Κατανοώντας την συσχέτιση
(Understanding correlation)

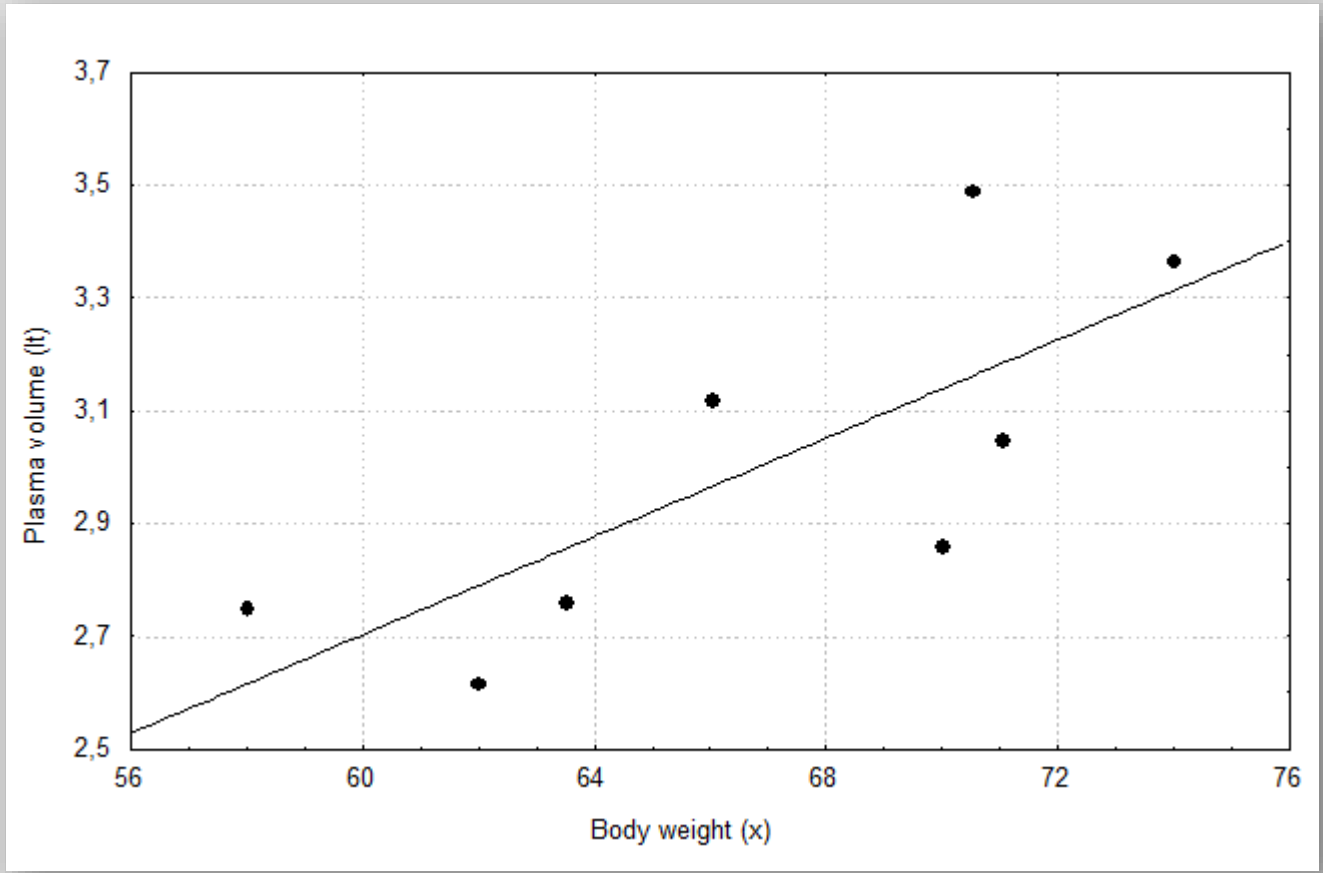
Ζιντζαράς Ηλίας, M.Sc., Ph.D.

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

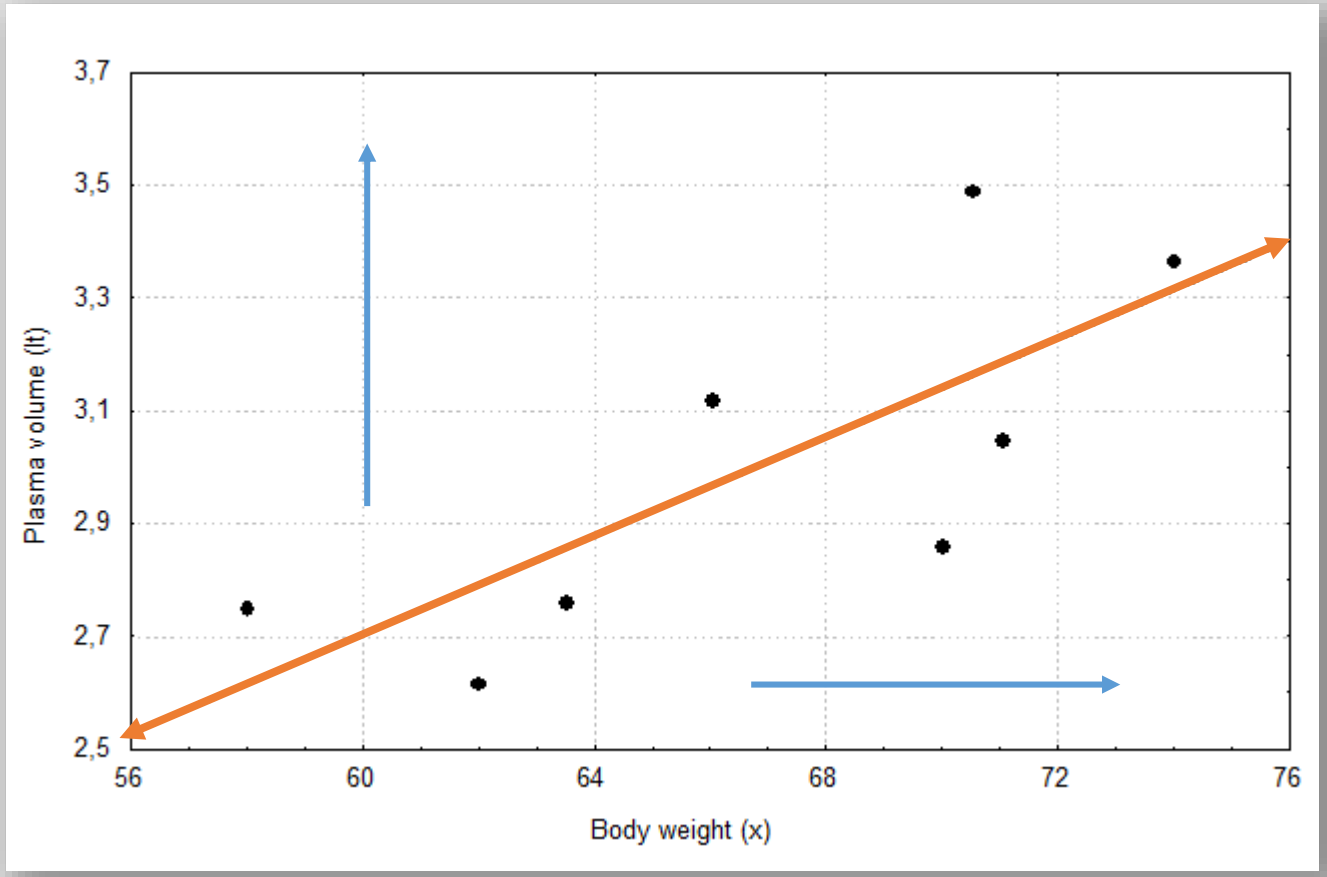
Θεόδωρος Μπρότσης, MSc, PhD
Εντεταλμένος Διδάσκων
(http://biomath.med.uth.gr)
Πανεπιστήμιο Θεσσαλίας
Email: tmprotsis@uth.gr

