Περιγραφική Στατιστική ΙΙ (*Descriptive Statistics*)

- **Οργάνωση** και **Γραφική αναπαράσταση** στατιστικών δεδομένων
- **Οπτικοποίηση** των δεδομένων
- Αριθμητικά **περιγραφικά μέτρα**

Φυλλογράμματα (*stem-leaf notes*)

- **Διατηρεί την ατομικότητα** των παρατηρήσεων (που «χάνεται» με τα ιστογράμματα).

- Κάθε παρατήρηση χωρίζεται σε **2 μέρη**:
  - το **στέλεχος** ή **οδηγός** (***stem***) και
  - το **φύλλο** (***leaf***)

- Υπάρχουν **εναλλακτικοί** τρόποι διαχωρισμού των δεδομένων, ορίζοντας τον **οδηγό** και το **φύλλο** ανάλογα με τον τύπο τους

Βήματα *κατασκευής φυλλογραμμάτων*

i. **Επιλέγουμε** πρώτα τα **στελέχη** (ή οδηγούντα ψηφία) και τα **φύλλα**

ii. **Διατάσσουμε** τα στελέχη κατά αύξουσα

iii. **Τοποθετούμε** τα (*διαφορετικά*) **φύλλα στην ίδια γραμμή** των αντίστοιχων **στελεχών**

iv. **Ελέγχουμε** εάν έχουν καταγραφεί όλα τα φύλλα (αριθμός τους ίσος με το συνολικό πλήθος παρατηρήσεων)

Παράδειγμα

Σύνολο δεδομένων

X={136, 111, 120, 105, 113, 116, 99, 110, 125, 139, 122, 96}

- **στέλεχος (*stem*)** τις **δεκάδες**
- **φύλλο (*leaf*)** τις **μονάδες**

Stem (δεκάδες)	***Leaf (μονάδες)***
9	6 9
10	5
11	0 1 3 6
12	0 2 5
13	6 9

2. Αριθμητικά περιγραφικά μέτρα

- Έχουμε στη διάθεσή μας ένα **σύνολο δεδομένων** (**δείγμα**).
- Εκτελούμε **υπολογισμούς πάνω στα δεδομένα**, εξάγοντας χρήσιμες ποσότητες οι οποίες:
  - χαρακτηρίζουν **μονοσήμαντα** το δείγμα, και
  - **παρέχουν τάσεις** και **ροπές** των δεδομένων
- Αποτελούν (*συνήθως*) τα **στατιστικά στοιχεία** του διαθέσιμου δείγματος

Αριθμητικά περιγραφικά μέτρα

[A]. Μέτρα θέσης ή κεντρικής τάσης (*central tendency*)

- Περιγράφουν την **θέση** της **κατανομής** ή του **κέντρου** των **δεδομένων**.

- Δημοφιλέστερα μέτρα τάσης είναι:
  - ***η μέση τιμή***,
  - ***η κορυφή*** και
  - ***η διάμεσος***

Ο **δειγματικός μέσος** (***mean***)

- Είναι ο (γνωστός) **μέσος όρος των διαθέσιμων παρατηρήσεων**

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

Κορυφή ή επικρατούσα τιμή (*mode*)

- είναι η επικρατέστερη τιμή του δείγματος, δηλ. αυτή με την **μέγιστη συχνότητα εμφάνισης.**

Διάμεσος (*median*)

- είναι η τιμή ***δ*** που χωρίζει το δείγμα σε *2* ίσα μέρη, ώστε ο αριθμός των παρατηρήσεων που είναι **≤ *δ*** να είναι ίσος (**50%**) με τον αριθμό των δεδομένων που είναι **≥ *δ***.
- Έτσι, αν διατάσουμε τις *n* παρατηρήσεις του δείγματος:

τότε $$\delta = \begin{cases} x_{(r)} & \alpha\nu \quad n = 2r-1 \\ \dfrac{x_{(r)} + x_{(r+1)}}{2} & \alpha\nu \quad n = 2r \end{cases}$$

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n-1)} \le x_{(n)}$$

- **Παρατήρηση**: Αν η **κατανομή είναι συμμετρική**, τότε ο μέσος, η κορυφή και η διάμεσος **συμπίπτουν** (όπως π.χ. στην **κανονική κατανομή**).

