ARCHIVES OF HELLENIC MEDICINE: ISSN 11-05-3992

ΕΦΑΡΜΟΣΜΕΝΗ ΙΑΤΡΙΚΗ ΕΡΕΥΝΑ
APPLIED MEDICAL RESEARCH

ΑΡΧΕΙΑ ΕΛΛΗΝΙΚΗΣ ΙΑΤΡΙΚΗΣ 2014, 31(2):221–243
ARCHIVES OF HELLENIC MEDICINE 2014, 31(2):221–243

Μονομεταβλητή ανάλυση επιδημιολογικών δεδομένων

Π. Γαλάνης

Εργαστήριο Οργάνωσης και Αξιολόγησης Υπηρεσιών Υγείας, Τμήμα Νοσηλευτικής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Αθήνα

Univariate analysis of epidemiological data

Abstract at the end of the article

Λέξεις ευρετηρίου

Ανάλυση δεδομένων
Κανονική κατανομή
Μέτρα διασποράς
Μέτρα θέσης
Μονομεταβλητή ανάλυση

1. Εισαγωγή
2. Μεταβλητές
2.1. Ποιοτικές μεταβλητές
2.1.1. Ονομαστικές μεταβλητές
2.1.2. Διατάξιμες μεταβλητές
2.2. Ποσοτικές μεταβλητές
2.2.1. Μεταβλητές διαστηματικής κλίμακας
2.2.2. Μεταβλητές κλίμακας λόγου
2.2.3. Συνεχείς και ασυνεχείς μεταβλητές
3. Παρουσίαση δεδομένων
3.1. Πίνακες συχνοτήτων
3.2. Γραφήματα
3.2.1. Ραβδογράμματα
3.2.2. Κυκλικά διαγράμματα
3.2.3. Ιστογράμματα
3.2.4. Διαγράμματα πλαισίου
3.2.5. Διαγράμματα σημείων
4. Μέτρα θέσης
4.1. Μέσος
4.2. Διάμεσος
4.3. Επικρατούσα τιμή
5. Μέτρα διασποράς
5.1. Εύρος
5.2. Ενδοτεταρτημοριακό εύρος
5.3. Διασπορά και τυπική απόκλιση
5.4. Συντελεστής μεταβλητότητας
6. Κανονική κατανομή
7. Έλεγχος κανονικότητας
7.1. Μέσος και διάμεσος
7.2. Συντελεστές ασυμμετρίας και κύρτωσης
7.3. Στατιστικοί έλεγχοι
7.4. Γραφήματα
7.5. Συμπεράσματα
8. Σύνοψη

1. ΕΙΣΑΓΩΓΗ

Η Στατιστική είναι ο επιστημονικός κλάδος που αφορά στη συλλογή, στην οργάνωση και στην ανάλυση (ή καλύτερα σύνθεση) δεδομένων που υπόκεινται σε τυχαία μεταβλητότητα.[1] Σημειώνεται ότι ο όρος «ανάλυση δεδομένων» είναι εσφαλμένος εννοιολογικά, καθώς οι παρατηρήσεις δεν είναι δεδομένα* και η επεξεργασία των παρατηρήσεων είναι σύνθεση και όχι ανάλυση. Στην ουσία, η ανάλυση των δεδομένων είναι το σύνολο της μαρτυρίας ή, αλλιώς, της ένδειξης που χρησιμοποιεί η εμπειρική έρευνα για τον έλεγχο της υπόθεσης.

Οι εφαρμογές της Βιοστατιστικής** συνεχώς αυξάνονται,

* Δεδομένα (data) ή εμπειρικά δεδομένα (empirical data) με την έννοια ότι υπάρχουν ανεξάρτητα από τη νόηση δεν υφίστανται. Αποτελούν ουσιαστικά παρατηρήσεις (observations), νοητικά φορτισμένες. Αυτό εξ άλλου διακρίνει τις παρατηρήσεις από τις εντυπώσεις (impressions).

** Η Στατιστική (statistics) που εφαρμόζεται στην επιστήμη της Βιολογίας καλείται Βιοστατιστική (biostatistics). Έχει επικρατήσει εξ άλλου ο όρος Βιοστατιστική να χρησιμοποιείται και στην περίπτωση κατά την οποία η Στατιστική εφαρμόζεται στις επιστήμες υγείας.

διευκολύνοντας έτσι σημαντικά την ανάλυση δεδομένων ακόμη και σε εξαιρετικά σύνθετες περιπτώσεις. Με την ανάπτυξη εξειδικευμένων στατιστικών προγραμμάτων, αρκετά από τα οποία είναι προσιτά στο μέσο χρήστη ηλεκτρονικών υπολογιστών, η ανάλυση των δεδομένων που προκύπτουν από τις επιδημιολογικές μελέτες, στις περισσότερες περιπτώσεις, είναι σχετικά απλή διαδικασία. Είναι σαφές ότι οι επιστήμονες υγείας δεν είναι στατιστικοί και δεν μπορούν βέβαια να πραγματοποιήσουν πολύπλοκες μαθηματικές διαδικασίες χωρίς τη βοήθεια στατιστικών προγραμμάτων. Αρκετοί θα μπορούσαν χωρίς τη βοήθεια στατιστικών προγραμμάτων να διερευνήσουν τη σχέση μεταξύ δύο διχοτόμων μεταβλητών, με τη δημιουργία του κατάλληλου τετράπτυχου πίνακα και την εφαρμογή του στατιστικού ελέγχου x^2, αλλά είναι πρακτικά αδύνατον να δημιουργήσουν πολυμεταβλητά μοντέλα παλινδρόμησης. Ακόμη και έμπειροι στατιστικοί θα συναντούσαν σημαντικές δυσκολίες στην περίπτωση δημιουργίας πολυμεταβλητών μοντέλων παλινδρόμησης. Σκοπός δεν είναι να μεταβληθούν οι επιστήμονες υγείας σε στατιστικούς, αλλά να μπορούν να αντιλαμβάνονται τα αποτελέσματα της ανάλυσης των επιδημιολογικών δεδομένων και να εξάγουν ορθά συμπεράσματα.

Η ανάλυση δεδομένων περιλαμβάνει (α) τα περιγραφικά στατιστικά μέτρα που συνοψίζουν τα δεδομένα μιας μελέτης και (β) τα διαλογισμικά στατιστικά μέτρα που χρησιμοποιούνται στα στατιστικά υποδείγματα ή, αλλιώς, μοντέλα για την εξαγωγή συμπερασμάτων σχετικά με το αντικείμενο μιας μελέτης που είναι, ουσιαστικά, η παράμετρος η οποία ενδιαφέρει τους ερευνητές. Η *περιγραφική Στατιστική* (descriptive statistics) χρησιμοποιείται για τη συνοπτική και εμπεριστατωμένη παρουσίαση των δεδομένων ή, καλύτερα, των παρατηρήσεων μιας επιδημιολογικής μελέτης, ενώ η συμπερασματολογική ή *επαγωγική Στατιστική* (inferential statistics) χρησιμοποιείται για τη διερεύνηση της ύπαρξης σχέσεων μεταξύ προσδιοριστών και εκβάσεων. Η περιγραφική Στατιστική αφορά ουσιαστικά στην παρουσίαση των δεδομένων των επιδημιολογικών μελετών (μονομεταβλητή ανάλυση), ενώ η επαγωγική Στατιστική αναφέρεται στη διμεταβλητή και στην πολυμεταβλητή ανάλυση.

Η μονομεταβλητή ανάλυση (univariate analysis) αφορά στην ξεχωριστή παρουσίαση κάθε μεταβλητής μιας μελέτης, η διμεταβλητή ανάλυση (bivariate analysis) αναφέρεται στη διερεύνηση της ύπαρξης σχέσης μεταξύ ενός προσδιοριστή* και μιας έκβασης και η πολυμεταβλητή ανάλυση (multivariate analysis) αφορά στη διερεύνηση της ύπαρξης σχέσης μεταξύ ενός προσδιοριστή και μιας έκβασης, λαμβάνοντας όμως υπ' όψη και την ύπαρξη τυχόν συγχυτών** και τροποποιητών.***

Η μονομεταβλητή ανάλυση στην περίπτωση των ονομαστικών μεταβλητών αφορά στην παράθεση των απόλυτων και των σχετικών συχνοτήτων, ενώ στην περίπτωση των ποσοτικών μεταβλητών αναφέρεται στην παράθεση των κατάλληλων μέτρων θέσης και διασποράς. Ιδιαίτερη περίσκεψη απαιτείται στην περίπτωση των διατάξιμων μεταβλητών όπου είναι δυνατή η παράθεση τόσο των απόλυτων και των σχετικών συχνοτήτων όσο και των μέτρων θέσης και διασποράς. Σημειώνεται ότι, στην περίπτωση των διατάξιμων μεταβλητών, η παράθεση των μέτρων θέσης και διασποράς έχει νόημα μόνο εφ' όσον οι μεταβλητές αυτές αντιμετωπιστούν από τους ερευνητές ως μεταβλητές κλίμακας λόγου. Από στατιστική άποψη, η συγκεκριμένη προσέγγιση δεν είναι η πλέον ενδεδειγμένη, αλλά εφ' όσον επιλεγεί πρέπει να αναφέρονται με σαφήνεια τα κριτήρια και οι προϋποθέσεις εφαρμογής της.

Τα *μέτρα θέσης* (measures of location) ή, αλλιώς, *μέτρα κεντρικής τάσης* (measures of central tendency) αφορούν στις τιμές αυτές γύρω από τις οποίες οι παρατηρήσεις τείνουν να συγκεντρώνονται σε μεγαλύτερο βαθμό, ενώ τα *μέτρα διασποράς* (measures of dispersion) ή, αλλιώς, *μέτρα μεταβλητότητας* (measures of variability) δηλώνουν το βαθμό στον οποίο διασπείρονται οι παρατηρήσεις. Τα σημαντικότερα μέτρα θέσης είναι ο μέσος, η διάμεσος και η επικρατούσα τιμή, ενώ τα σημαντικότερα μέτρα διασποράς είναι το εύρος, το ενδοτεταρτημοριακό εύρος, η διασπορά, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας.

2. ΜΕΤΑΒΛΗΤΕΣ

Ο εμπειρικός καθορισμός των τιμών μιας αριθμητικής συνάρτησης ή, αλλιώς, μιας ποσοτικής έννοιας καλείται μέτρηση. Ο όρος *μέτρηση* όμως χρησιμοποιείται συχνά και με ευρύτερη έννοια, περιλαμβάνοντας και την κατανομή των αντικειμένων μιας κατηγορίας σε τάξεις (ποιοτικές έννοιες).

* Παράγοντας κινδύνου (risk factor) ή έκθεση (exposure) ή *προσδιοριστής* (determinant), όπως τελικά επικράτησε να αναφέρεται σήμερα, είναι το χαρακτηριστικό (συγγενές, περιβαλλοντικό ή συμπεριφοράς) των ατόμων από το οποίο εξαρτάται (σχετίζεται ή συναρτάται) η συχνότητα εμφάνισης της μελετώμενης έκβασης.[2]

** Ένα χαρακτηριστικό για να θεωρηθεί *συγχυτής* (confounder) πρέπει (α) να σχετίζεται με τη συχνότητα εμφάνισης της μελετώμενης έκβασης, (β) να έχει σχέση με το μελετώμενο προσδιοριστή ή, με άλλη διατύπωση, να κατανέμεται ανισότιμα στις δύο (ή περισσότερες) κατηγορίες του μελετώμενου προσδιοριστή και (γ) να μην είναι αποτέλεσμα του μελετώμενου προσδιοριστή, να μην αποτελεί δηλαδή ενδιάμεσο στάδιο του μηχανισμού με τον οποίο ο μελετώμενος προσδιοριστής προκαλεί την έκβαση.[2]

*** *Τροποποιητής* (modifier) είναι το χαρακτηριστικό του ατόμου, το οποίο προκαλεί αλλαγή στο αποτέλεσμα του μελετώμενου προσδιοριστή στη συχνότητα εμφάνισης της έκβασης, μεταβάλλει δηλαδή τη σχέση μεταξύ του μελετώμενου προσδιοριστή και της συχνότητας εμφάνισης της έκβασης.[2]

Με τη διαδικασία της μέτρησης επιτυγχάνεται ουσιαστικά η συστηματική απόδοση αριθμών στα αντικείμενα και τις ιδιότητές τους, με αποτέλεσμα να διευκολύνεται σημαντικά η χρήση των μαθηματικών μεθόδων στη μελέτη και στην περιγραφή των αντικειμένων και των σχέσεών τους.

Διακρίνονται διάφορα είδη ή επίπεδα μέτρησης με τη χρήση των ίδιων όρων και για τα δεδομένα που αντιστοιχούν σε κάθε επίπεδο. Τα διάφορα αυτά είδη δεδομένων διαφέρουν τόσο στην ερμηνεία των αριθμητικών τιμών οι οποίες χρησιμοποιούνται όσο και στις στατιστικές μεθόδους που επιλέγονται για τη στατιστική ανάλυση. Οι μεταβλητές (και κατ' επέκταση και τα δεδομένα) ανάλογα με τα μαθηματικά τους χαρακτηριστικά διακρίνονται σε *ποιοτικές* (qualitative) και *ποσοτικές* (quantitative), με τις πρώτες να διακρίνονται σε ονομαστικές και διατάξιμες και τις δεύτερες σε μεταβλητές διαστηματικής κλίμακας και μεταβλητές κλίμακας λόγου (εικ. 1).[3–13] Οι ποσοτικές μεταβλητές εξ άλλου διακρίνονται σε συνεχείς και ασυνεχείς. Οι ποιοτικές μεταβλητές είναι γνωστές και ως *κατηγορικές μεταβλητές* (categorical variates), καθώς με τις «μετρήσεις» ένα ορισμένο αντικείμενο συνδέεται με μια ορισμένη κατηγορία ή τάξη (π.χ. άνδρας ή γυναίκα).

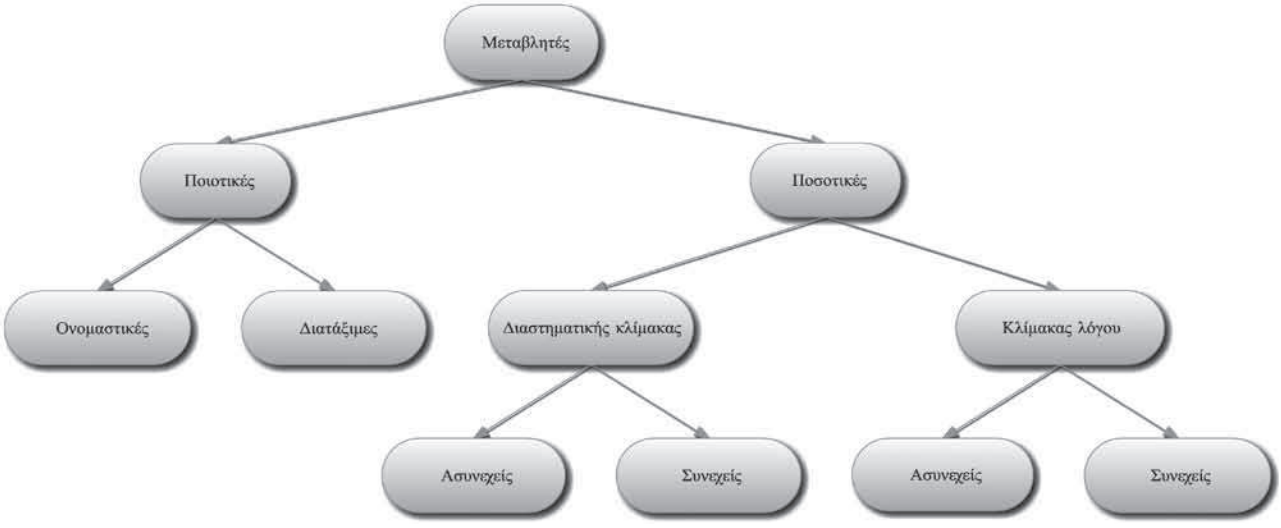
Σημειώνεται ότι οι μεταβλητές (ο αγγλικός όρος είναι variate και όχι variable) δεν υπάρχουν στη φύση και αποτελούν στατιστικές έννοιες, που σχεδιάζονται από τους ερευνητές, ενώ ως δεδομένα νοούνται οι τιμές που λαμβάνει μια μεταβλητή κατά τη μέτρησή της. Για παράδειγμα, το φύλο δεν είναι μεταβλητή. Στην περίπτωση δημιουργίας μιας μεταβλητής για το φύλο, οι άνδρες μπορούν να λάβουν την τιμή 0 και οι γυναίκες την τιμή 1. Οι τιμές αυτές 0 και 1 δεν έχουν αριθμητικό νόημα και βέβαια δεν αποτελούν τη μοναδική επιλογή. Είναι δυνατόν οι άνδρες να λάβουν την τιμή 1 και οι γυναίκες την τιμή 2 χωρίς να επηρεαστεί η ανάλυση των δεδομένων. Ενδεικτικά αναφέρεται ότι στην περίπτωση κατά την οποία η μεταβλητή είναι το φύλο, τα δεδομένα μπορούν να είναι είτε άνδρες είτε γυναίκες. Στην περίπτωση όπου η μεταβλητή είναι η ομάδα αίματος, τα δεδομένα μπορούν να είναι ομάδα αίματος Α, Β, ΑΒ ή Ο κ.λπ. (πίν. 1).

Πίνακας 1. Απεικόνιση των μεταβλητών και των δεδομένων μιας μελέτης.