Όγκος πλάσματος 8 υγιών ανδρών



Άτομο	Βάρος σε Kg (x)	Όγκος πλάσματος σε lt (y)
1	58.0	2.75
2	70.0	2.86
3	74.0	3.37
4	63.5	2.76
5	62.0	2.62
6	70.5	3.49
7	71.0	3.05
8	66.0	3.12

Όγκος πλάσματος 8 υγιών ανδρών



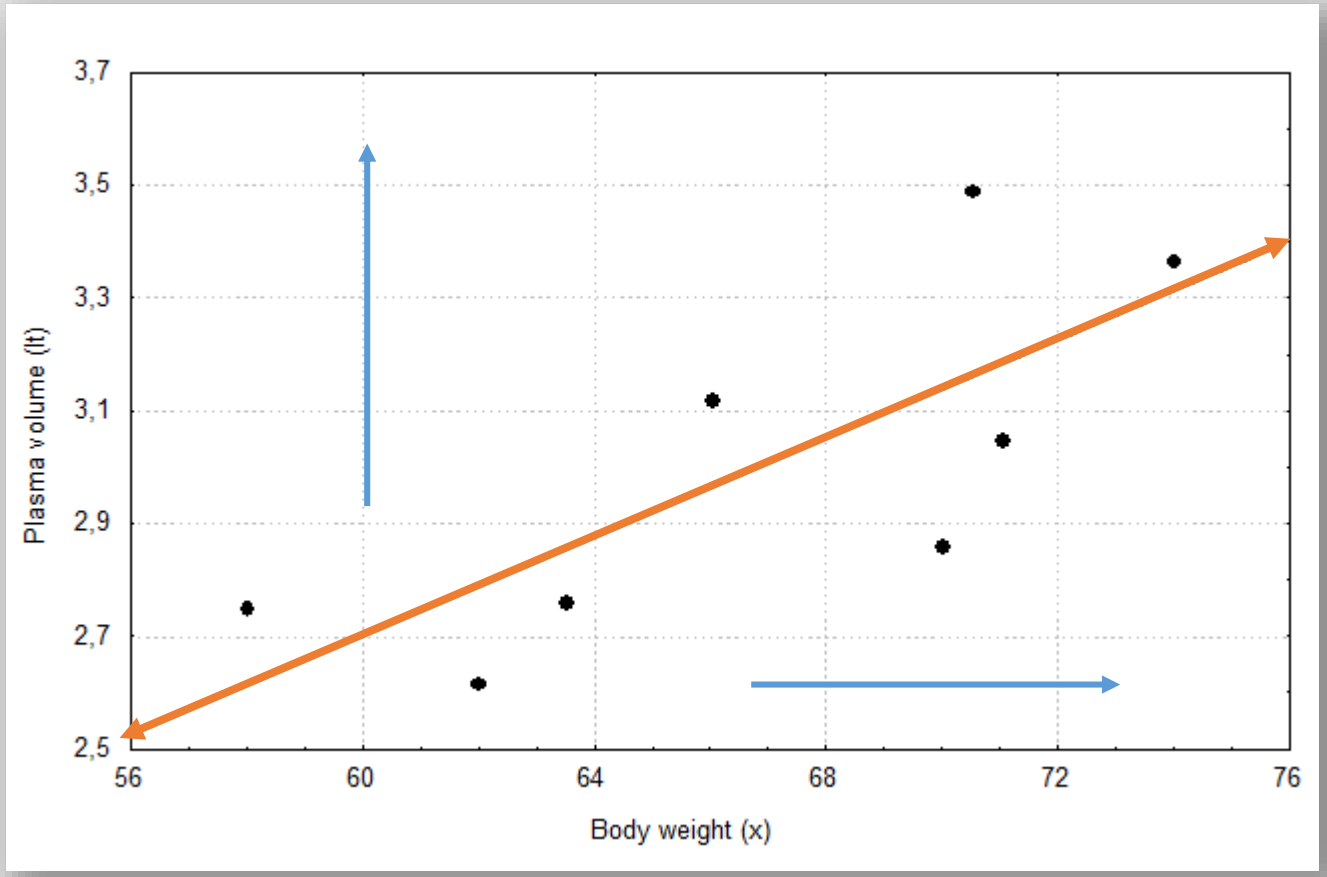
Πως θα περιγράφαμε το σχήμα ή το μοτίβο των σημείων των δεδομένων μας;

