Πληθυσμός - Μεταβλητές

Όπως αναφέρθηκε και προηγουμένως, αυτό που μας ενδιαφέρει είναι να εξετάσουμε τα στοιχεία ενός συνόλου ως προς ένα ή περισσότερα χαρακτηριστικά τους. Αυτό συμβαίνει, για παράδειγμα, όταν ενδιαφερόμαστε για:

α) τις προτιμήσεις των ψηφοφόρων εν όψει των προσεχών εκλογών
β) τον αριθμό των υπαλλήλων μιας επιχείρησης
γ) το ύψος, το βάρος, την ομάδα αίματος και το φύλο των μαθητών της Γ΄τάξης Λυκείου
δ) τις συνέπειες του καπνίσματος στην υγεία των καπνιστών κτλ.

Σε καθένα από τα παραδείγματα αυτά έχουμε ένα σύνολο και θέλουμε να εξετάσουμε τα στοιχεία του ως προς ένα ή περισσότερα χαρακτηριστικά τους. Ένα τέτοιο σύνολο λέγεται **πληθυσμός** (population). Τα στοιχεία του πληθυσμού συχνά αναφέρονται και ως μονάδες ή άτομα του πληθυσμού. Στο πρώτο παράδειγμα έχουμε το σύνολο των ψηφοφόρων και μας ενδιαφέρει η προτίμησή τους, ποιο "κόμμα" π.χ. υποστηρίζουν. Στο τρίτο παράδειγμα έχουμε το σύνολο των μαθητών της Γ΄ Λυκείου και μας ενδιαφέρουν τα τέσσερα χαρακτηριστικά τους: ύψος, βάρος, ομάδα αίματος και φύλο.
Τα χαρακτηριστικά ως προς τα οποία εξετάζουμε έναν πληθυσμό λέγονται **μεταβλητές** (variables) και τις συμβολίζουμε συνήθως με τα κεφαλαία γράμματα X , Y , Z, B, ... Οι δυνατές τιμές που μπορεί να πάρει μια μεταβλητή λέγονται **τιμές της μεταβλητής**. Από τη διαδοχική εξέταση των ατόμων του πληθυσμού ως προς ένα χαρακτηριστικό τους προκύπτει μια σειρά από δεδομένα, που λέγονται στατιστικά δεδομένα ή παρατηρήσεις. Τα στατιστικά δεδομένα δεν είναι κατ'ανάγκη διαφορετικά. Για παράδειγμα, αν εξετάζουμε την ομάδα αίματος δέκα ατόμων, τα στατιστικά δεδομένα ή παρατηρήσεις που θα προκύψουν μπορεί να είναι: A, A, B, A, AB, O, AB, AB, AB, O, B. Οι δυνατές όμως τιμές που μπορεί να πάρει η μεταβλητή "ομάδα αίματος" είναι οι εξής τέσσερις: A, B, AB και O.
Τις μεταβλητές τις διακρίνουμε:

1. Σε **ποιοτικές** ή **κατηγορικές** μεταβλητές, των οποίων οι τιμές τους δεν είναι αριθμοί. Τέτοιες είναι, για παράδειγμα, η ομάδα αίματος (με τιμές A, B, AB, O), το φύλο (με τιμές αγόρι, κορίτσι), οι συνέπειες του καπνίσματος (με τιμές καρδιακά νοσήματα, καρκίνος κτλ), όπως επίσης και η οικονομική κατάσταση και η υγεία των ανθρώπων (που μπορεί να χαρακτηριστεί ως κακή, μέτρια, καλή ή πολύ καλή), καθώς και το ενδιαφέρον των μαθητών για τη Στατιστική, που μπορεί να χαρακτηριστεί ως υψηλό, μέτριο, χαμηλό ή μηδαμινό.

2. Σε **ποσοτικές** μεταβλητές, των οποίων οι τιμές είναι αριθμοί και διακρίνονται:
   i) Σε **διακριτές** μεταβλητές, που παίρνουν μόνο "μεμονωμένες" τιμές.
   Τέτοιες μεταβλητές είναι, για παράδειγμα, ο αριθμός των υπαλλήλων μιας επιχείρησης (με τιμές 1,2,…), το αποτέλεσμα της ρίψης ενός ζαριού (με τιμές 1,2,…,6) κτλ.
   ii) Σε **συνεχείς** μεταβλητές, που μπορούν να πάρουν οποιαδήποτε τιμή ενός διαστήματος πραγματικών αριθμών (α, β). Τέτοιες μεταβλητές είναι το ύψος και το βάρος των μαθητών της Γ΄ Λυκείου, ο χρόνος που χρειάζονται οι μαθητές να απαντήσουν στα θέματα μιας εξέτασης, η διάρκεια μιας τηλεφωνικής συνδιάλεξης κτλ.

Πίνακες Κατανομής Συχνοτήτων

Ας υποθέσουμε ότι $x_1, x_2,..., x_\kappa$ είναι οι τιμές μιας μεταβλητής X, που αφορά τα άτομα ενός δείγματος μεγέθους ν, $\kappa \leq \nu$. Στην τιμή x_i αντιστοιχίζεται η (απόλυτη) **συχνότητα** (frequency) ν_i, δηλαδή ο φυσικός αριθμός που δείχνει πόσες φορές εμφανίζεται η τιμή x_i της εξεταζόμενης μεταβλητής X στο σύνολο των παρατηρήσεων. Είναι φανερό ότι το άθροισμα όλων των συχνοτήτων είναι ίσο με το μέγεθος ν του δείγματος, δηλαδή:

$$\nu_1 + \nu_2 + ... + \nu_\kappa = \nu \qquad (1)$$

Για παράδειγμα, για τη μεταβλητή X: "αριθμός αδελφών" του πίνακα 4 οι συχνότητες για τις τιμές $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$ είναι, αντίστοιχα, $\nu_1 = 8$, $\nu_2 = 22$, $\nu_3 = 7$, $\nu_4 = 3$ με $\nu_1 + \nu_2 + \nu_3 + \nu_4 = 40$. Ο υπολογισμός των συχνοτήτων γίνεται με τη **διαλογή** των παρατηρήσεων, όπως φαίνεται στον παρακάτω πίνακα 5. Διατρέχοντας με τη σειρά τη λίστα των δεδομένων καταγράφουμε κάθε παρατήρηση με συμβολικό τρόπο σαν μια γραμμή " | " στην αντίστοιχη τιμή της μεταβλητής.

Πίνακας 5

Κατανομή συχνοτήτων της μεταβλητής X: "αριθμός αδελφών" των μαθητών του πίνακα 4.

Αριθμός αδελφών x_i	Διαλογή	Συχνότητα ν_i	Σχετική Συχνότητα f_i	Σχετική Συχνότητα f_i %
0	𝍸 |||	8	0,200	20,0
1	𝍸 𝍸 𝍸 |	22	0,550	55,0
2	|||| ||	7	0,175	17,5
3	|||	3	0,075	7,5
Σύνολο		40	1,000	100,00

Αν διαιρέσουμε τη συχνότητα ν_i με το μέγεθος ν του δείγματος, προκύπτει η **σχετική συχνότητα** (relative frequency) f_i της τιμής x_i, δηλαδή

$$f_i = \frac{\nu_i}{\nu}, \quad i = 1,2,...,\kappa. \qquad (2)$$

Για τη σχετική συχνότητα ισχύουν οι ιδιότητες:

(i) $0 \le f_i \le 1$ για $i = 1,2,...,\kappa$ αφού $0 \le \nu_i \le \nu$.

(ii) $f_1 + f_2 + ... + f_\kappa = 1$, αφού

$$f_1 + f_2 + ... + f_\kappa = \frac{\nu_1}{\nu} + \frac{\nu_2}{\nu} + ... + \frac{\nu_\kappa}{\nu} = \frac{\nu_1 + \nu_2 + ... + \nu_\kappa}{\nu} = \frac{\nu}{\nu} = 1.$$

Συνήθως, τις σχετικές συχνότητες f_i τις εκφράζουμε επί τοις εκατό, οπότε συμβολίζονται με f_i %, δηλαδή f_i % $=100f_i$. Για παράδειγμα, οι σχετικές συχνότητες για τις τιμές $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$ της μεταβλητής X: "αριθμός αδελφών΄" είναι αντιστοίχως:

$$f_1 = \frac{8}{40} = 0{,}20, \quad f_2 = \frac{22}{40} = 0{,}55, \quad f_3 = \frac{7}{40} = 0{,}175 \quad \text{και} \quad f_4 = \frac{3}{40} = 0{,}075 \quad \text{με}$$
$$f_1 + f_2 + f_3 + f_4 = 0{,}20 + 0{,}55 + 0{,}175 + 0{,}075 = 1.$$

Συνεπώς f_1% = 20% , f_2% = 55%, f_3% =17,5% και f_4% = 7,5% με f_1% + f_2% + f_3% + f_4% = 100%.
Οι ποσότητες x_i, ν_i, f_i για ένα δείγμα συγκεντρώνονται σε ένα συνοπτικό πίνακα, που ονομάζεται **πίνακας κατανομής συχνοτήτων** ή απλά **πίνακας συχνοτήτων**.
Για μια μεταβλητή, το σύνολο των ζευγών (x_i, ν_i) λέμε ότι αποτελεί την **κατανομή συχνοτήτων** και το σύνολο των ζευγών (x_i, f_i), ή των ζευγών (x_i, f_i%), την **κατανομή των σχετικών συχνοτήτων**.

