

Κεφάλαιο 6

Διασπορά

6.1 Η έννοια της διασποράς

Ο μέσος αριθμητικός, ο μέσος γεωμετρικός, η διάμεσος, η επικρατούσα τιμή και τα υπόλοιπα άλλα μέτρα τα οποία έχουμε εξετάσει στο προηγούμενο κεφάλαιο έχουν ως αντικειμενικό σκοπό να αντιπροσωπεύουν έναν πληθυσμό σε μία μόνο παράμετρο, η οποία μας δίνει το σημείο στο οποίο τείνουν να συγκεντρωθούν οι τιμές της μεταβλητής του πληθυσμού που ερευνούμε.

Η αντιπροσώπευση όμως ενός πληθυσμού με μια από τις πιο πάνω παραμέτρους έχει αξία εφόσον ο πληθυσμός παρουσιάζει μεγάλη *ομοιογένεια*. Αντίθετα αν ο πληθυσμός παρουσιάζει μεγάλη *ανομοιογένεια*, τότε τα μέτρα της κεντρικής τάσης και θέσης θα πρέπει να μη χρησιμοποιούνται ως αντιπροσωπευτικοί αριθμοί ενός πληθυσμού.

Ο βαθμός κατά τον οποίο οι διάφορες τιμές ενός πληθυσμού τείνουν να είναι διεσπαρμένες γύρω από τον αριθμητικό μέσο ονομάζεται **διασπορά**.

Η παράμετρος που μας πληροφορεί αν οι τιμές των παρατηρήσεων είναι συγκεντρωμένες ή διασκορπισμένες σε σχέση με τον αριθμητικό μέσο ονομάζεται **διασπορά ή διακύμανση**.

6.2 Μέτρα διασποράς

Τα μέτρα διασποράς που χρησιμοποιούνται είναι:

- το εύρος μεταβολής,
- το ημιενδοτεταρτημορικό εύρος,
- η μέση απόκλιση,
- η τυπική απόκλιση και
- ο συντελεστής μεταβλητότητας.

6.2.1. Το εύρος μεταβολής

Ορισμός: Είναι μια παράμετρος που μας δίνει τη διαφορά ανάμεσα στη μεγαλύτερη και τη μικρότερη τιμή μιας σειράς παρατηρήσεων και συμβολίζεται με R .

Παράδειγμα: Η μέση θερμοκρασία σε 18 πόλεις της χώρας στις 20 Αυγούστου 1978 είχε ως εξής: 19, 20, 20, 23, 24, 24, 25, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35.

Το εύρος της μεταβολής θα είναι η διαφορά μεταξύ μεγαλύτερης και μικρότερης θερμοκρασίας. Δηλαδή: $R = M - E = 35 - 19 = 16$

όπου R - το εύρος, M - η μέγιστη και E - η ελάχιστη θερμοκρασία.

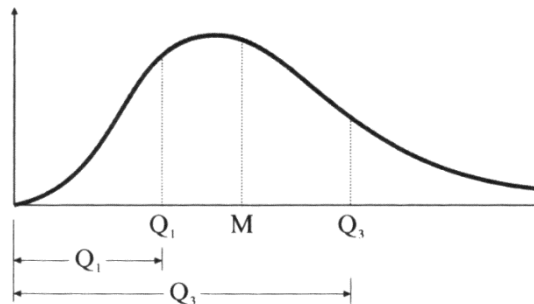
Η παράμετρος αυτή είναι πολύ εύκολο να υπολογιστεί. Όμως δεν είναι ικανοποιητική, γιατί εξαρτάται μόνο από δύο τιμές των παρατηρήσεών μας, τις ακραίες τιμές. Αν οι

δύο αυτές ακραίες τιμές είναι πολύ μακριά από το σύνολο των άλλων παρατηρήσεων, δίνουν ψεύτικη εικόνα για την ένταση της διασποράς.
Το εύρος χρησιμοποιείται συνήθως στην Μετεωρολογία και τα χρηματιστήρια.

6.2.2. Ημιενδοτεταρτημοριακό εύρος

Ορισμός: Η παράμετρος αυτή ορίζεται ως το ήμισυ της διαφοράς μεταξύ τρίτου και πρώτου τεταρτημορίου και συμβολίζεται με το Q , δηλαδή:

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



Όσο μικρότερη είναι η τιμή του Q τόσο μεγαλύτερη είναι η συγκέντρωση των τιμών της μεταβλητής. Το ημιενδοτεταρτημοριακό εύρος δεν επηρεάζεται από τις ακραίες τιμές όπως συμβαίνει με το εύρος μεταβολής.