Φαίνεται πως έχουν μία γραμμική σχέση

Όταν αυξάνεται το βάρος τι συμβαίνει στον όγκο πλάσματος;

Αυξάνει επίσης

Όγκος πλάσματος 8 υγιών ανδρών

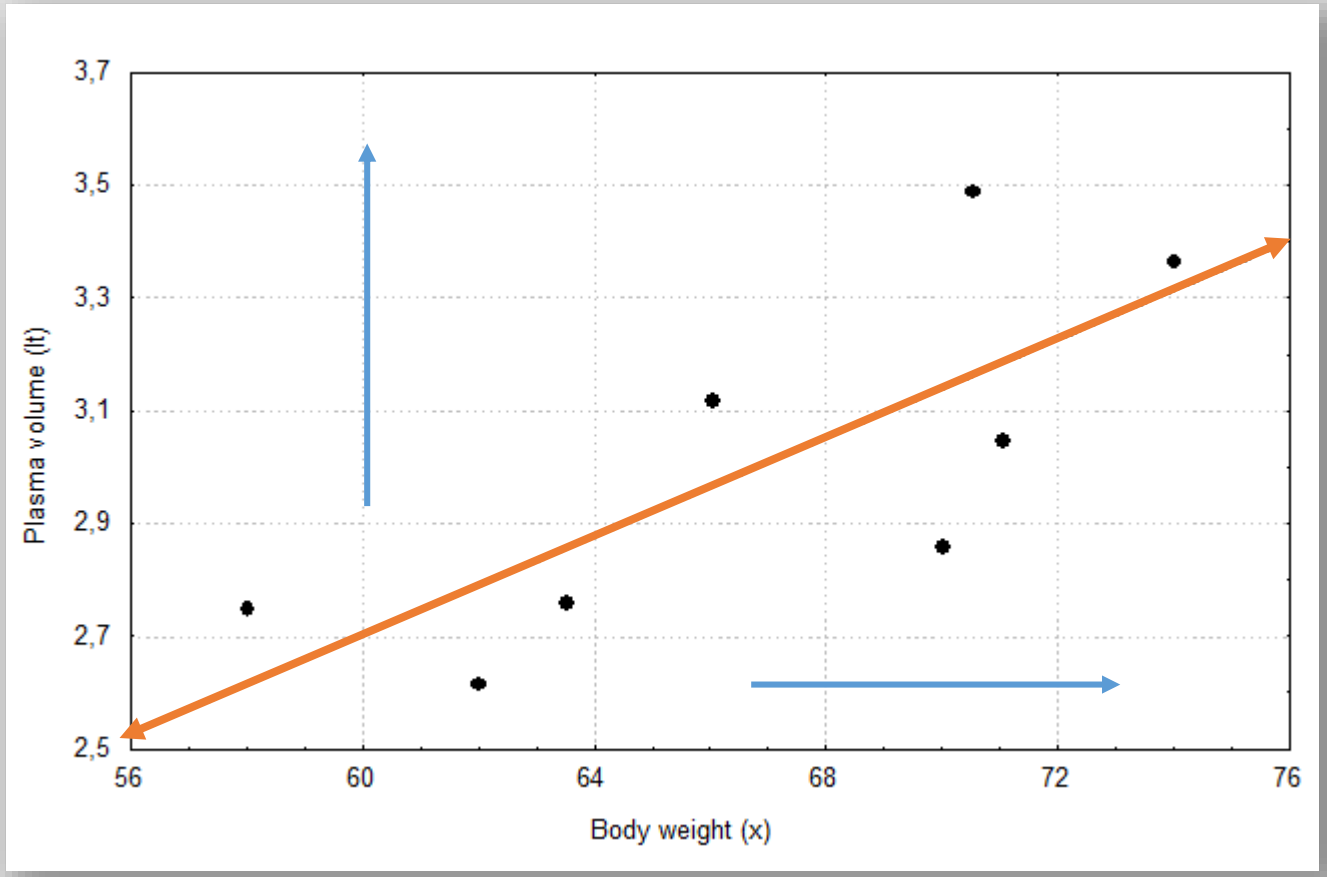


Ακολουθούν γραμμική σχέση!

Λέμε ότι δύο μεταβλητές που εμφανίζουν αυτού του είδους μοτίβο έχουν μια **θετική γραμμή σχέση**

Όταν μία μεταβλητή κινείται προς κάποια κατεύθυνση η άλλη κινείται προς την ίδια κατεύθυνση

Όγκος πλάσματος 8 υγιών ανδρών



Ακολουθούν γραμμική σχέση!

Αυτό ονομάζεται

ΣΥΝΔΙΑΚΥΜΑΝΣΗ
COVARIANCE

CO-vary

Πως αλλάζουν μαζί

Γραμμικές σχέσεις

- Η συνδιακύμανση είναι μια από τις οικογένειες στατιστικών μέτρων που χρησιμοποιούνται στην ανάλυση της γραμμικής σχέσης μεταξύ δύο ποσοτικών μεταβλητών
- Πως δύο μεταβλητές συμπεριφέρονται ως ζευγάρι;

Συνδιακύμανση (Covariance)

Συσχέτιση (Correlation)

Γραμμική παλινδρόμηση (Linear Regression)

