

§5.10 Υπολογισμός Τεταρτηρορίων

Όταν τα δεδομένα δεν είναι οραδοποιημένα σε τάξεις τότε ο υπολογισμός των Q_1 και Q_3 δίνεται όπως και για τη διάμεσο Q_2 προσαρμόζοντας κατάλληλα τα Q_1, Q_3 ώστε μέχρι τις τιμές των Q_1, Q_3 να περιλαμβάνεται αντίστοιχα το $\frac{1}{4}$ και τα $\frac{3}{4}$ των διατεταγμένων παρατηρήσεων. Όταν τα δεδομένα είναι οραδοποιημένα σε τάξεις συχνότητας, χρησιμοποιούμε τους ακόλουθους τύπους:

$$Q_1 = L_i + \frac{\delta_i}{f_i} \left(\frac{N}{4} - F_{i-1} \right) \text{ και}$$

$$Q_3 = L_i + \frac{\delta_i}{f_i} \left(\frac{3N}{4} - F_{i-1} \right)$$

όπου $i = n$ τάξη στα όρια της οποίας περιλαμβάνεται το πρώτο (τρίτο) τεταρτηρόριο

δ_i = το πλάτος της τάξης στην οποία εντοπίζεται το Q_1 (Q_3)

f_i = η απλή συχνότητα της τάξης στην οποία εντοπίζεται το Q_1 (Q_3)

L_i = το κατώτερο όριο της τάξης στην οποία εντοπίζεται το Q_1 (Q_3)

F_{i-1} = η αθροιστική συχνότητα της προηγούμενης τάξης από αυτήν που εντοπίζεται το Q_1 (Q_3).

Παράδειγμα. Χρησιμοποιώντας τα στοιχεία της κατανομής των γεννήσεων κατά ηλικίες της μητέρας (βλέπε §5.8, Πίνακας, σελ. 64) βρίσκουμε τις τιμές των δύο

τεταρτηγορίων ως εξής:

$$\text{Θέση του } Q_1 = \frac{N}{4} = \frac{24802}{4} \approx 6200. \text{ Άρα, το } Q_1 \text{ εντοπίζεται}$$

στην τρίτη τάξη της κατανομής και:

$$\text{Θέση του } Q_3 = \frac{3N}{4} = \frac{3 \cdot 24802}{4} \approx 18601. \text{ Άρα, το } Q_3 \text{ εντο-}$$

πίζεται στην τέταρτη τάξη της κατανομής.

Επομένως, έχουμε:

$$Q_1 = L_3 + \frac{\delta_3}{f_3} \left(\frac{N}{4} - F_2 \right) = 20 + \frac{5}{10338} (6200 - 3570) \\ = 21,27 \text{ έτη}$$

και

$$Q_3 = L_4 + \frac{\delta_4}{f_4} \left(\frac{3N}{4} - F_3 \right) = 25 + \frac{5}{6686} (18601 - 13908) \\ = 28,51 \text{ έτη}$$

Το ενδοτεταρτηγοριακό εύρος της κατανομής των γεννήσεων είναι: $IQR = Q_3 - Q_1 = 28,51 - 21,27 = 7,24$ έτη.

ΚΕΦΑΛΑΙΟ 6 ΜΕΤΡΑ ΥΨΜΑΝΣΗΣ

§6.1 Εισαγωγή

Ένα από τα βασικά γνωρίσματα των στατιστικών δεδομένων είναι η μεταβλητότητα (variability), δηλαδή η διαφοροποίηση των τιμών μιας μεταβλητής μεταξύ των στατιστικών μονάδων. Μικρός βαθμός διαφοροποίησης των τιμών των δεδομένων χαρακτηρίζει ομοιογενή δεδομένα. Ο όρος ομοιογένεια αναφέρεται στην "ομοιότητα" των μετρήσεων. Στην πράξη η περίπτωση πλήρους ομοιογένειας είναι σπανιότητα και αυτό έχει ως συνέπεια να είναι απαραίτητη η μέτρηση του βαθμού μεταβλητότητας η οποία αυξάνεται όσο περισσότερο διαφοροποιούνται

