κεφάλαιο 3

ΜΕΤΡΑ ΘΕΣΗΣ

- Ο Αριθμητικός Μέσος
- Ο Σταθμισμένος Μέσος
- Η Διάμεσος
- Η Επικρατούσα Τιμή
- Μέση σχετική μεταβολή. Ο Γεωμετρικός Μέσος

3.1 Γενικά

Τα μέτρα θέσης ή μέσες τιμές είναι αριθμοί οι οποίοι προσδιορίζουν περιληπτικά την τάση των δεδομένωννα συγκεντρώνονταιεπάνω στην ευθεία των πραγματικών αριθμών. Γι'αυτό και αναφέρονται εναλλακτικά και ως **μέτρα κεντρικής τάσης** (measures of central tendency). Κάθε μέτρο θέσης θ ικανοποιεί τις ακόλουθες δύο συνθήκες:

α. Βρίσκεται μέσα στα όρια των δεδομένων $x_1, x_2, ..., x_n$ δηλαδή ισχύει $\min_i\{x_i\} \le \theta \le \max_i\{x_i\},\ \ i = 1,...,n$

β. Αν σε κάθε τιμή x_i προστεθεί μία σταθερά α (*μετασχηματισμός αρχής*) και/ή κάθε τιμή x_i πολλαπλασιαστεί μία σταθερά β (*μετασχηματισμός κλίμακας*), τότε στο θ προστίθεται επίσης το α και/ή το θ πολλαπλασιάζεται με το β.

Τα κυριότερα μέτρα θέσης είναι ο αριθμητικός μέσος, η διάμεσος και η επικρατούσα τιμή τα οποία παρουσιάζουμε αμέσως στη συνέχεια. Μαζί θα παρουσιάσουμε και τον σταθμισμένο μέσο ο οποίος έχει πολλές εφαρμογές και τον αποκομμένο μέσο που είναι χρήσιμος όταν στα δεδομένα υπάρχουν ακραίες τιμές. Τέλος, στο κεφάλαιο αυτό θα παρουσιάσουμε και τον γεωμετρικό μέσο οοποίος είναι χρήσιμος στον υπολογισμό της μέσης ποσοστιαίας μεταβολής, να τονιστεί όμως ότι για τον μέσο αυτό δεν ισχύει η συνθήκη β.

3.2 Ο Αριθμητικός Μέσος

Ο αριθμητικός μέσος (arithmetic mean) ή απλώς **ο μέσος** (the mean) των τιμών $x_1, x_2, ..., x_n$ συμβολίζεται με $\bar{x}$ και είναι **ο μέσοςόρος** (the average) τους δηλαδή

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (3.2.1)$$

○○● **Σημείωση** Έχει καθιερωθεί διεθνώς, ο αριθμητικός μέσος παρατηρήσεων οι οποίες αναφέρονται σε ολόκληρο τον πληθυσμό να συμβολίζεται με μ.

Επαναλαμβανόμενες τιμές

Όταν στις n παρατηρήσεις των δεδομένων μας οι τιμές $x_1, x_2, ..., x_\kappa$ εμφανίζονται με συχνότητες, αντίστοιχα, $f_1, f_2, ..., f_\kappa$, με $\kappa<n$ και $\sum_{i=1}^{n} f_i = n$, τότε είναι πρακτικό, στον τύπο του αριθμητικού μέσου να πάρουμε f_i φορές την αντίστοιχη τιμή x_i ή, διαφορετικά, να υπολογίσουμε:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + ... + x_\kappa f_\kappa}{f_1 + f_2 + ... + f_\kappa} = \frac{\sum_{i=1}^{\kappa} x_i f_i}{\sum_{i=1}^{\kappa} f_i} = \frac{\sum_{i=1}^{\kappa} x_i f_i}{n} \qquad (3.2.2)$$

Παράδειγμα 3.2.1 Ρωτήσαμε 20 φοιτητές που είναι στο πτυχίο τον αριθμό Χ των μαθημάτων που "οφείλουν" και πήραμε τα ακόλουθα αποτελέσματα:

Αριθμός μαθημάτων x_i	Αριθμός φοιτητών f_i	$x_i f_i$
1	3	3
2	4	8
3	3	9
4	3	12
5	2	10
6	2	12
9	1	9
11	1	11
17	1	17
Άθροισμα	20	91

Επομένως ο μέσος αριθμός μαθημάτων για το πτυχίο, ανά φοιτητή ισούται με:

$$\bar{x} = \frac{\sum_{i=1}^{9} x_i f_i}{20} = \frac{91}{20} = 4.55 \text{ μαθήματα}$$

Παράδειγμα 3.2.2 Στον πίνακα που ακολουθεί δίνεται η κατανομή συχνοτήτων του αριθμού Χ των παιδιών που γέννησαν 653 γυναίκες, οι οποίες παντρεύτηκαν μετά τα 30 τους χρόνια. Στον ίδιο πίνακα δίνονται οι απαραίτητοι υπολογισμοί για τον προσδιορισμό του αριθμητικού μέσου

Αριθμός παιδιών x_i	Αριθμός γυναικών f_i	$x_i f_i$
0	123	0
1	152	152
2	154	308
3	96	288
4	53	212
5	30	150
6	17	102
7	10	70
8	6	48
9	5	45
10	7	70
Άθροισμα	653	1445

Ο μέσος αριθμός παιδιών ανά γυναίκα είναι

$$\bar{x} = \frac{\sum_{i=1}^{11} x_i f_i}{\sum_{i=1}^{11} f_i} = \frac{1445}{653} = 2.2$$

Ομαδοποιημένες παρατηρήσεις

Ο τύπος (3.2.2) μπορεί να εφαρμοστεί και για να υπολογίσουμε τον αριθμητικό μέσο παρατηρήσεων οι οποίες είναι ομαδοποιημένες σε μία κατανομή. Στην περίπτωση αυτή οι τιμές $x_1, x_2, ..., x_\kappa$ είναι οι κεντρικές τιμές των αντίστοιχων ταξικών διαστημάτων. Με τον τρόπο αυτό αντικαθιστούμε τις f_i τιμές κάθε ταξικού διαστήματος i με την κεντρική τιμή x_i. Επομένως, ο αριθμητικός μέσος που υπολογίζεται από την κατανομή συχνοτήτων αποτελεί *προσέγγιση*[1] του αριθμητικού μέσου των δεδομένων και θα πρέπει να αποφεύγεται όταν είναι διαθέσιμα τα πρωτογενή δεδομένα.