6.2.3. Μέση απόκλιση

Ορισμός: Η μέση απόκλιση ορίζεται ως ο αριθμητικός μέσος όλων των αποστάσεων των τιμών της μεταβλητής από τον αριθμητικό μέσο της μεταβλητής αυτής.

α) Αταξινόμητες παρατηρήσεις

Στην περίπτωση αταξινόμητων στοιχείων, η μέση απόκλιση δίνεται από τύπο:

$$M \cdot A = \frac{\sum_{i=1}^N |x_i - \mu|}{N}$$

όπου οι κατακόρυφες γραμμές σημαίνουν ότι παίρνουμε την απόλυτη διαφορά ανάμεσα στις τιμές της μεταβλητής και του μέσου αριθμητικού και N το πλήθος των παρατηρήσεων της μεταβλητής X .

Παράδειγμα: Δίνονται οι παρατηρήσεις: 3, 5, 7, 9, 16. Ο αριθμητικός μέσος θα είναι:

$$\mu = \frac{3+5+7+9+16}{5} = \frac{40}{5} = 8$$

Η μέση απόκλιση λοιπόν θα είναι:

$$M \cdot A = \frac{\sum_{i=1}^N |x_i - \mu|}{N} = \frac{|3-8| + |5-8| + |7-8| + |9-8| + |16-8|}{5} = \frac{18}{5} = 3,6$$

β) Ταξινομημένα στοιχεία σε μορφή κατανομών συχνοτήτων

Στην περίπτωση ταξινομημένων στοιχείων, η μέση απόκλιση δίνεται από τον τύπο:

$$M \cdot A = \frac{\sum_{i=1}^N f_i |x_i - \mu|}{\sum_{i=1}^N f_i}$$

Παράδειγμα: Στον παρακάτω πίνακα δίνεται η κατανομή των ημερών απουσίας 21 εργαζομένων λόγω ασθένειας.

Πίνακας					
Τάξεις	f_i	Κεντρικές τιμές x_i	$f_i x_i$	Αποκλίσεις	
				$ x_i - \mu $	$f_i x_i - \mu $
1 - 3	2	2	4	4	8
3 - 5	5	4	20	2	10
5 - 7	7	6	42	0	0
7 - 9	5	8	40	2	10
9 - 11	2	10	20	4	8
Σύνολο	21		126		36

Να υπολογιστεί η μέση απόκλιση.

Λύση:

Ο μέσος αριθμητικός είναι $\mu = \frac{\sum f_i x_i}{\sum f_i} = \frac{126}{21} = 6$.

Η μέση απόκλιση θα είναι $M \cdot A = \frac{\sum_{i=1}^N f_i |x_i - \mu|}{\sum_{i=1}^N f_i} = \frac{36}{21} = 17,7$.

Παρατηρούμε ότι για τον υπολογισμό της μέσης απόκλισης δεν λαμβάνουμε υπόψη μας τα σημεία των αποκλίσεων αλλά μόνο τις απόλυτες τιμές των διαφορών. Αυτό γίνεται από ανάγκη, γιατί αν λάβουμε το άθροισμα των αποκλίσεων μεταξύ των τιμών της μεταβλητής και του αριθμητικού μέσου, αυτό είναι πάντοτε μηδέν, δηλαδή:

$$\sum (x_i - \mu) = \sum x_i - \mu N = N\mu - \mu N = 0$$

6.2.4. Διακύμανση και τυπική απόκλιση

Για να αποφύγουμε τη διαδικασία εκείνη κατά την οποία δεν λαμβάνουμε υπόψη σημεία των αρνητικών διαφορών μεταξύ του αριθμητικού μέσου και των τιμών της μεταβλητής, μπορούμε να υψώσουμε τις διαφορές στο τετράγωνο για να μην έχουμε να ασχοληθούμε με θετικά ή αρνητικά σημεία των αποκλίσεων.

Ορισμός: Ο αριθμητικός μέσος των τετραγώνων των αποκλίσεων είναι επίσης μέτρο διασποράς και ονομάζεται **διακύμανση**. Επομένως, διακύμανση ενός πλήθους παρατηρήσεων ονομάζεται ο αριθμητικός μέσος των τετραγώνων των αποκλίσεων των τιμών των παρατηρήσεων από τον αριθμητικό μέσο.

Η διακύμανση εκφράζεται σε μονάδες, οι οποίες είναι τα τετράγωνα των αρχικών μονάδων, π.χ. αν το εισόδημα είναι σε δραχμές, η διακύμανση εκφράζεται με μονάδα τη δραχμή στο τετράγωνο, ομοίως αν η μεταβλητή εκφράζεται σε εκατοστά, η διακύμανση εκφράζεται σε εκατοστά στο τετράγωνο κ.λπ. Για να έχουμε όμως ένα

δείκτη ο οποίος να μετράει τη διασπορά και να εκφράζεται στις ίδιες μονάδες που εκφράζεται η μεταβλητή μας, παίρνουμε την τετραγωνική ρίζα της διακύμανσης. Το μέτρο αυτό ονομάζεται **τυπική απόκλιση** και είναι το μέτρο διασποράς που χρησιμοποιούμε συνήθως στην πράξη.

