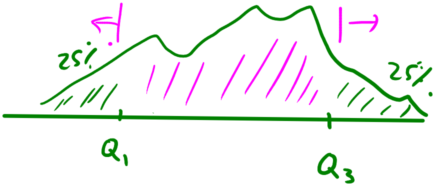
ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

3η εβδομάδα (διάλεξη θεωρίας)

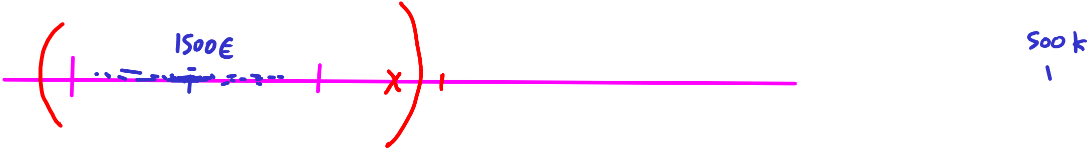
Ενδοτεταρτημοριακό Εύρος (Interquartile Range-IQR)



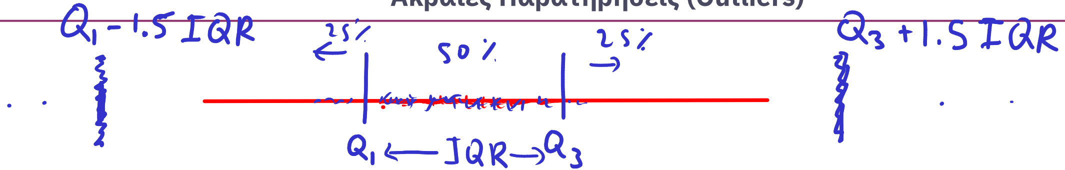
Η απόσταση μεταξύ του πρώτου και τρίτου τεταρτημορίου

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

Περιλαμβάνει το 50 % (κεντρικότερες) παρατηρήσεις του δείγματος



Ακραίες Παρατηρήσεις (Outliers)



- 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 $Q_1 - 1.5 * \mathrm{IQR}$ ή μεγαλύτερες από $Q_3 + 1.5 * \mathrm{IQR}$.

Newcomb.

Παράδειγμα - Μετρώντας τη ταχύτητα του φωτός

Χρόνος ταξιδιού:

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

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: 66

-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

66

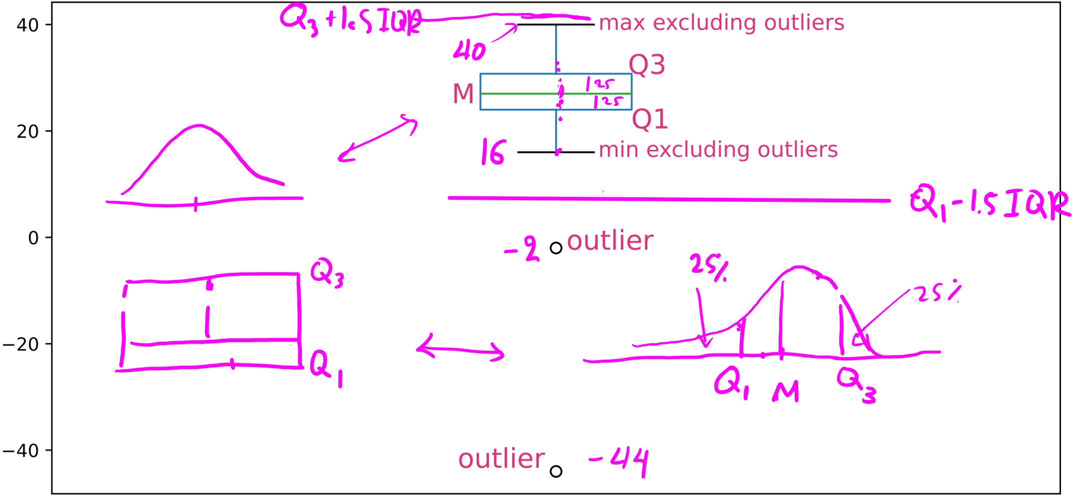
~~-44~~	~~-2~~	16	16	19	20	21	21	22	22	23
23	23	24	24	24	24	24	25	25	25	25
25	26	26	26	26	26	27	27	27	27	27
27	28	28	28	28	28	28	28	29	29	29
29	29	30	30	30	31	31	32	32	32	32
32	33	33	34	36	36	36	36	37	39	40

