Μέτρα θέσης και διασποράς

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

Μl. Κάσδαγλη, PhD

Εργαστήριο Υγιεινής, Επιδημιολογίας

και Ιατρικής Στατιστικής, Ιατρική Σχολή Αθηνών

Μεταβλητές

Ποιοτικές

Διατάξιμες:
θέση σε μια λίστα.
Οι κατηγορίες ιεραρχούνται π.χ.
Κατηγορίες ΔΜΣ
Βαρύτητα ασθένειας (ήπια/μεσαία/ σοβαρή)

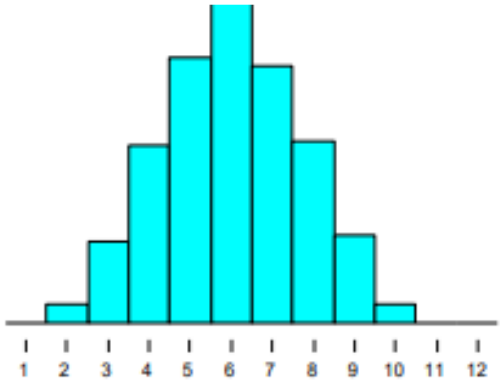
Ονομαστικές:
αριθμοί που προσδιορίζουν κατηγορίες ή τύπους πραγμάτων (χωρίς ταξινόμηση)
Π.χ. Φύλο (άντρας/γυναίκα)
Θάνατος (ναι/όχι)

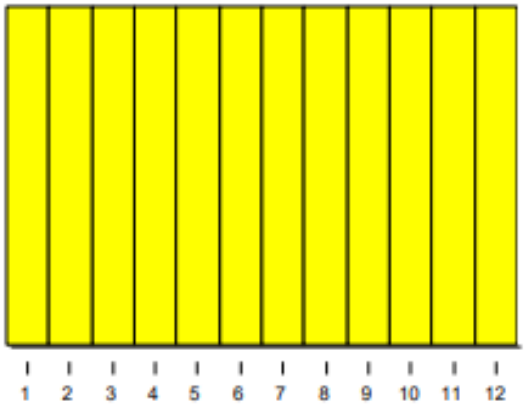
Ποσοτικές

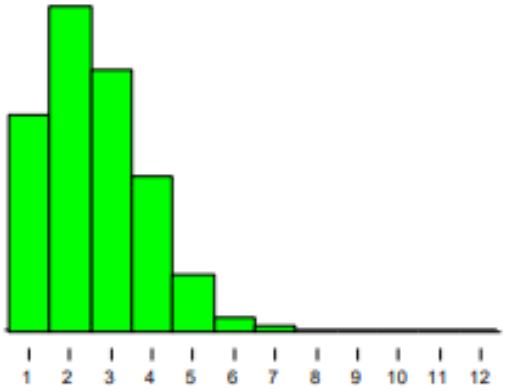
Συνεχείς:
Μπορεί να λάβει οποιεσδήποτε τιμές
Π.χ. Βάρος (kg)

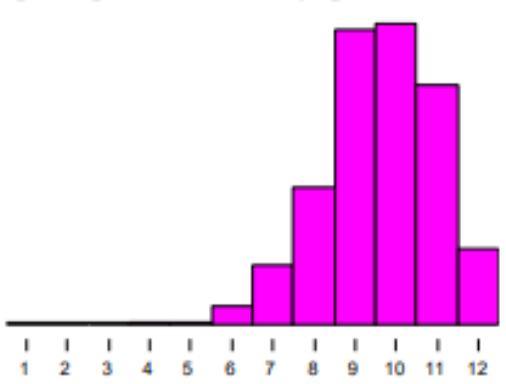
Διακριτές
Ακέραιες τιμές, συνήθως αριθμός μετρήσεων
πχ. Αριθμός καταγμάτων ή θανάτων

