ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

2η εβδομάδα (διάλεξη θεωρίας)

Παράδειγμα

Έστω ότι συλλέγουμε πληροφορίες για την ηλικία και το φύλο 20 φοιτητών/τριών που είναι εγγεγραμμένοι σε ένα μάθημα.

(37,M)	(18,M)	(19,F)	(22,F)	(30,M)
(24,F)	(22,M)	(19,F)	(28,M)	(20,F)
(22,F)	(21,F)	(34,F)	(19,M)	(22,M)
(20,M)	(18,F)	(33,F)	(19,F)	(24,M)

- Θέλουμε να μελετήσουμε τις ηλικίες.

Εύρος τιμών R

Ορίζεται ως η διαφορά της μικρότερης παρατήρησης/μέτρησης από την μεγαλύτερη.

$$R = \max_{n=1,\ldots,N} \{x_n\} - \min_{n=1,\ldots,N} \{x_n\}$$

- Για το παράδειγμά μας έχουμε $R = 37 - 18 = 19$.

Κανόνας του Sturges

- Είναι βασισμένος στην υπόθεση για δεδομένα που ακολουθούν την κανονική κατανομή.

$$K^{\text{opt}} = 1 + 3.322 * \log(N)$$

.

- Για $N = 20$ έχουμε $K^{\text{opt}} = 5.33$ κλάσεις. Θέλουμε ακέραιο πλήθος άρα θέτουμε $K^{\text{opt}} = 5$.
- Κάθε κλάση θα έχει εύρος $d \approx R/K = 19/5 = 3.8$. Συνήθως στρογγυλοποιούμε το πλάτος προς τα επάνω, άρα $d = 4$.
- Ξεκινώντας από την μικρότερη παρατήρηση ορίζουμε 5 κλάσεις με πλάτος 4.
  - πρώτη κλάση: 18,19,20,21 → [18,21] ή [18,22)
  - δεύτερη κλάση: 22,23,24,25 → [22,25] ή [22,25)
  - τρίτη κλάση: 26,27,28,29 → [26,29] ή [26,30)
  - τέταρτη κλάση: 30,31,32,33 → [30,33] ή [30,34)
  - πέμπτη κλάση: 34,35,36,37 → [34,37] ή [34,38)

- Αναπαράσταση κατάλληλη για διακριτές μεταβλητές.

Class	LB	UB	Midpoint (m)	Width (d)	Frequency (f)
[18,21]	17.5	21.5	(18+21)/2 = 19.5	$UB_1 - LB_1 = 4$	$f_1 = 9$
[22,25]	21.5	25.5	23.5	$UB_2 - LB_2 = 4$	$f_2 = 6$
[26,29]	25.5	29.5	27.5	$UB_3 - LB_3 = 4$	$f_3 = 1$
[30,33]	29.5	33.5	31.5	$UB_4 - LB_4 = 4$	$f_4 = 2$
[34,37]	33.5	37.5	35.5	$UB_5 - LB_5 = 4$	$f_5 = 2$
Total					$\sum_{i=1}^{5} f_i = 20$

- Αναπαράσταση καταλληλότερη για συνεχείς μεταβλητές.

Class	LB	UB	Midpoint (m)	Width (d)	Frequency (f)
[18,22)	18	22	(18+22)/2 = 20	$UB_1 - LB_1 = 4$	$f_1 = 9$
[22,26)	22	26	24	$UB_2 - LB_2 = 4$	$f_2 = 6$
[26,30)	26	30	28	$UB_3 - LB_3 = 4$	$f_3 = 1$
[30,34)	30	34	32	$UB_4 - LB_4 = 4$	$f_4 = 2$
[34,38)	34	38	36	$UB_5 - LB_5 = 4$	$f_5 = 2$
Total					$\sum_{i=1}^{5} f_i = 20$

- Το αριστερό όριο της πρώτης κλάσης μπορεί να στογγυλοποιηθεί προς τα κάτω και το δεξί όριο της τελευταίας κλάσης προς τα επάνω.
- Σε μια τέτοια περίπτωση πρέπει να αναπροσαρμοσθεί το εύρος R.