	Άτομο Α	**Άτομο Β**	**Άτομο Γ**	**Άτομο Δ**
Οι μεταβλητές	***Τα δεδομένα***			
Φύλο	Άνδρας	Γυναίκα	Άνδρας	Γυναίκα
Ηλικία (έτη)	34	43	29	23
Ομάδα αίματος	A	B	AB	O
Βάρος (kg)	85	67	90	55

2.1. Ποιοτικές μεταβλητές

2.1.1. Ονομαστικές μεταβλητές. Οι ονομαστικές μεταβλητές (nominal variates) αποτελούν την απλούστερη και πλέον συνήθη μορφή μεταβλητών.[3–13] Στην περίπτωση αυτή, τα αντικείμενα μιας ορισμένης κατηγορίας «διασπώνται» και εντάσσονται σε διάφορες ομάδες, με τους αριθμούς να αποτελούν ουσιαστικά ονόματα ή χαρακτηρισμούς τάξεων και χωρίς να έχουν αριθμητικό νόημα. Για παράδειγμα, οι άνδρες συμβολίζονται με 0 και οι γυναίκες με 1, χωρίς όμως οι αριθμοί να έχουν νόημα και χωρίς να υπάρχει διάταξη των δύο κατηγοριών. Θα μπορούσε να υπάρξει ο ακριβώς αντίθετος συμβολισμός, δηλαδή οι άνδρες να συμβολιστούν με 1 και οι γυναίκες με 0, χωρίς και πάλι οι αριθμοί να έχουν νόημα και δίχως να επηρεαστεί η ανάλυση των δεδομένων.



Εικόνα 1. Είδη μεταβλητών.

Στην περίπτωση των ονομαστικών μεταβλητών, για να είναι επιστημονικά χρήσιμες, θα πρέπει (α) οι τάξεις να μην έχουν κοινά στοιχεία, δηλαδή να έχουν σαφώς προσδιορισμένο πλάτος, ώστε να αποκλείεται να ανήκει ένα στοιχείο σε περισσότερες από μία τάξεις και (β) η κατανομή των τάξεων να εξαντλεί την περιοχή που αναλύεται. Σημειώνεται ότι τα δεδομένα που αφορούν στις ονομαστικές μεταβλητές δεν έχουν μονάδα μέτρησης. Χαρακτηριστικά παραδείγματα ονομαστικών μεταβλητών αποτελούν το φύλο, η ομάδα αίματος, η εθνικότητα κ.ά.

Όταν οι ονομαστικές μεταβλητές μπορούν να λάβουν μόνο μία από δύο συγκεκριμένες τιμές, όπως άνδρες και γυναίκες, τότε καλούνται *διχοτόμες* ή *δυαδικές* (dichotomous, binary). Οι διχοτόμες μεταβλητές (όπως π.χ. η εμφάνιση ή όχι μιας πάθησης, το θετικό ή το αρνητικό αποτέλεσμα μιας εργαστηριακής δοκιμασίας κ.ά.) χρησιμοποιούνται συχνά στις επιστήμες υγείας και γι' αυτό έχουν αναπτυχθεί ιδιαίτερες στατιστικές μέθοδοι ανάλυσης, με πλέον χαρακτηριστική τη λογιστική παλινδρόμηση.

2.1.2. Διατάξιμες μεταβλητές. Οι διατάξιμες μεταβλητές (ordinal variates) είναι εκείνες στις οποίες η σειρά ή, αλλιώς, η διάταξη μεταξύ των διαφόρων κατηγοριών έχει σημασία, έτσι ώστε οι μεγαλύτερες αριθμητικές τιμές να αντιπροσωπεύουν την παρουσία ενός χαρακτηριστικού σε μεγαλύτερο βαθμό και οι μικρότερες την παρουσία του ίδιου χαρακτηριστικού σε μικρότερο βαθμό.[3–13] Στην περίπτωση αυτή, τα αντικείμενα μιας ορισμένης κατηγορίας όχι μόνο «διασπώνται» και εντάσσονται σε διάφορες κατηγορίες ή τάξεις, αλλά είναι δυνατή και η διάταξη των εν λόγω τάξεων με τρόπο που να επιτρέπει τις μεταξύ τους συγκρίσεις. Τα δεδομένα που αφορούν στις διατάξιμες μεταβλητές δεν έχουν μονάδα μέτρησης, όπως ακριβώς συμβαίνει και στην περίπτωση των ονομαστικών μεταβλητών.

Χαρακτηριστικό παράδειγμα διατάξιμης μεταβλητής στις επιστήμες υγείας αποτελεί ο βαθμός εγκαύματος, που συνήθως λαμβάνει τιμές 1–4, με τις υψηλότερες τιμές να αντιπροσωπεύουν σοβαρότερη μορφή εγκαύματος. Ένα άλλο παράδειγμα αποτελεί η ταξινόμηση των τραυματισμών σύμφωνα με το επίπεδο σοβαρότητάς τους, με τη μεταβλητή αυτή να λαμβάνει, π.χ. τιμές 1–4, όπου 1 αντιστοιχεί σε ελαφρύ τραυματισμό, 2 σε μέτριο, 3 σε σοβαρό και 4 σε θανατηφόρο τραυματισμό. Και στις δύο περιπτώσεις η διάταξη των τάξεων ή, αλλιώς, των κατηγοριών πραγματοποιείται λογικά, αλλά δεν είναι δυνατόν να ποσοτικοποιηθεί η διαφορά μεταξύ των κατηγοριών και να καθοριστεί αν η διαφορά, π.χ., μεταξύ εγκαυμάτων πρώτου και δεύτερου βαθμού είναι ίδια με τη διαφορά μεταξύ εγκαυμάτων τρίτου και τέταρτου βαθμού.

Οι ευρύτατα χρησιμοποιούμενες κλίμακες Likert* οδηγούν στη συλλογή διατάξιμων δεδομένων. Η κλίμακα Likert είναι μια ψυχομετρική κλίμακα που χρησιμοποιείται στα ερωτηματολόγια εκτίμησης του βαθμού συμφωνίας (ή διαφωνίας) των συμμετεχόντων αναφορικά με διάφορες προτάσεις. Η *κλίμακα Likert* (Likert scale) πρέπει να διαχωρίζεται από τα στοιχεία Likert (Likert items). Η κλίμακα Likert είναι το άθροισμα των απαντήσεων των συμμετεχόντων στα διάφορα στοιχεία Likert που συνιστούν την κλίμακα. Κάθε στοιχείο Likert αποτελεί μια πρόταση στην οποία οι συμμετέχοντες καλούνται να δηλώσουν το βαθμό συμφωνίας τους (ή βαθμό διαφωνίας τους). Συνήθως υπάρχουν 5 (ή σπανιότερα 7 ή 9) απαντήσεις σε διατεταγμένη κλίμακα και οι συμμετέχοντες καλούνται να επιλέξουν αυτή που τους εκφράζει περισσότερο. Η τυπική δομή ενός στοιχείου Likert στο οποίο υπάρχουν 5 πιθανές απαντήσεις σε διατεταγμένη κλίμακα είναι η εξής: (α) Διαφωνώ τελείως, (β) διαφωνώ, (γ) ούτε διαφωνώ ούτε συμφωνώ, (δ) συμφωνώ, (ε) συμφωνώ τελείως.

Σε ορισμένες περιπτώσεις εξ άλλου χρησιμοποιείται η λεγόμενη «υποχρεωτική επιλογή» ("forced choice"), στην οποία σ' ένα στοιχείο Likert υπάρχουν 4 απαντήσεις σε διατεταγμένη κλίμακα, καθώς αφαιρείται η ενδιάμεση επιλογή («ούτε διαφωνώ, ούτε συμφωνώ»), έτσι ώστε οι συμμετέχοντες να «αναγκαστούν» να συμφωνήσουν ή να διαφωνήσουν με το συγκεκριμένο στοιχείο.

Σημειώνεται ότι τα δεδομένα των κλιμάκων Likert, μολονότι είναι διατάξιμα, σε αρκετές περιπτώσεις αντιμετωπίζονται ως δεδομένα κλίμακας λόγου διευκολύνοντας σημαντικά τη στατιστική ανάλυση. Από στατιστική άποψη, η προσέγγιση αυτή δεν είναι η πλέον ενδεδειγμένη, αλλά εφ' όσον επιλεγεί πρέπει να αναφέρονται με σαφήνεια τα κριτήρια και οι προϋποθέσεις εφαρμογής της.

Χαρακτηριστικά παραδείγματα διατάξιμων μεταβλητών αποτελούν η κλίμακα Γλασκόβης** (Glasgow scale) και η κλίμακα Apgar*** (Apgar scale). Η κλίμακα Γλασκόβης για ενήλικες χρησιμοποιείται για την εκτίμηση του επιπέδου συνείδησης και αποτελείται από τρία στοιχεία: Την αντίδρα-

* Οι κλίμακες Likert εισήχθησαν για πρώτη φορά στις επιστήμες υγείας το 1932 από τον Αμερικανό ψυχολόγο Rensis Likert (1903–1981).

** Η κλίμακα Γλασκόβης εισήχθη το 1974 από τους Graham Teasdale και Bryan Jennette, καθηγητές Νευροχειρουργικής στο Πανεπιστήμιο της Γλασκόβης.

*** Η κλίμακα Apgar εισήχθη το 1952 από την Αμερικανή αναισθησιολόγο Virginia Apgar (1909–1974) και χρησιμοποιείται για την εκτίμηση της γενικής κατάστασης των νεογνών, το πρώτο και το πέμπτο λεπτό μετά τη γέννησή τους. Η κλίμακα Apgar αποτελείται από πέντε στοιχεία (συχνότητα καρδιακών παλμών, αναπνευστική λειτουργία, μυϊκός τόνος, χρώμα δέρματος και αντανακλαστικά), καθένα από τα οποία λαμβάνει τιμές 0–2. Έτσι, η εν λόγω κλίμακα λαμβάνει τιμές 0–10, με τις μεγαλύτερες από αυτές να δηλώνουν καλύτερη κατάσταση του νεογνού.

ση των οφθαλμών, τη λεκτική και την κινητική αντίδραση. Αναφορικά με την αντίδραση των οφθαλμών λαμβάνονται τιμές 1–4 (1: απουσία ανοίγματος οφθαλμών, 2: άνοιγμα οφθαλμών ως αντίδραση στο άλγος, 3: άνοιγμα οφθαλμών ως αντίδραση στα λεκτικά παραγγέλματα, 4: αυτόματο άνοιγμα οφθαλμών), σχετικά με τη λεκτική αντίδραση λαμβάνονται τιμές 1–5 (1: καμιά λεκτική απάντηση, 2: ακατάληπτοι ήχοι, 3: ακατάληπτες λέξεις, 4: συγχυτική ομιλία, 5: προσανατολισμένη ομιλία) και όσον αφορά στην κινητική αντίδραση λαμβάνονται τιμές 1–6 (1: καμιά κινητική αντίδραση, 2: έκταση ως αντίδραση στο άλγος, 3: κάμψη ως αντίδραση στο άλγος, 4: απόσυρση ως αντίδραση στο άλγος, 5: εντοπισμός του σημείου του άλγους, 6: «υπακοή» σε λεκτικά παραγγέλματα). Έτσι, η κλίμακα Γλασκόβης λαμβάνει τιμές 3–15, με τις μεγαλύτερες τιμές να δηλώνουν υψηλότερο επίπεδο συνείδησης και μικρότερη βλάβη της εγκεφαλικής λειτουργίας.

Σε αρκετές περιπτώσεις, τα δεδομένα κλιμάκων, όπως η κλίμακα Γλασκόβης, μολονότι είναι διατάξιμα, αντιμετωπίζονται ως δεδομένα κλίμακας λόγου διευκολύνοντας έτσι σημαντικά τη στατιστική ανάλυση. Στην περίπτωση αυτή, όσο μεγαλύτερο είναι το εύρος των τιμών που μπορεί να λάβει μια κλίμακα τόσο πιο αξιόπιστη είναι η στατιστική ανάλυση. Για παράδειγμα, είναι αποτελεσματικότερο από στατιστική άποψη να αντιμετωπιστεί ως μεταβλητή κλίμακας λόγου η κλίμακα APACHE II,* που λαμβάνει τιμές 0–71, σε σχέση με την κλίμακα Γλασκόβης, η οποία λαμβάνει τιμές 3–15.

2.2. Ποσοτικές μεταβλητές

2.2.1. Μεταβλητές διαστηματικής κλίμακας. Στις μεταβλητές διαστηματικής κλίμακας (interval scale variates) έχει σημασία τόσο η διάταξη όσο και το μέγεθος, με τους αριθμούς να αντιπροσωπεύουν πραγματικές μετρήσιμες ποσότητες και όχι ονόματα ή χαρακτηρισμούς τάξεων.[11,13] Τα διαστήματα που χωρίζουν τη μια τιμή της μεταβλητής από την άλλη έχουν διαφορετική βαρύτητα στα διάφορα σημεία της κλίμακας. Χαρακτηριστικό παράδειγμα της συγκεκριμένης κατηγορίας αποτελεί η κλίμακα μέτρησης της έντασης των σεισμών ή, αλλιώς, κλίμακα Ρίχτερ, στην οποία οι τιμές της μεταβλητής μπορεί να απέχουν μεταξύ τους κατά ίση αριθμητική διαφορά, αλλά αυτό δεν συνεπάγεται κιόλας ότι οι εν λόγω διαφορές είναι ισοδύναμες. Στην περίπτωση της κλίμακας Ρίχτερ, η διαφορά της έντασης μεταξύ δύο σεισμών 4 και 5 βαθμών της κλίμακας Ρίχτερ δεν είναι ισοδύναμη με τη διαφορά μεταξύ δύο σεισμών 7 και 8 βαθμών της κλίμακας Ρίχτερ. Τα δεδομένα διαστηματικής κλίμακας είναι μαθηματικά αξιοποιήσιμα, αλλά δεν υφίσταται η σχέση του πολλαπλασιασμού μεταξύ των τιμών τους. Έτσι, έχει νόημα ότι ένας σεισμός 8 βαθμών της κλίμακας Ρίχτερ είναι ισχυρότερος από ένα σεισμό 4 βαθμών της κλίμακας Ρίχτερ, αλλά όχι ότι ο πρώτος είναι δύο φορές πιο ισχυρός από το δεύτερο.

Στις μεταβλητές διαστηματικής κλίμακας δεν υπάρχει πραγματικό μηδενικό σημείο έναρξης της μέτρησης. Για παράδειγμα, 0 βαθμοί Fahrenheit δεν σημαίνει ότι δεν υπάρχει θερμοκρασία. Ακριβώς επειδή οι μεταβλητές διαστηματικής κλίμακας δεν έχουν ένα πραγματικό μηδενικό σημείο έναρξης της μέτρησης, αλλά ένα αυθαίρετο, δεν είναι δυνατόν να πραγματοποιηθεί η πρόσθεση, η αφαίρεση, ο πολλαπλασιασμός και η διαίρεση των αντίστοιχων δεδομένων. Σημειώνεται ότι στην πράξη τα δεδομένα διαστηματικής κλίμακας χρησιμοποιούνται ελάχιστα στις επιστήμες υγείας.

2.2.2. Μεταβλητές κλίμακας λόγου. Οι μεταβλητές κλίμακας λόγου (ratio scale variates) υπερέχουν έναντι των μεταβλητών διαστηματικής κλίμακας στο ότι διαθέτουν ένα φυσικό μηδενικό σημείο έναρξης της μέτρησης. Οι περισσότερες μεταβλητές (όπως το βάρος, το ύψος, η ηλικία, το εισόδημα, οι δαπάνες υγείας κ.ά.) ανήκουν στην κατηγορία αυτή. Στην περίπτωση των μεταβλητών κλίμακας λόγου είναι δυνατόν να πραγματοποιηθεί η πρόσθεση, η αφαίρεση, ο πολλαπλασιασμός και η διαίρεση των δεδομένων, καθώς υπάρχει ένα πραγματικό σημείο μηδέν. Για παράδειγμα, ένα άτομο με ετήσιο εισόδημα ίσο με 30.000€ έχει διπλάσιο εισόδημα από ένα άτομο με ετήσιο εισόδημα ίσο με 15.000€. Επί πλέον, ένα άτομο σωματικού βάρους 100 kg είναι δύο φορές βαρύτερο από ένα άτομο σωματικού βάρους 50 kg.

2.2.3. Συνεχείς και ασυνεχείς μεταβλητές. Οι ποσοτικές μεταβλητές εξ άλλου διακρίνονται σε συνεχείς (continuous) και ασυνεχείς ή, αλλιώς, διακριτές (discrete).[3–13] Η διαφορά τους έγκειται στο γεγονός ότι οι ασυνεχείς μεταβλητές περιορίζονται στη λήψη μόνο ακέραιων τιμών, που διαφέρουν μεταξύ τους κατά συγκεκριμένες ποσότητες, ενώ οι συνεχείς μεταβλητές μπορούν να λάβουν οποιαδήποτε τιμή εντός των πραγματικών αριθμών, ακόμη και δεκαδική. Οι περισσότερες μεταβλητές διαστηματικής κλίμακας και κλίμακας λόγου αποτελούν συνεχείς μεταβλητές, γεγονός που διευκολύνει σημαντικά τη στατιστική ανάλυση. Παραδείγματα ασυνεχών μεταβλητών αποτελούν ο αριθμός των τροχαίων ατυχημάτων, ο αριθμός των γεννήσεων μιας γυναίκας, ο αριθμός των νέων περιπτώσεων μιας πάθησης που καταγράφηκαν σε μια χώρα σε ένα συγκεκριμένο χρονικό διάστημα κ.ά., ενώ παραδείγματα συνεχών

* Η κλίμακα APACHE II (acute physiology and chronic health evaluation II) χρησιμοποιείται για την εκτίμηση της βαρύτητας της κατάστασης των πασχόντων στις μονάδες εντατικής θεραπείας και λαμβάνει τιμές 0–71, με τις μεγαλύτερες από αυτές να δηλώνουν πιο βαριά κατάσταση και μεγαλύτερο κίνδυνο θανάτου.

μεταβλητών αποτελούν η ηλικία, το βάρος, το ύψος, το εισόδημα, η αρτηριακή πίεση, ο δείκτης μάζας σώματος κ.ά. Οι ποσοτικές μεταβλητές, σε αντίθεση με τις ποιοτικές, έχουν μονάδα μέτρησης, όπως π.χ. τα κιλά για το σωματικό βάρος, τα εκατοστά για το ύψος, τα έτη για την ηλικία κ.λπ.

Σημειώνεται ότι οι ποσοτικές μεταβλητές είναι δυνατόν να μετατραπούν σε διατάξιμες μεταβλητές, χρησιμοποιώντας κάποια διαχωριστικά όρια, ενώ το αντίθετο δεν είναι δυνατόν να συμβεί. Για παράδειγμα, ανάλογα με το δείκτη μάζας σώματος, τα άτομα μπορούν να ταξινομηθούν σε αδύνατα, φυσιολογικά, υπέρβαρα ή παχύσαρκα. Επί πλέον, οι τιμές της χοληστερόλης ορού είναι δυνατόν να ταξινομηθούν σε χαμηλές, φυσιολογικές, υψηλές και πολύ υψηλές τιμές ή απλά σε τιμές μεγαλύτερες ή μικρότερες ενός διαχωριστικού ορίου.

