# Περιγραφική Στατιστική

Παναγιώτα Λάλου

# 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.

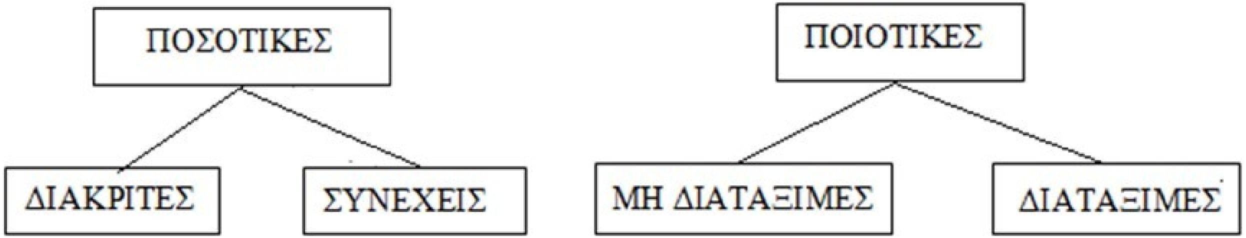
Τα στοιχεία του πληθυσμού, εξετάζονται ως προς ένα ή περισσότερα χαρακτηριστικά. Το χαρακτηριστικό ως προς το οποίο εξετάζουμε έναν πληθυσμό, ονομάζεται στατιστική **μεταβλητή**. Οι δυνατές τιμές που μπορεί να πάρει μια μεταβλητή λέγονται τιμές της μεταβλητής.

Τις μεταβλητές τις διακρίνουμε στις εξής κατηγορίες:

α) Σε <u>ποιοτικές ή κατηγορικές</u>: 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 <u>διατάξιμες</u> ποιοτικές μεταβλητές, π.χ. κατάσταση ασθενή (με τιμές κρίσιμη, κακή, μέτρια καλή), επίπεδο εκπαίδευσης (με τιμές δημοτικό, γυμνάσιο, λύκειο, πανεπιστήμιο). Διαφορετικά, μιλάμε για <u>μη διατάξιμες (ονομαστικές)</u> π.χ. χρώμα αυτοκινήτου.

β) Σε <u>ποσοτικές:</u> 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. <u>Διακριτές</u> 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.

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

$$\nu_1 + \nu_2 + \ldots + \nu_k = \boldsymbol{n}$$

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

$$f_i = \frac{\nu_i}{n} \quad i = 1, \ldots, n$$

Για την σχετική συχνότητα ισχύουν:

i) $0 \leq f_i \leq 1$, ii) $f_1 + f_2 + \ldots + f_k = 1$

Οι ποσότητες $x_i$, $\nu_i$, $f_i$ 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:

$$\boldsymbol{N_i} = \nu_1 + \nu_2 + \ldots + \nu_i$$

# Παρουσίαση Στατιστικών Δεδομένων

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

$$\boldsymbol{F_i} = f_1 + f_2 + \dots + f_i$$

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.

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.

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

**Λύση:**

| $x_i$ | $\nu_i$ | $f_i = \frac{\nu_i}{n}$ | $f_i\%$ | $N_i$ | $F_i$ | $F_i\%$ |
|---|---|---|---|---|---|---|
| 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%

## Παράδειγμα 2:

Στον πίνακα παρουσιάζονται τα επαγγέλματα και το ημερομίσθιο του πατέρα,δεδομένα προερχόμενα από 20 οικογένειες μιας περιοχής της Αθήνας. Να γίνουν οι πίνακες κατανομής συχνοτήτων για τις δύο μεταβλητές

| 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 |

**Λύση:**

**Επάγγελμα**

| $x_i$ | $\nu_i$ | $f_i\%$ | $N_i$ | $F_i\%$ |
|---|---|---|---|---|
| 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**

| $x_i$ | $\nu_i$ | $f_i\%$ | $N_i$ | $F_i\%$ |
|---|---|---|---|---|
| 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 | | |

# 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.

# Ομαδοποίηση Δεδομένων

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 3:

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 $\approx$ 5 (στρογγυλοποιούμε προς τα επάνω). Ο πίνακας κατανομής συχνοτήτων, σχετικών συχνοτήτων, σχετικών αθροιστικών συχνοτήτων που θα προκύψει είναι:

| Κλάσεις [ - ) | Κεντρικ ές τιμές | Συχν. | Σχετική Συχνότ ητα | Σχετική συχνότητ α ποσοστό | Αθρ. Συχν. | Αθρ. Σχετ. συχν | Αθρ. Σχετ. συχν. |
|---|---|---|---|---|---|---|---|
| [ , ) | $x_i$ | $\nu_i$ | $f_i = \frac{\nu_i}{n}$ | $f_i\%$ | $N_i$ | $F_i$ | $F_i\%$ |
| [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 |

# Μέτρα θέσης και Διασποράς

## 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:

1. η μέση τιμή,
2. η διάμεσος,
3. η επικρατούσα τιμή και
4. 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:

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

# 1. Μέτρα θέσης (συν.)

β) **η διάμεσος** $\delta$, 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.

$$Median(x) = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ περιττός}, \\ \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & n \text{ άρτιος}. \end{cases}$$

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.

