

Δεδομένα και στατιστική

Σκοπός των περισσότερων ερευνών είναι η συλλογή **δεδομένων**, για τη λήψη πληροφοριών ενός ορισμένου ερευνητικού τομέα. Τα δεδομένα μας περιλαμβάνουν **παρατηρήσεις** μίας ή περισσότερων μεταβλητών: κάθε ποσότητα που μεταβάλλεται ορίζεται ως **μεταβλητή**. Για παράδειγμα, μπορεί να συλλέξουμε βασικές κλινικές και δημογραφικές πληροφορίες για ασθενείς με μία ορισμένη ασθένεια. Οι μεταβλητές που εξετάζονται μπορεύνα περιλαμβάνουν το φύλο, την ηλικία και το ανάστημα των ασθενών.

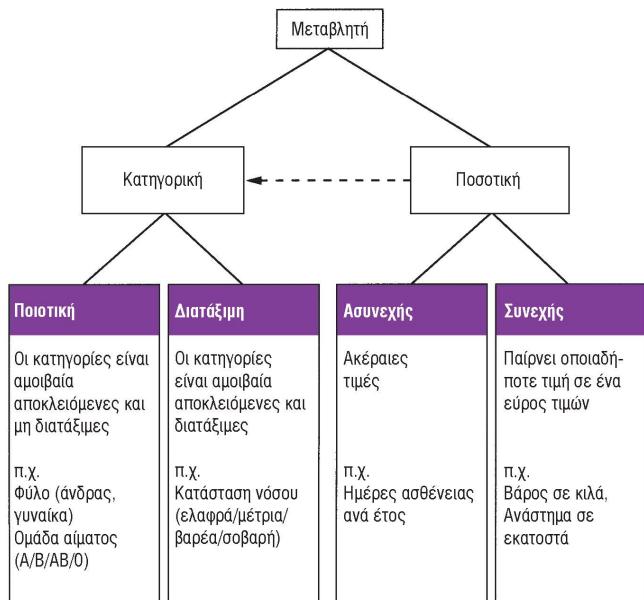
Τα δεδομένα μας συνήθως προέρχονται από ένα **δείγμα** ατόμων, το οποίο εκφράζει τον εξεταζόμενο **πληθυσμό**. Σκοπός μας είναι να συνοψίσουμε αυτά τα δεδομένα με έναν κατανοητό τρόπο και να εξαγουμένων μερικές πληροφορίες από αυτά. Η **στατιστική** περικλείει τις μεθόδους της συλλογής, της συνόψισης, της ανάλυσης και της εξαγωγής συμπερασμάτων από τα δεδομένα: εμείς χρησιμοποιούμε τις στατιστικές τεχνικές, για να πετύχουμε το σκοπό μας.

Τα δεδομένα μπορεύνα παίρνουν πολλές διαφορετικές μορφές. Προτού αποφασίσουμε ποιες θα είναι οι καταλληλότερες στατιστικές μέθοδοι που θα χρησιμοποιήσουμε, χρειάζεται να γνωρίζουμε ποια μορφή παίρνει η κάθε μεταβλητή. Κάθε μεταβλητή και τα προκύπτοντα δεδομένα μπορεύνανται να είναι από τα δύο είδη: **ποιοτικά** (categorical, qualitative) ή **ποσοτικά** (numerical) (Εικ. 1.1.).

Ποιοτικά δεδομένα ή κατηγορικά δεδομένα

Δεδομένα κατηγοριών είναι εκείνα για τα οποία κάθε άτομο ανήκει μόνο σε μία από τις πολλές διακεκριμένες κατηγορίες μίας μεταβλητής.

- **Ποιοτικά δεδομένα.** Οι κατηγορίες δεν διατάξιμες, απλώς έχουν ονόματα. Ως παράδειγμα αναφέρονται οι ομάδες αίματος (A, B, AB, O) και η οικογενειακή κατάσταση (παντρεμένος/χήρος/ανύπαντρος κ.λπ.). Στην τελευταία περίπτωση δεν συντρέχει λόγος να θεωρείται ότι το να είναι κανείς παντρεμένος είναι καλύτερο (ή χειρότερο) από το να είναι ανύπαντρος!
- **Διαβαθμιζόμενα (διατάξιμα) δεδομένα.** Οι κατηγορίες είναι κατά κάποιο



Εικόνα 1.1 Διάγραμμα στο οποίο παρουσιάζονται τα διαφορετικά είδη μίας μεταβλητής.

τρόπο διατάξιμες. Ως παραδείγματα αναφέρονται τα στάδια της νόσου (προχωρημένη, μέτρια, ελαφρά, ανύπαρκτη) και η ένταση του πόνου (οξύς, μέτριος, ελαφρύς, καθόλου).

Μια κατηγορική μεταβλητή είναι **δυαδική** (binary) ή **διχοτομική** (dichotomous), όταν υπάρχουν μόνο δύο πιθανές κατηγορίες. Ως παραδείγματα αναφέρονται «Ναι/Οχι», «Πεθαμένος/Ζωντανός» ή «Ασθενής/Μη ασθενής».

