1η ΑΣΚΗΣΗ «ΙΑΤΡΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ»

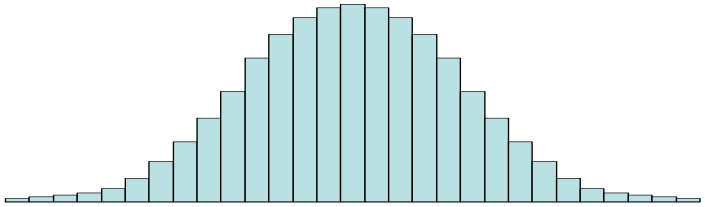
ΑΙΘΟΥΣΑ ΥΠΟΛΟΓΙΣΤΩΝ ΚΤΗΡΙΟ 14, 1ος ΟΡΟΦΟΣ

ΩΡΑ 11:00-13:00

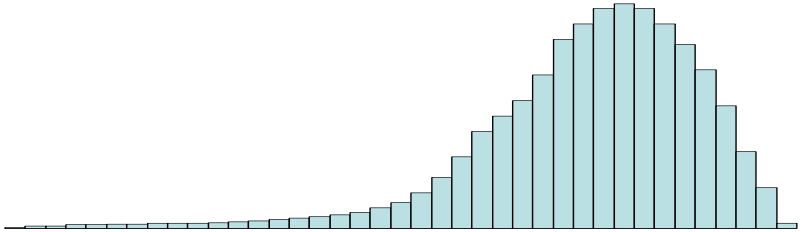
	Άσκηση	Ομάδα
Πέμπτη 10-10-2024	1η	1450202400001-1450202400050
Τρίτη 15-10-2024	1η	1450202400051-1450202400100
Πέμπτη 17-10-2024	1η	1450202400101-1450202400150
Τρίτη 22-10-2024	1η	1450202400151-1450202400200

Κατανομή Ποσοτικών Δεδομένων

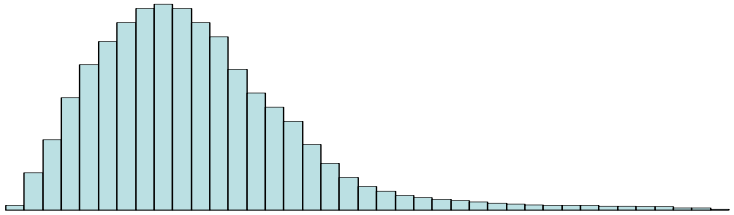
Συμμετρική, σχήμα καμπάνας
Κανονική (Gaussian) κατανομή



Αρνητικά λοξή
Λιγότερο συνηθισμένη

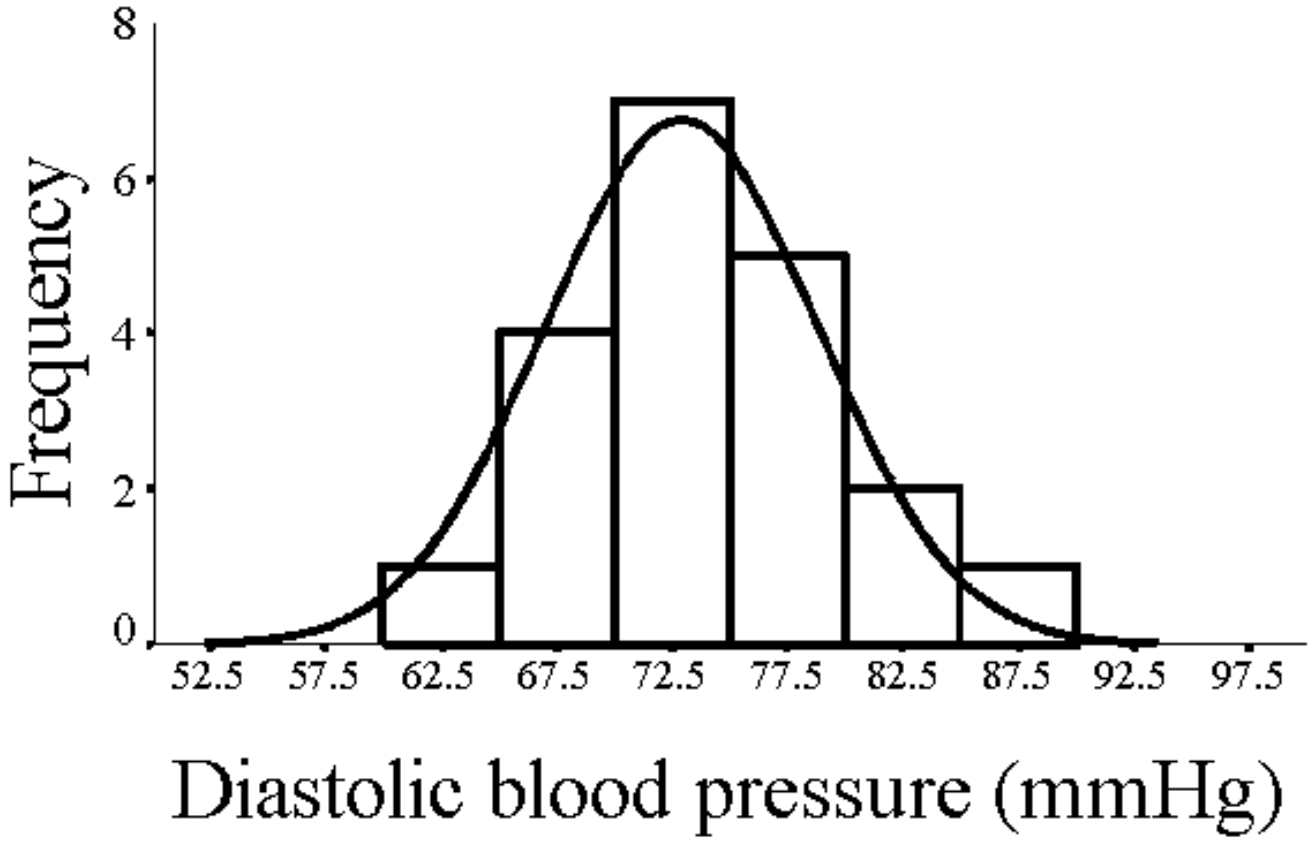


Θετικά λοξή
Σε εργαστηριακά δεδομένα
πχ. CD4 counts in HIV+



Ομοιόμορφη κατανομή
Ίδια πιθανότητα οποιασδήποτε τιμής
μέσα στο εύρος

Παράδειγμα κανονικής κατανομής



Βασικοί ορισμοί

- Πληθυσμός (population): ένα σύνολο ατόμων
- Παράμετρος (parameter): χαρακτηριστικό του πληθυσμού (π.χ. μ, σ, p)
- Δείγμα (sample): ένα μέρος του πληθυσμού
- Στατιστικό στοιχείο (statistic): υπολογίζεται από το δείγμα (π.χ. , SD, επιπολασμός)

- Το στατιστικό στοιχείο εξυπηρετεί 2 σκοπούς:
  - ✓ περιγραφή δείγματος
  - ✓ εκτίμηση παραμέτρου

Ποσοτικές μεταβλητές

- Περιγράφονται συνοπτικά και αποτελεσματικά με τη χρήση παραμέτρων

- Οι βασικές ομάδες παραμέτρων που χρησιμοποιούνται είναι οι αντιπροσωπευτικές τιμές θέσης και οι αντιπροσωπευτικές τιμές διασποράς

- **Έγκυρες** όταν τα δείγματα είναι αντιπροσωπευτικά ή τυχαία του πληθυσμού.

Αντιπροσωπευτικό (Τυχαίο) Δείγμα

Ένα δείγμα θεωρείται τυχαίο όταν κάθε μέλος του γενικού συνόλου υπό μελέτη έχει την ίδια πιθανότητα να περιληφθεί στο δείγμα.

Εισαγωγή συστηματικών σφαλμάτων στην έρευνα

Πχ Υπολογισμός μέσου ύψους αγοριών ηλικίας 7 ετών με δείγμα μόνο από Αθήνα.

