ΝΟΣΗΛΕΥΤΙΚΗ 2011, 50(2): 132–146 • NOSILEFTIKI 2011, 50(2): 132–146

ΓΕΝΙΚΟ ΑΡΘΡΟ – GENERAL ARTICLE

Διαχείριση Μεταβλητών και Δεδομένων στις Επιδημιολογικές Μελέτες

Πέτρος Γαλάνης

Management of Variates and Data

Abstract at the end of the article

Νοσηλευτής ΠΕ, MSc Δημόσιας Υγείας, Διδάκτωρ Πανεπιστημίου Αθηνών, Εργαστήριο Οργάνωσης και Αξιολόγησης Υπηρεσιών Υγείας, Τμήμα Νοσηλευτικής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Αθήνα

Τμήμα Νοσηλευτικής
Πανεπιστήμιο Αθηνών

Οι επιστήμονες υγείας καλούνται στις επιδημιολογικές μελέτες να συλλέξουν τα δεδομένα όσο το δυνατόν πιο αξιόπιστα και αντικειμενικά, έτσι ώστε η στατιστική ανάλυση να οδηγήσει σε ακριβή και έγκυρα συμπεράσματα. Η λανθασμένη καταγραφή των δεδομένων που αφορούν στις περιπτώσεις μιας μελέτης, η λανθασμένη εισαγωγή των δεδομένων στις κατάλληλες ηλεκτρονικές βάσεις δεδομένων και η άγνοια της διαχείρισης των βάσεων δεδομένων μειώνουν σημαντικά την αξιοπιστία της ανάλυσης των δεδομένων, οδηγώντας σε μη έγκυρα αποτελέσματα. Πριν αρχίσει η ανάλυση των δεδομένων, είναι απαραίτητη τόσο η σωστή εισαγωγή των στοιχείων στη βάση δεδομένων που πρόκειται να χρησιμοποιηθεί για την ανάλυση όσο και η σωστή διαχείριση και η κωδικοποίηση των μεταβλητών, προκειμένου να επισημανθούν και να διορθωθούν τυχόν λάθη και παραλείψεις. Οι μεταβλητές –και κατ' επέκταση και τα δεδομένα– ανάλογα με τα μαθηματικά τους χαρακτηριστικά διακρίνονται σε ποιοτικές και ποσοτικές, με τις πρώτες να διακρίνονται σε ονομαστικές και διατάξιμες και τις δεύτερες σε μεταβλητές διαστηματικής κλίμακας και μεταβλητές κλίμακας λόγου. Επιπλέον, οι ποσοτικές μεταβλητές διακρίνονται σε συνεχείς και ασυνεχείς. Ιδιαίτερη σημασία στην ανάλυση των επιδημιολογικών δεδομένων έχουν οι απούσες τιμές και οι απομακρυσμένες παρατηρήσεις. Απούσες τιμές καλούνται οι τιμές εκείνες στις οποίες απουσιάζουν οι παρατηρήσεις για τις διάφορες μεταβλητές. Όσο αυξάνεται το ποσοστό των απουσών τιμών σε μια ανάλυση τόσο μειώνεται η αξιοπιστία των αποτελεσμάτων της ανάλυσης. Οι απομακρυσμένες παρατηρήσεις αφορούν σε παρατηρήσεις, οι τιμές των οποίων διαφέρουν σημαντικά από τις τιμές των υπολοίπων παρατηρήσεων. Οι απομακρυσμένες παρατηρήσεις μπορεί να οφείλονται σε λανθασμένη καταγραφή των παρατηρήσεων κατά τη συλλογή των δεδομένων ή σε λανθασμένη εισαγωγή των παρατηρήσεων στη βάση δεδομένων ή να αποτελούν πραγματικές τιμές που απλά διαφέρουν σημαντικά από τις τιμές των υπολοίπων παρατηρήσεων.

Λέξεις ευρετηρίου: *Ανάλυση δεδομένων, απομακρυσμένες παρατηρήσεις, απούσες τιμές, βάσεις δεδομένων, ποιοτικές μεταβλητές, ποσοτικές μεταβλητές*

Υποβλήθηκε: 30.6.2010
Επανυποβλήθηκε: 22.11.2010
Εγκρίθηκε: 27.11.2010

Υπεύθυνος αλληλογραφίας:
Πέτρος Γαλάνης
Δίκης 14, 157 73 Αθήνα
Τηλ.: 210 77 81 044, 694 43 87 354
e-mail: pegalan@nurs.uoa.gr

Εισαγωγή

Η Στατιστική είναι ο επιστημονικός κλάδος που αφορά στη συλλογή, στην οργάνωση και στην ανάλυση (ή καλύτερα στη σύνθεση) δεδομένων που υπό-

κεινται σε τυχαία μεταβλητότητα.[1] Σημειώνεται, ότι ο όρος «ανάλυση δεδομένων» είναι εσφαλμένος εννοιολογικά, καθώς οι παρατηρήσεις δεν είναι δεδομένα[1*] και η επεξεργασία των παρατηρήσεων είναι σύνθεση και όχι ανάλυση.[2] Στην ουσία, η ανάλυση των δεδομένων είναι το σύνολο της μαρτυρίας ή, αλλιώς, της ένδειξης (evidence) που η εμπειρική έρευνα χρησιμοποιεί για τον έλεγχο της υπόθεσης.

Η λανθασμένη καταγραφή των δεδομένων που αφορούν στις περιπτώσεις μιας επιδημιολογικής μελέτης, η λανθασμένη εισαγωγή των δεδομένων στις κατάλληλες ηλεκτρονικές βάσεις δεδομένων και η άγνοια της διαχείρισης των βάσεων δεδομένων μειώνουν σημαντικά την αξιοπιστία της ανάλυσης των δεδομένων, οδηγώντας σε μη έγκυρα αποτελέσματα. Ο σχεδιασμός της μελέτης προϋποθέτει τη συνεργασία με το στατιστικό, έτσι ώστε να διασφαλίζεται η πλέον αποτελεσματική ανάλυση των δεδομένων. Για παράδειγμα, σε ορισμένες περιπτώσεις οι επιστήμονες υγείας μετατρέπουν την ηλικία σε ποιοτική μεταβλητή, καλώντας τους συμμετέχοντες να δηλώσουν την ηλικιακή ομάδα στην οποία ανήκουν (π.χ. 18–30 έτη, 30–50 έτη και > 50 έτη) και όχι την ακριβή τους ηλικία. Με τον τρόπο αυτόν όμως χάνεται πολύτιμη πληροφορία, καθώς μια ποσοτική μεταβλητή, όπως η ηλικία, μετατρέπεται σε ποιοτική. Επιπλέον, το γεγονός αυτό έχει καταλυτική σημασία στα αποτελέσματα που προκύπτουν από την ανάλυση των δεδομένων, καθώς είναι τελείως διαφορετικός ο τρόπος με τον οποίο αναλύονται οι ποσοτικές μεταβλητές σε σχέση με τις ποιοτικές. Είναι σαφές εξάλλου ότι καταγράφοντας την ακριβή ηλικία των συμμετεχόντων είναι δυνατή ακολούθως η κατηγοριοποίηση της ηλικίας με διάφορα μάλιστα όρια. Το αντίθετο βέβαια δεν μπορεί να συμβεί, καθώς καταγράφοντας την ηλικιακή ομάδα στην οποία ανήκει ένα άτομο, δεν είναι δυνατόν να καταστεί γνωστή η ακριβής ηλικία του. Οι επιστήμονες υγείας καλούνται να συλλέξουν τα δεδομένα όσο το δυνατόν πιο αξιόπιστα και αντικειμενικά, έτσι ώστε η στατιστική ανάλυση να οδηγήσει σε ακριβή και έγκυρα συμπεράσματα.

