MEM-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

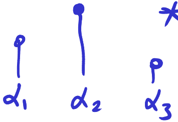
Κώστας Σμαραγδάκης (kesmarag@gmail.com)

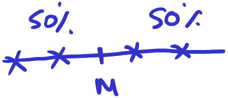
3η εβδομάδα (διάλεξη θεωρίας)

Μέτρα κεντρικής τάσης

- Μέση τιμή $\bar{x}$ $= \frac{1}{N}\sum_{n=1}^{N} x_n$
- Διάμεσος M
- Γεωμετρικός μέσος G $(x_1 \cdots x_n)^{1/n}$
- Επικρατέστερη τιμή M_0

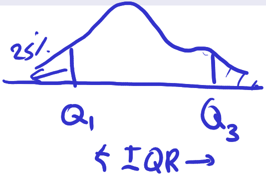
$$\bar{x} = \frac{\sum w_n x_n}{\sum w_n}$$

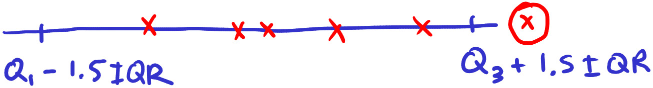




Μέτρα μεταβλητότητας

- Εύρος R
- Ενδοτεταρτημοριακό εύρος IQR $= Q_3 - Q_1$





$x_1, x_2, \ldots, x_N$ $\quad$ $\bar{x} = \frac{1}{N} \sum x_n$

- Μέση τιμή δείγματος: $\bar{x}$
- Μέση τιμή πληθυσμού: μ

$\bar{x} \xrightarrow[N \to \infty]{} \mu$

Έστω $x_1, x_2, \ldots, x_N$ παρατηρήσεις που αντιστοιχούν σε ένα τυχαίο δείγμα ενός πληθυσμού.
Έχουμε ορίσει ως μέση τιμή των παρατηρήσεων του δείγματος την ποσότητα:

$$\bar{x} = 1/N \sum_{n=1}^{N} x_n$$

Αυτή η μέση τιμή εκφράζει μόνο το δείγμα και όχι τον πληθυσμό, αν και για μεγάλο N προσεγγίζει την αντίστοιχη μέση τιμή μ του πληθυσμού.

Ανεξάρτητα των τιμών του δείγματος ισχύει η ανισότικη σχέση

$x_1, \dots, x_N$

$$\sum_{n=1}^{N}(x_n - \bar{x})^2 \leq \sum_{n=1}^{N}(x_n - \mu)^2$$

με ισότητα μόνο αν $\bar{x} = \mu$.

ορίζουμε

$$f(y) = \sum_{n=1}^{N}(x_n - y)^2$$

$$f'(y) = -2\sum_{n=1}^{N}(x_n - y)$$

ακρότατο αν $f'(y) = 0$

$$\sum_{n=1}^{N}(x_n - y) = 0 \Rightarrow \sum_{n=1}^{N} x_n = \sum_{n} y = Ny$$

$$y = \frac{\sum x_n}{N} = \bar{x}$$

$f''(y) = 2 > 0$ ελάχιστο.

$$\sum_{n}^{N}(x_n - \bar{x})^2 \leq \sum_{n=1}^{N}(x_n - \xi)^2 \quad \forall \xi \in \mathbb{R}$$

Παράδειγμα

Έστω το πείραμα της ρίψης ενός αμερόληπτου ζαριού.

$$\mu = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$

Ρίχνουμε το ζάρι 3 φορές και λαμβάνουμε τα αποτελέσματα: 3,2,6
Έχουμε $\bar{x} = 3.66$

$$\sum_{i=1}^{3} (x_i - \bar{x})^2 = 8.66 < 8.75 = \sum_{i=1}^{3} (x_i - \mu)^2$$

Διασπορά πληθυσμού

Ορίζεται ως η μέση τιμή του συνόλου τιμών

$$\{(x_n - \mu)^2\} \qquad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2$$

για κάθε παρατήρηση x του πληθυσμού. Η διασπορά του πληθυσμού συμβολίζεται με σ^2.

