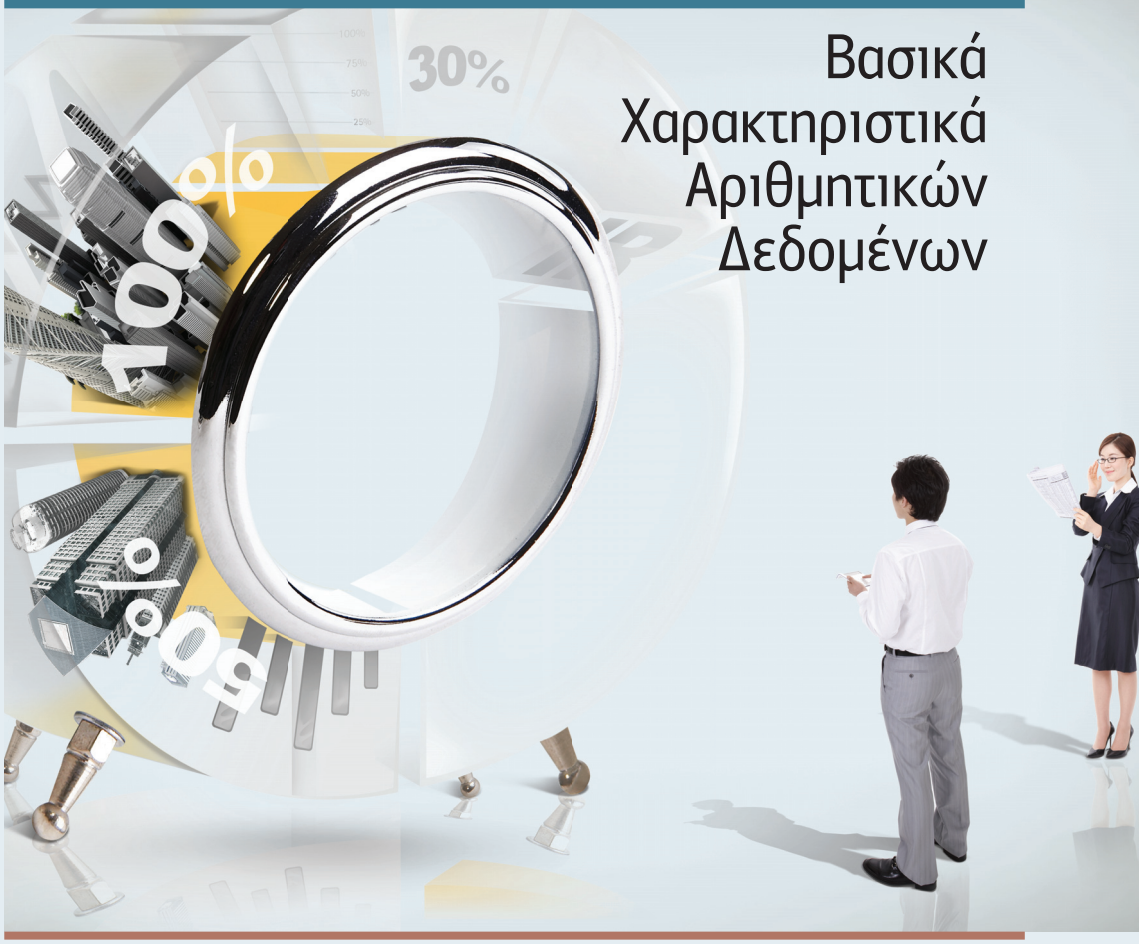
ΚΕΦΑΛΑΙΟ 3

Βασικά Χαρακτηριστικά Αριθμητικών Δεδομένων



Αντικείμενο του κεφαλαίου είναι: Να κατανοήσουμε τα βασικά χαρακτηριστικά των αριθμητικών δεδομένων (τάση, διασπορά, ασυμμετρία) και να περιγράψουμε τους τρόπους μέτρησής τους.

3.1. **Εισαγωγή**

Στο προηγούμενο Κεφάλαιο είδαμε πώς παρουσιάζουμε τα δεδομένα με τη βοήθεια των πινάκων και των διαγραμμάτων. Όμως, πώς θα αξιοποιήσουμε αυτές τις πληροφορίες; Αν και η παρουσίαση των δεδομένων αποτελεί βασικό στοιχείο της περιγραφικής στατιστικής, δεν αρκεί για να αποκαλύψει όλη την εικόνα που περιέχουν τα δεδομένα. Η πλήρης ανάλυση των δεδομένων δε βασίζεται μόνο στην παρουσίαση και παρατήρηση του τι προσπαθούν τα δεδομένα να αποκαλύψουν, αλλά περιλαμβάνει υπολογισμούς και εκτιμήσεις των βασικών χαρακτηριστικών, η ανάλυση των οποίων θα οδηγήσει και στην πλήρη κατανόηση της εικόνας που αποκαλύπτουν.

Έτσι, σκοπός μας είναι να δούμε με ποιες μεθόδους θα επιτύχουμε τη συνοπτική περιγραφή των δεδομένων και στη συνέχεια την ερμηνεία τους. Μην ξεχνάμε ότι η ανάλυση των δεδομένων αποτελεί σύστημα υποστήριξης αποφάσεων. Όσο πιο σωστά κατανοήσουμε τι αποκαλύπτουν τα δεδομένα τόσο πιο ορθές αποφάσεις θα πάρουμε.

Τρεις είναι οι βασικές ιδιότητες που χαρακτηρίζουν ένα σύνολο αριθμητικών δεδομένων

- η κεντρική τάση
- η διασπορά ή μεταβλητότητα
- το σχήμα της κατανομής τους

Η ανάλυση των δεδομένων περιλαμβάνει, μεταξύ άλλων, και μία σειρά μετρήσεων που θα περιγράφουν τις παραπάνω ιδιότητες: κεντρική τάση, διασπορά, και σχήμα της κατανομής. Εάν αυτές οι περιγραφικές μετρήσεις προκύπτουν από δεδομένα ενός δείγματος ονομάζονται **εκτιμήσεις** ή **στατιστικές** (estimations ή statistics). Ενώ, εάν υπολογίζονται για όλον τον πληθυσμό, τότε ονομάζονται **παράμετροι**.