Δυστυχώς, σε αρκετές περιπτώσεις, η άγνοια της διαχείρισης των ηλεκτρονικών βάσεων δεδομένων και η επιπολαιότητα αυτών που εισάγουν τα δεδομένα οδηγούν στη λανθασμένη εισαγωγή δεδομένων. Για το λόγο αυτόν, πριν αρχίσει η ανάλυση των δεδομένων είναι απαραίτητος ο λεγόμενος «καθαρισμός» των δεδομένων (data cleaning), έτσι ώστε να επισημανθούν και να διορθωθούν τυχόν λάθη και παραλείψεις.[3]

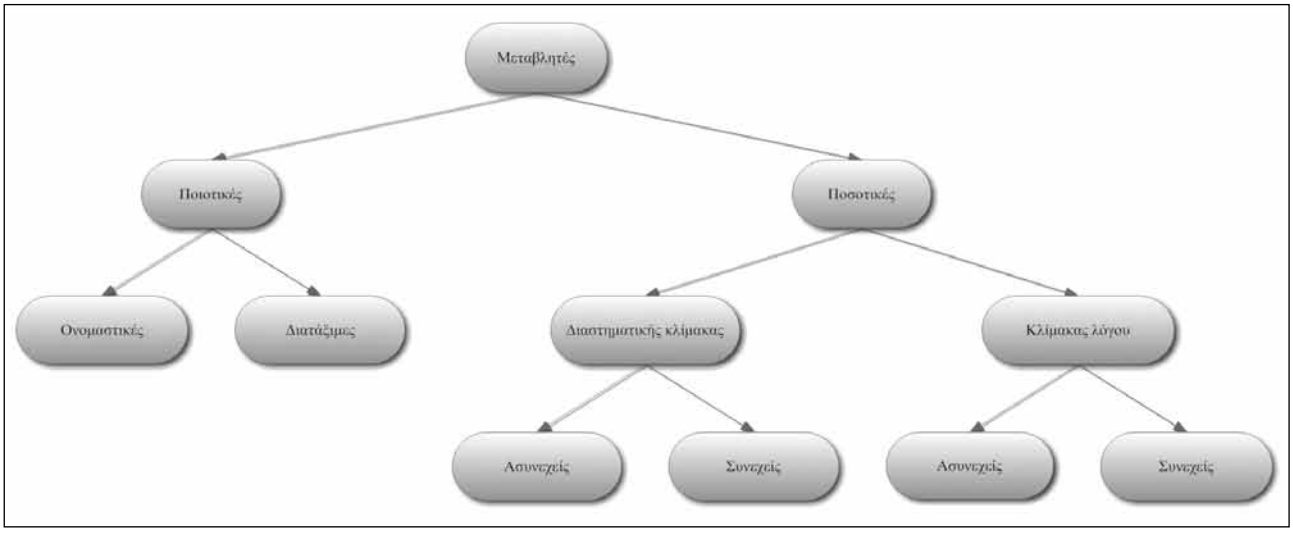
Μεταβλητές και δεδομένα

Ο εμπειρικός καθορισμός των τιμών μιας αριθμητικής συνάρτησης ή μιας ποσοτικής έννοιας καλείται μέτρηση. Ο όρος «μέτρηση» συχνά χρησιμοποιείται και με ευρύτερη έννοια, περιλαμβάνοντας και την κατανομή των αντικειμένων μιας κατηγορίας σε τάξεις (ποιοτικές έννοιες). 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 την 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 που 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 (εικόνα 1).[3–11] Οι ποσοτικές μεταβλητές εξάλλου 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 που η μεταβλητή είναι το φύλο, τα δεδομένα μπορούν να είναι είτε άνδρες είτε γυναίκες, ενώ στην περίπτωση που η μεταβλητή είναι η ομάδα αίματος, τα δεδομένα μπορούν να είναι ομάδα αίματος Α, Β, ΑΒ ή Ο κ.λπ. (πίνακας 1).

[1*] *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 εξάλλου διακρίνει τις παρατηρήσεις από τις εντυπώσεις (impressions).*[2]



Εικόνα 1. Είδη μεταβλητών.

Ποιοτικές μεταβλητές

Ονομαστικές μεταβλητές

Οι *ονομαστικές μεταβλητές* (nominal variates) αποτελούν την απλούστερη και πλέον συνήθη μορφή μεταβλητών.[3–11] 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 να συμβολιστούν οι άνδρες με 1 και οι γυναίκες με 0, χωρίς και πάλι οι αριθμοί να έχουν νόημα και χωρίς να επηρεαστεί η ανάλυση των δεδομένων.

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 ανα-

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

	Άτομο Α	Άτομο Β	Άτομο Γ	Άτομο Δ
Φύλο	Άνδρας	Γυναίκα	Άνδρας	Γυναίκα
Ηλικία (έτη)	34	43	29	23
Ομάδα αίματος	Α	Β	ΑΒ	Ο
Βάρος (kg)	85	67	90	55

Οι μεταβλητές

Τα δεδομένα

λύεται. 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.

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.[3–11] 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 τάξεων αυτών 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 λαμβάνει συνήθως 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. 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[2*] 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 αυτή. Κάθε στοιχείο Likert αποτελεί μια πρόταση, στην οποία οι συμμετέχοντες καλούνται να δηλώσουν το βαθμό συμφωνίας τους (ή το βαθμό διαφωνίας τους). Συνήθως, υπάρχουν 5 –ή σπανιότερα 7 ή 9– απαντήσεις σε διατεταγμένη κλίμακα και οι συμμετέχοντες καλούνται να επιλέξουν εκείνη που τους εκφράζει περισσότερο. Η τυπική δομή ενός στοιχείου Likert στο οποίο υπάρχουν 5 πιθανές απαντήσεις σε διατεταγμένη κλίμακα είναι η εξής:

1. Διαφωνώ τελείως
2. Διαφωνώ
3. Ούτε διαφωνώ, ούτε συμφωνώ
4. Συμφωνώ
5. Συμφωνώ τελείως.

Σε ορισμένες περιπτώσεις εξάλλου 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, εφόσον επιλεγεί, πρέπει να αναφέρονται με σαφήνεια τα κριτήρια και οι προϋποθέσεις εφαρμογής της.

Χαρακτηριστικά παραδείγματα διατάξιμων μεταβλητών αποτελούν η κλίμακα Γλασκόβης[3*] (Glasgow scale) και η κλίμακα Apgar[4*] (Apgar scale). Η κλίμακα Γλασκόβης για ενήλικες χρησιμοποιείται για την εκτίμηση του επιπέ-

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

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

[4]Η κλίμακα Apgar εισήχθη το 1952 από την Αμερικανίδα αναισθησιολόγο Virginia Apgar (1909–1974) και χρησιμοποιείται για την εκτίμηση της γενικής κατάστασης των νεογνών, το πρώτο και το πέμπτο λεπτό μετά από 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 τιμές να δηλώνουν καλύτερη κατάσταση του νεογνού.*

δου συνείδησης και αποτελείται από τρία στοιχεία: Την αντίδραση 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 στον πόνο, 3: άνοιγμα οφθαλμών ως αντίδραση στα λεκτικά παραγγέλματα, 4: αυτόματο άνοιγμα οφθαλμών), όσον αφορά στη λεκτική αντίδραση λαμβάνονται τιμές 1–5 (1: καμιά λεκτική απάντηση, 2: ακατάληπτοι ήχοι, 3: ακατάληπτες λέξεις, 4: συγχυτική ομιλία, 5: προσανατολισμένη ομιλία) και σχετικά με την κινητική αντίδραση λαμβάνονται τιμές 1–6 (1: καμιά κινητική αντίδραση, 2: έκταση ως αντίδραση στον πόνο, 3: κάμψη ως αντίδραση στον πόνο, 4: απόσυρση ως αντίδραση στον πόνο, 5: εντοπισμός του σημείου του πόνου, 6: «υπακοή» σε λεκτικά παραγγέλματα). Έτσι, η κλίμακα Γλασκόβης λαμβάνει τιμές 3–15, με τις μεγαλύτερες αυτών 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[5*] που λαμβάνει τιμές 0–71 σε σχέση με την κλίμακα Γλασκόβης που λαμβάνει τιμές 3–15.

Ποσοτικές μεταβλητές

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. 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 (Richter scale) στην οποία οι τιμές της μεταβλητής μπορεί να απέχουν μεταξύ τους κατά ίση αριθμητική διαφορά, αλλά αυτό δε συνεπάγεται κιόλας ότι οι σχετικές 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 δε 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 εφικτό 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.

Μεταβλητές κλίμακας λόγου

Οι *μεταβλητές κλίμακας λόγου* (ratio scale variates) υπερέχουν έναντι των μεταβλητών διαστηματικής κλίμακας λόγω του 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 μπορεί 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 €. Επιπλέον, ένα άτομο βάρους 100 kg είναι δύο φορές βαρύτερο από ένα άτομο βάρους 50 kg.

Συνεχείς και ασυνεχείς μεταβλητές

Οι ποσοτικές μεταβλητές εξάλλου διακρίνονται σε *συνεχείς* (continuous) και ασυνεχείς ή, αλλιώς, *διακριτές* (discrete).[3–11] Η διαφορά τους έγκειται στο γεγονός ότι οι ασυνεχείς μεταβλητές περιορίζονται στο να λαμβάνουν μόνο ακέραιες τιμές, που διαφέρουν μεταξύ τους κατά συγκεκριμένες ποσότητες, ενώ οι συνεχείς μεταβλητές μπορεί να λάβουν οποιαδήποτε τιμή των πραγματικών αριθμών, ακόμη και δεκαδική. Οι περισσότερες μεταβλητές διαστηματικής κλίμακας και κλίμακας λόγου αποτελούν

5 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 τιμές να δηλώνουν πιο βαριά κατάσταση και μεγαλύτερο κίνδυνο θανάτου.*

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 σ' ένα συγκεκριμένο χρονικό διάστημα κ.ά., ενώ παραδείγματα συνεχών 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 kg για το βάρος, τα cm για το ύψος, τα έτη για την ηλικία κ.λπ.