Κατανομή Ποσοτικών Δεδομένων

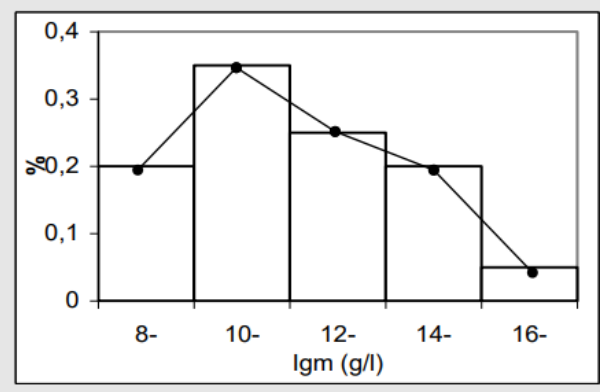


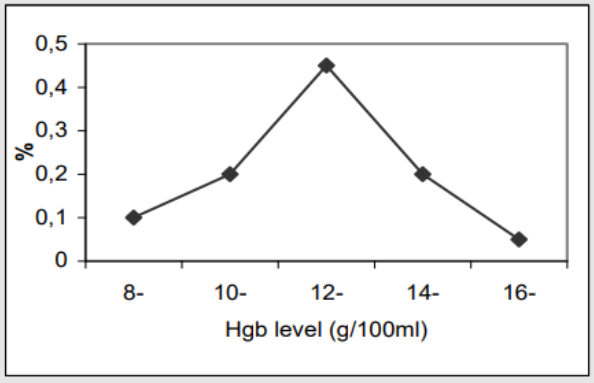


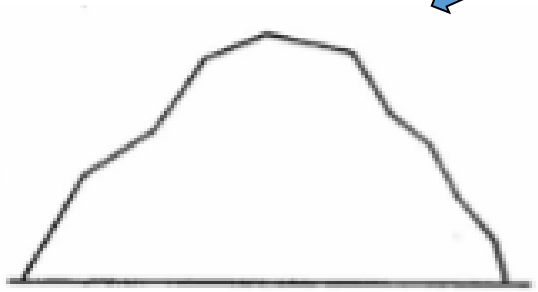


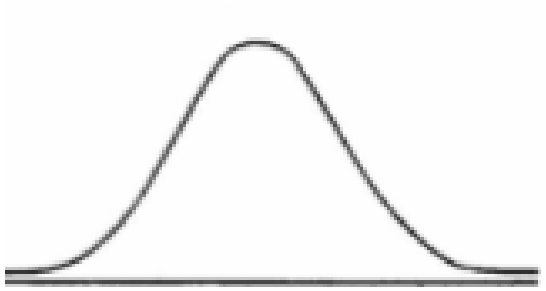


Κατανομή Ποσοτικών Δεδομένων









Αντιπροσωπευτικές τιμές

- Τιμές θέσης

  Επικρατούσα τιμή, μέση τιμή, διάμεσος

- Τιμές βαθμού διασποράς

  Τυπική απόκλιση, ακραίες τιμές, εκατοστημόρια

Μέτρα θέσης: Μέση τιμή (mean)

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

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

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

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

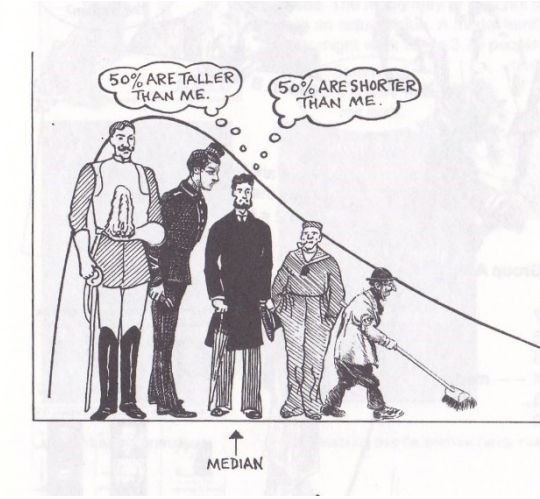
Μέτρα θέσης: Διάμεσος (median)

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

- Παίρνει την τιμή της παρατήρησης εκείνης που καθορίζεται από τον όρο **(N+1)/2**, μετά από ταξινόμηση των X_i κατ' 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
  - (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

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

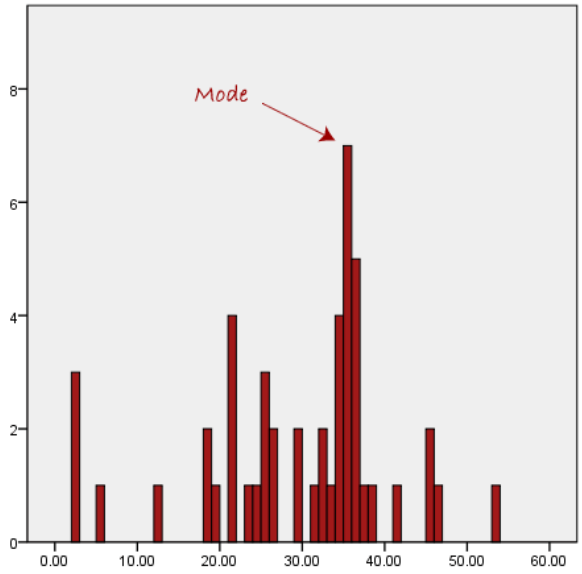
Μέτρα θέσης: Επικρατούσα τιμή (mode)

