Πρόλογος

Αυτό το εισαγωγικό βιβλίο βιοστατιστικής είναι σχεδιασμένο για προχωρημένους προπτυχιακούς φοιτητές ή απόφοιτους που ενδιαφέρονται για την ιατρική ή για άλλους τομείς σχετικούς με την υγεία. Δεν απαιτεί στατιστικό υπόβαθρο, ενώ σε επίπεδο μαθηματικών προϋποθέτει μόνο γνώσεις άλγεβρας.

Το βιβλίο *Θεμελιώδεις Αρχές Βιοστατιστικής* προήλθε από σημειώσεις που χρησιμοποίησα σε ένα μάθημα βιοστατιστικής το οποίο διδασκόταν σε προπτυχιακούς φοιτητές στο Πανεπιστήμιο του Χάρβαρντ, στην Ιατρική Σχολή του Χάρβαρντ και στους φοιτητές της Σχολής Δημόσιας Υγείας του Χάρβαρντ τα τελευταία 30 χρόνια. Έγραψα αυτό το βιβλίο για να παρακινήσω τους φοιτητές να μάθουν σε βάθος τις στατιστικές μεθόδους που χρησιμοποιούνται συχνότερα στην ιατρική βιβλιογραφία. Από την πλευρά των φοιτητών, είναι σημαντικό ότι το υλικό που χρησιμοποιείται στα παραδείγματα για την ανάπτυξη αυτών των μεθόδων είναι αντιπροσωπευτικό αυτού που υπάρχει στη βιβλιογραφία. Επομένως, τα περισσότερα παραδείγματα και οι ασκήσεις σε αυτό το βιβλίο βασίζονται είτε σε πραγματικά άρθρα της ιατρικής βιβλιογραφίας είτε σε πραγματικά ιατρικά ερευνητικά προβλήματα που έχω συναντήσει κατά τη διάρκεια της εμπειρίας μου ως επιστημονικού συμβούλου στην Ιατρική Σχολή του Χάρβαρντ.

Η προσέγγιση

Τα περισσότερα κείμενα στατιστικής είτε χρησιμοποιούν μια τελείως μη μαθηματική προσέγγιση (σαν βιβλίο με συνταγές) ή αναπτύσσουν το υλικό μέσα σε ένα αυστηρό, σύνθετο μαθηματικό πλαίσιο. Σε αυτό το βιβλίο, ωστόσο, ακολουθώ μια ενδιάμεση οδό, περιορίζοντας στο ελάχιστο τον αριθμό των μαθηματικών τύπων και επεξηγώντας πλήρως όλες τις σημαντικές έννοιες. Κάθε νέα έννοια αναπτύσσεται συστηματικά μέσω πλήρως λυμένων παραδειγμάτων από σύγχρονα ιατρικά ερευνητικά προβλήματα. Επιπλέον, εισάγω αποτελέσματα από υπολογιστή όπου αυτό ταιριάζει προκειμένου να επεξηγήσω τις έννοιες αυτές.

Αρχικά έγραψα αυτό το βιβλίο για το εισαγωγικό μάθημα της Βιοστατιστικής. Ωστόσο, ο τομέας αυτός άλλαξε δραματικά κατά τη διάρκεια των τελευταίων τριάντα ετών. Λόγω των αυξημένων δυνατοτήτων που διαθέτουν τα νεότερα στατιστικά πακέτα, είμαστε σε θέση τώρα, πιο πολύ από ποτέ, να εκτελούμε περισσότερο πολύπλοκες αναλύσεις δεδομένων. Επομένως, ένας δεύτερος στόχος του βιβλίου είναι να παρουσιάσει αυτές τις νέες τεχνικές σε εισαγωγικό επίπεδο, ώστε οι φοιτητές να μπορέσουν να εξοικειωθούν με αυτές χωρίς να χρειαστεί να ανατρέξουν σε εξειδικευμένα (και, συνήθως, πιο προχωρημένα) κείμενα στατιστικής.

Για να διαφοροποιήσω αυτούς τους δύο στόχους ακόμα περισσότερο, συμπεριέλαβα το μεγαλύτερο μέρος του περιεχομένου για το εισαγωγικό μάθημα στα πρώτα 12 κεφάλαια. Πιο προχωρημένες στατιστικές τεχνικές που χρησιμοποιούνται στις πρόσφατες επιδημιολογικές έρευνες καλύπτονται στο Κεφάλαιο 13, «Τεχνικές σχεδιασμού και ανάλυσης επιδημιολογικών μελετών», και στο Κεφάλαιο 14, «Έλεγχος υποθέσεων: Δεδομένα ανθρωποχρόνου».

Αλλαγές στην όγδοη έκδοση

Σε αυτή την έκδοση, έχουν προστεθεί τρεις νέες ενότητες και νέο περιεχόμενο σε τρεις άλλες ενότητες. Μεταξύ των νέων στοιχείων περιλαμβάνονται τα ακόλουθα:

- Τα αρχεία δεδομένων είναι τώρα διαθέσιμα στη συνοδευτική ιστοσελίδα του βιβλίου στη διεύθυνση www.cengage.com/statistics/rosner σε περισσότερους τύπους, συμπεριλαμβανομένων Excel, Minitab®, SPSS, JMP, SAS, Stata, R και ASCII.
- Τα δεδομένα και τα ευρήματα ιατρικής έρευνας στα παραδείγματα έχουν ενημερωθεί.
- Έχει προστεθεί νέα ή εκτεταμένη κάλυψη των ακόλουθων θεμάτων:

- Μέθοδος Bootstrap (Ενότητα 6.11)
- Συμπερασματολογία ενός δείγματος για τη Διωνυμική Κατανομή (Ενότητα 7.9)
- Έλεγχοι μεταθέσεων (Permutation Tests, Ενότητα 9.6)
- Υπολογισμός μεγέθους δείγματος για λογιστική παλινδρόμηση (Ενότητα 13.9)
- Εκτίμηση καμπύλης επιβίωσης: Ο εκτιμητής Kaplan-Meier (Ενότητα 14.9)
- Αναλυτικοί υπολογισμοί επιλεγμένων μαθηματικών τύπων (Ενότητες 7.12, 8.11, 10.9, 11.14, 12.11, 13.17, 14.15)

Οι νέες ενότητες και οι διευρυμένες ενότητες αυτής της έκδοσης σημειώνονται με έναν αστερίσκο στον πίνακα περιεχομένων.

Ασκήσεις - προβλήματα

Η έκδοση αυτή περιέχει 1.490 ασκήσεις· 171 από αυτές είναι καινούργιες. Τα δεδομένα και τα ευρήματα ιατρικής έρευνας στα προβλήματα έχουν ενημερωθεί όπου άρμοζε. Όλα τα προβλήματα που βασίζονται στα σύνολα δεδομένων συμπεριλαμβάνονται. Προβλήματα που σημειώνονται με αστερίσκο (*) στο τέλος κάθε κεφαλαίου έχουν αντίστοιχες σύντομες λύσεις στην ενότητα των απαντήσεων στο τέλος του βιβλίου. Με βάση αιτήματα φοιτητών για προβλήματα με πιο ολοκληρωμένες λύσεις, περίπου 600 επιπλέον προβλήματα με πλήρεις λύσεις παρουσιάζονται στον Οδηγό Μελέτης (Study Guide) που είναι διαθέσιμος στη συνοδευτική ιστοσελίδα της πρωτότυπης έκδοσης του βιβλίου. Επιπρόσθετα, περίπου 100 από αυτά τα προβλήματα συμπεριλαμβάνονται στην ενότητα Επιπλέον Προβλήματα (Miscellaneous Problems) και είναι τυχαία ταξινομημένα, ώστε να μην είναι συνδεδεμένα με κάποιο συγκεκριμένο κεφάλαιο του βιβλίου. Αυτό παρέχει στον φοιτητή τη δυνατότητα να εξασκηθεί περισσότερο στην επιλογή της μεθόδου που θα ακολουθήσει σε κάθε περίπτωση. Πλήρεις εκπαιδευτικές λύσεις σε όλες τις ασκήσεις είναι διαθέσιμες στη συνοδευτική εκπαιδευτική ιστοσελίδα cengage.com/statistics/rosner.

Υπολογιστική μέθοδος

Η μέθοδος διαχείρισης των υπολογισμών είναι παρόμοια με αυτή που χρησιμοποιήθηκε στην έβδομη έκδοση. Όλα τα ενδιάμεσα αποτελέσματα πραγματοποιούνται με πλήρη ακρίβεια (στους υπολογισμούς χρησιμοποιήθηκαν 10+ δεκαδικά ψηφία), αν και τα αποτελέσματα συνήθως παρουσιάζονται με λιγότερα δεκαδικά ψηφία (συνήθως 2 ή 3) στο κείμενο. Επομένως, σε μερικές περιπτώσεις, τα ενδιάμεσα αποτελέσματα μπορεί να φαίνονται ασυνεπή σε σχέση με τα τελικά αποτελέσματα· ωστόσο, κάτι τέτοιο δεν ισχύει.

Οργάνωση

Η όγδοη έκδοση του βιβλίου είναι οργανωμένη ως εξής:

Το **Κεφάλαιο 1** είναι ένα *εισαγωγικό κεφάλαιο*, όπου περιγράφεται η ανάπτυξη μιας πραγματικής ιατρικής μελέτης με την οποία είχα εμπλακεί. Παρέχει μια μοναδική αίσθηση του ρόλου της βιοστατιστικής στην ιατρική έρευνα.

Το **Κεφάλαιο 2** ασχολείται με την *περιγραφική στατιστική* και παρουσιάζει όλα τα σημαντικά αριθμητικά και γραφικά εργαλεία για την απεικόνιση ιατρικών δεδομένων. Το κεφάλαιο αυτό είναι εξαιρετικά σημαντικό τόσο για τους χρήστες όσο και για τους παραγωγούς της ιατρικής βιβλιογραφίας, διότι ένα μεγάλο μέρος των πληροφοριών μεταδίδεται στην πράξη μέσω περιγραφικού υλικού.

Τα **Κεφάλαια 3 έως 5** πραγματεύονται την *πιθανότητα*. Σε αυτά αναπτύσσονται οι βασικές αρχές των πιθανοτήτων και παρουσιάζονται οι πιο κοινές κατανομές πιθανοτήτων – όπως η διωνυμική και η κανονική κατανομή. Οι κατανομές αυτές χρησιμοποιούνται εκτενώς σε επόμενα κεφάλαια του βιβλίου. Επίσης, γίνεται μια εισαγωγή στις έννοιες των εκ των προτέρων πιθανοτήτων και των εκ των υστέρων πιθανοτήτων.

Τα **Κεφάλαια 6 έως 10** καλύπτουν κάποιες από τις βασικές μεθόδους της *στατιστικής συμπερασματολογίας*.

Το **Κεφάλαιο 6** μας εισάγει στην έννοια της επιλογής τυχαίου δείγματος από πληθυσμούς. Σε αυτό αναπτύσσεται η δύσκολη έννοια της κατανομής δειγματοληψίας και περιλαμβάνεται μια εισαγωγή στις πιο κοινά χρησιμοποιούμενες κατανομές δειγματοληψίας, όπως είναι η κατανομή *t* και η κατανομή *χι-τετράγωνο*. Παρουσιάζονται επίσης οι βασικές μέθοδοι *εκτίμησης*, συμπεριλαμβανομένης μιας εκτεταμένης συζήτησης για τα διαστήματα εμπιστοσύνης. *Επιπλέον, παρουσιάζεται για πρώτη φορά η μέθοδος bootstrap για την απόκτηση διαστημάτων εμπιστοσύνης*.

