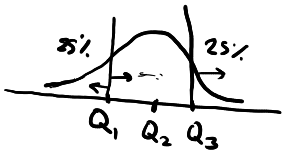
ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

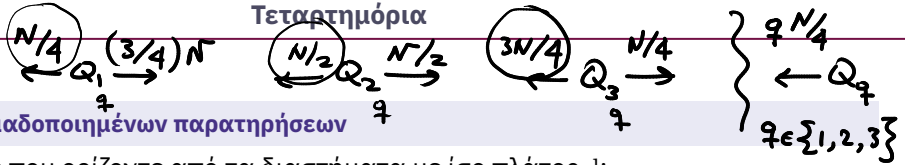
13-02-2023



$\mathrm{Q}_1 \equiv \mathrm{P}_{0.25}$ (Πρώτο Τεταρτημόριο)

$\mathrm{Q}_2 \equiv \mathrm{M} \equiv \mathrm{P}_{0.5}$ (Δεύτερο Τεταρτημόριο ή Διάμεσος)

$\mathrm{Q}_3 \equiv \mathrm{P}_{0.75}$ (Τρίτο Τεταρτημόριο)



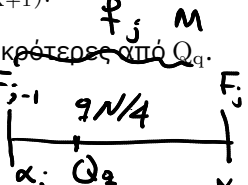
Τεταρτημόρια ομαδοποιημένων παρατηρήσεων

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

$$[\mathrm{a}_1, \mathrm{a}_2), [\mathrm{a}_2, \mathrm{a}_3), \ldots, [\mathrm{a}_\mathrm{j}, \mathrm{a}_{\mathrm{j}+1}), \ldots, [\mathrm{a}_\mathrm{K}, \mathrm{a}_{\mathrm{K}+1}).$$

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

$$\mathrm{F}_{\mathrm{j}-1} < \mathrm{qN}/4 \leq \mathrm{F}_\mathrm{j}.$$



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

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

Qq
$a_{j+1} - \alpha_j$

$Q_1 = ;$ $\quad N/4 = 5 \quad$ $Q_1 \in [x_2, x_3) = [1,2)$

Παράδειγμα - Τεταρτημόρια ομαδοποιημένων παρατηρήσεων

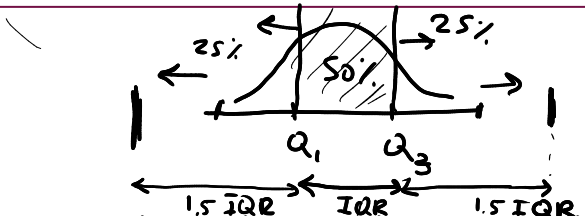
	f	F
[0,1)	3	3
[1,2)	4	7
[2,3)	5	12
[3,4)	2	14
[4,5)	4	18
[5,6)	2	20
Total	20	

$$Q_1 = 1 + 1 \cdot \frac{20/4 - 3}{4} = 1 + \frac{2}{4} = 1.5$$

$Q_3 = ;$ $\quad Q_3 \in [4,5)$

$3N/4 = 15$

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



Η απόσταση μεταξύ του πρώτου και τρίτου τεταρτημορίου

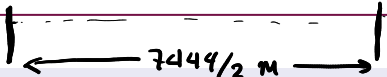
$$\text{IQR} = \text{Q}_3 - \text{Q}_1$$

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 $\mathrm{Q}_1 - 1.5 * \mathrm{IQR}$ ή μεγαλύτερες από $\mathrm{Q}_3 + 1.5 * \mathrm{IQR}$.



Παράδειγμα - Μετρώντας τη ταχύτητα του φωτός

Χρόνος ταξιδιού:

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

Απόσταση: $\approx 7444\ \mathrm{m}$
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:

$$t(x) = 24.8 + 0.001 * x \quad \text{nanoseconds.}$$

Απόσταση: ≈ 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

$Q_1 - 1.5\,IQR$ 13.875

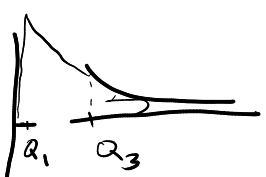
-44	-2	16	16	19	20	21	21	22	22	23
23	23	24	24	24	24	24	25	25	25	25
25	26	26	26	26	26	27	27	27	27	27
27	28	28	28	28	28	28	28	29	29	29
29	29	30	30	30	31	31	32	32	32	32
32	33	33	34	36	36	36	36	37	39	40