Δεδομένου, ότι η συνήθης πρακτική είναι να εξετάζουμε δείγματα και όχι ολόκληρους πληθυσμούς, που είναι χρονοβόρο και πολυδάπανο, η προσοχή μας θα συγκεντρωθεί περισσότερο στις εκτιμήσεις και λιγότερο στις παραμέτρους.

3.2. **Μέτρηση της Κεντρικής Τάσης**

Στις περισσότερες περιπτώσεις ένα σύνολο δεδομένων παρουσιάζει τάση συγκέντρωσης των τιμών του γύρω από μία κεντρική τιμή. Έτσι, για κάθε συγκεκριμένο σύνολο δεδομένων, είναι δυνατόν να επιλέξουμε κάποια τυπική τιμή ή **μέσο** που θα περιγράφει τη συμπεριφορά των τιμών. Με άλλα λόγια προσπαθούμε να βρούμε τον "εκπρόσωπο" των τιμών που θα τις αντιπροσωπεύει όποτε θα αναφερόμαστε σε αυτές.

Τρεις είναι οι συνηθέστεροι τρόποι μέτρησης της κεντρικής τάσης μιας ομάδας αριθμητικών δεδομένων: ο μέσος αριθμητικός, η διάμεσος και το σημείο μέγιστης συχνότητας (ή τύπος).

Μέσος αριθμητικός

Ο μέσος αριθμητικός (ή απλά μέσος) είναι ο συνηθέστερος τρόπος μέτρησης της κεντρικής τάσης. Υπολογίζεται από το άθροισμα όλων των τιμών διαιρούμενο με το πλήθος των παρατηρήσεων. Εάν συμβολίσουμε με X το χαρακτηριστικό που μετράμε (μεταβλητή) και με n το πλήθος των παρατηρήσεων, τότε οι n τιμές συμβολίζονται με $X_1, X_2, \ldots, X_n$. Ο δε μέσος αριθμητικός, που συμβολίζεται με $\bar{X}$, υπολογίζεται με τον εξής τύπο:

$$\bar{X} = \frac{X_1 + X_2 \ldots, + X_n}{n} \qquad (3.1)$$

Για να απλοποιήσουμε τον τύπο (3.1), μπορούμε να συμβολίσουμε το άθροισμα των n τιμών της μεταβλητής X με

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n \qquad (3.2)$$

που σημαίνει ότι αθροίζουμε όλες τις X_i τιμές. Και επειδή πάντα αναφερόμαστε στις n τιμές της X, παραλείπουμε τα όρια του αθροιστή Σ και το δείκτη i της X, και η (3.2) απλοποιείται σε

$$\sum X = X_1 + X_2 + \ldots + X_n \qquad (3.3)$$

Έτσι ο μέσος αριθμητικός $\bar{X}$ ισούται με

$$\bar{X} = \frac{\Sigma X}{n} \qquad (3.4)$$

Ο μέσος που υπολογίζεται με τον τύπο (3.4) **ονομάζεται απλός** ή **αστάθμητος μέσος αριθμητικός** (arithmetic mean), διότι προκύπτει από το απλό άθροισμα των τιμών της X. Για παράδειγμα, ο μέσος ετήσιος μισθός των 190 υπαλλήλων της εταιρίας ισούται με

$$\bar{X} = (11{,}340+11{,}448+11{,}664+\ldots+79{,}650+97{,}200)/190 = 27{,}189 \text{ χιλ. €}$$

Και τώρα θα ρωτήσετε "ποια είναι η ερμηνεία του μέσου; τι ακριβώς εκφράζει;". Ο μέσος αριθμητικός μας υποδεικνύει τη "μέση αμοιβή". Δηλαδή, εάν το συνολικό ποσό της μισθοδοσίας μοιραζόταν σε 190 ίσα ποσά, τότε κάθε εργαζόμενος θα αμειβόταν με 27,189 χιλ. € ετησίως. Το γεγονός ότι κάποιοι αμείβονται με μισθό μεγαλύτερο από το μέσο σημαίνει ότι κάποιοι άλλοι αμείβονται λιγότερο. Γι' αυτό ο μέσος βρίσκεται στο ενδιάμεσο των παρατηρήσεων. Εάν αυτοί που έχουν μισθό χαμηλότερο από το μέσο μισθό είναι περίπου ίσοι με αυτούς που έχουν μεγαλύτερο, λιγότεροι ή περισσότεροι, εξαρτάται από το σχήμα της κατανομής, όπως θα δούμε αργότερα.

Ας εξετάσουμε τώρα τον **σταθμικό μέσο αριθμητικό**. Το τμήμα προσωπικού της εταιρίας υπολογίζει ξεχωριστά για κάθε κατηγορία προσωπικού το μέσο επίπεδο αποδοχών, και παρουσιάζει τα εξής στοιχεία (Πίνακας 3.1).

Α/Α	Κατηγορία Εργαζομένων	Αριθμός Εργαζομένων	Μέσες Αποδοχές (χιλ. Ευρώ)
1	Υπάλληλοι	80	20,168
2	Εκπαιδευόμενοι	58	20,759
3	Ασφάλεια	7	23,081
4	Πτυχιούχοι	20	43,956
5	Πτυχιούχοι με MBA	15	48,278
6	Αναλυτές	4	46,800
7	Προϊστάμενοι	6	66,045
	Σύνολο	**190**	

