Αριθμητικά Μέτρα

Τα αριθμητικά μέτρα συνοψίζουν την πληροφορία μιας κατανομής συχνοτήτων (άρα και ενός συνόλου δεδομένων) χρησιμοποιώντας κάποιες αριθμητικές τιμές. Τα μέτρα θέσης περιγράφουν το "κέντρο" γύρω από το οποίο είναι συγκεντρωμένες οι παρατηρήσεις, ενώ τα μέτρα διασποράς (ή αλλιώς μέτρα μεταβλητότητας) το πόσο συγκεντρωμένες είναι οι παρατηρήσεις γύρω από το "κέντρο". Τα μέτρα ασυμμετρίας μετράνε αν οι παρατηρήσεις είναι συγκεντρωμένες περισσότερο δεξιά ή αριστερά από το "κέντρο". Τα μέτρα κύρτωσης μετράνε πόσο συγκεντρωμένες είναι οι παρατηρήσεις στα άκρα σε σχέση με το "κέντρο".

Το σύμβολο της πρόσθεσης, $\sum$

Έστω ότι έχουμε τις παρατηρήσεις $x_1, x_2, ..., x_n$.

- Τότε $\sum_{i=1}^{n} x_i$ συμβολίζει το άθροισμα $x_1 + x_2 + ... + x_n$.
- π.χ. $x_1 = 3$, $x_2 = 4$, $x_3 = 1$, $x_4 = -10$, $x_5 = 0$. Τότε $\sum_{i=1}^{n} x_i = \sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 3 + 4 + 1 - 10 + 0 = -2$.
- π.χ. $\sum_{i=1}^{6} i = 1 + 2 + 3 + 4 + 5 + 6 = 21$. (Σε αυτό το παράδειγμα $x_i = i$)
- π.χ. $\sum_{i=1}^{5} i^2 = 1 + 4 + 9 + 16 + 25 = 55$. (Σε αυτό το παράδειγμα $x_i = i^2$)

Κανόνες άθροισης

- Για οποιαδήποτε σταθερά c έχουμε $\sum_{i=1}^{n} c = n \cdot c$.
- Για οποιαδήποτε σταθερά c έχουμε $\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i$.
- Επίσης, αν $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ τότε θα έχουμε $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$.

Μέτρα Θέσης

- Αριθμητικός μέσος (μέσος όρος ή μέση τιμή)
- Επικρατούσα τιμή/κλάση
- Διάμεσος

- Τεταρτημόρια

Ο **αριθμητικός μέσος** ενός συνόλου παρατηρήσεων $x_1, x_2, \ldots, x_n$ είναι

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Επίσης, αν τα δεδομένα είναι διακριτά και αν $y_1, y_2, \ldots, y_k$ οι διαφορετικές (διακριτές) τιμές που παίρνει η μεταβλητή x_i (ποσοτικά διακριτά δεδομένα) τότε

$$\bar{x} = \frac{y_1 v_1 + y_2 v_2 + \ldots + y_k v_k}{n} = \frac{\sum_{i=1}^{k} y_i v_i}{n}$$

ή

$$\bar{x} = y_1 f_1 + y_2 f_2 + \ldots + y_k f_k = \sum_{i=1}^{k} y_i f_i$$

όπου v_i, f_i η συχνότητα και η σχετική συχνότητα της τιμής y_i.
Έστω, για παράδειγμα, η βαθμολογία 10 μαθητών σε ένα μάθημα:
13 13 14 15 15 15 15 16 16 18.
Ο αριθμητικός μέσος σύμφωνα με τον πρώτο τύπο θα είναι

$$\bar{x} = \frac{13 + 13 + 14 + 15 + 15 + 15 + 15 + 16 + 16 + 18}{10} = \frac{150}{10} = 15.$$

Εναλλακτικά αν έχουμε φτιάξει την κατανομή συχνοτήτων των δεδομένων

Βαθμός y_i	Συχνότητα v_i	Σχετ. Συχνότητα f_i	Αθρ. σχετ. συχνότητα F_i	$y_i v_i$	$y_i f_i$
13	2	0,2	0,2	26	2,6
14	1	0,1	0,3	14	1,4
15	4	0,4	0,7	60	6
16	2	0,2	0,9	32	3,2
18	1	0,1	1	18	1,8
Σύνολο	10	1		$\sum_{i=1}^{k} y_i v_i = 150$	$\sum_{i=1}^{k} y_i f_i = 15$