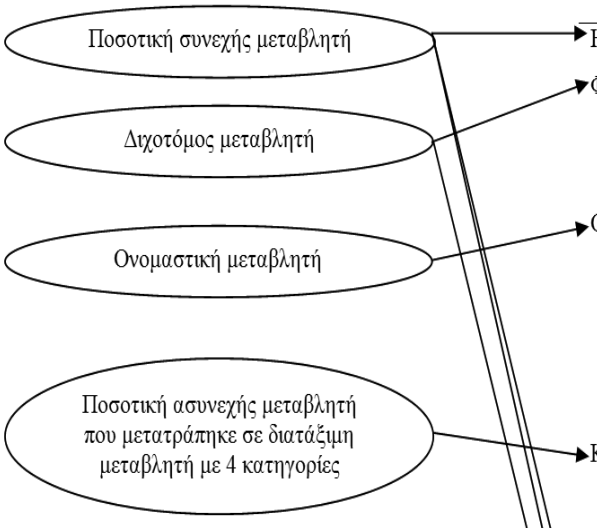
Σημειώνεται, ότι οι ποσοτικές μεταβλητές είναι δυνατόν να μετατραπούν σε διατάξιμες μεταβλητές χρησιμοποιώντας κάποια διαχωριστικά όρια, ενώ το αντίθετο δεν είναι δυνατόν να συμβεί. Για παράδειγμα, ανάλογα με το δείκτη μάζας σώματος, τα άτομα μπορούν να ταξινομηθούν σε αδύνατα, φυσιολογικά, υπέρβαρα ή παχύσαρκα. Επιπλέον, οι τιμές της χοληστερόλης του 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».

Βάσεις δεδομένων

Η ευρεία χρήση των ηλεκτρονικών υπολογιστών και των στατιστικών προγραμμάτων ανάλυσης δεδομένων, όπως το PASW (SPSS Inc., IBM Company, Chicago, Illinois, USA), το STATA (Stata Corp LP, Texas, USA), το SAS (SAS Institute Inc., North Carolina, USA), το Statgraphics (Statpoint Inc., Virginia, USA), το Statistica (Statsoft Inc., Tulsa, USA) κ.ά., έχει διευκολύνει σημαντικά την ανάλυση δεδομένων ακόμη και για αυτούς που δεν είναι ιδιαίτερα

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)
Καπνιστική 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*

εξοικειωμένοι με τις έννοιες της Στατιστικής. Τα στατιστικά αυτά προγράμματα παρέχουν τη δυνατότητα εισαγωγής, επεξεργασίας και ανάλυσης δεδομένων, έχοντας το καθένα διαφορετικά πλεονεκτήματα και μειονεκτήματα. Σημειώνεται ότι το PASW[12] (Predictive Analytics SoftWare) είναι από το 2009 η νέα ονομασία του SPSS[13–15] (Statistical Package for Social Sciences).

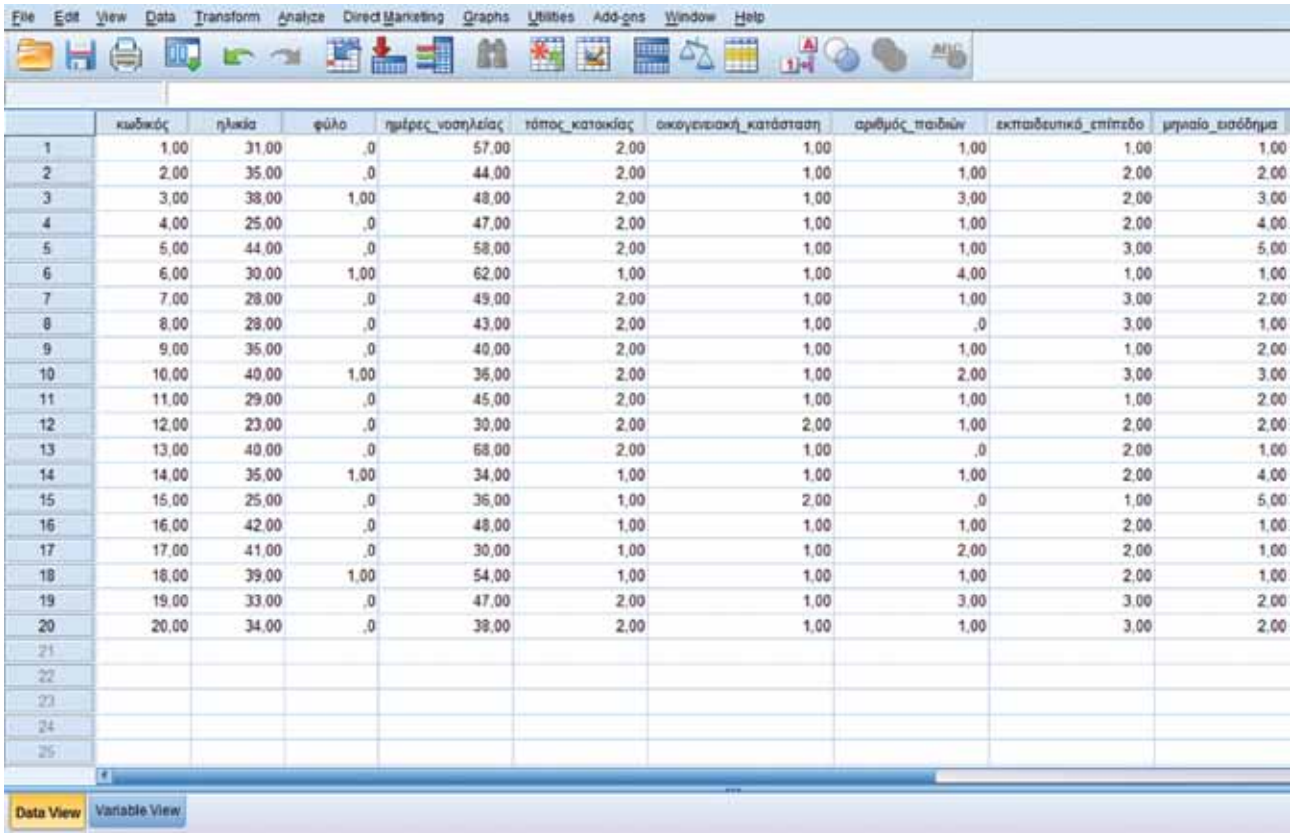
Είναι αρκετά σύνηθες εξάλλου, τα δεδομένα, αρχικά, να εισάγονται σ' ένα φύλλο εργασίας του Microsoft Office Excel ή σε μια βάση δεδομένων της Microsoft Office Access και στη συνέχεια να μεταφέρονται σ' ένα στατιστικό πρόγραμμα ανάλυσης δεδομένων. Στην περίπτωση αυτή πάντως απαιτείται ιδιαίτερη προσοχή κατά τη μεταφορά των δεδομένων, έτσι ώστε να αποφεύγονται τυχόν λάθη και παραλείψεις, καθώς τα στατιστικά προγράμματα χαρακτηρίζονται από διάφορες ιδιαιτερότητες που πρέπει να λαμβάνονται σοβαρά υπόψη. Ενδεικτικά αναφέρεται το γεγονός ότι η πρώτη γραμμή σ' ένα φύλλο εργασίας του Microsoft Office Excel αφορά στα ονόματα των μεταβλητών και όχι στα δεδομένα μιας συγκεκριμένης περίπτωσης, ενώ η πρώτη γραμμή σε μια βάση δεδομένων του PASW αφορά στα δεδομένα μιας συγκεκριμένης περίπτωσης. Έτσι, κατά τη μεταφορά μιας βάσης δεδομένων από το Microsoft Office Excel στο PASW, εάν στο φύλλο του Excel υπάρχουν n γραμμές, τότε στη βάση δεδομένων του PASW θα πρέπει να υπάρχουν $n-1$ γραμμές, εφόσον βέβαια η πρώτη γραμμή στο φύλλο εργασίας του Excel αφορά στα ονόματα των μεταβλητών (εικόνες 2 και 3).

Το PASW είναι σχετικά εύχρηστο για το μέσο χρήστη ηλεκτρονικών υπολογιστών και εξαιρετικά διαδεδομένο και γι' αυτό θα χρησιμοποιηθεί για τη δημιουργία μιας ενδεικτικής βάσης δεδομένων. Το PASW αποτελείται ουσιαστικά από δύο καρτέλες, που είναι τελείως διαφορετικές μεταξύ τους τόσο στη δομή όσο και στη λειτουργία. Στην καρτέλα «απεικόνιση δεδομένων» (data view) εισάγονται τα δεδομένα (εικόνα 3), ενώ στην καρτέλα «απεικόνιση μεταβλητών» (variable view) πραγματοποιείται η διαχείριση των μεταβλητών (εικόνα 4).