Τα **Κεφάλαια 7 και 8** περιέχουν τις βασικές αρχές των *ελέγχων υποθέσεων*. Οι πιο στοιχειώδεις έλεγχοι υποθέσεων για την κανονική κατανομή των δεδομένων, όπως είναι ο έλεγχος *t*, παρουσιάζονται επίσης αναλυτικά για προβλήματα με ένα και δύο δείγματα.

Το **Κεφάλαιο 9** καλύπτει τις βασικές αρχές της *μη παραμετρικής στατιστικής*. Οι υποθέσεις της κανονικότητας χαλαρώνουν και αναπτύσσονται αντιστοιχίσεις ελεύθερων κατανομών για τους ελέγχους των Κεφαλαίων 7 και 8. Οι *στατιστικοί έλεγχοι μεταθέσεων, οι οποίοι χρησιμοποιούνται ευρέως στις μελέτες γενετικής, εισάγονται για πρώτη φορά*.

Το **Κεφάλαιο 10** περιέχει τις βασικές αρχές του *ελέγχου υποθέσεων* όπως εφαρμόζονται σε κατηγορικά δεδομένα, συμπεριλαμβάνοντας κάποιες από τις ευρύτερα χρησιμοποιούμενες στατιστικές διαδικασίες όπως τον έλεγχο *χι-τετράγωνο* και τον ακριβή έλεγχο του Fisher.

Το **Κεφάλαιο 11** αναπτύσσει τις βασικές αρχές της *ανάλυσης παλινδρόμησης*. Η περίπτωση της απλής γραμμικής παλινδρόμησης καλύπτεται διεξοδικά, και παρέχονται προεκτάσεις αναφορικά με την περίπτωση της πολλαπλής παλινδρόμησης. Επίσης, εμπεριέχονται σημαντικές ενότητες για την καλή προσαρμογή των μοντέλων παλινδρόμησης. Επιπλέον, εισάγεται η έννοια της μη παραμετρικής συσχέτισης που βασίζεται στις τάξεις μεγέθους (δηλαδή στη σειρά κατάταξης της κάθε παρατηρούμενης τιμής), ενώ συμπεριλαμβάνονται μέθοδοι για τον υπολογισμό διαστημάτων εμπιστοσύνης για τη μη παραμετρική συσχέτιση τάξεων.

Το **Κεφάλαιο 12** παρουσιάζει τις βασικές αρχές της *ανάλυσης διακύμανσης* (ANOVA). Εξετάζεται η ανάλυση διακύμανσης κατά έναν παράγοντα μοντέλων σταθερής και τυχαίας επίδρασης. Επιπρόσθετα, καλύπτονται η ANOVA κατά δύο παράγοντες, η ανάλυση συνδιακύμανσης και τα μοντέλα μεικτής επίδρασης. Τέλος, εξετάζονται μη παραμετρικές προσεγγίσεις για την ANOVA κατά έναν παράγοντα. Επίσης, θα βρείτε λεπτομέρειες για τις μεθόδους πολλαπλών συγκρίσεων αλλά και υλικό σχετικό με την αναλογία εσφαλμένου εντοπισμού (False Discovery Rate).

Το **Κεφάλαιο 13** εξετάζει μεθόδους σχεδιασμού και ανάλυσης *επιδημιολογικών μελετών*. Στο κεφάλαιο αυτό παρουσιάζονται οι πιο σημαντικοί σχεδιασμοί μελετών, συμπεριλαμβανομένων της προοπτικής μελέτης (prospective study), της μελέτης ασθενών-μαρτύρων (case-control study), της συγχρονικής μελέτης (cross-sectional study), καθώς και του διασταυρούμενου σχεδιασμού (cross-over design) σε μια ιατρική μελέτη.

Παρουσιάζεται επίσης η έννοια του συγχυτικού παράγοντα (confounding variable) –δηλαδή, μιας μεταβλητής που συσχετίζεται και με την ασθένεια και με τη μεταβλητή έκθεσης– και εξετάζονται λεπτομερώς μέθοδοι για τον έλεγχο της συνεπίδρασης, που περιλαμβάνουν τον έλεγχο Mantel-Haenszel και την πολλαπλή λογιστική παλινδρόμηση. Επιπλέον, εξετάζονται επεκτάσεις μοντέλων λογιστικής παλινδρόμησης, συμπεριλαμβανομένων της υπό συνθήκη λογιστικής παλινδρόμησης, της πολυωνυμικής λογιστικής παλινδρόμησης και της διατάξιμης λογιστικής παλινδρόμησης. *Μέθοδοι εκτίμησης μεγέθους δείγματος για μοντέλα λογιστικής παλινδρόμησης παρέχονται για πρώτη φορά*. Η παρουσίαση αυτή ακολουθείται από τη διερεύνηση θεμάτων τρέχοντος ενδιαφέροντος στην ανάλυση επιδημιολογικών δεδομένων, στα οποία συμπεριλαμβάνονται: η μετα-ανάλυση (ο συνδυασμός αποτελεσμάτων από περισσότερες από μία μελέτες)· τεχνικές συσχετισμένων δίτιμων δεδομένων (τεχνικές που μπορούν να εφαρμοστούν όταν έχουμε επαναλαμβανόμενες μετρήσεις, όπως για παράδειγμα δεδομένα από πολλαπλά δόντια από τον ίδιο άνθρωπο, διαθέσιμες για ένα άτομο)· μέθοδοι μέτρησης σφαλμάτων (χρήσιμες όταν υπάρχει σημαντικό σφάλμα μέτρησης στα συλλεχθέντα δεδομένα έκθεσης)· μελέτες ισοδυναμίας (των οποίων σκοπός είναι να εγκαθιδρύσουν βιοϊσοδυναμία μεταξύ δύο θεραπειών, αντί η μία θεραπεία να υπερέχει της άλλης)· και μέθοδοι ελλειπουσών τιμών, για το πώς διαχειριζόμαστε τα ελλιπή δεδομένα στις επιδημιολογικές μελέτες. Η ανάλυση διαχρονικών δεδομένων και μέθοδοι εκτίμησης γενικευμένων εξισώσεων [generalized estimating equation (GEE)] εξετάζονται επίσης εν συντομία.

Το **Κεφάλαιο 14** παρουσιάζει μεθόδους ανάλυσης *δεδομένων ανθρωποχρόνου* (person-time). Οι μέθοδοι που καλύπτονται σε αυτό το κεφάλαιο περιλαμβάνουν αυτές που αφορούν δεδομένα ρυθμού επίπτωσης, καθώς και αρκετές μεθόδους ανάλυσης επιβίωσης: την εκτίμηση καμπύλης επιβίωσης των Kaplan-Meier, τον έλεγχο log-rank και το μοντέλο αναλογικών κινδύνων. Συμπεριλαμβάνονται επίσης μέθοδοι για τον έλεγχο των υποθέσεων του μοντέλου αναλογικών κινδύνων. Επιπλέον, εξετάζονται παραμετρικές μέθοδοι ανάλυσης επιβίωσης.

Σε όλη την έκταση του κειμένου –ιδιαίτερα στο Κεφάλαιο 13– αναφέρομαι στα στοιχεία του σχεδιασμού μελετών, μεταξύ των οποίων συμπεριλαμβάνονται οι έννοιες της εξομοίωσης , οι μελέτες κοόρτης, οι μελέτες ασθενών-μαρτύρων, οι αναδρομικές μελέτες, οι προοπτικές μελέτες, καθώς και η ευαισθησία, η ειδικότητα και η τιμή πρόβλεψης των προσυμπτωματικών ανιχνευτικών ελέγχων. Οι συγκεκριμένοι σχεδιασμοί παρουσιάζονται στο πλαίσιο πραγματικών παραδειγμάτων. Επιπλέον, τα Κεφάλαια 7, 8, 10, 11, 13 και 14 περιέχουν συγκεκριμένες ενότητες με την εκτίμηση μεγέθους δείγματος σε διαφορετικές στατιστικές περιπτώσεις.

Πραγματοποιήθηκαν δύο σημαντικές οργανωτικές αλλαγές στην παρουσίαση του υλικού στο βιβλίο. Πρώτον, η προέλευση των πιο πολύπλοκων εξισώσεων μεταφέρθηκε είτε μετά τη δήλωση μιας εξίσωσης ή σε ξεχωριστές ενότητες στο τέλος του κεφαλαίου, ώστε να παρασχεθεί στους φοιτητές η δυνατότητα να έχουν πιο άμεση πρόσβαση στα κύρια αποτελέσματα των εξισώσεων. Δεύτερον, υπάρχουν πολυάριθμες υποενότητες με τον τίτλο «Εκτέλεση συγκεκριμένου ελέγχου με τη χρήση υπολογιστή», ώστε να τονίζεται εμφανέστερα η χρήση του υπολογιστή για την εφαρμογή πολλών μεθόδων του κειμένου.

Ένα διάγραμμα ροής των κατάλληλων μεθόδων στατιστικής συμπερασματολογίας (βλ. σελ. 979-983) αποτελεί έναν χρήσιμο οδηγό αναφοράς στις μεθόδους που αναπτύσσονται σε αυτό το βιβλίο. Δίνονται επίσης σελίδες αναφοράς για κάθε κύρια μέθοδο που παρουσιάζεται στο βιβλίο. Στα Κεφάλαια 7 και 8 και στα Κεφάλαια 10-14, κάνω λόγο στους φοιτητές γι' αυτό το διάγραμμα ροής, για να τους δώσω κάποια προοπτική ως προς το πώς οι μέθοδοι που έχουν συζητηθεί σε ένα κεφάλαιο συνδέονται με όλες τις άλλες στατιστικές μεθόδους που παρουσιάστηκαν σε αυτό το βιβλίο.

Επιπρόσθετα, έχω παραθέσει ένα ευρετήριο εφαρμογών, ομαδοποιημένο ως προς την ιατρική ειδικότητα, συνοψίζοντας όλα τα παραδείγματα και τα προβλήματα που καλύπτει το βιβλίο.

Τέλος, παρέχουμε για πρώτη φορά ένα ευρετήριο λογισμικού υπολογιστή, ώστε να γίνεται πιο ξεκάθαρα η αναγνώριση των υπολογιστικών εντολών σε συγκεκριμένα υπολογιστικά προγράμματα που παρουσιάζονται στο βιβλίο.

Ευχαριστίες

Είμαι υποχρεωμένος στην Debra Sheldon, στην εκλιπούσα Marie Sheehan και στον Harry Taplin για την ανεκτίμητη βοήθειά τους στη δακτυλογράφηση του χειρογράφου, στον Dale Rinkel για την πολύτιμη βοήθεια στην πληκτρολόγηση των λύσεων των προβλημάτων και στη Marion McPhee για τη βοήθεια στην προετοιμασία των σετ δεδομένων στη συνοδευτική ιστοσελίδα (Companion Website). Είμαι επίσης υπόχρεος στον Roland Matsouaka για την ενημέρωση των λύσεων στα προβλήματα αυτής της έκδοσης, καθώς και στη Virginia Piaseczny για τη δακτυλογράφηση του Ευρετηρίου Εφαρμογών. Επιπλέον, θα ήθελα να ευχαριστήσω τους κριτικούς αναγνώστες του χειρογράφου, μεταξύ των οποίων συγκαταλέγονται οι εξής: Shouhao Zhou, Daniela Szatmari-Voicu, Jianying Gu, Raid Amin, Claus Wilke, Glen Johnson, Kara Zografos και Hui Zhao. Θα ήθελα επίσης να ευχαριστήσω τους συναδέλφους μου Nancy Cook, η βοήθεια της οποίας υπήρξε καθοριστική στην ανάπτυξη της Ενότητας 12.4 για την αναλογία εσφαλμένου εντοπισμού (false discovery rate), και Robert Glynn, που η βοήθειά του ήταν ανεκτίμητη στην ανάπτυξη της Ενότητας 13.16 για τις ελλείπουσες τιμές και της Ενότητας 14.11 για τον έλεγχο των υποθέσεων του μοντέλου αναλογικών κινδύνων.

