



Ευαγγελία-Γεωργία Κωστάκη

Βιοστατιστικός-Επιδημιολόγος, PhD
Μέλος ΕΔΙΠ
Εργαστήριο Προληπτικής Ιατρικής,
Επιδημιολογίας και Ιατρικής
Στατιστικής
Ιατρική Σχολή, Εθνικό και
Καποδιστριακό Πανεπιστήμιο
Αθηνών

ekostakh@med.uoa.gr

Περιγραφική Στατιστική (2): Μέτρα θέσης & διασποράς

Περιγραφική στατιστική

- Η περιγραφική στατιστική ανάλυση (descriptive statistics) ασχολείται με τη συνοπτική και περιγραφική παρουσίαση των δεδομένων μίας στατιστικής έρευνας.

Μεταβλητές

Ποιοτικές

Διατάξιμες:

Θέση σε μια λίστα.
Οι κατηγορίες
ιεραρχούνται π.χ.
Κατηγορίες ΔΜΣ,
βαρύτητα
ασθένειας
(ήπια/μεσαία/
σοβαρή)

Κατηγορικές:

Αριθμοί που
προσδιορίζουν
κατηγορίες ή
τύπους
πραγμάτων
(χωρίς
ταξινόμηση)
Π.χ. Φύλο
(άντρας/γυναίκα),
θάνατος (ναι/όχι)

Ποσοτικές

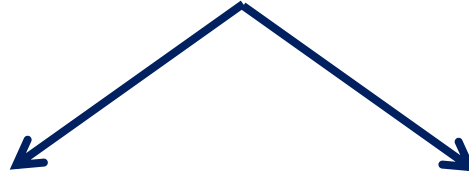
Συνεχείς:

Μπορεί να λάβει
οποιοσδήποτε
τιμές
Π.χ. Βάρος (kg)

Διακριτές

Ακέραιες τιμές,
συνήθως αριθμός
μετρήσεων
πχ. Αριθμός
καταγμάτων ή
θανάτων

Περιγραφή ποιοτικών δεδομένων



Αριθμητικά

Απόλυτες συχνότητες,
σχετικές συχνότητες (%)

Γραφικά

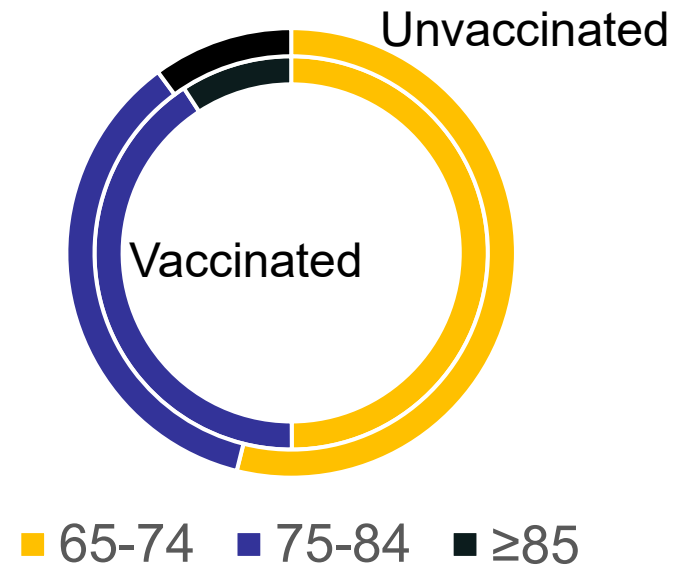
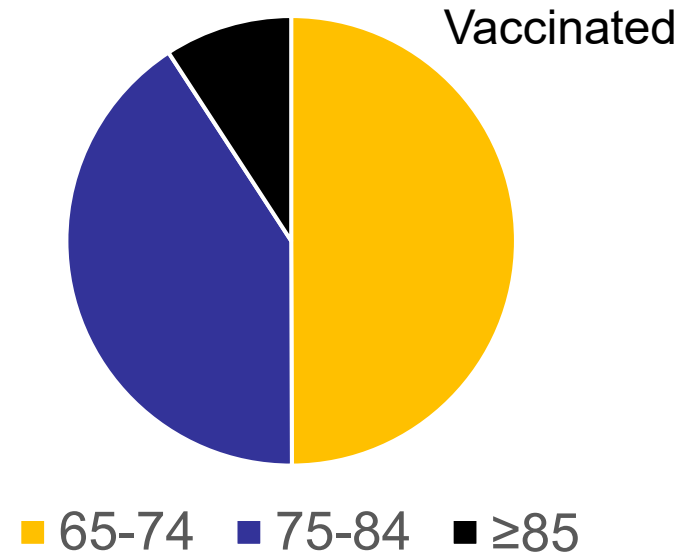
- Γράφημα πίτας
- Ραβδόγραμμα

Παράδειγμα

Characteristic	Vaccinated person-time, % (n = 205 472 person-years)	Unvaccinated person-time, % (n = 132 792 person-years)
Age group (yr)		
65–74	50.0	53.9
75–84	40.9	36.0
≥85	9.2	10.1
Male	42.7	41.9

Conditions defined by diagnosis codes assigned during the baseline period

Hypertension	26.9	23.9
Lung disease	25.4	21.5
Heart disease	23.8	20.2
Diabetes	13.5	11.8
Cancer	11.6	9.7
Lipid disorders	8.7	7.1
Atrial fibrillation	5.6	4.6
Dementia	3.2	3.9
Renal disease	2.3	2.1
Vasculitis or rheumatologic disease	2.2	1.8



Περιγραφή ποσοτικών δεδομένων



Αριθμητικά

Μέτρα θέσης

- Μέση τιμή
- Διάμεση τιμή
- Επικρατούσα τιμή

Μέτρα διασποράς

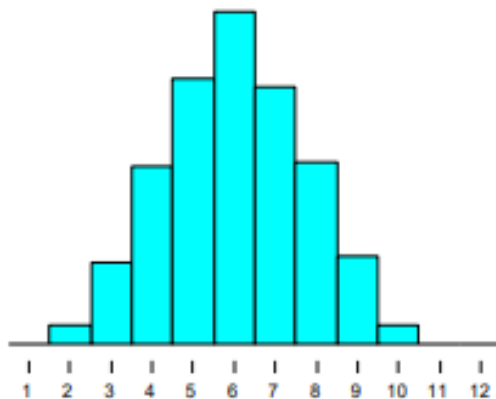
- Εύρος
- Διακύμανση
 - Τυπική απόκλιση
- Εκατοστημόρια
 - Ενδοτεταρτημοριακό εύρος

