Επαγωγική Στατιστική

Συσχέτιση – Συντελεστές συσχέτισης

Χαράλαμπος Γναρδέλλης

Πρόγραμμα Μεταπτυχιακών Σπουδών
Βιώσιμη Αλιεία, Υδατοκαλλιέργεια

Συσχέτιση

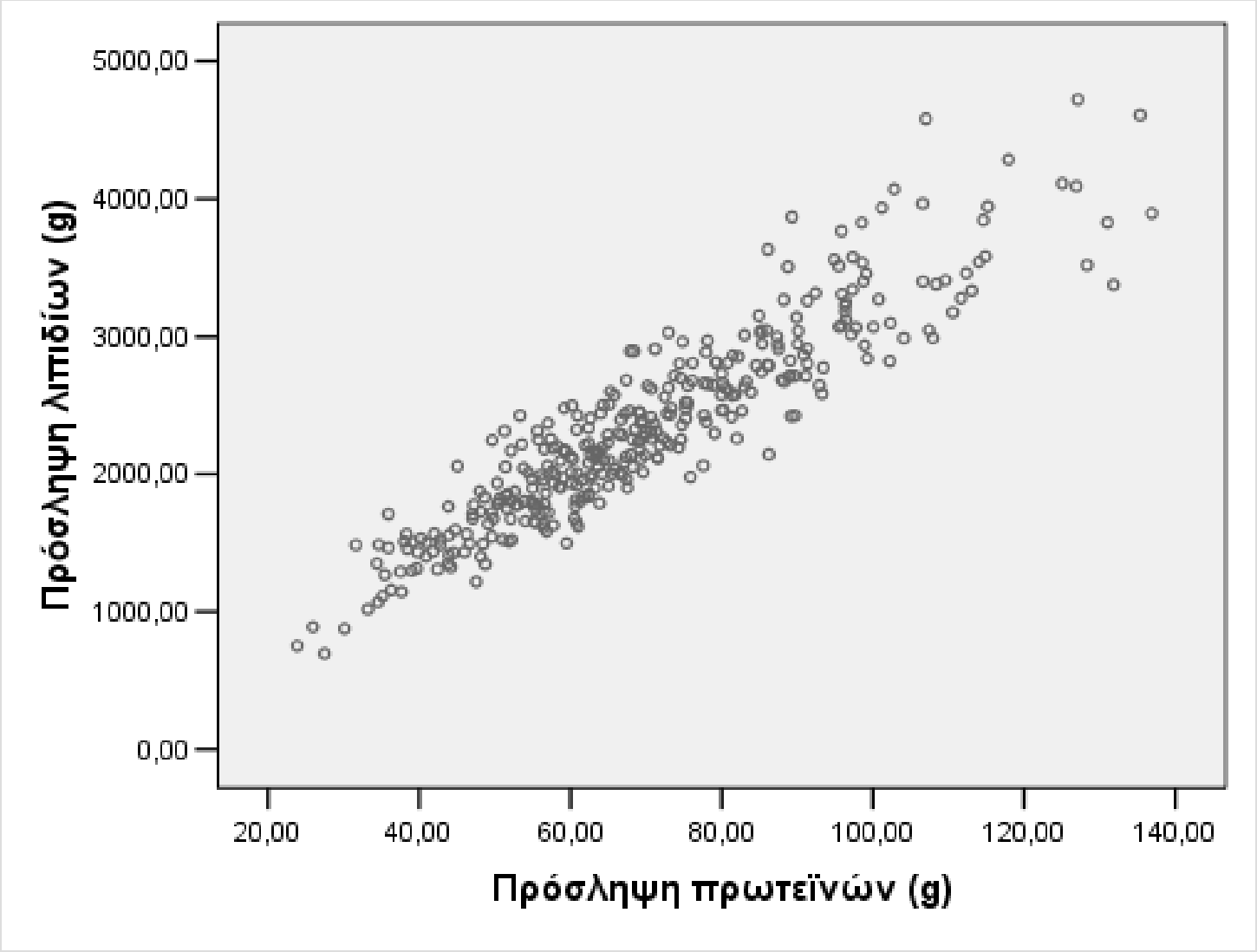
- **Με τον όρο *συσχέτιση* (*correlation*) εννοούμε το βαθμό στον οποίο συμμεταβάλλονται δύο ποσοτικές μεταβλητές υπό την προϋπόθεση ότι η σχέση τους είναι γραμμική.**