Επιπλέον, θα ήθελα να ευχαριστήσω τους Spencer Arritt και Jay Campbell, των οποίων η συνεισφορά ήταν καθοριστική ως προς την παροχή συμβουλών αναφορικά με την έκδοση και την προετοιμασία του αρχικού κειμένου του βιβλίου.

Είμαι επίσης υπόχρεος στους συναδέλφους μου στο Εργαστήριο Channing (Channing Laboratory) –ιδιαίτερα στον εκλιπόντα Edward Kass, στους Frank Speizer, Charles Hennekens, στον εκλιπόντα Frank Polk,

στους Ira Tager, Jerome Klein, James Taylor, Stephen Zinner, Scott Weiss, Frank Sacks, Walter Willett, Alvaro Munoz, Graham Colditz και στη Susan Hankinson– και στους άλλους συναδέλφους μου στην Ιατρική Σχολή του Χάρβαρντ (Harvard Medical School), ιδιαίτερα στον εκλιπόντα Frederick Mosteller, στους Eliot Berson, Robert Ackerman, Mark Abelson, Arthur Garvey, Leo Chylack, Eugene Braunwald και Arthur Dempster, που με ενέπνευσαν στη συγγραφή αυτού του βιβλίου. Θα ήθελα επίσης να εκφράσω την ευγνωμοσύνη μου στους John Hopper και Philip Landrigan για την παροχή των δεδομένων για τις μελέτες περιπτώσεων.

Τέλος, θα ήθελα να εκφράσω τις ευχαριστίες μου στους Leslie Miller, Andrea Wagner, Ithamar Jotkowitz, Loren Fishman και Frank Santopietro, χωρίς την κλινική βοήθεια των οποίων η τρέχουσα έκδοση αυτού του βιβλίου δεν θα ήταν εφικτή.

Bernard Rosner

Σχετικά με τον Συγγραφέα

Ευγενική παραχώρηση του Museum of Scienc, Βοστόνι.

Ο **Bernard Rosner** είναι καθηγητής της Ιατρικής (Βιοστατιστκής) στην Ιατρική Σχολή του Χάρβαρντ (Harvard Medical School) και Καθηγητής Βιοστατιστικής στη Σχολή Δημόσιας Υγείας του Χάρβαρντ (Harvard School of Public Health). Έλαβε το πρώτο πτυχίο (B.A.) στα Μαθηματικά από το Πανεπιστήμιο Κολούμπια (Columbia University) το 1967, Μεταπτυχιακό (M.S.) στη Στατιστική από το Πανεπιστήμιο του Στάνφορντ (Stanford University) το 1968, και Διδακτορικό (Ph.D.) στη Στατιστική από το Πανεπιστήμιο του Χάρβαρντ (Harvard University) το 1971.

Έχει περισσότερα από 30 χρόνια εμπειρία ως επιστημονικός σύμβουλος βιοστατιστικής μαζί με άλλους ερευνητές στην Ιατρική Σχολή του Χάρβαρντ. Ειδικοί τομείς ενδιαφέροντος περιλαμβάνουν την καρδιαγγειακή νόσο, την υπέρταση, τον καρκίνο του μαστού και την οφθαλμιατρική. Πολλά από τα παραδείγματα και τις ασκήσεις που χρησιμοποιούνται στο βιβλίο αντανακλούν δεδομένα που συλλέχθηκαν από πραγματικές μελέτες σε συνδυασμό με τη συμβουλευτική του εμπειρία. Επιπλέον, έχει αναπτύξει νέες βιοστατιστικές μεθόδους, κυρίως στα πεδία της ανάλυσης διαχρονικών δεδομένων, της ανάλυσης δεδομένων κατά συστάδες (όπως, για παράδειγμα, δεδομένων που συλλέχθηκαν μέσα σε οικογένειες ή από όργανα του σώματος που υπάρχουν κατά ζεύγη στο ίδιο άτομο), των μεθόδων μέτρησης σφαλμάτων και των μεθόδων εντοπισμού ελλειπουσών τιμών. Θα δείτε κάποιες από αυτές τις μεθόδους να παρουσιάζονται στο βιβλίο αυτό σε βασικό επίπεδο.

Ο Bernard Rosner παντρεύτηκε το 1972 τη σύζυγό του, Cynthia, και έχουν τρία παιδιά, τη Sarah, τον David και τη Laura, καθένα από τα οποία έχει συνεισφέρει παραδείγματα στο παρόν βιβλίο.

1 | Γενική Επισκόπηση

Στατιστική είναι η επιστήμη με την οποία εξάγονται συμπεράσματα σχετικά με συγκεκριμένα τυχαία φαινόμενα με βάση περιορισμένο αριθμό δειγματοληπτικού υλικού. Το πεδίο της στατιστικής έχει δύο κύριες περιοχές: τη μαθηματική στατιστική και την εφαρμοσμένη στατιστική. Η μαθηματική στατιστική αφορά την ανάπτυξη νέων μεθόδων στατιστικής συμπερασματολογίας και απαιτεί βαθιά γνώση των θεωρητικών μαθηματικών για την εφαρμογή της. Η εφαρμοσμένη στατιστική περιλαμβάνει την εφαρμογή των μεθόδων της μαθηματικής στατιστικής σε συγκεκριμένες ερευνητικές περιοχές, όπως στα οικονομικά, στην ψυχολογία και στη δημόσια υγεία. Η βιοστατιστική είναι ο κλάδος της εφαρμοσμένης στατιστικής που εφαρμόζει στατιστικές μεθόδους σε ιατρικά και βιολογικά προβλήματα. Φυσικά, οι περιοχές αυτές της στατιστικής αλληλοεπικαλύπτονται ως έναν βαθμό. Για παράδειγμα, σε κάποιες περιπτώσεις, με δεδομένη μια συγκεκριμένη βιοστατιστική εφαρμογή, οι τυποποιημένες μέθοδοι δεν εφαρμόζονται και απαιτείται να προσαρμοστούν. Στην περίπτωση αυτή, οι βιοστατιστικοί καλούνται να αναπτύξουν νέες μεθόδους.

Ένας καλός τρόπος για να αποκτήσει κάποιος γνώσεις για τη βιοστατιστική και τον ρόλο της στην ερευνητική διαδικασία είναι να ακολουθήσει τη ροή μιας ερευνητικής μελέτης από την έναρξή της, στη φάση σχεδιασμού, μέχρι την ολοκλήρωσή της, η οποία συνήθως συμπίπτει με τη δημοσίευση μιας χειρόγραφης αναφοράς των αποτελεσμάτων της μελέτης. Θα περιγράψω, ως παράδειγμα, μία τέτοια μελέτη στην οποία συμμετείχα.

Ένα πρωί μου τηλεφώνησε ένας φίλος και κατά τη διάρκεια της συζήτησής μας ανέφερε ότι είχε χρησιμοποιήσει πρόσφατα μια νέα, αυτόματη συσκευή μέτρησης της αρτηριακής πίεσης σαν κι αυτές που βλέπουμε σε πολλές τράπεζες, ξενοδοχεία και πολυκαταστήματα. Η συσκευή μέτρησε τη μέση διαστολική πίεση σε πολλές περιπτώσεις ως 115 mm Hg· η υψηλότερη μέτρηση ήταν 130 mm Hg. Ανησύχησα πολύ, γιατί αν αυτές οι μετρήσεις ήταν ακριβείς, ο φίλος μου μπορεί να διέτρεχε άμεσο κίνδυνο να πάθει εγκεφαλικό ή να αναπτύξει κάποια άλλη σοβαρή καρδιαγγειακή νόσο. Τον παρέπεμψα σε έναν συνάδελφό μου κλινικό γιατρό, ο οποίος, με τη χρήση του κλασικού πιεσόμετρου μπράτσου, μέτρησε τη διαστολική πίεση του φίλου μου ως 90 mm Hg. Η αντιφατικότητα των μετρήσεων μου κέντρισε το ενδιαφέρον και άρχισα να καταγράφω μετρήσεις από την ψηφιακή συσκευή όποτε περνούσα από το μηχάνημα στην τράπεζα της γειτονιάς μου. Είχα την ξεκάθαρη εντύπωση ότι το μεγαλύτερο ποσοστό των μετρήσεων που είχαν καταγραφεί ήταν μέσα στο εύρος της υπέρτασης. Παρότι θα ανέμενε κανείς να χρησιμοποιούν μια τέτοια συσκευή περισσότερο τα υπερτασικά άτομα, εξακολουθούσα να πιστεύω ότι οι μετρήσεις αρτηριακής πίεσης από τις συσκευές αυτές μπορεί να μην ήταν σύμφωνες με αυτές που παίρνουμε όταν χρησιμοποιούμε τις τυποποιημένες μεθόδους μέτρησης της πίεσης.

Μίλησα με τον Δρα B. Frank Polk, γενικό γιατρό στην Ιατρική Σχολή του Χάρβαρντ που ασχολούνταν με την υπέρταση, σχετικά με την υποψία μου και πέτυχα να του εξάψω το ενδιαφέρον για μια μικρού εύρους εκτίμηση αυτών των συσκευών. Αποφασίσαμε να στείλουμε έναν ανθρώπινο παρατηρητή, που ήταν πολύ καλά εκπαιδευμένος στις τεχνικές μέτρησης αρτηριακής πίεσης, σε πολλές από αυτές τις συσκευές. Θα προσφερόταν να καταβάλει στους συμμετέχοντες τα 50 σεντς, που ήταν το κόστος χρήσης του μηχανήματος, εάν συμφωνούσαν να συμπληρώσουν ένα σύντομο ερωτηματολόγιο και να δεχτούν να τους μετρήσει την αρτηριακή πίεση και ο άνθρωπος-παρατηρητής και η αυτόματη συσκευή.

Στο στάδιο αυτό έπρεπε να λάβουμε αρκετές σημαντικές αποφάσεις, καθεμία από τις οποίες αποδείχτηκε ζωτικής σημασίας για την επιτυχία της μελέτης. Οι αποφάσεις αυτές βασίστηκαν στα ακόλουθα ερωτήματα:

(1) Πόσα μηχανήματα θα πρέπει να ελέγξουμε;
(2) Πόσοι συμμετέχοντες θα πρέπει να ελεγχθούν σε κάθε συσκευή;
(3) Με ποια σειρά θα πρέπει να λάβουμε τις μετρήσεις; Δηλαδή, θα πρέπει ο άνθρωπος ή το μηχάνημα να λάβει την πρώτη μέτρηση; Κάτω από ιδανικές συνθήκες, θα παίρναμε τόσο την ανθρώπινη όσο και την τεχνητή μέτρηση ταυτόχρονα, αλλά αυτό πρακτικά ήταν αδύνατον.
(4) Τι δεδομένα θα πρέπει να συλλέξουμε στο ερωτηματολόγιο τα οποία ενδεχομένως να επηρεάζουν τη σύγκριση των δύο μεθόδων;
(5) Πώς θα πρέπει να καταγράψουμε τα δεδομένα ώστε να διευκολύνουμε τη μετέπειτα μηχανογράφηση;
(6) Πώς θα ελέγξουμε την ακρίβεια των μηχανογραφημένων δεδομένων;

Επιλύσαμε τα προβλήματα αυτά ως ακολούθως:

(1) και (2) Επειδή δεν ήμασταν σίγουροι εάν όλες οι συσκευές αρτηριακής πίεσης ήταν συγκρίσιμες ποιοτικά, αποφασίσαμε να ελέγξουμε τέσσερις από αυτές. Ωστόσο, θέλαμε να πάρουμε δείγμα αρκετών ατόμων από αυτές τις συσκευές, ώστε να εξασφαλίσουμε μια ακριβή σύγκριση μεταξύ της τυπικής και της αυτόματης μεθόδου για κάθε μηχάνημα. Προσπαθήσαμε να προβλέψουμε πόσο μεγάλη μπορεί να είναι η ασυμφωνία μεταξύ των δύο μεθόδων. Χρησιμοποιώντας τις μεθόδους του καθορισμού μεγέθους δείγματος που περιγράφονται σε αυτό το βιβλίο, υπολογίσαμε ότι θα χρειαστούμε 100 συμμετέχοντες σε κάθε τοποθεσία προκειμένου να πραγματοποιήσουμε ακριβή σύγκριση.

