Τύπος των δεδομένων

Η στατιστική επεξεργασία ενός συνόλου δεδομένων εξαρτάται από το είδος των δεδομένων και των μεταβλητών που τα απαρτίζουν. Τυχαία μεταβλητή είναι κάθε μετασχηματισμός που απεικονίζει κάθε ενδεχόμενο ενός πειράματος τύχης σε μια μέτρηση ή με απλά λόγια κάθε γνώρισμα ή ιδιότητα που χαρακτηρίζει κάποιο μέλος του πληθυσμού που μελετάμε και διαφέρει από μέλος σε μέλος. Ο κυριότερος διαχωρισμός των τυχαίων μεταβλητών είναι σε ποιοτικές (κατηγορικές και διάταξης) και ποσοτικές (συνεχείς και διακριτές) μεταβλητές.

Μια μεταβλητή που οι τιμές της είναι στην ουσία παρατηρήσεις οι οποίες κατανέμονται σε κατηγορίες ανάλογα με κάποια ποιοτικά χαρακτηριστικά, όπως το φύλλο, η οικογενειακή κατάσταση, η εθνικότητα κ.α., λέγεται κατηγορική μεταβλητή. Μια κατηγορική μεταβλητή της οποίας οι απαντήσεις παρουσιάζουν κάποια διάταξη (π.χ. με ερώτηση όπου ο ερωτώμενος εκφ΄ραζει μια προτίμηση με απαντήσεις του τύπου διαφωνώ πολύ – διαφωνώ – συμφωνώ – συμφωνό πολύ) λέγεται μεταβλητή διάταξης (ordinal).

Μια μεταβλητή που οι τιμές της είναι μετρήσεις οι οποίες εκφράζουν μια ποσότητα λέγεται ποσοτική μεταβλητή. Παραδείγματα ποσοτικών μεταβλητών είναι το ύψος, το βάρος, η ηλικία, η θερμοκρασία, το εισόδημα, ο αριθμός των παιδιών που έχει κάποιος κ.α

Οι ποσοτικές μεταβλητές χωρίζονται σε συνεχείς όπου το σύνολο των δυνατών τιμών είναι ένα συνεχές υποσύνολο των πραγματικών αριθμών όπως το ύψος, το βάρος κ.α. και σε διακριτές όπου το σύνολο τιμών αποτελείται από συγκεκριμμένες τιμές (συνήθως ακεραίους) π.χ. το πόσα αδέρφια έχει κάποιος κ.α.

Υπάρχουν τέσσερις κλίμακες μέτρησης των μεταβλητών : Ονομαστική κλίμακα (αναφέρεται σε κατηγορικές μεταβλητές), διατεταγμένη κλίμακα (αναφέρεται σε κατηγορικές διατεταγμένες μεταβλητές), κλίμακα διαστήματος και κλίμακα λόγου.

α) Ονομαστική κλίμακα (nominal scale) : Μια μεταβλητή είναι σε ονομαστική κλίμακα όταν τα στοιχεία της είναι απλώς ετικέτες οι οποίες προσδίδουν μια ιδιότητα σε κάθε μέλος του δείγματος. Παραδείγματος χάριν σε μια ερώτηση που διερευνά το φύλλο του ερωτώμενου υπάρχουν δυο δυνατές απαντήσεις “Άνδρας” ή “Γυναίκα”. Αυτές οι απαντήσεις μπορούν, για χάριν συντομίας αλλά και για να είναι δυνατή η αναγνώριση τους από τον υπολογιστή, να κωδικοποιηθούν με αριθμούς 1 για το ένα φύλλο και 0 για το άλλο. Οι αριθμοί αυτοί είναι απλώς ετικέτες και αντί αυτών θα μπορούσαμε να χρησιμοποιήσουμε οποιουσδήποτε άλλους αριθμούς (π.χ. 3234 και 8). Παραδείγματα μεταβλητών που είναι σε ονομαστική κλίμακα είναι τα παρακάτω.

- Οικογενειακή κατάσταση
- Θρήσκευμα
- Υπηκοότητα

Καμία αριθμητική πράξη (πρόσθεση, αφαίρεση, πολλαπλασιασμός, διαίρεση) δεν μπορεί να εφαρμοσθεί σε δεδομένα που βρίσκονται σε ονομαστική κλίμακα. Η μόνη πράξη που έχει νόημα για τα δεδομένα αυτά είναι, όπως θα δούμε και στη συνέχεια, ο υπολογισμός συχνοτήτων (πόσες φορές δίνεται μια συγκεκριμένη απάντηση) και ο υπολογισμός σχετικών συχνοτήτων (ποσοστών).

β) Διατεταγμένη κλίμακα (ordinal scale) : Οι μεταβλητές που βρίσκονται σε διατεταγμένη κλίμακα παρουσιάζουν τις ίδιες ιδιότητες με αυτές που βρίσκονται σε ονομαστική κλίμακα με τη διαφορά ότι οι δυνατές τιμές της μεταβλητής παρουσιάζουν

κάποια διάταξη π.χ. η ερώτηση " Συμφωνείτε με την τάδε ενέργεια της κυβέρνησης " η οποία έχει τις παρακάτω 5 πιθανές απαντήσεις είναι σε διατεταγμένη κλίμακα

α) Διαφωνώ πολύ
β) Διαφωνώ
γ) Ούτε διαφωνώ ούτε συμφωνώ
δ) Συμφωνώ
ε) Συμφωνώ πολύ

Είναι φανερό πως οι πιθανές απαντήσεις έχουν κάποια διάταξη και για την κωδικοποίηση μια τέτοιας μεταβλητής μπορούμε να χρησιμοποιήσουμε οποιουσδήποτε αριθμούς λαμβάνουν υπ'όψη τους
αυτή τη διάταξη π.χ.

1=Διαφωνώ πολύ	1=Διαφωνώ πολύ
2=Διαφωνώ	4=Διαφωνώ
3=Ούτε διαφωνώ ούτε συμφωνώ ή	8=Ούτε διαφωνώ ούτε συμφωνώ
4=Συμφωνώ	44=Συμφωνώ
5=Συμφωνώ πολύ	510=Συμφωνώ πολύ

Για λόγους ευκολίας χρησιμοποιείται η πρώτη κωδικοποίηση. Στη δεύτερη κωδικοποίηση οι αριθμοί είναι απλά ετικέτες που διατηρούν τη διάταξη των δεδομένων. Δεν σημαίνει ότι ο αριθμός 4 της δεύτερης πιθανής απάντησης είναι τέσσερις φορές μεγαλύτερος από τον αριθμό 1 της πρώτης πιθανής απάντησης και δυο φορές μικρότερος από τον αριθμό 8 της τρίτης πιθανής απάντησης.

γ) Κλίμακα διαστήματος (interval scale) : Οι μεταβλητές που βρίσκονται σε κλίμακα διαστήματος έχουν όλες τις ιδιότητες των μεταβλητών που βρίσκονται σε διατεταγμένη κλίμακα και επιπλέον το διάστημα ανάμεσα σε δυο τιμές δείχνει πόσο διαφέρει η μια τιμή από την άλλη. Επομένως το διάστημα ανάμεσα σε δυο τιμές έχει σημασία όχι όμως και ο λόγος μεταξύ των δυο αυτών τιμών. Το πιο κλασσικό παράδειγμα δεδομένων που βρίσκονται σε κλίμακα διαστήματος είναι η θερμοκρασία. Η Θερμοκρασία 30 βαθμών Κέλσιου δεν είναι 2 φορές μεγαλύτερη από τη θερμοκρασία 15 βαθμών Κελσίου. Αν αυτό ίσχυε, δηλαδή αν ο λόγος μεταξύ δυο τιμών είχε ουσιαστική σημασία τότε θα έπρεπε να πούμε ότι η θερμοκρασία 30 βαθμών Κελσίου είναι –2 φορές μεγαλύτερη από τη θερμοκρασία –15 βαθμών Κελσίου. Επίσης ο λόγος μεταξύ δυο θερμοκρασιών εκ της οποίας η μια θα ήταν 0 βαθμοί Κελσίου θα ήταν ή 0 ή άπειρος. Η μόνη σχέση μεταξύ των διαφόρων θερμοκρασιών είναι το διάστημα που υπάρχει μεταξύ τους και επομένως είναι σε κλίμακα διαστήματος. Τα δεδομένα που είναι σε κλίμακα διαστήματος μπορούμε να τα προσθέσουμε, να τα αφαιρέσουμε, να τα πολλαπλασιάσουμε ή να τα διαιρέσουμε με οποιονδήποτε αριθμό. Αυτό όμως που πρέπει να κοιτάμε πάντα στο τέλος είναι το διάστημα ανάμεσα στις διάφορες τιμές της μεταβλητής.

