιοσδήποτε υπολογισμός, θα πρέπει να καθοριστεί ο τύπος των μεταβλητών. Ανάλογα με το πώς μεταβάλλεται μια μεταβλητή διακρίνεται και το είδος στο οποίο ανήκει. Ο βασικός διαχωρισμός είναι μεταξύ των **ποσοτικών μεταβλητών** και των **ποιοτικών μεταβλητών** (Πίνακας 1). Οι ποσοτικές μεταβλητές μπορεί να είναι είτε **συνεχείς (continuous)** ή **ασυνεχείς (discrete)**.

Οι συνεχείς μεταβλητές, όπως είναι το βάρος ή η ηλικία, μπορούν θεωρητικά να πάρουν μια οποιαδήποτε αριθμητική τιμή από ένα εύρος τιμών μεταξύ ελάχιστης και μέγιστης τιμής. Οι ασυνεχείς μεταβλητές, όπως είναι ο αριθμός επισκέψεων σε ένα ιατρείο ή οι ημέρες νοσηλείας στη μονάδα εντατικής θεραπείας, μπορούν να πάρουν μόνο ορισμένες ακέραιες αριθμητικές τιμές και δεν «κτενίζουν» όλες τις δυνατές τιμές ανάμεσα σε ένα ελάχιστο και ένα μέγιστο όριο.

Οι ποιοτικές μεταβλητές συγκροτούν κατηγορίες κατά το δυνατόν ευδιάκριτες και σαφείς και είναι είτε **ονομαστικές (nominal),** χωρίς κάποια διάταξη ή **διατακτικές, διατάξιμες**, **τακτικές, διαβαθμιζόμενες (ordinal)**. Εάν οι ποιοτικές μεταβλητές έχουν δύο κατηγορίες αμοιβαία αποκλειόμενες, τότε ονομάζονται **διχότομες, δυαδικές (binary)**, μπορεί όμως να έχουν παραπάνω από δύο κατηγορίες. Παραδείγματα διχότομων ονομαστικών μεταβλητών είναι «Ναι/Όχι», «Άντρας/Γυναίκα», «Καπνιστής/Μη καπνιστής», «Ζωντανός/Νεκρός», και ονομαστικών μεταβλητών με περισσότερες κατηγορίες είναι η οικογενειακή κατάσταση «Παντρεμένος/Ελεύθερος/Διαζευγμένος/Χήρος», το χρώμα ματιών «καφέ/πράσινο/μπλε», η ομάδα αίματος «Α/Β/ΑΒ/Ο». Η διάταξη μπορεί να μην παίζει ρόλο σε ορισμένες ονομαστικές μεταβλητές που έχουν περισσότερες από δύο κατηγορίες. Για παράδειγμα, τα άτομα που έχουν ομάδα αίματος Β δεν θεωρούνται ότι βρίσκονται μεταξύ της Α και της ΑΒ ομάδας αίματος. Συχνά, όμως παίζει ρόλο η διάταξη και θα πρέπει να ληφθεί υπόψη, όπως συμβαίνει στα διάφορα στάδια του καρκίνου, όσο μεγαλύτερο το στάδιο τόσο χειρότερη η έκβαση του ασθενή. Ένα άλλο παράδειγμα είναι η ένταση του πόνου, ελάχιστος/μέτριος/έντονος.

Οι ποσοτικές μεταβλητές μπορούν να μετασχηματιστούν σε ποιοτικές μεταβλητές χωρίζοντας τα δεδομένα σε κατηγορίες. Για παράδειγμα, ο δείκτης μάζας σώματος (ΔΜΣ), συνήθως μετασχηματίζεται σε μια διατακτική ποιοτική μεταβλητή με τις εξής 3 κατηγορίες: φυσιολογικός θα θεωρείται κάποιος με ΔΜΣ < 25 kg/m^2, υπέρβαρος με ΔΜΣ 25–30 kg/m^2, και παχύσαρκος με ΔΜΣ ≥ 30 kg/m^2. Πολλές βιοχημικές παράμετροι μπορούν να ταξινομηθούν σε 2 κατηγορίες ως φυσιολογικές ή παθολογικές ανάλογα με το αν υπερβαίνουν ή όχι τις φυσιολογικές τιμές, για παράδειγμα η γλυκόζη μεταξύ 60–110 mg/dl θεωρείται φυσιολογική, αλλιώς παθολογική.

Πίνακας 1.1: Παραδείγματα από διάφορους τύπους δεδομένων

Ποσοτικές μεταβλητές	
Συνεχείς	**Ασυνεχείς**
Δείκτης μάζας σώματος (ΔΜΣ), ηλικία, βάρος, αρτηριακή πίεση	ο αριθμός επισκέψεων σε ένα ιατρείο, ημέρες νοσηλείας στη μονάδα εντατικής θεραπείας
Ποιοτικές μεταβλητές	
Ονομαστικές	**Διατακτικές**
Ναι/Όχι Φύλο (Άντρας/Γυναίκα) Ομάδα αίματος A, B, AB, O	Στάδια καρκίνου (I, II, III, IV, V) Ελάχιστος, μέτριος, έντονος πόνος

Γενικά, τα ποιοτικά δεδομένα συνοψίζονται και περιγράφονται πιο εύκολα, οπότε συχνά μετασχηματίζονται οι ποσοτικές μεταβλητές σε ποιοτικές για την περιγραφή και παρουσίαση των δεδομένων. Στη διάγνωση ενός ασθενούς, δεν χρειάζεται να γνωρίζει κάποιος την ακριβή τιμή της γλυκόζης του, αλλά για το αν κυμαίνεται μέσα στις φυσιολογικές τιμές και πιο εύκολα γίνεται αναφορά στο ποσοστό των ατόμων με παθολογική γλυκόζη παρά στις επιμέρους τιμές της γλυκόζης. Ωστόσο, η κατηγοριοποίηση μιας ποσοτικής μεταβλητής σε ποιοτική, περιορίζει τη διαθέσιμη πληροφορία των δεδομένων και τη στατιστική ισχύ στις αναλύσεις, το οποίο σημαίνει μειωμένη πιθανότητα να βρεθεί ένα πραγματικά στατιστικά σημαντικό αποτέλεσμα. Οπότε, η κατηγοριοποίηση μιας ποσοτικής συνεχούς μεταβλητής είναι χρήσιμη για τη σύνοψη και παρουσίαση των αποτελεσμάτων αλλά όχι για τη στατιστική ανάλυση των δεδομένων αυτής της μεταβλητής. Επίσης, υπάρχει το πρόβλημα της διαμόρφωσης των κατάλληλων κατηγοριών με τη χρήση των κατάλληλων διαχωριστικών ορίων και διαφορετικές κατηγορίες μπορεί να οδηγήσουν σε διαφορετικά αποτελέσματα στην ίδια βάση δεδομένων. Είναι προτιμότερο λοιπόν να καταγραφεί η πραγματική τιμή μιας ποσοτικής μεταβλητής, με όσο το δυνατόν μεγαλύτερη ακρίβεια και στη συνέχεια να χρησιμοποιηθεί για τις κατάλληλες στατιστικές δοκιμασίες ανάλογα με το ερευνητικό σκοπό, παρά να καταγραφεί μόνο ως ποιοτική. Ο λόγος είναι ότι εύκολα μετασχηματίζεται μια ποσοτική μεταβλητή σε ποιοτική για την ανάλυση, ενώ αν είχε καταγραφεί αρχικά μόνο ποιοτικά θα ήταν δύσκολο να βρεθεί αργότερα η πραγματική της τιμή. Επίσης, είναι καλύτερο να καταγραφούν για παράδειγμα οι ημερομηνίες γέννησης και εξέτασης έτσι ώστε να υπολογιστεί στη συνέχεια η ηλικία με ακρίβεια παρά να καταγραφεί αρχικά μόνο η ηλικία. Η ακρίβεια των μετρήσεων και ο τύπος των δεδομένων είναι απαραίτητα για τη κατάλληλη στατιστική ανάλυση των δεδομένων.

