Στατιστική για Χημικούς Μηχανικούς - ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Δημήτρης Κουγιουμτζής

5 Απριλίου 2013

Περιεχόμενα

- Περιγραφική στατιστική (Σημειώσεις: Κεφ. 1)
- Εκτίμηση παραμέτρων (Σημειώσεις: Κεφ. 2)
- Στατιστικοί έλεγχοι (Σημειώσεις: Κεφ. 3)
- Συσχέτιση, Παλινδρόμηση (Σημειώσεις: Κεφ. 4)

Ιστοσελίδα μαθήματος:
http://users.auth.gr/dkugiu/Teach/ChemicalEngineer/
http://blackboard.lib.auth.gr/
κωδικός μαθήματος 10U010

Εισαγωγικά

Βεβαιότητα

βέβαια φαινόμενα $\longrightarrow$ χρήση καθοριστικών μεθόδων / μοντέλων

- Πόση ώρα κάνει ένας ποδηλάτης από το ένα στο άλλο άκρο της Νέας Παραλίας Θεσσαλονίκης (3km), όταν πηγαίνει με σταθερή ταχύτητα 20km/h;
- Ποια θα είναι η θέση της γης ως προς τον ήλιο το θερινό ηλιοστάσιο 2013;

Αβεβαιότητα / τυχαιότητα

αβέβαια ή τυχαία φαινόμενα $\longrightarrow$ χρήση στατιστικών μεθόδων / μοντέλων

- Πόση ώρα κάνει κάποιος τη διαδρομή αεροδρόμιο – Πολυτεχνείο με ποδήλατο / αυτοκίνητο;
- Είναι ακριβές το δρομολόγιο του ΟΣΕ για Αθήνα – Θεσσαλονίκη; (π.χ. InterCity 12.18 – 17.41)

Παραδείγματα

- Ποιός είναι ο αριθμός οκτανίων σ΄ ένα μίγμα βενζίνης;
- Ποιός είναι ο χρόνος αντίδρασης για δύο διαφορετικές θερμοκρασίες σε μια χημική αντίδραση; Για περισσότερες; Μπορούμε να καθορίσουμε μια σχέση εξάρτησης;
- Πόσοι χώροι πάρκινγκ ανά διαμέρισμα έχει κατά μέσο όρο μια πολυκατοικία στην Τούμπα; Στην Καλαμαριά; Υπάρχει διαφορά στις δύο συνοικίες;
- Πως μεταβάλλεται η εκπομπή ρύπων (π.χ. νιτρικό οξύ) με την παλαιότητα ενός τύπου μηχανής; Μπορούμε να προβλέψουμε για κάποια μηχανή τέτοιου τύπου που γνωρίζουμε το χρόνο λειτουργίας της πόσους ρύπους εκπέμπει;

Πως θα απαντήσουμε; $\Longrightarrow$ **Στατιστική**

Διαδικασία Στατιστικής

1. Δειγματοληψία (συλλογή δεδομένων)
2. Περιγραφική Στατιστική (περιγραφή / παρουσίαση δεδομένων και συνοπτικών μέτρων)
3. **Στατιστική Συμπερασματολογία** ή **Στατιστική** (ανάλυση στατιστικών δεδομένων και ερμηνεία αποτελεσμάτων)

Παράδειγμα: Χώροι πάρκιγκ

1. Συλλέγουμε δεδομένα από 20 πολυκατοικίες στην Τούμπα και 19 στην Καλαμαριά. [X]
2. 'Βλέπουμε' τις παρατηρήσεις (πίνακες, γραφήματα) και υπολογίζουμε κάποια μέτρα (π.χ. μέσο όρο). [V]
3. Εφαρμόζουμε στατιστικές μεθόδους για να εκτιμήσουμε το μέσο χώρο πάρκιγκ στις δύο συνοικίες και για να ελέγξουμε αν υπάρχει σημαντική διαφορά. [V]

Χρήσιμοι όροι Στατιστικής

- **τυχαία μεταβλητή (τ.μ.)**: οποιοδήποτε χαρακτηριστικό του οποίου η τιμή αλλάζει στα διάφορα στοιχεία του πληθυσμού, π.χ. πορώδες ηλίου σε γαιάνθρακα.
- **δεδομένα**: ένα σύνολο τιμών μιας τ.μ. που έχουμε στη διάθεση μας, π.χ. μετρήσεις του πορώδους ηλίου (σε ποσοστά ύλης) σε γαιάνθρακες.
- **πληθυσμός**: μια ομάδα ή μια κατηγορία στην οποία αναφέρεται η τ.μ., π.χ. γαιάνθρακας από κάποια περιοχή.
- **δείγμα**: ένα υποσύνολο του πληθυσμού που μελετάμε, π.χ. 25 δοκίμια γαιάνθρακα.
- **παράμετρος**: ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στον πληθυσμό, π.χ. η μέση τιμή του πορώδους ηλίου του γαιάνθρακα.
- **στατιστικό**: ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στο δείγμα, π.χ. ο μέσος όρος του πορώδους ηλίου από τα 25 δοκίμια του γαιάνθρακα που μετρήσαμε.