Αθροιστικές Συχνότητες

Στην περίπτωση των **ποσοτικών μεταβλητών** εκτός από τις συχνότητες ν_i και f_i χρησιμοποιούνται συνήθως και οι λεγόμενες **αθροιστικές συχνότητες** (cumulative frequencies) N_i και οι **αθροιστικές σχετικές συχνότητες** (cumulative relative frequencies) F_i, οι οποίες εκφράζουν το πλήθος και το ποσοστό αντίστοιχα των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής x_i. Συχνά οι F_i πολλαπλασιάζονται επί 100 εκφραζόμενες έτσι επί τοις εκατό, δηλαδή F_i% = $100F_i$, βλέπε πίνακα 6. Αν οι τιμές x_1, x_2,..., μιας ποσοτικής μεταβλητής X είναι σε αύξουσα διάταξη, τότε η αθροιστική συχνότητα της τιμής x_i είναι $N_i = \nu_1 + \nu_2 + ... + \nu_i$. Όμοια, η αθροιστική σχετική συχνότητα είναι $F_i = f_1 + f_2 + ... + f_i$, για $i = 1,2,...,\kappa$. Για παράδειγμα, για τη μεταβλητή X: "αριθμός αδελφών" του πίνακα 4 είναι $N_1 = \nu_1 = 8$, $N_2 = \nu_1 + \nu_2 = 30$, $N_3 = \nu_1 + \nu_2 + \nu_3 = 37$ και $N_4 = \nu_1 + \nu_2 + \nu_3 + \nu_4 = \nu = 40$, οπότε
$F_1 = f_1 = 0{,}20$, $F_2 = f_1 + f_2 = 0{,}75$, $F_3 = f_1 + f_2 + f_3 = 0{,}925$ και $F_4 = f_1 + f_2 + f_3 + f_4 = 1$, οπότε F_1%=20%, F_2%=75% , F_3%=92,5% και F_4%=100%. Είναι φανερό ότι ισχύουν οι σχέσεις:

$$\nu_1 = N_1 \,,\ \nu_2 = N_2 - N_1 \,,..., \nu_\kappa = N_\kappa - N_{\kappa-1}$$

$$f_1 = F_1,\ f_2 = F_2 - F_1 \,,..., f_\kappa = F_\kappa - F_{\kappa-1}.$$

Αριθμός αδελφών x_i	Συχνότητα ν_i	Σχετ. Συχν. f_i	Σχετ. Συχν. f_i%	Άθροισ. Συχν. N_i	Αθροιστική Σχετ. Συχν. F_i	Αθροιστική Σχετ. Συχν. F_i%
0	8	0,200	20,0	8	0,200	20,0
1	22	0,550	55,0	30	0,750	75,0
2	7	0,175	17,5	37	0,925	92,5
3	3	0,075	7,5	40	1,000	100,0
Σύνολο	40	1,000	100,0	-	-	-

Γραφική Παράσταση Κατανομής Συχνοτήτων

Τα στατιστικά δεδομένα παρουσιάζονται πολλές φορές και υπό μορφή γραφικών παραστάσεων ή διαγραμμάτων. Οι γραφικές παραστάσεις παρέχουν πιο σαφή εικόνα του χαρακτηριστικού σε σχέση με τους πίνακες, είναι πολύ πιο ενδιαφέρουσες και ελκυστικές, χωρίς βέβαια να προσφέρουν περισσότερη πληροφορία από εκείνη που περιέχεται στους αντίστοιχους πίνακες συχνοτήτων. Επί πλέον με τα διαγράμματα διευκολύνεται η σύγκριση μεταξύ ομοειδών στοιχείων για το ίδιο ή για διαφορετικά χαρακτηριστικά. Υπάρχουν διάφοροι τρόποι γραφικής παρουσίασης, ανάλογα με το είδος των δεδομένων που έχουμε. Όπως όμως οι στατιστικοί πίνακες έτσι και τα στατιστικά διαγράμματα πρέπει να συνοδεύονται από α) τον τίτλο, β) την κλίμακα με τις τιμές των μεγεθών που απεικονίζονται, γ) το υπόμνημα που επεξηγεί συνήθως τις τιμές της μεταβλητήςκαι δ) την πηγή των δεδομένων.

α) Ραβδόγραμμα

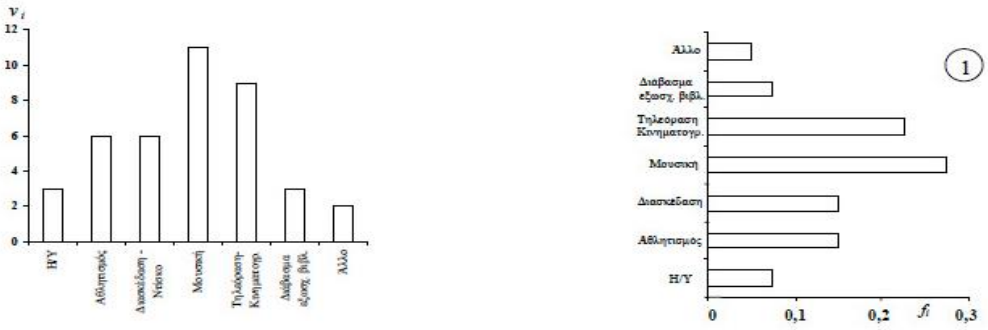
Το **ραβδόγραμμα** (barchart) χρησιμοποιείται για τη γραφική παράσταση των τιμών μιας ποιοτικής μεταβλητής. Το ραβδόγραμμα αποτελείται από ορθογώνιες στήλες που οι βάσεις τους βρίσκονται πάνω στον οριζόντιο ή τον κατακόρυφο άξονα. Σε κάθε τιμή της μεταβλητής Χ αντιστοιχεί μια ορθογώνια στήλη της οποίας το ύψος είναι ίσο με την αντίστοιχη συχνότητα ή σχετική συχνότητα. Έτσι έχουμε αντίστοιχα το **ραβδόγραμμα συχνοτήτων** και το **ραβδόγραμμα σχετικών** συχνοτήτων. Τόσο η απόσταση μεταξύ των στηλών όσο και το μήκος των βάσεών τους καθορίζονται αυθαίρετα. Στον πίνακα 7 έχουμε την κατανομή συχνοτήτων της μεταβλητής X: "απασχόληση στον ελεύθερο χρόνο" και στα σχήματα 1(α), (β) τα αντίστοιχα ραβδογράμματα συχνοτήτων και σχετικών συχνοτήτων.

Πίνακας 7

Κατανομή συχνοτήτων για την απασχόληση στον ελεύθερο χρόνο τους των μαθητών του πίνακα 4.

