2.3.2 Παράμετροι Διασποράς

Ένα μέτρο θέσης δεν είναι από μόνο του ικανό να αντιπροσωπεύσει ικανοποιητικά ένα σύνολο παρατηρήσεων. Χρειαζόμαστε και ένα μέτρο διασποράς, δηλαδή ένα μέτρο το οποίο μας δίνει με τρόπο περιληπτικό και αντικειμενικό τη μεταβλητότητα ή ανομοιογένεια, των παρατηρήσεων.

2.3.2.1 Εύρος

Το εύρος d της μεταβολής μιας μεταβλητής ορίζεται ως η διαφορά της μικρότερης από μεγαλύτερη τιμή, δηλαδή $d=X_n- X_1$

Το εύρος είναι ένα πολύ χονδρικό μέτρο διασποράς το οποίο όμως έχει ως πλεονεκτήματα ότι υπολογίζεται αμέσως και γίνεται εύκολα αντιληπτό και από τον μη ειδικό. Σε επαναληπτικές λήψεις μικρών δειγμάτων από έναν πληθυσμό, το εύρος των τιμών σε καθένα από τα δείγματα, χρησιμεύει ως δείκτης της μεταβλητότητας του πληθυσμού. Ιδιαίτερα στο στατιστικό έλεγχο της ποιότητας ενός προϊόντος συνηθίζεται η λήψη πάρα πολλών μικρών δειγμάτων, στα οποία είναι εύκολο να υπολογιστεί το εύρος των τιμών. Από το εύρος των δειγμάτων αυτών παίρνουμε μια εικόνα για τη μεταβλητότητα της ποιότητας του παραγόμενου προϊόντος.

Το εύρος όμως έχει και σημαντικά μειονεκτήματα, όπως: Επηρεάζεται σημαντικά από μία μόνον τιμή, η οποία μπορεί να βρίσκεται πολύ πιο πάνω ή πιο κάτω από τις υπόλοιπες. Η τιμή του συνήθως επηρεάζεται από το πλήθος των παρατηρήσεων. Γενικά, περιμένουμε το εύρος ενός μεγάλου δείγματος να είναι μεγαλύτερο από το εύρος ενός μικρού δείγματος.

2.3.2.2 Ενδοτεταρτημοριακό πλάτος

Το ενδοτεταρτημοριακό πλάτος Q της μεταβολής μιας μεταβλητής ορίζεται ως η διαφορά του πρώτου τεταρτημορίου Q_1 από το τρίτο τεταρτημόριο Q_3, δηλαδή $Q=Q_3- Q_1$

2.3.2.3 Σχετικό ενδοτεταρτημοριακό πλάτος

Το σχετικό ενδοτεταρτημοριακό πλάτος CQ της μεταβολής μιας μεταβλητής ορίζεται ως ο λόγος του ενδοτεταρτημοριακού πλάτους δια της διαμέσου M, δηλαδή $CQ=Q/M$

2.3.2.4 Διακύμανση – Τυπική Απόκλιση

Ο ορισμός της διακύμανσης ενός συνόλου δεδομένων διαφοροποιείται ανάλογα με το αν τα δεδομένα αποτελούν ένα δείγμα ή τον πληθυσμό. Όταν τα δεδομένα αποτελούν τον πληθυσμό τότε, το πλήθος τους συμβολίζεται με N, και ο αριθμητικός τους μέσος με μ. Η διακύμανση τους συμβολίζεται με σ^2 και ορίζεται ως ο αριθμητικός μέσος των τετραγωνικών αποκλίσεων από τη μέση τιμή. Έχουμε δηλαδή:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i-\mu\right)^2$$

Αν έχουμε ένα δείγμα παρατηρήσεων τότε η διακύμανση τους συμβολίζεται με s^2 και ισούται με το άθροισμα των η τετραγωνικών αποκλίσεων από το δειγματικό μέσο διαιρεμένο δια n-1. Είναι δηλαδή:

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

Παρατηρούμε ότι και στις δύο περιπτώσεις ο αριθμητής είναι το άθροισμα των τετραγωνικών αποκλίσεων των παρατηρήσεων από τον αριθμητικό τους μέσο. Στη διακύμανση όμως του δείγματος δε διαιρούμε το άθροισμα αυτό με n αλλά με n-1. Επειδή συνήθως επεξεργαζόμαστε δειγματικά σύνολα δεδομένων στη συνέχεια θα αναφερόμαστε σχεδόν αποκλειστικά στην δειγματική διακύμανση.