- **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.**

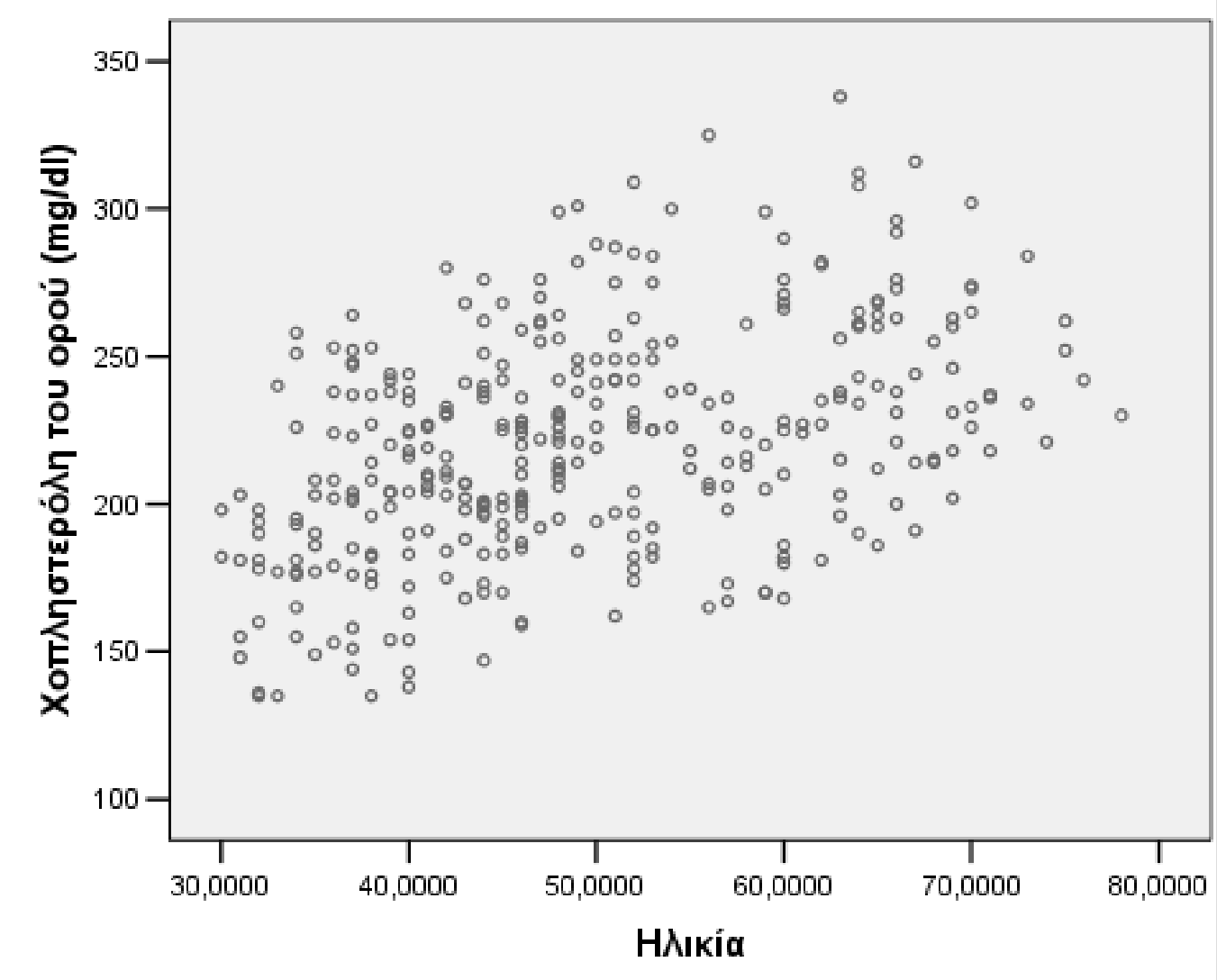
- **Η συνήθης παραδοχή που γίνεται για τη σχέση δύο ποσοτικών μεταβλητών *X* και *Y* είναι ότι αυτή είναι γραμμική (δηλαδή ότι οι δύο μεταβλητές συμμεταβάλλονται μονότονα).**

