Περιγραφική Στατιστική
(Descriptive Statistics)

Ζιντζαράς Ηλίας, M.Sc., Ph.D.

Καθηγητής Βιομαθηματικών-Βιομετρίας
Εργαστήριο Βιομαθηματικών
Τμήμα Ιατρικής
Πανεπιστήμιο Θεσσαλίας

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Θεόδωρος Μπρότσης, MSc, PhD
Εντεταλμένος Διδάσκων
(http://biomath.med.uth.gr)
Πανεπιστήμιο Θεσσαλίας
Email: tmprotsis@uth.gr

Βασικοί Στατιστικοί Όροι

Βασικοί Στατιστικοί Όροι

- Στην Ιατρική έρευνα και κλινική πρακτική συλλέγοντας δεδομένα από ένα **δείγμα** ατόμων εξάγουμε συμπεράσματα για έναν **πληθυσμό** στον οποίο ανήκει το δείγμα

- ***Παράδειγμα:*** Αν θέλουμε να ερευνήσουμε τη σχέση μεταξύ της αύξησης του βάρους της εγκύου κατά τη διάρκεια της κύησης και του βάρους του νεογέννητου πρέπει να μελετήσουμε ένα δείγμα από εγκύους. Ποτέ δεν μπορούμε να μελετήσουμε όλες τις εγκύους

Βασικοί Στατιστικοί Όροι

Μερικοί από τους πιο βασικούς όρους στη στατιστική μεθοδολογία είναι οι παρακάτω:

- **Πληθυσμός**: όλο το σύνολο των αντικειμένων που ερευνούμε, π.χ. όλες οι έγκυοι
- **Δείγμα**: κάθε υποσύνολο του πληθυσμού που ερευνάται, π.χ. 30 έγκυοι έχουν συλλεχθεί τυχαία από ένα νοσοκομείο.
- **Μεταβλητή** (συμβολίζεται με Χ ή x): κάθε χαρακτηριστικό που μπορεί να μετρηθεί ή να παρατηρηθεί, π.χ. αύξηση βάρους σώματος.
- **Παρατήρηση**: η τιμή της μεταβλητής για ένα συγκεκριμένο αντικείμενο που μελετάται, π.χ. η αύξηση βάρους της 12ης εγκύου κατά τη διάρκεια της κύησης ήταν 7Kg.

Μηδενική υπόθεση

- **Μηδενική υπόθεση (H_0):** Η γενική ιδέα της διαδικασίας στατιστικού ελέγχου υποθέσεων είναι η εξής:
  - θέτουμε ως μηδενική υπόθεση (H_0) αυτή για την οποία αμφιβάλουμε, αυτή που αμφισβητείται, και
  - εξετάζουμε αν ένα τυχαίο δείγμα που παίρνουμε από τον πληθυσμό συνηγορεί-δίνει αποδείξεις υπέρ της απόρριψής της, έναντι της εναλλακτικής (H_1)

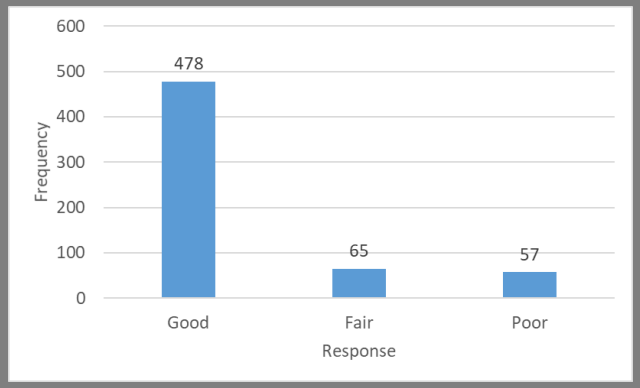
- Δηλαδή, η H_0, απορρίπτεται ή δεν απορρίπτεται με βάση το τι **παρατηρείται** στο τυχαίο δείγμα που πήραμε από τον πληθυσμό
- Πιο συγκεκριμένα, υποθέτοντας ότι η H_0 είναι αληθής, αν αυτό που παρατηρείται στο δείγμα είναι ακραίο, δηλαδή, αν έχει πολύ μικρή πιθανότητα να συμβεί, τότε απορρίπτου με την H_0

Είδη μεταβλητών

Μια **μεταβλητή** είναι **ποιοτική** όταν παίρνει ορισμένες διακριτές τιμές ή **ποσοτική** όταν παίρνει τιμές σε μία συνεχή κλίμακα

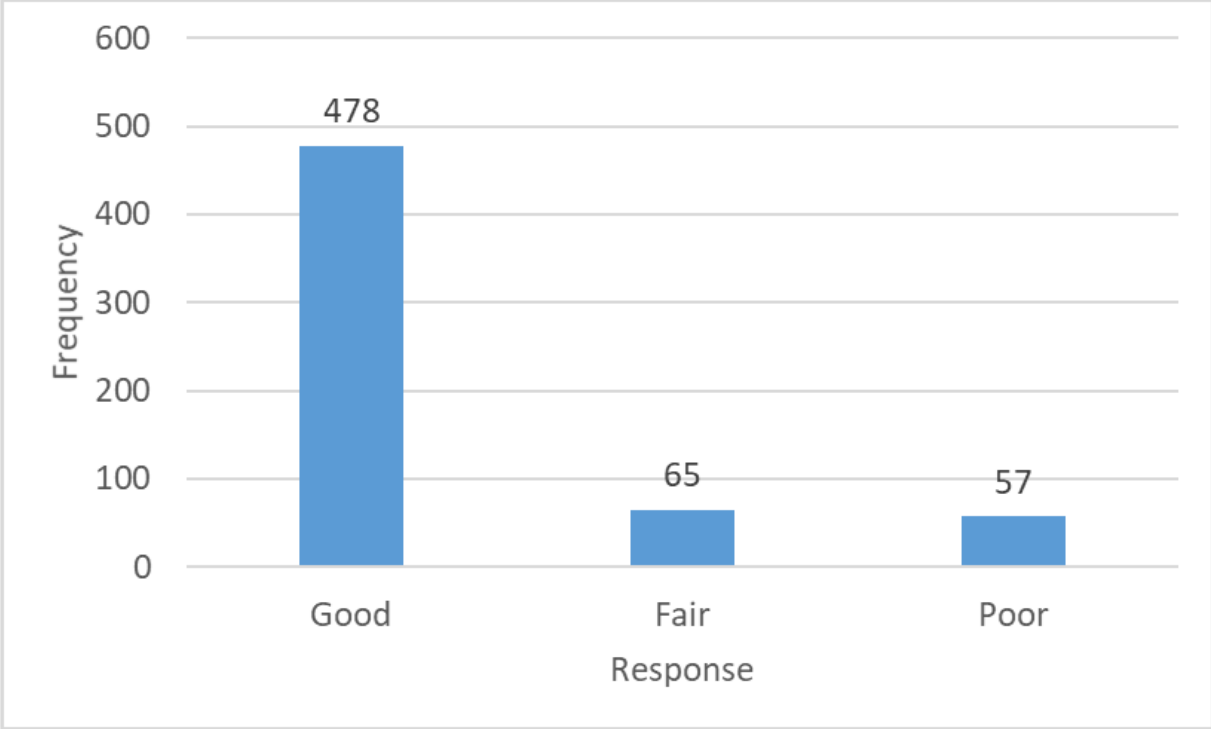
- Ποιοτικές μεταβλητές είναι:
  - το φύλο (άνδρας/γυναίκα),
  - η σταθεροποίηση αιμοσφαιρίνης σε νεφροπαθείς (ναι/όχι),
  - η επιβίωση (επιβιώνει/πεθαίνει)
  - το αποτέλεσμα θεραπείας (ναι/όχι)
  - Ποσότητα λήψης φαρμάκου (μικρή/μεσαία/μεγάλη)
- Ποσοτικές μεταβλητές (αριθμητικές) είναι:
  - το ύψος, το βάρος, η πίεση αίματος, η ηλικία κ.α.

Γραφικές Μέθοδοι Περιγραφής Δεδομένων



Περιγραφική Στατιστική

Είναι σημαντικό να αρχίζουμε πάντα την ανάλυσή μας εξετάζοντας τα δεδομένα μας με γραφικές μεθόδους

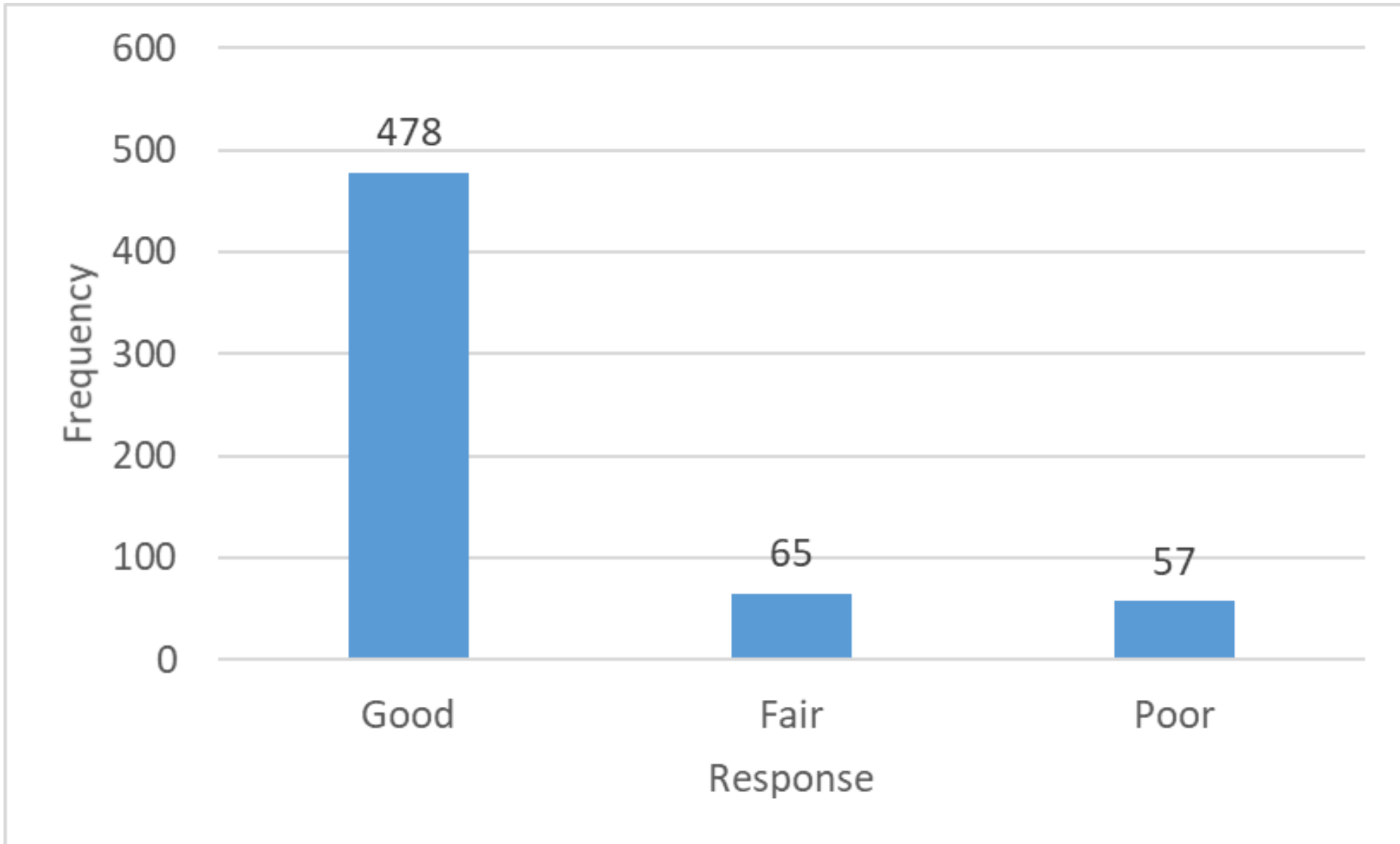


Ραβδογράμματα (bar charts)

Για να δείξουμε τη συχνότητα μιας **ποιοτικής** μεταβλητής χρησιμοποιούμε το **ραβδόγραμμα**.

