Παράρτημα Α: Περιγραφική Στατιστική και Οπτικοποίηση Δεδομένων για Μηχανική Μάθηση

Fundamentals of Machine Learning for Predictive Data Analytics

Αθανάσιος Σάκκας, ΟΠΑ

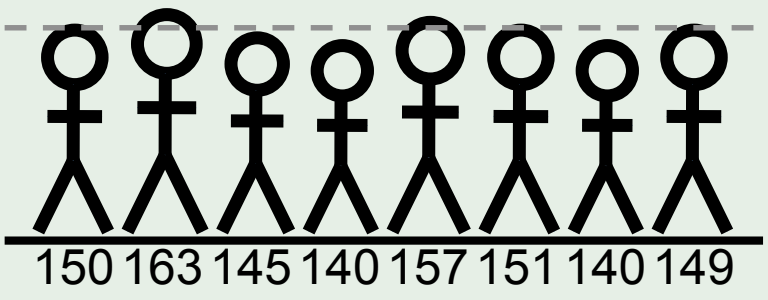
Περιγραφική Στατιστική

- Ο **αριθμητικός μέσος** (ή **δειγματικός μέσος** ή απλώς **μέσος**) ενός συνόλου n τιμών για ένα χαρακτηριστικό a, $a_1, a_2 \ldots a_n$, συμβολίζεται με το σύμβολο $\overline{a}$, και υπολογίζεται ως:

$$\overline{a} = \frac{1}{n} \sum_{i=1}^{n} a_i$$

Example

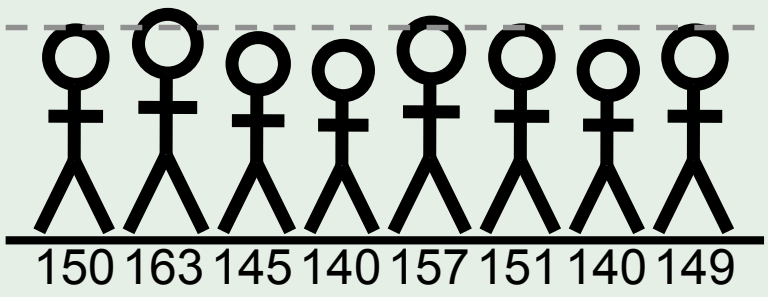
ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149



Σχήμα: Τα μέλη μιας σχολικής ομάδας μπάσκετ. Η διακεκομμένη γκρι γραμμή δείχνει τον αριθμητικό μέσο του ύψους των παικτών.

Example

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149



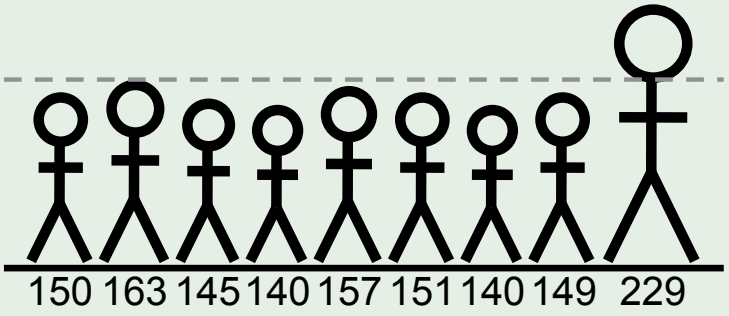
Σχήμα: Τα μέλη μιας σχολικής ομάδας μπάσκετ. Η διακεκομμένη γκρι γραμμή δείχνει τον αριθμητικό μέσο του ύψους των παικτών.

$$\begin{aligned}\overline{\text{Height}} &= \frac{1}{8} \times (150 + 163 + 145 + 140 + 157 + 151 + 140 + 149) \\ &= 149.375\end{aligned}$$

- Ο **αριθμητικός μέσος** είναι ένα μέτρο της **κεντρικής τάσης** ενός **δείγματος** (για τους σκοπούς μας, δείγμα είναι απλώς ένα σύνολο τιμών ενός χαρακτηριστικού σε έναν ABT).
- Κάθε μέτρο **κεντρικής τάσης**, ωστόσο, είναι απλώς μια προσέγγιση.

