ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

13-02-2020

Έστω παρατηρήσεις μιας μεταβλητής X. Ο γεωμετρικός μέσος G ορίζεται ως:

$$G = (x_1 \cdot x_2 \dots x_N)^{1/N}$$

Χρησιμοποιείται κυρίως σε οικονομικά και επιχειρηματικά προβλήματα για την μελέτη των ρυθμών μεταβολής οικονομικών μεγεθών με το χρόνο.
Τις περισσότερες φορές είναι ευκολότερο να υπολογίσουμε τον λογάριθμο του G.

$$\log G = \frac{1}{N} \sum_{n=1}^{N} \log x_n$$

Παράδειγμα

Να βρεθεί ο γεωμετρικός μέσος των παρατηρήσεων:
14, 5, 10, 20, 1

$$\log G = \frac{1}{5}\Big(\log(14) + \log(5) + \log(10) + \log(20) + \log(1)\Big) = \frac{4.146128}{5} = 0.829226$$

$$G = 10^{0.829226} = 6.748785$$

Έστω x_0 ένα αρχικό κεφάλαιο και x_j, $j = 1, \ldots, N$ το κεφάλαιο μετά από j έτη. Έστω επίσης ότι κάθε έτος έχουμε διαφορετικό επιτόκιο r_j εκφρασμένο ως δεκαδικό αριθμό.

- Μετά το N-οστό έτος θα έχουμε κεφάλαιο: $x_N = x_0 \prod_{n=1}^{N}(1 + r_n)$

Θέλουμε να βρούμε "μέσο επιτόκιο" r τέτοιο ώστε:

$$x_N = x_0(1 + r)^N$$

Έχουμε:

$$1 + r = \left((1 + r_1)(1 + r_2) \cdots (1 + r_N)\right)^{1/N}$$

Άρα

$$r = G - 1$$

όπου G ο γεωμετρικός μέσος των $\{(1 + r_n)\}_{n=1}^{N}$

Γνωρίζουμε ότι

$$1 + r_n = x_n/x_{n-1},\ n = 1, \ldots, N$$

Ο γεωμετρικός μέσος G των $1 + r_n$ ταυτίζεται με αυτό των x_n/x_{n-1} ως αποτέλεσμα

$$G = \left(\frac{x_1}{x_0}\frac{x_2}{x_1} \cdots \frac{x_{N-1}}{x_{N-2}}\frac{x_N}{x_{N-1}}\right)^{1/N} = \left(\frac{x_N}{x_0}\right)^{1/N}$$

και

$$r = \left(\frac{x_N}{x_0}\right)^{1/N} - 1$$

Το r θα το ονομάζουμε **μέσο ρυθμό μεταβολής** και εξαρτάται μόνο από την αρχική και την τελική τιμή μιας χρονολογικής σειράς.

Παράδειγμα

Το κεφάλαιο μιας επιχείρησης πενταπλασιάστηκε σε μια δεκαετία. Ποιος είναι ο μέσος ετήσιος ποσοστιαίος ρυθμός αύξήσης του κεφαλαίου;

$$r = \left(\frac{x_{10}}{x_0}\right)^{1/10} - 1 = \left(\frac{5 * x_0}{x_0}\right)^{1/10} - 1 = 0.1746$$

Παράδειγμα

Το κεφάλαιο μιας επιχείρησης υποπενταπλασιάστηκε σε μια δεκαετία. Ποιος είναι ο μέσος ετήσιος ποσοστιαίος ρυθμός μείωσης του κεφαλαίου;

$$r = \left(\frac{x_{10}}{x_0}\right)^{1/10} - 1 = \left(\frac{x_0/5}{x_0}\right)^{1/10} - 1 = -0.1487$$

Επικρατέστερη Τιμή (Mode)

- Είναι η τιμή της μεταβλητής με τη μεγαλύτερη συχνότητα εμφάνισης.
- Ορίζεται και για ποιοτικές μεταβλητές.
- Αν δυο ή περισσότερες τιμές έχουν την ίδια μέγιστη συχνότητα δεν ορίζεται επικρατέστερη τιμή.

Παράδειγμα

Έστω παρατηρήσεις: 2, 3, 4, 1, 2, 6, -2, 2
Το 2 με συχνότητα 3 είναι η επικρατέστερη τιμή του δείγματος.

Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων

Έστω οι κλάσεις που ορίζονται από τα διαστήματα με ίσο πλάτος d:

$$[a_1, a_2), [a_2, a_3), \ldots, [a_j, a_{j+1}), \ldots, [a_K, a_{K+1}).$$

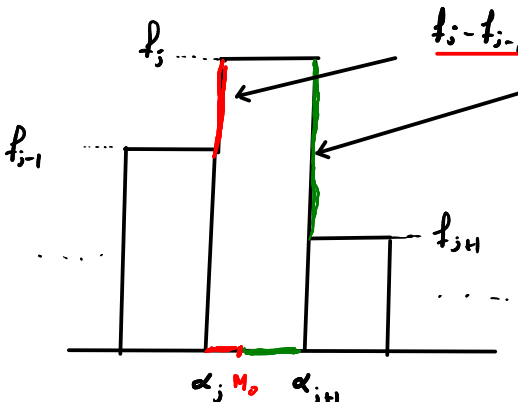
Εάν υπάρχει μοναδικός δείκτης j τέτοιος ώστε

$$f_j > f_k, \ \forall k \neq j.$$

Τότε $M_0 \in [a_j, a_{j+1})$.