Στον πίνακα 2 φαίνονται τα βασικά χαρακτηριστικά των συμμετεχόντων και τα διάφορα είδη μεταβλητών σε μια υποθετική μελέτη «ασθενών-μαρτύρων».

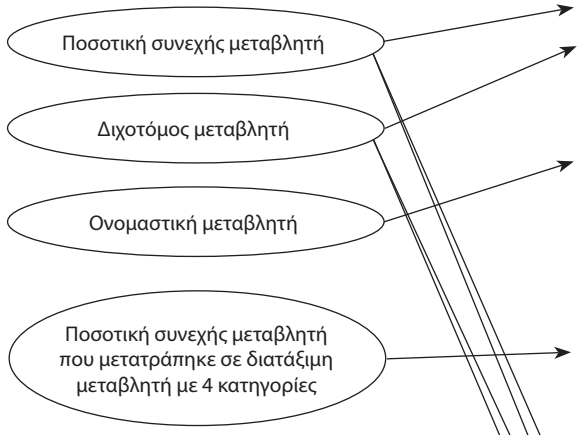
3. ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ

Η συνοπτική παρουσίαση των δεδομένων των επιδημιολογικών μελετών μπορεί να πραγματοποιηθεί είτε με τη χρήση πινάκων είτε με τη χρήση γραφημάτων ή, αλλιώς, γραφικών παραστάσεων.[3,4,6,7,9,11,14–22]

3.1. Πίνακες συχνοτήτων

Οι πίνακες συχνοτήτων (frequencies tables) χρησιμοποιούνται για τη συνοπτική παρουσίαση δεδομένων που αφορούν σε όλα τα είδη των μεταβλητών.[3,4,6,7,9,11,14–22] Συνήθως, στους πίνακες συχνοτήτων περιλαμβάνονται οι απόλυτες συχνότητες, οι σχετικές συχνότητες και οι αθροιστικές σχετικές συχνότητες. Η *απόλυτη συχνότητα* (absolute frequency) εκφράζεται ως απόλυτος αριθμός και αφορά στον αριθμό των παρατηρήσεων για μια συγκεκριμένη κατηγορία ή τάξη της μελετώμενης μεταβλητής. Η *σχετική συχνότητα* (relative frequency) εκφράζεται ως ποσοστό και προκύπτει από τη διαίρεση των παρατηρήσεων σε μια συγκεκριμένη κατηγορία με το συνολικό αριθμό των παρατηρήσεων που αφορούν στη μελετώμενη μεταβλητή. Η *αθροιστική σχετική συχνότητα* (cumulative relative frequency) εκφράζεται ως ποσοστό και για μια συγκεκριμένη κατηγορία της μελετώμενης μεταβλητής προκύπτει από την πρόσθεση των σχετικών συχνοτήτων για την εν λόγω κατηγορία και όλες τις προηγούμενες κατηγορίες. Οι σχετικές και οι αθροιστικές σχετικές συχνότητες είναι

Πίνακας 2. Τα βασικά χαρακτηριστικά των συμμετεχόντων και τα διάφορα είδη μεταβλητών σε μια υποθετική μελέτη «ασθενών-μαρτύρων».



	Ασθενείς (n=100)	«Μάρτυρες» (n=200)
Ηλικία*	43 (1,7)	44,5 (1,5)
Φύλο (n, %)		
Άνδρες	30 (30)	66 (33)
Γυναίκες	70 (70)	134 (67)
Ομάδα αίματος (n, %)		
A	38 (38)	74 (37)
B	14 (14)	26 (13)
AB	44 (44)	90 (45)
O	4 (4)	10 (5)
Καπνισματική συνήθεια (n, %)		
0 τσιγάρα/ημέρα	5 (5)	50 (25)
1–20 τσιγάρα/ημέρα	25 (25)	50 (25)
21–40 τσιγάρα/ημέρα	30 (30)	60 (30)
>40 τσιγάρα/ημέρα	40 (40)	40 (20)
Συστολική αρτηριακή πίεση*	157,8 (0,4)	129,2 (0,5)
Διαστολική αρτηριακή πίεση*	97,9 (0,3)	82 (0,3)
Διαβήτης (n, %)		
Ναι	20 (20)	20 (10)
Όχι	80 (80)	180 (90)

* Μέση τιμή (τυπική απόκλιση)

χρήσιμες για τη σύγκριση ομάδων με διαφορετικό αριθμό παρατηρήσεων.

Σημειώνεται ότι στην περίπτωση των ποσοτικών μεταβλητών είναι απαραίτητη η κατηγοριοποίησή τους σε τάξεις με όρια, τα οποία δεν επικαλύπτονται μεταξύ τους. Όσο αυξάνεται ο αριθμός των τάξεων τόσο αυξάνει και η πιθανότητα ο πίνακας συχνοτήτων να μη διευκολύνει τη συνοπτική απεικόνιση των δεδομένων. Επί πλέον, όσο μειώνεται ο αριθμός των τάξεων τόσο ελαττώνεται και η πληροφορία που παρέχει ο πίνακας συχνοτήτων.

Στον πίνακα 3 παρουσιάζονται οι απόλυτες, οι σχετικές και οι αθροιστικές σχετικές συχνότητες 200 ατόμων ηλικίας ≥18 ετών αναφορικά με το φύλο, την ηλικία, το βάρος, τον αριθμό των παιδιών και το εκπαιδευτικό επίπεδο. Η ηλικία κατηγοριοποιήθηκε σε 6 τάξεις, το βάρος σε 5, ο αριθμός των παιδιών σε 3 και το εκπαιδευτικό επίπεδο σε 3. Σε όλες τις περιπτώσεις, τα όρια των τάξεων δημιουργήθηκαν κατά τέτοιον τρόπο ώστε να μην επικαλύπτονται μεταξύ τους. Ενδεικτικά αναφέρεται ότι, αναφορικά με τον αριθμό των παιδιών, τα 200 άτομα ταξινομήθηκαν σε τρεις κατηγορίες: Σε αυτά που δεν είχαν παιδιά, σε αυτά που είχαν 1–3 παιδιά και σε αυτά που είχαν τουλάχιστον 4 παιδιά. Με βάση τον πίνακα 3 προκύπτει ότι το 30% των ατόμων δεν είχε παιδί, το 60% είχε 1–3 παιδιά και το 10% είχε τουλάχιστον 4 παιδιά. Η αθροιστική σχετική συχνότητα στην κατηγορία «1–3 παιδιά» (=90%) προκύπτει από την πρόσθεση των σχετικών συχνοτήτων για τη συγκεκριμένη κατηγορία και όλες τις προηγούμενες κατηγορίες, όπου σε αυτή την περίπτωση είναι μόνο η κατηγορία «0 παιδιά».

Πίνακας 3. Απόλυτες, σχετικές και αθροιστικές σχετικές συχνότητες 200 ατόμων ηλικίας ≥18 ετών αναφορικά με το φύλο, την ηλικία, το σωματικό βάρος, τον αριθμό των παιδιών και το εκπαιδευτικό επίπεδο.

	Απόλυτη συχνότητα (αριθμός περιπτώσεων)	**Σχετική συχνότητα (%)**	**Αθροιστική σχετική συχνότητα (%)**
Φύλο			
Άνδρες	90	45	45
Γυναίκες	110	55	100
Ηλικία (έτη)			
18–28	26	13	13
29–38	42	21	34
39–48	52	26	60
49–58	32	16	76
59–68	26	13	89
>68	22	11	100
Σωματικό βάρος (kg)			
<60	16	8	8
60–70	76	38	46
71–80	66	33	79
81–90	26	13	92
>90	16	8	100
Αριθμός παιδιών			
0	60	30	30
1–3	120	60	90
>3	20	10	100
Εκπαιδευτικό επίπεδο			
Κατώτερο	30	15	15
Μεσαίο	140	70	85
Ανώτερο	30	15	100

Όπως προαναφέρθηκε, οι σχετικές και οι αθροιστικές σχετικές συχνότητες είναι χρήσιμες για τη σύγκριση ομάδων με διαφορετικό αριθμό παρατηρήσεων, γεγονός που δεν μπορεί να επιτευχθεί με τη χρήση των απόλυτων συχνοτήτων. Στον πίνακα 4 φαίνονται οι απόλυτες, οι σχετικές και οι αθροιστικές σχετικές συχνότητες των 90 ανδρών και των 110 γυναικών του πίνακα 3 αναφορικά με την ηλικία, το βάρος, τον αριθμό των παιδιών και το εκπαιδευτικό επίπεδο. Παρατηρώντας τον πίνακα 4, αναφορικά με το εκπαιδευτικό επίπεδο, ο ίδιος αριθμός ανδρών (n=15) και γυναικών (n=15) ανήκουν στο κατώτερο εκπαιδευτικό επίπεδο, αλλά οι σχετικές συχνότητες διαφέρουν (για τους άνδρες η σχετική συχνότητα είναι 16,7 και για τις γυναίκες είναι 13,6), καθώς ο συνολικός αριθμός των παρατηρήσεων είναι μικρότερος στους άνδρες (n=90) σε σχέση με τις γυναίκες (n=110). Έτσι, με βάση τις σχετικές συχνότητες προκύπτει το συμπέρασμα ότι το ποσοστό των ανδρών που ανήκουν στο κατώτερο εκπαιδευτικό επίπεδο είναι μικρότερο από το αντίστοιχο ποσοστό των γυναικών.

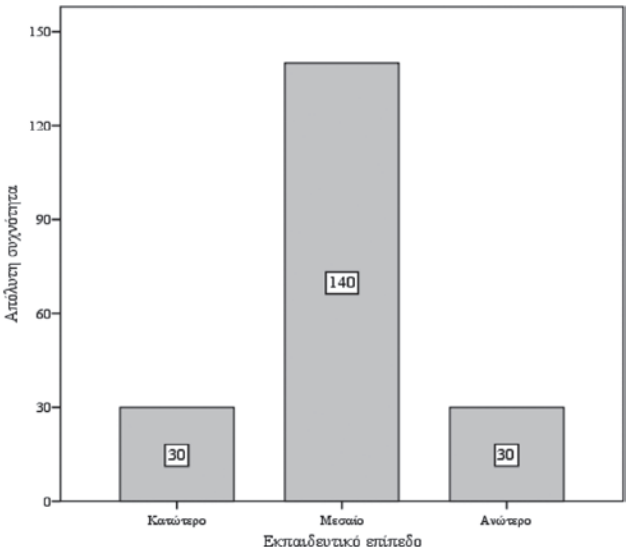
3.2. Γραφήματα

Τα γραφήματα συνήθως είναι ευκολότερα στην κατανόηση από τους πίνακες, αλλά χαρακτηρίζονται από μικρότερο βαθμό λεπτομέρειας. Τα στατιστικά προγράμματα ανάλυσης δεδομένων παρέχουν τη δυνατότητα δημιουργίας αρκετών γραφημάτων που διευκολύνουν σημαντικά τη συνοπτική παρουσίαση των δεδομένων των μελετών. Τα πλέον συνήθη γραφήματα είναι τα ραβδογράμματα, τα κυκλικά διαγράμματα, τα ιστογράμματα, τα διαγράμματα πλαισίου και τα διαγράμματα σημείων.[3,4,6,7,9,11,14–22]

3.2.1. Ραβδογράμματα. Τα ραβδογράμματα χρησιμοποιούνται στην περίπτωση των ποιοτικών μεταβλητών και ιδιαίτερα όταν ο αριθμός των κατηγοριών είναι σχετικά μικρός. Στην εικόνα 2 φαίνεται το ραβδόγραμμα (bar chart)

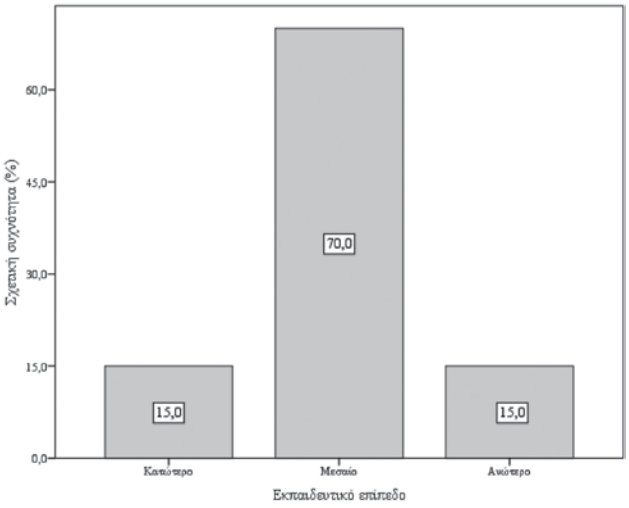
Πίνακας 4. Απόλυτες, σχετικές και αθροιστικές σχετικές συχνότητες των 90 ανδρών και των 110 γυναικών του πίνακα 3 αναφορικά με την ηλικία, το σωματικό βάρος, τον αριθμό των παιδιών και το εκπαιδευτικό επίπεδο.

	Άνδρες (n=90)			Γυναίκες (n=110)		
	Απόλυτη συχνότητα (αριθμός περιπτώσεων)	Σχετική συχνότητα (%)	Αθροιστική σχετική συχνότητα (%)	Απόλυτη συχνότητα (αριθμός περιπτώσεων)	Σχετική συχνότητα (%)	Αθροιστική σχετική συχνότητα (%)
Ηλικία (έτη)						
18–28	14	15,6	15,6	12	10,9	10,9
29–38	24	26,7	42,3	18	16,4	27,3
39–48	20	22,2	64,5	32	29,1	56,4
49–58	18	20,0	84,5	14	12,7	69,1
59–68	10	11,1	95,6	16	14,5	83,6
>68	4	4,4	100,0	18	16,4	100,0
Σωματικό βάρος (kg)						
<60	2	2,2	2,2	14	12,7	12,7
60–70	16	17,8	20,0	60	54,5	67,2
71–80	41	45,6	65,6	25	22,7	89,9
81–90	18	20,0	85,6	8	7,3	97,2
>90	13	14,4	100,0	3	2,8	100,0
Αριθμός παιδιών						
0	35	38,9	38,9	25	22,7	22,7
1–3	45	50,0	88,9	75	68,2	90,9
>3	10	11,1	100,0	10	9,1	100,0
Εκπαιδευτικό επίπεδο						
Κατώτερο	15	16,7	16,7	15	13,6	13,6
Μεσαίο	60	66,6	83,3	80	72,8	86,4
Ανώτερο	15	16,7	100,0	15	13,6	100,0

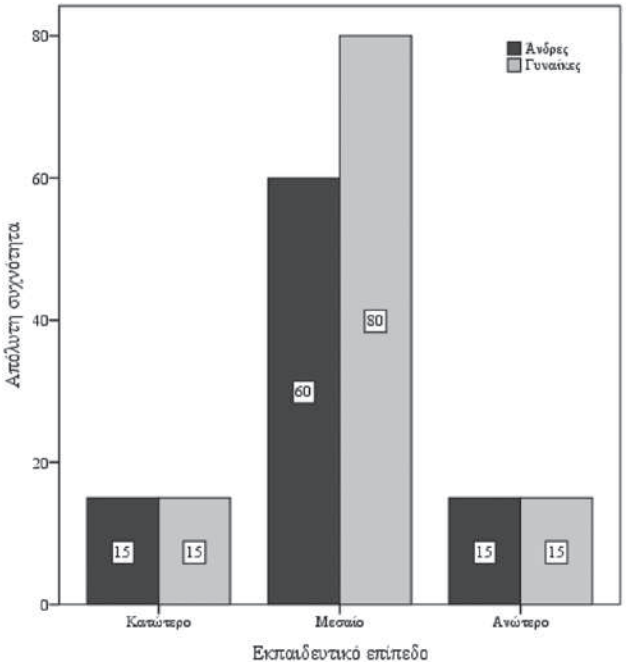


Εικόνα 2. Ραβδόγραμμα απόλυτων συχνοτήτων αναφορικά με το εκπαιδευτικό επίπεδο των 200 ατόμων του πίνακα 3.

απόλυτων συχνοτήτων αναφορικά με το εκπαιδευτικό επίπεδο των 200 ατόμων του πίνακα 3, ενώ στην εικόνα 3 φαίνεται το αντίστοιχο ραβδόγραμμα σχετικών συχνοτήτων. Στα ραβδογράμματα, οι διάφορες κατηγορίες στις οποίες ανήκουν οι παρατηρήσεις παρουσιάζονται στον οριζόντιο άξονα ή, αλλιώς, άξονα των x, ενώ στον κάθετο άξονα ή, αλλιώς, άξονα των y εμφανίζονται οι απόλυτες και οι σχετικές συχνότητες. Οι κάθετες στήλες ή, αλλιώς, ράβδοι που αντιστοιχούν στις διάφορες κατηγορίες του οριζόντιου άξονα σχεδιάζονται κατά τέτοιον τρόπο ώστε το ύψος τους να αντιπροσωπεύει την απόλυτη ή τη σχετική συχνότητα των παρατηρήσεων. Οι στήλες πρέπει να έχουν ίσο πλάτος και να υπάρχει απόσταση μεταξύ τους, έτσι ώστε να μην υποδηλώνεται συνέχεια. Στην εικόνα 4 φαίνεται το ραβδόγραμμα απόλυτων συχνοτήτων των 90 ανδρών και των 110 γυναικών του πίνακα 4 αναφορικά με το εκπαιδευτικό επίπεδο, ενώ στην εικόνα 5 φαίνεται το αντίστοιχο ραβδόγραμμα σχετικών συχνοτήτων.