Πίνακας 3.1
Αποδοχές Προσωπικού ανά Κατηγορία

Για να υπολογίσουμε τον μέσο ετήσιο μισθό για το σύνολο των εργαζομένων, θα χρησιμοποιήσουμε τον τύπο του σταθμικού μέσου. Δηλαδή, θα εκτιμήσουμε τον μέσο αριθμητικό των επτά μέσων αποδοχών των αντίστοιχων κατηγοριών εργαζομένων, σταθμίζοντας κάθε μέσο με τον αριθμό των εργαζομένων που εκπροσωπεί. Έτσι, εάν συμβολίσουμε με $\bar{X}_i$ το μέσο επίπεδο αποδοχών της i κατηγορίας εργαζομένων, και με n_i τον αριθμό των εργαζομένων της i κατηγορίας, ο σταθμικός μέσος αριθμητικός ισούται με

$$\bar{X} = \frac{\bar{X}_1 \times n_1 + \bar{X}_2 \times n_2 + \ldots + \bar{X}_7 \times n_7}{n_1 + n_2 + \ldots + n_7}$$

$$= (20{,}168 \times 80 + 20{,}759 \times 58 + \ldots + 66{,}045 \times 6) / (80 + 58 + \ldots + 6)$$

$$= 27{,}189 \text{ χιλ. €}$$

που συμπίπτει με τον απλό μέσο αριθμητικό όλων των ατομικών μισθών. Επομένως, για k κατηγορίες (ομάδες) δεδομένων, όπου $\bar{X}_i$ είναι ο μέσος της ομάδας, και n_i είναι ο αριθμός παρατηρήσεων της ομάδας, ο σταθμικός μέσος αριθμητικός ισούται με

$$\bar{X} = \frac{\bar{X}_1 \times n_1 + \bar{X}_2 \times n_2 + \ldots + \bar{X}_k \times n_k}{n_1 + n_2 + \ldots + n_k} \qquad (3.5)$$

ή

$$\bar{X} = \bar{X}_1 \times (n_1/n) + \bar{X}_2 \times (n_2/n) + \ldots + \bar{X}_k \times (n_k/n) \qquad (3.6)$$

όπου

$$n = n_1 + n_2 + \ldots + n_k$$

Οι συντελεστές στάθμισης n_i/n ονομάζονται σχετικοί ή ποσοστιαίοι συντελεστές στάθμισης. Στο συγκεκριμένο παράδειγμα εκφράζουν τα ποσοστά των εργαζομένων που ανήκουν στις διάφορες κατηγορίες εργαζομένων. Είναι προφανές ότι το άθροισμα των σχετικών συντελεστών στάθμισης ισούται με τη μονάδα.

Τώρα, μπορούμε να δώσουμε έναν γενικό τύπο του σταθμικού μέσου. Εάν X_1, X_2, …, X_k, είναι μία σειρά μετρήσεων με συντελεστές στάθμισης w_1, w_2, …, w_k, αντίστοιχα, τότε ο σταθμικός μέσος ισούται με

$$\bar{X} = \sum_{i=1}^{k} X_i w_i \qquad (3.7)$$

όπου

$$\sum_{i=1}^{k} w_i = 1$$

Ας δούμε τώρα πώς εφαρμόζεται ο τύπος του σταθμικού μέσου σε δεδομένα που έχουν ταξινομηθεί σε κατανομή συχνοτήτων και ως εκ τούτου τα πρωτογενή στοιχεία δεν είναι πλέον διαθέσιμα. Θα εξετάσουμε την κατανομή του Πίνακα 2.3. Με βάση τα στοιχεία του πίνακα, η μόνη πληροφορία που έχουμε είναι ο αριθμός των υπαλλήλων για κάθε εισοδηματική τάξη. Για να εκτιμήσουμε τον μέσο μισθό πρέπει να γνωρίζουμε το άθροισμα των εισοδημάτων κάθε διαστήματος τάξης. Από τον τύπο του απλού μέσου αριθμητικού $\bar{X} = \Sigma X/n$, προκύπτει η σχέση

$$\Sigma X = \bar{X} \times n \qquad (3.8)$$

Δηλαδή, το άθροισμα μιας ομάδας τιμών ισούται με το γινόμενο του πλήθους επί τον μέσο όρο. Επομένως, εάν υποθέσουμε ότι σε κάθε διάστημα εισοδημάτων η **κεντρική τιμή** ή **κεντρικός όρος** αντιπροσωπεύει το μέσο εισόδημα των ατόμων του διαστήματος, τότε το γινόμενο (συχνότητα) × (κεντρικός όρος) θα δίνει προσεγγιστικά το άθροισμα των ετήσιων μισθών των υπαλλήλων του διαστήματος. Με άλλα λόγια, η κεντρική τιμή κάθε διαστήματος (X_i) σταθμίζεται με τη συχνότητα του διαστήματος, που συμβολίζεται με f_i. Έτσι, για k διαστήματα τάξεως, και σύμφωνα με τον τύπο του σταθμικού μέσου (3.5), ο μέσος αριθμητικός ενός δείγματος n τιμών ταξινομημένων σε κατανομή συχνοτήτων ισούται με

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{\Sigma fX}{n} \qquad (3.9)$$

όπου,

$$\Sigma f = n$$

Ο Πίνακας 3.2 δείχνει τον τρόπο εργασίας για την εκτίμηση του σταθμικού μέσου των δεδομένων της κατανομής συχνοτήτων. Προσέξτε που ο σταθμικός μέσος των ταξινομημένων δεδομένων (26,937 χιλ. €) διαφέρει από τον μέσο που εκτιμήθηκε από τα αταξινόμητα (πρωτογενή) δεδομένα (27,189 χιλ. €). Αυτό ήταν αναμενόμενο, αφού οι κεντρικοί όροι των διαστημάτων τάξεως αποτελούν εκτιμήσεις των μέσων εισοδημάτων.

Σημειώνουμε, ότι η παραπάνω διαδικασία έχει εκπαιδευτικό σκοπό και δε χρησιμοποιείται πλέον στην πράξη. Η διάδοση της πληροφορικής και των προγραμμάτων ανάλυσης δεδομένων έχουν υποκαταστήσει τον χειρωνακτικό τρόπο, και όλες οι αριθμητικές πράξεις γίνονται με τη βοήθεια των ηλεκτρονικών υπολογιστών. Η μόνη περίπτωση να χρησιμοποιήσετε τον τύπο (3.9) είναι όταν δεν γνωρίζετε τα πρωτογενή δεδομένα και διαθέτετε μόνο ταξινομημένα δεδομένα. Για παράδειγμα, το Υπουργείο Οικονομικών δημοσιεύει κάθε χρόνο σε πίνακα κατανομής συχνοτήτων τα δηλωμένα εισοδήματα. Είναι προφανές ότι για να εκτιμήσουμε το μέσο εισόδημα πρέπει να χρησιμοποιήσουμε τον τύπο του σταθμικού μέσου.