Ποσοτικά δεδομένα

Ποσοτικά (numerical ή qualitative) δεδομένα είναι εκείνα που παίρνουν κάποια αριθμητική τιμή. Μπορούμε να διακρίνουμε τα ποσοτικά δεδομένα σε δύο είδη:

- **Ασυνεχής (discrete) δεδομένα** – είναι εκείνα για τα οποία η μεταβλητή μπορεί να πάρει μόνο ορισμένες ακέραιες αριθμητικές τιμές. Αυτές οι τιμές συχνά μετρούν αριθμούς γεγονότων, όπως ο αριθμός των επισκέψεων σε ένα γενικό ιατρό κατά τη διάρκεια ενός έτους ή ο αριθμός των επεισοδίων της νόσου σε ένα άτομο τα τελευταία πέντε χρόνια.
- **Συνεχής (continuous) δεδομένα** – είναι εκείνα για τα οποία η μεταβλητή μπορεί να πάρει οποιαδήποτε αριθμητική τιμή π.χ. βάρος ή ανάστημα, εκτός από αυτήν που τα περιορίζει κατά τη διάρκεια της μέτρησης.

Διαχωρισμός μεταξύ των ειδών των δεδομένων

Συχνά χρησιμοποιούμε διαφορετικές στατιστικές μεθόδους και αυτό εξαρτάται από το αν τα δεδομένα είναι κατηγορικά ή ποσοτικά. Αν και η διάκριση ανάμεσα στα κατηγορικά και ποσοτικά δεδομένα είναι συνήθως σαφής, σε μερικές περιπτώσεις μπορεί να γίνει ασαφής. Για παράδειγμα, όταν έχουμε μία μεταβλητή με ένα μεγάλο αριθμό από διαβαθμιζόμενες κατηγορίες (π.χ. Μία κλίμακα πάνου με επτά κατηγορίες), μπορεί να είναι δύσκολο να τη διακρίνουμε από μία ασυνεχή ποσοτική μεταβλητή. Η διάκριση ανάμεσα στα ασυνεχή και συνεχή ποσοτικά δεδομένα μπορεί να είναι ακόμη και πο ασαφής, αν και γενικώς αυτό θα έχει μικρή επίδραση στα αποτελέσματα των περισσότερων αναλύσεων. Η ηλικία είναι ένα παράδειγμα μεταβλητής, η οποία συχνά θεωρείται ως ασυνεχής, αν και αυτή είναι πράγματι συνεχής. Συνήθως αναφερόμαστε στην «ηλικία κατά τα τελευταία γενέθλια», παρά στην «ηλικία» απλώς, και συνεπώς μία γυναίκα, η οποία λέει ότι είναι 30 ετών, ίσως να έχει ακριβώς τα τριακοστά γενέθλιά της ή ίσως να έχει περίπου τα τριακοστά-πρώτα γενέθλιά της.

Να μην παρασυρόμαστε εξαρχής να καταγράψουμε ποσοτικά δεδομένα ως κατηγορικά (π.χ. καταγράφοντας μόνο το εύρος μέσα στο οποίο εμπεριέχεται η ηλικία του ασθενούς, από το να αναφέρουμε την πραγματική του ηλικία), διότι συνήθως χάνεται σημαντική πληροφορία. Είναι απλό να μετατραπούν ποσοτικά δεδομένα σε κατηγορικά που έχουν συλλεγεί παλαιότερα.

Προκύπτοντα δεδομένα

Μπορεύνα που συναντήσουμε έναν αριθμό άλλων ειδών δεδομένων στον ιατρικό χώρο. Αυτά είναι:

- **Εκατοστιαία ποσοστά** (percentages) – Αυτά εμφανίζονται όταν λαμβάνονται υπ' όψιν οι βελτιώσεις της κατάστασης υγείας, σε ασθενείς που ακολουθούν μία θεραπεία π.χ. η αναπνευστική λειτουργία ενός ασθενούς (FEV1) μπορεί να αυξηθεί κατά 24%, ακολουθώντας μία θεραπεία με ένα νέο φάρμακο. Στην περίπτωση αυτή, ενδιαφέρει περισσότερο το επίπεδο της βελτίωσης παρά η απόλυτη τιμή. Δείκτης αντιστοιχίας (ο αριθμητής δεν αποτελεί τιμή του παρονομαστή)
- **Λόγοι (ratios) ή πηλίκα (quotients)** – Μερικές φορές μπορεύνα που συναντήσετε τους λόγους ή τα πηλίκα δύο μεταβλητών. Για παράδειγμα, το σωματικό δείκτη παχυσαρκίας (BMI), ο οποίος υπολογίζεται από το βάρος

(kg) ενός ατόμου διαιρεμένο με το τετράγωνο του αναστήματός του (m^2) και ο οποίος συχνά χρησιμοποιείται για να προσδιορίσουμε αν το άτομο είναι υπέρβαρο ή όχι.

• **Αναλογίες (Rates)** – Δείκτες, στους οποίους ο αριθμός των περιπτώσεων ασθενειών που εμφανίζονται μεταξύ των ατόμων σε μία έρευνα διαιρείται με τη συνολική χρονική περίοδο της παρακολούθησης όλων των ατόμων σε αυτή την έρευνα, και οι οποίοι είναι συνήθεις στις επιδημιολογικές έρευνες (Κεφάλαιο 12).