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}

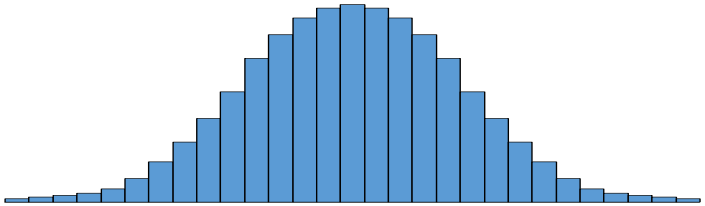
Επικρατούσα τιμή = 2

- ✓Πρόβλημα με λίγες παρατηρήσεις
- ✓Μόνο περιγραφική χρήση

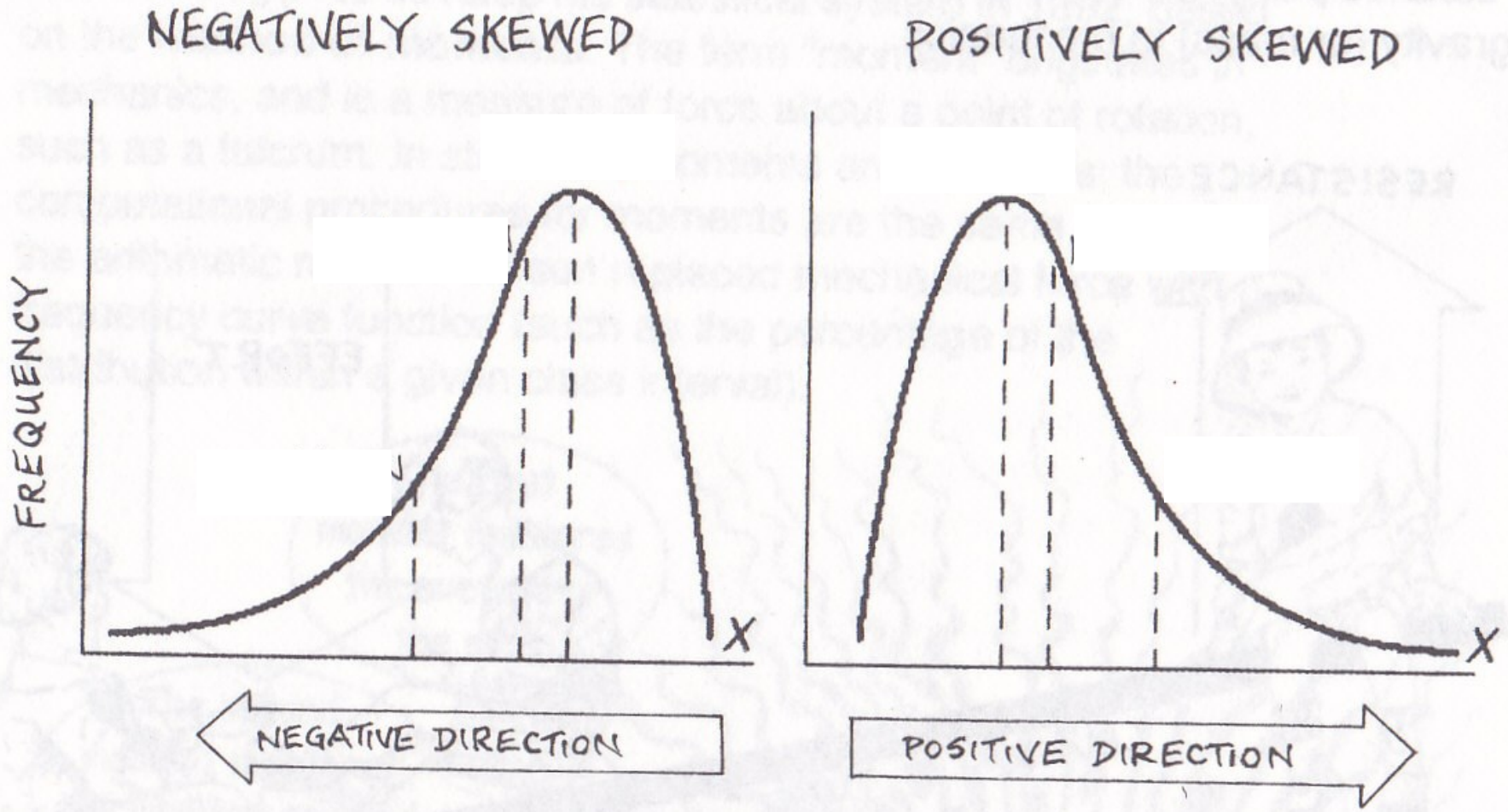


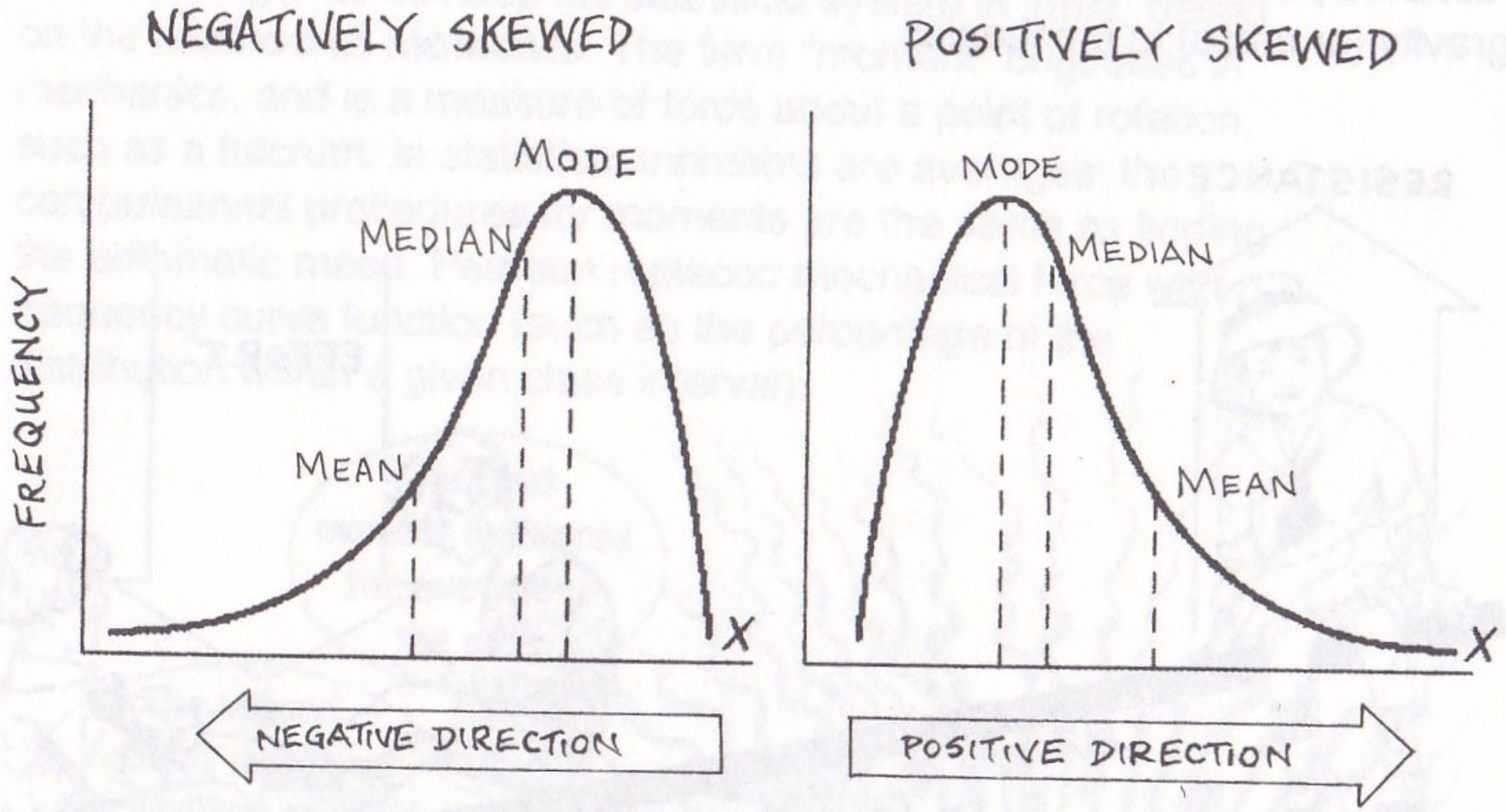
Σχέσεις -Ιδιότητες

- ✓ **Κανονική κατανομή:**
  - ✓ Μέση τιμή= Διάμεσο =Επικρατούσα



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