Ετήσιος Μισθός (χιλ. Ευρώ)	Κεντρικός Όρος (X)	Εργαζόμενοι (Συχνότητα: f)	f X
6 - 12	9	4	36
12 - 18	15	47	705
18 - 24	21	68	1428
24 - 30	27	25	675
30 - 36	33	7	231
36 - 42	39	11	429
42 - 48	45	6	270
48 - 54	51	9	459
54 - 60	57	4	228
60 - 66	63	3	189
66 - 72	69	2	138
72 - 78	75	2	150
78 - 84	81	1	81
84 - 90	87	0	0
90 - 96	93	0	0
96 - 102	99	1	99
Σύνολο		**190**	**5.118**

$\bar{X} = \Sigma f \cdot X/n = 5.118/190 = 26,937$ χιλ. €

Πίνακας 3.2
Εκτίμηση Μέσου Αριθμητικού Κατανομής Συχνοτήτων

Ο μέσος αριθμητικός είναι ο πιο χρήσιμος τρόπος μέτρησης της τάσης ενός δείγματος αριθμητικών δεδομένων. Βασικό του πλεονέκτημα είναι ότι η εκτίμησή του βασίζεται σε όλες τις τιμές του δείγματος. Γι' αυτό τον λόγο καλείται και **επαρκής** εκτιμητής της κεντρικής τάσης. Έχει όμως και μειονεκτήματα. Παρασύρεται από τις ακραίες τιμές και πολλές φορές οδηγεί σε λανθασμένα συμπεράσματα. Αν εξετάσετε με προσοχή τα πρωτογενή στοιχεία, θα διαπιστώσετε ότι τα εισοδήματα δεν είναι τόσο ψηλά όσο ο μέσος μισθός σάς αφήνει να πιστέψετε. Μόνο το 29% των εργαζομένων έχει ετήσιες αμοιβές πάνω από το μέσο εισόδημα των 27,189 χιλ. €. Οι υπόλοιποι παίρνουν χαμηλότερο μισθό.

Αυτό οφείλεται στους λίγους υψηλούς μισθούς των στελεχών που "παρασύρουν" προς το μέρος τους τον μέσο. Εάν εξαιρέσουμε τους μισθούς εννέα στελεχών που έχουν αμοιβές πάνω από 60 χιλ. €, οι μέσες αποδοχές των υπολοίπων 181 υπαλλήλων περιορίζονται σε 24,948 χιλ. €. Και εάν αποκλείσουμε και άλλα 23 στελέχη με μισθό πάνω από 40 χιλ. €, ο μισθός των υπολοίπων 158 υπαλλήλων μειώνεται στα 21,478 χιλ. €. Έτσι, για κατανομές που υπάρχουν ακραίες τιμές ο μέσος αριθμητικός δεν αποτελεί αξιόπιστο τρόπο μέτρησης της τάσης. Σε αυτές τις περιπτώσεις είναι προτιμότερο να χρησιμοποιήσουμε τη διάμεσο, που περιγράφουμε στη συνέχεια.

Διάμεσος

Η **διάμεσος** (Median) είναι η μεσαία τιμή μιας ομάδας τιμών ιεραρχημένων σε αύξουσα τάξη μεγέθους. Εάν δεν υπάρχουν **δεσμοί** (οι τιμές δεν συμπίπτουν μεταξύ τους), τότε οι μισές παρατηρήσεις είναι μικρότερες της διαμέσου και οι άλλες μισές μεγαλύτερες. Έτσι, η διάμεσος δείχνει την τιμή που χωρίζει τις παρατηρήσεις σε δύο ίσες υποομάδες. Ο υπολογισμός της τιμής της είναι εύκολος και το μόνο που προϋποθέτει είναι ότι οι τιμές βρίσκονται σε αύξουσα τάξη μεγέθους.

- Εάν ο αριθμός των παρατηρήσεων είναι περιττός (μονός) αριθμός, τότε η διάμεσος είναι η κεντρική τιμή. Δηλαδή, η (n+1)/2 παρατήρηση, εφόσον οι (n) τιμές του δείγματος τεθούν σε αύξουσα τάξη μεγέθους. Για παράδειγμα, ας εξετάσουμε τους εννέα πιο χαμηλόμισθους υπαλλήλους. Οι μισθοί τους είναι:

Μισθός:	11,340	11,448	11,664	11,880	**12,204**	12,744	13,068	13,500	14,148
Τάξη:	1	2	3	4	**5**	6	7	8	9

  Ο διάμεσος μισθός είναι η (n+1)/2=(9+1)/2=5η παρατήρηση, δηλαδή 12,204 χιλ. €.

- Για άρτιο (ζυγό) αριθμό παρατηρήσεων η διάμεσος ισούται με τον απλό μέσο αριθμητικό των δύο κεντρικών τιμών. Δηλαδή των n/2 και (n/2)+1 παρατηρήσεων. Για παράδειγμα, ας εξετάσουμε τους οκτώ πιο χαμηλόμισθους υπαλλήλους. Οι μισθοί τους είναι:

Μισθός:	11,340	11,448	11,664	**11,880**	**12,204**	12,744	13,068	13,500
Τάξη:	1	2	3	**4**	**5**	6	7	8

  Ο διάμεσος μισθός είναι ο μέσος όρος της 4ης και 5ης παρατήρησης [n/2=8/4=4, και (n/2)+1=(8/2)+1=5]. Δηλαδή, (11,880+12,204)/2 = 12,042 χιλ. €.