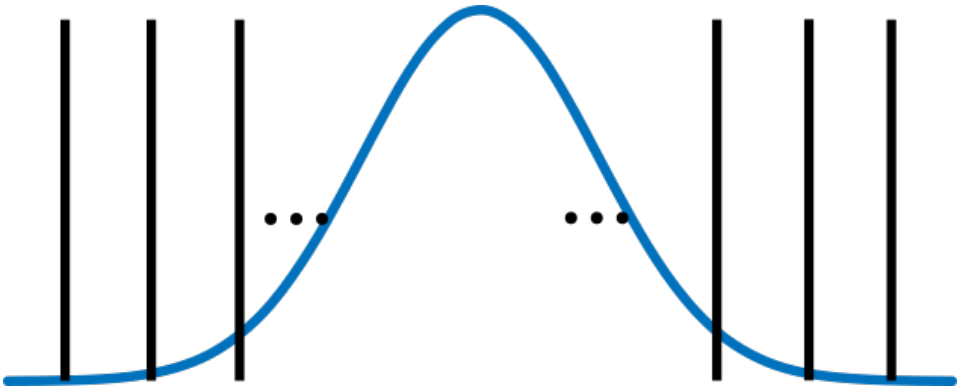
Παράδειγμα

1.1	1.14	0.25	0.05	1.1
1.18	0.34	2.1	1.0	3.75

- Έχουμε μικρότερη παρατήρηση το 0.05 και μεγαλύτερη το 3.75
- Το εύρος είναι R = 3.75-0.05 = 3.7
- Μπορούμε να θέσουμε το αριστερότερο όριο 0.0 και το δεξιότερο 4.0
- Αναπροσαρμόζουμε το R = 4.0 - 0.0 = 4.0
- Από τον κανόνα του Sturges έχουμε $1 + 3.322 * log(10) = 4.322$. Θέτουμε $K = 4$
- Το πλάτος κάθε κλάσης θα δοθεί από τη σχέση $d = R/K = 4/4 = 1$
- **Κλάσεις: [0,1), [1,2), [2,3), [3,4)**

Άσκηση

Κατασκευάστε τον πίνακα συχνοτήτων για τα δεδομένα του προηγούμενου παραδείγματος.



Άσκηση

Δίδονται τα παρακάτω ακατέργαστα δεδομένα.

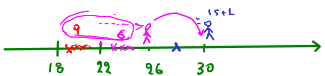
1.1	-3.8	0.2	3.3	-2.4	0.5	-2.1	4.7	-0.1	1.2
0.1	-2.3	2.5	3.5	-3.7	3.0	1.1	0.2	1.8	0.3
3.6	-1.7	0.1	-0.2	1.0	3.3	-1.5	0.9	-2.7	4.1

1. Κατασκευάστε κατάλληλο πίνακα συχνοτήτων χρησιμοποιώντας τον κανόνα του Sturges για τον καθορισμό του αριθμού των κλάσεων.
2. Πώς θα αλλάξουν τα όρια των κλάσεων εάν προσθέσετε σε όλες τις παρατηρήσεις τον αριθμό 2;

(1+3.322*log(30)=5.907)

Αθροιστική συχνότητα (Cumulative Frequency)

Η κατανομή αθροιστικών συχνοτήτων εκφράζει το πλήθος των παρατηρήσεων που είναι μικρότερες από το επάνω σύνορο κάθε κλάσης. Για την j-οστή κλάση συμβολίζεται με F_j.



$$F_j = \sum_{i=1}^{j} f_i,\ j = 1, \ldots, K$$

Class	LB	UB	m	f	F
[18,22)	18	22	20	$f_1 = 9$	$F_1 = f_1 = 9$
[22,26)	22	26	24	$f_2 = 6$	$F_2 = F_1 + f_2 = 15$
[26,30)	26	30	28	$f_3 = 1$	$F_3 = F_2 + f_3 = 16$
[30,34)	30	34	32	$f_4 = 2$	$F_4 = F_3 + f_4 = 18$
[34,38)	34	38	36	$f_5 = 2$	$F_5 = F_4 + f_5 = 20$
Total				20	

Σχετική συχνότητα και σχετική αθροιστική συχνότητα

$$rf_j = f_j / \sum_{i=1}^{K} f_j = \frac{f_j}{N}, \quad RF_j = F_j / \sum_{i=1}^{K} f_j = \frac{F_j}{N}$$

Class	LB	UB	m	f	rf	F	RF
[18,22)	18	22	20	9	0.45	9	0.45
[22,26)	22	26	24	6	0.3	15	0.75
[26,30)	26	30	28	1	0.05	16	0.8
[30,34)	30	34	32	2	0.1	18	0.9
[34,38)	34	38	36	2	0.1	20	1
Total				20	1		

Σχετική συχνότητα και σχετική αθροιστική συχνότητα

$$rf_j\% = rf_j * 100\%, \quad RF_j\% = RF_j * 100\%$$

Class	LB	UB	m	f	rf	rf%	F	RF	RF%
[18,22)	18	22	20	9	0.45	45	9	0.45	45
[22,26)	22	26	24	6	0.3	30	15	0.75	75
[26,30)	26	30	28	1	0.05	5	16	0.8	80
[30,34)	30	34	32	2	0.1	10	18	0.9	90
[34,38)	34	38	36	2	0.1	10	20	1	100
Total				20	1	100			

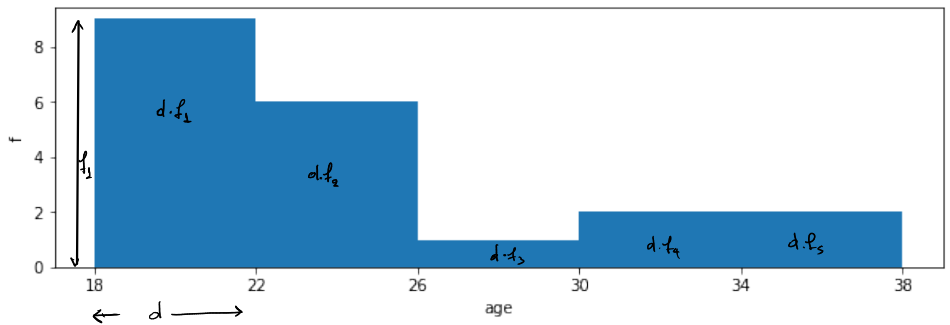
Άσκηση

Δίνονται οι παρακάτω μετρήσεις.

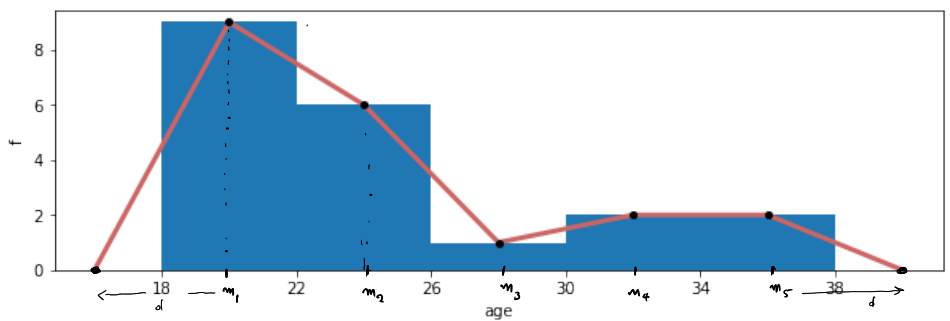
239.1	212.1	249.1	227.1	218.1	310.0	281.2	330.1	226.1
223.2	161.1	195.3	233.8	249.5	284.6	284.5	174.2	170.7
169.0	299.6	210.4	301.3	199.1	258.3	258.5	195.4	227.3
355.0	234.1	195.9	196.4	354.3	282.1	282.3	286.1	286.3
195.5	163.8	297.1	211.5	288.1	309.4	309.9	225.7	223.9
248.2	284.4	173.9	256.0	169.2	209.6	209.3	200.3	258.0

Ομαδοποιήστε τις τιμές και κατασκευάστε πίνακα συχνοτήτων, σχετικών συχνοτήτων, αθροιστικών συχνοτήτων και αθροιστικών σχετικών συχνοτήτων.
(1+3.322*log(60) = 6.907018)

- Κατασκευάζουμε ορθογώνια με βάσεις τα διαστήματα $[LB_j, UB_j]$(ομοιόμορφου πλάτους d) τών κλάσεων και με ύψη τις αντίστοιχες συχνότητες f_j.
- Το εμβαδόν κάθε ορθογωνίου είναι $d * f_j$.
- Το συνολικό εμβαδόν του ιστογράμματος (όλα τα ορθογώνια) είναι $d * \sum_{j=1}^{K} f_j = d * N$.



- Ενώνουμε με ευθύγραμμα τμήματα το σύνολο των σημείων $\{(m_j, f_j)\}_{j=1}^{K}$, όπου m_j η κεντρική τιμή της j-οστής κλάσης.
- Το εμβαδόν της περιοχής που ορίζεται από τα ευθύγραμμα τμήματα και τον οριζόντιο άξονα είναι πάντα μικρότερο ή ίσο από το εμβαδόν του αντιστοίχου ιστογράμματος.
- Το εμβαδόν γίνεται ίσο με αυτό του ιστογράμματος έαν θεωρήσουμε επιπλέον τα σημεία $(m_1 - d, 0), (m_K + d, 0)$



Κατανομές συχνοτήτων ποιοτικών δεδομένων

- Κάθε δυνατή τιμή μιας ποιοτικής μεταβλητής ορίζει μια κατηγορία.
- Η κατανομή συχνοτήτων για ποιοτικά δεδομένα απαριθμεί τα στοιχείων τα οποία ανήκουν σε κάθε κατηγορία.
- Για το παράδειγμα με τους φοιτητές μετρώντας τον αριθμό για το κάθε φύλο κατασκευάζουμε τον πίνακα

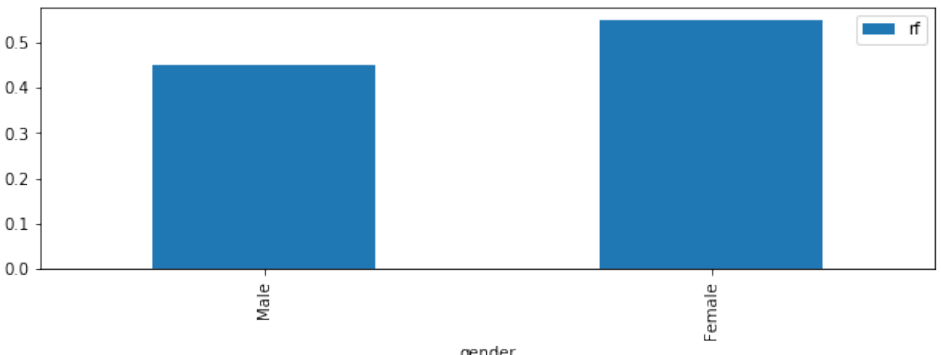
		Frequency (f)
Male (M)	𝍸 ||||	$f_1 = 9$
Female (F)	𝍸 𝍸 |	$f_2 = 11$
Total		$f_1 + f_2 = N = 20$

Σχετικές Συχνότητες

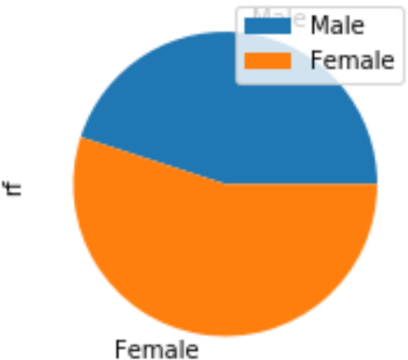
$$rf_k = \frac{f_k}{N},\ k = 1, 2, \ldots, K$$

	Frequency (f)	Relative Frequency (rf)	Percentage (rf%)
Male (M)	9	$rf_1 = 9/20 = 0.45$	$rf_1 * 100 = 45$
Female (F)	11	$rf_2 = 11/20 = 0.55$	$rf_2 * 100 = 55$
Total	20	$rf_1 + rf_2 = 1$	100

- Σαν το ιστόγραμμα αλλά για ποιοτικά δεδομένα.
- Κάθε ορθογώνιο αντιστοιχεί σε μια κατηγορία.
- Οι βάσεις των ορθογωνίων δεν εκφράζονται αριθμητικά, οπότε δεν ορίζεται εμβαδόν.



- Στην j-οστή κατηγορία αντιστοιχίζουμε γωνία $rf_j * 360^o$.
- Αυτές οι γωνίες θα είναι οι γωνίες των κυκλικών τμημάτων ενός κυκλικού δίσκου.



- Θελουμε να περιγράψουμε την κατανομή μιας τυχαίας μεταβλητής που περιγράφει μια μεταβλητή του στατιστικού πληθυσμού με ένα σύνολο από χαρακτηριστικούς αριθμούς.
- Αυτοί οι αριθμοί παρέχουν πληροφορίες για τις τάσεις των τιμών που λαμβάνει η μεταβλητή.
- Τα περιγραφικά μέτρα που θα εξετάσουμε διακρίνονται στις επόμενες κατηγορίες:
  1. Μέτρα κεντρικής τάσης: Προσδιορίζουν μια τιμή γύρω από την οποία τείνουν να συγκεντρώνονται οι τιμές της μεταβλητής.
  2. Μέτρα μεταβλητότητας: Ποσοτικοποιούν πόσο μακριά απλώνονται οι τιμές από κάποιο μέτρο θέσης.
  3. Μέτρα ασυμμετρίας: Εκφράζουν κατά πόσο υπάρχει συμμετρία των τιμών ως πρός ένα μέτρο θέσης.
  4. Μέτρα κύρτωσης: Περιγράφουν την οξυτήτα της κορυφής της κατανομής των τιμών μιας μεταβλητής.

Μέση τιμή (mean value)

Έστω $x_1, x_2, \ldots, x_N$ παρατηρήσεις μια μεταβλητής X. Η μέση τιμή $\bar{x}$ ορίζεται ως:

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

Γραμμικός μετασχηματισμός

Έστω $Y = aX + b$, όπου $a, b \in \mathbb{R}$ τότε $\bar{y} = a\bar{x} + b$.

Παράδειγμα

$$x_1 = 10,\ x_2 = 14,\ x_3 = 15,\ x_4 = 5,\ x_5 = 6, \quad \text{και} \quad Y = 2X - 3$$

$$\bar{x} = \frac{1}{5}(10 + 14 + 15 + 5 + 6) = 10, \quad \text{και} \quad \bar{y} = 2 * 10 - 3 = 17$$

Σταθμισμένη μέση τιμή (weighted mean value)

Σε κάποιες περιπτώσεις οι τιμές μιας μεταβλητής δεν έχουν την ίδια βαρύτητα για όλα τα στοιχεία του πληθυσμού. Εάν η βαρύτητα της παρατήρησης x_n καθορίζεται από ένα βάρος w_n τότε έχει νόημα ο υπολογισμός της σταθμισμένης μέσης τιμής.

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

Σταν μέση τιμή $w_n = 1 \ \forall n$

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

Παράδειγμα - Μέσο κόστος ανά τεμάχιο

Quality	Items	Unit price (Euro)
A	3 ← w_1	500
B	7 ← w_2	100
C	10 ← w_3	20

500, 500, 500, 100, 100, ..., 100 (7)
, 20 ..... 20 (10)

$\frac{20 + 100 + 500}{3}$ = 206 !

$$\bar{x} = \frac{3 * 500 + 7 * 100 + 10 * 20}{3 + 7 + 10} = 120$$

Γραμμικός μετασχηματισμός

Έστω $x_1, x_2, \ldots, x_N$ και αντίστοιχα βάρη $w_1, w_2, \ldots, w_N$. Εάν $Y = aX + b$ τότε:

$$\bar{y} = \frac{\sum_{n=1}^{N} w_n(ax_n + b)}{\sum_{n=1}^{N} w_n} = a\frac{\sum_{n=1}^{N} w_n x_n}{\sum_{n=1}^{N} w_n} + b\frac{\sum_{n=1}^{N} w_n}{\sum_{n=1}^{N} w_n} = a\bar{x} + b$$

Όταν έχουμε ομαδοποιημένα δεδομένα σε Κ κλάσεις η μέση τιμή δίνεται από τη παρακάτω σχέση:

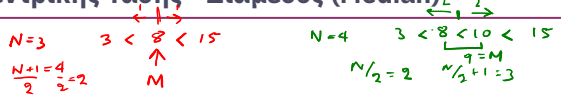
$$\bar{x} = \frac{\sum_{j=1}^{K} m_j f_j}{\sum_{j=1}^{K} f_j}$$

Παράδειγμα

Class	m	f	m * f
[18,22)	20	9	180
[22,26)	24	6	144
[26,30)	28	1	28
[30,34)	32	2	64
[34,38)	36	2	72
Total		20	488

$$\bar{x} = \frac{\sum_{j=1}^{K} m_j f_j}{\sum_{j=1}^{K} f_j} = \frac{488}{20} = 24.4$$

- Εάν υπολογίζαμε τη μέση τιμή στα ακατέργαστα δεδομένα θα είχαμε το ίδιο αποτέλεσμα;



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: $M = \frac{1}{2}\left(x_{N/2} + x_{(N/2+1)}\right)$

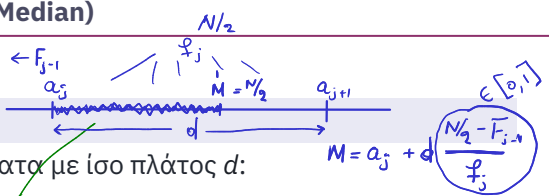
Διάμεσος

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

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

$M = P_{0.5}$

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

$[3.5] = 3$

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

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

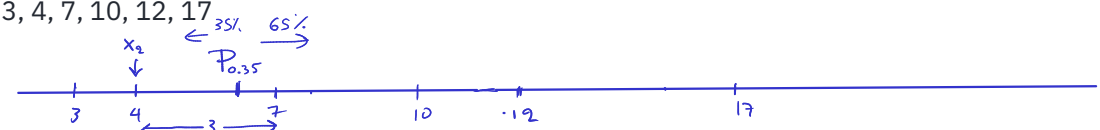
Στη 2η περίπτωση επιλέγουμε τιμή με γραμμική παρεμβολή.

$N=6 \quad p=0.35 \quad p(N-1) = 0.35 \cdot 5 = 1.75 \quad [1.75] = 1$

$P_{0.35} \in [X_2, X_3] = [4, 7]$

Παράδειγμα

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



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

$Q_1 \equiv P_{0.25}$ (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 $Q_q \in [a_j, a_{j+1})$. Υποθέτοντας ότι οι τιμές σε αυτό το διάστημα ακολουθούν ομοιόμορφη κατανομή έχουμε

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

Q_1 < Q_2 (M) < Q_3

$q=1$ $q=2$ $q=3$

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	

$q=1$ $q\frac{N}{4}=5$ $F_1 < 5 \le F_2$ $Q \in [a_2, a_3)$

$Q_1 = a_2 + \frac{5-3}{4} = 1 + \frac{1}{2} = 1.5$

a_2 a_1 5

Q_2 άσκηση

$q=3$ $q\frac{N}{4} = 3\frac{20}{4} = 15$ $F_4 < 15 \le F_5$

15

a_5 a_3

$Q_3 = a_5 + \frac{15-14}{4} = 4 + \frac{1}{4} = 4.25$