Πιθανότητες

Πιθανοθεωρία

- Τυχαία μεταβλητή τ.μ. X
- Συνάρτηση πυκνότητας πιθανότητας ή συνάρτηση μάζας πιθανότητας $f_X(x)$
- Αθροιστική συνάρτηση κατανομής $F_X(x)$
- Κύριες παράμετροι κατανομής:
  - Μέση (προσδοκώμενη) τιμή: $\mathrm{E}(X) = \mu$
  - Διασπορά (διακύμανση): $\mathrm{Var}(X) = \sigma^2$

Σκοπός Πιθανοθεωρίας

γνωρίζουμε κατανομή (παραμέτρους) $\longrightarrow$ μελετάμε την τ.μ.

Σκοπός Στατιστικής

γνωρίζουμε τιμές της τ.μ. $\longrightarrow$ συμπεράσματα για την τ.μ.

Με βάση τα δεδομένα θέλουμε να βγάλουμε συμπεράσματα:

από δείγμα $\longrightarrow$ για πληθυσμό

από στατιστικό $\longrightarrow$ για παράμετρο

Παράμετρος: σταθερή κι άγνωστη

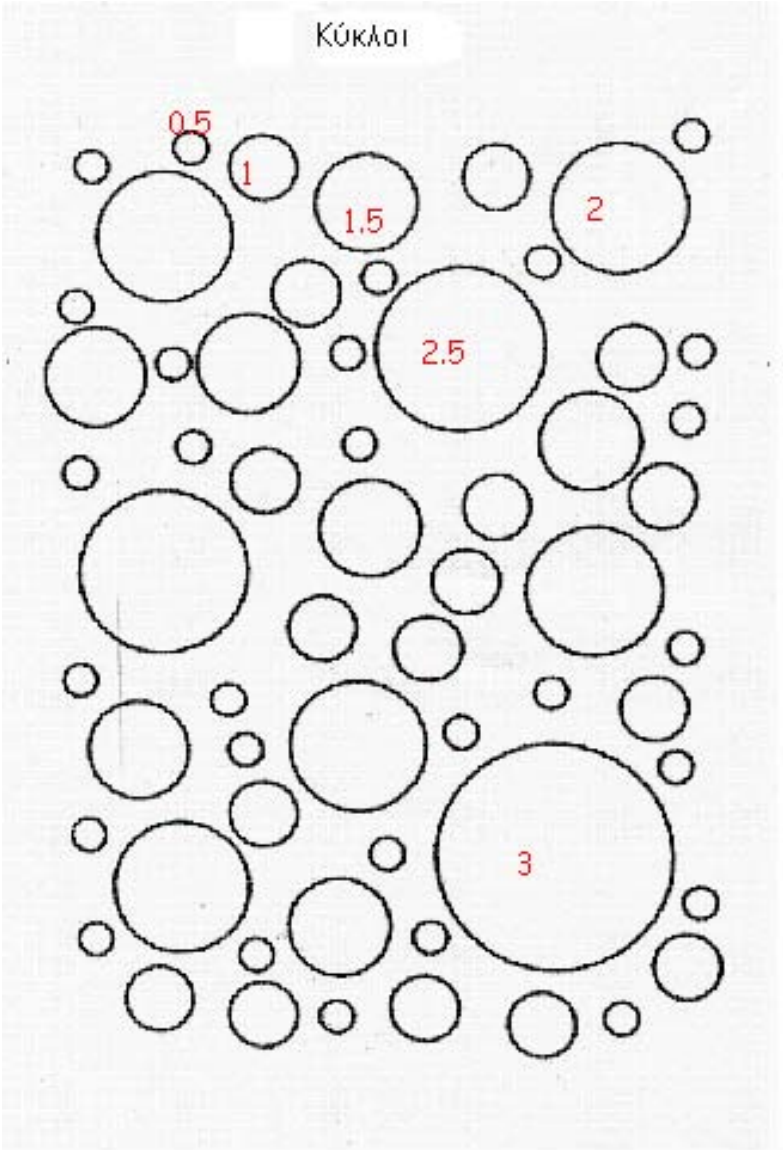
Στατιστικό: μεταβλητό και γνωστό

Δείγμα: τυχαίο και αντιπροσωπευτικό του πληθυσμού

Πείραμα: 'τυχαίο και αντιπροσωπευτικό δείγμα';

1. Πάρε ένα χαρτάκι με την εικόνα στο διπλανό σχήμα.
2. Διάλεξε τυχαίο και αντιπροσωπευτικό δείγμα πέντε κύκλων από τους 60 του πληθυσμού.
3. Υπολόγισε το μέσο όρο των ακτίνων των 5 κύκλων και γράψε το στο πίσω μέρος, π.χ.

$$\frac{1 + 0.5 + 1.5 + 1 + 2.5}{5} = 1.3$$



ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Περιγραφή Στατιστικών δεδομένων

1. Διακεκριμένες τιμές (k κατηγορίες ή λίγες αριθμητικές τιμές x_i, $i = 1, ..., k$, δείγμα μεγέθους n)

Συχνότητες

- συχνότητα εμφάνισης της τιμής x_i, f_i
- σχετική συχνότητα / ποσοστό, $p_i = \frac{f_i}{n}$
- αθροιστική συχνότητα $F_i = \sum_{j=1}^{i} f_j$ όπου $x_j \leq x_i$
- αθροιστική σχετική συχνότητα $P_i = \sum_{j=1}^{i} p_j$ όπου $x_j \leq x_i$

Παρουσίαση συχνοτήτων για τα x_i:

- **πίνακα συχνοτήτων**: μια γραμμή για κάθε τιμή, κάθε στήλη είναι ένας τύπος συχνότητας
- **ραβδόγραμμα**: μια ράβδος για τη συχνότητα κάθε τιμής
- κυκλικό διάγραμμα ('πίτα'), ...