Για μεγάλο αριθμό παρατηρήσεων επιστρατεύουμε τους ηλεκτρονικούς υπολογιστές. Έτσι, το διάμεσο εισόδημα των 190 υπαλλήλων είναι 22,140 χιλ. €, που σημαίνει ότι οι μισοί εργαζόμενοι έχουν αμοιβή κάτω από αυτό το ποσό, και οι άλλοι μισοί πάνω (μπορείτε να το επαληθεύσετε από τα ιεραρχημένα στοιχεία του Πίνακα 2.2).

Το πλεονέκτημα της διαμέσου είναι ότι δεν επηρεάζεται από τις ακραίες τιμές. Εάν εξαιρέσουμε τους μισθούς εννέα στελεχών που έχουν αμοιβές πάνω από 60 χιλ. €, οι διάμεσες αποδοχές των υπολοίπων 181 υπαλλήλων διαμορφώνονται στα 21,708 χιλ. €. Και εάν αποκλείσουμε και άλλα 23 στελέχη με μισθό πάνω από 40 χιλ. €, ο διάμεσος μισθός των υπολοίπων 158 υπαλλήλων γίνεται 20,952 χιλ. €. Συγκρίνετε αυτές τις μεταβολές της διαμέσου με τις μεταβολές του μέσου αριθμητικού που προέκυψαν από την εξαίρεση των ίδιων ακραίων τιμών και θα διαπιστώσετε πόσο ανεπηρέαστη παραμένει η διάμεσος.

Ας δούμε τώρα πώς προσδιορίζεται η διάμεσος σε δεδομένα που έχουν ταξινομηθεί σε κατανομή συχνοτήτων. Θα χρησιμοποιήσουμε το διάγραμμα της αθροιστικής πολυγωνικής γραμμής (Διάγραμμα 2.4) και θα εφαρμόσουμε τη μέθοδο της γραμμικής παρεμβολής. Από τον κάθετο άξονα των ποσοστιαίων αθροιστικών συχνοτήτων σύρετε μια οριζόντια γραμμή από την τιμή 50%, και το σημείο που συναντά την πολυγωνική γραμμή το προβάλλετε στον οριζόντιο άξονα των μισθών. Το Διάγραμμα 3.1 δείχνει τη μέθοδο της γραμμικής παρεμβολής.

Είναι εύκολο να αποδείξουμε ότι η παρακάτω διαγραμματική εκτίμηση της διαμέσου μπορεί να γίνει με τον εξής γενικό τύπο:

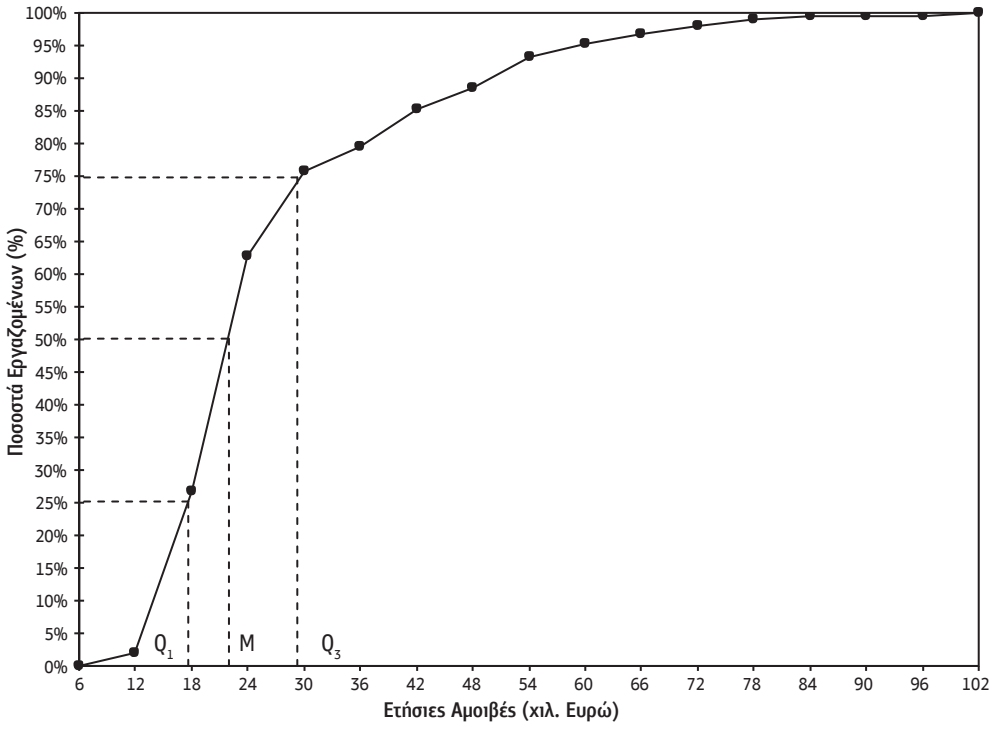
$$M = L_M + \delta \times (n/2 - F_{M-1})/f_M \qquad (3.10)$$

όπου

M = η διάμεσος
L_M = το κάτω όριο του διαστήματος που εντοπίζεται η διάμεσος
δ = το πλάτος του διαστήματος
F_{M-1} = η αμέσως προηγούμενη αθροιστική συχνότητα
f_M = η συχνότητα του διαστήματος που εντοπίζεται η διάμεσος

Η απόδειξη του τύπου βασίζεται στο γεγονός ότι το σημείο που αντιστοιχεί στη διάμεσο βρίσκεται μεταξύ δύο σημείων. Ένα που αντιστοιχεί σε σύνολο παρατηρήσεων μόλις μικρότερο από n/2 (F_{M-1} - με **F** συμβολίζουμε τις αθροιστικές συχνότητες) και ένα που αντιστοιχεί σε αμέσως μεγαλύτερο από n/2 (δηλαδή, την επόμενη αθροιστική συχνότητα). Η διαφορά τους ισούται με τον αριθμό των παρατηρήσεων (συχνότητα) του ενδιάμεσου διαστήματος τάξης. Αυτά όσον αφορά τον κάθετο άξονα των αθροιστικών συχνοτήτων. Όσον αφορά τον άξονα των τιμών, το σημείο της διαμέσου βρίσκεται μεταξύ των δύο διαδοχικών άνω ορίων που απέχουν κατά το πλάτος διαστήματος (**δ**). Από εκεί και πέρα τα πράγματα είναι εύκολα αρκεί πρώτα να εντοπίσουμε το διάστημα που βρίσκεται η διάμεσος.

