Συσχέτιση δύο μεταβλητών

Θεωρούμε δύο τυχαίες μεταβλητές X, Y και ν ζεύγη παρατηρήσεων (χ_1, y_1), (χ_2, y_2),...,(χ_ν, y_ν) από *τυχαίο δείγμα* μεγέθους ν.

Αναφερόμαστε, δηλαδή, σε **μη πειραματικά** δεδομένα (ο ερευνητής δεν προκαθορίζει-ελέγχει τις τιμές καμιάς από τις δύο μεταβλητές) όπως,

- *Χ το ύψος των φοιτητών ενός πανεπιστημιακού τμήματος και Υ το βάρος τους*
- *Χ οι ώρες μελέτης των φοιτητών ενός πανεπιστημιακού τμήματος και Υ η απόδοση τους σε ένα τεστ*
- *Χ οι εβδομάδες εμπειρίας ενός εργάτη σε μια επιχείρηση* και *Υ ο αριθμός των ελαττωματικών προϊόντων που παράγει*
- *Χ η κατάταξη δέκα προϊόντων από ένα κριτή και Υ η κατάταξη των ιδίων προϊόντων από έναν άλλο κριτή*
- *Χ ο αριθμός των πωλήσεων μουσικών CD σε μια περιοχή και Υ ο αριθμός των νέων στην ίδια περιοχή*.

Δεν αναφερόμαστε όμως σε περιπτώσεις όπως,

- *Χ ο αριθμός των ανοιχτών ταμείων ενός υποκαταστήματος τραπέζης (που καθορίζει ο διευθυντής)* και *Υ ο χρόνος αναμονής των πελατών*
- *Χ η ποσότητα λιπάσματος (που καθορίζει ο ερευνητής)* και *Υ η απόδοση του αγρού*
- *Χ το ύψος της διαφημιστικής δαπάνης ενός προϊόντος (που καθορίζει μια επιχείρηση)* και *Υ το ύψος των πωλήσεων του προϊόντος*.

Στις περιπτώσεις όπου από τον πληθυσμό επιλέγουμε ένα τυχαίο δείγμα και σε κάθε μονάδα του δείγματος μελετάμε δύο ή περισσότερα χαρακτηριστικά, είναι λογικό, να αναζητήσουμε μέτρα τα οποία να μπορούν να εκφράσουν και να ποσοτικοποιήσουν την πιθανή συμμεταβολή-συσχέτιση των χαρακτηριστικών. Για παράδειγμα, συσχετίζονται-συμμεταβάλλονται *ο μισθός* και *τα έτη σπουδών των εργαζομένων*; Πώς συμμεταβάλλονται; Δηλαδή, όταν αυξάνουν τα έτη σπουδών, αυξάνει ο μισθός του εργαζομένου; (μειώνεται μήπως;!). Ποσό ισχυρή είναι η συμμεταβολή των μεταβλητών *έτη σπουδών* και *μισθός*;

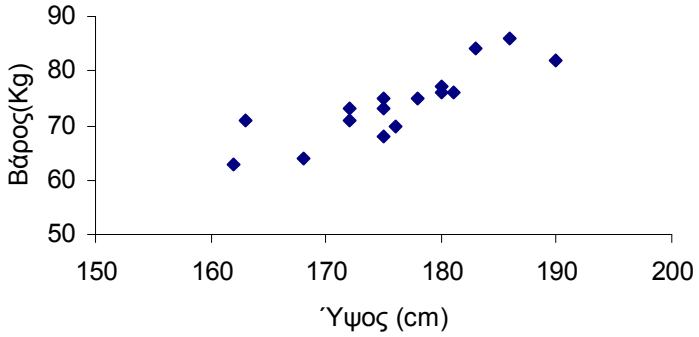
Ένας απλός τρόπος για να αποκτήσουμε μια πρώτη ιδέα για το αν και πώς δυο μεταβλητές συμμεταβάλλονται-συσχετίζονται, είναι να κατασκευάσουμε το ***διάγραμμα διασποράς (Scatter Diagram)***. Να αναπαραστήσουμε δηλαδή τα ζεύγη των παρατηρήσεων σε ένα διάγραμμα. Ας δούμε ένα παράδειγμα:

Στον πίνακα που ακολουθεί φαίνονται οι παρατηρήσεις που πήραμε για το ύψος και το βάρος 16 εργατών μιας βιομηχανίας.

