ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΘΕΩΡΙΑ ΑΠΟΦΑΣΕΩΝ

ΔΙΑΛΕΞΗ: ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΙΙ

Διδάσκουσα: Ε. Γάκη, Επίκ. Καθηγήτρια

Περιεχόμενα

- Αριθμητικές Μέθοδοι Σύνοψης Δεδομένων
  - για αταξινόμητα δεδομένα
  - για ταξινομημένα δεδομένα

Αριθμητικές Μέθοδοι Σύνοψης Δεδομένων

- Τα αριθμητικά μέτρα που αναφέρονται σε πληθυσμό ονομάζονται παράμετροι ενώ τα αντίστοιχα μέτρα που αναφέρονται σε δείγμα υπολογίζονται ως τιμές στατιστικών συναρτήσεων.

- Με τη βοήθεια των δειγματικών δεδομένων υπολογίζουμε στατιστικές συναρτήσεις που δίνουν μια θεωρητική εικόνα της δειγματικής κατανομής αποτελούν τη βάση της στατιστικής συμπερασματολογίας για τον πληθυσμό

- Τα αριθμητικά μέτρα χωρίζονται σε πέντε κατηγορίες:
  - Μέτρα Κεντρικής Τάσης
  - Μέτρα Σχετικής Θέσης
  - Μέτρα Διασποράς
  - Μέτρα Ασυμμετρίας
  - Μέτρα Κύρτωσης
  - Μέτρα Συγκέντρωσης

Μέτρα Κεντρικής Τάσης

- Τα μέτρα κεντρικής τάσης αποτελούν μια ένδειξη της τιμής γύρω από την οποία τείνουν να συσσωρεύονται τα δεδομένα.

- Θα εξετάσουμε τα μέτρα κεντρικής τάσης σε ταξινομημένα και αταξινόμητα δεδομένα

Αταξινόμητα Δεδομένα: Αριθμητικός Μέσος

x_i	Την τιμή της i μονάδας του πληθυσμού
Xi	Την τιμή της i παρατήρησης του δείγματος
N	Το πλήθος των μονάδων του πληθυσμού
n	Το πλήθος των παρατηρήσεων του δείγματος

Δειγματικός Α.Μ

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Α.Μ. Πληθυσμού

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Ο αριθμητικός μέσος

- Μπορεί να υπολογισθεί για οποιοδήποτε σύνολο δεδομένων.
- Κάθε σύνολο δεδομένων έχει έναν και μοναδικό μέσο αριθμητικό.
- Επιδέχεται μαθηματικούς χειρισμούς.
- Λαμβάνει υπόψη του όλα τα δεδομένα.

ΟΜΩΣ

- επηρεάζεται από τυχόν ακραίες τιμές των δεδομένων

Αταξινόμητα Δεδομένα
Σταθμικός Αριθμητικός Μέσος

- Για τον υπολογισμό του **σταθμικού αριθμητικού μέσου** (*weighted arithmetic mean*) οι παρατηρήσεις αθροίζονται αφού πρώτα πολλαπλασιασθούν με κάποιους συντελεστές στάθμισης, οι οποίοι εκφράζουν την σπουδαιότητα κάθε παρατήρησης στον προσδιορισμό της μέσης τιμής τους, και το άθροισμα τους διαιρείται με το άθροισμα των συντελεστών στάθμισης.

Δειγματικός Σταθμικός Α.Μ

$$\overline{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Πληθυσμού

Σταθμικός Α.Μ.

$$\overline{X} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

wi, I = 1, 2, … είναι οι συντελεστές στάθμισης.

Αταξινόμητα Δεδομένα: Διάμεσος

- Η **διάμεσος** (*median*) Μ ενός συνόλου μετρήσεων είναι η τιμή εκείνη που όταν οι παρατηρήσεις τοποθετηθούν σε σειρά τάξης μεγέθους τις χωρίζει σε δύο μέρη έτσι ώστε το πολύ 50% των μετρήσεων είναι μικρότερες από την τιμή αυτή και το πολύ 50% των μετρήσεων μεγαλύτερες από την τιμή αυτή.
- Η διάμεσος βρίσκεται στη **θέση** $\frac{n}{2}+\frac{1}{2}$, είναι δηλαδή η παρατήρηση $X_{\frac{n}{2}+\frac{1}{2}}$
- Η **τιμή** της προσδιορίζεται

$$M = X_{\frac{n}{2}+\frac{1}{2}} = X_{\frac{n}{2}} + \frac{1}{2}\left(X_{\frac{n}{2}+1} - X_{\frac{n}{2}} \right) = X_{\frac{n}{2}} + 0,5(X_{\frac{n}{2}+1} - X_{\frac{n}{2}})$$

Παράδειγμα

- Τα δεδομένα του προηγούμενου παραδείγματος σε αύξουσα σειρά φαίνονται στον πίνακα.

13	25	26	27	27	27	29	31	33	34
35	35	36	36	39	39	40	40	40	41
41	41	42	42	42	43	43	43	44	44
44	44	44	45	45	45	45	46	46	46
47	48	48	48	49	49	49	51	51	51
51	51	51	52	52	53	53	53	53	53
54	54	54	54	54	54	55	55	55	55
55	55	56	56	56	56	57	57	57	57
57	58	58	58	59	59	59	59	59	60
60	60	60	60	61	61	61	61	61	61
62	62	62	62	62	63	64	65	65	66
67	67	68	68	68	69	69	69	69	69
69	70	70	71	71	71	71	72	73	73
73	75	75	76	76	77	78	79	80	80
82	82	83	85	85	86	88	89	91	94

- Η διάμεσος είναι η παρατήρηση που βρίσκεται στη θέση $\left(\frac{n}{2}+\frac{1}{2}\right)=\frac{150}{2}+\frac{1}{2}=75,5_{η}$
- Είναι δηλαδή η παρατήρηση και βρίσκεται μεταξύ της 75ης και 76ης παρατήρησης.
- Η τιμή της είναι $M = X_{75,5} = X_{75} + 0,5\left(X_{76} - X_{75}\right) = 56 + 0,5(56-56) = 56$

Αταξινόμητα Δεδομένα:
Επικρατούσα Τιμή

- **Επικρατούσα τιμή** (*mode*) $\mathbf{T_o}$,είναι εκείνη η τιμή της μεταβλητής η οποία έχει τη μεγαλύτερη συχνότητα εμφάνισης.