Example

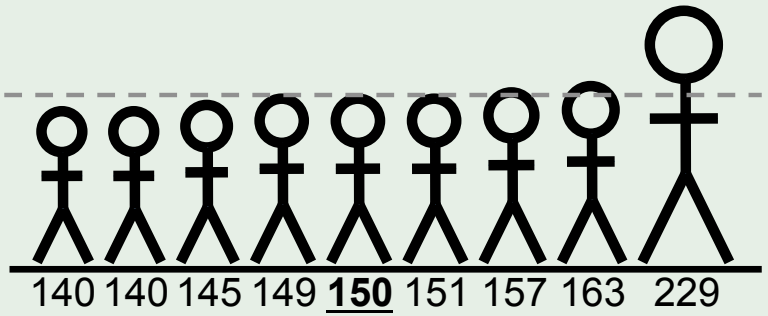
- Έστω ότι η ομάδα μπάσκετ καταφέρνει να αποκτήσει έναν *ringer* με ύψος 229cm



- Ο αριθμητικός μέσος για όλη την ομάδα είναι 158.235cm και πλέον δεν αντιπροσωπεύει την κεντρική τάση της ομάδας.
- Μια ασυνήθιστα μεγάλη ή μικρή τιμή όπως αυτή ονομάζεται **ακραία τιμή** (outlier) — ο αριθμητικός μέσος είναι πολύ ευαίσθητος στις ακραίες τιμές.

- Η **διάμεσος** ενός συνόλου τιμών μπορεί να υπολογιστεί ταξινομώντας τις τιμές από τη μικρότερη στη μεγαλύτερη και επιλέγοντας τη μεσαία τιμή.
- Αν υπάρχει άρτιος αριθμός τιμών στο δείγμα, τότε η διάμεσος προκύπτει υπολογίζοντας τον αριθμητικό μέσο των δύο μεσαίων τιμών.

Example



Σχήμα: Τα μέλη της σχολικής ομάδας μπάσκετ ταξινομημένα κατά ύψος· η διακεκομμένη γκρι γραμμή δείχνει τη **διάμεσο**.

ID	4	7	3	8	1	6	5	2	9
Height	140	140	145	149	**150**	151	157	163	229

- Μετράμε επίσης τη **διακύμανση** των δεδομένων μας.
- Ουσιαστικά, το μεγαλύτερο μέρος της στατιστικής, και κατ' επέκταση της αναλυτικής, αφορά την περιγραφή και την κατανόηση της διακύμανσης.

- Το απλούστερο μέτρο διακύμανσης είναι το **εύρος**:

$$\text{range} = max(a) - min(a)$$

Example

Ποιο είναι το εύρος των υψών των δύο ομάδων μπάσκετ;

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149

ID	1	2	3	4	5	6	7	8
Height	192	102	145	165	126	154	123	188

Example

Ποιο είναι το εύρος των υψών των δύο ομάδων μπάσκετ;

$$\text{range} = 163 - 140 = 23$$

$$\text{range} = 192 - 102 = 90$$

- Η **διακύμανση** ενός δείγματος μετρά τη μέση απόκλιση κάθε τιμής ενός δείγματος από τον μέσο του δείγματος.
- Η **διακύμανση** των n τιμών ενός χαρακτηριστικού a, $a_1, a_2 \ldots a_n$, συμβολίζεται $var(a)$ και υπολογίζεται ως:

$$var(a) = \frac{\sum_{i=1}^{n} (a_i - \overline{a})^2}{n-1}$$

Example

Ποια είναι η διακύμανση των υψών των δύο ομάδων μπάσκετ;

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149

ID	1	2	3	4	5	6	7	8
Height	192	102	145	165	126	154	123	188

Example

$$\textit{var}(\text{Height}) = \frac{(150 - 149.375)^2 + (163 - 149.375)^2 + \ldots + (149 - 149.375)^2}{8 - 1}$$
$$= 63.125$$

$$\textit{var}(\text{Height}) = \frac{(192 - 149.375)^2 + (102 - 149.375)^2 + \ldots + (188 - 149.375)^2}{8 - 1}$$
$$= 1,011.41071$$

