Πλαίσιο Εργασίας

Στατιστική είναι: η συλλογή, παρουσίαση, ανάλυση και εξαγωγή συμπερασμάτων για κάποιο χαρακτηριστικό που αφορά τις μονάδες ενός πληθυσμού, τις οποίες δεν τις μελετούμε όλες αλλά επιλέγουμε ένα αντιπροσωπευτικό δείγμα από αυτές.

Βασικές έννοιες

- Πληθυσμός
- Επιλογή (τυχαίου) δείγματος
- Το υπό μελέτη χαρακτηριστικό

Βασικοί κλάδοι της Στατιστικής

- Δειγματοληψία
- Περιγραφική Στατιστική
- Εκτιμητική
- Στατιστική Συμπερασματολογία

Στατιστικά Δεδομένα

1. Ποιοτικά Δεδομένα

προτίμηση σε κόμμα, ποδοσφαιρική ομάδα, αναψυκτικό

φύλο, σχολή, έτος φοίτησης

2. Ποσοτικά Δεδομένα

α. διακριτά χαρακτηριστικά

πλήθος κοριτσιών σε μια οικογένεια

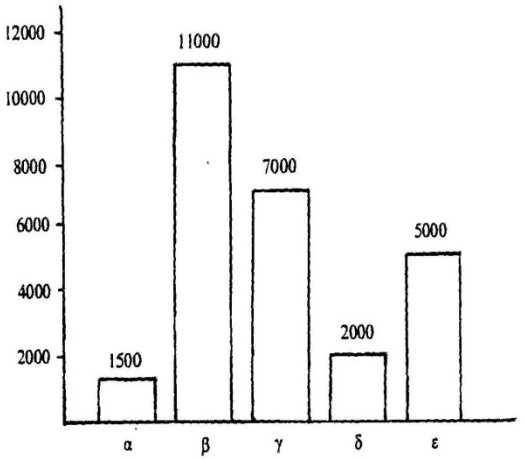
αφίξεις στην ουρά μιας τράπεζας, πλήθος ατυχημάτων

β. συνεχή χαρακτηριστικά

βάρος, ύψος, χρόνος ζωής λαμπτήρα, θερμοκρασία

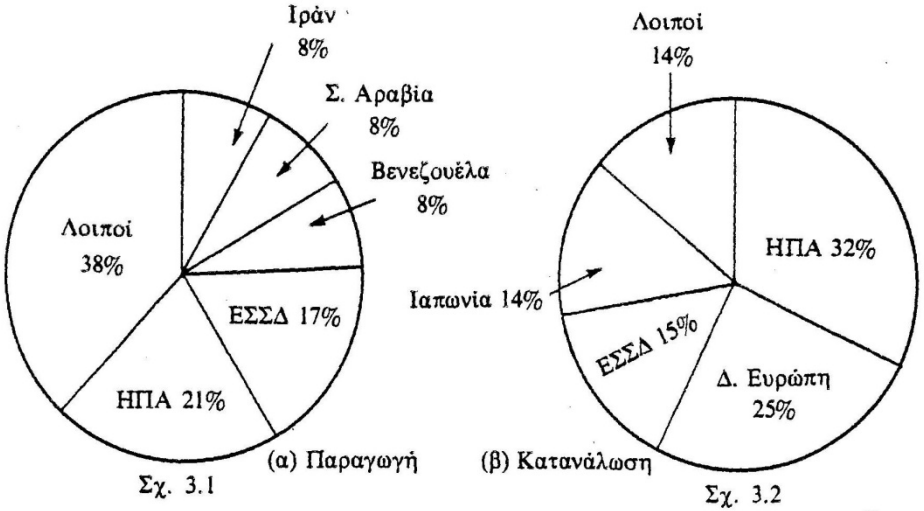
Παρουσίαση Ποιοτικών Δεδομένων

1. ραβδογράμματα (Bar Chart)
2. τομεογράμματα (Pie Chart)
3. Πίνακες συνάφειας (Contigency Tables)