μπορούμε να τον υπολογίσουμε ως:

$$\bar{x} = \frac{\sum_{i=1}^{k} y_i v_i}{n} = \frac{150}{10} = 15$$

ή

$$\bar{x} = \sum_{i=1}^{k} y_i f_i = 13\frac{2}{10} + 14\frac{1}{10} + 15\frac{4}{10} + 16\frac{2}{10} + 18\frac{1}{10} = 15.$$

Ο **διάμεσος** είναι η τιμή για την οποία το 50% των παρατηρήσεων $x_1, x_2, ..., x_n$ έχουν τιμή μικρότερη ή ίση και το (άλλο) 50% των παρατηρήσεων έχουν τιμή μεγαλύτερη ή ίση. Για να τον υπολογίσουμε φτιάχνουμε την διατεταγμένη λίστα των παρατηρήσεων ταξινομώντας τις από μικρότερη σε μεγαλύτερη

και παίρνουμε την μεσαία αν το πλήθος των παρατηρήσεων είναι περιττός αριθμός και τον μέσο όρο των δύο μεσαίων παρατηρήσεων αν το πλήθος είναι άρτιος αριθμός.
π.χ. 1 $\{12, 3, 4, 1, -10, 0, 5, 5, 10\}$ σε αύξουσα σειρά $(-10, 0, 1, 3, 4, 5, 5, 10, 12)$ οπότε ο διαμεσος είναι το 4. Δηλαδή είναι η παρατήρηση $(n+1)/2$ στην διατεταγμένη λίστα όταν n περιττός.
π.χ. 2 $\{3, 4, 1, -10, 0, 5, 5, 10\}$ σε αύξουσα σειρά $(-10, 0, 1, 3, 4, 5, 5, 10)$ οπότε ο διαμεσος είναι $\frac{3+4}{2} = 3.5$. Μέσος όρος παρατηρήσεων $n/2$ και $n/2+1$ στη διατεταγμένη λίστα όταν n άρτιος.
Η **επικρατούσα τιμή** είναι η παρατήρηση x_i η οποία έχει την μεγαλύτερη συχνότητα στα δεδομένα. Δεν είναι απαραίτητα μοναδική.
π.χ. 1 $\{1, 4, 7, 3, 3, 8, 8, 3, 7, 5, 3\}$ έχει επικρατούσα τιμή το 3.
π.χ. 2 $\{1, 4, 7, 3, 3, 8, 8, 7, 5\}$ έχει επικρατούσες τιμές $3, 7, 8$.
Έχει νόημα στα ποσοτικά διακριτά δεδομένα. Στα συνεχή θα μπορούσαμε να σκεφτούμε επικρατούσα κλάση, η οποία θα είναι η κλάση με την μεγαλύτερη συχνήτητα.
Τα **τεταρτημόρια** Q_1, Q_2, Q_3 χωρίζουν τις παρατηρήσεις -οργανωμένες σε αύξουσα σειρά- σε 4 ίσα (ή σχεδόν ίσα) μέρη. Έχουμε ότι Q_i για $i = 1, 2, 3$ είναι η παρατήρηση που βρίσκεται στην θέση $\lceil i \times 0.25 \times n \rceil$ της διατεταγμένης λίστας, όπου n ο αριθμός των παρατηρήσεων. Ο συμβολισμός $\lceil x \rceil$ συμβολίζει τον κοντινότερο στο x ακέραιο που είναι μεγαλύτερος ή ίσος του x. Αν το πλήθος των παρατηρήσεων είναι περιττός αριθμός τότε Q_2 είναι ο διάμεσος.

Μέτρα Διασποράς

- εύρος
- διακύμανση και τυπική απόκλιση
- συντελεστής μεταβλητότητας
- ενδοτεταρτημοριακό εύρος

Το **εύρος** είναι διαφορά μεταξύ της μεγαλύτερης και μικρότερης τιμής των δεδομένων.