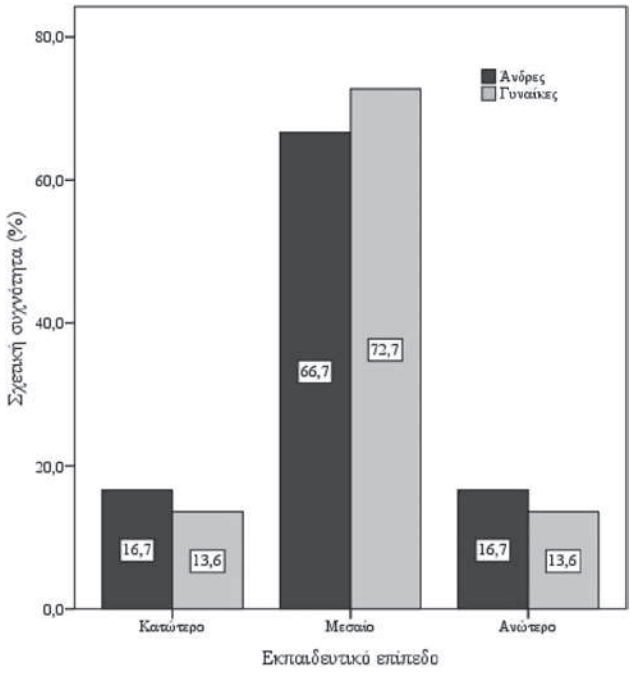
Εικόνα 3. Ραβδόγραμμα σχετικών συχνοτήτων αναφορικά με το εκπαιδευτικό επίπεδο των 200 ατόμων του πίνακα 3.



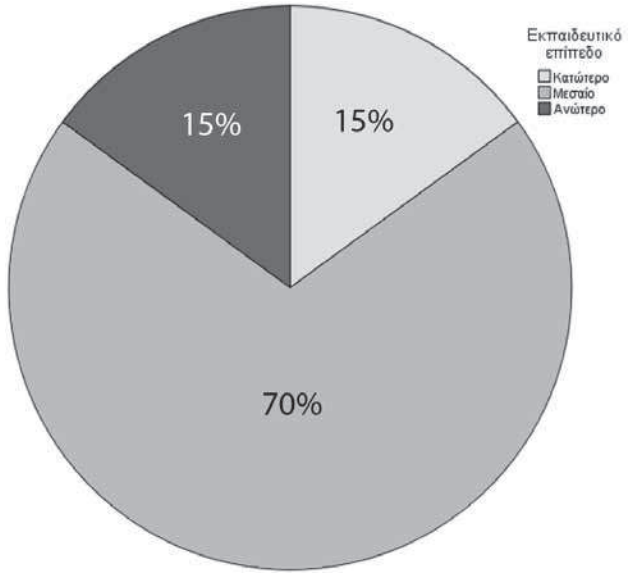
Εικόνα 4. Ραβδόγραμμα απόλυτων συχνοτήτων των 90 ανδρών και των 110 γυναικών του πίνακα 4 αναφορικά με το εκπαιδευτικό επίπεδο.

3.2.2. Κυκλικά διαγράμματα. Τα κυκλικά διαγράμματα ή, αλλιώς, διαγράμματα σε μορφή «πίτας» (pie charts) είναι εξαιρετικά απλά στο σχεδιασμό τους και χρησιμοποιούνται στην περίπτωση των ποιοτικών μεταβλητών, αλλά λιγότερο συχνά από τα αντίστοιχα ραβδογράμματα. Στην εικόνα 6 φαίνεται το κυκλικό διάγραμμα σχετικών συχνοτήτων αναφορικά με το εκπαιδευτικό επίπεδο των 200 ατόμων του πίνακα 3.

3.2.3. Ιστογράμματα. Τα ιστογράμματα χρησιμοποιούνται

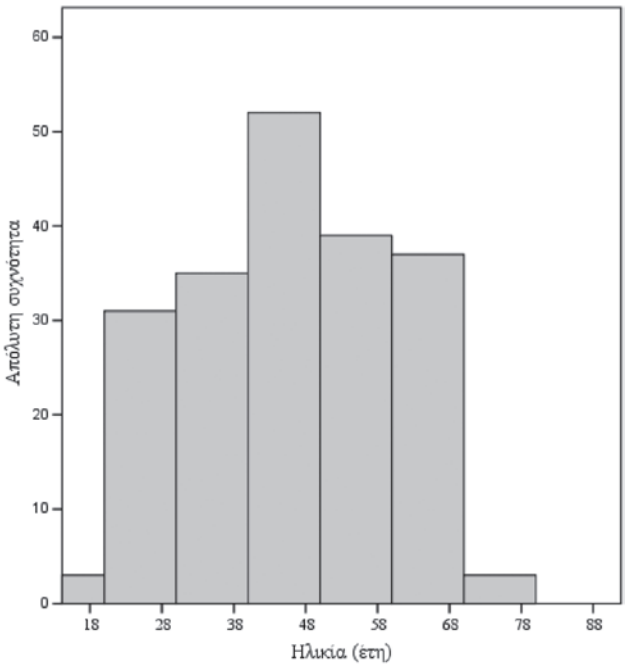


Εικόνα 5. Ραβδόγραμμα σχετικών συχνοτήτων των 90 ανδρών και των 110 γυναικών του πίνακα 4 αναφορικά με το εκπαιδευτικό επίπεδο.



Εικόνα 6. Κυκλικό διάγραμμα σχετικών συχνοτήτων αναφορικά με το εκπαιδευτικό επίπεδο των 200 ατόμων του πίνακα 3.

στην περίπτωση των ποσοτικών μεταβλητών. Στην εικόνα 7 φαίνεται το ιστόγραμμα (histogram) απόλυτων συχνοτήτων της ηλικίας των 200 ατόμων του πίνακα 3, διαιρώντας την ηλικία στον οριζόντιο άξονα ανά δεκαετή διαστήματα. Στον οριζόντιο άξονα παρουσιάζονται τα όρια των διαφόρων διαστημάτων στα οποία χωρίζεται η ποσοτική μεταβλητή, ενώ στον κάθετο άξονα παρουσιάζονται οι απόλυτες ή οι

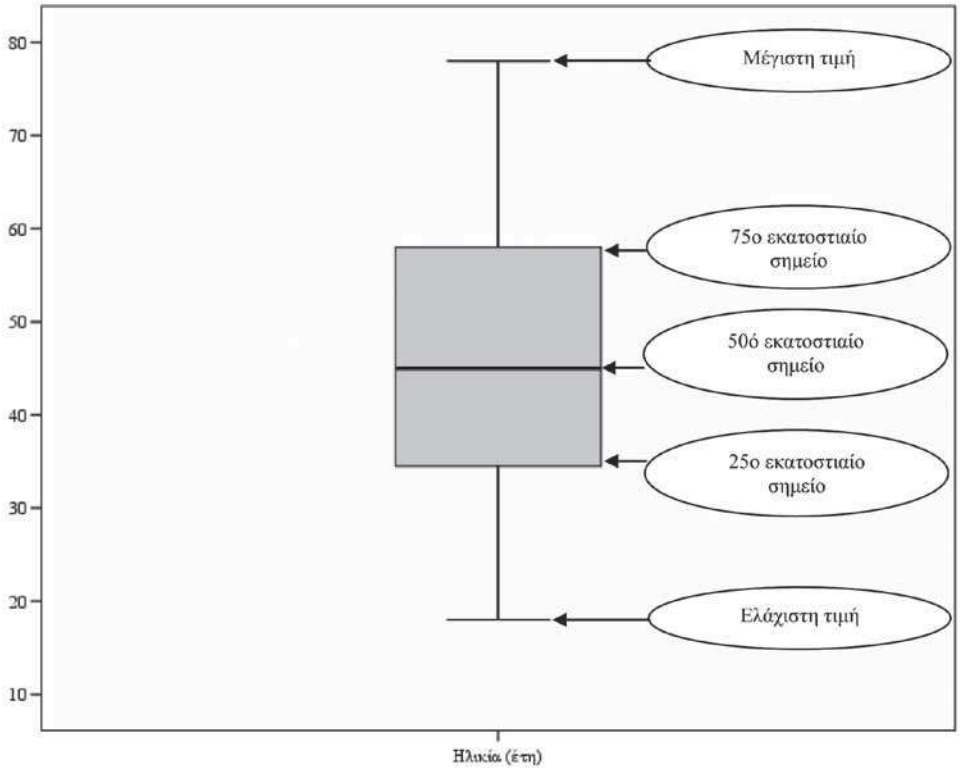


Εικόνα 7. Ιστόγραμμα απόλυτων συχνοτήτων της ηλικίας των 200 ατόμων του πίνακα 3, διαιρώντας την ηλικία στον οριζόντιο άξονα ανά δεκαετή διαστήματα.

σχετικές συχνότητες των παρατηρήσεων σε κάθε διάστημα. Η κλίμακα μέτρησης στον κάθετο άξονα πρέπει να αρχίζει από το 0. Οι στήλες πρέπει να έχουν ίσο πλάτος και να μην υπάρχει απόσταση μεταξύ τους, έτσι ώστε να υποδηλώνεται η συνέχεια της ποσοτικής μεταβλητής.

3.2.4. Διαγράμματα πλαισίου. Τα διαγράμματα πλαισίου χρησιμοποιούνται στην περίπτωση των ποσοτικών μεταβλητών και παρέχουν πολύ σημαντική πληροφορία. Πιο συγκεκριμένα, σε ένα διάγραμμα πλαισίου (box plot) απεικονίζονται η ελάχιστη και η μέγιστη τιμή, καθώς και το 25ο, το 50ό (ή, αλλιώς, διάμεσος) και το 75ο εκατοστιαίο σημείο των παρατηρήσεων. Γνωρίζοντας εξ άλλου το 25ο και το 75ο εκατοστιαίο σημείο των παρατηρήσεων είναι γνωστό και το ενδοτεταρτημοριακό εύρος. Επί πλέον, στα διαγράμματα πλαισίου απεικονίζονται τόσο οι απομακρυσμένες όσο και οι ακραίες παρατηρήσεις. Το κάτω και το άνω άκρο του πλαισίου συνιστούν το 25ο και το 75ο εκατοστιαίο σημείο, αντίστοιχα. Οι δύο οριζόντιες γραμμές που συνδέονται κάθετα με το πλαίσιο και ευρίσκονται η μια κάτω από το κάτω άκρο του πλαισίου και η άλλη πάνω από το άνω άκρο του πλαισίου συνιστούν, αντίστοιχα, την ελάχιστη και τη μέγιστη τιμή των παρατηρήσεων. Οι απομακρυσμένες παρατηρήσεις συμβολίζονται με κύκλους, ενώ οι ακραίες παρατηρήσεις με αστερίσκους.

Στην εικόνα 8 φαίνεται το διάγραμμα πλαισίου της ηλικίας των 200 ατόμων του πίνακα 3. Με βάση την εικόνα 8 προκύπτει ότι η ελάχιστη τιμή της ηλικίας ισούται με 18 έτη, η μέγιστη τιμή ισούται με 78, το 25ο εκατοστιαίο σημείο ισούται με 34, το 50ό εκατοστιαίο σημείο ισούται

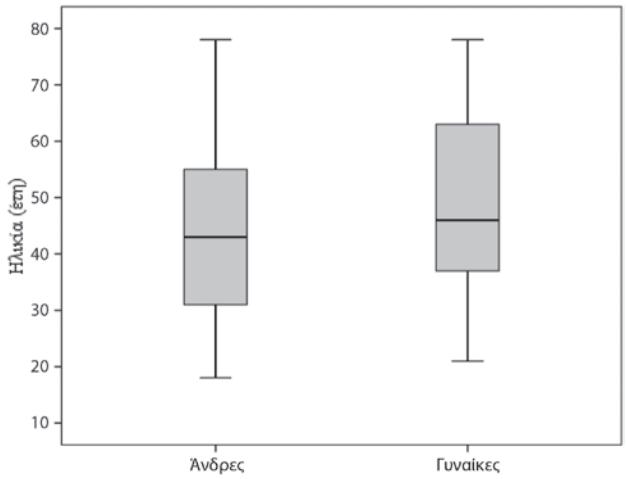


Εικόνα 8. Διάγραμμα πλαισίου της ηλικίας των 200 ατόμων του πίνακα 3.

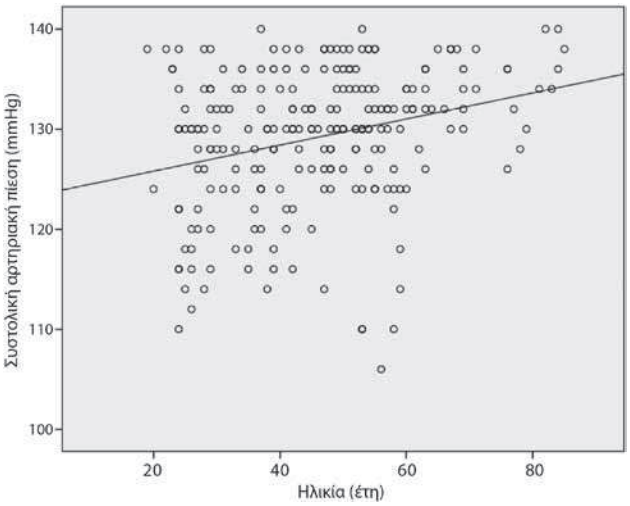
με 45 και το 75ο εκατοστιαίο σημείο ισούται με 58. Έτσι, η διάμεσος ισούται με 45 έτη και το ενδοτεταρτημοριακό εύρος ισούται με 24.

Σημειώνεται ότι τα διαγράμματα πλαισίου είναι ιδιαίτερα χρήσιμα για τη σύγκριση ομάδων με διαφορετικό αριθμό παρατηρήσεων. Στην εικόνα 9 φαίνονται τα διαγράμματα πλαισίου της ηλικίας των 90 ανδρών και των 110 γυναικών του πίνακα 4. Με βάση την εικόνα 9, αναφορικά με τους άνδρες, προκύπτει ότι η ελάχιστη τιμή της ηλικίας ισούται με 18 έτη, η μέγιστη τιμή ισούται με 78, το 25ο εκατοστιαίο σημείο ισούται με 31, το 50ό εκατοστιαίο σημείο ισούται με 43 και το 75ο εκατοστιαίο σημείο ισούται με 55. Αναφορικά με τις γυναίκες εξ άλλου προκύπτει ότι η ελάχιστη τιμή της ηλικίας ισούται με 21 έτη, η μέγιστη τιμή ισούται με 78, το 25ο εκατοστιαίο σημείο ισούται με 37, το 50ό εκατοστιαίο σημείο ισούται με 46 και το 75ο εκατοστιαίο σημείο ισούται με 63.

3.2.5. Διαγράμματα σημείων. Τα διαγράμματα σημείων ή, αλλιώς, διαγράμματα σκεδασμού χρησιμοποιούνται για την απεικόνιση της σχέσης μεταξύ δύο ποσοτικών μεταβλητών. Στον οριζόντιο άξονα ή, αλλιώς, άξονα των x απεικονίζεται η κλίμακα μέτρησης της μιας ποσοτικής μεταβλητής, ενώ στον κάθετο άξονα ή, αλλιώς, άξονα των y απεικονίζεται η κλίμακα μέτρησης της δεύτερης ποσοτικής μεταβλητής. Στην εικόνα 10 φαίνεται το διάγραμμα σημείων (scatter plot) 230 ατόμων που απεικονίζει τη σχέση μεταξύ της ηλικίας και της συστολικής αρτηριακής πίεσης. Στην εικόνα αυτή φαίνεται ότι υπάρχει γραμμική σχέση μεταξύ της ηλικίας και της συστολικής αρτηριακής πίεσης, καθώς η αύξηση της πρώτης μεταβλητής σχετίζεται γραμμικά με την αύξηση της δεύτερης.



Εικόνα 9. Διαγράμματα πλαισίου της ηλικίας των 90 ανδρών και των 110 γυναικών του πίνακα 4.



Εικόνα 10. Διάγραμμα σημείων 230 ατόμων, που απεικονίζει τη σχέση μεταξύ της ηλικίας και της συστολικής αρτηριακής πίεσης.

4. ΜΕΤΡΑ ΘΕΣΗΣ

4.1. Μέσος

Ο αριθμητικός μέσος (arithmetic mean) ή μέσος όρος (average) ή απλά μέσος (mean) αποτελεί το συνηθέστερο μέτρο θέσης.[3–7,9,11,14,15,17–21,23–25] Ο μέσος αποτελεί ουσιαστικά τη μέση τιμή των παρατηρήσεων αναφορικά με μια μεταβλητή και προκύπτει από τη διαίρεση του αθροίσματος των τιμών όλων των παρατηρήσεων ενός «δείγματος» με το συνολικό αριθμό των παρατηρήσεων:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

Σημειώνεται ότι, όταν τα μέτρα θέσης και διασποράς αφορούν σε πληθυσμούς συμβολίζονται με ελληνικούς χαρακτήρες (π.χ. ο μέσος ενός πληθυσμού συμβολίζεται με «μ»), ενώ όταν αφορούν σε «δείγματα» συμβολίζονται με λατινικούς χαρακτήρες (π.χ. ο μέσος ενός «δείγματος» συμβολίζεται με $\overline{x}$).

Εάν οι βαθμολογίες 5 μαθητών σε ένα μάθημα είναι 18, 19, 19, 19 και 20 και η κλίμακα βαθμολόγησης είναι 0–20, τότε η μέση βαθμολογία στο συγκεκριμένο «δείγμα», σύμφωνα με την ισότητα 1, είναι ίση με:

$$\overline{x} = \frac{1}{5}(18+19+19+19+20) = 19$$

Σημειώνεται ότι ο μέσος, σε αντίθεση με τη διάμεσο, επηρεάζεται σημαντικά από τις τιμές των απομακρυσμένων παρατηρήσεων. Εάν στο παραπάνω παράδειγμα η βαθμολογία του πέμπτου μαθητή δεν ήταν 20, αλλά ήταν 5, τότε η μέση βαθμολογία θα ήταν ίση με:

$$\overline{x} = \frac{1}{5}(18+19+19+19+5) = 16$$

Ο μέσος είναι το κατάλληλο μέτρο θέσης στην περίπτωση των ποσοτικών μεταβλητών που ακολουθούν την κανονική κατανομή. Εάν οι ποσοτικές μεταβλητές δεν ακολουθούν την κανονική κατανομή, τότε το κατάλληλο μέτρο θέσης είναι η διάμεσος. Στη συνέχεια, θα αναλυθεί εκτενώς ο τρόπος με τον οποίο εφαρμόζεται ο έλεγχος της κανονικότητας αναφορικά με τις ποσοτικές μεταβλητές. Στην περίπτωση των διατάξιμων μεταβλητών που αντιμετωπίζονται ως μεταβλητές κλίμακας λόγου το κατάλληλο μέτρο θέσης είναι η διάμεσος και όχι ο μέσος, ανεξάρτητα από το αν οι μεταβλητές ακολουθούν ή όχι την κανονική κατανομή.