- Όταν υπάρχουν δύο ή περισσότερες τιμές με την ίδια συχνότητα εμφάνισης τότε λέμε ότι τα δεδομένα έχουν δύο ή περισσότερες επικρατούσες τιμές και η κατανομή τους ονομάζεται **δικόρυφη (bimodal)** ή **πολυκόρυφη (multimodal)** αντίστοιχα.

ΠΑΡΑΔΕΙΓΜΑ

- Από τα δεδομένα του προηγούμενου παραδείγματος τα οποία έχουν τοποθετηθεί σε αύξουσα σειρά τάξης μεγέθους προκύπτει ότι έχουν τρεις επικρατούσες τιμές
  To =54, To=55 To=61

Παρατήρηση

Η επικρατούσα τιμή χρησιμοποιείται κυρίως όταν τα δεδομένα μας είναι ποιοτικά οπότε ο αριθμητικός μέσος και η διάμεσος δεν έχουν νόημα.

Ταξινομημένα Δεδομένα: Αριθμητικός Μέσος

Δειγματικός Αριθμητικός μέσος

$$X \cong \frac{\sum_{i=1}^{k} f_i m_i}{n}$$

- k το πλήθος των τάξεων στις οποίες έχουμε ταξινομήσει τις παρατηρήσεις
- fi η συχνότητα της i τάξης.
- mi ο κεντρικός όρος της i τάξης .
- n το πλήθος των παρατηρήσεων του δείγματος.

Αριθμητικός μέσος πληθυσμού

$$\mu = \frac{\sum_{i=1}^{k} f_i m_i}{N}$$

N το πλήθος των μονάδων του πληθυσμού.

Παραδοχή: ο κεντρικός όρος κάθε τάξης προσεγγίζει ικανοποιητικά τον αριθμητικό μέσο των μετρήσεων που ανήκουν στην τάξη αυτή.
Μια τέτοια υπόθεση ισχύει στη περίπτωση που οι μετρήσεις σε μια τάξη κατανέμονται συμμετρικά γύρω από τον κεντρικό της όρο.
Τότε το άθροισμα των μετρήσεων στην τάξη i προσεγγίζεται από το γινόμενο $f_i m_i$.

Μέτρα Σχετικής Θέσης

- Τα μέτρα σχετικής θέσης αναφέρονται στη σχετική θέση των δεδομένων μεταξύ τους. Τα σπουδαιότερα από τα μέτρα αυτά είναι τα ποσοστιαία σημεία (percentiles) (P1, P2, …, P99).

- Το p-ποσοστιαίο σημείο P (*p-percentile*) ενός συνόλου μετρήσεων είναι η τιμή εκείνη που όταν οι παρατηρήσεις τοποθετηθούν σε σειρά τάξης μεγέθους τις χωρίζει σε δύο μέρη έτσι ώστε το πολύ p% των μετρήσεων είναι μικρότερες από την τιμή αυτή και το πολύ (100-p)% των μετρήσεων είναι μεγαλύτερες από την τιμή αυτή.

- Τα πιο συχνά χρησιμοποιούμενα ποσοστιαία σημεία είναι τα τεταρτημόρια (*quartiles*) (Q1, Q2 , Q3) και τα δεκατημόρια (*deciles*) (D1,D2,…,D9) τα οποία χωρίζουν ένα σύνολο παρατηρήσεων σε τέταρτα και δέκατα αντίστοιχα.

Η διάμεσος (M) ταυτίζεται με το 2ο τεταρτημόριο (Q2), το 5ο δεκατημόριο (D5) και το 50ο εκατοστιαίο σημείο (P50).

Μέτρα Σχετικής Θέσης – Αταξινόμητα Δεδομένα

Τεταρτημόριο (Q_i)

Θέση $\left(\frac{i(n+1)}{4}\right)$ είναι δηλαδή η παρατήρηση $X_{\frac{i(n+1)}{4}}$

Τιμή $Q_i = X_{\frac{i(n+1)}{4}} = X_{A_Q} + \Delta_Q\left(X_{A_Q+1} - X_{A_Q}\right)$

A_Q=ακέραιο μέρος του $\left(\frac{i(n+1)}{4}\right)$ **και**

Δ_Q=δεκαδικό μέρος του $\left(\frac{i(n+1)}{4}\right)$

Δεκατημόριο (D_i)

Θέση $\left(\frac{i(n+1)}{10}\right)$ είναι δηλαδή η παρατήρηση $X_{\frac{i(n+1)}{10}}$

.

Τιμή $D_i = X_{\frac{i(n+1)}{10}} = X_{A_D} + \Delta_D\left(X_{A_{D+1}} - X_{A_D}\right)$

A_D= ακεραίο μέρος του $\left(\frac{i(n+1)}{10}\right)$ **και**

Δ_D=δεκαδικό μέρος του $\left(\frac{i(n+1)}{10}\right)$

Ποσοστιαίο Σημείο

Θέση (P_i)

Τιμή $P_i = X_{\frac{i(n+1)}{100}} = X_{A_P} + \Delta_P\left(X_{A_P+1} - X_{A_P}\right)$ είναι δηλαδή η παρατήρηση $X_{\frac{i(n+1)}{100}}$

A_p= ακέραιο μέρος του $\left(\frac{i(n+1)}{100}\right)$ **και**

Δ_p=δεκαδικό μέρος του $\left(\frac{i(n+1)}{100}\right)$

Παράδειγμα – Αταξινόμητα Δεδομένα

- Το $1_ο$ τεταρτημόριο είναι η παρατήρηση που βρίσκεται στη $\frac{(n+1)}{4} = \frac{150}{4} + \frac{1}{4} = 37,5 + 0,25 = 37,75_η$ θέση.

  Είναι δηλαδή η παρατήρηση $X_{37,75}$ και βρίσκεται μεταξύ της $37_ης$ και $38_ης$ παρατήρησης.