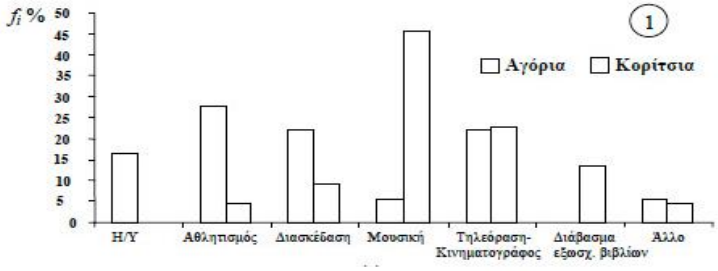
i	Απασχόληση x_i	Συχνότητα ν_i	Σχετική συχνότητα f_i	Σχετική συχνότητα f_i%
1	Υπολογιστές	3	0,075	7,5
2	Αθλητισμός	6	0,150	15,0
3	Διασκέδαση-ντίσκο	6	0,150	15,0
4	Μουσική	11	0,275	27,5
5	Τηλεόραση-Κινηματογράφος	9	0,225	22,5
6	Διάβασμα εξωσχ. Βιβλίων	3	0,075	7,5
7	Άλλο	2	0,050	5,0
Σύνολο		40	1,000	100,0

Μερικές φορές σε ένα ραβδόγραμμα συχνοτήτων ο ρόλος των δύο αξόνων είναι δυνατόν να αντιστραφεί, όπως φαίνεται στο σχήμα 1(β), που παριστάνεται το ραβδόγραμμα σχετικών συχνοτήτων της ίδιας μεταβλητής. Αν θέλουμε να συγκρίνουμε τον τρόπο που περνούν τον ελεύθερο χρόνο τους τα αγόρια και τα κορίτσια, τότε κατασκευάζουμε το ραβδόγραμμα σχετικών συχνοτήτων του σχήματος 1(γ), όπως προκύπτει από τον πίνακα 4.



(α) (β)

Ραβδόγραμμα συχνοτήτων (α) και σχετικών συχνοτήτων (β) για την απασχόληση των μαθητών του πίνακα 7.

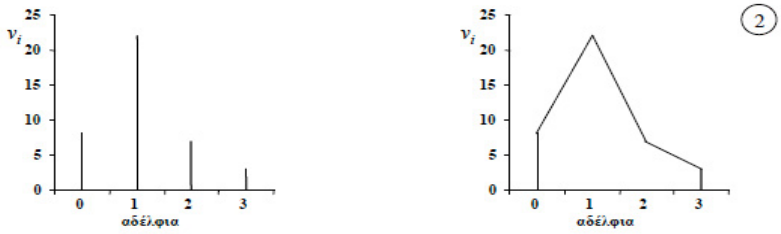


(γ)

Ραβδόγραμμα σχετικών συχνοτήτων για την απασχόληση των μαθητών του πίνακα 4 ανάλογα με το φύλο.

β) Διάγραμμα Συχνοτήτων

Στην περίπτωση που έχουμε μια ποσοτική μεταβλητή αντί του ραβδογράμματος χρησιμοποιείται το **διάγραμμα συχνοτήτων** (line diagram). Αυτό μοιάζει με το ραβδόγραμμα με μόνη διαφορά ότι αντί να χρησιμοποιούμε συμπαγή ορθογώνια υψώνουμε σε κάθε $\mathbf{x_i}$ (υποθέτοντας ότι $x_1 < x_2 <...< x_κ$) μία κάθετη γραμμή με μήκος ίσο προς την αντίστοιχη συχνότητα, όπως φαίνεται στο σχήμα 2(α). Μπορούμε επίσης αντί των συχνοτήτων v_i στον κάθετο άξονα να βάλουμε τις σχετικές συχνότητες $\mathbf{f_i}$, οπότε έχουμε το **διάγραμμα σχετικών συχνοτήτων**. Ενώνοντας τα σημεία (x_i, v_i) ή (x_i, f_i) έχουμε το λεγόμενο **πολύγωνο συχνοτήτων** ή **πολύγωνο σχετικών συχνοτήτων**, αντίστοιχα, που μας δίνουν μια γενική ιδέα για τη μεταβολή της συχνότητας ή της σχετικής συχνότητας όσο μεγαλώνει η τιμή της μεταβλητήςπου εξετάζουμε, βλέπε σχήμα 2(β).



(α) (β)

Διάγραμμα συχνοτήτων (α) και πολύγωνο συχνοτήτων (β) για τη μεταβλητή "αριθμός αδελφών" του πίνακα 4.

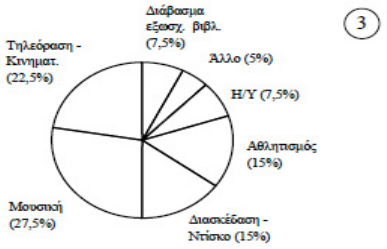
γ) Κυκλικό Διάγραμμα

Το **κυκλικό διάγραμμα** (piechart) χρησιμοποιείται για τη γραφική παράσταση τόσο των ποιοτικών όσο και των ποσοτικών δεδομένων, όταν οι διαφορετικές τιμές της μεταβλητής είναι σχετικά λίγες. Το κυκλικό διάγραμμα είναι ένας κυκλικός δίσκος χωρισμένος σε κυκλικούς τομείς, τα εμβαδά ή, ισοδύναμα, τα τόξα των οποίων είναι ανάλογα προς τις αντίστοιχες συχνότητες ν_i ή τις σχετικές συχνότητες f_i των τιμών x_i της μεταβλητής. Αν συμβολίσουμε με α_i το αντίστοιχο τόξο ενός κυκλικού τμήματος στο κυκλικό διάγραμμα συχνοτήτων, τότε

$$\alpha_i = \nu_i \frac{360^\circ}{\nu} = 360^\circ f_i \quad \text{για} \quad i=1,2,\ldots,\kappa .$$

Στο σχήμα 3 παριστάνεται το αντίστοιχο κυκλικό διάγραμμα σχετικών συχνοτήτων της "απασχόλησης των μαθητών" για τα δεδομένα του πίνακα 4.

Κυκλικό διάγραμμα σχετικών συχνοτήτων της απασχόλησης των μαθητών για τα δεδομένα του πίνακα 4.



(3)

Ομαδοποίηση των Παρατηρήσεων

Οι πίνακες συχνοτήτων και κατ' αναλογίαν τα αντίστοιχα διαγράμματα είναι δύσκολο να κατασκευαστούν, όταν το πλήθος των τιμών μιας μεταβλητής είναι αρκετά μεγάλο. Αυτό μπορεί να συμβεί είτε στην περίπτωση μιας διακριτής μεταβλητής είτε, πολύ περισσότερο, στην περίπτωση μιας συνεχούς μεταβλητής, όπου αυτή μπορεί να πάρει οποιαδήποτε τιμή στο διάστημα ορισμού της. Σ' αυτές τις περιπτώσεις είναι απαραίτητο να ταξινομηθούν (ομαδοποιηθούν) τα δεδομένα σε μικρό πλήθος ομάδων, που ονομάζονται και **κλάσεις** (class intervals), έτσι ώστε κάθε τιμή να ανήκει μόνο σε μία κλάση. Τα άκρα των κλάσεων καλούνται **όρια των κλάσεων** (class boundaries). Συνήθως υιοθετούμε την περίπτωση που μια κλάση περιέχει το κάτω άκρο της (κλειστή αριστερά) αλλά όχι το άνω άκρο της (ανοικτή δεξιά), δηλαδή που οι κλάσεις είναι της μορφής [,). Οι παρατηρήσεις κάθε κλάσης θεωρούνται όμοιες, οπότε μπορούν να "αντιπροσωπευθούν" από τις **κεντρικές τιμές**, τα κέντρα δηλαδή κάθε κλάσης.

- Το πρώτο βήμα στην ομαδοποίηση των δεδομένων είναι η εκλογή του αριθμού κ των ομάδων ή κλάσεων. Ο αριθμός αυτός συνήθως ορίζεται αυθαίρετα από τον ερευνητή σύμφωνα με την πείρα του. Γενικά όμως μπορεί να χρησιμοποιηθεί ως οδηγός ο παρακάτω πίνακας:

- Το πρώτο βήμα στην ομαδοποίηση των δεδομένων είναι η εκλογή του αριθμού κ των ομάδων ή κλάσεων. Ο αριθμός αυτός συνήθως ορίζεται αυθαίρετα από τον ερευνητή σύμφωνα με την πείρα του. Γενικά όμως μπορεί να χρησιμοποιηθεί ως οδηγός ο παρακάτω πίνακας:

Μέγεθος δείγματος ν	Αριθμός κλάσεων κ	Μέγεθος δείγματος ν	Αριθμός κλάσεων κ
<20	5	200-400	9
20-50	6	400-700	10
50-100	7	700-1000	11
100-200	8	≥1000	12

- Το δεύτερο βήμα είναι ο προσδιορισμός του πλάτους των κλάσεων. **Πλάτος μιας κλάσης** ονομάζεται η διαφορά του κατωτέρου από το ανώτερο όριο της κλάσης. Στην πλειονότητα των πρακτικών εφαρμογών οι κλάσεις έχουν το ίδιο πλάτος. Φυσικά υπάρχουν και περιπτώσεις όπου επιβάλλεται οι κλάσεις να έχουν άνισο πλάτος, όπως, για παράδειγμα, στις κατανομές εισοδήματος, ημερών απεργίας κτλ. Για να κατασκευάσουμε ισοπλατείς κλάσεις, χρησιμοποιούμε το **εύρος** (range) R του δείγματος, δηλαδή τη διαφορά της μικρότερης παρατήρησης από τη μεγαλύτερη παρατήρηση του συνολικού δείγματος. Τότε υπολογίζουμε το πλάτος c των κλάσεων διαιρώντας το εύρος R διά του αριθμού των κλάσεων κ, στρογγυλεύοντας, αν χρειαστεί για λόγους διευκόλυνσης, πάντα προς τα πάνω.