- Η **τυπική απόκλιση**, *sd*, ενός δείγματος υπολογίζεται παίρνοντας την τετραγωνική ρίζα της **διακύμανσης** του δείγματος:

$$sd(a) = \sqrt{var(a)} \tag{1}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} (a_i - \overline{a})^2}{n-1}} \tag{2}$$

Example

Ποια είναι η τυπική απόκλιση των υψών των δύο ομάδων μπάσκετ;

ID	1	2	3	4	5	6	7	8
Height	150	163	145	140	157	151	140	149

ID	1	2	3	4	5	6	7	8
Height	192	102	145	165	126	154	123	188

Example

$$\begin{aligned} sd(\text{Height}) &= \sqrt{63.125} \\ &= 7.9451\ldots \end{aligned}$$

$$\begin{aligned} sd(\text{Height}) &= \sqrt{1,011.41071} \\ &= 31.8026\ldots \end{aligned}$$

- Τα **ποσοστημόρια** (percentiles) είναι ένα ακόμη χρήσιμο μέτρο της διακύμανσης των τιμών ενός χαρακτηριστικού: ένα ποσοστό $\frac{i}{100}$ των τιμών σε ένα δείγμα λαμβάνει τιμές μικρότερες ή ίσες από το i^{th} ποσοστημόριο αυτού του δείγματος.

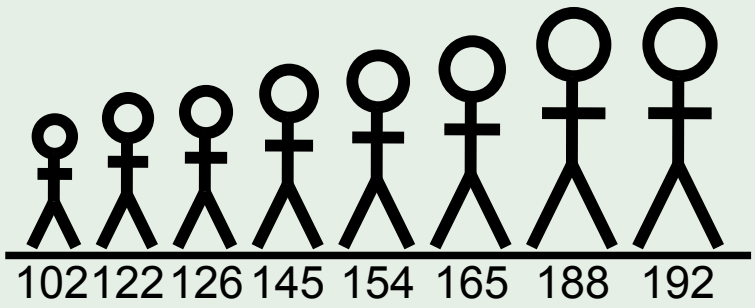
- Για να υπολογίσουμε το i^{th} ποσοστημόριο των n τιμών ενός χαρακτηριστικού a, $a_1, a_2 \ldots a_n$:
  - Πρώτα ταξινομούμε τις τιμές σε αύξουσα σειρά και μετά πολλαπλασιάζουμε το n με $\frac{i}{100}$ για να προσδιορίσουμε το *index*.
  - Αν το *index* είναι ακέραιος αριθμός, παίρνουμε την τιμή σε εκείνη τη θέση στη ταξινομημένη λίστα τιμών ως το i^{th} ποσοστημόριο.
  - Αν το *index* δεν είναι ακέραιος αριθμός, τότε **παρεμβάλλουμε** (interpolate) την τιμή για το i^{th} ποσοστημόριο ως:

    $$i^{th}\text{percentile} = (1 - index\_f) \times a_{index\_w} + index\_f \times a_{index\_w+1}$$

    όπου *index_w* είναι το ακέραιο μέρος του *index*, *index_f* είναι το κλασματικό μέρος του *index* και a_{index_w} είναι η τιμή στη ταξινομημένη λίστα στη θέση *index_w*.

Example

ID	2	7	5	3	6	4	8	1
Height	102	123	126	145	154	165	188	192



- Ποιο είναι το 25^{th} ποσοστημόριο των υψών της ομάδας μπάσκετ;
- Ποιο είναι το 80^{th} ποσοστημόριο των υψών της ομάδας μπάσκετ;

Example

- Για να υπολογίσουμε το 25^{th} ποσοστημόριο, πρώτα υπολογίζουμε το *index* ως $\frac{25}{100} \times 8 = 2$. Άρα, το 25^{th} ποσοστημόριο είναι η δεύτερη τιμή στην ταξινομημένη λίστα, δηλαδή 123.
- Για να υπολογίσουμε το 80^{th} ποσοστημόριο, πρώτα υπολογίζουμε το *index* ως $\frac{80}{100} \times 8 = 6.4$. Επειδή το *index* δεν είναι ακέραιος αριθμός, θέτουμε *index_w* ίσο με το ακέραιο μέρος του *index*, δηλαδή 6, και *index_f* ίσο με το κλασματικό μέρος, δηλαδή 0.4. Τότε μπορούμε να υπολογίσουμε το 80^{th} ποσοστημόριο ως:

$$(1 - 0.4) \times 165 + 0.4 \times 188 = 174.2$$

- Μπορούμε να χρησιμοποιήσουμε τα ποσοστημόρια για να περιγράψουμε ένα ακόμη μέτρο διακύμανσης, γνωστό ως **ενδοτεταρτημοριακό εύρος** (inter-quartile range).
- Το ενδοτεταρτημοριακό εύρος υπολογίζεται ως η διαφορά μεταξύ του 25^{th} ποσοστημορίου και του 75^{th} ποσοστημορίου.[1]

[1] Αυτά τα ποσοστημόρια είναι επίσης γνωστά ως **κάτω τεταρτημόριο** (ή 1^{st} τεταρτημόριο) και **άνω τεταρτημόριο** (ή 3^{rd} τεταρτημόριο), εξ ου και το όνομα ενδοτεταρτημοριακό εύρος.

Example

Για τα ύψη της πρώτης ομάδας μπάσκετ, το ενδοτεταρτημοριακό εύρος είναι $151 - 140 = 11$, ενώ για τη δεύτερη ομάδα είναι $165 - 123 = 42$.

- Για τα κατηγορικά χαρακτηριστικά μας ενδιαφέρουν κυρίως οι **μετρήσεις συχνότητας** και οι **αναλογίες**.
  - Η μέτρηση συχνότητας για κάθε επίπεδο ενός κατηγορικού χαρακτηριστικού υπολογίζεται μετρώντας πόσες φορές εμφανίζεται αυτό το επίπεδο στο δείγμα.
  - Η αναλογία για κάθε επίπεδο υπολογίζεται διαιρώντας τη συχνότητα αυτού του επιπέδου με το συνολικό μέγεθος του δείγματος.
  - Οι συχνότητες και οι αναλογίες παρουσιάζονται συνήθως σε έναν **πίνακα συχνοτήτων**.
- Η **επικρατούσα τιμή** (mode) είναι ένα μέτρο της κεντρικής τάσης ενός κατηγορικού χαρακτηριστικού και είναι απλώς το συχνότερο επίπεδο.
- Συχνά υπολογίζουμε και μια **δεύτερη επικρατούσα τιμή** (second mode), η οποία είναι το δεύτερο συχνότερο επίπεδο ενός χαρακτηριστικού.

Πίνακας: Ένα σύνολο δεδομένων που δείχνει τις θέσεις και τα εβδομαδιαία έξοδα προπόνησης μιας σχολικής ομάδας μπάσκετ.

ID	Position	Training Expenses	ID	Position	Training Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41

Πίνακας: Ένας πίνακας συχνοτήτων για το χαρακτηριστικό Position από το σύνολο δεδομένων της επαγγελματικής ομάδας μπάσκετ στον Πίνακα 4 [34].

Επίπεδο	Πλήθος	Αναλογία
guard	8	40%
forward	7	35%
center	5	25%

- Στη στατιστική είναι πολύ σημαντικό να κατανοούμε τη διαφορά μεταξύ **πληθυσμού** και **δείγματος**.
- Ο όρος πληθυσμός χρησιμοποιείται στη στατιστική για να αναπαραστήσει όλες τις δυνατές μετρήσεις ή εκβάσεις που μας ενδιαφέρουν σε μια συγκεκριμένη μελέτη ή ανάλυση.
- Ο όρος δείγμα αναφέρεται στο υποσύνολο του πληθυσμού που επιλέγεται για ανάλυση.
- Το **περιθώριο σφάλματος** που αναφέρεται στα αποτελέσματα δημοσκοπήσεων λαμβάνει υπόψη το γεγονός ότι το αποτέλεσμα βασίζεται σε δείγμα από έναν πολύ μεγαλύτερο πληθυσμό.

Πίνακας: Ορισμένα αποτελέσματα δημοσκοπήσεων στην πορεία προς τις προεδρικές εκλογές των ΗΠΑ το 2012.

Poll	Obama	Romney	Other	Date	Margin of Error	Sample Size
Pew Research	50	47	3	04-Nov	± 2.2	$2,709$
Gallup	49	50	1	04-Nov	± 2.0	$2,700$
ABC News/Wash Pos	50	47	3	04-Nov	± 2.5	$2,345$
CNN/Opinion Research	49	49	2	04-Nov	± 3.5	963
Pew Research	50	47	3	03-Nov	± 2.2	$2,709$
ABC News/Wash Post	49	48	3	03-Nov	± 2.5	$2,069$
ABC News/Wash Post	49	49	2	30-Oct	± 3.0	$1,288$

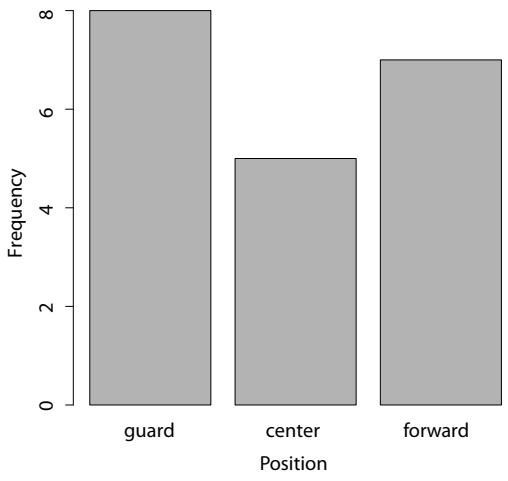
Οπτικοποίηση Δεδομένων

- Κατά τη διερεύνηση δεδομένων, η **οπτικοποίηση δεδομένων** μπορεί να βοηθήσει πάρα πολύ.
- Σε αυτή την ενότητα θα περιγράψουμε τρεις σημαντικές τεχνικές οπτικοποίησης που μπορούν να χρησιμοποιηθούν για να απεικονίσουμε τις τιμές ενός μόνο χαρακτηριστικού:
  - το **ραβδόγραμμα** (bar plot)
  - το **ιστόγραμμα** (histogram)
  - το **διάγραμμα κουτιού** (box plot)

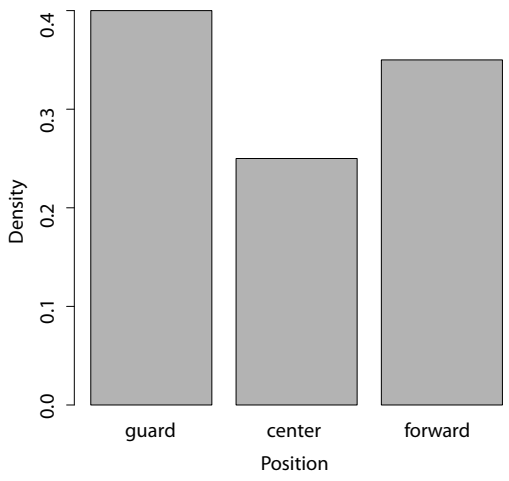
Πίνακας: Ένα σύνολο δεδομένων που δείχνει τις θέσεις και τα εβδομαδιαία έξοδα προπόνησης μιας σχολικής ομάδας μπάσκετ.

ID	Position	Training Expenses	ID	Position	Training Expenses
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41

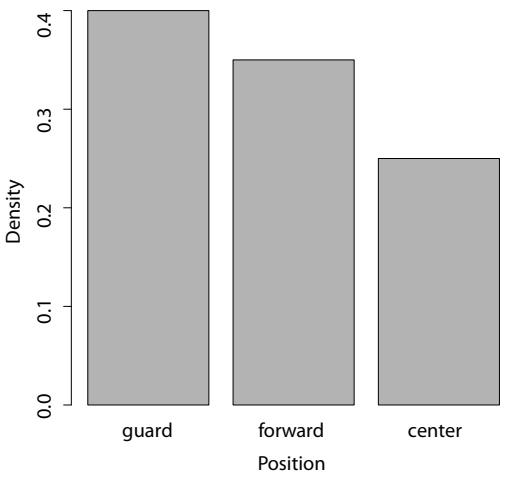
Τα ραβδογράμματα είναι εξαιρετικά για κατηγορικά χαρακτηριστικά



(α□) Συχνότητα

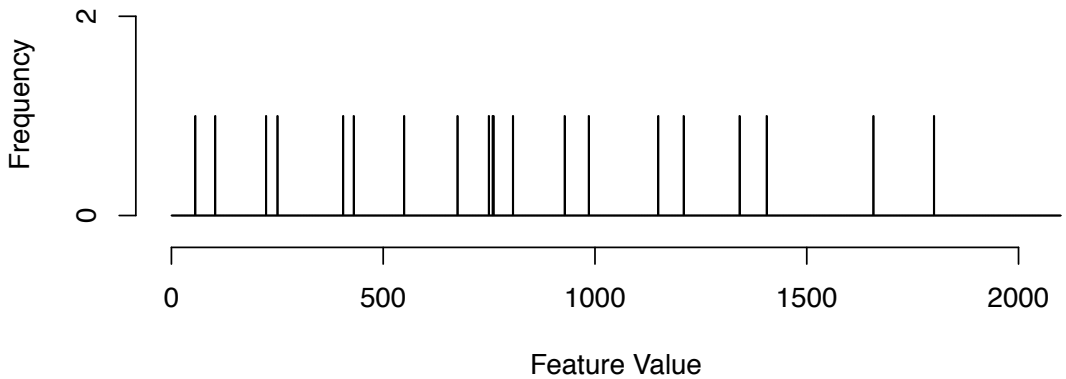


(β□) Αναλογία



(γ□) Ταξινομημένο

Τα ραβδογράμματα δεν λειτουργούν για συνεχή χαρακτηριστικά



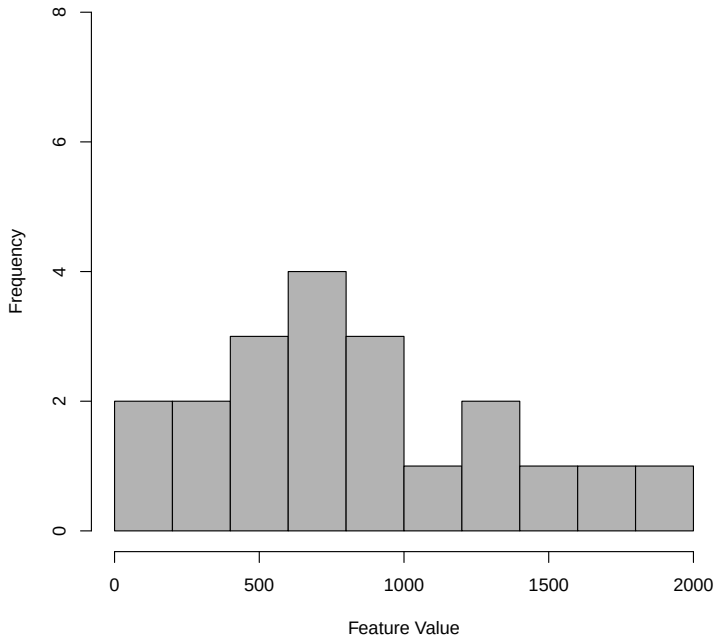
Διαιρώντας το εύρος μιας μεταβλητής σε διαστήματα, ή κλάσεις (bins), μπορούμε να δημιουργήσουμε **ιστογράμματα**

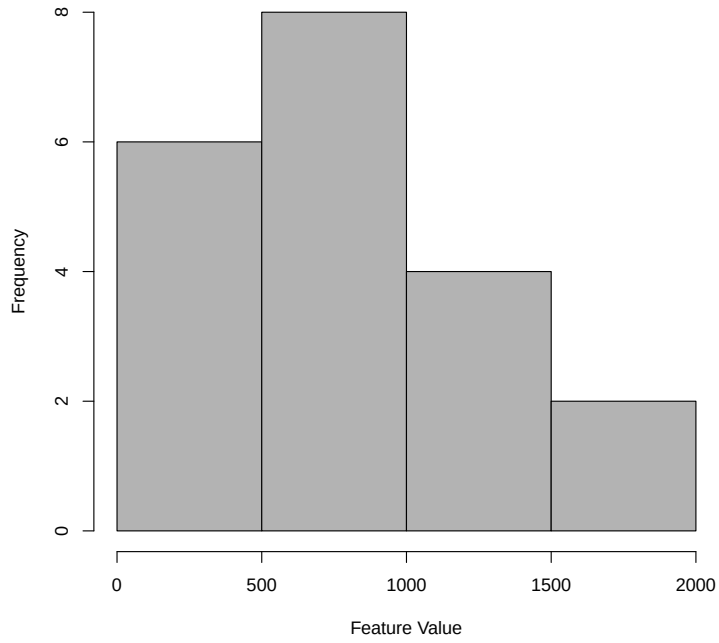
(α□) Διαστήματα 200 μονάδων

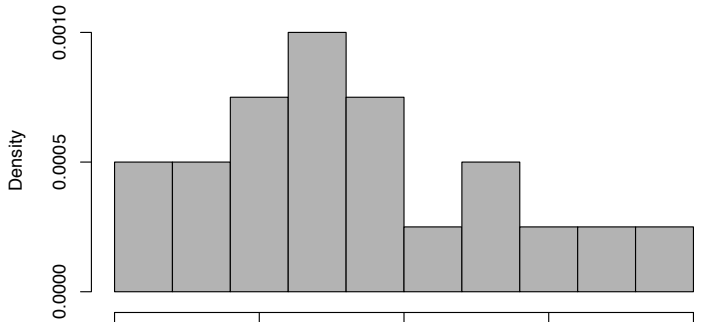
Διάστημα	**Πλήθος**	**Πυκνότητα**	**Πιθανότητα**
$[0, 200)$	2	0.0005	0.1
$[200, 400)$	2	0.0005	0.1
$[400, 600)$	3	0.00075	0.15
$[600, 800)$	4	0.001	0.2
$[800, 1000)$	3	0.00075	0.15
$[1000, 1200)$	1	0.00025	0.05
$[1200, 1400)$	2	0.0005	0.1
$[1400, 1600)$	1	0.00025	0.05
$[1600, 1800)$	1	0.00025	0.05
$[1800, 2000)$	1	0.00025	0.02

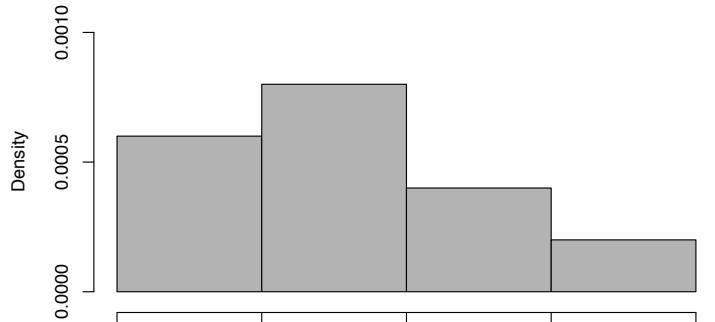
(β□) Διαστήματα 500 μονάδων

Διάστημα	**Πλήθος**	**Πυκνότητα**	**Πιθανότητα**
$[0, 500)$	6	0.0006	0.3
$[500, 1000)$	8	0.0008	0.4
$[1000, 1500)$	4	0.0004	0.2
$[1500, 2000)$	2	0.0002	0.1

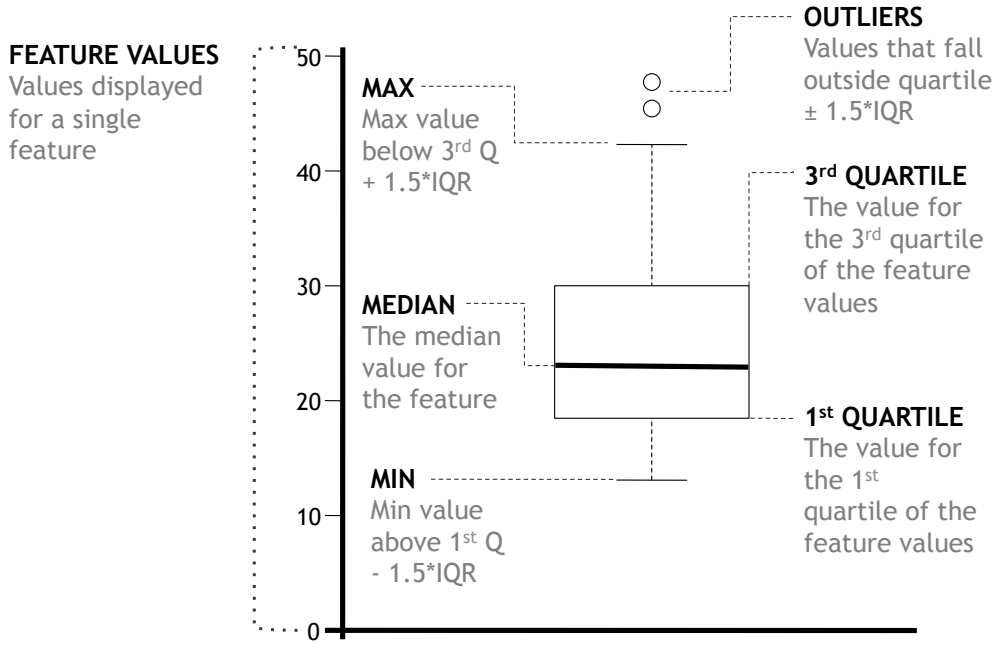

Frequency
0
2
4
6
8
0
500
1000
1500
2000
Feature Value


Frequency
0
2
4
6
8
0
500
1000
1500
2000
Feature Value

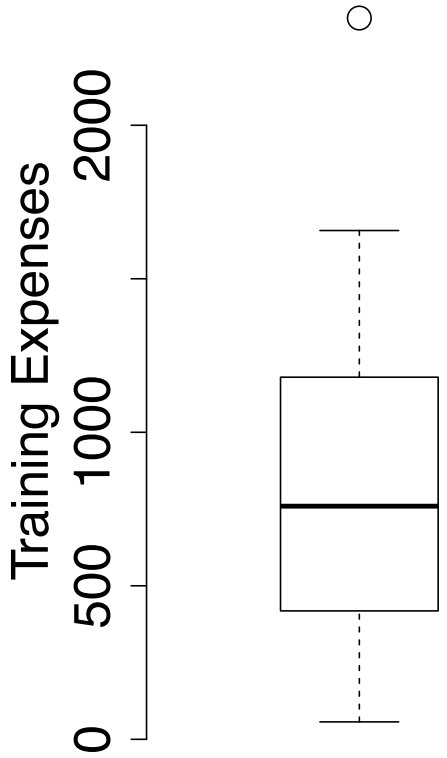

Density
0.0000
0.0005
0.0010


Density
0.0000
0.0005
0.0010

Τα διαγράμματα κουτιού είναι ένας ακόμη χρήσιμος τρόπος οπτικοποίησης συνεχών μεταβλητών



Σχήμα: Η δομή ενός διαγράμματος κουτιού.



Σχήμα: Ένα διάγραμμα κουτιού για το χαρακτηριστικό Training Expenses από το σύνολο δεδομένων της ομάδας μπάσκετ στον Πίνακα 4 [34].