Με τον ίδιο τρόπο εκτιμούμε και τα τεταρτημόρια (**Quartiles**), δηλαδή τις τιμές που χωρίζουν το σύνολο των παρατηρήσεων σε τέσσερα ίσα (από πλευράς παρατηρήσεων) μέρη. Έτσι, μέχρι την τιμή του πρώτου τεταρτημόριου (Q_1) βρίσκεται το 25% των παρατηρήσεων. Από το πρώτο τεταρτημόριο μέχρι το δεύτερο (που συμπίπτει με



Διάγραμμα 3.1
Εκτίμηση της Διαμέσου και των Τεταρτημορίων με Γραμμική Παρεμβολή

τη διάμεσο, δηλαδή M= Q_2) έχουμε το επόμενο 25% των παρατηρήσεων, από Q_2 έως το τρίτο τεταρτημόριο (Q_3) έχουμε το επόμενο 25% των τιμών και, τέλος, πάνω από Q_3 έχουμε το τελευταίο 25% των παρατηρήσεων. Η διάμεσος και τα τεταρτημόρια ονομάζονται στατιστικές **θέσης** (location), με την έννοια ότι δείχνουν πού συγκεντρώνονται (θέτονται) συγκεκριμένα ποσοστά των παρατηρήσεων.

Οι τύποι υπολογισμού των τεταρτημορίων έχουν ως εξής:

$$Q_1 = L_{Q1} + \delta \times (n/4 - F_{Q1-1})/f_{Q1} \qquad (3.11)$$

και

$$Q_3 = L_{Q3} + \delta \times (3\times n/4 - F_{Q3-1})/f_{Q3} \qquad (3.12)$$

Οι τύποι (3.11) και (3.12) προκύπτουν με την ίδια λογική που προέκυψε ο τύπος της διαμέσου (3.10). Η μόνη διαφορά είναι ότι στην περίπτωση του Q_1 αναζητούμε την (n/4) παρατήρηση, ενώ για το Q_3 την (3 × n/4) παρατήρηση.

Με βάση τον Πίνακα κατανομής συχνοτήτων 2.3 έχουμε

$$Q_1 = L_{Q1} + \delta \times (n/4 - F_{Q1-1})/f_{Q1} = 12+6\times(190/4-4)/47 = 17{,}553 \text{ χιλ. €}$$
$$M = Q_2 = L_M + \delta \times (n/2 - F_{M-1})/f_M = 18+6\times(190/2-51/68 = 21{,}882 \text{ χιλ. €}$$
$$Q_3 = L_{Q3} + \delta \times (3\times n/4 - F_{Q3-1})/f_{Q3} = 24+6\times(3\times190/4-119)/25 = 29{,}640 \text{ χιλ. €}$$

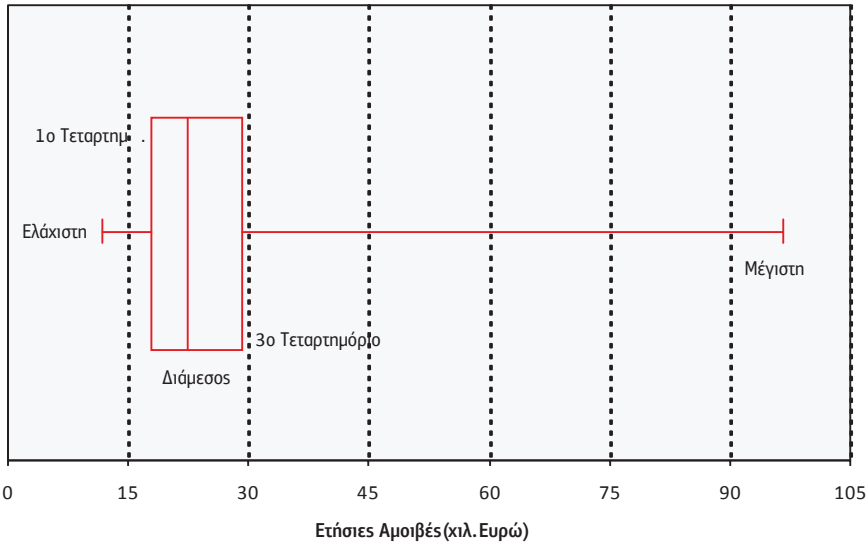
Ενώ, από τα αταξινόμητα (πρωτογενή) δεδομένα του Πίνακα 2.2 προκύπτουν οι εξής εκτιμήσεις:

$$Q_1 = 17{,}604 \text{ χιλ. €}$$
$$M = Q_2 = 22{,}140 \text{ χιλ. €}$$
$$Q_3 = 29{,}025 \text{ χιλ. €}$$

που είναι λογικό να διαφέρουν από εκείνες που προέκυψαν από τα ταξινομημένα δεδομένα. Στα ταξινομημένα δεδομένα του Πίνακα 2.3, η γραμμική παρεμβολή προϋποθέτει ότι οι παρατηρήσεις που ανήκουν σε ένα διάστημα τιμών είναι ομοιόμορφα κατανεμημένες. Για παράδειγμα, οι 4 μισθοί του διαστήματος 6 - 12, διαφέρουν η μία από την άλλη κατά (12-6)/4 χιλ. € (δηλαδή 1,5 χιλ. €). Οι 47 του διαστήματος 12 - 18, ξεκινούν από 12 χιλ. € και αυξάνονται διαδοχικά κατά (18-12)/47 χιλ. € (δηλαδή περίπου 128 €), κ.ο.κ. Σε αυτήν την υπόθεση στηρίζονται και οι τύποι 3.10, 3.11, και 3.12.

