.

 **ΣΤΑΤΙΣΤΙΚΗ**

 ***Εισαγωγή***

**Στατιστική** είναι η εφαρμοσμένη επιστήμη που συγκεντρώνει και παρουσιάζει πληροφορίες, αλλά ταυτόχρονα μελετά και αναλύει τις παρατηρήσεις, που αναφέρονται στην οικονομία, στη μετεωρολογία, στην υγεία, στην κοινωνιολογία, στον αθλητισμό κτλ.

Η Στατιστική περιλαμβάνει τόσο τις μεθόδους συλλογής και επεξεργασίας στοιχείων όσο και τις μεθόδους ανάλυσης και μελέτης αυτών. Από τη μελέτη αυτή προκύπτουν οι σχέσεις που υπάρχουν στα διάφορα φαινόμενα και διατυπώνονται συμπεράσματα που είναι χρήσιμα για τη λήψη αποφάσεων.

Ως ορισμό της «Στατιστικής» θα δώσουμε τον συνηθέστερο και πλέον γνωστό ορισμό του R.A. Fisher (1890-1962), πατέρα της σύγχρονης Στατιστικής:

**Στατιστική** είναι ένα σύνολο αρχών και μεθοδολογιών για:

* + τον σχεδιασμό της διαδικασίας συλλογής δεδομένων
	+ τη συνοπτική και αποτελεσματική παρουσίασή τους
	+ την ανάλυση και εξαγωγή αντίστοιχων συμπερασμάτων.

Ο κλάδος της Στατιστικής που ασχολείται με τον πρώτο στόχο λέγεται **σχεδιασμός πειραμάτων**, ενώ με τον δεύτερο ασχολείται η **περιγραφική στατιστική**, που αποτελεί και το αντικείμενο μελέτης μας στη συνέχεια. Τέλος, η **επαγωγική στατιστική ή στατιστική συμπερασματολογία** περιλαμβάνει τις μεθόδους με τις οποίες γίνεται η προσέγγιση των χαρακτηριστικών ενός μεγάλου συνόλου δεδομένων, με τη μελέτη των χαρακτηριστικών ενός μικρού υποσυνόλου των δεδομένων.

# ΕΝΟΤΗΤΑ 2.1 : ΠΛΗΘΥΣΜΟΣ - ΔΕΙΓΜΑ - ΜΕΤΑΒΛΗΤΕΣ

***Ορισμοί***

* + Το σύνολο, στο οποίο επικεντρωνόμαστε σε μια έρευνα, ονομάζεται **πληθυσμός** της έρευνας. Τα στοιχεία του πληθυσμού συχνά αναφέρονται ως **άτομα** του πληθυσμού.
	+ Τις περισσότερες φορές είναι πρακτικά πολύ δύσκολο ή και αδύνατο να προ- σεγγίσουμε καθένα από τα άτομα του πληθυσμού, για τεχνικούς λόγους ή λόγω χρόνου, κόστους κτλ. Στις περιπτώσεις αυτές επιλέγουμε ένα μέρος του πληθυσμού που είναι αντιπροσωπευτικό του πληθυσμού και το εξετάζουμε. Το μέρος αυτό του πληθυσμού λέγεται **δείγμα**.
	+ Τα χαρακτηριστικά ως προς τα οποία εξετάζουμε τα άτομα ενός πληθυσμού λέγονται **μεταβλητές** και συμβολίζονται με ένα κεφαλαίο γράμμα Χ, Υ, Ζ, ... κτλ.

***Διάκριση μεταβλητών***

Οι μεταβλητές διακρίνονται σε:

* + **Ποιοτικές** αν αναφέρονται σε ένα ποιοτικό χαρακτηριστικό του πληθυσμού, όπως είναι: τα χρώματα των αυτοκινήτων, το φύλο των μαθητών κτλ.
	+ **Ποσοτικές** αν αναφέρονται σε ένα ποσοτικό χαρακτηριστικό του πληθυσμού, όπως είναι: ο ετήσιος αριθμός των τροχαίων ατυχημάτων, το ύψος των μαθητών κτλ.
		- **Διακριτές** αν παίρνουν μεμονωμένες τιμές, όπως είναι: ο αριθμός των παιδιών των οικογενειών, το νούμερο των γυναικείων παπουτσιών ανά μισό πόντο κτλ.
		- **Συνεχείς** αν μπορούν να πάρουν οποιαδήποτε τιμή ενός διαστήματος, όπως είναι: το βάρος των μαθητών, ο χρόνος σε λεπτά που χρειάζονται οι μαθητές για να απαντήσουν σε ένα διαγώνισμα κτλ.

**ΠΟΣΟΤΙΚΕΣ**

**ΜΕΤΑΒΛΗΤΕΣ**

**ΣΥΝΕΧΕΙΣ**

**ΔΙΑΚΡΙΤΕΣ**

**ΠΟΙΟΤΙΚΕΣ**

Η διαδικασία επιλογής ενός δείγματος από τον πληθυσμό ονομάζεται **δειγματοληψία**. Όπως είπαμε, προκειμένου να γενικεύσουμε τα συμπεράσματα που θα εξαγάγουμε από το δείγμα σε ολόκληρο τον πληθυσμό, έχει ιδιαίτερη σημασία **το δείγμα να είναι αντιπροσωπευτικό του πληθυσμού**. Αυτό σημαίνει ότι τα χαρακτηριστικά που θεωρούμε ότι επηρεάζουν τις μεταβλητές που μας ενδιαφέρουν κατανέμονται με τον ίδιο τρόπο στο δείγμα και στον πληθυσμό.

Ένας απλός τρόπος δειγματοληψίας που επιτυγχάνει στατιστικά την αντιπροσω- πευτικότητα είναι η απλή τυχαία δειγματοληψία, στην οποία επιλέγουμε τυχαία τα μέλη του πληθυσμού που θα συμπεριληφθούν στο δείγμα. Όσο μεγαλύτερο είναι το μέγεθος του δείγματος, τόσο πιο βέβαιοι μπορούμε να είμαστε ότι τα αποτελέσματα που θα πάρουμε από το δείγμα μας γενικεύονται στον πληθυσμό.

# ΕΝΟΤΗΤΑ 2.2 : ΠΑΡΟΥΣΙΑΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

##  Διερεύνηση

Μια εταιρεία αθλητικών ειδών, πριν επενδύσει σε αθλητικά είδη που προτιμούν οι μαθητές, αποφάσισε να κάνει μια έρευνα. Για τον λόγο αυτό επέλεξε, με τυχαίο τρόπο, δείγμα τριακοσίων μαθητών απ’ όλη την Ελλάδα. Ο υπεύθυνος που έκανε την έρευνα, μετά την επεξεργασία των στοιχείων που συγκέντρωσε, παρουσίασε

στον διευθυντή της εταιρείας τον παρακάτω πίνακα και τα δυο διαγράμματα.

**Άθλημα που προτιμούν οι μαθητές**

120

100

80

60

40

20

0

**Βόλεϊ**

**Μπάσκετ Ποδόσφαιρο Άλλο**

|  |  |  |
| --- | --- | --- |
| **Άθλημα** | **Πλήθος μαθητών** | **Ποσοστό** |
| Βόλεϊ | 48 | 16% |
| Μπάσκετ | 93 | 31% |
| Ποδόσφαιρο | 108 | 36% |
| Άλλο | 51 | 17% |
| **Σύνολο** | **300** | **100%** |

**Άθλημα που προτιμούν οι μαθητές**

**Άλλο Βόλεϊ**

**17% 16%**

**Ποδόσφαιρο**

**36%**

**Μπάσκετ**

**31%**

1. Ποια είναι η **μεταβλητή** της έρευνας και ποιο το είδος της;
2. Πώς προκύπτουν τα αντίστοιχα **ποσοστά** για κάθε άθλημα;

**Πλήθος μαθητών**

1. Ποιο είναι το ύψος της κάθε μπάρας στο **ραβδόγραμμα**;
2. Ποια είναι η γωνία του κάθε κυκλικού τομέα στο **κυκλικό διάγραμμα**;

##  Βασικές μαθηματικές έννοιες

 ***Ορισμοί***

Έστω Χ μια μεταβλητή με τιμές x1 , x2 , x3 , ..., xκ .

* **Συχνότητα νi** μιας τιμής xi λέγεται ο φυσικός αριθμός που εκφράζει πόσες φορές εμφανίζεται η τιμή αυτή στο σύνολο των παρατηρήσεων του δείγματος. Είναι φανερό ότι το άθροισμα όλων των συχνοτήτων μας δίνει το **μέγεθος ν** του δείγματος. Δηλαδή:

**ν1 + ν2 + ν3 +… + νκ = ν**

* **H σχετική συχνότητα fi** μιας τιμής xi ορίζεται ως ο λόγος της αντίστοιχης συχνότητας vi

προς το μέγεθος ν του δείγματος. Δηλαδή:

fi = $\frac{ν\_{i}}{ν}$ , i = 1, 2, 3, . . ., κ

Η σχετική συχνότητα μπορεί να εκφραστεί και ως ποσοστό fi% : fi% = 100fi

Σύμφωνα με τους παραπάνω ορισμούς, ο πίνακας του παραδείγματος που αναφέραμε με

μεταβλητή «το άθλημα που προτιμούν οι μαθητές» μετατρέπεται όπως φαίνεται στη

συνέχεια και λέγεται πίνακας **κατανομής συχνοτήτων και σχετικών συχνοτήτων**.

*Η σχετική συχνότητα* fi *μπορεί να εκφραστεί και ως ποσοστό* fi%.

|  |  |  |  |
| --- | --- | --- | --- |
| **Άθλημα****xi** | **Συχνότητα****vi** | **Σχετική συχνότητα fi** | **Σχετική συχνότητα % fi%** |
| Βόλεϊ | 48 | 0,16 | 16 |
| Μπάσκετ | 93 | 0,31 | 31 |
| Ποδόσφαιρο | 108 | 0,36 | 36 |
| Άλλο | 51 | 0,17 | 17 |
| **Σύνολο** | **300** | **1,00** | **100** |

f1 = $\frac{ν\_{1}}{ν}$ f1 = $\frac{48}{300}$ f1 = 0,16 , οπότε f1% = 100f1 = 100∙0,16 = 16%

f2 = $\frac{ν\_{2}}{ν}$ f2 = $\frac{93}{300}$ f2 = 0,31 , οπότε f2% = 100f2 = 100∙0,31 = 31%

f3 = $\frac{ν\_{3}}{ν}$ f3 = $\frac{108}{300}$ f3 = 0,36 , οπότε f3% = 100f3 = 100∙0,36 = 36%

f4 = $\frac{ν\_{4}}{ν}$ f4 = $\frac{51}{300}$ f4 = 0,17 , οπότε f4% = 100f4 = 100∙0,17 = 17%

 ***Γραφικές παραστάσεις***

Οι πλέον συνηθισμένοι τρόποι γραφικής παρουσίασης ποιοτικών αλλά και ποσοτικών διακριτών δεδομένων είναι το **ραβδόγραμμα συχνοτήτων** και το **κυκλικό διάγραμμα**.