Παράδειγμα: Παρτίδες χημικών προϊόντων

Δίνεται ο αριθμός των κυβωτίων σε 120 παρτίδες παραγγελίας σε μια αποθήκη χημικών προϊόντων.

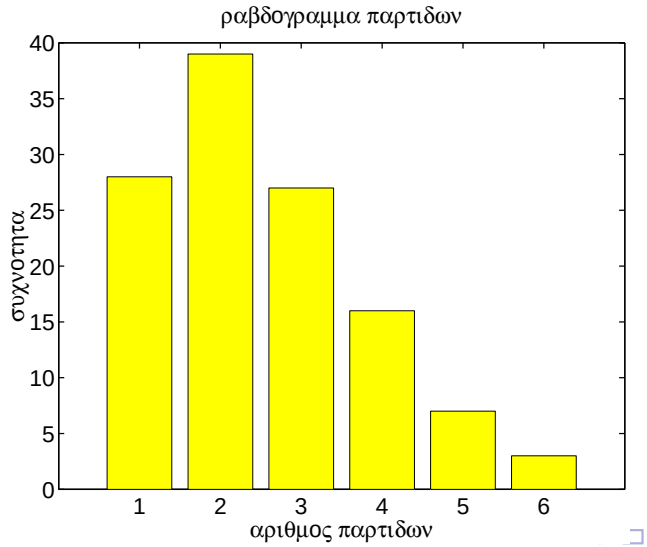
1	4	2	2	2	3	4	3	1	1	3	3
1	2	1	2	1	1	2	3	3	5	1	2
2	3	2	1	1	4	3	4	1	1	6	2
1	3	2	1	2	2	3	2	4	3	3	5
1	3	5	3	1	2	2	3	1	2	6	4
1	2	5	4	3	1	2	4	2	1	3	4
2	2	2	3	2	1	3	3	4	2	1	5
2	2	3	3	2	4	6	3	2	3	1	3
2	1	5	1	1	4	4	2	5	4	2	2
4	2	1	2	2	2	3	2	3	2	1	4

Παράδειγμα (συνέχεια)

Πίνακας συχνοτήτων

x_i	f_i	p_i	F_i	P_i
1	28	0.23	28	0.23
2	39	0.33	67	0.56
3	27	0.23	94	0.78
4	16	0.13	110	0.92
5	7	0.06	117	0.97
6	3	0.03	120	1.00
Άθροισμα	120	1.00		

Ραβδόγραμμα



2. Αριθμητικές τιμές (πολλές διακεκριμένες τιμές, τιμές σε διάστημα)

Ομαδοποίηση

Χωρίζουμε τα δεδομένα σε k ομάδες $\longrightarrow$ πίνακες / γραφήματα όπως πριν για τις k τιμές (ομάδες)

Χωρισμός σε ομάδες: (ίδιο εύρος τιμών r σε κάθε ομάδα)

Εύρος δεδομένων: $R = x_{\max} - x_{\min}$

$$R/k \quad \simeq \quad r$$

Το πρώτο διάστημα πρέπει να περιέχει το $x_{\min}$
Το τελευταίο διάστημα πρέπει να περιέχει το $x_{\max}$

Ραβδόγραμμα (ενώνοντας τις πλευρές) $\longrightarrow$ **ιστόγραμμα**

Άλλα γραφήματα: φυλλογράφημα, σημειογράφημα, ...

Θέμα 1

Προσδιορισμός αριθμού ομάδων ή εύρος διαστήματος ιστογράμματος (number of bins or bin width): Μέθοδοι και περιορισμοί.

Θέμα 2

Φυλλογράφημα (stem and leaf plot): Παρουσίαση, πλεονεκτήματα και παράδειγμα.

Παράδειγμα: πορώδες ηλίου γαιάνθρακα

Μετρήθηκε το πορώδες ηλίου (σε ποσοστό) σε δύο δείγματα δοκιμίων γαιανθράκων από δύο περιοχές A και B.

A/A	κοίτασμα A	κοίτασμα B
1	5.3	5.0
2	4.5	4.2
3	5.7	5.4
4	5.8	5.5
5	4.8	4.6
6	6.4	6.1
7	6.4	6.1
8	5.6	5.3
9	5.8	5.5
10	5.7	5.4
11	5.5	5.2
12	6.1	5.8
13	5.2	4.9
14	7.0	6.7
15	5.5	5.2
16	5.7	5.4
17	6.3	6.0
18	5.6	5.3
19	5.5	5.2
20	5.0	4.8
21	5.8	
22	4.7	
23	6.1	
24	6.7	
25	5.1	
Σύνολο	141.8	107.6

Παράδειγμα (συνέχεια)

X πορώδες ηλίου γαιάνθρακα από ένα κοίτασμα Α

Χωρισμός σε ομάδες

$x_{\min} = 4.5 \qquad x_{\max} = 7.0$

$R = x_{\max} - x_{\min} = 7.0 - 4.5 = 2.5$

Διαλέγουμε να χωρίσουμε τα δεδομένα σε 10 ομάδες ($k = 10$)

$$r = \frac{R}{k} = \frac{2.5}{10} = 0.25$$

Διαλέγουμε η πρώτη ομάδα (διάστημα) να αρχίζει από την τιμή 4.5

ομάδα 1: $4.50 - 4.75$

ομάδα 2: $4.75 - 5.00$

...

