Βιοστατιστική
Συσχέτιση – Συντελεστές συσχέτισης

Χαράλαμπος Γναρδέλλης

Τμήμα Ζωικής Παραγωγής Αλιείας και Υδατοκαλλιεργειών Πανεπιστημίου Πατρών

Συσχέτιση

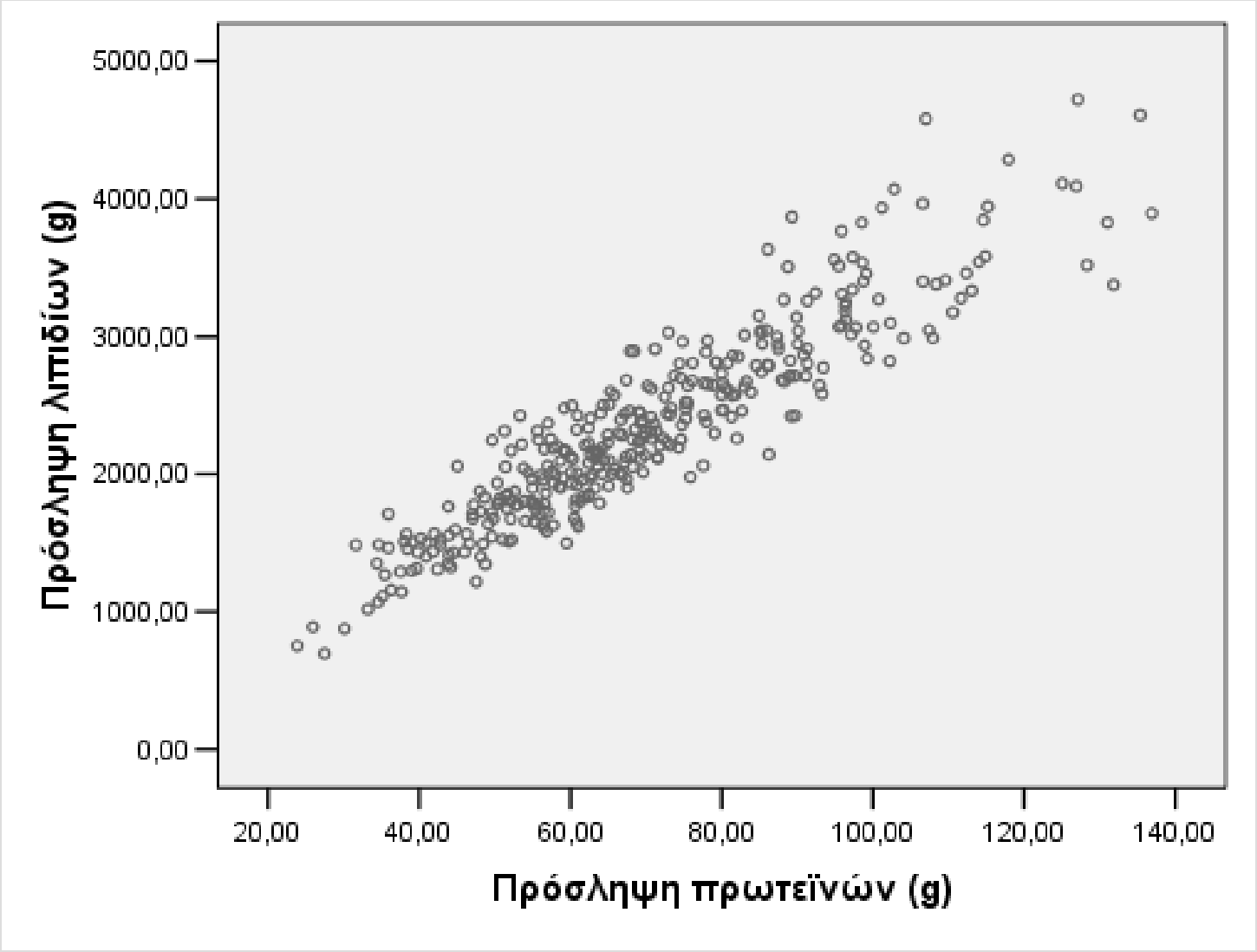
- **Με τον όρο *συσχέτιση* (*correlation*) εννοούμε το βαθμό στον οποίο συμμεταβάλλονται δύο ποσοτικές μεταβλητές υπό την προϋπόθεση ότι η σχέση τους είναι γραμμική.**