	1	2	3	4	5	6	7	8
Ύψος (cm)	183	162	172	181	180	168	176	180
Βάρος (Kg)	84	63	71	76	77	64	70	76

	9	10	11	12	13	14	15	16
Ύψος (cm)	190	175	178	175	186	172	175	163
Βάρος (Kg)	82	68	75	73	86	73	75	71

Από το *διάγραμμα διασποράς* φαίνεται ότι οι εργάτες στο δείγμα που έχουν μεγαλύτερο ύψος έχουν και μεγαλύτερο βάρος. Φαίνεται, δηλαδή, να υπάρχει μια ανάλογη σχέση μεταξύ του ύψους και του βάρους των εργατών.



Πόσο ισχυρή είναι όμως αυτή η συσχέτιση; Πώς μπορεί, δηλαδή, να μετρηθεί; Στο πλαίσιο του μαθήματος, θα ασχοληθούμε με τρία μέτρα συσχέτισης: με το *συντελεστή γραμμικής συσχέτισης του Pearson*, με το *συντελεστή γραμμικής συσχέτισης του Spearman* και με το *δείκτη Kendall.*

1. Συντελεστής Γραμμικής Συσχέτισης του Pearson

Ο ***δειγματικός συντελεστής γραμμικής συσχέτισης του Pearson*** συμβολίζεται με ***r*** και ορίζεται από τον τύπο:

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

όπου,

$$s_{xy} = Cov(X,Y) = \frac{\sum_{i=1}^{\nu}(\chi_i - \overline{\chi}) \cdot (y_i - \overline{y})}{\nu - 1} = \frac{\sum_{i=1}^{\nu}\chi_i y_i - \nu \cdot \overline{\chi} \cdot \overline{y}}{\nu - 1}$$

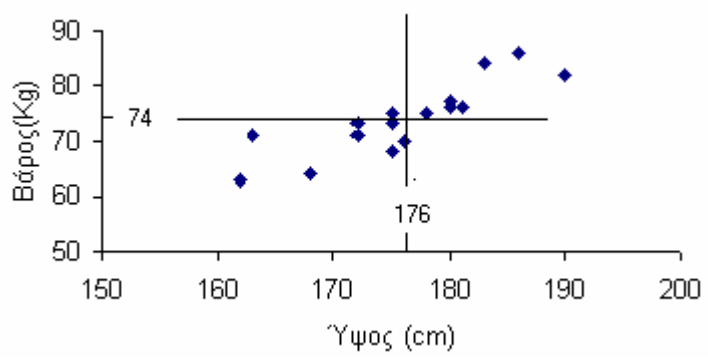
$$s_x = \sqrt{\frac{1}{\nu - 1}\sum_{i=1}^{\nu}(\chi_i - \overline{\chi})^2} \quad \text{και} \quad s_y = \sqrt{\frac{1}{\nu - 1}\sum_{i=1}^{\nu}(y_i - \overline{y})^2}$$

Επομένως

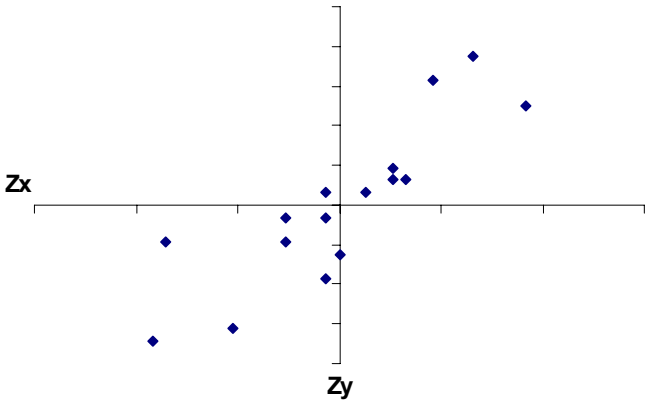
$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{\nu}(\chi_i - \overline{\chi}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{\nu}(\chi_i - \overline{\chi})^2} \cdot \sqrt{\sum_{i=1}^{\nu}(y_i - \overline{y})^2}} = \frac{\sum_{i=1}^{\nu} x_i y_i - \nu \cdot \overline{\chi} \cdot \overline{y}}{\sqrt{\sum_{i=1}^{\nu}\chi_i^2 - \nu \cdot \overline{\chi}^2} \cdot \sqrt{\sum_{i=1}^{\nu} y_i^2 - \nu \cdot \overline{y}^2}}$$

Ο ***πληθυσμιακός συντελεστής γραμμικής συσχέτισης του Pearson*** ορίζεται ανάλογα και συμβολίζεται με ρ.