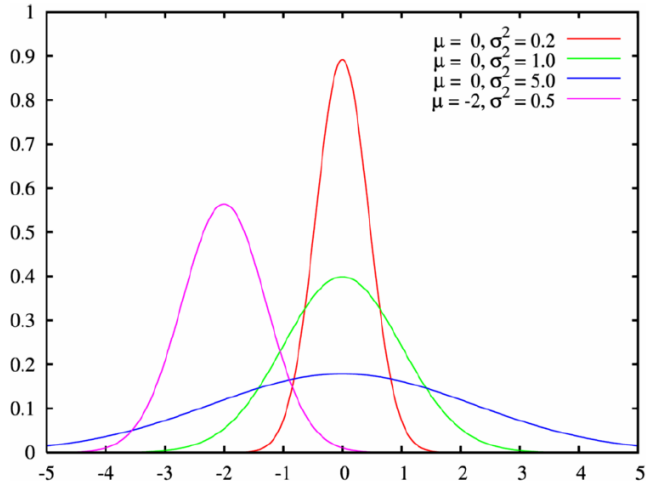
Αντιπροσωπευτικές τιμές

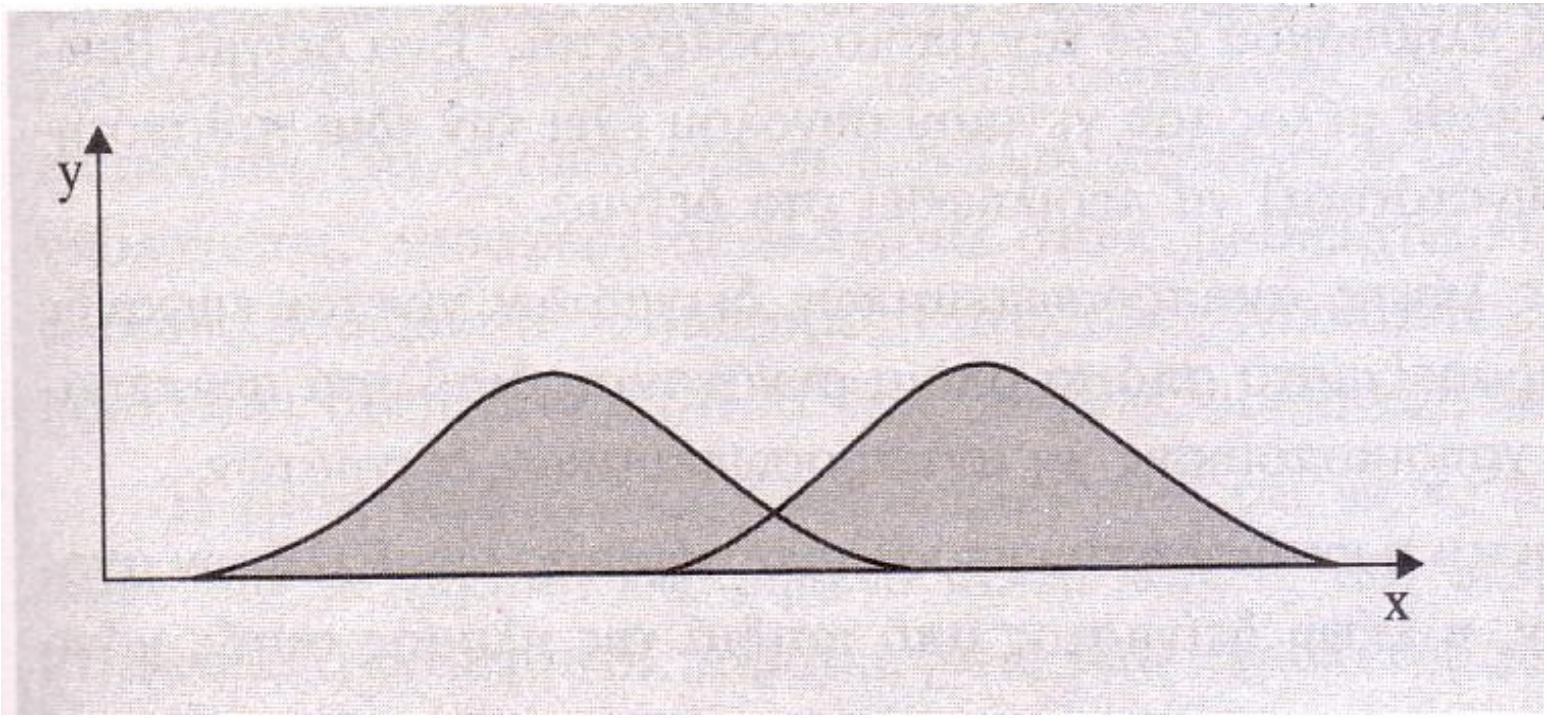
- Τιμές θέσης

  Επικρατούσα τιμή, μέση τιμή, διάμεσος

- Τιμές βαθμού διασποράς

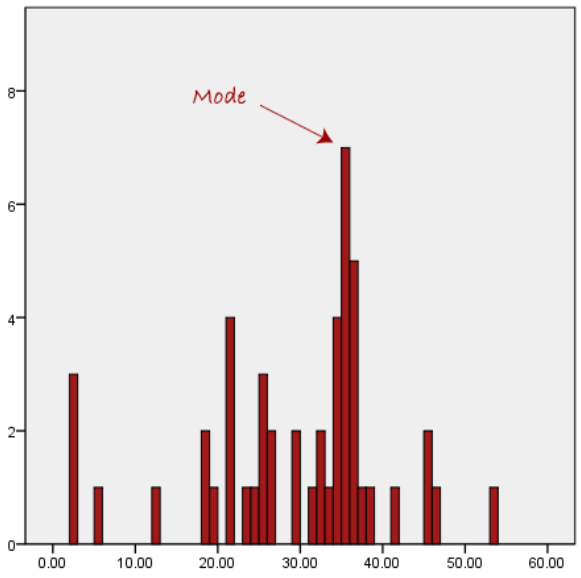
  Τυπική απόκλιση, ακραίες τιμές, εκατοστημόρια

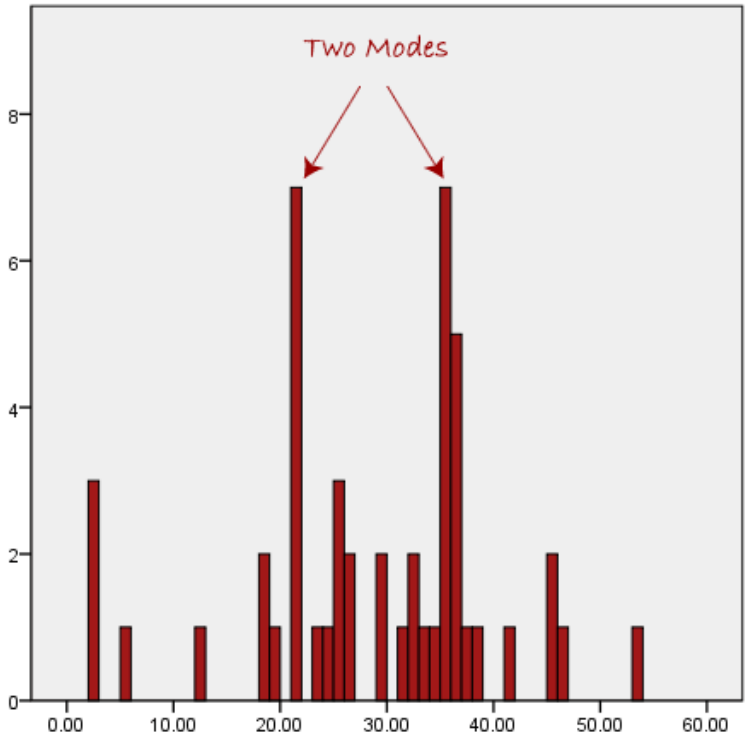



y
x

Αντιπροσωπευτικές τιμές θέσης

- ✓ **Επικρατούσα τιμή** είναι η τιμή στην οποία σημειώθηκαν οι περισσότερες παρατηρήσεις.
  - ✓ Πρόβλημα με λίγες παρατηρήσεις
  - ✓ Μόνο περιγραφική χρήση





Αντιπροσωπευτικές τιμές θέσης

- **Μέση τιμή** είναι το αλγεβρικό άθροισμα όλων των μετρήσεων διαιρεμένο με το πλήθος αυτών.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

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

όπου i η ερευνητική μονάδα και Xi η μέτρηση.

Παράδειγμα

- Παρατηρήσεις βάρους (kg) 5 παιδιών: 52, 45, 33, 40, 28

$$\overline{X} = 198/5=39,6 \text{ kg}$$

- Παρατηρήσεις διαφοράς στην αρτηριακή πίεση (mm Hg) πριν και μετά θεραπεία σε 6 άτομα: -2, +4, -5, +8, -9, -3

$$\overline{X} = -7/6=-1,2 \text{ mmHg}$$

Αντιπροσωπευτικές τιμές θέσης

- **Μέση τιμή** είναι το αλγεβρικό άθροισμα όλων των μετρήσεων διαιρεμένο με το πλήθος αυτών.

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

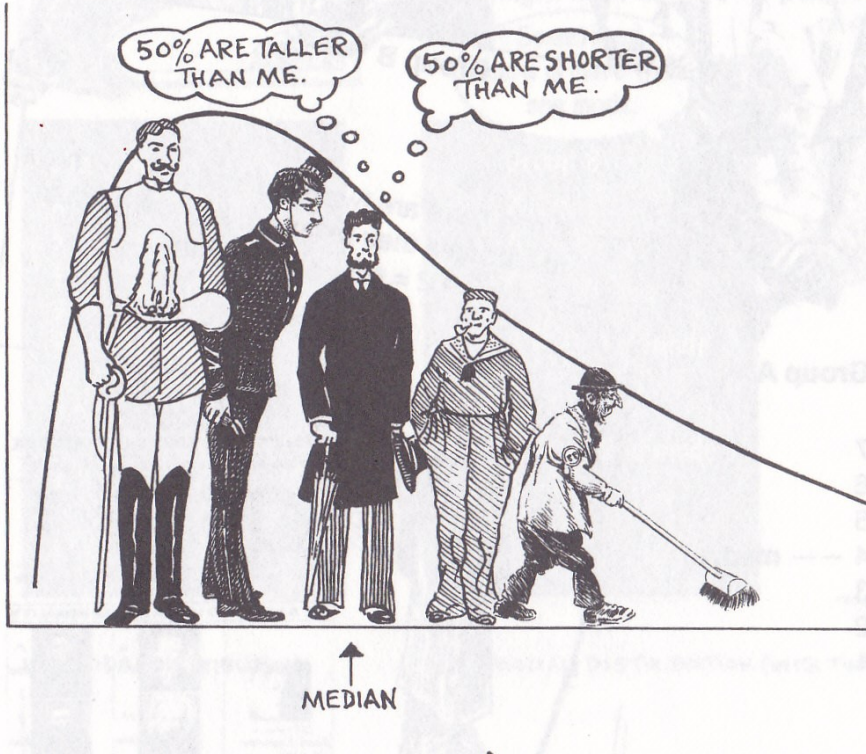
όπου i η ερευνητική μονάδα και Xi η μέτρηση.

ΙΔΙΟΤΗΤΑ: $\sum_i (X_i - \bar{X}) = 0$

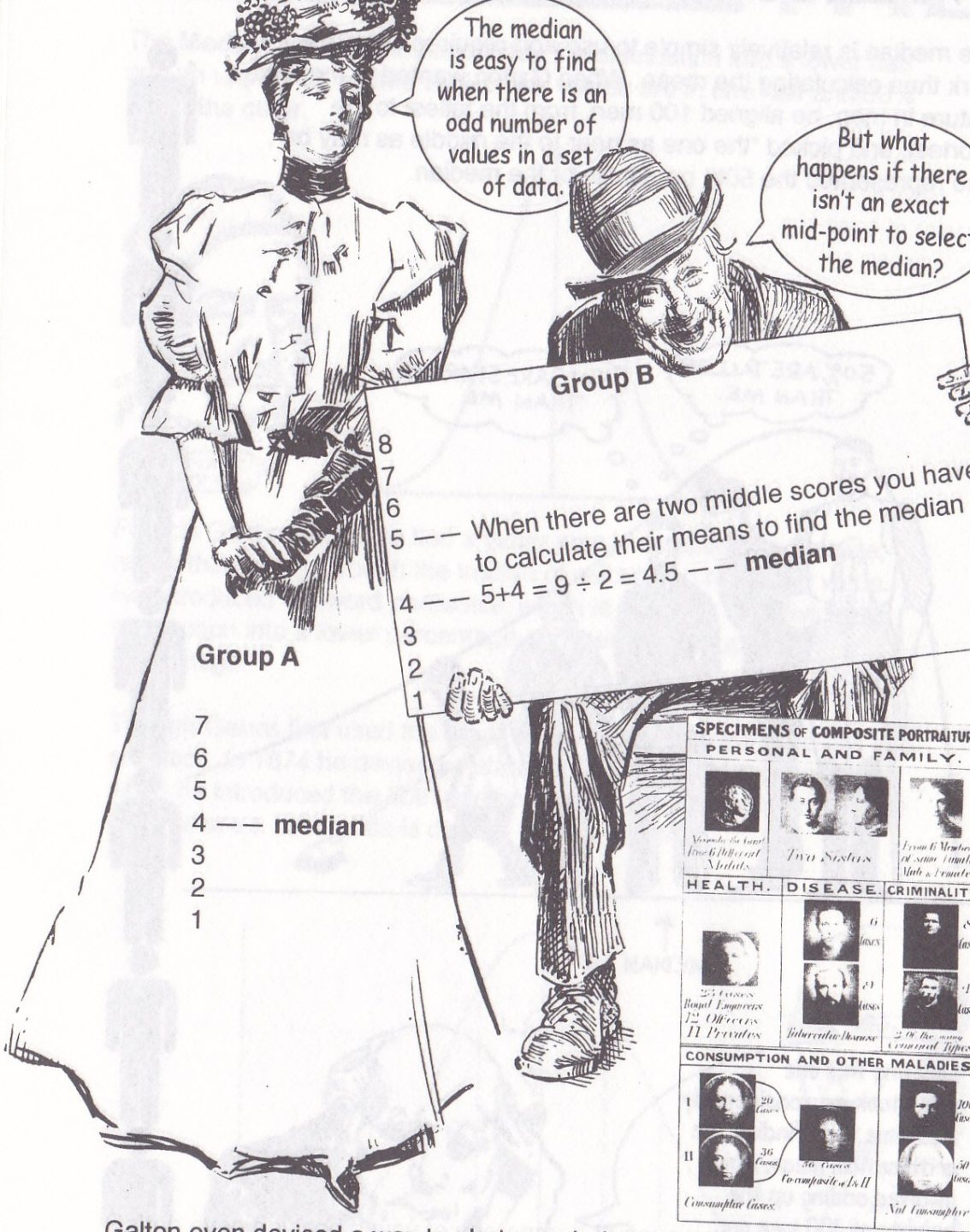
Αντιπροσωπευτικές τιμές θέσης

✓**Διάμεσος** είναι η τιμή που είναι συγχρόνως μεγαλύτερη από τις μισές μετρήσεις και μικρότερη από τις άλλες μισές.