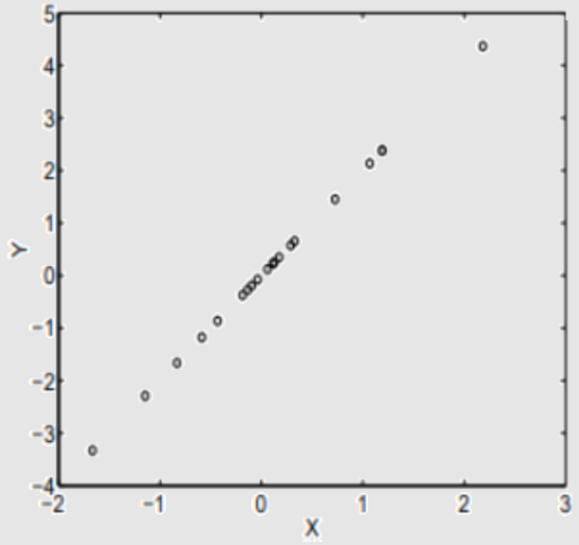
Συνδιακύμανση έναντι Συσχέτισης

- Η **συνδιακύμανση** μας δείχνει την ΚΑΤΕΥΘΥΝΣΗ (θετική, αρνητική, κοντά στο μηδέν) της γραμμικής σχέσης μεταξύ δύο ποσοτικών μεταβλητών
  - Ενώ η **συσχέτιση** μας δείχνει την ΚΑΤΕΥΘΥΝΣΗ και την ΙΣΧΥ
- Η **συνδιακύμανση** δεν έχει πάνω και κάτω όριο και το μέγεθός της εξαρτάται από την κλίμακα των μεταβλητών
  - Ενώ η **συσχέτιση** είναι πάντα μεταξύ -1 και $+1$ και η κλίμακά της είναι *ανεξάρτητη* από την κλίμακα των μεταβλητών
- Η **συνδιακύμανση** δεν είναι τυποποιημένη
  - Ενώ η **συσχέτισή** είναι τυποποιημένη (θυμηθείτε τις $z-$ τιμές)

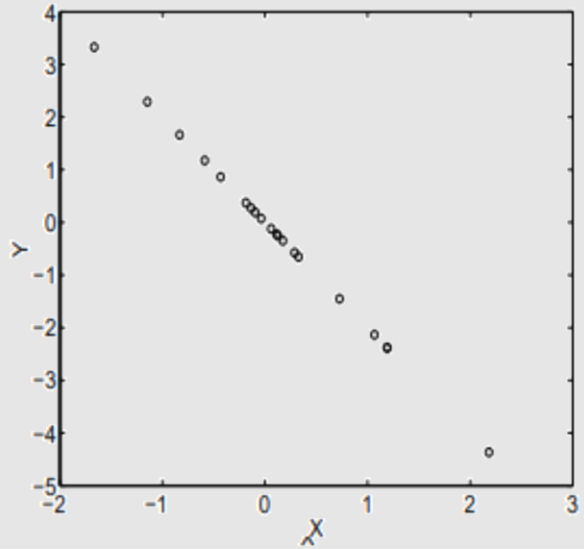
Επιφυλάξεις συσχέτισης

1. Πριν ξεκινήσουμε να υπολογίζουμε συσχετίσεις κοιτάμε το διάγραμμα συσχέτισης των δεδομένων. Τι μοτίβο (αν υπάρχει) παρουσιάζουν;
2. Η συσχέτιση είναι εφαρμόσιμη μόνο σε ΓΡΑΜΜΙΚΕΣ σχέσεις. Υπάρχουν και άλλες σχέσεις μεταξύ δύο μεταβλητών
3. Η συσχέτιση ΔΕΝ είναι αιτιολογική
    1. Η συσχέτιση δεν μπορεί να χρησιμοποιηθεί για να συναχθεί η ύπαρξη αιτιώδους σχέσης μεταξύ των μεταβλητών
    # Συσχέτιση μεταξύ της διάθεσης και της υγείας σε άτομα είναι λιγότερο αιτιολογικά διαφανείς. Η βελτιωμένη διάθεση οδηγεί σε βελτίωση της υγείας, ή η καλή υγεία οδηγεί σε καλή διάθεση, ή και τα δύο; Ή μήπως κάποιος άλλος παράγοντας αποτελεί τη βάση για και τα δύο;
4. Η ισχύς της συσχέτισης δεν σημαίνει απαραίτητα πως η συσχέτιση είναι στατιστικά σημαντική· εξαρτάται από το μέγεθος του δείγματος