ομάδα 10: $6.75 - 7.00$

Η τελευταία ομάδα περιλαμβάνει το $x_{\max} = 7.0$

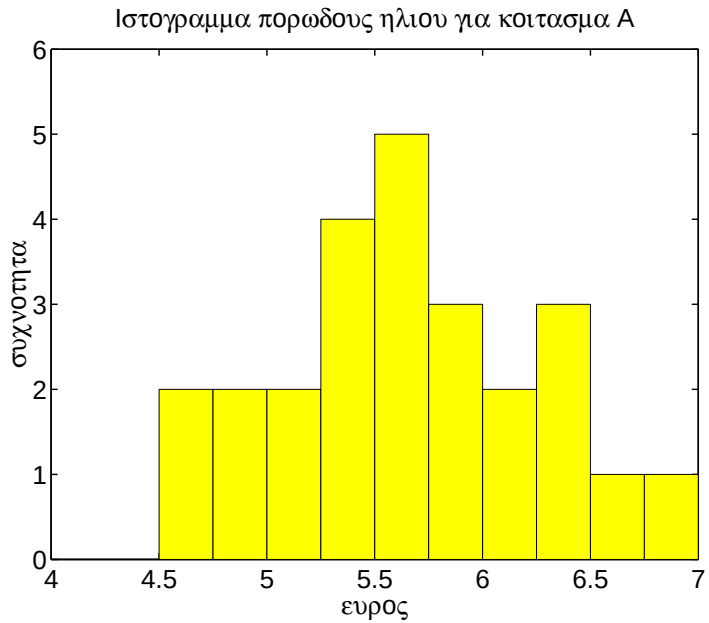
Παράδειγμα (συνέχεια)

Πίνακας συχνοτήτων

Διάστημα τιμών	f_i	p_i	F_i	P_i
4.50 – 4.75	2	0.08	2	0.08
4.75 – 5.00	2	0.08	4	0.16
5.00 – 5.25	2	0.08	6	0.24
5.25 – 5.50	4	0.16	10	0.40
5.50 – 5.75	5	0.20	15	0.60
5.75 – 6.00	3	0.12	18	0.72
6.00 – 6.25	2	0.08	20	0.80
6.25 – 6.50	3	0.12	23	0.92
6.50 – 6.75	1	0.04	24	0.96
6.75 – 7.00	1	0.04	25	1.00
Άθροισμα	25	1.00		

Παράδειγμα (συνέχεια)

Ιστόγραμμα



Μπορούμε να δεχτούμε ότι η X ακολουθεί κάποια γνωστή κατανομή;
Είναι σημαντικό για την στατιστική ανάλυση να είναι η κατανομή κανονική

Περιγραφικά Μέτρα Στατιστικών Δεδομένων

- **μέτρα θέσης**: προσδιορίζουν χαρακτηριστικές θέσεις μέσα στο εύρος των δεδομένων

- **μέτρα μεταβλητότητας**: δίνουν περιληπτικά τη διασκόρπιση και μεταβλητότητα των δεδομένων

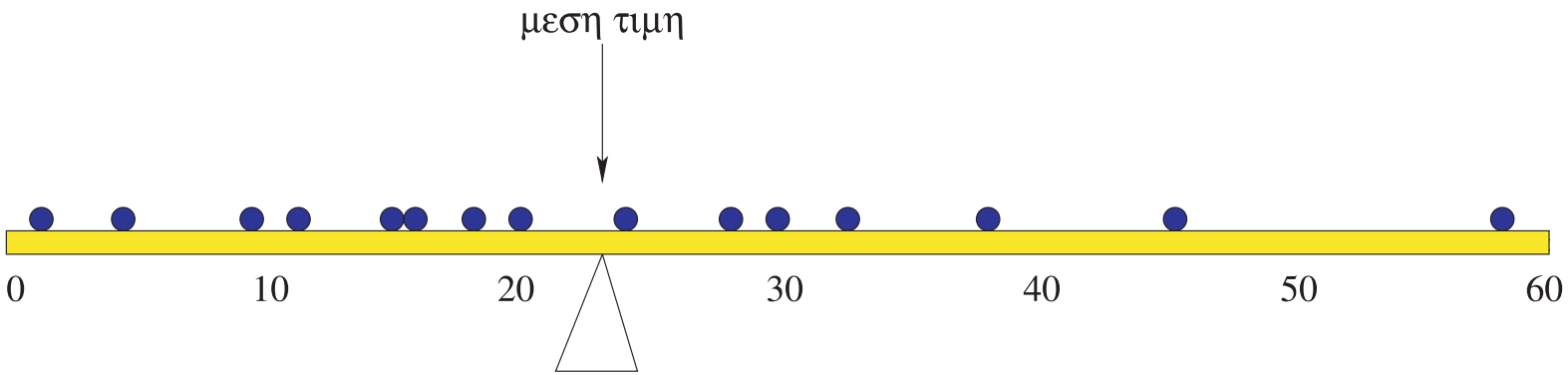
$x_1, x_2, \ldots, x_n$: παρατηρήσεις του δείγματος

