Περιγραφική Στατιστική

Παναγιώτα Λάλου

1. Βασικές έννοιες

Ορισμός: Στατιστικός **πληθυσμός** ονομάζεται το σύνολο των πειραματικών μονάδων π.χ άνθρωποι, ζώα, επιχειρήσεις κ.λπ, οι οποίες συμμετέχουν στην έρευνα που πραγματοποιείται.

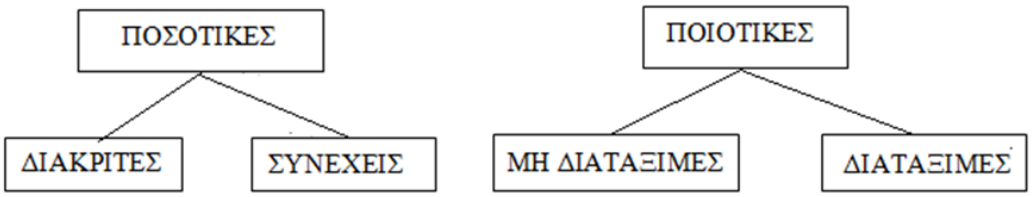
Τα στοιχεία του πληθυσμού, εξετάζονται ως προς ένα ή περισσότερα χαρακτηριστικά. Το χαρακτηριστικό ως προς το οποίο εξετάζουμε έναν πληθυσμό, ονομάζεται στατιστική **μεταβλητή**. Οι δυνατές τιμές που μπορεί να πάρει μια μεταβλητή λέγονται τιμές της μεταβλητής. Τις μεταβλητές τις διακρίνουμε στις εξής κατηγορίες:

α) Σε ποιοτικές ή κατηγορικές: των οποίων οι τιμές δεν είναι αριθμοί. Τέτοιες είναι για παράδειγμα: το φύλο (αγόρι-κορίτσι) και η ποδοσφαιρική ομάδα προτίμησης. Διακρίνουμε δύο είδη ανάλογα με το αν υπάρχει η έννοια της διάταξης στις τιμές των μεταβλητών ή όχι. Αν οι τιμές μπορούν να διαταχθούν μιλάμε για διατάξιμες ποιοτικές μεταβλητές, π.χ. κατάσταση ασθενή (με τιμές κρίσιμη, κακή, μέτρια καλή), επίπεδο εκπαίδευσης (με τιμές δημοτικό, γυμνάσιο, λύκειο, πανεπιστήμιο). Διαφορετικά, μιλάμε για μη διατάξιμες π.χ. χρώμα αυτοκινήτου.

β) Σε ποσοτικές: τον οποίων οι τιμές είναι αριθμοί. Παράδειγμα: το ύψος του μαθητή, ο αριθμός υπαλλήλων μιας επιχείρησης. Οι ποσοτικές μεταβλητές διακρίνονται σε διακριτές μεταβλητές και σε συνεχείς. Διακριτές μεταβλητές λέγονται αυτές που παίρνουν μόνο " μεμονωμένες" τιμές, συνήθως ακέραιες.

Π.χ το αποτέλεσμα ρίψης ενός ζαριού, ο αριθμός αδερφών. Συνεχείς λέγονται οι μεταβλητές που μπορούν να πάρουν οποιαδήποτε τιμή ενός διαστήματος πραγματικών αριθμών (α,β). Τέτοια είναι για παράδειγμα το ύψος των μαθητών ενός τμήματος.

ΕΙΔΗ ΜΕΤΑΒΛΗΤΩΝ



Σε πολλές περιπτώσεις η εξέταση όλων των μονάδων του πληθυσμού είναι δύσκολη ή ακόμα και αδύνατη. Δύσκολη, κυρίως για οικονομικούς λόγους, αφού ο χρόνος και το κόστος της έρευνας αυξάνεται καθώς αυξάνονται οι μονάδες του πληθυσμού. Επίσης, αν για παράδειγμα κάποια προϊόντα πρέπει να εξεταστούν για την αντοχή τους που συνίσταται στην εκτίμηση σημείου κάμψης ή του σημείου πέραν του οποίου σπάνε, μελέτη όλου του πληθυσμού θα σήμαινε καταστροφή του συνόλου της παραγωγής. Στις παραπάνω περιπτώσεις είναι προτιμότερο να μελετήσουμε ένα μέρος – υποσύνολο του πληθυσμού, τα συμπεράσματα από το οποίο μπορούν να γενικευτούν για το σύνολο του πληθυσμού. Το υποσύνολο αυτό του πληθυσμού, ονομάζεται **δείγμα**. Τα συμπεράσματα όμως που θα προκύψουν από την μελέτη του δείγματος θα είναι αξιόπιστα, θα ισχύουν δηλαδή με ικανοποιητική ακρίβεια για το σύνολο του πληθυσμού, αν η επιλογή του δείγματος γίνει με σωστό τρόπο. Τότε λέμε ότι το δείγμα είναι «αντιπροσωπευτικό». Οι αρχές και οι μέθοδοι επιλογής του κατάλληλου δείγματος είναι αντικείμενο του κλάδου της Στατιστικής που ονομάζεται Δειγματοληψία.

2. Παρουσίαση Στατιστικών Δεδομένων

Η παρουσίαση των στατιστικών δεδομένων γίνεται με στατιστικούς πίνακες και γραφικές παραστάσεις . Η παρουσίαση των στατιστικών δεδομένων σε πίνακες γίνεται με την κατάλληλη τοποθέτηση των πληροφοριών σε γραμμές και στήλες, με τρόπο που να διευκολύνεται η σύγκριση των στοιχείων και η καλύτερη ενημέρωση του αναγνώστη.

Ορισμοί:

Το πλήθος όλων των παρατηρήσεων του δείγματος ονομάζεται **μέγεθος n** του δείγματος.

Αν $x_1, x_2, \dots, x_k$ είναι οι τιμές μιας μεταβλητής Χ, δείγματος μεγέθους n, ο φυσικός αριθμός ν_i που δείχνει πόσες φορές εμφανίζεται η τιμή x_i στο δείγμα ονομάζεται **συχνότητα** της x_i. Προφανώς ισχύει:

$$\nu_1 + \nu_2 + \dots + \nu_k = n$$

Αν διαιρέσουμε τη συχνότητα ν_i με το μέγεθος του δείγματος n , προκύπτει η **σχετική συχνότητα** $\boldsymbol{f_i}$ της τιμής x_i. Δηλαδή:

$$f_i = \frac{\nu_i}{n} \qquad i = 1, \dots, n$$

Για την σχετική συχνότητα ισχύουν:

i) $0 \leq f_i \leq 1$

ii) $f_1 + f_2 + \dots + f_k = 1$

Οι ποσότητες x_i, ν_i, f_i μπορούν να συγκεντρωθούν σε έναν πίνακα ο οποίος ονομάζεται **πίνακας κατανομής συχνοτήτων**.

Το πλήθος των παρατηρήσεων που είναι μικρότερες ή ίσες της συγκεκριμένης τιμής x_i της μεταβλητής, ονομάζεται **αθροιστική συχνότητα** $\boldsymbol{N_i}$ της x_i. Προφανώς ισχύει:

$$N_i = \nu_1 + \nu_2 + \dots + \nu_i$$

Το ποσοστό των παρατηρήσεων που είναι μικρότερες ή ίσες της συγκεκριμένης τιμής x_i της μεταβλητής, ονομάζεται **σχετική αθροιστική συχνότητα F_i** της x_i.

$$F_i = f_1 + f_2 + \ \ldots + f_i$$

Η αθροιστική συχνότητα και η σχετική αθροιστική συχνότητα ορίζονται μόνο για ποσοτικές μεταβλητές.

Παράδειγμα 1:

Δίνεται ο αριθμός των ημερών απουσίας 10 εργαζομένων λόγω ασθενείας.