Γραφικά

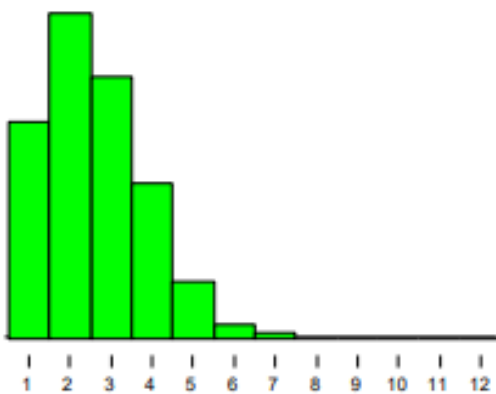
- Ιστόγραμμα
- Θηκόγραμμα

Κατανομή ποσοτικών δεδομένων I

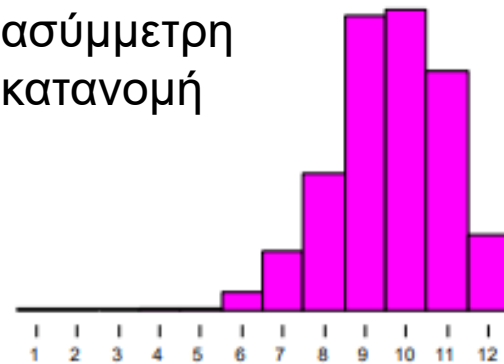
Κανονική
κατανομή



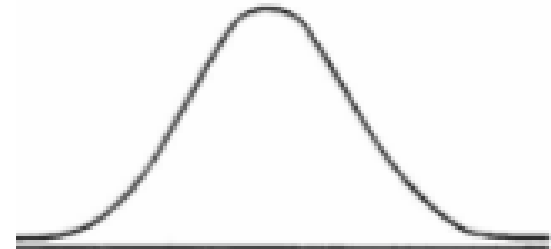
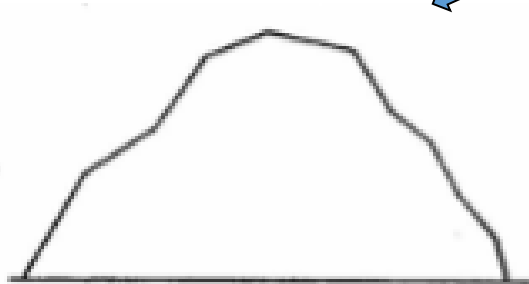
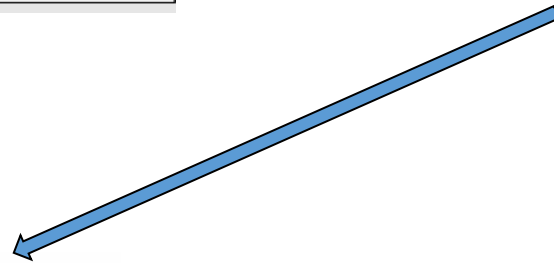
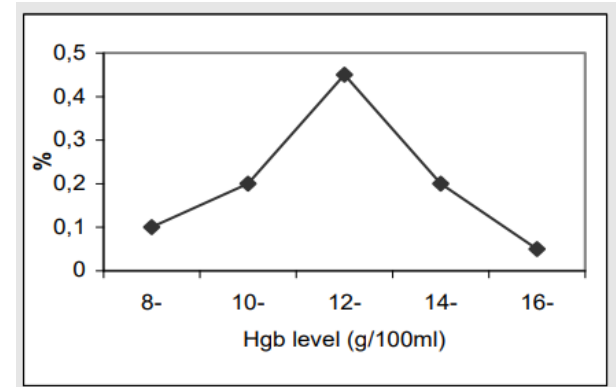
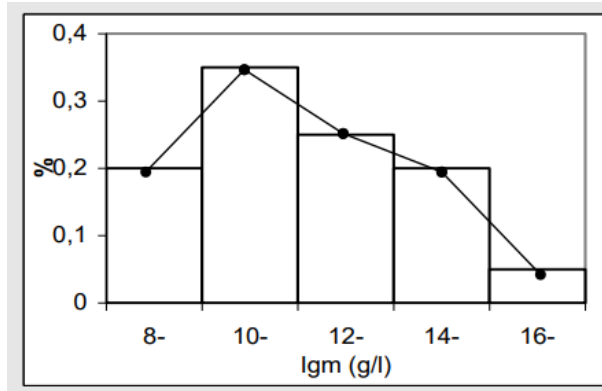
Θετικά
ασύμμετρη
κατανομή



Αρνητικά
ασύμμετρη
κατανομή



Κατανομή ποσοτικών δεδομένων II



Καμπύλη
συχνοτήτων

Μέτρα θέσης: Μέση τιμή (mean)

- Το **αλγεβρικό άθροισμα** όλων των μετρήσεων διαιρεμένο με το πλήθος αυτών:

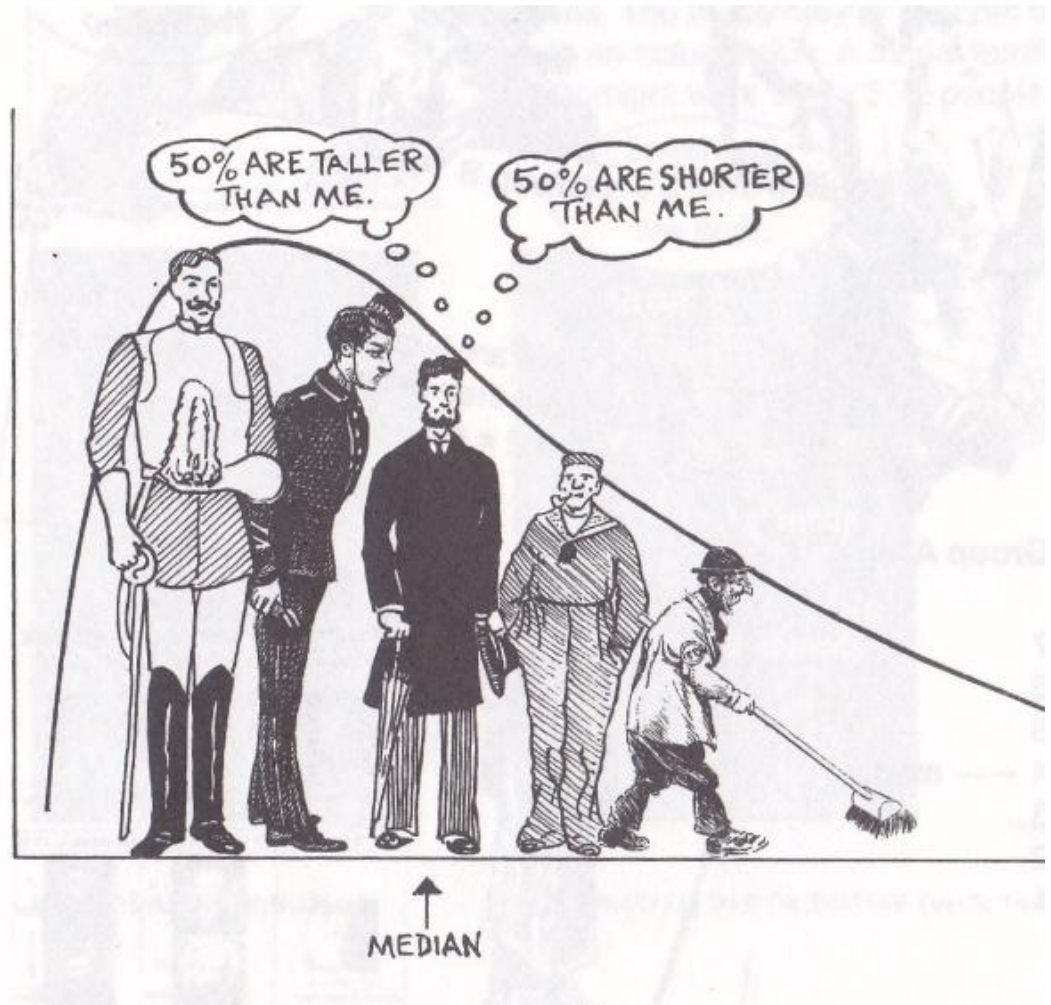
$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