4.2. Διάμεσος

Η διάμεσος (median) αποτελεί το 50ό εκατοστιαίο σημείο μιας ομάδας παρατηρήσεων.[3–7,9,11,14,15,17–21,23–25] Τοποθετώντας μια ομάδα παρατηρήσεων στη σειρά, από τη μικρότερη στη μεγαλύτερη τιμή, η διάμεσος αποτελεί τη μεσαία τιμή. Έτσι, οι μισές τιμές των παρατηρήσεων θα είναι μεγαλύτερες ή ίσες με τη διάμεσο και οι άλλες μισές θα είναι μικρότερες ή ίσες με τη διάμεσο.

Όταν υπάρχουν n παρατηρήσεις και το n είναι περιττός ή, αλλιώς, μονός αριθμός, τότε η διάμεσος είναι η μεσαία παρατήρηση, δηλαδή η παρατήρηση $\frac{n+1}{2}$. Εάν, π.χ., οι βαθμολογίες 5 μαθητών σε ένα μάθημα είναι 18, 19, 19, 19 και 20 και η κλίμακα βαθμολόγησης είναι 0–20, τότε η διάμεσος είναι η τρίτη παρατήρηση $\left(\frac{n+1}{2}=\frac{5+1}{2}=3\right)$, δηλαδή είναι η τιμή 19.

Στην περίπτωση εξ άλλου που υπάρχουν n παρατηρήσεις, αλλά το n είναι ζυγός ή, αλλιώς άρτιος αριθμός, τότε η διάμεσος είναι ο αριθμητικός μέσος των δύο μεσαίων παρατηρήσεων, δηλαδή των παρατηρήσεων $\frac{n}{2}$ και $\frac{n+2}{2}$. Εάν, π.χ., οι βαθμολογίες 6 μαθητών σε ένα μάθημα είναι 16, 17, 18, 19, 20 και 20 και η κλίμακα βαθμολόγησης είναι 0–20, τότε η διάμεσος είναι ο αριθμητικός μέσος της τρίτης και της τέταρτης παρατήρησης $\left(\frac{n}{2}=\frac{6}{2}=3 \text{ και } \frac{n+2}{2}=\frac{6+2}{2}=4\right)$, δηλαδή είναι η τιμή 18,5 $\left(=\frac{18+19}{2}\right)$.

Σημειώνεται ότι η διάμεσος επηρεάζεται πολύ λιγότερο από τις τιμές των απομακρυσμένων παρατηρήσεων σε σχέση με το μέσο. Για παράδειγμα, εάν οι βαθμολογίες 5 μαθητών σε ένα μάθημα είναι 18, 19, 19, 19 και 20 και η κλίμακα βαθμολόγησης είναι 0–20, τότε τόσο ο μέσος όσο και η διάμεσος ισούνται με 19. Εάν όμως η βαθμολογία του πέμπτου μαθητή δεν ήταν 20, αλλά ήταν 5, τότε ο μέσος θα ήταν ίσος με 16, θα μειωνόταν δηλαδή κατά 3 μονάδες, ενώ η διάμεσος θα ήταν και πάλι ίση με 19.

Η διάμεσος είναι το κατάλληλο μέτρο θέσης στην περίπτωση των ποσοτικών μεταβλητών που δεν ακολουθούν την κανονική κατανομή, καθώς και στην περίπτωση των διατάξιμων μεταβλητών.

4.3. Επικρατούσα τιμή

Η επικρατούσα τιμή (mode) είναι η τιμή που συμβαίνει πιο συχνά σε μια ομάδα παρατηρήσεων και χρησιμοποιείται συνήθως στην περίπτωση των διατάξιμων και των ονομαστικών μεταβλητών και σπανιότερα στην περίπτωση των ποσοτικών μεταβλητών.[3–7,9,11,14,15,17–21,23–25] Για παράδειγμα, εάν οι βαθμολογίες 5 μαθητών σε ένα μάθημα είναι 18, 19, 19, 19 και 20 και η κλίμακα βαθμολόγησης είναι 0–20, τότε η επικρατούσα τιμή είναι η τιμή 19, καθώς εμφανίζεται τις περισσότερες φορές.

5. ΜΕΤΡΑ ΔΙΑΣΠΟΡΑΣ

5.1. Εύρος

Το εύρος (range) αποτελεί το απλούστερο μέτρο διασποράς και είναι ίσο με τη διαφορά μεταξύ της ελάχιστης και της μέγιστης τιμής μιας ομάδας παρατηρήσεων.[3–7,9,11,14,15,17–21,23–25] Το εύρος υπολογίζεται εξαιρετικά απλά, αλλά δεν χρησιμοποιείται ιδιαίτερα, καθώς λαμβάνει υπ' όψη του μόνο την ελάχιστη και τη μέγιστη τιμή μιας ομάδας παρατηρήσεων και επί πλέον επηρεάζεται σε μεγάλο βαθμό από τις τιμές των απομακρυσμένων παρατηρήσεων. Για παράδειγμα, εάν οι τιμές πέντε παρατηρήσεων είναι 80, 82, 85, 87 και 90, τότε το εύρος είναι ίσο με 10 (=90–80). Σε ορισμένες περιπτώσεις, το εύρος χρησιμοποιείται ως μέτρο διασποράς για την παρουσίαση των ποσοτικών μεταβλητών που δεν ακολουθούν την κανονική κατανομή, καθώς και των διατάξιμων μεταβλητών.

5.2. Ενδοτεταρτημοριακό εύρος

Το ενδοτεταρτημοριακό εύρος (interquartile range) επηρεάζεται πολύ λιγότερο από τις τιμές των απομακρυσμένων παρατηρήσεων σε σχέση με το εύρος και προκύπτει από την αφαίρεση του 25ου εκατοστιαίου σημείου των παρατηρήσεων από το 75ο.[3–7,9,11,14,15,17–21,23–25] Το ενδοτεταρτημοριακό εύρος περιλαμβάνει το κεντρικό 50% των παρατηρήσεων.

Το ενδοτεταρτημοριακό εύρος είναι το κατάλληλο μέτρο διασποράς στην περίπτωση των ποσοτικών μεταβλητών

που δεν ακολουθούν την κανονική κατανομή, καθώς και στην περίπτωση των διατάξιμων μεταβλητών. Στην περίπτωση αυτή, το ενδοτεταρτημοριακό εύρος πρέπει να παρουσιάζεται σε συνδυασμό με τη διάμεσο.

5.3. Διασπορά και τυπική απόκλιση

Η διασπορά (variance) και η τυπική απόκλιση (standard deviation) αποτελούν τα συνηθέστερα μέτρα διασποράς και ποσοτικοποιούν τη μεταβλητότητα των παρατηρήσεων γύρω από το μέσο του «δείγματος».[3–7,9,14,15,17–21,23–25] Στην περίπτωση των πληθυσμών, η διασπορά και η τυπική απόκλιση συμβολίζονται με «σ^2» και «σ», αντίστοιχα, ενώ στην περίπτωση των «δειγμάτων» συμβολίζονται με s^2 και s, αντίστοιχα.

Η διασπορά σε ένα «δείγμα» προκύπτει από την αφαίρεση του μέσου μιας ομάδας παρατηρήσεων από κάθε μια από τις παρατηρήσεις, υψώνοντας στο τετράγωνο τις διαφορές αυτές, πραγματοποιώντας έπειτα την πρόσθεσή τους και διαιρώντας το αποτέλεσμα με το συνολικό αριθμό των παρατηρήσεων μειωμένο κατά μία:

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (2)$$

Για παράδειγμα, εάν οι βαθμολογίες 5 μαθητών σε ένα μάθημα είναι 18, 19, 19, 19 και 20 και η κλίμακα βαθμολόγησης είναι 0–20, τότε η μέση βαθμολογία στο συγκεκριμένο «δείγμα», σύμφωνα με την ισότητα 1, είναι ίση με 19, ενώ η διασπορά, σύμφωνα με την ισότητα 2, είναι ίση με:

$$s^2 = \frac{1}{(5-1)}\left[(18-19)^2 + (19-19)^2 + (19-19)^2 + (19-19)^2 + (20-19)^2\right]$$

$$s^2 = \frac{1}{4}\left[(-1)^2 + 0^2 + 0^2 + 0^2 + 1^2\right]$$

$$s^2 = 0,5.$$

Η τυπική απόκλιση σε ένα «δείγμα» ισούται με την τετραγωνική ρίζα της διασποράς:

$$s = \sqrt{s^2} \qquad (3)$$

Έτσι, στο παραπάνω παράδειγμα, η τυπική απόκλιση είναι ίση με 0,71 $\left(=\sqrt{0,5}\right)$.

Στην πράξη, η τυπική απόκλιση χρησιμοποιείται συχνότερα σε σχέση με τη διασπορά και αυτό οφείλεται στο γεγονός ότι η τυπική απόκλιση έχει τις ίδιες μονάδες μέτρησης με το μέσο. Συγκρίνοντας δύο ομάδες παρατηρήσεων, η ομάδα με τη μικρότερη τυπική απόκλιση έχει πιο ομοιογενείς παρατηρήσεις και παρουσιάζει μικρότερη μεταβλητότητα.

Η τυπική απόκλιση είναι το κατάλληλο μέτρο διασποράς στην περίπτωση των ποσοτικών μεταβλητών που ακολουθούν την κανονική κατανομή. Στην περίπτωση αυτή, η τυπική απόκλιση πρέπει να παρουσιάζεται σε συνδυασμό με το μέσο.

5.4. Συντελεστής μεταβλητότητας

Ο συντελεστής μεταβλητότητας (coefficient of variation) είναι ένα μέτρο σχετικής μεταβλητότητας και εκφράζεται ως ποσοστό.[3–7,9,11,14,15,17–21,23–25] Προκύπτει από τη διαίρεση της τυπικής απόκλισης με το μέσο και τον πολλαπλασιασμό του προϊόντος της διαίρεσης με το 100:

$$Συντελεστής\ μεταβλητότητας = \frac{s}{\bar{x}} \times 100\% \qquad (4)$$

Η τυπική απόκλιση και ο μέσος έχουν τις ίδιες μονάδες μέτρησης, οπότε ο συντελεστής μεταβλητότητας είναι καθαρός αριθμός, δεν έχει δηλαδή μονάδες μέτρησης, και για το λόγο αυτόν μπορεί να χρησιμοποιηθεί για τη σύγκριση μεταξύ μεταβλητών με διαφορετικές μονάδες μέτρησης.

6. ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

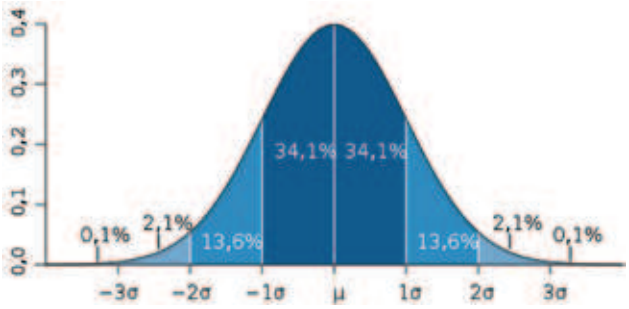
Κάθε τυχαία μεταβλητή έχει μια αντίστοιχη *κατανομή πιθανότητας* (probability distribution). Η κατανομή πιθανότητας στην περίπτωση των ασυνεχών μεταβλητών επιτρέπει τον προσδιορισμό όλων των πιθανών αποτελεσμάτων της τυχαίας μεταβλητής, καθώς και της πιθανότητας να συμβεί το καθένα από αυτά, ενώ στην περίπτωση των συνεχών μεταβλητών επιτρέπει τον καθορισμό των πιθανοτήτων που σχετίζονται με συγκεκριμένο εύρος τιμών. Υπάρχουν διάφορες κατανομές πιθανότητας, όπως η διωνυμική κατανομή (binomial distribution), η κατανομή Poisson (Poisson distribution) κ.ά., με την κανονική κατανομή να αποτελεί μια από τις σημαντικότερες.

Η *κανονική κατανομή* (normal distribution) ή, αλλιώς, κατανομή του Gauss* (Gaussian distribution) αφορά στις συνεχείς μεταβλητές, τις μεταβλητές δηλαδή που μπορούν να λάβουν οποιαδήποτε τιμή εντός των πραγματικών αριθμών, ακόμη και δεκαδική.[3,4,6,7,11,14,15,19–21,23,26] Ο μέσος (μ) και η τυπική απόκλιση (σ) ορίζουν πλήρως μια κανονική κατανομή. Η κανονική κατανομή έχει μια κορυφή και είναι συμμετρική γύρω από το μέσο της. Στην περίπτωση αυτή, ο μέσος, η διάμεσος και η επικρατούσα τιμή έχουν περίπου την ίδια τιμή. Η τυπική απόκλιση δηλώνει το μέγεθος της μεταβλητότητας γύρω από το μέσο. Μια κανονική κατανομή είναι δυνατόν να έχει έναν άπειρο αριθμό τι-

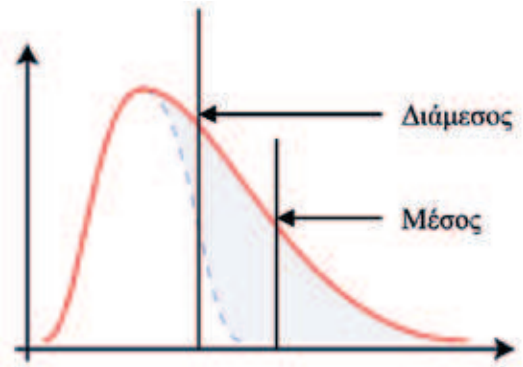
* Η έννοια της κανονικής κατανομής εισήχθη για πρώτη φορά το 1809 από το Γερμανό μαθηματικό Carl Gauss (1777–1855).

μών για το μέσο και την τυπική της απόκλιση. Η κανονική κατανομή στην οποία ο μέσος ισούται με μηδέν (μ=0) και η τυπική απόκλιση ισούται με 1 (σ=1) είναι γνωστή ως *τυπική κανονική κατανομή* (standard normal distribution). Στην κανονική κατανομή, κατά προσέγγιση, το 68,2% των παρατηρήσεων ευρίσκονται εντός μιας τυπικής απόκλισης (±1) από το μέσο, το 95,4% των παρατηρήσεων ευρίσκονται εντός δύο τυπικών αποκλίσεων (±2) από το μέσο και το 99% των παρατηρήσεων ευρίσκονται εντός τριών τυπικών αποκλίσεων (±3) από το μέσο (εικ. 11).

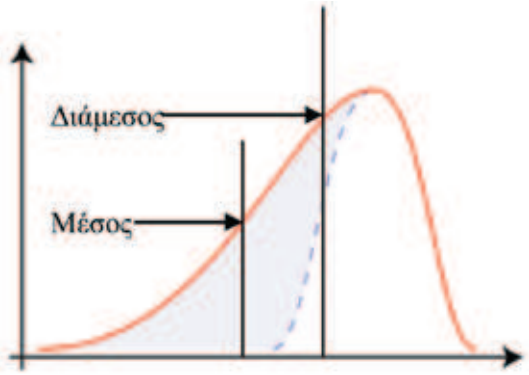
Όπως προαναφέρθηκε, στην περίπτωση της κανονικής κατανομής, ο μέσος, η διάμεσος και η επικρατούσα τιμή έχουν περίπου την ίδια τιμή, γεγονός που δεν συμβαίνει στις κατανομές με δεξιά και αριστερή ασυμμετρία. Στην εικόνα 12 φαίνεται μια κατανομή με δεξιά ασυμμετρία, στην οποία ο μέσος είναι μεγαλύτερος από τη διάμεσο, ενώ στην εικόνα 13 φαίνεται μια κατανομή με αριστερή ασυμμετρία, στην οποία ο μέσος είναι μικρότερος από τη διάμεσο.



Εικόνα 11. Κανονική κατανομή στην οποία, κατά προσέγγιση, το 68,2% των παρατηρήσεων ευρίσκονται εντός μίας τυπικής απόκλισης από το μέσο, το 95,4% των παρατηρήσεων ευρίσκονται εντός δύο τυπικών αποκλίσεων από το μέσο και το 99% των παρατηρήσεων ευρίσκονται εντός τριών τυπικών αποκλίσεων από το μέσο.



Εικόνα 12. Κατανομή με δεξιά ασυμμετρία, στην οποία ο μέσος είναι μεγαλύτερος από τη διάμεσο.



Εικόνα 13. Κατανομή με αριστερή ασυμμετρία, στην οποία ο μέσος είναι μικρότερος από τη διάμεσο.

7. ΕΛΕΓΧΟΣ ΚΑΝΟΝΙΚΟΤΗΤΑΣ

Η μονομεταβλητή ανάλυση των συνεχών μεταβλητών πρέπει οπωσδήποτε να περιλαμβάνει τον έλεγχο κανονικότητας, καθώς στην περίπτωση που ακολουθούν την κανονική κατανομή είναι δυνατή η εφαρμογή των *παραμετρικών μεθόδων* (parametric methods) στη διμεταβλητή και στην πολυμεταβλητή ανάλυση, ενώ στην περίπτωση που δεν ακολουθούν την κανονική κατανομή είναι δυνατή η εφαρμογή των *μη παραμετρικών μεθόδων* (non-parametric methods) ή, αλλιώς, των μεθόδων οι οποίες είναι «ελεύθερες» κατανομής (distribution-free methods).[3,4,6,11,15,26] Είναι σαφές ότι, ανάλογα με το αν οι συνεχείς μεταβλητές ακολουθούν ή όχι την κανονική κατανομή, θα χρησιμοποιηθούν και οι κατάλληλες στατιστικές μέθοδοι στη διμεταβλητή και στην πολυμεταβλητή ανάλυση.