Ποσοστημόρια (*quantiles*): μέτρο σχετικής θέσης

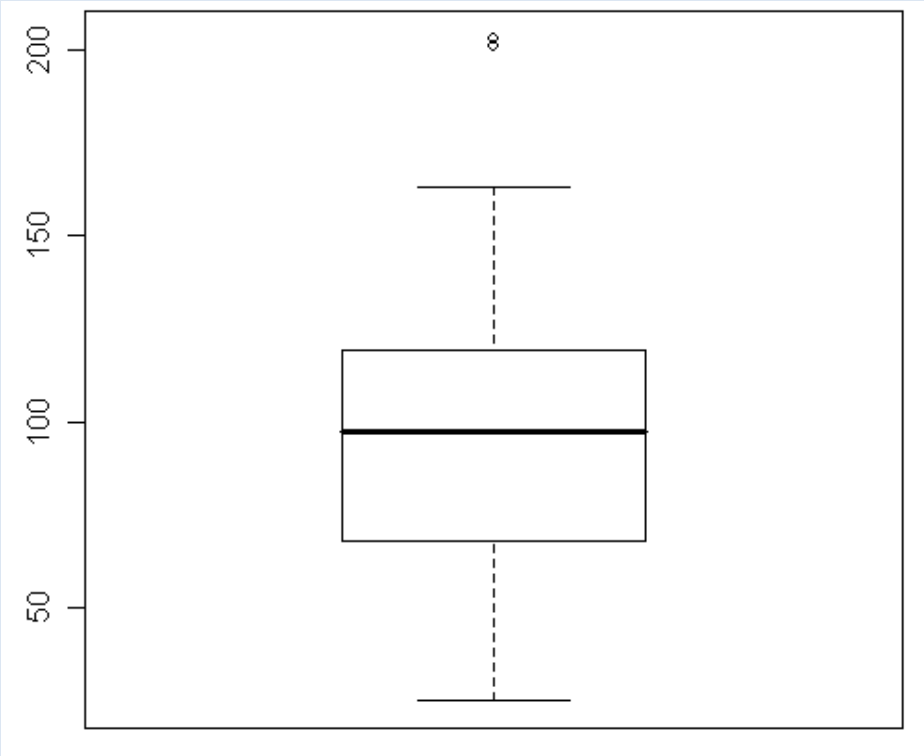
- Γενίκευση της διαμέσου (a=50%)
- Το ***a-οστό ποσοστημόριο*** είναι η τιμή για την οποία το *a%* των τιμών είναι **μικρότερο** και το *(100 - a)%* είναι **μεγαλύτερο** από την τιμή αυτή.
- Για ***a = {1, 2, …, 99}*** έχουμε **εκατοστημόρια** (***quantiles***).
- Για ***a = {10, 20, …, 90}*** έχουμε **δεκατημόρια**
- Για ***a = {25, 50, 75}*** έχουμε **τεταρτημόρια (*quartiles*)**

a=25 : Q_1 πρώτο τεταρτημόριο

a=75 : Q_3 τρίτο τεταρτημόριο

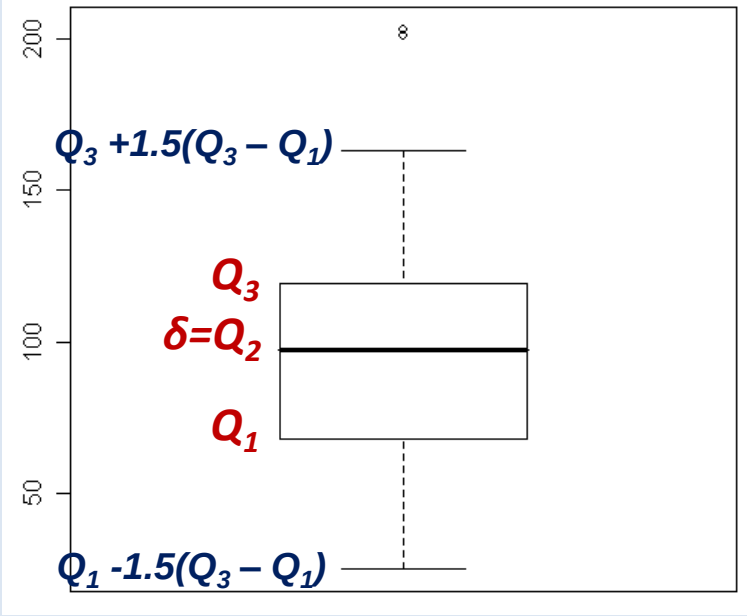
a=50 : Q_2 δεύτερο τεταρτημόριο (δηλ. η **διάμεσος**).

