ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

10-02-2020

Διάμεσος

Η διάμεσος ενός δείγματος είναι η τιμή που χωρίζει τις παρατηρήσεις έτσι ώστε τουλάχιστον το 50% αυτών να είναι μικρότερες ή ίσες και τουλάχιστον το 50% μεγαλύτερες ή ίσες από αυτήν.

Διάμεσος διατεταγμένων παρατηρήσεων

Έστω $x_1, x_2, \ldots, x_N$ διατεταγμένες παρατηρήσεις μιας μεταβλητής X τότε η διάμεσος δίνεται:

1. Εάν το N είναι περιττός αριθμός: $M = x_{(N+1)/2}$.
2. Εάν το N είναι άρτιος αριθμός: $M = \frac{1}{2}\left(x_{N/2} + x_{(N/2+1)}\right)$

Διάμεσος ομαδοποιημένων παρατηρήσεων

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

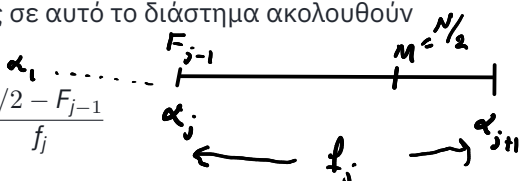
$$[a_1, a_2), [a_2, a_3), \ldots, [a_j, a_{j+1}), \ldots, [a_K, a_{K+1}).$$

$N/2 =$ Ο αριθμός των παρατηρήσεων που πρέπει να είναι μικρότερες από M.
Υπάρχει μοναδικός δείκτης j τέτοιος ώστε

$$F_{j-1} < N/2 \leq F_j.$$

Άρα το $M \in [a_j, a_{j+1})$. Υποθέτοντας ότι οι τιμές σε αυτό το διάστημα ακολουθούν ομοιόμορφη κατανομή έχουμε

$$M = a_j + d\frac{N/2 - F_{j-1}}{f_j}$$



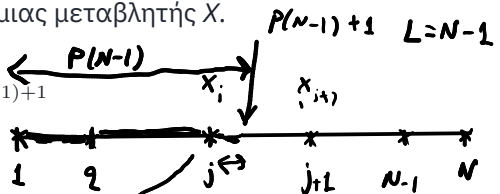
Έστω $p \in (0,1)$. Ορίζουμε το $100 * p$ -οστό ποσοστημόριο του δείγματος ως την τιμή P_p για την οποία τουλάχιστον $100 * p$ % των παρατηρήσεων είναι μικρότερες ή ίσες και τουλάχιστόν $100 * (1-p)$ % είναι μεγαλύτερες ή ίσες από αυτήν. Για $p = 0.5$ έχουμε τον ορισμό της διαμέσου, δηλαδή $P_{0.5} = M$.

100*p-οστό ποσοστημόριο διατεταγμένων παρατηρήσεων

Έστω $x_1, x_2, \ldots, x_N$ διατεταγμένες παρατηρήσεις μιας μεταβλητής X.

1. Εάν $p(N-1) \in \mathbb{Z}$ τότε:

$$P_p = x_{p(N-1)+1}$$



2. Διαφορετικά $P_p \in [x_{[p(N-1)]+1}, x_{[p(N-1)]+2}]$:

$$P_p = x_{[p(N-1)]+1} + u(x_{[p(N-1)]+2} - x_{[p(N-1)]+1})$$

όπου u το δεκαδικό μέρος του $p(N-1)$, δηλαδή $u = p(N-1) - [p(N-1)]$.

Στη 2η περίπτωση επιλέγουμε τιμή με γραμμική παρεμβολή.

Παράδειγμα

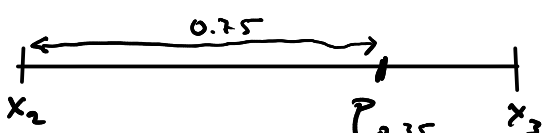
Να βρεθεί το 35-οστό ποσοστημόριο των διατεταγμένων παρατηρήσεων:
3, 4, 7, 10, 12, 17

$N=6 \quad p=0.35$

$P_{0.35} = 6.25$

$p(N-1) = 0.35 * 5 = 1.75 \notin \mathbb{Z}$

$P_{0.35} \in [x_{[1.75]+1}, x_{[1.75]+2}] = [x_2, x_3]$



$P_{0.35} = x_2 + 0.75 * (x_3 - x_2) = 4 + 0.75 * (7-4) = 4 + 0.75 * 3 = 6.25$

$Q_1 \equiv P_{0.25}$ (Πρώτο Τεταρτημόριο)

$Q_2 \equiv M \equiv P_{0.5}$ (Δεύτερο Τεταρτημόριο ή Διάμεσος)

$Q_3 \equiv P_{0.75}$ (Τρίτο Τεταρτημόριο)