δ) Κλίμακα λόγου (Ratio scale) : Οι μεταβλητές που βρίσκονται σε κλίμακα λόγου έχουν όλες τις ιδιότητες των μεταβλητών που βρίσκονται σε κλίμακα διαστήματος και επιπλέον ο λόγος μεταξύ των διαφόρων τιμών έχει σημασία. Τα ύψος, το βάρος, η βαθμολογία, ο χρόνος, το μήκος, το κόστος, η αμοιβή, η ηλικία αποτελούν παραδείγματα δεδομένων που βρίσκονται σε κλίμακα λόγου. Ένα κύριο χαρακτηριστικό αυτών των δεδομένων είναι ότι μια μηδενική τιμή δηλώνει την απουσία τιμής σε εκείνο το στοιχείο π.χ. μηδενικό ύψος δεν υπάρχει.

ΠΟΙΟΤΙΚΑ ΔΕΔΟΜΕΝΑ

Η μόνη δυνατή επεξεργασία των ποιοτικών δεδομένων είναι η μέτρηση των παρατηρήσεων που εμπίπτουν σε κάθε κατηγορία (συχνότητα) και το ποσοστό των παρατηρήσεων που αναλογεί σε κάθε κατηγορία (σχετική συχνότητα).

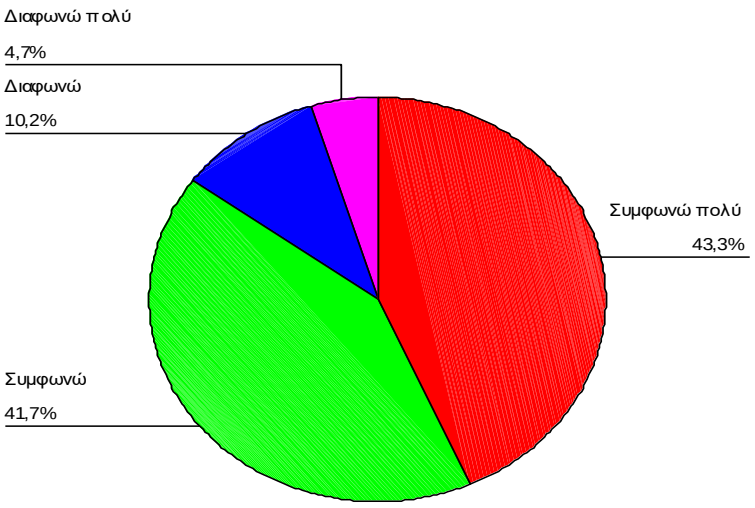
Συχνότητα και σχετική συχνότητα (frequency and relative frequency): Υποθέτουμε ότι τα δεδομένα μας είναι ποιοτικά και έχουμε Ν μονάδες που ταξινομούνται στις διάφορες διακριτές μονάδες της ποιοτικής μεταβλητής (κλάσεις). Για να βρούμε τη συχνότητα των δεδομένων μετράμε τον αριθμό των μονάδων που περιέχονται σε κάθε κλάση. Η σχετική συχνότητα της κλάσης ισούται με τη συχνότητα της κλάσης διαιρεμένης με τον αριθμό των παρατηρήσεων Ν. Ο καλύτερος τρόπος για να παραστήσει γραφικά κανείς ποιοτικά δεδομένα είναι είτε με τη χρησιμοποίηση ενός διαγράμματος πίτας είτε με τη χρησιμοποίηση ενός ραβδογράμματος.

Κυκλικό διάγραμμα – διάγραμμα πίτας (Pie chart) : Τα κυκλικό διάγραμμα είναι ένας κύκλος ο οποίος υποδιαιρείται σε κυκλικούς τομείς που αντιπροσωπεύουν τις διάφορες κατηγορίες ανάλογα με τις συχνότητες με τις οποίες παρατηρούνται. Είναι από τα πιο παραστατικά διαγράμματα για ποιοτικά δεδομένα ειδικά όταν ο αριθμός των κατηγοριών είναι μικρός.

Παράδειγμα : Στα πλαίσια μιας έρευνας στη Μεγάλη Βρετανία στην ερώτηση “Πιστεύεται ότι η κυβέρνηση πρέπει να ελέγχει τις τιμές των προϊόντων επιβάλλοντας ανώτερες και κατώτερες τιμές ” οι απαντήσεις που δόθηκαν από τα 822 άτομα συνοψίζονται στον πίνακα που ακολουθεί.

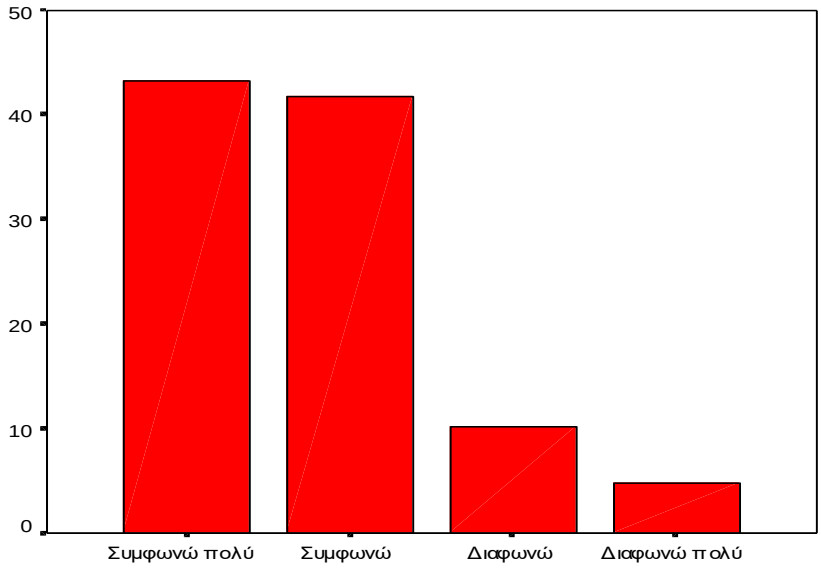
Απαντήσεις	*Συχνότητα*	*Σχετική συχνότητα (Ποσοστό %)*
Συμφωνώ πολύ	356	43.3
Συμφωνώ	343	41.7
Διαφωνώ	84	10.2
Διαφωνώ πολύ	39	4.7

Το διάγραμμα που ακολουθεί αποτυπώνει τα αποτελέσματα του πίνακα.



Ραβδογράμματα (Bar charts) : Στα ραβδογράμματα έχουμε ένα σύστημα αναφοράς με δυο κάθετους άξονες όπου ο ένας άξονας αναπαριστά τις κατηγορίες της ποιοτικής μεταβλητής και ο άλλος άξονας δείχνει τη συχνότητα ή τη σχετική συχνότητα της ποιοτικής μεταβλητής. Σε κάθε κατηγορία αναπαριστούμε τη συχνότητα ή τη σχετική συχνότητα με μια ορθογώνια ράβδο της οποίας το ύψος είναι ανάλογο της συχνότητας της.

Παράδειγμα 2: Στα δεδομένα του προηγούμενου παραδείγματος κατασκευάζουμε το ραβδόγραμμα των δεδομένων.



Συνάφεια (Crosstabulation): Εκτός από την κατανομή των απαντήσεων σε μια ερώτηση που φαίνεται από τις αντίστοιχες συχνότητες, συνήθως μας ενδιαφέρει και ο τρόπος που κατανέμονται οι απαντήσεις σε μια ερώτηση σε σχέση με τις απαντήσεις σε κάποια άλλη ερώτηση. Στο προηγούμενο παράδειγμα θα είχε ενδιαφέρον να δούμε πως κατανέμονται οι απαντήσεις στην ερώτηση σε σχέση με μια άλλη ερώτηση που θα διερευνούσε την κομματική προτίμηση των ερωτώμενων ή με μια άλλη ερώτηση που θα διερευνούσε το εισόδημα τους.