- Το επόμενο βήμα είναι η κατασκευή των κλάσεων. Ξεκινώντας από την μικρότερη παρατήρηση, ή για πρακτικούς λόγους λίγο πιο κάτω από την μικρότερη παρατήρηση, και προσθέτοντας κάθε φορά το πλάτος c δημιουργούμε τις κ κλάσεις. Αυτονόητο είναι ότι η μεγαλύτερη τιμή του δείγματος θα (πρέπει να) ανήκει οπωσδήποτε στην τελευταία κλάση.

- Τέλος, γίνεται η **διαλογή** των παρατηρήσεων. Το πλήθος των παρατηρήσεων ν_i που προκύπτουν από τη διαλογή για την κλάση i καλείται **συχνότητα της κλάσης** αυτής ή **συχνότητα της κεντρικής τιμής** x_i, i = 1,2,..., κ.
  Έστω, για παράδειγμα, ότι από τα δεδομένα του πίνακα 4 εξετάζουμε το ύψος των μαθητών. Το ύψος των μαθητών, όπως έχει καταγραφεί με τη σειρά, δίνεται στον παρακάτω πίνακα 8.

Πίνακας 8

Το ύψος (σε cm) των μαθητών της Γ' Λυκείου, όπως έχει καταγραφεί στον πίνακα 4.
Σε αγκύλες έχουμε τη μικρότερη και τη μεγαλύτερη τιμή.

170	180	178	165	170	168	175	175	173	162
160	170	167	177	180	170	182	178	165	178
[156]	175	172	173	167	187	170	180	178	[191]
176	169	167	166	179	178	180	164	170	173

Παρατηρούμε ότι το εύρος του δείγματος είναι $R = 191 - 156 = 35$. Επειδή έχουμε $\nu = 40$ παρατηρήσεις, χρησιμοποιούμε $\kappa = 6$ κλάσεις.

Το πλάτος των κλάσεων είναι $c = R / \kappa = 35 / 6 = 5{,}83 \approx 6$. Αν θεωρήσουμε ως αρχή της πρώτης κλάσης το 156, θα έχουμε τον επόμενο πίνακα 9.
Πρέπει να προσεχτεί ότι:

- Καμία παρατήρηση δεν μπορεί να μείνει έξω από κάποια κλάση.
- Οι κεντρικές τιμές διαφέρουν μεταξύ τους όσο και το πλάτος των κλάσεων, που εδώ είναι ίσο με 6.
- Μία παρατήρηση που συμπίπτει με το άνω άκρο μιας κλάσης θα τοποθετηθεί κατά τη διαλογή στην αμέσως επόμενη κλάση. Για παράδειγμα, ο μαθητής με ύψος 180 θα τοποθετηθεί στην πέμπτη κλάση [180, 186).

Πίνακας 9

Κατανομές συχνοτήτων (απόλυτων, σχετικών, αθροιστικών) για τα δεδομένα του πίνακα 8.

Κλάσεις [-)	Κεντρικές τιμές x_i	Διαλογή	Συχν. ν_i	Σχετική Συχνότητα f_i%	Αθρ. συχν. N_i	Αθρ. Σχετ. Συχν. F_i%
156-162	159	||	2	5,0	2	5,0
162-168	165	𝍸 |||	8	20,0	10	25,0
168-174	171	𝍸 𝍸 |	12	30,0	22	55,0
174-180	177	𝍸 |||| |	11	27,5	33	82,5
180-186	183	||||	5	12,5	38	95,0
186-192	189	||	2	5,0	40	100,0
	Σύνολο	-	40	100	-	-

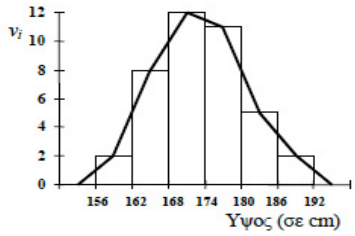
Ιστόγραμμα Συχνοτήτων

Η αντίστοιχη γραφική παράσταση ενός πίνακα συχνοτήτων με ομαδοποιημένα δεδομένα γίνεται με το λεγόμενο **ιστόγραμμα** (histogram) συχνοτήτων. Στον οριζόντιο άξονα ενός συστήματος ορθογωνίων αξόνων σημειώνουμε, με κατάλληλη κλίμακα, τα όρια των κλάσεων. Στη συνέχεια, κατασκευάζουμε διαδοχικά ορθογώνια (ιστούς), από καθένα από τα οποία έχει βάση ίση με το πλάτος της κλάσης και ύψος τέτοιο, ώστε το **εμβαδόν του ορθογωνίου να ισούται μετη συχνότητατης κλάσης αυτής**.

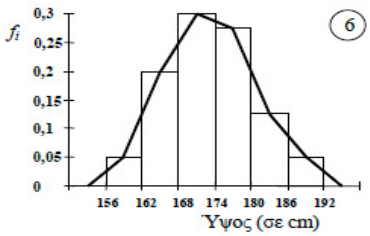
α) Κλάσεις Ίσου Πλάτους

Θεωρώντας το πλάτος c ως μονάδα μέτρησης του χαρακτηριστικού στον οριζόντιο άξονα, το ύψος κάθε ορθογωνίου είναι ίσο προς τη συχνότητα της αντίστοιχης κλάσης, έτσι ώστε να ισχύει πάλι ότι το εμβαδόν των ορθογωνίων είναι ίσο με τις αντίστοιχες συχνότητες.

Επομένως, στον κατακόρυφο άξονα σε ένα ιστόγραμμα συχνοτήτων βάζουμε τις συχνότητες. Με ανάλογο τρόπο κατασκευάζεται και το **ιστόγραμμα σχετικών συχνοτήτων**, οπότε στον κάθετο άξονα βάζουμε τις σχετικές συχνότητες.
Αν στα ιστογράμματα συχνοτήτων θεωρήσουμε δύο ακόμη υποθετικές κλάσεις, στην αρχή και στο τέλος, με συχνότητα μηδέν και στη συνέχεια ενώσουμε τα μέσα των άνω βάσεων των ορθογωνίων, σχηματίζεται το λεγόμενο **πολύγωνο συχνοτήτων** (frequency polygon). Το εμβαδόν του χωρίου που ορίζεται από το πολύγωνο συχνοτήτων και τον οριζόντιο άξονα είναι ίσο με το άθροισμα των συχνοτήτων, δηλαδή με το μέγεθος του δείγματος ν. Όμοια κατασκευάζεται από το ιστόγραμμα σχετικών συχνοτήτων και το **πολύγωνο σχετικών συχνοτήτων** με εμβαδόν ίσο με 1, (βλέπε σχήμα 6).



(α)

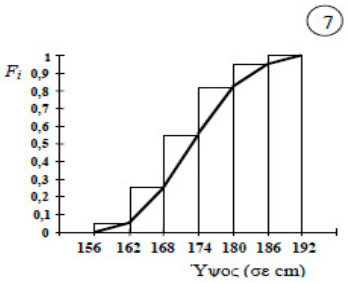


6

(β)

Ιστόγραμμα και πολύγωνο (α) συχνοτήτων και (β) σχετικών συχνοτήτων για τα δεδομένα του πίνακα 9.

Με τον ίδιο τρόπο κατασκευάζονται και τα **ιστογράμματα αθροιστικών συχνοτήτων** και **αθροιστικών σχετικών συχνοτήτων**. Αν ενώσουμε σε ένα ιστόγραμμα αθροιστικών συχνοτήτων τα **δεξιά άκρα** (όχι μέσα) των άνω βάσεων των ορθογωνίων με ευθύγραμμα τμήματα βρίσκουμε το **πολύγωνο αθροιστικών συχνοτήτων** (ogive) της κατανομής. Στο σχήμα 7 παριστάνεται το ιστόγραμμα και το πολύγωνο αθροιστικών σχετικών συχνοτήτων για το ύψος των μαθητών του πίνακα 9.