- Παράδειγμα: Έστω οι $n=5$ παρατηρήσεις $\{2,3,9,5,2\}$:

$$\bar{X} = \frac{2 + 3 + 9 + 5 + 2}{5} = 4,2$$

Μέτρα θέσης: Διάμεση τιμή (median) I

- Συγχρόνως μεγαλύτερη από τις μισές παρατηρήσεις (50%) και μικρότερη από τις άλλες μισές (50%).



Μέτρα θέσης: Διάμεση τιμή (median) II

- Παίρνει την τιμή της παρατήρησης εκείνης που καθορίζεται από τον τύπο $(n+1)/2$, μετά από **ταξινόμηση** των X_i κατά αύξουσα σειρά.
- Υπολογίζεται δυσκολότερα όταν έχουμε ισοψηφίες ή πολλές παρατηρήσεις.
- Μεγαλύτερη χρήση στη περιγραφική στατιστική.

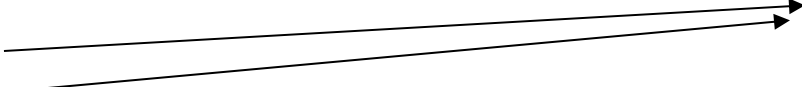
Πρώτα υπολογίζεται η «θέση» που καταλαμβάνει η διάμεσος στην αύξουσα σειρά

- Η **θέση** υπολογίζεται με τον τύπο: $\frac{n + 1}{2}$
- Άρα αν έχουμε 11 παρατηρήσεις:
 - $(11+1)/2 = 6$ (η έκτη παρατήρηση)
- Αν έχουμε 16 παρατηρήσεις:
 - $(16+1)/2 = 8,5$ (παίρνουμε το μέσο όρο της 8^{ης} και της 9^{ης} παρατήρησης)

Σειρά Διάρκεια νόσου (ημέρες)

1	12
2	14
3	14
4	15
5	16
6	17
7	17
8	18
9	20
10	21
11	22
12	24
13	26
14	28
15	>61
16	>61
17	>61
18	>61

Παράδειγμα


$$\frac{20 + 21}{2} = 20.5$$

Έχουμε 18
παρατηρήσεις:
➤ $(18+1)/2 = 9,5$
(παίρνουμε το μέσο
όρο της 9^{ης} και της
10^{ης} παρατήρησης)

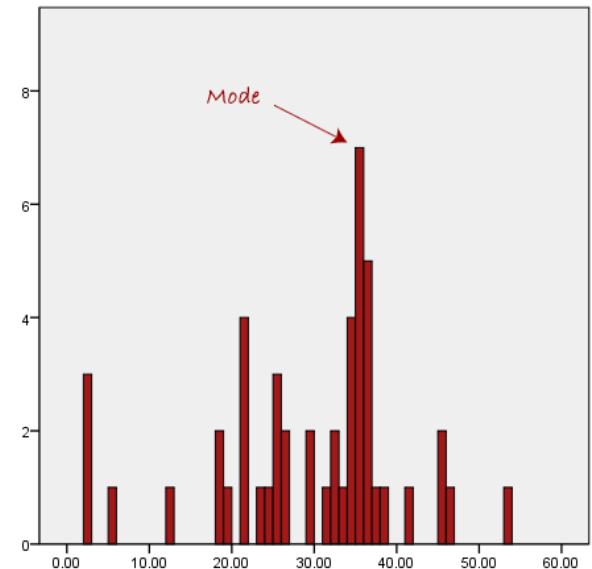
Μέτρα θέσης: Επικρατούσα τιμή (mode)

Είναι η τιμή με τη **μεγαλύτερη συχνότητα εμφάνισης** (εφαρμόζεται μόνο εάν η μεταβλητή παίρνει λίγες επαναλαμβανόμενες διακριτές τιμές).

Παράδειγμα: Έστω οι παρατηρήσεις {2,3,9,5,2}

Επικρατούσα τιμή = 2

- Πρόβλημα με λίγες παρατηρήσεις
- Μόνο περιγραφική χρήση

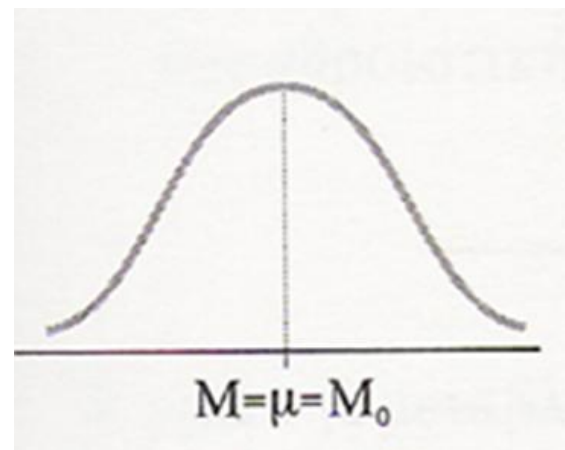
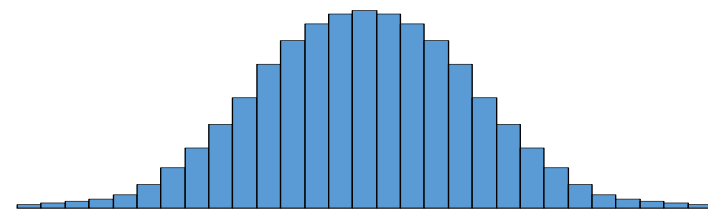