Διασπορά στατιστικού δείγματος

$$x_n^2 - 2x_n\bar{x} + (\bar{x})^2$$

$$s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2 \qquad \sum x_n^2 - 2\bar{x}\sum x_n + N(\bar{x})^2$$

$$= \sum x_n^2 - 2\frac{1}{N}\left(\sum x_n\right)^2 + \cancel{N}\frac{1}{N^2}\left(\sum x_n\right)^2$$

Μπορούμε να γράψουμε ισοδύναμα:

$$= \sum x_n^2 - \frac{1}{n}\left(\sum x_n\right)^2$$

$$s^2 = \frac{\sum_{n=1}^{N} x_n^2 - \frac{\left(\sum_{n=1}^{N} x_n\right)^2}{N}}{N-1}$$

Όσο το N αυξάνεται έχουμε $s^2 \to \sigma^2$.

Διασπορά στατιστικού δείγματος

$$\frac{1}{N}\sum_n (x_n-\bar{x})^2 \le \underbrace{\frac{1}{N}\sum_n (x_n-\mu)^2}_{\sigma^2}$$

$$s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

Γιατί διαιρούμε με $N-1$ και όχι απλά με N;

$$\bar{x} \neq \mu \qquad \frac{1}{N}\sum_n (x_n-\bar{x})^2 < \sigma^2$$

$$x_1, \dots, x_N \qquad \bar{x} = \frac{1}{N}\sum_n^N x_n \Rightarrow \bar{x} = \frac{1}{N}\sum_{n=1}^{N-1} x_n + \frac{x_N}{N}$$

$$\boxed{x_N = N\bar{x} - \sum_{n=1}^{N-1} x_n}$$

δειγματική μέση τιμή

$N-1$ παρατηρήσεις

$\boxed{1-3} \rightarrow m_1 = 2$

$\boxed{4-6} = m_2 = 5$

$\boxed{7-9} \quad m_3 = 8$

Διασπορά ομαδοποιημένων δεδομένων

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{K} f_j(m_j - \bar{x})^2$$

Μπορούμε να γράψουμε ισοδύναμα:

$$s^2 = \frac{\sum_{j=1}^{K} m_j^2 f_j - \frac{(\sum_{j=1}^{K} m_j f_j)^2}{N}}{N-1}$$

$$s^2 = \frac{\sum_{j=1}^{K} m_j^2 f_j - \frac{\left(\sum_{j=1}^{K} m_j f_j\right)^2}{N}}{\underbrace{N-1}_{19}}$$

$\sum m_j^2 f_j$ $\sum m_j f_j$

Άσκηση - Διασπορά ομαδοποιημένων δεδομένων

	f	m	mf	m^2	$m^2 f$
[0,2)	3	1	3	1	3
[2,4)	4	3	12	9	27
[4,6)	5	5	25	25	
[6,8)	2	7	14	49	
[8,10)	4	9	36	81	
[10,12)	2	11	22	121	
Total	20		$\sum m_j f_j$		$\sum m_j^2 f_j$

Τυπική Απόκλιση (Standard Deviation)

$[x_i] = m \qquad \bar{x} = \frac{1}{N}\sum x_i \qquad [\bar{x}] = m \qquad s^2 = \frac{1}{N-1}\sum (x_i - \bar{x})^2 \qquad [s^2] = m^2$

$[s] = m$

Αποτελεί το πιο συχνά χρησιμοποιούμενο μέτρο μεταβλητότητας. Ορίζεται ως η τετραγωνική ρίζα της διασποράς.

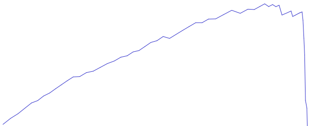
- Τυπική απόκλιση πληθυσμού:

$$\sigma = \sqrt{\sigma^2}$$

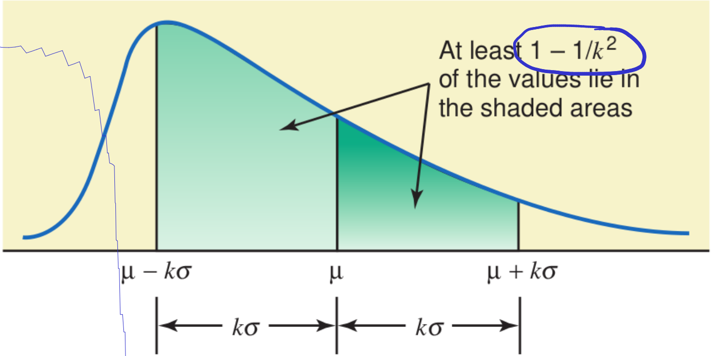
- Τυπική απόκλιση δείγματος:

$$s = \sqrt{s^2}$$

Η τυπική απόκλιση εκφράζεται στην ίδια μονάδα μέτρησης με τη μεταβλητή που αναφέρεται.