  Η τιμή του Q_1 είναι: $X_{37,75} = X_{37} + 0,75(X_{38} - X_{37}) = 45 + 0,75(46 - 45) = 45,75$

- Το 3ο δεκατημόριο είναι η παρατήρηση που βρίσκεται ανάμεσα στη $\frac{3(n+1)}{10} = \frac{450}{10} + \frac{3}{10} = 45 + 0,3 = 45,3_η$ θέση.

  Είναι δηλαδή η παρατήρηση $X_{45,3}$ και βρίσκεται μεταξύ της 45ης και της 46ης παρατήρησης.

  Η τιμή της D_3 είναι: $X_{45,3} = X_{45} + 0,3(X_{46} - X_{45}) = 49 + 0,3(49 - 49) = 49$

- Το 90ο ποσοστιαίο σημείο είναι η παρατήρηση που βρίσκεται στη $\frac{90(n+1)}{100} = \frac{13500}{100} + \frac{90}{100} = 135 + 0,9 = 135,9_η$ θέση.

  Είναι δηλαδή η παρατήρηση $X_{135,9}$ και βρίσκεται μεταξύ της $135_ης$ και της $136_ης$ παρατήρησης.

  Η τιμή του P90 είναι: $X_{135,9} = X_{135} + 0.9(X_{136} - X_{135}) = 76 + 0,9(77 - 76) = 76,9$

Μέτρα Σχετικής Θέσης – Ταξινομημένα Δεδομένα

Τεταρτημόριο (Q_i)

Θέση Είναι η τάξη που περιέχει την παρατήρηση $\frac{ni}{4}$

Τιμή $$Q_1 = L_{Q_1} + \frac{\delta}{f_{Q_1}}\left(\frac{n}{4} - F_{Q_1-1}\right) = 40 + \frac{10}{31}(37,5 - 16) = 46,94$$

Δεκατημόριο (D_i)

Θέση Είναι η τάξη που περιέχει την παρατήρηση $\frac{ni}{10}$

Τιμή $$D_i = L_{D_i} + \frac{\delta}{f_{D_i}}\left(\frac{ni}{10} - F_{D_i-1}\right)$$

Ποσοστιαίο Σημείο (P_i)

Θέση Είναι η τάξη που περιέχει την παρατήρηση $\frac{ni}{100}$

Τιμή $$P_i = L_{P_i} + \frac{\delta}{f_{P_i}}\left(\frac{ni}{100} - F_{P_i-1}\right)$$

Σημείωση: Για την εύρεση της τάξης του εκάστοτε ποσοστιαίου σημείου χρησιμοποιείται η δεξιόστροφη αθροιστική συχνότητα

Μέτρα Σχετικής Θέσης – Ταξινομημένα Δεδομένα

L_{Q_i} Κατώτερο όριο της τάξης i του τεταρτημορίου

f_{Q_i} Συχνότητα της τάξης i τεταρτημορίου

F_{Q_i-1} Δεξιόστροφη αθροιστική συχνότητα της τάξης που προηγείται εκείνης στην οποία εντοπίζεται το i τεταρτημόριο.

L_{D_i} Κατώτερο όριο της τάξης i του δεκατημορίου

f_{D_i} Συχνότητα της τάξης i δεκατημορίου

F_{D_i-1} Δεξιόστροφη αθροιστική συχνότητα της τάξης που προηγείται εκείνης στην οποία εντοπίζεται το i τεταρτημόριο

L_{Pi} Κατώτερο όριο της τάξης του i ποσοστιαίου σημείου

f_{Pi} Συχνότητα της τάξης του i ποσοστιαίου σημείου

F_{P_i-1} Δεξιόστροφη αθροιστική συχνότητα της τάξης που προηγείται εκείνης στην οποία εντοπίζεται το i ποσοστιαίο σημείο

δ Πλάτος της τάξης της κατανομής

n Πλήθος μετρήσεων

Παράδειγμα – Ταξινομημένα Δεδομένα

ΤΑΞΕΙΣ	m_i	f_i	F_{M-1}
10-20	15	1	1
20-30	25	6	7
30-40	35	9	16
40-50	45	31	47
50-60	55	42	89
60-70	65	32	121
70-80	75	17	138
80-90	85	10	148
90-100	95	2	150

Πρώτο Τεταρτημόριο (Q_1)

Θέση: Είναι η τάξη 40-50 γιατί περιέχει την $\frac{ni}{4} = \frac{150*1}{4} = 37,5$ παρατήρηση

Τιμή $Q_1 = L_{Q_1} + \frac{\delta}{f_{Q_1}}\left(\frac{n}{4} - F_{Q_1-1}\right) = 40 + \frac{10}{31}(37,5-16) = 46,94$

Τρίτο Δεκατημόριο (D_3)

Θέση: Είναι η τάξη 40-50 γιατί περιέχει την $\frac{ni}{10} = \frac{150*3}{10} = 45$ παρατήρηση

Τιμή $D_1 = L_{D_3} + \frac{\delta}{f_{D_3}}\left(\frac{3n}{10} - F_{D_3-1}\right) = 40 + \frac{10}{31}(45-16) = 49,35$

90ο Ποσοστιαίο Σημείο (P_{90})

Θέση: Είναι η τάξη 70-80 γιατί περιέχει την $\frac{ni}{100} = \frac{150*90}{4} = 135$ παρατήρηση

Τιμή $P_{90} = L_{P_{90}} + \frac{\delta}{f_{P_{90}}}\left(\frac{90n}{100} - F_{P_{90}-1}\right) = 70 + \frac{10}{17}(135-121) = 78,24$

Μέτρα Διασποράς

- Τα μέτρα κεντρικής τάσης δίνουν μια ένδειξη της τιμής γύρω από την οποία τείνουν να **συσσωρεύονται τα δεδομένα**.

- Είναι απαραίτητο να γνωρίζουμε και το **πώς κατανέμονται** οι παρατηρήσεις **γύρω** από ένα μέτρο κεντρικής τάσης.

- Την πληροφορία αυτή μας την παρέχουν τα **μέτρα διασποράς**

Αταξινόμητα Δεδομένα:
Εύρος

- Το **εύρος** (*range*) ενός συνόλου παρατηρήσεων ορίζεται ως η διαφορά μεταξύ της μεγίστης και της ελαχίστης τιμής του συνόλου των παρατηρήσεων

R = Xmax - Xmin

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Πλεονέκτημα: ευκολία του υπολογισμού του

Μειονέκτημα: εξαρτάται **μόνο** από τις δύο ακραίες τιμές του συνόλου των παρατηρήσεων χωρίς να λαμβάνει υπόψη του τις άλλες.