- 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

Παράδειγμα

- 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:

$$\overset{\alpha_1}{-13},\ \overset{\alpha_2}{-4},\ \overset{\alpha_3}{0},\ \boxed{1,\ 3},\ \overset{\alpha_6}{5},\ \overset{\alpha_7}{6},\ 15$$

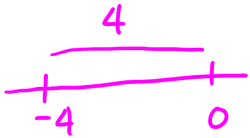
1

$M = 2$

$[1.75] + 1 = 2$

$Q_1 = ;$ $\quad P = 0.25$ $\quad \frac{1}{4}\cdot(8-1) = \frac{7}{4} = 1.75$

$Q_3 = ;$



$Q_1 \in [\alpha_2, \alpha_3]$

$Q_1 = -4 + 0.75 * 4 = -1$

$P = 0.75$ $\quad \frac{3}{4}(8-1) = \frac{21}{4} = 5.25$ $\quad [5.25] + 1 = 6$

$Q_3 = 5 + \frac{1}{4} = 5.25$

Άσκηση

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

$$\cancel{-13}, -4, 0, 1, 3, 5, 6, \cancel{15}$$

-10.375 14.625

$IQR = 5.25 - (-1) = 6.25$

$5.25 + 1.5 \cdot 6.25 = 14.625$

$-1 - 1.5 \cdot 6.25 = -10.375$

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

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

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

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

$G = 10^{0.829226} = 6.748785$

$x_j \xrightarrow{r_{j+1}} x_{j+1}$ $x_{j+1} = x_j + r_{j+1} x_j = (1 + r_{j+1}) x_j = (1+r_{j+1})(1+r_j) x_{j-1}$

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:

$$\boxed{x_N = x_0(1 + r)^N}$$

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

$x_n = (1+r_n)x_{n-1}$ $\qquad G = \big((1+r_1)(1+r_2)\cdots(1+r_N)\big)$

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

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

και

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

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 αύξησης του κεφαλαίου;

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

Παράδειγμα

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

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

- 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 (Mode)

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

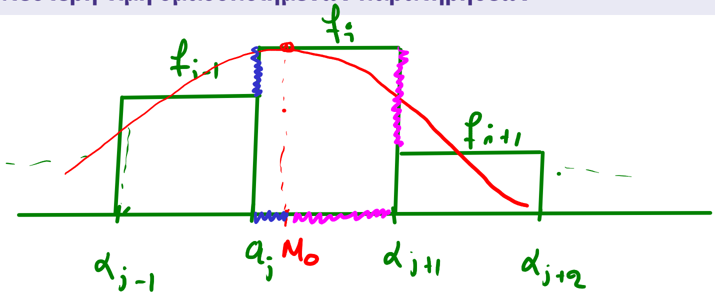
$$f_j > f_k, \ \forall k \neq j.$$

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 (Mode)

Επικρατέστερη τιμή ομαδοποιημένων παρατηρήσεων



$$\frac{f_i - f_{i-1}}{M_0 - a_j} = \frac{f_i - f_{i+1}}{a_{j+1} - M_0} = \frac{(f_i - f_{i-1}) + (f_i - f_{i+1})}{(\cancel{M_0} - a_j) + (a_{j+1} - \cancel{M_0})} = \frac{}{d}$$

Επικρατέστερη Τιμή (Mode)

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

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

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