NEGATIVELY SKEWED
POSITIVELY SKEWED
FREQUENCY
X
X
NEGATIVE DIRECTION
POSITIVE DIRECTION

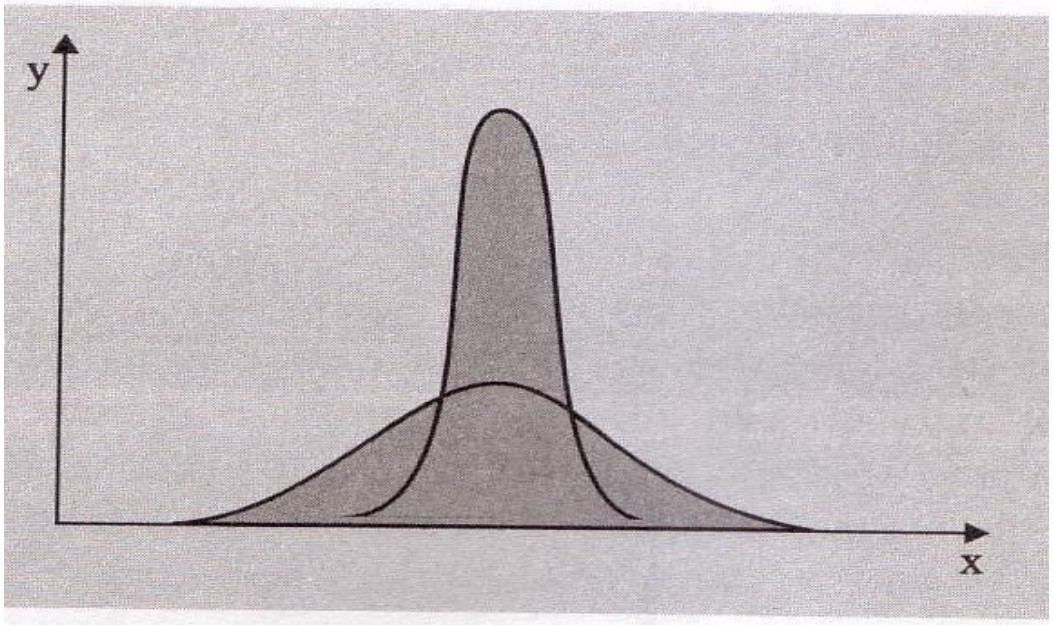

NEGATIVELY SKEWED
POSITIVELY SKEWED
FREQUENCY
MODE
MEDIAN
MEAN
X
MODE
MEDIAN
MEAN
X
NEGATIVE DIRECTION
POSITIVE DIRECTION

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, 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) 151, 153, 157, 156, 154 έχουν μικρότερη διασπορά από τις