Σχέσεις - Ιδιότητες

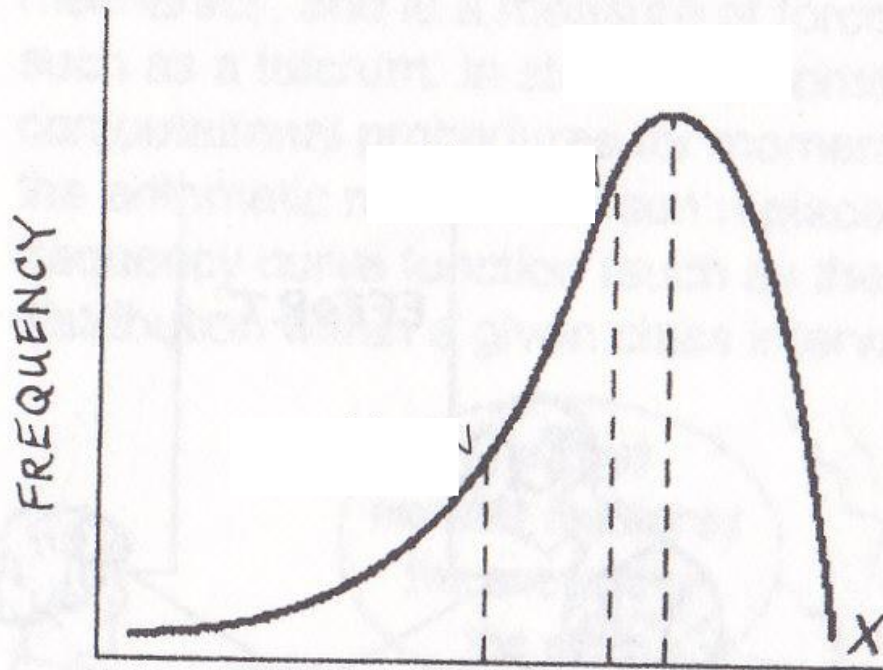
✓ Κανονική κατανομή:

Μέση = Διάμεση = Επικρατούσα

- Κεντρική τιμή για κανονικές κατανομές.
- Οι περισσότερες παρατηρήσεις X_i τείνουν να συγκεντρώνονται γύρω από αυτή στις κανονικές κατανομές.
- Μέση τιμή: Ευαίσθητη σε ακραίες τιμές (outliers).