Μέτρα θέσης

δειγματική μέση τιμή ή αριθμητικός μέσος ή μέσος όρος

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

Η μέση τιμή είναι το 'κέντρο ισορροπίας' των δεδομένων



Μέτρα θέσης

δειγματική διάμεσος

Είναι η κεντρική τιμή όταν διατάξουμε τα δεδομένα σε αύξουσα σειρά

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & n = 2k+1 \\ \frac{x_{n/2}+x_{n/2+1}}{2} & n = 2k \end{cases}$$

π.χ. δείγμα $x_1 \leq x_2 \leq x_3 \longrightarrow \tilde{x} = x_2$

π.χ. δείγμα $x_1 \leq x_2 \leq x_3 \leq x_4 \longrightarrow \tilde{x} = \frac{x_2+x_3}{2}$

δειγματική επικρατούσα τιμή

Είναι η τιμή που εμφανίζεται με τη μεγαλύτερη συχνότητα

Μέτρα θέσης

Παρατηρήσεις:

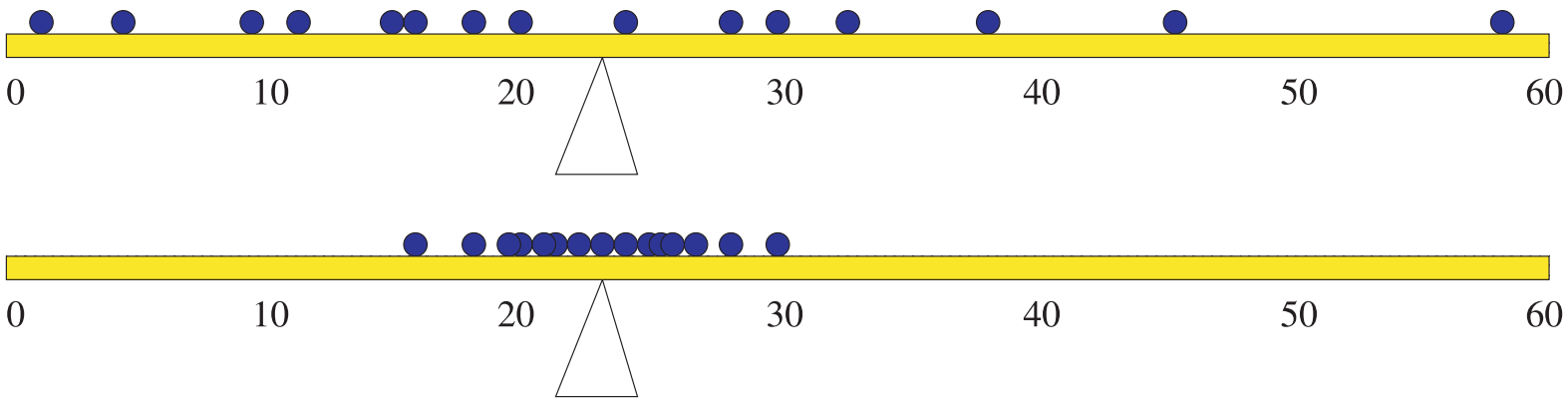
- $\bar{x}$ είναι το πιο σημαντικό μέτρο και αποτελεί εκτίμηση του μ
- για τον υπολογισμό του $\bar{x}$ χρησιμοποιούνται όλες οι παρατηρήσεις, για τη διάμεσο μόνο η τάξη τους
- το $\bar{x}$ επηρεάζεται από μακρινές τιμές, η διάμεσος όχι

Θέμα 3

Το συνοπτικό μέτρο του περικομμένου μέσου (trimmed mean): Παρουσίαση, ιδιότητες και παράδειγμα.

Μέτρα μεταβλητότητας

Η μεταβλητότητα ή διασπορά των παρατηρήσεων είναι ένα δεύτερο σημαντικό χαρακτηριστικό του δείγματος



Τα κυριότερα μέτρα διασποράς είναι:

δειγματικό εύρος R

Δεν είναι ανθεκτικό μέτρο και υπολογίζεται μόνο από τις δύο ακραίες τιμές

Μέτρα μεταβλητότητας

δειγματική διασπορά ή δειγματική διακύμανση s^2

Μετράει τη μεταβλητότητα των παρατηρήσεων γύρω από τη μέση τιμή.

Απόκλιση μιας x_i από τη μέση τιμή : $x_i - \bar{x}$

Το άθροισμα όλων των αποκλίσεων είναι 0!