• **Κλίμακες (Scores)** – μερικές φορές χρησιμοποιούμε μία αυθαίρετη τιμή, π.χ. Μία βαθμολογία (score), όταν δεν μπορούμε να μετρήσουμε ένα μέγεθος. Για παράδειγμα, μία σειρά απαντήσεων σε ερωτήσεις για την ποιότητα ζωής μπορεί να αθροιστούν και να δώσουν μία συνολική βαθμολογία της ποιότητας ζωής του κάθε ατόμου.

Όλες αυτές οι μεταβλητές μπορεί να θεωρηθούν ως ποσοτικές μεταβλητές για τις περισσότερες αναλύσεις. Όπου η μεταβλητή προκύπτει από περισσότερες της μίας τιμές (π.χ. ο αριθμητής και ο παρονομαστής ενός εκατοστιαίου ποσοστού), είναι σημαντικό να καταγραφούν όλες οι τιμές που χρησιμοποιήθηκαν. Για παράδειγμα, μία βελτίωση κατά 10% ενός δείκτη μετά από θεραπεία μπορεί να έχει διαφορετική σημασία, επειδή εξαρτάται από το επίπεδο του δείκτη πριν τη θεραπεία.

Αποκομμένα δεδομένα

Μπορεί να συναντήσουμε **αποκομμένα δεδομένα** (censored) στις περιπτώσεις που φαίνονται στα παρακάτω παραδείγματα:

• Αν μετρούμε εργαστηριακές τιμές χρησιμοποιώντας ένα όργανο, το οποίο μπορεί να ανιχνεύει επίπεδα τιμών πάνω από μία ορισμένη τιμή (cut-off value), τότε κάθε τιμή κάτω από αυτήν την ορισμένη τιμή δεν θα ανιχνευθεί, δηλ. είναι αποκομμένη. Για παράδειγμα, όταν μετρούμε επίπεδα μικροβίων, εκείνα τα οποία είναι κάτω από το όριο της ανιχνευσιμότητας, θα μπορούσαν να αναφέρονται ως «μη ανιχνεύσιμα», ακόμη και αν δυνητικά υπάρχουν μικρότια στο δείγμα. Στην περίπτωση αυτή, αν η κατώτερη ορισμένη τιμή ενός οργάνου είναι π.χ. τα αποτελέσματα μπορεί να αναφέρονται ως <x. Ομοίως, μερικά όργανα μπορεί αξιόπιστα να προσδιορίσουν επίπεδα τιμών κάτω από μία ορισμένη κατώτερη τιμή, π.χ. y, οποιεσδήποτε τιμές πάνω από εκείνη την τιμή, θα είναι επίσης αποκομμένη, και το αποτέλεσμα της δοκιμής μπορεί να αναφέρεται ως >y.

• Μπορεί να συναντήσουμε αποκομμένα δεδομένα, όταν για παράδειγμα παρακολουθώντας ασθενείς σε ένα κλινικό πείραμα, μερικοί ασθενείς «φεύγουν» από το πείραμα πριν αυτό τελειώσει. Αυτό το είδος των δεδομένων συζητείται λεπτομερέστερα στο Κεφάλαιο 44.

Όταν διεξάγετε μία έρευνα, σχεδόν πάντοτε απαιτείται η είσοδος των δεδομένων σε έναν ηλεκτρονικό υπολογιστή. Οι ηλεκτρονικοί υπολογιστές προσφέρουν ανεκτίμητες υπηρεσίες στη βελτίωση της ακρίβειας και της ταχύτητας της συλλογής και της ανάλυσης δεδομένων, κάνοντας εύκολο τον έλεγχο λαθών, παρέχοντας περιλήψεις των δεδομένων και δημιουργώντας νέες μεταβλητές. Αδιέτοι οι ιδεές για την εργασία στην είσοδο δεδομένων.

Μορφοποιήσεις για είσοδο δεδομένων

Υπάρχουν πολλοί τρόποι με τους οποίους τα δεδομένα μπορούν να εισαχθούν και να αποθηκευτούν σε έναν ηλεκτρονικό υπολογιστή. Πολλά στατιστικά πακέτα σας επιτρέπουν να εισάγετε τα δεδομένα απευθείας. Εντούτοις, ο περιορισμός αυτής της προσέγγισης είναι ότι μερικές φορές δεν μπορείται να μεταφέρεται δεδομένα σε ένα άλλο πακέτο. Ένας απλός εναλλακτικός τρόπος είναι να αποθηκεύονται τα δεδομένα είτε σε ένα λογιστικό φύλλο ή σε ένα πακέτο βάσης δεδομένων. Δυστυχώς, σε αυτά οι στατιστικές διαδικασίες είναι συχνά περιορισμένες και είναι συνήθως απαραίτητο να εισάγονται τα δεδομένα σε ένα ειδικό στατιστικό πακέτο, ώστε να μπορεί να γίνονται και στατιστικές αναλύσεις.