Παράδειγμα: Θεωρούμε τον παρακάτω πίνακα που περιέχει τα αποτελέσματα από δυο ερωτήσεις. Οι ερωτήσεις είναι "Είστε καπνιστής" και "Πάσχετε από κίρρωση του ήπατος" Το δείγμα μας αποτελείται από 1000 άτομα από τα οποία τα 500 πάσχουν από την ασθένεια ενώ τα υπόλοιπα 500 όχι. Οι απαντήσεις που έδωσαν τα άτομα που συμμετείχαν στο δείγμα συνοψίζονται στον πίνακα που ακολουθεί.

Ερώτηση	*Καπνιστής*		*Σύνολο*
κίρρωση του ήπατος	Ναι	Οχι	
Ναι	350	150	500
Όχι	200	300	500
Σύνολο	550	450	1000

Ο πίνακας μας δείχνει ότι 500 άτομα πάσχουν από κίρρωση του ήπατος από τα οποία τα 350 είναι καπνιστές (70%). Από τους 500 που δεν πάσχουν από κίρρωση του ήπατος οι 200 είναι καπνιστές (40%). Εικάζουμε ότι αν καπνίζεις αυξάνεις τις πιθανότητες να πάθεις κίρρωση του ήπατος. Η μελέτη του πίνακα συνάφειας μας βοηθάει να δούμε τη σχέση μεταξύ του καπνίσματος και της κίρρωσης του ήπατος. Τη συνάφεια μεταξύ των δυο μεταβλητών την ελέγξαμε με το μάτι, υπάρχει και ένας μαθηματικός τρόπος με τον οποίο μπορούμε να ελέγξουμε αν δυο μεταβλητές έχουν σημαντική συνάφεια (x^2-έλεγχος ανεξαρτησίας).

Η συνάφεια δεν σημαίνει απαραίτητα και αιτιώδη σχέση (causal relationship). Από ότι γνωρίζουμε μια από τις αιτίες τις κίρρωσης του ήπατος είναι τα οινοπνευματώδη ποτά και όχι το κάπνισμα επομένως τίθεται το ερώτημα πως δημιουργήθηκε η συνάφεια μεταξύ του καπνίσματος και τις κίρρωσης του ήπατος αφού αυτή δεν υπάρχει; Αν χωρίσουμε τα άτομα του δείγματος σε "άτομα που καταναλώνουν αρκετό αλκοόλ" και σε " άτομα που δεν καταναλώνουν αρκετό αλκοόλ" και ξαναδημιουργήσουμε τους πίνακες συνάφειας για το κάπνισμα και την κίρρωση του ήπατος παίρνουμε τα ακόλουθα αποτελέσματα.

ΑΛΚΟΟΛ						
	ΝΑΙ			ΟΧΙ		
ΚΙΡΡΩΣΗ	ΚΑΠΝΙΣΜΑ			ΚΑΠΝΙΣΜΑ		
ΗΠΑΤΟΣ	ΝΑΙ	ΟΧΙ		ΝΑΙ	ΟΧΙ	
ΝΑΙ	320	80	400	30	120	150
ΟΧΙ	100	20	100	100	280	350
	400	100	500	100	400	500

Βλέπουμε ότι η συνάφεια μεταξύ του καπνίσματος και τη κίρρωσης του ήπατος δεν υπάρχει για τους ανθρώπους που δεν καταναλώνουν πολύ αλκοόλ. Επομένως είναι το αλκοόλ η μεταβλητή ή οποία είναι εξαρτημένη με τη κίρρωση του ήπατος. Ο λόγος που βρήκαμε να υπάρχει εξάρτηση μεταξύ της κίρρωσης του ήπατος και του καπνίσματος είναι ότι το κάπνισμα έχει ισχυρή εξάρτηση με το αλκοόλ που είναι η πραγματική αιτία της κίρρωσης του ήπατος. Οι περισσότεροι άνθρωποι που πίνουν αλκοόλ είναι συγχρόνως και καπνιστές και αυτό δημιούργησε και το πρόβλημα.

ΚΩΔΙΚΟΠΟΙΗΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Για τα ποσοτικά δεδομένα η είσοδος τους στον υπολογιστή είναι απλή. Απλά καταγράφουμε κάθε τιμή σε ένα φύλλο εργασίας είτε στο Excel είτε σε ένα στατιστικό πρόγραμμα (SPSS, Minitab). Σε κάθε περίπτωση, στο φύλλο εργασίας, οι στήλες αντιστοιχούν στις μεταβλητές και οι γραμμές στις μονάδες του δείγματος. Προβλήματα προκύπτουν όταν έχουμε ποιοτικές μεταβλητές. Μπορεί μια ερώτηση σε ένα ερωτηματολόγιο να είναι

Ποια είναι η οικογενειακή σας κατάσταση;

Α) Έγγαμος/η Β)Άγαμος/η Γ)Διαζευγμένος/η Δ)Χήρος/α

Δεν μπορούμε να καταγράψουμε αυτές τις απαντήσεις στον υπολογιστή. Τα δεδομένα μας βρίσκονται σε ονομαστική κλίμακα και απλά πρέπει να δώσουμε σε κάθε απάντηση έναν αριθμό ο οποίος θα είναι μια ετικέτα η οποία θα δίνει σε κάθε μονάδα του δείγματος μια ιδιότητα. Για λόγους ευκολίας είθισται να δίνουμε ακέραιες διαδοχικές τιμές αρχής γενομένης από τη μονάδα. Επομένως θα δώσουμε τις τιμές Α-1, Β-2, Γ-3 και Δ-4.

Ας υποθέσουμε τώρα ότι έχουμε την εξής ερώτηση.
Ποιες από τις παρακάτω συσκευές έχετε στο σπίτι σας ;

Συσκευές	*Ναι*	*Όχι*
Τηλεόραση Η/Υ DVD		

Σε αυτήν την περίπτωση δεν μπορούμε να ακολουθήσουμε τον προηγούμενο συλλογισμό. Αν και η ερώτηση είναι μια εμείς θα φτιάξουμε τρεις στήλες σε κάποιο φύλλο εργασίας, μια για την τηλεόραση, μια για τον Η/Υ και μια για το DVD. Επομένως έχουμε τρεις μεταβλητές αν και η ερώτηση είναι 1. Σε ποιοτικές μεταβλητές με δυο δυνατές απαντήσεις είθισται να δίνουμε τις τιμές 0 και 1 για τις 2 δυνατές απαντήσεις.

ΠΟΣΟΤΙΚΑ ΔΕΔΟΜΕΝΑ – ΣΥΝΕΧΕΙΣ ΜΕΤΑΒΛΗΤΕΣ (QUALITATIVE DATA – CONTINUOUS RANDOM VARIABLES)

Μέτρα Κεντρικής Θέσης

Τα μέτρα κεντρικής θέσης δίνουν μια συνοπτική εικόνα για την κεντρική τιμή μιας τυχαίας μεταβλητής

1) Αριθμητικός Μέσος : Το πιο διαδεδομένο μέτρο είναι ο αριθμητικός μέσος. Υπολογίζεται από τον λόγο του αθροίσματος των τιμών μιας τυχαίας μεταβλητής προς το σύνολο τιμών της. Αν υποθέσουμε ότι έχουμε ν τιμές μιας τυχαίας μεταβλητής χ (που συμβολίζονται $\chi_1, \chi_2, ..., \chi_\nu$) τότε ο αριθμητικός μέσος (συμβολίζεται $\bar{\chi}$) δίνεται από τον τύπο $\bar{\chi} = \frac{\sum_{\iota=1}^{\nu} \chi_\iota}{\nu}$

Παράδειγμα : Αν υποθέσουμε ότι σε κάποιο διαγώνισμα παρατηρήσαμε τους εξής βαθμούς 11, 13, 17, 18, 19 τότε ο μέσος όρος ($\bar{x} = \frac{11+13+17+18+19}{5} = 15.6$)

Πολλές φορές τα δεδομένα μας είναι ομαδοποιημένα είτε για εξοικονομία χώρου είτε για να είναι η δομή των δεδομένων πιο ευδιάκριτη και να έχουμε μια καλύτερη εικόνα της κατανομής τους. Σε αυτήν την περίπτωση κάθε τιμή χ_ι της τυχαίας μεταβλητής Χ συνοδεύεται και από τη συχνότητα ν_ι (ή τον αριθμό των φορών που εμφανίζεται). Ο μέσος όρος δίνεται από τον τύπο $\bar{\chi} = \frac{\sum_{\iota=1}^{\nu} \chi_\iota \nu_\iota}{\sum_{\iota=1}^{\nu} \nu_\iota}$