Γενικά, οι παραμετρικές μέθοδοι προτιμώνται έναντι των μη παραμετρικών, καθώς χαρακτηρίζονται από μεγαλύτερη στατιστική ισχύ εφ' όσον ο αριθμός των παρατηρήσεων δεν είναι πολύ μικρός. Επί πλέον, οι παραμετρικές μέθοδοι είναι περισσότερες και απλούστερες στην κατανόηση και στην παρουσίαση των αποτελεσμάτων σε σχέση με τις μη παραμετρικές. Οι μη παραμετρικές μέθοδοι εξ άλλου βασίζονται στις διατάξεις και όχι στις πραγματικές τιμές των παρατηρήσεων, με αποτέλεσμα στην περίπτωση αυτή να μη χρησιμοποιείται όλη η πληροφορία που είναι γνωστή για μια μεταβλητή. Εν τούτοις, η χρήση των διατάξεων καθιστά τις μη παραμετρικές μεθόδους λιγότερο ευαίσθητες στο σφάλμα μέτρησης, επιτρέποντας παράλληλα την εφαρμογή των εν λόγω μεθόδων και στις διατάξιμες μεταβλητές.

Σημειώνεται ότι σε ορισμένες περιπτώσεις είναι δυνατή η εφαρμογή μαθηματικών μετασχηματισμών, όπως π.χ. ο λογαριθμικός μετασχηματισμός, οι οποίοι οδηγούν στη δημιουργία νέων συνεχών μεταβλητών που ακολουθούν πλέον την κανονική κατανομή, ενώ οι αρχικές μη

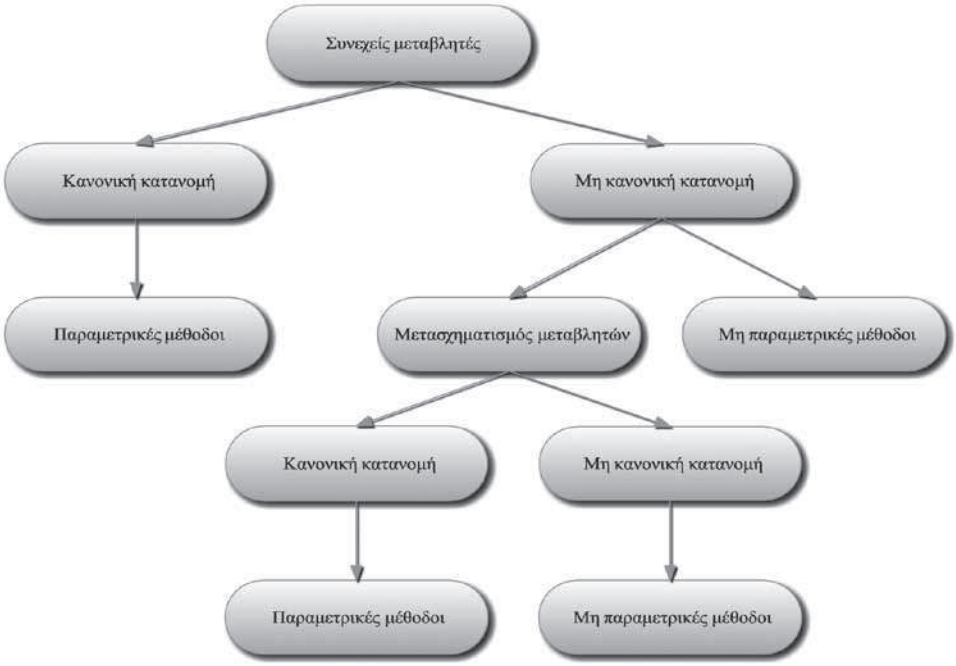
μετασχηματισμένες μεταβλητές δεν την ακολουθούσαν. Εάν πάντως ο αριθμός των παρατηρήσεων είναι αρκετά μικρός (συνήθως <30 παρατηρήσεις), κρίνεται σκόπιμο να εφαρμόζονται μη παραμετρικές μέθοδοι.

Στην εικόνα 14 φαίνεται ο τρόπος εφαρμογής των παραμετρικών και των μη παραμετρικών μεθόδων στην περίπτωση των συνεχών μεταβλητών ανάλογα με το αν ακολουθούν ή όχι την κανονική κατανομή.

Ο έλεγχος κανονικότητας μπορεί να πραγματοποιηθεί με τους εξής τρόπους: (α) Σύγκριση των τιμών μεταξύ του μέσου και της διαμέσου, (β) εκτίμηση των συντελεστών ασυμμετρίας και κύρτωσης, (γ) εφαρμογή κατάλληλων στατιστικών ελέγχων και (δ) χρήση γραφημάτων.[3,4,6,11,15,26] Ιδιαίτερη σημασία έχει η αναγνώριση των απομακρυσμένων και των ακραίων παρατηρήσεων, καθώς και ο τρόπος με τον οποίο επηρεάζουν τα αποτελέσματα της στατιστικής ανάλυσης.

7.1. Μέσος και διάμεσος

Οι τιμές του μέσου και της διαμέσου μιας συνεχούς μεταβλητής μπορούν να χρησιμοποιηθούν για μια αδρή εκτίμηση της κανονικότητας. Πιο συγκεκριμένα, όσο πλησιέστερα ευρίσκονται ο μέσος και η διάμεσος τόσο πιο πιθανό είναι η ποσοτική μεταβλητή να ακολουθεί την κανονική κατανομή. Στον πίνακα 5 φαίνονται οι τιμές των μέσων και των διαμέσων της ηλικίας, του δείκτη μάζας σώματος και της διάρκειας νοσηλείας 223 ατόμων, οι απόλυτες και οι σχετικές διαφορές μεταξύ των μέσων και των διαμέσων και η ερμηνεία των αποτελεσμάτων αναφορικά με το αν οι μεταβλητές ακολουθούν ή όχι την κανονική κατανομή. Η σχετική διαφορά εκφράζεται ως ποσοστό και προκύπτει από τη διαίρεση της απόλυτης διαφοράς μεταξύ του μέσου και της διαμέσου με την τιμή του μέσου $\left(\text{σχετική διαφορά} = \frac{|\text{μέσος} - \text{διάμεσος}|}{\text{μέσος}}\right)$. Αναφορικά με την



Εικόνα 14. Εφαρμογή των παραμετρικών και των μη παραμετρικών μεθόδων στην περίπτωση των συνεχών μεταβλητών ανάλογα με το αν ακολουθούν ή όχι την κανονική κατανομή.

Πίνακας 5. Οι τιμές των μέσων και των διαμέσων της ηλικίας, του δείκτη μάζας σώματος και της διάρκειας νοσηλείας 223 ατόμων, οι απόλυτες και οι σχετικές διαφορές μεταξύ των μέσων και των διαμέσων, καθώς και η ερμηνεία των αποτελεσμάτων αναφορικά με το αν οι μεταβλητές ακολουθούν ή όχι την κανονική κατανομή.

Μεταβλητή	**Μέσος**	**Διάμεσος**	**Απόλυτη διαφορά** (|*μέσος–διάμεσος*|)	**Σχετική διαφορά (%)** $\left(\frac{\|\text{μέσος–διάμεσος}\|}{\text{μέσος}}\right)$	**Ερμηνεία**
Ηλικία (έτη)	46	47	|-1|	2,2	Ισχυρή ένδειξη κανονικής κατανομής
Δείκτης μάζας σώματος (kg/m^2)	26	25,8	0,2	0,8	Ισχυρή ένδειξη κανονικής κατανομής
Διάρκεια νοσηλείας (ημέρες)	11,8	9	2,8	23,7	Ισχυρή ένδειξη κατανομής με δεξιά ασυμμετρία

ηλικία και το δείκτη μάζας σώματος, οι τιμές των μέσων και των διαμέσων ευρίσκονται πολύ κοντά, γεγονός που σημαίνει ότι οι δύο αυτές μεταβλητές είναι πολύ πιθανό να ακολουθούν την κανονική κατανομή. Πιο συγκεκριμένα, στην περίπτωση της ηλικίας η σχετική διαφορά είναι 2,2%, ενώ στην περίπτωση του δείκτη μάζας σώματος είναι 0,8%. Όσο πλησιέστερα στο 0 ευρίσκεται η απόλυτη και η σχετική διαφορά, τόσο πιο πιθανό είναι η συνεχής μεταβλητή να ακολουθεί την κανονική κατανομή. Στην περίπτωση της διάρκειας νοσηλείας, η μεγάλη σχετική διαφορά (=23,7%) σε συνδυασμό με το γεγονός ότι ο μέσος είναι μεγαλύτερος από τη διάμεσο δηλώνουν ότι η κατανομή της μεταβλητής αυτής εμφανίζει δεξιά ασυμμετρία.

Όπως προαναφέρθηκε, στην κανονική κατανομή, κατά προσέγγιση, το 95,4% των παρατηρήσεων ευρίσκονται εντός δύο τυπικών αποκλίσεων (±2) από το μέσο, οπότε οι περισσότερες παρατηρήσεις ευρίσκονται σε απόσταση δύο τυπικών αποκλίσεων από το μέσο (εικ. 11). Διπλασιάζοντας την τυπική απόκλιση μιας συνεχούς μεταβλητής και στη συνέχεια αφαιρώντας και προσθέτοντας την τιμή αυτή με το μέσο, προκύπτουν δύο τιμές που συνιστούν το εκτιμώμενο εύρος στο οποίο αναμένεται να ευρίσκεται περίπου το 95% των παρατηρήσεων. Στην περίπτωση κατά την οποία η συνεχής μεταβλητή ακολουθεί την κανονική κατανομή, το εκτιμώμενο αυτό εύρος αναμένεται να ευρίσκεται εντός των ορίων του πραγματικού εύρους, εντός δηλαδή του εύρους που προκύπτει από την ελάχιστη και τη μέγιστη τιμή της μεταβλητής. Στον πίνακα 6 φαίνεται το 95% εκτιμώμενο εύρος, καθώς και η ελάχιστη και η μέγιστη τιμή της ηλικίας, του δείκτη μάζας σώματος και της διάρκειας νοσηλείας των 223 ατόμων του πίνακα 5. Αναφορικά με το δείκτη μάζας σώματος, το 95% εκτιμώμενο εύρος (18,6–33,4) ευρίσκεται εντός των ορίων του πραγματικού εύρους, όπως αυτό ορίζεται από την ελάχιστη (=17) και τη μέγιστη τιμή (=36). Επομένως, ο δείκτης μάζας σώματος είναι πολύ πιθανό να ακολουθεί την κανονική κατανομή. Εν τούτοις, στην περίπτωση της διάρκειας νοσηλείας, το 95% αναμενόμενο εύρος δεν ευρίσκεται εντός των ορίων του πραγματικού εύρους, γεγονός που δηλώνει ότι η μεταβλητή αυτή είναι πολύ πιθανό να μην ακολουθεί την κανονική κατανομή. Στην περίπτωση αυτή, μάλιστα, η τιμή του μικρότερου ορίου του 95% αναμενόμενου εύρους είναι αρνητική, γεγονός που δηλώνει μεγάλο βαθμό ασυμμετρίας και διαφοροποίησης από την κανονική κατανομή. Επί πλέον, αναφορικά με τη διάρκεια νοσηλείας, οι δύο εκτιμώμενες τιμές είναι αρκετά μικρότερες από την πραγματική ελάχιστη και μέγιστη τιμή που λαμβάνει η μεταβλητή, γεγονός το οποίο δηλώνει την ύπαρξη κατανομής με δεξιά ασυμμετρία. Αντίθετα, εάν οι δύο εκτιμώμενες τιμές της συνεχούς μεταβλητής είναι αρκετά μεγαλύτερες από την πραγματική ελάχιστη και μέγιστη τιμή που λαμβάνει η μεταβλητή, τότε πρόκειται για κατανομή με αριστερή ασυμμετρία. Αναφορικά με την ηλικία, η τιμή του μικρότερου ορίου του 95% αναμενόμενου εύρους ευρίσκεται εκτός των ορίων του πραγματικού εύρους, ενώ η τιμή του μεγαλύτερου ορίου του αναμενόμενου εύρους ευρίσκεται εντός των ορίων του πραγματικού εύρους, γεγονός που δεν επιτρέπει την εξαγωγή ασφαλών συμπερασμάτων και καθιστά πιθανή την ύπαρξη κανονικής κατανομής.

Γενικά, μεταβλητές με τυπική απόκλιση μεγαλύτερη από το ήμισυ του μέσου είναι πολύ πιθανό να μην ακολουθούν την κανονική κατανομή. Χαρακτηριστικά, στον πίνακα 6, η τυπική απόκλιση (=9,7) της διάρκειας νοσηλείας είναι πολύ μεγαλύτερη από το ήμισυ του μέσου (=11,8/2=5,9), γεγονός το οποίο δηλώνει την ύπαρξη μεγάλου βαθμού ασυμμετρίας, ενώ η τυπική απόκλιση (=3,7) του δείκτη μάζας σώματος είναι πολύ μικρότερη από το ήμισυ του μέσου (=26/2=13), γεγονός που δηλώνει την ύπαρξη κανονικής κατανομής.

7.2. Συντελεστές ασυμμετρίας και κύρτωσης

Οι συντελεστές ασυμμετρίας και κύρτωσης χρησιμοποιούνται για την εκτίμηση της ασυμμετρίας (skewness) και της κύρτωσης (kurtosis), αντίστοιχα, και προσφέρουν πολύτιμη πληροφορία σχετικά με το αν μια συνεχής μεταβλητή ακολουθεί ή όχι την κανονική κατανομή.

Αναλυτικότερα, οι συντελεστές ασυμμετρίας και κύρτωσης λαμβάνουν τιμές από -3 έως 3, με τις τιμές από -1 έως 1 να δηλώνουν την ύπαρξη κανονικής κατανομής και τις τιμές από -1 έως -3 ή από 1–3 να δηλώνουν τη μη ύπαρξη κανονικής κατανομής. Όσο πλησιέστερα είναι

Πίνακας 6. Το 95% εκτιμώμενο εύρος, η ελάχιστη και η μέγιστη τιμή της ηλικίας, του δείκτη μάζας σώματος και της διάρκειας νοσηλείας των 223 ατόμων του πίνακα 5, καθώς και η ερμηνεία των αποτελεσμάτων αναφορικά με το αν οι μεταβλητές ακολουθούν ή όχι την κανονική κατανομή.

Μεταβλητή	Μέση τιμή±2 τυπικές αποκλίσεις	95% εκτιμώμενο εύρος	Ελάχιστη και μέγιστη τιμή	Ερμηνεία
Ηλικία (έτη)	46±2×15	16–76	19 και 85	Μέτρια ένδειξη κανονικής κατανομής
Δείκτης μάζας σώματος (kg/m^2)	26±2×3,7	18,6–33,4	17 και 36	Ισχυρή ένδειξη κανονικής κατανομής
Διάρκεια νοσηλείας (ημέρες)	11,8±2×9,7	-7,6–31,2	1 και 74	Ισχυρή ένδειξη κατανομής με δεξιά ασυμμετρία

οι συντελεστές ασυμμετρίας και κύρτωσης στο 0, τόσο πιθανότερο είναι μια συνεχής μεταβλητή να ακολουθεί την κανονική κατανομή. Όταν οι συντελεστές ασυμμετρίας και κύρτωσης λαμβάνουν τιμές ≤3 ή >3, τότε υπάρχει ισχυρή ένδειξη ότι η συνεχής μεταβλητή δεν ακολουθεί την κανονική κατανομή.

Επί πλέον, διαιρώντας τους συντελεστές ασυμμετρίας και κύρτωσης με τα αντίστοιχα τυπικά σφάλματα προκύπτουν δύο τιμές, που χρησιμοποιούνται για τον έλεγχο της κανονικότητας των συνεχών μεταβλητών. Εάν οι τιμές που προκύπτουν από τη διαίρεση των συντελεστών ασυμμετρίας και κύρτωσης με τα αντίστοιχα τυπικά σφάλματα είναι ≤2 ή >2, τότε υπάρχει ένδειξη ότι η μεταβλητή δεν ακολουθεί την κανονική κατανομή.

Στον πίνακα 7 φαίνονται οι συντελεστές ασυμμετρίας και κύρτωσης και τα αντίστοιχα τυπικά σφάλματα αναφορικά με την ηλικία, το δείκτη μάζας σώματος και τη διάρκεια νοσηλείας των 223 ατόμων του πίνακα 5. Σύμφωνα με τα αποτελέσματα του πίνακα 7, αναφορικά με το δείκτη μάζας σώματος, οι συντελεστές ασυμμετρίας (=0,148) και κύρτωσης (=-0,233) είναι αρκετά πλησίον του 0, γεγονός που αποτελεί ισχυρή ένδειξη κανονικής κατανομής. Επί πλέον, διαιρώντας τους συντελεστές ασυμμετρίας και κύρτωσης με τα αντίστοιχα τυπικά σφάλματα προκύπτουν δύο τιμές (0,91 και -0,72, αντίστοιχα), οι οποίες ευρίσκονται μεταξύ -2 και 2, γεγονός που αποτελεί και πάλι ισχυρή ένδειξη κανονικής κατανομής. Αναφορικά με την ηλικία, οι συντελεστές ασυμμετρίας (=0,374) και κύρτωσης (=-0,359) ευρίσκονται μεταξύ -1 και 1, γεγονός που αποτελεί ένδειξη κανονικής κατανομής. Επί πλέον, διαιρώντας τους συντελεστές ασυμμετρίας και κύρτωσης με τα αντίστοιχα τυπικά σφάλματα προκύπτουν δύο τιμές (2,23 και -1,11, αντίστοιχα), με τη μία τιμή να μην ευρίσκεται μεταξύ -2 και 2, γεγονός που αποτελεί ένδειξη μέτριας ασυμμετρίας. Αναφορικά με τη διάρκεια νοσηλείας, οι συντελεστές ασυμμετρίας (=3,910) και κύρτωσης (=18,64) ευρίσκονται εκτός του διαστήματος -3 έως 3, γεγονός που αποτελεί ισχυρή ένδειξη κατανομής με ασυμμετρία. Επί πλέον, διαιρώντας τους συντελεστές ασυμμετρίας και κύρτωσης με τα αντίστοιχα τυπικά σφάλματα προκύπτουν δύο τιμές (24 και 81,5, αντίστοιχα), οι οποίες ευρίσκονται εκτός του διαστήματος -2 έως 2, γεγονός που αποτελεί και πάλι ισχυρή ένδειξη κατανομής με ασυμμετρία.