Η τετραγωνική ρίζα s της δειγματικής διακύμανσης ονομάζεται δειγματική τυπική απόκλιση.

Παράδειγμα 2.6

Να υπολογιστούν η διασπορά και η τυπική απόκλιση των δεδομένων του παραδείγματος 2.1

Για να απλοποιήσουμε την υπολογιστική διαδικασία κατασκευάζουμε τον παρακάτω πίνακα:

α/α	X_i	$X_i-\overline{X}$	$\left(X_i-\overline{X}\right)^2$
1	33	10	100
2	36	13	169
3	5	-18	324
4	8	-15	225
5	36	13	169
6	30	7	49
7	12	-11	121
8	13	-10	100
9	4	-19	361
10	34	11	121
11	40	17	289
12	65	42	1764
13	4	-19	361
14	2.5	-20.5	420.25
15	2.5	-20.5	420.25
16	43	20	400
Σύνολο	368		5393.5

Συνεπώς η διασπορά και η τυπική απόκλιση υπολογίζονται ως εξής:

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\frac{5393.5}{15}=359.57 \text{ και } s=\sqrt{s^2}=\sqrt{359.57}=18.96$$

Όταν τα δεδομένα είναι ομαδοποιημένα σε μια κατανομή με κ τάξεις, υπολογίζουμε τη διακύμανση κάνοντας τη γνωστή υπόθεση ότι όλες οι τιμές που ανήκουν σε μια τάξη μπορούν να αντιπροσωπευτούν από την κεντρική τιμή της. Έτσι έχουμε:

$$s^2=\frac{1}{n-1}\sum_{i=1}^{k}f_i\left(X_i-\overline{X}\right)^2, \quad X_i=\frac{e_{i-1}+e_i}{2}$$

Είναι προφανές ότι ο υπολογισμός της διακύμανσης με βάση τους τύπους αυτούς είναι προσεγγιστικός, όπως προσεγγιστικός είναι και ο υπολογισμός των μέτρων θέσης από τις ομαδοποιημένες παρατηρήσεις.

Παράδειγμα 2.7

Να βρεθεί η διασπορά της ηλικίας του δείγματος που παρουσιάζεται στον πίνακα 2.8.

Στον πίνακα 2.8 θα προσθέσουμε τέσσερις νέες στήλες: Την στήλη με επικεφαλίδα X_i στην οποία θα καταγράψουμε το κέντρο της κάθε κλάσης, τη στήλη $X_i - \overline{X}$ στην οποία θα υπολογίσουμε τη διαφορά των κέντρων των κλάσεων από τη μέση τιμή, τη στήλη $\left(X_i - \overline{X}\right)^2$ και τέλος τη στήλη $f_i\left(X_i - \overline{X}\right)^2$ στη οποία θα υπολογίσουμε το γινόμενο των συχνοτήτων επί των τετραγωνικών αποκλίσεων.

Πίνακας 2.16 Μέσης τιμή ομαδοποιημένης ποσοτικής μεταβλητής

ΗΛΙΚΙΑ	f_i	X_i	$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$	$f_i\left(X_i - \overline{X}\right)^2$
[0, 6)	166	3	-32.14	1032.86	171455.5
[6, 12)	146	9	-26.14	683.21	99748.09
[12 ,18)	178	15	-20.14	405.55	72187.47
[18, 24)	229	21	-14.14	199.89	45774.59
[24, 30)	229	27	-8.14	66.23	15166.78
[30, 36)	221	33	-2.14	4.57	1010.4
[36, 42)	217	39	3.86	14.91	3236.21
[42, 48)	213	45	9.86	97.25	20715.29
[48, 54)	213	51	15.86	251.60	53590.02
[54, 60)	139	57	21.86	477.94	66433.35
[60, 66)	115	63	27.86	776.28	89272.12
[66, 72)	66	69	33.86	1146.62	75676.97
[72, 78)	52	75	39.86	1588.96	82626.03
[78, 84)	23	81	45.86	2103.30	48375.98
[84, 90)	3	87	51.86	2689.65	8068.935
[90, 96)	4	93	57.86	3347.99	13391.95
Σύνολο	2214				866729.7

