

Μάθημα 13

§ 6.4 Διαύφανση και Τυπική Απόκλιση

Άλλος τρόπος υπολογισμού της συνοχικής διασποράς των δεδομένων γύρω από τον αριθμητικό τους μέσο επιτυγχάνεται αν αντί των αποκλίτων τιμών πάρουμε τα τετράγωνα των αποκλίσεων των μετρήσεων από το μέσο τους. Ορίζονται, μ' αυτόν τον τρόπο δύο στατιστικά μέτρα, η διαύφανση (variance) και η τυπική απόκλιση (standard deviation) του πληθυσμού ή του δείγματος τα οποία είναι τα συχνότερα χρησιμοποιούμενα μέτρα μεταβλητότητας. Η διαύφανση του πληθυσμού (population variance) συμβολίζεται με σ^2 ή $V(X)$ και η δειγματική διαύφανση (sample variance) με s^2 . Η τυπική απόκλιση του πληθυσμού (population standard deviation) και η δειγματική τυπική απόκλιση (sample standard deviation) συμβολίζονται με σ και s αντίστοιχα.

Αν $x_1, \dots, x_N$ πληθυσμός με μέσο μ γνωστό μέσο μ , η διαύφανση σ^2 των x_i ορίζεται ως:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N},$$

δηλαδή ως ο αριθμητικός μέσος των τετράγωνων των αποκλίσεων των τιμών της μεταβλητής από τον αριθμητικό τους μέσο. Η θετική τετραγωνική ρίζα της διαύφανσης ορίζεται ως τυπική απόκλιση του πληθυσμού (σ), δηλαδή:

Μεγάλες τιμές αυτών των μέτρων δηλώνουν μεγάλη μεταβλητότητα των τιμών του πληθυσμού και συνεπώς μεγάλη απόσταση των x_i από τον πληθυσμιακό μέσο.

Από τον αρχικό τύπο της διαύφανσης έχουμε:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{\sum_{i=1}^N (x_i^2 - 2\mu x_i + \mu^2)}{N} = \frac{\sum_{i=1}^N x_i^2 - 2\mu \sum_{i=1}^N x_i + N\mu^2}{N}$$

και επειδή από τον ορισμό του αριθμητικού μέσου είναι $\sum_{i=1}^N x_i = N\mu$

Λαμβάνουμε:

$$\sigma^2 = \frac{\sum_{i=1}^N x_i^2 - 2N\mu^2 + N\mu^2}{N} = \frac{\sum_{i=1}^N x_i^2 - N\mu^2}{N}$$

Τελικά,

$$\sigma^2 = \frac{\sum_{i=1}^N x_i^2}{N} - \mu^2 = \frac{\sum_{i=1}^N x_i^2}{N} - \left(\frac{\sum_{i=1}^N x_i}{N} \right)^2$$
$$= \frac{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2}{N^2}$$

ή απόδοση δίνεται ως:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N} - \mu^2} = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N} - \left(\frac{\sum_{i=1}^N x_i}{N} \right)^2}$$
$$= \sqrt{\frac{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2}{N^2}}$$

Παράδειγμα θεωρούμε τα βάρη 5 ατόρων σε κιλά:

82, 77, 65, 86, 70, θεωρώντας ότι αυτές οι 5 παρατηρήσεις συνιστούν πληθυσμό μεγέθους $N=5$. Έχουμε:

$$\mu = \frac{82 + 77 + 65 + 86 + 70}{5} = 76 \text{ και}$$

$$s^2 = \frac{(82-76)^2 + (77-76)^2 + (65-76)^2 + (86-76)^2 + (70-76)^2}{5} = \frac{82^2 + 77^2 + 65^2 + 86^2 + 70^2}{5} - 76^2 = 58,8 \text{ κιλά.}$$

Η τυπική απόκλιση είναι $\sigma = \sqrt{58,8} = 7,668 \text{ κιλά.}$

Αν, αντί για πληθυσμό έχουμε δείγμα μεγέθους n , η δειγματική απόκλιση υπολογίζεται ως:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right), \text{ όπου}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \text{ ο δειγματικός μέσος. Διαιρούμε τις τετραγωνικές αποκλίσεις διά } n-1 \text{ και όχι διά } n \text{ για θεωρητικούς λόγους}$$

διότι αποδεικνύεται ότι η έκφραση με $n-1$ για το s^2 δίνει καλύτερη εκτίμηση της σ^2 του πληθυσμού από την έκφραση με n . Η τυπική απόκλιση ενός δείγματος $x_1, \dots, x_n$ αποτελεί δείκτη του βαθμού διασποράς των τιμών $x_i, i=1, \dots, n$ γύρω από τον αριθμητικό τους μέσο.