Όσο μεγαλύτερη είναι η τυπική απόκλιση, τόσο μεγαλύτερη είναι η διασπορά των παρατηρήσεων από τον αριθμητικό μέσο.

Υπολογισμός της διακύμανσης και της τυπικής απόκλισης

Η διακύμανση συμβολίζεται με σ^2 ή $Var(X)$ και η τυπική απόκλιση με σ .

α) Αταξινόμητες παρατηρήσεις

Υποθέτουμε ότι έχουμε τις N παρατηρήσεις $x_1, x_2, \dots, x_N$, που ο αριθμητικός μέσος τους είναι ο μ .

Η διακύμανση των παραπάνω παρατηρήσεων δίνεται από τον τύπο:

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

και η τυπική απόκλιση από τον τύπο:

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

Παράδειγμα: Να υπολογιστεί η διακύμανση και η τυπική απόκλιση στις παρακάτω παρατηρήσεις: $x_i : 2, 3, 5, 8, 12$

Λύση:

Επειδή οι παρατηρήσεις δεν παρουσιάζουν συχνότητα (δηλαδή η κάθε τιμή απαντάται μία φορά), θα χρησιμοποιήσουμε τον παρακάτω τύπο για τον υπολογισμό της διακύμανσης:

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

Υπολογίζουμε πρώτα τον αριθμητικό μέσο:

$$\mu = \frac{2+3+5+8+12}{5} = 6$$

Μετά σχηματίζουμε τον πίνακα:

x_i	$x_i - \mu$	$(x_i - \mu)^2$
2	-4	16
3	-3	9
5	-1	1
8	2	4
12	6	36
Σύνολο		66

Οπότε έχουμε:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{66}{5} = 13,2$$

Η τυπική απόκλιση θα είναι:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} = \sqrt{13,2} = 3,63$$

β) Ομαδοποιημένες παρατηρήσεις

Στην περίπτωση που οι παρατηρήσεις μας δίνονται σε μορφή κατανομής συχνοτήτων, η διακύμανση υπολογίζεται με τους παρακάτω τύπους:

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

$$\sigma^2 = \frac{\sum_{i=1}^k f_i x_i^2}{\sum f_i} - \mu^2$$

$$\sigma^2 = \delta \left[\frac{\sum f_i \xi_i^2}{\sum f_i} - \left(\frac{\sum f_i \xi_i}{\sum f_i} \right)^2 \right]$$

Παράδειγμα: Στον παρακάτω πίνακα δίνεται η ταχύτητα με την οποία 10 αυτοκίνητα πέρασαν μια επικίνδυνη διασταύρωση.

Πίνακας	
Τάξεις	f_i
0-10	4
10-20	3
20-30	2
30-40	1
Σύνολο	10

Ζητείται η διακύμανση και η τυπική απόκλιση με τρεις διαφορετικούς τρόπους.

Λύση:

1^{ος} τρόπος

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

Ο μέσος (αριθμητικός) είναι: $\mu = \frac{150}{10} = 15$

Για να κάνουμε τους υπολογισμούς αυτούς, μετασχηματίζουμε τον πίνακά μας σε:

1 ^η	2 ^η	3 ^η	4 ^η	5 ^η	6 ^η	7 ^η	8 ^η
Τάξεις	f_i	x_i	$x_i - \mu$	$(x_i - \mu)^2$	$f_i(x_i - \mu)^2$	x_i^2	$f_i x_i^2$
0-10	4	5	-10	100	400	25	100
10-20	3	15	0	0	0	225	675
20-30	2	25	10	100	200	625	1.250
30-40	1	35	20	400	400	1.225	1.225
Σύνολο	10				1.000	25	3.250

Η διακύμανση είναι:

$$\sigma^2 = \frac{\sum_{i=1}^N f_i (x_i - \mu)^2}{\sum f_i} = \frac{1.000}{10} = 100$$

η τυπική απόκλιση είναι:

$$\sigma = \sqrt{\sigma^2} = \sqrt{100} = 10$$

2^{ος} τρόπος:

$$\sigma^2 = \frac{\sum_{i=1}^k f_i x_i^2}{\sum f_i} - \mu^2$$

Ο μέσος (αριθμητικός) είναι: $\mu = \frac{150}{10} = 15$

Για να κάνουμε τους υπολογισμούς αυτούς, μετασχηματίζουμε τον πίνακά μας δημιουργώντας τις στήλες 7^η και 8^η