  - 151, 175, 167, 189, 188



Μέτρα Διασποράς: Εύρος

- Εύρος: η μεγαλύτερη – η μικρότερη τιμή

- 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

Αντιπροσωπευτικές τιμές Διασποράς

✓ Διακύμανση (Variance)

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

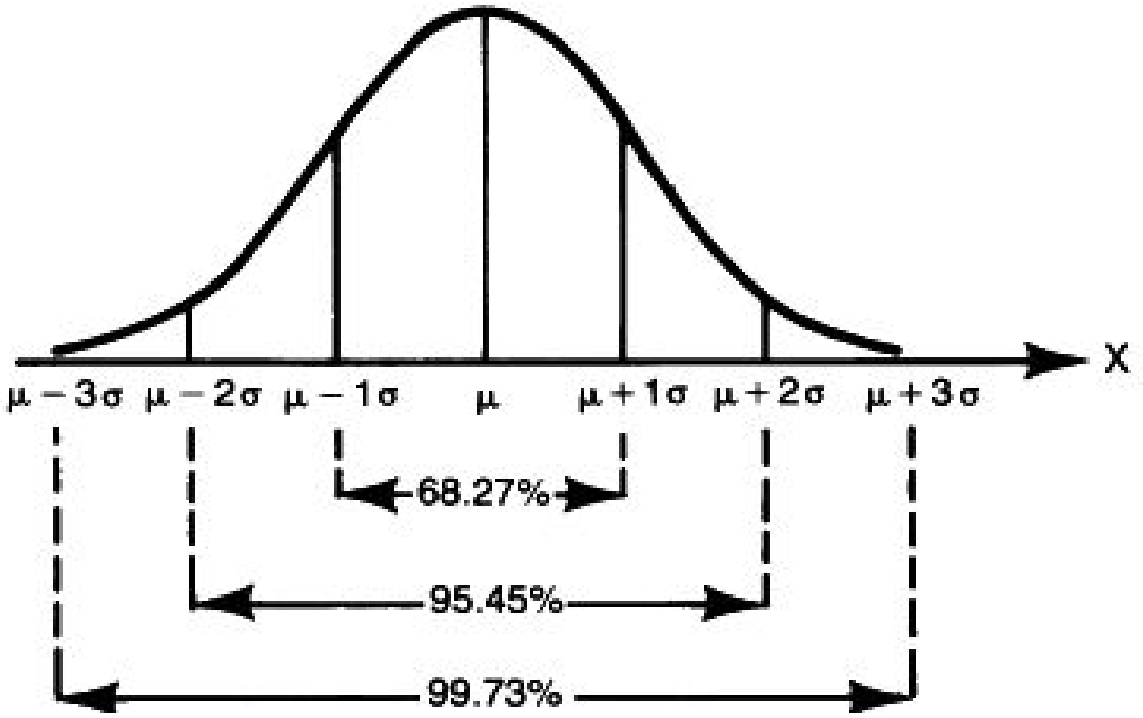
✓Τυπική / σταθερή απόκλιση (Standard Deviation)