Εργαζόμενος	Ημέρες ασθεν.
1	5
2	4
3	3
4	5
5	2
6	5
7	3
8	2
9	1
10	1

i) Να γίνει ο πίνακας κατανομής συχνοτήτων και αθροιστικών συχνοτήτων.

ii) Ποιό είναι το ποσοστό των εργαζομένων που ασθένησαν 5 ημέρες και ποιο αυτών που ασθένησαν το πολύ 3 ημέρες.

Λύση:

x_i	ν_i	$f_i = \frac{\nu_i}{n}$	$f_i\%$	N_i	F_i	$F_i\%$
1	2	0,2	20	2	0,2	20
2	2	0,2	20	4	0,4	40
3	2	0,2	20	6	0,6	60
4	1	0,1	10	7	0,7	70
5	3	0,3	30	10	1	100
Σύνολο	10	1	100			

ii) 30%, 60%

2.1 Ομαδοποίηση Δεδομένων

Όταν το πλήθος των παρατηρήσεων είναι μεγάλο δημιουργούνται δυσχέρειες στη συλλογή πληροφοριών. Αυτό συμβαίνει κυρίως στην περίπτωση μιας συνεχούς μεταβλητής, όπου αυτή μπορεί να πάρει οποιαδήποτε τιμή στο διάστημα ορισμού της αλλά συμβαίνει και στην περίπτωση διακριτών μεταβλητών. Τότε θα πρέπει να ομαδοποιηθούν τα δεδομένα σε μικρό πλήθος ομάδων (διαστημάτων), που ονομάζονται και κλάσεις (class intervals), έτσι ώστε κάθε τιμή να ανήκει μόνο σε μία κλάση. Τα άκρα των κλάσεων καλούνται όρια των κλάσεων (class boundaries).

Είναι φανερό ότι μικρό πλήθος κλάσεων σημαίνει και μεγάλη απώλεια πληροφορίας από τα αρχικά (πρωτογενή) δεδομένα. Μεγάλο πλήθος (άνω των είκοσι) κλάσεων δεν συνηθίζεται γιατί έχει δυσκολία στους υπολογισμούς. Όσον αφορά στο πλάτος των κλάσεων θα πρέπει να γνωρίζουμε ότι δεν είναι απαραίτητο να είναι το ίδιο σε όλες τις κλάσεις. Τις περισσότερες φορές όμως είναι πιο χρήσιμο να έχουμε ίδιο πλάτος, αφού έτσι ‘διαβάζουμε’ υπολογίζονται μέτρα θέσεως και διασποράς με μεγαλύτερη ευκολία.

Σε κάθε κλάση διακρίνουμε το κάτω και το άνω άκρο της το ημιάθροισμα των δυο άκρων μας δίνει την κεντρική τιμή η οποία χρησιμοποιείται για τον υπολογισμό των μέτρων θέσεως και διασποράς που θα δούμε πιο κάτω. Οι κλάσεις που θα ασχοληθούμε θα έχουν τη μορφή [,).

Η διαδικασία της ομαδοποίησης δεδομένων ακολουθεί τα εξής βήματα:

α) Κατατάσσουμε τις παρατηρήσεις κατά σειρά. Από τη μικρότερη προς τη μεγαλύτερη.

β) Βρίσκουμε το εύρος (τη διαφορά μεταξύ μεγαλύτερης x_{max} και μικρότερης x_{min} παρατήρησης)

$$R = x_{max} - x_{min}$$

γ) Διαιρούμε το R με το πλήθος των κλάσεων που επιθυμούμε να έχουμε και βρίσκουμε το πλάτος c κάθε κλάσης.

δ) Εντάσσουμε κάθε παρατήρηση στην κλάση που ανήκει (συχνότητες των κλάσεων).

Παράδειγμα 2:

Στον πίνακα που ακολουθεί δίνονται οι θερμοκρασίες σε μία Μεσογειακή πόλη. Να κατασκευαστεί πίνακας με 6 κλάσεις-διαστήματα.

11	17	30	32
12	23	29	33
15	25	29	36
17	22	27	38
18	25	24	40
16	27	25	39
19	26	23	37
20	25	26	35
18	28	29	33
21	30	31	32

Το Εύρος των παρατηρήσεων είναι: 40-11=29. Το πλάτος των κλάσεων θα είναι: 29/6=4,8333 ≈ 5 (στρογγυλοποιούμε προς τα επάνω). Ο πίνακας κατανομής συχνοτήτων, σχετικών συχνοτήτων, σχετικών αθροιστικών συχνοτήτων που θα προκύψει είναι:

Κλάσεις [-)	Κεντρικές τιμές	Συχν.	Σχετική Συχνότητα	Σχετική συχνότητα ποσοστό	Αθρ. Συχν.	Αθρ. Σχετ. συχν	Αθρ. Σχετ. συχν.
[,)	x_i	ν_i	$f_i = \frac{\nu_i}{n}$	f_i %	N_i	F_i	F_i %
[11, 16)	13,5	3	0,075	7,5	3	0,075	7,5
[16, 21)	18,5	7	0,175	17,5	10	0,25	25
[21, 26)	23,5	9	0,225	22,5	19	0,475	47,5
[26, 31)	28,5	10	0,25	25	29	0,725	72,5
[31, 36)	33,5	6	0,15	15	35	0,875	87,5
[36, 41)	38,5	5	0,125	12,5	40	1	100

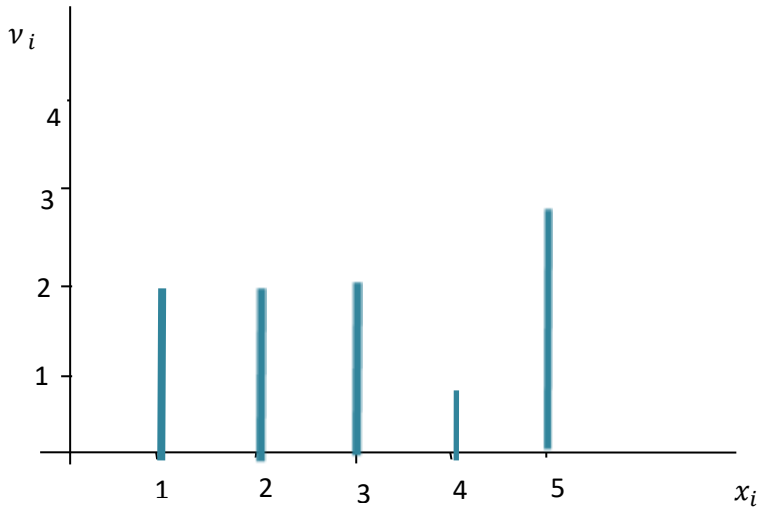
3. Γραφικές Παραστάσεις

Σε πολλές περιπτώσεις η γραφική παρουσίαση της κατανομής συχνοτήτων παρέχει περιεκτικά αλλά με σαφήνεια όλες τις δυνατές πληροφορίες σχετικά με την κατανομή. Υπάρχουν πολλοί τύποι γραφικής παρουσίασης των δεδομένων κάποιοι από τους οποίους είναι τα διαγράμματα που θα εξετάσουμε παρακάτω: α) διάγραμμα συχνοτήτων (line diagram) β) ιστόγραμμα (histogram), γ) πολύγωνο συχνοτήτων (frequency polygon) δ) πίτα συχνοτήτων ή κυκλικό διάγραμμα (pie chart) ε) ραβδόγραμμα (bar chart).