- **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



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**

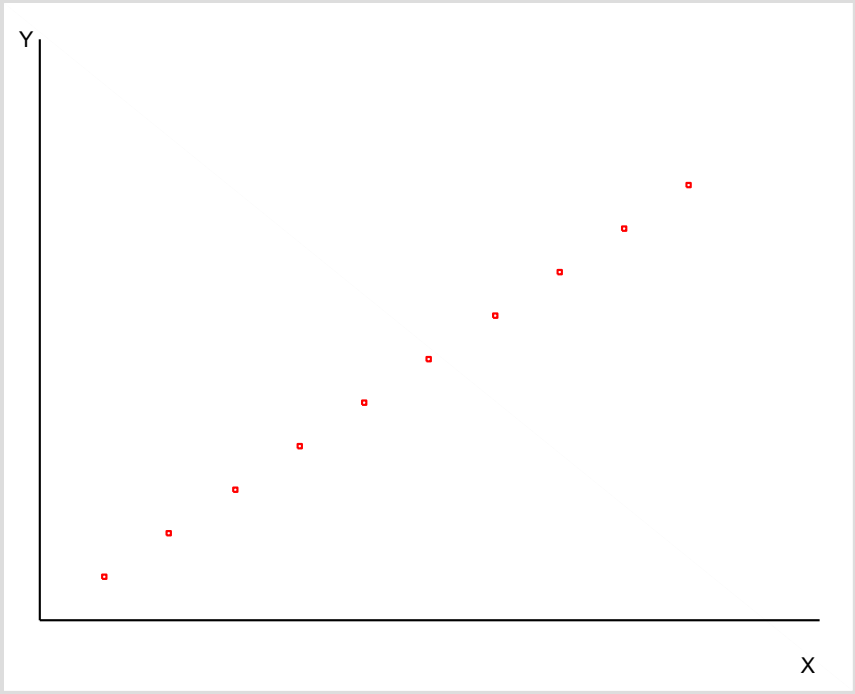
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

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.

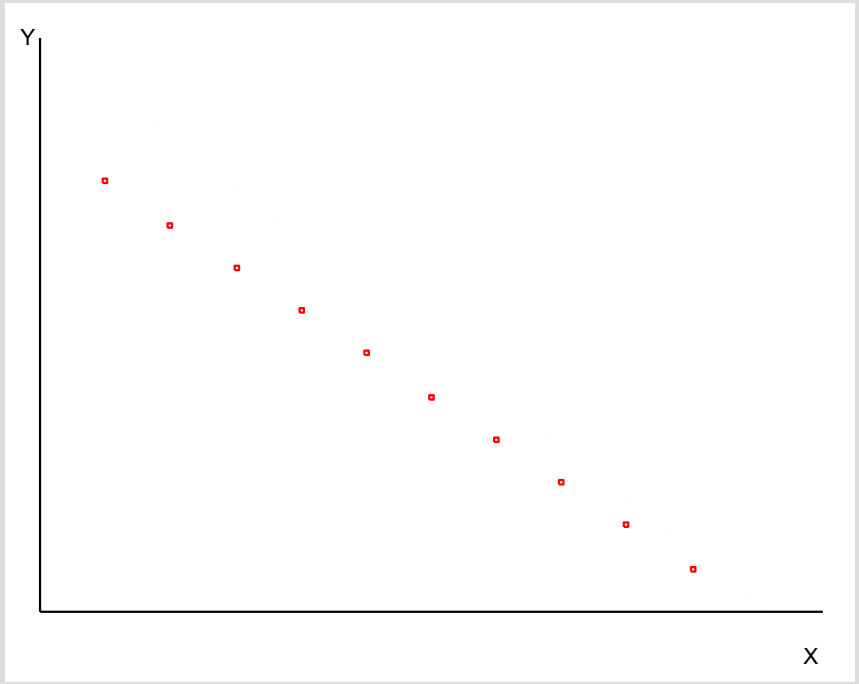
- **Ο συντελεστής συσχέτισης του Pearson είναι ανεξάρτητος μονάδων και το εύρος των δυνατών τιμών του είναι το διάστημα [-1, 1]. Οι τιμές r = -1 και r =1 προκύπτουν όταν υπάρχει πλήρης γραμμική σχέση μεταξύ των δύο μεταβλητών X και Y. Όταν, δηλαδή, τα σημεία του αντίστοιχου διαγράμματος διασποράς που 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.**