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

Γι αυτό χρησιμοποιούμε τα τετράγωνα των αποκλίσεων

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

ισοδύναμα

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

Μέτρα μεταβλητότητας

Δύσκολο να ερμηνεύσουμε τη δειγματική διασπορά s^2

δειγματική τυπική απόκλιση s

Η δειγματική τυπική απόκλιση s είναι πιο κατάλληλο μέτρο γιατί επιδέχεται φυσική ερμηνεία

p-εκατοστιαία σημεία

p-εκατοστιαίο σημείο: ποσοστό παρατηρήσεων το πολύ $p\%$ είναι μικρότερες απ' αυτήν την
παρατήρηση $(0 \leq p < 1)$

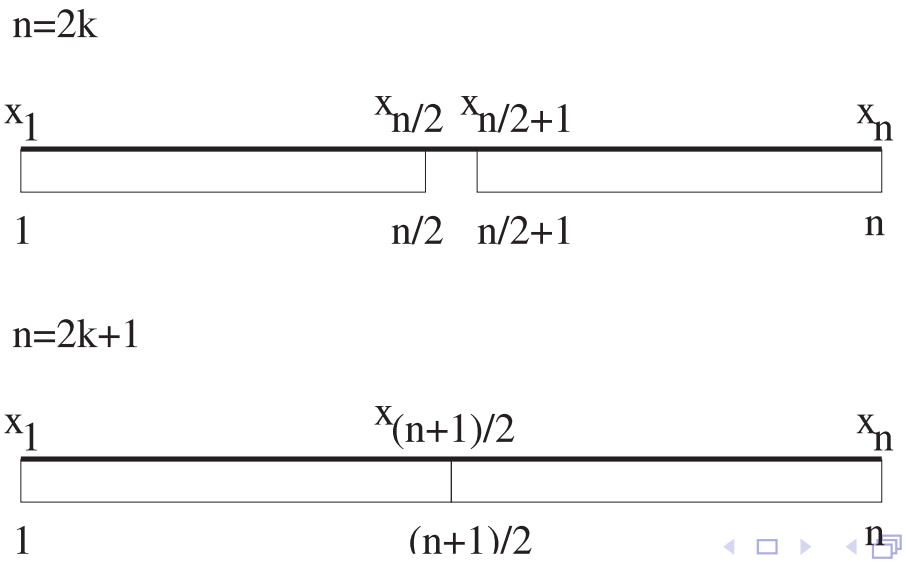
Η διάμεσος είναι το 50-εκατοστιαίο σημείο

Μέτρα μεταβλητότητας

Χαρακτηριστικά εκατοστιαία σημεία είναι τα *τεταρτομόρια*

- **πρώτο ή κατώτερο τεταρτομόριο** Q_1: το 25-εκατοστιαίο σημείο
- **τρίτο ή ανώτερο τεταρτομόριο** Q_3: το 75-εκατοστιαίο σημείο

Q_1 και Q_3 ορίζονται όπως η διάμεσος αλλά περιορίζοντας το σύνολο των δεδομένων στα αντίστοιχα υποσύνολα (κατώτερο ή ανώτερο μισό).

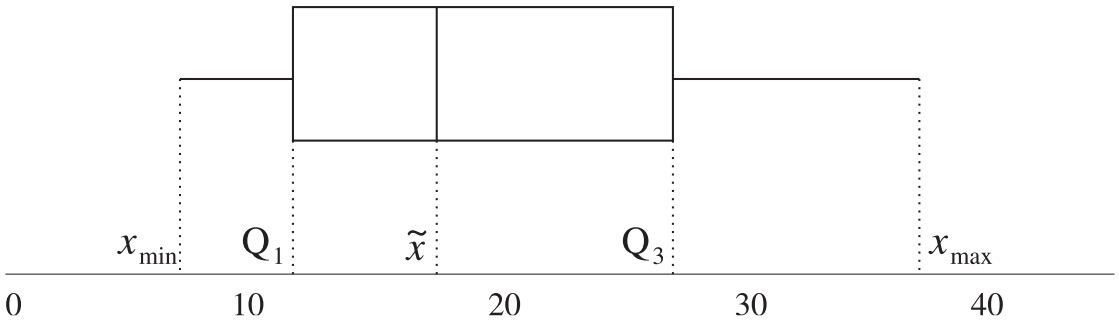


Μέτρα μεταβλητότητας

ενδοτεταρτομοριακό εύρος I

$I = Q_3 - Q_1$ είναι το εύρος που καλύπτουν τα μισά από τα δεδομένα που είναι πιο κοντά διάμεσο

σύνοψη των 5 αριθμών - θηκόγραμμα



Συνθήκες για αποδοχή κανονικής κατανομής από θηκόγραμμα

- $\tilde{x}$ όχι κοντά στο Q_1 ή στο Q_3
- το εύρος των τιμών στα δύο ακραία τεταρτομόρια να μη διαφέρει σημαντικά
- να μην υπάρχουν ακραίες τιμές

Παράδειγμα: περιεκτικότητα σε ραδιενέργεια του χάλυβα

Δίνεται η περιεκτικότητα σε ραδιενέργεια του χάλυβα σε 10 δοκίμια από ένα εργοστάσιο Α

Α/Α	εργοστάσιο Α	εργοστάσιο Β
1	0.40	0.11
2	0.51	0.13
3	0.51	0.26
4	0.54	0.27
5	0.55	0.33
6	0.59	0.37
7	0.63	0.52
8	0.67	0.65
9	0.75	
10	2.10	
Σύνολο	7.25	2.64

Παράδειγμα (συνέχεια)

δειγματική μέση τιμή:

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{7.25}{10} = 0.725$$

δειγματική διάμεσος:

$$\tilde{x} = \frac{x_{n/2} + x_{n/2+1}}{2} = \frac{x_5 + x_6}{2} = \frac{0.55 + 0.59}{2} = 0.57$$