(3) Έπρεπε μετά να αποφασίσουμε με ποια σειρά να πάρουμε τις μετρήσεις σε κάθε άτομο. Σύμφωνα με κάποιες μελέτες, ένα πρόβλημα με τη λήψη επαναλαμβανομένων μετρήσεων αρτηριακής πίεσης είναι ότι τα άτομα τείνουν να αγχώνονται κατά την πρώτη μέτρηση, με αποτέλεσμα να παρατηρούνται αυξημένες τιμές πίεσης. Επομένως, δεν θα θέλαμε να έχουμε πάντα είτε την αυτόματη είτε τη χειροκίνητη μέθοδο πρώτη, διότι η επίδραση της μεθόδου θα συγχεόταν με την επίδραση της σειράς μέτρησης. Μία βολική τεχνική που χρησιμοποιήσαμε εδώ ήταν να **τυχαιοποιήσουμε** τη σειρά με την οποία ελήφθησαν οι μετρήσεις, ώστε να είναι εξίσου πιθανό η πρώτη μέτρηση για οποιοδήποτε άτομο να ληφθεί είτε από το μηχάνημα είτε από άνθρωπο. Αυτό το τυχαίο μοτίβο μπορούσε να εφαρμοστεί είτε με το στρίψιμο ενός νομίσματος είτε, το πιθανότερο, με τη χρήση ενός πίνακα **τυχαίων αριθμών** παρόμοιου με τον Πίνακα 4 στο Παράρτημα.

(4) Πιστεύαμε ότι ο κύριος εξωτερικός παράγοντας που θα μπορούσε να επηρεάσει το αποτέλεσμα ήταν οι διαστάσεις του σώματος (θα αντιμετωπίζαμε μεγαλύτερη δυσκολία στο να λάβουμε ακριβείς μετρήσεις από άτομα με πιο παχιά μπράτσα παρά από άτομα με λεπτότερα μπράτσα). Θέλαμε επίσης να έχουμε μια ιδέα για τον τύπο των ατόμων που χρησιμοποιούν τις συγκεκριμένες συσκευές. Για τον λόγο αυτό, κάναμε ερωτήσεις για την ηλικία, το φύλο και το ιστορικό υπέρτασης.

(5) Για την καταγραφή των στοιχείων, αναπτύξαμε μια φόρμα κωδικοποίησης που μπορούσε να συμπληρωθεί επιτόπου και από την οποία τα δεδομένα μπορούσαν εύκολα να εισαχθούν σε έναν υπολογιστή για περαιτέρω ανάλυση. Σε κάθε άτομο της μελέτης αντιστοιχίστηκε ένας μοναδικός αριθμός αναγνώρισης (ID – από το «identification») μέσω του οποίου ο υπολογιστής μπορούσε να ταυτοποιήσει το άτομο. Τα δεδομένα στις φόρμες κωδικοποίησης εν συνεχεία συσχετίστηκαν μοναδικά και επαληθεύτηκαν. Δηλαδή, η ίδια φόρμα εισήχθη δύο φορές και οι δύο εγγραφές συγκρίθηκαν προκειμένου να διαπιστωθεί ότι ήταν ίδιες. Εάν οι εγγραφές δεν είχαν αντιστοίχιση, τα στοιχεία της φόρμας εισάγονταν ξανά.

(6) Ο έλεγχος κάθε στοιχείου σε κάθε φόρμα ήταν αδύνατος λόγω του μεγάλου πλήθους δεδομένων που συμπεριλάμβανε. Αντί γι' αυτό, μετά την εισαγωγή των δεδομένων εκτελέσαμε κάποια προγράμματα επεξεργασίας ώστε να διασφαλίσουμε την ακρίβεια των δεδομένων. Τα προγράμματα αυτά έλεγχαν ότι οι τιμές των μεμονωμένων μεταβλητών έπεφταν μέσα σε προκαθορισμένα διαστήματα τιμών και εκτύπωναν παρεκκλίνουσες τιμές για χειροκίνητο έλεγχο. Για παράδειγμα, ελέγξαμε ότι όλες οι μετρήσεις αρτηριακής πίεσης ήταν τουλάχιστον 50 mm Hg και όχι μεγαλύτερες από 300 mm Hg, και εκτυπώναμε όλες τις μετρήσεις που έπεφταν έξω από αυτό το διάστημα. Επίσης, εκτελέσαμε προγράμματα για τον εντοπισμό ακραίων τιμών όπως αναφέρονται αργότερα στο βιβλίο.

Μετά την ολοκλήρωση των φάσεων συλλογής, εισαγωγής και επεξεργασίας δεδομένων, ήμασταν έτοιμοι

να δούμε τα αποτελέσματα της μελέτης. Το πρώτο βήμα αυτής της διαδικασίας είναι να αποκτήσουμε μια εικόνα για τα δεδομένα συνοψίζοντας την πληροφορία με τη μορφή αρκετών περιγραφικών στατιστικών. Αυτό το περιγραφικό υλικό μπορεί να είναι είτε αριθμητικό είτε οπτικό. Εάν είναι αριθμητικό, τότε μπορεί να έχει τη μορφή μερικών συνοπτικών στατιστικών δεικτών, οι οποίοι είναι δυνατόν να παρουσιάζονται πινακοποιημένοι, ή εναλλακτικά με τη μορφή **κατανομής συχνοτήτων**, με την οποία καταγράφεται κάθε τιμή των δεδομένων και η συχνότητα εμφάνισής τους. Εάν το υλικό είναι οπτικό, τότε τα δεδομένα παρουσιάζονται με τη μορφή εικόνας (διαγραμματικά) και μπορούν να παρουσιαστούν σε ένα ή σε περισσότερα διαγράμματα. Ο κατάλληλος τύπος περιγραφικού υλικού που θα χρησιμοποιηθεί ποικίλλει ανάλογα με τον τύπο κατανομής που έχουμε. Εάν η κατανομή είναι **συνεχής** –δηλαδή, εάν υπάρχει ουσιαστικά ένα άπειρο πλήθος πιθανών τιμών, όπως είναι η περίπτωση της αρτηριακής πίεσης–, τότε οι μέσες τιμές και οι τυπικές αποκλίσεις μπορεί να είναι τα κατάλληλα περιγραφικά στατιστικά. Ωστόσο, εάν η κατανομή είναι **διακριτή** –δηλαδή, εάν υπάρχουν μόνο μερικές πιθανές τιμές, όπως είναι η περίπτωση του φύλου–, τότε το κατάλληλο περιγραφικό μέτρο είναι τα ποσοστά των ατόμων που παίρνουν την κάθε τιμή. Σε μερικές περιπτώσεις και οι δύο τύποι περιγραφικών στατιστικών χρησιμοποιούνται για τις συνεχείς κατανομές, συγκεντρώνοντας το εύρος των πιθανών τιμών σε μικρότερες ομάδες και δίνοντας το ποσοστό των ατόμων που ανήκουν σε κάθε ομάδα (π.χ. τα ποσοστά των ατόμων που έχουν αρτηριακή πίεση μεταξύ 120 και 129 mm Hg, μεταξύ 130 και 139 mm Hg, κ.ο.κ.).

Σε αυτή τη μελέτη αποφασίσαμε πρώτα να δούμε τη μέση αρτηριακή πίεση για κάθε μέθοδο σε καθεμία από τις τέσσερις τοποθεσίες. Ο Πίνακας 1.1 συνοψίζει αυτή τη πληροφορία [1].

Από τον πίνακα αυτό μπορεί να παρατηρήσετε ότι δεν λάβαμε σημαντικά δεδομένα και από τα 100 άτομα που ερευνήσαμε σε κάθε τοποθεσία. Αυτό συνέβη επειδή δεν μπορέσαμε να λάβουμε έγκυρες τιμές από τη συσκευή για αρκετά άτομα. Αυτό το πρόβλημα των ελλειπουσών τιμών είναι πολύ συνηθισμένο στη βιοστατιστική και θα πρέπει να το θεωρούμε αναμενόμενο στο στάδιο του σχεδιασμού, όταν αποφασίζουμε για το μέγεθος του δείγματος (κάτι το οποίο δεν έγινε στην παρούσα μελέτη).

ΠΙΝΑΚΑΣ 1.1 Μέση αρτηριακή πίεση και διαφορές μεταξύ μετρήσεων συσκευών και μετρήσεων από άνθρωπο (ερευνητή) σε τέσσερις τοποθεσίες

		Συστολική αρτηριακή πίεση (mm Hg)					
		Συσκευή		Άνθρωπος		Διαφορά	
Τοποθεσία	Πλήθος ατόμων	Μέση τιμή	Τυπική απόκλιση	Μέση τιμή	Τυπική απόκλιση	Μέση τιμή	Τυπική απόκλιση
A	98	142.5	21.0	142.0	18.1	0.5	11.2
B	84	134.1	22.5	133.6	23.2	0.5	12.1
C	98	147.9	20.3	133.9	18.3	14.0	11.7
D	62	135.4	16.7	128.5	19.0	6.9	13.6

Πηγή: Βασισμένο στην American Heart Association, Inc.

Το επόμενο βήμα στη μελέτη ήταν να αποφασίσουμε εάν οι εμφανείς διαφορές στην αρτηριακή πίεση μεταξύ των μετρήσεων από τις συσκευές και των ανθρώπινων μετρήσεων σε δύο τοποθεσίες (C, D) ήταν «αληθινές» κατά μία έννοια ή ήταν «τυχαίο γεγονός». Αυτού του είδους το ερώτημα υπάγεται στην περιοχή της **επαγωγικής στατιστικής**. Κατανοήσαμε ότι, αν και υπήρχε μία διαφορά 14 mm Hg στη μέση συστολική αρτηριακή πίεση μεταξύ των δύο μεθόδων για τα 98 άτομα που εξετάσαμε στην τοποθεσία C, η διαφορά αυτή μπορεί να μην εξακολουθούσε να ισχύει εάν εξετάζαμε 98 διαφορετικά άτομα στην ίδια τοποθεσία σε διαφορετικό χρόνο, και θέλαμε να έχουμε μια ιδέα για το **σφάλμα εκτίμησης** του 14 mm Hg. Στη στατιστι-

κή γλώσσα, αυτή η ομάδα των 98 ατόμων αντιπροσωπεύει ένα **δείγμα** από τον **πληθυσμό** όλων των ατόμων που μπορούν ενδεχομένως να χρησιμοποιήσουν αυτή τη συσκευή. Ενδιαφερόμασταν για τον πληθυσμό, επομένως θέλαμε να χρησιμοποιήσουμε το δείγμα για να μας βοηθήσει να μάθουμε κάτι σχετικά με τον πληθυσμό. Συγκεκριμένα, θέλαμε να μάθουμε εάν ήταν ακόμα πιθανό να μην υπήρχε καμία υποκείμενη διαφορά μεταξύ των δύο μεθόδων και ότι τα αποτελέσματά μας οφείλονταν στην τύχη. Η διαφορά των 14 mm Hg στην ομάδα μας των 98 ατόμων αναφέρεται ως μια **εκτίμηση** της αληθινής μέσης διαφοράς (difference, d) στον πληθυσμό. Το πρόβλημα της επαγωγής χαρακτηριστικών ενός πληθυσμού από ένα δείγμα είναι το κύριο αντικείμενο ενδιαφέροντος της επαγωγικής στατιστικής και αποτελεί το βασικό θέμα του κειμένου αυτού. Για να επιτύχουμε τον σκοπό αυτό, χρειαζόταν να αναπτύξουμε ένα **μοντέλο πιθανοτήτων**, το οποίο θα μας έλεγε πόσο πιθανό είναι να πάρουμε μία διαφορά 14 mm Hg μεταξύ των δύο μεθόδων σε ένα δείγμα 98 ατόμων, εάν δεν υπήρχε πραγματική διαφορά μεταξύ των δύο μεθόδων σε ολόκληρο τον πληθυσμό των χρηστών της συσκευής. Εάν η πιθανότητα αυτή ήταν αρκετά μικρή, τότε θα αρχίζαμε να πιστεύουμε ότι υπήρχε μια πραγματική διαφορά μεταξύ των δύο μεθόδων.

