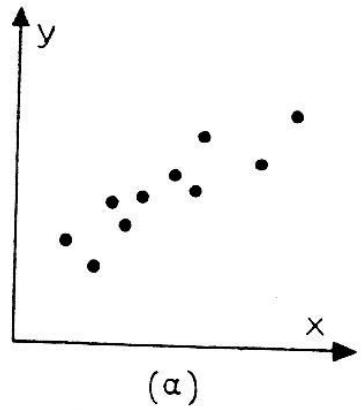
ΓΡΑΜΜΙΚΗ ΣΥΣΧΕΤΙΣΗ

Παναγιώτα Λάλου, Αλέξανδρος Γρυπάρης

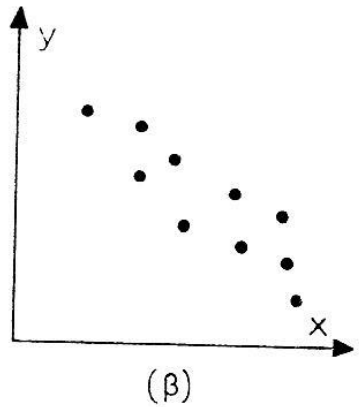
Συντελεστής συσχέτισης του Pearson: r

- Ο επόμενος έλεγχος που θα μελετήσουμε διερευνά αν υπάρχει γραμμική σχέση μεταξύ 2 ποσοτικών μεταβλητών.

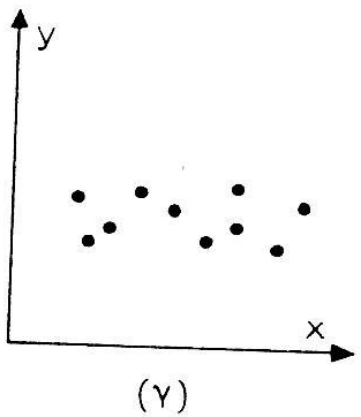
Διαγράμματα συσχέτισης:
(α) θετική συσχέτιση, (β) αρνητική συσχέτιση,
(γ) και (δ) απουσία γραμμικής συσχέτισης



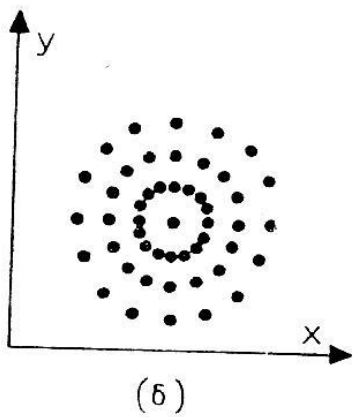
(α)



(β)



(γ)