7.3. Στατιστικοί έλεγχοι

Οι στατιστικοί έλεγχοι που χρησιμοποιούνται για τον έλεγχο της κανονικότητας είναι ο έλεγχος Kolmogorov-Smirnov και ο έλεγχος Shapiro-Wilk. Και στις δύο περιπτώσεις, η μηδενική υπόθεση είναι ότι οι παρατηρήσεις του «δείγματος» προέρχονται από έναν πληθυσμό στον οποίο η κατανομή της συνεχούς μεταβλητής είναι κανονική. Έτσι, εάν προκύψουν τιμές p ή, αλλιώς, τιμές παρατηρούμενου επιπέδου στατιστικής σημαντικότητας που είναι <0,05, τότε απορρίπτεται η μηδενική υπόθεση και θεωρείται ότι η συνεχής μεταβλητή δεν ακολουθεί την κανονική κατανομή. Αντίθετα, εάν προκύψουν τιμές p που είναι >0,05, τότε δεν απορρίπτεται η μηδενική υπόθεση και θεωρείται ότι η συνεχής μεταβλητή ακολουθεί την κανονική κατανομή. Σημειώνεται ότι ο έλεγχος Shapiro-Wilk έχει μεγαλύτερη στατιστική ισχύ από τον έλεγχο Kolmogorov-Smirnov στην αναγνώριση μεταβλητών που δεν ακολουθούν την κανονική κατανομή.

Δυστυχώς, οι έλεγχοι Kolmogorov-Smirnov και Shapiro-Wilk είναι αρκετά «ευαίσθητοι» στον αριθμό των παρατηρήσεων του «δείγματος». Πιο συγκεκριμένα, σε «δείγματα» με μικρό αριθμό παρατηρήσεων (συνήθως <30 παρατηρήσεις) η πιθανότητα απόρριψης της μηδενικής υπόθεσης είναι μικρή, ενώ σε «δείγματα» με μεγάλο αριθμό παρατηρήσεων (συνήθως >100 παρατηρήσεις), ακόμη και μια μικρή απόκλιση από την κανονικότητα μπορεί να οδηγήσει σε απόρριψη της μηδενικής υπόθεσης. Γενικά, στην περίπτωση «δειγμάτων» με μικρό αριθμό παρατηρήσεων (<30 παρατηρήσεις) θεωρείται ότι η συνεχής μεταβλητή

Πίνακας 7. Συντελεστές ασυμμετρίας και κύρτωσης, καθώς και τα αντίστοιχα τυπικά σφάλματα αναφορικά με την ηλικία, το δείκτη μάζας σώματος και τη διάρκεια νοσηλείας των 223 ατόμων του πίνακα 5, καθώς και η ερμηνεία των αποτελεσμάτων αναφορικά με το αν οι μεταβλητές ακολουθούν ή όχι την κανονική κατανομή.

Μεταβλητή	Συντελεστής ασυμμετρίας (τυπικό σφάλμα)	Συντελεστής ασυμμετρίας / Τυπικό σφάλμα	Συντελεστής κύρτωσης (τυπικό σφάλμα)	Συντελεστής ασυμμετρίας / Τυπικό σφάλμα	Ερμηνεία
Ηλικία (έτη)	0,374 (0,163)	2,23	-0,359 (0,324)	-1,11	Μέτρια ένδειξη κανονικής κατανομής
Δείκτης μάζας σώματος (kg/m²)	0,148 (0,163)	0,91	-0,233 (0,324)	-0,72	Ισχυρή ένδειξη κανονικής κατανομής
Διάρκεια νοσηλείας (ημέρες)	3,910 (0,163)	24,0	18,64 (0,324)	81,5	Ισχυρή ένδειξη κατανομής με δεξιά ασυμμετρία

δεν ακολουθεί την κανονική κατανομή, ανεξάρτητα από τις τιμές p που προκύπτουν στους ελέγχους Kolmogorov-Smirnov και Shapiro-Wilk. Επί πλέον, στην περίπτωση «δειγμάτων» με αριθμό παρατηρήσεων μεταξύ 30–100, εάν οι στατιστικοί έλεγχοι οδηγήσουν σε απόρριψη της μηδενικής υπόθεσης, τότε θεωρείται ότι η μεταβλητή δεν ακολουθεί την κανονική κατανομή, ενώ εάν οι στατιστικοί έλεγχοι δεν οδηγήσουν σε απόρριψη της μηδενικής υπόθεσης, τότε είναι απαραίτητο να διεξαχθούν και οι άλλες μέθοδοι ελέγχου της κανονικότητας που αναλύονται στο παρόν άρθρο, έτσι ώστε να επιβεβαιωθεί η ύπαρξη της κανονικότητας. Στην περίπτωση «δειγμάτων» εξ άλλου με μεγάλο αριθμό παρατηρήσεων (>100 παρατηρήσεις), εάν οι στατιστικοί έλεγχοι δεν οδηγήσουν σε απόρριψη της μηδενικής υπόθεσης, τότε θεωρείται ότι η μεταβλητή ακολουθεί την κανονική κατανομή, ενώ εάν οι στατιστικοί έλεγχοι οδηγήσουν σε απόρριψη της μηδενικής υπόθεσης, τότε είναι απαραίτητο να διεξαχθούν και οι άλλες μέθοδοι ελέγχου της κανονικότητας που αναλύονται στο άρθρο αυτό, έτσι ώστε να επιβεβαιωθεί η απουσία της κανονικότητας.

Στον πίνακα 8 εμφανίζονται οι τιμές p που προκύπτουν από τους ελέγχους Kolmogorov-Smirnov και Shapiro-Wilk αναφορικά με την ηλικία, το δείκτη μάζας σώματος και τη διάρκεια νοσηλείας των 223 ατόμων του πίνακα 5. Στην περίπτωση του δείκτη μάζας σώματος, και οι δύο στατιστικοί έλεγχοι δεν οδήγησαν σε απόρριψη της μηδενικής υπόθεσης (τιμές $p>0,05$), γεγονός που αποτελεί ισχυρή ένδειξη κανονικής κατανομής. Αντίθετα, αναφορικά με την ηλικία και τη διάρκεια νοσηλείας, οι δύο στατιστικοί έλεγχοι οδήγησαν σε απόρριψη της μηδενικής υπόθεσης (τιμές $p<0,05$), γεγονός που αποτελεί ισχυρή ένδειξη ασύμμετρης κατανομής.

7.4. Γραφήματα

Τα γραφήματα χρησιμοποιούνται επικουρικά για τον έλεγχο της κανονικότητας, καθώς δεν υπάρχουν σαφείς κανόνες και σε μεγάλο βαθμό υπεισέρχεται η υποκειμενική κρίση. Τα γραφήματα που συνήθως χρησιμοποιούνται είναι τα ιστογράμματα, τα κανονικά διαγράμματα Q-Q και τα διαγράμματα πλαισίου.

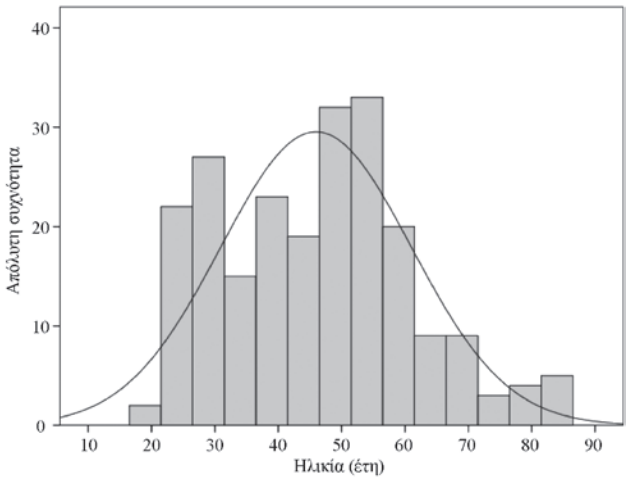
Τα ιστογράμματα παρέχουν μια άμεση και αδρή οπτική εκτίμηση της κατανομής που ακολουθεί μια συνεχής μεταβλητή, ενώ τα *κανονικά διαγράμματα Q-Q* (normal Q-Q plots) απεικονίζουν τις παρατηρηθείσες τιμές μιας μεταβλητής σε σχέση με τις αναμενόμενες τιμές, οι οποίες θα προέκυπταν εάν η μεταβλητή ακολουθούσε την κανονική κατανομή. Στα κανονικά διαγράμματα Q-Q, εάν η συνεχής μεταβλητή ακολουθεί την κανονική κατανομή, τότε τα σημεία τα οποία αντιστοιχούν στις παρατηρήσεις συγκεντρώνονται πολύ κοντά στην ευθεία γραμμή που δηλώνει την ύπαρξη κανονικότητας. Αποκλίσεις από την ευθεία αυτή γραμμή δηλώνουν την ύπαρξη ασυμμετρίας. Επί πλέον, στα διαγράμματα πλαισίου, η ύπαρξη αρκετών ακραίων παρατηρήσεων, καθώς και η μη κεντρική θέση της διαμέσου στο πλαίσιο αποτελούν ενδείξεις ασυμμετρίας. Εάν η διάμεσος είναι πιο κοντά στο κάτω άκρο του πλαισίου, τότε η κατανομή της μεταβλητής εμφανίζει δεξιά ασυμμετρία, ενώ εάν η διάμεσος είναι πλησιέστερα στο άνω άκρο του πλαισίου, τότε η κατανομή της μεταβλητής εμφανίζει αριστερή ασυμμετρία.

Στις εικόνες 15, 16 και 17 φαίνονται το ιστόγραμμα, το κανονικό διάγραμμα Q-Q και το διάγραμμα πλαισίου, αντίστοιχα, της ηλικίας των 223 ατόμων του πίνακα 5. Με βάση το ιστόγραμμα, η ηλικία φαίνεται να ακολουθεί την κανονική κατανομή, στοιχείο όμως το οποίο δεν επιβεβαιώνεται και από το κανονικό διάγραμμα Q-Q και το διάγραμμα πλαισίου, που παρέχουν ενδείξεις μέτριας αριστερής ασυμμετρίας. Πιο συγκεκριμένα, στο κανονικό διάγραμμα Q-Q, τόσο στο άνω όσο και στο κάτω άκρο οι παρατηρήσεις ευρίσκονται σε απόσταση από την ευθεία γραμμή, γεγονός που δηλώνει την ύπαρξη κανονικότητας, ενώ στο διάγραμμα πλαισίου η διάμεσος δεν ευρίσκεται στο κέντρο του πλαισίου, αλλά είναι πλησιέστερα στο άνω άκρο του πλαισίου.

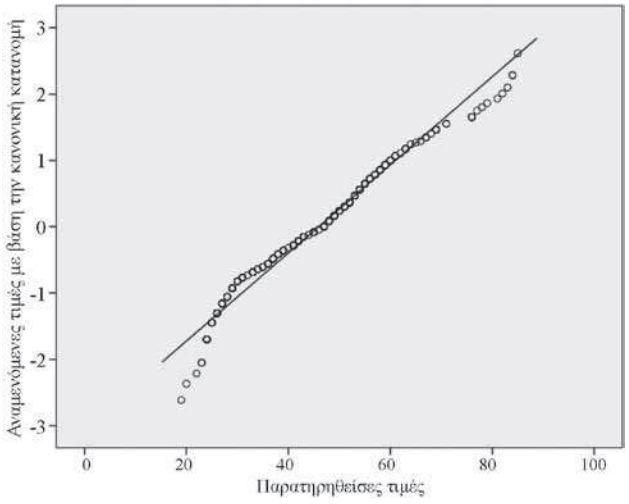
Στις εικόνες 18, 19 και 20 απεικονίζονται το ιστόγραμμα, το κανονικό διάγραμμα Q-Q και το διάγραμμα πλαισίου, αντίστοιχα, του δείκτη μάζας σώματος των 223 ατόμων

Πίνακας 8. Τιμές p που προκύπτουν από τους ελέγχους Kolmogorov-Smirnov και Shapiro-Wilk αναφορικά με την ηλικία, το δείκτη μάζας σώματος και τη διάρκεια νοσηλείας των 223 ατόμων του πίνακα 5, καθώς και η ερμηνεία των αποτελεσμάτων αναφορικά με το αν οι μεταβλητές ακολουθούν ή όχι την κανονική κατανομή.

Μεταβλητή	Τιμή p		Ερμηνεία
	Έλεγχος Kolmogorov-Smirnov	Έλεγχος Shapiro-Wilk	
Ηλικία (έτη)	0,013	<0,001	Ισχυρή ένδειξη ασύμμετρης κατανομής
Δείκτης μάζας σώματος (kg/m^2)	>0,200	0,475	Ισχυρή ένδειξη κανονικής κατανομής
Διάρκεια νοσηλείας (ημέρες)	<0,001	<0,001	Ισχυρή ένδειξη ασύμμετρης κατανομής



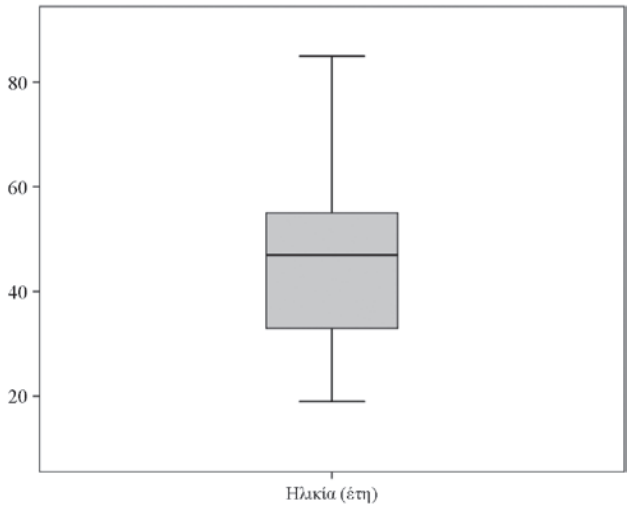
Εικόνα 15. Ιστόγραμμα της ηλικίας των 223 ατόμων του πίνακα 5.



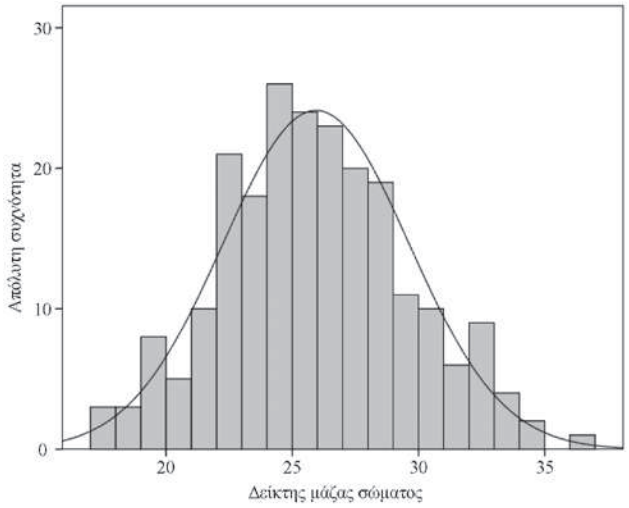
Εικόνα 16. Κανονικό διάγραμμα Q-Q της ηλικίας των 223 ατόμων του πίνακα 5.

του πίνακα 5. Και τα τρία γραφήματα παρέχουν ενδείξεις ύπαρξης κανονικής κατανομής. Πιο συγκεκριμένα, στο κανονικό διάγραμμα Q-Q, οι περισσότερες παρατηρήσεις ευρίσκονται πολύ κοντά στην ευθεία γραμμή, στοιχείο που δηλώνει την ύπαρξη κανονικότητας, ενώ στο διάγραμμα πλαισίου η διάμεσος ευρίσκεται στο κέντρο του πλαισίου.

Στις εικόνες 21, 22 και 23 φαίνονται το ιστόγραμμα, το κανονικό διάγραμμα Q-Q και το διάγραμμα πλαισίου, αντίστοιχα, της διάρκειας νοσηλείας των 223 ατόμων του πίνακα 5. Και τα τρία γραφήματα παρέχουν ενδείξεις δεξιάς ασυμμετρίας. Πιο συγκεκριμένα, στο κανονικό διάγραμμα Q-Q, σχεδόν όλες οι παρατηρήσεις ευρίσκονται μακριά από την ευθεία γραμμή, γεγονός που δηλώνει την ύπαρξη κανονικότητας, ενώ στο διάγραμμα πλαισίου υπάρχουν πολλές απομακρυσμένες και ακραίες παρατηρήσεις.



Εικόνα 17. Διάγραμμα πλαισίου της ηλικίας των 223 ατόμων του πίνακα 5.

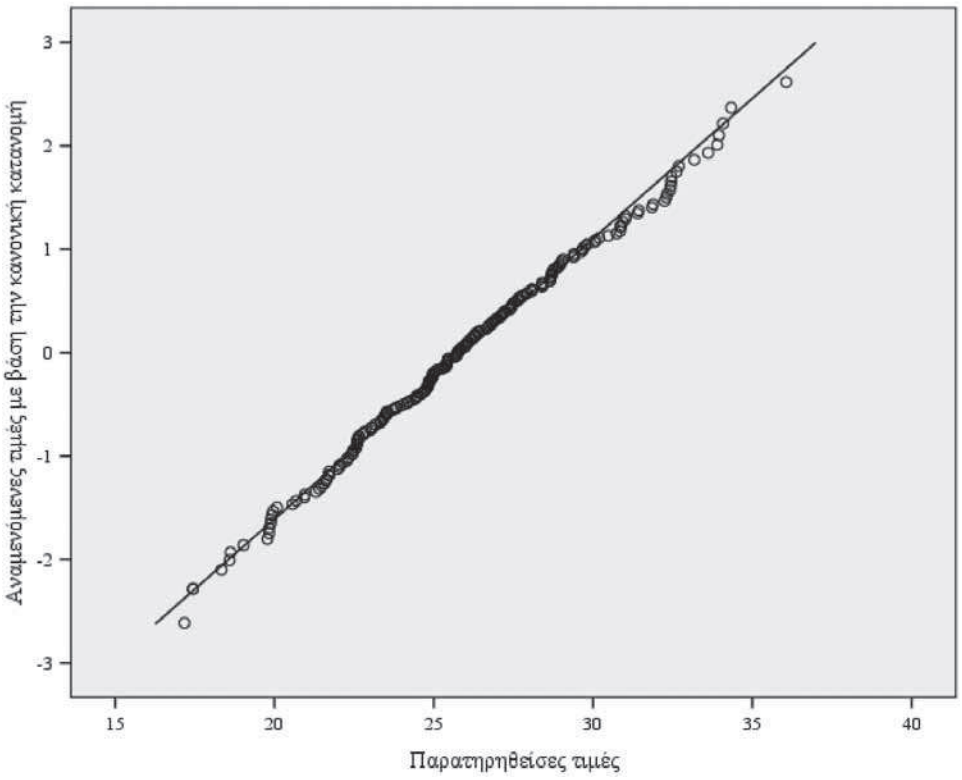


Εικόνα 18. Ιστόγραμμα του δείκτη μάζας σώματος των 223 ατόμων του πίνακα 5.