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

- Η σταθερή απόκλιση είναι μέτρο του βαθμού διασποράς, όχι πάντα του τρόπου

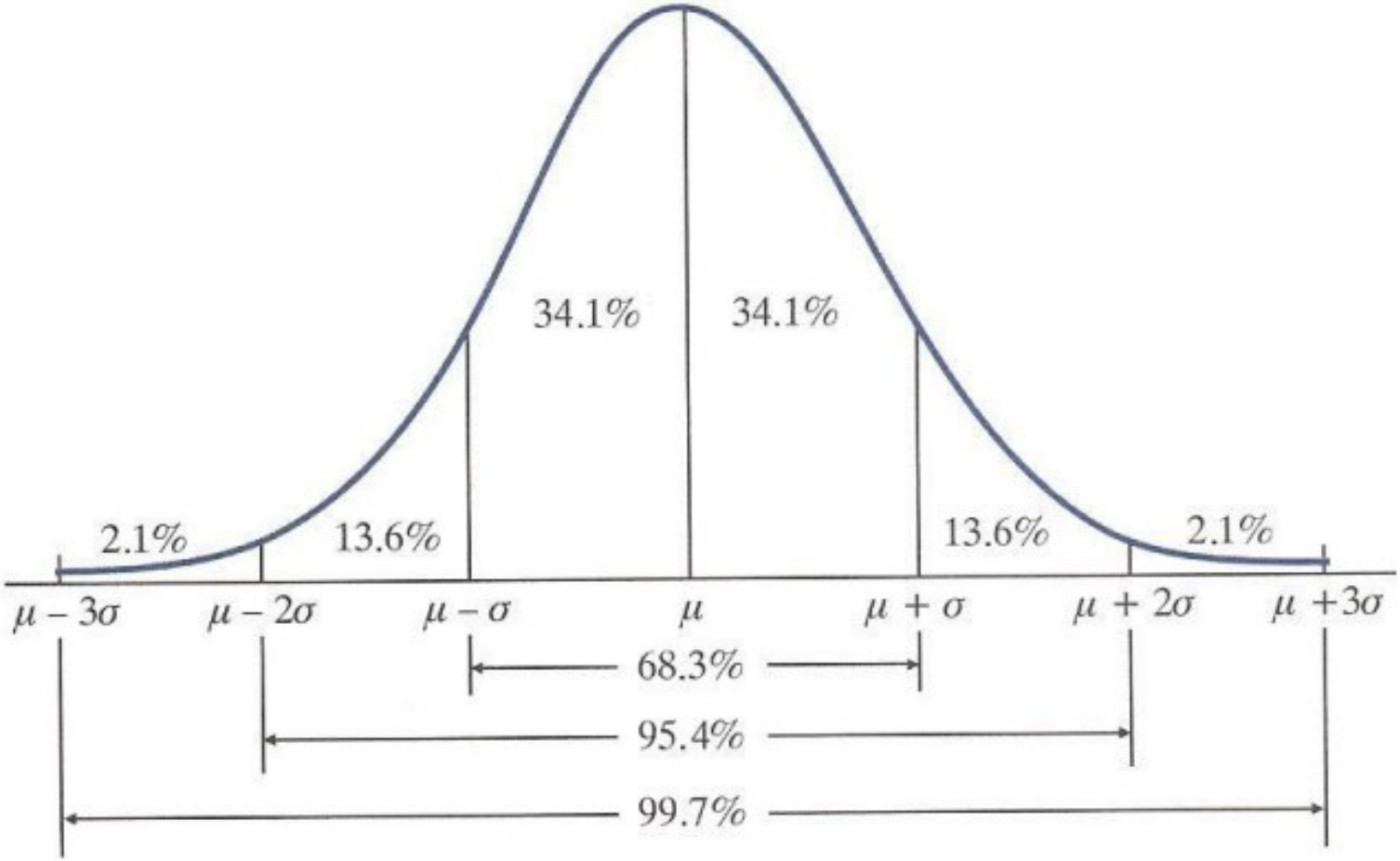
- Ειδικά στις κανονικές κατανομές μας δείχνει και τον τρόπο της διασποράς