Τεταρτημόρια ομαδοποιημένων παρατηρήσεων

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

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

$qN/4 =$ Ο αριθμός των παρατηρήσεων που πρέπει να είναι μικρότερες από Q_q.
Υπάρχει μοναδικός δείκτης j τέτοιος ώστε

$$F_{j-1} < qN/4 \leq F_j.$$

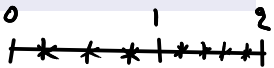
Άρα το $M \in [a_j, a_{j+1})$. Υποθέτοντας ότι οι τιμές σε αυτό το διάστημα ακολουθούν ομοιόμορφη κατανομή έχουμε

$$Q_q = a_j + d\frac{qN/4 - F_{j-1}}{f_j},\ q = 1, 2, 3$$

Παράδειγμα - Τεταρτημόρια ομαδοποιημένων παρατηρήσεων

	f	F
[0,1)	3	3
[1,2)	4	7
[2,3)	5	12
[3,4)	2	14
[4,5)	4	18
[5,6)	2	20
Total	20	

$q = 1, 2, 3 \quad Q_1 \quad M \quad Q_3$



$q=1 \quad q\frac{N}{4} = \frac{20}{4} = 5 \quad Q_1 \in [1, 2) \quad j = 2$

$Q_1 = \alpha_{\nu_2} + 1 \cdot \frac{5-3}{4} = 1 + \frac{1}{2} = 1.5$

$\nu_2 = 1$

$N/2 = 10 \quad j = 3$

$Q_2 = M = \alpha_j^{(2)} + 1 \cdot \frac{10-7}{5} = 2 + \frac{3}{5} = 2.6$

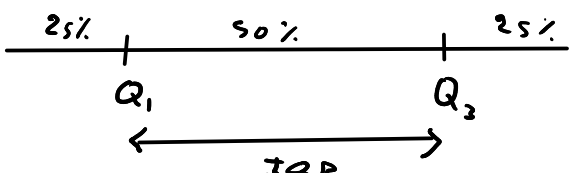
$q\frac{N}{4} = 3\frac{N}{4} = 3\frac{20}{4} = 3 \cdot 5 = 15 \quad j = 5$

$Q_3 = \alpha_5 + 1 \cdot \frac{15-14}{4} = 4 + \frac{1}{4} = 4.25$

Η απόσταση μεταξύ του πρώτου και τρίτου τεταρτημορίου

$$\text{IQR} = Q_3 - Q_1$$

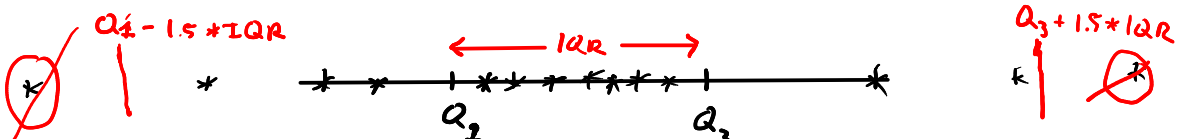
Περιλαμβάνει το 50 % (κεντρικότερες) παρατηρήσεις του δείγματος



- Ως ακραία παρατήρηση χαρακτηρίζεται εκείνη που διαφέρει σημαντικά από τις περισσότερες παρατηρήσεις.
- Μια ακραία παρατήρηση μπορεί να οφείλεται σε μεταβολές των συνθηκών μέτρησης ή μπορεί να υποδηλώνει κάποιο πειραματικό σφάλμα.

Κριτήριο 1.5*IQR για αναγνώριση Ακραίων τιμών

Το κριτήριο αναγνωρίζει ως ακραίες τις παρατηρήσεις οι οποίες είναι μικρότερες από $Q_1 - 1.5 * \text{IQR}$ ή μεγαλύτερες από $Q_3 + 1.5 * \text{IQR}$.



Παράδειγμα - Μετρώντας τη ταχύτητα του φωτός

Χρόνος ταξιδιού:

$$24.8 + 0.001 * x \quad \text{nanoseconds.}$$

Απόσταση: $\approx 7444\, m$
Μετρήσεις του x:

28	26	33	24	34	-44	27	16	40	-2	29
22	24	21	25	30	23	29	31	19	24	20
36	32	36	28	25	21	28	29	37	25	28
26	30	32	36	26	30	22	36	23	27	27
28	27	31	27	26	33	26	32	32	24	39
28	24	25	32	25	29	27	28	29	16	23

Παράδειγμα - Μετρώντας τη ταχύτητα του φωτός

Χρόνος ταξιδιού:

$$24.8 + 0.001 * x \quad \text{nanoseconds.}$$

Απόσταση: $\approx 7444\ m$
Διατεταγμένες μετρήσεις του x:

-44	-2	16	16	19	20	21	21	22	22	23
23	23	24	24	24	24	24	25	25	25	25
25	26	26	26	26	26	27	27	27	27	27
27	28	28	28	28	28	28	28	29	29	29
29	29	30	30	30	31	31	32	32	32	32
32	33	33	34	36	36	36	36	37	39	40

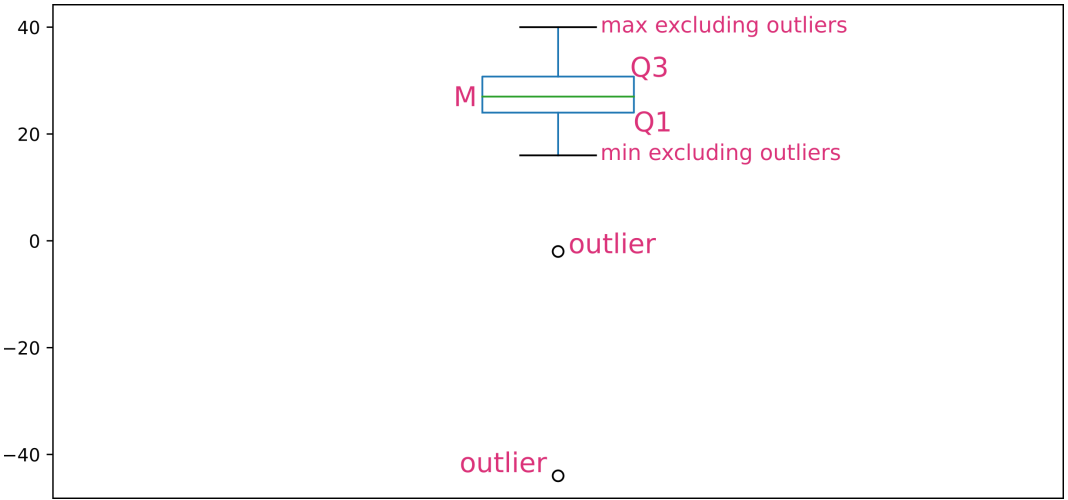
Παράδειγμα

-44	-2	16	16	19	20	21	21	22	22	23
23	23	24	24	24	24	24	25	25	25	25
25	26	26	26	26	26	27	27	27	27	27
27	28	28	28	28	28	28	28	29	29	29
29	29	30	30	30	31	31	32	32	32	32
32	33	33	34	36	36	36	36	37	39	40

- Μέση τιμή $\bar{X} = 26.21$
- Διάμεσος $M = 27.0$
- Πρώτο τεταρτημόριο $Q_1 = 24.0$, Τρίτο τεταρτημόριο $Q_3 = 30.75$
- Ενδοτεταρτημορικό εύρος $IQR = Q_3 - Q_1 = 30.75 - 24.0 = 6.75$
- $(Q_1 - 1.5 * \text{IQR}, Q_3 + 1.5 * \text{IQR}) = (13.875, 40.875)$
- Ακραίες τιμές κατά 1.5*IQR : -44 και -2

- Προσέγγιστική τιμή της ταχύτητας του φωτός σήμερα: $299792\ km/s$
- Προσέγγιση με τη μέση τιμή των παρατηρήσεων: $299844\ km/s$
- Προσέγγιση με τη διάμεσο των παρατηρήσεων: $299835\ km/s$
- Προσέγγιση με τη μέση τιμή εκτός των ακραίων παρατηρήσεων: $299809\ km/s$

- Για το παράδειγμα υπολογισμού της ταχυτητας του φωτός.



Άσκηση

Κατασκευάστε το γράφημα box-and-whisker για τις διατεταγμένες παρατηρήσεις:

x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8

-13, -4, 0, 1, 3, 5, 6, 15

$N = 8$

$P(N-1) = \frac{1}{4} \cdot 7 = 1.75$

$M = \frac{1}{2}(1+3) = 2$

$Q_1 = P_{0.25} = x_2 + 0.75 \cdot (x_3 - x_2) =$

$P(N-1)+1 = 2.75$

$= -4 + 0.75 \cdot 4 = -1$

$Q_3 = P_{0.75}$

$0.75 \cdot 7 = 5.25$

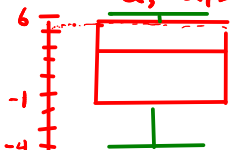
$0.75 \cdot 7 + 1 = 6.25$

$Q_3 \in (x_6, x_7)$

$Q_3 = x_6 + 0.25 \cdot (x_7 - x_6) = 5 + 0.25 \cdot 1 = 5.25$

$IQR = Q_3 - Q_1 = 5.25 + 1 = 6.25$

$(Q_1 - 1.5\,IQR,\ Q_3 + 1.5\,IQR) = (-10.375,\ 14.625)$



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

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

Χρησιμοποιήται 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:

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

Άρα

$$r = G - 1$$

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