Για κάθε $k > 1$, τουλάχιστον $(1 - 1/k^2)$ των παρατηρήσεων ανοίκουν στο διάστημα $[\mu - k\sigma, \mu + k\sigma]$



Τουλάχιστον $\left(1-\frac{1}{k^2}\right)\cdot 100\%$ $[\mu - k\sigma, \mu + k\sigma]$

Άσκηση

Η μέση συστολική αρτηριακή πίεση 4000 γυναικών που υποβλήθηκαν σε εξέταση για υψηλή πίεση αίματος βρέθηκε να είναι 187 mm Hg με τυπική απόκλιση 22. Χρησιμοποιώντας το Θεώρημα του Chebyshev βρείτε το ελάχιστο ποσοστό των γυναικών αυτής της ομάδας με συστολική αρτηριακή πίεση μεταξύ 143 και 231 mm Hg.

$\mu = 187$ mm Hg $\quad \sigma = 22$

$[143, 231] = [187-44, 187+44] = [187-2\cdot 22, 187+2\cdot 22]$

$k=2$ Τουλάχιστον το $\left(1-\frac{1}{4}\right)\cdot 100\% = 75\%$ θα έχουν πίεση $\in [143, 231]$

- Είναι το πηλίκο της τυπικής απόκλισης δια της μέσης τιμής. Συμβολίζεται ως CV:

$$\mathrm{CV} = \frac{s}{\bar{x}}$$

- Είναι χρήσιμος για τη σύγκριση της ομοιογένειας δυο συσχετισμένων μεταβλητών με διαφορετικές μονάδες μέτρησης ή στο να συγκρίνουμε την ομοιογένεια μεταβλητών με ίδιες μονάδες μέτρησης αλλά με διαφορετικές μέσες τιμές.
- Επίσης χρησιμοποιείται για το χαρακτηρισμό ένος δείγματος ως αν ομοιογενές ($\mathrm{CV} \geq 0.1$) ή ~~α~~νομοιογενές ($\mathrm{CV} < 0.1$) .

Παράδειγμα

Έστω δείγματα με τις ημερήσιες μετρήσεις θερμοκρασίας 2 πολέων στη διάρκεια ενός έτους. Για την πόλη A η μέση θερμοκρασία ήταν 20 βαθμούς ^oC και η τυπική απόκλιση 2, ενώ για την B η μέση θερμοκρασία ήταν 15 βαθμούς ^oC και η τυπική απόκλιση 1.8

A: $\mu_A = 20^\circ C$ $\quad \sigma_A = 2^\circ C$ $\quad CV_A = \frac{\sigma_A}{\mu_A} = \frac{2}{20} = \frac{1}{10} = 0.1$ ✓

B: $\mu_B = 15^\circ C$ $\quad \sigma_B = 1.8^\circ C$ $\quad CV_B = \frac{\sigma_B}{\mu_B} = \frac{1.8}{15} = 0.12$

Παράδειγμα

Σε δυο γραπτές δοκιμασίες οι μαθητές μιας τάξης είχαν επιδόσεις που περιγράφονται παρακάτω:
δοκιμασία A (κλίμακα 0-20): μέση τιμή 14, τυπική απόκλιση 1.4
δοκιμασία B (κλίμακα 0-100): μέση τιμή 70, τυπική απόκλιση 3.5

$\mu_A = 14 \quad \sigma_A = 1.4$

$\mu_B = 70 \quad \sigma_B = 3.5$

$$CV_A = \frac{\sigma_A}{\mu_A} = \frac{1.4}{14} = 0.1 > CV_B = \frac{\sigma_B}{\mu_B} = \frac{3.5}{70} = 0.05$$