Το πρώτο βήμα στη δημιουργία μιας βάσης δεδομένων στο PASW αποτελεί η κωδικοποίηση των μεταβλητών, που πραγματοποιείται μέσω της καρτέλας «απεικόνιση μεταβλητών». Στην εικόνα 4, στην καρτέλα «απεικόνιση μεταβλητών» φαίνονται αναλυτικά τα χαρακτηριστικά και η κωδικοποίηση κάθε μεταβλητής της βάσης δεδομένων της εικόνας 3. Στη συγκεκριμένη βάση δεδομένων συμπεριλαμβάνονται 9 μεταβλητές (κωδικός περίπτωσης, ηλικία, φύλο, ημέρες νοσηλείας, τόπος κατοικίας, οικογενειακή κατάσταση, αριθμός παιδιών, εκπαιδευτικό επίπεδο και μηνιαίο εισόδημα) και 20 περιπτώσεις. Σχεδόν πάντα, η πρώτη μεταβλητή σε μια βάση δεδομένων αφορά στον κωδικό των περιπτώσεων, επιτρέποντας με τον τρόπο αυτόν τη σχετικά εύκολη και άμεση εύρεση μιας

	A	B	C	D	E	F	G	H	I
1	κωδικός	ηλικία	φύλο	ημέρες νοσηλείας	τόπος κατοικίας	οικογενειακή κατάσταση	αριθμός παιδιών	εκπαιδευτικό επίπεδο	μηνιαίο εισόδημα
2	1	31	0	57	2	1	1	1	1
3	2	35	0	44	2	1	1	2	2
4	3	38	1	48	2	1	3	2	3
5	4	25	0	47	2	1	1	2	4
6	5	44	0	58	2	1	1	3	5
7	6	30	1	62	1	1	4	1	1
8	7	28	0	49	2	1	1	3	2
9	8	28	0	43	2	1	0	3	1
10	9	35	0	40	2	1	1	1	2
11	10	40	1	36	2	1	2	3	3
12	11	29	0	45	2	1	1	1	2
13	12	23	0	30	2	2	1	2	2
14	13	40	0	68	2	1	0	2	1
15	14	35	1	34	1	1	1	2	4
16	15	25	0	36	1	2	0	1	5
17	16	42	0	48	1	1	1	2	1
18	17	41	0	30	1	1	2	2	1
19	18	39	1	54	1	1	1	2	1
20	19	33	0	47	2	1	3	3	2
21	20	34	0	38	2	1	1	3	2
22									
23									

Εικόνα 2. Φύλλο εργασίας του Microsoft Office Excel, στο οποίο υπάρχουν 20 περιπτώσεις. Οι 21 διαφορετικές γραμμές αντιπροσωπεύουν τα δεδομένα 20 διαφορετικών περιπτώσεων, καθώς η πρώτη γραμμή αντιπροσωπεύει τα ονόματα των μεταβλητών. Οι 9 διαφορετικές στήλες αντιπροσωπεύουν 9 διαφορετικές μεταβλητές (κωδικός, ηλικία, φύλο, ημέρες νοσηλείας, τόπος κατοικίας, οικογενειακή κατάσταση, αριθμός παιδιών, εκπαιδευτικό επίπεδο και μηνιαίο εισόδημα).



	κωδικός	ηλικία	φύλο	ημέρες_νοσηλείας	τόπος_κατοικίας	οικογενειακή_κατάσταση	αριθμός_παιδιών	εκπαιδευτικό_επίπεδο	μηνιαίο_εισόδημα
1	1,00	31,00	,0	57,00	2,00	1,00	1,00	1,00	1,00
2	2,00	35,00	,0	44,00	2,00	1,00	1,00	2,00	2,00
3	3,00	38,00	1,00	48,00	2,00	1,00	3,00	2,00	3,00
4	4,00	25,00	,0	47,00	2,00	1,00	1,00	2,00	4,00
5	5,00	44,00	,0	58,00	2,00	1,00	1,00	3,00	5,00
6	6,00	30,00	1,00	62,00	1,00	1,00	4,00	1,00	1,00
7	7,00	28,00	,0	49,00	2,00	1,00	1,00	3,00	2,00
8	8,00	28,00	,0	43,00	2,00	1,00	,0	3,00	1,00
9	9,00	35,00	,0	40,00	2,00	1,00	1,00	1,00	2,00
10	10,00	40,00	1,00	36,00	2,00	1,00	2,00	3,00	3,00
11	11,00	29,00	,0	45,00	2,00	1,00	1,00	1,00	2,00
12	12,00	23,00	,0	30,00	2,00	2,00	1,00	2,00	2,00
13	13,00	40,00	,0	68,00	2,00	1,00	,0	2,00	1,00
14	14,00	35,00	1,00	34,00	1,00	1,00	1,00	2,00	4,00
15	15,00	25,00	,0	36,00	1,00	2,00	,0	1,00	5,00
16	16,00	42,00	,0	48,00	1,00	1,00	1,00	2,00	1,00
17	17,00	41,00	,0	30,00	1,00	1,00	2,00	2,00	1,00
18	18,00	39,00	1,00	54,00	1,00	1,00	1,00	2,00	1,00
19	19,00	33,00	,0	47,00	2,00	1,00	3,00	3,00	2,00
20	20,00	34,00	,0	38,00	2,00	1,00	1,00	3,00	2,00
21									
22									
23									
24									
25									

Εικόνα 3. Βάση δεδομένων του PASW (καρτέλα «απεικόνιση δεδομένων») που προέκυψε από τη μεταφορά των δεδομένων του φύλλου εργασίας του Microsoft Office Excel της εικόνας 2. Οι 20 διαφορετικές γραμμές αντιπροσωπεύουν τα δεδομένα 20 διαφορετικών περιπτώσεων. Οι 9 διαφορετικές στήλες αντιπροσωπεύουν 9 διαφορετικές μεταβλητές (κωδικός, ηλικία, φύλο, ημέρες νοσηλείας, τόπος κατοικίας, οικογενειακή κατάσταση, αριθμός παιδιών, εκπαιδευτικό επίπεδο και μηνιαίο εισόδημα).

συγκεκριμένης περίπτωσης σε οποιαδήποτε στιγμή της ανάλυσης των δεδομένων.

Στην καρτέλα «απεικόνιση μεταβλητών», η πρώτη στήλη έχει τίτλο «όνομα» (name) και αφορά στο όνομα κάθε μεταβλητής. Στη στήλη αυτή αναφέρεται με συντομία το όνομα κάθε μεταβλητής, ενώ υπάρχουν και ορισμένοι περιορισμοί, όπως π.χ. δεν επιτρέπονται κενά μεταξύ των χαρακτήρων, ο πρώτος χαρακτήρας πρέπει να είναι γράμμα, δεν επιτρέπεται το ίδιο όνομα για δύο μεταβλητές κ.ά. Τα ονόματα των μεταβλητών που εισάγονται στην πρώτη στήλη («όνομα») της καρτέλας «απεικόνιση μεταβλητών» (εικόνα 4) είναι ταυτόχρονα και τα ονόματα των στηλών στην καρτέλα «απεικόνιση δεδομένων» (εικόνα 3). Έτσι, στην καρτέλα «απεικόνιση μεταβλητών», το όνομα της μεταβλητής στην πρώτη γραμμή αντιστοιχεί στο όνομα της πρώτης στήλης των δεδομένων στην καρτέλα «απεικόνιση δεδομένων», το όνομα της μεταβλητής στη δεύτερη γραμμή αντιστοιχεί στο όνομα της δεύτερης στήλης των δεδομένων κ.ο.κ.

Στην καρτέλα «απεικόνιση μεταβλητών», η δεύτερη στήλη έχει τίτλο «είδος» (type) και αφορά στο είδος των δεδομένων. Η επιλογή «αριθμητική τιμή» (numeric) είναι προεπιλεγμένη από το PASW, καθώς τα δεδομένα τις περισσότερες φορές είναι αριθμητικά. Επιλέγεται η «σειρά χαρακτήρων» (string), όταν τα δεδομένα είναι σε μορφή χαρακτήρων ή απλώς είναι γράμματα. Η τρίτη στήλη έχει τίτλο «πλάτος» (width) και αφορά στο μέγιστο πλήθος των ψηφίων που μπορούν να λάβουν τα αριθμητικά δεδομένα, ενώ η τέταρτη στήλη έχει τίτλο «δεκαδικά ψηφία» (decimals) και αφορά στο μέγιστο πλήθος των δεκαδικών ψηφίων που μπορούν να λάβουν τα αριθμητικά δεδομένα. Η πέμπτη στήλη έχει τίτλο «ετικέτα» (label) και αφορά στην πλήρη ονομασία των μεταβλητών.

Η έκτη στήλη έχει τίτλο «τιμές» (values) και αφορά στην κωδικοποίηση των ποιοτικών μεταβλητών. Στην εικόνα 4, αναφορικά με τη μεταβλητή «φύλο», επιλέγοντας τις «τιμές», προκύπτει το μικρότερο «παράθυρο» στο οποίο γίνεται η κωδικοποίηση των κατηγοριών της ποιοτικής

Εικόνα 4. Η καρτέλα «απεικόνιση μεταβλητών» στο PASW, μέσω της οποίας πραγματοποιείται η διαχείριση των μεταβλητών της βάσης δεδομένων.