3.1 Διάγραμμα Συχνοτήτων

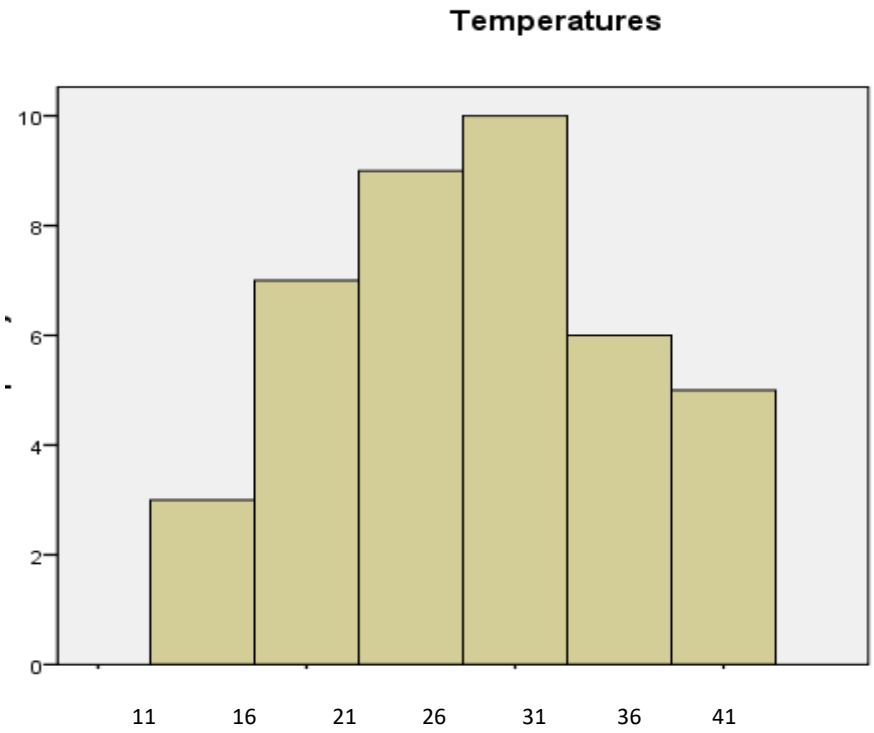
Το διάγραμμα συχνοτήτων είναι ένας τύπος γραφικής παράστασης που χρησιμοποιείται για την απεικόνιση των τιμών μιας ποσοτικής μεταβλητής. Σε

κάθε τιμή x_i (υποθέτουμε οτι οι τιμές x_i είναι ταξινομημένες κατά αύξουσα σειρά) υψώνουμε μια κάθετη γραμμή με μήκος ίσο με την αντίστοιχη συχνότητα. Μπορούμε επίσης αντί των συχνοτήτων ν_i στον κάθετο άξονα να βάλουμε τις σχετικές συχνότητες f_i , οπότεέχουμε διάγραμμα σχετικών συχνοτήτων. Το διάγραμμα συχνοτήτων που ακολουθεί είναι για την μεταβλητή ημερών απουσίας του παραδείγματος 1.



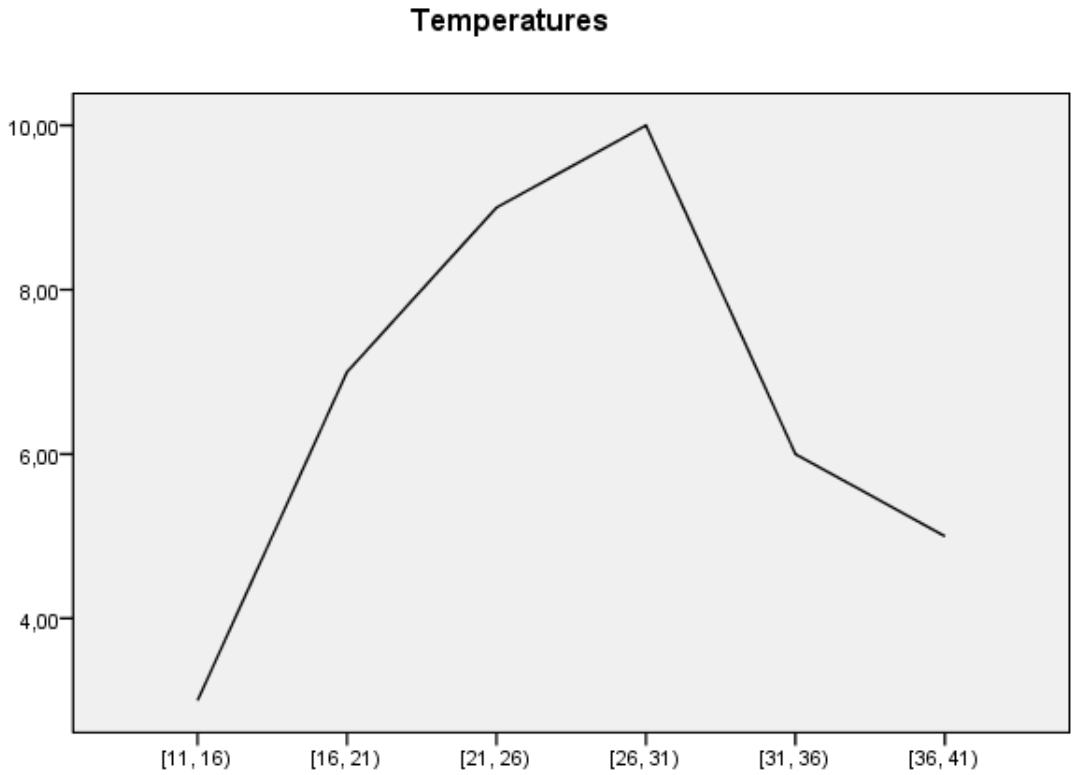
3.2 Ιστόγραμμα

Η γραφική παράσταση ενός πίνακα συχνοτήτων με ομαδοποιημένα δεδομένα γίνεται με το λεγόμενο ιστόγραμμα (histogram) συχνοτήτων. Στον κάθετο άξονα αναγράφονται οι συχνότητες ή οι σχετικές συχνότητες και στον οριζόντιο άξονα η μεταβλητή καθώς και τα όρια των κλάσεων. Η ονομασία του διαγράμματος προέρχεται από τη λέξη ιστός δηλαδή ορθογώνιο. Αποτελείται από διαδοχικά ορθογώνια καθένα από τα οποία έχει βάση ίση με το πλάτος της κλάσης και ύψος ίσο με τη συχνότητα της κλάσης. Το ιστόγραμμα που ακολουθεί είναι για τα δεδομένα του παραδείγματος με τις θερμοκρασίες της μεσογειακής πόλης.



3.3 Πολύγωνο συχνοτήτων

Το πολύγωνο συχνοτήτων προκύπτει με την ένωση των σημείων που αντιστοιχούν στην κεντρική τιμή κάθε κλάσης στο ιστόγραμμα. Στο παρακάτω σχήμα εμφανίζεται το πολύγωνο συχνοτήτων για τα δεδομένα του παραδείγματος με τις θερμοκρασίες.



3.4 Πίτα συχνοτήτων ή Κυκλικό διάγραμμα

Η πίτα συχνοτήτων (pie chart) χρησιμοποιείται και για ποσοτικές (συνεχείς ομαδοποιημένες) και για ποιοτικές μεταβλητές. Το διάγραμμα είναι μία πίτα και κάθε κομάτι της είναι ανάλογο με τη συχνότητας της κλάσης (αν πρόκειται για συνεχή ομαδοποιημένη μεταβλητή) ή της συχνότητας της τιμή της μεταβλητής (αν πρόκειται για ποσοτική κατηγορική η ποιοτική μεταβλητή)

Αν συμβολίσουμε με α_i το αντίστοιχο τόξο ενός κυκλικού τμήματος στο κυκλικό διάγραμμα συχνοτήτων, τότε

$$a_i = \nu_i \cdot \frac{360^0}{n} = 360^0 \cdot f_i$$

Έστω ότι έχουμε την κατανομή συχνοτήτων:

Κλάσεις [-)	Κεντρικές τιμές	Συχν.	Σχετική Συχνότητα	Σχετική συχνότητα ποσοστό
[,)	x_i	ν_i	$f_i = \frac{\nu_i}{n}$	$\alpha_i = \frac{360}{n}\,\nu_i = 360\,f_i$
[11, 16)	13,5	3	0,075	$0{,}075 \cdot 360^0 = 27^0$
[16, 21)	18,5	7	0,175	$0{,}175 \cdot 360^0 = 63^0$
[21, 26)	23,5	9	0,225	$0{,}225 \cdot 360^0 = 81^0$
[26, 31)	28,5	10	0,25	$0{,}25 \cdot 360^0 = 90^0$
[31, 36)	33,5	6	0,15	$0{,}15 \cdot 360^0 = 54^0$
[36, 41)	38,5	5	0,125	$0{,}125 \cdot 360^0 = 45^0$

Το κυκλικό διάγραμμα που προκύπτει από τα δεδομένα του πίνακα, είναι:

3.5 Ραβδόγραμμα

Το ραβδόγραμμα (bar chart) μοιάζει πολύ με το ιστόγραμμα. Οι συχνότητες των τιμών της μεταβλητής παρουσιάζονται και εδώ με ορθογώνια. Η διαφορά είναι ότι χρησιμοποιείται για ποιοτικές μεταβλητές. Για ποιοτικές μεταβλητές (οι τιμές αναγράφονται στον οριζόντιο άξονα) δεν παίζει ρόλο η σειρά των τιμών ούτε το πλάτος των ράβδων.