$Q_3 + 1.5\,IQR$

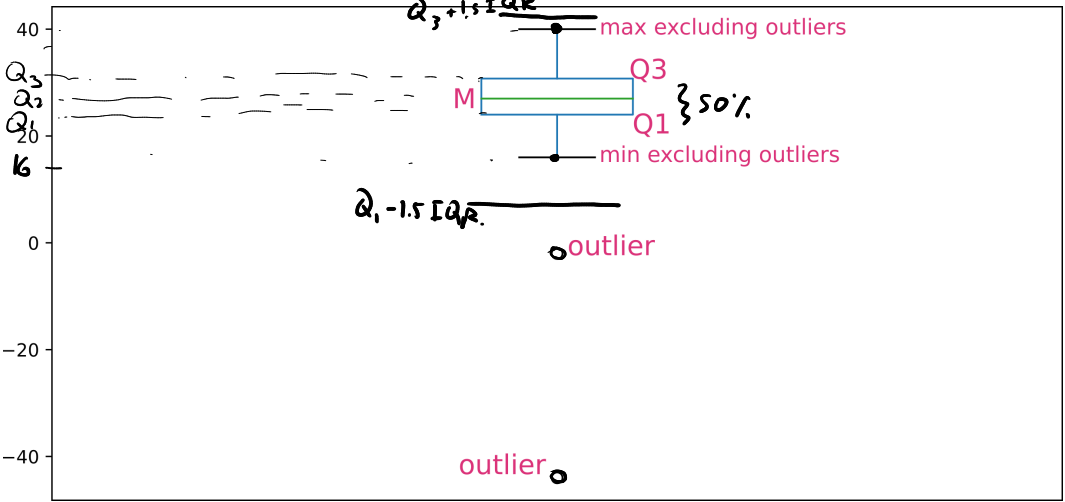
- Μέση τιμή $\bar{X} = 26.21$
- Διάμεσος $M = 27.0$ $= Q_2$
- Πρώτο τεταρτημόριο $Q_1 = 24.0$, Τρίτο τεταρτημόριο $Q_3 = 30.75$
- Ενδοτεταρτημορικό εύρος $IQR = Q_3 - Q_1 = 30.75 - 24.0 = 6.75$
- $(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR) = (13.875, 40.875)$
- Ακραίες τιμές κατά 1.5*IQR : -44 και -2

$$\frac{R}{t(x) \cdot 10^{-9}} = \frac{R}{t(x)} \cdot 10^{9} \ m/s$$

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



- Για το παράδειγμα υπολογισμού της ταχυτητας του φωτός.



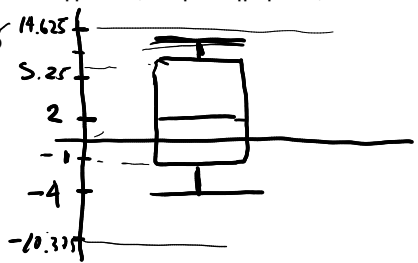
Άσκηση

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

$\alpha_1\ \alpha_2\ \alpha_3\ \alpha_4\ \alpha_5\ \alpha_6\ \alpha_7$

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

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



$Q_1 = P_{0.25}$

$Q_3 = P_{0.75}$

$0.25 \cdot (N-1) = 0.25 \cdot 7 = 1.75$

$Q_1 = -4 + [0 - (-4)] \cdot 0.75 = -4 + 4 \cdot 0.75 = -1$

$Q_3 =$

$0.75 \cdot (N-1) = 5.25$

$Q_3 = 5 + 1 \cdot 0.25 = 5.25$

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

$[-1 - 6.25 \cdot 1.5,\ 5.25 + 6.25 \cdot 1.5]$

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

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

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

$x_1 = x_0(1+r_1) \qquad x_2 = x_1(1+r_2)$

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 $\{(1+r_n)\}_{n=1}^{N}$

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:

$$[\mathrm{a}_1, \mathrm{a}_2), [\mathrm{a}_2, \mathrm{a}_3), \ldots, [\mathrm{a}_\mathrm{j}, \mathrm{a}_{\mathrm{j}+1}), \ldots, [\mathrm{a}_\mathrm{K}, \mathrm{a}_{\mathrm{K}+1}).$$