- **Στην πραγματικότητα υπάρχουν διάφοροι τρόποι με τους οποίους μπορούν να σχετίζονται οι τιμές δύο ποσοτικών μεταβλητών και είναι απαραίτητο, προτού γίνει οποιοσδήποτε προσδιορισμός της σχέσης τους, να οριστεί πρώτα η συναρτησιακή της μορφή.**

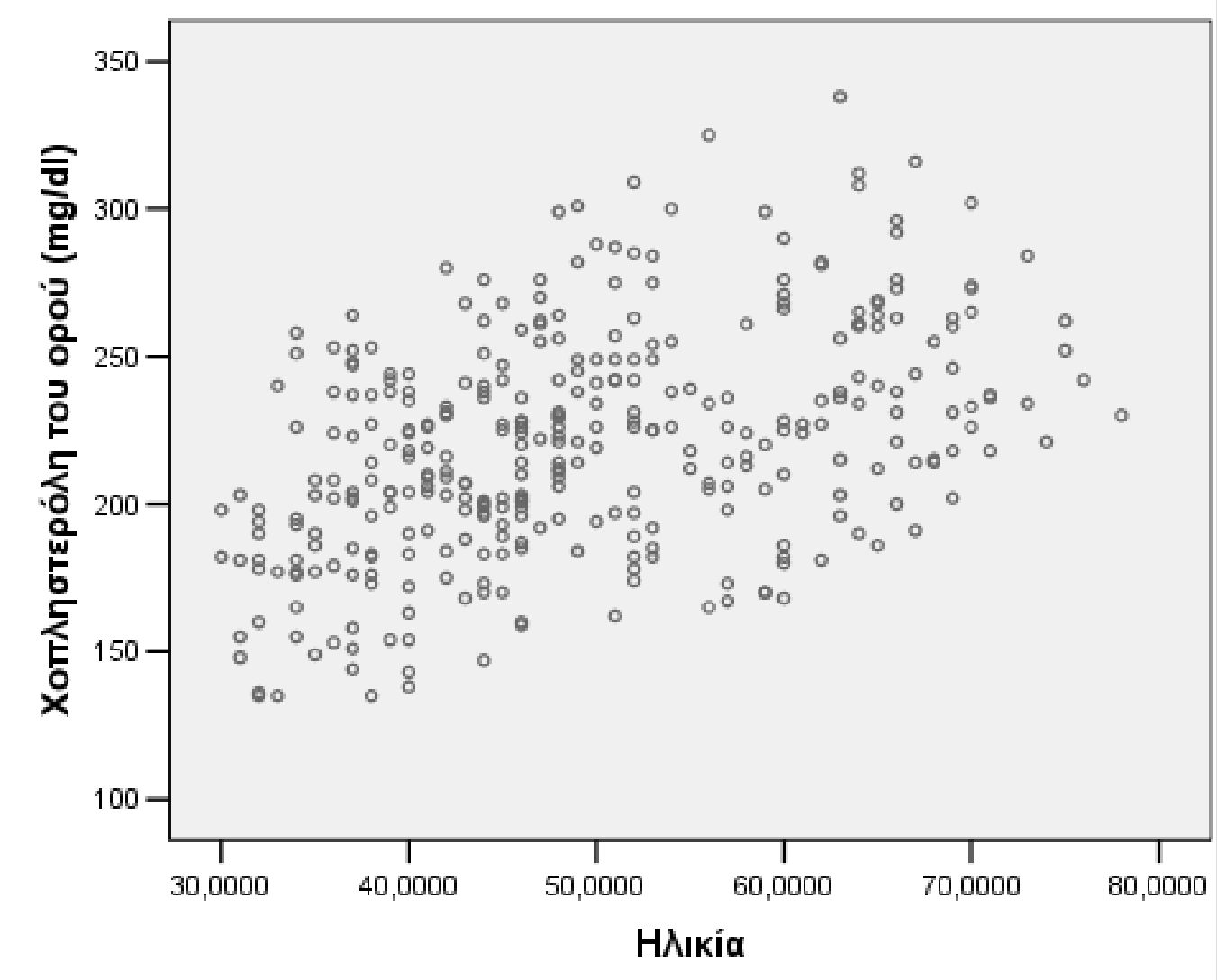
- **Η συνήθης παραδοχή που γίνεται για τη σχέση δύο ποσοτικών μεταβλητών *X* και *Y* είναι ότι αυτή είναι γραμμική (δηλαδή ότι οι δύο μεταβλητές συμμεταβάλλονται μονότονα).**