Η πιο ευέλικτη προσέγγιση είναι να έχετε τα δεδομένα σας διαθέσιμα σε ένα ASCII αρχείο ή σε ένα αρχείο σε μορφή κειμένου. Εκείνα τα δεδομένα σε ASCII αρχείο μπορούν να διαβαστούν από τα περισσότερα πακέτα. Η διαμόρφωση ενός ASCII αρχείου απλώς περιέχει γραμμές κειμένου, τις οποίες μπορείτε να δείτε στην οθόνη του ηλεκτρονικού υπολογιστή. Συνήθως, κάθε μεταβλητή στο αρχείο έχωριζε από την επόμενη με έναν οριοθέτη (delimiter), που συχνά είναι ένα κενό ή το κόμμα. Αυτό είναι γνωστό ως ελεύθερη διαμόρφωση (free format).

Ο απλούστερος τρόπος εισαγωγής δεδομένων σε ASCII διαμόρφωση είναι να πληκτρολογούνται τα δεδομένα κατευθείαν σε αυτήν τη διαμόρφωση κειμένου ή σε ένα πρόγραμμα διόρθωσης. Χρησιμοποιώντας είτε τη μία είτε την άλλη προσέγγιση, είναι συνηθισμένο σε κάθε γραμμή των δεδομένων να αντιστοιχεί ένα διαφορετικό άτομο της έρευνας, και σε κάθε στήλη να αντιστοιχεί μία διαφορετική μεταβλητή, αν και συχνά είναι απαραίτητο να συνεχίζουμε σε επόμενες σειρές, αν για κάθε άτομο έχει συλλεγεί μεγάλος αριθμός μεταβλητών.

Σχεδιασμός εισόδου δεδομένων

Όταν συλλέγετε δεδομένα για μία έρευνα, συχνά θα χρειαστεί να χρησιμοποιήσετε μία φόρμα ή ένα ερωτηματολόγιο για την καταγραφή των δεδομένων. Αν αυτά είναι σχεδιασμένα προσεκτικά, μπορεί να μειωθεί το μέγεθος της εργασίας που απαιτείται για την είσοδο των δεδομένων. Γενικώς, αυτές οι φόρμες/ερωτηματολόγια περιέχουν μία σειρά από κουτάκια, στα οποία καταγράφονται τα δεδομένα – είναι συνηθισμένο να έχουμε ένα χωριστό κουτί για κάθε πιθανό ψηφίο της απάντησης.

Κατηγορικά δεδομένα

Μερικά στατιστικά πακέτα παρουσιάζουν προβλήματα στη διαχείριση των μη αριθμητικών δεδομένων. Συνεπώς, μπορεί να χρειαστείτε να ορίσετε αριθμητικούς κωδικούς για κατηγορικά δεδομένα, προτού εισάγετε τα δεδομένα στον ηλεκτρονικό υπολογιστή. Πα παράδειγμα, μπορεί να διαλέξετε να ορίσετε τους κωδικούς 1, 2, 3 και 4 για τις κατηγορίες «όχι πόνος», «ελαφρύς πόνος», «μέτριος πόνος» και «οξύς πόνος», αντίστοιχα. Αυτοί οι κωδικοί μπορεί να προστεθούν στις φόρμες, όταν συλλέγονται τα δεδομένα. Για δυαδικά (binary) δεδομένα, π.χ. ναι/όχι απαντήσεις, είναι συχνά κατάλληλο να ορίζονται οι κώδικες 1 (π.χ. για «ναι») και 0 (για «όχι»).

• **Μονο-κωδικοποιούμενες μεταβλητές** – υπάρχει μία μόνο δυνατή απάντηση σε μία ερώτηση, π.χ. «είναι ο ασθενής νεκρός;». Δεν είναι δυνατό σε αυτή την ερώτηση να απαντήσουμε και τα δύο (και ναι και όχι).

• **Πολυ-κωδικοποιούμενες μεταβλητές** – είναι δυνατές περισσότερες από μία απαντήσεις για κάθε ερώτηση. Για παράδειγμα, ποια συμπτώματα εμφανίσεις αυτός ο ασθενής; Στην περίπτωση αυτή, ένα άτομο μπορεί να έχει εμφανίσει έναν οποιοδήποτε αριθμό συμπτωμάτων. Υπάρχουν δύο τρόποι αντιμετώπισης αυτού του τύπου δεδομένων, οι οποίοι εξαρτώνται από τις παρακάτω περιπτώσεις:

- **Υπάρχουν μόνο μερικά ενδεχόμενα συμπτώματα και τα άτομα μπορεί να εμφανίσουν πολλά από αυτά.** Μπορεί να δημιουργηθεί ένας αριθμός διαφορετικών δυαδικών μεταβλητών, οι οποίες αντιστοιχούν στο αν ο ασθενής απάντησε ναι ή όχι στην παρουσία κάθε ενδεχόμενου συμπτώματος. Για παράδειγμα, «έχει ο ασθενής βήχα», «Έχει ο ασθενής πονόλαιμο»

- **Υπάρχει ένας πολύ μεγάλος αριθμός ενδεχόμενων συμπτωμάτων, αλλά κάθε ασθενής αναμένεται να υποφέρει μόνο από μερικά από αυτά.** Μπορεί να δημιουργηθεί ένας αριθμός διαφορετικών ποιοτικών μεταβλητών, που σας επιτρέπει να ονομάζετε ένα σύμπτωμα από το οποίο υποφέρει ο ασθενής. Για παράδειγμα, «ποιο ήταν το πρώτο σύμπτωμα από το οποίο υπέφερε ο ασθενής;»; «Ποιο ήταν το δεύτερο σύμπτωμα;» Θα χρειαστεί να αποφασίσετε εκ των προτέρων μεγαλύτερο αριθμό των συμπτωμάτων από τα οποία νομίζετε ότι ένας ασθενής ενδέχεται να υποφέρει.