* + - **Το ραβδόγραμμα συχνοτήτων** αποτελείται από ορθογώνιες στήλες, μια για κάθε τιμή της μεταβλητής, όπου το ύψος της κάθε στήλης είναι ίσο με την αντίστοιχη συχνότητα. Αν αντί για τις συχνότητες έχουμε τις σχετικές συχνότητες, τότε λέγεται **ραβδόγραμμα σχετικών συχνοτήτων**.
		- **Το κυκλικό διάγραμμα** χρησιμοποιείται επίσης για τη γραφική παράσταση δεδομένων, κυρίως όταν αυτά παίρνουν λίγες τιμές. Στο κυκλικό διάγραμμα χωρίζουμε έναν κύκλο σε κυκλικούς τομείς, έτσι ώστε η γωνία του κάθε κυκλικού τομέα να είναι ανάλογη της αντίστοιχης σχετικής συχνότητας. Δηλαδή:

αi = 360° ∙ fi για i = 1, 2, 3,..., κ

Για το παραπάνω παράδειγμα με μεταβλητή «το άθλημα που προτιμούν οι μαθητές» έχουμε ότι:

− ο κυκλικός τομέας που αντιστοιχεί στην τιμή «Βόλεϊ» είναι γωνίας α1 = 360ο∙f1 = 360ο∙0,16 = 57,6o

− ο κυκλικός τομέας που αντιστοιχεί στην τιμή «Μπάσκετ» είναι γωνίας α2 = 360ο∙f2 = 360ο∙0,31 = 111,6o

− ο κυκλικός τομέας που αντιστοιχεί στην τιμή «Ποδόσφαιρο» είναι γωνίας α3 = 360ο∙f3 = 360ο∙0,36 = 129,6o

− ο κυκλικός τομέας που αντιστοιχεί στην τιμή «Άλλο» είναι γωνίας α4 = 360ο∙f4 = 360ο∙0,17 = 61,2o

* + - * Όταν έχουμε λίγες παρατηρήσεις, τότε η κατανομή τους μπορεί να περιγραφεί με το **σημειόγραμμα** στο οποίο οι τιμές παριστάνονται με σημεία υπεράνω ενός άξονα.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |  |  |  |  |  |  |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

* Το **χρονόγραμμα** χρησιμοποιείται για την γραφική απεικόνιση της εξέλιξης σε σχέση με το χρόνο ενός μεγέθους, συνήθως οικονομικού ή δημογραφικού.

**ποσοστό ανεργίας**

**ποσοστά ανεργίας στην Ελλάδα την περίοδο 2004-2018**

30,00%

25,00%

20,00%

15,00%

10,00%

5,00%

0,00%

2004

2005

2006

2007

2008

2009

2010

2011

2012

2013

2014

2015

2016

2017

2018

***Ομαδοποίηση παρατηρήσεων - Δεδομένα σε κλάσεις***

Στην περίπτωση που έχουμε ποσοτικά συνεχή δεδομένα με πολλές διαφορετικές τιμές, τότε, για την καλύτερη παρουσίασή τους, γίνεται **ομαδοποίηση** αυτών σε **κλάσεις**, δηλαδή σε διαστήματα, συνήθως ίσου πλάτους. Σχετικός είναι ο παρακάτω πίνακας με τους χρόνους που χρειάστηκαν οι 20 μαθητές ενός τμήματος για να απαντήσουν σε ένα πρόβλημα.

|  |  |  |  |
| --- | --- | --- | --- |
| **Χρόνος σε λεπτά των μαθητών****Κλάσεις** | **Συχνότητα****vi** | **Σχετική συχνότητα****fi** | **Σχετική συχνότητα %****fi%** |
| [6,8) | 2 | 0,10 | 10 |
| [8,10) | 6 | 0,30 | 30 |
| [10,12) | 9 | 0,45 | 45 |
| [12,14) | 3 | 0,15 | 15 |
|  **Σύνολο** | **20** | **1,00** | **100** |

* Η γραφική παρουσίαση ομαδοποιημένων στατιστικών δεδομένων γίνεται με το **ιστόγραμμα συχνοτήτων**. Στο παρακάτω σχήμα φαίνεται η γραφική πα- ρουσίαση των δεδομένων του προηγουμένου πίνακα. Αν αντί για τις συχνότη- τες έχουμε τις σχετικές συχνότητες, τότε λέγεται **ιστόγραμμα σχετικών συ- χνοτήτων**.

|  |  |
| --- | --- |
| 10 |  |
|  |
| 9 |
|  |  |  |
| 8 |
|  |  |
| 7 |
|  |  |
| 6**Πλήθος μαθητών** |
|  |  |  |
| 5 |
|  |  |
| 4 |
|  |  |
| 3 |
|  |  |  |
| 2 |
|  |  |  |
| 1 |
|  |  |
| 0 |
|  |  | [6,8) | [8,10) [10,12)**Χρόνος σε λεπτά** | [12,14) |  |

* Αν θεωρήσουμε δυο επιπλέον κλάσεις ίσου πλάτους, μια στην αρχή και μια στο τέλος, με συχνότητα 0 και ενώσουμε τα μέσα των άνω βάσεων, τότε προ- κύπτει το λεγόμενο **πολύγωνο συχνοτήτων**. Σχετικό είναι το παρακάτω σχή- μα. Αν αντί για τις συχνότητες έχουμε τις σχετικές συχνότητες, τότε λέγεται **πολύγωνο σχετικών συχνοτήτων**.

**Πλήθος μαθητών**

[6,8) [8,10) [10,12) [12,14)

**Χρόνος σε λεπτά**

10

9

8

7

6

5

4

3

2

1

0

**ΕΝΟΤΗΤΑ 2.3: ΜΕΤΡΑ ΘΕΣΗΣ ΚΑΙ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ, ΘΗΚΟΓΡΑΜΜΑ, ΣΥΝΤΕΛΕΣΤΗΣ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ**

##  Βασικές μαθηματικές έννοιες

* Η **μέση τιμή x** ενός συνόλου ν παρατηρήσεων t1, t2, t3, …, tν ορίζεται ως το άθροισμα των

 παρατηρήσεων διά του πλήθους τους.

x = t1 + t2 + t3 +...+ tν

ν

Αν για παράδειγμα οι βαθμοί ενός μαθητή σε δέκα μαθήματα είναι:

18, 17, 19, 17, 18, 16, 18, 17, 16, 18, τότε:

$\overbar{x}$ = $\frac{18+17+19+17+18+16+18+17+16+18}{10}$ = $\frac{174}{10}$ = 17,4

 Αν οι διαφορετικές τιμές των ν παρατηρήσεων είναι x1, x2, . . . ,xκ με αντίστοιχες συχνότητες ν1, ν2, . . . ,νκ, τότε:

$\overbar{x}$ = $\frac{x\_{1}∙ν\_{1}+x\_{2}∙ν\_{2}+ . . . +x\_{κ}∙ν\_{κ}}{ν}$

|  |  |  |
| --- | --- | --- |
| xi | νi | xi  νi |
| 16 | 2 | 32 |
| 17 | 3 | 51 |
| 18 | 4 | 72 |
| 19 | 1 | 19 |
| Σύνολο | 10 | 174 |

 Για το παράδειγμα με τους βαθμούς του μαθητή στα δέκα μαθήματα έχουμε:

x= 16  2+17  3+18  4+19 1 = 10

174

10

= 17,4

* + - * Προκειμένου να υπολογίσουμε τη **διάμεσο** ενός δείγματος ν παρατηρήσεων, αρχικά διατάσσουμε τις παρατηρήσεις σε αύξουσα σειρά. Η **διάμεσος δ** ορίζεται ως:
* η μεσαία παρατήρηση του διατεταγμένου δείγματος, όταν το ν είναι περιττός αριθμός,
* το ημιάθροισμα των δύο μεσαίων παρατηρήσεων του διατεταγμένου δείγματος, όταν το ν είναι άρτιος αριθμός.

Για το παράδειγμα με τους βαθμούς του μαθητή στα δέκα μαθήματα, αν τους διατάξουμε σε αύξουσα σειρά, έχουμε: 16, 16, 17, 17, 17, 18, 18, 18, 18, 19. Επομένως, αφού το μέγεθος του δείγματος είναι 10, δηλαδή άρτιος, η διάμεσος είναι ίση με το ημιάθροισμα των δύο μεσαίων παρατηρήσεων, δηλαδή είναι:

δ = 17 +18 = 17,5

2

Αν έχουμε τις 7 παρατηρήσεις 4, 4, 5, 5, 6, 7, 39 σε αύξουσα σειρά, τότε, αφού το μέγεθος του δείγματος είναι 7, δηλαδή περιττός, η διάμεσος είναι ίση με τη μεσαία παρατήρηση, δηλαδή είναι δ = 5 ενώ η μέση τιμή είναι:

$\overbar{x}$ = $\frac{4+4+5+5+6+7+39}{7}$ = $\frac{70}{7}$ = 10

Παρατηρούμε ότι, στην παραπάνω περίπτωση, η διάμεσος επηρεάζεται λιγότερο από την ακραία παρατήρηση «39».

Είναι φανερό ότι το πολύ 50% των παρατηρήσεων είναι μικρότερες και το πολύ 50% είναι μεγαλύτερες από τη διάμεσο δ.

* + Τα **τεταρτημόρια** συμβολίζονται με Q1, Q2 και Q3.
		- Για το Q1 έχουμε αριστερά το πολύ 25% και δεξιά το πολύ 75% των παρατηρήσεων.
		- Για το Q2 έχουμε Q2 = δ, δηλαδή συμπίπτει με τη διάμεσο.
		- Για το Q3 έχουμε αριστερά το πολύ 75% και δεξιά το πολύ 25% των παρατηρήσεων.

Για τον ευκολότερο υπολογισμό των Q1 και Q3 έχουμε:

* + - Όταν το πλήθος των διατεταγμένων παρατηρήσεων είναι άρτιος αριθμός, διακρίνουμε το πρώτο μισό, του οποίου η διάμεσος είναι το Q1 και το δεύτερο μισό, του οποίου η διάμεσος είναι το Q3.
		- Όταν το πλήθος των διατεταγμένων παρατηρήσεων είναι περιττός αριθμός, αφαιρούμε από το δείγμα τη διάμεσο και διακρίνουμε το πρώτο μισό, του οποίου η διάμεσος είναι το Q1 και το δεύτερο μισό, του οποίου η διάμεσος είναι το Q3.

Για το παράδειγμα με τους βαθμούς του μαθητή στα δέκα μαθήματα είναι:

Q1 = 17 , Q2 = δ =17,5 , Q3 = 18

* + **Επικρατούσα τιμή Μ0** είναι η παρατήρηση με τη μεγαλύτερη συχνότητα.

Στο παράδειγμα με τους βαθμούς του μαθητή στα δέκα μαθήματα, επικρατούσα τιμή είναι ίση με Μ0 = 18.

Η επικρατούσα τιμή μπορεί να μην είναι μοναδική.

 Όταν όλες οι παρατηρήσεις είναι διαφορετικές, τότε λέμε ότι δεν υπάρχει επικρατούσα τιμή.