Παράδειγμα*:* Σε μία κλινική μελέτη συμμετείχαν 600 ασθενείς και η ανταπόκριση τους στην θεραπεία ήταν: **good, fair** και **poor**. Η ποιοτική μεταβλητή που ερευνάται είναι η ανταπόκριση (response) στη θεραπεία

Response (ανταπόκριση)	N
Good	478
Fair	65
Poor	57
Total	**600**



Το ύψος κάθε στήλης είναι ίσο με την αντίστοιχη συχνότητα

Σχετικές και Ποσοστιαίες Συχνότητες

- Ο προηγούμενος πίνακας συχνοτήτων μας δίνει κάποιες πληροφορίες, όπως για παράδειγμα, ότι η τιμή **Good** έχει συχνότητα 478 (δηλαδή 478 ασθενείς είχαν καλή ανταπόκριση στη θεραπεία)
- Η συχνότητα όμως αυτή (δηλαδή ο αριθμός 478) δεν έχει καμιά αξία μόνη της, αν δεν αναφερθεί ο αριθμός των ασθενών που συμμετείχαν στη θεραπεία
- Για να βρούμε τη **σχετική συχνότητα** μιας τιμής, **διαιρούμε τη συχνότητα της τιμής αυτής με το πλήθος όλων των παρατηρήσεων**
- Στη συνέχεια, μπορούμε να εκφράσουμε τον αριθμό αυτό ως ποσοστό επί τοις εκατό (%)

Σχετικές και Ποσοστιαίες Συχνότητες

$$\text{Σχετική συχνότητα} = \frac{\text{Συχνότητα τιμής}}{n}$$

$$\text{Ποσοστιαία συχνότητα} = \frac{\text{Συχνότητα τιμής}}{n} \cdot 100$$

Σχετικές και Ποσοστιαίες Συχνότητες

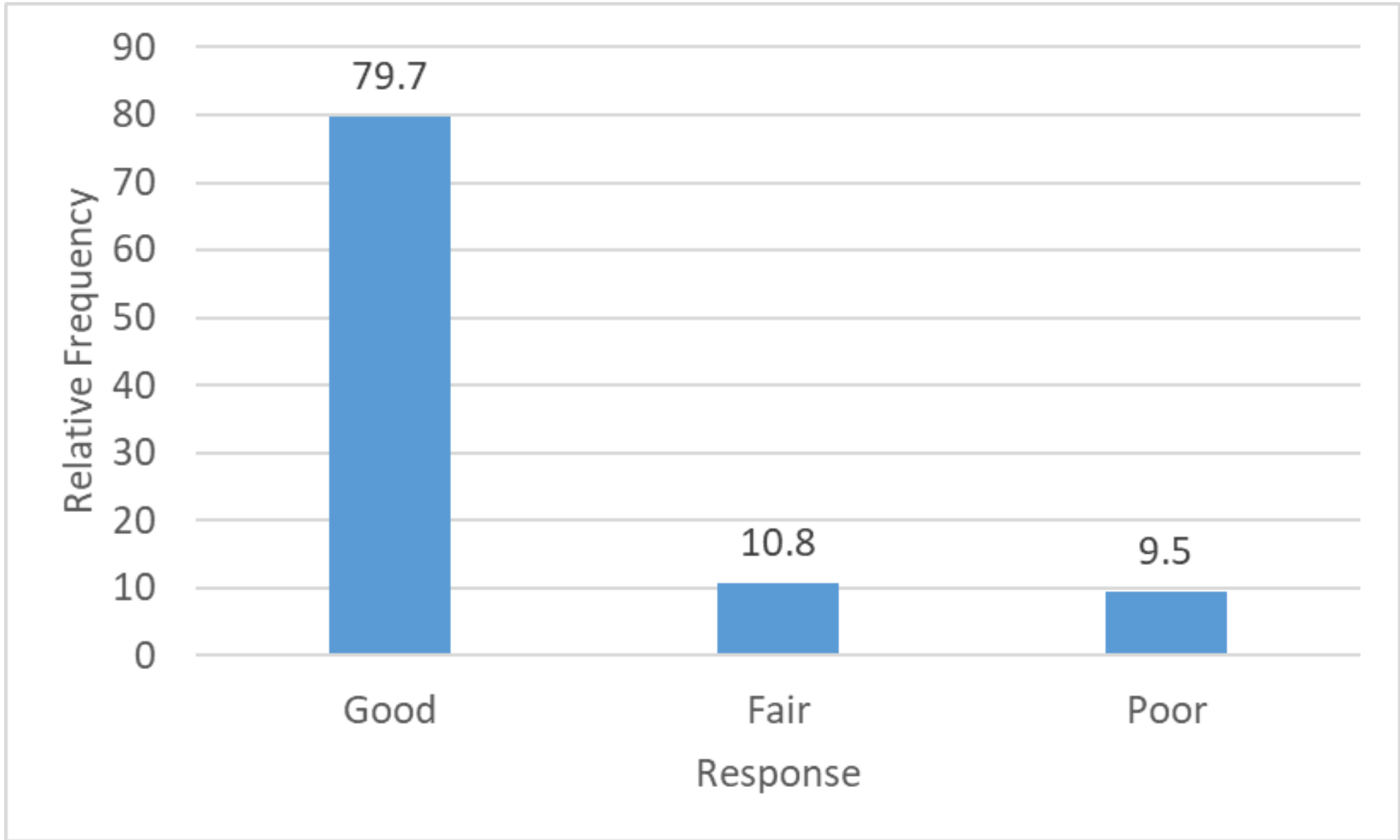
$$\text{Σχετική συχνότητα “Good”} = \frac{478}{600}$$

$$\text{Σχετική συχνότητα “Good”} = 0.797$$

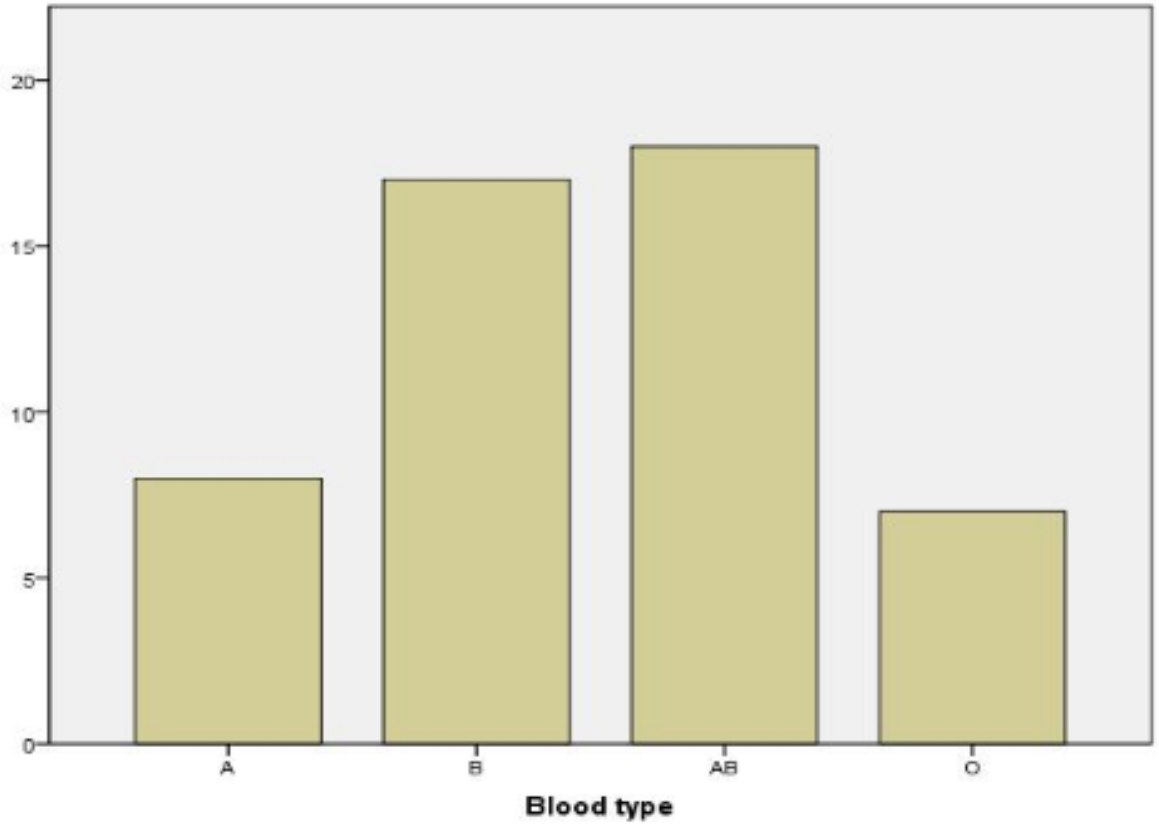
$$\text{Ποσοστιαία συχνότητα “Good”} = 0.797 \cdot 100$$

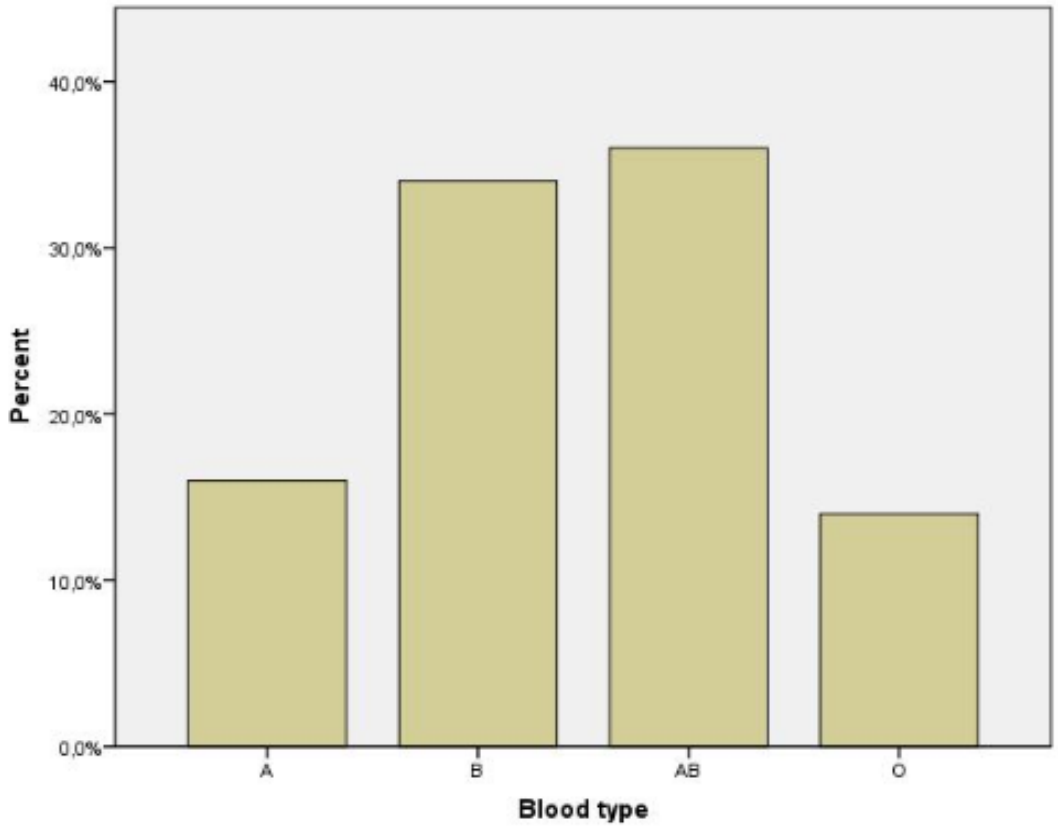
$$\text{Ποσοστιαία συχνότητα “Good”} = 79.7\%$$

Response	N	Σχετικές συχνότητες (%)
Good	478	79.70%
Fair	65	10.80%
Poor	57	9.50%
Total	**600**	**100%**