Αριθμητικά δεδομένα

Τα αριθμητικά δεδομένα θα πρέπει να εισάγονται με την ίδια ακρίβεια με την οποία αυτά μετριούνται και η μονάδα μέτρησης θα πρέπει να είναι σταθερή για διάφορες τις παρατηρήσεις μίας μεταβλητής. Για παράδειγμα, το βάρος θα πρέπει να καταγραφεί σε χιλιόγραμμα ή σε λίβρες (pounds), αλλά όχι και στα δύο εναλλακτικά.

Πολλαπλές φόρμες για κάθε ασθενή

Μερικές φορές συλλέγονται πληροφορίες για τον ίδιο ασθενή σε περισσότερες από μία περιπτώσεις. Είναι σημαντικό ότι υπάρχει κάποια μοναδική καθοριστική παράμετρος (π.χ. αύξοντας αριθμός) που σχετίζεται με το άτομο και η οποία θα σας δώσει τη δυνατότητα να συνδέσετε όλα τα δεδομένα από ένα άτομο στην έρευνα.

Προβλήματα με τις ημερομηνίες και τους χρόνους

Οι ημερομηνίες και οι χρόνοι θα πρέπει να εισάγονται με ένα σταθερό τρόπο, π.χ. είτε σε ημέρα/μήνας/έτος ή μήνας/ημέρα/έτος, αλλά όχι εναλλακτικά. Είναι σημαντικό να βρεθεί ποια διαμόρφωση αρχείου μπορεί να διαβαστεί από το στατιστικό πακέτο.

Κωδικοποίηση ελλειπόντων τιμών

Θα πρέπει να λάβετε υπ' όψιν τι θα κάνετε με τις ελλειπόντωσης τιμές, προτού εισάγετε τα δεδομένα. Στις περισσότερες περιπτώσεις χρειάζεται να χρησιμοποιήσετε μερικά σύμβολα για να εκφράσετε μία ελλειπόντωση τιμής. Τα στατιστικά πακέτα αντιμετωπίζουν τις ελλειπόντωσης τιμές με διαφορετικούς τρόπους. Μερικά χρησιμοποιούν ειδικούς χαρακτήρες (π.χ. Μία τελεία ή ένα αστεράκι) για να δείξουν τις ελλειπόντωσης τιμές, ενώ άλλα απαιτούν εσείς να καθορίσετε τον δικό σας κωδικό για μία ελλειπόντωση τιμή (οι συνήθως χρησιμοποιούμενες τιμές είναι 9,999 ή -99). Η τιμή που θα επιλεγεί θα πρέπει να είναι τέτοια, που να μην είναι ενδεχόμενη για τη μεταβλητή αυτή. Για παράδειγμα, όταν εισάγεται μία κατηγορική μεταβλητή με τέσσερις κατηγορίες (κωδικοποιημένες ως 1, 2, 3 και 4), μπορεί να επιλέξετε την τιμή 9 για να παραστήσετε τις ελλειπόντωσης τιμές. Ενώ, αν η μεταβλητή είναι «ηλικία του παιδιού», τότε θα επιλεγεί ένας διαφορετικός κωδικός (π.χ. 99). Τα ελλείποντα δεδομένα θα συζητηθούν λεπτομερέστερα στο Κεφάλαιο 3.

Παράδειγμα

Ποιοτικές μεταβλητές – οχι διατάξιμες καπηγορίες		Ασυνεχής/Διακριτή										Συνεχής μεταβλητή		Ποιοτική Διατάξιμη		
		Αποτιμώνες παρεμβολές κατά τη διάρκεια της εγκυμοσύνης										ΗΜΕΡΟ-ΜΗΝΙΑ				
Αριθμός σε θενάρια	Αιμορραγική συντήρηση	Φύλο μαρού	Ηλικία κυήσης (εβδομάδες)	Εισπίεσ- μενος αέρας	Ενδο μυϊκή έγκυηση ρεθιδίνη	Ενδοφλέβια έγκυηση ρεθιδίνη	Επισκόπη- ρίδιος	Βαθμολογία Αρραγ	kg	lb	oz	Ημερο- μηνία γέννησης	Ηλικία μητέρας (επ) κατά τη γέννηση του παιδιού	Ωμόδια αιματος	Συχνότητα αιμορραγίας σε άλλων	
47	3	3	08/08/74	.	3	6
33	3	.	41	0	1	0	1	.	6	13	11/08/52	27.26	1	4		
34	3	1	39	1	0	0	0	.	7	14	04/02/53	22.12	1	1		
43	3	1	41	1	1	0	0	.	8	0	26/02/54	27.51	3	33		
23	3	2	.	0	0	0	0	10/1-10/	11.19	.	.	29/12/65	36.58	1	3	
49	3	3	09/08/57	.	1	5	
51	3	3	21/06/51	.	3	5	
20	2	41	0	1	0	0	.	.	7	12	15/08/96	25.61	3	3		
64	4	.	1	1	0	0	0	10/11/51	24.61	3	2	
27	3	1	14	1	0	0	0	ok	8	8	02/12/71	22.45	1	1		
38	3	2	38	1	0	0	0	9/1-9/5	.	6	10	12/11/61	31.60	1	1	
50	3	2	40	0	0	0	0	.	5	11	06/02/68	18.75	1	6		
54	4	1	41	0	1	0	0	.	7	4	17/10/59	24.62	3	2		
7	1	1	40	0	0	0	1	.	6	5	17/12/65	20.35	2	5		
9	1	2	38	0	1	0	0	.	5	4	12/12/96	28.49	3	3		
17	1	4	15/05/71	26.81	1	5	
53	3	2	40	0	0	1	0	.	8	7	07/03/41	31.04	1	3		
56	4	2	40	0	0	0	0	.	.	0	16/11/57	37.86	3	3		
58	4	1	40	0	1	0	1	.	8	0	17/06/47	22.32	3	Y		
14	1	1	38	0	0	0	1	.	7	12	04/05/61	19.12	4	2		