Σε αυτή τη συγκεκριμένη περίπτωση, χρησιμοποιώντας ένα μοντέλο πιθανοτήτων βασισμένο στην κατανομή *t*, συμπεράναμε ότι η πιθανότητα αυτή ήταν λιγότερο από 1 στις 1000 για καθεμία από τις συσκευές στις τοποθεσίες C και D. Η πιθανότητα αυτή ήταν αρκετά μικρή για εμάς ώστε να συμπεράνουμε ότι υπήρχε πραγματική διαφορά μεταξύ της αυτόματης και της χειροκίνητης (ανθρώπινης) μέτρησης αρτηριακής πίεσης για δύο από τις τέσσερις συσκευές που ελέγχθηκαν.

Χρησιμοποιήσαμε ένα στατιστικό πακέτο (package) για να πραγματοποιήσουμε την προηγούμενη ανάλυση δεδομένων. Ένα πακέτο είναι μια συλλογή στατιστικών προγραμμάτων που περιγράφουν δεδομένα και εκτελούν διάφορους στατιστικούς ελέγχους (test) στα δεδομένα. Επί του παρόντος, τα πιο ευρέως χρησιμοποιούμενα στατιστικά πακέτα είναι τα: SAS, SPSS, Stata, R, MINITAB και Excel.

Το τελικό βήμα σε αυτή τη μελέτη, μετά την ολοκλήρωση της στατιστικής ανάλυσης, ήταν η συγκέντρωση των αποτελεσμάτων σε ένα χειρόγραφο που να μπορεί να δημοσιευθεί. Αναπόφευκτα, λόγω του περιορισμένου χώρου, αφαιρέσαμε αρκετό από το υλικό που αναπτύξαμε κατά τη φάση της ανάλυσης δεδομένων και παρουσιάσαμε μόνο τα ουσιαστικά στοιχεία, που ήταν κατάλληλα για δημοσίευση.

Η ανασκόπηση της μελέτης μας για την αρτηριακή πίεση θα πρέπει να σας δώσει μια ιδέα για το τι σημαίνει ιατρική έρευνα και για τον ρόλο της βιοστατιστικής σε αυτή τη διαδικασία. Το υλικό σε αυτό το βιβλίο προχωρά παράλληλα με την περιγραφή της φάσης ανάλυσης δεδομένων της μελέτης. Το Κεφάλαιο 2 συνοψίζει τους διάφορους τύπους περιγραφικής στατιστικής. Στα Κεφάλαια 3 έως 5 παρουσιάζονται ορισμένες βασικές αρχές των πιθανοτήτων και διάφορα μοντέλα πιθανοτήτων για χρήση σε μεταγενέστερα ζητήματα επαγωγικής στατιστικής. Τα Κεφάλαια 6 έως 14 συζητούν τα κύρια θέματα της επαγωγικής στατιστικής όπως εφαρμόζεται στη βιοϊατρική πρακτική. Θέματα σχεδιασμού μελέτης ή συλλογής δεδομένων θίγονται μόνο εφόσον σχετίζονται με άλλα θέματα που συζητούνται στο κείμενο.

Βιβλιογραφικές αναφορές

[1] Polk, B. F., Rosner, B., Feudo, R., & Vandenburgh, M. (1980). An evaluation of the Vita-Stat automatic blood pressure measuring device. *Hypertension, 2*(2), 221–227.

2 | Περιγραφική Στατιστική

2.1 ΕΙΣΑΓΩΓΗ

Όταν κοιτάζουμε δεδομένα, το πρώτο βήμα είναι να περιγράψουμε τα διαθέσιμα στοιχεία με κάποιον συνοπτικό τρόπο. Στις μικρότερες μελέτες αυτό το βήμα μπορεί να επιτευχθεί με την καταγραφή κάθε σημείου δεδομένων. Γενικά, όμως, η διαδικασία αυτή είναι κουραστική ή ανέφικτη και, ακόμα και εάν ήταν εφικτή, δεν θα παρείχε μια συνολική εικόνα του πώς δείχνουν τα δεδομένα.

ΠΑΡΑΔΕΙΓΜΑ 2.1

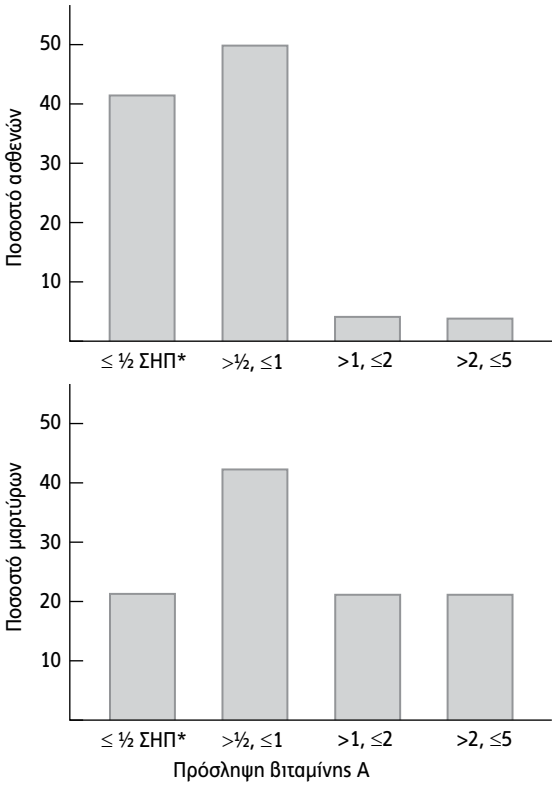
Καρκίνος, Διατροφή Κάποιοι ερευνητές έχουν υποστηρίξει ότι η πρόσληψη βιταμίνης Α προστατεύει από τον καρκίνο. Για να ελέγξουμε τη θεωρία αυτή, μπορούμε να χρησιμοποιήσουμε ένα διατροφικό ερωτηματολόγιο για τη συλλογή δεδομένων αναφορικά με την κατανάλωση βιταμίνης Α μεταξύ 200 νοσηλευόμενων ασθενών με καρκίνο (περιπτώσεις/ασθενείς – cases) και 200 μαρτύρων (controls). Οι μάρτυρες θα είναι αντιστοιχισμένοι αναφορικά με την ηλικία και το φύλο με τους καρκινοπαθείς ασθενείς και θα βρίσκονται στο νοσοκομείο το ίδιο χρονικό διάστημα για μία ασθένεια άσχετη με τον καρκίνο. Τι θα πρέπει να γίνει με αυτά τα δεδομένα μετά τη συλλογή τους;

Προτού γίνει οποιαδήποτε επίσημη απόπειρα να απαντηθεί το συγκεκριμένο ερώτημα, θα πρέπει να περιγραφεί η πρόσληψη βιταμίνης Α μεταξύ των ασθενών και των μαρτύρων. Ας δούμε την Εικόνα 2.1. Τα **ραβδογράμματα** (**bar graphs**) δείχνουν ότι οι μάρτυρες προσλαμβάνουν περισσότερη βιταμίνη Α απ' ό,τι οι ασθενείς, ιδιαίτερα σε επίπεδα κατανάλωσης που υπερβαίνουν τη Συνιστώμενη Ημερήσια Πρόσληψη (ΣΗΠ – RDA: Recommended Daily Allowance).

ΠΑΡΑΔΕΙΓΜΑ 2.2

Πνευμονική Νόσος Ιατρικοί ερευνητές έχουν πολλές φορές υποπτευθεί ότι οι παθητικοί καπνιστές –άτομα που τα ίδια δεν καπνίζουν αλλά ζουν ή εργάζονται σε ένα περιβάλλον στο οποίο οι άλλοι καπνίζουν– μπορεί να παρουσιάσουν ως αποτέλεσμα επιβαρυμένη πνευμονική λειτουργία. Το 1980 μια ερευνητική ομάδα στο Σαν Ντιέγκο δημοσίευσε αποτελέσματα που υποδήλωναν ότι οι παθητικοί καπνιστές είχαν πράγματι σημαντικά μειωμένη πνευμονική λειτουργία συγκριτικά με μη καπνιστές που δεν εργάζονταν σε περιβάλλοντα γεμάτα καπνό [1]. Ως αποδεικτικό στοιχείο, οι συγγραφείς μέτρησαν τις συγκεντρώσεις μονοξειδίου του άνθρακα (CO) στα περιβάλλοντα εργασίας παθητικών καπνιστών και μη καπνιστών των οποίων οι εταιρείες δεν επέτρεπαν το κάπνισμα στον χώρο εργασίας, προκειμένου να διαπιστώσουν εάν η σχετική συγκέντρωση CO άλλαζε κατά τη διάρκεια της ημέρας. Τα αποτελέσματα αυτά παρουσιάζονται στο **διάγραμμα διασποράς** (**scatter plot**) της Εικόνας 2.2.

Η Εικόνα 2.2 (σελ. 31) δείχνει καθαρά ότι οι συγκεντρώσεις CO στα δύο περιβάλλοντα εργασίας είναι περίπου οι ίδιες νωρίς μέσα στη μέρα, αλλά αποκλίνουν ευρέως κατά το μέσο της ημέρας και μετά συγκλίνουν πάλι μετά το τέλος της εργάσιμης ημέρας, στις 7 μ.μ.



ΣΗΠ* (RDA) = Συνιστώμενη Ημερήσια Πρόσληψη.

ΕΙΚΟΝΑ 2.1 Ημερήσια πρόσληψη βιταμίνης Α μεταξύ των ασθενών με καρκίνο και των μαρτύρων

Οι διαγραμματικές απεικονίσεις καταδεικνύουν τον σημαντικό ρόλο των περιγραφικών στατιστικών, ο οποίος είναι να παρέχουν στον ερευνητή μια άμεση εικόνα των δεδομένων αναφορικά με τις κυρίαρχες τάσεις που υπάρχουν σε αυτά, καθώς και μια ιδέα για το εάν θα άξιζε τον κόπο να προβεί σε πιο λεπτομερή περιγραφή των δεδομένων με τη χρήση των μεθόδων της επαγωγικής στατιστικής.