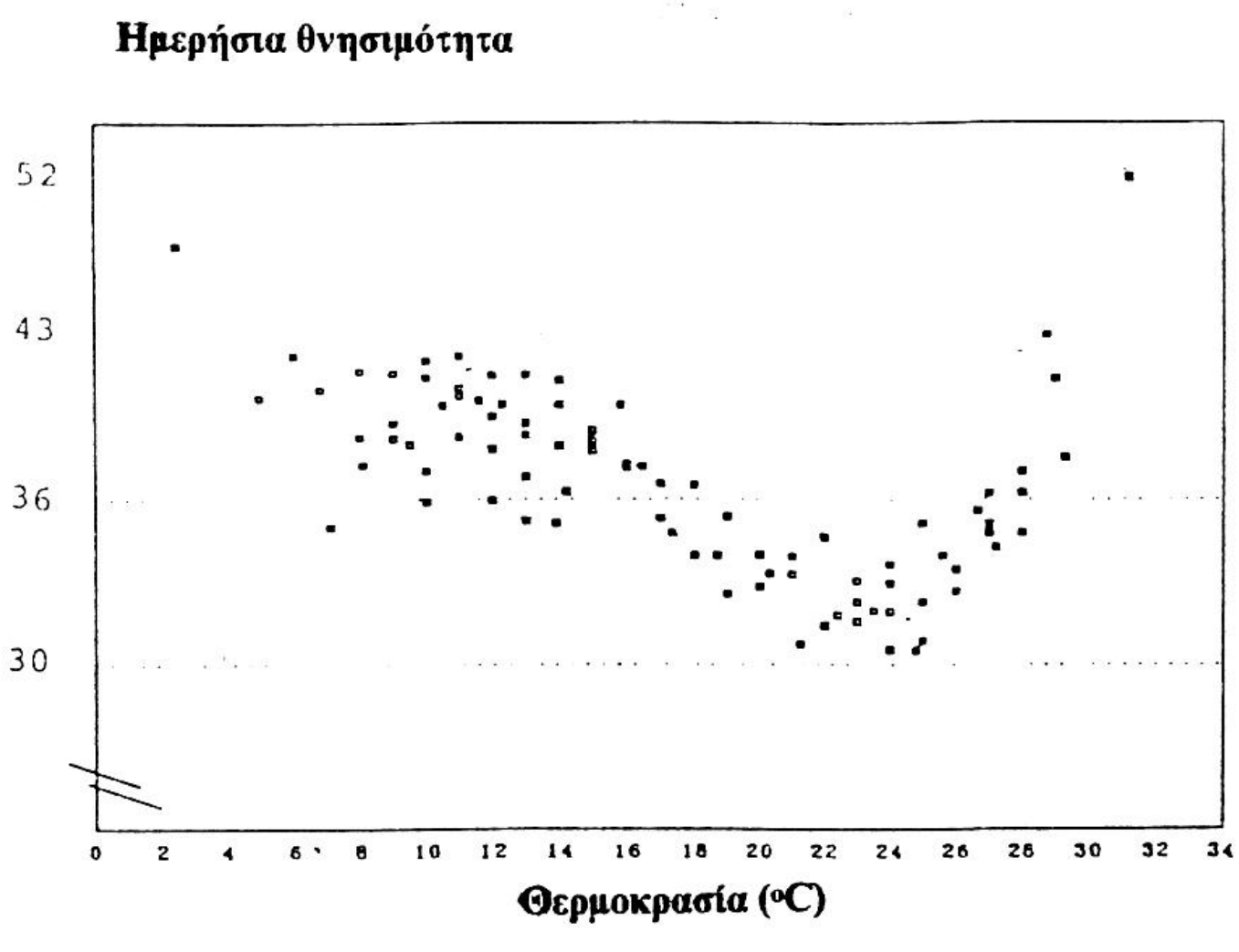
(δ)

Έλλειψη γραμμικής σχέσης;

- Έλλειψη γραμμικής σχέσης δεν συνεπάγεται και απουσία σχέσης.
  - Δηλαδή, μπορεί να υπάρχει σχέση μεταξύ 2 μεταβλητών, απλά να μην είναι γραμμική!

- Στο επόμενο σχήμα παρουσιάζεται το στικτόγραμμα της μέσης ημερήσιας θερμοκρασίας με τον μέσο ημερήσιο αριθμό θανάτων από όλες τις αιτίες.

Σχέση μεταξύ μέσης ημερήσιας θνησιμότητας και μέσης ημερήσιας θερμοκρασίας



(συν.)

- Παρατηρούμε μια μη-γραμμική σχέση μεταξύ θνησιμότητας και θερμοκρασίας:
  - Μεγαλύτερη θνησιμότητα παρατηρείται κατά τις πολύ κρύες ή πολύ ζεστές ημέρες.
  - Στην περίπτωση αυτή, ο συντελεστής συσχέτισης του Pearson δεν θα μας ήταν χρήσιμος.

Ιδιότητες του συντελεστή συσχέτισης

1. Είναι καθαρός αριθμός

2. Παίρνει τιμές από -1 ως +1 (αρνητική ή θετική συσχέτιση).