Το ύψος κάθε στήλης είναι ίσο με την αντίστοιχη σχετική συχνότητα

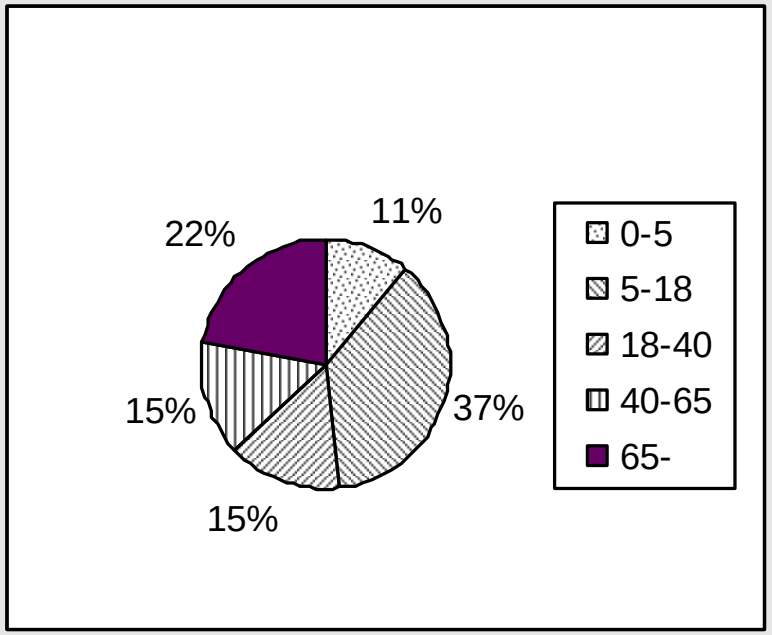

20
15
10
5
0
A
B
AB
O
Blood type

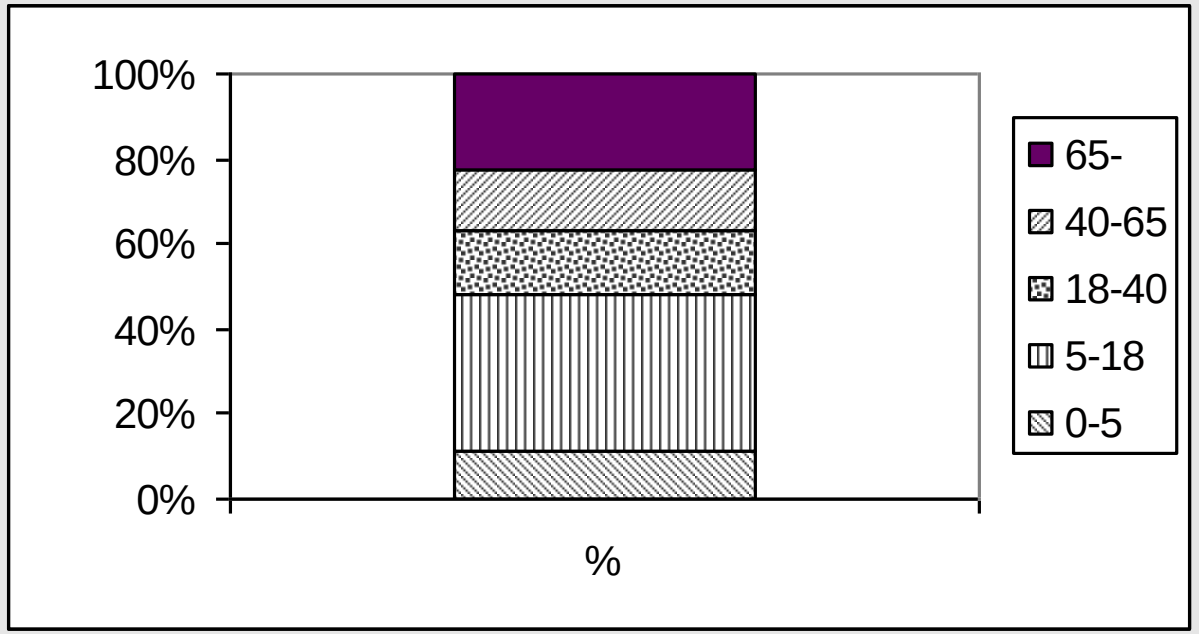

40,0%
30,0%
20,0%
10,0%
0,0%
Percent
A
B
AB
O
Blood type

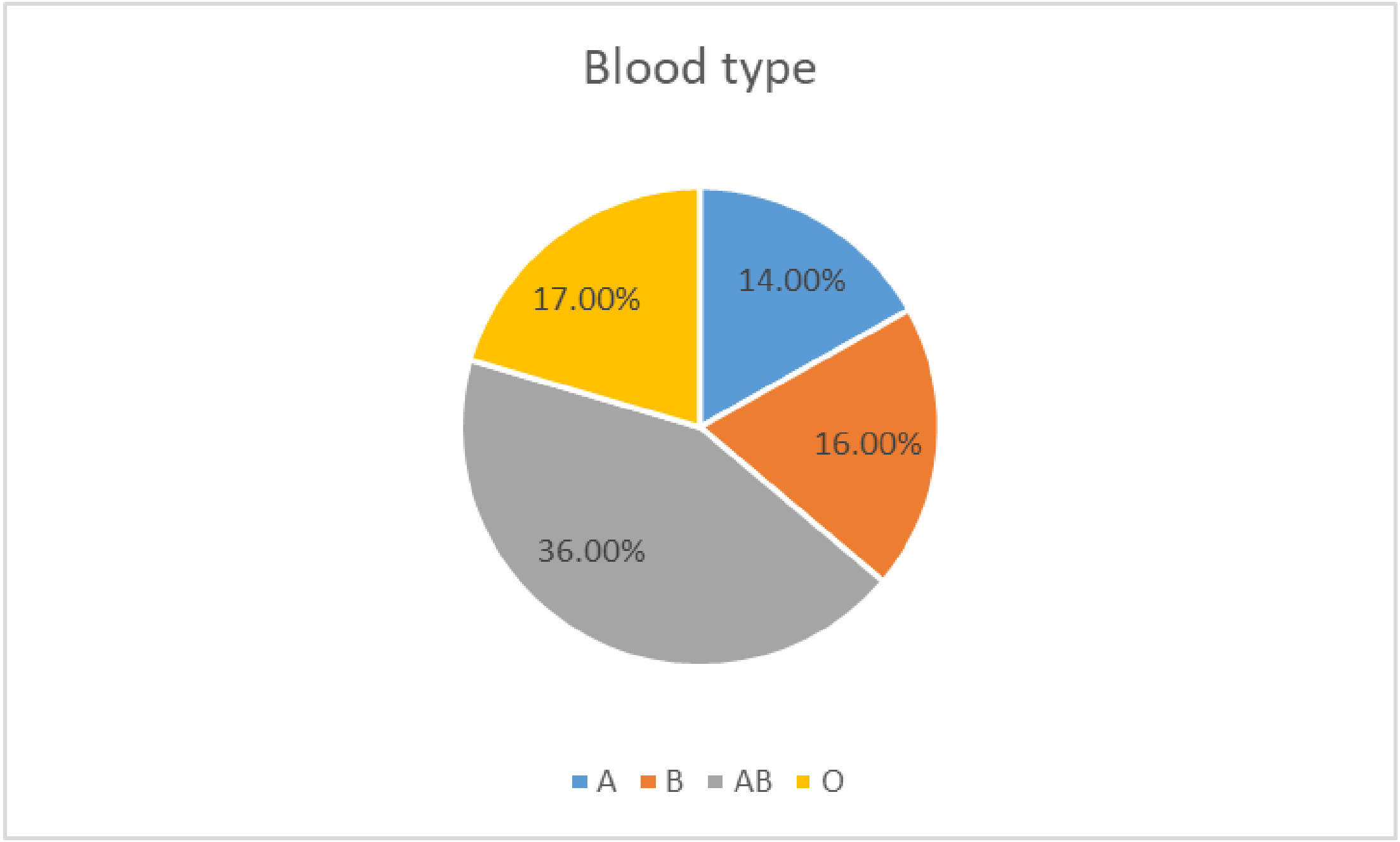
Διάγραμμα Πίτας ή Ράβδου

Η πίτα είναι ένας κύκλος που είναι χωρισμένος σε κομμάτια και κάθε κομμάτι αντιπροσωπεύει τιμές μιας μεταβλητής για διάφορες κατηγορίες

Παράδειγμα: Η παρακάτω πίτα δείχνει τη κατανομή της ηλικίας στον πληθυσμό μιας πόλης






Blood type
14.00%
16.00%
36.00%
17.00%
A
B
AB
O

Ιστογράμματα

- Όταν η μεταβλητή που μελετάμε είναι **ποσοτική** τότε κατασκευάζουμε μία **κατανομή των συχνοτήτων** που περιγράφεται με ένα **ιστόγραμμα**
- Αν οι τιμές είναι πολλές τότε τις ομαδοποιούμε σε 5-8 ομάδες (bins)
- **Οριζόντια** (x-axis) βρίσκεται η **μεταβλητή ενδιαφέροντος** π. χ. αιμοσφαιρίνη, η ηλικία κ.λπ.
- Στον **κάθετο άξονα** βάζουμε τις **απλές συχνότητες**, τις **σχετικές συχνότητες** ή τις **ποσοστιαίες συχνότητες**

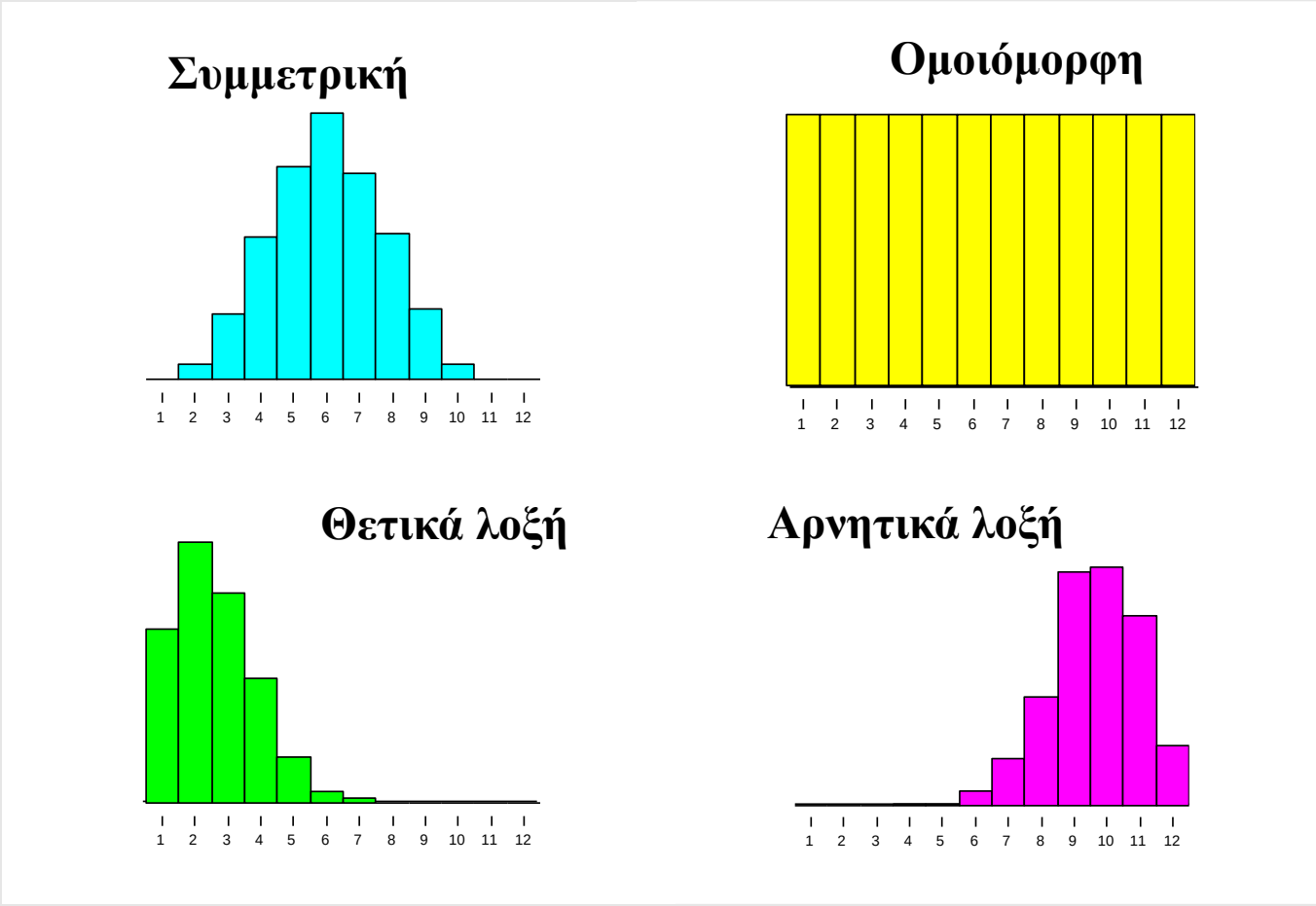
Ιστογράμματα

- Αποτελεί την **ποιο χρήσιμη** γραφική απεικόνιση **ποσοτικών δεδομένων**
- Δείχνει το **σχήμα** της κατανομής
- Κάθετο ορθογώνιο για κάθε ομάδα (bin)
- Το **ύψος** καθορίζεται από τις απλές συχνότητες, τις σχετικές συχνότητες ή τις ποσοστιαίες συχνότητες του αριθμού των παρατηρήσεων σε αυτήν την ομάδα
- **Δεν υπάρχουν κενά** μεταξύ των ράβδων

Κοιλότητα ιστογραμμάτων

Το ιστόγραμμα υποδεικνύει την **κοιλότητα** ή την **συμμετρία** των παρατηρήσεων

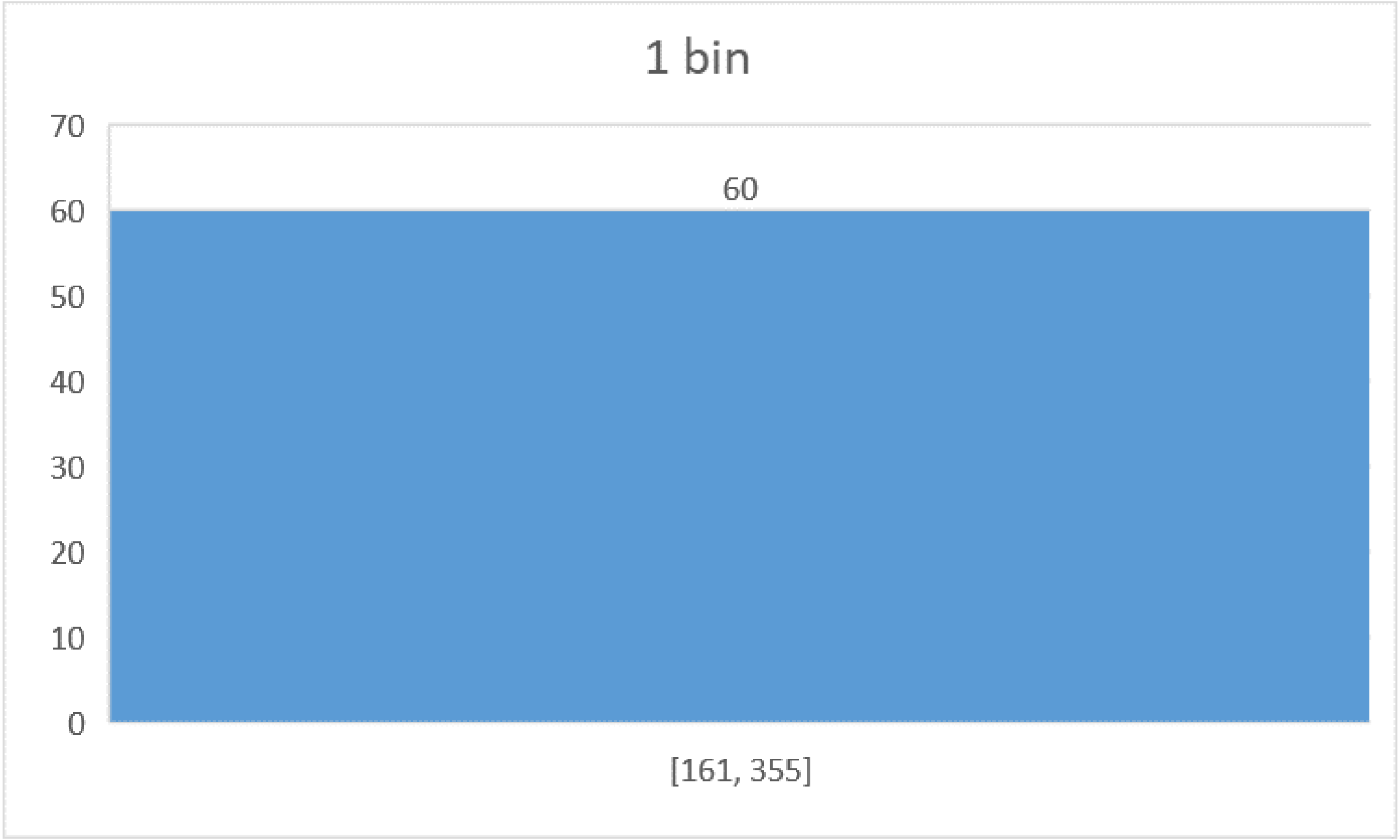
Παράδειγμα*:* Η κατανομή των τιμών της συστολικής αρτηριακής πίεσης ατόμων προχωρημένης ηλικίας είναι θετικά λοξή ενώ η κατανομή της αιμοσφαιρίνης Hgb των 20 γυναικών είναι συμμετρική