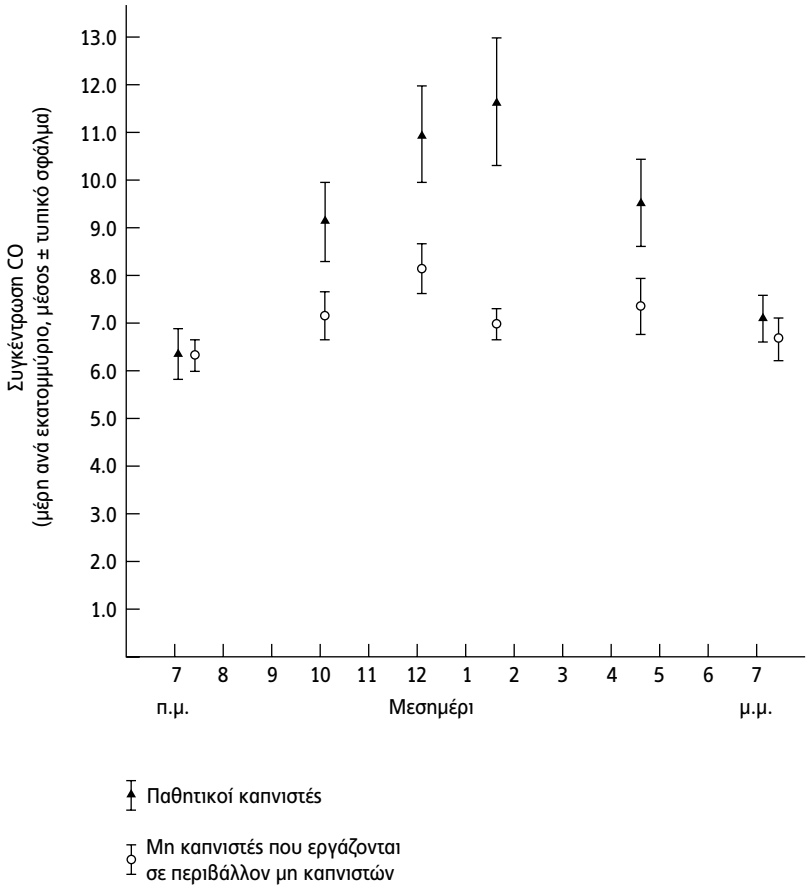
Τα περιγραφικά στατιστικά είναι ιδιαίτερα σημαντικά στη μεταφορά των αποτελεσμάτων των μελετών στις γραπτές δημοσιεύσεις. Οι περισσότεροι ερευνητές δεν θα έχουν τον χρόνο να αξιολογήσουν με κριτικό τρόπο την εργασία των άλλων –εκτός εάν εμπίπτει στο πλαίσιο των βασικών ενδιαφερόντων τους–, αλλά θα επηρεαστούν κυρίως από την παρουσίαση των περιγραφικών στατιστικών.

Τι καθιστά μια διαγραμματική ή αριθμητική απεικόνιση σωστή; Η βασική οδηγία είναι ότι το υλικό θα πρέπει να είναι όσο το δυνατόν πιο αυτόνομο και κατανοητό χωρίς να χρειάζεται να διαβάσει κάποιος το κείμενο. Τα χαρακτηριστικά αυτά απαιτούν σαφή σήμανση. Οι λεζάντες, οι μονάδες μέτρησης και οι άξονες στα διαγράμματα θα πρέπει να δηλώνονται ξεκάθαρα, και οι στατιστικοί όροι που χρησιμοποιούνται στους πίνακες και στα διαγράμματα θα πρέπει να είναι σωστά καθορισμένοι. Η ποσότητα του υλικού που παρουσιάζεται είναι εξίσου σημαντική. Εάν έχουν κατασκευαστεί ραβδογράμματα, τότε θα πρέπει να δοθεί προσοχή ώστε να μην απεικονίζονται ούτε υπερβολικά πολλές ούτε υπερβολικά λίγες ομάδες. Το ίδιο ισχύει και για το πινακοποιημένο υλικό.

Πολλές μέθοδοι είναι διαθέσιμες για τη σύνοψη δεδομένων, τόσο σε αριθμητική όσο και σε διαγραμματική μορφή. Σε αυτό το κεφάλαιο παρουσιάζονται αυτές οι μέθοδοι και επισημαίνονται τα δυνατά και αδύναμα σημεία τους.

2.2 ΜΕΤΡΑ ΘΕΣΗΣ

Το βασικό πρόβλημα της στατιστικής μπορεί να οριστεί ως εξής: Θεωρήστε ένα δείγμα δεδομένων $x_1, \ldots, x_n$, όπου το x_1 αντιστοιχεί στο πρώτο σημείο του δείγματος και το x_n αντιστοιχεί στο n-οστό σημείο του δείγματος.



ΕΙΚΟΝΑ 2.2 Μέση συγκέντρωση μονοξειδίου του άνθρακα (± τυπικό σφάλμα) ανά ώρα της ημέρας όπως μετρήθηκε στο εργασιακό περιβάλλον των παθητικών καπνιστών και των μη καπνιστών που εργάζονται σε ένα περιβάλλον μη καπνιστών.

Πηγή: Βασισμένο σε δημοσίευση του *The New England Journal of Medicine, 302,* 720–723, 1980.

Αν υποθέσουμε ότι το δείγμα λήφθηκε από έναν πληθυσμό P, τι συμπεράσματα μπορούν να εξαχθούν για τον πληθυσμό P από το δείγμα;

Πριν μπορέσει να απαντηθεί αυτό το ερώτημα, τα δεδομένα θα πρέπει να συνοψιστούν με όσο το δυνατόν μεγαλύτερη σαφήνεια – αυτό ισχύει επειδή το πλήθος των δειγματικών σημείων είναι συνήθως μεγάλο και είναι εύκολο να χαθεί η συνολική εικόνα όταν εξετάζει κανείς μεμονωμένα δειγματικά σημεία. Ένα είδος μέτρησης χρήσιμο για τη σύνοψη δεδομένων είναι αυτό που ορίζει το κέντρο, ή το μέσο, του δείγματος. Αυτό το είδος μέτρησης είναι ένα **μέτρο θέσης** (**measure of location**).

Ο αριθμητικός μέσος

Μπορεί να φαίνεται προφανές πώς θα ορίσουμε το «κέντρο» ενός δείγματος, αλλά όσο περισσότερο το σκεφτόμαστε, τόσο λιγότερο προφανές γίνεται. Ας υποθέσουμε ότι το δείγμα αποτελείται από το βάρος που είχαν κατά τη γέννησή τους όλα τα ζώντα βρέφη που γεννήθηκαν σε ένα ιδιωτικό νοσοκομείο στο Σαν Ντιέγκο, στην Καλιφόρνια των ΗΠΑ, κατά τη διάρκεια μίας εβδομάδας. Το δείγμα αυτό παρουσιάζεται στον Πίνακα 2.1.

Ένα μέτρο θέσης για το δείγμα αυτό είναι ο αριθμητικός μέσος (αυτός που στην καθομιλουμένη αποκαλούμε μέσο όρο ή μέση τιμή). Ο αριθμητικός μέσος (ή μέσος ή δειγματικός μέσος) συνήθως συμβολίζεται με $\bar{x}$.

ΠΙΝΑΚΑΣ 2.1 Δείγμα βαρών κατά τη γέννηση (γρ.) ζώντων βρεφών που γεννήθηκαν σε ιδιωτικό νοσοκομείο στο Σαν Ντιέγκο, σε διάστημα 1 εβδομάδας

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

ΟΡΙΣΜΟΣ 2.1

Ο **αριθμητικός μέσος** (**arithmetic mean**) είναι το άθροισμα όλων των παρατηρήσεων διαιρεμένο με το πλήθος των παρατηρήσεων. Με στατιστικούς όρους γράφεται ως:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Το σύμβολο Σ (σίγμα) στον Ορισμό 2.1 είναι το σύμβολο του αθροίσματος. Η έκφραση

$$\sum_{i=1}^{n} x_i$$

είναι απλώς ένας σύντομος τρόπος για να γράψουμε την ποσότητα $(x_1 + x_2 + \ldots + x_n)$.

Εάν a και b είναι ακέραιοι, όπου $a < b$, τότε

$$\sum_{i=a}^{b} x_i$$

σημαίνει $x_a + x_{a+1} + \ldots + x_b$.

Εάν a = b, τότε $\sum_{i=a}^{b} x_i = x_a$. Μία ιδιότητα των συμβόλων αθροίσματος είναι ότι εάν κάθε όρος μέσα στο άθροισμα είναι πολλαπλάσιο της ίδιας σταθεράς c, τότε η c μπορεί να βγει εκτός του αθροίσματος, δηλαδή:

$$\sum_{i=1}^{n} cx_i = c\left(\sum_{i=1}^{n} x_i\right)$$

ΠΑΡΑΔΕΙΓΜΑ 2.3

Αν $x_1 = 2 \quad x_2 = 5 \quad x_3 = -4$

βρείτε τα $\sum_{i=1}^{3} x_i \quad \sum_{i=2}^{3} x_i \quad \sum_{i=1}^{3} x_i^2 \quad \sum_{i=1}^{3} 2x_i$

Λύση:

$$\sum_{i=1}^{3} x_i = 2 + 5 - 4 = 3 \quad \sum_{i=2}^{3} x_i = 5 - 4 = 1$$

$$\sum_{i=1}^{3} x_i^2 = 4 + 25 + 16 = 45 \quad \sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 6$$

Είναι σημαντικό να εξοικειωθείτε με τα σύμβολα της άθροισης, διότι χρησιμοποιούνται εκτεταμένα σε όλο το υπόλοιπο βιβλίο.

ΠΑΡΑΔΕΙΓΜΑ 2.4 Ποιος είναι ο αριθμητικός μέσος για το δείγμα των βαρών γέννησης του Πίνακα 2.1;

Λύση: x = (3265 + 3260 + ... + 2834)/20 = 3166.9 γρ.

Ο αριθμητικός μέσος είναι, γενικά, ένα πολύ λογικό μέτρο θέσης. Παρ' όλα αυτά, ένας από τους βασικούς περιορισμούς του είναι ότι είναι εξαιρετικά ευαίσθητος στις ακραίες τιμές. Στην περίπτωση αυτή, μπορεί να μην είναι αντιπροσωπευτικός της τοποθεσίας της πλειονότητας των δειγματικών σημείων. Για παράδειγμα, εάν το πρώτο βρέφος του Πίνακα 2.1 τύχαινε να ήταν ένα πρόωρο βρέφος που ζύγιζε 500 γρ. αντί για 3265 γρ., τότε ο αριθμητικός μέσος του δείγματος θα έπεφτε στα 3028.7γρ. Στην περίπτωση αυτή, 7 μετρήσεις βάρους θα ήταν μικρότερες από τον αριθμητικό μέσο και 13 θα ήταν υψηλότερες από τον αριθμητικό μέσο. Είναι πιθανό να υπάρξουν ακραίες περιπτώσεις όπου όλα τα δειγματικά σημεία, εκτός από ένα, να βρίσκονται στη μία πλευρά του αριθμητικού μέσου. Σε αυτού του τύπου τα δείγματα, ο αριθμητικός μέσος είναι ένα «κακό» μέτρο κεντρικής θέσης, διότι δεν αντικατοπτρίζει το κέντρο του δείγματος. Ωστόσο, ο αριθμητικός μέσος είναι σαφώς το πιο ευρέως χρησιμοποιούμενο μέτρο κεντρικής θέσης.

Η διάμεσος

Ένα εναλλακτικό μέτρο θέσης, ενδεχομένως το δεύτερο σε δημοτικότητα μετά τον αριθμητικό μέσο, είναι η **διάμεσος** (**median**) ή, για να είμαστε πιο ακριβείς, η **δειγματική διάμεσος** (**sample median**).