Παράδειγμα : Σε ένα διαγώνισμα σε μια τάξη 28 μαθητών παρατηρήθηκαν οι βαθμοί που δίνονται στον Πίνακα 1. Ο μέσος όρος είναι $\bar{\chi} = \frac{\sum_{\iota=1}^{\nu} \chi_\iota \nu_\iota}{\sum_{\iota=1}^{\nu} \nu_\iota} = \frac{426}{28} \cong 15.2$

Πίνακας 1 : Κατανομή Βαθμολογίας

Βαθμολογία	Αριθμός Μαθητών	
χ_i	ν_ι	$\chi_\iota \nu_\iota$
11	1	11
12	3	36
13	2	26
14	4	56
15	5	75

16	5	80
17	4	68
18	3	54
20	1	20
Σύνολο	28	426

2) Σταθμισμένος μέσος όρος : Ο σταθμισμένος μέσος όρος δίνεται από τον τύπο $\bar{\chi} = \frac{\sum_{\iota=1}^{\nu} \chi_\iota w_\iota}{\sum_{\iota=1}^{\nu} w_\iota}$ όπου w_i είναι η στάθμιση που παίρνει η κάθε παρατήρηση. Αν $w_i = \frac{1}{n}$ τότε ο σταθμισμένος μέσος όρος ισούται με τον αριθμητικό μέσο. Συνήθως οι σταθμίσεις δίνονται ως ποσοστά οπότε $\sum_{i=1}^{\nu} w_\iota = 1$ και $\bar{\chi} = \sum_{\iota=1}^{\nu} \chi_\iota w_\iota$. Ας υποθέσουμε ότι δίνουμε εξετάσεις σε 4 μαθήματα Α, Β, Γ και Δ και οι βαθμοί μας είναι 12, 13, 15 και 16 αντίσοιχα δίνοντας έναν αριθμητικό μέσο ίσο με 14. Ας υποθέσουμε ότι όλα μετράνε το ίδιο με εξαίρεση το Δ όπου έχει δπλάσια βαρύτητα. Ο σταθμισμένος μέσος όρος θα είναι $0.2 \times (12+13+15) + 0.4 \times 16 = 14.4$.

3) Διάμεσος : Είναι εκείνη η αριθμητική τιμή η οποία χωρίζει τα δεδομένα μας σε δυο κομμάτια με τέτοιο τρόπο ώστε το 50% των τιμών των δεδομένων μας να είναι μεγαλύτερο από αυτή την τιμή και το 50% μικρότερο. Για τον υπολογισμό της διαμέσου ακολουθούμε τα παρακάτω βήματα

I. Ταξινομούμε τις ν παρατηρήσεις σε αύξουσα σειρά
II. Αν το σύνολο των παρατηρήσεων ν είναι μονός αριθμός η διάμεσος είναι η τιμή που βρίσκεται στην (ν+1)/2 θέση ενώ αν το ν είναι ζυγός αριθμός η διάμεσος είναι ο μέσος όρος των τιμών που βρίσκονται στην ν/2 και στην (ν+1)/2 θέση

Παράδειγμα : Σε μια τάξη παρατηρούμε τους παρακάτω βαθμούς 13, 17, 12, 14, 19
Για να υπολογίσουμε τη διάμεσο τους ταξινομούμε σε αύξουσα σειρά 12, 13, 14, 17, 19. Το σύνολο των παρατηρήσεων είναι ν=5 (μονός αριθμός) επομένως η διάμεσος είναι η τιμή που βρίσκεται στην (ν+1)/2=(5+1)/2=3 θέση δηλαδή 14. Αν υποθέσουμε ότι είχαμε μια ακόμα παρατήρηση με την τίμη 16 τότε οι βαθμοί σε αύξουσα σειρά γίνονται 12, 13, 14, 16, 17 και 19, το σύνολο των τιμών είναι ζυγός αριθμός ν=6 και η διάμεσος είναι ο μέσος όρος των τιμών που βρίσκονται στην ν/2 και στην (ν+1)/2 θέση, δηλαδή στην τρίτη και στην τεταρτη θέση, επομένως (14+16)/2=15

4) Επικρατούσα τιμή : Είναι η τιμή που έχει τη μεγαλύτερη συχνότητα (που εμφανίζεται πιο συχνά). Στον Πίνακα 1, Επικρατούσες τιμές είναι οι τιμές 15 και 16 που παρατηρούνται 5 φορές. Η επικρατούσα τιμή χρησιμοποιείται κυρίως για κατηγορικά(ή ποιοτικά) δεδομένα.

5) α%-Περικομμένος μέσος : Είναι ο μέσος όρος των τιμών μιας τυχαίας μεταβλητής αφού αγνοήσουμε το α% των παρατηρήσεων με τις μεγαλύτερες και τις μικρότερες τιμές. Παράδειγμα : Αν έχουμε 10 βαθμολογίες 13, 16, 12, 18, 14, 17, 15, 14, 19 και 13. Ο 10% περικομμένος μέσος είναι ο μέσος όρος των τιμών αφού αφαιρέσουμε τη μεγαλύτερη τιμή (19) και τη μικρότερη τιμή (12) και η τιμή που παίρνουμε είναι 15.

Πλεονεκτήματα – Μειονεκτήματα μέτρων κεντρικής θέσεως

Το πιο διαδεδομένο μέτρο είναι ο αριθμητικός μέσος. Στα πλεονεκτήματα αυτού του μέτρου συγκαταλέγονται η ευκολία στον υπολογισμό του, η φήμη του, το ευκολονόητο περιεχόμενό του και το γεγονός ότι χρησιμοποιεί το σύνολο των δεδομένων. Το βασικό μειονέκτημά του είναι ότι επηρεάζεται από ακραίες τιμές. Η διάμεσος και ο α% περικομμένος μέσος δεν επηρεάζονται από ακραίες τιμές. Το βασικό μειονέκτημα της διαμέσου είναι ότι χρησιμοποιεί μόνο μια ή δυο τιμές απο τα δεδομένα μας χάνοντας, με αυτόν τον τρόπο, πολύτιμη πληροφορία. Η επικρατούσα τιμή χρησιμοποιείται κυρίως σε ποιοτικά δεδομένα όπου θέλουμε να δούμε ποιά κατηγορία εμφανίστηκε περισσότερες φορές. Εαν η κατανομή των δεδομένων μας παρουσιάζει ασυμμετρία, ο αριθμητικός μέσος δεν είναι αξιόπισος.

Μέτρα Κεντρικής Διασποράς

Αυτά τα μέτρα δίνουν μια εικόνα για το πως κατανέμονται οι τιμές μιας τυχαίας μεταβλητης και το πως απλώνονται στο πεδίο ορισμού των τιμών της.

Ελάχιστη τιμή : Είναι η μικρότερη τιμή της τυχαίας μεταβλητής

Μέγιστη τιμή : Είναι η μεγαλύτερη τιμή της τυχαίας μεταβλητής

Εύρος : Είναι η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής

α% ποσοστιαίο σημείο : Είναι το σημείο εκείνο από το οποίο το α% των παρατηρήσεων έχει μικρότερη τιμή και το (1-α)% έχει μεγαλύτερες τιμές. Αν έχουμε ν τιμές, για να υπολογίσουμε το α% ποσοστιαίο σημείο, ταξινομούμε τις παρατηρήσεις μας σε αύξουσα σειρά, υπολογίζουμε την ποσότητα $\Pi = \nu \times \alpha / 100.$
Αν το Π είναι ακέραιος αριθμός τότε το α% ποσοστιαίο σημείο είναι ο μέσος όρος των τιμών που βρίσκονται στην Π και στην Π+1 θέση. Αν το Π είναι δεκαδικός αριθμός τότε το στρογγυλοποιούμε στο πλησιέστερο ακέραιο και το α% ποσοστιαίο σημείο είναι η τιμή σε αυτή τη θέση