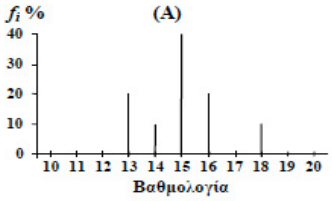
7

Εισαγωγή

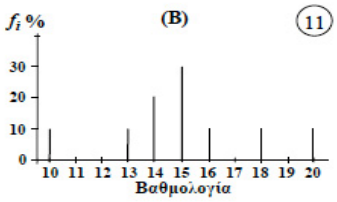
Εκτός από τους στατιστικούς πίνακες και τα διαγράμματα υπάρχουν και αριθμητικά μέτρα με τα οποία μπορούμε να περιγράψουμε με συντομία μια κατανομή συχνοτήτων. Η γνώση των μέτρων αυτών διευκολύνει και την παραπέρα στατιστική επεξεργασία των δεδομένων. Έστω, για παράδειγμα, ένας καθηγητής ο οποίος, για να συγκρίνει δύο διαφορετικά τμήματα Α και Β της ίδιας τάξης ως προς την επίδοσή τους σε ένα μάθημα, πήρε τυχαία 10 μαθητές από κάθε τμήμα. Η βαθμολογία τους στο μάθημα αυτό ήταν:

Τμήμα Α: 13 13 14 15 15 15 15 16 16 18
Τμήμα Β: 10 13 14 14 15 15 15 16 18 20.

Τα διαγράμματα σχετικών συχνοτήτων δίνονται στα σχήματα 11(α), (β).



(α)



(β)

11

Παρατηρούμε ότι η βαθμολογία και των δύο τμημάτων είναι συγκεντρωμένη γύρω στο 15, αλλά το δεύτερο τμήμα παρουσιάζει μεγαλύτερη διασπορά βαθμών από το πρώτο. Δηλαδή, οι βαθμοί του Β΄ τμήματος είναι περισσότερο διασκορπισμένοι γύρω από μια "κεντρική" τιμή. Οι έννοιες "κεντρική τιμή" και "διασπορά των παρατηρήσεων" μας δίνουν το ερέθισμα για έναν ακόμα πιο σύντομο τρόπο περιγραφής της κατανομής ενός συνόλου δεδομένων. Για να ορίσουμε δηλαδή κάποια **μέτρα** (αριθμητικά μεγέθη), που να μας δίνουν α) τη θέση του "κέντρου" των παρατηρήσεων στον οριζόντιο άξονα και β) τη διασπορά των παρατηρήσεων, δηλαδή πόσο αυτές εκτείνονται γύρω από το "κέντρο" τους. Τα πρώτα τα καλούμε **μέτρα θέσης** της κατανομής (location measures), ενώ τα δεύτερα **μέτρα διασποράς** ή **μέτρα μεταβλητότητας** (measures of variability).

Μέτρα Θέσης

Τα πιο συνηθισμένα μέτρα που χρησιμοποιούνται για την περιγραφή της θέσης ενός συνόλου δεδομένων πάνω στον οριζόντιο άξονα οx, εκφράζοντας την "κατά μέσο όρο" απόστασή τους από την αρχή των αξόνων, είναι ο αριθμητικός μέσος ή μέση τιμή (arithmetic mean or average), η διάμεσος (median) και η κορυφή ή επικρατούσα τιμή (mode).

α) Μέση Τιμή $(\bar{x})$

Η μέση τιμή ενός συνόλου ν παρατηρήσεων αποτελεί το σπουδαιότερο και χρησιμότερο μέτρο της Στατιστικής και ορίζεται ως το άθροισμα των παρατηρήσεων διά του πλήθους των παρατηρήσεων.
Όταν σε ένα δείγμα μεγέθους ν οι παρατηρήσεις μιας μεταβλητής X είναι $t_1,t_2,...,t_v$, τότε η μέση τιμή συμβολίζεται με $\bar{x}$ και δίνεται από τη σχέση:

$$\bar{x} = \frac{t_1 + t_2 + \ldots + t_v}{v} = \frac{\sum_{i=1}^{v} t_i}{v} = \frac{1}{v}\sum_{i=1}^{v} t_i \qquad (1)$$

όπου το σύμβολο $\sum_{i=1}^{v} t_i$ παριστάνει μια συντομογραφία του αθροίσματος $t_1+t_2+...+t_v$ και διαβάζεται "άθροισμα των t_1 από i = 1 έως ν". Συχνά, όταν δεν υπάρχει πρόβλημα σύγχυσης, συμβολίζεται και ως $\sum t_i$ ή ακόμα πιο απλά με $\sum t$.
Σε μια κατανομή συχνοτήτων, αν x_1, x_2 ,..., $x_κ$ είναι οι τιμές της μεταβλητής X με συχνότητες $v_1,v_2,...,v_κ$ αντίστοιχα, η μέση τιμή ορίζεται ισοδύναμα από τη σχέση:

$$\bar{x} = \frac{x_1v_1 + x_2v_2 + \ldots + x_κv_κ}{v_1 + v_2 + \ldots + v_κ} = \frac{\sum_{i=1}^{κ} x_iv_i}{\sum_{i=1}^{κ} v_i} = \frac{1}{v}\sum_{i=1}^{κ} x_iv_i \qquad (2)$$

Η παραπάνω σχέση ισοδύναμα γράφεται:

$$\bar{x} = \sum_{i=1}^{\kappa} x_i \frac{\nu_i}{\nu} = \sum_{i=1}^{\kappa} x_i f_i$$

όπου f_i οι σχετικές συχνότητες. Για παράδειγμα, η μέση επίδοση των μαθητών στο τμήμα A θα είναι σύμφωνα με την (1)

$$\bar{x}_A = \frac{13+13+14+\ldots+18}{10} = \frac{150}{10} = 15$$

ή ισοδύναμα από τον αντίστοιχο πίνακα συχνοτήτων σύμφωνα με την (2).

Βαθμός x_i	Συχνότητα ν_i	$x_i\nu_i$
13	2	26
14	1	14
15	4	60
16	2	32
18	1	18
Σύνολο	$\nu_A = 10$	$\Sigma x_i\nu_i = 150$

$$\bar{x}_A = \frac{\Sigma x_i \nu_i}{\nu_A} = \frac{150}{10} = 15.$$

Ομοίως, υπολογίζεται και η μέση επίδοση για το τμήμα Β, η οποία είναι πάλι

$$\bar{x}_B = 15.$$

Επίσης, το μέσο ύψος των 40 μαθητών της Γ΄ Λυκείου του πίνακα 8, σύμφωνα με τη σχέση (1) είναι $\bar{x} = \frac{6918}{40} = 172{,}95 \text{ cm}$.

Για ευκολότερο όμως υπολογισμό χρησιμοποιούμε τον πίνακα συχνοτήτων, όπως αυτός δίνεται παρακάτω, ομαδοποιώντας τα δεδομένα σε $\kappa = 6$ κλάσεις. Αν x_i είναι το κέντρο της i κλάσης και ν_i η αντίστοιχη συχνότητα, τότε σύμφωνα με τη σχέση (2) η μέση τιμή θα είναι:

$$\bar{x} = \frac{\Sigma x_i \nu_i}{\nu} = \frac{6903}{40} = 172{,}6 \text{cm}.$$

Παρατηρούμε ότι οι δύο μέσες τιμές του ίδιου συνόλου δεδομένων δεν είναι ακριβώς οι ίδιες. Πού οφείλεται αυτή η, έστω και μικρή, διαφορά;

Η διαφορά αυτή οφείλεται στο γεγονός ότι κατά την ομαδοποίηση υποθέσαμε ότι οι παρατηρήσεις κάθε κλάσης είναι ομοιόμορφα κατανεμημένες και ότι οι τιμές της μεταβλητής σε κάθε κλάση εκπροσωπούνται από την αντίστοιχη κεντρική τιμή x_i. Η υπόθεση αυτή σημαίνει απώλεια πληροφοριών για τις αρχικές τιμές. Χάνουμε λοιπόν λίγο ως προς στην ακρίβεια κερδίζουμε όμως χρόνο!