Έστω ότι υπάρχουν n παρατηρήσεις στο δείγμα. Εάν οι παρατηρήσεις αυτές ταξινομηθούν από τη μικρότερη στη μεγαλύτερη, τότε η διάμεσος ορίζεται ως εξής:

ΟΡΙΣΜΟΣ 2.2

Η **δειγματική διάμεσος** είναι:

(1) Η $\left(\frac{n+1}{2}\right)$ στη σειρά παρατήρηση, εάν το n είναι περιττός αριθμός

(2) Ο μέσος όρος της $\left(\frac{n}{2}\right)$ και της $\left(\frac{n}{2}+1\right)$ στη σειρά παρατήρησης, εάν το n είναι άρτιος αριθμός.

Η λογική πίσω από αυτούς τους ορισμούς είναι να εξασφαλίσουμε ίσο αριθμό δειγματικών σημείων μεταξύ των δύο πλευρών της δειγματικής διαμέσου. Η διάμεσος ορίζεται διαφορετικά όταν το n είναι άρτιος και όταν είναι περιττός αριθμός, διότι είναι αδύνατον να πετύχουμε τον στόχο με έναν ενιαίο ορισμό. Δείγματα με περιττό πλήθος αριθμών έχουν ένα μοναδικό κεντρικό σημείο– για παράδειγμα, για δείγματα μεγέθους 7, το τέταρτο μεγαλύτερο σημείο είναι το κεντρικό σημείο με την έννοια ότι 3 σημεία είναι μικρότερα από αυτό και 3 σημεία μεγαλύτερα. Δείγματα με άρτιο πλήθος αριθμών δεν έχουν ένα μοναδικό κεντρικό σημείο, οπότε από τις δύο ενδιάμεσες τιμές θα πρέπει να υπολογιστεί ο μέσος όρος. Επομένως, για δείγματα μεγέθους 8, από το τέταρτο και από το πέμπτο μεγαλύτερο σημείο θα πρέπει να υπολογίσουμε τον μέσο για να πάρουμε τη διάμεσο, διότι κανένα από τα δύο δεν είναι το κεντρικό σημείο.

ΠΑΡΑΔΕΙΓΜΑ 2.5 Υπολογίστε τη δειγματική διάμεσο του δείγματος στον Πίνακα 2.1.

Λύση: Πρώτα ταξινομήστε το δείγμα σε αύξουσα διάταξη:

2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

Επειδή το n είναι άρτιο,

Δειγματική διάμεσος = μέση τιμή μεταξύ της 10ης και 11ης μεγαλύτερης παρατήρησης = (3245 + 3248)/2 = 3246.5 γρ.

Για τον υπολογισμό της μέσης τιμής και της διαμέσου, μπορούμε επίσης να χρησιμοποιήσουμε κάποιο υπολογιστικό πακέτο. Για τον σκοπό αυτό χρησιμοποιήσαμε το υπολογιστικό πακέτο R. Πρώτα εισάγουμε τα στοιχεία από τον Πίνακα 2.1 σε ένα λογιστικό φύλλο και το αποθηκεύουμε ως αρχείο κειμένου με τις ονομασίες των μεταβλητών (στην περίπτωση αυτή δύο μεταβλητών, που ονομάζονται id για τον κωδικό ταυτοποίησης και birthwt για το βάρος κατά τη γέννηση) στην πρώτη γραμμή. Εν συνεχεία, διαβάζουμε τα δεδομένα μέσα στην R χρησιμοποιώντας την εντολή read.xlsx και θέτοντας ως όνομα στο σετ δεδομένων το bwt. Μετά χρησιμοποιούμε την εντολή attach για να αναφερθούμε στις μεταβλητές του σετ δεδομένων με το όνομά τους. Χρησιμοποιούμε επίσης την εντολή names για να καθορίσουμε ποια είναι τα ονόματα που υπάρχουν μέσα στο σετ δεδομένων. Τέλος, χρησιμοποιούμε τις εντολές mean και median για τον υπολογισμό του αριθμητικού μέσου και της διαμέσου, αντίστοιχα. Τα αποτελέσματα δίνονται στον Πίνακα 2.2.

ΠΙΝΑΚΑΣ 2.2 Χρήση της R για τον υπολογισμό του μέσου και της διαμέσου των δεδομένων που αφορούν τα βάρη κατά τη γέννηση του Πίνακα 2.1

```
> bwt<-read.xlsx("E:/Rosner/DataCh2/fob.8thedition.
table2.1.xlsx",2, header=TRUE)
> attach(bwt)
> names(bwt)
[1]  "id" "birthwt"
birthwt
[1] 3265 3260 3245 3484 4146 3323 3649 3200 3031 2069 2581 2841 3609 2838
3541
[16] 2759 3248 3314 3101 2834
> mean(birthwt)
[1] 3166.9
> median(birthwt)
[1] 3246.5
```

ΠΑΡΑΔΕΙΓΜΑ 2.6

Λοιμώδης Ασθένεια Θεωρήστε το σετ δεδομένων του Πίνακα 2.3, το οποίο αποτελείται από τις μετρήσεις των λευκών αιμοσφαιρίων που λήφθηκαν κατά την εισαγωγή όλων των ασθενών σε ένα μικρό νοσοκομείο στην πόλη Allentown της Πενσιλβάνια, μια συγκεκριμένη μέρα. Υπολογίστε τη διάμεσο των λευκών αιμοσφαιρίων.

ΠΙΝΑΚΑΣ 2.3 Δείγμα μέτρησης λευκών αιμοσφαιρίων (x 1000) για όλους τους ασθενείς που εισάγονται σε νοσοκομείο στην Allentown της Πενσιλβάνια, μία συγκεκριμένη μέρα

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

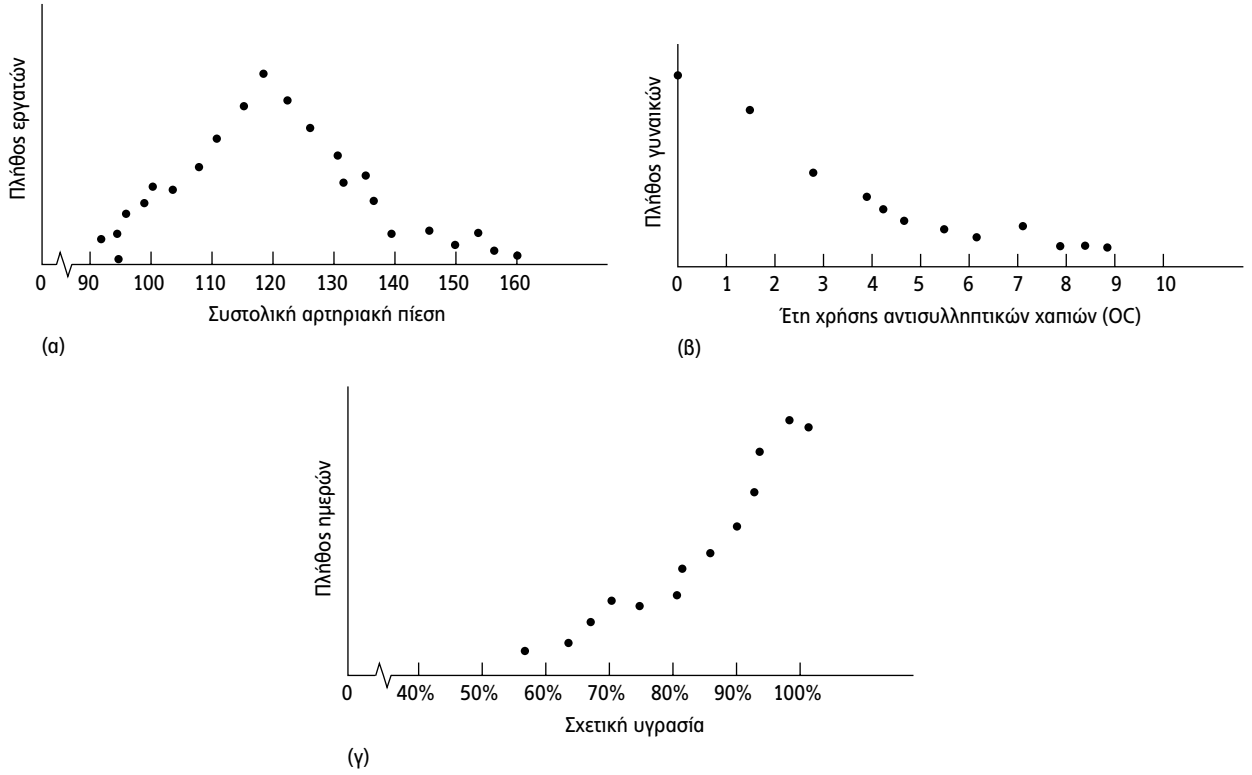
Λύση: Αρχικά, ταξινομήστε το δείγμα ως εξής: 3, 5, 7, 8, 8, 9, 10, 12, 35. Επειδή το n είναι περιττός αριθμός, η δειγματική διάμεσος δίνεται από το 5ο μεγαλύτερο σημείο, το οποίο ισούται με 8 ή 8000 στην αρχική κλίμακα.

Το βασικό πλεονέκτημα της δειγματικής διαμέσου είναι ότι δεν επηρεάζεται από πολύ μεγάλες ή πολύ μικρές τιμές. Συγκεκριμένα, εάν ο δεύτερος ασθενής του Πίνακα 2.3 είχε μια μέτρηση λευκών αιμοσφαιρίων 65.000 αντί για 35.000, η δειγματική διάμεσος θα παρέμενε αμετάβλητη, διότι η πέμπτη μεγαλύτερη τιμή θα εξακολουθούσε να είναι το 8000. Αντίθετα, ο αριθμητικός μέσος θα αυξανόταν σημαντικά από 10.778 του αρχικού δείγματος σε 14.111 στο νέο δείγμα. Η κύρια αδυναμία της δειγματικής διαμέσου είναι ότι καθορίζεται κυρίως από τα ενδιάμεσα σημεία του δείγματος και ότι είναι λιγότερο ευαίσθητη στις πραγματικές τιμές των υπόλοιπων σημείων του δείγματος.

Σύγκριση του αριθμητικού μέσου με τη διάμεσο

Εάν η κατανομή είναι **συμμετρική** (**symmetric**), τότε η σχετική θέση των σημείων σε κάθε πλευρά της δειγματικής διαμέσου είναι η ίδια. Ένα παράδειγμα κατανομής που αναμένουμε να είναι σχετικά συμμετρική είναι η κατανομή των μετρήσεων της συστολικής αρτηριακής πίεσης που λήφθηκαν από όλους τους εργάτες 30-39 ετών ενός συγκεκριμένου εργοστασίου στον χώρο εργασίας (Εικόνα 2.3α).

Εάν μια κατανομή έχει **θετική ασυμμετρία** (**positively skewed** – ασυμμετρία με ουρά δεξιά), τότε τα σημεία πάνω από τη διάμεσο τείνουν να βρίσκονται πιο μακριά από τη διάμεσο κατ' απόλυτη τιμή απ' ό,τι τα σημεία κάτω από τη διάμεσο. Ένα παράδειγμα κατανομής με θετική ασυμμετρία είναι τα έτη χρήσης αντισυλληπτικού χαπιού (oral contraceptive – OC) σε γυναίκες ηλικίας 20-29 ετών (Εικόνα 2.3β). Ομοίως, εάν μία κατανομή έχει **αρνητική ασυμμετρία** (**negatively skewed** –ασυμμετρία με ουρά αριστερά), τότε τα σημεία κάτω από τη διάμεσο τείνουν να βρίσκονται πιο μακριά από τη διάμεσο κατ' απόλυτη τιμή απ' ό,τι τα σημεία πάνω από τη διάμεσο. Ένα παράδειγμα κατανομής με αρνητική συμμετρία είναι οι μετρήσεις της σχετικής υγρασίας που παρατηρείται σε υγρά κλίματα την ίδια ώρα της ημέρας κατά τη διάρκεια μερικών ημερών. Στην περίπτωση αυτή, οι περισσότερες μετρήσεις υγρασίας βρίσκονται πάνω ακριβώς ή κοντά στο 100%, με λίγες χαμηλές μετρήσεις κάποιες ξηρές ημέρες (Εικόνα 2.3γ).