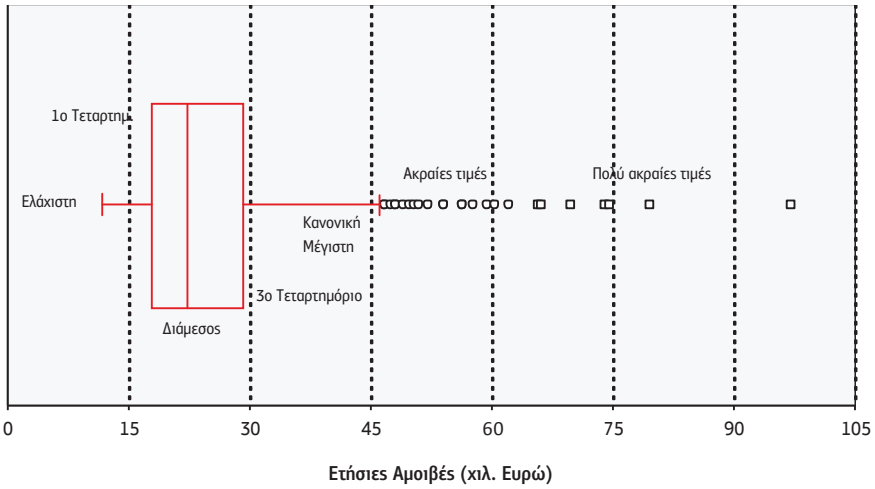
Ένα χρήσιμο διάγραμμα που απεικονίζει τις τιμές των τεταρτημορίων είναι το διάγραμμα Box and Whisker ή Διάγραμμα Πλαισίου και Απολήξεων (Διάγραμμα 3.2). Είναι τόσο χρήσιμο και δημοφιλές μεταξύ των αναλυτών, που όλα τα στατι-

(α) Box and Whisker Διάγραμμα των Ετήσιων Αποδοχών (χωρίς ακραίες τιμές)

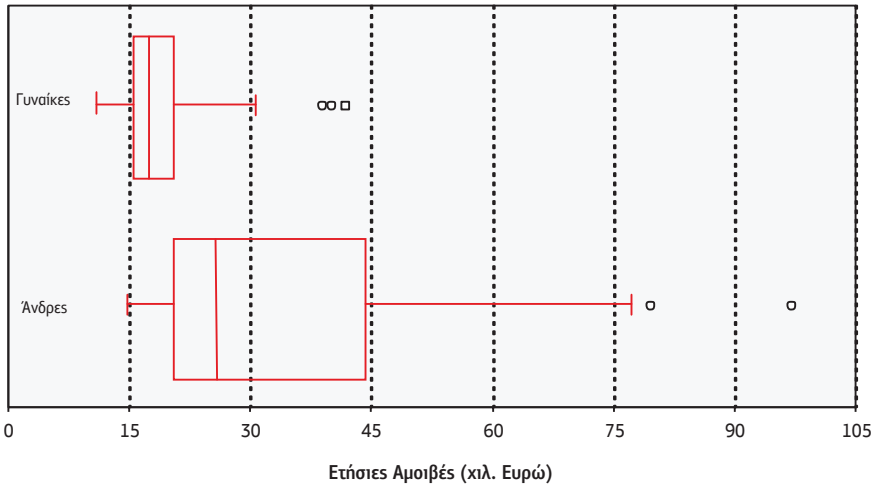


(α)

(β) Box and Whisker Διάγραμμα των Ετήσιων Αποδοχών (με απεικόνιση των ακραίων τιμών)



(γ) Πολλαπλό Box and Whisker Διάγραμμα των Ετήσιων Αποδοχών
(με απεικόνιση των ακραίων τιμών)



Διάγραμμα 3.2
Box and Whisker Διαγράμματα των Ετήσιων Αποδοχών

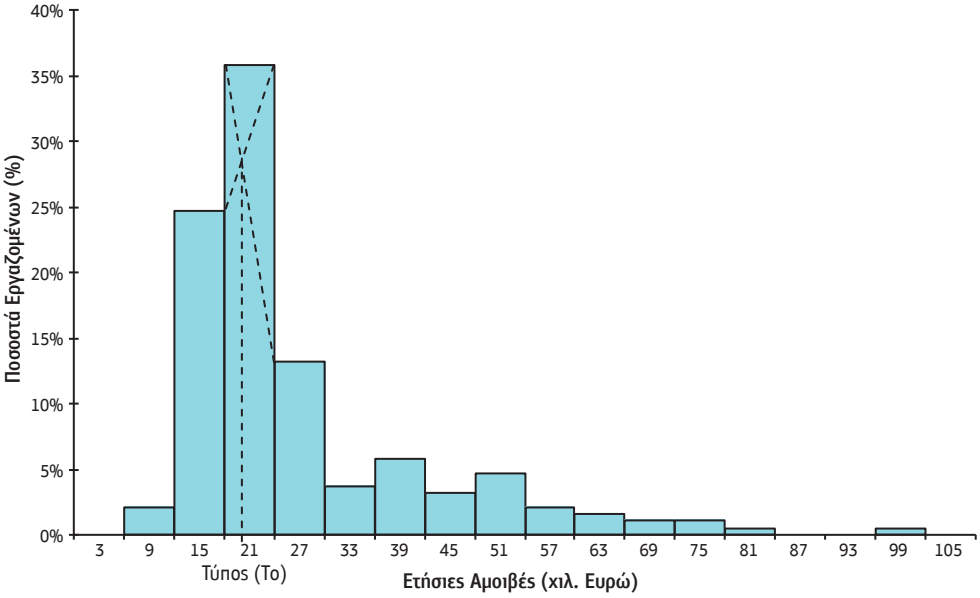
στικά προγράμματα το περιλαμβάνουν στις επιλογές τους. Είναι σαν να βλέπετε το ιστόγραμμα (Διάγραμμα 2.1) από πάνω (κάτοψη). Στις δύο άκρες αποτυπώνονται η ελάχιστη και η μέγιστη τιμή. Ενώ, από το 1ο έως το 3ο τεταρτημόριο επεκτείνεται ένα ορθογώνιο που απεικονίζει το εύρος των τιμών που συγκεντρώνεται το κεντρικό 50% των παρατηρήσεων. Τέλος, η κάθετη γραμμή αντιστοιχεί στην τιμή της διαμέσου. Με το διάγραμμα Box and Whisker μπορούμε να εντοπίσουμε και τις ακραίες τιμές. Μία τιμή χαρακτηρίζεται ως **ακραία** (outlier) εάν είναι μικρότερη από το Q_1 ή μεγαλύτερη από το Q_3 περισσότερο από 1,5 φορές την τεταρτημοριακή απόκλιση (Q_3-Q_1), ανάλογα με το άν οι ακραίες τιμές είναι προς τα κάτω ή προς τα πάνω.