- Υπολογίζεται δυσκολότερα όταν έχουμε ισοψηφίες ή πολλές παρατηρήσεις.
- Μεγαλύτερη χρήση στη περιγραφική στατιστική.



- Για να υπολογιστεί πρέπει πρώτα να διαταχθούν οι παρατηρήσεις κατά αύξουσα (ή φθίνουσα) σειρά
- Όταν το πλήθος των παρατηρήσεων είναι μονός αριθμός, τότε είναι η μεσαία παρατήρηση.
- Όταν το πλήθος των παρατηρήσεων είναι ζυγός αριθμός, τότε είναι ο μέσος όρος των δύο μεσαίων παρατηρήσεων



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)

Παράδειγμα υπολογισμού διαμέσου

- Έστω ότι έχουμε τις παρατηρήσεις για το ανάστημα 9 ατόμων σε εκατοστά:
- 154, 189, 173, 192, 151, 161, 165, 189, 170
- Σε αύξουσα σειρά
- 151, 154, 161, 165, 170, 173, 189, 189, 192
- Η θέση της διαμέσου (9+1)/2= $5^{η}$ παρατήρηση
- Άρα το 170

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

Θέση Διαμέσου: $\frac{18+1}{2} = 9,5$

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

Διάταξη παρατηρήσεων και προσδιορισμός της θέσης (της σειράς) κάθε παρατήρησης

Παρατήρηση (βάρος kg)	Διατεταγμένη σε αύξουσα σειρά	Θέση ή Σειρά
45	41	1
56	45	2
41	53	3
72	56	4,5 (μ.ο.)*
86	56	4,5 (μ.ο.)*
69	69	6
53	72	7
56	86	8

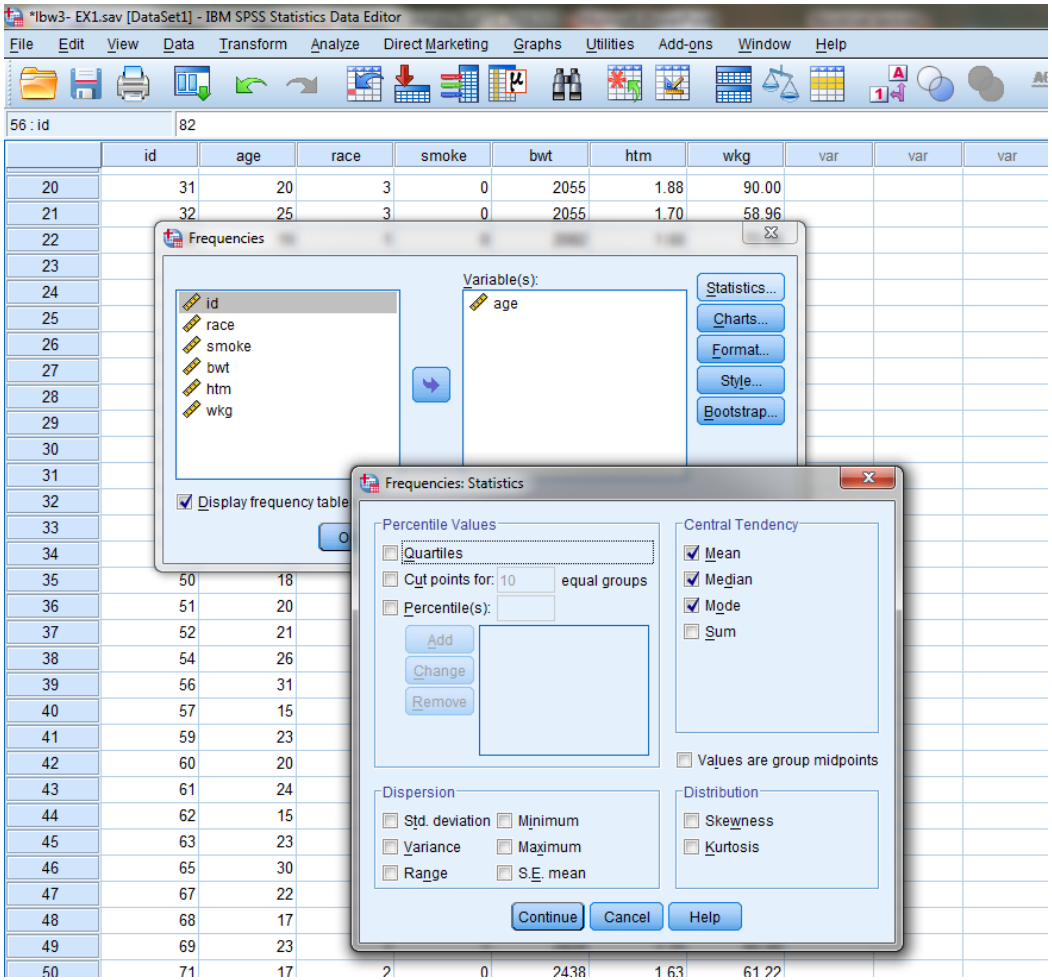
* Μέσος όρος των θέσεων που θα καταλάμβαναν αν δεν ισοψηφούσαν

Προσοχή

- Οι αντιπροσωπευτικές τιμές θέσης εκφράζονται σε τιμές του μεγέθους που μελετάμε

- Σε κανονικές κατανομές, η καλύτερη αντιπροσωπευτική τιμή θέσης είναι η μέση τιμή

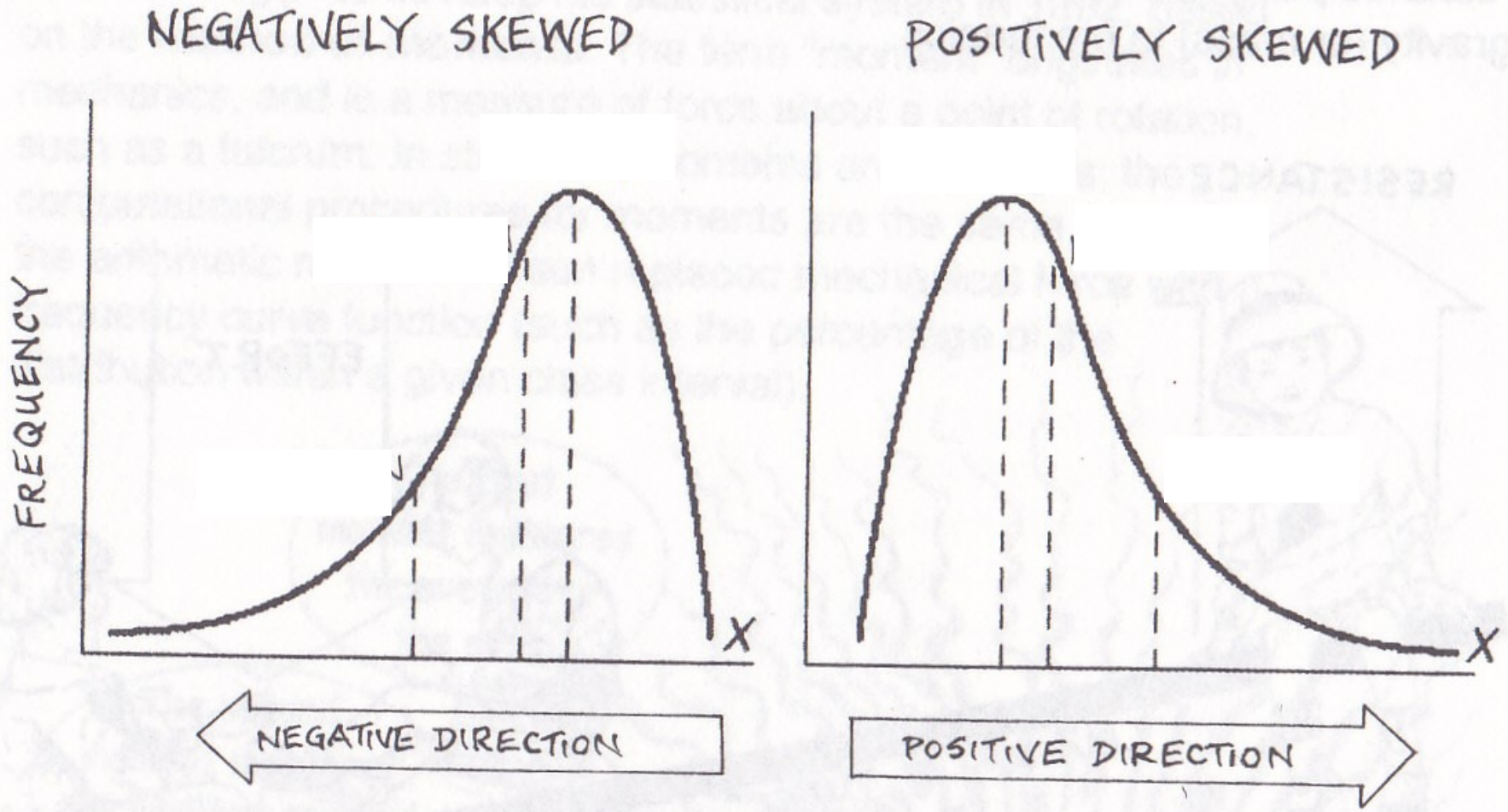
- Σε ασύμμετρες κατανομές, είναι η διάμεση, επειδή η μέση τιμή επηρεάζεται πολύ από τις ακραίες τιμές.