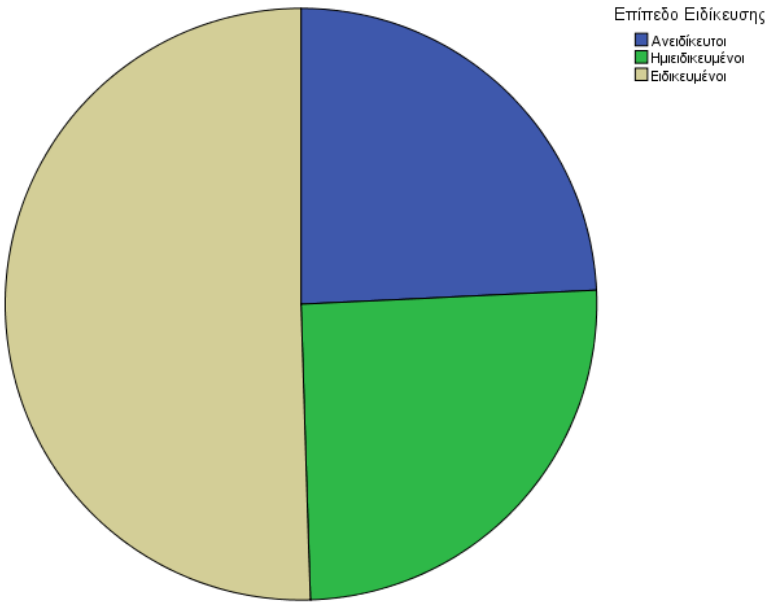
Παράδειγμα 3:

Ο παρακάτω πίνακας δίνει τον αριθμό των εργατών μιας βιομηχανικής επιχείρησης, σύμφωνα με το επίπεδο ειδικεύσεώς τους.

Επίπεδο ειδικεύσεως	Αριθμός εργαζομένων
Ειδικευμένοι	50
Ημιειδικευμένοι	25
Ανειδίκευτοι	25

Να απεικονισθούν τα παραπάνω δεδομένα:
α. Με κυκλικό διάγραμμα
β. Με ραβδόγραμμα

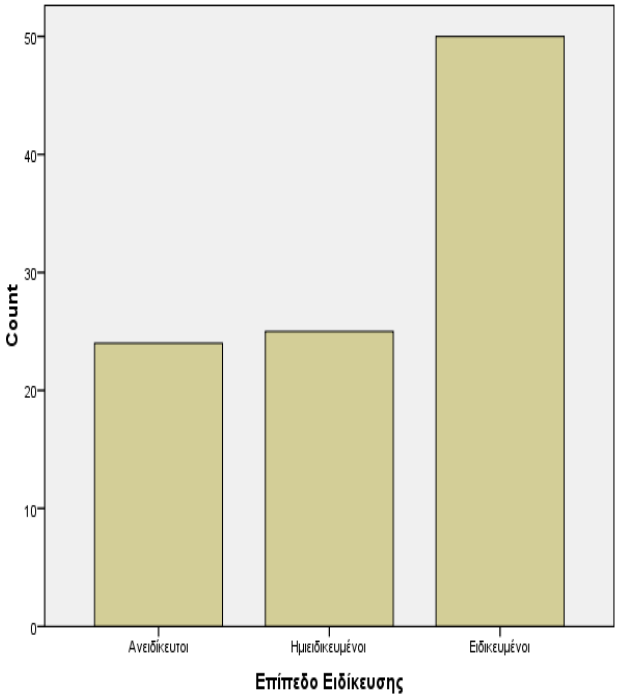
Λύση:



(Για να βρούμε τις μοίρες του κάθε τόξου:

$$\alpha_1 = \frac{\nu_1}{n} \cdot 360 = \frac{50}{100}\ 360 = 180\,, \qquad \alpha_2 = \frac{\nu_2}{n} \cdot 360 = \frac{25}{100}\ 360 = 90, \quad \alpha_3 = 90)$$

β)



Παράδειγμα 4

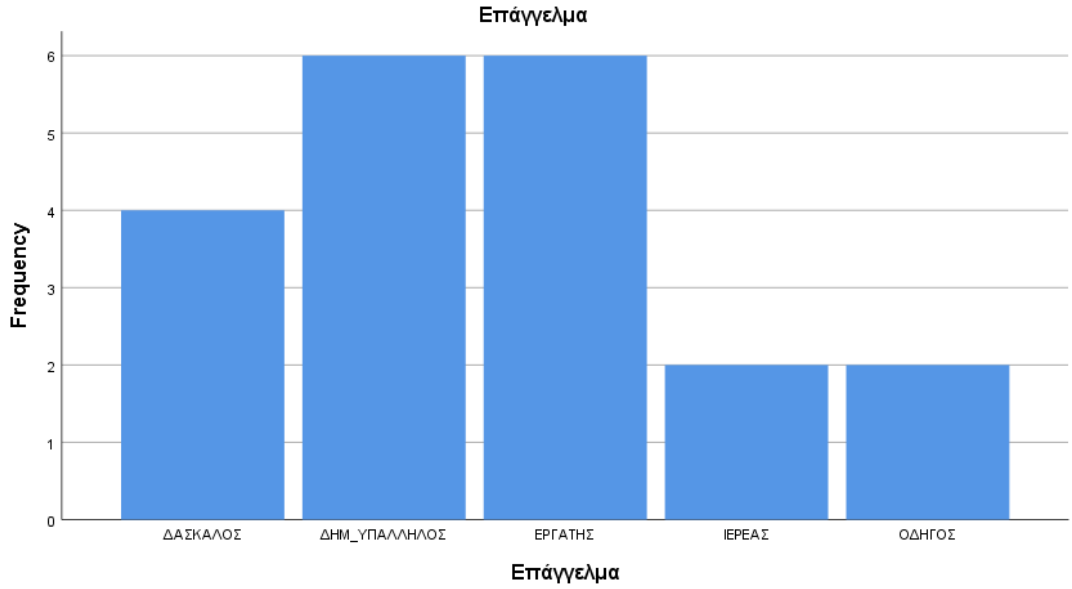
Στον παρακάτω πίνακα παρουσιάζονται τα επαγγέλματα και το ημερομίσθιο του πατέρα, δεδομένα προερχόμενα από 20 οικογένειες μιας περιοχής της Αθήνας.

Επάγγελμα	Ημερομίσθιο
Εργάτης	70
Οδηγός	75
Εργάτης	80
Δημόσιος υπάλληλος	70
Δημόσιος υπάλληλος	80
Δημόσιος υπάλληλος	50
Δάσκαλος	90
Ιερέας	100
Οδηγός	60
Εργάτης	60
Δάσκαλος	70
Εργάτης	60
Εργάτης	80
Δημόσιος υπάλληλος	70
Ιερέας	90
Δάσκαλος	100
Εργάτης	90
Δημόσιος υπάλληλος	65
Δάσκαλος	75
Δημόσιος υπάλληλος	80

Να γίνουν οι πίνακες κατανομής συχνοτήτων για τις δύο μεταβλητές, το ραβδόγρμμα της μεταβλητής επάγγελμα και το κυκλικό διάγραμμμα της μεταβλητής ημερομίσθιο.