εύρος τιμών δείγματος:

$$x_{\min} = 0.40 \qquad x_{\max} = 2.10 \quad \longrightarrow \quad R = 1.70$$

δειγματική διασπορά (Πρώτα το άθροισμα τετραγώνων)

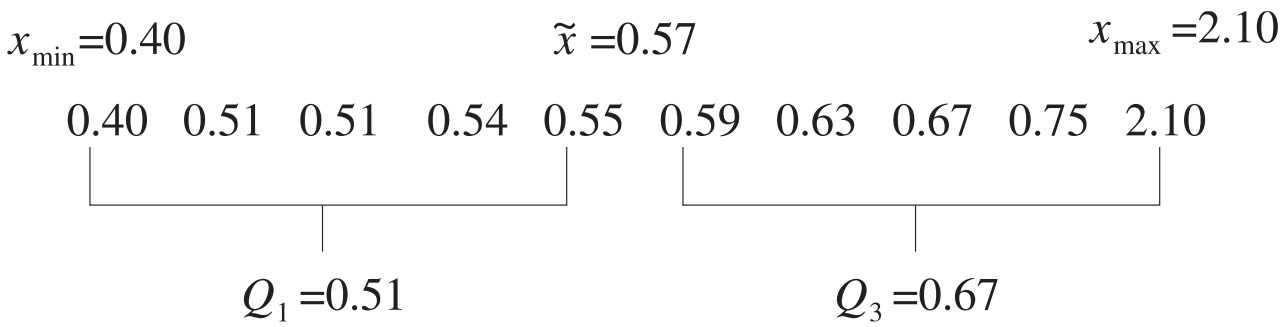
$$\sum_{i=1}^{10} x_i^2 = 0.40^2 + 0.51^2 + \cdots + 0.75^2 + 2.10^2 = 7.44$$

$$s^2 = \frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2\right) = \frac{1}{9}\left(7.44 - 10 \cdot 0.725^2\right) = 0.243$$

Παράδειγμα (συνέχεια)

δειγματική τυπική απόκλιση $s = \sqrt{s^2} = \sqrt{0.243} = 0.493$

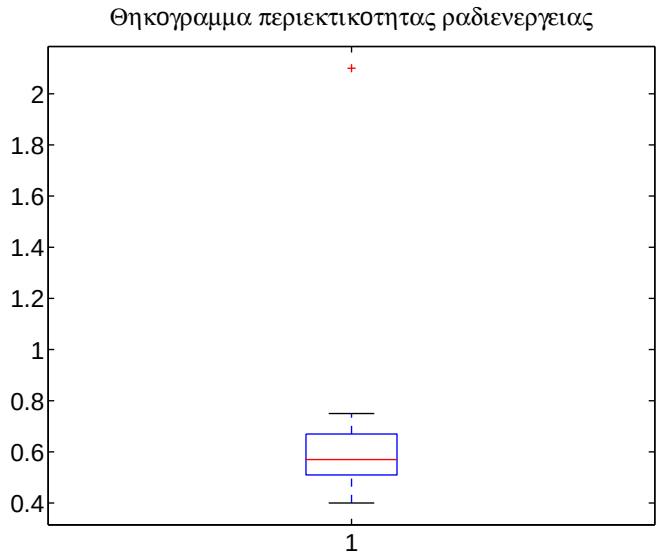
τεταρτομόρια



ενδοτεταρτομοριακό εύρος

$I = Q_3 - Q_1 = 0.67 - 0.51 = 0.16$

Θηκόγραμμα



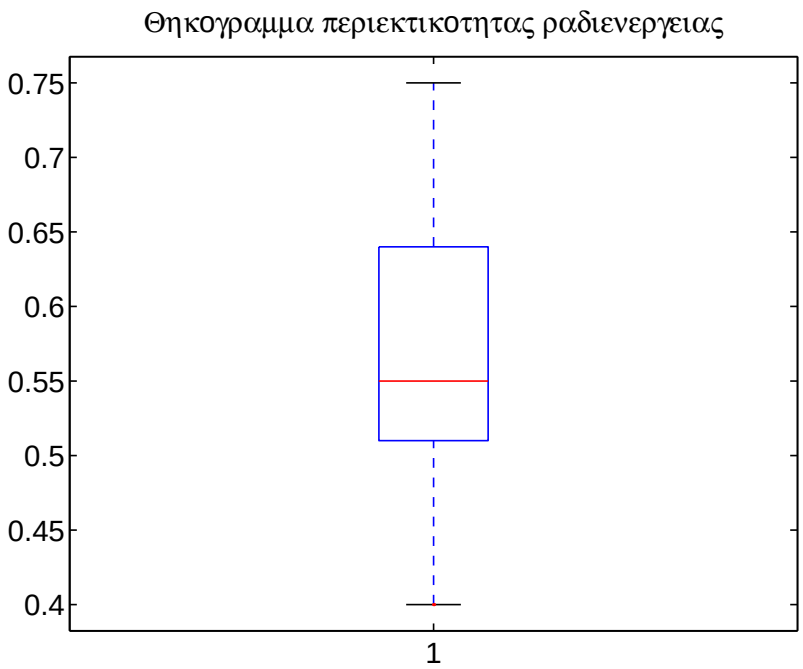
Παράδειγμα (συνέχεια)

