

Στατιστική Ι

η προεγγραφή σφίλι

$\hat{Y}_i$ η πρόβλεψη της τιμής του Y το X

Σημειώσεις - Λογμένα θέματα
e-students.gr
☎ 698 93 94 741

Απλή Γραμμική Παλινδρόμηση

Η Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression) χρησιμοποιείται για να περιγράψουμε και να προβλέψουμε τη σχέση ανάμεσα σε δύο μεταβλητές:

- Ανεξάρτητη μεταβλητή (x): προβλεπτική / ερμηνευτική,
- Εξαρτημένη μεταβλητή (y): τιμή που θέλουμε να προβλέψουμε

Η υπόθεση είναι ότι η σχέση τους μπορεί να προσεγγιστεί από μία ευθεία:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

β_0 : σταθερός όρος

- β_0 «κατεβάζει» ή «ανεβάζει» όλη την συνθήκη στον άξονα y

β_1 : συντελεστής κλίση

- $\beta_1 > 0$: θετική σχέση (όσο αυξάνει το x , αυξάνει και το y)
- $\beta_1 < 0$: αρνητική σχέση (όσο αυξάνει το x , μειώνεται το y)
- $\beta_1 = 0$: δεν υπάρχει γραμμική σχέση

ε_i : σφάλμα (τυχαία απόκλιση)

Εισαγωγή στη Στατιστική

- Στατιστική: Επιστήμη που ασχολείται με τη συλλογή, οργάνωση, παρουσίαση, ανάλυση και ερμηνεία δεδομένων.
- Δεδομένα: Οι πληροφορίες που συγκεντρώνονται για μελέτη
- Πληθυσμός (Population): Το σύνολο όλων των στοιχείων που μελετάμε (π.χ. όλοι οι μαθητές ενός σχολείου).
- Δείγμα (Sample): Ένα υποσύνολο του πληθυσμού, που επιλέγεται για να μελετηθεί πιο εύκολα.
- Μεταβλητή (Variable): Ένα χαρακτηριστικό που μπορεί να πάρει διαφορετικές τιμές (π.χ. ύψος, ηλικία, βάρος).
 - Οι μεταβλητές χωρίζονται σε:
 1. Ποιοτικές (Qualitative): Δεν μετρώνται αριθμητικά (π.χ. φύλο, χρώμα ματιών).
 2. Ποσοτικές (Quantitative): Μετρώνται με αριθμούς (π.χ. βάρος, ηλικία).
 - Διακριτές: Παίρνουν συγκεκριμένες τιμές (π.χ. αριθμός παιδιών).
 - Συνεχείς: Παίρνουν οποιαδήποτε τιμή σε ένα διάστημα (π.χ. ύψος, χρόνος).
- Τύποι
 - Μέγεθος πληθυσμού: N
 - Μέγεθος δείγματος: n
 - Συχνότητα (frequency): Ο αριθμός φορών που εμφανίζεται μια τιμή (ν_i).
 - Σχετική συχνότητα:

$$f_i = \frac{\nu_i}{n}$$

όπου ν_i είναι η συχνότητα της τιμής και n το μέγεθος δείγματος.

- Αθροιστική συχνότητα (N_i)

$$N_i = \sum_{j=1}^i \nu_j$$

- Αθροιστική σχετική συχνότητα (F_i)

$$F_i = \frac{N_i}{n}$$

Γραφικές Παραστάσεις

1. Ραβδόγραμμα (Bar chart): Κάθε τιμή της μεταβλητής στον οριζόντιο άξονα, ύψος ράβδου = συχνότητα.
2. Κυκλικό διάγραμμα (Pie chart): Η κυκλική επιφάνεια χωρίζεται σε τομείς ανάλογους με τις σχετικές συχνότητες.
➤ Γωνία τομέα:

$$\theta_i = f_i \times 360^\circ$$



3. Ιστόγραμμα (Histogram): Για συνεχείς μεταβλητές, χρησιμοποιούνται διαστήματα τιμών (κλάσεις) αντί για μεμονωμένες τιμές.

Άσκηση

Σε μια τάξη 10 μαθητών οι ηλικίες είναι: 15, 15, 16, 14, 17, 15, 16, 14, 16, 15.

Ζητούνται:

1. Να βρεθούν οι συχνότητες, οι σχετικές συχνότητες των ηλικιών, η αθροιστική συχνότητα και η αθροιστική σχετική συχνότητα.
2. Να συμπληρωθεί ο πίνακας συχνοτήτων.
3. Να σχεδιαστεί ραβδόγραμμα.

Λύση

1. Συχνότητες v_i
Διαφορετικές τιμές (ταξινομημένες): 14, 15, 16, 17
Το 14: εμφανίζεται 2 φορές $\rightarrow v_{14}=2$
Το 15: εμφανίζεται 4 φορές $\rightarrow v_{15}=4$
Το 16: εμφανίζεται 3 φορές $\rightarrow v_{16}=3$
Το 17: εμφανίζεται 1 φορά $\rightarrow v_{17}=1$
Συνολικό μέγεθος δείγματος: $n=10$

Σχετικές συχνότητες $f_i = v_i / n$
 $f_{14} = v_{14} / n = 2 / 10 = 0,2 = 20\%$.
 $f_{15} = v_{15} / n = 4 / 10 = 0,4 = 40\%$.
 $f_{16} = v_{16} / n = 3 / 10 = 0,3 = 30\%$.
 $f_{17} = v_{17} / n = 1 / 10 = 0,1 = 10\%$.
 $\Sigma f_i = 0,2 + 0,4 + 0,3 + 0,1 = 1,0 = 100\%$

Αθροιστική συχνότητα N_i

$$N_{14}=v_{14}=2$$

(Πράξη: 2)

$$N_{14}+v_{15}=2+4=6.$$

(Πράξη: $2 + 4 = 6$)

$$N_{16}=N_{15}+v_{16}=6+3=9.$$

(Πράξη: $6 + 3 = 9$)

$$N_{17}=N_{16}+v_{17}=9+1=10.$$

(Πράξη: $9 + 1 = 10$)

Αθροιστική σχετική συχνότητα $F_i=N_i/n$

$$F_{14}=N_{14}/n=2/10=0,2=20\%.$$

$$F_{15}=N_{15}/n=6/10=0,6=60\%.$$

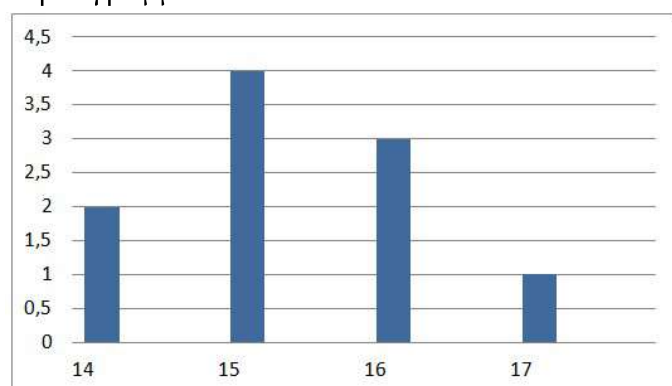
$$F_{16}=N_{16}/n=9/10=0,9=90\%.$$

$$F_{17}=N_{17}/n=10/10=1=100\%.$$

2. Πίνακας συχνοτήτων

Ηλικία (x_i)	Συχνότητα (v_i)	Σχετική συχνότητα (f_i)	Αθροιστική συχνότητα (N_i)	Αθροιστική σχετική (F_i)
14	2	0.20 (20%)	2	0.20
15	4	0.40 (40%)	6	0.60
16	3	0.30 (30%)	9	0.90
17	1	0.10 (10%)	10	1.00

3. Ραβδόγραμμα



Αριθμητικός Μέσος (Μέσος Όρος)

- Ο αριθμητικός μέσος (ή μέσος όρος) είναι το πηλίκο του αθροίσματος όλων των τιμών του δείγματος προς το πλήθος τους.
- Τύπος

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

$\bar{x}$: μέσος όρος,

x_i : η i -οστή παρατήρηση,

n : το πλήθος παρατηρήσεων,

$\sum x_i$: άθροισμα όλων των παρατηρήσεων.

Άσκηση

Σε ένα τμήμα 4 μαθητών οι βαθμοί ενός διαγωνίσματος ήταν: 5, 7, 8 και 10.
Να βρεθεί ο μέσος όρος των βαθμών.

Λύση

Καταγράφουμε τις τιμές: $x_1=5$, $x_2=7$, $x_3=8$, $x_4=10$

Υπολογίζουμε το άθροισμα:

$$\sum x_i = 5 + 7 + 8 + 10 = 30$$

Το πλήθος είναι $n=4$

Εφαρμόζουμε τον τύπο: $\bar{x} = \sum x_i / n = 30 / 4 = 7,5$

Ο μέσος όρος είναι 7,5.

Διάμεσος (Median)

- Η διάμεσος είναι η τιμή που χωρίζει τις παρατηρήσεις σε δύο ίσα μέρη αφού πρώτα ταξινομηθούν σε αύξουσα σειρά.
- Αν n είναι περιττός:

$$\text{Διάμεσος} = x_{\frac{n+1}{2}}$$

- Αν n είναι άρτιος:

$$\text{Διάμεσος} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

Άσκηση

Οι ηλικίες 5 παιδιών είναι: 2, 5, 7, 8, 10.

Να βρεθεί η διάμεσος.

Λύση

Το πλήθος είναι $n=5$ (περιττός αριθμός).

Θέση διαμέσου: $(n+1)/2=5+1/2=3$

Η 3^η τιμή της ταξινομημένης σειράς είναι 7.

Η διάμεσος είναι 7

Άσκηση

Οι μισθοί 4 εργαζομένων είναι: 3, 4, 6, 8 (σε χιλιάδες €).

Να βρεθεί η διάμεσος.

Λύση

Το πλήθος είναι $n=4$ (άρτιος αριθμός).

Η διάμεσος είναι ο μέσος όρος των τιμών στις θέσεις $n/2=2$ και $(n/2)+1=3$

Οι τιμές αυτές είναι 4 και 6.

Διάμεσος $= (4+6)/2=5$

Η διάμεσος είναι 5.

Επικρατούσα Τιμή

- Η επικρατούσα τιμή (ή τρόπος) είναι η τιμή που εμφανίζεται με τη μεγαλύτερη συχνότητα.

Άσκηση

Δίνονται οι τιμές: 2, 4, 4, 6, 7, 7, 7, 8.

Να βρεθεί η επικρατούσα τιμή.

Λύση

Το 2 εμφανίζεται 1 φορά, το 4 εμφανίζεται 2 φορές, το 7 εμφανίζεται 3 φορές, το 8 εμφανίζεται 1 φορά.

Η μεγαλύτερη συχνότητα είναι 3, που αντιστοιχεί στην τιμή 7. Η επικρατούσα τιμή είναι 7.

Εύρος (R)

- Τι μετράει: Δείχνει τη διαφορά ανάμεσα στη μεγαλύτερη και τη μικρότερη τιμή του δείγματος.
- Χρησιμότητα: Επηρεάζεται πολύ από ακραίες τιμές.
- Τύπος:

$$R = X_{\max} - X_{\min}$$

όπου:

$X_{\max}$ = μέγιστη τιμή, $X_{\min}$ = ελάχιστη τιμή

Άσκηση

Οι βαθμολογίες 5 μαθητών είναι: 5, 7, 9, 12, 15. Να βρεθεί το εύρος.

Λύση

$$X_{\max} = 15, X_{\min} = 5$$

$$R = 15 - 5 = 10$$

Το εύρος είναι 10 μονάδες.

Διακύμανση (σ^2 ή s^2)

- Τι μετράει: Το πόσο διασκορπισμένες είναι οι τιμές γύρω από τον μέσο όρο.
- Χρησιμότητα: Δίνει πιο ολοκληρωμένη εικόνα από το εύρος γιατί λαμβάνει υπόψη όλες τις τιμές.
- Τύποι:

Για πληθυσμό:

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

Για δείγμα:

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

όπου:

x_i = κάθε τιμή δεδομένων

μ = μέσος όρος πληθυσμού

$\bar{x}$ = μέσος όρος δείγματος

N = μέγεθος πληθυσμού

n = μέγεθος δείγματος

Τυπική Απόκλιση (σ ή s)

- Τι μετράει: Είναι η «μέση απόσταση» των τιμών από τον μέσο όρο.
- Χρησιμότητα: Εκφράζεται στις ίδιες μονάδες με τα δεδομένα, γι' αυτό είναι πιο κατανοητή από τη διακύμανση.
- Τύπος:

$$\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2}$$