Λύση

	Επάγγελμα	ν_i	$f_i\%$	N_i	$F_i\%$
x_i	ΔΑΣΚΑΛΟΣ	4	20	4	20
	ΔΗΜ_ΥΠΑΛΛΗΛΟΣ	6	30	10	50
	ΕΡΓΑΤΗΣ	6	30	16	80
	ΙΕΡΕΑΣ	2	10	18	90
	ΟΔΗΓΟΣ	2	10	20	100
	Total	20	100		



ΗΜΕΡΟΜΙΣΘΙΟ

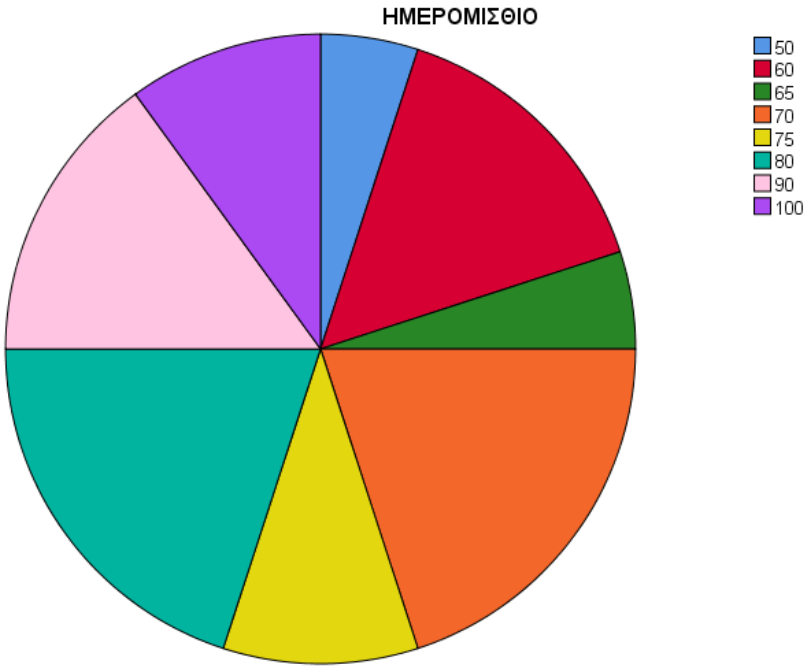
x_i	ν_i	$f_i\%$	N_i	$F_i\%$
50	1	5	1	5
60	3	15	4	20
65	1	5	5	25
70	4	20	9	45
75	2	10	11	55
80	4	20	15	75
90	3	15	18	90
100	2	10	20	100
Σύνολο	20	100		

Για το κυκλικό διάγραμμα :

$\alpha_1 = \frac{1}{20} \cdot 360° = 18°$ $\alpha_2 = \frac{3}{20} \cdot 360° = 54°$ $\alpha_3 = \frac{1}{20} \cdot 360°$

$\alpha_4 = \frac{4}{20} \cdot 360° = 72°$ $\alpha_5 = \frac{2}{20} \cdot 360° = 36°$ $\alpha_6 = \frac{4}{20} \cdot 360° = 72°$

$\alpha_7 = \frac{3}{20} \cdot 360° = 54°$ $\alpha_8 = \frac{2}{20} \cdot 360° = 36°$



4. Μέτρα θέσης και Διασποράς

4.1 Μέτρα θέσης

Τα μέτρα θέσης, είναι κάποια αριθμητικά μεγέθη τα οποία χρησιμοποιούνται για την περιγραφή της θέσης των παρατηρήσεων στον οριζόντιο άξονα. Τα κυριότερα μέτρα θέσης είναι: η μέση τιμή, η διάμεσος, η επικρατούσα τιμή και τα τεταρτημόρια

α**)** **η μέση τιμή**, $\bar{x}$ αποτελεί το χρησιμότερο μέτρο της Στατιστικής και ορίζεται ως το άθροισμα των παρατηρήσεων δια του πλήθους των παρατηρήσεων. Δηλαδή:

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Περιπτώσεις ταξινομημένων ή ομαδοποιημένων δεδομένων:

Αν οι τιμές (κεντρικές τιμές των κλάσεων) $x_1, x_2, \dots, x_k$ μιας μεταβλητής Χ αντιστοιχούν στις συχνότητες $\nu_1, \nu_2, \dots, \nu_k$, ο αριθμητικός μέσος υπολογίζεται από τη σχέση:

$$\bar{x} = \frac{\nu_1 \cdot x_1 + \nu_2 \cdot x_2 + ..\nu_k \cdot x_k}{n} = \frac{\sum_{i=1}^{k} \nu_i \cdot x_i}{\sum_{i=1}^{k} \nu_i}$$

β) **η διάμεσος** δ, ενός δείγματος n παρατηρήσεων οι οποίες έχουν διαταχθεί σε αύξουσα σειρά ορίζεται ως η μεσαία παρατήρηση, όταν το n είναι περιττός αριθμός, ή το ημιάθροισμα των δύο μεσαίων παρατηρήσεων όταν το n είναι άρτιος αριθμός.

Παρατήρηση:
Η διάμεσος δεν επηρεάζεται από παρατηρήσεις οι οποίες βρίσκονται πολύ μακριά από τον κύριο όγκο των δεδομένων. Το αντίθετο συμβαίνει με τον αριθμητικό μέσο του οποίου η τιμή είναι ευαίσθητη σε τέτοιες παρατηρήσεις.

Περίπτωση ομαδοποιημένων δεδομένων

Η διάμεσος ομαδοποιημένων δεδομένων προκύπτει από τον πίνακα συχνοτήτων από τον τύπο:

$$\delta = \alpha_i + \frac{c}{\nu_i}\left(\frac{n}{2} - N_{i-1}\right)$$

όπου:
α_i είναι κατώτερο όριο της κλάσης που ανήκει η διάμεσος ,
c είναι το πλάτος της κλάσης i
ν_i είναι η συχνότητα της κλάσης i
N_{i-1} είναι η αθροιστική συχνότητα της προηγούμενης από την i κλάσης και
n είναι το πλήθος των παρατηρήσεων (μέγεθος του δείγματος).

γ) **η επικρατούσα τιμή ή κορυφή** M_o, ορίζεται ως η παρατήρηση με την μεγαλύτερη συχνότητα.

Περίπτωση ομαδοποιημένων δεδομένων:

Η επικρατούσα τιμή ομαδοποιημένων δεδομένων προκύπτει από τον πίνακα συχνοτήτων από τον τύπο:

$$M_0 = a_i + \frac{c \cdot \Delta_1}{\Delta_1 + \Delta_2}$$

όπου
a_i είναι το κατώτερο όριο της κλάσης που αντιστοιχεί στη μεγαλύτερη συχνότητα
c είναι το πλάτος της κλάσης i,
το Δ_1 υπολογίζεται αν από τη μεγαλύτερη συχνότητα αφαιρέσουμε την συχνότητα της προηγούμενης κλάσης, δηλαδή: $\Delta_1 = \nu_i - \nu_{i-1}$ και
το Δ_2 υπολογίζεται αν από τη μεγαλύτερη συχνότητα αφαιρέσουμε την συχνότητα της επόμενης κλάσης: $\Delta_2 = \nu_i - \nu_{i+1}$.

δ) **Το πρώτο τεταρτημόριο Q_1** διαιρεί τα δείγμα σε δύο μέρη, έτσι ώστε, όταν οι παρατηρήσεις είναι διατεταγμένες σε αύξουσα σειρά, το μέρος με τις μικρότερες παρατηρήσεις να αντιστοιχεί στο 25% των παρατηρήσεων.