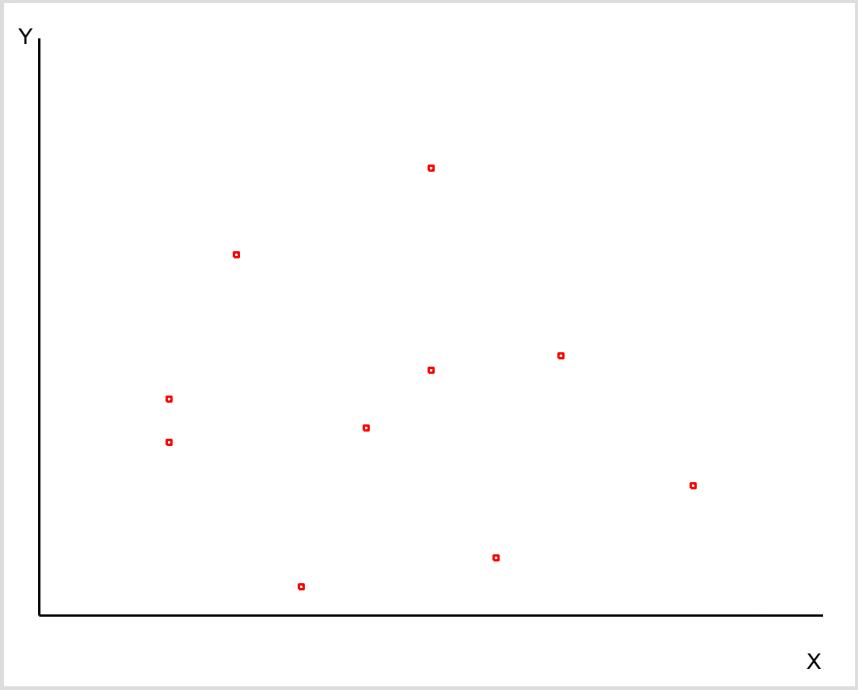
- **Όταν οι τιμές της *Y* τείνουν να αυξάνουν όσο αυξάνουν και οι αντίστοιχες τιμές της *X*, η τιμή του *r* είναι θετική και οι μεταβλητές χαρακτηρίζονται *θετικά συσχετιζόμενες*.**