μεταβλητής. Στη συγκεκριμένη περίπτωση, η μεταβλητή «φύλο» μπορεί να λάβει δύο πιθανές τιμές (άνδρες και γυναίκες), οπότε επιλέγεται οι γυναίκες να συμβολίζονται με 0 και οι άνδρες με 1. Η κωδικοποίηση θα μπορούσε να γίνει και με διαφορετικό τρόπο, π.χ. συμβολίζοντας τις γυναίκες με 1 και τους άνδρες με 0, χωρίς φυσικά να επηρεαστεί η ανάλυση των δεδομένων. Αναφορικά με τη μεταβλητή «εκπαιδευτικό επίπεδο», υπάρχουν 3 κατηγορίες (κατώτερο, μεσαίο και ανώτερο επίπεδο), οπότε επιλέγεται τα άτομα με κατώτερο επίπεδο να συμβολίζονται με 1, τα άτομα με μεσαίο επίπεδο να συμβολίζονται με 2 και τα άτομα με ανώτερο επίπεδο να συμβολίζονται με 3. Η μεταβλητή «εκπαιδευτικό επίπεδο» είναι διατάξιμη, με την αύξηση των τιμών να δηλώνει υψηλότερο εκπαιδευτικό επίπεδο. Η μεταβλητή «μηνιαίο εισόδημα» εξάλλου είναι ποσοτική μεταβλητή, αλλά έχει μετατραπεί στη συγκεκριμένη βάση δεδομένων σε διατάξιμη μεταβλητή με 5 κατηγορίες, οπότε επιλέγεται τα άτομα με μηνιαίο εισόδημα < 1.000 € να συμβολίζονται με 1, τα άτομα με μηνιαίο εισόδημα 1.000–2.000 € να συμβολίζονται με 2, τα άτομα με μηνιαίο εισόδημα 2.000–3.000 € να συμβολίζονται με 3, τα άτομα με μηνιαίο εισόδημα 3.000–4.000 € να συμβολίζονται με 4 και τα άτομα με μηνιαίο εισόδημα > 4.000 € να συμβολίζονται με 5.

Η έβδομη στήλη έχει τίτλο «απούσες τιμές» (missing values) και αφορά στην περίπτωση κατά την οποία απουσιάζουν τα δεδομένα για ορισμένες μεταβλητές, όπως π.χ. άτομα που δεν έχουν δηλώσει την ηλικία τους. Η όγδοη στήλη έχει τίτλο «στήλες» (columns) και αφορά στο μέγεθος των στηλών ή, αλλιώς, στον αριθμό των χαρακτήρων που εμφανίζονται στις στήλες στην καρτέλα «απεικόνιση των δεδομένων», ενώ η ένατη στήλη έχει τίτλο «στοίχιση» (align) και αφορά στη στοίχιση (αριστερή, κεντρική ή δεξιά) των δεδομένων στις στήλες. Η δέκατη στήλη έχει τίτλο «μέτρηση» (measure) και αφορά στο είδος των μεταβλητών και κατ' επέκταση και των δεδομένων, με την επιλογή «κλίμακα» (scale) να δηλώνει τις ποσοτικές μεταβλητές (κωδικός περίπτωσης, ηλικία, ημέρες νοσηλείας και αριθμός παιδιών), την επιλογή «ονομαστικός» (nominal) να δηλώνει τις ονομαστικές μεταβλητές (φύλο, τόπος κατοικίας και οικογενειακή κατάσταση) και την επιλογή «διατάξιμος» (ordinal) να δηλώνει τις διατάξιμες μεταβλητές (εκπαιδευτικό επίπεδο και μηνιαίο εισόδημα). Η ενδέκατη στήλη έχει τίτλο «ρόλος» (role) και αφορά

στην απόδοση συγκεκριμένου ρόλου (π.χ. προσδιοριστής, έκβαση κ.ά.) σε μια μεταβλητή.

Στην εικόνα 3, στην καρτέλα «απεικόνιση δεδομένων», οι διαφορετικές γραμμές αντιπροσωπεύουν τα δεδομένα διαφορετικών περιπτώσεων, ενώ οι διαφορετικές στήλες αντιπροσωπεύουν διαφορετικές μεταβλητές. Για παράδειγμα, η πρώτη στήλη αφορά στον κωδικό των περιπτώσεων που πρόκειται να συμπεριληφθούν στην ανάλυση, η δεύτερη στήλη αφορά στην ηλικία, η τρίτη στήλη αφορά στο φύλο κ.λπ. Η πρώτη γραμμή εξάλλου αφορά στην πρώτη περίπτωση, η δεύτερη γραμμή στη δεύτερη περίπτωση, η τρίτη γραμμή στην τρίτη περίπτωση κ.ο.κ.

Όλα τα δεδομένα που εισάγονται είναι αριθμητικά. Τα δεδομένα που αφορούν σε μια συγκεκριμένη περίπτωση καταλαμβάνουν μόνο μία γραμμή (row), ενώ κάθε στήλη (column) περιέχει δεδομένα της ίδιας μεταβλητής για όλες τις περιπτώσεις. Έτσι, στην εικόνα 3, η συγκεκριμένη βάση δεδομένων περιλαμβάνει 20 περιπτώσεις (20 γραμμές) και δεδομένα αναφορικά με 9 μεταβλητές (9 στήλες). Σημειώνεται, ότι όταν η ίδια μεταβλητή μετριέται περισσότερες από μία φορές για κάθε περίπτωση (π.χ. μέτρηση βάρους πριν και μετά από την εφαρμογή μιας διαιτητικής παρέμβασης), τότε πρέπει να εισάγονται διαφορετικές στήλες με διαφορετικά ονόματα.

Απούσες τιμές

Απούσες τιμές (missing values) καλούνται οι τιμές εκείνες, στις οποίες απουσιάζουν οι παρατηρήσεις για τις διάφορες μεταβλητές.[15–18] Όσο αυξάνεται το ποσοστό των απουσών τιμών σε μια ανάλυση τόσο μειώνεται η αξιοπιστία των αποτελεσμάτων της ανάλυσης. Ιδιαίτερη σημασία έχει η εύρεση των αιτιών που οδηγούν στις απούσες τιμές, καθώς και η κατανομή των απουσών τιμών μεταξύ των διαφόρων μεταβλητών. Εάν, π.χ., σε μια βάση δεδομένων απουσιάζει > 30% των δεδομένων αναφορικά με μια συγκεκριμένη μεταβλητή, φαίνεται περισσότερο λογικό η εν λόγω μεταβλητή να μη συμπεριληφθεί στην ανάλυση των δεδομένων. Επιπλέον, εάν σε ορισμένες περιπτώσεις απουσιάζει > 30% των παρατηρήσεων αναφορικά με το σύνολο των μεταβλητών, φαίνεται ότι είναι αποδοτικότερο οι περιπτώσεις αυτές να μη συμπεριληφθούν στην ανάλυση των δεδομένων. Γενικά, δεν υπάρχουν συγκεκριμένοι κανόνες σχετικά με το ποσοστό των απουσών τιμών που είναι ανεκτό σε μια ανάλυση δεδομένων, με αρκετούς ερευνητές πάντως να θεωρούν ότι οι απούσες τιμές δεν πρέπει να υπερβαίνουν το 10% του συνόλου των δεδομένων. Σε μελέτες με μεγάλο αριθμό συμμετεχόντων, οι απούσες τιμές επηρεάζουν λιγότερο την αξιοπιστία της ανάλυσης σε σχέση με μελέτες με μικρότερο αριθμό συμμετεχόντων. Χαρακτηριστικό παράδειγμα αποτελούν οι κλινικές δοκιμές με μικρό αριθμό συμμετεχόντων, στις οποίες η ύπαρξη απουσών τιμών ακόμη και σε μικρό ποσοστό μπορεί να οδηγήσει σε απώλεια σημαντικού ποσοστού περιπτώσεων, μειώνοντας σε μεγάλο βαθμό την αξιοπιστία των αποτελεσμάτων της ανάλυσης.