Γενικά μοτίβα συσχέτισης

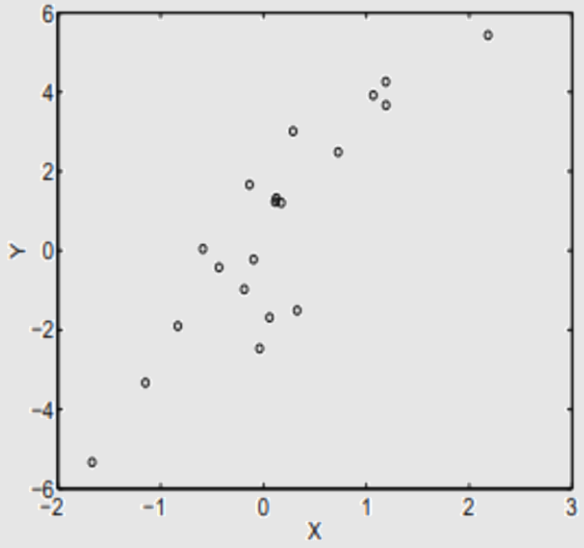


Τέλεια θετική γραμμική συσχέτιση, $+1$

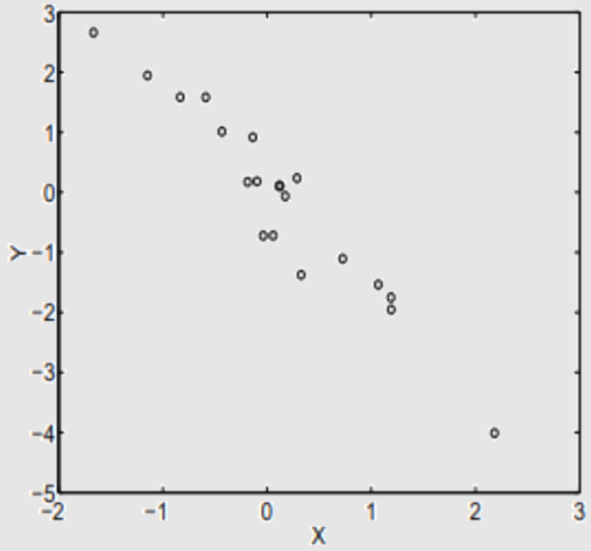


Τέλεια αρνητική γραμμική συσχέτιση, -1

Γενικά μοτίβα συσχέτισης

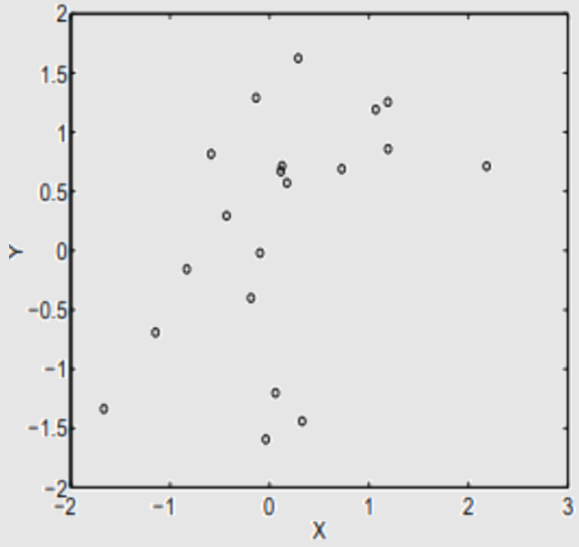


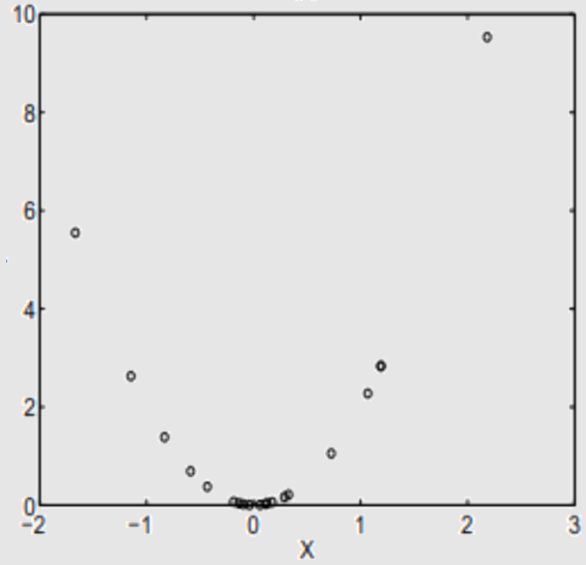
Θετική γραμμική συσχέτιση, κοντά στο $+1$



Αρνητική γραμμική συσχέτιση, κοντά στο -1

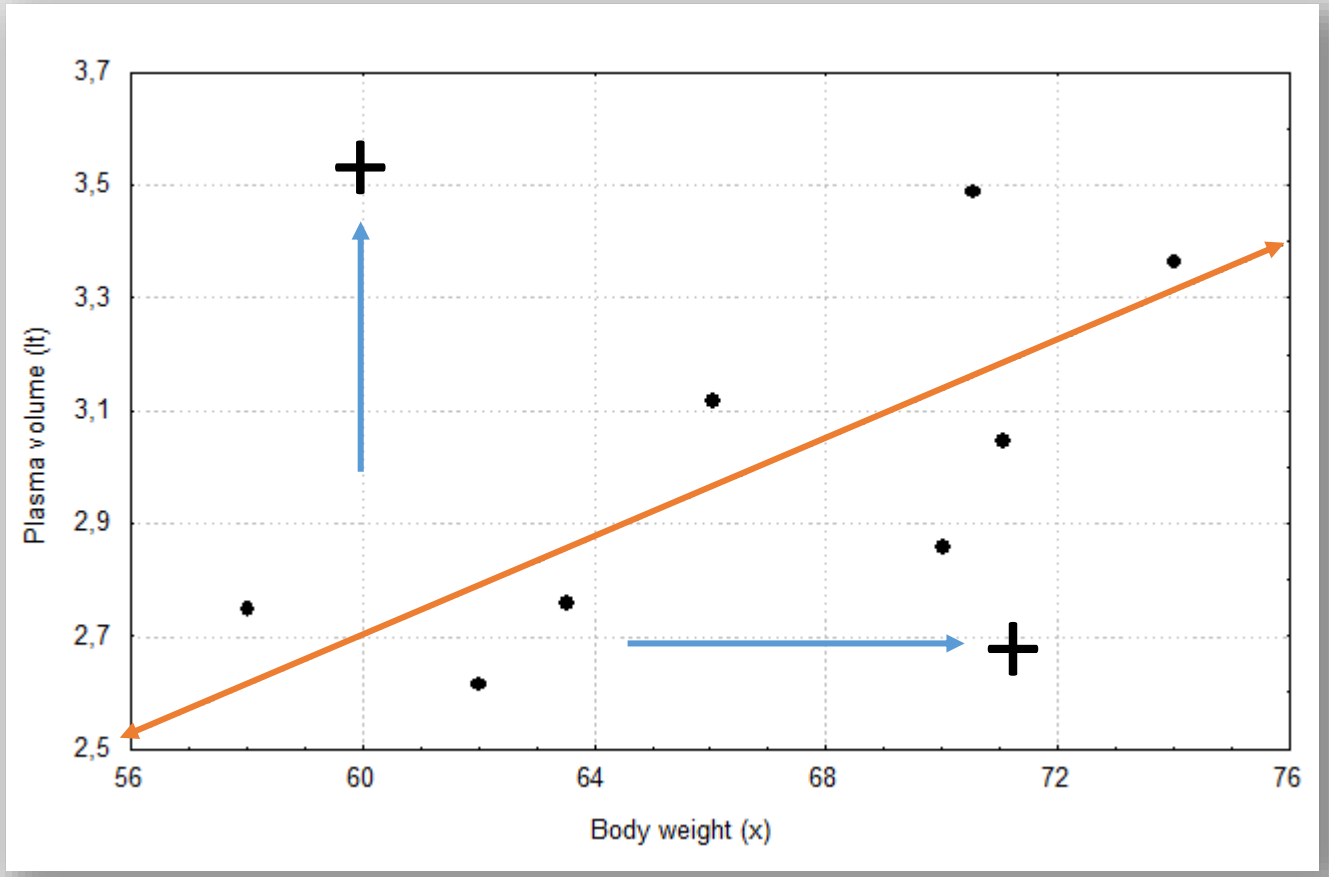
Καμία γραμμική συσχέτιση





Πάντα βλέπουμε πρώτα το διάγραμμα συσχέτισης των δεδομένων μας!!!