Σήμεραπάντως με την εκτεταμένη χρήση Η/Υ και τη δυνατότητα μεταφοράς και επεξεργασίας μεγάλου όγκου δεδομένων, είναι λίγες οι περιπτώσεις που θα χρειαστεί να υπολογίσουμε τον αριθμητικό μέσο αλλά και τα άλλα περιγραφικά μέτρα από την κατανομή-συνήθως έχουμε πρόσβαση στα πρωτογενήδεδομένα.

[1]Πάντως, η διαφορά ανάμεσα στις δύο τιμές δεν μπορεί να είναι μεγαλύτερη από το μισό του ταξικού εύρους

Ιδιότητες του αριθμητικού μέσου

Ο αριθμητικός μέσος $\bar{x}$ των τιμών $x_1, x_2, ..., x_n$ είναι μέτρο θέσης και άρα ισχύει:

a. $\min_i\{x_i\} \le \bar{x} \le \max_i\{x_i\}$, i= 1,…,n

b. Η αντικατάσταση των τιμών x_i με τις $x_i' = a + bx_i$ έχει ως αποτέλεσμα τον ίδιο μετασχηματισμό του $\bar{x}$

$$\bar{x}' = \bar{x}_{a+bx} = a + b\bar{x}$$

Ο $\bar{x}$ έχει επιπλέον τις ακόλουθες ιδιότητες:

c. Το άθροισμα των αποκλίσεων των τιμών $x_1, x_2, ..., x_n$ από το μέσο αριθμητικό τους $\bar{x}$ ισούται με μηδέν δηλαδή

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

d. Αν $\bar{x}_1$ ο αριθμητικός μέσος n_1 τιμών και $\bar{x}_2$ ο αριθμητικός μέσος n_2 τιμών τότε ο αριθμητικός μέσος των $n_1 + n_2$ τιμών ισούται με

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Ηιδιότητα αυτή με μαθηματική επαγωγή γενικεύεται για r>2, (r πεπερασμένο) σύνολα τιμών ως εξής

$$\bar{x} = \frac{n_1\bar{x}_1 + ... + n_r\bar{x}_r}{n_1 + ... + n_r} = \frac{\sum_{i=1}^{r} n_i\bar{x}_i}{\sum_{i=1}^{r} n_i}$$

e. Το άθροισμα των τετραγωνικών αποκλίσεων των τιμών $x_1, ..., x_n$ από τον αριθμητικό τους μέσο είναι ελάχιστο. Δηλαδή ισχύει:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - \alpha)^2, \quad \forall\alpha > 0$$

Η ιδιότητα αυτή έχει ως αποτέλεσμα ο αριθμητικός να χρησιμοποιείται ως εκτίμηση των παρατηρήσεων $x_i, ..., x_n$, όταν στόχος της εκτίμησης είναι η

ελαχιστοποίηση των τετραγωνικών αποκλίσεων (βλέπε και Β τόμο αυτού του βιβλίου).

Πλεονεκτήματα και μειονεκτήματα του αριθμητικού μέσου

Τα πιο σημαντικά **πλεονεκτήματα** του αριθμητικού μέσου είναι τα εξής:

- Είναι μία παράμετρος θέσης, η οποία είναι καλά ορισμένη, υπολογίζεται εύκολα και γίνεται κατανοητή και από τον μη ειδικό
- Παίρνει υπόψη όλες τις τιμές των δεδομένων και επηρεάζεται απ' αυτές
- Μπορεί να χρησιμοποιηθεί για περαιτέρω στατιστική ανάλυση. Όπως, ειδικότερα, θα δούμε σε επόμενα κεφάλαια, η διαδικασία εκτίμησης της μέσης τιμής πληθυσμού με βάση τον αριθμητικό μέσο τυχαίου δείγματος επηρεάζεται από τις διακυμάνσεις της δειγματοληψίας λιγότερο από οποιαδήποτε άλλη διαδικασία εκτίμησης.

Ο αριθμητικός μέσος όμως έχει και ορισμένα **μειονεκτήματα**. Τα σημαντικότερα απ' αυτά είναι τα εξής:

- Επηρεάζεται πολύ από τις ακραίες τιμές. Έτσι π.χ. ο αριθμητικός μέσος των τιμών 3, 5, 7, 7, 9, 11 ισούται με 7. Αν όμως η τελευταία τιμή δεν ήταν 11 αλλά 65, τότε ο μέσος αριθμητικός θα ήταν ίσος με 16. Έτσι μία ή περισσότερες ακραίες τιμές μπορεί να μειώσουν σημαντικά την αντιπροσωπευτικότητα του αριθμητικού μέσου και να τον κάνουν ακατάλληλο μέτρο κεντρικής τάσης. Γενικά, ο αριθμητικός μέσος είναι ανεπαρκής για να προσδιορίσει τη θέση της κατανομής όταν αυτή είναι πολύ λοξή.
- Μπορεί να πάρει μια τιμή η οποία να μην αποτελεί δυνατή τιμή για τη μεταβλητή στην οποία αναφέρονται τα δεδομένα. Έτσι ο μέσος αριθμητικός ενός συνόλου παρατηρήσεων μιας ασυνεχούς μεταβλητής μπορεί να μην έχει και ουσιαστικό περιεχόμενο (όπως π.χ. μέσος αριθμητικός παιδιών ανά οικογένεια ίσος με 2.75).


Οι περισσότεροι άνθρωποι σ' αυτή την πόλη έχουν περισσότερα πόδια από το μέσο όρο

 Στατιστικός διεθνούς κύρους πνίγηκε σε πισίνα με μέσο βάθος ένα μέτρο!