1 = Αιμορραφία Α
2 = Αιμορραφία Β
3 = Νόσος του Von Willebrand
4 = Ανεπάρκεια F XI

0 = Οχι
1 = Ναι
1 = Άνδρας
2 = Γυναίκα
3 = Αποβολή
4 = Θυηριγενές

1 = 0+ νε
2 = 0- νε
3 = A+ νε
4 = A- νε
5 = B+ νε
6 = B- νε
7 = AB+ νε
8 = AB- νε
1 = Περισσότερο από φορά
την ημέρα
2 = Μία φορά την ημέρα
3 = Μία φορά την εβδομάδα
4 = Μία φορά τον μήνα
5 = Λιγότερο συχνά
6 = Ποτέ

Εικόνα 2.1 Μέρος του λογιστικού φύλλου που δείχνει τα δεδομένα που συλλέχθηκαν από ένα δείγμα 64 γυναικών, με κληρονομικές αιμορροφιλικές διαταραχές.

Συλλέχθηκαν δεδομένα από ένα δείγμα 64 γυναικών, που καταχωρίστηκαν, σε ένα απλό αιμορροφιλικό κέντρο στο Λονδίνο, ως ένα μέρος μιάς μελέτης διαταραχών κατά την εγκυμοσύνη και τη γέννηση του παιδιού. Οι γυναικες ρωτήθηκαν σχετικά με τις αιμορραγικές τους διαταραχές και την πρώτη τους εγκυμοσύνη (ή την τρέχουσα εγκυμοσύνη, αν ήταν έγκυες για πρώτη φορά κατά την ημερομηνία της συνέντευξης). Η Εικόνα 2.1 δείχνει ένα μέρος από τα δεδομένα του λογιστικού φύλλου,

αλλά προτού τα δεδομένα αυτά ελεγχθούν για λάθη. Η διάταξη της κωδικοποίησης για τις κατηγορικές μεταβλητές φαίνεται στο κάτω μέρος της Εικόνας 2.1. Κάθε γραμμή του λογιστικού φύλλου δείχνει ένα ξεχωριστό άτομο της μελέτης κάθε στήλη δείχνει μία διαφορετική μεταβλητή. Όπου η γυναίκα είναι σε κύηση, η ηλικία της γυναικας κατά τον χρόνο γέννησης έχει υπολογιστεί από την κατ' εκτίμηση ημερομηνία γέννησης του παιδιού. Δεδομένα σχετικά με γεννημένα ζωντανά παιδιά δίνονται στο Κεφάλαιο 34.

Τα δεδομένα αποτελούν ευγενική παραχώρηση του Dr. R.A. Kadir, University Department of Obstetrics and Gynecology, και του καθηγητή C.A. Lee, Haemophilia Center and Haemostasis Unit, Royal Free Hospital, London.

Σε κάθε ερευνητική μελέτη πάντοτε υπάρχει το ενδεχόμενο της ύπαρξης λαθών στα δεδομένα, είτε όταν γίνονται οι μετρήσεις είτε όταντα δεδομένα συλλέγονται, μεταφέρονται και εισάγονται στον Η/Υ. Είναι δύσκολο να ελαχιστοποιηθούν όλα αυτά τα λάθη. Ωστόσο, μπορείτε να μειώσετε τον αριθμό των λαθών μεταφοράς και πληκτρολόγησης ελέγχοντας τα δεδομένα προσεκτικά, όταν αυτά έχουν εισαχθεί στον Η/Υ. Απλώς, διαβάζοντας με μία ματιά τα δεδομένα, συχνά θα αναγνωρίσετε τιμές οι οποίες προφανώς είναι λαθεμένες. Συνιστούμε μία σειρά από άλλες προσεγγίσεις, τις οποίες μπορείτε να χρησιμοποιήσετε για τον έλεγχο των δεδομένων.