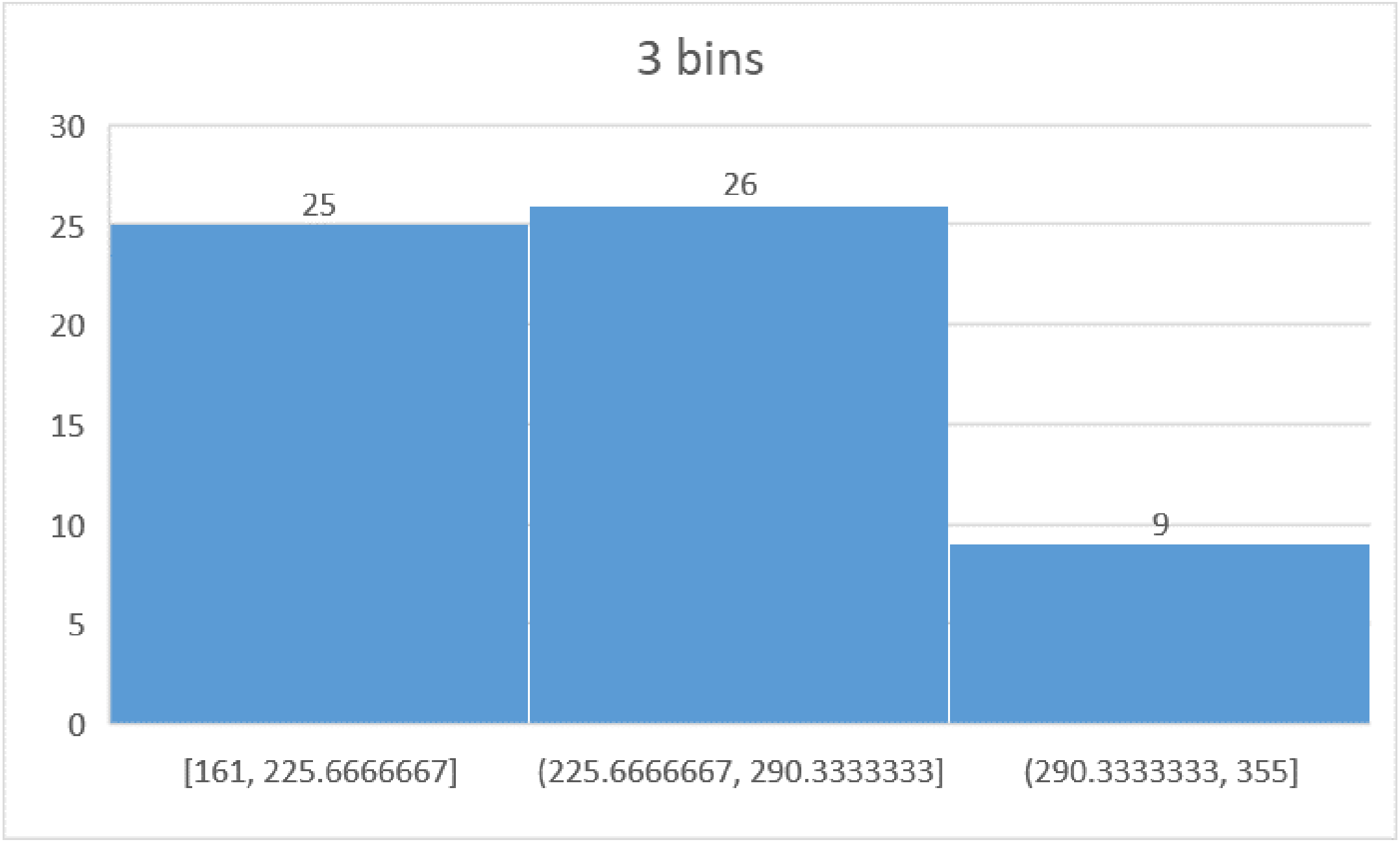


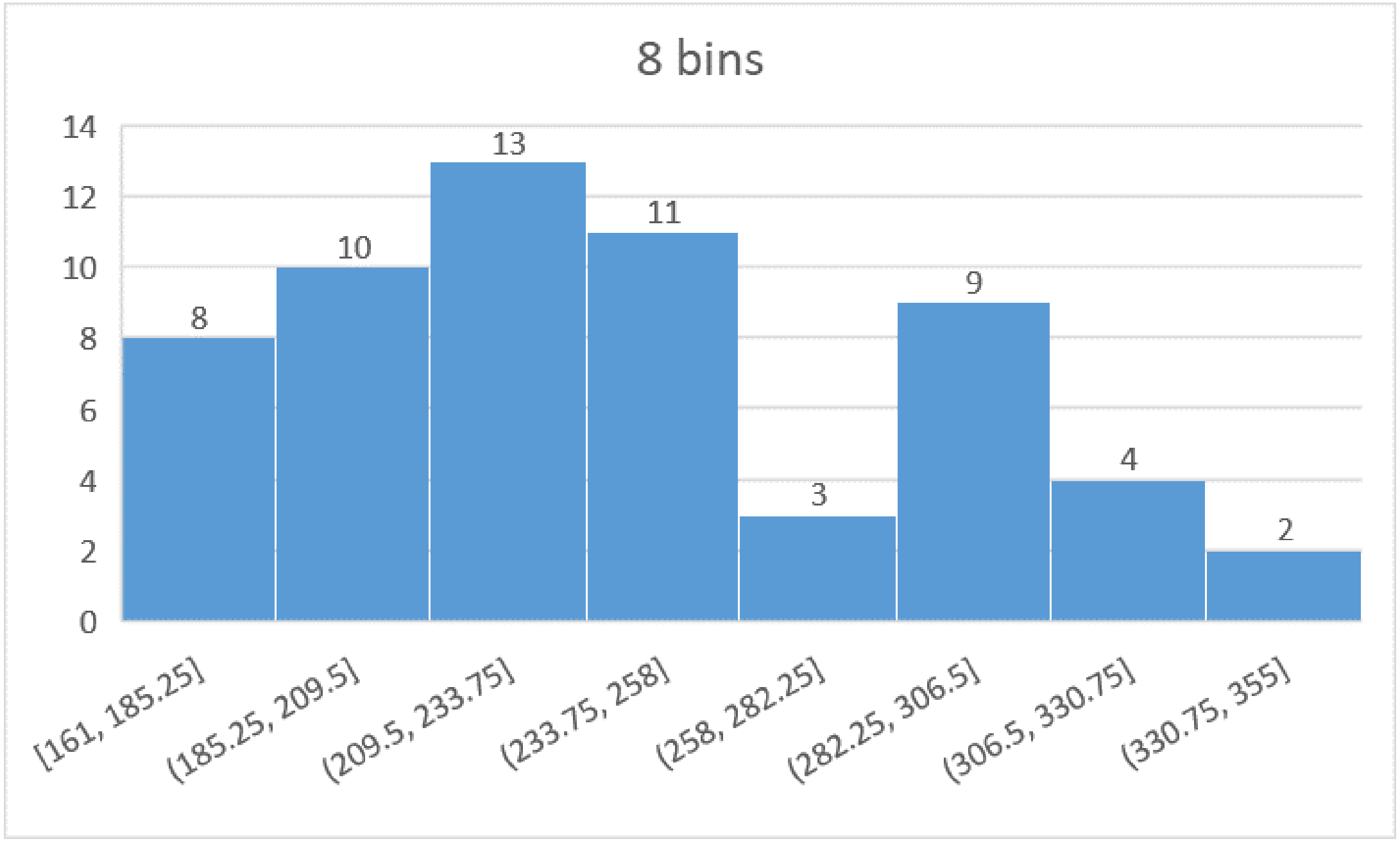
Πως να δημιουργήσουμε ένα ιστόγραμμα

Παράδειγμα: Μετρήθηκαν τα επίπεδα χοληστερόλης 60 ατόμων που έλαβαν μέρος σε μία κλινική δοκιμή.

212	249	227	218	310	281	330	226
233	223	161	195	233	249	284	284
174	170	256	169	299	210	301	199
258	258	195	227	244	355	234	195
196	354	282	282	286	286	176	195
163	297	211	228	309	309	225	223
195	248	284	173	256	169	209	209
200	258	284	239				


1 bin
70
60
50
40
30
20
10
0
60
[161, 355]


3 bins
30
25
20
15
10
5
0
25
26
9
[161, 225.6666667]
(225.6666667, 290.3333333]
(290.3333333, 355]


8 bins
14
12
10
8
6
4
2
0
8
10
13
11
3
9
4
2
[161, 185.25]
(185.25, 209.5]
(209.5, 233.75]
(233.75, 258]
(258, 282.25]
(282.25, 306.5]
(306.5, 330.75]
(330.75, 355]

Ομαδοποίηση δεδομένων σε 8 διαστήματα τάξης

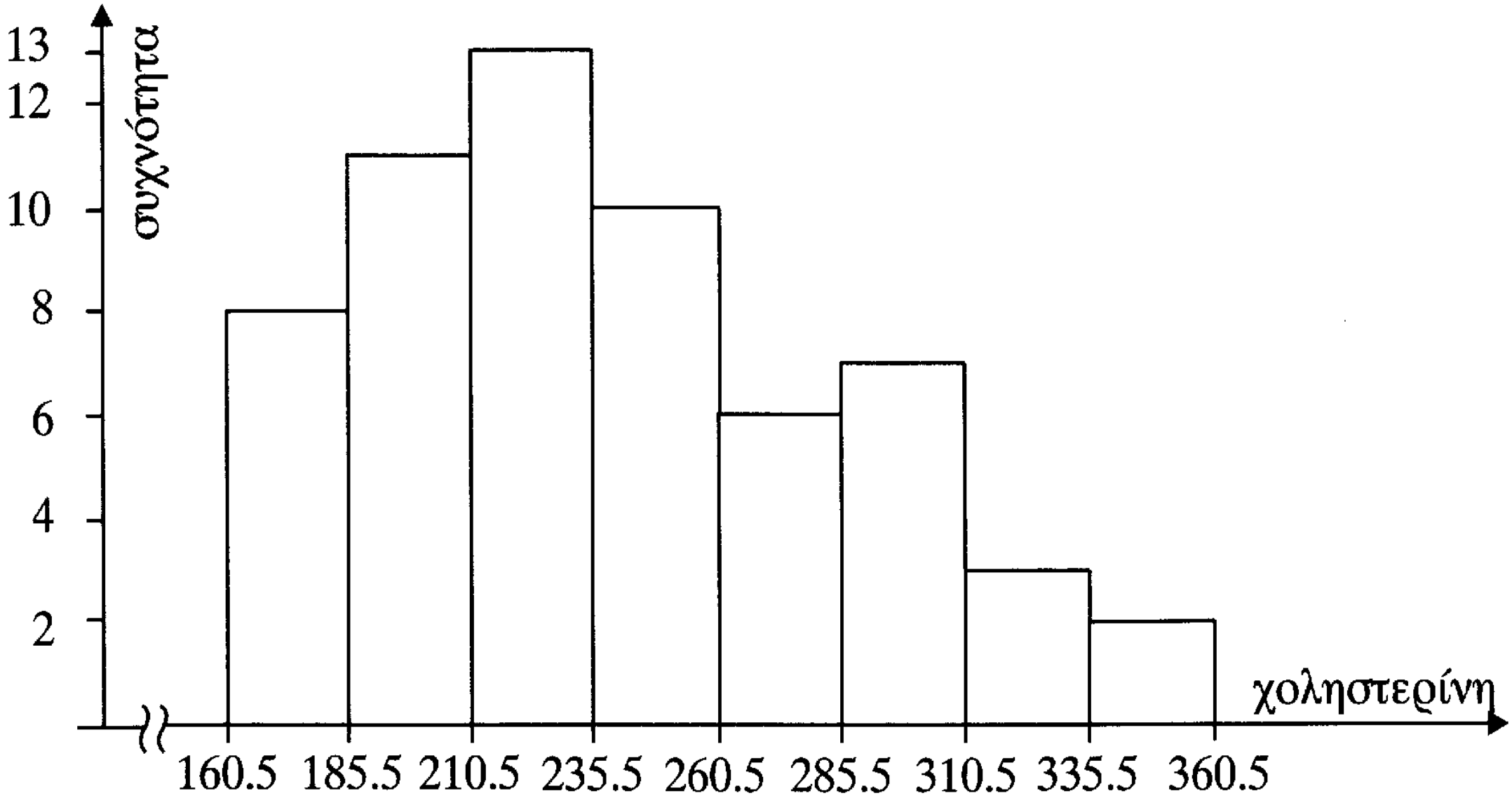
- Επιλογή του αριθμού διαστημάτων τάξης, $\boldsymbol{k = 8}$
- Προσδιορισμός του πλάτους των διαστημάτων τάξης

$$\boldsymbol{c = \frac{R}{k} = \frac{(355-161)}{8} = \frac{194}{8} = 24.25 \cong 25.}$$

- Καθορισμός των διαστημάτων τάξης

Μικρότερη τιμή $x_{min} = 161$
Μεγαλύτερη τιμή $x_{max} = 355$

Διαστήματα τάξης	Κεντρική τιμή	Συχνότητα τάξης	Σχετική Συχνότητα (%)
[160.5—185.5)	173	8	13.33
[185.5—210.5)	198	11	18.33
[210.5—235.5)	223	13	21.67
[235.5—260.5)	248	10	16.67
[260.5—285.5)	273	6	10.00
[285.5—310.5)	298	7	11.67
[310.5—335.5)	323	3	5.00
[335.5—360.5)	348	2	3.33
Σύνολο		**60**	**100**