 Όταν ο υπουργός υγείας διάβασε σε μια αναφορά ότι πέρυσι πέθαναν από γρίπη 12.5 άτομα ανά 1000 κατοίκους, αναρωτήθηκε πώς είναι δυνατό να έχουν πεθάνει δώδεκα και μισό άτομα. Ο Γενικός Γραμματέας του υπουργείου του απάντησε: *"όταν οι στατιστικοί λένε ότι πέθαναν 12.5 άτομα εννοούν ότι στην πραγματικότητα πέθαναν 12 και ο 13^{ος} χαροπαλεύει"*

(παραλλαγή από Rao, 1989)

 Τι δύσκολα θέματα έβαλε τον Ιούνιο! Να φανταστείς όλοι γράψαμε κάτω από το μέσο όρο!

(σχόλιο φοιτητή)

Ο αποκομμένος μέσος

Είδαμε ότι εξαιρετικά μεγάλες ή εξαιρετικά μικρές τιμές στα δεδομένα επηρεάζουν τον αριθμητικό μέσο έτσι ώστε να μη σηματοδοτεί ικανοποιητικά τη θέση του κυρίως όγκου των δεδομένων. Γι' αυτό σε πολλές πρακτικές εφαρμογές υπολογίζεται ο **αποκομμένος μέσος** (trimmed mean), ο οποίος υπολογίζεται αφαιρώντας από τα δεδομένα ορισμένο ποσοστό μεγαλύτερων και μικρότερων τιμών. Έτσι π.χ. ο κατά 20% αποκομμένος μέσος είναι ο αριθμητικός μέσος ο οποίος υπολογίζεται αν αγνοήσουμε το 20% των μεγαλύτερων και το 20% των μικρότερων τιμών. Ομοίως, ο κατά 10% αποκομμένος μέσος είναι ο μέσος που υπολογίζεται μετά την αφαίρεση του 10% των μεγαλύτερων και του 10% των μικρότερων τιμών των δεδομένων.

☼☼☼**Κουΐζ** Στις δημοσιευμένες λίστες του North Carolina University για τις αποδοχές που (κατά μέσο όρο) απολαμβάνουν οι απόφοιτοί του, ανά ειδικότητα, πρώτη εμφανίζεται η ειδικότητα του Γεωγράφου. Πώς μπορεί να ερμηνεύσει κανείς αυτή την απρόσμενη πρωτιά; (*Η απάντηση στο κάτω μέρος της σελίδας*[1])

[1]Στους αποφοίτους με αυτή την ειδικότητα συγκαταλέγεται και ο μπασκετμπολίσταςτου NBA Μάικλ Τζόρνταν!

3.3 Ο Σταθμισμένος Μέσος

Ο **σταθμισμένος μέσος** (weighted mean or average) των τιμών $x_1, x_2, ..., x_n$ με αντίστοιχους **συντελεστές στάθμισης** ή **συντελεστές βαρύτητας** (weights) $w_1, w_2, ..., w_n$ συμβολίζεται με $\bar{x}_w$ και ισούται με:

$$\bar{x}_w = \mu_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

Η τιμή του συντελεστή στάθμισης w_i της x_i προσδιορίζεται ανάλογα με την σπουδαιότητα της x_i στο σύνολο των τιμών.

Παράδειγμα 3.3.1 Για ένα υποθετικό Τμήμα ΑΕΙ δίνεται ο αριθμός των εγγεγραμμένων σε κάθε έτος φοιτητών και ο αντίστοιχος αριθμός των μαθημάτων.

Έτος	Αριθμός φοιτητών x_i	Αριθμός μαθημάτων w_i	$x_i w_i$
1	1200	14	16800
2	1050	17	17850
3	720	12	8640
4	530	15	7950
Άθροισμα	3500	58	51240

Για να υπολογίσουμε το μέσο αριθμό φοιτητών ανά μάθημα σταθμίζουμε τον αριθμό x_i των φοιτητών του έτους ίμε τον αντίστοιχο αριθμό μαθημάτων και υπολογίζουμε τον σταθμικό μέσο:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} = \frac{51240}{58} = 883.44$$

Δηλαδή σε κάθε μάθημα αντιστοιχούν κατά μέσο όρο 883.44 φοιτητές. Αν στο τμήμα υπάρχουν 28 διδάσκοντες (μέλη ΔΕΠ και ειδικοί επιστήμονες) θα υπολογίσουμε την αναλογία διδασκόντων προς διδασκόμενους ως εξής: Σε κάθε διδάσκοντα αντιστοιχούν ετησίως 58/28=2.07 μαθήματα κατά μέσο όρο, επομένως αντιστοιχούν (2.07)(883.44) = 1829.9 φοιτητές.

Ο Σταθμισμένος μέσος ως αριθμητικός μέσος

Όταν από δύο ή περισσότερα δείγματα ή πληθυσμούς διαθέτουμε τον αριθμητικό μέσο και θέλουμε να υπολογίσουμε τον αντίστοιχο μέσο της ενοποίησής τους τότε, σύμφωνα και με την **ιδιότητα c** του αριθμητικού μέσου, δεν μπορούμε να πάρουμε απλώς τον μέσο όρο τους. Ο κάθε αριθμητικός μέσος θα πρέπει να σταθμιστεί με το αντίστοιχο μέγεθος δείγματος και έτσι η επίδρασή του στο τελικό αποτέλεσμα να είναι ανάλογη με αυτό. Επομένως ο τύπος

$$\bar{x} = \frac{n_1\bar{x}_1 + \ldots + n_r\bar{x}_r}{n_1 + \ldots + n_r} = \frac{\sum_{i=1}^{r} n_i\bar{x}_i}{\sum_{i=1}^{r} n_i}$$

ορίζει έναν σταθμισμένο μέσο με συντελεστές στάθμισης τα μεγέθη των επιμέρους δειγμάτων.

Με την ίδια συλλογιστική συνδυάζουμε ποσοστά που υπολογίζονται σε δείγματα διαφορετικού μεγέθους. Ετσι, όταν ενοποιούμε δείγματα μεγέθους $n_1,\ldots,n_r$ στα οποία η αναλογία κάποιου δίτιμου χαρακτηριστικού είναι, αντίστοιχα, $p_1,\ldots,p_r$ τότε η αναλογία στο ενοποιημένο δείγμα υπολογίζεται ως εξής:

$$p = \frac{n_1p_1 + \ldots + n_rp_r}{n_1 + \ldots + n_r}$$

δηλαδή ως σταθμισμένος μέσος των επιμέρους αναλογιών.

Μια εφαρμογή του σταθμισμένου μέσου στη δειγματοληψία[1]

Έστω ότι στη δημόσια Μέση Εκπαίδευση το 85% των καθηγητών ξένων γλωσσών είναι της Αγγλικής, το 10% της Γαλλικής και το 5% της Γερμανικής γλώσσας και θέλουμε να καταγράψουμε την άποψή τους για θέματα που τους αφορούν, σε ένα τυχαίο δείγμα. Οι πόροι μας υπαγορεύουν δείγμα 200 ατόμων και μια *αναλογικά στρωματοποιημένη τυχαία δειγματοληψία* (βλέπε κεφάλαιο 17) υπαγορεύει να πάρουμε περίπου 10 καθηγητές της Γερμανικής και 20 της Γαλλικής. Αυτά είναι πολύ μικρά δείγματα για να βγάλουμε αξιόπιστα συμπεράσματα για όλους τους

[1]Από διδακτορική διατριβή που υποβλήθηκε στο Τμήμα Γαλλικής Γλώσσας του ΑΠΘ. Τα νούμερα για λόγους ευκολίας έχουν τροποποιηθεί

καθηγητές των αντίστοιχων γλωσσών. Στην περίπτωση αυτή ο ερευνητής μπορεί να διαλέξει να πάρει π.χ. 40 (=4x10) της Γερμανικής και ομοίως 40 (=2x20) της Γαλλικής. Όμως αν θελήσουμε να εκτιμήσουμε π.χ. τη μέση ηλικία σε όλους τους καθηγητές ξένων γλωσσών, επειδή οι αντίστοιχοι της Γαλλικής και της Γερμανικής υπερ-αντιπροσωπεύονται στο δείγμα, θα πρέπει ο αριθμητικός μέσος του δείγματος της κάθε γλώσσας i να σταθμιστεί με τον ακόλουθο συντελεστή:

$$w_i = p_i \frac{n}{f_i}$$

όπου f_i η συχνότητα της αντίστοιχης γλώσσας στο δείγμα και η p_i αναλογία της στον πληθυσμό. Ετσι π.χ. επειδή πήραμε $f_i = 40$ καθηγητές της Γερμανικής στο δείγμα των 400 αντί για (0.05)200 = 10, ο αντίστοιχος συντελεστής στάθμισης θα είναι

$$w_{\Gamma\varepsilon\rho} = (0.05)\frac{200}{40} = 0.25$$

ενώ οι συντελεστές στάθμισης της Γαλλικής και της Αγγλικής θα ισούνται, αντίστοιχα, με:

$$w_{\Gamma} = (0.10)\frac{200}{40} = 0.50$$

και

$$w_{A} = (0.85)\frac{200}{120} = 1.42$$

Έτσι, επειδή οι καθηγητές της Αγγλικής υπο-αντιπροσωπεύονται, ο αριθμητικός τους μέσος θα πολλαπλασιαστεί με το 1.42.

3.4 Η Διάμεσος

Η **Διάμεσος** ή **Διχοτόμος** (median) ενός συνόλου n τιμών, συμβολίζεται με m στις παρατηρήσεις δείγματος και με M στις παρατηρήσεις πληθυσμού και είναι η τιμή της μεσαίας παρατήρησης όταν όλες οι παρατηρήσεις είναι ταξινομημένες σε αύξουσα ή φθίνουσα σειρά. Έτσι, μισές παρατηρήσεις έχουν τιμή μεγαλύτερη ή ίση με τη διάμεσο και μισές έχουν τιμή μικρότερη ή ίση μ' αυτήν.

Η Διάμεσος n παρατηρήσεων οι οποίες είναι ταξινομημένες σε αύξουσα σειρά θα βρίσκεται στη θέση της (n+1)/2 παρατήρησης. Όταν το n είναι αριθμός περιττός στη θέση αυτή βρίσκεται η μεσαία παρατήρηση. Όταν το n είναι ζυγός αριθμός τότε το πηλίκον (n+1)/2 δεν είναι ακέραιος ή, αντίστοιχα η θέση (n+1)/2 θα βρίσκεται στα μισά του διαστήματος μεταξύ της θέσης n/2 και της (n/2)+1. Στην περίπτωση αυτή έχει καθιερωθεί να παίρνεται ως διάμεσος το ημιάθροισμα των παρατηρήσεων που βρίσκονται στις αντίστοιχες θέσεις.

Παράδειγμα 3.4.1 Η διάμεσος των 9 παρατηρήσεων

28 16 17 **21** 33 33 35 37

είναι η τιμή της $5^{ης}$ παρατήρησης, δηλαδή m= 21. Σημειώνεται ότι η τιμή αυτή είναι η $5^{η}$ μεγαλύτερη και η $5^{η}$ μικρότερη.

Ως διάμεσο των παρατηρήσεων

100 150 170 **220** **230** 270 380 400

παίρνουμε την

$$m = \frac{220 + 230}{2} = 225$$

και εδώ έχουμε 4 παρατηρήσεις μικρότερες και 4 μεγαλύτερες από τη διάμεσο.

Ένας τρόπος να προσδιορίσουμε τη Διάμεσο ενός συνόλου παρατηρήσεων χωρίς να τις ξαναγράψουμε σε αύξουσα σειρά είναι ο εξής: ενώνουμε την μεγαλύτερη με την μικρότερη τιμή, στη συνέχεια ανά ζεύγη τις δύο αμέσως μικρότερη και αμέσως μεγαλύτερη κ.ο.κ. μέχρις ότου μείνει μια μεσαία (και αυτή είναι η διάμεσος) ή δύο μεσαίες (και διάμεσος είναι το ημιάθροισμά τους). Ο τρόπος όμως αυτός είναι αποτελεσματικός μόνον όταν το μέγεθος των παρατηρήσεων δεν είναι πολύ μεγάλο

$$m = \frac{1}{2}(220 + 230) = 225$$

Προσδιορισμός της Διαμέσου από την κατανομή συχνοτήτων

Σε ορισμένες περιπτώσεις δεν διαθέτουμε σε ηλεκτρονική μορφή το αρχείο με τα πρωτογενή δεδομένα οπότε είμαστε υποχρεωμένοι να προσδιορίσουμε την διάμεσο απο την κατανομή συχνοτήτων. Διακρίνουμε δύο περιπτώσεις:

- **Η κατανομή αφορά παρατηρήσεις $x_1,...,x_κ$, οι οποίες επαναλαμβάνονται με συχνότητες $f_1,...,f_κ$.** Η διάμεσος είναι η πρώτη τιμή στην οποία αντιστοιχεί αθροιστική συχνότητα μεγαλύτερη ή ίση με (n+1)/2.

Παράδειγμα 3.4.2 Δίνεται η κατανομή συχνοτήτων και η κατανομή αθροιστικών συχνοτήτων του αριθμού των παιδιών που γέννησαν 653 γυναίκες οι οποίες παντρεύτηκαν μετά τα 30 τους χρόνια.

x_i	f_i	F_i
0	123	123
1	152	275
2	154	429
3	96	525
4	53	578
5	30	608
6	17	625
7	10	635
8	6	641
9	5	646
10	7	653
Άθροισμα	653	

Επειδή n= 653 είναι περιττός η διάμεσος είναι η τιμή της (653+1)/2 = $327^{ης}$ παρατήρησης.
Από την κατανομή αθροιστικών συχνοτήτων προκύπτει ότι υπάρχουν 275 παρατηρήσεις με τιμή μέχρι και 1 και 429 παρατηρήσεις με τιμή μέχρι και 2. Επομένως η τιμή της $327^{ης}$ παρατήρησης ισούται με 2 ή m= 2. Δηλαδή, οι μισές από τις 653 γυναίκες γέννησαν μέχρι 2 παιδιά και οι υπόλοιπες μισές δύο και περισσότερα.

Από το προηγούμενο παράδειγμα γίνεται φανερό ότιστα διακριτά δεδομένα, την τιμή της διαμέσου μπορεί να έχουν περισσότερες από μια παρατηρήσεις οπότε το πλήθος των παρατηρήσεων που είναι μικρότερες από την διάμεσο και το πλήθος αυτών που είναι μεγαλύτερες να μην είναι ίδιο και μάλιστα μπορεί να διαφέρουν σημαντικά.

◘ **Η κατανομή αφορά παρατηρήσεις ομαδοποιημένες σε μια κατανομή με κ τάξεις:** Για τον εντοπισμό της διαμέσου θα χρησιμοποιήσουμε παρεμβολή ως εξής:

- Η διάμεσος ανήκει στην πρώτη τάξη της οποίας η αθροιστική συχνότητα είναι ίση ή μεγαλύτερη από (n+1)/2.
- Υποθέτοντας ότι οι f_i τιμές της τάξης αυτής κατανέμονται ομοιόμορφα, η τιμή της διαμέσου προσδιορίζεται, κατ'εκτίμηση, από τον τύπο:

$$m = L + \frac{c}{f_i}\left(\frac{n+1}{2} - F_{i-1}\right) \qquad (3.4.1)$$

όπου: L = το κατώτερο όριο του ταξικού διαστήματος που περιέχει τη διάμεσο
c = το εύρος του ταξικού διαστήματος
f_i= η συχνότητα του ταξικού διαστήματος
F_{i-1} = η αθροιστική συχνότητα του προηγούμενου ταξικού διαστήματος

Παράδειγμα 3.4.3 Η κατανομή συχνοτήτων των παρατηρήσεων του βάρους γεννήσεων 60 βρεφών δίνεται στον πίνακα που ακολουθεί.

Τάξεις	f_i	F_i	F_i'
1400–1800	2	2	60
1800–2200	6	8	58
2200–2600	5	13	52
2600–3000	9	22	47
3000–3400	11	33	38
3400–3800	13	46	27
3800–4200	8	54	14
4200–4600	6	60	6
Άθροισμα	60		

Για να προσδιορίσουμε τη Διάμεσο με παρεμβολή προσδιορίζουμε την πρώτη τάξη στην οποία η αθροιστική συχνότητα είναι μεγαλύτερη από 30: είναι η τάξη 3000–3400. Επομένως L= 3000, c= 400, f_i= 11 και F_{i-1} = 22. Αντικαθιστώντας στον τύπο (3.4.1) παίρνουμε:

$$m = 3000 + \frac{400}{11}(30.5 - 22) = 3309.1$$

Δηλαδή τα μισά από τα 60 βρέφη έχουν βάρος μέχρι 3309.1 gr.

Ιδιότητες και Πλεονεκτήματα της Διαμέσου

Η διάμεσος είναι μέτρο θέσης και άρα ισχύει

- $\min_i\{x_i\} \le m \le \max_i\{x_i\},\ i= 1,\ldots,n$
- $m_{a+bx} = a + bm$

Η διάμεσος έχει την εξής επιπλέον ιδιότητα

- Για κάθε σταθερά c>0 ισχύει:

$$\sum|x_i - m| \le \sum|x_i - c|$$

Το σημαντικότερο **πλεονέκτημα** της διαμέσου είναι ότι δεν επηρεάζεται από τις ακραίες ή ασυνήθιστες τιμές. Έτσι π.χ. στις ακόλουθες παρατηρήσεις

2	7	3	8	14

έχουμε m= 7 και $\bar{x} = 6.8$ και τα δύο μέτρα σηματοδοτούν με τρόπο ικανοποιητικό την θέση των δεδομένων. Αν όμως λόγω σφάλματος η τελευταία τιμή αντικατασταθεί από την 140 τότε η διάμεσος εξακολουθεί να είναι ίση με 7 ενώ ο αριθμητικός μέσος γίνεται ίσος με 32 και επομένως δεν μπορεί να αντιπροσωπεύσει τα δεδομένα. Λέμε ότι η διάμεσος είναι **ανθεκτική** (robust) στις ακραίες τιμές.