Λάθη πληκτρολόγησης

Κατά την εισαγωγή των δεδομένων στον Η/Υ, η πιο συχνή πηγή λαθών είναι εκείνη της πληκτρολόγησης. Αντο πλήθος των δεδομένων είναι μικρό, μπορείτε να ελέγχετε τα πληκτρολόγημένα δεδομένα με τα αρχικά ερωτηματολόγια (φόρμες), για να δείτε αν υπάρχουν λάθη πληκτρολόγησης. Ωστόσο, αυτός ο τρόπος είναι χαμένος χρόνος αν το πλήθος των δεδομένων είναι μεγάλο. Είναι δυνατόν να πληκτρολογηθούντα δεδομένα δύο φορές και να γίνει σύγκριση των δύο αρχείων, χρησιμοποιώντας πρόγραμμα Η/Υ. Οι όποιες διαφορές μεταξύ των δύο αρχείων θα οφείλονται στα λάθη πληκτρολόγησης. Αν και αυτή η προσέγγιση δεν αποκλείει το ενδεχόμενο το ίδιο λάθος να γίνει και στις δύο περιπτώσεις ή τη τιμή στο ερωτηματολόγιο να μην είναι σωστή, εντούτοις ελαχιστοποιεί τον αριθμό των λαθών. Το μειονέκτημα της μεθόδου αυτής είναι ότι απαιτείται διπλάσιος χρόνος για την είσοδο των δεδομένων στον Η/Υ, με συνέπεια το μεγαλύτερο κόστος ή τη χρονική καθυστέρηση.

Έλεγχος λαθών

- Κατηγορικά δεδομένα** – Ο έλεγχος των κατηγορικών δεδομένων είναι σχετικά εύκολος, επειδή οι απαντήσεις για κάθε μεταβλητή μπορεί να παίρνουν μία μόνο τιμή από ένα πλήθος περιορισμένων τιμών. Συνεπώς, τιμές που δεν επιτρέπονται πρέπει να είναι λαθεμένες.
- Ποσοτικά δεδομένα** – Τα ποσοτικά δεδομένα συχνά είναι δύσκολο να ελέγχουν, αλλά είναι επιρρεπή σε λάθη. Για παράδειγμα, όταν εισάγονται ποσοτικά δεδομένα είναι απλό να μετατοπίζονται ψηφία ή να τοποθετείται λαθεμένα ένα δεκαδικό σημείο. Τα ποσοτικά δεδομένα μπορεί να ελέγχονται με το **εύρος των τιμών** τους, δηλαδή για κάθε μεταβλητή μπορεί να καθοριστούν ανώτερα και κατώτερα όρια. Αν μία τιμή βρίσκεται εκτός των ορίων αυτών, τότε αυτή ελέγχεται περαιτέρω.
- Ημερομηνίες** – Είναι συχνά δύσκολο να ελεγχθεί η ακρίβεια των ημερομηνιών, αν και μερικές φορές μπορεί να γνωρίζετε ότι αυτές οι ημερομηνίες πρέπει να περιλαμβάνονται μέσα σε ορισμένες χρονικές περιόδους. Οι ημερομηνίες πρέπει να ελέγχονται, ώστε να είστε σίγουροι ότι αυτές ισχύουν. Για παράδειγμα, η 30ή Φεβρουαρίου πρέπει να είναι λαθεμένη, όπως και κάθε ημέρα του μήνα μεγαλύτερη από 31, και κάθε μήνας μεγαλύτερος από 12. Μπορεί, επίσης, να διενεργηθούν ορισμένοι λογικοί έλεγχοι. Για παράδειγμα, η ημερομηνία γέννησης ενός ασθενούς θα πρέπει να αντιστοιχεί στην ηλικία του και οι ασθενείς θα πρέπει να έχουν γεννηθεί πριν την έναρξη της μελέτης (τουλάχιστον στις περισσότερες έρευνες). Επιπλέον, ασθενείς που έχουν πεθάνει δεν μπορούν να εμφανίζονται σε επόμενες επισκέψεις παρακολούθησης!

Σε όλους τους έλεγχους λαθών, μία τιμή πρέπει να διορθώνεται μόνο αν υπάρχουν στοιχεία ότι έχει γίνει λάθος. Δεν πρέπει να αλλάζονται οι τιμές έτσι απλώς, επειδή αυτές φαίνονται ασυνήθιστες.

Χειρισμός ελλειπόντων δεδομένων

Τυχαίνει μερικά δεδομένα να είναι ελλιπή (missing). Ανένα μεγάλο ποσοστό δεδομένων είναι ελλιπές, τότε τα αποτελέσματα είναι απίθανο να είναι αξιόπιστα. Θα πρέπει πάντοτε να εξετάζονται οι λόγοι για τους οποίους δεδομένα είναι ελλιπή – αν τα ελλιπή δεδομένα τείνουν να συγκεντρώνονται σε μία ιδιαίτερη μεταβλητή ή/και σε μία ιδιαίτερη υποομάδα απόμων, αυτό μπορεί να σημαίνει ότι η μεταβλητή δεν είναι κατάλληλη ή δεν έχει