- **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*.**

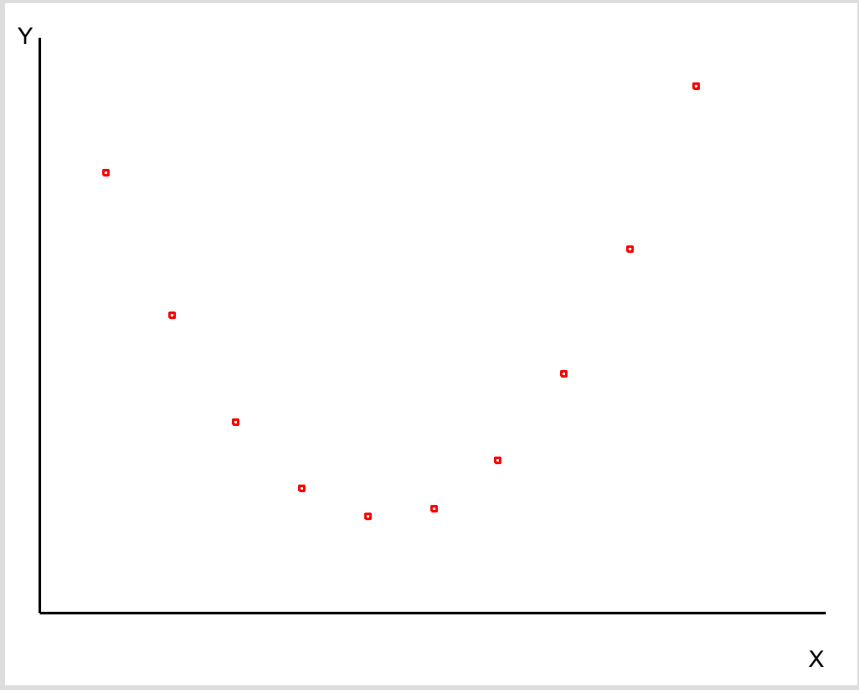
- **Αν η τιμή του συντελεστή συσχέτισης είναι *r* = 0, τότε μεταξύ των δύο μεταβλητών δεν υπάρχει γραμμική σχέση. Σε μια τέτοια περίπτωση, όμως, μπορεί να υπάρχει μη γραμμική σχέση μεταξύ των δύο μεταβλητών.**

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

$r = 0$



$r = 0$



- **Επειδή ο συντελεστής συσχέτισης *r* υπολογίζεται από δειγματικά δεδομένα, εναπομένει να αξιολογηθεί επαγωγικά.**
- **Δηλαδή, όπως έχουμε τη δυνατότητα να συγκρίνουμε έναν πληθυσμιακό μέσο *μ* ως προς μια προκαθορισμένη αριθμητική μ_0 βασιζόμενοι σε δειγματικά δεδομένα, έτσι και στην περίπτωση του πληθυσμιακού συντελεστή συσχέτισης *ρ* έχουμε τη δυνατότητα να διατυπώσουμε κάποιο συμπέρασμα για την τιμή του, βασιζόμενοι στο δειγματικό συντελεστή *r*.**
- **Ο έλεγχος του πληθυσμιακού συντελεστή συσχέτισης γίνεται ως προς την τιμή 0, διότι η τιμή αυτή υποδηλώνει την απουσία γραμμικής σχέσης μεταξύ των δύο μεταβλητών.**

Η μηδενική υπόθεση που ελέγχεται είναι η

$$H_0 : \rho = 0$$

έναντι της εναλλακτικής

$$H_A : \rho \neq 0$$

Ο έλεγχος γίνεται με τη βοήθεια του κριτηρίου

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

το οποίο ακολουθεί την κατανομή *t* με *n*-2 βαθμούς ελευθερίας, εφόσον οι μεταβλητές *X* και *Y* είναι κανονικά κατανεμημένες.

Αν, επομένως, η τιμή της ποσότητας *t* είναι αρκετά μεγάλη ώστε να ορίζει συμμετρικά στη δεξιά και την αριστερή ουρά της αντίστοιχης κατανομής *t* δύο περιοχές συνολικού εμβαδού μικρότερου από 0,05 (γεγονός που σημαίνει ότι η πιθανότητα να προκύψει μια τόσο μεγάλη τιμή είναι $p<0,05$), τότε η υπόθεση της ισότητας του πληθυσμιακού συντελεστή συσχέτισης με το 0 απορρίπτεται.