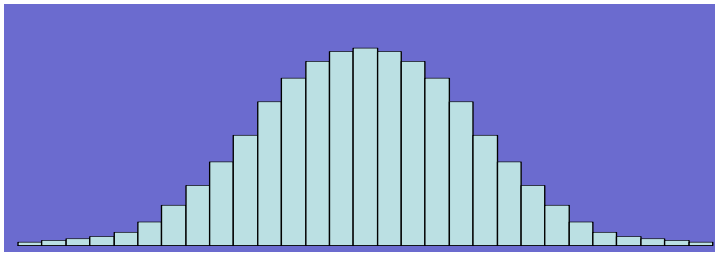
Statistics

age

N	Valid	187
	Missing	0
Mean		23.29
Median		23.00
Mode		20


NEGATIVELY SKEWED
POSITIVELY SKEWED
FREQUENCY
X
X
NEGATIVE DIRECTION
POSITIVE DIRECTION

Σχέσεις -Ιδιότητες



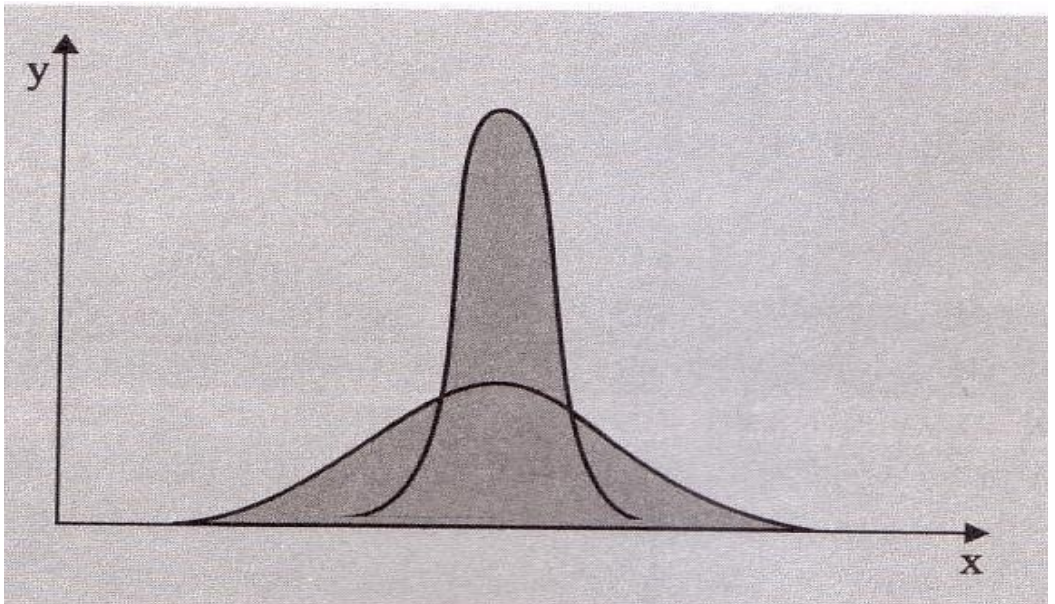
- ✓ **Κανονική κατανομή:**
  - ✓ Μέση τιμή= Διάμεσο =Επικρατούσα

- ✓ **Σε ασύμμετρες κωδωνοειδής ισχύει εμπειρικά:**

Επικρατούσα – Διάμεσος = 2* (Διάμεσος – Μέση)

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



Αντιπροσωπευτικές τιμές Διασποράς

- Εύρος/ ακραίες τιμές
- Εκατοστημόρια
- Σταθερή απόκλιση

Εύρος/ ακραίες τιμές

- 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) $$V = \sigma^2 = \frac{\sum_i (X_i - \mu)^2}{N}$$
- ✓ Τυπική / σταθερή απόκλιση (Standard Deviation)

$$SD = \sigma = \sqrt{\frac{\sum_i (X_i - \mu)^2}{N}} = \sqrt{\frac{\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{N}}{N}}$$

$$\sigma = \sqrt{\frac{\sum_{i}(X_i^2 + \mu^2 - 2\mu X_i)}{N}} = \sqrt{\frac{\sum_{i}(X_i^2) + n * \mu^2 - \sum_{i} 2\mu X_i)}{N}}$$

$$= \sqrt{\frac{\sum_{i} X_i^2 - \frac{(\sum_{i} X_i)^2}{N}}{N}}$$

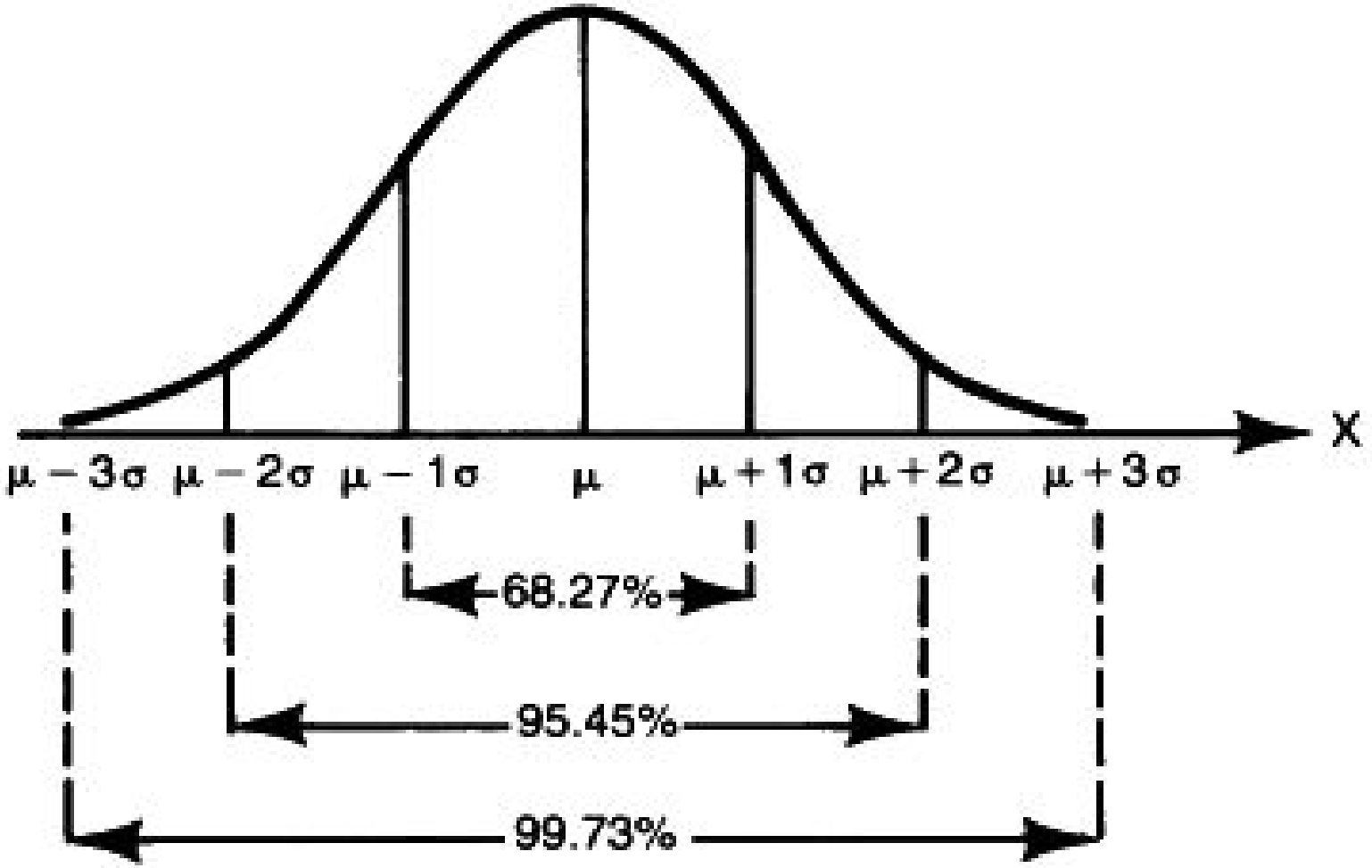
Ο παρονομαστής αντικαθίσταται με n-1 στα δείγματα

$$SD = \sqrt{\frac{\sum_{i}(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{\sum_{i} X_i^2 - \frac{(\sum_{i} X_i)^2}{n}}{n-1}}$$

- 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)



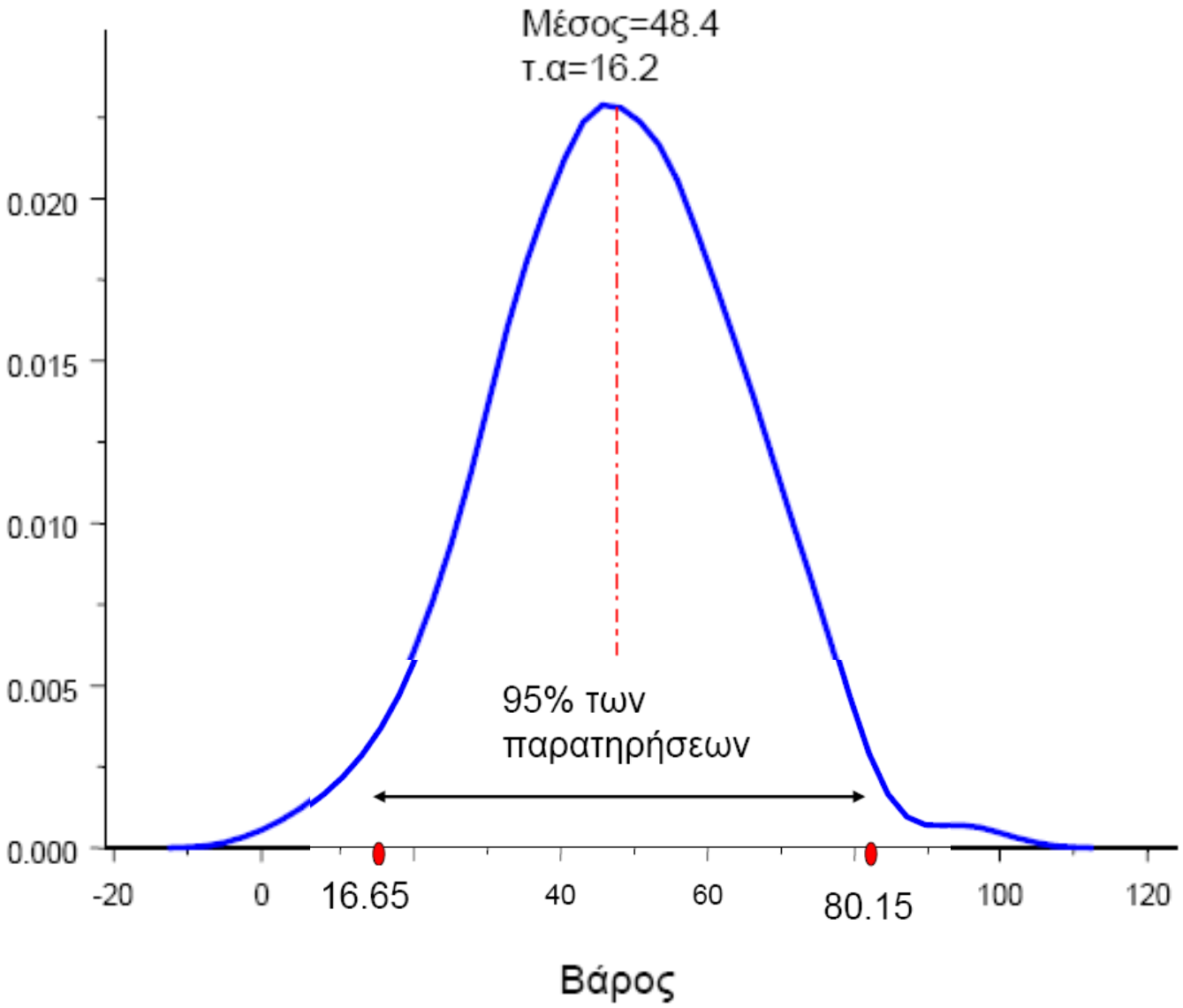
ΠΙΝΑΚΑΣ Π1

Κανονική κατανομή. **Το εκατοστιαίο ποσοστό των παρατηρήσεων των οποίων οι τιμές περιλαμβάνονται μεταξύ των τιμών $\bar{x} + Zs$ και $\bar{x} - Zs$ και το εκατοστιαίο ποσοστό των παρατηρήσεων των οποίων οι τιμές βρίσκονται εκτός των παραπάνω τιμών (δηλαδή είναι μεγαλύτερες της $\bar{x} + Zs$ ή μικρότερες της $\bar{x} - Zs$). Με $\bar{x}$ παριστάνεται η μέση τιμή, με s η σταθερή απόκλιση και με Z το κάθε φορά υποπολλαπλάσιο ή πολλαπλάσιο της σταθερής απόκλισης (standard normal deviate).**