- **Αυτό πρακτικά σημαίνει ότι η συνδυασμένη απεικόνιση των δύο μεταβλητών σε ένα διάγραμμα διασποράς, ορίζει ένα σύνολο σημείων τα οποία τείνουν να συσσωρεύονται κατά μήκος μιας ευθείας γραμμής .**

Διάγραμμα διασποράς δύο γραμμικά συσχετιζόμενων ποσοτικών μεταβλητών

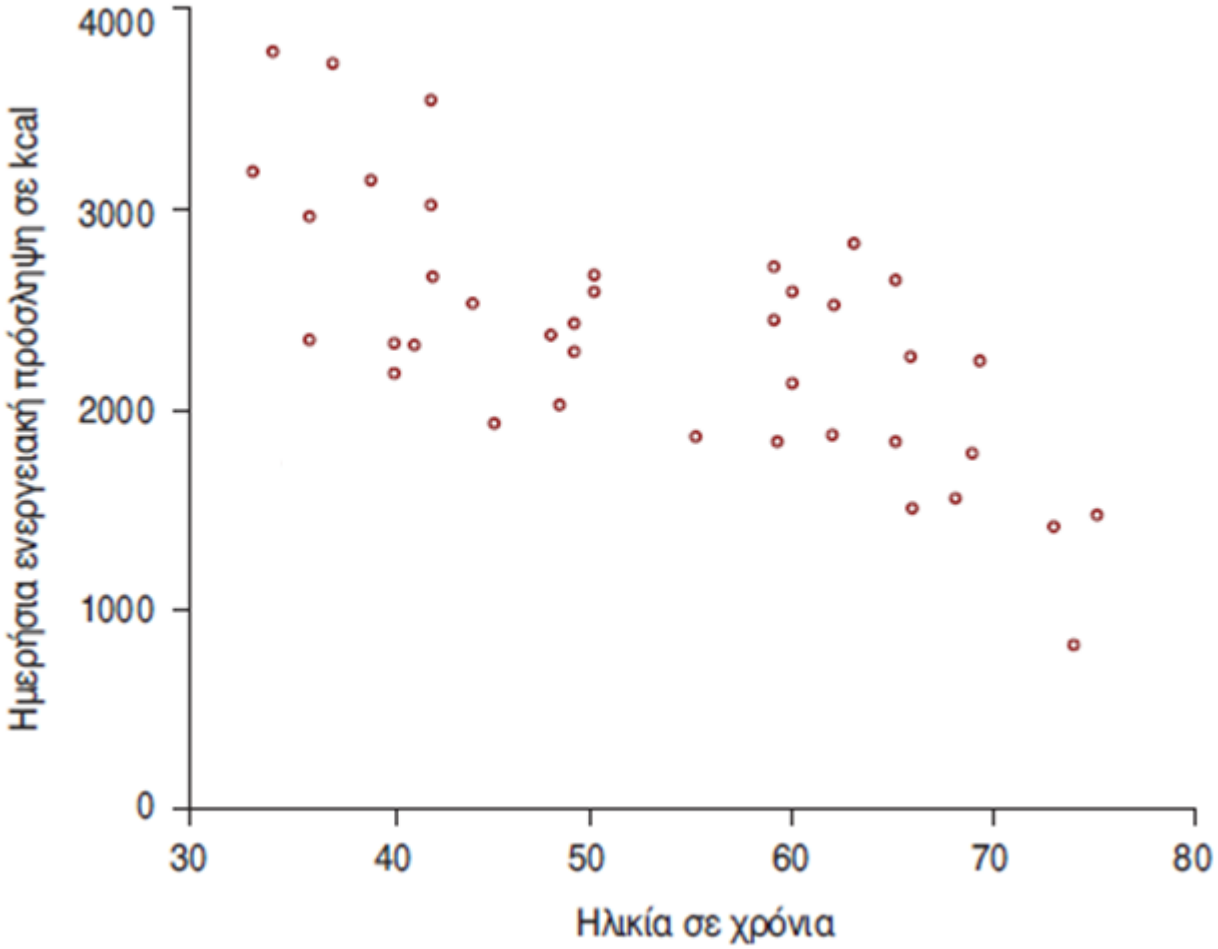


Ασθενής γραμμική σχέση μεταξύ δύο ποσοτικών μεταβλητών



Αρνητική συσχέτιση μεταξύ δύο ποσοτικών μεταβλητών

Διάγραμμα διασποράς της ηλικίας και της ημερήσιας ενεργειακής πρόσληψης 40 ενηλίκων



Συντελεστής συσχέτισης του Pearson

- **Η συσχέτιση δύο ποσοτικών μεταβλητών *X* και *Y* προσδιορίζεται αριθμητικά μέσω του *συντελεστή συσχέτισης του Pearson (Pearson's correlation coefficient)*. Ο συντελεστής συσχέτισης του Pearson ορίζεται από τη σχέση**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

όπου x_i και y_i , i=1,2, …,n είναι οι τιμές των δύο μεταβλητών *X* και *Y* και s_x, s_y, οι τυπικές τους αποκλίσεις.