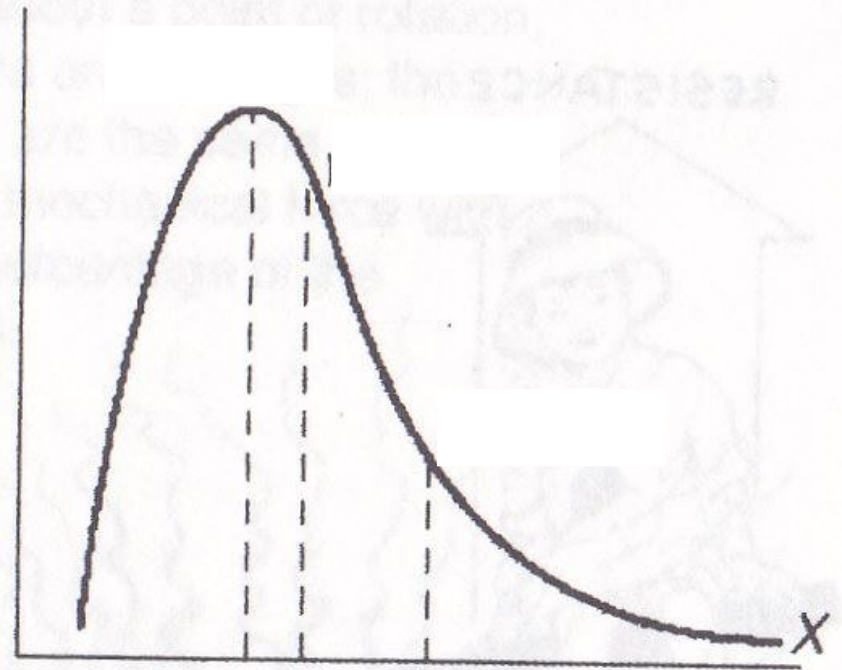


NEGATIVELY SKEWED



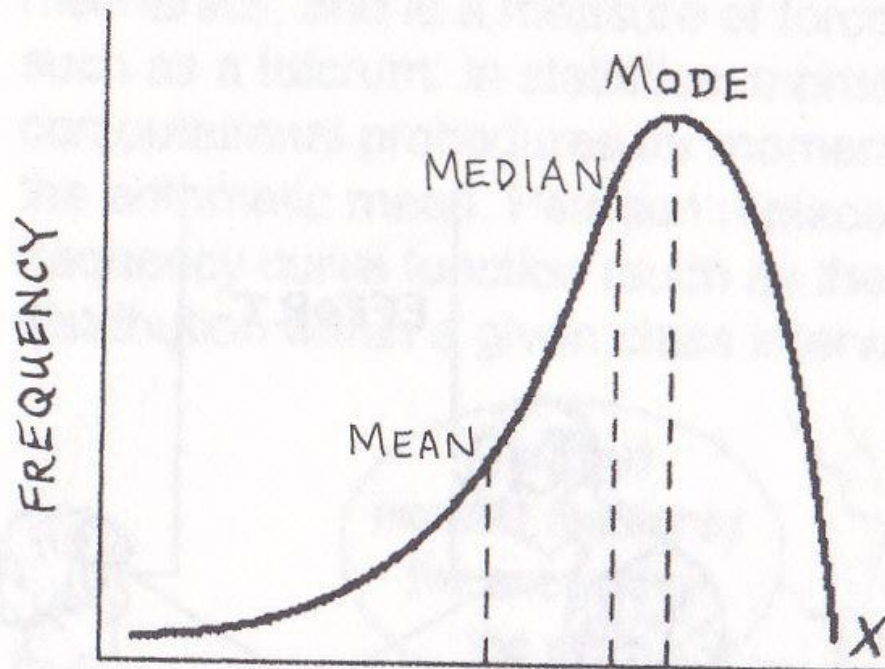
← NEGATIVE DIRECTION

POSITIVELY SKEWED

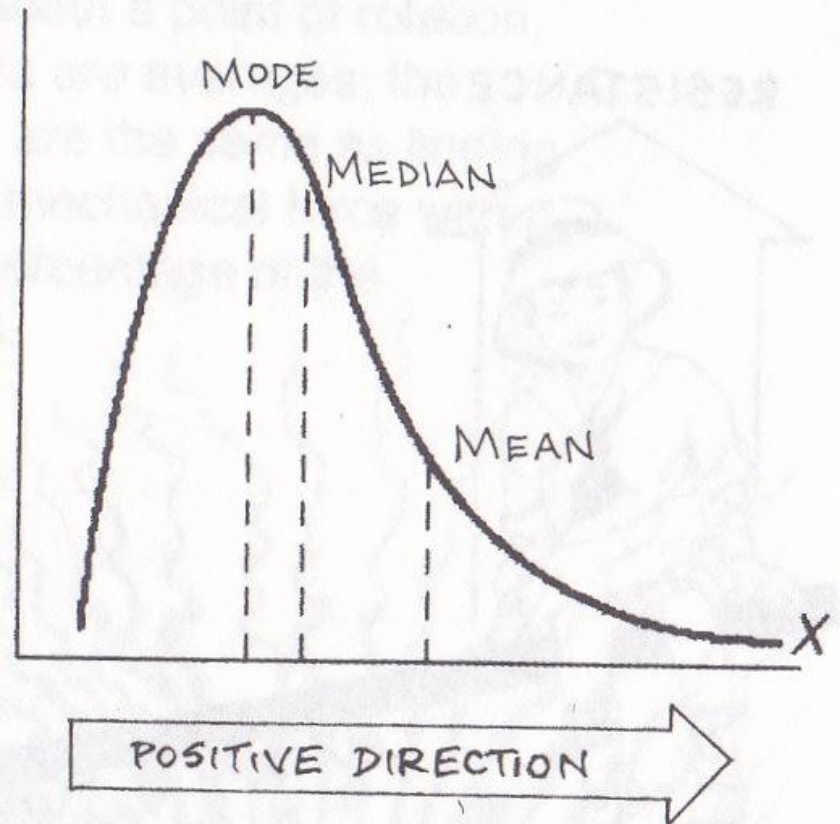


POSITIVE DIRECTION →

NEGATIVELY SKEWED



POSITIVELY SKEWED



Προσοχή!

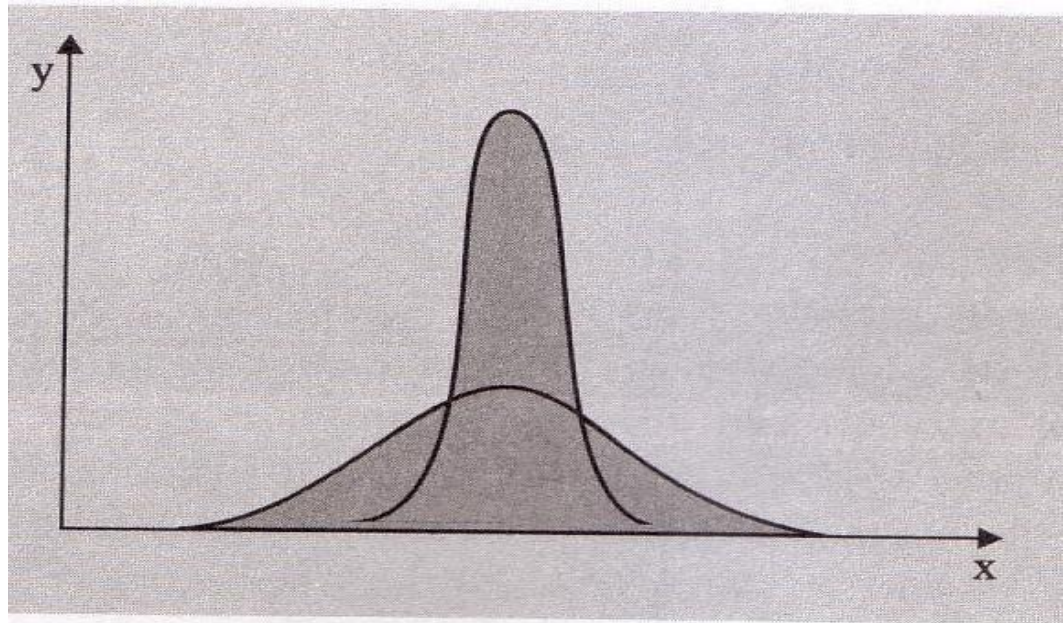
- Οι αντιπροσωπευτικές τιμές θέσης εκφράζονται σε τιμές του μεγέθους που μελετάμε.
- Σε **κανονικές κατανομές**, η καλύτερη αντιπροσωπευτική τιμή θέσης είναι η **μέση τιμή**.
- Σε **ασύμμετρες** κατανομές, είναι η **διάμεση τιμή**, επειδή η μέση τιμή επηρεάζεται πολύ από τις ακραίες τιμές.

Τι είναι η διασπορά μιας κατανομής?

□ Είναι μια έκφραση της ποικιλίας των παρατηρήσεων.

Πχ. Οι παρατηρήσεις αναστήματος (εκατοστά):

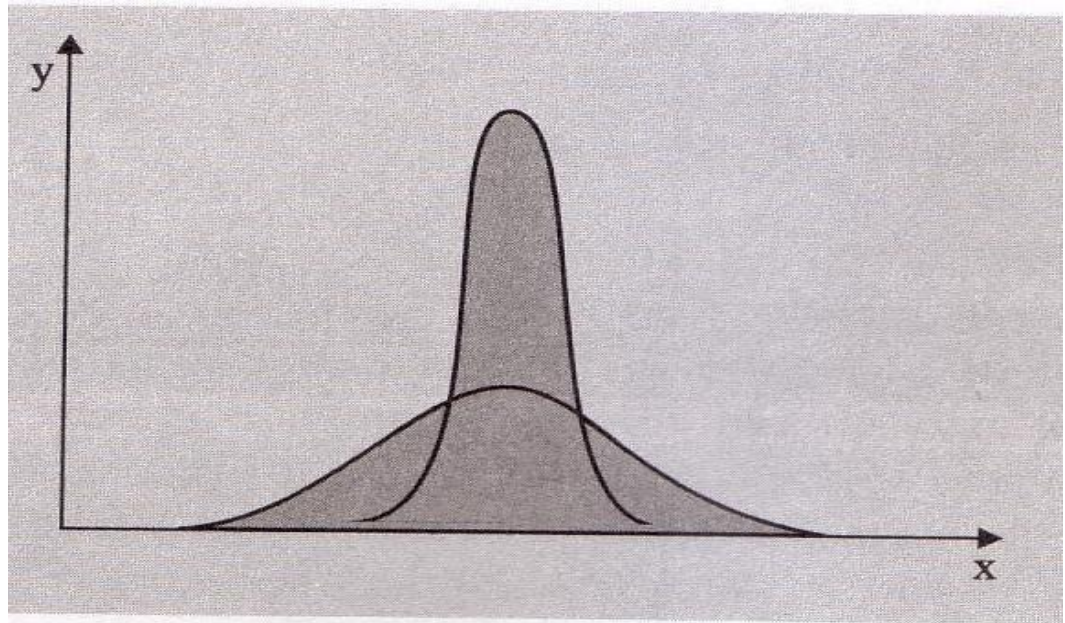
- $n=5$: 151, 153, 157, 156, 154 (μικρότερη διασπορά)
- $n=5$: 151, 175, 167, 189, 188