Όγκος πλάσματος 8 υγιών ανδρών



Correlations

		Body weight	Plasma volume (lt)
Body weight	Pearson Correlation	1	.759*
	Sig. (2-tailed)		0.029
	N	8	8
Plasma volume (lt)	Pearson Correlation	.759*	1
	Sig. (2-tailed)	0.029	
	N	8	8

*. Correlation is significant at the 0.05 level (2-tailed).

Τι μας λέει το SPSS;

$$r = 0.759$$

Τύπος συσχέτισης

Το r ονομάζεται συντελεστής συσχέτισης (Pearson)

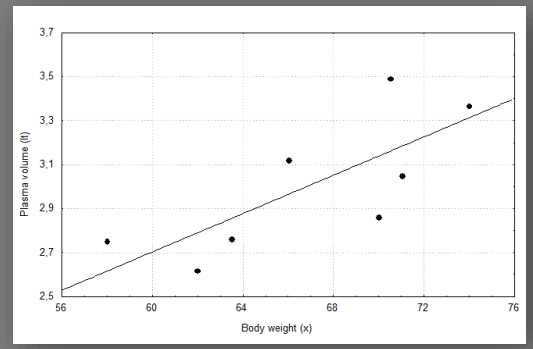
$$r = \frac{Covariance(x, y)}{Standard\ Deviation(x) \times Standard\ Deviation(y)}$$

$$r = \frac{Cov\ (x, y)}{S_x S_y}$$

1. Συνδιακύμανση μεταξύ των δύο μεταβλητών
2. Διαιρεμένο με τον πολλαπλασιασμό των τυπικών αποκλίσεων τους

Σημείωση: Αν γνωρίζουμε το r και τις τυπικές αποκλίσεις, μπορούμε να βρούμε την συνδιακύμανση!

Παράδειγμα


3,7
3,5
3,3
3,1
2,9
2,7
2,5
Plasma volume (l)
56
60
64
68
72
76
Body weight (x)

Σχέση αναστήματος 10 έγγαμων ζευγαριών

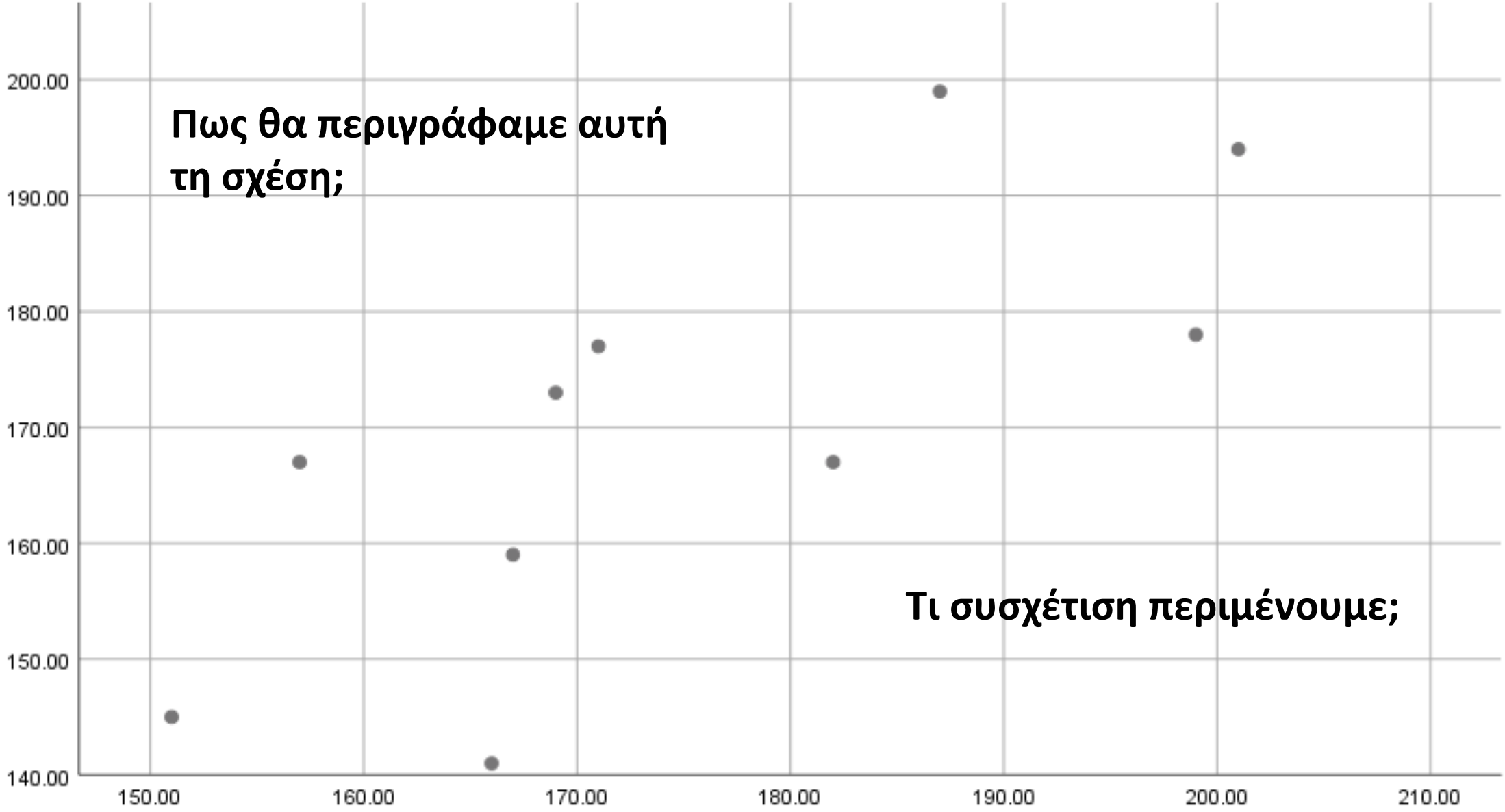
Έστω πως θέλουμε να μελετήσουμε τη σχέση μεταξύ του αναστήματος των συζύγων σε 10 έγγαμα ζευγάρια