Ιδιότητες Κανονικής Κατανομής (πώς η σταθερή 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% των παρατηρήσεων, στο διάστημα από ($\mu-2\sigma$) μέχρι ($\mu+2\sigma$) περιλαμβάνεται το 95% των παρατηρήσεων και στο διάστημα από ($\mu-3\sigma$) μέχρι ($\mu+3\sigma$) περιλαμβάνεται το 99% των παρατηρήσεων

Ιδιότητες Κανονικής Κατανομής (πώς η σταθερή απόκλιση δείχνει τον τρόπο διασποράς των παρατηρήσεων)



Σε μια έρευνα σε δείγμα 1500 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.

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

Το 68% έχει ηλικία από 31-39 έτη (από $(\mu-\sigma)$ μέχρι $(\mu+\sigma)$)
Το 95% έχει ηλικία από 27-43 έτη (από $(\mu-2\sigma)$ μέχρι $(\mu+2\sigma)$)
Το 99% έχει ηλικία από 23-47 έτη (από $(\mu-3\sigma)$ μέχρι $(\mu+3\sigma)$)

Τι ποσοστό του δείγματος έχει ηλικία:
Α) πάνω από 47 έτη (= $\mu+3\sigma$): 0,15%
Β) κάτω από 23 έτη (= $\mu-3\sigma$): 0,15%
Γ) από 23 μέχρι 39 έτη (από $(\mu-3\sigma)$ μέχρι $(\mu+\sigma)$): περίπου 84%
Δ) μέχρι 35 έτη (= μ): 50% (μέση τιμή =διάμεση)

Εκατοστημόρια

- Το Κ εκατοστημόριο είναι εκείνη η τιμή της κατανομής με βάση την οποία ποσοστό ίσο με το Κ% των παρατηρήσεων βρίσκεται κάτω ή πάνω από την τιμή αυτή.

- Πχ το 5ο εκατοστημόριο είναι η τιμή από την οποία 5% των παρατηρήσεων είναι μικρότερες (και 95% μεγαλύτερες)

- Το 75ο εκατοστημόριο είναι η τιμή από την οποία 75% των παρατηρήσεων είναι μικρότερες (και 25% μεγαλύτερες)

- Το 50ο εκατοστημόριο?

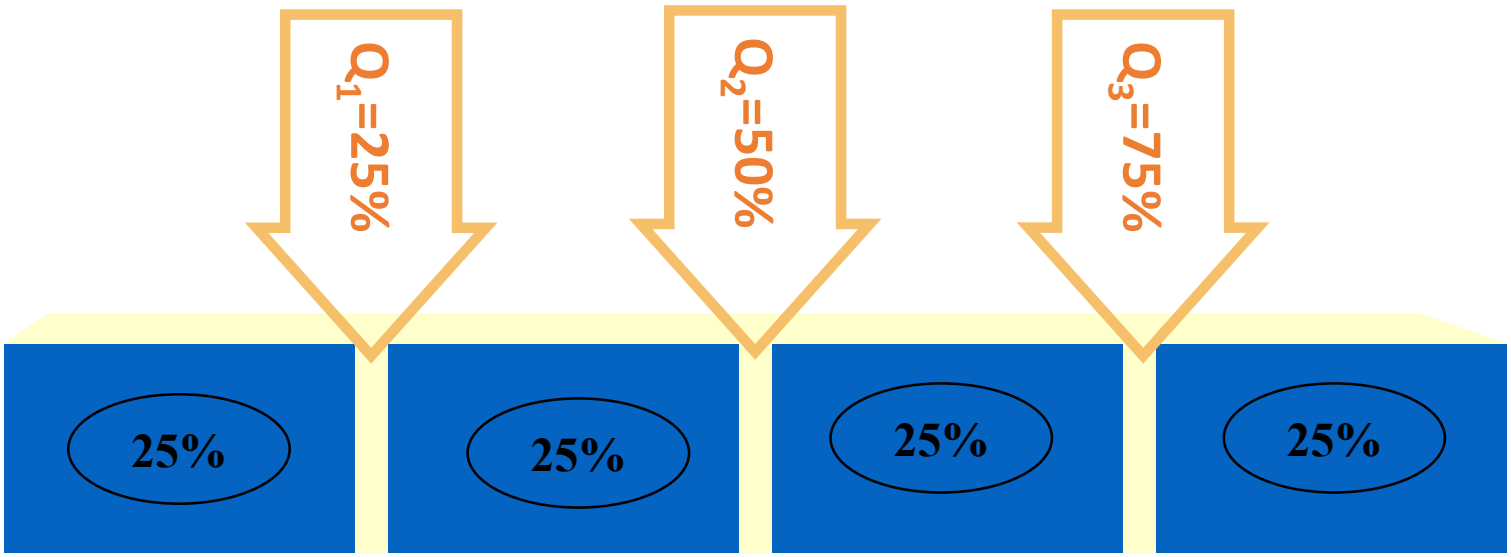
Πώς υπολογίζονται τα εκατοστημόρια?