   Θετική συσχέτιση: σε υψηλές τιμές του ενός μεγέθους αντιστοιχούν υψηλές τιμές του άλλου.
   Αρνητική: σε υψηλές τιμές του ενός μεγέθους αντιστοιχούν χαμηλές τιμές του άλλου.

   - Όσο ο *r* πλησιάζει την τιμή +1 (ή –1) τόσο ισχυρότερη είναι η συσχέτιση, ενώ όσο πιο κοντά είναι στο 0 τόσο πιο αδύναμη η γραμμική συσχέτιση.

3. Μετρά μόνο την γραμμική συσχέτιση

Οι τιμές του συντελεστή …

Εμπειρική κατηγοριοποίηση του συντελεστή συσχέτισης.

Ο συντελεστής συσχέτισης r του Pearson

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) * \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

Έλεγχος υπόθεσης για τον συντελεστή συσχέτισης του Pearson

- Όταν έχουμε 2 ποσοτικά μεγέθη, μας ενδιαφέρει να ελέγξουμε αν υπάρχει ή όχι γραμμική συσχέτιση μεταξύ τους.
- Αυτό είναι ισοδύναμο με το να ελέγξουμε αν ο r είναι ίσος με 0 (οπότε δεν υπάρχει γραμμική συσχέτιση), ή όχι.
- Ο έλεγχος υπόθεσης που χρησιμοποιούμε είναι ο εξής:

— H_0: r=0 (οι μεταβλητές μας δεν συσχετίζονται γραμμικά)

— H_1: r≠0 (οι μεταβλητές μας συσχετίζονται γραμμικά)

— Η κρίσιμη περιοχή είναι η: $R = \{t > t_{n-2;\,\alpha}\}$

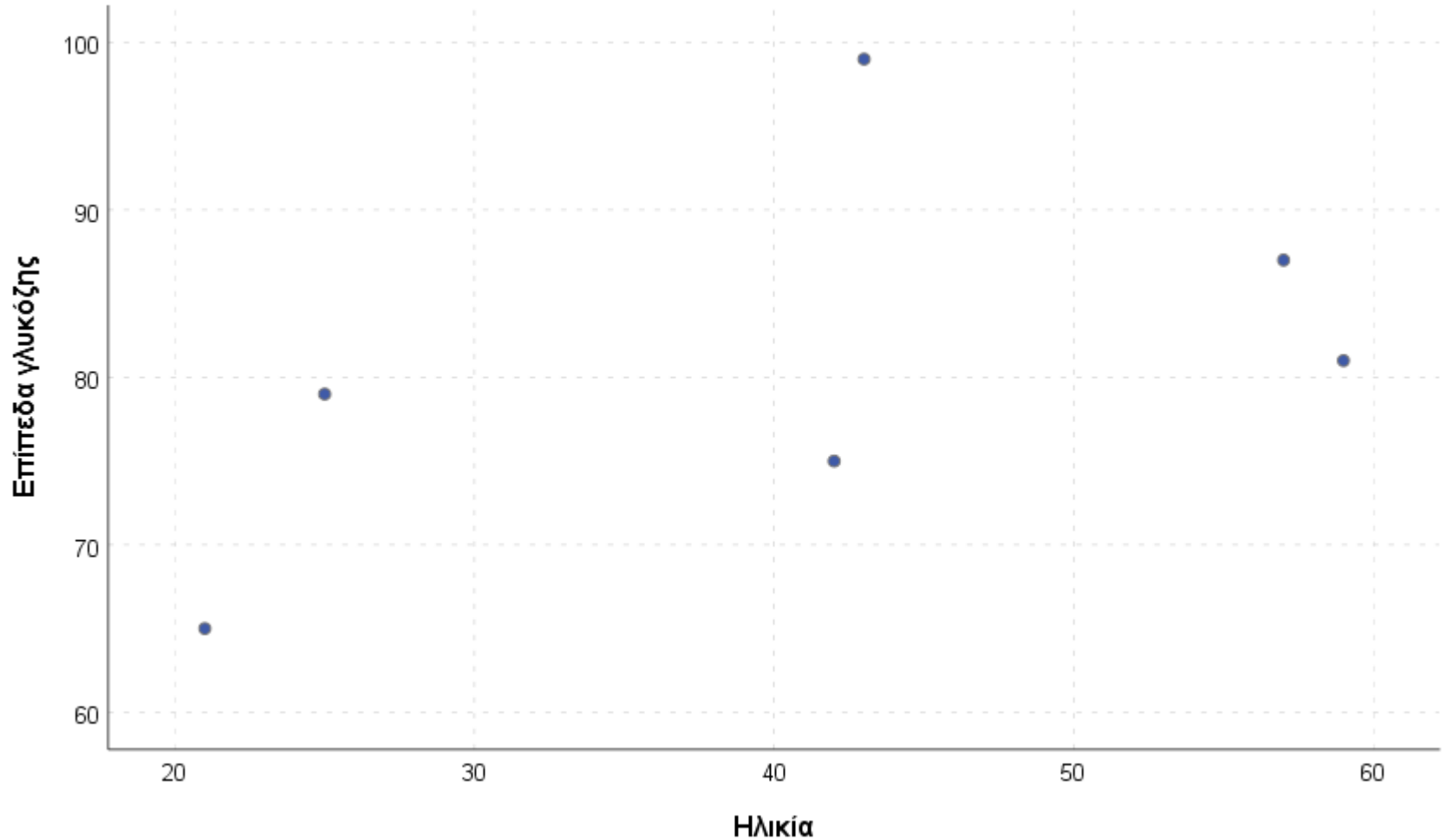
— Το στατιστικό κριτήριο: $t = r * \sqrt{\frac{n-2}{1-r^2}}$