Τεταρτημόρια : Τα τεταρτημόρια χωρίζουν τις παρατηρήσεις μας σε 4 μέρη. Το πρώτο τεταρτημόριο (Q1) είναι το 25/% ποσοστιαίο σημείο, το δεύτερο τεταρτημόριο (Q2) είναι το 50% ποσοστιαίο σημείο ή η διάμεσος, το τρίτο τεταρτημόριο (Q3) είναι το 75% ποσοστιαίο σημείο και το τέταρτο τεταρτημόριο (Q4) είναι το 100% ποσοστιαίο σημείο ή η μέγιστη τιμή. Για να τα υπολογίσουμε ταξινομούμε τα δεδομένα σε αύξουσα σειρά και το πρώτο τεταρτημόριο είναι η παρατήρηση που βρίσκεται στην (ν+1)/4 θέση, η διάμεσος η παρατήρηση που βρίσκεται στη (ν+1)/2 θέση και το τρίτο τεταρτημόριο η παρατήρηση η παρατήρηση που βρίσκεται στην $3\times(\nu+1)/4$ θέση. Αν το πρώτο τεταρτημόριο βρίσκεται μεταξύ των θέσεων x_i και x_{i+1} θέση παίρνουμε το σταθμισμένο μέσο όρο τους $0.75\times x_i + 0.25\times x_{i+1}$, ομοίως για το τρίτο τεταρτημόριο $0.25\times x_i + 0.75\times x_{i+1}$.

Ενδοτεταρτομοριακό εύρος (Interquartile Range): Είναι η διαφορά μεταξύ του τρίτου και του πρώτου τεταρτημορίου IQ=Q3-Q1. Το ενδοτεταρτημοριακό εύρος, σε αντίθεση με το εύρος, δεν επηρεάζεται από ακραίες τιμές.

Διακύμανση ή Διασπορά : Είναι το βασικό μέτρο κεντρικής διασποράς και δείχνει πως κατανέμονται τα δεδομένα γύρω από τη μέση τους τιμή. Η διακύμανση δίνεται από τον τύπο

$$\sigma^2 = \frac{\sum_{i=1}^{\nu}(\chi_i - \bar{\chi})^2}{\nu}$$

Τυπική απόκλιση : Δίνεται από την ρίζα της διακύμανσης $\sigma = \sqrt{\sigma^2}$

Συντελεστής διασποράς : Δίνεται από τον τύπο $CV = \frac{\sigma}{\bar{\chi}}$ και χρησιμοποιείται για συγκρίσεις μεταξύ μεταβλητών που είτε έχουν μεγάλη διαφορά στους αριθμητικούς μέσους τους είτε μετριούνται σε διαφορετική κλίμακα π.χ βάρος σε γραμμάρια και κιλά, χρήματα σε ευρώ και δολλάρια κλπ

Παράδειγμα : Έχουμε τα ύψη και τα βάρη 5 ατόμων

Άτομο	*1*	*2*	*3*	*4*	*5*
Ύψος (σε εκ.)	177	180	185	173	184
Βάρος (σε κιλά)	85	78	90	80	87

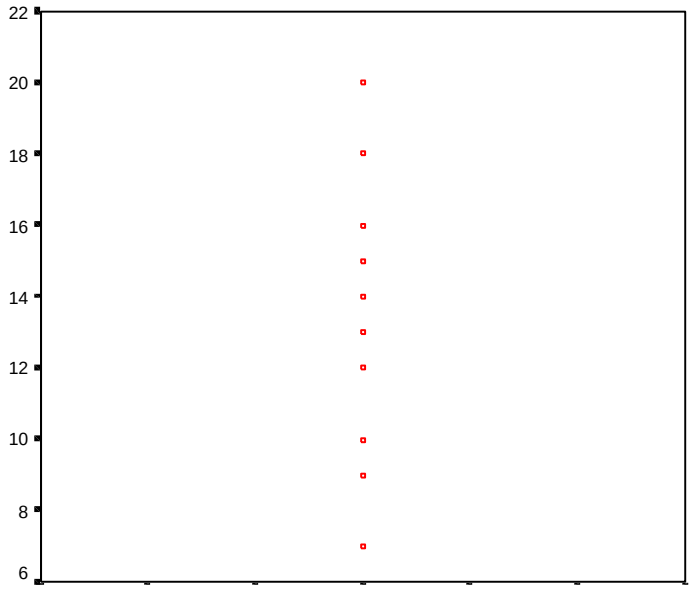
Η διακύμανση του ύψους (24.7) είναι σχεδόν ίδια με αυτή του ύψους (24.5). Δεν μπορούμε να υποθέσουμε ότι οι δυο μεταβλητές έχουν την ίδια διακύμανση γιατί έχουν διαφορετικές μονάδες μέτρησεις(εκατοστά και κιλά). Η μέση τιμή του ύψους είναι αρκετά υψηλότερη από αυτή του βάρους και ο συντελεστής μεταβλητότητας για το ύψος είναι 2.7 ενώ ο συντελεστής μεταβλητότητα του βάρους είναι 5.9. Επομένως η διασπορά του βάρους είναι μεγαλύτερη από αυτή του ύψους.

Διαγράμματα για ποσοτικά δεδομένα

Διάγραμμα σημείων και διάγραμμα διασποράς (Dot plot and scatter plot) : Για να κατασκευάσουμε ένα διάγραμμα σημείων απλά τοποθετούμε τις τιμές της μεταβλητής σε ένα σύστημα αξόνων. Το διάγραμμα σημείων είναι εξαιρετικά κατατοπιστικό όταν έχουμε μικρό αριθμό παρατηρήσεων, διαφορετικά είναι πυκνό και δυσνόητο.

Παράδειγμα : Δέκα φοιτητές έγραψαν τους εξής βαθμούς σε ένα διαγώνισμα μαθηματικών : 13 18 9 16 20 15 7 10 12 14

Το διάγραμμα σημείων δίνεται παρακάτω

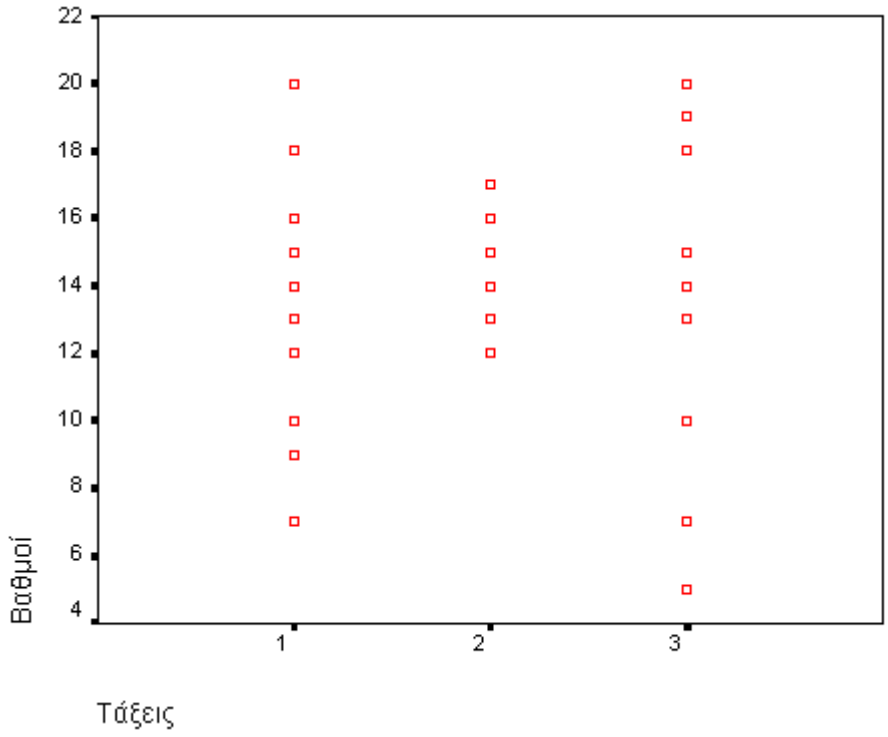


Το διάγραμμα σημείων είναι ιδιαίτερα χρήσιμο όταν έχουμε να συγκρίνουμε μικρά δείγματα από διαφορετικούς πληθυσμούς όπως θα φανεί και στο παράδειγμα που ακολουθεί.