Περίπτωση ομαδοποιημένων δεδομένων

Το πρώτο τεταρτημόριο σε ομαδοποιημένα δεδομένα προκύπτει από τον πίνακα συχνοτήτων από τον τύπο:

$$Q_1 = \alpha_i + \frac{c}{\nu_i}\left(\frac{n}{4} - N_{i-1}\right)$$

όπου

α_i είναι κατώτερο όριο της κλάσης που ανήκει το πρώτο τεταρτημόριο ,
c είναι το πλάτος της κλάσης i
ν_i είναι η συχνότητα της κλάσης i
N_{i-1} είναι η αθροιστική συχνότητα της προηγούμενης από την i κλάσης και
n είναι το πλήθος των παρατηρήσεων (μέγεθος του δείγματος).

ε) Το **τρίτο τεταρτημόριο Q_3** διαιρεί τα δείγμα σε δύο μέρη, έτσι ώστε, όταν οι παρατηρήσεις είναι διατεταγμένες σε αύξουσα σειρά, το μέρος με τις μικρότερες παρατηρήσεις να αντιστοιχεί στο 75% των παρατηρήσεων.

Περίπτωση ομαδοποιημένων δεδομένων

Το τρίτο τεταρτημόριο σε ομαδοποιημένα δεδομένα προκύπτει από τον πίνακα συχνοτήτων από τον τύπο:

$$Q_3 = \alpha_i + \frac{c}{\nu_i}\left(\frac{3n}{4} - N_{i-1}\right)$$

4.2 Μέτρα Διασποράς

Μέτρα διασποράς ονομάζονται τα μέτρα που εκφράζουν τις αποκλίσεις των τιμών μιας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης. Τα κυριότερα μέτρα διασποράς είναι:

α) **Εύρος:**

R=Μεγαλύτερη τιμή – Μικρότερη τιμή

β)**Ενδοτεταρτημοριακό εύρος**:

$$Q = Q_3 - Q_1$$

όπου Q_3 , Q_1 το 3ο και 1ο τεταρτημόριο αντίστοιχα.

γ) **Διακύμανση:**

Διακρίνουμε και εδώ 2 περιπτώσεις: Μη ομαδοποιημένων και ομαδοποιημένων δεδομένων.

Αν οι τιμές της μεταβλητής X είναι $x_1, x_2, \dots, x_n$, η διακύμανση (Variance) υπολογίζεται από τη σχέση:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Ο τύπος αυτός μετά από πράξεις γίνεται:

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

Περίπτωση ομαδοποιημένων δεδομένων

Αν οι τιμές (κεντρικές τιμές των κλάσεων) $x_1, x_2, \dots, x_k$ μιας μεταβλητής Χ αντιστοιχούν στις συχνότητες $\nu_1, \nu_2, \dots, \nu_k$, για τη διακύμανση ισχύουν οι παρακάτω τύποι:

$$s^2 = \frac{\sum_{i=1}^{k} \nu_i \cdot (x_i - \bar{x})^2}{n-1}$$

$$s^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{k} \nu_i \cdot x_i^2 - \frac{(\sum_{i=1}^{k} \nu_i \cdot x_i)^2}{n} \right)$$

δ) **Τυπική απόκλιση:**

Η διακύμανση εκφράζεται σε μονάδες που αντιστοιχούν στα τετράγωνο των αρχικών μονάδων. Η ανάγκη να για ένα μέτρο διασποράς που να εκφράζεται στις ίδιες μονάδες με τις αρχικές (ούτως ώστε να μπορεί να συνεκτιμάται σε συνδυασμό και με τη μέση τιμή) οδήγησε στην χρησιμοποίηση της τετραγωνικής ρίζας της διακύμανσης η οποία ονομάζεται τυπική απόκλιση:

$$s = \sqrt{s^2}$$

4.3 Συντελεστής Μεταβλητότητας

Ας θεωρήσουμε τους μηνιαίους μισθούς, σε ευρώ, πέντε υπαλλήλων δύο εταιριών Α και Β. Για την Α έχουμε τους μισθούς 10100, 10050, 10020, 10000, 10010 και για την Β έχουμε τους μισθούς 1100, 1050, 1020, 1000, 1010. Παρατηρούμε ότι και οι δύο εταιρίες έχουν ίδια μέτρα διασποράς (π.χ τυπική απόκλιση, εύρος κ.λπ). Παρόλα αυτά αν κάποιος υπάλληλος της πρώτης εταιρίας υποστεί μείωση μισθού 1000 ευρώ τότε αυτό θα έχει για αυτόν πολύ μικρότερες συνέπειες από ότι μια αντίστοιχη μείωση στο μισθό ενός υπαλλήλου της εταιρίας Β.

Από το παραπάνω παράδειγμα φαίνεται η ανάγκη να οριστεί ένα καινούργιο μέτρο το οποίο δεν θα αντικατοπτρίζει μόνο τη διασπορά των δεδομένων, αλλά και τις επιπτώσεις που έχει αυτή η διασπορά στην πειραματική μονάδα. Το μέτρο αυτό συμβολίζεται με CV, ονομάζεται Συντελεστής Μεταβλητότητας και δίνεται από τον λόγο της τυπικής απόκλισης (s) προς την μέση τιμή. Δηλαδή έχουμε τον τύπο:

$$CV = \frac{s}{|\bar{x}|}$$

Εάν η μέση τιμή είναι κοντά στο μηδέν ή πολύ μεγάλη, τότε ο συντελεστής μεταβλητότητας καθίσταται αναξιόπιστος.

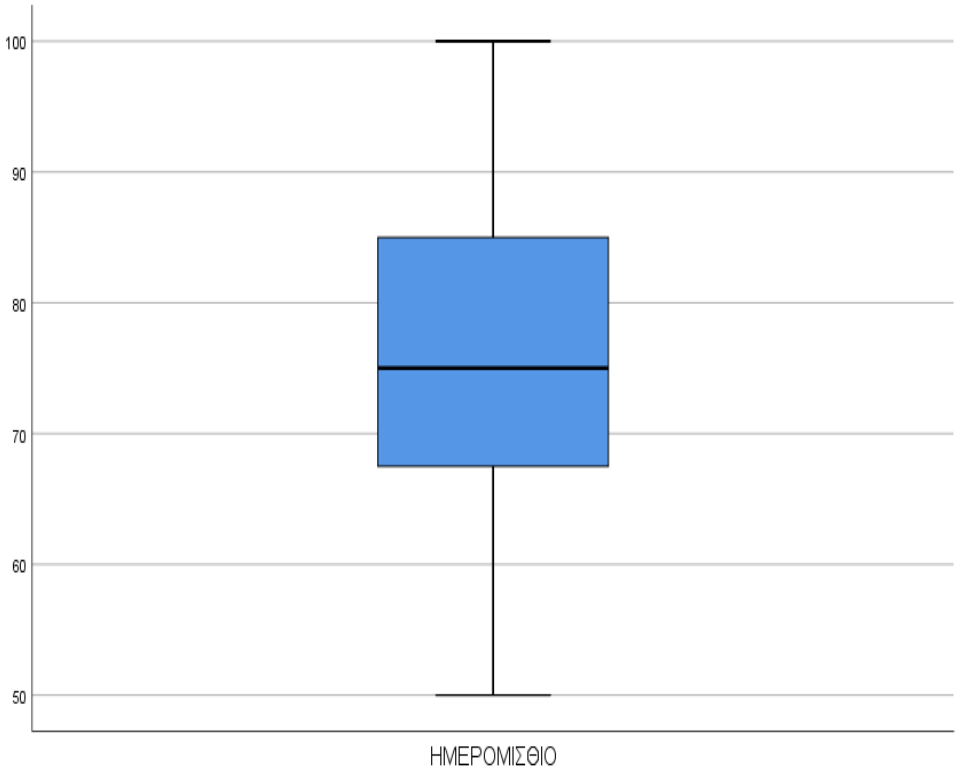
Ο συντελεστής μεταβλητότητας είναι ανεξάρτητος από τις μονάδες μέτρησης, εκφράζεται επί τοις εκατό και εκφράζει τη μεταβλητότητα των δεδομένων απαλλαγμένη από την επίδραση της μέσης τιμής.