Παράδειγμα: Το εύρος του προηγούμενου παραδείγματος είναι:

R = Xmax - Xmin = 94 – 13 = 81

Αταξινόμητα Δεδομένα: Ενδοτεταρτημοριακό Εύρος

- **Το ενδοτεταρτημοριακό εύρος** ορίζεται ως

  Q3 – Q1

  και περιλαμβάνει μόνο το *50% των παρατηρήσεων* που βρίσκονται *γύρω* από τη *διάμεσο*.

- Δίνει δηλαδή μια ένδειξη του εύρους του διαστήματος που ορίζεται εκατέρωθεν της διαμέσου και περιλαμβάνει το 50% του συνολικού αριθμού των παρατηρήσεων.

Ουσιαστικά αποκλείει 25% των παρατηρήσεων σε κάθε άκρο της κατανομής

Παράδειγμα: $Q_1 = X_{37.75} = X_{37} + 0.75(X_{38} - X_{37}) = 45 + 0.75(46 - 45) = 45.75$

$$Q_3 = X_{113.25} = X_{113} + 0.25(X_{114} - X_{113}) = 68 + 0.25(68 - 68) = 68$$

- Μπορούμε να χρησιμοποιήσουμε και την **ενδοτεταρτημοριακή απόκλιση**

$$Q_3 - Q_1 = 68 - 45.75 = 22.25 \qquad Q = \frac{Q_3 - Q_1}{2}$$

Αταξινόμητα Δεδομένα: Ενδοποσοστιαία Εύρη

- Το **ενδοποσοστιαίο εύρος** ορίζεται ως η διαφορά δύο ποσοστιαίων σημείων

$$P_2 - P_1 = X_{P_1} - X_{P_2}$$

- *Εκφράζει* το ποσοστό των παρατηρήσεων του δείγματος που περιλαμβάνεται μεταξύ των δυο αυτών ποσοστιαίων σημείων.

$$P_{90} - P_{10} = X_{\frac{90(n+1)}{100}} - X_{\frac{10(n+1)}{100}} = X_{135.9} - X_{15.1}$$

$$= 76 + 0.9(77 - 76) - 39 + 0.1(39 - 39)$$

$$= (76 + 0.9) - 39 = 37.9$$

Αταξινόμητα Δεδομένα: Μέση Απόλυτη Απόκλιση

- **Η Μέση Απόλυτη Απόκλιση** λαμβάνει υπόψη το μέσο όρο των αποκλίσεων όλων των παρατηρήσεων από τον αριθμητικό μέσο

$$\text{M.A.A.} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n} \qquad \text{στην περίπτωση δείγματος ή}$$

$$\text{M.A.A.} = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} \qquad \text{στην περίπτωση πληθυσμού}$$

- Γιατί χρησιμοποιούμε απόλυτη τιμή???
  - *Το άθροισμα των αποκλίσεων των παρατηρήσεων από τον μέσο αριθμητικό τους είναι πάντοτε μηδέν (αφού οι θετικές αποκλίσεις αντισταθμίζονται από τις αρνητικές). Κατά συνέπεια η μέση απόκλιση θα είναι πάντοτε μηδέν*

- Η χρήση της μέσης απόλυτης απόκλισης είναι περιορισμένη επειδή είναι συνάρτηση απόλυτων τιμών.

Αταξινόμητα Δεδομένα: Διακύμανση

- Η **Διακύμανση** (*Variance*) βασίζεται στην έννοια της απόστασης μιας παρατήρησης από τον μέσο αριθμητικό των παρατηρήσεων.
  - *Ξεπερνά το πρόβλημα του μηδενικού αθροίσματος αποκλίσεων χρησιμοποιώντας τα τετράγωνα των αποκλίσεων τα οποία έχουν πάντοτε μη αρνητικές τιμές*

Διακύμανση Πληθυσμού

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

Διακύμανση Δείγματος

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{n}$$

Εδώ χρησιμοποιήσουμε ως διαιρέτη το n-1 αντί του n γιατί η τιμή του s2 στην οποία καταλήγουμε είναι καλύτερη εκτίμηση, από την προηγούμενη

$$\hat{s}^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{n-1}$$

Στη πράξη συνήθως χρησιμοποιούμε τους ακόλουθους τύπους

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \left[\frac{\sum_{i=1}^{N} x_i}{N}\right]^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

$$\hat{s}^2 = \frac{\sum_{i=1}^{N} X_i^2}{n-1} - \frac{\left[\sum_{i=1}^{N} X_i\right]^2}{(n-1)n}$$

Αταξινόμητα Δεδομένα: Τυπική Απόκλιση

- Η διακύμανση στις πρακτικές εφαρμογές είναι δύσκολο να ερμηνευτεί δεδομένου ότι εκφράζεται σε τετράγωνα των μονάδων των παρατηρήσεων.

- Η **τυπική απόκλιση** ορίζεται ως η θετική τετραγωνική ρίζα της διακύμανσης.

Πληθυσμό

$$\sigma = +\sqrt{\sigma^2}$$

Δείγμα

$$s = +\sqrt{s^2} \qquad \text{και} \qquad \hat{s} = +\sqrt{\hat{s}^2}$$

Ταξινομημένα Δεδομένα: Ενδοτεταρτομοριακό Εύρος

- Το **ενδοτεταρτομοριακό εύρος** υπολογίζεται από τον τύπο

  Q3 – Q1

- Η **τεταρτημοριακή απόκλιση** η οποία ορίζεται ως εξής

$$Q = \frac{Q_3 - Q_1}{2}$$

Παράδειγμα: Για το παράδειγμα των ταξινομημένων δεδομένων των βαθμολογιών των υποψηφίων γραμματέων στην σε μια εταιρεία (παράδειγμα προηγούμενου μαθήματος) έχουμε

$$Q_3 - Q_1 = 67{,}34 - 46{,}94 = 20{,}40$$

και

$$Q = \frac{Q_3 - Q_1}{2} = \frac{67{,}34 - 46{,}94}{2} = 10{,}20$$

Ταξινομημένα Δεδομένα: Ενδοποσοστιαία Εύρη

- P1, P2 είναι δύο ποσοστιαία σημεία ενός συνόλου παρατηρήσεων, το **ενδοποσοστιαίο εύρος** υπολογίζεται από τον τύπο,

$$P_2 - P_1 = X_{P_1} - X_{P_2}$$

- Θυμηθείτε ότι

$$P_i = L_{P_i} + \frac{\delta}{f_{P_i}}\left(\frac{ni}{100} - F_{P_i-1}\right)$$