ποτέ μετρηθεί για αυτή την ομάδα απόμων. Στην τελευταία περίπτωση, η ομάδα των απόμων θα έπρεπε να εξαιρεθεί από κάθε ανάλυση για αυτή τη μεταβλητή. Μπορεί να συναντήσουμε συγκεκριμένα προβλήματα όταν η πιθανότητα ύπαρξης ελλειπόντων δεδομένων συνδέεται στενά με τη μεταβλητή ενδιαφέροντος στη μελέτη μας (π.χ. το αποτέλεσμα σε μία ανάλυση παλινδρόμησης, Κεφάλαιο 27). Σε αυτήν την περίπτωση, τα αποτελέσματα μας μπορεί να είναι σοβαρά εσφαλμένα (Κεφάλαιο 12). Πα παράδειγμα, έστω ότι μας ενδιαφέρει μία μέτρηση που αντικατοπτρίζει την κατάσταση υγείας των ασθενών και η πληροφορία αυτή λείπει για ορισμένους ασθενείς, διότι δεν ήταν αρκετά καλά για να προσέλθουν στο κλινικό ραντεβού τους είναι πιθανότερο ότι θα έχουμε μία συνολικά πιο αισιόδοξη θώρηση της υγείας των ασθενών, αν δεν υπολογίσουμε τα ελλιπή δεδομένα στην ανάλυση. Μπορεί να είναι πιθανόν να μειωθεί το σφάλμα αυτό, χρησιμοποιώντας κατάλληλες στατιστικές μεθόδους ή υπολογίζοντας τα ελλιπή δεδομένα κατά κάποιο τρόπο, αλλά μία προτιμώμενη εναλλακτική είναι να ελαχιστοποιήσουμε το πλήθος των ελλειπόντων δεδομένων από την αρχή.

Ακραίες τιμές

Τι σημαίνει ακραίες τιμές;

Ακραίες τιμές είναι οι παρατηρήσεις εκείνες που ξεχωρίζουν από τον κύριο κορμό των δεδομένων και δεν είναι συμβατές με τις υπόλοιπες τιμές. Αυτές οι τιμές μπορεί να είναι γνωστές παρατηρήσεις από άτομα με πολύ ακραία επίπεδα τιμών για αυτήν τη μεταβλητή. Όμως, μπορεί να είναι αποτέλεσμα λάθους πληκτρολόγησης και συνεπώς κάθε ύποπτη τιμή θα πρέπει να ελέγχεται. Είναι σημαντικό να διαπιστώνεται αν οι τιμές αυτές είναι ακραίες στα συγκεκριμένα δεδομένα, επειδή μπορεί αυτές να έχουν αξιόλογη επίδραση στα αποτελέσματα που προέρχονται από ορισμένου είδους αναλύσεις.

Για παράδειγμα, μία γνωστά η οποία έχει ανάστημα 210 cm θα εμφανίζοταν πιθανώς ως ακραία τιμή. Όμως, αν και η τιμή αυτή είναι φανερά πολύ υψηλή συγκρινόμενη με τα συνηθισμένα αναστήματα γυναικών, ενδέχεται να είναι πραγματική και η γυναίκα αυτή μπορεί απλώς να είναι πολύ ψηλή. Στην περίπτωση αυτή, θα έπρεπε η συγκρινόμενη τιμή να διερευνηθεί περισσότερο, ενδεχομένως ελέγχοντας άλλες μεταβλητές, όπως την ηλικία και το βάρος της, προτού ληφθεί οποιαδήποτε απόφαση για την εγκυρότητα αυτού του αποτελέσματος. Η τιμή αυτή θα έπρεπε να αλλάξει μόνο όταν πράγματι υπάρχει ένδειξη ότι δεν είναι σωστή.

Έλεγχος για ακραίες τιμές

Ένας απλός τρόπος ελέγχου είναι να τυπωθούν τα δεδομένα και να ελέγχουν «με το μάτι». Αυτός ο τρόπος είναι κατάλληλος, αν ο αριθμός των παρατηρήσεων δεν είναι πολύ μεγάλος και οι ενδεχόμενες ακραίες τιμές είναι πολύ μικρότερες ή μεγαλύτερες από τις υπόλοιπες τιμές των δεδομένων. Εναλλακτικά, μπορεί να γίνει γραφική παράσταση των δεδομένων (Κεφάλαιο 4) – οι ακραίες τιμές μπορούν να εξακριβωθούν σε ιστογράμματα και διαγράμματα σημειών (βλέπε, επίσης, Κεφάλαιο 29 για μια συζήτηση σχετικά με ακραίες τιμές στην ανάλυση εξάρτησης).

Χειρισμός ακραίων τιμών

Είναι σημαντικό να μην αφαιρεθεί ένα άτομο από την ανάλυση απλώς επειδή οι τιμές του είναι μεγαλύτερες ή μικρότερες από τις αναμενόμενες. Συνεπώς, όταν εφαρμόζονται κάποιες στατιστικές τεχνικές, ενδέχεται να παραμονή των ακραίων τιμών να επιδρά στα αποτελέσματα. Μία απλή προσέγγιση είναι η επανάληψη της ανάλυσης, με ή χωρίς τις ακραίες αυτές τιμές. Αν τα αποτελέσματα είναι παρόμοια, τότε οι ακραίες τιμές δεν έχουν καμία επίδραση στα αποτελέσματα. Ωστόσο, αυτά αποτελέσματα αλλάζουν σημαντικά, είναι σημαντικό να εφαρμοστούν κατάλληλες μεθόδους, οι οποίες δεν επηρεάζονται από τις ακραίες τιμές στην ανάλυση των δεδομένων. Στις μεθόδους αυτές συμπεριλαμβάνονται οι μετασχηματισμοί (Κεφάλαιο 9) και οι μη παραμετρικές δοκιμασίες (Κεφάλαιο 17).