Μέτρα Διασποράς: Εύρος (range)

- Εύρος: Μεγαλύτερη τιμή – Μικρότερη τιμή στο δείγμα.
- Πχ. Στο παράδειγμα με τις τιμές αναστημάτων, η πρώτη κατανομή είχε εύρος: $157 - 151 = 6$ εκατοστά και η δεύτερη: $189 - 151 = 38$ εκατοστά.

Μειονέκτημα: Δεν δίνει πληροφορία για τις άλλες παρατηρήσεις.



Μέτρα Διασποράς: Διακύμανση (variance)

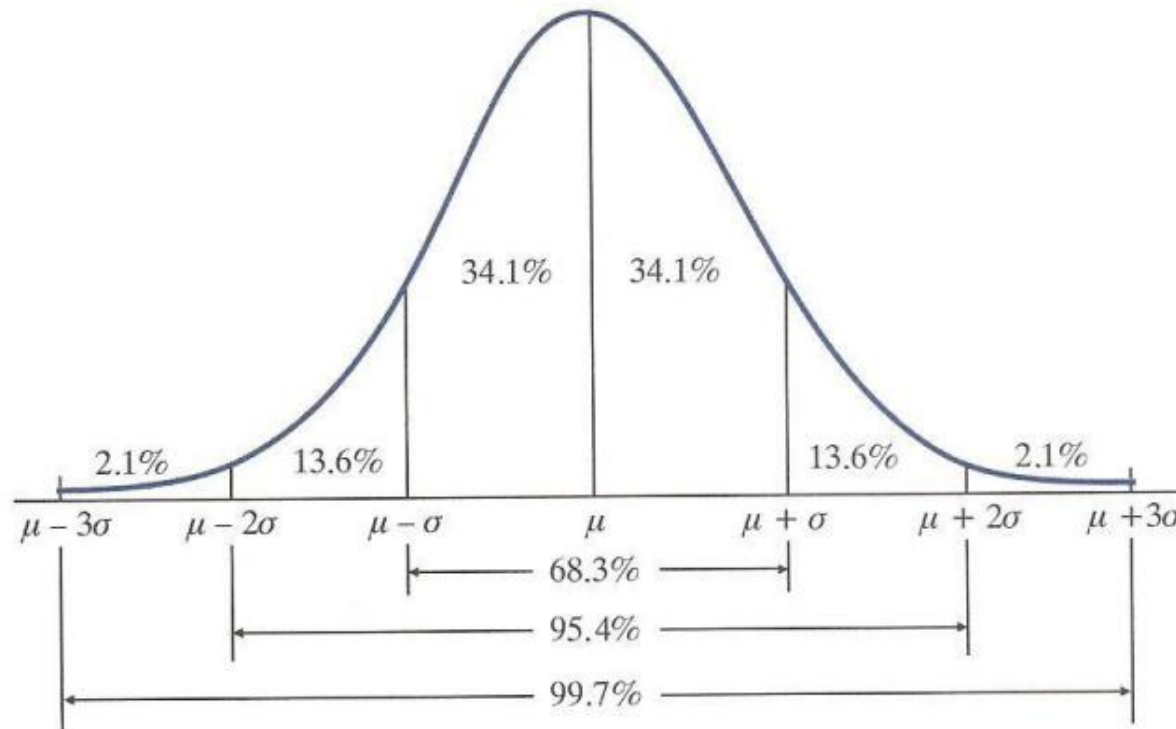
✓ Διακύμανση:
$$V = \frac{\sum_i (X_i - \bar{X})^2}{n}$$

✓ Τυπική / σταθερή απόκλιση (standard deviation):

$$SD = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n}} = \sqrt{\frac{\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n}}{n}}$$

- Η τυπική απόκλιση είναι μέτρο του **βαθμού διασποράς**, όχι πάντα του τρόπου.
- Ειδικά στις **κανονικές κατανομές** μας δείχνει και τον **τρόπο της διασποράς**.

Ιδιότητες κανονικής κατανομής (πώς η τυπική απόκλιση δείχνει τον τρόπο διασποράς των παρατηρήσεων;)



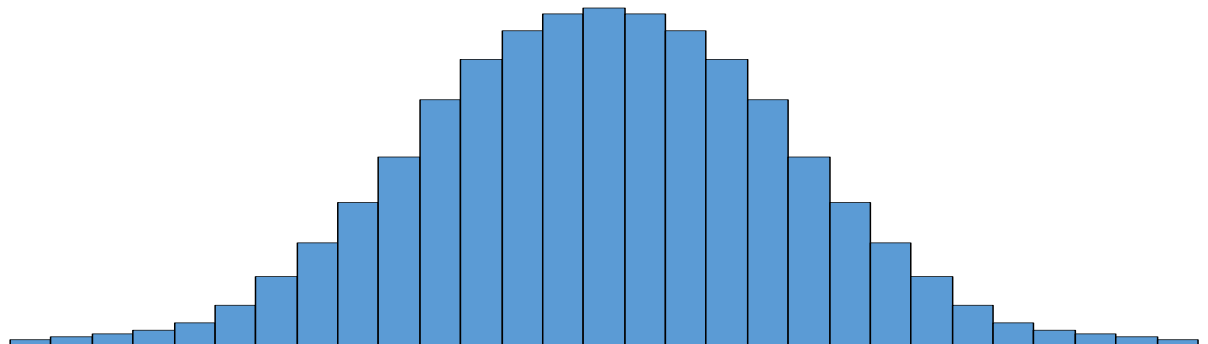
- Στο διάστημα από $(\mu - \sigma)$ μέχρι $(\mu + \sigma)$, όπου μ =μέση τιμή και σ =τυπική απόκλιση, περιλαμβάνεται το **68,3%** των παρατηρήσεων.
- Στο διάστημα από $(\mu - 2\sigma)$ μέχρι $(\mu + 2\sigma)$ περιλαμβάνεται το **95,4%** των παρατηρήσεων.
- Στο διάστημα από $(\mu - 3\sigma)$ μέχρι $(\mu + 3\sigma)$ περιλαμβάνεται το **99,7%** των παρατηρήσεων.