$$M_0 = a_j + d\frac{f_j - f_{j-1}}{(f_j - f_{j-1}) + (f_j - f_{j+1})}$$

Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων



$$\frac{M_o - \alpha_j}{f_j - f_{j-1}} = \frac{\overbrace{\alpha_{j+1} - \alpha_j}^{d}}{(f_j - f_{j+1}) + (f_j - f_{j-1})}$$

$$M_o - \alpha_j = d \, \frac{f_j - f_{j-1}}{(f_j - f_{j+1}) + (f_j - f_{j-1})}$$

Παράδειγμα - Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων

	f
[0,1)	3
[1,2)	4
[2,3)	5
[3,4)	2
[4,5)	4
[5,6)	2
Total	20

$j = 3$

$$M_0 = \alpha_3 + 1 \cdot \frac{f_3 - f_2}{(f_3 - f_2) + (f_3 - f_4)} =$$

$$= 2 + \frac{1}{1+3} = 2 + \frac{1}{4} = 2.25.$$

Μέτρα κεντρικής τάσης

- Μέση τιμή $\bar{X}$
- Διάμεσος M
- Γεωμετρικός μέσος G
- Επικρατέστερη τιμή M_0

Μέτρα μεταβλητότητας

- Εύρος R
- Ενδοτεταρτημορικό εύρος IQR

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

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

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

με ισότητα μόνο αν $\bar{X} = \mu$.

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

$f'(x^*) = 0 \iff \sum_{n=1}^{N}(x_n - x^*) = 0 \iff \sum_{n=1}^{N} x_n - N x^* = 0 \iff x^* = \frac{1}{N}\sum_{n=1}^{N} x_n = \bar{X}$

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

$$\{(x-\mu)^2\}$$

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

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

Διασπορά στατιστικού δείγματος

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

Μπορούμε να γράψουμε ισοδύναμα:

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

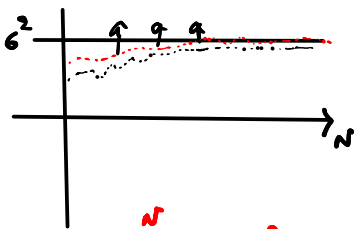
Όσο το N αυξάνεται έχουμε $s^2 \to \sigma^2$.

Διασπορά στατιστικού δείγματος

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

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

$$\frac{1}{N}\sum (x_n - \bar{x})^2 \leq \frac{1}{N}\sum (x_n - \mu)^2 \to \sigma^2$$



$$\lim_{N\to\infty} \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2 = \sigma^2$$

Όταν γνωρίσουμε το μ. $x_1, x_2, x_3, \ldots, x_N$

$\bar{x}$ $x_1, x_2, x_3, \ldots, x_{N-1}, x_N$

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\sum_{n=1}^{N-1} x_n + \frac{1}{N}x_N \Leftrightarrow x_N = N\bar{x} - \sum_{n=1}^{N-1} x_n$$

Διασπορά ομαδοποιημένων δεδομένων

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

Μπορούμε να γράψουμε ισοδύναμα:

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

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

Άσκηση - Διασπορά ομαδοποιημένων δεδομένων

	f	m	m^2	mf	$m^2 f$
[0,2)	3	1	1	3	3
[2,4)	4	3	9	12	
[4,6)	5	5	25		
[6,8)	2	7			
[8,10)	4	9			
[10,12)	2	11			
Total	$N=20$			$\sum m_j f_j$	$\sum m_j^2 f_j$

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

- Τυπική απόκλιση πληθυσμού:

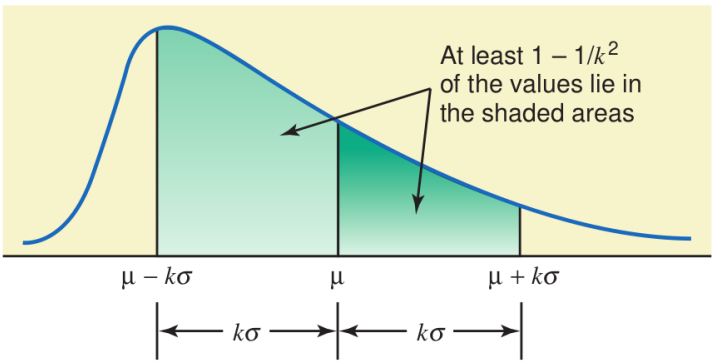
$$\sigma = \sqrt{\sigma^2}$$

- 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 $[\mu - k\sigma, \mu + k\sigma]$



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.

$231 - 187 = 44$ $\qquad 44 = k \cdot 6 = 22k$

$187 - 143 = 44$ $\qquad k = 2$

Τουλάχιστον $\left(1 - \frac{1}{2^2}\right) \cdot 100\%$ των παρατηρήσεων $\in [143, 231]$

$\left(1 - \frac{1}{4}\right) \cdot 100\%$ $\qquad \frac{3}{4} \cdot 100\% = 75\%$