Συντελεστής συσχέτισης του Pearson

- **Ο συντελεστής συσχέτισης του Pearson μπορεί να τροποποιηθεί ως εξής:**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sqrt{\frac{1}{(n-1)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sqrt{\frac{1}{(n-1)^2}} \cdot \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

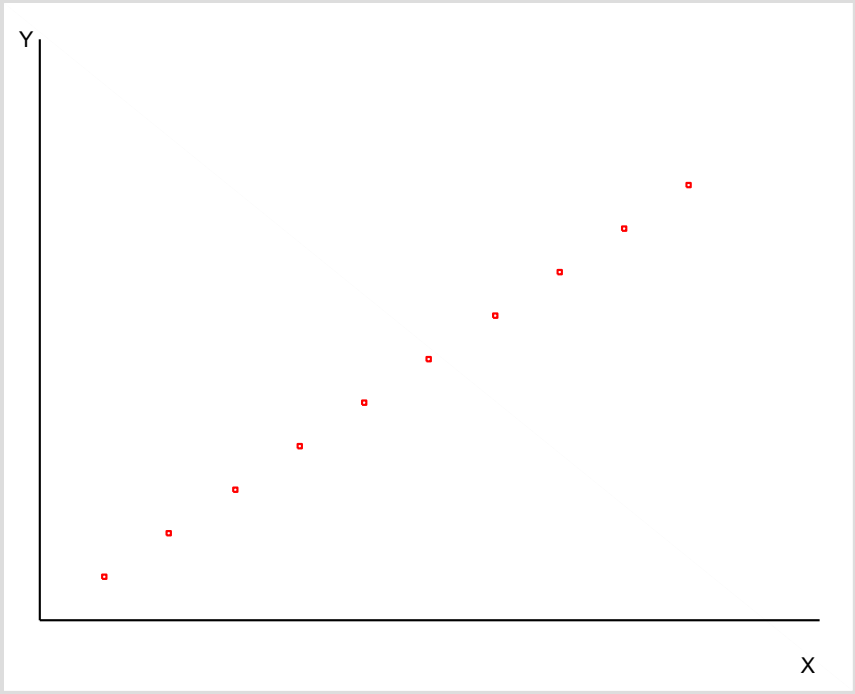
Άρα, εναλλακτικά ο συντελεστής συσχέτισης του Pearson μπορεί να υπολογιστεί και από τον τύπο:

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

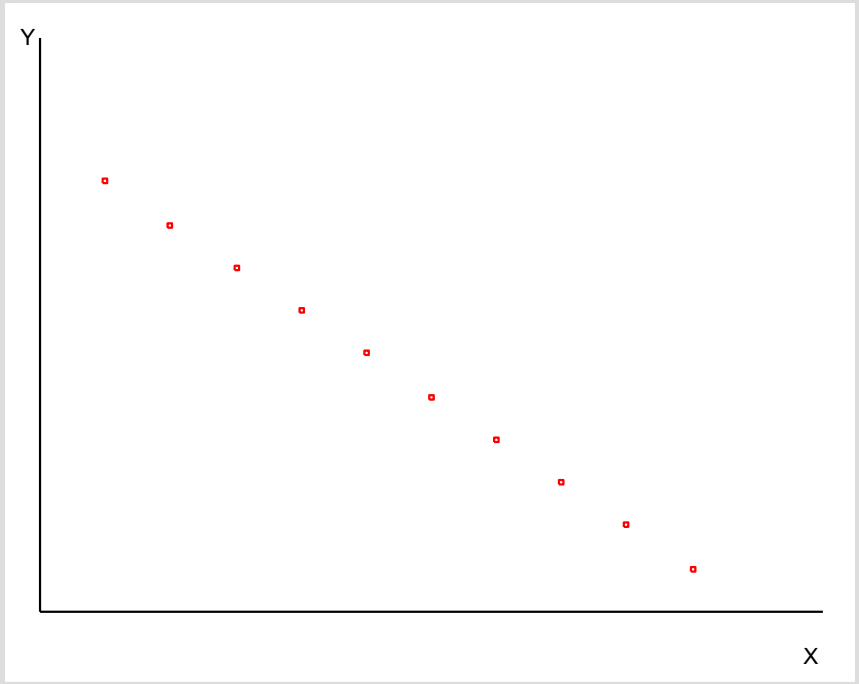
- **Ο συντελεστής συσχέτισης είναι ανεξάρτητος μονάδων και το εύρος των δυνατών τιμών του είναι το διάστημα [-1, 1]. Οι τιμές r = -1 και r =1 προκύπτουν όταν υπάρχει πλήρης γραμμική σχέση μεταξύ των δύο μεταβλητών X και Y. Όταν, δηλαδή, τα σημεία του αντίστοιχου διαγράμματος διασποράς που ορίζεται από τα ζεύγη των τιμών (x_i, y_i), βρίσκονται κατά μήκος μιας ευθείας γραμμής.**