Ύψος σε cm	Κεντρικές τιμές x_i	Συχνότητα ν_i	$x_i\nu_i$
156-162	159	2	318
162-168	165	8	1320
168-174	171	12	2025
174-180	177	11	1947
180-186	183	5	915
186-192	189	2	378
	Σύνολο	$\Sigma\nu_i = 40$	$\Sigma x_i\nu_i = 6903$

γ) Διάμεσος (δ)

Οι χρόνοι (σε λεπτά) που χρειάστηκαν 9 μαθητές, για να λύσουν ένα πρόβλημα είναι: 3, 5, 5, 36, 6, 7, 4, 7, 8 με μέση τιμή $\bar{x}=9$. Παρατηρούμε όμως ότι οι οκτώ από τις εννέα παρατηρήσεις είναι μικρότερες του 9 και μία (ακραία τιμή), η οποία επηρεάζει και τη μέση τιμή είναι, αρκετά μεγαλύτερη του 9. Αυτό σημαίνει ότι η μέση τιμή δεν ενδείκνυται ως μέτρο θέσης ("κέντρο") των παρατηρήσεων αυτών. Αντίθετα, ένα άλλο μέτρο θέσης που δεν επηρεάζεται από ακραίες παρατηρήσεις είναι η **διάμεσος** (median), η οποία ορίζεται ως εξής:

> Διάμεσος (δ) ενός δείγματος ν παρατηρήσεων οι οποίες έχουν διαταχθεί σε αύξουσα σειρά ορίζεται ως η μεσαία παρατήρηση, όταν το ν είναι περιττός αριθμός, ή ο μέσος όρος (ημιάθροισμα) των δύο μεσαίων παρατηρήσεων όταν το ν είναι άρτιος αριθμός.

Για παράδειγμα, για να βρούμε τη διάμεσο των δεδομένων:
α) 3, 4, 0, 6, 5, 8, 1, 1, 6, 1, 2, 8, 9
β) 3, 4, 0, 6, 5, 8, 1, 1, 6, 1, 2, 8, 9, 9
εργαζόμαστε ως εξής:

α) Έχουμε $\nu = 13$ παρατηρήσεις, οι οποίες σε αύξουσα σειρά είναι:

0 1 1 1 2 3 4 5 6 6 8 8 9.

Άρα, η διάμεσος είναι η μεσαία παρατήρηση (έβδομη στη σειρά), $\delta = 4$.

β) Έχουμε $\nu = 14$ παρατηρήσεις οι οποίες σε αύξουσα σειρά είναι:

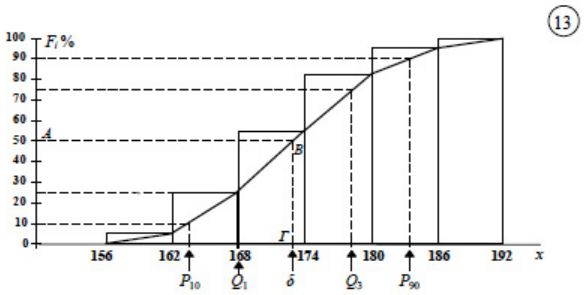
0 1 1 1 2 3 4 5 6 6 8 8 9 9.

Άρα, η διάμεσος είναι το ημιάθροισμα των δύο μεσαίων παρατηρήσεων (της έβδομης και όγδοης στη σειρά), δηλαδή $\delta=\frac{4+5}{2}=4{,}5$.

Παρατηρούμε ότι, η διάμεσος είναι η τιμή που χωρίζει ένα σύνολο παρατηρήσεων σε δύο ίσα μέρη όταν οι παρατηρήσεις αυτές τοποθετηθούν με σειρά τάξης μεγέθους. Ακριβέστερα, η διάμεσος είναι η τιμή για την οποία το πολύ 50% των παρατηρήσεων είναι μικρότερες από αυτήν και το πολύ 50% των παρατηρήσεων είναι μεγαλύτερες από την τιμή αυτήν.

Διάμεσος σε Ομαδοποιημένα Δεδομένα

Θεωρούμε τα δεδομένα του ύψους των μαθητών στον πίνακα 9 και το αντίστοιχο ιστόγραμμα αθροιστικών σχετικών συχνοτήτων με την πολυγωνική γραμμή, σχήμα 13. Η διάμεσος, όπως ορίστηκε, αντιστοιχεί στην τιμή $x = \delta$ της μεταβλητής X (στον οριζόντιο άξονα), έτσι ώστε το 50% των παρατηρήσεων να είναι μικρότερες ή ίσες του δ. Δηλαδή, η διάμεσος θα έχει αθροιστική σχετική συχνότητα F_i = 50% . Εφόσον στον κάθετο άξονα έχουμε τις αθροιστικές σχετικές συχνότητες, από το σημείο A (50% των παρατηρήσεων) φέρουμε την $AB // 0x$ και στη συνέχεια τη $B\Gamma \perp 0x$. Τότε, στο σημείο Γ αντιστοιχεί η διάμεσος δ των παρατηρήσεων. Δηλαδή, $\delta \approx 173$.



δ) Εκατοστημόρια ($P_κ$)

Όπως ορίσαμε τη διάμεσο δ, έτσι ώστε το πολύ 50% των παρατηρήσεων να είναι μικρότερες του δ και το πολύ 50% των παρατηρήσεων να είναι μεγαλύτερες του δ, μπορούμε ανάλογα να ορίσουμε και τα **εκατοστημόρια** (percentiles) $P_κ$, $κ$ = 1,2,...,99 . Οι τιμές P_1, P_2,..., P_{99} χωρίζουν τη συνολική συχνότητα σε 100 ίσα μέρη. Επομένως, αναλόγως και με τον ορισμό της διαμέσου, ορίζουμε ως $κ$-εκατοστιαίο σημείο ή $P_κ$ εκατοστημόριο ενός συνόλου παρατηρήσεων την τιμή εκείνη για την οποία το πολύ κ% των παρατηρήσεων είναι μικρότερες του $P_κ$ και το πολύ (100-$κ$)% των παρατηρήσεων είναι μεγαλύτερες από την τιμή αυτήν.

Ειδική περίπτωση εκατοστημορίων είναι τα P_{25}, P_{50}, P_{75}, τα οποία καλούνται **τεταρτημόρια** (quartiles) και συμβολίζονται με Q_1, Q_2 και Q_3, αντίστοιχα. Για το Q_1 έχουμε αριστερά το πολύ 25% των παρατηρήσεων και δεξιά το πολύ 75%. Όμοια για το Q_3 έχουμε αριστερά το πολύ 75% των παρατηρήσεων και δεξιά το πολύ 25% των παρατηρήσεων. Προφανώς το $Q_2 = P_{50}$ συμπίπτει και με τη διάμεσο, δηλαδή $Q_2 = \delta$. Τα μέτρα αυτά χρησιμοποιούνται αρκετά συχνά για τη μελέτη ενός συνόλου δεδομένων.

Συχνά για ευκολία ο υπολογισμός των τεταρτημορίων Q_1 και Q_3 ενός συνόλου δεδομένων γίνεται κατά προσέγγιση υπολογίζοντας τις διαμέσους του πρώτου και του δεύτερου μισού των διατεταγμένων παρατηρήσεων, αντίστοιχα. Για παράδειγμα, προκειμένου να υπολογίσουμε τα τεταρτημόρια των δεδομένων 3, 4, 0, 6, 5, 8, 1, 1, 6, 1, 2, 8, 9, εργαζόμαστε ως εξής:

- Διατάσσουμε τις παρατηρήσεις σε αύξουσα σειρά μεγέθους:
  Έχουμε $ν$ = 13 παρατηρήσεις, οι οποίες σε αύξουσα σειρά είναι:

  0 1 1 1 2 3 4 5 6 6 8 8 9.

- Υπολογίζουμε τη διάμεσο, όπως προαναφέραμε:
  Η διάμεσος είναι η έβδομη στη σειρά παρατήρηση, δηλαδή δ = 4.
- Υπολογίζουμε τη διάμεσο του πρώτου μισού των διατεταγμένων παρατηρήσεων, δηλαδή των παρατηρήσεων που είναι αριστερά του δ. Η τιμή αυτή είναι το Q_1:

  Η διάμεσος των παρατηρήσεων που είναι αριστερά του δ, δηλαδή των 0 1 1 1 2 3, είναι το $Q_1 = \frac{1+1}{2} = 1$.
- Υπολογίζουμε τη διάμεσο του δεύτερου μισού των διατεταγμένων παρατηρήσεων, δηλαδή των παρατηρήσεων που είναι δεξιά του δ. Η τιμή αυτή είναι το Q_3.

  Η διάμεσος των παρατηρήσεων που είναι δεξιά του δ, δηλαδή των 5 6 6 8 8 9, είναι το $Q_3 = \frac{6+8}{2} = 7$.