Ερώτηση:
Ποια είναι η βασική ιδέα στον ορισμό του συντελεστή του Pearson;
(Στο παράδειγμα, το μέσο ύψος είναι 176 cm και το μέσο βάρος 74Kg. Παρατηρείστε ότι οι εργάτες που έχουν ύψος πάνω από το μέσο ύψος έχουν (στις περισσότερες περιπτώσεις) και βάρος πάνω από το μέσο βάρος. Ανάλογα, οι εργάτες που έχουν ύψος κάτω από το μέσο ύψος έχουν (στις περισσότερες περιπτώσεις) και βάρος κάτω από το μέσο βάρος).



Δείτε και το ακόλουθο διάγραμμα:



Παρατηρείστε επίσης ότι $r = \dfrac{\sum_{i=1}^{\nu} z_{x_i} \cdot z_{y_i}}{\nu - 1}$ *(όπου* $z_{x_i} = \dfrac{\chi_i - \overline{\chi}}{s_x}$ *και* $z_{y_i} = \dfrac{y_i - \overline{y}}{s_y}$*).*

Ας δούμε δύο αριθμητικά παραδείγματα υπολογισμού του συντελεστή γραμμικής συσχέτισης του *Pearson*.

Παράδειγμα-1

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})\,.\,(y_i - \overline{y})$
1	4	-2	4	4	16	-8
2	2	-1	2	1	4	-2
3	0	0	0	0	0	0
4	-2	1	-2	1	4	-2
5	-4	2	-4	4	16	-8
$\sum x_i = 15$	$\sum y_i = 0$			$\sum (x_i - \overline{x})^2 = 10$	$\sum (y_i - \overline{y})^2 = 40$	$\sum (x_i - \overline{x}) \cdot (y_i - \overline{y}) = -20$

$\overline{x} = 3,\ \overline{y} = 0$

$$r = \frac{\sum_{i=1}^{5} (\chi_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{5} (\chi_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{5} (y_i - \overline{y})^2}} = \frac{-20}{\sqrt{10} \cdot \sqrt{40}} = -1$$

Παράδειγμα –2

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	-2	1	4	-2
3	0	9	0	0
5	1	25	1	5
7	3	49	9	21
9	5	81	25	45
10	6	100	36	60
12	8	144	64	96
13	10	169	100	130
$\sum x_i = 60$	$\sum y_i = 31$	$\sum x_i^2 = 578$	$\sum y_i^2 = 239$	$\sum x_i \cdot y_i = 355$

$\bar{x} = 7{,}5 \;\; και \;\; \bar{y} = 3{,}9$

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{\nu} x_i y_i - \nu \cdot \bar{\chi} \cdot \bar{y}}{\sqrt{\sum_{i=1}^{\nu} \chi_i^2 - \nu \cdot \bar{\chi}^2} \cdot \sqrt{\sum_{i=1}^{\nu} y_i^2 - \nu \cdot \bar{y}^2}} = \frac{355 - 8 \cdot 7{,}5 \cdot 3{,}9}{\sqrt{578 - 8 \cdot 7{,}5^2} \cdot \sqrt{239 - 8 \cdot 3{,}9^2}} = 0{,}99$$

Ερμηνεία και ιδιότητες του συντελεστή γραμμικής συσχέτισης *r*

- Ο συντελεστής **γραμμικής** συσχέτισης r δίνει ένα μέτρο του μεγέθους της **γραμμικής** συσχέτισης μεταξύ δύο μεταβλητών.
- Παίρνει τιμές στο κλειστό διάστημα **[-1 , 1]**

Απόδειξη:

Για κάθε $\lambda \in R$ ισχύει $\sum \left((x_i - \bar{x}) + \lambda (y_i - \bar{y}) \right)^2 \geq 0$.