Όσο μικρότερη είναι η τιμή του CV, τόσο μεγαλύτερη ομοιογένεια θεωρείται ότι έχει το δείγμα. Γενικά, δεχόμαστε ότι ένα δείγμα τιμών μιας μεταβλητής είναι «***ομοιογενές***», εάν ο συντελεστής μεταβλητότητας δεν ξεπερνά το 10%.

4.4 Θηκόγραμμα (Box-plot)

Το θηκόγραμμα είναι μια γραφική παράσταση της κατανομής των παρατηρήσεων, που για την κατασκευή του χρειάζεται ο υπολογισμός της διαμέσου δ, καθώς και του πρώτου και τρίτου τεταρτημορίου (Q_1 , Q_3). Αποτελείται από ένα ευθύγραμμο τμήμα με αρχή την ελάχιστη και τέλος την μέγιστη τιμή του δείγματος, και ανάμεσα τοποθετείται ένα ορθογώνιο με κάτω βάση που αντιστοιχεί στην τιμή Q_1 και άνω βάση που αντιστοιχεί στην τιμή Q_3. Ενδιάμεσα τοποθετείται ένα ευθύγραμμο τμήμα που αντιστοιχεί στην διάμεσο.

Παρακάτω βλέπετε το θηκόγραμμα των τιμών της μεταβλητής ημερομίσθιο του παραδείγματος 4:



Είναι φανερό οτι από την ελάχιστη και τη μέγιστη τιμή μπορούμε να βρούμε το εύρος και από το πρώτο και τρίτο τεταρτημόριο το ενδοτεταρτημοριακό εύρος. Το ενδιάμεσο ορθογώνιο περιγράφει το διάστημα το οποίο περιέχει τις 50% μεσαίες τιμές του δείγματος.

Θα πρέπει να σημειώσουμε οτι αν υπάρχουν τιμές που είναι είτε 'πολύ μεγαλύτερες' είτε 'πολύ μικρότερες' απο τις υπόλοιπες τιμές του δείγματος, δηλαδή αν υπάρχουν «ακραίες» τιμές, τότε το ευθύγραμμο τμήμα δεν έχει όρια την ελάχιστη και την μέγιστη τιμή όπως αναφέρθηκε παραπάνω. Τα όρια στο θηκόγραμμα υπολογίζονται από τους τύπους;

Κάτω άκρο: $\max\{$ ελάχιστη τιμή , $Q_1 - 1{,}5(Q_3 - Q_1)\}$

Άνω άκρο: $\min\{$ μέγιστη τιμή , $Q_3 + 1{,}5(Q_3 - Q_1)\}$

Παράδειγμα 5:

Στο παρακάτω δείγμα να βρεθούν όλα τα μέτρα θέσης και διασποράς και να εξεταστεί ως προς την ομοιογένεια:

$$3,\ 4,\ 0,\ 6,\ 5,\ 8,\ 1,\ 1,\ 6,\ 1,\ 2,\ 7,\ 8$$

Λύση:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{52}{13} = 4$$

Διατάσουμε τις παρατηρήσεις σε αύξουσα σειρά:

$$0,\ 1,\ 1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 6,\ 7,\ 8,\ 8$$

$\delta = x_7 = 4$

$M_o=1$

$Q_1 = \frac{x_3 + x_4}{2} = \frac{1+1}{2} = 1$

$Q_2 = \delta = 4$

$Q_3 = \frac{x_{10} + x_{11}}{2} = \frac{6+7}{2} = 6{,}5$

$R = 8 - 0 = 8$

$s^2 = \frac{\sum_{i=1}^{13}(x_i - \bar{x})^2}{12}$

$= \frac{(0-4)^2 + (1-4)^2 \cdot 3 + (2-4)^2 + (3-4)^2 + (5-4)^2 + (6-4)^2 \cdot 2 + (7-4)^2 + (8-4)^2 2}{12}$

$= \frac{98}{12} = 8{,}17$

$s = \sqrt{8{,}17} = 2{,}86$

$CV = \frac{s}{|\bar{x}|} = \frac{2{,}86}{4} = 0{,}715$ ή 71.5%. Άρα το δείγμα δεν είναι ομοιογενές.

Παράδειγμα 6:

Στο παράδειγμα 1 (αριθμός απουσιών εργαζομένων) να υπολογιστούν:

i) μέση τιμή $\bar{x}$

ii) διάμεσος δ

iii) κορυφή M_o

iv) τυπική απόκλιση s

v) CV

Λύση:

Για την μέση τιμή κ διακύμανση είναι χρήσιμος ο πίνακας:

x_i	ν_i	$x_i\nu_i$	x_i^2	$x_i^2\nu_i$
1	2	2	1	2
2	2	4	4	8
3	2	6	9	18
4	1	4	16	16
5	3	15	25	75
		31		119

i) $\bar{x} = \frac{\sum_{i=1}^{k} \nu_i \cdot x_i}{\sum_{i=1}^{k} \nu_i} = \frac{31}{10}$

ii) είναι n=10 άρα $\delta = \frac{x_5 + x_6}{2} = \frac{3+3}{2} = 3$

iii) $M_o = 5$

iv) $s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{k} x_i^2\nu_i - \frac{\left(\sum_{i=1}^{k} x_i\nu_i\right)^2}{n}\right\} = \frac{1}{9}\left\{119 - \frac{(31)^2}{10}\right\} = 2{,}54$

άρα $s = \sqrt{2{,}54} = 1{,}594$

v) $CV = \frac{1{,}594}{3{,}1} = 0{,}514$

Παράδειγμα 7:

Για την μεταβλητή «ημερομίσθιο» του παραδείγματος 4, να υπολογιστούν:

i) μέση τιμή $\bar{x}$

ii) διάμεσος δ

iii) τα τεταρτημόρια Q_1, Q_2, Q_3

iv) κορυφή M_o

v) τυπική απόκλιση s

vi) CV

Λύση:

x_i	ν_i	$f_i\%$	N_i	$F_i\%$	$x_i \cdot \nu_i$	x_i^2	$x_i^2 \nu_i$
50	1	5	1	5	50	2500	2500
60	3	15	4	20	180	3600	10800
65	1	5	5	25	65	4225	4225
70	4	20	9	45	280	4900	19600
75	2	10	11	55	150	5625	11250
80	4	20	15	75	320	6400	25600
90	3	15	18	90	270	8100	24300
100	2	10	20	100	200	10000	20000
Σύνολο	20	100			1515		118275

i) $\bar{x} = \frac{\sum_{i=1}^{k} \nu_i \cdot x_i}{n} = \frac{1515}{20} = 75{,}75$

ii) είναι n=20, άρα $\delta = \frac{x_{10}+x_{11}}{2} = \frac{75+75}{2} = 75$

iii) Το Q_2 ταυτίζεται με τη διάμεσο άρα $Q_2 = 75$

n=20. Χωρίζουμε το δείγμα σε 2 ίσα τμήματα των 10 παρατηρήσεων. Το Q_1 είναι η διάμεσος του 1ου τμήματος, δηλαδή $Q_1 = \frac{x_5+x_6}{2} = \frac{65+70}{2} = 67{,}5$ και το Q_3 του 2ου τμήματος: $Q_3 = \frac{80+90}{2} = 85$

iv) έχουμε δύο κορυφές: $M_{o(1)} = 70$ και $M_{o(2)} = 80$

v) $s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{k} x_i^2 \nu_i - \frac{\left(\sum_{i=1}^{k} x_i \nu_i\right)^2}{n}\right\} = \frac{1}{19}\left\{118275 - \frac{(1515)^2}{20}\right\} = 184{,}934$

άρα $s = \sqrt{184{,}934} = 13{,}599$

vi) $CV = \frac{13{,}599}{75{,}75} = 0{,}1795$

Παράδειγμα 8:

Να βρεθούν τα κυριότερα μέτρα θέσεως και διασποράς για τα δεδομένα των θερμοκρασιών της μεσογειακής πόλης του παραδείγματος 2.