Εκατοστημόρια σε Ομαδοποιημένα Δεδομένα

Ο υπολογισμός των εκατοστημορίων (ή τεταρτημορίων) σε ομαδοποιημένα δεδομένα γίνεται όπως και στη διάμεσο από το πολύγωνο αθροιστικών σχετικών συχνοτήτων. Στο σχήμα 13 δίνονται τα Q_1, $Q_2 = \delta$, Q_3 και P_{10}, P_{90} για τα δεδομένα του πίνακα 9, από το οποίο βρίσκουμε κατά προσέγγιση:

$$P_{10} = 162{,}5,\ Q_1 = 168,\ \delta = 173,\ Q_3 = 178 \text{ και } P_{90} = 184.$$

Μέτρα Διασποράς

Στα προηγούμενα είδαμε ότι τα μέτρα θέσης παρέχουν κάποια πληροφορία για την κατανομή ενός πληθυσμού. Αυτά όμως δεν επαρκούν, για να περιγράψουν πλήρως την κατανομή, όπως διαπιστώσαμε στην αρχή της § 2.3 συγκρίνοντας τις βαθμολογίες των μαθητών δύο τμημάτων *A* και *B* στα σχήματα 11(α), (β). Ενώ οι βαθμολογίες των δύο τμημάτων *A* και *B* έχουν ίσες μέσες τιμές $\bar{x}_A = \bar{x}_B = 15$ και ίσες διαμέσους $\delta_A = \delta_B = 15$, είναι φανερό ότι οι κατανομές τους διαφέρουν σημαντικά ως προς τη μεταβλητότητά τους. Οι βαθμοί του τμήματος Α είναι περισσότερο "συγκεντρωμένοι" γύρω από τη μέση τιμή, ενώ, αντίθετα, οι βαθμοί του τμήματος Β διασπείρονται περισσότερο, έχουν δηλαδή μεγάλες αποκλίσεις γύρω από τη μέση τιμή τους.
Παράλληλα λοιπόν με τα μέτρα θέσης κρίνεται απαραίτητη και η εξέταση κάποιων μέτρων διασποράς ή μεταβλητότητας, δηλαδή μέτρων που εκφράζουν τις αποκλίσεις των τιμών μιας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης. Τέτοια μέτρα λέγονται **μέτρα διασποράς** (measures of variation, dispersion measures). Τα σπουδαιότερα μέτρα διασποράς είναι το εύρος, η ενδοτεταρτημοριακή απόκλιση, η διακύμανση και η τυπική απόκλιση.

α) Εύρος (R)

Το απλούστερο από τα μέτρα διασποράς είναι το **εύρος** ή **κύμανση** (range) (*R*), που ορίζεται ως η διαφορά της ελάχιστης παρατήρησης από τη μέγιστη παρατήρηση, δηλαδή:

Εύρος R = Μεγαλύτερη παρατήρηση-Μικρότερη παρατήρηση

Έτσι, για τη βαθμολογία του τμήματος Α το εύρος είναι $R_A = 18 - 13 = 5$, ενώ για το τμήμα Β, $R_B = 20 - 10 = 10$, τιμές που επιβεβαιώνουν ότι πράγματι στο τμήμα Β έχουμε μεγαλύτερη διασπορά βαθμολογίαςπαρά στο τμήμα Α.
Όταν έχουμε ομαδοποιημένα δεδομένα, το εύρος δίνεται από τη διαφορά του κατώτερου ορίου της πρώτης κλάσης από το ανώτερο όριο της τελευταίας κλάσης. Το εύρος των υψών των μαθητών του δείγματος στον πίνακα 9 είναι $R = 192 - 156 = 36$. Προφανώς, το εύρος σε ομαδοποιημένα δεδομένα μπορεί να διαφέρει ελαφρώς από τα αντίστοιχα δεδομένα πριν αυτά ομαδοποιηθούν. Για παράδειγμα, το εύρος των υψών στον πίνακα 8, πριν αυτά ομαδοποιηθούν, βρήκαμε ότι είναι $R = 191 - 156 = 35$.
Το εύρος είναι ένα αρκετά απλό μέτρο, που υπολογίζεται εύκολα δε θεωρείται όμως αξιόπιστο μέτρο διασποράς, γιατί βασίζεται μόνο στις δυο ακραίες παρατηρήσεις.

β) Ενδοτεταρτημοριακό Εύρος (Q)

Το **ενδοτεταρτημοριακό εύρος** (interquartile range) είναι η διαφορά του πρώτου τεταρτημορίου Q_1 από το τρίτο τεταρτημόριο Q_3, δηλαδή:

$Q = Q_3 - Q_1$

Στο μεταξύ τους διάστημα περιλαμβάνεται το 50% των παρατηρήσεων. Επομένως όσο μικρότερο είναι αυτό το διάστημα, τόσο μεγαλύτερη θα είναι η συγκέντρωση των τιμών και άρα μικρότερη η διασπορά των τιμών της μεταβλητής.
Από τα δεδομένα του σχήματος 13 βρήκαμε κατά προσέγγιση $Q_1 = 168$, $Q_3 = 178$ επομένως το ενδοτεταρτημοριακό εύρος είναι $Q = 10$. Δηλαδή το 50% των μαθητών έχουν ύψος μεταξύ 168 και 178 cm.

γ) Διακύμανση (s^2)

Ένας άλλος τρόπος για να υπολογίσουμε τη διασπορά των παρατηρήσεων $t_1, t_2, \ldots, t_v$ μιας μεταβλητής X θα ήταν να αφαιρέσουμε τη μέση τιμή $\bar{x}$ από κάθε παρατήρηση και να βρούμε τον αριθμητικό μέσο των διαφορών αυτών, δηλαδή τον αριθμό:

$$\frac{(t_1 - \bar{x}) + (t_2 - \bar{x}) + \ldots + (t_v - \bar{x})}{v} = \frac{\sum_{i=1}^{v}(t_i - \bar{x})}{v}.$$

Ο αριθμός όμως αυτός είναι ίσος με μηδέν, αφού

$$\frac{(t_1 - \bar{x}) + (t_2 - \bar{x}) + \ldots + (t_v - \bar{x})}{v} = \frac{t_1 + t_2 + \ldots + t_v}{v} - \frac{v\bar{x}}{v} = \bar{x} - \bar{x} = 0.$$

Γι'αυτό, ως ένα μέτρο διασποράς παίρνουμε τον μέσο όρο των τετραγώνων των αποκλίσεων των t_1 από τη μέση τιμή τους $\bar{x}$. Το μέτρο αυτό καλείται **διακύμανση** ή **διασπορά** (variance) και ορίζεται από τη σχέση

$$s^2 = \frac{1}{v}\sum_{i=1}^{v}(t_i - \bar{x})^2 \qquad (1)$$

Ο τύπος αυτός αποδεικνύεται ότι μπορεί να πάρει την ισοδύναμη μορφή:

$$s^2 = \frac{1}{v}\left\{\sum_{i=1}^{v} t_i^2 - \frac{\left(\sum_{i=1}^{v} t_i\right)^2}{v}\right\} \qquad (2)$$

η οποία διευκολύνει σημαντικά τους υπολογισμούς κυρίως όταν η μέση τιμή $\bar{x}$ δεν είναι ακέραιος αριθμός. Όταν έχουμε πίνακα συχνοτήτων ή ομαδοποιημένα δεδομένα, η διακύμανση ορίζεται από τη σχέση:

$$s^2 = \frac{1}{v}\sum_{i=1}^{\kappa}(x_i - \bar{x})^2 v_i \qquad (3)$$

ή την ισοδύναμη μορφή:

$$s^2 = \frac{1}{v}\left\{\sum_{i=1}^{\kappa} x_i^2 v_i - \frac{\left(\sum_{i=1}^{\kappa} x_i v_i\right)^2}{v}\right\}. \qquad (4)$$

όπου $x_1, x_2, ..., x_\kappa$ οι τιμές της μεταβλητής (ή τα κέντρα των κλάσεων) με αντίστοιχες συχνότητες $v_1, v_2, ..., v_\kappa$.
Για παράδειγμα, η διακύμανση της βαθμολογίας των μαθητών του τμήματος Α είναι σύμφωνα με την (1)

$$s_A^2 = \frac{(13-15)^2 + (13-15)^2 + (14-15)^2 + ... + (18-15)^2}{10} = \frac{20}{10} = 2,$$