- **Ο έλεγχος σημαντικότητας για το συντελεστή συσχέτισης του Pearson απαιτεί την κανονικότητα των κατανομών των δύο μεταβλητών *X* και *Y*. Εκτός τούτου, πρέπει να επισημανθεί ότι όταν υπάρχουν ακραίες τιμές στις δύο μεταβλητές ή υπάρχουν ζεύγη παρατηρήσεων των οποίων τα σημεία διαφοροποιούνται πολύ των υπολοίπων επί του διαγράμματος διασποράς, η τιμή του συντελεστή *r* πρέπει να αντιμετωπίζεται με επιφύλαξη.**

- **Γενικά, ο συγκεκριμένος συντελεστής συσχέτισης είναι πολύ ευαίσθητος στην ύπαρξη ακραίων τιμών και, αν υπάρχουν ένα ή περισσότερα ζεύγη ακραίων τιμών (x_i, y_i), μπορεί να οδηγήσουν σε αλλοίωση του αριθμητικού περιεχομένου του.**

Συντελεστής συσχέτισης του Spearman

Ο *συντελεστής συσχέτισης του Spearman (Spearman's rank correlation coefficient)*, είναι ουσιαστικά ο συντελεστής συσχέτισης του Pearson, υπολογιζόμενος, όμως, όχι στις δειγματικές τιμές x_i και y_i, i=1,2, …,n των δύο μεταβλητών X και Y, αλλά στις σχετικές θέσεις (ranks) αυτών των τιμών. Αποδεικνύεται ότι ο συντελεστής συσχέτισης του Spearman, μπορεί να υπολογιστεί από τη σχέση

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

όπου n είναι ο αριθμός των παρατηρήσεων του δείγματος και d_i είναι η διαφορά των σχετικών θέσεων x_{ri} και y_{ri}, i=1,2, …,n των X και Y αντίστοιχα.

Όπως ο συντελεστής συσχέτισης του Pearson, έτσι και ο συντελεστής του Spearman παίρνει τιμές από -1 μέχρι 1. Τιμές του συντελεστή συσχέτισης πλησίον των τιμών -1 ή 1, υποδηλώνουν υψηλό βαθμό συσχέτισης μεταξύ των *X* και *Y*, ενώ τιμές πλησίον του 0 ορίζουν έλλειψη γραμμικής σχέσης μεταξύ των δύο μεταβλητών.

Ο συντελεστής συσχέτισης του Spearman μπορεί να χρησιμοποιηθεί για τον έλεγχο του αντίστοιχου πληθυσμιακού συντελεστή συσχέτισης ρ_s ως προς την τιμή 0. Το κριτήριο του ελέγχου είναι της ίδιας μορφής με αυτό που χρησιμοποιήθηκε και στην περίπτωση του συντελεστή συσχέτισης του Pearson. Δηλαδή

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

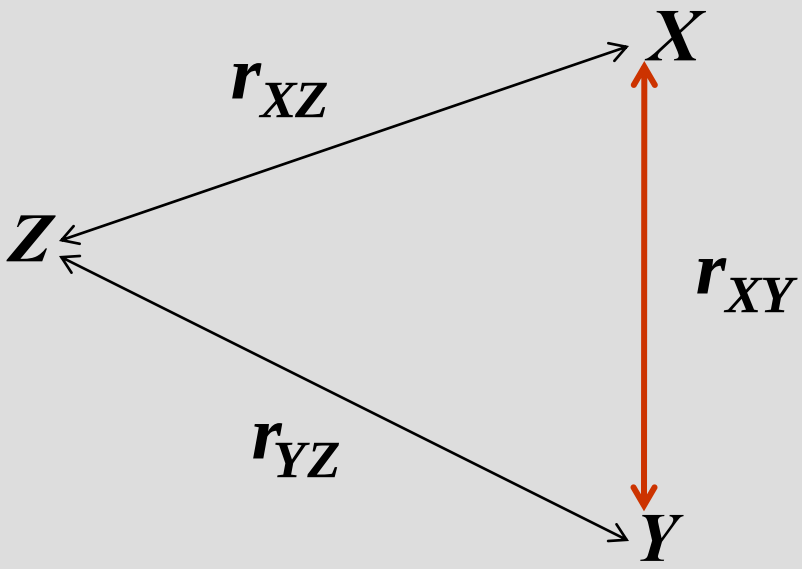
Η παραπάνω ποσότητα, υπό την υπόθεση $H_0: \rho_s = 0$, ακολουθεί την κατανομή *t* με *n*-2 βαθμούς ελευθερίας.