Για παράδειγμα η τεταρτημοριακή απόκλιση των μισθών ισούται με Q_3-Q_1 = 29,025 - 17,604 = 11,421 χιλ. €. Επομένως, επειδή παρατηρούνται ορισμένοι ασυνήθιστα υψηλοί μισθοί, όποιος μισθός είναι υψηλότερος από Q_3+1,5·(Q_3- Q_1) = 29,025+1,5·(29,025-17,604) = 46,157 χιλ. € (που αποτελεί την κανονική μέγιστη τιμή) απεικονίζεται με ένα κύκλο (○). Και εάν μία τιμή είναι μεγαλύτερη από το Q_3 περισσότερο από 3 φορές την τεταρτημοριακή απόκλιση (Q_3-Q_1), δηλαδή Q_3+3·(Q_3-Q_1) = 29,025+3·(29,025-17,604) = 63,288 χιλ. €, χαρακτηρίζεται ως **πολύ ακραία** τιμή (extreme value) και απεικονίζεται με ένα τετράγωνο (□). Επειδή οι μισθοί παρουσιάζουν έντονη ασυμμετρία είναι λογικό να συναντούμε τόσο ακραίες όσο και πολύ ακραίες τιμές. Όμως, προσέξτε το 3ο διάγραμμα Box and Whisker στο οποίο οι μισθοί διακρίνονται σε δύο κατηγορίες, ανάλογα με το φύλο των εργαζομένων, και ως εκ τούτου ονομάζεται πολλαπλό Box and Whisker. Τώρα η εικόνα είναι διαφορετική. Επειδή οι μισθοί ανά φύλο έχουν μεγαλύτερη ομοιογένεια, δεν συναντάμε πολλές ακραίες τιμές και στην περίπτωση των ανδρών έχουμε δύο ακραίες τιμές, και στους μισθούς των γυναικών (που είναι σαφώς χαμηλότεροι των ανδρών) διακρίνουμε δύο ακραίες τιμές και μία πολύ ακραία τιμή.

Σημείο Μέγιστης Συχνότητας

Το **σημείο μέγιστης συχνότητας** (ή **επικρατούσα τιμή** ή **τύπος**), που συμβολίζεται με T_0, είναι η τιμή με τη μεγαλύτερη συχνότητα σε ένα σύνολο μετρήσεων. Εάν τα στοιχεία είναι αταξινόμητα, τότε η επικρατούσα τιμή είναι εκείνη που εμφανίζεται συχνότερα. Για παράδειγμα, στο χαρακτηριστικό (μεταβλητή) "αριθμός παιδιών ανά οικογένεια" η συχνότερα απαντώμενη τιμή είναι το 2. Στην περίπτωση του αριθμού των παιδιών η εκτίμηση του τύπου είναι εύκολη, διότι η συγκεκριμένη μεταβλητή είναι ασυνεχής (ακέραιος αριθμός).

Όμως, στην περίπτωση των συνεχών μεταβλητών, και για αταξινόμητα δεδομένα, ο προσδιορισμός του τύπου είναι αδύνατος, διότι είναι πολύ πιθανό όλες οι τιμές να διαφέρουν μεταξύ τους. Σε αυτές τις περιπτώσεις πρώτα κατασκευάζουμε την κατανομή συχνοτήτων και στη συνέχεια εκτιμούμε τον τύπο (επικρατούσα τιμή) T_0. Η τιμή του εντοπίζεται στο διάστημα με τη μεγαλύτερη συχνότητα και προσεγγίζεται διαγραμματικά από το ιστόγραμμα (Διάγραμμα 3.3).



Διάγραμμα 3.3
Εκτίμηση του Τύπου (Σημείο Μέγιστης Συχνότητας)

3.3. **Μέτρηση της Διασποράς**

Η δεύτερη σημαντική ιδιότητα που χαρακτηρίζει ένα σύνολο αριθμητικών δεδομένων είναι η διασπορά ή μεταβλητότητα. Η διασπορά είναι το μέγεθος της ανομοιογένειας μεταξύ των τιμών, δηλαδή πόσο διαφέρουν μεταξύ τους ή πόσο "διεσπαρμένες" είναι οι τιμές. Οι κατανομές των αρχικών και σημερινών μισθών που παρουσιάζονται στο Διάγραμμα 2.3 διαφέρουν σε μεταβλητότητα. Οι αρχικοί μισθοί είναι πιο συγκεντρωμένοι μεταξύ τους, δηλαδή έχουν μικρότερη διασπορά σε σύγκριση με τους σημερινούς μισθούς.

Πέντε είναι οι συνηθέστεροι τρόποι μέτρησης της διασποράς: το εύρος, η τεταρτημοριακή απόκλιση, η διακύμανση, η τυπική απόκλιση, και ο συντελεστής μεταβλητότητας.

Εύρος

Το **εύρος** (**R**ange) είναι η διαφορά μεταξύ της μεγαλύτερης (X_{max}) και της μικρότερης τιμής (X_{min}) των δεδομένων.

Επομένως, το εύρος των σημερινών μισθών του Πίνακα 2.2 είναι:

$$R = X_{max} - X_{min} = 97,200 - 11,340 = 85,860 \text{ χιλ. } €$$

Ενώ, για τους αρχικούς μισθούς το εύρος είναι: