

Παρουσίαση

βασισμένη στο βιβλίο

Στατιστική Τόμος Α

ΤΩΝ

Βασίλη Αγγελή & Κατερίνας Δημάκη

Αθήνα, 2011

Να κατασκευαστεί το διάγραμμα σημείων των τελικών βαθμολογιών (με άριστα το 100) ενός τυχαίου δείγματος 150 φοιτητών ενός Πανεπιστημιακού Τμήματος, στη θεματική ενότητα «Ποσοτικές μέθοδοι», κατά το ακαδημαϊκό έτος 2009-10. (Τα δεδομένα παρουσιάστηκαν στη χθεσινή διάλεξη)



Να κατασκευαστεί το διάγραμμα μίσχου φύλλου των τελικών βαθμολογιών (με άριστα το 100) ενός τυχαίου δείγματος 150 φοιτητών ενός Πανεπιστημιακού Τμήματος, στη θεματική ενότητα «Ποσοτικές μέθοδοι», κατά το ακαδημαϊκό έτος 2009-10. (Τα δεδομένα παρουσιάστηκαν στη χθεσινή διάλεξη)

Stem-and-Leaf Display: Βαθμολογία 150 Φοιτητών

Stem-and-leaf of C1 N = 150

Leaf Unit = 1.0

1	1	3
7	2	567779
16	3	134556699
47	4	0001112223334444455556667888999
(42)	5	111111223333344444455555566667777788899999
61	6	00000111111222223455677888999999
29	7	00111123335566789
12	8	0022355689
2	9	14

3. Αριθμητικές μέθοδοι σύνοψης δεδομένων

- Τα αριθμητικά μέτρα που αναφέρονται σε πληθυσμό ονομάζονται παράμετροι
- Τα αντίστοιχα αριθμητικά μέτρα που αναφέρονται σε δείγμα υπολογίζονται ως τιμές στατιστικών συναρτήσεων.
- Τα αριθμητικά μέτρα εντάσσονται σε επτά κατηγορίες:
 - Μέτρα Κεντρικής Τάσης
 - Μέτρα Σχετικής Θέσης
 - Μέτρα Διασποράς
 - Μέτρα Σχετικής Διασποράς
 - Μέτρα Ασυμμετρίας
 - Μέτρα Κύρτωσης
 - Μέτρα Συγκέντρωσης

Μη Ομαδοποιημένα ή
Αταξινόμητα
Δεδομένα

Μέτρα Κεντρικής Τάσης

- ☀ Τα μέτρα κεντρικής τάσης προσδιορίζουν τη θέση του κέντρου της κατανομής.
- ☀ Κάθε τέτοιο μέτρο είναι μια χαρακτηριστική τιμή, υπό την έννοια ότι μπορεί να χρησιμοποιηθεί ως «εκπρόσωπος» καθεμιάς από τις μεμονωμένες τιμές μιας σειράς παρατηρήσεων.
- ☀ Επιπλέον, δίνει μια ένδειξη της τιμής γύρω από την οποία τείνουν να σωρεύονται τα δεδομένα.
- ☀ Υπάρχουν πολλά μέτρα κεντρικής τάσης, καθένα από τα οποία έχει ξεχωριστές ιδιότητες που το χαρακτηρίζουν.

Αριθμητικός Μέσος

x_i	Την τιμή της i μονάδας του πληθυσμού
X_i	Την τιμή της i παρατήρησης του δείγματος
N	Το πλήθος των μονάδων του πληθυσμού
n	Το πλήθος των παρατηρήσεων του δείγματος

Α.Μ. Πληθυσμού

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

Α.Μ Δείγματος

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

Ο αριθμητικός μέσος

- μπορεί να υπολογισθεί για οποιοδήποτε σύνολο δεδομένων.
- επιδέχεται μαθηματικούς χειρισμούς.
- λαμβάνει υπόψη του όλα τα δεδομένα.

ΟΜΩΣ

- επηρεάζεται από τυχόν ακραίες τιμές των δεδομένων

Επιπλέον, κάθε σύνολο δεδομένων έχει έναν και μοναδικό μέσο αριθμητικό.

Παράδειγμα Δίνονται οι τελικές βαθμολογίες (με άριστα το 100) ενός τυχαίου δείγματος 15 φοιτητών ενός Πανεπιστημιακού Τμήματος που εξετάστηκαν στη θεματική ενότητα «Ποσοτικές μέθοδοι» κατά το ακαδημαϊκό έτος 2009-10:

56 67 34 89 67 89 78 57 48 47 89 80 59 89 94.

Να υπολογισθεί η μέση βαθμολογία των φοιτητών αυτών στη συγκεκριμένη ενότητα.

Λύση

Ο αριθμητικός μέσος της βαθμολογίας του τυχαίου δείγματος των 15 φοιτητών ενός Πανεπιστημιακού Τμήματος είναι

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \Rightarrow \bar{x} = \frac{56 + 67 + \dots + 94}{15} = \frac{1043}{15} = 69.53 .$$

Σταθμικός Αριθμητικός Μέσος

- Για τον υπολογισμό του σταθμικού αριθμητικού μέσου οι παρατηρήσεις αθροίζονται αφού πρώτα πολλαπλασιασθούν με κάποιους συντελεστές στάθμισης, οι οποίοι εκφράζουν την σπουδαιότητα κάθε παρατήρησης στον προσδιορισμό της μέσης τιμής τους, και το άθροισμα τους διαιρείται με το άθροισμα των συντελεστών στάθμισης.

Σταθμικός Αριθμητικός Μέσος

$$\bar{X} = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$$

$w_i, i = 1, 2, \dots, n$ είναι οι συντελεστές στάθμισης.

Παράδειγμα Η βαθμολογία ενός φοιτητή στις δύο προόδους και στην τελική εξέταση του μαθήματος της Στατιστικής (με άριστα το 100) είναι 50, 70 και 90 μονάδες αντίστοιχα και επιπλέον η συνεισφορά της κάθε προόδου και της τελικής εξέτασης στον τελικό βαθμό είναι 20% και 60% αντίστοιχα, τότε ο μέσος βαθμός του φοιτητή θα είναι

$$\overline{X}_w = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i} \Rightarrow \bar{x}_w = \frac{0.2 * 50 + 0.2 * 70 + 0.6 * 90}{0.2 + 0.2 + 0.6} = 78.$$

Αντίθετα, αν η συνεισφορά στη διαμόρφωση του τελικού βαθμού ήταν η ίδια και για τις τρεις εξετάσεις, τότε ο μέσος βαθμός του φοιτητή θα ήταν

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n} \Rightarrow \bar{x} = \frac{50 + 70 + 90}{3} = \frac{210}{3} = 70.$$

Διάμεσος

Η **διάμεσος** M , είναι η τιμή που χωρίζει ένα σύνολο δεδομένων περίπου στη μέση, εφόσον τα δεδομένα τοποθετηθούν σε αύξουσα σειρά. Ουσιαστικά, η διάμεσος ενός συνόλου παρατηρήσεων είναι η τιμή για την οποία ισχύει ότι το πολύ το 50% των μετρήσεων είναι μικρότερες και το πολύ το 50% των μετρήσεων είναι μεγαλύτερες από αυτήν.

Προσδιορισμός Της Διαμέσου

- Αν n περιττός,

M : η τιμή της παρατήρησης στη θέση $\frac{n}{2} + \frac{1}{2}$ δηλαδή $M = X_{\left(\frac{n}{2} + \frac{1}{2}\right)}$.

- Αν n άρτιος,

$$M = \frac{1}{2} \left(X_{\left(\frac{n}{2} + 1\right)} + X_{\left(\frac{n}{2}\right)} \right)$$

Παράδειγμα Δίνονται οι τελικές βαθμολογίες (με άριστα το 100) ενός τυχαίου δείγματος 15 φοιτητών ενός Πανεπιστημιακού Τμήματος που εξετάστηκαν στη θεματική ενότητα «Ποσοτικές μέθοδοι» κατά το ακαδημαϊκό έτος 2009-10:

56 67 34 89 67 89 78 57 48 47 89 80 59 89 94.

Να υπολογισθεί η διάμεσος βαθμολογία των φοιτητών αυτών στη συγκεκριμένη ενότητα.

Λύση

Προκειμένου να υπολογίσουμε τη διάμεσο των βαθμολογιών, τοποθετούμε τις παρατηρήσεις κατά αύξουσα τάξη μεγέθους:

34 47 48 56 57 59 67 **67** 78 80 89 89 89 89 94.

Το πλήθος των παρατηρήσεων είναι 15, δηλαδή περιττός αριθμός. Άρα, η διάμεσος των βαθμολογιών θα είναι η τιμή της παρατήρησης που βρίσκεται στη θέση $\frac{n+1}{2} = \frac{15+1}{2} = 8$,

δηλαδή $M = X_{(8)} = 67$.

Παράδειγμα Δίνονται οι τελικές βαθμολογίες (με άριστα το 100) ενός τυχαίου δείγματος 14 φοιτητών ενός Πανεπιστημιακού Τμήματος που εξετάστηκαν στη θεματική ενότητα «Ποσοτικές μέθοδοι» κατά το ακαδημαϊκό έτος 2009-10:

56 67 89 67 89 78 57 48 47 89 80 59 89 94.

Να υπολογισθεί η διάμεσος βαθμολογία των φοιτητών αυτών στη συγκεκριμένη ενότητα.

Λύση

Προκειμένου να υπολογίσουμε τη διάμεσο των βαθμολογιών, τοποθετούμε τις παρατηρήσεις κατά αύξουσα τάξη μεγέθους:

47 48 56 57 59 67 **67 78** 80 89 89 89 89 94.

Το πλήθος των παρατηρήσεων είναι 14, δηλαδή άρτιος αριθμός. Επομένως η διάμεσος των παρατηρήσεων θα είναι ανάμεσα στις θέσεις $\frac{n}{2} = \frac{14}{2} = 7$ και $\frac{n}{2} + 1 = 7 + 1 = 8$, δηλαδή ο μέσος της 7ης και της 8ης παρατήρησης. Συνεπώς, η διάμεσος βαθμολογία θα είναι

$$M = \frac{1}{2} \left(X_{\left(\frac{n}{2}+1\right)} + X_{\left(\frac{n}{2}\right)} \right) = \frac{1}{2} \left(X_{(8)} + X_{(7)} \right) = \frac{1}{2} (78 + 67) = \frac{1}{2} * 145 = 72.5$$

Επικρατούσα Τιμή

Ως **επικρατούσα τιμή**, T_0 , ενός συνόλου δεδομένων χαρακτηρίζουμε εκείνη την παρατήρηση με τη μεγαλύτερη συχνότητα εμφάνισης.

Παράδειγμα Δίνονται οι τελικές βαθμολογίες (με άριστα το 100) ενός τυχαίου δείγματος 15 φοιτητών ενός Πανεπιστημιακού Τμήματος που εξετάστηκαν στη θεματική ενότητα «Ποσοτικές μέθοδοι» κατά το ακαδημαϊκό έτος 2009-10:

56 67 34 89 67 89 78 57 48 47 89 80 59 89 94.

Να υπολογισθεί η επικρατούσα τιμή της βαθμολογίας των φοιτητών αυτών.

Λύση

Η επικρατούσα τιμή είναι η βαθμολογία 89 που εμφανίζεται 4 φορές, δηλαδή $T_0 = 89$.

Όταν τα δεδομένα έχουν μία επικρατούσα τιμή, η κατανομή τους ονομάζεται **μονοκόρυφη**. Στην περίπτωση που υπάρχουν δύο ή περισσότερες τιμές με την ίδια συχνότητα εμφάνισης, τότε λέμε ότι τα δεδομένα έχουν δύο ή περισσότερες επικρατούσες τιμές. Η κατανομή τους τότε ονομάζεται **δικόρυφη** ή **πολυκόρυφη** αντίστοιχα.

Πρέπει να σημειωθεί ότι η επικρατούσα τιμή χρησιμοποιείται συνήθως στις περιπτώσεις ποιοτικών δεδομένων.

Μέτρα Σχετικής Θέσης

Τεταρτημόρια

Σε πολλές περιπτώσεις τα μέτρα κεντρικής τάσης δεν εξασφαλίζουν ικανοποιητική περιγραφή των δεδομένων. Στις περιπτώσεις αυτές προσδιορίζουμε κάποια άλλα μέτρα, που αναφέρονται στη σχετική θέση των δεδομένων και ονομάζονται ποσοστιαία σημεία. Τα πιο συχνά χρησιμοποιούμενα ποσοστιαία σημεία είναι τα **τεταρτημόρια**, (Q_1, Q_2, Q_3) , τα οποία χωρίζουν ένα σύνολο παρατηρήσεων σε τέταρτα. Για παράδειγμα το πρώτο τεταρτημόριο, Q_1 , είναι η τιμή η οποία χωρίζει το σύνολο των παρατηρήσεων σε δύο μέρη, έτσι ώστε το πολύ 25% από αυτές να είναι μικρότερες και το πολύ 75% μεγαλύτερες από τη συγκεκριμένη τιμή. Όπως και στην περίπτωση της διαμέσου, ο προσδιορισμός των τεταρτημορίων προϋποθέτει διάταξη των παρατηρήσεων κατά αύξουσα τάξη μεγέθους και στη συνέχεια εντοπισμό της θέσης τους.

Προσδιορισμός Τεταρτημορίων

Το $i = 1, 2, 3$ τεταρτημόριο (Q_i) βρίσκεται στην $\left[\frac{i(n+1)}{4} \right]$ θέση.

Η τιμή του $i = 1, 2, 3$ τεταρτημορίου (Q_i) είναι

$$Q_i = X_{(A_Q)} + \Delta_Q \left[X_{(A_Q+1)} - X_{(A_Q)} \right]$$

όπου $A_Q =$ το ακέραιο μέρος του πηλίκου $\left[\frac{i(n+1)}{4} \right]$

και $\Delta_Q =$ το δεκαδικό μέρος του πηλίκου $\left[\frac{i(n+1)}{4} \right]$.

Παράδειγμα Να υπολογισθούν το πρώτο και το τρίτο τεταρτημόριο των βαθμολογιών των 15 φοιτητών ενός Πανεπιστημιακού Τμήματος (*προηγούμενο παράδειγμα*)

Λύση

Προκειμένου να υπολογίσουμε τα τεταρτημόρια των βαθμολογιών, τοποθετούμε τις παρατηρήσεις σε αύξουσα τάξη μεγέθους:

34 47 48 56 57 59 67 67 78 80 89 89 89 89 94.

Το πρώτο τεταρτημόριο είναι η παρατήρηση που βρίσκεται στη θέση $\frac{(n+1)}{4} = \frac{16}{4} = 4$.

Στη συγκεκριμένη περίπτωση $A_Q = 4$ και $\Delta_Q = 0$. Άρα, η τιμή του πρώτου τεταρτημορίου είναι $Q_1 = X_{(4)} + 0 * (X_{(5)} - X_{(4)}) = X_{(4)} = 56$.

Το τρίτο τεταρτημόριο είναι η παρατήρηση που βρίσκεται στη θέση $\frac{3(n+1)}{4} = \frac{3*16}{4} = 12$.

Στη συγκεκριμένη περίπτωση $A_Q = 12$ και $\Delta_Q = 0$. Άρα, η τιμή του τρίτου τεταρτημορίου είναι $Q_3 = X_{(12)} + 0 * (X_{(13)} - X_{(12)}) = X_{(12)} = 89$.

Μέτρα Διασποράς

- Τα μέτρα κεντρικής τάσης δίνουν μια ένδειξη της τιμής γύρω από την οποία τείνουν να συσσωρεύονται τα δεδομένα.
- Είναι απαραίτητο να γνωρίζουμε και το πώς κατανέμονται οι παρατηρήσεις γύρω από ένα μέτρο κεντρικής τάσης.
- Την πληροφορία αυτή μας την παρέχουν τα μέτρα διασποράς

Εύρος

Το **εύρος**, R , ενός συνόλου παρατηρήσεων ορίζεται ως η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής του συνόλου των παρατηρήσεων.

$$R = X_{\max} - X_{\min}$$

όπου $X_{\max}$ η μέγιστη τιμή των παρατηρήσεων

και $X_{\min}$ η ελάχιστη τιμή των παρατηρήσεων.

Παράδειγμα Να υπολογισθεί το εύρος των βαθμολογιών των 15 φοιτητών ενός Πανεπιστημιακού Τμήματος (*προηγούμενο παράδειγμα*)

Λύση

Με βάση τον ορισμό προκύπτει ότι $R = X_{\max} - X_{\min} = 94 - 34 = 60$.

Ενδοτεταρτημοριακό Εύρος

Το **ενδοτεταρτημοριακό εύρος**, IR , ενός συνόλου παρατηρήσεων ορίζεται ως

$$IR = Q_3 - Q_1$$

όπου Q_1 , Q_3 το πρώτο και το τρίτο τεταρτημόριο αντίστοιχα και περιλαμβάνει το 50% των παρατηρήσεων που βρίσκονται γύρω από τη διάμεσο.

Πολλές φορές αντί του ενδοτεταρτημοριακού εύρους χρησιμοποιείται η **τεταρτημοριακή απόκλιση** η οποία ορίζεται ως εξής:

$$Q = \frac{Q_3 - Q_1}{2}.$$

Παράδειγμα Να υπολογισθεί το ενδοτεταρτημοριακό εύρος των βαθμολογιών των 15 φοιτητών ενός Πανεπιστημιακού Τμήματος (*προηγούμενο παράδειγμα*)

Λύση

Με βάση τον ορισμό και χρησιμοποιώντας τα αποτελέσματα του Παραδείγματος 3.4 προκύπτει ότι $IR = Q_3 - Q_1 = 89 - 56 = 33$.

Διακύμανση

Η **διακύμανση**, S_{op}^2 , ενός συνόλου παρατηρήσεων είναι ένα μέτρο διασποράς το οποίο βασίζεται στην έννοια της απόκλισης καθεμιάς από τις παρατηρήσεις αυτές από τον αριθμητικό τους μέσο.

Η διακύμανση ενός πληθυσμού N τιμών $x_1, x_2, \dots, x_N$, με μέσο αριθμητικό μ , ορίζεται ως η μέση τιμή των τετραγώνων των αποκλίσεων των N τιμών από τον μέσο μ του πληθυσμού. Δηλαδή

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}.$$

Αντίστοιχα, στην περίπτωση δείγματος n παρατηρήσεων με δειγματικό μέσο $\bar{X}$, η δειγματική διακύμανση ορίζεται ως η μέση τιμή των τετραγώνων των αποκλίσεων των n τιμών του δείγματος από τον αντίστοιχο δειγματικό μέσο και συμβολίζεται με S_{op}^2 .

Δηλαδή

$$S_{op}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}.$$

Στις εφαρμογές, **οι οποίες δεν έχουν μόνο περιγραφικό χαρακτήρα**, υπολογίζουμε, συνήθως, τη δειγματική διακύμανση με βάση τον τύπο

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

Ο υπολογισμός της δειγματικής διακύμανσης μπορεί να γίνει και με άλλους εναλλακτικούς και ισοδύναμους τύπους, οι οποίοι σε πολλές περιπτώσεις είναι πιο απλοί στη χρήση τους. Συγκεκριμένα

$$S^2 = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1} = \frac{\sum_{i=1}^n X_i^2}{n-1} - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n(n-1)}.$$

Παράδειγμα Δίνονται οι τελικές βαθμολογίες (με άριστα το 100) ενός τυχαίου δείγματος 15 φοιτητών ενός Πανεπιστημιακού Τμήματος που εξετάστηκαν στη θεματική ενότητα «Ποσοτικές μέθοδοι» κατά το ακαδημαϊκό έτος 2009-10:

56 67 34 89 67 89 78 57 48 47 89 80 59 89 94.

Να υπολογισθεί η διακύμανση της βαθμολογίας των φοιτητών αυτών.

Λύση

Αντικαθιστώντας στον τύπο $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ τις τιμές των δεδομένων και την τιμή του $\bar{X}$, όπως έχει ήδη υπολογιστεί, προκύπτει ότι

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{4993.7}{14} = 356.69.$$

Τυπική Απόκλιση

Η διακύμανση είναι αναμφισβήτητα ένα πολύ χρήσιμο μέτρο διασποράς, αλλά δύσκολο να ερμηνευτεί, δεδομένου ότι εκφράζεται στο τετράγωνο των μονάδων των παρατηρήσεων. Για να ξεπεράσουμε το πρόβλημα αυτό, εισάγουμε την **τυπική απόκλιση**, S_{op} ή S , η οποία ορίζεται ως η θετική τετραγωνική ρίζα της διακύμανσης.

Δηλαδή

$$S_{op} = +\sqrt{S_{op}^2} \quad \text{ή} \quad S = +\sqrt{S^2}$$

και εκφράζεται στις μονάδες των παρατηρήσεων.

Παράδειγμα Να υπολογισθεί η τυπική απόκλιση των βαθμολογιών των 15 φοιτητών ενός Πανεπιστημιακού Τμήματος (*προηγούμενο παράδειγμα*)

Λύση

Με βάση τον ορισμό της τυπικής απόκλισης και χρησιμοποιώντας την τιμή της διακύμανσης που έχει ήδη υπολογιστεί, προκύπτει ότι

$$S = +\sqrt{S^2} = \sqrt{356.69} = 18.89 .$$

Μέτρα Σχετικής Διασποράς

- Όλα τα μέτρα της διασποράς εκφράζονται στην **μονάδα μέτρησης του χαρακτηριστικού** που περιγράφουν.
- Δεν μπορούν λοιπόν να χρησιμοποιηθούν για **συγκρίσεις** ομάδων τιμών οι οποίες είτε εκφράζονται σε διαφορετικές μονάδες είτε εκφράζονται μεν στην ίδια μονάδα αλλά έχουν διαφορετικούς αριθμητικούς μέσους.
- Σε περιπτώσεις του τύπου αυτού χρησιμοποιούμε τα μέτρα σχετικής διασποράς.

Συντελεστής Μεταβλητότητας

- Ο συντελεστής μεταβλητότητας συμβολίζεται με CV και εκφράζεται ως ποσοστό της τυπικής απόκλισης προς τον αντίστοιχο αριθμητικό μέσο:

για το δείγμα $CV = \frac{S_{op}}{\bar{X}} \times 100$ $CV = \frac{S}{\bar{X}} \times 100$

για τον πληθυσμό $CV = \frac{\sigma}{\mu} \times 100$

- *Ο συντελεστής μεταβλητότητας είναι ανεξάρτητος από την μονάδα μέτρησης του χαρακτηριστικού και εκφράζει την τυπική απόκλιση ως ποσοστό του αριθμητικού μέσου.*

Παράδειγμα . Η επίδοση ενός τυχαίου δείγματος 100 φοιτητών του Τμήματος Διοίκησης Επιχειρήσεων ενός Πανεπιστημίου στη Στατιστική και τα Μαθηματικά βαθμολογείται στην κλίμακα 0-100. Κατά την τελική εξέταση η μέση βαθμολογία τους στη Στατιστική ήταν 70 μονάδες με τυπική απόκλιση 5 μονάδες, ενώ στα Μαθηματικά ήταν 50 μονάδες με τυπική απόκλιση 4 μονάδες. Να συγκριθεί η μεταβλητότητα των βαθμολογιών στα Μαθηματικά και στη Στατιστική των 100 φοιτητών του Τμήματος Διοίκησης Επιχειρήσεων του Πανεπιστημίου.

Λύση

Σύμφωνα με τα δεδομένα του Παραδείγματος, η βαθμολογία των φοιτητών στα δύο μαθήματα έδωσε τα εξής αποτελέσματα:

$$\text{Στατιστική: } \bar{X}_S = 70 \text{ και } S_S = 5 \quad \text{Μαθηματικά: } \bar{X}_M = 50 \text{ και } S_M = 4$$

Επειδή για τις τυπικές αποκλίσεις των βαθμολογιών ισχύει ότι $S_S = 5 > 4 = S_M$, μπορεί κάποιος να οδηγηθεί στο συμπέρασμα ότι η διασπορά στις επιδόσεις των φοιτητών στη Στατιστική είναι μεγαλύτερη από εκείνη στις επιδόσεις των Μαθηματικών. Το συμπέρασμα αυτό, όμως, είναι λανθασμένο, δεδομένου ότι τα δύο σύνολα παρατηρήσεων έχουν διαφορετικούς μέσους και, κατά συνέπεια, **δεν επιτρέπεται** η σύγκριση της μεταβλητότητας με την τυπική απόκλιση. Το πρόβλημα της σύγκρισης που ανακύπτει στην περίπτωση αυτή αντιμετωπίζεται με τη βοήθεια του συντελεστή μεταβλητότητας, ο οποίος υπολογίζει τη μεταβλητότητα των τιμών της κατανομής σε σχετικούς όρους.

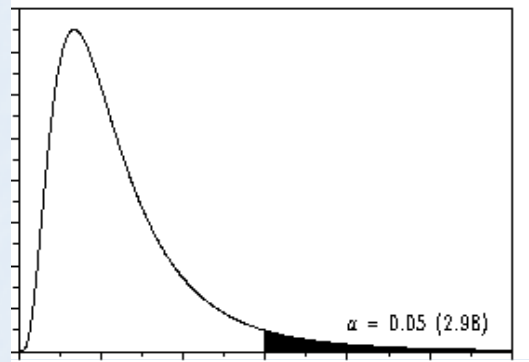
Υπολογίζοντας κατά τα γνωστά το συντελεστή μεταβλητότητας για τις βαθμολογίες στα Μαθηματικά και στη Στατιστική έχουμε:

$$\begin{array}{ll} \text{Μαθηματικά: } C V_M = \frac{S_M}{\bar{X}_M} = \frac{4}{50} = 0.08 & \text{Στατιστική: } C V_S = \frac{S_S}{\bar{X}_S} = \frac{5}{70} = 0.07 \\ C V_M = 8\% & C V_S = 7\% \end{array}$$

Συμπερασματικά λοιπόν, μπορούμε να πούμε ότι η σχετική ανομοιογένεια (μεταβλητότητα) των επιδόσεων των φοιτητών στη Στατιστική (7%) είναι μικρότερη από εκείνη των επιδόσεών τους στα Μαθηματικά (8%).

Μέτρα Ασυμμετρίας

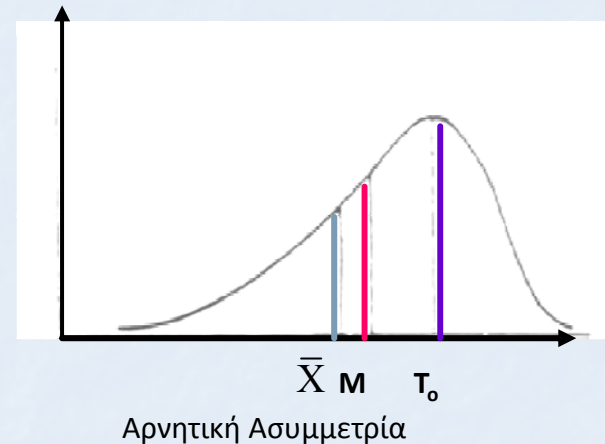
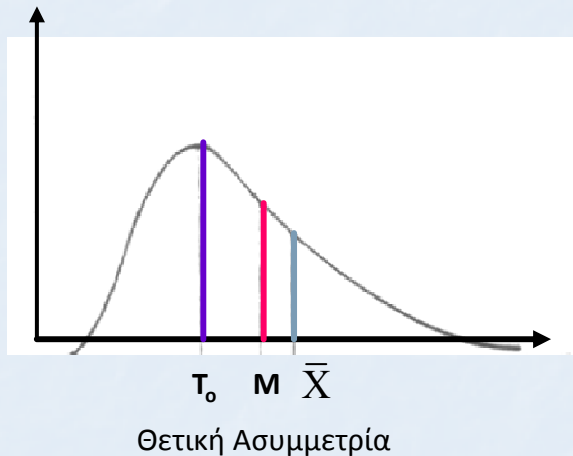
- Τα μέτρα κεντρικής τάσης και διασποράς μιας κατανομής δίνουν μία εικόνα της μορφής της.
- Χρειάζεται όμως να προσδιορίσουμε πόσο και προς ποια κατεύθυνση αποκλίνει η κατανομή από την πλήρως συμμετρική κατανομή.
- Η ασυμμετρία μπορεί να είναι θετική ή αρνητική.
 - Θετικά ασυμμετρική είναι μια κατανομή όταν παρουσιάζει εξόγκωση προς τα αριστερά και η μεγάλη συγκέντρωση των παρατηρήσεων βρίσκεται στις μικρές τιμές της μεταβλητής.



- Αντίθετα, αρνητικά ασυμμετρική είναι μία κατανομή όταν παρουσιάζει εξόγκωση προς τα δεξιά και η μεγάλη συγκέντρωση των παρατηρήσεων βρίσκεται στις μεγάλες τιμές της μεταβλητής

Μέτρα Ασυμμετρίας που βασίζονται στη Μέση Τιμή

- Το μέτρο αυτό βασίζεται στην σχέση που υπάρχει μεταξύ του αριθμητικού μέσου ($\bar{X}$), της διαμέσου (M), και της επικρατούσας τιμής (T_o).
- Στην περίπτωση μιας **μονοκόρυφης συμμετρικής** κατανομής τα τρία αυτά μέτρα κεντρικής τάσης συμπίπτουν.
- Στην περίπτωση όμως ασυμμετρίας ο μέσος απομακρύνεται από την επικρατούσα τιμή και τείνει προς την κατεύθυνση της ασυμμετρίας ενώ η διάμεσος παραμένει πάντα ανάμεσα τους



Κατά συνέπεια η απόσταση μεταξύ μέσου και επικρατούσας τιμής και το πρόσημο της, μπορεί να χρησιμοποιηθεί ως μέτρο ασυμμετρίας.

Μέτρα Ασυμμετρίας που βασίζονται στη Μέση Τιμή

Ως μέτρο ασυμμετρίας προτάθηκε από τον K. Pearson ο **συντελεστής ασυμμετρίας** S_p που βασίζεται στη σχέση η οποία υπάρχει μεταξύ του αριθμητικού μέσου ($\bar{X}$) και της επικρατούσας τιμής (T_0). Συγκεκριμένα

$$S_p = \frac{\bar{X} - T_0}{S}.$$

Η διαίρεση του $(\bar{X} - T_0)$ με την τυπική απόκλιση καθιστά το μέτρο αυτό ανεξάρτητο από τη μονάδα μέτρησης του χαρακτηριστικού και επιτρέπει τη χρήση του για σύγκριση κατανομών που είτε αναφέρονται σε χαρακτηριστικά με διαφορετικές μονάδες μέτρησης, είτε έχουν διαφορετική διασπορά.

Ισχύει:

$$S_p \begin{cases} > 0 & \text{Θετική ασυμμετρία} \\ = 0 & \text{Συμμετρία} \\ < 0 & \text{Αρνητική ασυμμετρία} \end{cases}$$

Να επισημάνουμε ότι το μέτρο μπορεί να εφαρμοστεί μόνο στην περίπτωση μονοκόρυφων κατανομών.

Παράδειγμα Να υπολογιστεί ο συντελεστής ασυμμετρίας S_p των βαθμολογιών των 15 φοιτητών του γνωστού Πανεπιστημιακού Τμήματος.

Λύση

Με βάση τον ορισμό και τα μέτρα που έχουν ήδη υπολογιστεί για τα δεδομένα του Παραδείγματος έχουμε :

$$S_p = \frac{\bar{X} - T_0}{S} = \frac{69.53 - 89}{18.89} = -1.03.$$

Άρα, η κατανομή της βαθμολογίας παρουσιάζει αρνητική ασυμμετρία.

Μέτρα Ασυμμετρίας που βασίζονται στη διάμεσο και τα τεταρτημόρια

- Το μέτρο αυτό βασίζεται στη σχέση που υπάρχει μεταξύ του πρώτου τεταρτημορίου (Q_1), του τρίτου τεταρτημορίου (Q_3) και της Διαμέσου (M).
- Σε μία πλήρως **συμμετρική κατανομή** η απόσταση του 3ου τεταρτημορίου από τη διάμεσο είναι όση η απόσταση της διαμέσου από το 1ο τεταρτημόριο
- Το μέτρο αυτό παίρνει τιμές στο διάστημα $[-1, +1]$

$$S_B = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1}$$

$$S_B = \begin{cases} 0 & \text{στην περίπτωση πλήρως συμμετρικής κατανομής} \\ > 0 & \text{στην περίπτωση θετικής ασυμμετρίας} \\ < 0 & \text{στην περίπτωση αρνητικής ασυμμετρίας} \end{cases}$$

Ροπή Κατανομής

- Ως **ροπή** (μ_k) k τάξης μιας σειράς παρατηρήσεων γύρω από το μέσο τους ορίζεται η συνάρτηση:

$$\mu_k = \frac{\sum_{i=1}^n (X_i - \bar{X})^k}{n}$$

- Για $k=2$
- $$\mu_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} = S_{op}^2$$

Μέτρα Ασυμμετρίας που βασίζονται στις ροπές μιας κατανομής

- Ως μέτρο ασυμμετρίας της κατανομής η παρατηρήσεων ορίζεται ο συντελεστής

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\left[\sum_{i=1}^n (X_i - \bar{X}) \right]^2}{(s^2)^3}$$

- Το μέτρο αυτό είναι γνωστό ως συντελεστής ασυμμετρίας β_1 του Pearson. Αξίζει να παρατηρήσουμε ότι το είδος της ασυμμετρίας με τη μέθοδο αυτή προσδιορίζεται από το πρόσημο της ροπής μ_3 .

$$\mu_3 = \begin{cases} 0 & \text{στην περίπτωση πλήρως συμμετρικής κατανομής} \\ > 0 & \text{στην περίπτωση θετικά συμμετρικής κατανομής} \\ < 0 & \text{στην περίπτωση αρνητικά συμμετρικής κατανομής} \end{cases}$$

- Πολλές φορές χρησιμοποιούμε αντί του β_1 ως συντελεστή ασυμμετρίας το β_3 το οποίο ορίζεται από τον τύπο

- $\beta_3 = \frac{\mu_3}{s^3}$

Ο συντελεστής β_3 κυμαίνεται μεταξύ περίπου -1 και +1 και δίνει πληροφορίες τόσο για το είδος όσο και το μέτρο της ασυμμετρίας.

$\beta_3 > 0$ θετική ή δεξιά ασυμμετρία

$\beta_3 < 0$ αρνητική ή αριστερή ασυμμετρία

Παράδειγμα Να υπολογιστεί ο συντελεστής ασυμμετρίας β_3 των βαθμολογιών των 15 φοιτητών του γνωστού Πανεπιστημιακού Τμήματος.

Λύση

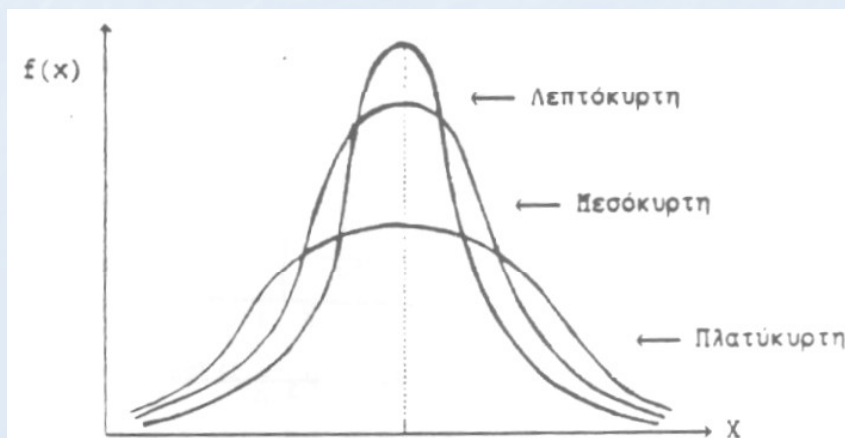
Με βάση τον ορισμό και τα μέτρα που έχουν ήδη υπολογιστεί, για τα δεδομένα του Παραδείγματος έχουμε:

$$\beta_3 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{S^3} = \frac{-25032.25}{(18.89)^3} = \frac{-1668.82}{6740.56} = -0.25.$$

Άρα, η κατανομή της βαθμολογίας παρουσιάζει ελαφρά αρνητική ασυμμετρία.

Μέτρα Κύρτωσης

- Δύο μονοκόρυφες κατανομές μπορεί να έχουν τον ίδιο αριθμητικό μέσο, την ίδια τυπική απόκλιση να είναι ασυμμετρικές αλλά παρόλα αυτά να διαφέρουν ως προς την *κύρτωση* στην περιοχή του αριθμητικού μέσου.
- ✓ Το χαρακτηριστικό αυτό έχει τοπικό χαρακτήρα. Δεν αναφέρεται στην κύρτωση ολόκληρης της κατανομής, η οποία εκφράζεται από τα μέτρα διασποράς, αλλά στην συγκέντρωση των τιμών και στην αιχμηρότητα της κατανομής στην περιοχή του αριθμητικού μέσου.
- Με βάση το χαρακτηριστικό αυτό μια κατανομή μπορεί να θεωρηθεί:
 - λεπτόκυρτη
 - πλατύκυρτη
 - μεσόκυρτηανάλογα με το αν παρουσιάζει μεγάλη, μικρή ή φυσιολογική συγκέντρωση τιμών (δηλαδή αιχμηρότητα) στην περιοχή του αριθμητικού μέσου.



Μέτρο κύρτωσης που βασίζονται στις ροπές

- Το μέτρο αυτό είναι συνάρτηση της 4^{ης} ροπής γύρω από το μέσο όρο

$$\mu_4 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{n}$$

- Μέτρο κύρτωσης της κατανομής η παρατηρήσεων κατά Pearson ορίζεται

$$\beta_4 = \frac{\mu_4}{s^4} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^4}{s^4}$$

$$\beta_4 = \begin{cases} = 3 & \text{στην περίπτωση μεσόκυρτης κατανομής} \\ > 3 & \text{στην περίπτωση λεπτόκυρτης κατανομής} \\ < 3 & \text{στην περίπτωση πλατύκυρτης κατανομής} \end{cases}$$

Παράδειγμα Να υπολογιστεί ο συντελεστής κύρτωσης β_4 των βαθμολογιών των 15 φοιτητών του γνωστού Πανεπιστημιακού Τμήματος.

Λύση

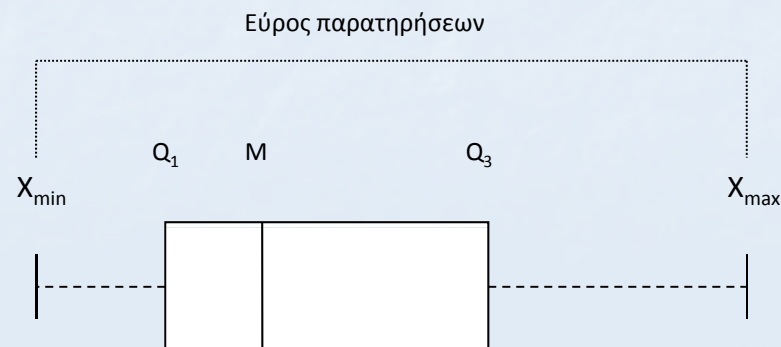
Με βάση τον ορισμό και τα μέτρα που έχουν ήδη υπολογιστεί για τα δεδομένα του Παραδείγματος, έχουμε:

$$\beta_4 = \frac{\frac{\sum_{i=1}^n (X_i - \bar{X})^4}{n}}{S^4} = \frac{\frac{3087183}{15}}{(18.898)^4} = \frac{205812.2}{127329.15} = 1.62 .$$

Άρα, η κατανομή της βαθμολογίας είναι πλατύκυρτη.

Διαγράμματα Πλαισίου Απολήξεων

- Απεικονίζει τις τιμές της διαμέσου και των τεταρτημορίων (μέτρα που δεν επηρεάζονται από ακραίες τιμές των δεδομένων).
- Συνδυάζει αυτά τα μέτρα με τις πληροφορίες που περιέχονται στις ακραίες τιμές των δεδομένων δίνοντας έτσι μια πληρέστερη εικόνα της κατανομής τους.
- Το Διάγραμμα Πλαισίου - Απολήξεων χρησιμοποιεί πέντε χαρακτηριστικές τιμές των δεδομένων:
 - X_{min} : Ελάχιστη τιμή των δεδομένων
 - Q_1 : Πρώτο τεταρτημόριο
 - M : Διάμεσος
 - Q_3 : Τρίτο τεταρτημόριο
 - X_{max} : Μέγιστη τιμή των δεδομένων



Μέτρα Συγκέντρωσης

Η έννοια της συγκέντρωσης εστιάζεται
στον προσδιορισμό του βαθμού απόκλισης
της όλης κατανομής
από την κατάσταση ισοκατανομής.

Συντελεστής Gini

- Ο συντελεστής Gini συμβολίζεται με g και υπολογίζεται

$$g = \frac{d}{2\bar{X}}$$

$\bar{X}$ ο αριθμητικός μέσος της κατανομής
 d η μέση διαφορά Gini

- Ο συντελεστής Gini παίρνει τιμές στο διάστημα $[0, 1]$
 - Το 0 αντιστοιχεί στην περίπτωση της τέλει ισοκατανομής
 - το 1 στην περίπτωση της μέγιστης δυνατής ανισοκατανομής
- Η μέση διαφορά Gini d , που χρησιμοποιείται στον τύπο για τον υπολογισμό του g είναι ο αριθμητικός μέσος των απολύτων τιμών των διαφορών (d_{ij}) τις οποίες μπορούμε να σχηματίσουμε συνδυάζοντας ανά δύο με όλους τους δυνατούς τρόπους τις τιμές μιας μεταβλητής.
 - Αν οι δυνατές τιμές μιας μεταβλητής είναι n οι δυνατές διαφορές είναι n^2 .

$$d = \frac{\sum_i \sum_j |d_{ij}|}{n^2}$$

Ομαδοποιημένα ή
Ταξινομημένα
Δεδομένα

Παράδειγμα

Ο Πίνακας που ακολουθεί περιέχει τις τελικές βαθμολογίες (με άριστα το 100) ενός τυχαίου δείγματος 150 φοιτητών ενός Πανεπιστημιακού Τμήματος, στη θεματική ενότητα «Ποσοτικές μέθοδοι», κατά το ακαδημαϊκό έτος 2009-10.

13	25	26	27	27	27	29	31	33	34
35	35	36	36	39	39	40	40	40	41
41	41	42	42	42	43	43	43	44	44
44	44	44	45	45	45	45	46	46	46
47	48	48	48	49	49	49	51	51	51
51	51	51	52	52	53	53	53	53	53
54	54	54	54	54	54	55	55	55	55
55	55	56	56	56	56	57	57	57	57
57	58	58	58	59	59	59	59	59	60
60	60	60	60	61	61	61	61	61	61
62	62	62	62	62	63	64	65	65	66
67	67	68	68	68	69	69	69	69	69
69	70	70	71	71	71	71	72	73	73
73	75	75	76	76	77	78	79	80	80
82	82	83	85	85	86	88	89	91	94

Αριθμητικός Μέσος

Δειγματικός Αριθμητικός μέσος $\bar{X} \cong \frac{\sum_{i=1}^k f_i m_i}{n}$

- k το πλήθος των τάξεων στις οποίες έχουμε ταξινομήσει τις παρατηρήσεις
- f_i η συχνότητα της i τάξης.
- m_i ο κεντρικός όρος της i τάξης .
- n το πλήθος των παρατηρήσεων του δείγματος.

Αριθμητικός μέσος πληθυσμού $\mu = \frac{\sum_{i=1}^k f_i m_i}{N}$

N το πλήθος των μονάδων του πληθυσμού.

Παραδοχή: ο κεντρικός όρος κάθε τάξης προσεγγίζει ικανοποιητικά τον αριθμητικό μέσο των μετρήσεων που ανήκουν στην τάξη αυτή.

Μια τέτοια υπόθεση ισχύει στη περίπτωση που οι μετρήσεις σε μια τάξη κατανέμονται συμμετρικά γύρω από τον κεντρικό της όρο.

Τότε το άθροισμα των μετρήσεων στην τάξη i προσεγγίζεται από το γινόμενο $f_i m_i$.

Παράδειγμα Να υπολογιστεί ο αριθμητικός μέσος των βαθμολογιών των 150 φοιτητών, έτσι όπως ταξινομήθηκαν σε κατανομή συχνοτήτων στον Πίνακα:

Κατανομή συχνοτήτων της τελικής βαθμολογίας 150 φοιτητών ενός Πανεπιστημιακού Τμήματος

ΤΑΞΕΙΣ	ΚΕΝΤΡΙΚΟΣ ΟΡΟΣ m_i	ΣΥΧΝΟΤΗΤΑ f_i
10-20	15	1
20-30	25	6
30-40	35	9
40-50	45	31
50-60	55	42
60-70	65	32
70-80	75	17
80-90	85	10
90-100	95	2
		150

Λύση

Προκειμένου να υπολογίσουμε τον δειγματικό αριθμητικό μέσο για τα δεδομένα του Παραδείγματος 2.9, δημιουργούμε τον παρακάτω Πίνακα και στη συνέχεια εφαρμόζουμε

τον τύπο $\bar{X} \cong \frac{\sum_{i=1}^k f_i m_i}{n}$.

ΤΑΞΕΙΣ	ΚΕΝΤΡΙΚΟΣ ΟΡΟΣ m_i	ΣΥΧΝΟΤΗΤΑ f_i	$f_i m_i$
10-20	15	1	15
20-30	25	6	150
30-40	35	9	315
40-50	45	31	1395
50-60	55	42	2310
60-70	65	32	2080
70-80	75	17	1275
80-90	85	10	850
90-100	95	2	190
ΣΥΝΟΛΑ		150	8580

Το αποτέλεσμα που προκύπτει είναι $\bar{X} \cong \frac{\sum_{i=1}^k f_i m_i}{n} \Rightarrow \bar{x} = \frac{8580}{150} = 57.2$.

Διακύμανση

Πληθυσμός

$$\sigma^2 = \frac{\sum_{i=1}^k f_i (m_i - \mu)^2}{N}$$

Δείγμα

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

k το πλήθος των τάξεων στις οποίες έχουμε ταξινομήσει τις παρατηρήσεις

f_i η συχνότητα της i τάξης

m_i ο κεντρικός όρος της i τάξης

$\bar{X}$ ο δειγματικός αριθμητικός μέσος

μ ο αριθμητικός μέσος του πληθυσμού

ΤΑΞΕΙΣ	m_i	f_i	$m_i - \bar{X}$	$(m_i - \bar{X})^2$	$f_i (m_i - \bar{X})^2$
10 - 20	15	1	-42,2	1780,84	1780,84
20 - 30	25	6	-32,2	1036,84	6221,04
30 - 40	35	9	-22,2	492,84	4435,56
40 - 50	45	31	-12,2	148,84	4614,04
50 - 60	55	42	-2,2	4,84	203,28
60 - 70	65	32	7,8	60,84	1946,88
70 - 80	75	17	17,8	316,84	5386,28
80 - 90	85	10	27,8	772,84	7728,40
90 - 100	95	2	37,8	1428,84	2857,68
		150			35174,00

$$s^2 = 236,07$$

Τυπική Απόκλιση

Πληθυσμός

$$\sigma = +\sqrt{\sigma^2}$$

Δείγμα

$$S_{op} = +\sqrt{S_{op}^2}$$

και

$$S = +\sqrt{S^2}$$

Η τυπική απόκλιση μας επιτρέπει να προσδιορίζουμε με αρκετή ακρίβεια το διάστημα, γύρω από τον μέσο μιας κατανομής, στο οποίο συγκεντρώνονται οι περισσότερες τιμές της.