Παράδειγμα : Έστω δύο επιπλέον τάξεις με 10 φοιτητές των οποίων οι βαθμοί είναι Τάξη 2 : 14 13 17 15 14 16 12 17 15 14

Τάξη 3 : 13 19 18 15 7 5 20 18 10 14

Το διάγραμμα σημείων που ακολουθεί συνοψίζει τα αποτελέσματα και από τις 3 τάξεις.



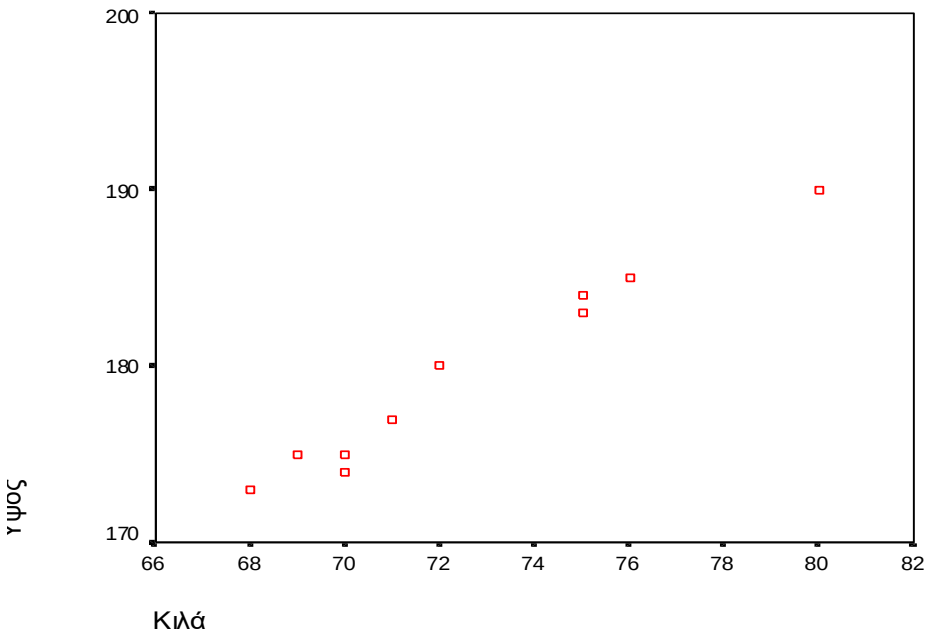
Είναι φανερό από το διάγραμμα σημείων ότι οι μαθητές της δεύτερης τάξης δεν έχουν μεγάλο εύρος βαθμών και οι βαθμοί τους κυμαίνονται γύρω από το 14-15. Η δεύτερη τάξη έχει κάποιους καλούς μαθητές και κάποιους αδύνατους ενώ η τρίτη τάξη έχει το μεγαλύτερο άπλωμα βαθμών από οποιαδήποτε άλλη τάξη.

Το διάγραμμα σημείων είναι και ένα πρώτο εργαλείο για να διαπιστώσουμε τι είδους σχέση συνδέει 2 μεταβλητές τοποθετώντας τις τιμές της μιας στον οριζόντιο άξονα και τις τιμές της άλλης στον κάθετο. Συνήθως σε αυτές τις περιπτώσεις όπου έχουμε 2 μεταβλητές το διάγραμμα σημείων λέγεται διάγραμμα διασποράς (scatter plot).

Παράδειγμα : Μετράμε το ύψος και το βάρος 10 ατόμων για να δούμε αν υπάρχει κάποια σχέση μεταξύ τους. Τα αποτελέσματα που πήραμε είναι :

Άτομο	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Ύψος (σε εκ)	180	175	173	184	190	174	175	183	185	177
Βάρος	72	69	68	75	80	70	70	75	76	71

Το διάγραμμα σημείων για τις δυο μεταβλητές είναι :



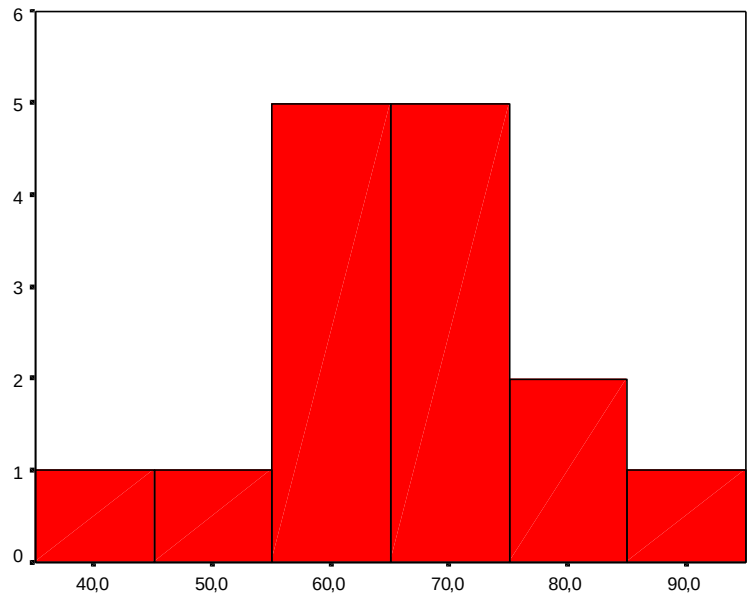
Είναι φανερό ότι όσο αυξάνει το βάρος αυξάνει και το ύψος. Αυτό είναι ένα λογικό συμπέρασμα και ίσως δεν χρειαζόταν το διάγραμμα σημείων για να το διαπιστώσουμε. Από το διάγραμμα σημείων όμως διαπιστώνουμε ότι η σχέση που συνδέει τις δυο μεταβλητές είναι γραμμική δηλαδή ότι τα δεδομένα μας μπορούν να αποτυπωθούν σε μια γραμμή.

Ιστόγραμμα (Histogram): Ο πιο γνωστός ίσως τρόπος γραφικής παρουσίασης μιας ποσοτικής μεταβλητής είναι το ιστόγραμμα συχνοτήτων και το ιστόγραμμα σχετικών συχνοτήτων. Για την κατασκευή του ιστογράμματος προχωράμε ως εξής: Διαιρούμε το εύρος των τιμών της ποσοτικής μεταβλητής σε έναν αριθμό κλάσεων. Στη συνέχεια μετράμε τον αριθμό ή το ποσοστό των παρατηρήσεων που περιέχονται σε κάθε κλάση. Τώρα η κατασκευή του ιστογράμματος είναι ίδια όπως και η κατασκευή του ραβδογράμματος για ποιοτικά δεδομένα. Η διαφορά όμως με το ραβδόγραμμα είναι ότι στα ραβδογράμματα οι κλάσεις είναι προκαθορισμένες ενώ στα ιστογράμματα θα πρέπει να καθορίσουμε το εύρος κάθε κλάσης. Αν έχουμε καθορίσει τον αριθμό των κλάσεων τότε το εύρος κάθε κλάσης ισούται με το εύρος των τιμών διαιρεμένο με τον αριθμό των κλάσεων. Το εύρος κάθε κλάσης ή καλύτερα ο αριθμός των κλάσεων θα πρέπει να καθοριστεί με τέτοιο τρόπο ώστε να απεικονίζει όσον το δυνατό καλύτερα την κατανομή των δεδομένων. Επίσης μπορούμε να

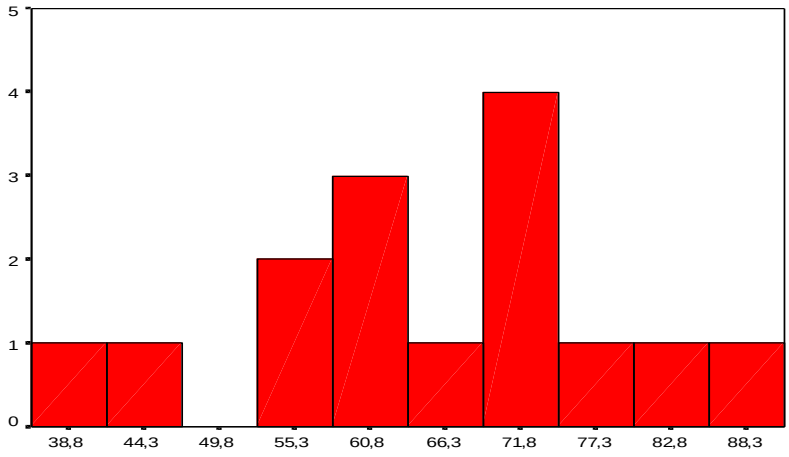
καθορίσουμε ως αρχή της πρώτης κλάσης όχι την ελάχιστη τιμή των δεδομένων αλλά κάποια μικρότερη τιμή όπως επίσης μπορούμε να καθορίσουμε ως το τέλος της τελευταίας κλάσης όχι τη μέγιστη τιμή αλλά κάποια μεγαλύτερη τιμή. Στον υπολογισμό του εύρους της κλάσης θα πρέπει όμως να συμπεριληφθούν οι καινούριες τιμές. Στο παράδειγμα που ακολουθεί φαίνονται πιο καθαρά τα διάφορα ερωτήματα που τίθενται στην κατασκευή ενός ιστογράμματος.