Παράδειγμα

Σε 6 άτομα που έπασχαν από διαβήτη τύπου 2, μετρήθηκαν οι τιμές γλυκόζης στο αίμα. Το ερώτημα που μας απασχολεί είναι αν τα επίπεδα αυτά σχετίζονται με την ηλικία τους. Ο παρακάτω πίνακας παρουσιάζει τα στοιχεία που έχουμε:

Ηλικία (X)	Επίπεδα Γλυκόζης (Y)
43	99
21	65
25	79
42	75
57	87
59	81

Γράφημα (στικτόγραμμα)



Θέλουμε να ελέγξουμε αν:

— H_0: Τα επίπεδα γλυκόζης δεν συσχετίζονται γραμμικά με την ηλικία

— H_1: Τα επίπεδα γλυκόζης συσχετίζονται γραμμικά με την ηλικία

	Ηλικία (Χ)	Επίπεδα γλυκόζης (Υ)	ΧΥ	$Χ^2$	$Υ^2$
	43	99	4257	1849	9801
	21	65	1365	441	4225
	25	79	1975	625	6241
	42	75	3150	1764	5625
	57	87	4959	3249	7569
	59	81	4779	3481	6561
ΣΥΝΟΛΟ	***247***	***486***	**20485**	***11409***	***40022***

$$\bar{x} = \frac{247}{6} = 41{,}17$$

$$\bar{y} = \frac{486}{6} = 81$$

Ο συντελεστής συσχέτισης r του Pearson δίνεται από τη σχέση:

$$\mathrm{r} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) * (\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}} = \frac{20.485 - 20 * 41{,}17 * 81}{\sqrt{(11.409 - 6 * 41{,}17^2)(40.022 - 6 * 81^2)}} =$$

$$= 0{,}53$$

Το στατιστικό κριτήριο δίνεται από τη σχέση:

$$t = r * \sqrt{\frac{n-2}{1-r^2}} = 0{,}53 * \sqrt{\frac{6-2}{1-0{,}53^2}} = 0{,}53 * \sqrt{5{,}56} = 0{,}53 * 2{,}35 = 1{,}25$$

Η κρίσιμη τιμή είναι η $t_{n-2;\alpha} = t_{4;0,05}$=2,132

Η κρίσιμη περιοχή είναι η: $R = \{t > t_{n-2;\,\alpha}\}$

Έχουμε 1,25=t < $t_{n-2;\,\alpha}$ =2,132. Άρα δεν βρισκόμαστε στην κρίσιμη περιοχή.