Άρα $\sum \left((x_i - \bar{x})^2 + \lambda^2 (y_i - \bar{y})^2 + 2\lambda (x_i - \bar{x})(y_i - \bar{y}) \right) \geq 0$ ή

$\sum (x_i - \bar{x})^2 + \lambda^2 \sum (y_i - \bar{y})^2 + 2\lambda \sum (x_i - \bar{x})(y_i - \bar{y}) \geq 0$ ή

$\lambda^2 \cdot \sum (y_i - \bar{y})^2 + \lambda \cdot 2 \sum (x_i - \bar{x})(y_i - \bar{y}) + \sum (x_i - \bar{x})^2 \geq 0$.

Επειδή η τελευταία ανισότητα ισχύει για κάθε $\lambda \in R$, θα είναι $\beta^2 - 4\alpha\gamma \leq 0$ και άρα

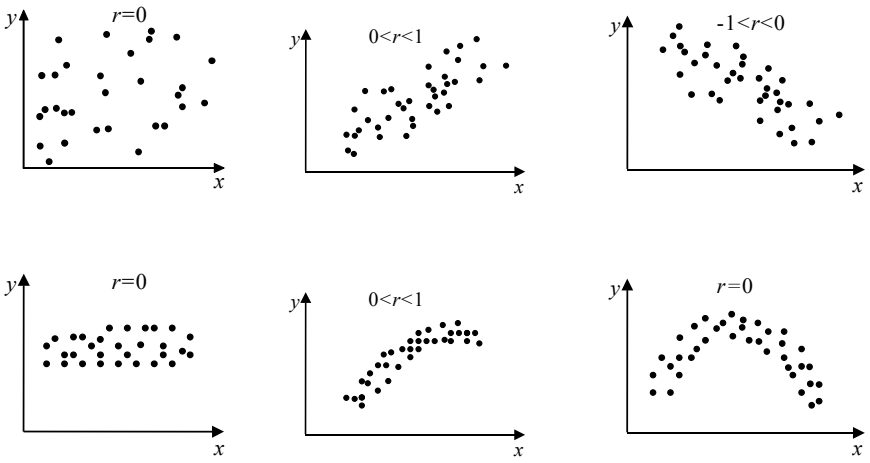
$$4 \cdot \left(\sum (x_i - \bar{x})(y_i - \bar{y}) \right)^2 \leq 4 \cdot \sum (y_i - \bar{y})^2 \cdot \sum (x_i - \bar{x})^2 \Leftrightarrow$$

$$\left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2} \sqrt{\sum (x_i - \bar{x})^2}} \right)^2 \leq 1 \Leftrightarrow r^2 \leq 1 \Leftrightarrow |r| \leq 1 \Leftrightarrow -1 \leq r \leq 1.$$

- Αν $r = \pm 1$ υπάρχει **τέλεια γραμμική** συσχέτιση.
  Αν $-0{,}3 \leq r < 0{,}3$ **δεν υπάρχει γραμμική** συσχέτιση. Αυτό, όμως, δεν σημαίνει ότι δεν υπάρχει άλλου είδους συσχέτιση μεταξύ των δύο μεταβλητών.
  Αν $-0{,}5 < r \leq -0{,}3$ *ή* $0{,}3 \leq r < 0{,}5$ υπάρχει **ασθενής γραμμική** συσχέτιση.
  Αν $-0{,}7 < r \leq -0{,}5$ *ή* $0{,}5 \leq r < 0{,}7$ υπάρχει **μέση γραμμική** συσχέτιση.
  Αν $-0{,}8 < r \leq -0{,}7$ *ή* $0{,}7 \leq r < 0{,}8$ υπάρχει **ισχυρή γραμμική** συσχέτιση.
  Αν $-1 < r \leq -0{,}8$ *ή* $0{,}8 \leq r < 1$ υπάρχει **πολύ ισχυρή γραμμική** συσχέτιση.
- Θετικές τιμές του r δεν υποδηλώνουν, κατ' ανάγκην μεγαλύτερο βαθμό γραμμικής συσχέτισης από το βαθμό γραμμικής συσχέτισης που υποδηλώνουν αρνητικές τιμές του r. Ο βαθμός γραμμικής συσχέτισης καθορίζεται από την απόλυτη τιμή του r και όχι από το πρόσημο του r. Το πρόσημο του r καθορίζει το είδος, μόνο, της συσχέτισης (θετική ή αρνητική). Μας πληροφορεί δηλαδή για το αν αύξηση της μιας μεταβλητής αντιστοιχεί σε αύξηση ή σε μείωση της άλλης