L_{Q_i} Κατώτερο όριο της τάξης i του τεταρτημορίου

f_{Q_i} Συχνότητα της τάξης i τεταρτημορίου

F_{Q_i-1} Δεξιόστροφη αθροιστική συχνότητα της τάξης που προηγείται εκείνης στην οποία εντοπίζεται το i τεταρτημόριο.

L_{D_i} Κατώτερο όριο της τάξης i του δεκατημορίου

L_{P_i} Συχνότητα της τάξης i δεκατημορίου

f_{P_i} Δεξιόστροφη αθροιστική συχνότητα της τάξης που προηγείται εκείνης στην οποία εντοπίζεται το i τεταρτημόριο.

F_{D_i-1} Κατώτερο όριο της τάξης του i ποσοστιαίου σημείου.

δ Συχνότητα της τάξης του i ποσοστιαίου σημείου.

n Δεξιόστροφη αθροιστική συχνότητα της τάξης που προηγείται εκείνης στην οποία εντοπίζεται το i ποσοστιαίο σημείο.

F_{P_i-1} Πλάτος της τάξης της κατανομής.

Πλήθος μετρήσεων.

Στο προηγούμενο παράδειγμα το ενδοποσοστιαίο εύρος $P_{90} - P_{10}$, υπολογίζεται ως εξής:

$$P_{90} - P_{10} = 78.24 - 38.89 = 39.35$$

Ταξινομημένα Δεδομένα: Μέση Απόλυτη Απόκλιση

Πληθυσμός

$$\text{M.A.A.} \cong \frac{\sum_{i-1}^{k} f_i \left| m_i - \mu \right|}{N}$$

Δείγμα

$$\text{M.A.A.} \cong \frac{\sum_{i=1}^{k} f_i \left| m_i - \overline{X} \right|}{n}$$

k το πλήθος των τάξεων στις οποίες έχουμε ταξινομήσει τις παρατηρήσεις
f_i η συχνότητα της i τάξης
m_i ο κεντρικός όρος της i τάξης
$\overline{X}$ ο δειγματικός αριθμητικός μέσος
μ ο αριθμητικός μέσος του πληθυσμού
n το πλήθος των παρατηρήσεων του δείγματος
N το πλήθος των μονάδων του πληθυσμού

Ταξινομημένα Δεδομένα:
Μέση Απόλυτη Απόκλιση (παράδειγμα)

ΤΑΞΕΙΣ	m_i	f_i	$m_i - \bar{X}$	$\lvert m_i - \bar{X} \rvert$	$f_i \lvert m_i - \bar{X} \rvert$
10 - 20	15	1	-42,2	42,2	42,2
20 – 30	25	6	-32,2	32,2	193,2
30 – 40	35	9	-22,2	22,2	199,8
40 – 50	45	31	-12,2	12,2	378,2
50 – 60	55	42	-2,2	2,2	92,4
60 – 70	65	32	7,8	7,8	249,6
70 – 80	75	17	17,8	17,8	302,6
80 – 90	85	10	27,8	27,8	278,0
90 – 100	95	2	37,8	37,8	75,6
		150			**1811,6**

$$\text{Μ.Α.Α.} = \frac{1811,6}{150} = 12,08$$

Ταξινομημένα Δεδομένα: Διακύμανση

Πληθυσμός

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i \left(m_i - \mu\right)^2}{N}$$

Δείγμα

$$\hat{s}^2 = \frac{\sum_{i=1}^{k} f_i \left(m_i - \bar{X}\right)^2}{n-1}$$

k το πλήθος των τάξεων στις οποίες έχουμε ταξινομήσει τις παρατηρήσεις

fi η συχνότητα της i τάξης

mi ο κεντρικός όρος της i τάξης

$\bar{X}$ ο δειγματικός αριθμητικός μέσος

μ ο αριθμητικός μέσος του πληθυσμού

ΤΑΞΕΙΣ	m_i	f_i	$m_i - \bar{X}$	$\left(m_i - \bar{X}\right)^2$	$f_i \left(m_i - \bar{X}\right)^2$
10 - 20	15	1	-42,2	1780,84	1780,84
20 - 30	25	6	-32,2	1036,84	6221,04
30 - 40	35	9	-22,2	492,84	4435,56
40 - 50	45	31	-12,2	148,84	4614,04
50 - 60	55	42	-2,2	4,84	203,28
60 - 70	65	32	7,8	60,84	1946,88
70 - 80	75	17	17,8	316,84	5386,28
80 - 90	85	10	27,8	772,84	7728,40
90 - 100	95	2	37,8	1428,84	2857,68
		150			**35174,00**

$$\hat{s}^2 = 236,07$$

Ταξινομημένα Δεδομένα: Τυπική Απόκλιση

Πληθυσμός

$$\sigma = +\sqrt{\sigma^2}$$

Δείγμα

$$s = +\sqrt{s^2} \qquad \text{και} \qquad \hat{s} = +\sqrt{\hat{s}^2}$$

Η τυπική απόκλιση μας επιτρέπει να προσδιορίζουμε με αρκετή ακρίβεια το διάστημα, γύρω από τον μέσο μιας κατανομής, στο οποίο συγκεντρώνονται οι περισσότερες τιμές της.

Στην περίπτωση της κωδωνοειδούς και συμμετρικής κατανομής έχουμε

- 68% των παρατηρήσεων βρίσκεται στο διάστημα $\bar{X} \pm S$
- 95% των παρατηρήσεων βρίσκεται στο διάστημα $\bar{X} \pm 2S$
- 99% των παρατηρήσεων βρίσκεται στο διάστημα $\bar{X} \pm 3S$

Μέτρα Σχετικής Διασποράς

- Όλα τα μέτρα της διασποράς εκφράζονται στην **μονάδα μέτρησης του χαρακτηριστικού** που περιγράφουν.