Η διακύμανση είναι:

$$\sigma^2 = \frac{\sum_{i=1}^N f_i x_i^2}{\sum f_i} - \mu^2 = \frac{3.250}{10} - 15^2 = 325 - 225 = 100$$

η τυπική απόκλιση είναι:

$$\sigma = \sqrt{\sigma^2} = \sqrt{100} = 10$$

3^{ος} τρόπος:

Υπολογισμός της διακύμανσης με τον τύπο:

$$\sigma^2 = \delta \left[\frac{\sum f_i \xi_i^2}{\sum f_i} - \left(\frac{\sum f_i \xi_i}{\sum f_i} \right)^2 \right]$$

Για να κάνουμε τους υπολογισμούς αυτούς, μετασχηματίζουμε τον αρχικό πίνακα σε:

1 ^η	2 ^η	3 ^η	4 ^η	5 ^η	6 ^η	7 ^η
Τάξεις	f_i	x_i	ξ_i	ξ_i^2	$f_i \xi_i$	$f_i \xi_i^2$
0-10	4	5	-1	1	-4	4
10-20	3	15	0	0	0	0
20-30	2	25	1	1	2	2
30-40	1	35	2	4	2	4
Σύνολο	10				0	10

Έτσι:

$$\sigma^2 = s \left[\frac{\sum f_i \xi_i^2}{\sum f_i} - \left(\frac{\sum f_i \xi_i}{\sum f_i} \right)^2 \right] = 10^2 \left[\frac{10}{10} - \left(\frac{0}{10} \right) \right] = 100$$

και

$$\sigma = \sqrt{\sigma^2} = \sqrt{100} = 10$$

Ιδιότητες της διακύμανσης

1. Αν $x_1 = x_2 = \dots = x_N = a$, τότε η διακύμανση είναι μηδέν, δηλαδή η διακύμανση μιας σταθεράς είναι μηδέν.

Απόδειξη: άσκηση για το σπίτι

2. Αν όλες οι τιμές μιας μεταβλητής X πολλαπλασιαστούν με τον ίδιο αριθμό a , η διακύμανση πολλαπλασιάζεται με το τετράγωνο του αριθμού αυτού, δηλαδή:

$$\text{Var}(aX) = a^2 \cdot \text{Var}(X)$$

Απόδειξη: άσκηση για το σπίτι

3. Αν όλες οι τιμές μιας μεταβλητής X αυξηθούν κατά μια σταθερά a , τότε η διακύμανση δεν μεταβάλλεται, δηλαδή:

$$\text{Var}(X + a) = \text{Var}(X)$$

6.2.5. Συντελεστής μεταβλητότητας

Παράδειγμα: Έστω ότι από ένα δείγμα 20 μαθητών της Α' Γυμνασίου βρήκαμε μέσο βάρος $\mu_1 = 40 \text{ kg}$ και τυπική απόκλιση $\sigma_1 = 6 \text{ kg}$, ενώ από ένα δεύτερο δείγμα 30 μαθητών της Γ' Λυκείου βρήκαμε μέσο βάρος $\mu_2 = 75 \text{ kg}$ και τυπική απόκλιση $\sigma_2 = 6 \text{ kg}$. 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 kg στο μέσο βάρος των 40 kg είναι διαφορετική από αυτήν που έχουν στο μέσο βάρος των 75 kg.

Παράδειγμα: Ακόμη, ας υποθέσουμε ότι ο μέσος μισθός των υπαλλήλων μιας εταιρείας Α είναι $\mu_A = 350.000 \text{ €}$ με τυπική απόκλιση $\sigma_A = 42.000 \text{ €}$, ενώ για τους υπαλλήλους μιας εταιρείας Β είναι $\mu_B = 1.400 \text{ €}$ με τυπική απόκλιση $\sigma_B = 350 \text{ €}$. 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**, ο οποίος ορίζεται (για $\mu \neq 0$) από το λόγο:

$$\boxed{CV = \frac{\sigma}{\mu}} \quad \text{ή} \quad \boxed{CV = \frac{\sigma}{\mu} \cdot 100\%}$$

Αν $\mu < 0$ τότε αντί του μ χρησιμοποιούμε την απόλυτη τιμή του μέσου, δηλ $|\mu|$.

Ο συντελεστής μεταβολής είναι ανεξάρτητος από τις μονάδες μέτρησης, εκφράζεται επί τοις εκατό 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:

$$CV_A = \frac{\sigma}{\mu} = \frac{6}{40} = 0,15 = 15\%$$

$$CV_B = \frac{\sigma}{\mu} = \frac{6}{75} = 0,08 = 8\%$$

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%.

ΚΑΛΗ ΕΠΙΤΥΧΙΑ !!!