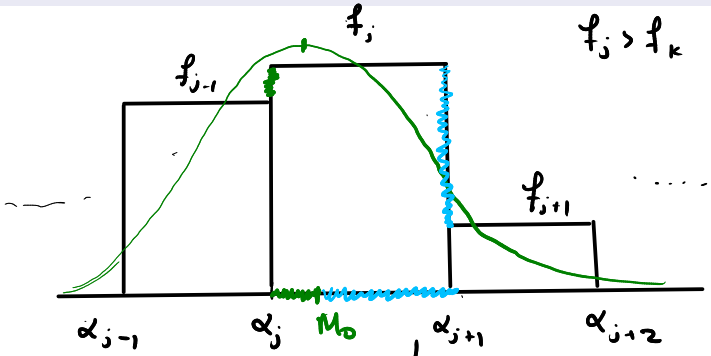
Εάν υπάρχει μοναδικός δείκτης j τέτοιος ώστε

$$\mathrm{f}_\mathrm{j} > \mathrm{f}_\mathrm{k}, \ \forall \mathrm{k} \neq \mathrm{j}.$$

Τότε $\mathrm{M}_0 \in [\mathrm{a}_\mathrm{j}, \mathrm{a}_{\mathrm{j}+1})$.

$$\mathrm{M}_0 = \mathrm{a}_\mathrm{j} + \mathrm{d}\frac{\mathrm{f}_\mathrm{j} - \mathrm{f}_{\mathrm{j}-1}}{(\mathrm{f}_\mathrm{j} - \mathrm{f}_{\mathrm{j}-1}) + (\mathrm{f}_\mathrm{j} - \mathrm{f}_{\mathrm{j}+1})}$$

Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων



$$f_j > f_k \quad \forall k \neq j$$

$$\frac{a}{b} = \frac{\gamma}{\delta} = \frac{\alpha+\gamma}{\beta+\delta}$$

$$\frac{M_0 - \alpha_j}{f_j - f_{j-1}} = \frac{\alpha_{j+1} - M_0}{f_j - f_{j+1}} = \frac{\overbrace{\alpha_{j+1} - \alpha_j}^{d}}{(f_j - f_{j-1}) + (f_j - f_{j+1})}$$

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

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

$$M_0 = 2 + 1 \cdot \frac{5-4}{(5-4)+(5-2)} = 2 + \frac{1}{4} = 2.25$$

Μέτρα κεντρικής τάσης

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

Μέτρα μεταβλητότητας

- Εύρος R
- Ενδοτεταρτημορικό εύρος IQR ← της κεντρικότερης 50%

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

Έστω $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{N}$ παρατηρήσεις που αντιστοιχούν σε ένα τυχαίο δείγμα ενός πληθυσμού.
Έχουμε ορίσει ως μέση τιμή των παρατηρήσεων του δείγματος την ποσότητα:

$$\bar{\mathrm{X}} = 1/\mathrm{N} \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{n}$$

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

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

με ισότητα μόνο αν $\bar{X} = \mu$.

$f(x) = \sum_{n=1}^{N}(x_n - x)^2 \Rightarrow f'(x) = -2\sum_{n=1}^{N}(x_n - x) = 0 \Rightarrow$

$\Rightarrow \sum_{n=1}^{N} x_n = \sum_{n=1}^{N} x = Nx \Rightarrow x = \frac{1}{N}\sum_{n=1}^{N} x_n = \bar{X}$

$f''(x) > 0$ άρα $\bar{X}$ είναι ελαχιστο.

Εαν $\mu \neq \bar{X}$ τοτε ισχύει η $*$ με $<$

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

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

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

Διασπορά στατιστικού δείγματος

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

Μπορούμε να γράψουμε ισοδύναμα:

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

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

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;

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
[0,2)	3
[2,4)	4
[4,6)	5
[6,8)	2
[8,10)	4
[10,12)	2
Total	20