- **Δεν** μπορούν λοιπόν να χρησιμοποιηθούν για **συγκρίσεις** ομάδων τιμών οι οποίες είτε εκφράζονται σε διαφορετικές μονάδες είτε εκφράζονται στην ίδια μονάδα αλλά έχουν σημαντικά διαφορετικούς αριθμητικούς μέσους.

- Σε περιπτώσεις του τύπου αυτού χρησιμοποιούμε μέτρα όχι απόλυτης αλλά σχετικής διασποράς των τιμών

Συντελεστής Μεταβλητότητας

- Ο **συντελεστής μεταβλητότητας** (*coefficient of variation*) συμβολίζεται με C και εκφράζεται ως ποσοστό της τυπικής απόκλισης προς τον αντίστοιχο αριθμητικό μέσο:

για το δείγμα $C = \frac{s}{\bar{x}} \times 100$

$$C = \frac{\hat{s}}{\bar{x}} \times 100$$ Εκτίμηση του s

για τον πληθυσμό $C = \frac{\sigma}{\mu} \times 100$

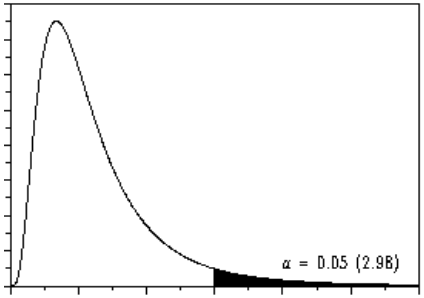
Ο C είναι ανεξάρτητος από την μονάδα μέτρησης του χαρακτηριστικού και εκφράζει την τυπική απόκλιση ως ποσοστό του αριθμητικού μέσου.

Μέτρα Ασυμμετρίας

- Τα μέτρα κεντρικής τάσης και διασποράς μιας κατανομής δίνουν μία εικόνα της μορφής της.

- Χρειάζεται όμως να προσδιορίσουμε πόσο και προς ποια κατεύθυνση αποκλίνει η κατανομή από την πλήρως συμμετρική κατανομή.

- Η ασυμμετρία μπορεί να είναι θετική ή αρνητική.
  - Θετικά ασυμμετρική είναι μια κατανομή όταν παρουσιάζει εξόγκωση προς τα αριστερά και η μεγάλη συγκέντρωση των παρατηρήσεων βρίσκεται στις μικρές τιμές της μεταβλητής.



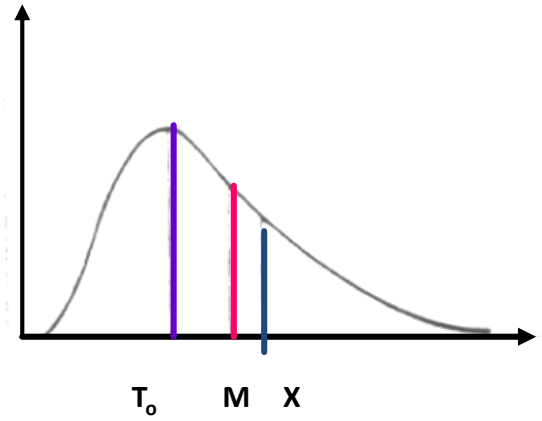
  - Αντίθετα, αρνητικά ασυμμετρική είναι μία κατανομή όταν παρουσιάζει εξόγκωση προς τα δεξιά και η μεγάλη συγκέντρωση των παρατηρήσεων βρίσκεται στις μεγάλες τιμές της μεταβλητής

Μέτρα Ασυμμετρίας που βασίζονται στη Μέση Τιμή

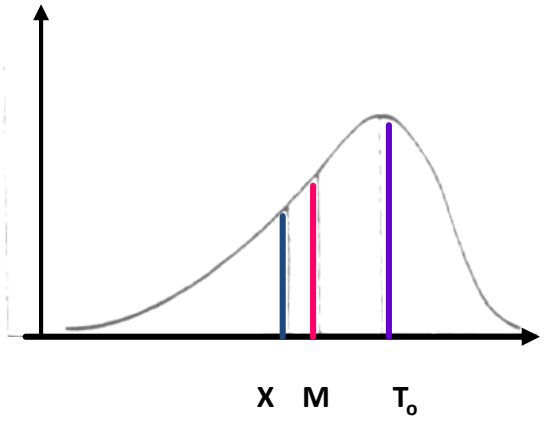
- Το μέτρο αυτό βασίζεται στην σχέση που υπάρχει μεταξύ του *αριθμητικού μέσου* (Χ), *της διαμέσου* (Μ), και της *επικρατούσας τιμής* (Το).

- Στην περίπτωση μιας μονοκόρυφης συμμετρικής κατανομής τα τρία αυτά μέτρα κεντρικής τάσης συμπίπτουν.

- Στην περίπτωση όμως ασυμμετρίας ο μέσος απομακρύνεται από την επικρατούσα τιμή και τείνει προς την κατεύθυνση της ασυμμετρίας ενώ η διάμεσος παραμένει πάντα ανάμεσα τους



Θετική Ασυμμετρία



Αρνητική Ασυμμετρία

Κατά συνέπεια η απόσταση μεταξύ μέσου και επικρατούσας τιμής και το πρόσημο της, μπορεί να χρησιμοποιηθεί ως μέτρο ασυμμετρίας.