 ***Ορισμοί για τα μέτρα διασποράς ποσοτικών δεδομένων***

* + Το **εύρος R** είναι το απλούστερο από τα μέτρα διασποράς και ορίζεται ως η

διαφορά της ελάχιστης xmin παρατήρησης από τη μέγιστη xmax παρατήρηση.

Δηλαδή:

R = μέγιστη παρατήρηση - ελάχιστη παρατήρηση = xmax - xmin

* Το **ενδοτεταρτημοριακό εύρος Q** είναι η διαφορά του πρώτου τεταρτημορίου Q1 από το τρίτο τεταρτημόριο Q3. Δηλαδή:

 Q = Q3 - Q1

Όσες παρατηρήσεις βρίσκονται έξω από το διάστημα [Q1 - 1,5∙Q , Q3 + 1,5∙Q] ονομάζονται **ακραίες τιμές**. Η διάμεσος είναι ένα μέτρο θέσης το οποίο δεν επηρεάζεται από το μέγεθος των ακραίων τιμών.

Το εύρος και το ενδοτεταρτημοριακό εύρος είναι πολύ απλά μέτρα διασποράς και δεν εκφράζουν την απόκλιση που έχει κάθε παρατήρηση από το «κέντρο» των παρατηρήσεων. Για τον λόγο αυτό χρησιμοποιούμε και τα επόμενα μέτρα διασποράς.

* Η **διακύμανση s2** ορίζεται ως η μέση τιμή των τετραγώνων των διαφορών της μέσης τιμής

των παρατηρήσεων από τις παρατηρήσεις. Δηλαδή:

s2 = $\frac{(t\_{1}-\overbar{x})^{2}+(t\_{2}-\overbar{x})^{2}+(t\_{3}-\overbar{x})^{2}+ … +(t\_{ν}-\overbar{x})^{2}}{ν}$

 Για παράδειγμα, αν οι βαθμοί έξι μαθητών στα Μαθηματικά είναι: 18, 15, 18, 18, 15, 18, τότε:

$\overbar{x}$ = $\frac{18+15+18+18+15+18}{6}$ = $\frac{102}{6}$ = 17 και

s2 = $\frac{(18-17)^{2}+(15-17)^{2}+(18-17)^{2}+ (18-17)^{2}+(15-17)^{2}+(18-17)^{2}}{6}$ = $\frac{12}{6}$ = 2

Αν οι τιμές των παρατηρήσεων είναι xi με αντίστοιχες συχνότητες νi, τότε:

s2 = $\frac{(x\_{1}-\overbar{x})^{2}∙ν\_{1}+(x\_{2}-\overbar{x})^{2}∙ν\_{2}+(x\_{3}-\overbar{x})^{2}∙ν\_{3}+ … +(x\_{κ}-\overbar{x})^{2}∙ν\_{κ}}{ν}$

Για το παράδειγμα των βαθμών των έξι μαθητών στα Μαθηματικά, σχετικός είναι ο επόμενος πίνακας:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| xi | vi | xi ∙ vi | xi - $\overbar{x}$ |  (xi - $\overbar{x})$2 | (xi - $\overbar{x})$2 vi |
| 15 | 2 | 30 | -2 | 4 | 8 |
| 18 | 4 | 72 | 1 | 1 | 4 |
| Σύνολο | 6 | 102 |  |  | 12 |

$\overbar{x}$ = $\frac{102}{6}$ = 17 και s2 = $\frac{12}{6}$ = 2

Η διακύμανση δεν εκφράζεται με τις μονάδες που εκφράζονται οι παρατηρήσεις.

* Η **τυπική απόκλιση s** είναι: s= $\sqrt{s^{2}}$ .

Στο παραπάνω παράδειγμα είναι: s = $\sqrt{2}$  1,41 .

Η τυπική απόκλιση έχει το πλεονέκτημα ότι εκφράζεται με τις μονάδες, που εκφρά-

ζονται οι παρατηρήσεις.

***Θηκόγραμμα***

* + Έστω ότι έχουμε τον αριθμό των εργαζομένων σε 20 βιοτεχνίες.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10 | 14 | 25 | 7 | 31 | 8 | 12 | 19 | 10 | 24 | 9 | 13 | 5 | 28 | 24 | 19 | 26 | 51 | 68 | 14 |

Κατατάσσουμε τις παρατηρήσεις σε αύξουσα σειρά:

5, 7, 8, 9, 10, 10, 12, 13, 14, 14, 19, 19, 24, 24, 25, 26, 28, 31, 51, 68

Είναι:

 Q1 = $\frac{10+10}{2}$ = 10 , Q2 = δ = $\frac{14+19}{2}$ = 16,5 , Q3 = $\frac{25+26}{2}$ = 25,5

Το ενδοτεταρτημοριακό εύρος είναι: Q = 25,5 - 10 = 15,5 και 1,5 ∙ Q = 23,25 Επομένως Q1 - 1,5 Q , Q3 +1,5 Q = -13,25 , 48,75 . Οπότε ακραίες παρατη- ρήσεις είναι το 51 και το 68.

Τα μέτρα αυτά απεικονίζονται στο επόμενο διάγραμμα που λέγεται **θηκό- γραμμα**.

***Συντελεστής μεταβλητότητας***

**ακραία τιμή**

**ακραία τιμή**

**Q3**

**x min Q1 Q2=δ**

**2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72**

Από τα μέσα των πλευρών, που παριστάνουν το πρώτο και το τρίτο τεταρτημό- ριο (δηλαδή τα Q1 και Q3 αντιστοίχως) φέρουμε ευθύγραμμα τμήματα με μήκος που προσδιορίζεται ως εξής:

Η μικρότερη παρατήρηση που είναι μεγαλύτερη του Q1-1,5Q είναι το 5, επο- μένως το αριστερό άκρο του τμήματος που θα φέρουμε από το Q1 είναι το 5 και δεν υπάρχουν παρατηρήσεις μικρότερες του Q1-1,5Q για να σημειωθούν χωριστά.

Η μεγαλύτερη παρατήρηση που είναι μικρότερη του Q3+1,5Q είναι η 31, επο- μένως το δεξιό άκρο του τμήματος που θα φέρουμε από το Q3 είναι το 31 και θα σημειώσουμε χωριστά τις ακραίες τιμές 51 και 68 οι οποίες είναι μεγαλύτερες του Q3+1,5Q.

* Δυο όμιλοι επιχειρήσεων, που ο καθένας αποτελείται από 5 επιχειρήσεις, εί- χαν ετήσιες δαπάνες για το οικονομικό έτος 2017 τα ποσά, σε χιλιάδες ευρώ, που αναγράφονται στον παρακάτω πίνακα.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Όμιλος Α** | 200 | 250 | 300 | 300 | 350 |
| **Όμιλος Β** | 5.000 | 5.050 | 5.100 | 5.100 | 5.150 |

Οι μέσες τιμές των δαπανών για κάθε όμιλο είναι:

$\overbar{x}$Α = $\frac{200+250+300+300+350}{5}$ = 280 χιλιάδες ευρώ

$\overbar{x}$Β = $\frac{5.000+5.050+5.100+5.100+5.150}{5}$ = 5.080 χιλιάδες ευρώ

Οι διακυμάνσεις των δαπανών για κάθε όμιλο είναι:

 (200 - 280)2 +(250 - 280)2 +(300 - 280)2 +(300 - 280)2 + (350 - 280)2

$s\_{A}^{2}$$ $= 5

s2 = 2.600

A

s2 =

B

(5.000 - 5.080)2 + (5.050 - 5.080)2 + (5.100 - 5.080)2 + (5.100 - 5.080)2 + (5.150 - 5.080)2

5

s2 = 2.600

Β

Οι τυπικές αποκλίσεις των δαπανών για κάθε όμιλο είναι ίσες με:

sA = sB =  51 χιλιάδες ευρώ.

2.600

Μπορούμε άραγε να ισχυριστούμε ότι οι 51 χιλιάδες ευρώ των ίσων τυπικών αποκλίσεων έχουν την ίδια βαρύτητα για τους ομίλους Α και Β, που έχουν διαφο-

ρετικές μέσες τιμές

xΑ = 280 χιλιάδες ευρώ και

xΒ = 5.080 χιλιάδες ευρώ; Ένας

τέτοιος ισχυρισμός θα ήταν λανθασμένος, γιατί η σχέση τιμών της μεταβλητής

«ετήσιες δαπάνες» με την τιμή της τυπικής απόκλισης, είναι διαφορετική για τις επιχειρήσεις των δυο ομίλων.

Ένα μέτρο, που μας βοηθάει να αντιμετωπίσουμε τέτοια προβλήματα και μας δίνει τη δυνατότητα συγκρίσεων, είναι ο **συντελεστής μεταβλητότητας CV**, ο οποίος είναι καθαρός αριθμός και εκφράζεται συνήθως ως ποσοστό. Ορίζεται ως ακολούθως:

CV =

s

x

Στο παράδειγμά μας ο συντελεστής μεταβλητότητας για τις δυο επιχειρήσεις εί- ναι: CVΑ = $\frac{51}{280}$  0,1821= 18,21% και CVΒ = $\frac{51}{5.080}$  0,01004  1,00%

Όσο μικρότερο είναι το ποσοστό αυτό, τόσο περισσότερη ομοιογένεια υπάρ- χει στις τιμές της μεταβλητής που μας ενδιαφέρει. Στο παράδειγμά μας είναι CVA  CVB, επομένως η διασπορά των τιμών στον όμιλο Α σε σχέση με τη μέση τιμή είναι μεγαλύτερη από την αντίστοιχη διασπορά στον όμιλο Β.

# ΕΝΟΤΗΤΑ 2.4 : ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

##  Διερεύνηση

Το παρακάτω ιστόγραμμα και το αντίστοιχο πολύγωνο σχετικών συχνοτήτων πα- ρουσιάζουν τον χρόνο, σε λεπτά, που χρειάζονται οι μαθητές για να πάνε από το σπίτι τους στο σχολείο. Η μέση τιμή είναι 10 λεπτά και η τυπική απόκλιση 2 λεπτά.

40

35

30

25

20

15

10

5

0

4 6 8 10 12 14 16

**Χρόνος σε λεπτά**

40

35

30

25

20

15

10

5

0

4

6 8 10 12 14

**Χρόνος σε λεπτά**

16

**Mαθητές %**

**Mαθητές %**

1. Τι έχετε να παρατηρήσετε για τη συμμετρία της κατανομής;
2. Τι ποσοστό μαθητών χρειάζονται το πολύ 10 λεπτά και πόσοι τουλάχιστον 10 λεπτά για να πάνε στο σχολείο;
3. Τι ποσοστό μαθητών χρειάζονται από 8 έως 12 λεπτά για να πάνε στο σχο- λείο;
4. Τι ποσοστό μαθητών χρειάζονται τουλάχιστον 8 λεπτά για να πάνε στο σχο- λείο;

##  Βασικές μαθηματικές έννοιες

 ***Κανονική κατανομή***

Στο παρακάτω ιστόγραμμα σχετικών συχνοτήτων επί τοις εκατό (%) παριστάνεται η κατανομή 10.000 νεογέννητων στις ΗΠΑ, ως προς το βάρος τους (σε κιλά). Τα νεογέννητα κατανέμονται σε κλάσεις ίσου πλάτους. Από αυτό το ιστόγραμμα αντλούμε αρκετές πληροφορίες, όπως π.χ. ότι η κλάση με τα περισσότερα νεο- γέννητα είναι από 3,2 έως 3,4 κιλά, η οποία περιέχει γύρω στο 9% του δείγματος, δηλαδή 900 νεογέννητα περίπου.