Παράδειγμα I

Σε μια έρευνα σε δείγμα 1.500 ενήλικων ατόμων καταγράφηκε, μεταξύ άλλων, η ηλικία (σε έτη) των συμμετεχόντων.

Η μέση ηλικία των συμμετεχόντων ήταν **35 έτη** και η τυπική απόκλιση **4 έτη**.

Υποθέτουμε πως η ηλικία ακολουθεί προσεγγιστικά κανονική κατανομή στο δείγμα μας.



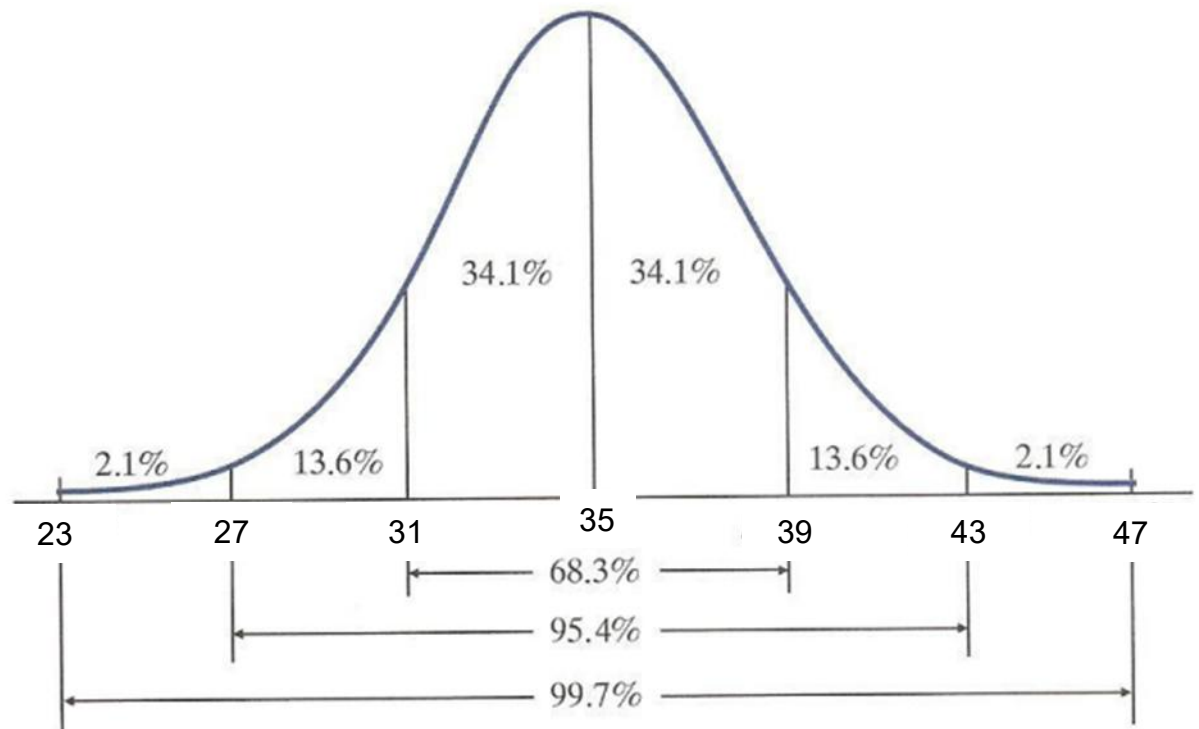
Παράδειγμα II

Υποθέτοντας πως η ηλικία ακολουθεί προσεγγιστικά κανονική κατανομή στο δείγμα μας, τότε:

Το 68,3% έχει ηλικία από 31-39 έτη (από $(\mu - \sigma)$ μέχρι $(\mu + \sigma)$).

Το 95,4% έχει ηλικία από 27-43 έτη (από $(\mu - 2\sigma)$ μέχρι $(\mu + 2\sigma)$).

Το 99,7% έχει ηλικία από 23-47 έτη (από $(\mu - 3\sigma)$ μέχρι $(\mu + 3\sigma)$).



Μέτρα Διασποράς: Εκατοστημόρια (percentiles)

- Το K εκατοστημόριο είναι εκείνη η τιμή της κατανομής με βάση την οποία ποσοστό ίσο με το **$K\%$ των παρατηρήσεων** βρίσκεται **κάτω από την τιμή αυτή**.
- Πχ το 5° εκατοστημόριο είναι η τιμή από την οποία 5% των παρατηρήσεων είναι μικρότερες (και 95% μεγαλύτερες).
- Πχ Το 75° εκατοστημόριο είναι η τιμή από την οποία 75% των παρατηρήσεων είναι μικρότερες (και 25% μεγαλύτερες).
- Το 50° εκατοστημόριο;

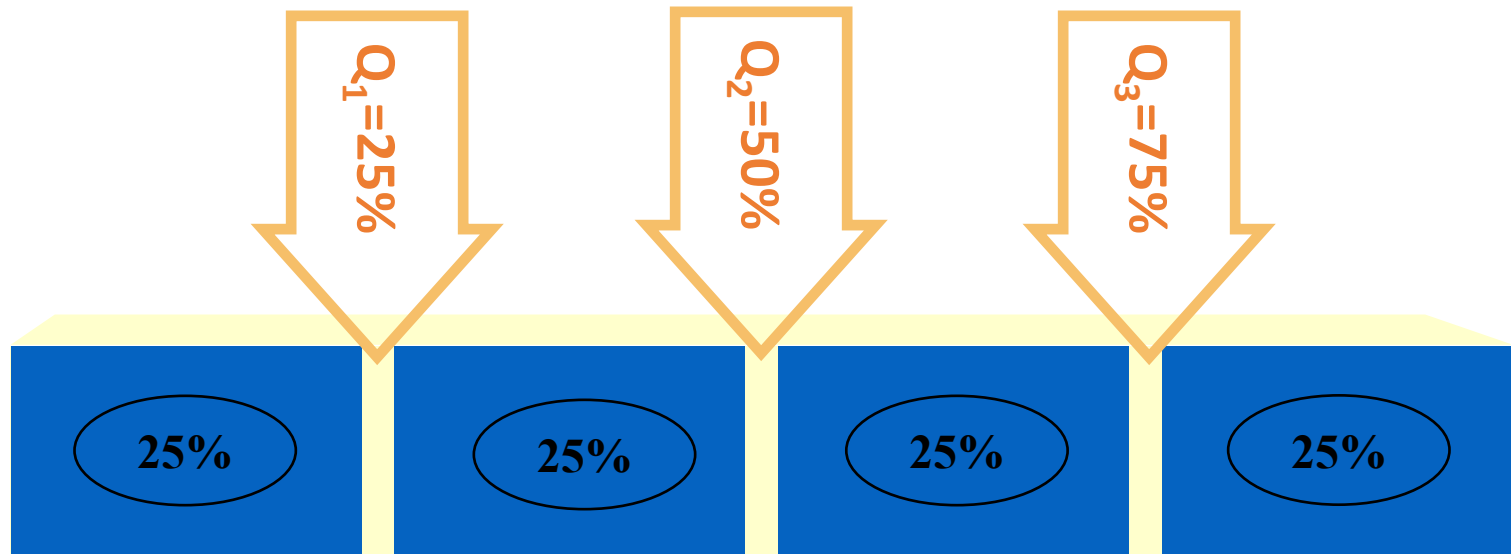
Πώς υπολογίζονται τα εκατοστημόρια;