(α) Ραβδόγραμμα αποφοίτων

Παράδειγμα 3.1. Τα δύο παρακάτω διαγράμματα παριστάνουν την παραγωγή (α) και την κατανάλωση (β) πετρελαίου στην πενταετία 1966 - 1970



(α) Παραγωγή

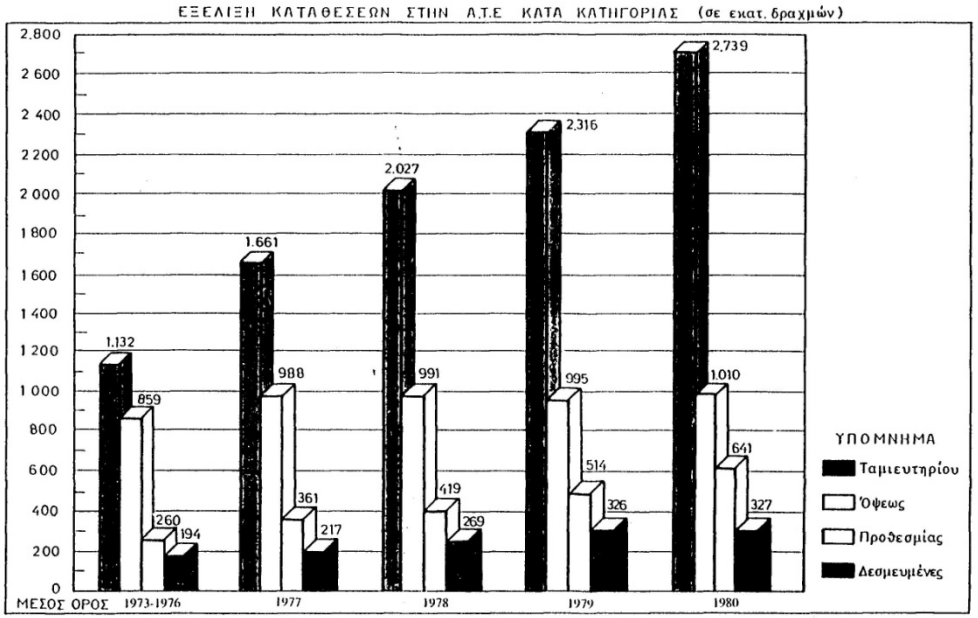
Σχ. 3.1

(β) Κατανάλωση

Σχ. 3.2

Από τους 100 φοιτητές μιας αίθουσας 30 είναι αγόρια και 70 είναι κορίτσια. Από αυτούς 40 είναι καπνιστές εκ των οποίων 20 είναι κορίτσια. Να δοθεί πίνακας συνάφειας των δεδομένων.

	Αγόρια	Κορίτσια	Σύνολο
Καπνιστές	20	20	40
Μη Καπνιστές	10	50	60
Σύνολο	30	70	100



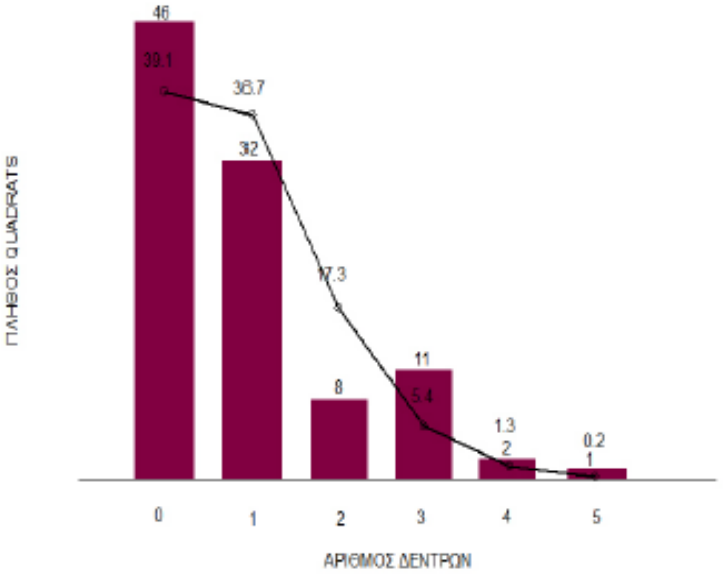
Περιγραφική Στατιστική (διακριτό χαρακτηριστικό)

Το πλήθος των δέντρων της ποικιλίας Lacistema aggregatum σε 100 συστηματικά τοποθετημένα γειτονικά τετράγωνα γης "quadrats" σε ένα τροπικό δάσος.