Ακολουθεί το αντίστοιχο πολύγωνο σχετικών συχνοτήτων (%) της μεταβλητής

10

9

8

7

6

5

4

3

2

1

0

**Iστόγραμμα σχετικών συχνοτήτων των βαρών 10.000 νεογέννητων στις ΗΠΑ**

**Βάρος νεογέννητου σε κιλά**

**Νεογέννητα %**

«βάρος νεογέννητου στις ΗΠΑ», για το ίδιο δείγμα των 10.000. Παρατηρούμε ότι η μορφή του είναι σχεδόν συμμετρική ως προς μία κατακόρυφη ευθεία, η οποία είναι κάθετη στον άξονα των βαρών, κοντά στα 3,2 κιλά. Από τη μορφή του πολυ- γώνου συχνοτήτων καταλαβαίνουμε ότι:

* + σχεδόν το 50% των 10.000 νεογέννητων έχει βάρος μικρότερο των 3,2 κιλών και σχεδόν το 50% έχει βάρος μεγαλύτερο των 3,2 κιλών,
	+ η κατανομή των βαρών «γύρω από τα 3,2 κιλά» γίνεται με συμμετρικό τρόπο,
	+ καθώς κατευθυνόμαστε από τις ακραίες τιμές βάρους προς τα 3,2 κιλά (δηλα- δή προς το «κέντρο» της κατανομής), τα νεογέννητα που κατανέμονται στις κλάσεις είναι περισσότερα.

10

9

8

7

6

5

4

3

2

1

0

**Πολύγωνο σχετικών συχνοτήτων των βαρών 10.000 νεογέννητων στις ΗΠΑ**

**Βάρος νεογέννητου σε κιλά**

**Νεογέννητα %**

Στη φύση συχνά συναντούμε κατανομές συχνοτήτων μεγάλων δειγμάτων, των οποίων τα ιστογράμματα και τα πολύγωνα σχετικών συχνοτήτων έχουν παρόμοια μορφή και ιδιότητες με το προηγούμενο παράδειγμα. Σε τέτοιες περιπτώσεις, όταν το πλήθος είναι αρκετά μεγάλο και το πλήθος των κλάσεων κατάλληλα μεγάλο, το πολύγωνο σχετικών συχνοτήτων τείνει να μοιάσει με τη γραφική παράσταση της

 συνάρτησης y = $\frac{e^{- \frac{1}{2}\left(\frac{x-μ}{σ}\right)^{2}}}{σ\sqrt{2π}}$ , για κάποιες τιμές των παραμέτρων μ και σ (η σ είναι θετική).

και σ (η σ είναι θετική). Αυτή τη γραφική παράσταση την ονομάζουμε γκαουσιανή καμπύλη και φαίνεται στο παρακάτω σχήμα.

μ

Η ευθεία x = μ είναι ο άξονας συμμετρίας της γραφικής παράστασης της συνάρ- τησης. Συνεπώς, μιλώντας κάπως πρακτικά, η τιμή του σ καθορίζει πόσο «απλω- μένη» ή «μαζεμένη» είναι η καμπύλη γύρω από το μ και ποιο είναι το μέγιστο

«ύψος» της. Η τιμή του μ επηρεάζει τη «θέση» της καμπύλης καθώς καθορίζει τη θέση του άξονα συμμετρίας της. Όταν το σ μεγαλώνει, το σχήμα της μοιάζει να είναι πιο «απλωμένο» πάνω από τον οριζόντιο άξονα και το μέγιστο «ύψος» της καμπύλης μικραίνει. Τα αντίστροφα συμβαίνουν όταν το σ μικραίνει. Ο ρόλος των τιμών των παραμέτρων μ και σ στη μορφή της γραφικής παράστασης φαίνεται και στα παρακάτω παραδείγματα.



|  |  |  |
| --- | --- | --- |
|  |  |  |

y = $\frac{e^{- \frac{1}{2}\left(\frac{x-μ}{σ}\right)^{2}}}{σ\sqrt{2π}}$

|  |  |  |
| --- | --- | --- |
|  |  |  |

Η κανονική κατανομή είναι το μαθηματικό μοντέλο που χρησιμοποιούμε για να περιγράψουμε πώς κατανέμονται σε έναν ιδεατό, άπειρο πληθυσμό οι τιμές με- ταβλητών, όπως εκείνες στο παράδειγμά μας. Για την κανονική κατανομή, το μ εκφράζει το κέντρο συμμετρίας της και το σ είναι ένας δείκτης διασποράς της.

Έστω μία μεταβλητή που μοντελοποιείται από την κανονική κατανομή με μ και σ. Αν επιλέξουμε τυχαία ένα στοιχείο/άτομο από τον άπειρο πληθυσμό και κοιτάξου- με την τιμή της μεταβλητής για αυτό το στοιχείο/άτομο, τότε η πιθανότητα αυτή η τιμή να βρίσκεται σε ένα διάστημα (α, β) ισούται με το εμβαδόν μεταξύ της γκα- ουσιανής καμπύλης και του άξονα x*'*x ανάμεσα στα α και β. Αποδεικνύεται ότι:

* η πιθανότητα του ενδεχομένου «η τιμή της μεταβλητής είναι μεγαλύτερη από μ» ισούται με 0,5 και, λόγω συμμετρίας, η πιθανότητα του ενδεχομένου «η τιμή της μεταβλητής είναι μικρότερη από μ» ισούται επίσης με 0,5.
* η πιθανότητα του ενδεχομένου «η τιμή της μεταβλητής είναι στο διάστημα (μ - σ, μ + σ) »

 ισούται κατά προσέγγιση με 0,68 ή 68%.

* η πιθανότητα του ενδεχομένου «η τιμή της μεταβλητής είναι στο διάστημα (μ - 2σ , μ + 2σ) » ισούται κατά προσέγγιση με 0,95 ή 95%
* η πιθανότητα του ενδεχομένου «η τιμή της μεταβλητής είναι στο διάστημα (μ - 3σ , μ + 3σ)» ισούται κατά προσέγγιση με 0,997 ή 99,7%.

**μ-3σ μ-2σ μ-σ μ μ+σ μ+2σ μ+3σ**

Από τα προηγούμενα και λόγω συμμετρίας της κανονικής κατανομής, υπολογίζουμε τις πιθανότητες και άλλων ενδεχομένων. Π.χ. η πιθανότητα του ενδεχομένου «η τιμή της μεταβλητής είναι στο διάστημα (μ - σ , μ)» ισούται κατά προσέγγιση με 0,68 : 2 = 0,34.

Λόγω στατιστικής ομαλότητας, σε ένα μεγάλο δείγμα (ή σε ολόκληρο τον υπαρκτό πληθυσμό) μπορούμε να εκτιμήσουμε ότι:

* το ποσοστό των στοιχείων/ατόμων για τα οποία η τιμή της μεταβλητής είναι στο διάστημα

(μ - σ , μ + σ), είναι περίπου το 68% του πληθυσμού

* το ποσοστό των στοιχείων/ατόμων για τα οποία η τιμή της μεταβλητής είναι στο διάστημα

(μ - 2σ , μ + 2σ), είναι περίπου το 95% του πληθυσμού

* το ποσοστό των στοιχείων/ατόμων για τα οποία η τιμή της μεταβλητής είναι στο διάστημα

(μ - 3σ , μ + 3σ), είναι περίπου το 99,7% του πληθυσμού.

Λόγω συμμετρίας μπορούμε να κάνουμε και άλλες παρόμοιες εκτιμήσεις.

Επίσης, στην περίπτωση μεγάλου δείγματος (ή ολόκληρου του υπαρκτού πληθυσμού) η δειγματική μέση τιμή της μεταβλητής είναι κοντά στο μ και η δειγματική τυπική απόκλισή της είναι κοντά στο σ.

Για παράδειγμα, στην περίπτωση του δείγματος των νεογέννητων, η μεταβλητή «βάρος νεογέννητου στις ΗΠΑ» μοντελοποιείται από την κανονική κατανομή με μ = 3,18 και σ = 0,91 (βλέπε το παρακάτω σχήμα). Η μέση τιμή του βάρος των νεογέννητων του δείγματος είναι κοντά στο 3,18 και η τυπική απόκλισή του είναι κοντά στο 0,91. Το ίδιο ισχύει και για τη μέση τιμή και την τυπική απόκλιση του βάρους ολόκληρου του υπαρκτού πληθυσμού των νεογέννητων στις ΗΠΑ.

0 1 2 3 4 5 6

# ΕΝOΤΗΤΑ 2.5 ΠΙΝΑΚΕΣ ΣΥΝΑΦΕΙΑΣ ΚΑΙ ΡΑΒΔΟΓΡΑΜΜΑΤΑ

##  Βασικές μαθηματικές έννοιες

 ***Πίνακες συνάφειας συχνοτήτων***

Στα διάφορα προβλήματα που εξετάσαμε έως τώρα στη Στατιστική ασχοληθήκαμε κάθε φορά με μία μόνο μεταβλητή ξεχωριστά, ποιοτική ή ποσοτική, π.χ. με το φύλο των μαθητών, με τον χρόνο των αθλητών σε έναν αγώνα δρόμου, τον μηνιαίο μισθό των υπαλλήλων μιας εταιρείας. Μια πραγματική στατιστική ανάλυση δεν περιορίζεται συνήθως στη μελέτη μιας μεταβλητής, αλλά απαιτείται η μελέτη της σχέσης μεταξύ δύο ή περισσοτέρων μεταβλητών. Στην παράγραφο αυτή θα ασχοληθούμε με την αναζήτηση σχέσεων ανάμεσα σε δύο ποιοτικές μεταβλητές.

Ας δούμε τις βασικές ιδέες της διερεύνησης ενός προβλήματος με δύο ποιοτικές μεταβλητές μέσα από ένα παράδειγμα. Σε μια μελέτη που αφορά τη χρήση ζώνης ασφαλείας από οδηγούς ηλικίας 18 – 24 ετών μετρήθηκαν 198 οδηγοί σύμφωνα με το φύλο και το ενδεχόμενο να είχαν ή όχι κάποιο τροχαίο ατύχημα τα τελευταία πέντε χρόνια. Η εύρεση της πιθανής σχέσης μεταξύ του φύλου (χαρακτηριστικό Α) και τροχαίου ατυχήματος (χαρακτηριστικό Β) επιτυγχάνεται μέσω της κατα- σκευής του **πίνακα συνάφειας** **συχνοτήτων**, ο οποίος είναι ένας πίνακας συχνοτήτων της μορφής:

|  |  |
| --- | --- |
|  | **Τροχαίο Ατύχημα** |
| **Ναι** | **Όχι** | **Σύνολο** |
| **Φύλο** | **Άνδρες** | **69** | **58** | **127** |
| **Γυναίκες** | **27** | **44** | **71** |
| **Σύνολο** | **96** | **102** | **198** |

*Πίνακας 1: Πίνακας Συνάφειας απόλυτων συχνοτήτων φύλου και συμμετοχής σε ατύχημα.*

Οι γραμμές του πίνακα συνάφειας (Πίνακας 1) αντιστοιχούν στις κατηγορίες της μίας ποιοτικής μεταβλητής (π.χ. του φύλου) και οι στήλες στις κατηγορίες της άλλης ποιοτικής μεταβλητής (π.χ. του ατυχήματος), ενώ στο εσωτερικό του παρα- τίθενται οι συχνότητες που αντιστοιχούν σε όλους τους δυνατούς συνδυασμούς των κατηγοριών των δύο μεταβλητών.

Το πλήθος των ατόμων στις κατηγορίες του Πίνακα 1 είναι διαφορετικό, οπότε δεν μπορούμε να εξηγήσουμε ποια είναι η σχέση μεταξύ φύλου και συμμετοχής σε ατύχημα. Προκειμένου να κατανοήσουμε καλύτερα αυτή τη σχέση, θα πρέπει να εξετάσουμε τα ποσοστά αντί για τις συχνότητες. Ο πίνακας 1 μπορεί να μετατραπεί σε πίνακα σχετικών συχνοτήτων (%) ως προς το σύνολο των παρατηρήσεων του δείγματος ν (Πίνακας 2), ως προς το πλήθος των παρατηρήσεων του χαρακτηριστικού Α (Πίνακας 3) και ως προς το πλήθος των παρατηρήσεων του χαρακτηριστικού Β (Πίνακας 4). Από τον Πίνακα 2, προκύπτει ότι :

* + - Στο δείγμα της έρευνάς μας συμμετείχαν περισσότεροι άνδρες οδηγοί από γυναίκες οδηγούς.
		- Οι περισσότεροι συμμετέχοντες στην έρευνα δε συμμετείχαν σε ατύχημα, ωστόσο δεν είναι και λίγοι εκείνοι που συμμετείχαν.

|  |  |
| --- | --- |
|  | **Τροχαίο Ατύχημα** |
| **Ναι** | **Όχι** | **Σύνολο** |
| **Φύλο** | **Άνδρες** | **34,9%** | **29,3%** | **64,2%** |
| **Γυναίκες** | **13,6%** | **22,2%** | **35,8%** |
| **Σύνολο** | **48,5%** | **51,5%** | **100%** |

*Πίνακας 2: Πίνακας Συνάφειας σχετικών συχνοτήτων φύλου και ατυχήματος ως προς το σύνολο των παρατηρήσεων του δείγματος*

|  |  |
| --- | --- |
|  | **Τροχαίο Ατύχημα** |
| **Ναι** | **Όχι** | **Σύνολο** |
| **Φύλο** | **Άνδρες** | **54,3%** | **45,7%** | **100%** |
| **Γυναίκες** | **38%** | **62%** | **100%** |
| **Σύνολο** |  |  |  |

Η εμφάνιση των ποσοστών στο εσωτερικό ενός πίνακα συνάφειας μας δίνει τη δυνατότητα να διακρίνουμε τη μορφή της σχέσης που ενδεχομένως υπάρχει μεταξύ των μεταβλητών ενός πίνακα. Στην περίπτωση δύο μεταβλητών, η μια μεταβλητή μπορεί να παίξει τον ρόλο της ανεξάρτητης μεταβλητής (το φύλο) και η άλλη να παίξει τον ρόλο της εξαρτημένης (από το φύλο) μεταβλητής (το τροχαίο ατύχημα).

*Πίνακας 3: Πίνακας Συνάφειας σχετικών συχνοτήτων φύλου και συμμετοχής σε ατύχημα ως προς το φύλο*

Από τον Πίνακα 3 φαίνεται ότι οι άνδρες οδηγοί τους δείγματος είχαν σε μεγαλύτερο ποσοστό ατυχήματα τα τελευταία πέντε χρόνια, σε σχέση με τις γυναίκες οδηγούς. Φαίνεται πως στο δείγμα μας υπάρχει μια σχέση ανάμεσα στο φύλο και στη συμμετοχή σε ατύχημα στο ιστορικό του οδηγού: οι άνδρες οδηγοί του δείγματος είναι πιο πιθανό να είχαν ατύχημα τα τελευταία 5 χρόνια. Το αν αυτή η σχέση που ανακαλύψαμε στο δείγμα μας θα μπορούσε να γενικευθεί στον πληθυσμό όλων των οδηγών δεν είναι προφανές. Εξαρτάται πρώτα απ’ όλα από το αν το δείγμα που επιλέξαμε είναι αντιπροσωπευτικό του πληθυσμού. Ακόμα κι αν αυτό έχει εξασφαλιστεί με τη δειγματοληψία, το εύρημά μας θα μπορούσε να είναι τυχαίο (έτυχε ανάμεσα στους άνδρες οδηγούς που επιλέξαμε στο δείγμα να υπάρχουν πολλοί με ατύχημα τα τελευταία χρόνια σε σχέση με τον πληθυσμό των ανδρών ή κάτι αντίστοιχο για τις γυναίκες). Όσο μεγαλύτερο είναι το μέγεθος του δείγματος και όσο πιο μεγάλη είναι η διαφορά που βρήκαμε στο δείγμα μας, τόσο λιγότερο πιθανό είναι το εύρημα να είναι τυχαίο και με μεγαλύτερη εμπιστοσύνη μπορούμε να ισχυριστούμε ότι το εύρημα στο δείγμα μας αντανακλά μια διαφορά στον πληθυσμό.

|  |  |
| --- | --- |
|  | **Τροχαίο Ατύχημα** |
| **Ναι** | **Όχι** | **Σύνολο** |
| **Φύλο** | **Άνδρες** | **71,9%** | **56,9%** |  |
| **Γυναίκες** | **28,1%** | **43,1%** |  |
| **Σύνολο** | **100%** | **100%** |  |

*Πίνακας 4: Πίνακας Συνάφειας σχετικών συχνοτήτων φύλου και της συμμετοχής σε ατύχη- μα ως προς τη συμμετοχή σε ατύχημα*

Από τον πίνακα 4 φαίνεται ότι τόσο στο σύνολο των οδηγών που είχαν ατύχημα όσο και στο σύνολο των οδηγών που δεν είχαν ατύχημα, οι περισσότεροι οδηγοί ήταν άνδρες.

 ***Στοιβαγμένο Ραβδόγραμμα – Ομαδοποιημένο ραβδόγραμμα***

Όπως είδαμε στην περίπτωση της μίας ποιοτικής μεταβλητής οι πληροφορίες που αντλούνται από τους πίνακες συχνοτήτων ή σχετικών συχνοτήτων αναπαριστώνται με τα ραβδογράμματα και τα κυκλικά διαγράμματα. Στην περίπτωση των δύο ποιοτικών μεταβλητών, οπτικοποιούμε τα αποτελέσματα ενός πίνακα συνάφειας συχνοτήτων με το **στοιβαγμένο ραβδόγραμμα** και τα αποτελέσματα ενός πίνακα συνάφειας σχετικών συχνοτήτων με το **ομαδοποιημένο ραβδόγραμμα**.

***Στοιβαγμένο Ραβδόγραμμα***

Το στοιβαγμένο ραβδόγραμμα συχνοτήτων (σχήμα 1) του πίνακα 2 αποτελείται από ορθογώνιες στήλες, οι βάσεις των οποίων βρίσκονται πάνω στον οριζόντιο άξονα, στον οποίο αντιστοιχούν οι κατηγορίες του ενός χαρακτηριστικού π.χ. του φύλου. Το πλήθος των ορθογώνιων στηλών είναι ίσο με το πλήθος των κατηγοριών του φύλου. Κάθε ορθογώνια στήλη διαιρείται σε τόσα τμήματα διαφορετικού χρώματος, όσες και οι κατηγορίες του άλλου χαρακτηριστικού (ατυχήματος). Συνήθως, στον οριζόντιο άξονα τοποθετούμε τις κατηγορίες της ανεξάρτητης μεταβλητής (φύλου, στο παράδειγμά μας) και στον κατακόρυφο άξονα τις κατηγορίες της εξαρτημένης μεταβλητής (ατυχήματος, στο παράδειγμά μας) ώστε να ελέγξουμε εάν το φύλο σχετίζεται με τα τροχαία ατυχήματα.

140

**27**

**44**

**69**

**58**

120

100

80

60

40

20

0

Άνδρες

Γυναίκες

Ναι Όχι

*Σχήμα 1: Στοιβαγμένο ραβδόγραμμα συχνοτήτων φύλου και συμμετοχής σε ατύχημα του Πίνακα 1*

***Ομαδοποιημένο Ραβδόγραμμα***

Το ομαδοποιημένο ραβδόγραμμα σχετικών συχνοτήτων αποτελείται από ορθογώνιες στήλες, οι βάσεις των οποίων βρίσκονται πάνω στον οριζόντιο άξονα στον οποίο αντιστοιχούν οι κατηγορίες του ενός χαρακτηριστικού π.χ. του φύλου. Σε κάθε κατηγορία του φύλου αντιστοιχούν τόσες ορθογώνιες στήλες διαφορετικού χρώματος όσες το πλήθος των κατηγοριών του άλλου χαρακτηριστικού π.χ. της συμμετοχής σε ατύχημα. Το ύψος κάθε στήλης είναι ίσο με την αντίστοιχη σχετική συχνότητα.

40%

35%

30%

25%

20%

15%

10%

5%

0%

**34,9%**

**29,3%**

**22,2%**

**13,6%**

Άνδρες Γυναίκες

Ναι Όχι

*Σχήμα 2: Ομαδοποιημένο ραβδόγραμμα σχετικών συχνοτήτων φύλου και ατυχήματος του πίνακα 2*

Στο σχήμα 2 φαίνονται όλα τα ποσοστά του πίνακα 2. Για παράδειγμα, εάν αθροίσετε όλα τα ποσοστά που φαίνονται στο σχήμα προκύπτει το 100%, ενώ εάν αθροίσετε τα ποσοστά των δύο ράβδων στους άνδρες οδηγούς προκύπτει το 64,2% που είναι το ποσοστό τους στο δείγμα. Επίσης, παρατηρούμε ότι οι ράβδοι σε κάθε κατηγορία του φύλου δεν έχουν το ίδιο μοτίβο. Δηλαδή, στους άνδρες οδηγούς οι περισσότεροι είχαν ατύχημα, ενώ στις γυναίκες οδηγούς οι περισσότερες δεν είχαν ατύχημα.

Στα σχήματα 3 και 4 φαίνονται τα ομαδοποιημένα ραβδογράμματα σχετικών συχνοτήτων ως προς το φύλο και τη συμμετοχή σε ατύχημα, αντίστοιχα. Στο σχήμα 3 τα ποσοστά των ράβδων έχουν άθροισμα 100% τόσο ανάμεσα στους άνδρες όσο και ανάμεσα στις γυναίκες οδηγούς. Βλέπουμε, δηλαδή, στους άνδρες (γυναίκες) τα ποσοστά των οδηγών που είχαν/δεν είχαν ατύχημα στο σύνολο των ανδρών (γυναικών) οδηγών.

70%

**62%**

**54,3%**

**45,7%**

**38%**

60%

50%

40%

30%

20%

10%

0%

Άνδρες

Γυναίκες

Ναι Όχι

*Σχήμα 3: Ομαδοποιημένο ραβδόγραμμα σχετικών συχνοτήτων φύλου και της συμμετοχής σε ατύχημα ως προς το φύλο (αντ. Πίνακα 3)*

Από το σχήμα 3, το μοτίβο των ράβδων σε κάθε κατηγορία του φύλου είναι δια- φορετικό, οπότε φαίνεται να υπάρχει μια σχέση ανάμεσα στο φύλο και στα ατυ- χήματα. Εάν το μοτίβο των ράβδων σε κάθε κατηγορία του φύλου ήταν περίπου ίδιο, αυτό θα ήταν μία ένδειξη ότι οι μεταβλητές φύλο και ατύχημα τα τελευταία πέντε χρόνια δε σχετίζονται.

Στο σχήμα 4 τα ποσοστά των ομοιόμορφα χρωματισμένων ράβδων έχουν άθροισμα 100%. Βλέπουμε, δηλαδή, τα ποσοστά των μπλε ράβδων να αντιστοιχούν στους άνδρες και γυναίκες οδηγούς που είχαν ατύχημα στο σύνολο των οδηγών που είχαν ατύχημα, ενώ τα ποσοστά των ροζ ράβδων αντιστοιχούν στους άνδρες και γυναίκες οδηγούς που δεν είχαν ατύχημα στο σύνολο των οδηγών που δεν είχαν ατύχημα. Από αυτό το σχήμα φαίνεται πιο εύκολα ότι το φύλο και η συμμετοχή σε ατύχημα σχετίζονται, αφού το ύψος των ράβδων σε κάθε κατηγορία του φύλου δεν είναι περίπου το ίδιο.

80%

70%

60%

50%

40%

30%

20%

10%

0%

71,9%

56,9%

43,1%

28,1%

Άνδρες Γυναίκες

Ναι Όχι

*Σχήμα 4: Ομαδοποιημένο ραβδόγραμμα σχετικών συχνοτήτων φύλου και συμμετοχής σε ατύχημα ως προς τη συμμετοχή σε ατύχημα (αντ. Πίνακα 4)*

***Σχέση Αιτιότητας***

Με ανάλογο τρόπο προκύπτουν στοιβαγμένα και ομαδοποιημένα ραβδογράμματα, εάν στον οριζόντιο άξονα τοποθετήσουμε τις κατηγορίες του ατυχήματος. Στο σημείο αυτό, καλό είναι να γίνει η παρακάτω διευκρίνιση. Το γεγονός ότι δύο μεταβλητές φαίνεται να σχετίζονται, όπως στο παράδειγμα μας, είναι λάθος να ερμηνεύεται με όρους αιτίου και αποτελέσματος, δηλαδή η ανεξάρτητη μεταβλητή να είναι η αιτία που προκλήθηκε η εξαρτημένη μεταβλητή (το αποτέλεσμα). Στο παράδειγμά μας δε θα ήταν σωστό για μια ασφαλιστική εταιρεία να ερμηνεύσει το αποτέλεσμα ως ότι το φύλο ενός οδηγού είναι υπεύθυνο για την πρόκληση ατυχημάτων. Με δεδομένο, λοιπόν, ότι δύο μεταβλητές Α και Β σχετίζονται, αυτό μπορεί να σημαίνει ότι:

* η μεταβλητή Α είναι η αιτία για τη μεταβλητή Β
* η μεταβλητή Β είναι η αιτία για τη μεταβλητή Α
* υπάρχει ένας τρίτος (συγχυτικός) παράγοντας ο οποίος να είναι η αιτία τόσο για το Α, όσο και για το Β.
* είναι απλά μια σύμπτωση, διότι απλά ένα τυχαίο γεγονός συνέβη στο δείγμα μας, ενώ στην πραγματικότητα, δηλαδή στον πληθυσμό, οι δύο μεταβλητές είναι ασυσχέτιστες.

Για παράδειγμα, μια έρευνα έδειξε ότι τα μικρά παιδιά που κοιμούνται με αναμμένο το φως είναι πολύ πιθανότερο να εμφανίσουν μυωπία αργότερα. Μπορούμε άραγε να συμπεράνουμε ότι ο ύπνος με αναμμένο το φως προκαλεί μυωπία; Μεταγενέστερη έρευνα έδειξε ότι υπάρχει μια ισχυρή σχέση μεταξύ της γονικής μυωπίας και της εμφάνισης μυωπίας στα παιδιά τους, σημειώνοντας επίσης ότι οι μυωπικοί γονείς είναι πιο πιθανό να αφήσουν ένα φως στην κρεβατοκάμαρα των παιδιών τους. Ο συγχυτικός παράγοντας εδώ είναι η γονική μυωπία που επηρεάζει τόσο τη μεταβλητή «αναμμένο φως» όσο και τη μεταβλητή «μυωπία παιδιών».

 ***Κατανομή μαθητών/μαθητριών σε ομάδες προσανατολισμού σπουδών***

Σε ένα Γενικό Λύκειο καταγράφηκαν το φύλο και η ομάδα προσανατολισμού σπουδών που επέλεξαν οι 50 μαθητές/μαθήτριες της Β’ τάξης. Συγκεκριμένα, 11 αγόρια επέλεξαν την ομάδα προσανατολισμού Ανθρωπιστικών Σπουδών και 18 αγόρια επέλεξαν την ομάδα προσανατολισμού Θετικών Σπουδών. Αντίστοιχα, 9 κορίτσια επέλεξαν την ομάδα προσανατολισμού Ανθρωπιστικών Σπουδών και 12 κορίτσια την ομάδα προσανατολισμού Θετικών Σπουδών.

**α)** Να οργανώσετε τα παραπάνω δεδομένα σε έναν πίνακα συνάφειας (όπως δίνεται παρακάτω) που να περιέχει τις απόλυτες συχνότητες και τις σχετικές συχνότητες:

* + ως προς το σύνολο των μαθητών/τριών της Β΄ Λυκείου
	+ ως προς το φύλο (γραμμές)
	+ ως προς την ομάδα προσανατολισμού Σπουδών (στήλες)

|  |  |
| --- | --- |
|  | **Ομάδα Προσανατολισμού Σπουδών** |
| **Ανθρωπιστικών** | **Θετικών** | **Σύνολο** |
| **Φύλο** | **Αγόρια** |  |  |  |
| **Κορίτσια** |  |  |  |
| **Σύνολο** |  |  |  |

**β)** Να κατασκευάσετε κατάλληλα γραφήματα συχνοτήτων και σχετικών συχνοτήτων

* + ως προς το σύνολο των μαθητών/τριών της Β Λυκείου
	+ ως προς το φύλο (γραμμές)
	+ ως προς την ομάδα προσανατολισμού Σπουδών (στήλες)

**γ)** Θα μπορούσατε να πείτε εάν το φύλο παίζει ρόλο στην επιλογή της ομάδας προσανατολισμού σπουδών;

# ΕΝOΤΗΤΑ 2.6 ΣΥΓΚΡΙΣΗ ΠΟΣΟΤΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΣΤΙΣ ΚΑΤΗΓΟΡΙΕΣ ΕΝΟΣ ΠΟΙΟΤΙΚΟΥ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ

##  Βασικές μαθηματικές έννοιες

Σε προηγούμενη παράγραφο είδαμε αναλυτικά τα μέτρα θέσης και μεταβλητότητας για μια μεταβλητή. Εδώ, θα τα χρησιμοποιήσουμε για να συγκρίνουμε ομάδες δεδο- μένων.

Για παράδειγμα, παρακάτω βλέπουμε τα γκολ που έχουν σημειώσει οι παίκτες δύο ομάδων χάντμπωλ σε δέκα αγώνες. Η ομάδα Α έχει 20 παίκτες, ενώ η ομάδα Β έχει 18.

Ομάδα Α: 0, 0, 0, 1, 2, 5, 7, 8, 10, 11, 11, 17, 18, 18, 19, 19, 24, 26, 27, 27.

Ομάδα Β: 0, 0, 1, 4, 5, 7, 8, 10, 13, 15, 17, 20, 21, 22, 25, 26, 27, 28.

Οι δύο ομάδες έχουν σημειώσει συνολικά περίπου τα ίδια γκολ (250 η ομάδα Α και 249 η ομάδα Β). Επίσης, ο μέσος όρος των γκολ της ομάδας Α είναι 12,5 και της ομάδας Β είναι 13,83, περίπου. Η διαφορά των μέσων όρων προκύπτει από το γεγονός ότι η ομάδα Β, αν και έχει δύο παίκτες λιγότερους, έχει πετύχει σχεδόν τα ίδια γκολ. Μπορούμε, όμως με μεγαλύτερη λεπτομέρεια να δούμε, πώς η ομάδα Β το καταφέρνει αυτό; Στην προσπάθειά μας να συγκρίνουμε τις δύο ομάδες με κριτήριο τα γκολ που πέτυχε κάθε παίκτης, το παρακάτω θηκόγραμμα δίνει μία αρκετά σαφή εικόνα. Από τη θέση των δύο ορθογωνίων φαίνεται ότι οι «μεσαίοι σκόρερ» της ομάδας Β (δηλαδή εκείνοι που πετυχαίνουν έναν ούτε μεγάλο ούτε μικρό αριθμό γκολ») έχουν πετύχει περισσότερα γκολ από τους «μεσαίους σκόρερ» της ομάδας Α. Αυτό, ενδεχομένως, εξηγεί και γιατί η ομάδα Β με δύο λιγότερους παίκτες έχει πετύχει σχεδόν τα ίδια γκολ. Επίσης, συγκρίνοντας τα μήκη των ορθογωνίων βλέπουμε ότι για τους «μεσαίους σκόρερ» υπάρχει μεγαλύτερη μεταβολή των γκολ στην ομάδα Β (μεγαλύτερο μήκος), ενώ από τα μήκη των θηκογραμμάτων φαίνεται ότι το εύρος των γκολ της ομάδας Β είναι κατά 1 γκολ μεγαλύτερο (κάτι που οφείλεται στον παίκτη που σημείωσε 28 γκολ).



###### 0 5 10 15 20 25 30

Σε ένα άλλο παράδειγμα, ο οργανισμός συγκοινωνιών σε ένα νησί του Αιγαίου μετρά τη διάρκεια ενός συγκεκριμένου δρομολογίου σε λεπτά (min) καθημερινά για τους μήνες Μάιο και Ιούλιο του ίδιου έτους, προκειμένου να συγκρίνει το επίπεδο υπηρεσιών του μεταξύ Μαΐου και Ιουλίου. Τα δειγματικά στατιστικά μέτρα για κάθε μήνα φαίνονται στον επόμενο πίνακα 1.

|  |  |  |
| --- | --- | --- |
|  | **Μάιος** | **Ιούλιος** |
| x | 106,67 min | 107,33 min |
| **δ(Q2)** | 101 min | 106 min |
| **Q1** | 95 min | 95 min |
| **Q3** | 108 min | 118 min |
| **xmin** | 81 min | 85 min |
| **xmax** | 167 min | 155 min |
| **S** | 23,57 min | 16,15 min |

*Πίνακας 1: Δειγματικά στατιστικά μέτρα της διάρκειας ενός συγκεκριμένου δρομολογίου για τους μήνες Μάιο και Ιούλιο*

Η δειγματική μέση τιμή της διάρκειας του συγκεκριμένου δρομολογίου του Μαΐου είναι 106,67 λεπτά, ενώ η αντίστοιχη δειγματική μέση τιμή για τον Ιούλιο είναι 107,33 λεπτά. Εκ πρώτης όψεως θα μπορούσε κάποιος να πει ότι συγκρίνοντας τις δύο δειγματικές μέσες τιμές η μέση διάρκεια του συγκεκριμένου δρομολογίου δεν άλλαξε αισθητά, οπότε η εταιρεία δε χρειάζεται να προχωρήσει σε κάποια βελτιωτική ενέργεια.

Η δειγματική διάμεσος για τον μήνα Μάιο είναι 101 λεπτά, ενώ για τον Ιούλιο η αντίστοιχη τιμή είναι 106 λεπτά. Αυτό σημαίνει ότι τον Μάιο τα μισά περίπου συγκεκριμένα δρομολόγια εκτελούνται σε χρόνο το πολύ 101 λεπτά, ενώ τα υπόλοιπα σε χρόνο τουλάχιστον 101 λεπτών. Αντίστοιχα, τον Ιούλιο τα μισά περίπου δρομολόγια εκτελούνται το πολύ σε 106 λεπτά, ενώ τα υπόλοιπα σε 106 λεπτά και περισσότερο.

Το εύρος της διάρκειας του δρομολογίου τον Μάιο είναι 86 λεπτά και τον Ιούλιο είναι 70 λεπτά, κάτι που σημαίνει ότι η διαφορά της διάρκειας του συντομότερου με το μακρύτερο δρομολόγιο τον Μάιο είναι μεγαλύτερη. Από την άλλη μεριά, το ενδοτεταρτημοριακό εύρος της διάρκειας του δρομολογίου για το Μάιο είναι 13 λεπτά, ενώ για τον Ιούλιο είναι 23 λεπτά, κάτι που δηλώνει ότι το διάστημα [Q1, Q3] είναι κατά 10 λεπτά μεγαλύτερο, από το αντίστοιχο του Μαΐου. Επομένως, παρατηρείται πιο σημαντική μεταβολή για τη διάρκεια των δρομολογίων του Ιουλίου, που βρίσκονται στο διάστημα [Q1, Q3], δηλαδή για τα δρομολόγια, τα οποία μπορούν να χαρακτηριστούν ως «μέσης διάρκειας» δρομολόγια, σε σύγκριση με τα αντίστοιχα δρομολόγια του Μαΐου.

Αν δει κανείς, επιπλέον, τις τιμές των δύο δειγματικών τυπικών αποκλίσεων (23,57 λεπτά και 16,15 λεπτά), φαίνεται ότι οι διάρκειες των συνολικών δρομολογίων του Μαΐου ήταν πιο «διάσπαρτες» γύρω από τη μέση διάρκεια. Τέλος, αν υπολογίσουμε τους συντελεστές μεταβλητότητας για κάθε μήνα, τότε προκύπτουν CVM = 22,1% περίπου και CVI = 15%, περίπου. Άρα, φαίνεται ότι η διάρκεια των δρομολογίων του Μαΐου, συνολικά, παρουσιάζει μεγαλύτερη μεταβλητότητα και μικρότερη ομοιογένεια από αυτά του Ιουλίου.

Βλέπουμε ότι συμπεράσματα που βγαίνουν από τη μελέτη των δειγμάτων είναι αρκετά και σε διαφορετικά επίπεδα. Για να τα αποτυπώσουμε, δίνοντας μία πλήρη και εύληπτη εικόνα, μπορούμε να σχεδιάσουμε δύο θηκογράμματα. Το θηκόγραμμα είναι μια γραφική παράσταση που μας βοηθάει να απεικονίσουμε την κατανομή μιας ποσοτικής μεταβλητής στις κατηγορίες μιας ποιοτικής μεταβλητής. Από τη διάμεσο μπορούμε να πάρουμε μια ιδέα για την κεντρική τάση των δεδομένων, ενώ από το μήκος του (ολόκληρου, αλλά και του ορθογωνίου) μπορούμε να δούμε πόσο ποικίλλουν οι τιμές (μεταβλητότητα). Έτσι, λοιπόν, στο σχήμα 1 φαίνονται σχηματικά όσα περιγράψαμε προηγουμένως για τα δρομολόγια Μαΐου και Ιουνίου. Η μεταβλητότητα της διάρκειας του δρομολογίου τον Ιούλιο είναι μεγαλύτερη από τον μήνα Μάιο και για τα δρομολόγια «μέσης διάρκειας» (δηλαδή εντός ενδοτεταρτημοριακού εύρους), όπως φαίνεται από το μήκος του ορθογωνίου, αλλά και για όλα τα «τυπικά» δρομολόγια, όπως φαίνεται από το συνολικό μήκος των οριζοντίων γραμμών των θηκογραμμάτων. Επίσης, μπορού- με να εντοπίσουμε τις ακραίες τιμές της διάρκειας των δρομολογίων. Τον Ιούλιο έχουμε μόνο μία, ενώ τον Μάιο έχουμε τρεις ακραίες τιμές (από τις οποίες οι δύο είναι μεγαλύτερες από αυτή του Μαΐου). Ενδεχομένως, να συνέβη κάτι έκτακτο σε εκείνα τα δρομολόγια. Λόγω των ακραίων τιμών, το εύρος της διάρκειας των δρομολογίων τον Μάιο είναι μεγαλύτερο, κάτι που φαίνεται και στα θηκογράμματα.

Κοιτώντας μόνο τα θηκογράμματα, φαίνεται τα «τυπικά» δρομολόγια τον Ιούλιο να κρατάνε λίγο περισσότερο και η διάρκειά τους να έχει λίγο μεγαλύτερη διασπορά.

**60 80 100 120 140 160 180**

*Σχήμα 1: Θηκόγραμμα της διάρκειας ενός συγκεκριμένου δρομολογίου για τους μήνες Μάιο και Ιούλιο*

Αυτό, ενδεχομένως συμβαίνει γιατί τον Ιούλιο το λεωφορείο είναι περισσότερο γεμάτο και χρειάζεται περισσότερος χρόνος για την επιβίβαση/αποβίβαση, ενώ και η κίνηση στο νησί μπορεί να είναι μεγαλύτερη και να κάνει τη διάρκεια των δρομολογίων απρόβλεπτη. Ωστόσο, κοιτώντας τους μέσους χρόνους των δρομολογίων τους δύο μήνες, βλέπουμε ότι δε διαφέρουν πολύ. Άρα, αν συγκρίναμε μόνο τους μέσους χρόνους, χωρίς τα θηκογράμματα, δε θα είχαμε αυτή την εικόνα. Ο λόγος που οι μέσοι χρόνοι δε διαφέρουν πολύ είναι διότι επηρεάζονται από τις ακραίες τιμές και τον Μάιο είχαμε περισσότερες, όπως είπαμε.

# ΕΝΟΤΗΤΑ 2.7 ΓΡΑΜΜΙΚΗ ΣΥΣΧΕΤΙΣΗ ΠΟΣΟΤΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ ΚΑΙ ΔΙΑΓΡΑΜΜΑΤΑ ΔΙΑΣΠΟΡΑΣ

##  Βασικές Μαθηματικές Έννοιες

Έχουμε δει μέχρι τώρα ότι η σχέση δύο ποιοτικών μεταβλητών περιγράφεται με τη βοήθεια των πινάκων συνάφειας και των κατάλληλων ραβδογραμμάτων, ενώ υπολογίζοντας τα μέτρα θέσης και μεταβλητότητας μιας ποσοτικής μεταβλητής στις κατηγορίες μιας ποιοτικής μπορούμε να συγκρίνουμε ομάδες δεδομένων. Τι συμβαίνει εάν έχουμε δύο ποσοτικές μεταβλητές; Πώς μπορούμε να περιγράψουμε τη σχέση δύο ποσοτικών μεταβλητών; Για παράδειγμα:

* Η ηλικία και το βάρος ενός παιδιού έχουν κάποια θετική συσχέτιση μεταξύ τους, με την έννοια ότι όσο πιο μεγάλο είναι το παιδί τόσο μεγαλύτερο βάρος θα έχει.
* Η διάρκεια ζωής των ζώντων οργανισμών σε μια περιοχή και ένας δείκτης μόλυνσης της περιοχής έχουν αρνητική συσχέτιση μεταξύ τους, με την έννοια ότι όσο πιο μεγάλος είναι ο δείκτης μόλυνσης της περιοχής τόσο μικρότερη είναι η διάρκεια ζωής των οργανισμών που ζουν στην περιοχή.
* Το ύψος των αποδοχών των υπαλλήλων μιας εταιρείας δε συσχετίζεται με το βάρος τους.

Στην ενότητα αυτή θα εξετάσουμε μόνο την περίπτωση της γραμμικής συσχέτισης μεταξύ δύο ποσοτικών μεταβλητών. Για παράδειγμα, καταγράφουμε το ύψος (Χ σε cm) και το βάρος (Y σε kg) 18 μαθητών της Β΄ Λυκείου. Αυτό που μας ενδιαφέρει είναι τι συμβαίνει με το βάρος των παιδιών όταν αλλάζει το ύψος τους ή τι συμβαίνει με το ύψος των παιδιών όταν αλλάζει το βάρος τους. Αν μας ενδιαφέρει η πρώτη περίπτωση, θεωρούμε ως ανεξάρτητη μεταβλητή το ύψος των μαθητών και ως εξαρτημένη (από το ύψος) μεταβλητή το βάρος των μαθητών. Στη δεύτερη περίπτωση συμβαίνει το αντίθετο.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Μαθητής** | **Ύψος (Χ)** | **Βάρος (Υ)** | **Μαθητής** | **Ύψος (Χ)** | **Βάρος (Υ)** |
| **Α** | **170** | **58** | **Κ** | **178** | **79** |
| **Β** | **172** | **60** | **Λ** | **179** | **76** |
| **Γ** | **173** | **67** | **Μ** | **180** | **79** |
| **Δ** | **175** | **72** | **Ν** | **180** | **80** |
| **Ε** | **176** | **65** | **Ξ** | **180** | **83** |
| **Ζ** | **177** | **81** | **Ο** | **180** | **85** |
| **Η** | **178** | **73** | **Π** | **182** | **89** |
| **Θ** | **178** | **74** | **Ρ** | **187** | **90** |
| **Ι** | **178** | **78** | **Σ** | **191** | **92** |

*Πίνακας 1: Ύψη (Χ) και Βάρη (Υ) 18 μαθητών της Β΄ Λυκείου*

Για κάθε άτομο έχουμε δύο μετρήσεις: το ύψος και το βάρος του. Το δείγμα μας δηλαδή από-τελείται από τα ζεύγη τιμών (xi , yi) , i = 1,2,…,18 των συνεχών μεταβλητών Χ και Υ.