Τα τελευταία χρόνια, που με τη χρήση Η/Υ γίνεται επεξεργασία μεγάλου όγκου δεδομένων, αυξάνει η πιθανότητα μιας λανθασμένης καταχώρησης και συγχρόνως μειώνεται η ευκολία εντοπισμού της. Αυτό καθιστά την ανθεκτικότητα στις ακραίες τιμές μια πολύ επιθυμητή ιδιότητα. Έτσι, στις ΗΠΑ οι κρατικές υπηρεσίες αναφέρουν το διάμεσο εισόδημα πο συχνά από ότι το μέσο εισόδημα. Αντίστοιχα, έχουν αναπτυχθεί τεχνικές στατιστικής συμπερασματολογίας, οι οποίες βασίζονται στη διάμεσο και είναι μέρος των λεγόμενων **μη παραμετρικών** (non-parametric) μεθόδων όπως ειδικότερα θα δούμε σε επόμενα κεφάλαια.

Η διάμεσος πάντως έχει και ορισμένα **μειονεκτήματα**:

- Οι διάμεσοι δύο μερών δεν μπορούν να συνδυαστούν ώστε να προσδιοριστεί η διάμεσος ολόκληρου του συνόλου των δεδομένων.
- Επηρεάζεται πιο πολύ από τις διακυμάνσεις της δειγματοληψίας από τον μέσο αριθμητικό και άρα είναι λιγότερο ικανοποιητική για εκτιμήσεις.

3.5 Η Επικρατούσα Τιμή

Η **Επικρατούσα** ή **Τυπική τιμή** (mode) ενός συνόλου παρατηρήσεων συμβολίζεται με **τ** και ορίζεται ως η τιμή με τη μεγαλύτερη συχνότητα. Στις ομαδοποιημένες παρατηρήσεις το ταξικό διάστημα με τη μεγαλύτερη συχνότητα που ονομάζεται **Τυπική** ή **Επικρατούσα Τάξη** (modal class), προσδιορίζεται άμεσα. Το πολύγωνο συχνοτήτων έχει κορυφή στο μέσον της τυπικής τάξης και γι' αυτό συνήθως λαμβάνεται ως επικρατούσα τιμή η κεντρική τιμή της τυπικής τάξης. Η τιμή αυτή πάντως θα πρέπει να εννοείται ως το σημείο στο οποίο υπάρχει η μεγαλύτερη συγκέντρωση παρατηρήσεων και όχι ως η πιο συχνά παρατηρούμενη τιμή. Είναι προφανές ότι, για να έχει ουσιαστικό περιεχόμενο η έννοια της τυπικής τάξης, θα πρέπει τα ταξικά διαστήματα να είναι ίσα.

Είναι δυνατόν μια κατανομή να έχει περισσότερες από δύο τυπικές τιμές οπότε αναφέρεται ως **δίτυπη** ή **δίκορφη** (bimodal) σε αντιδιαστολή με την **μονότυπη** ή **μονόκορφη** (unimodal) (βλέπε και *Ερμηνεύοντας το ιστόγραμμα* στο μέρος 2.5). Όταν στις ομαδοποιημένες παρατηρήσεις υπάρχουν περισσότερες από μια κορυφές δοκιμάζουμε άλλη ομαδοποίηση η οποία μπορεί να αλλάξει την εικόνα. Αν δοκιμάζοντας διαφορετικές ομαδοποιήσεις η κατανομή εξακολουθεί να έχει περισσότερες από μία κορυφές τότε έχουμε ισχυρές ενδείξεις ότι τα δεδομένα μας είναι ανομοιογενή ως προς κάποιον παράγοντα και ίσως θα πρέπει να εξετάζονται χωριστά.

3.6 Επιλογή του κατάλληλου μέτρου θέσης

- Όταν τα δεδομένα είναι **ονομαστικά** τότε η επικρατούσα τιμή δηλαδή αυτή με τη μεγαλύτερη συχνότητα είναι το μόνο μέτρο θέσης που μπορούμε να χρησιμοποιήσουμε.
- Όταν τα δεδομένα είναι **διατακτικά** έχει νόημα να προσδιορίσουμε εκτός από την επικρατούσα και την διάμεσο τιμή.
- Όταν τα δεδομένα είναι **μετρήσεις** μπορούμε να χρησιμοποιήσουμε την επικρατούσα τιμή, τη διάμεσο και τον αριθμητικό μέσο.

Το μέτρο που τελικά θα επιλέξουμε εξαρτάται από τη φύση του προβλήματος και τους σκοπούς της έρευνας. Ειδικότερα, αν στα δεδομένα έχουμε λίγες, διακριτές τιμές x_i, $i = 1,\ldots,n$ οι οποίες επαναλαμβάνονται με συχνότητα, αντίστοιχα f_i τότε η επικρατούσα τιμή και η διάμεσος δίνουν καλύτερη πληροφόρηση στον μέσο αναγνώστη από τον αριθμητικό μέσο, ο οποίος μπορεί να

πάρει και μια τιμή η οποία είναι μη παρατηρήσιμη. Έτσι π.χ. αν τα δεδομένα είναι παρατηρήσεις του αριθμού παιδιών ανά οικογένεια ο μέσος αναγνώστης αντιλαμβάνεται καλύτερα τις προτάσεις "η τυπική οικογένεια έχει 2 παιδιά" (δηλαδή τ = 2) και "οι μισές οικογένειες έχουν μέχρι ένα παιδί" (δηλαδή m= 1) από την πρόταση "ο μέσος αριθμός ανά οικογένεια ισούται με 1.55" (δηλαδή $\bar{x} = 1.55$).

Όταν όμως πρόκειται να γίνει περαιτέρω στατιστική ανάλυση, ο αριθμητικός μέσος θα προτιμηθεί στις περισσότερες περιπτώσεις.

Εκτίμηση των παρατηρήσεων με ένα μέτρο θέσης

Σε πολλές πρακτικές εφαρμογές θα χρειαστεί να κάνουμε εκτιμήσεις για παρατηρήσεις, επιπλέον αυτών που διαθέτουμε. Ως κανόνα εκτίμησης συνήθως υιοθετούμε κάποιο μέτρο θέσης και σημαντικό κριτήριο για την επιλογή είναι τοαναμενόμενο σφάλμα που συνεπάγεται το κάθε μέτρο.