Έστω ότι η ακραία τιμή οφείλεται σε σφάλμα μέτρησης (δεν είναι πραγματική)

Απαλοιφή ακραίας τιμής ($\longrightarrow$ 9 παρατηρήσεις)

δειγματική μέση τιμή	$\bar{x} = \frac{5.15}{9} = 0.572$
δειγματική διάμεσος	$\tilde{x} = x_{(n+1)/2} = x_5 = 0.55$
Διασπορά	$s^2 = \frac{1}{8}\left(5.15 - 9 \cdot 0.572^2\right) = 0.010$
Τυπική απόκλιση	$s = \sqrt{0.010} = 0.10$
Ελάχιστη τιμή	$x_{\min} = 0.40$
Μέγιστη τιμή	$x_{\max} = 0.75$
Εύρος	$R = 0.75 - 0.40 = 0.35$
Πρώτο τεταρτομόριο	(διάμεσος των $\{x_1, \ldots, x_5\}$) $Q_1 = x_3 = 0.51$
Τρίτο τεταρτομόριο	(διάμεσος των $\{x_5, \ldots, x_9\}$) $Q_3 = x_7 = 0.63$
Ενδοτεταρτομοριακό εύρος	$I = 0.63 - 0.51 = 0.12$

Παράδειγμα (συνέχεια)

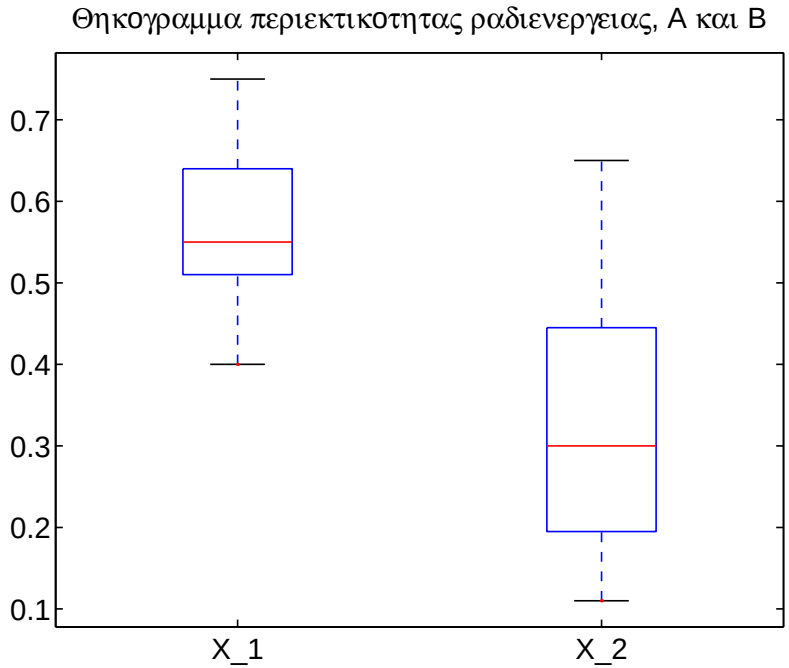


Παράδειγμα (συνέχεια)

Σύγκριση περιεκτικότητας ραδιενέργειας σε χάλυβα από δύο εργοστάσια

Μέτρο	X_1	X_2
Μέση τιμή	$\bar{x}_1 = 0.572$	$\bar{x}_2 = 0.33$
Διάμεσος	$\tilde{x}_1 = 0.55$	$\tilde{x}_2 = 0.30$
Διασπορά	$s_1^2 = 0.010$	$s_2^2 = 0.034$
Τυπική απόκλιση	$s_1 = 0.10$	$s_2 = 0.18$
Ελάχιστη τιμή	$x_{1,\min} = 0.40$	$x_{2,\min} = 0.11$
Μέγιστη τιμή	$x_{1,\max} = 0.75$	$x_{1,\max} = 0.65$
Εύρος	$R_1 = 0.35$	$R_2 = 0.54$
Πρώτο τεταρτομόριο	$Q_{1,1} = 0.51$	$Q_{2,1} = 0.195$
Τρίτο τεταρτομόριο	$Q_{1,3} = 0.63$	$Q_{2,3} = 0.445$
Ενδοτεταρ. εύρος	$I_1 = 0.12$	$I_2 = 0.250$

Παράδειγμα (συνέχεια)



- Είναι η κατανομή ίδια;
- Είναι η διασπορά ίδια;
- Είναι η μέση τιμή ίδια;

Άσκηση

Έγιναν 15 μετρήσεις της συγκέντρωσης διαλυμένου οξυγόνου (Δ.Ο.) σε ένα ποτάμι (σε mg/l)

1.8	2.0	2.1	1.7	1.2	2.3	2.5	2.9	1.6	2.2	2.3	1.8	2.4	1.6	1.9

1. Υπολογίστε τα μέτρα θέσης και μεταβλητότητας για τα δεδομένα του δείγματος και σχηματίστε το κατάλληλο θηκόγραμμα.
2. Σχολιάστε αν φαίνεται η συγκέντρωση Δ.Ο. στο νερό του ποταμού να ακολουθεί κανονική κατανομή.