Μέτρα Ασυμμετρίας που βασίζονται στη Μέση Τιμή

Αταξινόμητα Δεδομένα

$$S_P = \frac{3(\bar{X} - M)}{\hat{s}}$$

Ταξινομημένα Δεδομένα

$$S_p = \frac{3\left(\bar{X} - M\right)}{\hat{s}}$$

Συντελεστής Ασυμμετρίας του K. Pearson
Το SP παίρνει τιμές στο διάστημα [-3, +3] σπάνια όμως οι τιμές του είναι έξω από το διάστημα [-1, +1].

$$s_p = \begin{cases} 0 \text{ στην περίπτωση πλήρως συμμετρικής κατανομής} \\ > 0 \text{ στην περίπτωση θετικά συμμετρικής κατανομής} \\ < 0 \text{ στην περίπτωση αρνητικά συμμετρικής κατανομής} \end{cases}$$

Μέτρα Ασυμμετρίας που βασίζονται στη διάμεσο και τα τεταρτημόρια

- Το μέτρο αυτό βασίζεται στη σχέση που υπάρχει μεταξύ του πρώτου τεταρτημορίου (Q1), του τρίτου τεταρτημορίου (Q3) και της Διαμέσου (M).

- Σε μία πλήρως συμμετρική κατανομή η απόσταση του 3ου τεταρτημορίου από τη διάμεσο είναι όση η απόσταση της διαμέσου από το 1ο τεταρτημόριο

- Το μέτρο αυτό παίρνει τιμές στο διάστημα [-1, +1]

Αταξινόμητα Δεδομένα

$$S_B = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1}$$

Ταξινομημένα Δεδομένα

$$S_B = \frac{\left(Q_3 - M\right) - \left(M - Q_1\right)}{Q_3 - Q_1}$$

$$s_B = \begin{cases} 0 \text{ στην περίπτωση πλήρως συμμετρικής κατανομής} \\ > 0 \text{ στην περίπτωση θετικά συμμετρικής κατανομής} \\ < 0 \text{ στην περίπτωση αρνητικά συμμετρικής κατανομής} \end{cases}$$

Ροπή Κατανομής

- Ως **ροπή** (*moment*) k τάξης (μ_k) μιας σειράς παρατηρήσεων γύρω από το μέσο τους ορίζεται η συνάρτηση:

Αταξινόμητα δεδομένα

$$\mu_k = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^k}{n}$$

Ταξινομημένα δεδομένα

$$\mu_k = \frac{\sum_{i=1}^{n} f_i (m_i - \bar{x})^k}{n}$$

- Για k=2

$$\mu_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} = s^2$$

Μέτρα Ασυμμετρίας που βασίζονται στις ροπές μιας κατανομής

- Ως μέτρο ασυμμετρίας της κατανομής η παρατηρήσεων ορίζεται ο συντελεστής

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\left[\sum_{\iota=1}^{\nu}(X_\iota - \bar{X})\right]^2}{\left(s^2\right)^3}$$

- Το μέτρο αυτό είναι γνωστό ως συντελεστής ασυμμετρίας β1 του Pearson. Αξίζει να παρατηρήσουμε ότι το είδος της ασυμμετρίας με τη μέθοδο αυτή προσδιορίζεται από το πρόσημο της ροπής μ_3.

$$\mu_3 = \begin{cases} 0 \text{ στην περίπτωση πλήρως συμμετρικής κατανομής} \\ > 0 \text{ στην περίπτωση θετικά συμμετρικής κατανομής} \\ < 0 \text{ στην περίπτωση αρνητικά συμμετρικής κατανομής} \end{cases}$$

- Πολλές φορές χρησιμοποιούμε αντί του β1 ως συντελεστή ασυμμετρίας το β1* το οποίο ορίζεται από τον τύπο

$$\beta 1^* = \frac{\mu_3}{s^3}$$

Ο συντελεστής β1* κυμαίνεται μεταξύ περίπου –1 και +1 και δίνει πληροφορίες τόσο για το είδος όσο και το μέτρο της ασυμμετρίας.

β1>0 θετική ή δεξιά ασυμμετρία

β1<0 αρνητική ή αριστερή ασυμμετρία

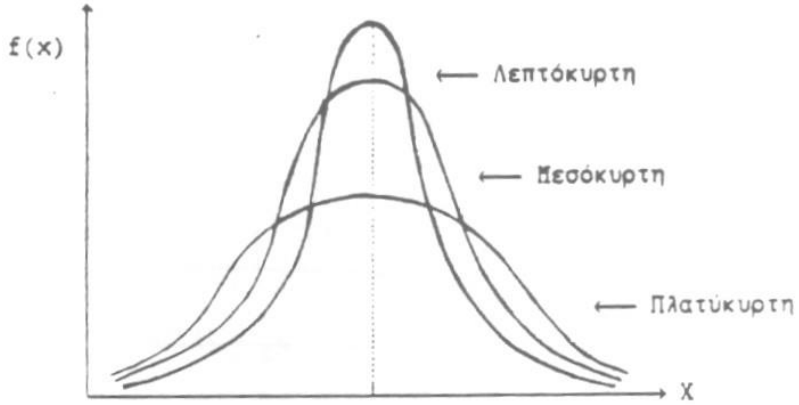
Μέτρα Κύρτωσης

- Δύο μονοκόρυφες κατανομές μπορεί να έχουν τον ίδιο αριθμητικό μέσο, την ίδια τυπική απόκλιση, να είναι συμμετρικές αλλά παρόλα αυτά *να διαφέρουν ως προς την **κύρτωση** στην περιοχή του αριθμητικού μέσου*.

  - ✓ **<u>Το χαρακτηριστικό αυτό έχει τοπικό χαρακτήρα</u>**. Δεν αναφέρεται στην κύρτωση ολόκληρης της κατανομής, η οποία εκφράζεται από τα μέτρα διασποράς, αλλά στην συγκέντρωση των τιμών και στην αιχμηρότητα της κατανομής στην περιοχή του αριθμητικού μέσου.

- Με βάση το χαρακτηριστικό αυτό μια κατανομή μπορεί να θεωρηθεί:
  - λεπτόκυρτη
  - πλατύκυρτη
  - μεσόκυρτη

  ανάλογα με το αν παρουσιάζει μεγάλη, μικρή ή φυσιολογική συγκέντρωση τιμών (δηλαδή αιχμηρότητα) στην περιοχή του αριθμητικού μέσου.