μεταβλητής. Για παράδειγμα η τιμή $r = -0{,}9$ δείχνει ισχυρότερη γραμμική συσχέτιση από την τιμή $r = 0{,}8$ ενώ οι τιμές $r = -0{,}6$ *και* $r = 0{,}6$ δείχνουν ίδιο βαθμό γραμμικής συσχέτισης αλλά αντίθετο είδος.

- Στην πράξη, υπολογίζουμε το συντελεστή γραμμικής συσχέτισης στις περιπτώσεις μόνο που το διάγραμμα διασποράς (στικτό διάγραμμα) έχει σχήμα *επιμήκους κεκλιμένης έλλειψης* ή *πλατυσμένου* ***J***. Αν, όμως, τον υπολογίσουμε και σε περιπτώσεις που το διάγραμμα διασποράς έχει άλλη μορφή, η τιμή του η οποία θα είναι μικρή, δεν συνεπάγεται <u>μη συσχέτιση</u> αλλά <u>μη γραμμική συσχέτιση</u>. Είναι, δηλαδή, δυνατόν να υπάρχει μεγάλη μη γραμμική συσχέτιση.



- Ο *συντελεστής γραμμικής συσχέτισης r* χρησιμοποιείται ως μια εκτιμήτρια του πληθυσμιακού συντελεστή γραμμικής συσχέτισης ρ, μόνο όταν τα ζεύγη (χ_1, y_1), (χ_2, y_2),...,(χ_ν, y_ν) προέρχονται από τυχαία δειγματοληψία. Δεν έχει, επομένως, μεγάλη χρησιμότητα σε πειραματικές έρευνες, όπου οι τιμές της μιας μεταβλητής ελέγχονται-καθορίζονται από τον ερευνητή.
- Συσχέτιση δε σημαίνει αιτιότητα

  Όταν σε μια **μη** πειραματική έρευνα (δειγματοληψία) δύο μεταβλητές *X* και *Y* βρίσκονται συσχετισμένες αυτό σημαίνει μόνο ότι οι μεταβλητές αυτές συνδέονται με κάποια σχέση. Δε συνεπάγεται, κατ' ανάγκη, **αιτιότητα**. Οι δύο μεταβλητές μπορεί βεβαία να συνδέονται με σχέση αιτιότητας, μπορεί όμως, όχι. Για παράδειγμα, μπορεί και οι δύο να επηρεάζονται από μια τρίτη μεταβλητή. Ας δούμε δύο παραδείγματα:

  1) Παρατηρήθηκε ότι το *ύψος των μαθητών* ενός σχολείου, ηλικίας 6 έως 13 ετών, έχει ισχυρή θετική γραμμική συσχέτιση με την *αντιληπτική ικανότητα των μαθητών*. Προφανώς η αντιληπτική ικανότητα των μαθητών δεν επηρεάζεται από το ύψος τους. Απλώς τόσο η πνευματική όσο και η φυσική ανάπτυξη των μικρών μαθητών επηρεάζονται παράλληλα από άλλους παράγοντες.
  2) Παρατηρήθηκε ότι οι πωλήσεις ταχύπλοων στο Sidney είχαν, για μια μακρά περίοδο, ισχυρή θετική συσχέτιση με τις πωλήσεις έγχρωμων τηλεοράσεων στη Melbourne. Προφανώς, τόσο οι πωλήσεις ταχύπλοων όσο και οι πωλήσεις έγχρωμων τηλεοράσεων ήταν συνάρτηση γενικότερων ευνοϊκών οικονομικών παραγόντων.