Θηκογράμματα (*box plots*)



- Τρόπος παρουσίασης των κυριότερων χαρακτηριστικών μιας κατανομής μέσω ενός γραφήματος

Βήματα κατασκευής boxplot

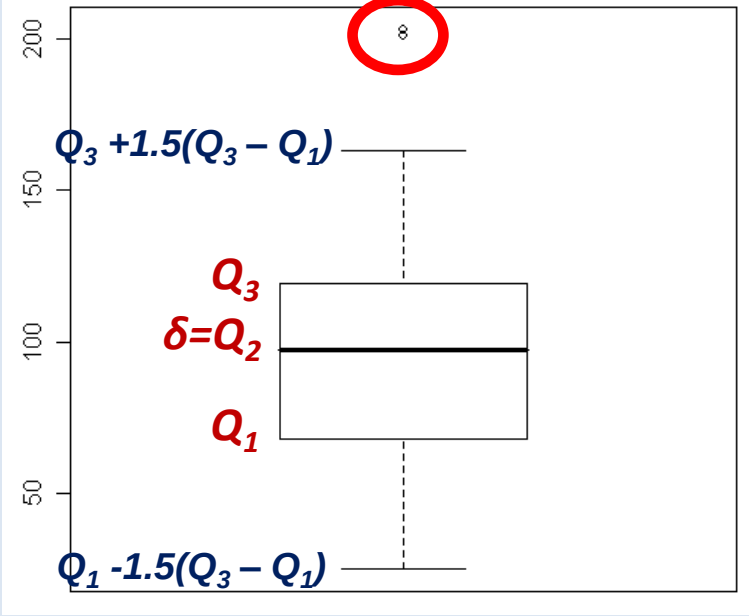


1. Αρχικά βρίσκουμε τα δύο **τεταρτημόρια *Q1*, *Q3*** και τη διάμεσο ***δ*** **(***δηλ. το* Q_2**)**.
2. Κατασκευάζουμε ένα ***ορθογώνιο*** με κάτω πλευρά το ***Q1*** και πάνω πλευρά το ***Q3***. Η διάμεσος παριστάνεται ως ευθύγραμμο τμήμα μέσα στο ορθογώνιο παράλληλο με τις βάσεις.
3. Φέρουμε διακεκομμένες κάθετες γραμμές από τα μέσα των βάσεων του ορθογωνίου μέχρι τις **οριακές τιμές** που προκύπτουν.

Q3 + 1.5(Q2-Q1) : **άνω οριακή**

Q3 - 1.5(Q2-Q1) : **κάτω οριακή**

Βήματα κατασκευής boxplot



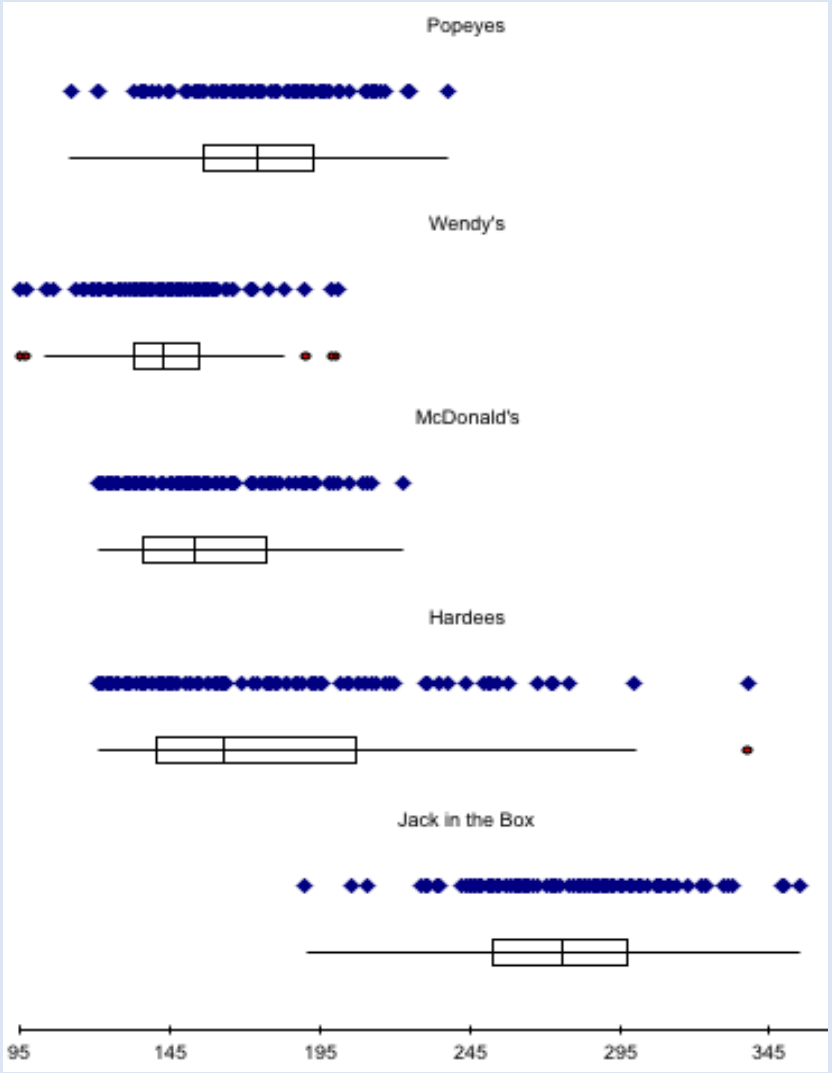
1. Αρχικά βρίσκουμε τα δύο τεταρτημόρια ***Q1***, ***Q3*** και τη διάμεσο ***δ*** **(***δηλ. το* Q_2**)**.
2. Κατασκευάζουμε ένα ***ορθογώνιο*** με κάτω πλευρά το ***Q1*** και πάνω πλευρά το ***Q3***. Η διάμεσος παριστάνεται ως ευθύγραμμο τμήμα μέσα στο ορθογώνιο παράλληλο με τις βάσεις.
3. Φέρουμε διακεκομμένες γραμμές από τα μέσα των βάσεων του ορθογωνίου μέχρι τις **οριακές τιμές** που προκύπτουν.
4. Κάθε σημείο που πέφτει έξω από το εύρος των δύο οριακών τιμών λέγεται **ακραία τιμή** (***outlier***) και παριστάνεται με ένα ιδιαίτερο σύμβολο (π.χ. *)

Πλεονεκτήματα των θηκογραμμάτων

- Τα θηκογράμματα μας δίνουν το διάστημα τιμών του **50%** των **συχνότερων παρατηρήσεων** – μεταξύ του 1ου και 3ου τεταρτημορίου (Q1, Q3)

- Οι επεκτεινόμενες γραμμές και η θέση της διαμέσου μας δίνουν ένα **βαθμό συμμετρικότητας** της κατανομής

- Δυνατότητα **μελέτης** των **ακραίων τιμών** και πιθανώς του βαθμού επίδρασής τους



Μέτρα διασποράς

Εκφράζουν **αποκλίσεις των τιμών** μιας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης

- **Εύρος (*range*)** τιμών ή κύμανση = $\boldsymbol{max\{ x_i \} - min\{ x_i \}}$
  - εύκολο στον υπολογισμό, αλλά μικρής αξιοπιστίας (*ακραίες τιμές*)
- **Ενδοτεταρτημοριακή απόκλιση (*interquartile range*)** = $\boldsymbol{Q_3 - Q_1}$
  - Μετράει το **άπλωμα του 50%** των μεσαίων (συχνότερων) τιμών των παρατηρήσεων.
  - Μεγάλες τιμές αυτής της στατιστικής υποδεικνύει μεταβλητότητας. υψηλό επίπεδο
  - Μικρές τιμές (διάστημα) σημαίνει μεγάλη συγκέντρωση τιμών και άρα μικρότερη διασπορά τιμών.

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

- **Μέση απόκλιση (*mean deviation*)**

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

Ο αριθμητικός μέσος των αποκλίσεων των τιμών από το μέσον τους

- **Δειγματική διασπορά ή διακύμανση (*variance*)**

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

<u>Παρατήρηση:</u>

στον παρονομαστή έχουμε ***n-1*** (και **όχι *n***) για καλύτερη εκτίμηση του (***αμερόληπτος εκτιμητής***)

- **Δειγματική τυπική απόκλιση (*standard deviation*)** $S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

[Γ]. Μέτρα σχετικής μεταβλητότητας

Συντελεστής μεταβλητότητας (*coefficient of variation*)

$$v = \frac{s}{\bar{x}} = \frac{\text{τυπική απόκλιση}}{\text{δειγματικός μέσος}} \left(\times 100\%\right)$$

- Μέτρο **σχετικής μεταβλητότητας τιμών**.
- Χρησιμοποιείται **για συγκρίσεις** ανάμεσα σε δείγματα που είτε εκφράζονται σε διαφορετικές μονάδες μέτρησης, είτε έχουν διαφορετικές μέσες τιμές.
- Δεχόμαστε ότι δύο δείγματα θα είναι **ομογενή** αν ο συντελεστής μεταβλητότητας τους διαφέρει **το πολύ 10%.**
- Μειονέκτημα όταν ο μέσος πλησιάζει στο μηδέν (τότε **δεν** πρέπει να χρησιμοποιείται).

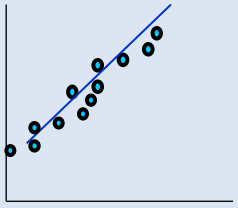
Συνδιακύμανση (***covariance***):

- μέτρο κατευθυντικότητας δύο μεταβλητών

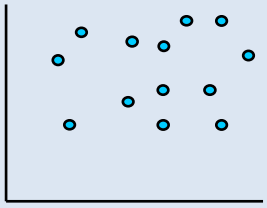
$$\text{Sample covariance} = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

Συντελεστής συσχέτισης (***correlation coefficient***):

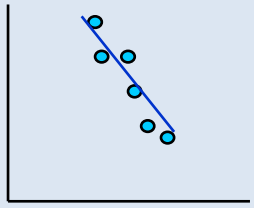
- μέτρο γραμμικότητας μεταξύ των δύο μεταβλητών



r -> 1



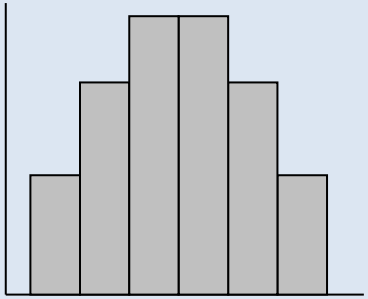
r -> 0



r -> -1

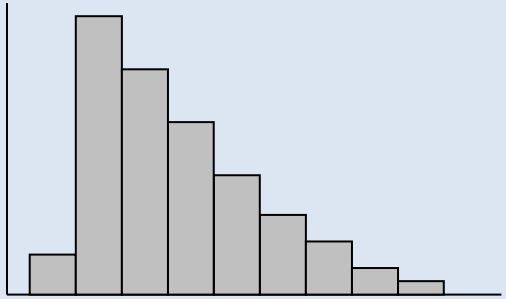
$$r = \frac{s_{xy}}{s_x s_y} \in [-1, 1]$$

[Δ]. Μέτρα ασυμμετρίας



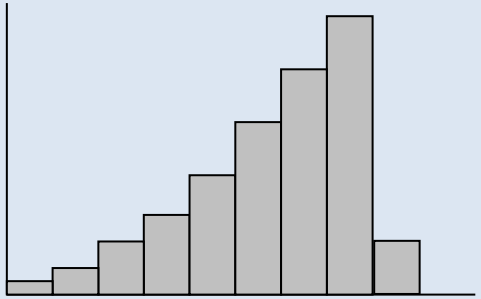
$\bar{x} = \delta = M_0$

Συμμετρική κατανομή
Η κορυφή (***M_0***), ο μέσος και η διάμεσος συμπίπτουν



$M_0 < \delta < \bar{x}$

Θετική συμμετρία
Οι περισσότερες παρατηρήσεις είναι δεξιά της κορυφής (***M_0***).