Με βάση τις ιδιότητες των μέτρων θέσης που είδαμε λοιπόν, συνοψίζουμε:

- Αν κάθε παρατήρηση εκτιμάται με την επικρατούσα τιμή τότε θα έχουμε τα περισσότερα μηδενικά σφάλματα
- Αν εκτιμάται με τη διάμεσο τότε θα έχουμε ελάχιστο το άθροισμα των απόλυτων τιμών των σφαλμάτων
- Αν χρησιμοποιήσουμε τον αριθμητικό μέσο για να εκτιμήσουμε την κάθε παρατήρηση τότε το άθροισμα των σφαλμάτων θα ισούται με μηδέν. Με μια πιο καθημερινή έκφραση θα λέγαμε ότι στην περίπτωση αυτή, θα έχουμε κατά μέσο όρο μηδενικό σφάλμα. Επίσης στην περίπτωση αυτή θα έχουμε ελάχιστο το άθροισμα των τετραγωνικών σφαλμάτων.

3.7 Μέση σχετική μεταβολή. Ο Γεωμετρικός Μέσος

Όταν καταγράφονται οι τιμές ενός μεγέθους στο χρόνο, τότε τα δεδομένα μας αποτελούν μια **χρονική σειρά** (time series) δηλαδή παρατηρήσεις του μεγέθους αυτού σε ισαπέχοντα[1] χρονικά σημεία. Μια παρατήρηση χρονικής σειράς συμβολίζεται συνήθως με x_t και ο δείκτης t χρησιμοποιείται για να τονιστεί η ενυπάρχουσα διάταξη ως προς το χρόνο.

[1]Οι χρονικές σειρές στις οποίες οι παρατηρήσεις μπορεί να αναφέρονται σε άνισα χρονικά διαστήματα είναι έξω απο τα όρια αυτού του βιβλίου

Η τιμή ενός μεγέθους x_t στην περίοδο t ως ποσοστό της τιμής του στην αμέσως προηγούμενη περίοδο x_{t-1} ονομάζεται **συντελεστής μεταβολής** (coefficient of change) από την t–1 στην περίοδο t και συμβολίζεται με c_t. Δηλαδή έχουμε:

$$c_t = \frac{x_t}{x_{t-1}} \tag{3.7.1}$$

Αν μας ενδιαφέρει ο μέσος συντελεστής μεταβολήςαπό μια αρχική περίοδο έστω 0ώς την περίοδο n, τότε ο κατάλληλος μέσος δεν είναι ο αριθμητικός αλλά ο **Γεωμετρικός μέσος** (geometric mean) των $c_1,...,c_n$ο οποίος συμβολίζεται με G και ορίζεται ως η n–οστή ρίζα του γινομένου τους. Έχουμε δηλαδή:

$$G = \sqrt[n]{c_1...c_n} \tag{3.7.2}$$

Αντικαθιστώντας στη σχέση αυτή την (3.7.1) και με απλοποίηση, προκύπτει οτι

$$G = \sqrt[n]{\frac{x_n}{x_0}} \tag{3.7.3}$$

Πολλές φορές είναι χρήσιμο, η μεταβολή του μεγέθους από την περίοδο t–1 στην περίοδο t (δηλαδή η $x_t - x_{t-1}$) να εκφραστεί ως ποσοστό της τιμής του στην περίοδο t–1 οπότε προκύπτει η **ποσοστιαία μεταβολή** (proportional change) η οποία συνήθως συμβολίζεται με r_t. Έχουμε δηλαδή:

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1$$

$$\Leftrightarrow \frac{x_t}{x_{t-1}} = 1 + r_t$$

$$\Leftrightarrow c_t = 1 + r_t \tag{3.7.4}$$

Παράδειγμα 3.7.1 Στον πίνακα που ακολουθεί δίνεται η μέση κατά κεφαλήν (ακαθάριστη) αμοιβή μιας υποθετικής κατηγορίας εργαζομένων σε ελληνική ΔΕΚΟ, από το 2003–2008. Στην τρίτη και τέταρτη στήλη δίνονται ο συντελεστής μεταβολής και η σχετική μεταβολή αντίστοιχα.

Έτος	Αμοιβή σε ευρώ x_t	Συντελεστής Μεταβολής $\frac{x_t}{x_{t-1}} = 1 + r_t$	Σχετική Μεταβολή $r_t = \frac{x_t - x_{t-1}}{x_{t-1}}$
2003	822.00	--	--
2004	931.60	1.13333	0.13333
2006	1150.80	1.23529	0.23529
2007	1972,80	1.71429	0.71429
2008	2685.20	1.36111	0.36111
2009	3288.00	1.22449	0.22449

Σύμφωνα με τα παραπάνω ο μέσος συντελεστής μεταβολής του μισθού στο χρονικό διάστημα 1979–1984 ισούται με τον γεωμετρικό μέσο των 5 συντελεστών μεταβολής, δηλαδή

$$G = \sqrt[5]{(1.13333)(1.23529)(1.71429)(1.36111)(1.22449)}$$
$$= \sqrt[5]{3.99999} = 1.3195$$

Ο ίδιος συντελεστής υπολογίζεται εναλλακτικά από την (3.7.4) ως

$$G = \sqrt[5]{\frac{3288}{822}} = 1.3195 = 1 + r$$

Επομένως ο μέσος συντελεστής αύξησης του μισθού στο χρονικό διάστημα 1979 – 1984 ήταν 1.3195 και η αντίστοιχη ποσοστιαία αύξηση ίση με 1.3195 – 1 = 0.3195 ή 31.95%.

Το μέσο επιτόκιο

Έστω ότι κεφάλαιο K_0 τοκίζεται για ένα χρόνο με επιτόκιο r_1. Στο τέλος του χρόνου η αξία του θα ισούται με

$$K_1 = K_0 + K_0 r_1 = K_0(1 + r_1)$$

και η σχετική του μεταβολή θα ισούται με

$$c_1 = \frac{K_1}{K_0} = 1 + r_1$$

Αν το κεφάλαιο ανατοκίζεται για n συνολικά χρόνια με μεταβαλόμενο επιτόκιο αντίστοιχα $r_1,..., r_n$ θα ισχύει, γενκά,

$$c_n = \frac{K_n}{K_{n-1}} = 1 + r_n \tag{3.7.5}$$

Για να υπολογίσουμε το μέσο επιτόκιο σε όλο το διάστημα των nπεριόδων αντικαθιστούμε την (3.7.5) στην (3.7.2) για να υπολογίσουμε πρώτα τη μέση σχετική μεταβολή

$$G \equiv 1 + r = \sqrt[n]{(1 + r_1)...(1 + r_n)} \tag{3.7.6}$$

είτε στην (3.7.3) οπότε

$$G = \sqrt[n]{\frac{K_n}{K_0}} \tag{3.7.7}$$

Αφαιρώντας τη μονάδας από τον G, προκύπτει το μέσο επιτόκιο r.