Z	Μεταξύ των $\bar{x} + Zs$ και $\bar{x} - Zs$	Εκτός των $\bar{x} + Zs$ και $\bar{x} - Zs$
0,00	0,00	100,00
0,10	7,97	92,03
0,20	15,85	84,15
0,30	23,58	76,42
0,40	31,08	68,92
0,50	38,29	61,71
0,60	45,15	54,85
0,70	51,61	48,39
0,80	57,63	42,37
0,90	63,19	36,81
1,00	68,27	31,73
1,10	72,87	27,13
1,20	76,99	23,01
1,30	80,64	19,36
1,40	83,85	16,15
1,50	86,64	13,36
1,60	89,04	10,96
1,645	**90,00**	**10,00**
1,70	91,09	8,91
1,80	92,81	7,19
1,90	94,26	5,74
1,960	**95,00**	**5,00**
2,00	95,44	4,56
2,10	96,43	3,57
2,20	97,22	2,78
2,30	97,85	2,15
2,40	98,36	1,64
2,50	98,76	1,24
2,576	**99,00**	**1,00**
2,60	99,07	0,93
2,70	99,31	0,69
2,80	99,49	0,51
2,90	99,63	0,37
3,00	99,73	0,27
3,10	99,81	0,19
3,20	99,86	0,14
3,30	99,90	0,10

Συχνότητες

Διάστημα τιμών	Τιμή	Απόλυτη συχνότητα	Σχετική συχνότητα	Απόλυτη αθροιστική συχνότητα	Σχετική αθροιστική συχνότητα
0.00- 9.99	5	1	0.01	1	0.01
10.00-19.99	15	3	0.03	4	0.04
20.00-29.99	25	8	0.08		
30.00-39.99	35	18	0.18		
40.00-49.99	45	24	0.24		
50.00-59.99	55	22	0.22		
60.00-69.99	65	15	0.15		
70.00-79.99	75	8	0.08		
80.00-89.99	85	0	0.00		
90.00-99.99	95	1	0.01		



Εκατοστημόρια

- Είναι σύνολο τιμών που το καθένα διαχωρίζει το αντίστοιχο ποσοστό παρατηρήσεων από το σύνολο των τιμών

- 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

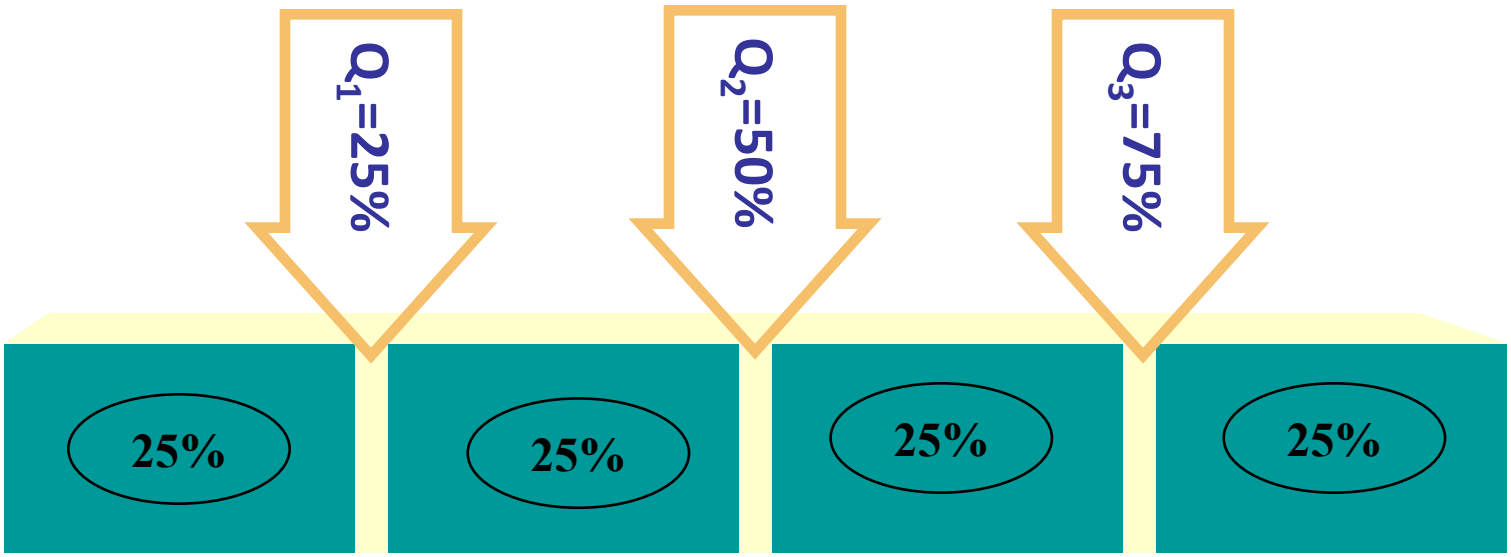
$$\frac{(n+1)*k}{100}$$

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

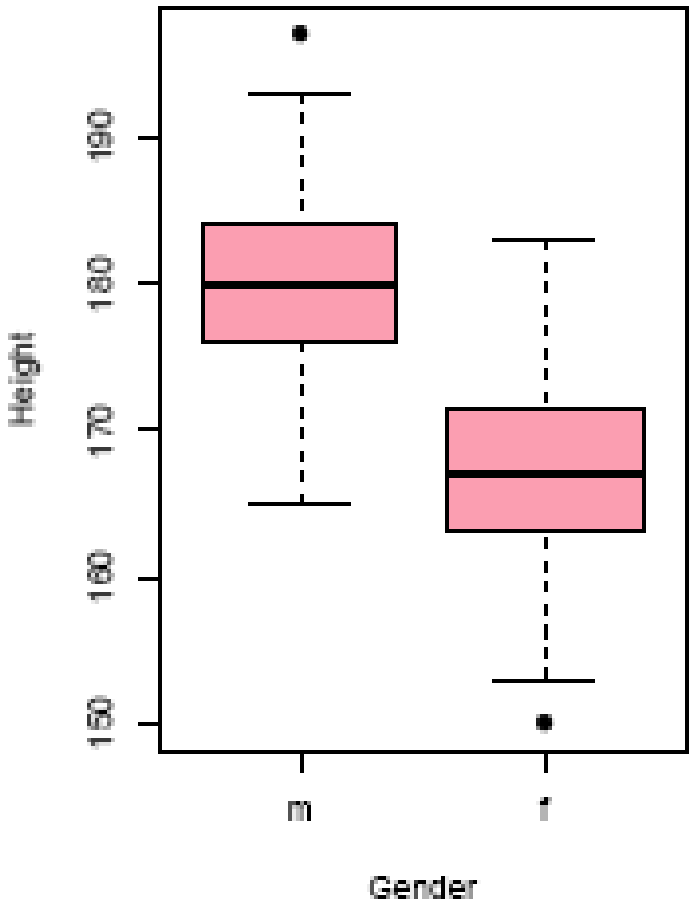
age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	14	2	1.1	1.1	1.1
	15	3	1.6	1.6	2.7
	16	7	3.7	3.7	6.4
	17	12	6.4	6.4	12.8
	18	10	5.3	5.3	18.2
	19	16	8.6	8.6	26.7
	20	18	9.6	9.6	36.4
	21	12	6.4	6.4	42.8
	22	12	6.4	6.4	49.2
	23	13	7.0	7.0	56.1
	24	13	7.0	7.0	63.1
	25	15	8.0	8.0	71.1
	26	8	4.3	4.3	75.4
	27	3	1.6	1.6	77.0
	28	9	4.8	4.8	81.8
	29	7	3.7	3.7	85.6
	30	7	3.7	3.7	89.3
	31	5	2.7	2.7	92.0
	32	6	3.2	3.2	95.2
	33	3	1.6	1.6	96.8
	34	1	.5	.5	97.3
	35	2	1.1	1.1	98.4
	36	2	1.1	1.1	99.5
	45	1	.5	.5	100.0
	Total	187	100.0	100.0	

Τεταρτημόρια (Quartiles)



Θηκόγραμμα (Boxplot)



← ≤ Q3+1.5(Q3-Q1)

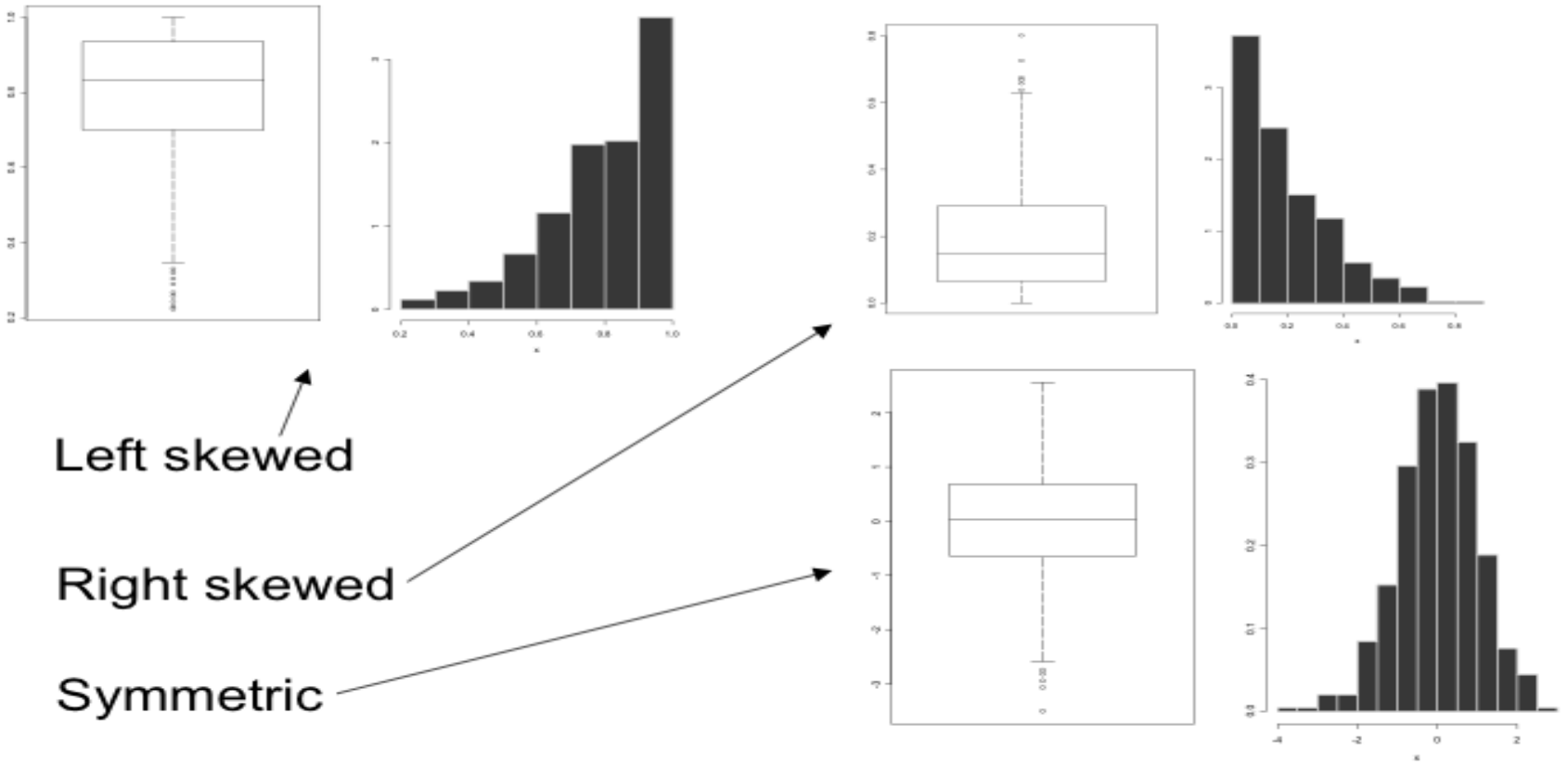
← 75ο εκατοστημόριο (3ο τεταρτημόριο-Q3)