ενώ για τους μαθητές του τμήματος Β βρίσκουμε $s_B^2 = 6{,}6$, που επιβεβαιώνει τη διαπίστωσή μας ότι η βαθμολογία των μαθητών του τμήματος Β παρουσιάζει μεγαλύτερη μεταβλητότητα από τη βαθμολογία των μαθητών του τμήματος Α.
Ομοίως, η διακύμανση του ύψους των μαθητών για τα ομαδοποιημένα δεδομένα του πίνακα 9, υπολογίζεται σύμφωνα με τον τύπο (3), όπως φαίνεται στον επόμενο πίνακα:

Κλάσεις [-)	Κεντρικές τιμές x_i	Συχνότητα v_i	x_i^2	$x_i v_i$	$x_i^2 v_i$
156-162	159	2	25281	318	50562
162-168	165	8	27225	1320	217800
168-174	171	12	29241	2052	350892
174-180	177	11	31329	1942	344619
180-186	183	5	33489	915	167445
186-192	189	2	35721	378	71442
Σύνολο		$v = 40$	-	$\Sigma x_i v_i = 6930$	$\Sigma x_i^2 v_i = 1202760$

Επομένως:

$$s^2 = \frac{1}{v}\left\{\sum_{i=1}^{\kappa} x_i^2 v_i - \frac{\left(\sum_{i=1}^{\kappa} x_i v_i\right)^2}{v}\right\} = \frac{1}{40}\left\{1202776 - \frac{6930^2}{40}\right\} = 53{,}4$$

Εάν υπολογίσουμε τη διακύμανση από τα μη ομαδοποιημένα δεδομένα του πίνακα 8, βρίσκουμε $s^2 = 50{,}9$. Η διαφορά αυτή οφείλεται στην απώλεια πληροφορίας λόγω ομαδοποίησης των παρατηρήσεων.

δ) Τυπική Απόκλιση (s)

Η διακύμανση είναι μια αξιόπιστη παράμετρος διασποράς, αλλά έχει ένα μειονέκτημα. Δεν εκφράζεται με τις μονάδες με τις οποίες εκφράζονται οι παρατηρήσεις. Για παράδειγμα, αν οι παρατηρήσεις εκφράζονται σε cm, η διακύμανση εκφράζεται σε cm_2. Αν όμως πάρουμε τη θετική τετραγωνική ρίζα της διακύμανσης, θα έχουμε ένα μέτρο διασποράς που θα εκφράζεται με την ίδια μονάδα μέτρησης του χαρακτηριστικού, όπως ακριβώς είναι και όλα τα άλλα μέτρα θέσης, που εξετάσαμε έως τώρα. Η ποσότητα αυτή λέγεται **τυπική απόκλιση** (standard deviation), συμβολίζεται με s και δίνεται από τη σχέση:

$$s = \sqrt{s^2}$$

Η τυπική απόκλιση για το ύψος των μαθητών του πίνακα 4 είναι από το προηγούμενο παράδειγμα $s = \sqrt{53,4} = 7,3$ cm αν αυτή υπολογιστεί από τα ομαδοποιημένα δεδομένα του πίνακα 9, ή $s = \sqrt{50,9} = 7,13$ cm, αν υπολογιστεί από τα μη ομαδοποιημένα δεδομένα του πίνακα 8.
Αξίζει να σημειωθεί ότι αν η καμπύλη συχνοτήτων για το χαρακτηριστικό που εξετάζουμε είναι κανονική ή περίπου κανονική, τότε η τυπική απόκλιση s έχει τις παρακάτω ιδιότητες:

i) το 68% περίπου των παρατηρήσεων βρίσκεται στο διάστημα

$$(\bar{x} - s,\ \bar{x} + s)$$

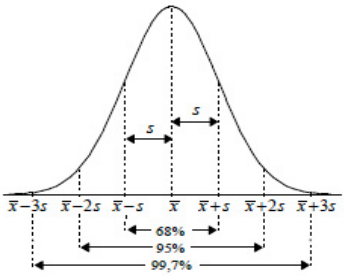
ii) το 95% περίπου των παρατηρήσεων βρίσκεται στο διάστημα

$$(\bar{x} - 2s,\ \bar{x} + 2s)$$

iii) το 99,7% περίπου των παρατηρήσεων βρίσκεται στο διάστημα

$$(\bar{x} - 3s,\ \bar{x} + 3s)$$

iv) το εύρος ισούται περίπου με έξι τυπικές αποκλίσεις, δηλαδή $R \approx 6s$.



Συντελεστής Μεταβολής (CV)

Έστω ότι από ένα δείγμα είκοσι μαθητών της Α΄ Γυμνασίου βρήκαμε μέσο βάρος $\bar{x}_A = 40$ kgr και τυπική απόκλιση s_A = 6 kgr, ενώ από ένα δεύτερο δείγμα τριάντα μαθητών της Γ' Λυκείου βρήκαμε μέσο βάρος $\bar{x}_B = 75$ kgr και τυπική απόκλιση s_B = 6 kgr. Όπως αντιλαμβανόμαστε, είναι λάθος να πούμε ότι το βάρος των μαθητών του Λυκείου έχει τον ίδιο βαθμό μεταβλητότητας με το βάρος των μαθητών του Γυμνασίου, καθόσον η βαρύτητα που έχουν τα 6 kgr στο μέσο βάρος των 40 kgr είναι διαφορετική από αυτήν που έχουν στο μέσο βάρος των 75 kgr.
Ακόμη, ας υποθέσουμε ότι ο μέσος μισθός των υπαλλήλων μιας εταιρείας Α είναι $\bar{x}_A = 250.000$ δρχ. με τυπική απόκλιση s_A = 42.000 δρχ., ενώ για τους υπαλλήλους μιας εταιρείας Β είναι $\bar{x}_B = \$1.400$ με τυπική απόκλιση s_B = $350. Στην περίπτωση αυτή έχουμε διαφορετικές μονάδες μέτρησης του μισθού, επομένως οι διασπορές των παρατηρήσεων δεν είναι άμεσα συγκρίσιμες.
Ένα μέτρο με το οποίο μπορούμε να ξεπεράσουμε τις παραπάνω δυσκολίες και το οποίο μας βοηθά στη σύγκριση ομάδων τιμών, που είτε εκφράζονται σε διαφορετικές μονάδες μέτρησης είτε εκφράζονται στην ίδια μονάδα μέτρησης, αλλά έχουν σημαντικά διαφορετικές μέσες τιμές, είναι ο **συντελεστής μεταβολής** ή **συντελεστής μεταβλητότητας** (coefficient of variation), ο οποίος ορίζεται από το λόγο:

$$CV = \frac{\text{τυπική απόκλιση}}{\text{μέση τιμή}} \cdot 100\% = \frac{s}{\bar{x}} \cdot 100\%.$$

Ο συντελεστής μεταβολής εκφράζεται επί τοις εκατό, είναι συνεπώς ανεξάρτητος από τις μονάδες μέτρησης και παριστάνει ένα μέτρο **σχετικής διασποράς** των τιμών και όχι της απόλυτης διασποράς, όπως έχουμε δει έως τώρα. Εκφράζει, δηλαδή, τη μεταβλητότητα των δεδομένων απαλλαγμένη από την επίδραση της μέσης τιμής. Για το πρώτο παράδειγμα του βάρους έχουμε συντελεστή μεταβολής για τις δύο ομάδες μαθητών:

$$CV_A = \frac{s_A}{\bar{x}_A} \cdot 100\% = \frac{6}{40} \cdot 100\% = 15\% \quad \text{και}$$

$$CV_B = \frac{s_B}{\bar{x}_B} \cdot 100\% = \frac{6}{75} \cdot 100\% = 8\%$$

δηλαδή, ο βαθμός διασποράς του βάρους των μαθητών Γυμνασίου είναι μεγαλύτερος από το βαθμό διασποράς του βάρους των μαθητών Λυκείου (για τα συγκεκριμένα δείγματα).
Ανάλογα συμπεράσματα βγάζουμε και για το δεύτερο παράδειγμα, όπου βρίσκουμε CV_A = 12% και CV_B = 25%.

Παρ' όλο που η τυπική απόκλιση των μισθών στην εταιρεία Α είναι μεγαλύτερη από την τυπική απόκλιση στην εταιρεία Β, ο συντελεστής μεταβολής δίνει μεγαλύτερη σχετική διασπορά στην εταιρεία Β. Αυτό μεταφράζεται στο να λέμε ότι έχουμε μεγαλύτερη **ομοιογένεια** μισθών στην εταιρεία Α παρά στη Β.
Γενικά δεχόμαστε ότι ένα δείγμα τιμών μιας μεταβλητής θα είναι ομοιογενές, εάν ο συντελεστής μεταβολής δεν ξεπερνά το 10%.