Για να γίνει αυτό παίρνουμε ένα δείγμα 10 μετρήσεων

x = ανάστημα του άνδρα

y = ανάστημα της γυναίκας

Ανάστημα ανδρών έναντι ανάστημα γυναικών



Πως θα περιγράφαμε αυτή τη σχέση;

Τι συσχέτιση περιμένουμε;

#Ζεύγος	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$		
1	151	145	-24	-25	576	625		
2	166	141	-9	-29	81	841		
3	167	159	-8	-11	64	121		
4	157	167	-18	-3	324	9		
5	187	199	12	29	144	841		
6	201	194	26	24	676	576		
7	199	178	24	8	576	64		
8	182	167	7	-3	49	9		
9	169	173	-6	3	36	9		
10	171	177	-4	7	16	49		
	$\bar{x} = 175$	$\bar{y} = 170$			$\sum = 2542$	$\sum = 3144$	$s_x = \sqrt{\frac{2542}{9}} = 16.81$	$s_y = \sqrt{\frac{3144}{9}} = 18.69$

#Ζεύγος	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	151	145	-24	-25	600
2	166	141	-9	-29	261
3	167	159	-8	-11	88
4	157	167	-18	-3	54
5	187	199	12	29	348
6	201	194	26	24	624
7	199	178	24	8	192
8	182	167	7	-3	-21
9	169	173	-6	3	-18
10	171	177	-4	7	-28
	$\bar{x} = 175$	$\bar{y} = 170$			$\sum = 2100$