← Διάμεσος

← 25ο εκατοστημόριο (1ο τεταρτημόριο-Q1)

← ≥ Q1-1.5(Q3-Q1)

Συμμετρία: ιστόγραμμα και θηκόγραμμα



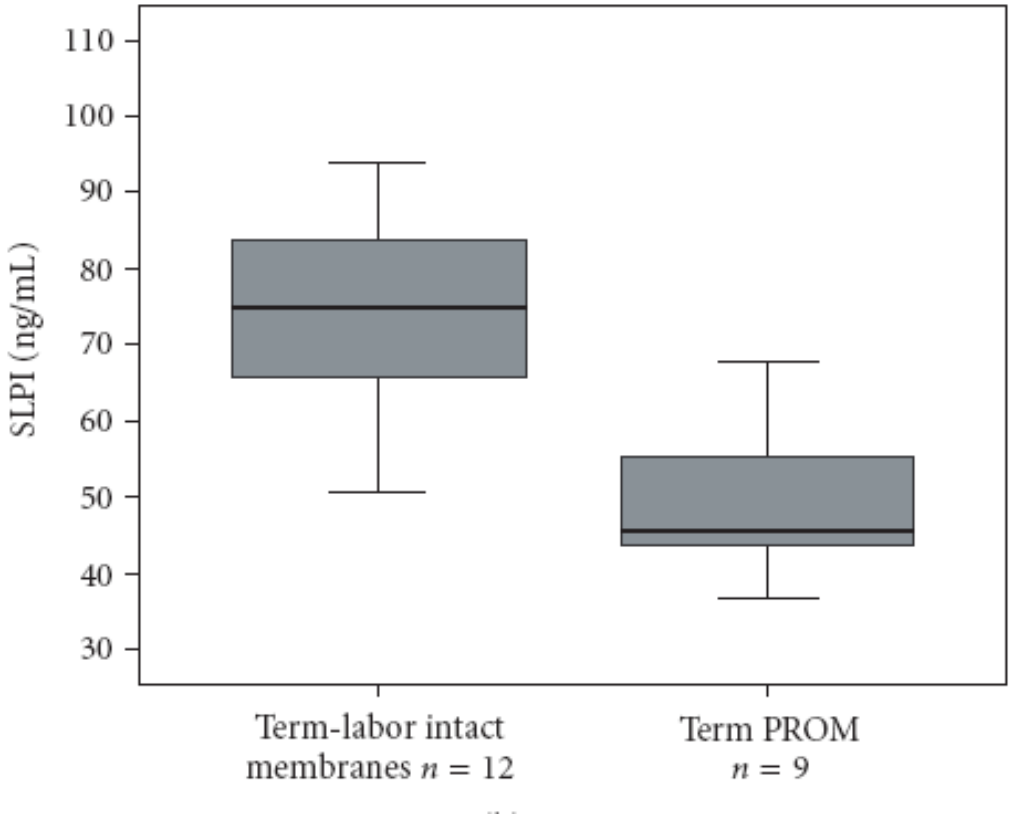
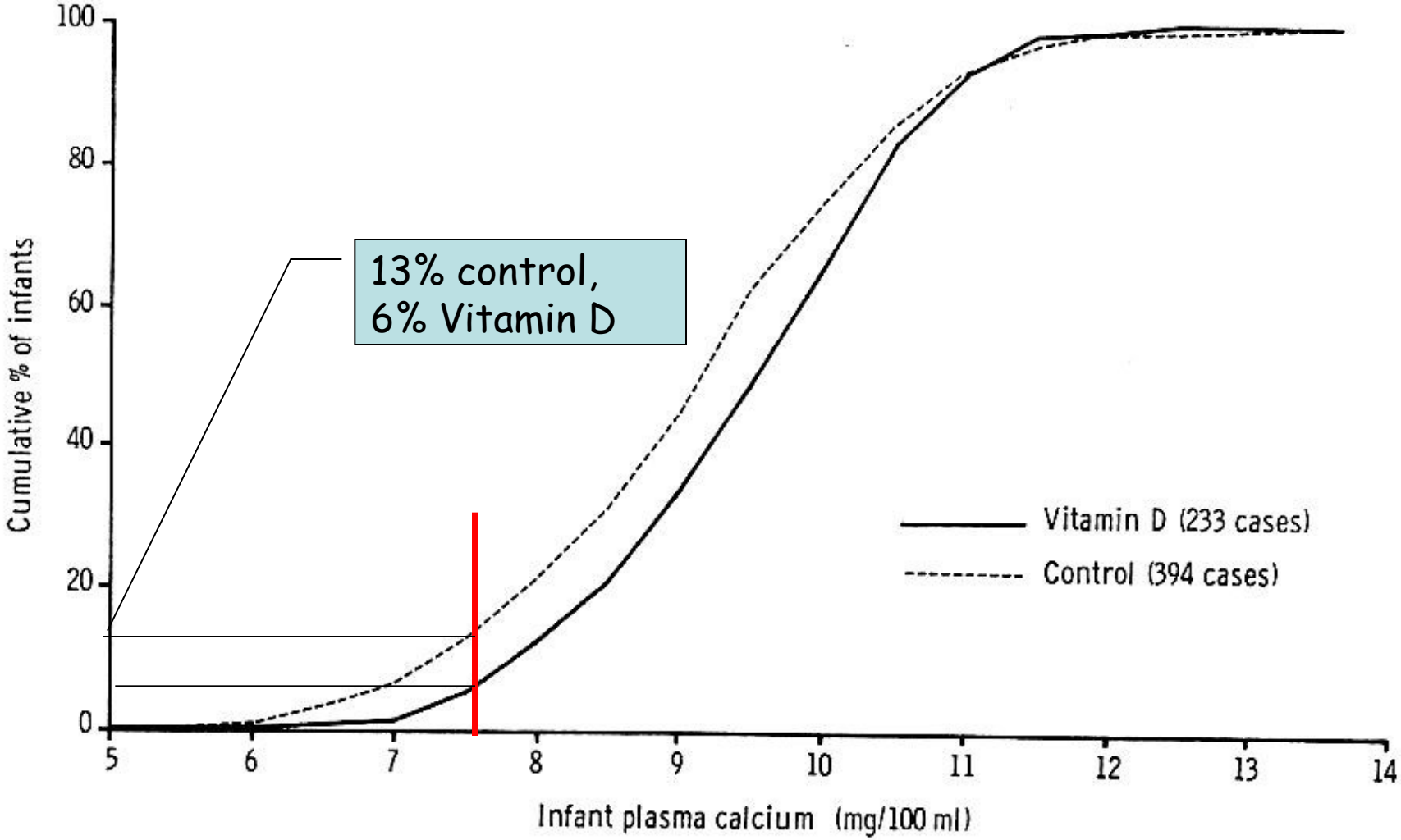


FIGURE 1: Levels of SLPI in women, who delivered either preterm (a) or at term (b) with intact membranes (a) $n = 7$ and (b) $n = 12$ or premature rupture of membranes (PROM) (a) $n = 6$ and (b) $n = 9$, respectively.

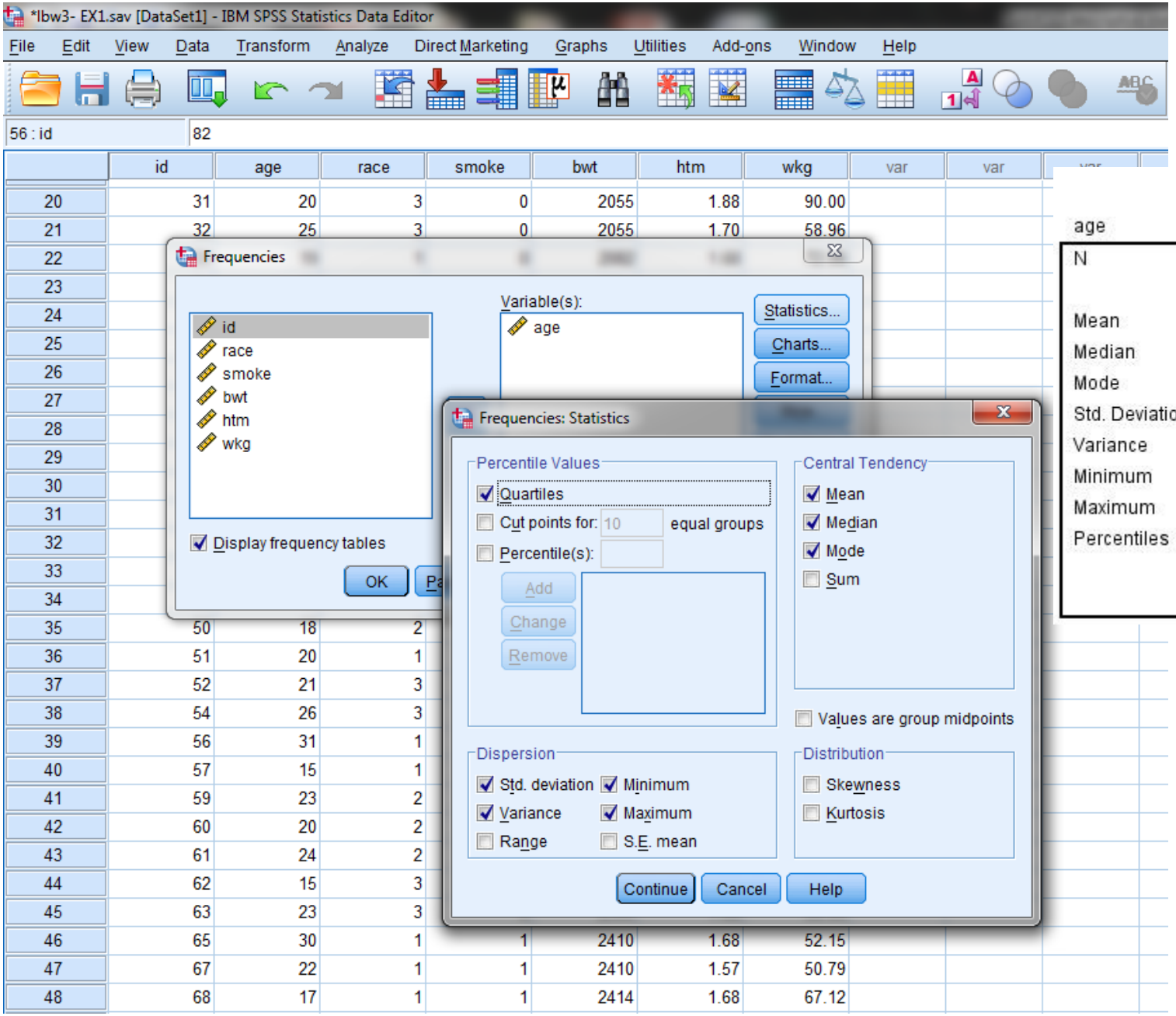
Πηγή: Malamitsi-Puchner et al. 2006

Διάγραμμα αθροιστικής κατανομής



Παράδειγμα παρουσίασης ποσοτικών δεδομένων κλινικής δοκιμής για πρόληψη βρεφικής υπο-ασβεστιαιμίας