οι τιμές των μετρήσεων μεταξύ τους. Άλλες ονομασίες της μεταβλητότητας είναι διασπορά (dispersion), εξάπλωση (spread) και διασκόρπιση (scatter). Θα ασχοληθούμε με την αριθμητική εκτίμηση της μεταβλητότητας των δεδομένων. Τα συννηθέστερα μέτρα για την μέτρηση της μεταβλητότητας είναι το εύρος, η διακύμανση, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας.

§ 6.2 Εύρος παρατηρήσεων

Το εύρος ενός συνόλου μετρήσεων ορίζεται ως η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής των δεδομένων. Η χρήση του εύρους ως περιγραφικού μέτρου είναι περιορισμένη. Γνωρίζοντας απλά τις τελικές τιμές των παρατηρήσεων δεν μπορούμε να ανταλήσουμε "πληροφορία" γύρω από την "πυκνότητα" με την οποία κατανέμονται οι μετρήσεις εντός του εύρους. Για παράδειγμα, αν πάρουμε όλους τους φοιτητές ενός Πανεπιστημίου με ύψη μεταξύ **175cm** και **190cm** με εξαίρεση ένα φοιτητή ύψους **156cm** και ένα φοιτητή με ύψος **206cm**, το εύρος σ' αυτή την περίπτωση θα δώσει τη διαφορά μεταξύ των δύο ακραίων τιμών ($206 - 156 = 50\text{cm}$) και όχι την "τυπική" διαφορά των **15cm**. Αν, ακόμη, θεωρήσουμε συσχετίσεις με αριθμό παιδιών: 0, 1, 1, 2, 2, 3, 3 και 9 το εύρος των δεδομένων είναι $R=9$ αλλά όλες οι συσχετίσεις εκτός από μία έχουν 0, 1, 2 ή 3 παιδιά. Δηλαδή, το εύρος είναι μεγάλο κάτι που δεν ανταποκρίνεται στη "ρίκη" μεταβλητότητα των δεδομένων. Παρατηρούμε λοιπόν, ότι το εύρος είναι εξαιρετικά "ευαίσθητο" στις ακραίες παρατηρήσεις και αυτό μας οδηγεί

συχνά σε εσφαλμένα συμπεράσματα για τη μεταβλητότητα των δεδομένων. Χρησιμότερα μέτρα συχνοτικής με το εύρος μπορούν να θεωρηθούν το ενδοτεταρτηροριακό εύρος (IQR) που ορίζεται ως: $IQR = Q_3 - Q_1$ και το ενδοδεκατηροριακό εύρος: $D = Q_9 - Q_1$, όπου Q_9 είναι το ένατο και Q_1 το πρώτο δεκατημόριο, αντιστοίχα, δηλαδή με Q_9 συμβολίζουμε την τιμή της μεταβλητής μέχρι την οποία περιλαμβάνεται το 90% των διατεταγμένων τιμών του δείγματος και με Q_1 συμβολίζουμε την τιμή της μεταβλητής μέχρι την οποία περιλαμβάνεται το 10% των διατεταγμένων τιμών του δείγματος, (δηλαδή το Q_1 έχει διαφορετική ερμηνεία από αυτή που έχει στο IQR). Το ενδοτεταρτηροριακό εύρος (IQR) και το ενδοδεκατηροριακό εύρος (D) δεν είναι τόσο "ευαίσθητα" στις αλλαγές του πληθους των παρατηρήσεων και δεν επηρεάζονται ιδιαίτερα από τις ακραίες τιμές των δεδομένων.

§6.3 Μέση Απόκλιση

Ορίζεται ως ο αριθμητικός μέσος των αποχών τιμών των αποκλίσεων των τιμών των δεδομένων από τον αριθμητικό τους μέσο. Για ένα πληθυσμό $x_1, \dots, x_N$ με μέσο μ , η μέση απόκλιση (A_1) είναι:

$$A_1 = \frac{\sum_{i=1}^N |x_i - \mu|}{N} \quad \text{και κατ'αναλογία για δείγμα με}$$

μέγεθος n με δειγματικό μέσο $\bar{x}$, η δειγματική μέση απόκλιση (a_1) είναι: $a_1 = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$. Η μέση από-

η μέση απόκλιση θα βάνει μεγάλες τιμές στην περίπτωση που τα $x_1, \dots, x_n$ βρίσκονται "ραβδιά" από το μέσο τους. Η μέση απόκλιση επηρεάζεται λιγότερο από την παρουσία αιραίων τιμών στο δείγμα από ότι η τυπική απόκλιση που θα δούρε παρακάτω. Αν στον ορισμό της μέσης απόκλισης αντικαταστήσουμε τον αριθμητικό μέσο με τη διάμεσο προκύπτουν οι ακόλουθες εκφράσεις που ισχύουν για πληθυσμό και δείγμα, αντίστοιχα.

$$A_1' = \frac{\sum_{i=1}^N |x_i - M|}{N} \quad \text{και} \quad a_1' = \frac{\sum_{i=1}^n |x_i - m|}{n}$$

Παράδειγμα: Έστω οι μετρήσεις: 9, 10, 12, 18, 26
Ο αριθμητικός μέσος των δεδομένων είναι: $\bar{x} = \frac{75}{5} = 15$
και η διάμεσος είναι η τρίτη τιμή του διατεταγμένου δείγματος, δηλαδή $m = 12$.

Το άθροισμα των απόλυτων αποκλίσεων από το μέσο είναι:

$$\sum_{i=1}^5 |x_i - \bar{x}| = |9 - 15| + |10 - 15| + |12 - 15| + |18 - 15| + |26 - 15| = 28$$

Το άθροισμα των απόλυτων αποκλίσεων από τη διάμεσο είναι:

$$\sum_{i=1}^5 |x_i - m| = |9 - 12| + |10 - 12| + |12 - 12| + |18 - 12| + |26 - 12| = 25$$

Επομένως, $a_1 = \frac{28}{5} = 5,6$ και $a_1' = \frac{25}{5} = 5$

Παράδειγμα Δίνεται η κατανομή 77 μετρήσεων πίας συν-
νεχούς μεταβλητής:

Κλάσεις	w_i	f_i	$f_i w_i - \bar{x} $	$f_i w_i - m $
(1)	(2)	(3)	(4)	(5)
0,5-1,5	1	10	10 1-3,299	10 1-3,175
1,5-2,5	2	15	15 2-3,299	15 2-3,175
2,5-3,5	3	20	20 3-3,299	20 3-3,175
3,5-4,5	4	15	15 4-3,299	15 4-3,175
4,5-5,5	5	10	10 5-3,299	10 5-3,175
5,5-6,5	6	5	5 6-3,299	5 6-3,175
6,5-7,5	7	2	2 7-3,299	2 7-3,175
Σύνολο		77	96,887	95,275

Ο μέσος της κατανομής είναι:

$$\bar{x} = \frac{1 \times 10 + 2 \times 15 + \dots + 7 \times 2}{77} = 3,299$$

Η διάμεσος της κατανομής είναι

$$m = 2,5 + \frac{1}{20} (38,5 - 25) = 3,175$$

Από τα στοιχεία της στήλης (4) έχουμε:

$$a_1 = \frac{\sum_{i=1}^7 f_i | w_i - 3,299 |}{\sum_{i=1}^7 f_i} = \frac{96,887}{77} = 1,258$$

και από τα στοιχεία της στήλης (5) λαμβάνουμε:

$$a'_1 = \frac{\sum_{i=1}^7 f_i | w_i - 3,175 |}{\sum_{i=1}^7 f_i} = \frac{95,275}{77} = 1,237$$