Ανάλογα με τη θέση της σε ένα στατιστικό μοντέλο η μεταβλητή διακρίνεται σε **ανεξάρτητη** και **εξαρτημένη**. **Ανεξάρτητη** καλείται η μεταβλητή όταν επιδρά επάνω σε άλλη ή άλλες μεταβλητές, οι οποίες χαρακτηρίζονται εξαρτημένες. Δηλαδή η ανεξάρτητη μεταβλητή έχει αυτόνομη δράση και η τιμή της επηρεάζει και διαμορφώνει τις τιμές των εξαρτημένων μεταβλητών. **Εξαρτημένη** χαρακτηρίζεται η μεταβλητή όταν η τιμή της επηρεάζεται ή προσδιορίζεται από τις μεταβολές της ανεξάρτητης μεταβλητής, δηλαδή εκφράζει το αποτέλεσμα της δράσης τους. Βέβαια τα πράγματα δεν είναι πάντα τόσο σαφή και η διάκριση των μεταβλητών σε εξαρτημένες και ανεξάρτητες δεν είναι τόσο καλά οριοθετημένη. Υπάρχουν επίσης μεταβλητές που υπεισέρχονται στην έρευνα και την επηρεάζουν προς κάποια κατεύθυνση χωρίς να είναι εύκολο να εντοπισθούν και να ελεγχθούν, δηλαδή έχουν μια «αόρατη, υπόγεια» δράση. Αυτές οι μεταβλητές λέγονται **συγχυτικές (confounding variables)**.

1.2 Κλίμακες μέτρησης

Τα δεδομένα προέρχονται από τη μέτρηση των μεταβλητών στις στατιστικές μονάδες, με τη βοήθεια των κλιμάκων μέτρησης. Οι κλίμακες μέτρησης προσδιορίζουν την ακρίβεια της πληροφορίας για τις μεταβλητές και την ακρίβεια των δεδομένων. Το είδος των δεδομένων έχει εξαιρετική σημασία διότι καθορίζει τη στατιστική μέθοδο ανάλυσης. Υπάρχουν κλίμακες μέτρησης οι οποίες κατά αύξουσα σειρά πολυπλοκότητας είναι:

1. Ονομαστικές ή κατηγορικές κλίμακες (nominal scales)
2. Διατακτικές κλίμακες (ordinal scales)
3. Αριθμητικές ή ισοδιαστημικές κλίμακες (interval scales)
4. Αναλογικές κλίμακες (ratio scales)

Οι δύο πρώτες ονομάζονται ποιοτικές κλίμακες μέτρησης, ενώ οι δύο επόμενες ποσοτικές κλίμακες μέτρησης. Ονομαστικές είναι οι κλίμακες μέτρησης στις οποίες η κατάταξη των στατιστικών μονάδων γίνεται σε καλά προσδιορισμένες, αμοιβαία αποκλειόμενες και διακρίσιμες φυσικές κατηγορίες. Οι τιμές εκφράζονται με λέξεις και τα δεδομένα ονομάζονται ονομαστικά ή κατηγορικά. Για παράδειγμα, άνδρας/γυναίκα στον προσδιορισμό του φύλου.

Διατακτικές κλίμακες μέτρησης είναι εκείνες στις οποίες η ένταξη των στατιστικών μονάδων γίνεται σε κατηγορίες σαφείς, ισοδύναμες και διατεταγμένες. Οι μετρήσεις με αυτές τις κλίμακες είναι απλές και εκφράζουν ένα είδος ποιοτικού χαρακτηριστικού και οι τιμές στις μεταβλητές μπορούν να διαταχθούν από «μικρότερο σε μεγαλύτερο», από «χαμηλότερο σε υψηλότερο» από «αρνητικό σε θετικό», χωρίς όμως να γνωρί-

ζουμε την ακριβή απόσταση μεταξύ των κατηγοριών ή από κάποιο σημείο αναφοράς. Για παράδειγμα, ένας γιατρός μπορεί να χρησιμοποιήσει μια κλίμακα από το 0 μέχρι το 10 για να καταγράψει τη βελτίωση σε κάποιο πρόβλημα υγείας ενός ασθενή, όπου το 0 θα υποδηλώνει «καμία βελτίωση» και το 10 θα υποδηλώνει «τέλεια βελτίωση» δηλαδή εξαφάνιση του προβλήματος.

Αριθμητικές ή ισοδιαστημικές είναι οι κλίμακες μέτρησης στις οποίες οι στατιστικές μονάδες εντάσσονται σε σαφώς καθορισμένες, αμοιβαία αποκλειόμενες, διατακτικές κατηγορίες και χρησιμοποιούν σταθερή μονάδα μέτρησης. Το μηδέν δεν υποδηλώνει «απουσία» της μέτρησης. Για παράδειγμα, η θερμοκρασία σε βαθμούς Κελσίου αποτελεί μια αριθμητική κλίμακα, όπου το μηδέν δεν σημαίνει «καμία» θερμοκρασία, αλλά αποτελεί τιμή θερμοκρασίας.

Οι αναλογικές κλίμακες διατηρούν όλα τα χαρακτηριστικά των αριθμητικών κλιμάκων και επιπλέον διαθέτουν πραγματικό σημείο αναφοράς που αντιστοιχεί στο απόλυτο μηδέν. Οι μετρήσεις όπως το ύψος και το βάρος αποτελούν αναλογικές κλίμακες όπου το απόλυτο μηδέν σημαίνει κανένα ύψος ή κανένα βάρος.