Μέτρο κύρτωσης που βασίζονται στις ροπές

- Το μέτρο αυτό είναι συνάρτηση της 4ης ροπής γύρω από το μέσο όρο

Αταξινόμητα δεδομένα

$$\mu_k = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k}{n}$$

Ταξινομημένα δεδομένα

$$\mu_\kappa = \frac{\sum_{i=1}^{n} f_i\left(m_i - \overline{X}\right)^k}{n}$$

- Μέτρο κύρτωσης της κατανομής n παρατηρήσεων κατά Pearson ορίζεται

$$\beta_2 = \frac{\mu_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{s^4}$$

$$\beta_2 = \frac{\mu_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n} f_i\left(m_i - \overline{X}\right)^k}{s^4}$$

$$\beta_2 = \begin{cases} = 3 \text{ στην περίπτωση μεσόκυρτης κατανομής} \\ > 3 \text{ στην περίπτωση λεπτόκυρτης κατανομής} \\ < 3 \text{ στην περίπτωση πλατύκυρτης κατανομης} \end{cases}$$

Μέτρο κύρτωσης που βασίζεται σε ενδοποσοστιαία εύρη

Το μέτρο αυτό είναι συνάρτηση της τερταμοριακής απόκλισης και του ενδοποσοστιαίου εύρους P90 – P10 . Καλείται **συντελεστής κύρτωσης k** και δίνεται από τον τύπο:

$$k = \frac{\frac{Q_3 - Q_1}{2}}{P_{90} - P_{10}}$$

Ταξινομημένα/Αταξινόμητα δεδομένα

(α) όσο πιο λεπτόκυρτη είναι η κατανομή τόσο μικρότερη γίνεται η διαφορά (P90 – P10) – (Q3 – Q1) .

Στην οριακή δε περίπτωση που το εύρος της κατανομής τείνει στο μηδέν και η διαφορά αυτή τείνει στο μηδέν (δηλαδή P90 – P10 = Q3 – Q1) το k παίρνει την μεγαλύτερη δυνατή τιμή του k = 0.5.

(β) όσο πιο πλατύκυρτη είναι η κατανομή τόσο μεγαλύτερη γίνεται η διαφορά (P90 – P10) – (Q3 – Q1) .

Στην οριακή δε περίπτωση μιας επίπεδης κατανομής που το εύρος της τείνει στο άπειρο και η διαφορά αυτή τείνει στο άπειρο (δηλαδή P90 – P10 >> Q3 – Q1) και επομένως το k = 0.

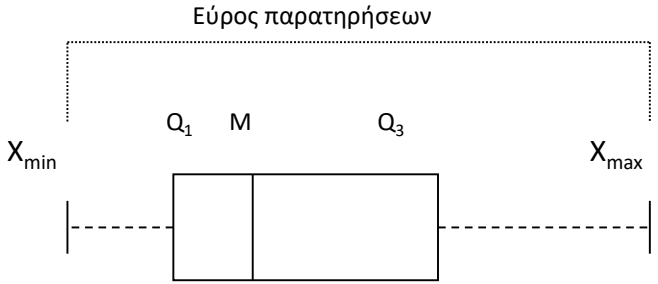
(γ) Για μια μεσόκυρτη κατανομή k = 0.25.

Διαγράμματα Πλαισίου Απολήξεων

- Απεικονίζει τις τιμές της διαμέσου και των τεταρτημορίων (μέτρα που δεν επηρεάζονται από ακραίες τιμές των δεδομένων).

- Συνδυάζει αυτά τα μέτρα με τις πληροφορίες που περιέχονται στις ακραίες τιμές των δεδομένων δίνοντας έτσι μια πληρέστερη εικόνα της κατανομής τους.

- Το Διάγραμμα Πλαισίου – Απολήξεων χρησιμοποιεί πέντε χαρακτηριστικές τιμές των δεδομένων:
  - X_{min}: Ελάχιστη τιμή των δεδομένων
  - Q_1: Πρώτο τεταρτημόριο
  - M: Διάμεσος
  - Q_3: Τρίτο τεταρτημόριο
  - X_{max} : Μέγιστη τιμή των δεδομένων



Μέτρα Συγκέντρωσης

Η έννοια της συγκέντρωσης εστιάζεται στον προσδιορισμό του **βαθμού απόκλισης** της όλης κατανομής από τη **κατάσταση ισοκατανομής.**

Αταξινόμητα Δεδομένα: Συντελεστής Gini

- **Ο συντελεστής Gini** συμβολίζεται με g και υπολογίζεται

$$g = \frac{d}{2\overline{X}}$$

$\overline{X}$ ο αριθμητικός μέσος της κατανομής
d η μέση διαφορά Gini

- Ο συντελεστής Gini παίρνει τιμές στο διάστημα [0, 1]
  - Το 0 αντιστοιχεί στην περίπτωση της τέλειας ισοκατανομής
  - το 1 στην περίπτωση της μέγιστης δυνατής ανισοκατανομής

- Η μέση διαφορά Gini d, που χρησιμοποιείται στον τύπο για τον υπολογισμό του g είναι ο αριθμητικός μέσος των απολύτων τιμών των διαφορών (d_{ij}) τις οποίες μπορούμε να σχηματίσουμε συνδυάζοντας ανά δύο με όλους τους δυνατούς τρόπους τις τιμές μιας μεταβλητής.
  - Αν οι δυνατές τιμές μιας μεταβλητής είναι n οι δυνατές διαφορές είναι n^2.

$$d = \frac{\sum_i \sum_j |d_{ij}|}{n^2}$$

Ταξινομημένα Δεδομένα:
Συντελεστής Gini

- Ο συντελεστής Gini όπως ορίστηκε για τα αταξινόμητα δεδομένα ισχύει και στην περίπτωση που τα δεδομένα είναι ταξινομημένα σε κατανομή συχνοτήτων.

$$g = \frac{d}{2\overline{X}}$$

- Η η μέση διαφορά Gini d, υπολογίζεται από τον τύπο:

$$d = \frac{2\delta}{n^2}\sum_{i=1}^{k}\left(n - \varphi_i\right)\varphi_i$$

δ	το πλάτος των τάξεων (σταθερό για όλες τις τάξεις)
ϕ_i	η δεξιόστροφη αθροιστική συχνότητα της i τάξης
n	το πλήθος των παρατηρήσεων
k	το πλήθος των τάξεων

Που απευθύνομαι?????

Ελένη Γάκη,
Τηλ. 2271035161
Email: e.gaki@aegean.gr
Υλικό μαθήματος
https://eclass.aegean.gr/