Εάν οι απούσες τιμές κατανέμονται τυχαία και το ποσοστό τους είναι σχετικά μικρό, τότε οι περιπτώσεις που αφορούν στις απούσες τιμές δε συμπεριλαμβάνονται στην ανάλυση αναφορικά με συγκεκριμένες μεταβλητές. Για παράδειγμα, εάν σε μια περίπτωση είναι γνωστή η ηλικία, αλλά απουσιάζει το εκπαιδευτικό επίπεδο, τότε κάθε φορά που η διερεύνηση σχέσεων θα αφορά και στην ηλικία, θα συμπεριλαμβάνεται στην ανάλυση και η συγκεκριμένη περίπτωση, ενώ κάθε φορά που η διερεύνηση σχέσεων θα αφορά και στο εκπαιδευτικό επίπεδο, δε θα συμπεριλαμβάνεται στην ανάλυση και η συγκεκριμένη περίπτωση. Σημειώνεται, ότι ο εν λόγω τρόπος αντιμετώπισης των απουσών τιμών, που αποτελεί μάλιστα και προεπιλογή στα περισσότερα στατιστικά προγράμματα ανάλυσης δεδομένων (ανάμεσά τους και το PASW), οδηγεί σε μείωση της στατιστικής ισχύος. Το πρόβλημα γίνεται ακόμη μεγαλύτερο όταν οι απούσες τιμές δεν κατανέμονται τυχαία, καθώς στην περίπτωση αυτή, εκτός από τη μείωση της στατιστικής ισχύος, περιορίζεται σημαντικά και η δυνατότητα γενίκευσης των συμπερασμάτων. Για παράδειγμα, εάν σε μια μελέτη τα άτομα με υψηλότερο εισόδημα δε δηλώνουν σε υψηλό ποσοστό το εισόδημά τους, τότε τα αποτελέσματα της μελέτης αναφορικά με το εισόδημα δεν μπορούν να γενικευτούν και στην κατηγορία των ατόμων με υψηλό εισόδημα. Σημειώνεται, ότι σε ορισμένες περιπτώσεις, επιλέγεται η αντικατάσταση μιας απούσας τιμής με μια εκτιμώμενη τιμή, επιτυγχάνοντας με τον τρόπο αυτόν τη διατήρηση της στατιστικής ισχύος της μελέτης. Στην περίπτωση των ποσοτικών μεταβλητών που ακολουθούν την κανονική κατανομή, η εκτιμώμενη τιμή που αντικαθιστά μια απούσα τιμή είναι ο μέσος της μεταβλητής, ενώ στην περίπτωση ποσοτικών μεταβλητών που δεν ακολουθούν την κανονική κατανομή, η εκτιμώμενη τιμή είναι η διάμεσος της μεταβλητής.

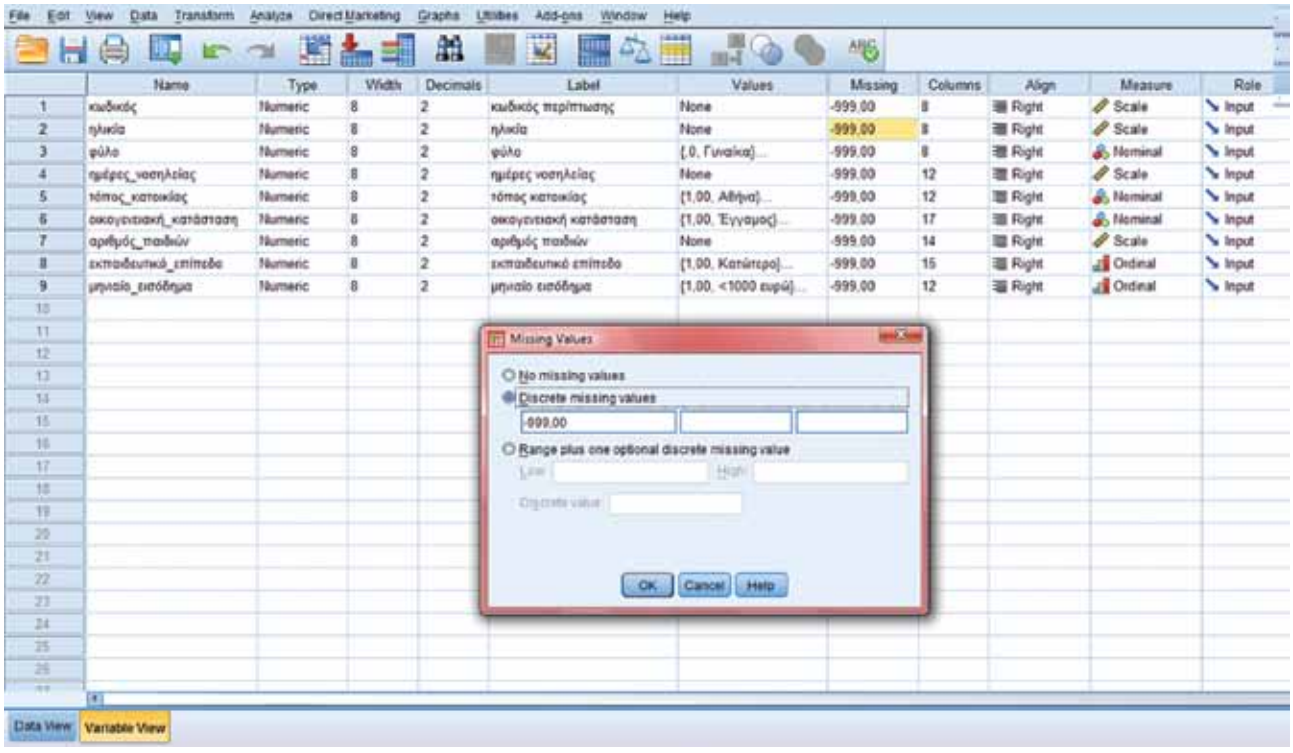
Η κωδικοποίηση των απουσών τιμών στα στατιστικά προγράμματα ανάλυσης δεδομένων πρέπει να γίνεται με σαφήνεια και ακρίβεια, έτσι ώστε να μη δημιουργείται σύγχυση στην ανάλυση των δεδομένων. Το PASW αναγνωρίζει ως απούσες τιμές τις παρατηρήσεις που σημειώνονται με «τελεία», με την προϋπόθεση ότι στην καρτέλα «απεικόνιση μεταβλητών» στη στήλη «είδος» (type) έχει επιλεγεί η «αριθμητική τιμή» (numeric) και

όχι η «σειρά χαρακτήρων» (string). Έτσι, μη συμπληρώνοντας ουσιαστικά ένα κελί στο PASW, η παρατήρηση στο κελί αυτό αναγνωρίζεται ως απούσα τιμή. Επιπλέον, το PASW έχει τη δυνατότητα αναγνώρισης συγκεκριμένων τιμών, που ορίζονται από τους ερευνητές, ως απούσες τιμές. Πιο συγκεκριμένα, στην εικόνα 5, στην καρτέλα «απεικόνιση μεταβλητών», επιλέγοντας το κελί που αντιστοιχεί στη στήλη «απούσες τιμές» (missing) αναφορικά με τη μεταβλητή «ηλικία» προκύπτει το μικρότερο παράθυρο ορισμού συγκεκριμένων τιμών που το PASW θα αναγνωρίζει πλέον ως απούσες τιμές. Στη συγκεκριμένη περίπτωση, επιλέχθηκε η τιμή «-999» να αναγνωρίζεται από το PASW ως απούσα τιμή, κάθε φορά που εισάγεται στην καρτέλα «απεικόνιση δεδομένων», στη στήλη της ηλικίας.

Απομακρυσμένες παρατηρήσεις

Οι *απομακρυσμένες παρατηρήσεις* (outliers) αφορούν σε παρατηρήσεις, οι τιμές των οποίων διαφέρουν σημαντικά από τις τιμές των υπολοίπων παρατηρήσεων.[15,16] Οι απομακρυσμένες παρατηρήσεις μπορεί να οφείλονται σε λανθασμένη καταγραφή των παρατηρήσεων κατά τη συλλογή των δεδομένων ή σε λανθασμένη εισαγωγή των παρατηρήσεων στη βάση δεδομένων ή να αποτελούν πραγματικές τιμές που απλά διαφέρουν σημαντικά από τις τιμές των υπολοίπων παρατηρήσεων.

Δεν υπάρχουν αυστηροί κανόνες αναφορικά με το πώς πρέπει να αντιμετωπίζονται οι απομακρυσμένες τιμές, αλλά αποτελεί γενική παραδοχή το γεγονός ότι οι τιμές αυτές πρέπει να αναγνωρίζονται και να εφαρμόζονται οι κατάλληλες στατιστικές μέθοδοι. Η αναγνώριση των απομακρυσμένων τιμών πραγματοποιείται με διάφορες στατιστικές μεθόδους, όπως οι τιμές επιρροής (leverage values), οι «αποστάσεις Cook» (Cook's distances), οι «αποστάσεις Mahalanobis» (Mahalanobis's distances κ.ά.). Επιπλέον, στην περίπτωση των ποσοτικών μεταβλητών που ακολουθούν την κανονική κατανομή, το 99% των παρατηρήσεων βρίσκεται εντός ±3 τυπικών αποκλίσεων από το μέσο, με τις παρατηρήσεις που δε βρίσκονται στο εν λόγω εύρος να θεωρούνται ως απομακρυσμένες παρατηρήσεις. Για παράδειγμα, εάν το μέσο βάρος ενός «δείγματος» ενηλίκων είναι 80 kg και η τυπική απόκλιση είναι 5, τότε το 99% των παρατηρήσεων του «δείγματος» θα πρέπει να βρίσκεται μεταξύ 65–95 kg, οπότε παρατηρήσεις με βάρος < 65 kg ή > 95 kg θεωρούνται ως απομακρυσμένες παρατηρήσεις.



Εικόνα 5. Στο PASW, στην καρτέλα «απεικόνιση μεταβλητών», επιλέγεται η τιμή «-999» να αναγνωρίζεται ως απούσα τιμή, κάθε φορά που εισάγεται στην καρτέλα «απεικόνιση δεδομένων», στη στήλη της ηλικίας.