Η γραφική παράσταση των σημείων σε ένα ορθογώνιο σύστημα αξόνων ονομάζεται **διάγραμμα διασποράς** των μεταβλητών Χ και Υ.

Στο διάγραμμα παρατηρούμε μια διασπορά των σημείων που αντιστοιχούν στους μαθητές που εξετάζουμε. Από το διάγραμμα διασποράς μπορούμε να εντοπίσουμε την ύπαρξη συσχέτισης που ενδεχομένως να υπάρχει μεταξύ των μεταβλητών που εξετάζουμε. Η γενική εικόνα που παίρνουμε από το διάγραμμα διασποράς είναι ότι οι ψηλότεροι μαθητές είναι συνήθως και βαρύτεροι. Για παράδειγμα, ο μαθητής Ζ είναι ψηλότερος και βαρύτερος από τον μαθητή Ε, ο Η είναι ψηλότερος και βαρύτερος από τον Ε, αλλά υπάρχουν και εξαιρέσεις, όπως ο Η είναι ψηλότερος από τον Ζ αλλά ο Ζ είναι βαρύτερος από τον Η. Το άπλωμα των σημείων στο διάγραμμα θα μπορούσαμε να πούμε ότι γίνεται κοντά σε μια νοητή ευθεία γραμμή, οπότε η σχέση θεωρείται γραμμική.

90

85 Ζ

Βάρος μαθητών (Υ, kg)

80

75

70 Γ(173, 67) Η

65

60 Ε

55

50

Σ(191, 92)

169 171 173 175 177 179 181 183 185 187 189 191 193

Ύψος μαθητών (Χ, cm)

*Σχήμα 1: Διάγραμμα Διασποράς Ύψους και Βάρους 18 μαθητών της Β΄ λυκείου και ευθεία προσαρμοσμένη «με το μάτι» για τα δεδομένα του πίνακα 1*

***Συντελεστής γραμμικής συσχέτισης των ποσοτικών μεταβλητών Χ και Υ***

Ένα μέτρο που μας δίνει το μέγεθος της γραμμικής σχέσης των δύο ποσοτικών μεταβλητών Χ και Υ είναι ο συντελεστής γραμμικής συσχέτισης των Χ και Υ. Ο συντελεστής γραμμικής συσχέτισης δύο ποσοτικών μεταβλητών Χ και Υ ορίζεται με βάση ένα δείγμα ν ζευγών παρατηρήσεων (xi , yi) , i = 1,2,…,ν, συμβολίζεται με r (X , Y) ή απλά r και δίνεται από τον τύπο:

xiyi - νx y

r = i=1

νsXsY

ν

 (1)

όπου $\overbar{x}$ και sX η δειγματική μέση τιμή και η δειγματική τυπική απόκλιση της μεταβλητής Χ και $\overbar{y}$ και sY η δειγματική μέση τιμή και η δειγματική τυπική απόκλιση της μεταβλητής Y. Ωστόσο, σε πραγματικές μελέτες με πολλά δεδομένα, η χρήση λογιστικού φύλλου ή κάποιου στατιστικού λογισμικού κρίνεται απαραίτητη.

Ο συντελεστής γραμμικής συσχέτισης δύο ποσοτικών μεταβλητών Χ και Υ είναι καθαρός αριθμός, δηλαδή δεν εκφράζεται σε συγκεκριμένες μονάδες μέτρησης, επομένως είναι ανεξάρτητος των χρησιμοποιούμενων μονάδων μέτρησης των μεταβλητών Χ και Υ.

Ισχύει πάντοτε:

-1 ≤ r ≤ 1

Πιο συγκεκριμένα όταν:

* 0 < r < 1, τότε οι Χ και Υ είναι **θετικά γραμμικά συσχετισμένες**. Αυτό ση- μαίνει ότι, όταν οι τιμές της μεταβλητής Χ αυξάνονται, οι τιμές της Υ τείνουν να αυξάνονται. Όσο ο συντελεστής πλησιάζει το 1 τόσο πιο συγκεντρωμέ- νες γύρω από μια νοητή ευθεία είναι οι παρατηρήσεις. Επομένως, τόσο πιο ισχυρή είναι η θετική γραμμική συσχέτιση.
* -1 < r < 0, τότε οι Χ και Υ είναι **αρνητικά γραμμικά συσχετισμένες**. Αυτό σημαίνει ότι, όταν οι τιμές της μεταβλητής Χ αυξάνονται, οι τιμές της Υ τείνουν να μειώνονται. Όσο ο συντελεστής πλησιάζει το -1 τόσο πιο συγκε- ντρωμένες γύρω από μια νοητή ευθεία είναι οι παρατηρήσεις. Επομένως, τόσο πιο ισχυρή είναι η αρνητική γραμμική συσχέτιση.

* r = 1, τότε οι Χ και Υ είναι **τέλεια θετικά γραμμικά συσχετισμένες** και όλα τα σημεία βρίσκονται πάνω σε μια ευθεία με θετική κλίση.
* r = -1, τότε οι Χ και Υ είναι **τέλεια αρνητικά γραμμικά συσχετισμένες** και όλα τα σημεία βρίσκονται πάνω σε μια ευθεία με αρνητική κλίση.
* r = 0, τότε οι Χ και Υ είναι **γραμμικά ασυσχέτιστες, χωρίς αυτό να σημαί- νει ότι δεν μπορεί να έχουν κάποια άλλη σχέση, μη γραμμική**.

Από τα δεδομένα του παραδείγματος με τη χρήση αριθμομηχανής ή λογιστικού φύλλου έχουμε:

 $\sum\_{i=1}^{18}x\_{i}y\_{i}$ = 247324, $\overbar{x}$ = 178, 56, $\overbar{y}$ = 76, 72, sX = 4, 809, sY = 9, 473 ,

οπότε ο συντελεστής γραμμικής συσχέτισης του ύψους (Χ) και του βάρους (Υ) 18 μαθητών της Β Λυκείου είναι r = 0,90, που σημαίνει ότι το ύψος και το βάρος των μαθητών είναι ισχυρά θετικά γραμμικά συσχετισμένες μεταβλητές.

r = 0,50

r = 0,90



|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

r = -0,50

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

r = -0,90

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

r = 1



|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |   |  |  |
|  |  |  |  |

r = -1

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

r = 0

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

***Γραμμική συσχέτιση ποσοτικών μεταβλητών και αιτιότητα***

Στο σημείο αυτό θα υπενθυμίσουμε ότι η ισχυρή συσχέτιση μεταξύ δύο μεταβλητών δε συνεπάγεται υποχρεωτικά μια αιτιολογική σχέση ανάμεσα στις δύο μεταβλητές. Υπάρχει περίπτωση η αλλαγή της μεταβλητής Χ να προκαλεί άμεση αλλαγή της μεταβλητής Υ. Αλλά πολύ συχνά οι αλλαγές των δύο μεταβλητών Χ και Υ οφείλονται σε κάποιες άλλες μεταβλητές ή σε κάποιους αστάθμητους παράγοντες. Για παράδειγμα, το μέγεθος παπουτσιού που φορά ένας μαθητής και το μέγεθος των αρχείων που αποθηκεύονται στον υπολογιστή του, ενώ έχουν εξαι- ρετικά μεγάλη γραμμική συσχέτιση, προφανώς δεν έχουν αιτιώδη σχέση, αλλά επηρεάζονται από έναν τρίτο παράγοντα, την ηλικία (συγχυτικός παράγοντας).

Τα σημεία (xi, yi), i=1,2,…,18 είναι συγκεντρωμένα γύρω από μια ευθεία, όπως φαίνεται στο σχήμα 1. Στην πραγματικότητα, σχεδόν ποτέ δεν υπάρχει ευθεία που να περνάει απ’ όλα τα σημεία των δεδομένων. Για παράδειγμα, στο πρόβλημά μας με το ύψος και το βάρος των μαθητών, η αλήθεια είναι ότι η μεταβλητή Υ (το βάρος των μαθητών) μπορεί να επηρεάζεται και από άλλους παράγοντες εκτός της Χ (το ύψος των μαθητών), όπως, π.χ. τις διατροφικές συνήθειες. Γι’ αυτό μαθητές με το ίδιο ύψος μπορεί να έχουν διαφορετικό βάρος.

 ***Χάραξη ευθείας στο διάγραμμα διασποράς «με το μάτι»***

Ψάχνουμε, λοιπόν, για εκείνη την ευθεία που περνάει κοντά από τα περισσότερα σημεία των δεδομένων. Όπως γνωρίζουμε, η εξίσωση μιας ευθείας δίνεται από τη σχέση:

 y = α+βx, (2)

όπου α και β είναι παράμετροι τις οποίες θέλουμε να υπολογίσουμε ή, όπως λέμε, «να εκτιμήσουμε», έτσι ώστε η ευθεία που θα προκύψει από τη (2) να περιγράφει την αναμενόμενη τιμή y της μεταβλητής Υ, όταν η τιμή της μεταβλητής Χ είναι ίση με x.

Η παράμετρος α μας δίνει το σημείο (0, α) όπου η ευθεία αυτή τέμνει τον άξονα yʹy , ενώ η παράμετρος β παριστάνει τον συντελεστή διεύθυνσης της ευθείας.

Ο πιο εύκολος τρόπος χάραξης της ευθείας είναι αυτός που γίνεται «με το μάτι». Μια τέτοια ευθεία έχουμε φέρει στο διάγραμμα διασποράς του σχήματος 1. Για να βρούμε τα α και β επιλέγουμε δύο σημεία, έστω τα Γ(173,67) και Σ(191,92) πάνω στην ευθεία που φέραμε «με το μάτι» και καταλήγουμε στην εξίσωση:

 y = -173,28+1,39x (3)

Επομένως, η ευθεία που κατά τη γνώμη μας προσαρμόζεται καλύτερα στα σημεία του διαγράμματος διασποράς διέρχεται από το σημείο (0,-173,28) και έχει συντελεστή διεύθυνσης 1,39.

Την εξίσωση της ευθείας (3) μπορούμε να τη χρησιμοποιήσουμε για να εκτιμήσουμε μια τιμή της μεταβλητής Υ, από κάποια τιμή της μεταβλητής Χ που ανήκει στο εύρος τιμών της τελευταίας. Αυτό συμβαίνει, διότι –για τιμές μικρότερες ή μεγαλύτερες από αυτές που παρατηρήθηκαν– δεν είμαστε σίγουροι ότι η γραμμική σχέση παραμένει. Για παράδειγμα: ένας μαθητής με ύψος 1,81 εκτιμάται ότι θα έχει βάρος ίσο με y =-173,28+1,39∙181 = 78,1 κιλά.

Ο πιο εύκολος τρόπος χάραξης