Οπότε δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Έτσι, συμπεραίνουμε ότι η ηλικία και τα επίπεδα της γλυκόζης δεν σχετίζονται γραμμικά.

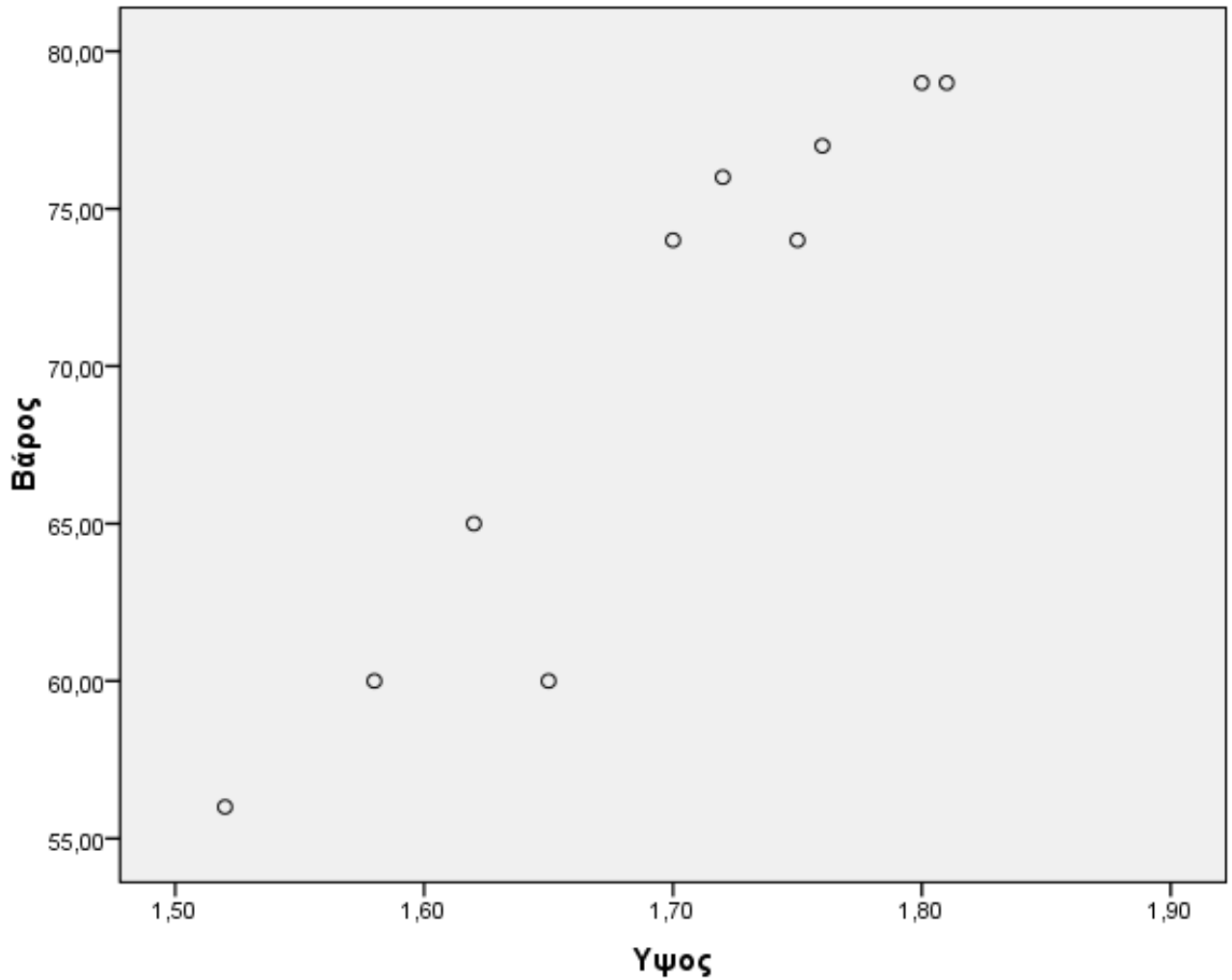
DF	A = 0.1	0.05	0.025	0.01	0.005	0.001	0.0005
∞	ta = 1.282	1.645	1.96	2.326	2.576	3.091	3.291
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.12	2.583	2.921	3.686	4.015
17	1.333	1.74	2.11	2.567	2.898	3.646	3.965
18	1.33	1.734	2.101	2.552	2.878	3.61	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.85
21	1.323	1.721	2.08	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.5	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.06	2.485	2.787	3.45	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.66
30	1.31	1.697	2.042	2.457	2.75	3.385	3.646
60	1.296	1.671	2	2.39	2.66	3.232	3.46
120	1.289	1.658	1.98	2.358	2.617	3.16	3.373
1000	1.282	1.646	1.962	2.33	2.581	3.098	3.3