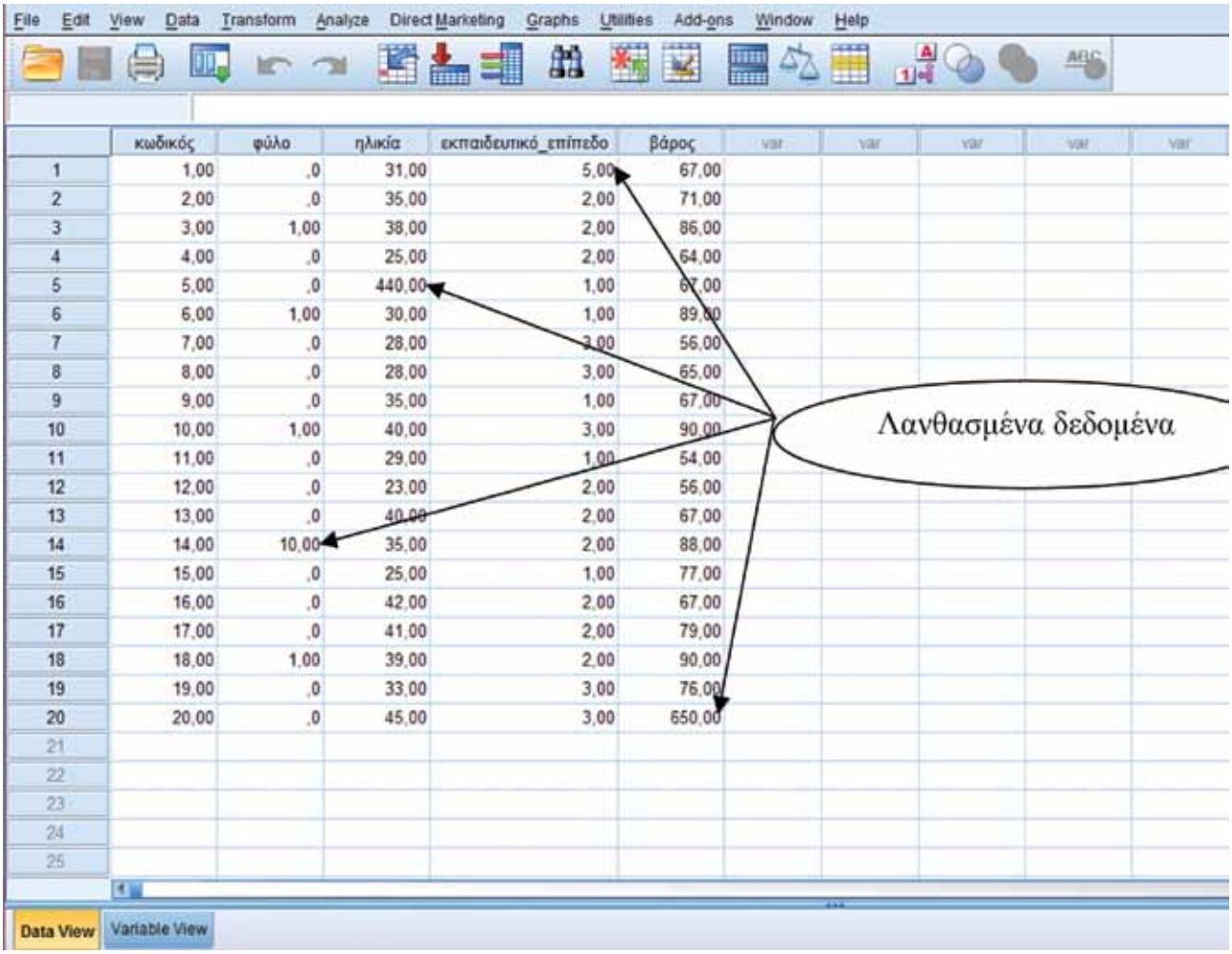
«Καθαρισμός» δεδομένων

Όπως προαναφέρθηκε, πριν από την έναρξη της ανάλυσης των δεδομένων είναι αναγκαίος ο «καθαρισμός» των δεδομένων, έτσι ώστε να επισημανθούν και να διορθωθούν τυχόν λάθη και παραλείψεις.[3,16,18,19] Ο πλέον αποτελεσματικός τρόπος για τον «καθαρισμό» των δεδομένων είναι η δημιουργία των πινάκων συχνοτήτων για κάθε μεταβλητή ξεχωριστά.[6,7,16] Στην εικόνα 6, φαίνεται μια βάση δεδομένων του PASW με 5 μεταβλητές (κωδικός περίπτωσης, φύλο, ηλικία, εκπαιδευτικό επίπεδο και βάρος) και 20 περιπτώσεις ατόμων. Οι πίνακες 3, 4, 5 και 6 είναι οι πίνακες απόλυτων και σχετικών συχνοτήτων που προκύπτουν στο PASW για το φύλο, την ηλικία, το εκπαιδευτικό επίπεδο και το βάρος, αντίστοιχα.

Στη βάση δεδομένων της εικόνας 6, η μεταβλητή «φύλο» έχει κωδικοποιηθεί ως «0» για τις γυναίκες και ως «1» για τους άνδρες. Στον πίνακα 3, φαίνεται ότι μία από τις 20 παρατηρήσεις δεν είναι ούτε γυναίκα ούτε άνδρας, καθώς έχει λάβει την τιμή «10». Είναι σαφές ότι, αναφορικά με τη μεταβλητή «φύλο», η παρατήρηση που έχει εισαχθεί στη βάση δεδομένων του PASW ως «10» αποτελεί λανθασμένη παρατήρηση, που μπορεί να οφείλεται είτε στη λανθασμένη καταγραφή των παρατηρήσεων κατά τη

Πίνακας 3. Απόλυτες και σχετικές συχνότητες για τη μεταβλητή «φύλο», που προκύπτουν από την ανάλυση της βάσης δεδομένων της εικόνας 6.

Φύλο	Απόλυτη συχνότητα (n)	Σχετική συχνότητα (%)
Γυναίκες	14	75
Άνδρες	5	20
10	1	5
Σύνολο	20	100



Εικόνα 6. Λανθασμένα δεδομένα στην καρτέλα «απεικόνιση δεδομένων», στο PASW.

Πίνακας 4. Απόλυτες και σχετικές συχνότητες για τη μεταβλητή «ηλικία», που προκύπτουν από την ανάλυση της βάσης δεδομένων της εικόνας 6.

Ηλικία (έτη)	Απόλυτη συχνότητα (n)	Σχετική συχνότητα (%)
23	1	5
25	2	10
28	2	10
29	1	5
30	1	5
31	1	5
33	1	5
35	3	15
38	1	5
39	1	5
40	2	10
41	1	5
42	1	5
45	1	5
440	1	5
Σύνολο	20	100

Πίνακας 5. Απόλυτες και σχετικές συχνότητες για τη μεταβλητή «εκπαιδευτικό επίπεδο», που προκύπτουν από την ανάλυση της βάσης δεδομένων της εικόνας 6.

Εκπαιδευτικό επίπεδο	Απόλυτη συχνότητα (n)	Σχετική συχνότητα (%)
Κατώτερο	5	25
Μεσαίο	9	45
Ανώτερο	5	25
5	1	5
Σύνολο	20	100

συλλογή των δεδομένων είτε στη λανθασμένη εισαγωγή των παρατηρήσεων στη βάση δεδομένων. Παρατηρώντας τη βάση δεδομένων της εικόνας 6, διαπιστώνεται ότι η περίπτωση «14» έχει κωδικοποιηθεί ως «10» αναφορικά με τη μεταβλητή «φύλο». Επομένως, πρέπει να αναζητηθεί το αρχείο της περίπτωσης «14» και να ελεγχθεί η παρατήρηση αναφορικά με τη μεταβλητή «φύλο», έτσι ώστε να γίνει και η απαραίτητη διόρθωση στη βάση δεδομένων.

Επιπλέον, στον πίνακα 4, φαίνεται ότι μία από τις 20 παρατηρήσεις της βάσης δεδομένων της εικόνας 6, αναφορικά με τη μεταβλητή «ηλικία», έχει ηλικία ίση με 440 έτη, που είναι σαφές ότι πρόκειται για λανθασμένη παρατήρηση. Παρατηρώντας τη βάση δεδομένων της εικόνας 6, διαπιστώνεται ότι στην περίπτωση «5», η ηλικία έχει δηλωθεί ίση με 440 έτη. Επομένως, πρέπει να αναζητηθεί το αρχείο της περίπτωσης «5» και να ελεγχθεί η παρατήρηση αναφορικά με τη μεταβλητή «ηλικία», έτσι ώστε να γίνει και η απαραίτητη διόρθωση στη βάση δεδομένων.