Απαντά στην ερώτηση "πόσο ραβδιά από το κέντρο;" βρίσκονται οι τιμές στο συχνευρισμένο υπό μέγεθος σύνολο δεδομένων. Μικρή ή μεγάλη τυπική απόκλιση (ή διακύμανση) μπορεί από μόνη της να είναι "απομαλυντική" αν έχουμε πλήρη γνώση της ιδέας της μέτρησης των δεδομένων. Τα δύο αυτά μέτρα της διακύμανσης και της τυπικής απόκλισης μπορούν να χρησιμοποιηθούν για συχρίσεις μεταξύ δειγμάτων ή πληθυσμών στην περίπτωση που οι παρατηρήσεις αναφέρονται σε νοι-

νές υγιήρεις πέτρους.

§ 6.5 Ιδιότητες της Διακύμανσης

1. Η διακύμανση μίας σταθερής μεταβλητής είναι ίση με μηδέν δηλαδή αν $x_1 = x_2 = \dots = x_N = a \Rightarrow \sigma^2 = 0$

Απόδειξη Αν $x_1 = x_2 = \dots = x_N = a$ τότε $\mu = a$. Άρα

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N (x_i - a)^2 = \frac{1}{N} \sum_{i=1}^N (a - a)^2 = 0.$$

2. Αν η μεταβλητή Y είναι γραμμικός συνδυασμός της μεταβλητής X , δηλαδή αν $Y = a + bX$ με $a, b \in \mathbb{R}$ τότε είναι $V(Y) = V(a + bX) = b^2 V(X)$ όπου $V(X), V(Y)$ οι διακυμάνσεις των X και Y , αντίστοιχα.

Απόδειξη Αν $Y = a + bX$ με $a, b \in \mathbb{R}$ τότε $\mu_Y = a + b\mu_X$, όπου

μ_X, μ_Y οι μέσοι των X, Y αντίστοιχα. Άρα, η διακύμανση

της μεταβλητής Y θα είναι:

$$\begin{aligned} V(Y) &= V(a + bX) = \frac{1}{N} \sum_{i=1}^N (y_i - \mu_Y)^2 \\ &= \frac{1}{N} \sum_{i=1}^N [(a + bx_i) - (a + b\mu_X)]^2 = b^2 \frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)^2 \\ &= b^2 V(X). \end{aligned}$$

Για ειδικές τιμές των a και b έχουμε τα ακόλουθα αποτελέσματα:

$$\text{Αν } b = \pm 1 \Rightarrow Y = a \pm X \Rightarrow V(Y) = V(a \pm X) = V(X)$$

$$\text{Αν } a = 0 \Rightarrow Y = bX \Rightarrow V(Y) = V(bX) = b^2 V(X)$$

$$\text{Αν } a = 0, b = \pm 1 \Rightarrow Y = \pm X \Rightarrow V(Y) = V(X)$$

$$\text{Αν } a = 0, b = \frac{1}{\delta} \Rightarrow Y = \frac{1}{\delta} X \Rightarrow V(Y) = V\left(\frac{X}{\delta}\right) = \frac{1}{\delta^2} V(X)$$

3. Το άθροισμα των τετραγώνων των αποκλίσεων μιας μεταβλητής X από δοθέντα παρατηρητικό αριθμό a γίνεται ελάχιστο, αν $a = \bar{x}$.

Απόδειξη Έστω $\phi(x, a) = \sum_{i=1}^N (x_i - a)^2$. Τότε

$$\frac{\partial \phi(x, a)}{\partial a} = -2 \sum_{i=1}^N (x_i - a)$$

Εξισώνοντας με το μηδέν βρίσκουμε:

$$\sum_{i=1}^N (x_i - a) = 0 \Rightarrow \sum_{i=1}^N x_i = Na \Rightarrow a = \frac{\sum_{i=1}^N x_i}{N} = \bar{x}, \text{ δηλαδή}$$

για $a = \bar{x}$ η $\phi(x, a)$ γίνεται ελάχιστη. Γι' αυτό το λόγο η διακύμανση $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$ αποτελεί την ελάχιστη μέση τετραγωνική απόκλιση.

Παράδειγμα Η μεταβλητή X έχει μέσο $\mu_X = 5$ και διακύμανση $\sigma_X^2 = 3$. Η μεταβλητή $Y = 4 - 0,5X$ έχει μέσο

$$\mu_Y = 4 - 0,5\mu_X = 4 - 0,5 \times 5 = 1,5 \text{ και διακύμανση}$$

$$\sigma_Y^2 = (0,5)^2 \sigma_X^2 = 0,25 \times 3 = 0,75 \text{ και τυπική απόκλιση}$$

$$\sigma_Y = \sqrt{0,75} = 0,866.$$