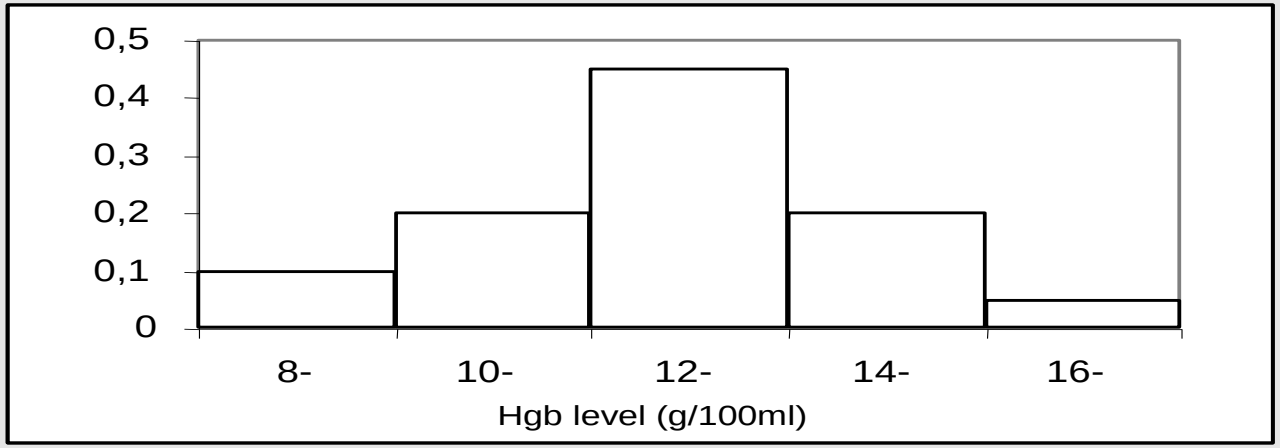

συχνότητα
13
12
10
8
6
4
2
χοληστερίνη
160.5
185.5
210.5
235.5
260.5
285.5
310.5
335.5
360.5

Άλλο παράδειγμα Ιστογράμματος

Παράδειγμα: Τα επίπεδα αιμοσφαιρίνης (g/100ml) 20 γυναικών έχουν μετρηθεί και είναι:

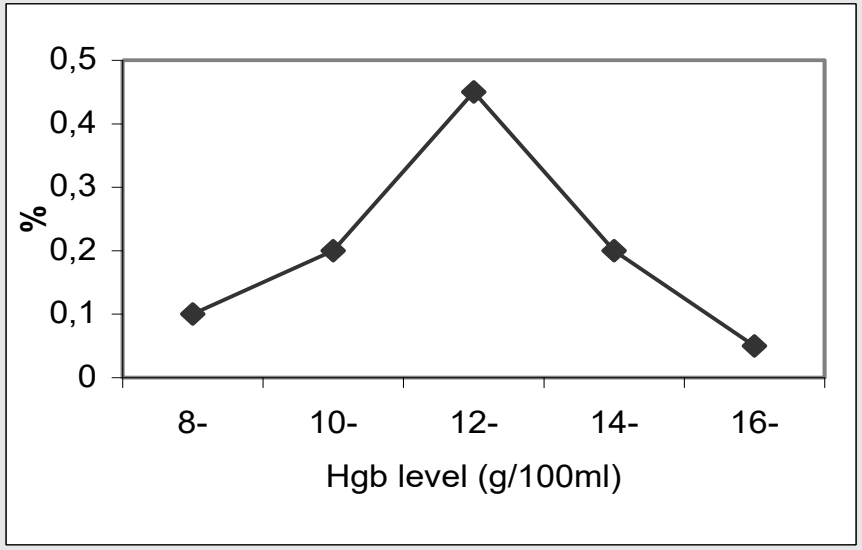
Hgb levels	
8,8	12,9
9,3	12,9
10,5	12,9
10,6	13,3
11,1	13,4
11,4	14,5
12	14,6
12	14,6
12,1	15,1
12,1	16,1

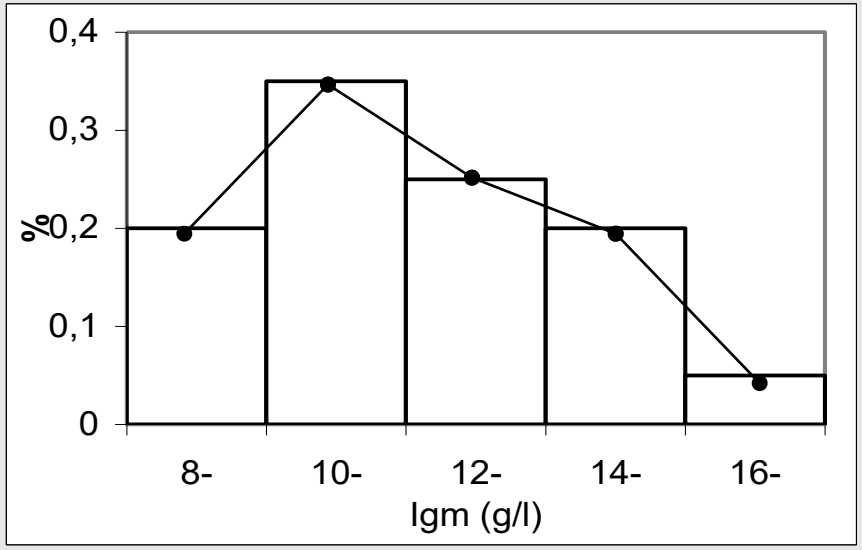
Hgb	Frequency	Proportion
8-	2	0,1
10-	4	0,2
12-	9	0,45
14-	4	0,2
16-	1	0,05
Σύνολο	20	



Καμπύλη Συχνοτήτων

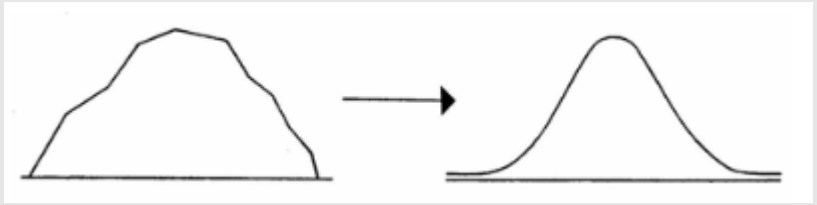
Το **ιστόγραμμα** για πρακτικούς λόγους παριστάνεται με μία καμπύλη που κατασκευάζεται αν ενώσουμε τα διαδοχικά **μέσα** των πάνω βάσεων των ορθογωνίων του ιστογράμματος (**καμπύλη συχνοτήτων**)





Κανονική Κατανομή

Αυξάνοντας το μέγεθος του δείγματος και κατασκευάζοντας το ιστόγραμμα με ολοένα και μικρότερου πλάτους κλάσεις, το αντίστοιχο πολύγωνο προσεγγίζει μια ομαλή-λεία καμπύλη.



- Η κανονική καμπύλη έχει **κωδωνοειδή μορφή**, είναι **συμμετρική** και οι «ουρές» της πλησιάζουν τον οριζόντιο άξονα ομαλά. Η μέση τιμή και η διάμεσος ταυτίζονται
- Η περιοχή που παρουσιάζει τη μεγαλύτερη πυκνότητα, βρίσκεται και αυτή στο μέσο της κατανομής. Δηλαδή, όταν οι τιμές μιας μεταβλητής είναι κανονικά κατανεμημένες, τότε γύρω από τη μέση τιμή τους υπάρχουν σχετικά πολλές τιμές ενώ μακριά από τη μέση τιμή βρίσκονται σχετικά λίγες τιμές

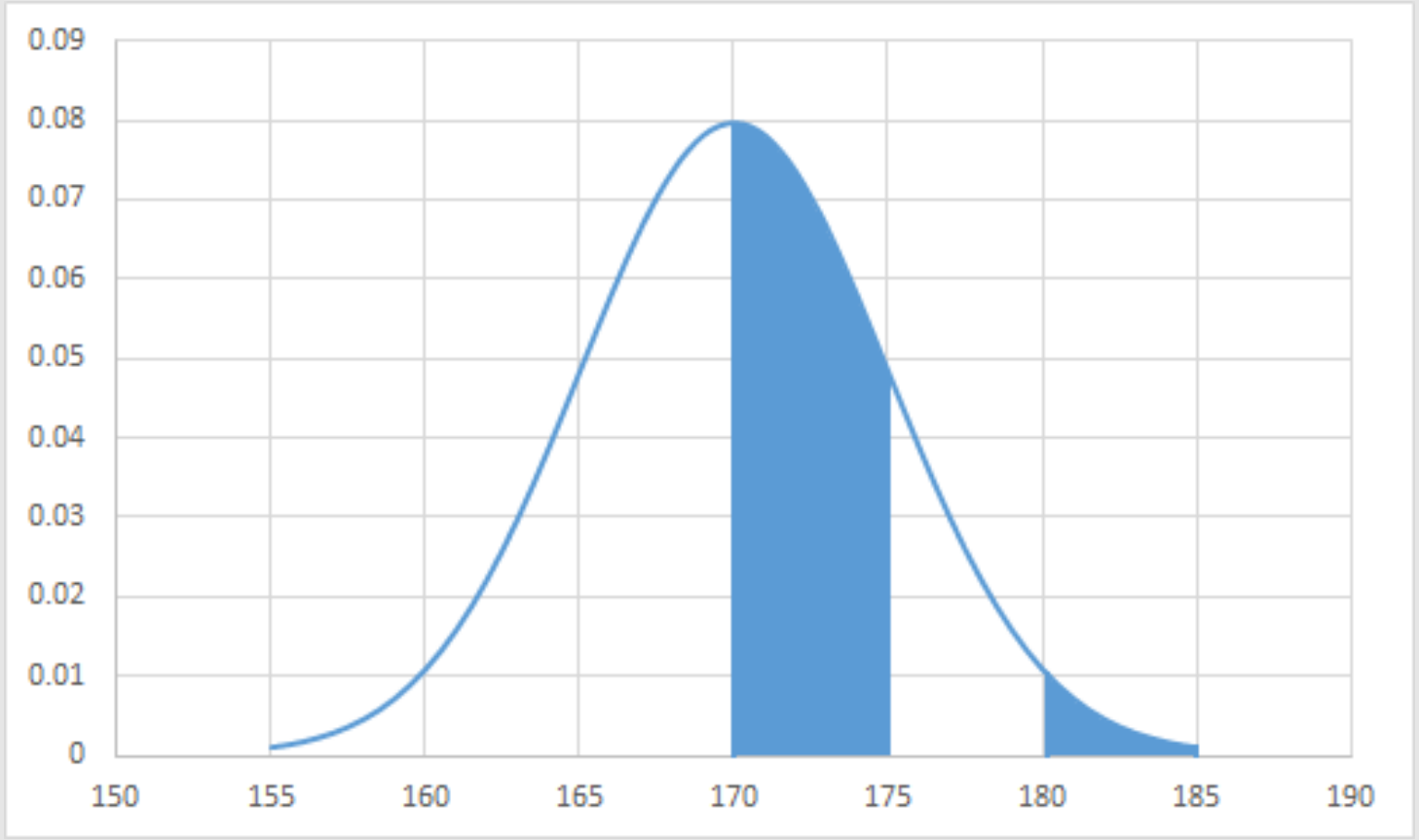
Γιατί έχουν σημασία οι κανονικές κατανομές;

- Όλα τα είδη των μεταβλητών που συναντάμε στις φυσικές και κοινωνικές επιστήμες **κατανέμονται κανονικά ή περίπου κανονικά**. Μερικά παραδείγματα αυτών των μεταβλητών είναι το ύψος, το βάρος γέννησης και η ικανοποίηση από την εργασία
- Επειδή οι **κανονικά κατανεμημένες μεταβλητές** είναι τόσο συχνές, πολλές στατιστικές δοκιμές έχουν σχεδιαστεί για κανονικά κατανεμημένους πληθυσμούς
- Η κατανόηση των ιδιοτήτων των κανονικών κατανομών σημαίνει ότι μπορείτε να χρησιμοποιήσετε επαγωγική στατιστική για να συγκρίνετε διαφορετικές ομάδες και να κάνετε εκτιμήσεις για πληθυσμούς χρησιμοποιώντας δείγματα

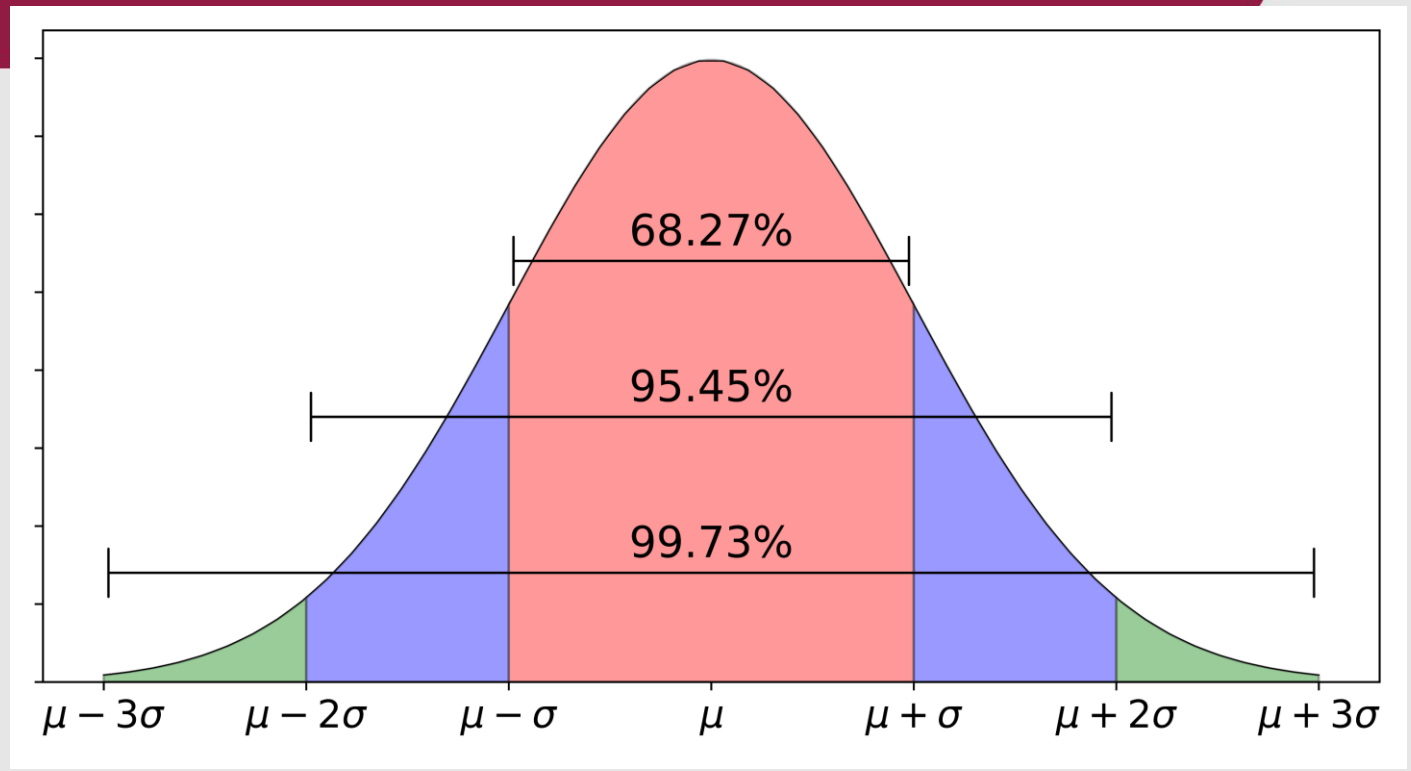
Κανονική Κατανομή

Παράδειγμα:

- **Ύψος των Ελλήνων**, ηλικίας 18 έως 25 ετών
- **Κανονικά κατανεμημένο**
- **Μέση τιμή:** 170 cm
- **Τυπική απόκλιση:** 5 cm
- Τότε, μεταξύ 170 cm και 175 cm βρίσκονται περισσότερα άτομα από όσα βρίσκονται μεταξύ 180 cm και 185 cm.
- Επίσης, πολύ λίγα άτομα έχουν ύψος μεγαλύτερο από 185 cm ή μικρότερο από 155 cm.