Από τον παραπάνω πίνακα εύκολα καταλήγουμε στον υπολογισμό της διασποράς της ηλικίας:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i\left(X_i - \overline{X}\right)^2 = \frac{866729.7}{2213} = 391.65$$

Προφανώς η δειγματική τυπική απόκλιση θα είναι: s=19.79.

2.3.2.5 Δείκτης Μεταβλητότητας (σχετική τυπική απόκλιση)

Έστω ότι θέλουμε να συγκρίνουμε τη μεταβλητότητα σε δύο σύνολα παρατηρήσεων. Όταν οι παρατηρήσεις και στα δύο σύνολα εκφράζονται στις ίδιες μονάδες μέτρησης και έχουν τον ίδιο ή περίπου τον ίδιο αριθμητικό μέσο, τότε η σύγκριση της μεταβλητότητας μπορεί να γίνει με τη σύγκριση των τυπικών τους αποκλίσεων.

Όταν οι παρατηρήσεις στα δύο σύνολα εκφράζονται στις ίδιες μονάδες αλλά ο αριθμητικός μέσος τους διαφέρει σημαντικά, τότε η σύγκριση της μεταβλητότητας μπορεί να γίνει με το δείκτης μεταβλητότητας που ονομάζεται και σχετική τυπική απόκλιση. Ο δείκτης μεταβλητότητας CV μίας μεταβλητής ορίζεται ως ο λόγος της τυπικής απόκλισης s δια τη μέση τιμή $\overline{X}$, δηλαδή:

$$CV = \frac{s}{\overline{X}}$$

2.3.2.6 Αλλαγή κλίμακας

Αν προσθέσουμε στις τιμές μίας μεταβλητής την ποσότητα x_0 τότε ο αριθμητικός μέσος της μεταβλητής που προκύπτει μεταβάλλεται επίσης κατά x_0 ενώ η διασπορά παραμένει αμετάβλητη. Αν πολλαπλασιαστούν επί $h \neq 0$ οι τιμές μιας μεταβλητής τότε ο αριθμητικός μέσος πολλαπλασιάζεται επί h ενώ η διασπορά πολλαπλασιάζεται επί h^2 Από τα παραπάνω προκύπτει ότι αν ανάμεσα στις μεταβλητές X, Y ισχύει:

$$Y = x_0 + hX \Leftrightarrow X = \frac{Y - x_0}{h}$$

Τότε:

$$\overline{Y} = x_0 + h\overline{X} \Leftrightarrow \overline{X} = \frac{\overline{Y} - x_0}{h}$$

$$s^2(Y) = h^2 s^2(X) \Rightarrow s^2(X) = \frac{s^2(Y)}{h^2}$$

2.3.2.7 Τυποποιημένες τιμές

Η τυπική απόκλιση χρησιμοποιείται για να αξιολογήσουμε πόσο μεγάλη είναι η απόκλιση κάθε παρατήρησης από τον αριθμητικό μέσο. Για τον σκοπό αυτό αρκεί να διαιρέσουμε κάθε απόκλιση από τη μέση τιμή με την τυπική απόκλιση προκύπτει η παρακάτω μεταβλητή:

$$Z = \frac{X - \overline{X}}{s}$$

Σύμφωνα με τα παραπάνω η μεταβλητή Z, ανεξάρτητα από το ποια είναι η μέση τιμή και η διασπορά της μεταβλητής X, θα έχει πάντα μέση τιμή 0 και διασπορά 1. Οι τιμές της μεταβλητής Z λέγονται τυποποιημένες τιμές. Οι τυποποιημένες τιμές εκφράζουν τις διαφορές σε τυπικές αποκλίσεις και για αυτό επιτρέπουν τις συγκρίσεις ανάμεσα σε μεταβλητές των οποίων η κλίμακες τιμών διαφέρουν.