1.3 Παρουσίαση και σύνοψη ποιοτικών δεδομένων

Τα ποιοτικά δεδομένα συνοψίζονται με **πίνακες συχνοτήτων,** όπου υπολογίζεται ο αριθμός των παρατηρήσεων μέσα σε κάθε κατηγορία. Για παράδειγμα σε ένα δείγμα αιμοδοτών καταγράφηκε η ομάδα αίματός τους και τα αποτελέσματα παρουσιάζονται στο παρακάτω πίνακα:

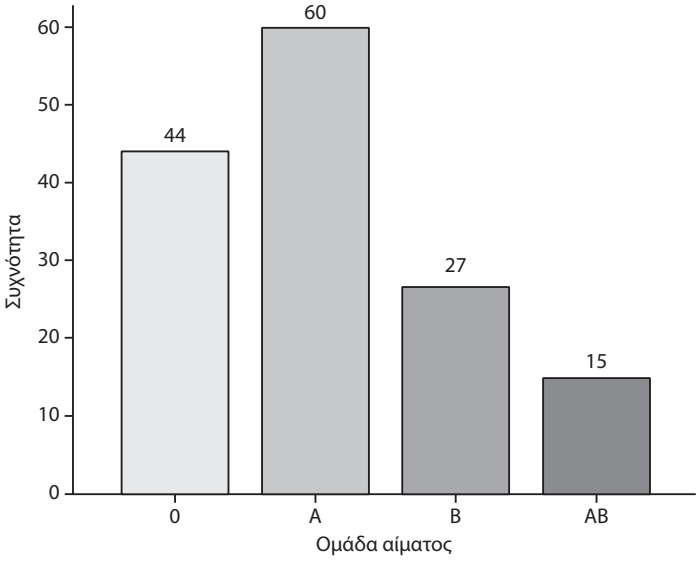
Πίνακας 1.2: Κατανομή ομάδα αίματος σε ένα δείγμα 146 αιμοδοτών

Ομάδα αίματος	Συχνότητα	Σχετική συχνότητα	Αθροιστική Συχνότητα	Αθροιστική σχετική συχνότητα
Ο	44	0,30	44	0,30
Α	60	0,41	104	0,71
Β	27	0,19	131	0,90
ΑΒ	15	0,10	146	1,00
Σύνολο	146	1,00		

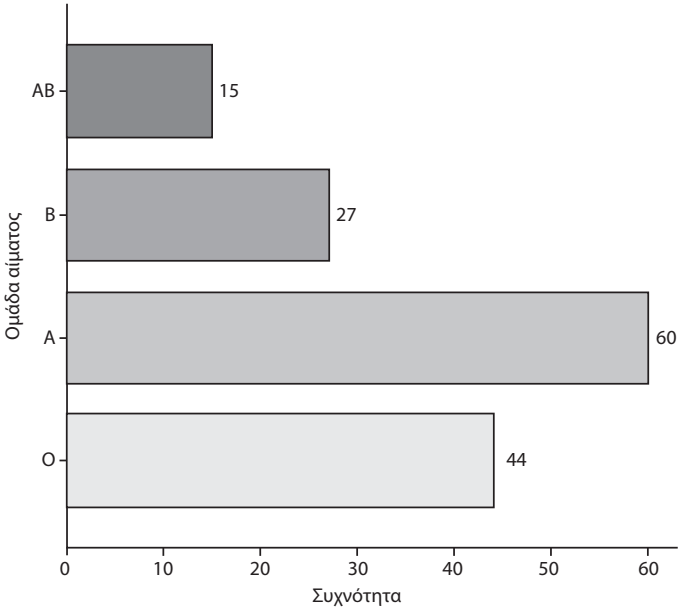
Συνήθως, τα ποιοτικά δεδομένα περιγράφονται με τη **συχνότητα,** τον αριθμό παρατηρήσεων μέσα σε κάθε κατηγορία και τη **σχετική συχνότητα,** που είναι το ποσοστό και υπολογίζεται, για παράδειγμα για την ομάδα αίματος Ο: 44/146 = 0,30 ή 30%. Σε μερικές περιπτώσεις, παρουσιάζονται η **αθροιστική συχνότητα** και η **αθροι-**

στική σχετική συχνότητα, τα οποία υπολογίζονται σε κάθε κατηγορία με την άθροιση των τιμών των προηγούμενων κατηγοριών. Για παράδειγμα, στην ομάδα αίματος Β η αθροιστική συχνότητα 131 προέκυψε από την άθροιση των τιμών 44 + 60 + 27 = 131.

Γραφικά, τα ποιοτικά δεδομένα παρουσιάζονται είτε με **ραβδογράμματα (bar charts)** ή **διαγράμματα σε τομείς,** δηλαδή τις **πίτες (pie charts),** που απεικονίζουν είτε τις συχνότητες ή τα ποσοστά κάθε ομάδας. Ένα παράδειγμα ραβδογράμματος για την κατανομή της ομάδας αίματος του Πίνακα 1.2 αποτελεί το Σχήμα 1.1. Για κάθε ομάδα αίματος σχεδιάζεται μια κάθετη ράβδος που το μήκος της είναι ανάλογο με τη συχνότητα εμφάνισης της συγκεκριμένης ομάδας αίματος. Μεταξύ των ράβδων υπάρχουν μικρά κενά που υποδηλώνουν ότι τα δεδομένα είναι ποιοτικά και έτσι ξεχωρίζει το ραβδόγραμμα από το ιστόγραμμα. Επίσης, αυτό το ραβδόγραμμα μπορεί να απεικονιστεί και οριζόντια όπως στο Σχήμα 1.2.

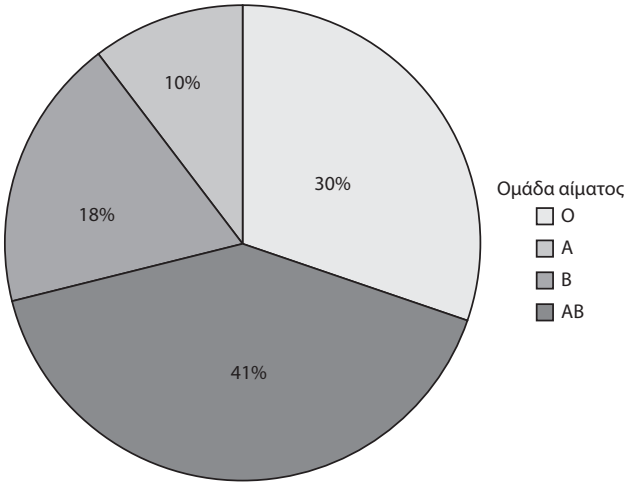


Σχήμα 1.1: Ραβδόγραμμα της κατανομής της ομάδας αίματος σε συχνότητες.



Σχήμα 1.2: Ραβδόγραμμα της κατανομής της ομάδας αίματος σε συχνότητες με οριζόντιες ράβδους.

Το αντίστοιχο γράφημα αλλά σε πίτα και ποσοστά παρουσιάζεται στο Σχήμα 1.3. Κάθε κομμάτι της πίτας είναι ανάλογο σε μέγεθος με τη συχνότητα εμφάνισης της συγκεκριμένης ομάδας αίματος.



Σχήμα 1.3: Πίτα της κατανομής της ομάδας αίματος σε ποσοστά.

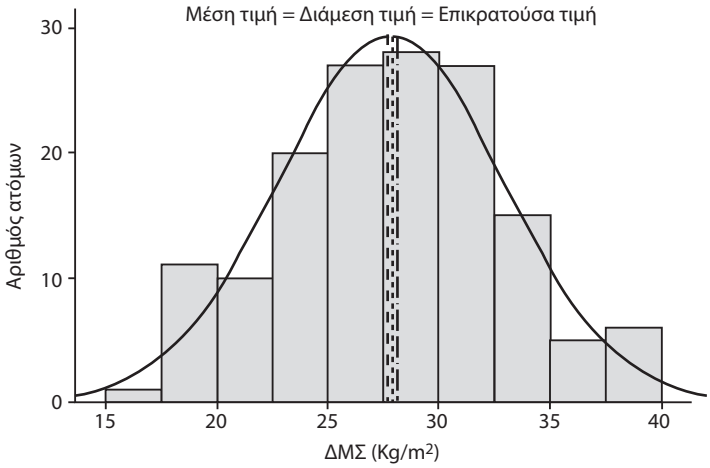
1.4 Παρουσίαση και σύνοψη ποσοτικών δεδομένων

Η γραφική παρουσίαση των ποσοτικών δεδομένων θα πρέπει να προηγείται από την εφαρμογή οποιασδήποτε στατιστικής ανάλυσης, διότι βοηθάει στην αναγνώριση των ακραίων τιμών και της κατανομής που έχουν τα δεδομένα. Επίσης, είναι ένας τρόπος για την επιλογή των κατάλληλων μέτρων περιγραφής των δεδομένων, όπως και της κατάλληλης στατιστικής ανάλυσης.

1.4.1 Ιστόγραμμα

Το **ιστόγραμμα** χρησιμοποιείται συχνά και είναι ένας απλός τρόπος για τον οπτικό έλεγχο της κατανομής των τιμών. Στην ουσία είναι ένα γράφημα που χρησιμοποιείται για την απεικόνιση της κατανομής συχνοτήτων των τιμών μιας μεταβλητής. Αρχικά τα δεδομένα ταξινομούνται με αύξουσα σειρά και μετά χωρίζονται σε τάξεις ή κλάσεις δηλαδή σε διαστήματα με το ίδιο εύρος τιμών. Ο αριθμός των διαστημάτων και το εύρος των τάξεων συνήθως είναι αυθαίρετα. Πολλά στατιστικά πακέτα έχουν ως προεπιλογή τα 10 διαστήματα.

Για παράδειγμα, επιλέχτηκε ένα τυχαίο δείγμα 150 ατόμων, καταγράφηκε το ύψος και το βάρος και υπολογίστηκε ο δείκτης μάζας σώματος (ΔΜΣ). Ο ΔΜΣ είναι μια συνεχής ποσοτική μεταβλητή διότι μπορεί να πάρει μια οποιαδήποτε αριθμητική τιμή από ένα εύρος τιμών. Οι τιμές του ΔΜΣ κυμαινόταν από 16 μέχρι 39 kg/m^2. Τα δεδομένα χωρίστηκαν σε 10 διαστήματα του ίδιου εύρους (οριζόντιος άξονας), καταμετρήθηκε ο αριθμός των ατόμων που ανήκουν σε κάθε διάστημα τιμών (κάθετος άξονας) και δημιουργήθηκε το ιστόγραμμα που παρουσιάζεται στο Σχήμα 1.4.

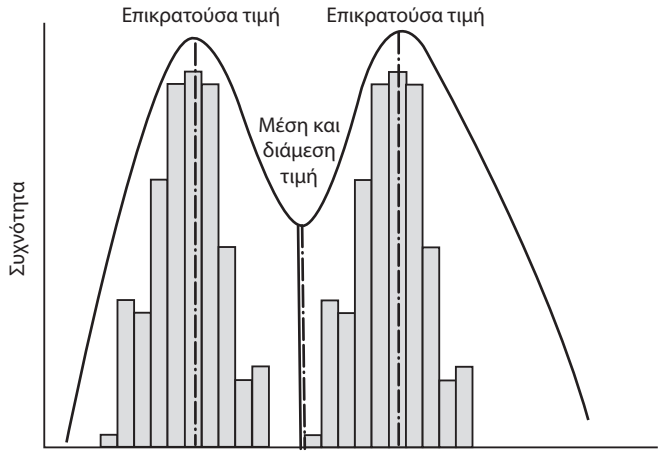


Σχήμα 1.4: Ιστόγραμμα του δείκτη μάζας σώματος (ΔΜΣ) 150 ατόμων.

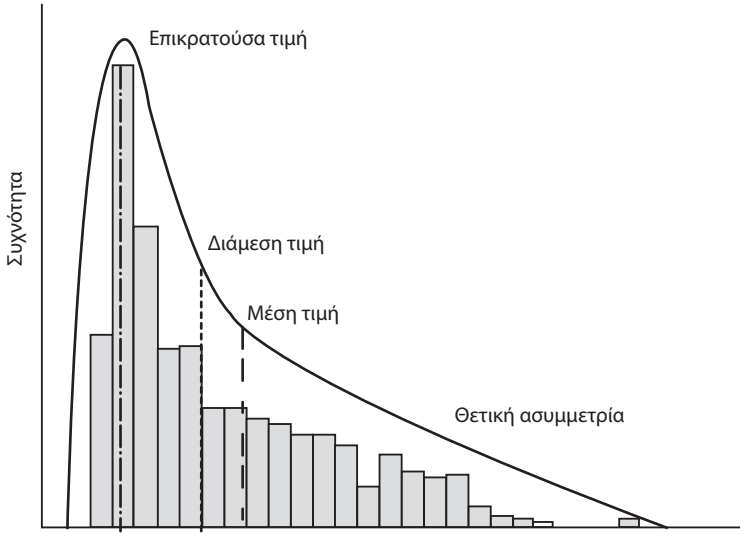
Τα ιστογράμματα βοηθούν τους ερευνητές να καταλάβουν την κατανομή των δεδομένων και στη συνέχεια να αποφασίσουν για την επιλογή των κατάλληλων στατιστικών αναλύσεων. Συνήθως, η κατανομή των δεδομένων είναι **μονοκόρυφη (unimodal),** όπως το ιστόγραμμα του Σχήματος 1.4 και απεικονίζει την κατανομή δεδομένων που έχουν πολλές βιολογικές και φυσιολογικές μεταβλητές. Η κατανομή των τιμών ακολουθεί σχεδόν το σχήμα καμπάνας όπως οριοθετείται με τη γραμμή που περνά πάνω από το ιστόγραμμα. Επίσης, το σχήμα της καμπάνας είναι συμμετρικό, όπου το ένα μισό της καμπάνας αντικατοπτρίζει το άλλο μισό και η μέση τιμή της μεταβλητής συμπίπτει με τη διάμεση και την επικρατούσα τιμή της. Αυτά είναι τα χαρακτηριστικά μια κανονικής κατανομής, που αποκαλείται κατανομή του Gauss. Τα δεδομένα που ακολουθούν την κανονική κατανομή αναλύονται με παραμετρικές μεθόδους.

Όταν η κατανομή των δεδομένων δεν είναι συμμετρική, για παράδειγμα έχει **δύο κορυφές (δικόρυφη, bimodal,** Σχήμα 1.5) ή και **παραπάνω κορυφές (πολυκόρυφη, mutlimodal)** ή είναι μονοκόρυφη αλλά **θετικά ασύμμετρη, λοξή προς τα δεξιά (right skewness,** Σχήμα 1.6), ή **αρνητικά ασύμμετρη, λοξή προς τα αριστερά (left skewness,** Σχήμα 1.7), τότε τα δεδομένα δεν ακολουθούν την κανονική κατανομή.

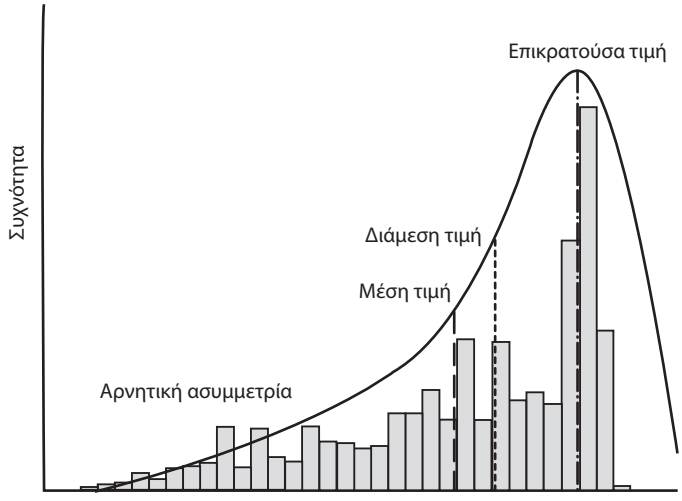
Τα χαρακτηριστικά μιας δικόρυφης κατανομής είναι ότι η μέση τιμή συμπίπτει με τη διάμεση τιμή αλλά έχει δύο κορυφές, δηλαδή δύο επικρατούσες τιμές (Σχήμα 1.5). Συνήθως, μια δικόρυφη κατανομή προκύπτει από το συνδυασμό δύο μονοκόρυφων κατανομών, π.χ. ύψος ανδρών και γυναικών. Επίσης, αποτελεί ένδειξη ετερογένειας στα δεδομένα και ότι δεν έχει ληφθεί υπόψη και η επίδραση κάποιου άλλου παράγοντα στη κατανομή των τιμών π.χ. το φύλο ή η ηλικία.



Σχήμα 1.5: Χαρακτηριστικά μιας δικόρυφης κατανομής.



Σχήμα 1.6: Θετικά ασύμμετρη κατανομή.



Σχήμα 1.7: Αρνητικά ασύμμετρη κατανομή.

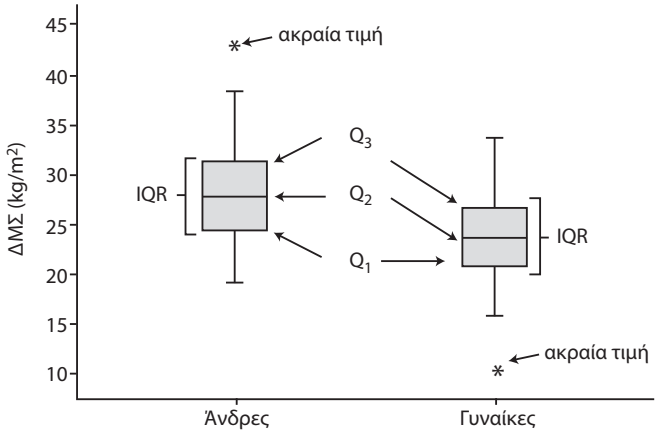
Μια θετικά ασύμμετρη κατανομή χαρακτηρίζεται από μια μακριά ουρά στα δεξιά αλλά η πλειοψηφία των δεδομένων βρίσκονται στην αριστερή πλευρά, δηλαδή υπάρχουν αναλογικά λίγες υψηλές τιμές που μπορεί να είναι και ακραίες (Σχήμα 1.6). Η κατανομή αυτή παρουσιάζει δεξιά λοξότητα και πολλές φορές ονομάζεται και *λοξή προς*

τα δεξιά ή *θετικά λοξή κατανομή*. Σ' αυτές τις περιπτώσεις, η μέση τιμή βρίσκεται πιο πέρα (δεξιά) από τη διάμεση τιμή στην ουρά και συνήθως η μέση τιμή είναι μεγαλύτερη από τη διάμεση τιμή, ενώ η επικρατούσα τιμή έχει την μικρότερη τιμή. Τέτοιου είδους κατανομές, συναντώνται συχνά στην ιατρική. Αντιθέτως, μια αρνητικά ασύμμετρη κατανομή χαρακτηρίζεται από μια μακριά ουρά στα αριστερά αλλά η πλειοψηφία των δεδομένων βρίσκονται στη δεξιά πλευρά, εδώ οι χαμηλές τιμές είναι αναλογικά πιο λίγες (Σχήμα 1.7). Δηλαδή, η κατανομή παρουσιάζει αριστερή λοξότητα και πολλές φορές ονομάζεται και *λοξή προς τα αριστερά* ή *αρνητικά λοξή κατανομή*. Εκτός από κάποιες εξαιρέσεις, η μέση τιμή είναι μικρότερη (αριστερά) από τη διάμεση τιμή, ενώ η επικρατούσα τιμή έχει την μεγαλύτερη τιμή. Οι αρνητικά ασύμμετρες κατανομές είναι πιο σπάνιες στην ιατρική.

Συχνά, στην ιατρική βιβλιογραφία, τα δεδομένα που δεν ακολουθούν την κανονική κατανομή αναλύονται με παραμετρικές μεθόδους, το οποίο είναι λάθος εκτός και αν το μέγεθος δείγματος είναι πολύ μεγάλο (N > 200). Αυτά τα δεδομένα θα πρέπει να αναλυθούν με αντίστοιχες μη παραμετρικές μεθόδους ή να μετασχηματιστούν πρώτα έτσι ώστε να αποκτήσουν την κανονική κατανομή για να αναλυθούν στη συνέχεια με παραμετρικές δοκιμασίες.

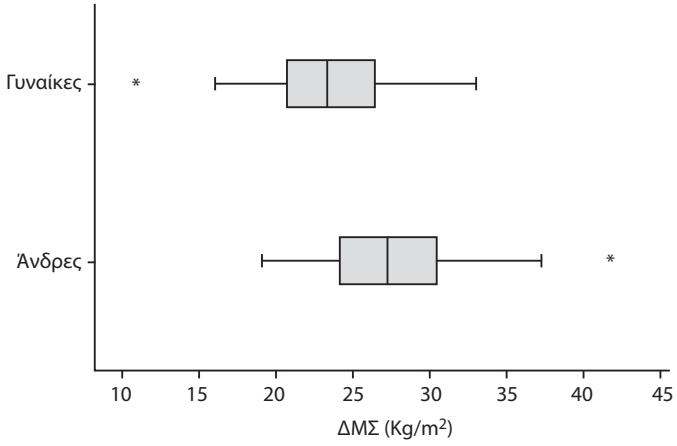
1.4.2 Θηκόγραμμα (Box-and-whisker plot, Box-plot)

Ένα ακόμη γράφημα που χρησιμοποιείται συχνά είναι το **θηκόγραμμα** (Σχήμα 1.8). Αυτό το γράφημα απεικονίζει τα τεταρτημόρια (Q1 = 25% των παρατηρήσεων ή 25^{ο} εκατοστημόριο, Q2 = 50% παρατηρήσεων ή 50^{ο} εκατοστημόριο, Q3 = 75% των παρατηρήσεων ή 75^{ο} εκατοστημόριο), σε ένα ορθογώνιο κουτί (θήκη). Η κάτω και η επάνω άκρη του κουτιού αντιστοιχούν στο 1^{ο} και 3^{ο} τεταρτημόριο (Q1 και Q3), η απόσταση μεταξύ των δύο άκρων αποτελούν το **ενδοτεταρτημοριακό εύρος (interquartile range) IQR = Q3 – Q1,** όπου μέσα σε αυτό περιέχεται το 50% των παρατηρήσεων. Η γραμμή περίπου στο μέσο του κουτιού είναι η **διάμεση τιμή (Q2),** δηλαδή το 50^{ο} εκατοστημόριο, όπου τα μισά δεδομένα κατανέμονται κάτω από τη διάμεση τιμή και τα άλλα μισά πάνω από τη διάμεση τιμή. Οι κάθετες γραμμές που εκτείνονται εκτός του κουτιού αντιστοιχούν στο 5^{ο} και 95^{ο} εκατοστημόριο. Οι τιμές εκτός των κάθετων γραμμών θεωρούνται **ακραίες τιμές (outliers, extreme values)** και συμβολίζονται με αστεράκια. Οποιαδήποτε τιμή που βρίσκεται σε μια απόσταση μεγαλύτερη κατά μιάμιση φορά από το IQR (1,5* IQR), θεωρείται ακραία.



Σχήμα 1.8: Θηκόγραμμα του δείκτη μάζας σώματος (ΔΜΣ) σε άνδρες και γυναίκες.

Το θηκόγραμμα, όπως και το ιστόγραμμα είναι χρήσιμο για τον οπτικό έλεγχο της κατανομής των τιμών. Όταν το κουτί είναι συμμετρικό με τη διάμεση τιμή ακριβώς στο κέντρο του τότε υπάρχει ένδειξη ότι τα δεδομένα κατανέμονται κανονικά. Το ιστόγραμμα δίνει καλύτερη πληροφορία για μια ομάδα δεδομένων, αλλά όταν υπάρχουν πολλές ομάδες δεδομένων τα θηκογράμματα είναι πιο εύχρηστα και προτιμούνται. Επίσης, τα θηκογράμματα είναι χρήσιμα για τη γραφική παρουσίαση των δεδομένων που δεν κατανέμονται κανονικά και αναλύονται με **μη παραμετρικές** στατιστικές μεθόδους. Αυτά τα γραφήματα μπορούν να παρουσιαστούν και οριζόντια, όπως στο Σχήμα 1.9.



Σχήμα 1.9: Οριζόντιο θηκόγραμμα του δείκτη μάζας σώματος (ΔΜΣ) σε άνδρες και γυναίκες.

1.4.3 Φυλλόγραμμα (Stem and leaf plot, stemplot)

Το φυλλόγραμμα, όπως και το ιστόγραμμα βοηθάει στον οπτικό έλεγχο της κατανομής των δεδομένων και χρησιμοποιόταν πιο συχνά παλιότερα διότι ήταν εύκολο να σχεδιαστεί με το χέρι. Πλέον, με την εξέλιξη της τεχνολογίας και των γραφικών μεθόδων στους ηλεκτρονικούς υπολογιστές, η χρήση του δεν είναι και τόσο συχνή. Σε αντίθεση με τα ιστογράμματα, τα φυλλογράμματα, σχεδιάζονται με τα αρχικά δεδομένα με τουλάχιστον 2 ψηφία, τα οποία ταξινομούνται κατά αύξοντα αριθμό.

Ένα κλασικό φυλλόγραμμα αποτελείται από δύο στήλες που χωρίζονται ενδιάμεσα από μια κάθετη γραμμή. Η στήλη αριστερά περιέχει τον «κορμό» και η στήλη δεξιά περιέχει τα «φύλλα».

Για παράδειγμα, σε μια νεογνολογική κλινική μετρήθηκε η περίμετρος κεφαλής σε 15 βρέφη που γεννήθηκαν μια συγκεκριμένη μέρα και οι τιμές της σε εκατοστά ήταν οι παρακάτω:

39,1 , 33,8 , 36,3 , 33,8 , 31,1 , 34,5, 35,4, 34,2, 39,2, 37,6, 36,3, 35,2, 38,1, 34,6, 34,2

Για να σχεδιαστεί το φυλλόγραμμα οι τιμές θα πρέπει να ταξινομηθούν κατά αύξουσα σειρά:

31,1, 33,8, 33,8, 34,2, 34,2, 34,5, 34,6, 35,2, 35,4, 36,3, 36,3, 37,6, 38,1, 39,1, 39,2

Μετά, θα πρέπει να αποφασιστεί ποια ψηφία θα αποτελούν τον κορμό και ποια τα φύλλα. Συνήθως, τα φύλλα αποτελούνται από το τελευταίο ψηφίο του αριθμού και ο κορμός αποτελείται από όλα τα άλλα ψηφία. Στις περιπτώσεις που υπάρχουν πολύ μεγάλοι αριθμοί ή πολλά δεκαδικά ψηφία, για απλούστευση, οι τιμές μπορούν να στρογγυλοποιηθούν. Στο συγκεκριμένο παράδειγμα, τα φύλλα αποτελούνται από τα δεκαδικά και ο κορμός αποτελείται από τις δεκάδες.

Το φυλλόγραμμα σχεδιάζεται με δύο στήλες που χωρίζονται από μια κάθετη γραμμή. Ο κορμός σχεδιάζεται από τα αριστερά και είναι σημαντικό ότι κάθε κορμός εμφανίζεται μόνο μια φορά στη στήλη και ότι κανένας ενδιάμεσος αριθμός δεν παραλείπεται ακόμη και αν δεν περιέχει φύλλα. Τα φύλλα σχεδιάζονται σε αύξουσα σειρά σε μια γραμμή στα δεξιά κάθε κορμού, όπως το Σχήμα 1.10.

31	|	1
32	|	
33	|	8 8
34	|	2 2 5 6
35	|	2 4
36	|	3 3
37	|	6
38	|	1
39	|	1
39	|	2

Σχήμα 1.10: Φυλλόγραμμα της περιμέτρου κεφαλής 15 νεογνών.

Με μια απλή ματιά στο φυλλόγραμμα, καταλαβαίνουμε το εύρος των τιμών, όπου η ελάχιστη τιμή είναι το 31,1 και η μέγιστη 39,2 εκατοστά.

1.5 Μετασχηματισμός δεδομένων

Συνήθως, ο μετασχηματισμός των δεδομένων εφαρμόζεται σε ποσοτικές μεταβλητές για να γίνει η κατανομή των δεδομένων πιο «συμμετρική» και να σταθεροποιηθεί η διακύμανση (ομοιογένεια της διασποράς), παραδοχές που θα πρέπει να πληρούνται για την εφαρμογή παραμετρικών δοκιμασιών. Επίσης, σε άλλες περιπτώσεις απαιτείται να υπάρχει γραμμική σχέση μεταξύ δύο μεταβλητών που και αυτό μπορεί να επιτευχθεί ύστερα από τον κατάλληλο μετασχηματισμό των δεδομένων. Ωστόσο, πολλές φορές αποφεύγεται ο μετασχηματισμός των δεδομένων διότι είναι δύσκολη η ερμηνεία των αποτελεσμάτων μετά από την ανάλυση των μετασχηματισμένων μεταβλητών. Επίσης, δεν είναι σωστό να δοκιμάζονται διάφοροι μετασχηματισμοί των δεδομένων και να επιλέγεται αυτός που θα καταλήξει με στατιστικά σημαντικό αποτέλεσμα.

Ο μετασχηματισμός των δεδομένων προκύπτει με την εφαρμογή μιας μαθηματικής συνάρτησης στα δεδομένα, για παράδειγμα λογαριθμοποίηση κάθε τιμής της μεταβλητής. Ο μετασχηματισμός εφαρμόζεται όταν τα δεδομένα παρουσιάζουν σημαντική ασυμμετρία, είτε θετική ή αρνητική όπως είδαμε στην προηγούμενη ενότητα (Σχήμα 1.6 και 1.7, αντίστοιχα). Ανάλογα με την αρχική κατανομή των δεδομένων εφαρμόζονται διάφοροι τύποι μετασχηματισμών, οι πιο γνωστοί από τους οποίους είναι ο λογαριθμικός, ο αντίστροφος και της τετραγωνικής ρίζας.

Ο λογαριθμικός μετασχηματισμός Z = logx:

Ο λογαριθμικός μετασχηματισμός είναι ο πιο συχνός και χρησιμοποιείται συνήθως για δείκτες ανάπτυξης που κατανέμονται εκθετικά ή για μεταβλητές που παρουσιάζουν θετική ασυμμετρία. Επίσης, εφαρμόζεται στις περιπτώσεις που η μεταβλητότητα αυξάνει με την αύξηση των τιμών της μεταβλητής. Ως λογαριθμικός μετασχηματισμός επιλέγεται είτε ο κοινός λογάριθμος με βάση το 10 ($\log_{10}x$) ή ο φυσικός (νεπέρειος) λογάριθμος με βάση το e = 2,718 ($\log_e x = \ln x$). Δεν υπάρχει διαφορά ποιος από τους δύο θα χρησιμοποιηθεί, αφού οι λογάριθμοι της μιας βάσης είναι ένα πολλαπλάσιο της άλλης, αλλά οι λογάριθμοι με βάση το 10 είναι λίγο πιο κατανοητοί διότι $\log_{10}1 = 0$ (10^0), $\log_{10}10 = 1$ (10^1), $\log_{10}100 = 2$ (10^2) κ.ο.κ. Ο λογαριθμικός μετασχηματισμός μπορεί να εφαρμοστεί μόνο σε θετικές τιμές ($x > 0$). Μια χαρακτηριστική μεταβλητή που χρησιμοποιείται συνήθως σε λογαριθμική κλίμακα με βάση το 10 είναι το ιϊκό φορτίο, λόγω της μεγάλης μεταβλητότητας των τιμών της (παίρνει τιμές από 1 μέχρι 100.000 αντίγραφα/mL), οπότε μετά τον μετασχηματισμό το ιϊκό φορτίο εκφράζεται σε Log_{10} αντίγραφα/mL.

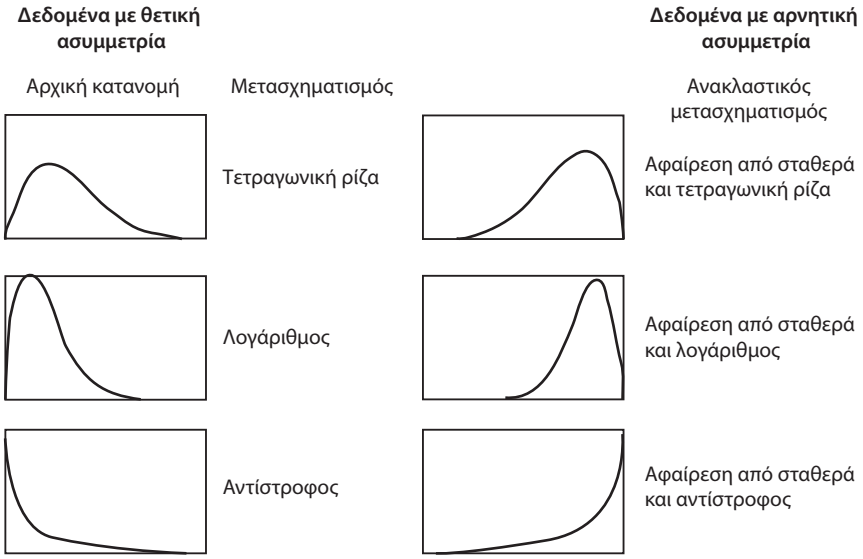
Ο μετασχηματισμός της τετραγωνικής ρίζας Z = $\sqrt{x}$

Συχνά, ο μετασχηματισμός αυτός εφαρμόζεται σε δεδομένα που μετράνε τη συχνότητα εμφάνισης ενός γεγονότος, π.χ. αριθμός βακτηριδίων σε ένα δείγμα νερού. Αυτά τα δεδομένα ακολουθούν την κατανομή Poisson και με τη χρήση του μετασχηματισμού της τετραγωνικής ρίζας ακολουθούν την κανονική κατανομή. Δεν μπορεί να εφαρμοστεί σε μεταβλητές που παίρνουν αρνητικές τιμές.

Ο αντίστροφος μετασχηματισμός Z = 1/x

Συνήθως, ο αντίστροφος μετασχηματισμός χρησιμοποιείται σε δεδομένα που μετράνε χρόνο εκδήλωσης ενός γεγονότος, π.χ. χρόνος ενζυμικής αντίδρασης. Μετά τον μετασχηματισμό σταθεροποιείται η διακύμανση και τα δεδομένα ακολουθούν την κανονική κατανομή. Δεν μπορεί να εφαρμοστεί σε μεταβλητές με μηδενικές τιμές διότι το αντίστροφο του μηδενός δεν ορίζεται.

Ο κατάλληλος μετασχηματισμός ανάλογα με το βαθμό ασυμμετρίας παρουσιάζεται στο **Σχήμα 1.11**. Τα δεδομένα με αρνητική ασυμμετρία απαιτούν πρώτα έναν «ανακλαστικό μετασχηματισμό» (*reflected transformation*), για να μετατραπούν σε δεδομένα με θετική ασυμμετρία. Δηλαδή, θα πρέπει πρώτα να δημιουργηθεί μια νέα μεταβλητή όπου κάθε τιμή της αρχικής μεταβλητής θα αφαιρεθεί από μια σταθερά (K). Η σταθερά υπολογίζεται προσθέτοντας 1 στη μέγιστη τιμή της αρχικής μεταβλητής και μετά εφαρμόζεται ο αντίστοιχος κατάλληλος μετασχηματισμός στη νέα μεταβλητή.



Σχήμα 1.11: Κατανομή τιμών και προτεινόμενοι μετασχηματισμοί (Tabachnick & Fidell 2007).

Συνοψίζοντας το παραπάνω σχήμα με λόγια και τύπους:

Κατανομή δεδομένων	**Προτεινόμενος Μετασχηματισμός**
Μέτρια θετική ασυμμετρία	Τετραγωνική ρίζα, $Z = \sqrt{x}$
Σημαντική θετική ασυμμετρία	Λογαριθμικός, Z = logx
Σημαντική θετική ασυμμετρία (με μηδενικές τιμές)	Λογαριθμικός, Z = log(x + C)
Μέτρια αρνητική ασυμμετρία	Τετραγωνική ρίζα, $Z = \sqrt{K - x}$
Σημαντική αρνητική ασυμμετρία	Λογαριθμικός, Z = log(K – x)

Όπου:

x = οι τιμές της μετρηθείσας μεταβλητής

C = μια σταθερά που προστίθεται σε κάθε τιμή έτσι ώστε η ελάχιστη τιμή να είναι 1

K = μια σταθερά από την οποία αφαιρείται κάθε τιμή έτσι ώστε η μικρότερη τιμή να είναι 1, συνήθως είναι ίση με τη μέγιστη τιμή + 1

Ασκήσεις

1. Ποιες από τις παρακάτω μεταβλητές είναι ποιοτικές;
   α) αριθμός επεισοδίων μιας νόσου σε έναν ασθενή κατά τη διάρκεια ενός έτους
   β) Αιμοσφαιρίνη
   γ) Φύλο
   δ) Αξιολόγηση του πόνου (πολύ/μέτριο/λίγο/καθόλου)
   ε) Μείωση της πίεσης ύστερα από τη χρήση αντιυπερτασικού φαρμάκου

2. Ποιες από τις παρακάτω μεταβλητές είναι ποσοτικές;
   α) Τριγλυκερίδια
   β) Οικογενειακή κατάσταση
   γ) Φυλή
   δ) Ηλικία
   ε) Κάπνισμα

3. Γραφικά πως θα μπορούσε να παρουσιαστεί μια συνεχής μεταβλητή;
   α) Ραβδόγραμμα
   β) Θηκόγραμμα
   γ) Ιστόγραμμα
   δ) Πίτα
   ε) Φυλλόγραμμα

4. Όταν τα δεδομένα κατανέμονται κανονικά ισχύουν τα παρακάτω:
   α) Η μέση τιμή συμπίπτει με τη διάμεση τιμή
   β) Η μέση τιμή συμπίπτει με την ελάχιστη τιμή
   γ) Η επικρατούσα τιμή συμπίπτει με τη μέγιστη τιμή
   δ) Η διάμεση τιμή συμπίπτει με την επικρατούσα τιμή
   ε) Η μέση τιμή είναι μεγαλύτερη από τη διάμεση τιμή

5. Όταν τα δεδομένα παρουσιάζουν αρνητική ασυμμετρία ισχύουν τα παρακάτω:
   α) Η μέση τιμή είναι μεγαλύτερη από τη διάμεση τιμή
   β) Η μέση τιμή συμπίπτει με την ελάχιστη τιμή
   γ) Η επικρατούσα τιμή είναι μεγαλύτερη από τη μέση τιμή
   δ) Η διάμεση τιμή συμπίπτει με την επικρατούσα τιμή
   ε) Η μέση τιμή είναι μικρότερη από τη διάμεση τιμή

6. Για ποιο λόγο εφαρμόζεται η τεχνική του μετασχηματισμού των δεδομένων και ποια είδη μετασχηματισμού χρησιμοποιούνται συχνότερα;