Κανονική Κατανομή



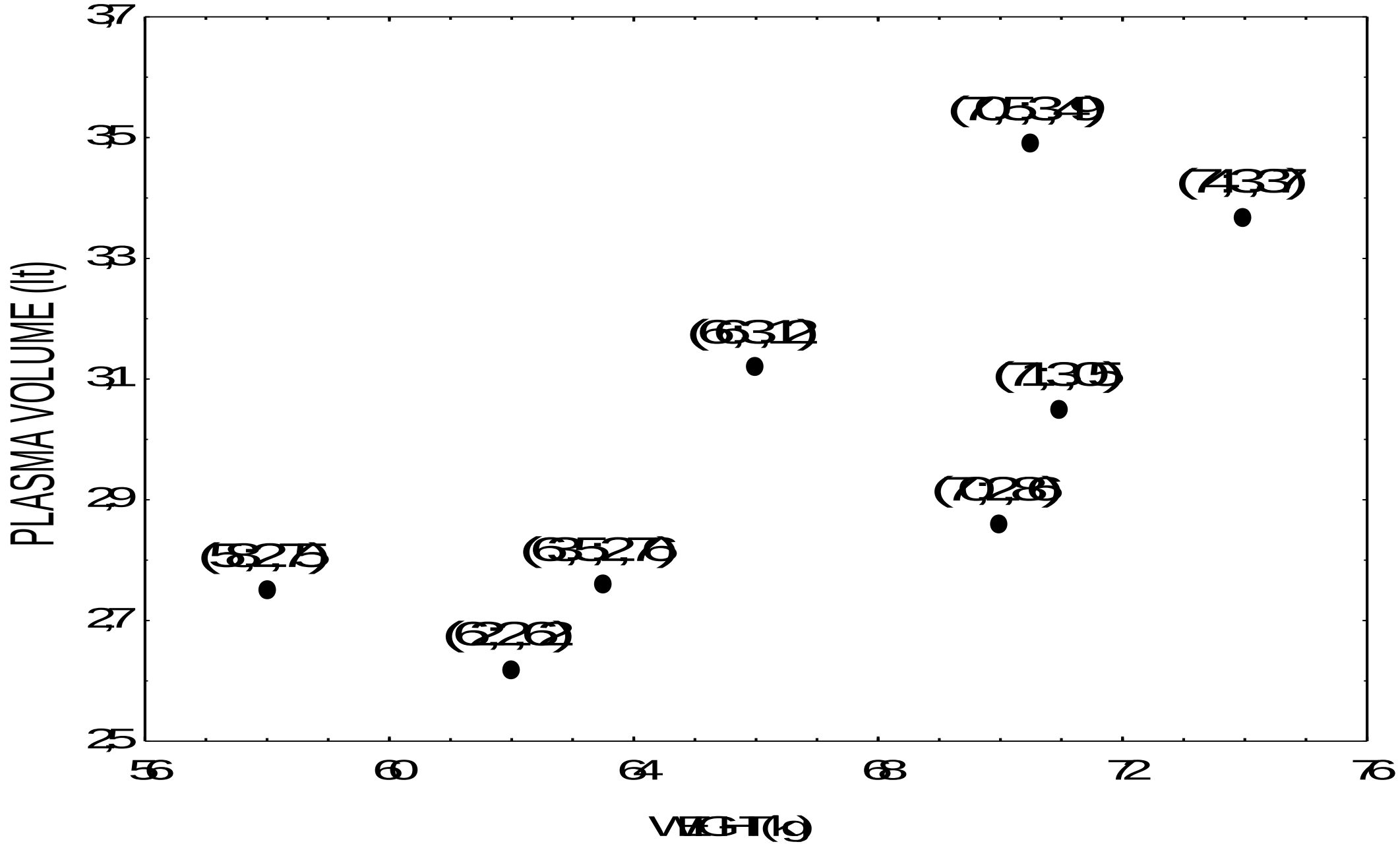
- Στο διάστημα από $(\mu - \sigma)$ μέχρι $(\mu + \sigma)$, όπου μ=μέση τιμή και σ=τυπική απόκλιση, περιλαμβάνεται περίπου το 68% των παρατηρήσεων,
- ενώ στο διάστημα από $(\mu - 2\sigma)$ μέχρι $(\mu + 2\sigma)$ περιλαμβάνεται περίπου το 95% των παρατηρήσεων

Διάγραμμα Συσχέτισης

Όταν υπάρχουν παρατηρήσεις από **δύο ποσοτικές μεταβλητές** και μας ενδιαφέρει η **σχέση** που έχουν μεταξύ τους τότε τα δεδομένα παρουσιάζονται με ένα διάγραμμα συσχέτισης.

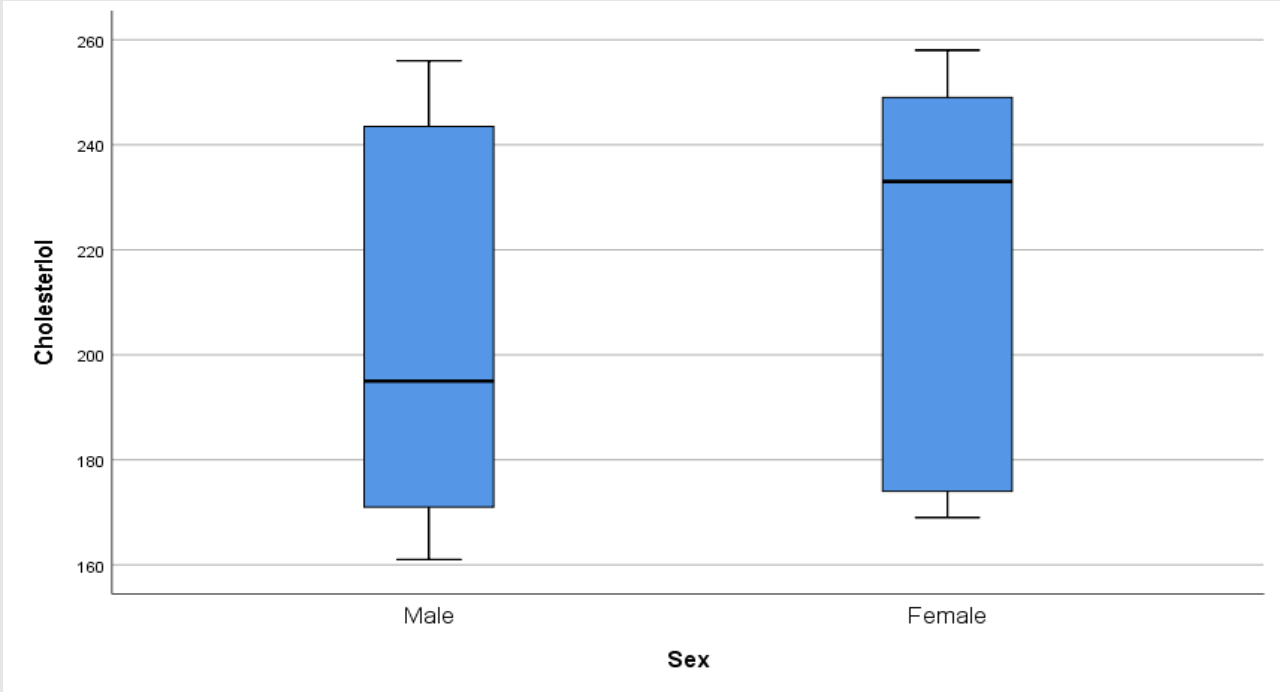
Παράδειγμα: Το βάρος του σώματος και ο όγκος πλάσματος από 8 υγιείς άνδρες είναι:

Άτομο	βάρος σε Kg (x)	Όγκος πλάσματος σε lt (y)
1	58	2.75
2	70	2.86
3	74	3.37
4	63.5	2.76
5	62	2.62
6	70.5	3.49
7	71	3.05
8	66	3.12


PLASMA VOLUME (lt)
WEIGHT (kg)
37
35
33
31
29
27
25
56
60
64
68
72
76
(70,5;34)
(74;33)
(66;31)
(71;30)
(70,2;26)
(58;27,5)
(63,5;27,6)
(62;26)

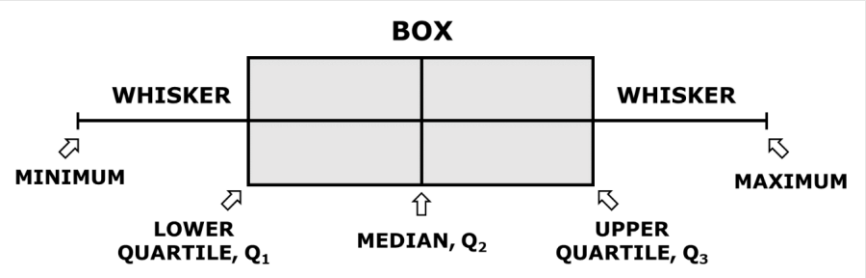
Θηκόγραμμα (Box-plot)

- Αποτελεί εύκολο τρόπο της γραφικής απεικόνισης του **σχήματος** των δεδομένων
- **Εύκολο στην ερμηνεία**
- Μπορούμε να δούμε εάν τα δεδομένα **τραβιούνται** προς μία κατεύθυνση (**skewed**)
- Αποτελεί εύκολο τρόπο εύρεσης **ακραίων τιμών**
- Κάνει εύκολη τη **σύγκριση** χαρακτηριστικών στα δεδομένα **μεταξύ κατηγοριών**

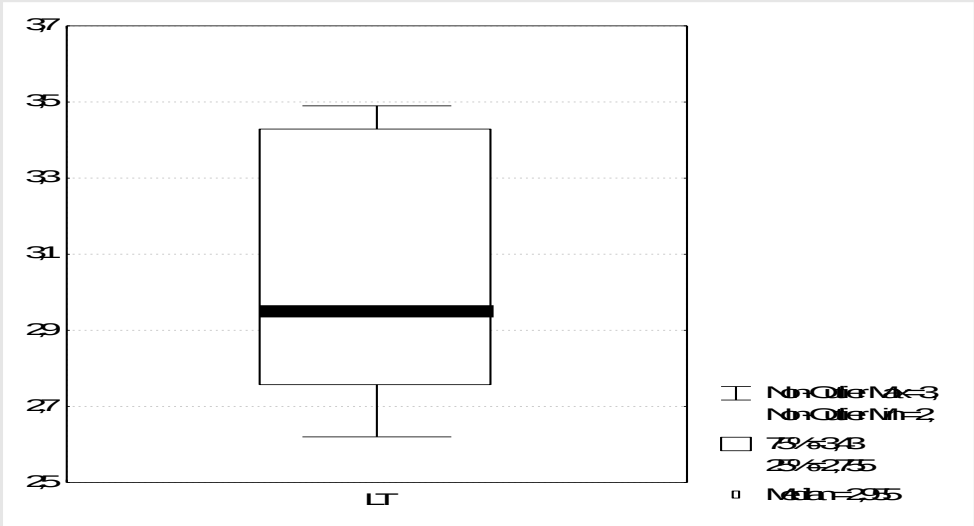


Θηκόγραμμα (Box-plot)

- Το διάγραμμα κουτιού παρουσιάζει τα δεδομένα σαν ένα **κουτί** με ένα **«μουστάκι»**
- Το πάνω μέρος του κουτιού δείχνει το **75ο εκατοστημόριο** και το κάτω μέρος του κουτιού το **25ο εκατοστημόριο**
- Η **διάμεσος** δίνεται με μία οριζόντια γραμμή στο κουτί
- Οι άκρες του μουστακιού δείχνουν την μέγιστη και την ελάχιστη τιμή