- **Ο συντελεστής συσχέτισης του Spearman, λόγω του ότι ο υπολογισμός του βασίζεται στις σχετικές θέσεις των τιμών και όχι στις τιμές αυτές καθ' αυτές, είναι πολύ λιγότερο ευαίσθητος από ότι ο συντελεστής του Pearson στην ύπαρξη ακραίων δειγματικών τιμών.**
- **Επιπλέον, μπορεί να υπολογίζεται όχι μόνο για ποσοτικές μεταβλητές, αλλά και για διατεταγμένες.**
- **Βασικό του μειονέκτημα είναι ότι δεν εξαντλεί κατά τον υπολογισμό του όλη τη διαθέσιμη δειγματική πληροφορία, εφόσον δεν υπολογίζεται από τις πραγματικές τιμές των μεταβλητών, αλλά από τις σχετικές θέσεις των τιμών στην εσωτερική τους διάταξη.**

Μερική συσχέτιση – Partial correlation

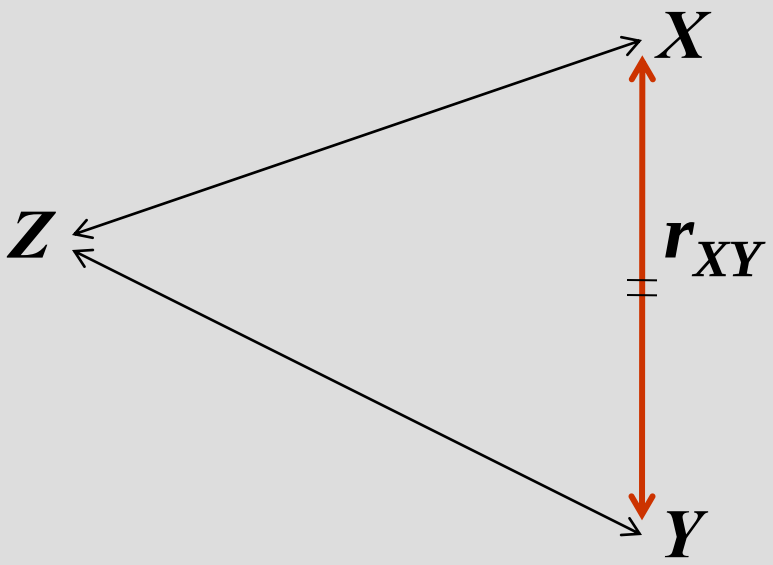
Ο συντελεστής μερικής συσχέτισης εκφράζει τη γραμμική σχέση που υπάρχει μεταξύ δύο ποσοτικών μεταβλητών *X* και *Y* όταν από αυτή αφαιρεθούν οι γραμμικές επιδράσεις μίας ή περισσοτέρων άλλων μεταβλητών.

Έστω ότι η μεταβλητή *Z* συσχετίζεται (στην πραγματικότητα) θετικά με τη μεταβλητή *X* και με τη μεταβλητή *Y*