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

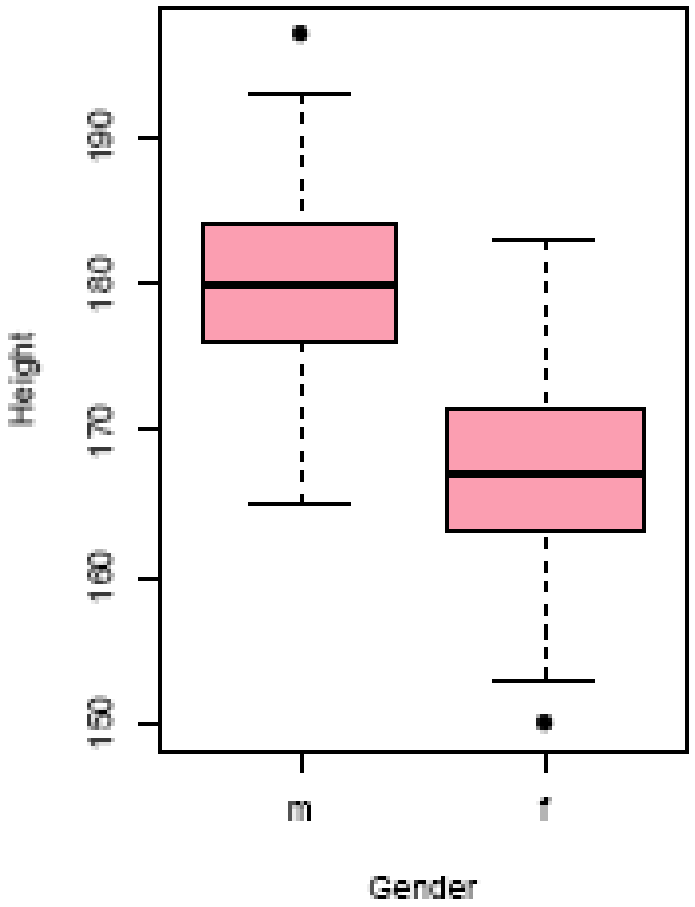
$$\frac{(n+1)*k}{100}$$

- Τέλος βρίσκουμε την τιμή που αντιστοιχεί σ' αυτήν τη θέση

Τεταρτημόρια (Quartiles)



Θηκόγραμμα (Boxplot)



≤Q3+1.5(Q3-Q1)

75ο εκατοστημόριο (3ο τεταρτημόριο-Q3)

Διάμεσος

25ο εκατοστημόριο (1ο τεταρτημόριο-Q1)

≥Q1-1.5(Q3-Q1)

Παρουσίαση Ποσοτικών δεδομένων

- Κατά προσέγγιση κανονικές κατανομές
  - Μέση Τιμή
  - Σταθερή απόκλιση

- Ασύμμετρες κατανομές
  - Διάμεσο
  - Τεταρτημόρια (25^o, 75^o), Εκατοστημόρια (10^o,90^o)

ΕΦΑΡΜΟΓΗ

Σκοπός της προκείμενης ερευνητικής εργασίας είναι η διερεύνηση παραγόντων που πιθανόν σχετίζονται με χαμηλό βάρος γέννησης (≤2500 gr).Τα δεδομένα προέρχονται από το νοσοκομείο Baystate Medical Center, Springfield, Massachusetts(Applied Logistic Regression- DAVID W. HOSMER, JR, STANLEY LEMESHOW, RODNEY X. STURDIVANT) και περιέχει στοιχεία από 189 γυναίκες.

ΠΕΡΙΓΡΑΦΗ ΜΕΤΑΒΛΗΤΩΝ

- **Id**:Patient's ID l
- **lbw**: Birth Weight (1:BWT≤2500g, 0:BWT>2500g)
- **age**: Age of Mother (yrs)
- **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)

ΕΡΜΗΝΕΙΑ

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 (τ.α) 5,3 έτη.
- Η διάμεση τιμή είναι τα 23 έτη
- Η ελάχιστη τιμή της ηλικίας 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 ηλικία ακολουθεί την κανονική κατανομή;