Η μεταβλητή «εκπαιδευτικό επίπεδο» έχει κωδικοποιηθεί ως «1» για το κατώτερο εκπαιδευτικό επίπεδο, ως «2» για το μεσαίο εκπαιδευτικό επίπεδο και ως «3» για το ανώτερο εκπαιδευτικό επίπεδο. Στον πίνακα 5, φαίνεται ότι μία από τις 20 παρατηρήσεις δεν ανήκει σε ένα από τα τρία εκπαιδευτικά επίπεδα, καθώς έχει λάβει την τιμή «5». Είναι σαφές ότι, αναφορικά με τη μεταβλητή «εκπαιδευτικό επίπεδο», η παρατήρηση που έχει εισαχθεί στη βάση δεδομένων του PASW ως «5» αποτελεί λανθασμένη παρατήρηση, που μπορεί να οφείλεται είτε στη λανθασμένη καταγραφή των παρατηρήσεων κατά τη συλλογή των δεδομένων είτε στη λανθασμένη εισαγωγή των παρατηρήσεων στη βάση δεδομένων. Παρατηρώντας τη βάση δεδομένων της εικόνας 6, διαπιστώνεται ότι η περίπτωση «1» έχει κωδικοποιηθεί ως «5» αναφορικά με τη μεταβλητή «εκπαιδευτικό επίπεδο». Επομένως, πρέπει να αναζητηθεί το αρχείο της περίπτωσης «1» και να ελεγχθεί η παρατήρηση αναφορικά με τη μεταβλητή «εκπαιδευτικό επίπεδο», έτσι ώστε να γίνει και η απαραίτητη διόρθωση στη βάση δεδομένων.

Τέλος, στον πίνακα 6, φαίνεται ότι μία από τις 20 παρατηρήσεις της βάσης δεδομένων της εικόνας 6, αναφορικά με τη μεταβλητή «βάρος», έχει βάρος ίσο με 650 kg, που είναι σαφές ότι πρόκειται για λανθασμένη παρατήρηση. Παρατηρώντας τη βάση δεδομένων της εικόνας 6, διαπιστώνεται ότι στην περίπτωση «20» το βάρος έχει δηλωθεί

Πίνακας 6. Απόλυτες και σχετικές συχνότητες για τη μεταβλητή «βάρος», που προκύπτουν από την ανάλυση της βάσης δεδομένων της εικόνας 6.

Βάρος (kg)	Απόλυτη συχνότητα (n)	Σχετική συχνότητα (%)
54	1	5
56	2	10
64	1	5
65	1	5
67	5	25
71	1	5
76	1	5
77	1	5
79	1	5
86	1	5
88	1	5
89	1	5
90	2	10
650	1	5
Σύνολο	20	100

ίσο με 650 έτη. Επομένως, πρέπει να αναζητηθεί το αρχείο της περίπτωσης «20» και να ελεγχθεί η παρατήρηση αναφορικά με τη μεταβλητή «βάρος», έτσι ώστε να γίνει και η απαραίτητη διόρθωση στη βάση δεδομένων.

Σύνοψη

Οι επιστήμονες υγείας καλούνται στις επιδημιολογικές μελέτες να συλλέξουν τα δεδομένα με τέτοιον τρόπο ώστε από τη στατιστική ανάλυση αυτών να προκύπτουν ακριβή και έγκυρα αποτελέσματα. Η ανάλυση που πραγματοποιείται με τα διάφορα στατιστικά προγράμματα οδηγεί σε έγκυρα αποτελέσματα μόνο εφόσον τα δεδομένα στα οποία στηρίζεται είναι έγκυρα. Δυστυχώς, όμως, σε αρκετές περιπτώσεις, η άγνοια της διαχείρισης των ηλεκτρονικών βάσεων δεδομένων και η έλλειψη προσοχής κατά την εισαγωγή των δεδομένων οδηγούν στη λανθασμένη καταχώρηση αυτών στη βάση δεδομένων. Για το λόγο αυτόν, πριν αρχίσει η ανάλυση των δεδομένων, είναι αναγκαία τόσο η σωστή εισαγωγή των στοιχείων στη βάση δεδομένων που πρόκειται να χρησιμοποιηθεί για την ανάλυση όσο και η σωστή διαχείριση και η κωδικοποίηση των μεταβλητών, έτσι ώστε να επισημανθούν και να διορθωθούν τυχόν λάθη και παραλείψεις.

ABSTRACT

The Management of Data and Variates in Epidemiological Studies

Petros Galanis

MSc in Public Health, PhD, RN, Centre for Health Services Management and Evaluation, Faculty of Nursing, National and Kapodistrian University of Athens, Athens, Greece

In epidemiological studies, health scientists should collect data that are as reliable and objective as possible, in order for the statistical analysis to lead to precise and valid conclusions. Mistakes in the way that data concerning the cases in a study are collected, and in the input of the data into the appropriate electronic databases, along with limited knowledge of database management, may all result in a considerable decrease in the reliability of data analysis, leading to invalid results. Before the start of statistical analysis, the correct input of the data into database that is going to be used for the analysis is essential. This involves the correct management and coding of the variates, in order to ascertain and rectify possible errors and omissions. The variates (and by extension the data as a whole) are divided, according to their mathematical properties, into qualitative and quantitative; the former are divided into nominal and ordinal, and the latter into interval scale variates and ratio scale variates. Quantitative variates are also divided into continuous and discrete. Missing values and outliers have a particular meaning in data analysis. Observations that have not been made on some variates are called missing values. An increase of the proportion of missing values in the analysis leads to decrease in credibility of the results of the analysis. Outliers are values of observations that are surprisingly extreme when compared to the values of the other observations. Outliers may be due to either errors in observation recorded during data collection or incorrect input of the observations into the database, or they may be actual values that just differ considerably from those of the other observations. ***NOSILEFTIKI* 2010, 50 (2): 132–146.**

Key-words: *data analysis, databases, missing values, outliers, qualitative variates, quantitative variates*

✉ **Corresponding Author:** Petros Galanis, 14 Dikis street, GR-157 73 Athens, Greece, tel.: +30 210 77 81 044, +30 694 43 87 354, e-mail: pegalan@nurs.uoa.gr

Βιβλιογραφία

1. Dodge Y. *The concise encyclopedia of statistics.* Springer Science & Business Media, Berlin, 2008:518–520
2. Γαλάνης ΠΑ, Σπάρος ΛΔ. *Εγχειρίδιο επιδημιολογίας*. Εκδόσεις ΒΗΤΑ, Αθήνα, 2010:27–29
3. Peat J, Barton B. *Medical statistics. A guide to data analysis and critical appraisal.* BMJ Books, Massachusetts, 2005:1–23
4. Bowers D. *Medical statistics from scratch. An introduction for health professionals.* 2nd ed. John Wiley & Sons, New Jersey, 2008:1–9
5. Bowers D, House A, Owens D. *Understanding clinical papers.* 2nd ed. John Wiley & Sons, New Jersey, 2006:63–67
6. Stewart A. *Basic statistics and epidemiology: A practical guide.* Radcliffe Medical Press, Oxford, 2002:11–22
7. Chernick M, Friis R. *Introductory biostatistics for the health sciences.* John Wiley & Sons, New Jersey, 2003:46–67
8. Shasha D, Wilson M. *Statistics is easy.* Morgan & Claypool Publishers. Washington, 2008:1–11
9. Rugg G. *Using statistics: A gentle introduction.* 3rd ed. Open University Press, Berkshire, 2007:1–20

10. Brase CH, Brase CP. *Understanding basic statistics*. 4th ed. Houghton Mifflin Co, Boston, 2007:2–31
11. Γέμτος Π. *Μεθοδολογία των κοινωνικών επιστημών. Μεταθεωρία και ιδεολογική κριτική των επιστημών του ανθρώπου*. 3η έκδοση, 2ος τόμος. Εκδόσεις Παπαζήση, Αθήνα, 1987:180–197
12. Γναρδέλλης Χ. *Ανάλυση δεδομένων με το PASW Statistics 17.0*. Εκδόσεις Παπαζήση, Αθήνα, 2009
13. Γναρδέλλης Χ. *Ανάλυση δεδομένων με το SPSS 14.0 for Windows*. Εκδόσεις Παπαζήση, Αθήνα, 2006
14. Field A. *Discovering statistics using SPSS*. 2nd ed. SAGE Publications, London, 2005
15. Leech N, Barrett K, Morgan G. *SPSS for intermediate statistics: Use and interpretation*. 2nd ed. Lawrence Erlbaum Associates, London, 2005
16. Boslaugh S, Watters P. *Statistics in a nutshell*. O'Reilly Media Inc., Cambridge, 2008:41–53
17. Molenberghs G, Beunckens C, Jansen I, Thijs H, Verbeke G, Kenward M. Missing data. In: Ahrens W, Pigeot I (eds) *Handbook of epidemiology: Missing data*. Springer Science & Business Media, Berlin, 2006:767–828
18. Rothman KJ, Greenland S, Lash TL. *Modern epidemiology*. 3rd ed. Lippincott Williams & Wilkins, Philadelphia, 2008:214–238
19. Myatt GJ. *Making sense of data. A practical guide to exploratory data analysis and data mining*. John Wiley & Sons, New Jersey, 2007:17–53