$$Cov(x, y) = s_{xy} = \frac{2100}{n - 1}$$

$$\frac{2100}{9}$$

$$Cov(x, y) = 233.33$$

$$s_x = 16.81 \qquad s_y = 18.69$$

Υπολογισμός συσχέτισης

$$r = \frac{Cov\,(x,y)}{s_x s_y}$$

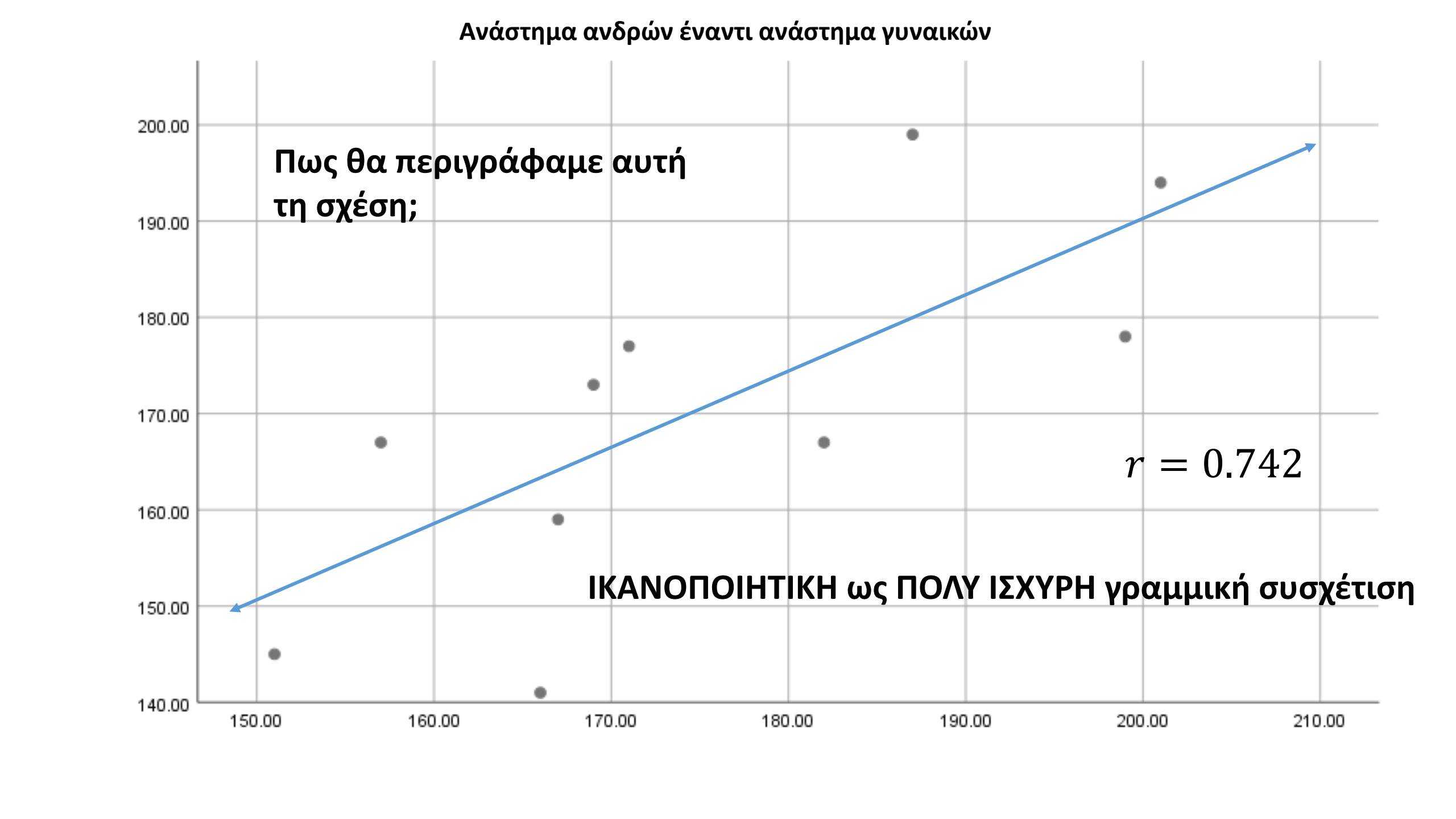
$$r = \frac{s_{xy}}{s_x s_y}$$

$$r = \frac{233.33}{16.81 \times 18.69}$$

$$r = \frac{233.33}{314.18}$$

$$r = 0.742$$

Το αποτέλεσμα μπορεί να διαφέρει λίγο από το SPSS λόγω στρογγυλοποίησης


Ανάστημα ανδρών έναντι ανάστημα γυναικών
200.00
190.00
180.00
170.00
160.00
150.00
140.00
150.00
160.00
170.00
180.00
190.00
200.00
210.00
Πως θα περιγράφαμε αυτή τη σχέση;
$r = 0.742$
ΙΚΑΝΟΠΟΙΗΤΙΚΗ ως ΠΟΛΥ ΙΣΧΥΡΗ γραμμική συσχέτιση

Ισχύς συντελεστή συσχέτισης

- Όταν $R = 1$, έχουμε **τέλεια θετική γραμμική συσχέτιση**.
- Όταν $R = -1$, έχουμε **τέλεια αρνητική γραμμική συσχέτιση**.
- Όταν $R = 0$, **δεν** έχουμε γραμμική σχέση.
- Όταν $0.7 < |r| < 1$, έχουμε ικανοποιητική ως πολύ ισχυρή συσχέτιση
- Όταν $0.5 < |r| < 0.7$, έχουμε μέτρια έως ικανοποιητική συσχέτιση
- Όταν $0.3 < |r| < 0.5$, έχουμε ασθενής έως μέτρια συσχέτιση
- Όταν $R > 0$, **έχουμε θετική σχέση** μεταξύ των δύο μεταβλητών, δηλαδή οι x και y συμμεταβάλλονται προς την ίδια κατεύθυνση.
- Όταν $R < 0$, **έχουμε αρνητική σχέση** μεταξύ των δύο μεταβλητών, δηλαδή οι x και y συμμεταβάλλονται προς την αντίθετη κατεύθυνση.
- Όσο μεγαλύτερη είναι απόλυτη τιμή του συντελεστή pearson |R| τόσο «καλύτερη» (ισχυρότερη) είναι η γραμμική σχέση μεταξύ των x και y.

Κανόνας συσχέτισης

Πώς μπορούμε να δηλώσουμε πιο αντικειμενικά αν υπάρχει σχέση μεταξύ δύο μεταβλητών

Κανόνας

$$\textbf{Αν } |r| \geq \frac{2}{\sqrt{n}} \textbf{ τότε υπάρχει συσχέτιση}$$

Κανόνας συσχέτισης

Οπότε για το πρόβλημά μας

$$|r| \geq \frac{2}{\sqrt{10}}$$

$$0.742 \geq 0.632$$

Η τιμή 0.632 είναι το κατώφλι του κανόνα

Επομένως υπάρχει συσχέτιση

Έλεγχος στατιστικής σημαντικότητας

Για να ελέγξουμε αν η τιμή της γραμμική
σημαντική, δηλ. διαφορετική από το 0, χ

$$t =$$

Αν η απόλυτη τιμή του t είναι μεγαλύτερ με $n-2$
(n είναι το πλήθος των ζευγών) β. ε. (βαθ
σημαντικότητας 5% ($p = 0.05$) τότε το

	Percentace points of the t distribution		
	p-value		
df(=n-1)	0.05	0.01	0.001
1	12.71	63.66	636.62
2	4.3	9.92	31.6
3	3.18	5.84	12.92
4	2.78	4.6	8.61
5	2.57	4.03	6.87
6	2.45	3.71	5.96
7	2.36	3.5	5.41
8	2.31	3.36	5.04
9	2.26	3.25	4.78
10	2.23	3.17	4.59
11	2.2	3.11	4.44
12	2.18	3.05	4.32
13	2.16	3.01	4.22
14	2.14	2.98	4.14
15	2.13	2.95	4.07
20	2.09	2.85	3.85
30	2.04	2.75	3.65
40	2.02	2.7	3.55
120	1.98	2.62	3.37
∞	1.96	2.58	3.29

$t - test$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t = \frac{0.742\sqrt{10-2}}{\sqrt{1-0.742^2}}$$

$$t = \frac{0.742 \times 2.828}{\sqrt{0.449}}$$

$$t = \frac{2.0984}{0.670}$$

$$t = 3.139$$

Η τιμή $t = 3.139$ είναι μεγαλύτερη από την τιμή της t −κατανομής με 8 $(n-2)$ β. ε. (df) σε $p = 0.05$ που είναι 2.31.

Συνεπώς, υπάρχει στατιστικά σημαντική συσχέτιση $(p < 0.05)$ μεταξύ του ύψους των ανδρών και του ύψους των γυναικών.

Ανασκόπηση

- Ο συντελεστής συσχέτισης **Pearson** ανιχνεύει τη γραμμική σχέση δύο ποσοτικών μεταβλητών, για συνεχείς μεταβλητές ή διακριτές αριθμητικές τιμές (π. χ. οικογένεια με ένα παιδί, οικογένεια με δύο παιδιά, οικογένεια με τρία παιδιά)
- Συμπληρώνει το διάγραμμα διασποράς ή συσχέτισης
- Ο συντελεστής συσχέτισης είναι παραμετρικός, προϋποθέτει δηλ. πως οι τιμές προέρχονται από κανονικούς πληθυσμούς
- Η κανονικότητα τεκμηριώνεται χρησιμοποιώντας το ιστόγραμμα
- Αν δεν υπάρχει κανονικότητα ή οι μεταβλητές είναι διακριτές με σχετικά λίγες τιμές τότε καλύτερα είναι να υπολογιστεί ο συντελεστής **Spearman** που είναι το μη παραμετρικό ανάλογο του συντελεστή Pearson