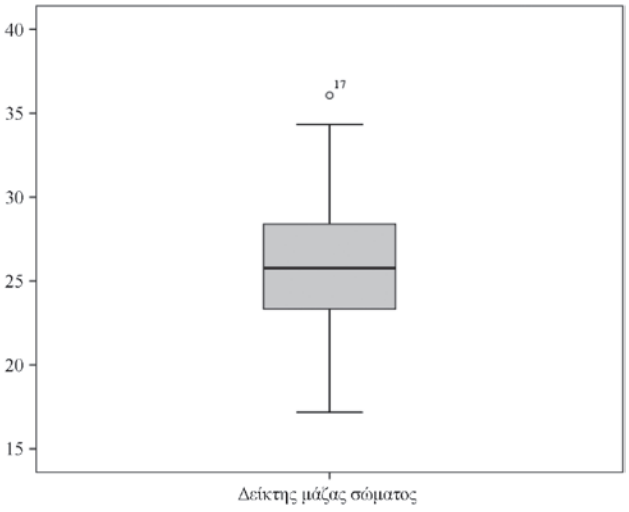
7.5. Συμπεράσματα

Είναι σαφές ότι οι μέθοδοι που χρησιμοποιούνται για τον έλεγχο της κανονικότητας των συνεχών μεταβλητών δεν οδηγούν πάντοτε στο ίδιο συμπέρασμα, με αποτέλεσμα σε ορισμένες περιπτώσεις να δημιουργείται σύγχυση. Για το λόγο αυτόν, είναι απαραίτητο να λαμβάνεται υπ' όψη η πληροφορία που προέρχεται από όλες τις μεθόδους ελέγχου της κανονικότητας, έτσι ώστε να εξάγονται ασφαλέστερα συμπεράσματα.

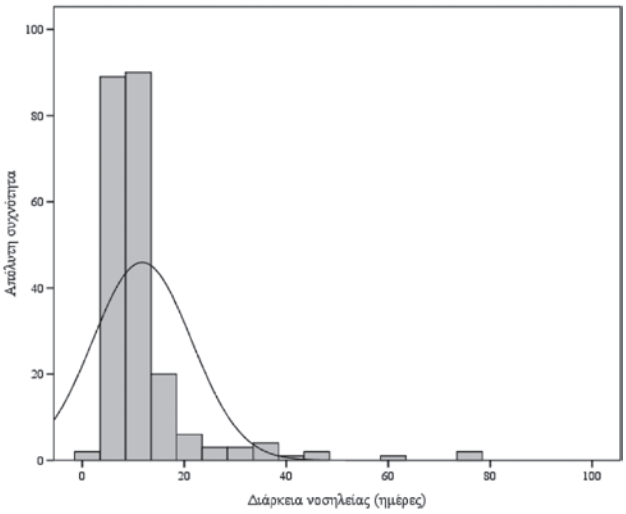
Το παράδειγμα που αναλύθηκε εκτενώς στους πίνακες 5–8 και στις εικόνες 15–23 είναι ενδεικτικό της δυσκολίας που υφίσταται στον έλεγχο της κανονικότητας των συνεχών μεταβλητών. Στον πίνακα 9 φαίνεται ο αλγόριθμος της λήψης



Εικόνα 19. Κανονικό διάγραμμα Q-Q του δείκτη μάζας σώματος των 223 ατόμων του πίνακα 5.



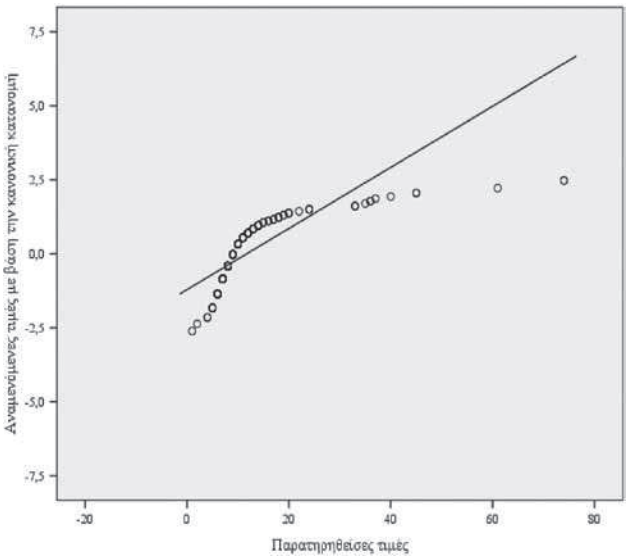
Εικόνα 20. Διάγραμμα πλαισίου του δείκτη μάζας σώματος των 223 ατόμων του πίνακα 5.



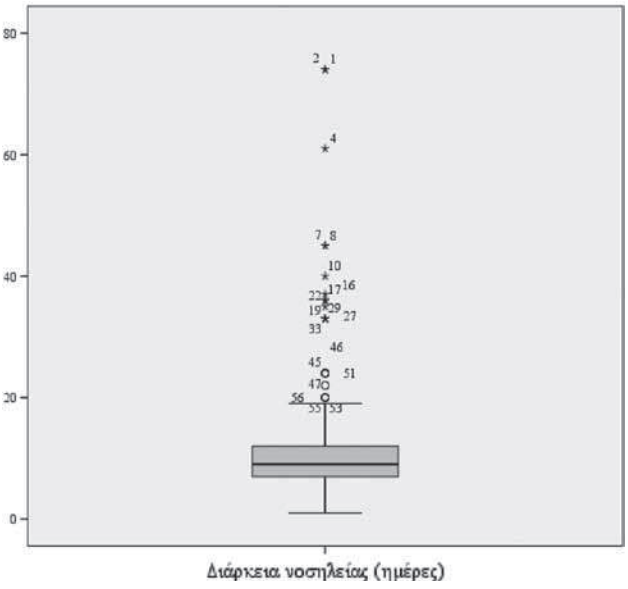
Εικόνα 21. Ιστόγραμμα της διάρκειας νοσηλείας των 223 ατόμων του πίνακα 5.

απόφασης σχετικά με το αν η ηλικία, ο δείκτης μάζας σώματος και η διάρκεια νοσηλείας, αναφορικά με τα 223 άτομα του πίνακα 5, ακολουθούν ή όχι την κανονική κατανομή, με το «ναι» να δηλώνει ισχυρή ένδειξη της ύπαρξης κανονικής κατανομής, το «πιθανώς» να δηλώνει μέτρια ένδειξη της ύπαρξης κανονικής κατανομής και το «όχι» να δηλώνει ισχυρή ένδειξη της ύπαρξης ασυμμετρίας. Σύμφωνα με τον πίνακα 9, ο δείκτης μάζας σώματος ακολουθεί την κανονική κατανομή, η ηλικία ακολουθεί κατά προσέγγιση την κανονική κατανομή, παρουσιάζοντας ελαφρά ασυμμετρία, και η διάρκεια νοσηλείας δεν ακολουθεί την κανονική κατανομή.

Σημειώνεται ότι οι παραμετρικές μέθοδοι είναι «ανθεκτικές» σε μικρές αποκλίσεις από την κανονικότητα, εφ' όσον ο αριθμός των παρατηρήσεων του «δείγματος» είναι



Εικόνα 22. Κανονικό διάγραμμα Q-Q της διάρκειας νοσηλείας των 223 ατόμων του πίνακα 5.



Εικόνα 23. Διάγραμμα πλαισίου της διάρκειας νοσηλείας των 223 ατόμων του πίνακα 5.

σχετικά μεγάλος (συνήθως >100 παρατηρήσεις). Εάν όμως ο αριθμός των παρατηρήσεων του «δείγματος» είναι μικρός (συνήθως <30 παρατηρήσεις), τότε είναι ασφαλέστερο να εφαρμόζονται μη παραμετρικές μέθοδοι.

8. ΣΥΝΟΨΗ

Η μονομεταβλητή ανάλυση στην περίπτωση των ονομαστικών μεταβλητών αφορά στην παράθεση των απόλυτων και των σχετικών συχνοτήτων, ενώ στην περίπτωση των ποσοτικών μεταβλητών αναφέρεται στην παράθεση των κατάλληλων μέτρων θέσης και διασποράς. Ιδιαίτερη περίσκεψη απαιτείται στην περίπτωση των διατάξιμων μεταβλητών, όπου είναι δυνατή η παράθεση τόσο των απόλυτων και των σχετικών συχνοτήτων όσο και των μέτρων θέσης και διασποράς. Σημειώνεται ότι στην περίπτωση των διατάξιμων μεταβλητών η παράθεση των μέτρων θέσης και διασποράς έχει νόημα μόνο εφ' όσον οι μεταβλητές αυτές αντιμετωπιστούν από τους ερευνητές ως μεταβλητές κλίμακας λόγου.

Στην περίπτωση των ποσοτικών μεταβλητών που ακολουθούν την κανονική κατανομή θα πρέπει να παρουσιάζονται ο μέσος και η τυπική απόκλιση, ενώ στην περίπτωση που δεν ακολουθούν την κανονική κατανομή θα πρέπει να παρουσιάζονται η διάμεσος και το ενδοτεταρτημοριακό εύρος ή το εύρος.

Στην περίπτωση των διατάξιμων μεταβλητών, ανεξάρτητα από το αν ακολουθούν ή όχι την κανονική κατανομή, θα πρέπει να παρουσιάζονται η διάμεσος και το ενδοτεταρτημοριακό εύρος ή το εύρος.

Η μονομεταβλητή ανάλυση των ποσοτικών μεταβλητών πρέπει οπωσδήποτε να περιλαμβάνει τον έλεγχο κανονικότητας, καθώς στην περίπτωση που ακολουθούν την κανονική κατανομή είναι δυνατή η εφαρμογή των παραμετρικών μεθόδων στη διμεταβλητή και την πολυμεταβλητή ανάλυση, ενώ στην περίπτωση που δεν ακολουθούν την κανονική κατανομή είναι δυνατή η εφαρμογή των μη παραμετρικών μεθόδων.

Πίνακας 9. Ο αλγόριθμος της λήψης απόφασης σχετικά με το αν η ηλικία, ο δείκτης μάζας σώματος και η διάρκεια νοσηλείας, αναφορικά με τα 223 άτομα του πίνακα 5, ακολουθούν ή όχι την κανονική κατανομή, με το «ναι» να δηλώνει ισχυρή ένδειξη της ύπαρξης κανονικής κατανομής, το «πιθανώς» να δηλώνει μέτρια ένδειξη της ύπαρξης κανονικής κατανομής και το «όχι» να δηλώνει ισχυρή ένδειξη της ύπαρξης ασυμμετρίας.

Μεταβλητή	Σύγκριση μεταξύ μέσου και διαμέσου	Μέση τιμή ±2 τυπικές αποκλίσεις	Συντελεστές ασυμμετρίας και κύρτωσης	Στατιστικοί έλεγχοι	Γραφήματα	Συμπέρασμα
Ηλικία (έτη)	Ναι	Πιθανώς	Πιθανώς	Όχι	Πιθανώς	Ναι
Δείκτης μάζας σώματος (kg/m^2)	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι
Διάρκεια νοσηλείας (ημέρες)	Όχι	Όχι	Όχι	Όχι	Όχι	Όχι

ABSTRACT

Univariate analysis of epidemiological data

P. GALANIS

Center for Health Services Management and Evaluation, Department of Nursing, National and Kapodistrian University of Athens, Athens, Greece

Archives of Hellenic Medicine 2014, 31(2):221–243

Descriptive statistics are used for the concise and detailed presentation of data in epidemiological studies, while inferential statistics are applied for the investigation of relationships between determinants and outcomes. Descriptive statistics concern the presentation of epidemiological data (univariate analysis), and inferential statistics include bivariate and multivariate analysis. Univariate analysis permits the separate presentation of each variable of a study, bivariate analysis the investigation of the relationship between a determinant and an outcome and multivariate analysis the investigation of the relationship between a determinant and an outcome, taking into consideration the effect of potential confounding and modifying factors. Univariate analysis permits the presentation of absolute and relative frequencies of nominal variates and the presentation of appropriate measures of location and dispersion of quantitative variates. Particular consideration is required in the case of ordinal variates, where the presentation is feasible of both absolute and relative frequencies and measures of location and dispersion. Measures of location are the values around which the observations appear to mass to a higher degree, while measures of dispersion capture the degree to which the observations are spread out. The most important measures of location are the mean, the median and the mode. The most important measures of dispersion are the range, the interquartile range, the variance, the standard deviation and the coefficient of variation. In the case of continuous variates, univariate analysis must include a check of normality of distribution. When continuous variates follow a normal distribution, the application of parametric methods in bivariate and multivariate analysis is feasible, but when they do not follow a normal distribution the application of non-parametric methods is required. A check of normality can be made as follows: (a) comparison between the mean and the median values, (b) estimation of the coefficient of skewedness and kurtosis, (c) application of appropriate statistical tests (e.g., Kolmogorov-Smirnov, Shapiro-Wilk), and (d) use of graphs (histograms, normal Q-Q plots and box plots). Quantitative variates that follow a normal distribution should be presented as mean and standard deviation, while those that do not follow normal distribution should be presented as median and interquartile range or range. Ordinal variates, regardless of whether or not they follow a normal distribution, should be presented as median and interquartile range or range.

Key words: Data analysis, Measures of dispersion, Measures of location, Normal distribution, Univariate analysis

Βιβλιογραφία

1. DODGE Y. *The concise encyclopedia of statistics.* Springer Science & Business Media, Berlin, 2008:518–520
2. ΓΑΛΑΝΗΣ ΠΑ, ΣΠΑΡΟΣ ΛΔ. *Εγχειρίδιο Επιδημιολογίας*. Εκδόσεις ΒΗΤΑ, Αθήνα, 2010:61–78, 151–156
3. PEAT J, BARTON B. *Medical statistics. A guide to data analysis and critical appraisal.* BMJ Books, Massachusetts, 2005:1–50
4. BOWERS D. *Medical statistics from scratch. An introduction for health professionals.* 2nd ed. John Willey & Sons, New Jersey, 2008:1–68
5. BOWERS D, HOUSE A, OWENS D. *Understanding clinical papers.* 2nd ed. John Willey & Sons, New Jersey, 2006:63–91
6. STEWART A. *Basic statistics and epidemiology. A practical guide.* Radcliffe Medical Press, Oxford, 2002:1–38
7. CHERNICK M, FRIIS R. *Introductory biostatistics for the health sciences.* John Willey & Sons, New Jersey, 2003:46–91, 121–132
8. SHASHA D, WILSON M. *Statistics is easy.* Morgan & Claypool Publishers, Washington, 2008:1–11
9. RUGG G. *Using statistics: A gentle introduction.* Open University Press, Berkshire, 2007:1–45
10. BRASE CH, BRASE CP. *Understanding basic statistics.* 4th ed. Houghton Mifflin Company, Boston, 2007:2–31
11. ΓΑΛΑΝΗΣ Π. *Μεθοδολογία ανάλυσης δεδομένων στις επιστήμες υγείας. Εφαρμογές με το IBM SPSS Statistics.* Broken Hill Publishers Ltd, Αθήνα, 2014
12. ΤΡΙΧΟΠΟΥΛΟΣ Δ, ΤΖΩΝΟΥ Α, ΚΑΤΣΟΥΓΙΑΝΝΗ Κ. *Βιοστατιστική.* Εκδόσεις Παρισιάνου, Αθήνα, 2000:1–30
13. ΓΑΛΑΝΗΣ Π. Στατιστικές μέθοδοι ανάλυσης δεδομένων. *Αρχ Ελλ Ιατρ* 2009, 26:699–711
14. DAWSON B, TRAPP R. *Basic and clinical biostatistics.* 4th ed. McGraw-Hill, New York, 2004:24–93

15. BOSLAUGH S, WATTERS P. *Statistics in a nutshell.* O'Reilly Media Inc, Cambridge, 2008:54–84, 125–150
16. GOOD P, HARDIN J. *Common errors in statistics.* 2nd ed. John Willey & Sons, New Jersey, 2006:125–144
17. RUMSEY D. *Statistics for dummies*. Wiley Publishing Inc, Indianapolis, 2003:59–114
18. MYATT GJ. *Making sense of data. A practical guide to exploratory data analysis and data mining.* John Willey & Sons, New Jersey, 2007:17–62
19. UTTS J. *Seeing through statistics.* 2nd ed. Duxbury Press, Pacific Grove, 2005:107–156
20. VAN BELLE G. *Statistical rules of thumb.* 2nd ed. John Willey & Sons, New Jersey, 2008:193–216
21. JAISINGH L. *Statistics for the utterly confused.* McGraw-Hill, New York, 2000:1–81, 144–188
22. TUFTE E. *The visual display of quantitative information.* 2nd ed. Graphic Press, Connecticut, 2001:13–90
23. HANRAHAN E, MADUPU G. *Epidemiology and biostatistics for the USMLE.* Appleton & Lange, Connecticut, 1994:49–56
24. SARDANELLI F, DI LEO G. *Biostatistics for radiologists.* Springer-Verlag Italia, Milan, 2009:41–61
25. HAND J. *Statistics: A very short introduction.* Oxford University Press, Oxford, 2008:21–35
26. MATTHEWS D, FAREWELL V. *Using and understanding medical statistics.* 4th ed. Karger, Basel, 2007:76–110

Corresponding author:

P. Galanis, 14 Dikis street, GR-157 73 Athens, Greece
e-mail: pegalan@nurs.uoa.gr