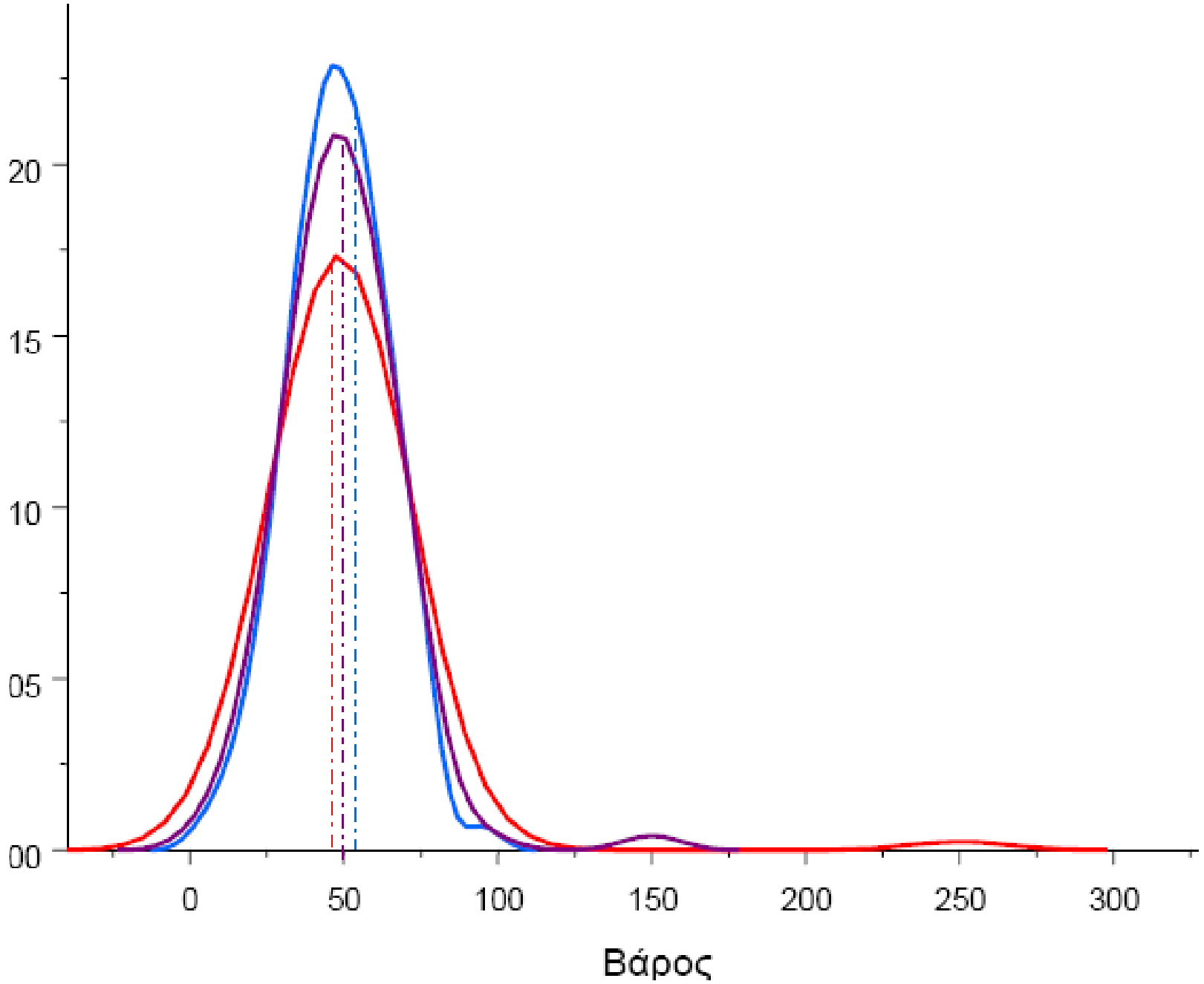
		Ασβέστιο στο πλάσμα στην 6η μέρα (mg/100ml)	
Αγωγή	Αριθμός ασθενών	Μέση τιμή	Σταθερή Απόκλιση (SD)
Vitamin D	233	9,36	1,15
Placebo	394	9,01	1,33



Statistics

age

N	Valid	187
	Missing	0
Mean		23.29
Median		23.00
Mode		20
Std. Deviation		5.283
Variance		27.908
Minimum		14
Maximum		45
Percentiles	25	19.00
	50	23.00
	75	26.00


20
15
10
05
00
0
50
100
150
200
250
300
Βάρος

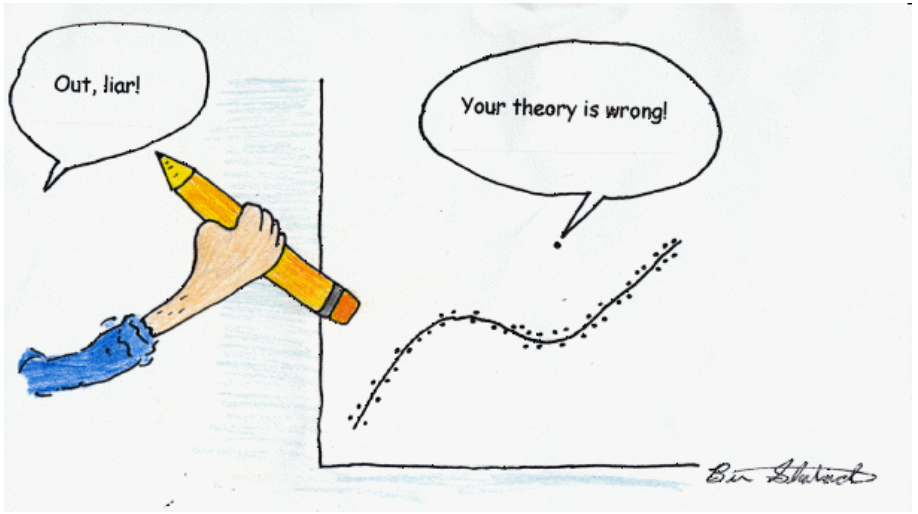
Περιγραφική Στατιστική Δεδομένων

- Κατά προσέγγιση κανονικές κατανομές
  - Μέση Τιμή
  - Σταθερή απόκλιση

- Ασύμμετρες κατανομές
  - Διάμεσο
  - Τεταρτημόρια (25^o, 75^o), Εκατοστημόρια (10^o,90^o)
  - Ενδοτεταρτομοριακό εύρος (Interquartile range-IQR)

Ακραίες Τιμές (Outliers)

“παρατηρήσεις που ξεχωρίζουν από τα υπόλοιπα δεδομένα”



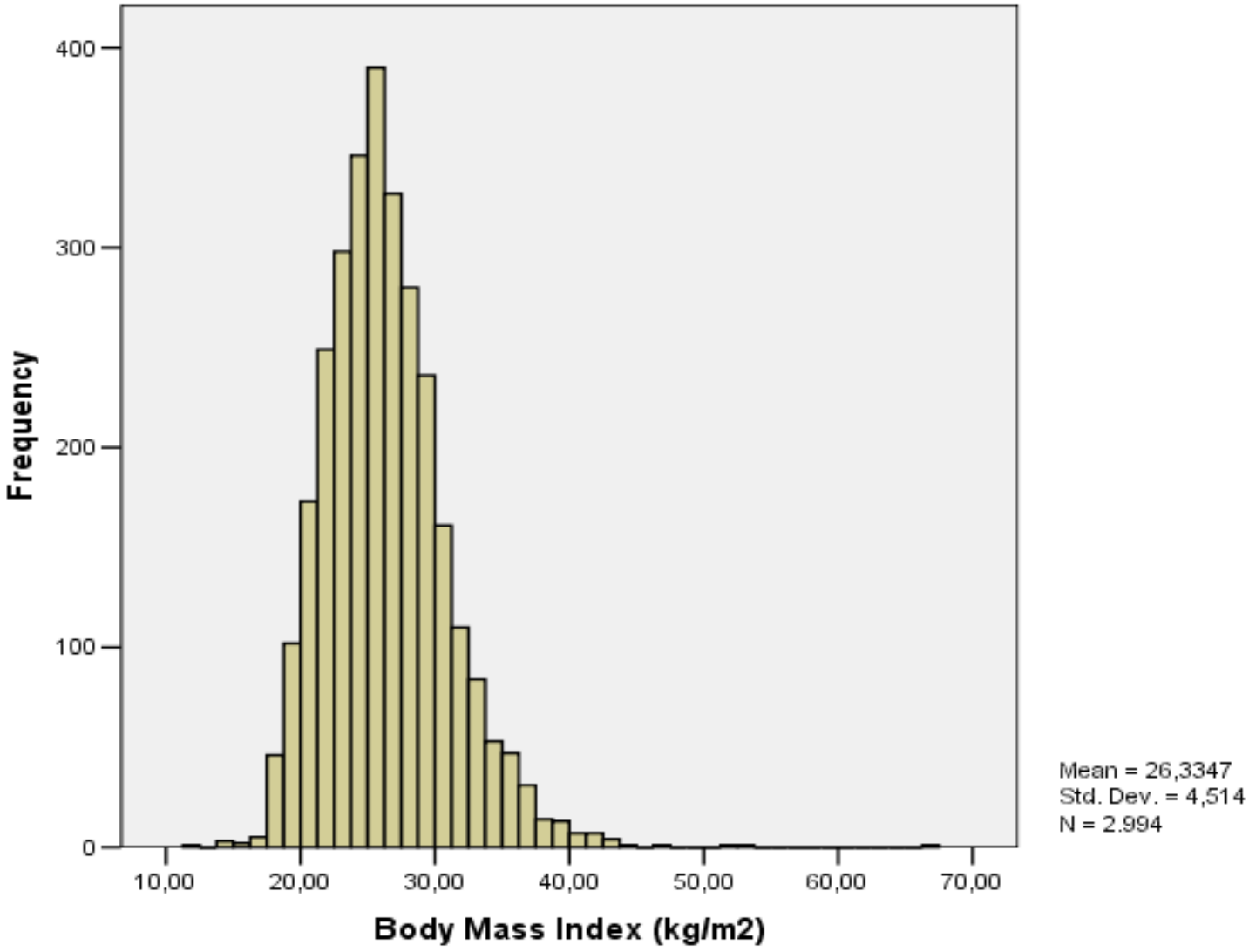
Προέρχονται από:

- Πραγματικές ακραίες τιμές, πχ υπέρβαρα άτομα
- Λάθος κατά την εισαγωγή δεδομένων