Παράδειγμα : Σε 1 διαγώνισμα οι βαθμοί των μαθητών είναι : 90 45 56 62 37 74 78 84 67 63 72 55 59 70 73. Ο ερευνητής θεωρεί ότι 6 κλάσεις είναι αρκετές για να απεικονίσουν την κατανομή των δεδομένων. Καθορίζει επίσης ως αρχή της πρώτης κλάσης την τιμή 35 και ως το τέλος της τελευταίας κλάσης τον αριθμό 95. Επομένως το εύρος κάθε κλάσης είναι (95-35)/6=10. Στη συνέχεια μετράει τον αριθμό των παρατηρήσεων που ανήκουν σε κάθε κατηγορία και παίρνει τα εξής αποτελέσματα.

Κλάσεις	*Τιμές*	*Συχνότητα*
1	35-44	1
2	45-54	1
3	55-64	5
4	65-74	5
5	75-84	2
6	85-94	1



Το ιστόγραμμα φαίνεται αρκετά απλό και ο ερευνητής έχει την πεποίθηση ότι στα δεδομένα μας υπάρχει μεγαλύτερη “δομή” από αυτήν που δείχνει το ιστόγραμμα. Επομένως αποφασίζει να δημιουργήσει 10 κλάσεις και να θεωρήσει ως αρχή της πρώτης κλάσης και πέρας της τελευταίας κλάσης την ελάχιστη και τη μέγιστη τιμή αντίστοιχα. Επομένως το εύρος κάθε κλάσης είναι (90-37)/10=5.3. Το ιστόγραμμα με τις 10 κλάσεις είναι πιο περίπλοκο και δείχνει περισσότερη πληροφορία από το ιστόγραμμα με τις 6 κλάσεις. Φαίνεται καθαρά η συσσώρευση βαθμολογιών στο διάστημα 69-74 όπως και η απουσία βαθμολογιών στο διάστημα 48-53.



Θηκόγραμμα/Διάγραμμα πλαισίου-απολήξεων (Box-Whisker plot)

Το διάγραμμα πλαισίου-απολήξεων χρησιμοποιεί πέντε στοιχεία. Την ελάχιστη τιμή, τη μέγιστη τιμή, το πρώτο τεταρτημόριο, το τρίτο τεταρτημόριο και τη διάμεσο. Οι απολήξεις του διαγράμματος αντιστοιχούν στην ελάχιστη και τη μέγιστη τιμή. Επομένως η απόσταση μεταξύ αυτών των σημείων είναι το εύρος των παρατηρήσεων. Τα άκρα του πλαισίου αντιστοιχούν στο πρώτο και τρίτο τεταρτημόριο αντίστοιχα και επομένως η απόσταση μεταξύ τους είναι το ενδοτεταρτημοριακό εύρος των παρατηρήσεων. Το ευθύγραμμο τμήμα μέσα στο πλαίσιο αντιστοιχεί στη διάμεσο των παρατηρήσεων.

Εάν τα δεδομένα είναι συμμετρικά τότε

1) Η απόσταση του Q_1 από τη διάμεσο θα ήταν ίση με την απόσταση του Q_3 από τη διάμεσο.
2) Η απόσταση της ελάχιστης τιμής από το Q_1 θα ήταν ίση με την απόσταση της μέγιστης τιμής από το Q_3.

Εάν τα δεδομένα έχουν δεξιά ασυμμετρία τότε

1) Η διάμεσος θα είναι πιο κοντά στο Q_1 από ότι στο Q_3.
2) Η απόσταση της μέγιστης τιμής από το Q_3 θα είναι μεγαλύτερη από αυτήν της ελάχιστης τιμής από το Q_1.

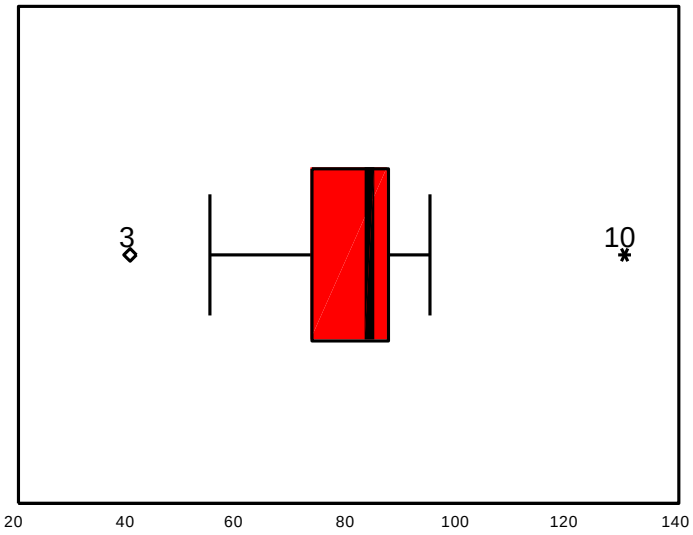
Εάν τα δεδομένα έχουν αριστερή ασυμμετρία τότε

1) Η διάμεσος θα είναι πιο κοντά στο Q_3 από ότι στο Q_1.
2) Η απόσταση της ελάχιστης τιμής από το Q_1 θα είναι μεγαλύτερη από αυτήν της μέγιστης τιμής από το Q_3.

Το διάγραμμα πλαισίου-απολήξεων εκτός του ότι δίνει μια εικόνα της κατανομής των δεδομένων είναι εξαιρετικά χρήσιμο γιατί μας βοηθάει να εντοπίσουμε τις ακραίες τιμές(outliers) στα δεδομένα μας. Οποιοδήποτε στατιστικό πρόγραμμα που έχει την επιλογή του διαγράμματος πλαισίου-απολήξεων θα μας δείξει τόσο τις ήπιες ακραίες τιμές (mild outliers) όσο και τις εξαιρετικά ακραίες τιμές (extreme outliers). Μια τιμή είναι μια ήπια ακραία τιμή αν απέχει από τα άκρα του πλαισίου περισσότερο από 1.5 φορές το μέγεθος του ενδοτεταρτημοριακού εύρους. Δηλαδή αν είναι μεγαλύτερη από την ποσότητα Q_3 +1.5IQR ή μικρότερη από την ποσότητα Q_1 -1.5IQR. Μια τιμή είναι εξαιρετικά ακραία τιμή αν απέχει από τα άκρα του πλαισίου περισσότερες από 3 φορές το μέγεθος του ενδοτεταρτημοριακού

εύρους. Δηλαδή αν είναι μεγαλύτερη από την ποσότητα Q_3 +3IQR ή μικρότερη από την ποσότητα Q_1 -3IQR.

Παράδειγμα: Έχουμε τα εξής δεδομένα : 80, 75, 40, 55, 72, 95,84, 88, 79, 130, 87, 69, 84, 86, 91. Θέλουμε να δούμε την κατανομή των δεδομένων και να διερευνήσουμε αν υπάρχουν ή όχι ακραίες τιμές στα δεδομένα μας. Από το διάγραμμα πλαισίου-απολήξεων που δίνεται παρακάτω βλέπουμε ότι η απόσταση της ελάχιστης τιμής από τη διάμεσο είναι μεγαλύτερη από αυτήν της διαμέσου με τη μέγιστη τιμή. Επομένως τα δεδομένα μας έχουν αριστερή ασυμμετρία. Ο κύκλος στο γράφημα αναφέρεται σε μια ήπια ακραία τιμή ενώ ο κύκλος σε μια εξαιρετικά ακραία τιμή.



ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ PEARSON (PEARSON CORRELATION)

Έστω ότι έχουμε δυο ποσοτικές τυχαίες μεταβλητές x και y. Ο συντελεστής συσχέτισης Pearson δίνεται από τον τύπο

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

Ο συντελεστής συσχέτισης παίρνει τιμές από –1 έως +1. Όταν η τιμή του είναι κοντά στο +1 υποδηλώνει θετική συσχέτιση (positive correlation) ανάμεσα στις μεταβλητές x και y. Δηλαδή όταν αυξάνεται η μεταβλητή x θα αυξάνεται και η μεταβλητή y και όταν μειώνεται η μεταβλητή x θα μειώνεται και η μεταβλητή y. Όταν η τιμή του συντελεστή είναι ακριβώς +1 τότε υπάρχει απόλυτα θετική συσχέτιση μεταξύ των μεταβλητών και σημαίνει ότι κατά τον ίδιο τρόπο και ποσό που μεταβάλλεται η μια μεταβλητή κατά τον ίδιο τρόπο και ποσό θα μεταβάλλεται και η άλλη μεταβλητή.

Αν ο συντελεστής συσχέτισης παίρνει τιμές κοντά στο –1 τότε έχουμε αρνητική συσχέτιση (negative correlation) και όταν αυξάνεται η μια μεταβλητή, θα μειώνεται η άλλη και αντίστροφα. Αν ο συντελεστής συσχέτισης είναι ακριβώς –1 τότε υπάρχει τέλεια αρνητική συσχέτιση και όταν μια μεταβλητή μεταβάλλεται προς μια κατεύθυνση κατά ένα συγκεκριμένο ποσό, η άλλη μεταβλητή θα μεταβάλλεται κατά το ίδιο ποσό προς την αντίθετη κατεύθυνση.

Αν ο συντελεστής συσχέτισης παίρνει τιμές κοντά στο 0, τότε οι δυο μεταβλητές είναι ασυσχέτιστες (uncorrelated) και οι μεταβολές κάποιας από αυτές δεν επηρεάζουν την άλλη.

Ο συντελεστής συσχέτισης φανερώνει μόνο τη γραμμική σχέση. Ακόμα και αν ο συντελεστής είναι κοντά στο 0 οι μεταβλητές μπορεί να σχετίζονται αλλά όχι με γραμμικό τρόπο

Παράδειγμα 19: Τα παρακάτω δεδομένα αναφέρονται στον ποσοστό ανεργίας και το ποσοστό αύξησης του πληθωρισμού για τις Η.Π.Α. για τα έτη 1977-1988

ΕΤΟΣ	*ΠΟΣΟΣΤΟ ΑΝΕΡΓΙΑΣ*	*ΠΛΗΘΩΡΙΣΜΟΣ*
1977	7.1	6.7
1978	6.1	9
1979	5.8	13.3
1980	7.1	12.5
1981	7.6	8.9
1982	9.7	3.8
1983	9.6	3.8
1984	7.5	3.9
1985	7.2	3.8
1986	7	1.1
1987	6.2	4.4
1988	5.5	4.4

Να υπολογιστεί ο συντελεστής συσχέτισης και ο συντελεστής προσδιορισμού.

Ο μέσος για το ποσοστό ανεργίας είναι 7.2% και ο μέσος για τον πληθωρισμό είναι 6,3%.

ΕΤΟΣ	*ΠΟΣΟΣΤΟ ΑΝΕΡΓΙΑΣ* x_i	*ΠΛΗΘΩΡΙΣΜΟΣ* y_i	*(1)* $x_i - \bar{x}$	*(2)* $y_i - \bar{y}$	*(1)*(2)*	*(1)*(1)*	*(2)*(2)*
1977	7.1	6.7	-0.1	0.4	-0.04	0.01	0.16
1978	6.1	9	-1.1	2.7	-2.97	1.21	7.29
1979	5.8	13.3	-1.4	7	-9.8	1.96	49
1980	7.1	12.5	-0.1	6.2	-0.62	0.01	38.44
1981	7.6	8.9	0.4	2.6	1.04	0.16	6.76
1982	9.7	3.8	2.5	-2.5	-6.25	6.25	6.25
1983	9.6	3.8	2.4	-2.5	-6	5.76	6.25
1984	7.5	3.9	0.3	-2.4	-0.72	0.09	5.76
1985	7.2	3.8	0	-2.5	0	0	6.25
1986	7	1.1	-0.2	-5.2	1.04	0.04	27.04
1987	6.2	4.4	-1	-1.9	1.9	1	3.61
1988	5.5	4.4	-1.7	-1.9	3.23	2.89	3.61
Σύνολο					-19.19	19.38	160.42

Άρα ο συντελεστής συσχέτισης είναι $r_{xy} = \frac{-19.19}{\sqrt{19.38 * 160.42}} = -0.344$ και ο συντελεστής προσδιορισμού είναι $r_{xy}^2 = 0.12$. Επομένως το 12% της μεταβολής του πληθωρισμού

εξηγείται από τις μεταβολές της ανεργίας και αντίστροφα. Ο συντελεστής συσχέτισης είναι κοντά στο 0 και επομένως δεν υπάρχει σημαντική συσχέτιση ανάμεσα στις δυο μεταβλητές.

ΑΣΚΗΣΕΙΣ

1) Ένας ερευνητής ποιότητας ελέγχου βρήκε τους ακόλουθους αριθμούς ελαττωματικών προϊόντων σε 15 μέρες που έλεγξε το εργοστάσιο.

8 7 10 10 5 7 15 6 4 10 6 7 7 7 4

Υπολογίστε το μέσο, τη διάμεσο, την επικρατούσα τιμή, το εύρος, το ενδοτεταρτημοριακό εύρος, τη διακύμανση και το συντελεστή διασποράς.

2) Μια εταιρεία κατασκευής πλυντηρίων θέλει να ελέγξει την αντοχή των πλυντηρίων που κατασκευάζει προκειμένου να ορίσει την περίοδο της εγγύησης που θα προσφέρει. Ελέγχει 30 πλυντήρια και ο χρόνος (σε μήνες) της πρώτης βλάβης που παρουσιάστηκε στα πλυντήρια δίνεται παρακάτω.

73 93 85 55 98 64 93 75 71 59 87 61 67 86 69 94 96 87 64 84 79 82
91 96 76 64 96 72 92 75

Να υπολογιστεί ο μέσος, η διάμεσος, το εύρος, η τυπική απόκλιση και ο συντελεστής διασποράς ;

3) Η τυπική απόκλιση της απόστασης όπου ένας ακοντιστής ρίχνει το ακόντιο είναι ίδια με την τυπική απόκλιση της απόστασης όπου μια ακοντίστρια ρίχνει το ακόντιο. Η μέση απόσταση όμως είναι 80 για τους άνδρες ενώ είναι 60 για τις γυναίκες. Μια ακοντίστρια ρίχνει το ακόντιο στα 70 μέτρα ενώ ένας άνδρας το ρίχνει στα 88 μέτρα. Ποιος είναι καλύτερος ακοντιστής ο άνδρας ή η γυναίκα (σε σχέση με την κατηγορία του).

4) Οι τιμές μιας τυχαίας μεταβλητής με 10 παρατηρήσεις κυμαίνονται από 123 έως 351 και έχουν μέσο όρο 274. Αν αφαιρέσουμε την ελάχιστή και τη μέγιστη τιμή ποιος θα είναι ο μέσος όρος των εναπομεινάντων τιμών ;

5) Ένας ερευνητής μελέτησε το μέσο διαθέσιμο εισόδημα και τη μέση κατανάλωση για τα έτη 1995-2001. Τα αποτελέσματα που πήρε συνοψίζονται στον πίνακα που ακολουθεί :

Έτος	*Εισόδημα (σε εκ)*	*Κατανάλωση(σε εκ)*
1995	3,5	3
1996	3,8	3,3
1997	4,5	4
1998	5	4,5
1999	5,1	4,5
2000	4,8	4,1
2001	4,9	4,2

Να κατασκευασθεί ένα διάγραμμα σημείων και να υπολογιστεί ο συντελεστής συσχέτισης Pearson.

6) Έστω το δείγμα μεγέθους 8 με παρατηρήσεις 27, 25, 20, 15, 30, 34, 20 και 25. Υπολογίστε το εύρος και το ενδοτεταρτημοριακό εύρος. Υπολογίστε τη διακύμανση και την τυπική απόκλιση.