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

# 1. Μέτρα θέσης (συν.)

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.

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.

$n$ περιττός

Χωρίζουμε τις παρατηρήσεις μας σε δύο ίσα τμήματα, *χωρίς να περιλαμβάνουμε τη διάμεσο σε κανένα από αυτά,* δηλαδή:

$$\underbrace{3,5,7,8}_{1ο\ \text{τμήμα}}, \underbrace{12}_{Q_2}, \underbrace{13,14,18,21}_{2ο\ \text{τμήμα}}$$

Το πρώτο τεταρτημόριο προσδιορίζεται πλέον ως η διάμεσος του 1ου τμήματος

$$Q_1 = \frac{5+7}{2} = 6$$

και το τρίτο τεταρτημόριο ως η διάμεσος του 2ου τμήματος, δηλαδή

$$Q_3 = \frac{14+18}{2} = 16.$$

# 1. Μέτρα θέσης (συν.)

$\boxed{n \text{ άρτιος}}$

Χωρίζουμε τις παρατηρήσεις μας σε δύο ίσα τμήματα, ***περιλαμβάνοντας τη διάμεσο στα δύο από αυτά***, δηλαδή:

$$\underbrace{3,5,7,8,12}_{1o\ \tau\mu\acute{\eta}\mu\alpha}, \quad \underbrace{12,13,14,18,21}_{2o\ \tau\mu\acute{\eta}\mu\alpha}$$

Το πρώτο τεταρτημόριο προσδιορίζεται πλέον ως η διάμεσος του 1ου τμήματος

$$Q_1 = 7$$

και το τρίτο τεταρτημόριο ως η διάμεσος του 2ου τμήματος, δηλαδή

$$Q_3 = 14.$$

## 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**:

$$\boldsymbol{Q} = Q_3 - Q_1$$

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

γ) **Διακύμανση:**

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:

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

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:

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

## 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. Θηκόγραμμα (Box-plot)

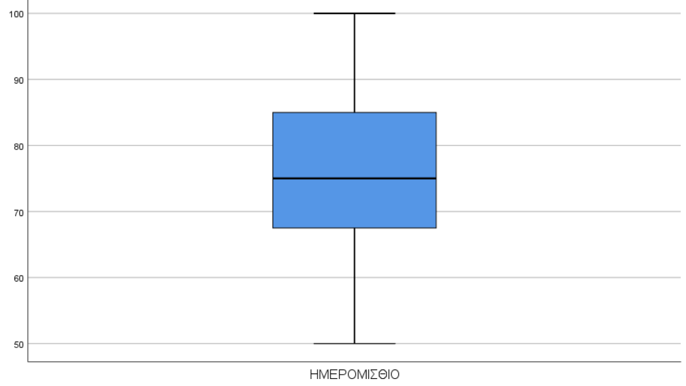
Το θηκόγραμμα είναι μια γραφική παράσταση της κατανομής των παρατηρήσεων, που για την κατασκευή του χρειάζεται ο υπολογισμός της διαμέσου δ, καθώς και του πρώτου και τρίτου τεταρτημορίου $(Q_1, Q_3)$. Αποτελείται από ένα ευθύγραμμο τμήμα

- **με αρχή την ελάχιστη και**
- **τέλος την μέγιστη τιμή του δείγματος,**

και ανάμεσα τοποθετείται ένα ορθογώνιο

- **με κάτω βάση που αντιστοιχεί στην τιμή $Q1$ και**
- **άνω βάση που αντιστοιχεί στην τιμή $Q3$.**
- **Ενδιάμεσα τοποθετείται ένα ευθύγραμμο τμήμα που αντιστοιχεί στην διάμεσο.**

**Θηκόγραμμα των τιμών της μεταβλητής ημερομίσθιο (Παράδειγμα 2)**




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

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

## Παράδειγμα 4:

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:

# Μέτρα θέσης

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

$$\delta = x_7 = 4$$

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

$M_o = 1$

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

0, 1, 1, 1, 2, 3, ~~4~~, 5, 6, 6, 7, 8, 8

0, 1, 1, 1, 2, 3, ~~4~~, 5, 6, 6, 7, 8, 8

## Παράδειγμα 4:

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:**

# Μέτρα διασποράς

$$R = 8 - 0 = 8$$

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

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

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

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

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

## Παράδειγμα 5:

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

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

**Λύση:**

Για την μέση τιμή και διακύμανση είναι χρήσιμος ο πίνακας:

| $x_i$ | $\nu_i$ | $x_i\nu_i$ | $x_i^2$ | $x_i^2\nu_i$ |
|---|---|---|---|---|
| 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) $\bar{x} = \dfrac{\sum_{i=1}^{k} \nu_i \cdot x_i}{\sum_{i=1}^{k} \nu_i} = \dfrac{31}{10}$

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