Παράδειγμα 3.7.2 Κεφάλαιο 100 χιλ. ευρώ. τοκίζεται για 8 χρόνια και το επιτόκιο για κάθε ένα απ' αυτά δίνεται στην δεύτερη στήλη του πίνακα που ακολουθεί.

Χρόνος	Επιτόκιο r_t %	Συντελεστής μεταβολής	Αξία στο τέλος του χρόνου (σε χιλ. ευρώ.)
1	2.0	1.020	102.00
2	3.5	1.035	105.57
3	6.0	1.060	111.90
4	5.5	1.055	118.06
5	4.0	1.040	122.77
6	7.0	1.070	131.36
7	7.0	1.070	140.56
8	12.0	1.120	157.43

Το το μέσο επιτόκιο για το χρονικό διάστημα των εν λόγω 8 υπολογίζεται ως ο γεωμετρικός μέσος των 8 συντελεστών μεταβολής, δηλαδή

$$G = \sqrt[8]{(1.020)(1.035)...(1.070)(1.120)} = 1.05837$$

Είτε, ισοδύναμα ως

$$G = \sqrt[8]{\frac{157.43}{100}} = 1.05837$$

Συμπεραίνουμε ότι το μέσο επιτόκιο είναι το

$$r = 1.05837 - 1 = 0.05837 \quad \text{ή} \quad 5.837\%$$

Που σημαίνει οτι αν το αρχικό κεφάλαιο των $K_0 = 100$ χιλ. δρχ. ανατοκίζεται επί 8 έτη με σταθερό επιτόκιο 5.837% τότε στο τέλος του $8^{ου}$ χρόνου θα έχει αξία ίση με:

$$\begin{aligned} K_8 = K_0(1+r)^8 &= 100(1+0.05837)^8 \\ &= 100(1.05837)^8 \\ &= 100(1.5743) \\ &= 157.43 \end{aligned}$$

○○● **Σημείωση** Αν ο υπολογισμός της n–οστής ρίζας μας δυσκολεύει μπορούμε να χρησιμοποιήσουμε λογαρίθμους. Ειδικότερα ο γεωμετρικός μέσος των τιμών $\alpha_1, \alpha_2, ..., \alpha_n$ γράφεται και ως

$$G = (\alpha_1 \cdot \alpha_2 \cdot ... \cdot \alpha_n)^{\frac{1}{n}}$$

Παίρνοντας λογαρίθμους έχουμε:

$$\begin{aligned} \log G &= \frac{1}{n}(\log\alpha_1 + \log\alpha_2 + ... + \log\alpha_n) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\alpha_i \end{aligned}$$

$$\Leftrightarrow G = \text{antilog}\left(\frac{1}{n}\sum_{i=1}^{n}\log\alpha_i\right)$$

Προσοχή στα ποσοστά (Ξανά!)

Από την ανάλυση αυτού του κεφαλαίου προκύπτει ότι:

◘ Όταν θέλουμε να υπολογίσουμε τη μέση σχετική μεταβολή μιας χρονικής σειράς πρέπει να χρησιμοποιήσουμε τον γεωμετρικό μέσο των αντίστοιχων σχετικών μεταβολών. Σημειώνεται πάντως οτι ακόμα και επίσημοι φορείς χρησιμοποιούν τον αριθμητικό μέσο για να υπολογίσουν τη μέση σχετική μεταβολή μισθών, πληθωρισμού, ανεργίας και άλλων μακροοικονομικών μεγεθών. Επειδή ο αριθμητικός μέσος είναι πάντα μεγαλύτερος ή ίσος με τον γεωμετρικό

μέσο (οι δύο μέσοι είναι ίσοι μόνον όταν οι σχετικοί αριθμοί είναι ίσοι) με τον τρόπο αυτό η μέση σχετική μεταβολή υπερεκτιμάται.

◘ Όταν έχουμε ένα σύνολο από διαφορετικά ποσοστά και θέλουμε να υπολογίσουμε τη μέση τιμή τους τότε θα χρησιμοποιήσουμε τον σταθμισμένο μέσο: κάθε ποσοστό θα σταθμίζεται με το αντίστοιχο μέγεθος-βάση.

☼☼☼ **Κουΐζ 1** Έχω τους αριθμούς Α, Β και Γ. Αν $\bar{x}_1 = \frac{A+B}{2}$ και $\bar{x}_2 = \frac{B+\Gamma}{2}$, τότε $\frac{\bar{x}_1 + \Gamma}{2}$ είναι το ίδιο με $\frac{\bar{x}_2 + A}{2}$;

☼☼☼ **Κουΐζ 2** Διάβασα ότι από μια έρευνα προκύπτει οτι ο μέσος άνδρας έχει 6 συντρόφους στη ζωή του ενώ η μέση γυναίκα μόνον δύο. Πώς μπορεί να ισχύει αυτό;

Απάντηση. *(Επιλέγετε μία από τις τρεις ή και τις τρεις)*

α. Μην ξεχνάτε ότι μέση τιμή είναι ο αριθμητικός, η διάμεσος αλλά και η επικρατούσα τιμή. Εδώ ο όρος "μέσος" δηλώνει τον τυπικό, τον πιο συχνά παρατηρούμενο

β. Στην έρευνα συμμετείχαν πολλοί ομοφυλόφιλοι.

γ. Μετά απο όσα ακούσατε για τον τρόπο που απαντούν οι φίλοι σας, ακόμη πιστεύετε στις έρευνες για τη σεξουαλική συμπεριφορά;