Πλήθος δέντρων	0	1	2	3	4	5	Σύνολο
Πλήθος τετραγώνων	46	32	8	11	2	1	100

Lacistema aggregatum

quadrats



$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 0.94$$

Ραβδόγραμμα

Ζητήθηκε από 30 φοιτητές να κάνουν όσο το δυνατόν περισσότερες έλξεις στο μονόζυγο.
Τα αποτελέσματα ήταν τα εξής:

10 11 5 9 7 6 8 6 5 8 10 7 7 8 5 6 4 7 9 7 4 8 10 11 7 4 9 5 8 9

Μια πρώτη ενέργεια είναι να διατάξουμε τις παρατηρήσεις:

4 4 4 5 5 5 5 6 6 6 7 7 7 7 7 7 8 8 8 8 8 9 9 9 9 10 10 10 11 11

Το πλήθος των φορών που εμφανίζεται μια συγκεκριμένη τιμή, x, του χαρακτηριστικού ονομάζεται συχνότητα, και συμβολίζεται με f_x.
Το πλήθος των παρατηρήσεων με τιμή ≤ x, ονομάζεται αθροιστική συχνότητα, και συμβολίζεται με F_x.

Εάν $x_1 < x_2 < \ldots < x_k$ οι διαφορετικές τιμές του χαρακτηριστικού που εμφανίστηκαν σε δείγμα μεγέθους n και συμβολίζουμε με f_i και F_i την αντίστοιχη συχνότητα και αθροιστική συχνότητα τότε ισχύει ότι:

$$\sum_{i=1}^{k} f_i = n$$

$$F_i = f_1 + f_2 + \ldots + f_{i-1} + f_i$$
$$= F_{i-1} + f_i$$

Εδώ παίρνουμε τον εξής πίνακα:

Τιμή x_i	4	5	6	7	8	9	10	11	Σύνολο
Συχνότητα f_i	3	4	3	6	5	4	3	2	30
Αθροιστική Συχνότητα F_i	3	7	10	16	21	25	28	30	

Περιγραφική Στατιστική (συνεχές χαρακτηριστικό)

Ύψη 45 φοιτητών

175.7	173.6	176.4	173.4	171.2	175.3	173.6	175.7	173.1	170.3
171.9	174.2	172.8	178.7	174.6	174.4	174.6	174.3	173.2	171.3
165.7	161.8	163.8	163.8	165.2	166.1	167.6	165.4	161.9	166.2
163.9	166.9	162.5	162.8	165.2	164.3	166.8	158.8	167.7	167.2
168.3	166.0	169.9	169.3	162.4					

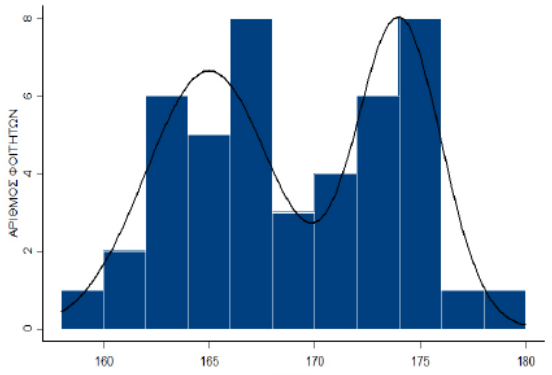
Διατεταγμένο δείγμα

158.8	161.8	161.9	162.4	162.5	162.8	163.8	163.8	163.9	164.3
165.2	165.2	165.4	165.7	166.0	166.1	166.2	166.8	166.9	167.2
167.6	167.7	168.3	169.3	169.9	170.3	171.2	171.3	171.9	172.8
173.1	173.2	173.4	173.6	173.6	174.2	174.3	174.4	174.6	174.6
175.3	175.7	175.7	176.4	178.7					

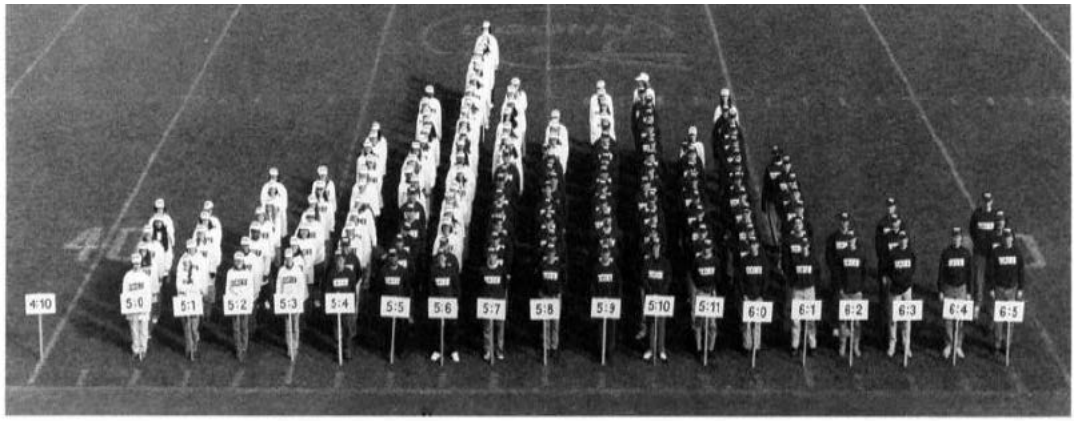
Q_1 Q_2 Q_3

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 169.06\text{cm}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = 24.64 = 4.96^2\text{cm}^2$$



Ιστόγραμμα συχνοτήτων των υψών 45 φοιτητών



Σχήμα : "Ζωντανό" ιστόγραμμα των υψών 143 φοιτητών στο Πανεπιστήμιο του Connecticut

Δείγμα από τη συγκέντρωση ενός συγκεκριμένου ρύπου (σε mgr/cm^3) σε δείγματα αέρος από 57 πόλεις

68 63 42 27 30 36 28 32 79 27 22 23 24 25 24

65 43 25 74 51 36 42 28 31 28 25 45 12 57 51

12 32 49 38 42 27 31 50 38 21 16 24 69 47 23

22 43 27 49 48 23 12 19 46 30 49 49

Διατεταγμένο δείγμα

12 12 12 16 19 21 22 22 23 23 23 24 24 24 25

25 25 27 27 27 27 28 28 28 30 30 31 31 32 32

36 36 38 38 42 42 42 43 43 45 46 47 48 49 49

49 49 50 51 51 57 63 65 68 69 74 79

Εμπειρικός τύπος για τον υπολογισμό του πλήθους και του εύρους των κλάσεων

$k=1+3.32*\log_{10}(n)= 1+1.44*\ln(n)$

$c=(x_{max}-x_{min})/k$

Εδώ $k=1+3.32*\log_{10}(57)= 1+3.32/\ln(10)*\ln(57)$

$= 6.8219\sim7$

$c=(x_{max}-x_{min})/k=(79-12)/7= 9.5714\sim10$

Όρια κλάσης L_i -U_i	Κεντρική τιμή x_i	Συχνότητα f_i	Αθροιστική Συχνότητα F_i
[10-20)	15	5	5
[20-30)	25	19	24
[30-40)	35	10	34
[40-50)	45	13	47
[50-60)	55	4	51
[60-70)	65	4	55
[70-80]	75	2	57
		57	

Μέτρα Θέσης και Διασποράς

Για την επεξεργασία των δεδομένων χρειάζονται ορισμένα χαρακτηριστικά στατιστικά μέτρα που καλούνται παράμετροι της κατανομής του πληθυσμού.

Οι παράμετροι καθορίζουν

1. Θέση
2. Διασπορά
3. Τη μορφή της κατανομής

Μέτρα Θέσης ή Κεντρικής Τάσης

Μέση Τιμή (Mean)

- Για δείγμα n παρατηρήσεων με τιμές $x_1, x_2, \dots, x_n$ η μέση τιμή είναι $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Αν έχουμε τις τιμές $x_1, x_2, \dots, x_k$ με αντίστοιχες συχνότητες $f_1, f_2, \dots, f_k$ τότε $\bar{x} = \sum_{i=1}^{k} x_i f_i \Big/ \sum_{i=1}^{k} f_i$
- Σε ομαδοποιημένες παρατηρήσεις, γνωρίζω μόνο ότι ανάμεσα στα όρια της i-κλάσης είναι f_i παρατηρήσεις χωρίς να ξέρω τις αντίστοιχες τιμές τους. Θεωρώ ότι όλες οι παρατηρήσεις έχουν την κεντρική τιμή της κλάσης η οποία ορίζεται ως το ημιάθροισμα των δύο άκρων της κλάσης

Διάμεσος (Median) δ

- Ενός πεπερασμένου συνόλου τιμών μιας μεταβλητής ή παρατηρήσεων είναι εκείνη η τιμή που χωρίζει το σύνολο σε δύο ίσα μέρη

  Δηλ. 50% παρατηρήσεων $\geq \delta$

  $\leq \delta$
- Διατάσσουμε τις παρατηρήσεις κατά αύξουσα σειρά μεγέθους. Εάν το πλήθος των παρατηρήσεων n είναι περιττό τότε η διάμεσος είναι η μεσαία τιμή. Εάν το n είναι άρτιο δ = ημιάθροισμα των δύο μεσαίων παρατηρήσεων
- Σε ομαδοποιημένες παρατηρήσεις η διάμεσος δίνεται από τον τύπο

$$\delta = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} \cdot c$$

L_i: το κάτω αληθινό όριο της κλάσης που βρίσκεται η διάμεσος

n: πλήθος παρατηρήσεων

$F_{i-1} = \sum_{j=1}^{i-1} f_j$ το άθροισμα των συχνοτήτων μέχρι την προηγούμενη κλάση

c: εύρος της κλάσης

με i τέτοιο ώστε $F_{i-1} < n/2 \leq F_i$

Το βάρος (σε mgr) της δραστικής ουσίας σε 7 διαφορετικά δισκία ενός φαρμάκου είναι:

1.4 2.0 2.4 1.2 1.6 1.8 2.2

Να βρεθεί η μέση τιμή, η διάμεσος και η τυπική απόκλιση της δραστικής ουσίας

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1.4+2.0+\ldots+1.8+2.2}{7} = \frac{12.6}{7} = 1.8 \text{ mgr}$$

Διατάσσουμε τις παρατηρήσεις

1.2 1.4 1.6 1.8 2.0 2.2 2.4

και επειδή το πλήθος τους είναι περιττό

$$\delta = x_{(4)}$$

επίσης, βλέπετε σελ. 10

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{2\times(0.2^2+0.4^2+0.6^2)}{6}$$

$$= 0.1866 = (0.432)^2 \text{ mgr}^2 \Rightarrow$$

$$S = 0.432 \text{ mgr}$$

Το πλήθος των δένδρων σε 40 περιοχές αστικού πρασίνου, εκτάσεως 10 m^2 η κάθε μία, δίδονται στον επόμενο πίνακα:

πλήθος δένδρων	1	2	3	4	5
περιοχές πρασίνου	4	10	14	8	4

Να βρεθούν το μέσο πλήθος δένδρων ανά περιοχή πρασίνου και η διασπορά.

$$\bar{x} = \sum_{i=1}^{k} \frac{x_i f_i}{\sum_{i=1}^{k} f_i} = \frac{1\times 4+2\times 10+\ldots+5\times 4}{4+10+\ldots+4} = \frac{118}{40} = 2.95$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i-\bar{x})^2 f_i$$

$$= \frac{(1-2.95)^2\times 4+(2-2.95)^2\times 10+\ldots+(5-2.95)^2\times 4}{39}$$

$$= \frac{49.9}{39} = 1.279$$

Ποια είναι η διάμεσος και η κορυφή; (βλέπετε σελ. 9)

$$\delta = \frac{x_{(20)}+x_{(21)}}{2} = 3$$

$M_0 = 3$ αφού $f_3 = 14$ μέγιστη συχνότητα

Κορυφή, τύπος ή επικρατούσα τιμή M_0

ορίζεται σαν η τιμή ευείαν της μεταβλητής στην οποία αντιστοιχεί η μεγαλύτερη συχνότητα των παρατηρήσεων

- Σε ομαδοποιημένες παρατηρήσεις η κορυφή δίνεται από τον τύπο

$$M_0 = L_i + \frac{\Delta_1}{\Delta_1+\Delta_2} \cdot c$$

$\Delta_1 = f_i - f_{i-1}$

$\Delta_2 = f_i - f_{i+1}$

με i τέτοιο ώστε $f_i > f_j$ για κάθε $j \neq i$

Εκατοστιαία σημεία (Percentiles)
ή ποσοστημόρια
Δεκατημόρια (Deciles)
Τεταρτημόρια (Quartiles)

Το κ-στο εκατοστημόριο έστω p_κ ενός συνόλου τιμών μιας μεταβλητής, είναι η τιμή ευείαν της μεταβλητής ώστε το κ% των παρατηρήσεων $\leq p_\kappa$

$(100-\kappa)\%$ —»— $\geq p_\kappa$

- Για ομαδοποιημένες παρατηρήσεις

$$p_\kappa = L_i + \left(\frac{\frac{\kappa n}{100} - F_{i-1}}{f_i}\right) c$$

με i τέτοιο ώστε $F_{i-1} < k*n/100 \leq F_i$

- 1ο τεταρτημόριο $Q_1 = L_i + \left(\frac{\frac{n}{4} - F_{i-1}}{f_i}\right) c$
- 2ο τεταρτημόριο (διάμεσος) $\delta = L_i + \left(\frac{\frac{n}{2} - F_{i-1}}{f_i}\right) c$
- 3ο τεταρτημόριο $Q_3 = L_i + \left(\frac{\frac{3n}{4} - F_{i-1}}{f_i}\right) c$

Το σημαντικότερο από τα μέτρα θέσης είναι ο μέσος γιατί χρησιμοποιώ για τον υπολογισμό του όλες τις πληροφορίες του δείγματος. Επηρεάζεται όμως πολύ από τις ακραίες τιμές.

Η διάμεσος και η κορυφή δεν επηρεάζονται από ακραίες τιμές, όμως δεν χρησιμοποιούν στον υπολογισμό τους όλες τις δεδομένες τιμές.

Μέτρα Διασποράς

I	1	3	10	17	19
II	4	9	10	16	16
III	8	9	10	11	12

Ενώ οι πληθυσμοί I, II, III έχουν τους ίδιους μέσους και διαμέσους, υπάρχει διαφορετική μεταβλητότητα, κατάληξη από τη μέση τιμή. Μέτρα που μετρούν αυτή τη μεταβλητότητα είναι

- Κύμανση – Εύρος $R = X_{max} - X_{min}$
- Ενδοτεταρτημοριακή απόκλιση $Q_3 - Q_1$

- **Διασπορά, τυπική απόκλιση**

Διασπορά ή Διακύμανση

Για δείγμα n παρατηρήσεων με μέσο $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ η διασπορά S^2 δίδεται από τον τύπο:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right\}$$

Ο συντελεστής $\frac{1}{n-1}$ χρησιμοποιείται για να έχει η διασπορά την ιδιότητα ως αμερόληπτη δηλ. $E(S^2) = \sigma^2$

Σε ομαδοποιημένες παρατηρήσεις

x_i : κεντρική τιμή κλάσεως

f_i : συχνότητα (# παρατηρήσεων) στην i κλάση

$$S^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{k} x_i^2 f_i - \frac{(\sum_{i=1}^{k} x_i f_i)^2}{n}\right\}$$

n: μέγεθος δείγματος

k: πλήθος των κλάσεων

Το S ονομάζεται τυπική απόκλιση

• Η κατανομή του συριγμού όξεος σε 267 υγιείς άρρενες βρέθηκε (σε mg/100ml)

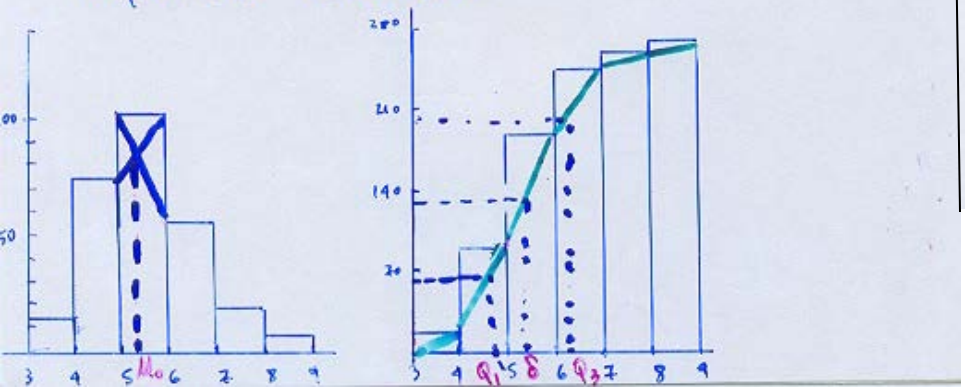
Συριγμώδη όξη	Συχνότητα f_i	Αθρ. σωρ. συχν. F_i	Κεντρ. τιμή x_i	$x_i f_i$	$x_i^2 f_i$
3.0 – 4.0	17	17	3.5	59.5	208.25
4.0 – 5.0	73	90	4.5	328.5	1478.25
5.0 – 6.0	101	191	5.5	555.5	3055.25
6.0 – 7.0	54	245	6.5	351.0	2281.50
7.0 – 8.0	18	263	7.5	135.0	1012.50
8.0 – 9.0	4	267	8.5	34.0	289.00
	267			1463.5	8324.75

Να υπολογιστούν

α. i) η μέση τιμή και η διασπορά

ii) η διάμεσος, η κορυφή και το 1ο και 3ο τεταρτημόριο.

β. Να κατασκευαστούν τα ιστογράμματα συχνοτήτων και αθροιστικών συχνοτήτων.



α. i) $\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n} = \frac{\sum_{i=1}^{6} x_i f_i}{267} = \frac{1463.5}{267} = 5.48$

$$S^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{k} x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n}\right\}$$

$$= \frac{1}{266}\left\{8324.75 - \frac{(1463.5)^2}{267}\right\} = 1.13 > 0 \quad \text{πάντοτε}$$

ii) για τη διάμεσο δ $\quad \frac{n}{2} = 133.5$ βρίσκεται στην 3η τάξη

$$\delta = L_3 + \frac{\frac{n}{2} - F_2}{f_3} \cdot c = 5 + \frac{133.5 - 90}{101} \cdot 1 = 5.43$$

για την κορυφή M_0 $\quad \max\{f_i\} = 101$ στην 3η τάξη

$$M_0 = L_3 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot c = 5 + \frac{28}{28+47} \cdot 1 = 5.37$$

όπου $\Delta_1 = f_3 - f_2 = 101 - 73 = 28$

$\Delta_2 = f_3 - f_4 = 101 - 54 = 47$

για το Q_1 $\quad \frac{n}{4} = 66.75$ στην 2η τάξη

$$Q_1 = L_2 + \frac{\frac{n}{4} - F_1}{f_2} \cdot c = 4 + \frac{66.75 - 17}{73} \cdot 1 = 4.68$$

για το Q_3 $\quad \frac{3n}{4} = 200.25$ στην 4η τάξη

$$Q_3 = L_4 + \frac{\frac{3n}{4} - F_3}{f_4} \cdot c = 6 + \frac{200.25 - 191}{54} \cdot 1 = 6.17$$