Προβλήματα αν είναι πραγματικές τιμές γιατί επηρεάζουν δυσανάλογα τα αποτελέσματα. Αντιμετώπιση:

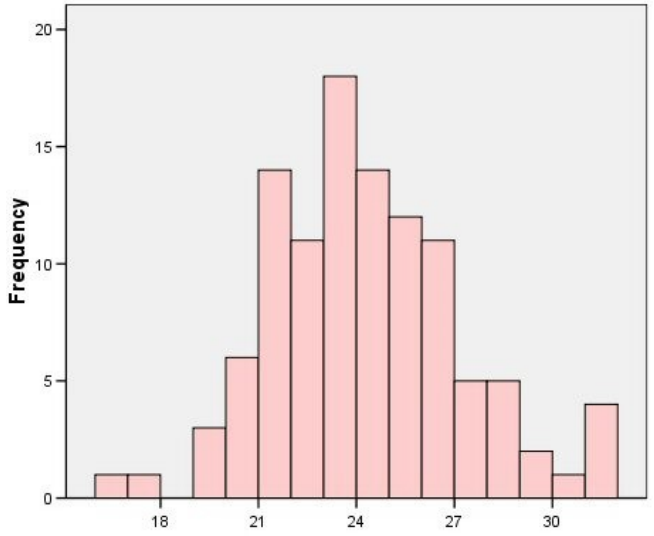
- Ανάλυση με και χωρίς αυτά τα άτομα
- Αν διαφορετικά αποτελέσματα μετασχηματισμός / μη παραμετρικές μεθόδους

Ακραίες Τιμές (Outliers)

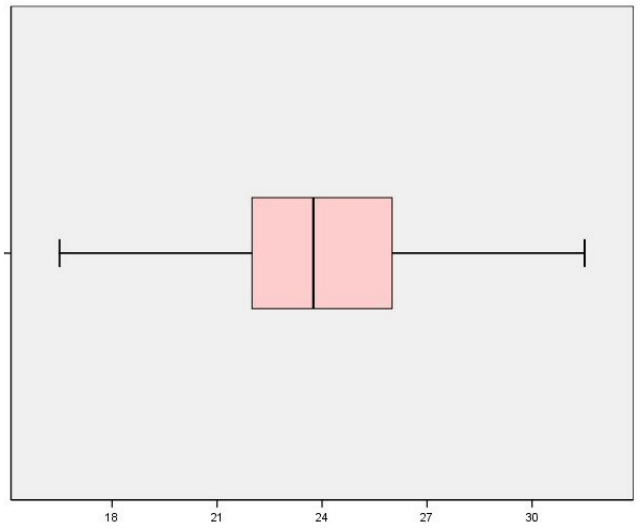


Εκτίμηση κανονικότητας

- Ιστόγραμμα - Θηκόγραμμα
- Περιγραφικά δεδομένα



Τι κάνουμε αν τα δεδομένα δεν κατανέμονται κανονικά;



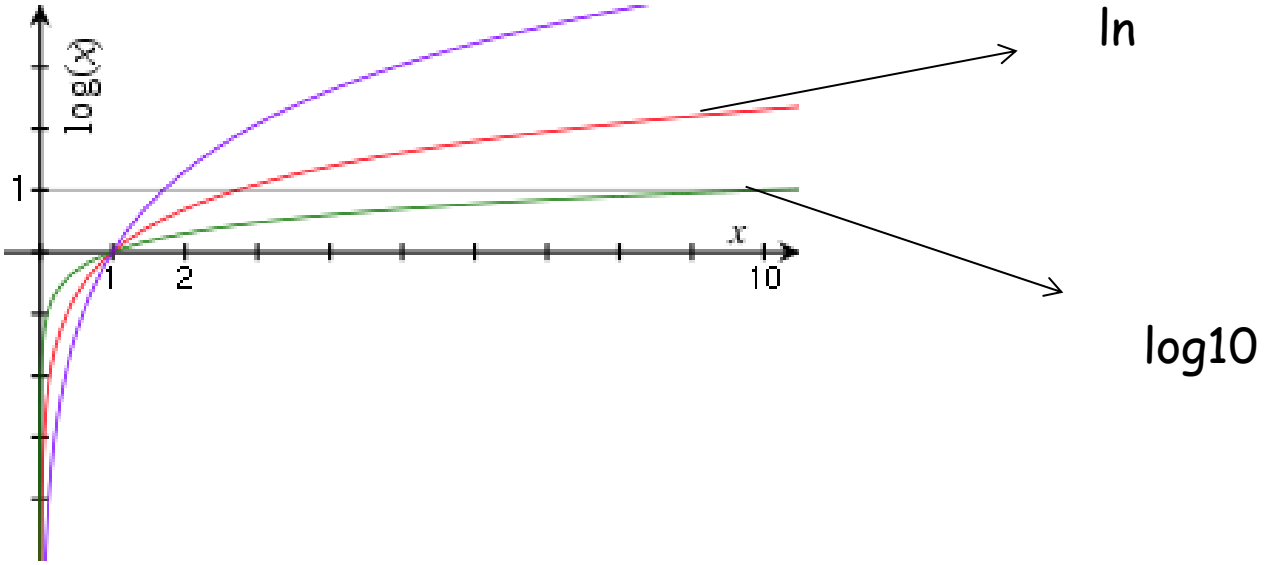
Πότε να μετασχηματίσουμε;

Όταν παραβιάζεται η κανονικότητα (προϋπόθεση πολλών στατιστικών δοκιμασιών), …για να ικανοποιήσουμε τις προϋποθέσεις εφαρμογής στατιστικών δοκιμασιών.

Ποιος Μετασχηματισμός;

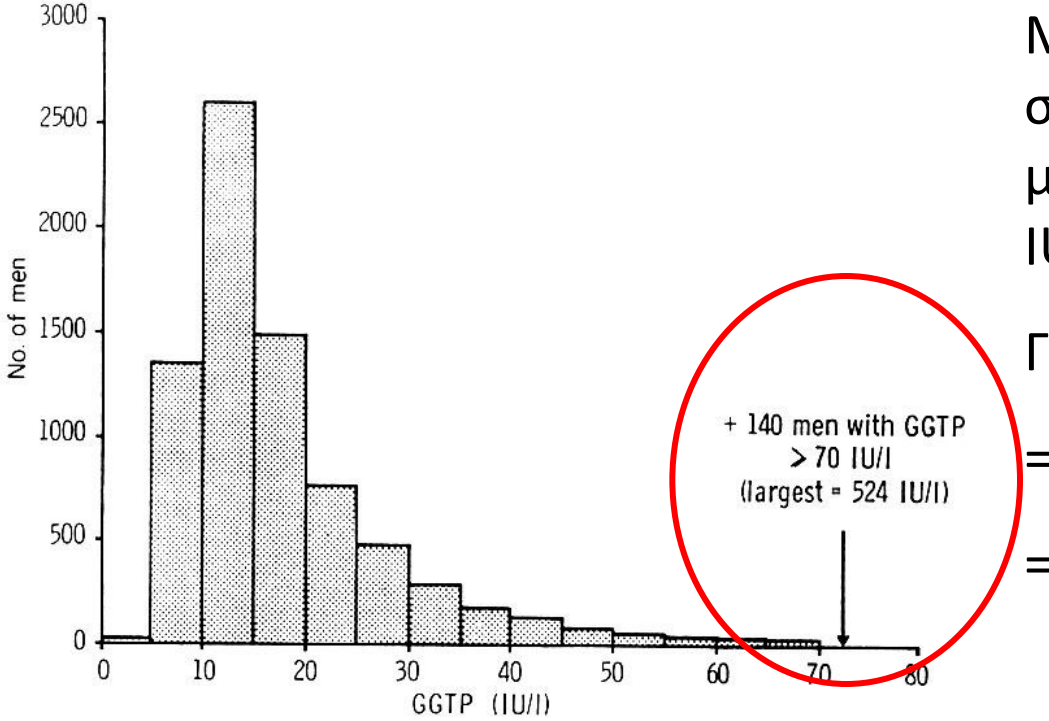
Το λογαριθμικό μετασχηματισμό $z=\log(y)$ για θετικά λοξές κατανομές, (ln, $\log_{10}$)

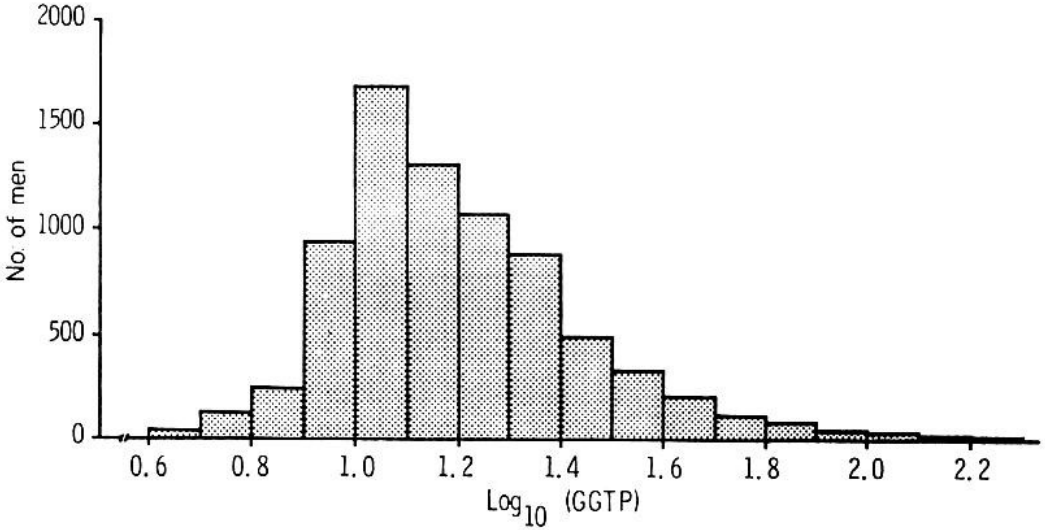
Λογαριθμική συνάρτηση για x>0



Μεταβλητή	x_1	x_2	x_3	x_4	x_5
Αρχικές τιμές	10	100	1.000	10.000	100.000
log10	1	2	3	4	5

Λογαριθμικός μετασχηματισμός





Μέτρηση δείκτη ήπατος σε σχέση με αλκοόλ σε 7.613 άντρες μέσης ηλικίας. Μέση τιμή 19,2 IU/lt

Γεωμετρικός μέσος

=αντιλογάριθμος (mean(log))

= 15.6 IU/lt

Πηγή: Shaper et al. 1983

Φυσιολογικές τιμές

$$\bar{x} \pm 2 * SD$$

Διάκριση φυσιολογικές / παθολογικές

Μεταβλητότητα βιολογικού μεγέθους που δεν οφείλεται σε διαπιστωμένους εξωγενείς ή εργαστηριακούς παράγοντες.

Όχι προσδιοριστικό

πιθανολογικό χαρακτήρα

οποιαδήποτε τιμή μπορεί να ανήκει σε υγιές άτομο, αλλά η **πιθανότητα** είναι **μικρότερη**, όταν η **απόσταση** από μέση τιμή είναι **μεγαλύτερη.**

Για Ασύμμετρες κατανομές

Θέλουμε το 95% του πληθυσμού να ανήκει μέσα στο διάστημα

Εκατοστημόρια (2.5, 97.5)