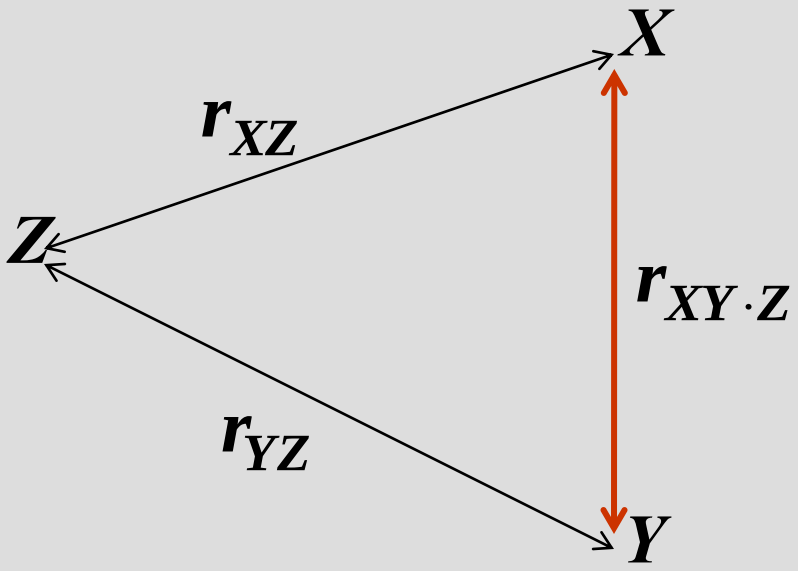


Λόγω της συσχέτισης αυτής, οι μεταβλητές *X* και *Y*, από τα διαθέσιμα δειγματικά δεδομένα, εμφανίζονται επίσης να συσχετίζονται θετικά μεταξύ τους με έναν υψηλό συντελεστή συσχέτισης r_{XY} .

Αν επιχειρήσουμε να ερμηνεύσουμε τη σχέση της *X* με τη *Y* αγνοώντας την επίδραση της *Z* –χρησιμοποιώντας δηλαδή το δειγματικό συντελεστή συσχέτισης r_{XY} –, ενδέχεται το συμπέρασμα στο οποίο θα καταλήξουμε να είναι εντελώς εσφαλμένο. Υπάρχει π.χ. το ενδεχόμενο, οι μεταβλητές *X* και *Y* να μην έχουν στην πραγματικότητα καμία σχέση, ενώ λόγω της κοινής θετικής σχέσης που έχουν με τη *Z* να εμφανίζουν μια υψηλή συσχέτιση μεταξύ τους.



Στην περίπτωση αυτή, αντί του απλού συντελεστή r_{XY}, μέτρο της πραγματικής γραμμικής σχέσης που υπάρχει μεταξύ των μεταβλητών X και της Y είναι ο μερικός συντελεστής συσχέτισης $r_{XY \cdot Z}$, ο οποίος συνοψίζει τη συσχέτιση των δύο μεταβλητών όταν από αυτήν απομακρυνθεί η γραμμική σχέση της Z τόσο με τη X όσο και με τη Y.



Ο συντελεστής μερικής συσχέτισης μεταξύ της X και Y, ελέγχοντας ως προς τις επιδράσεις της Z (απομακρύνοντας δηλαδή τις επιδράσεις της Z), δίνεται από την έκφραση

$$r_{XY \cdot Z} = \frac{\left(r_{XY} - r_{XZ} r_{YZ}\right)}{\sqrt{\left(1 - r_{XZ}^2\right)\left(1 - r_{YZ}^2\right)}}$$

όπου r_{XY} είναι ο συντελεστής συσχέτισης της X με τη Y και r_{XZ} , r_{YZ} οι συντελεστές συσχέτισης της X με τη Z και της Y με τη Z αντίστοιχα.

Το παράδειγμα που αναφέρθηκε επεκτείνεται και στην περίπτωση όπου οι μεταβλητές ως προς τις οποίες ελέγχεται (διορθώνεται) ο συντελεστής συσχέτισης δύο μεταβλητών, είναι περισσότερες από μία. Μπορεί, δηλαδή, να υπολογιστεί ο συντελεστής μερικής συσχέτισης δύο μεταβλητών *X* και *Y*, αφού πρώτα αφαιρεθούν από τη γραμμική σχέση των δύο μεταβλητών οι επιδράσεις ενός συνόλου άλλων μεταβλητών.