$\bar{x} < \delta < M_0$

Αρνητική συμμετρία
Οι περισσότερες παρατηρήσεις είναι αριστερά της κορυφής (***M_0***).

- **Συντελεστής ασυμμετρίας *Pearson***

$$Y_1 = \frac{\bar{x} - M_0}{s} \qquad Y_2 = \frac{3(\bar{x} - \delta)}{s}$$

Αν *Y=0* => συμμετρία
Αν *Y<0* => αρνητική συμμετρία
Αν *Y>0* => θετική συμμετρία

[E]. Μετασχηματισμοί δεδομένων

Z-score $$z = \frac{x - \bar{x}}{s}$$

- Έχουμε τον μετασχηματισμό: $x_i \to z_i = \frac{x_i - \bar{x}}{s}$
- Μετασχηματισμός κανονικότητας των δεδομένων
- Ισχύει ότι $x_i = \bar{x} + sz_i$

✓ Δηλαδή, το z_i εκφράζει τον **αριθμό των τυπικών αποκλίσεων** που το x_i διαφέρει από το μέσον του

Παράδειγμα

- Ο αριθμός των ελαττωματικών μπαταριών που βρέθηκαν σε 72 σωρούς παραγωγής των 500 μπαταριών ήταν

3	7	24	6	9	7	1	19
9	0	6	15	4	5	7	11
5	11	1	13	2	4	3	3
17	2	14	4	22	3	10	12
26	7	8	11	1	10	21	7
2	20	9	2	0	1	20	9
13	18	5	14	12	3	8	1
1	5	2	17	15	13	3	16
4	12	4	6	3	8	22	5

(α) να παραστούν σε μορφή φυλλογραφήματος

(β) να υπολογιστούν: (i) ο **δειγματικός μέσος**, (ii) η **διάμεσος**, (iii) η **κορυφή**, (iv) η **διασπορά**, (v) ο **συντελεστής μεταβλητότητας**.

(γ) να κατασκευαστεί το **θηκόγραμμα**

(δ) να κατασκευαστούν το **ιστόγραμμα** σχετικών συχνοτήτων.

Λύση

(α) φυλλογράφημα (*stem-leaf notes*)

stem	*leaf*
0	0 0 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4 4 5 5 5 5 5 6 6 6 7 7 7 7 7 8 8 8 9 9 9 9
1	0 0 1 1 1 2 2 2 3 3 3 4 4 5 5 6 7 7 8 9
2	0 0 1 2 2 4 6

(β) **δειγματικός μέσος**: 8 65

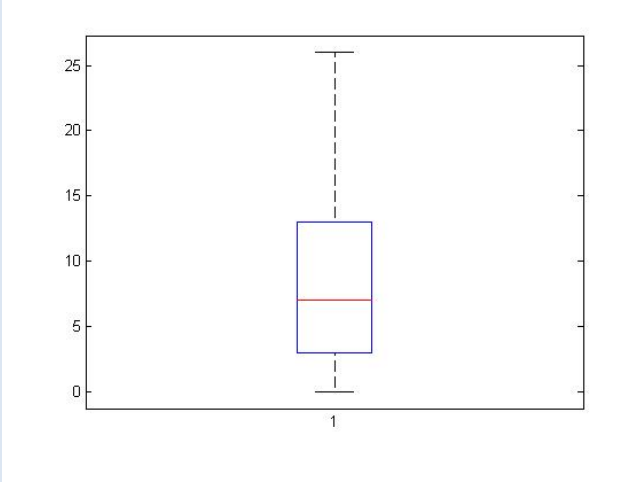
διάμεσος = 7

κορυφή = 3

δειγματική διακύμανση = 43.61

συντελεστής μεταβλητότητας = s/μ = τυπική απόκλιση / μέσος = 0.763

(γ) **θηκόγραμμα** (*boxplot(x))*



(δ) **Ιστόγραμμα** σχετικών συχνοτήτων (*με 10 ή 20 bins*)