Παράδειγμα 2

- Ο παρακάτω πίνακας δίνει τα ύψη και τα αντίστοιχα βάρη 10 μαθητών

Υψος σε m (X)	Βάρος σε κιλά (Y)
1,52	56
1,58	60
1,62	65
1,65	60
1,70	74
1,72	76
1,75	74
1,76	77
1,8	79
1,81	79

- Υπάρχει γραμμική συσχέτιση του ύψους με το βάρος;


80,00
75,00
70,00
65,00
60,00
55,00
Βάρος
1,50
1,60
1,70
1,80
1,90
Υψος

- Ho: Το ύψος δεν συσχετίζεται γραμμικά με το βάρος
- H1: Το ύψος συσχετίζεται γραμμικά με το βάρος

	Υψος σε m (Χ)	Βάρος σε κιλά (Υ)	ΧΥ	$Χ^2$	$Υ^2$
	1,52	56	85,12	2,3104	3136
	1,58	60	94,8	2,4964	3600
	1,62	65	105,3	2,6244	4225
	1,65	60	99	2,7225	3600
	1,70	74	125,8	2,89	5476
	1,72	76	130,72	2,9584	5776
	1,75	74	129,5	3,0625	5476
	1,76	77	135,52	3,0976	5929
	1,8	79	142,2	3,24	6241
	1,81	79	142,99	3,2761	6241
ΣΥΝΟΛΟ	16.91	700	1190,95	28,6783	49700

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{16{,}91}{10} = 1{,}691 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{700}{10} = 70$$

- Συντελεστής συσχέτισης Pearson:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} = \frac{1190{,}95 - 10 \cdot 1{,}691 \cdot 70}{\sqrt{28{,}6783 - 10 \cdot 1{,}691^2} \cdot \sqrt{49700 - 10 \cdot 70^2}}$$

$$= \frac{7{,}25}{0{,}289 \cdot 26{,}458} = 0{,}948$$

Το στατιστικό κριτήριο : $$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} = 0{,}948 \cdot \sqrt{\frac{10-2}{1-0{,}948^2}} = 0{,}948 \cdot 8{,}887 = 8{,}425$$

Η κρίσιμη τιμή: $t_{n-2;\alpha} = t_{8;0,05} = 1{,}86$

Η κρίσιμη περιοχή: $R = \{t > t_{n-2;\alpha}\}$

Άρα βρισκόμαστε στην κρίσιμη περιοχή, κι επομένως απορρίπτουμε την μηδενική υπόθεση. Δηλαδή το ύψος σχετίζεται γραμμικά με το βάρος.