Λύση:

Κλάσεις [,)	x_i	x_i^2	ν_i	$\nu_i \cdot x_i$	$\nu_i \cdot x_i^2$
[11, 16)	13,5	182,25	3	40,5	546,75
[16, 21)	18,5	342,25	7	129,5	2395,75
[21, 26)	23,5	552,25	9	211,5	4970,25
[26, 31)	28,5	812,25	10	285	8122,5
[31, 36)	33,5	1122,25	6	201	6733,5
[36, 41)	38,5	1482,25	5	192,5	7411,25
	156	4493,5		1060	30180

i)) Μέτρα θέσεως:

Εφαρμόζοντας τον τύπο της μέσης τιμής έχουμε $\bar{x} = \frac{\sum_{i=1}^{k} \nu_i \cdot x_i}{n}$ =26,5

Η διάμεσος θα βρεθεί από τον τύπο

$$\delta = \alpha_i + \frac{c}{\nu_i}\left(\frac{n}{2} - N_{i-1}\right)$$

Παρατηρούμε ότι η διάμεσος ανήκει στην κλάση [26, 31) άρα α_i=26 (το κατώτερο όριο της κλάσης που ανήκει η διάμεσος). Ακόμα το πλάτος της κλάσης i, όπως και όλων των κλάσεων είναι 5, η συχνότητα της κλάσης ν_i είναι ίση με 10, το πλήθος των παρατηρήσεων είναι n=40 και η αθροιστική συχνότητα της προηγούμενης από την i κλάσης είναι ίση με 19. Επομένως έχουμε:

$$\delta = 26 + \frac{5}{10}(\frac{40}{2} - 19) \Leftrightarrow \delta = 26{,}5$$

Η επικρατούσα τιμή θα βρεθεί από τον τύπο $M_0 = a_i + \frac{c \cdot \Delta_1}{\Delta_1 + \Delta_2}$

όπου α_i είναι το κατώτερο όριο της κλάσης που αντιστοιχεί στη μεγαλύτερη συχνότητα δηλαδή στην [26, 31). Επομένως α_i=26 και όπως και πριν c=5. Ακόμα Δ_1= 10-9=1 και Δ_2=10-6=4. Επομένως

$$M_0 = a_i + \frac{c \cdot \Delta_1}{\Delta_1 + \Delta_2} \Leftrightarrow M_0 = 26 + \frac{5 \cdot 1}{4 + 1} \Leftrightarrow M_0 = 27$$

Το 1ο τεταρτημόριο θα είναι η 10η τιμή. Επομένως η κλάση που περιέχει το Q_1 είναι η 2η. Άρα:

$$Q_1 = \alpha_i + \frac{c}{\nu_i}\left(\frac{n}{4} - N_{i-1}\right) = 16 + \frac{5}{7}\left(\frac{40}{4} - 3\right) = 21$$

Το 3ο τεταρτημόριο βρίσκεται στην 5η κλάση (η 30η παρατήρηση) άρα:

$$Q_3 = \alpha_i + \frac{c}{\nu_i}\left(\frac{3n}{4} - N_{i-1}\right) = 31 + \frac{5}{6}\left(\frac{120}{4} - 29\right) = 31{,}83$$

ii) Μέτρα διασποράς:

Η μεγαλύτερη κεντρική τιμή είναι το 38,5 και η μικρότερη το 13,5 οπότε το εύρος είναι 25.

Για ευκολία ο υπολογισμός της διακύμανσης γίνεται με τη χρήση του παραπάνω πίνακα. Εφαρμόζοντας τον τύπο παίρνουμε:

$$s^2 = \frac{1}{n-1}\cdot\left(\sum_{i=1}^{k} \nu_i \cdot x_i^2 - \frac{\left(\sum_{i=1}^{k} \nu_i \cdot x_i\right)^2}{n}\right) \Leftrightarrow s^2 = \frac{1}{39}\cdot\left(30180 - \frac{(1060)^2}{40}\right) \Leftrightarrow s^2 = 53{,}59$$

Ενώ για την τυπική απόκλιση:

$$s = \sqrt{53{,}59} = 7{,}32$$

ΑΣΚΗΣΕΙΣ

1) Οι ημέρες νοσηλείας 10 ασθενών ενός νοσοκομείου είναι:

8 9 8 6 7 8 7 7 5 5

Να βρεθούν η μέση τιμή, η διασπορά, και η διάμεσος του δείγματος των ημερών νοσηλείας.

2) Η βαθμολογία 20 φοιτητών στις εξετάσεις ενός μαθήματος είναι

8 9 8 6 7 8 7 7 8 7
6 8 9 8 7 9 8 9 7 8

i) Να γίνει ο πίνανακας κατανομής συχνοτητων.
ii) Να σχεδιαστεί το κυκλικό διάγραμμα.
iii) Να βρεθούν η μέση τιμή, η διάμεσος, το 1ο και 3 τεταρτημόριο του δείγματος των βαθμολογιών.
iv) Να εξεταστεί αν το δείγμα είναι ομοιογενές.

3) Στον πίνακα έχουμε το βάρος από 60 ροδάκινα σε γραμμάρια. Να υπολογιστούν:

i) Η επικρατούσα τιμή, η διάμεσος
ii) Ο συντελεστης μεταβλητότητας.

Βάρος(γρ) Ροδάκινων	Ροδάκινα
[100,120)	6
[120,140)	12
[140,160)	18
[160,180)	16
[180,200)	8

4) Ο αριθμός απουσιών 160 φοιτητών, φαίνονται στον πίνακα.
α) Να βρεθούν:
i) Ο μέσος αριθμός απουσιών.
ii) Η τυπική απόκλιση του δείγματος.
β) Να εξετάσετε αν το δείγμα είναι ομοιογενές.

Απουσίες	Άτομα
[7,11)	24
[11,15)	40
[15,19)	48
[19,23)	32
[23,27)	16

5) Η μηνιαία χρήση ενός συγκεκριμένου δρομολογίου πλοίου από 20 επιβάτες ήταν:

1	2	1	10	1	2	3	6	2	2
10	6	1	6	2	2	1	2	2	3

i) Να κατασκευαστεί ο πίνακας κατανομής συχνοτήτων.
ii) Να βρεθεί η μέση τιμή, η διάμεσος, η επικρατούσα τιμή και η τυπική απόκλιση.
iii) Να εξεταστεί το δείγμα ως προς την ομοιογένεια.

6) Να συμπληρωθεί ο παρακάτω πίνακας κατανομών.

x_i	ν_i	f_i	N_i	F_i
1	3	0,03		
2			15	

3			25	
4				0,65
5	10			
6				
ΣΥΝΟΛΟ				

7) Ζυγίστηκαν 30 αθλητές και τα βάρη του σε κιλά ήταν:

55	70	69	73	72	59	54	71	67	62
60	54	63	52	80	73	74	70	63	64
65	58	53	45	56	50	48	57	60	62

i) Να ομαδοποιηθούν οι τιμές σε 6 κλάσεις.
ii) Να βρεθούν η διάμεσος και η επικρατούσα τιμή.
iii) Να εξεταστεί το δείγμα ως προς την ομοιογένεια.

8) Η αρτηριακή πίεση 40 ασθενών που νοσηλεύονται στο χειρουργικό τμήμα ενός νοσοκομείου είναι

10	10	13	13	15	17	17	19	19	17
17	10	19	13	13	15	13	15	15	17
17	13	19	13	19	13	13	19	19	19
15	15	19	15	19	15	15	19	19	17

i) Να κατασκευαστεί ο πίνακας κατανομής συχνοτήτων
ii) Να εξεταστεί το δείγμα ως προς την ομοιογένεια.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1) Βιοστατιστική και Εφαρμογές, Έφη Παπαγεωργίου, Εκδοσεις Νέων Τεχνολογιών.

2) Επαγωγική Στατιστική, Μιλτιάδης Χαλικιάς, Συγχρονη Εκδοτική.

3) Στατιστική Επιχειρήσεων με Εφαρμογές σε SPSS και LISRER, Ευστάθιος Δημητριάδης, Εκδόσεις Κριτική.