$$R = \max\{x_1, x_2, \ldots, x_n\} - \max\{x_1, x_2, \ldots, x_n\}.$$

Είναι ευαίσθητο σε ακραίες τιμές και αγνοεί όλες τις ενδιάμεσες τιμές.
π.χ. $\{12, 3, 4, 1, -10, 0, 5, 5, 10\}$, το εύρος είναι $10 - (-10) = 20$.
Το ενδοτεταρτημοριακό εύρος ορίζεται ως:

$$Q = Q_3 - Q_1$$

Η **διακύμανση** (σ^2 ή s^2) είναι ο μέσος όρος των τετραγωνικών αποκλίσεων από τον "μέσο" :

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}.$$

Δείχνει πόσο συγκεντρωμένες είναι οι τιμές γύρω από τον αριθμητικό μέσο.
Έχουμε ότι $(x_i - \bar{x})^2 = x_i^2 + \bar{x}^2 - 2x_i\bar{x}$ οπότε $\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n} = \frac{\sum_{i=1}^n x_i^2}{n} + \bar{x}^2 - 2\bar{x}^2 = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2$.
Τυπική απόκλιση (σ ή s) είναι η ρίζα της διακύμανσης οπότε μετράται στις ίδιες μονάδες μέτρησης με τα δεδομένα (άρα και τον αριθμητικό μέσο). Αν πολλαπλασιάσουμε κάθε παρατήρηση με c το s θα πολλαπλασιαστεί με c. Είναι ευαίσθητη στις ακραίες τιμές.
Ο **συντελεστής μεταβλητότητας**

$$CV = \frac{\text{Τυπική Απόκλιση}}{\text{Αριθμητικός Μέσος}} = \frac{s}{|\bar{x}|}$$

είναι ένα μέτρο διασποράς ανεξάρτητο από την μονάδα μέτρησης των δεδομένων. Είναι δηλαδή μέτρο της σχετικής μεταβλητότητας των τιμών και όχι της απόλυτης μεταβλητότητας όπως τα προηγούμενα μέτρα.

Μέτρα Ασυμμετρίας και Κύρτωσης

Η κατανομή των δεδομένων λέμε ότι είναι συμμετρική όταν οι τιμές είναι συγκεντρωμένες με συμμετρικό τρόπο (δεξιά και αριστερά) γύρω από τον αριθμητικό μέσο. Θετική ασυμμετρία έχουμε όταν περισσότερες τιμές είναι συγκεντρωμένες δεξιά από τον αριθμητικό μέσο και αρνητική συμμετρία σε αντίθετη περίπτωση. Ένα μέτρο που δείχνει αν μια κατανομή είναι ασυμμετρική είναι το παρακάτω, που παίρνει θετικές και αρνητικές τιμές. Αν είναι > 0 (< 0) τότε έχουμε θετική (αρνητική) ασυμμετρία που σημαίνει ότι οι παρατηρήσεις x_i που είναι μεγαλύτερες (μικρότερες) από τον (αριθμητικό) μέσο είναι ποσοτικά κυριάρχες στο σύνολο των παρατηρήσεων.

$$S_3 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^3$$

Η κύρτωση δείχνει πόσο μεγάλη είναι η συγκέντρωση των δεδομένων στις ουρές τις κατανομής, δηλαδή πόσο παχιές είναι οι ουρές μιας κατανομής. Ένα μέτρο είναι το παρακάτω:

$$S_4 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^4.$$

Τα παραπάνω μέτρα εξαρτώνται από τις μονάδες μέτρησης. Συγκεκριμένα αν πολλαπλασιάσουμε κάθε παρατήρηση με μια σταθερά c το μέτρο ασυμμετρίας θα πολλαπλασιαστεί με c^3, το μέτρο κύρτωσης με c^4 κ.ο.κ. Μπορούμε να χρησιμοποιήσουμε ωστόσο νέα μέτρα που δεν εξαρτώνται από τις μονάδες μέτρησης. Συγκεκριμένα έχουμε τους ακόλουθους συντελεστές ασυμμετρίας και κύρτωσης αντίστοιχα:

$$a = \frac{S_3}{s^3}, \quad b = \frac{S_4}{s^4}$$

Ο συντελεστής κύρτωσης b συγκρίνεται με το 3 που αντιστοιχεί στον συντελεστή που προκύπτει από δεδομένα που προέρχονται από την κανονική κατανομή που θα δούμε αργότερα (για την οποία $a = 0$ και $b = 3$). Όταν $b > 3$ η κατανομή των δεδομένων λέμε ότι είναι λεπτόκυρτη, ενώ όταν $b < 3$ λέμε ότι είναι πλατύκυρτη.