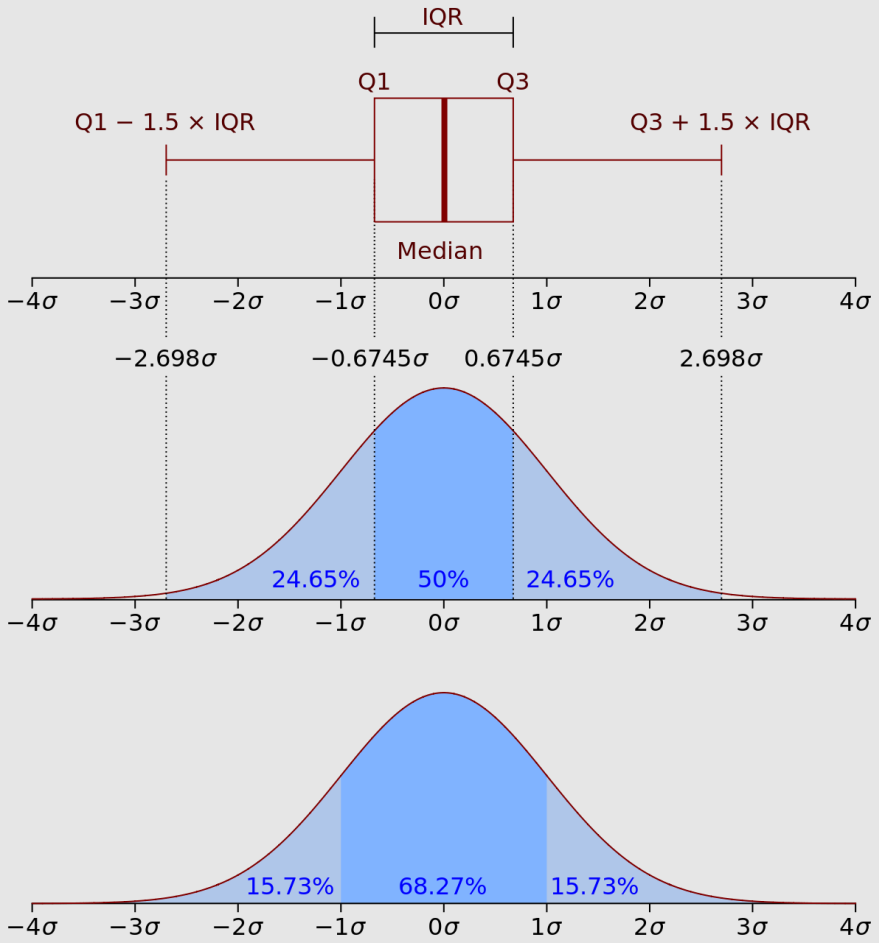


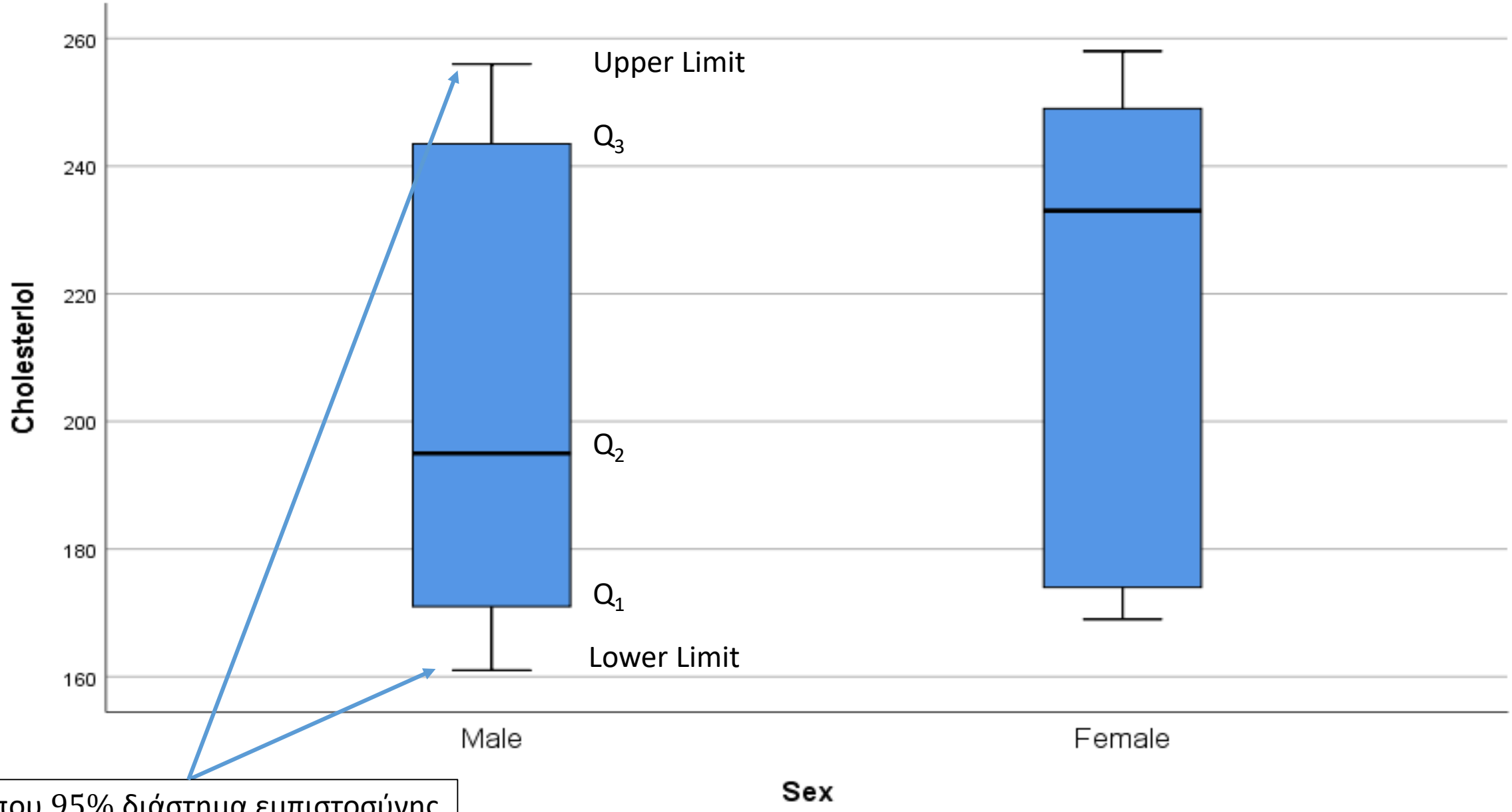
Παράδειγμα*:* Με τα δεδομένα από τους όγκους πλάσματος παράγεται το παρακάτω διάγραμμα κουτιού (η τιμή 7.32 θεωρείται ακραία και αγνοείται):



Θηκόγραμμα (Box-plot) έναντι Κανονική Κατανομή

Boxplot (με Interquartile range - IQR) και μία συνάρτηση πυκνότητας πιθανότητας (pdf) από έναν κανονικό $N(0, \sigma^2)$ πληθυσμό





Περίπου 95% διάστημα εμπιστοσύνης

Ποσοτικές Μέθοδοι Περιγραφής Δεδομένων

Αριθμητικά περιγραφικά μέτρα

- Μέτρα κεντρικής τάσης

Διάμεσος (Median)

Μέση τιμή (Mean)

Τεταρτημόρια (Quartiles)

Εκατοστημόρια (Percentiles)

- Μέτρα μεταβλητότητας

Εύρος (Range)

Διασπορά (Variance)

Τυπική απόκλιση (Std. Deviation, SD)

Μέση Τιμή

Ο πιο απλός τρόπος για να περιγραφεί ένα σύνολο παρατηρήσεων από μία **συνεχή μεταβλητή** είναι η μέση τιμή: το άθροισμα όλων των παρατηρήσεων δια τον αριθμό των παρατηρήσεων.

Παράδειγμα: Οι όγκοι πλάσματος (x) από 8 υγιείς ενήλικες άνδρες είναι:

2.75 lt	2.86 lt	3.37 lt	2.76 lt	2.62 lt	3.49 lt	3.05 lt	3.12 lt

x_1=2.75, x_2=2.86, x_3=3.37, x_4=2.76, x_5=2.62, x_6=3.49, x_7=3.05, x_8=3.12

Μέση Τιμή

Το άθροισμα των τιμών είναι:

$\sum x = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 2.75 + 2.86 + 3.37 + 2.76 + 2.62 + 3.49 + 3.05 + 3.12 = 24.02$

Ο αριθμός των παρατηρήσεων είναι n = 8

Οπότε η μέση τιμή είναι

$$\bar{x} = \frac{\sum x}{n} = \frac{(2.75 + 2.86 + 3.37 + 2.76 + 2.62 + 3.49 + 3.05 + 3.12)}{8} = \frac{24.02}{8} = 3.0025$$

Διάμεσος (median) - εκατοστημόρια (percentiles)

- Όταν υπάρχουν **μία** ή **περισσότερες υπερβολικά μικρές ή μεγάλες παρατηρήσεις**, η μέση τιμή δεν αντιπροσωπεύει τα δεδομένα
- Τότε οι παρατηρήσεις περιγράφονται καλύτερα από τη **διάμεσο** ή **50ο εκατοστημόριο**

- Οι παρατηρήσεις **ταξινομούνται** κατά αύξουσα σειρά
- Αν ο αριθμός των παρατηρήσεων είναι **μονός** η **κεντρική παρατήρηση** είναι η **διάμεσος**
- Αν ο αριθμός των παρατηρήσεων είναι **ζυγός** τότε η **διάμεσος** είναι η **μέση τιμή των δύο κεντρικών τιμών**

Διάμεσος (median)

Παράδειγμα*:* Η μέγιστη εισπνευστική πίεση, σε cmH_20, (PImax) 9 ασθενών με κυστική ίνωση είναι:

1	2	3	4	5	6	7	8	9
80	85	110	95	95	100	45	95	130

Ταξινομούμε τις τιμές κατά αύξουσα σειρά:

1	2	3	4	5	6	7	8	9
45	80	85	95	95	95	100	110	130

Τότε η διάμεσος είναι η κεντρική παρατήρηση που είναι η 5η τιμή $\left(\frac{9}{2} = 4.5 \cong 5\right)$, δηλ. διάμεσος = 95.

1	**2**	**3**	**4**	**5**	**6**	7	8	9
45	80	85	95	95	95	100	110	130

Διάμεσος (median)

Παράδειγμα*:* Οι όγκοι πλάσματος, σε lt, από 8 υγιείς ενήλικες άνδρες είναι:

1	2	3	4	5	6	7	8
2.75	2.86	3.37	2.76	2.62	3.49	7.32	3.05

Ταξινομούμε τις τιμές κατά αύξουσα σειρά:

1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

Τότε η διάμεσος είναι η μέση τιμή της $4^{ης}$ και $5^{ης}$ τιμής δηλ.

$$\text{διάμεσος} = (2.86 + 3.05)/2 = 2.96$$

Εκατοστημόρια (percentiles)

25ο (Q_1) 50ο (Q_2) 75ο (Q_3)

1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

$$L = \frac{p}{100}(n+1)$$

Όπου:

L η θέση στα ταξινομημένα δεδομένα
p το εκατοστημόριο που ψάχνουμε
n ο αριθμός των παρατηρήσεών μας

Εκατοστημόρια (percentiles)

$$L_{25} = \frac{25}{100}(8+1) = 2.25 \qquad L_{50} = \frac{50}{100}(8+1) = 4.5 \qquad L_{75} = \frac{75}{100}(8+1) = 6.75$$

25ο (Q_1) 50ο (Q_2) 75ο (Q_3)

1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

$$L_{50} = 4.5$$

Βλέπουμε πως η διάμεσος είναι στα μισά (0.5) ανάμεσα στην 4η και 5η παρατήρηση των οποίων οι τιμές είναι 2.86 και 3.05 αντίστοιχα.

$$Q_2 = 2.86 + 0.5(3.05 - 2.86) = 2.955 = 2.96$$

Διάμεσος (median)-εκατοστημόρια (percentiles)

25ο (Q_1) 50ο (Q_2) 75ο (Q_3)

1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

L_{25}=2.25

Βλέπουμε πως το 1ο εκατοστημόριο (Q_1) είναι ένα τέταρτο (0.25) ανάμεσα στην 2η και 3η παρατήρηση των οποίων οι τιμές είναι 2.75 και 2.76 αντίστοιχα.

$$Q_1 = 2.75 + 0.25(2.76 - 2.75) = 2.7525 = 2.75$$

Διάμεσος (median)-εκατοστημόρια (percentiles)

25ο (Q_1) 50ο (Q_2) 75ο (Q_3)

1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

L_{75}=6.75

Βλέπουμε πως το 3ο εκατοστημόριο (Q_3) είναι τρία τέταρτα (0.75) ανάμεσα στην 6η και 7η παρατήρηση των οποίων οι τιμές είναι 3.37 και 3.49 αντίστοιχα.

$$Q_3 = 3.37 + 0.75(3.49 - 3.37) = 3.46$$

Διάμεσος (median)-εκατοστημόρια (percentiles)

25ο (Q1) 50ο (Q2) 75ο (Q1)

1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

2.75 2.96 3.46

Επομένως:

Κατά προσέγγιση 25% από τους 8 υγιείς ενήλικες άνδρες έχουν όγκο πλάσματος 2.75 ή κάτω από 2.75.
Κατά προσέγγιση 50% από τους 8 υγιείς ενήλικες άνδρες έχουν όγκο πλάσματος 2.96 ή κάτω από 2.96.
Κατά προσέγγιση 75% από τους 8 υγιείς ενήλικες άνδρες έχουν όγκο πλάσματος 3.46 ή κάτω από 3.46.