ΕΙΚΟΝΑ 2.3 Διαγραμματική απεικόνιση (α) συμμετρικής κατανομής, (β) κατανομής με θετική ασυμμετρία, (γ) κατανομής με αρνητική ασυμμετρία

Σε πολλά δείγματα, η σχέση μεταξύ του αριθμητικού μέσου και της δειγματικής διαμέσου μπορεί να χρησιμοποιηθεί για την εκτίμηση της συμμετρίας της κατανομής. Συγκεκριμένα, στις συμμετρικές κατανομές ο αριθμητικός μέσος είναι περίπου ίσος με τη διάμεσο. Σε κατανομές με θετική ασυμμετρία, ο αριθμητικός μέσος τείνει να είναι μεγαλύτερος από τη διάμεσο· σε κατανομές με αρνητική ασυμμετρία, ο αριθμητικός μέσος τείνει να είναι μικρότερος από τη διάμεσο.

Η επικρατούσα τιμή

Ένα άλλο, ευρέως χρησιμοποιούμενο μέτρο θέσης είναι η επικρατούσα τιμή.

ΟΡΙΣΜΟΣ 2.3

Η **επικρατούσα τιμή** (**mode**) είναι η τιμή με τη μεγαλύτερη συχνότητα εμφάνισης από όλες τις παρατηρήσεις ενός δείγματος.

ΠΑΡΑΔΕΙΓΜΑ 2.7

Γυναικολογία Θεωρήστε ένα δείγμα χρονικών διαστημάτων (ημερών) μεταξύ δύο διαδοχικών έμμηνων ρύσεων μιας ομάδας 500 φοιτητριών ηλικίας 18-21 ετών, σύμφωνα με τον Πίνακα 2.4. Η στήλη της συχνότητας δίνει το πλήθος των γυναικών που δήλωσαν την αντίστοιχη διάρκεια. Η επικρατούσα τιμή είναι 28 διότι είναι η τιμή με τη μεγαλύτερη συχνότητα εμφάνισης.

ΠΙΝΑΚΑΣ 2.4 Δείγμα με χρονικά διαστήματα μεταξύ διαδοχικών έμμηνων ρύσεων (μέρες) σε νεαρές γυναίκες (φοιτήτριες)

Τιμή	Συχνότητα	Τιμή	Συχνότητα	Τιμή	Συχνότητα
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

ΠΑΡΑΔΕΙΓΜΑ 2.8

Υπολογίστε την επικρατούσα τιμή της κατανομής του Πίνακα 2.3.

Λύση: Η επικρατούσα τιμή είναι 8000, διότι εμφανίζεται πιο συχνά από οποιαδήποτε άλλη τιμή λευκών αιμοσφαιρίων.

Κάποιες κατανομές έχουν περισσότερες από μία επικρατούσες τιμές. Ένας τρόπος διαχωρισμού των κατανομών είναι ανάλογα με το πλήθος των επικρατουσών τιμών (δηλαδή των κορυφών στο διάγραμμα συχνοτήτων) που έχουν. Μια κατανομή με μία επικρατούσα τιμή (κορυφή) ονομάζεται **μονοκόρυφη** (**unimodal**), με δύο επικρατούσες τιμές ονομάζεται **δικόρυφη** (**bimodal**), με τρεις ονομάζεται **τρικόρυφη** (**trimodal**) κ.ο.κ.

ΠΑΡΑΔΕΙΓΜΑ 2.9 Υπολογίστε την επικρατούσα τιμή της κατανομής του Πίνακα 2.1.

Λύση: Δεν υπάρχει επικρατούσα τιμή, καθώς όλες οι τιμές εμφανίζονται ακριβώς μόνο μία φορά.

Το Παράδειγμα 2.9 παρουσιάζει ένα πολύ συχνό πρόβλημα της επικρατούσας τιμής: δεν αποτελεί χρήσιμο μέτρο θέσης εάν υπάρχει μεγάλο πλήθος πιθανών τιμών, καθεμία από τις οποίες εμφανίζεται σπάνια. Σε αυτές τις περιπτώσεις η επικρατούσα τιμή είτε θα βρίσκεται μακριά από το κέντρο του δείγματος, είτε σε ακραίες περιπτώσεις δεν θα υπάρχει, όπως στο Παράδειγμα 2.9.

Ο γεωμετρικός μέσος

Πολλά είδη εργαστηριακών δεδομένων, ιδιαίτερα δεδομένων με τη μορφή συγκεντρώσεων μιας ουσίας μέσα σε μία άλλη, προσδιορισμένων μέσω τεχνικών διαδοχικών αραιώσεων, μπορούν να εκφραστούν είτε ως πολλαπλάσια του 2 είτε ως μία σταθερά πολλαπλασιασμένη με δυνάμεις του 2· με άλλα λόγια, τα αποτελέσματα μπορούν να είναι μόνο της μορφής $2^k c$, $k = 0, 1, \ldots$, για μία σταθερά c. Για παράδειγμα, τα δεδομένα του Πίνακα 2.5 αντιπροσωπεύουν την ελάχιστη ανασταλτική συγκέντρωση (MIC) πενικιλίνης G στα ούρα για τον εντοπισμό γονόρροιας (*N. gonorrhoeae*) σε 74 ασθενείς [2]. Ο αριθμητικός μέσος δεν είναι κατάλληλος ως μέτρο θέσης στη συγκεκριμένη περίπτωση, διότι η κατανομή έχει έντονη ασυμμετρία.

Ωστόσο, τα δεδομένα έχουν ένα συγκεκριμένο μοτίβο διότι οι μοναδικές δυνατές τιμές είναι της μορφής $2^k(0.03125)$ για k= 0, 1, 2, ... Μία λύση είναι να δουλέψουμε με την κατανομή των λογαρίθμων των συγκεντρώσεων. Οι λογάριθμοι των συγκεντρώσεων έχουν την ιδιότητα ότι πιθανές διαδοχικές συγκεντρώσεις διαφέρουν κατά μία σταθερά– με άλλα λόγια, $\log(2^{k+1}c) - \log(2^k c) = \log(2^{k+1}) + \log c - \log(2^k) - \log c = (k + 1)\log 2 - k \log 2 = \log 2$. Επομένως, τα log των συγκεντρώσεων είναι ισομερώς κατανεμημένα το ένα από το άλλο, και η κατανομή που προκύπτει δεν είναι τόσο ασύμμετρη όσο οι ίδιες οι αρχικές συγκεντρώσεις. Ο αριθμητικός μέσος μπορεί να υπολογιστεί ακολούθως σε λογαριθμική κλίμακα, δηλαδή ως:

ΠΙΝΑΚΑΣ 2.5 Κατανομή της ελάχιστης ανασταλτικής συγκέντρωσης (MIC) πενικιλίνης G για τον εντοπισμό γονόρροιας σε 74 ασθενείς

Συγκέντρωση (μg/mL)	Συχνότητα	Συγκέντρωση (μg/mL)	Συχνότητα
$0.03125 = 2^0(0.03125)$	21	$0.250 = 2^3(0.03125)$	19
$0.0625 = 2^1(0.03125)$	6	$0.50 = 2^4(0.03125)$	17
$0.125 = 2^2(0.03125)$	8	$1.0 = 2^5(0.03125)$	3

Πηγή: Βασισμένο στο *JAMA*, *220*, 205–208, 1972.

$$\overline{\log x} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

και να χρησιμοποιηθεί ως μέτρο θέσης. Ωστόσο, συνήθως είναι προτιμότερο να δουλεύουμε στην αρχική κλίμακα παίρνοντας τον αντιλογάριθμο του $\log \bar{x}$ για τον σχηματισμό του γεωμετρικού μέσου, το οποίο οδηγεί στον ακόλουθο ορισμό:

ΟΡΙΣΜΟΣ 2.4

Ο **γεωμετρικός μέσος** (**geometric mean**) είναι ο αντιλογάριθμος του $\log \bar{x}$, όπου

$$\overline{\log x} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

Μπορεί να χρησιμοποιηθεί οποιαδήποτε βάση για τον υπολογισμό λογαρίθμων για τον γεωμετρικό μέσο. Ο γεωμετρικός μέσος είναι ο ίδιος ανεξάρτητα από τη βάση που θα χρησιμοποιηθεί. Το μόνο προαπαιτούμενο είναι ότι οι λογάριθμοι και αντιλογάριθμοι του Ορισμού 2.4 θα πρέπει να έχουν την ίδια βάση. Οι βάσεις που χρησιμοποιούνται συνήθως είναι η βάση του 10 και η βάση του *e*.

ΠΑΡΑΔΕΙΓΜΑ 2.10 **Λοιμώδης Νόσος** Υπολογίστε τον γεωμετρικό μέσο για το δείγμα του Πίνακα 2.5.

Λύση: (1) Για ευκολία, χρησιμοποιήστε τη βάση του 10 για να υπολογίσετε τους λογαρίθμους και τους αντιλογαρίθμους του συγκεκριμένου παραδείγματος

(2) Υπολογίστε

$$\overline{\log x} = \begin{bmatrix} 21\log_{10}(0.03125) + 6\log_{10}(0.0625) + 8\log_{10}(0.125) \\ + 19\log_{10}(0.250) + 17\log_{10}(0.50) + 3\log_{10}(1.0) \end{bmatrix} \Big/ 74 = -0.846$$

(3) Ο γεωμετρικός μέσος = ο αντιλογάριθμος του $-0.846 = 10^{-0.846} = 0.143$.
Ομοίως, μπορούμε να κάνουμε αυτούς τους υπολογισμούς στην R ως εξής:

```
< mic<-na.omit(read.xlsx("E:/Rosner/DataCh2/fob.8thedition.
table2.4.xlsx",1, header=TRUE))
< names(mic)
[1] "Concentration" "Frequency"
< logs<-sum(mic$Frequency*log10(mic$Concentration))/
sum(mic$Frequency)
< geo_mean=10^logs
< geo_mean
[1] 0.1425153
```

2.3 ΟΡΙΣΜΕΝΕΣ ΙΔΙΟΤΗΤΕΣ ΤΟΥ ΑΡΙΘΜΗΤΙΚΟΥ ΜΕΣΟΥ

Θεωρήστε ένα δείγμα $x_1, \ldots, x_n$, το οποίο θα αναφέρεται ως το αρχικό δείγμα. Για τη δημιουργία ενός **μετατοπισμένου δείγματος (translated sample)** $x_1 + c, \ldots, x_n + c$, προσθέστε μία σταθερά c σε κάθε σημείο δεδομένων. Έστω $y_i = x_i + c$, $i = 1, \ldots, n$. Ας υποθέσουμε ότι θέλουμε να υπολογίσουμε τον αριθμητικό μέσο του μετατοπισμένου δείγματος. Μπορούμε να δείξουμε ότι ισχύει η ακόλουθη σχέση:

ΕΞΙΣΩΣΗ 2.1

Εάν $y_i = x_i + c, \quad i = 1, \ldots, n$
τότε $\bar{y} = \bar{x} + c$

Επομένως, για να βρούμε τον αριθμητικό μέσο των y, υπολογίζουμε τον αριθμητικό μέσο των x και προσθέτουμε τη σταθερά c.

Η αρχή αυτή είναι χρήσιμη επειδή μερικές φορές είναι βολικό να αλλάξει η «προέλευση» των δειγμα-