Διαγράμματα διασποράς που εμφανίζουν ακριβείς γραμμικές σχέσεις

Πλήρης θετική συσχέτιση r = 1



Πλήρης αρνητική συσχέτιση r = -1



- **Όσο η σχέση μεταξύ των *X* και *Y* αποκλίνει από την πλήρη γραμμικότητα, η τιμή του *r* τείνει να απομακρύνεται από τις τιμές -1 και 1 και να πλησιάζει το 0.**

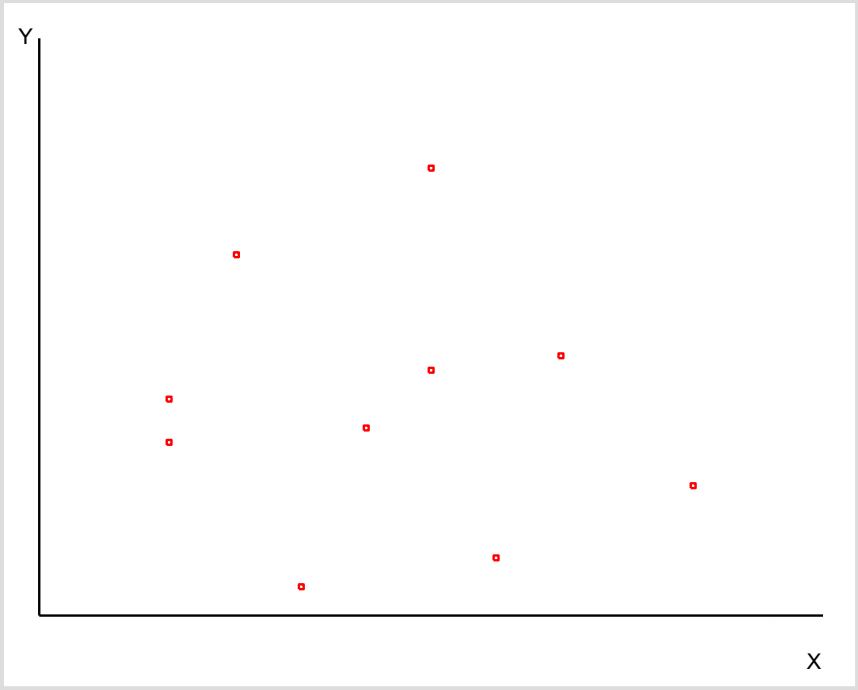
- **Όταν οι τιμές της *Y* τείνουν να αυξάνουν όσο αυξάνουν και οι αντίστοιχες τιμές της *X*, η τιμή του *r* είναι θετική και οι μεταβλητές χαρακτηρίζονται *θετικά συσχετιζόμενες*.**

- **Στην αντίστροφη περίπτωση, όπου οι τιμές της *Y* ελαττώνονται όσο οι τιμές της *X* αυξάνουν, ο συντελεστής συσχέτισης *r* παίρνει αρνητικές τιμές και οι δύο μεταβλητές χαρακτηρίζονται *αρνητικά συσχετιζόμενες*.**

- **Αν η τιμή του συντελεστή συσχέτισης είναι *r* = 0, τότε μεταξύ των δύο μεταβλητών δεν υπάρχει γραμμική σχέση. Σε μια τέτοια περίπτωση, όμως, μπορεί να υπάρχει μη γραμμική σχέση μεταξύ των δύο μεταβλητών.**

Διαγράμματα διασποράς που απεικονίζουν την απουσία γραμμικής σχέσης μεταξύ των δύο μεταβλητών

$r = 0$



$r = 0$