Χρησιμοποιούμε τον όρο κατά προσέγγιση επειδή οι τιμές δεν βρίσκονται μέσα στα δεδομένα

Τρόποι μέτρησης της διασποράς

- Όμως χρειαζόμαστε και ένα μέτρο της διασποράς των δεδομένων
- Από **μόνη της η μέση τιμή δεν** μας επιτρέπει να **διαφοροποιήσουμε δείγματα**

Εύρος

- Το εύρος είναι η διαφορά μεταξύ της μεγαλύτερης και της μικρότερης παρατήρησης
- Δεν δείχνει όμως πώς κατανέμονται οι υπόλοιπες παρατηρήσεις μεταξύ αυτών των δύο

Παράδειγμα*:* Οι όγκοι πλάσματος x από 8 υγιείς ενήλικες άνδρες είναι:

2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12 lt

Εύρος = max-min = 3.49 - 2.62 = 0.87

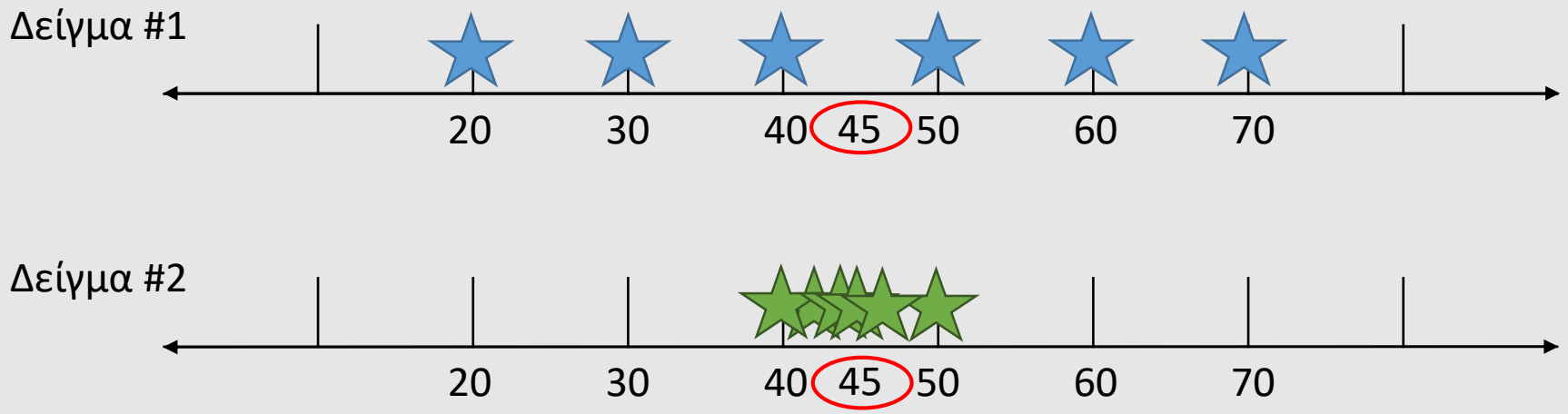
Διακύμανση (σ^2 ή s^2) (variance) και τυπική απόκλιση (standard deviation)

	Δείγμα #1	Δείγμα #2
1	20	40
2	30	43
3	40	44
4	50	46
5	60	47
6	70	50
	Μέση τιμή #1	Μέση τιμή #2
	$\bar{x} = 45$	$\bar{x} = 45$

Ίδια μέση τιμή

Αλλά διαφορετική διασπορά

Διακύμανση (σ^2 ή s^2) (variance) και τυπική απόκλιση (standard deviation)



Ερώτηση: Τι μας λένε τα δύο αυτά διαγράμματα σχετικά με την διακύμανση των δεδομένων;

Απάντηση: Ενώ και τα δύο έχουν την ίδια μέση τιμή, το Δείγμα #1 παρουσιάζει μεγαλύτερη μεταβλητότητα

Διακύμανση (σ^2 ή s^2) (variance) και τυπική απόκλιση (standard deviation)

- **Πόσο μακριά** βρίσκεται κάθε σημείο από τη μέση τιμή; (ΑΠΟΣΤΑΣΗ)
  - Αυτή είναι η ερώτηση που μας βοηθάει να απαντήσουμε η διακύμανση και η τυπική απόκλιση
- Η τυπική απόκλιση είναι απλά η τετραγωνική ρίζα της διακύμανσης. Οπότε είναι εύκολη να υπολογιστεί
- Αν κάποια σημεία είναι κοντά στη μέση τιμή, η διακύμανση και η τυπική απόκλιση θα είναι μικρότερη από αυτή για σημεία που βρίσκονται πιο μακριά από τη μέση τιμή
- Η μέση τιμή, η διακύμανση και η τυπική απόκλιση είναι πολύ σημαντικές όταν συγκρίνουμε σύνολα δεδομένων (data sets) (t-test, anova) ή όταν συγκρίνουμε ένα σύνολο δεδομένων (data set) με μία θεωρητική τιμή (one sample t-test)
- Καθώς η διακύμανση έχει το μειονέκτημα ότι είναι το τετράγωνο των παρατηρήσεων εκφράζουμε τη διασπορά με την τυπική απόκλιση
- Συμβολισμός
  - Μέση τιμή $\bar{x}$ (x-bar)
  - Διακύμανση σ^2 ή s^2
  - Τυπική απόκλιση σ ή s

Σημείωση: Κάνουμε χρήση διακύμανσης δείγματος και όχι πληθυσμού

Διακύμανση και τυπική απόκλιση

Διακύμανση

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Τυπική απόκλιση

$$s = \sqrt{s^2}$$

Όγκοι πλάσματος: διακύμανση και τυπική απόκλιση (σ^2 και σ)

	Όγκος πλάσματος x	Μέση τιμή $\bar{x}$	Όγκος πλάσματος – Μέση τιμή $x - \bar{x}$	$(x - \bar{x})^2$
1	2.75	3.0025	-0.2525	0.063756
2	2.86	3.0025	-0.1425	0.020306
3	3.37	3.0025	0.3675	0.135056
4	2.76	3.0025	-0.2425	0.058806
5	2.62	3.0025	-0.3825	0.146306
6	3.49	3.0025	0.4875	0.237656
7	3.05	3.0025	0.0475	0.002256
8	3.12	3.0025	0.1175	0.013806
		$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$		$s^2 = (0.063756 + 0.020306 + 0.135056 + 0.058806 + 0.146306 + 0.237656 + 0.002256 + 0.013806) / 7 = 0.09685$
				$s = \sqrt{0.09685} = 0.311207$

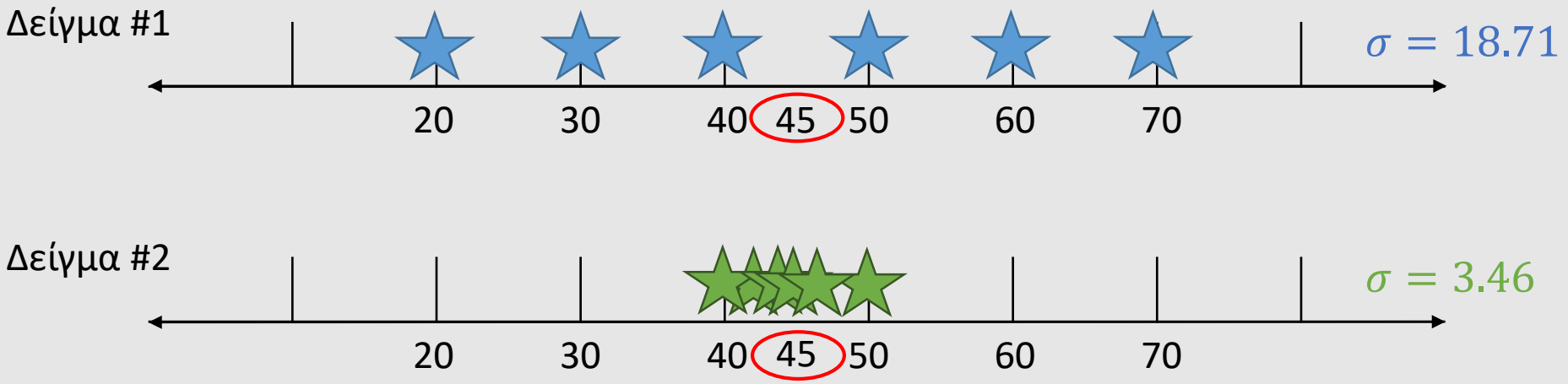
Δείγμα #1: διακύμανση και τυπική απόκλιση (σ^2 και σ)

	x	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
1	20	45	-25	625
2	30	45	-15	225
3	40	45	-5	25
4	50	45	5	25
5	60	45	15	225
6	70	45	25	625
	$\sigma^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$			$\sigma^2 = (625 + 225 + 25 + 25 + 225 + 625) \,/\, 5 = 350$
				$\sigma = \sqrt{350} = 18.71$

Δείγμα #2: διακύμανση και τυπική απόκλιση (σ^2 και σ)

	x	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
1	40	45	-5	25
2	43	45	-2	4
3	44	45	-1	1
4	46	45	1	1
5	47	45	2	4
6	50	45	5	25
	$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$			$\sigma^2 = (25 + 4 + 1 + 1 + 4 + 25) \,/\, 5 = 12$
				$\sigma = \sqrt{12} = 3.46$

Διακύμανση και τυπική απόκλιση (σ^2 και σ)

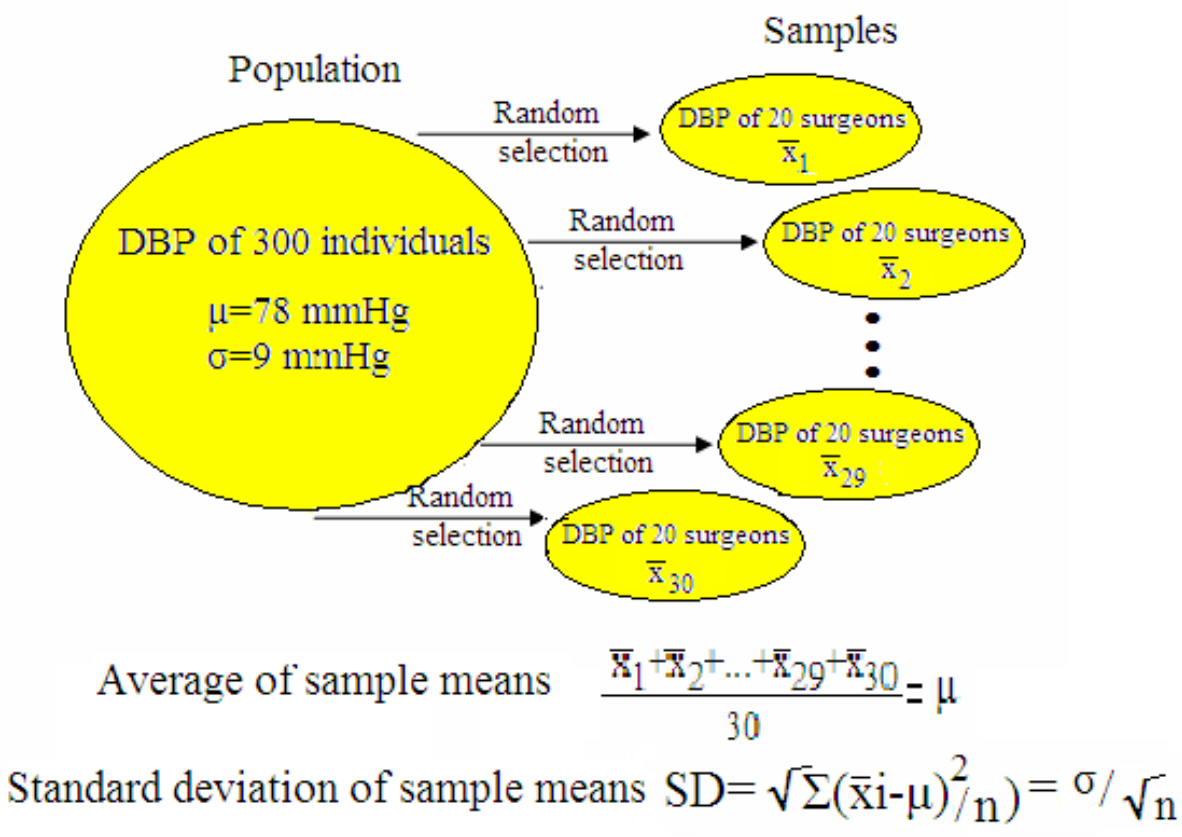


Τυπικό Σφάλμα (Standard Error)

- Εξάγουμε συμπεράσματα για έναν πληθυσμό συλλέγοντας ένα **αντιπροσωπευτικό δείγμα**
- Συνεπώς, η μέση τιμή (x) και η τυπική απόκλιση (*s*) ενός **δείγματος**, χρησιμοποιούνται για να **εκτιμήσουμε** τη μέση τιμή και την τυπική απόκλιση του **πληθυσμού** από τον οποίο προέρχεται το δείγμα, που συμβολίζονται με μ και σ, αντίστοιχα
- Η μέση τιμή ενός δείγματος **είναι απίθανο να είναι ακριβώς ίδια** με αυτή του πληθυσμού από όπου προέρχεται
- Ένα διαφορετικό δείγμα θα έδινε διαφορετική μέση τιμή και η διαφορά οφείλεται στη **δειγματοληπτική διακύμανση**

Τυπικό Σφάλμα (Standard Error)

- Αν συλλέξουμε **πολλά ανεξάρτητα δείγματα** του ίδιου μεγέθους και υπολογίσουμε τη **μέση τιμή** και την **τυπική απόκλιση** του καθενός τότε η μέση τιμή των μέσων τιμών των δειγμάτων είναι η **μέση τιμή του πληθυσμού**
- Η **τυπική απόκλιση** των **μέσων τιμών των δειγμάτων** είναι ίση με: $\frac{\sigma}{\sqrt{n}}$



Τυπικό Σφάλμα (Standard Error)

$$se = \frac{\sigma}{\sqrt{n}}$$

- Η **ποσότητα** $\frac{\sigma}{\sqrt{n}}$ λέγεται **τυπικό σφάλμα** της μέσης τιμής του δείγματος και μετρά πόσο καλά η μέση τιμή του πληθυσμού εκτιμάται από τη μέση τιμή του δείγματος
- Το SE είναι συνάρτηση της διακύμανσης και του μεγέθους του δείγματος (ή της κλινικής μελέτης)
- Ένα μεγάλο δείγμα με μικρή διακύμανση παράγει μικρό σφάλμα
- Επειδή σπάνια γνωρίζουμε την τυπική απόκλιση του πληθυσμού σ, χρησιμοποιούμε στη θέση της την τυπική απόκλιση του δείγματος s
- Συνεπώς, το τυπικό σφάλμα της μέσης τιμής εκτιμάται από την ποσότητα $se = \frac{s}{\sqrt{n}}$

Τυπικό Σφάλμα (Standard Error)

Παράδειγμα: Οι όγκοι πλάσματος, σε lt, από 8 υγιείς ενήλικες άνδρες είναι:

1	2	3	4	5	6	7	8
2.75	2.86	3.37	2.76	2.62	3.49	7.32	3.05

Μέση τιμή, $\bar{x}$	3.025
Τυπική απόκλιση, s	0.311
Το τυπικό σφάλμα της μέσης τιμής, $se(\bar{x})$	$se(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{0.311}{\sqrt{8}} = 0.111$

Αν το μέγεθος του δείγματος προσεγγίζει το μέγεθος του πληθυσμού τότε το se τείνει το 0 (μηδέν)