- Για να υπολογιστούν πρέπει πρώτα να **διαταχθούν** οι παρατηρήσεις **κατά αύξουσα** (ή φθίνουσα) σειρά.
- Μετά υπολογίζεται η **θέση** του εκατοστημορίου με τον παρακάτω τύπο, όπου k το αντίστοιχο ποσοστό:

$$\frac{(n + 1) * k}{100}$$

- Τέλος, βρίσκουμε την **τιμή** που αντιστοιχεί σε αυτήν τη θέση.

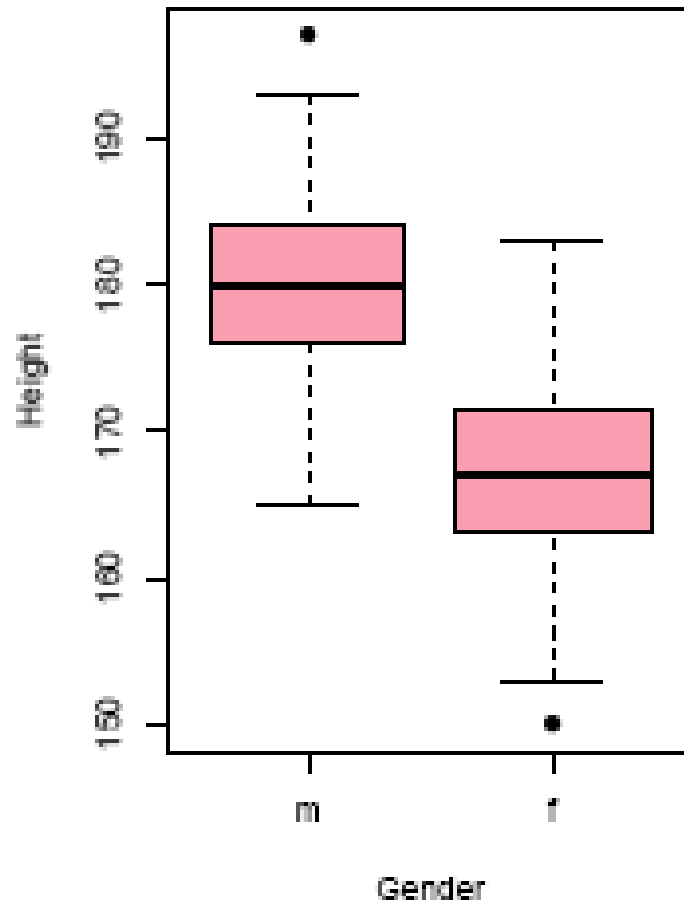
Τεταρτημόρια (Quartiles)



Ενδοτεταρτημοριακό εύρος:

$$IQR = Q_3 - Q_1$$

Θηκόγραμμα (boxplot)



← $\leq Q3 + 1,5(Q3 - Q1)$

← 75° εκατοστημόριο (3° τεταρτημόριο, Q3)

← Διάμεση τιμή

← 25° εκατοστημόριο (1° τεταρτημόριο, Q1)

← $\geq Q1 - 1,5(Q3 - Q1)$

Συνοπτικά: Παρουσίαση ποσοτικών δεδομένων

- Κατά προσέγγιση κανονικές κατανομές:

- Μέση τιμή

- Τυπική απόκλιση

- Ασύμμετρες κατανομές:

- Διάμεση τιμή

- Ενδοτεταρτημοριακό εύρος (25°, 75°)

Εφαρμογή – Ερευνητική εργασία

Σκοπός: Η διερεύνηση παραγόντων που πιθανόν σχετίζονται με χαμηλό βάρος γέννησης (≤ 2.500 gr).

Τα δεδομένα προέρχονται από το νοσοκομείο Baystate Medical Center, Springfield, Massachusetts, και αφορούν στοιχεία από 189 γυναίκες.

Περιγραφή μεταβλητών

- **Id**: Patient's ID
- **lbw**: Birth weight (1: $BWT \leq 2.500\text{gr}$, 0: $BWT > 2.500\text{gr}$)
- **age**: Age of mother (years)
- **lwt**: Weight of mother at last menstrual period (pounds)
- **race**: Race of mother (1: White, 2: Black, 3: Other)
- **smoke**: Smoking status of mother (1: Smoker, 0: Non-smoker)
- **histHyper** : History of hypertension (1: Yes, 0: No)

Ερμηνεία

→ Frequencies

Statistics

		Age of Mother (yrs)	Weight of Mother at Last Menstrual Period (pounds)
N	Valid	189	189
	Missing	0	0
Mean		23.24	129.81
Median		23.00	121.00
Std. Deviation		5.299	30.579
Minimum		14	80
Maximum		45	250
Percentiles	25	19.00	110.00
	50	23.00	121.00
	75	26.00	140.50

Age of mother

- Η μέση ηλικία του δείγματος είναι 23,24 έτη με τυπική απόκλιση 5,3 έτη.
- Η διάμεση τιμή είναι 23 έτη.
- Η ελάχιστη τιμή είναι τα 14 έτη και η μέγιστη τα 45 έτη.
- Το 25% των παρατηρήσεων είναι κάτω από την ηλικία των 19 ετών.
 - Το 75% ???
 - Το 50% ???
- Η μεταβλητή ακολουθεί κανονική κατανομή;