  Είναι, κατά συνέπεια, φανερό ότι η πρόχειρη ή επιπόλαιη ερμηνεία και χρήση του *r* οδηγεί πολλές φορές σε παρερμηνείες ή και σε λανθασμένα συμπεράσματα. Για αιτιολογικά συμπεράσματα, σχέδον πάντοτε, απαιτείται πειραματισμός. Σε κάθε περίπτωση, αιτιώδη σχέση (αλληλεξάρτηση) μεταξύ δύο

μεταβλητών δεχόμαστε μόνον όταν υπάρχει επιστημονική ή λογική βάση που την υπαγορεύει.

- Με s_{xy} συμβολίζουμε τη ***δειγματική συνδιασπορά*** των μεταβλητών *X* και *Y*. Η ***πληθυσμιακή συνδιασπορά*** ορίζεται ανάλογα και συμβολίζεται με σ_{xy}. Εκφράζει τη *συμμεταβολή-συσχέτιση* δύο μεταβλητών μέσω του αθροίσματος των γινομένων των αποκλίσεων των τιμών τους από τους αντίστοιχους μέσους. Μεγάλες τιμές της υποδηλώνουν ότι υπάρχει *συμμεταβολή-συσχέτιση* ενώ μικρές τιμές της υποδηλώνουν ότι δεν υπάρχει *συμμεταβολή-συσχέτιση*. Όμως, δε χρησιμοποιείται ως μέτρο συσχέτισης δύο μεταβλητών διότι επηρεάζεται από τις μονάδες στις οποίες εκφράζονται οι μεταβλητές.

2. Συντελεστής συσχέτισης του Spearman (*rho)*

Δίνει το μέγεθος της γραμμικής συσχέτισης ***ποιοτικών μεταβλητών διάταξης***.

$$rho = 1 - \frac{6 \cdot \sum_{i=1}^{\nu} \delta_i^2}{\nu \cdot (\nu^2 - 1)}$$

όπου,

ν το μέγεθος του δείγματος και $\delta_i = x_i - y_i$, $i = 1, 2, ..., \nu$.

Παράδειγμα:

Φοιτητές *i*	*Σειρά κατάταξης στη Στατιστική (X)*	*Σειρά κατάταξης στα Μαθηματικά (Y)*	δ_i	δ_i^2
Α	1ος	4ος	-3	9
Β	2ος	2ος	0	0
Γ	3ος	3ος	0	0
Δ	4ος	5ος	-1	1
Ε	5ος	1ος	4	16
ΣΤ	6ος	6ος	0	0
Ζ	7ος	8ος	-1	1
Η	8ος	7ος	1	1
				$\sum \delta_i^2 = 28$

$$rho = 1 - \frac{6 \cdot 28}{8 \cdot (8^2 - 1)} = 0{,}667$$

Ιδιότητες-χρήσεις του Συντελεστή Γραμμικής Συσχέτισης *rho*

- Παίρνει τιμές στο κλειστό διάστημα [-1 , 1]. Αν συμφωνούν πλήρως οι δύο κατατάξεις είναι $rho = 1$, ενώ όταν η μια διάταξη είναι ριζικά διαφορετική από την άλλη (για παράδειγμα, αν το μέγεθος δείγματος είναι 8 τότε το *X είναι 1 όταν το Y είναι 8,* το *X είναι 2 όταν το Y είναι 7, κ.ο.κ)* είναι $rho = -1$. Η τιμή 0 δείχνει το μικρότερο βαθμό συσχέτισης.
- Αν στην κατάταξη έχουμε ισοβαθμίες, δίνουμε, ως θέση, σε όλες τις θέσεις που ισοβαθμούν, τη μέση τιμή τους. Για παράδειγμα, αν η βαθμολογία οκτώ φοιτητών στα Μαθηματικά είναι: 10, 9, 9, 8, 7, 6, 6, 6 τότε η κατάταξη γίνεται ως εξής:

Φοιτητές *i*	*Βαθμός στα Μαθηματικά*	Κατάταξη
Α	10	1
Β	9	2,5
Γ	9	2,5
Δ	8	4
Ε	7	5
ΣΤ	6	7
Ζ	6	7
Η	6	7

- Όταν υπάρχουν πολλές ισοβαθμίες ο συντελεστής *rho* δεν είναι αξιόπιστος. Σε αυτή την περίπτωση ενδείκνυται ο δείκτης *Kendall W*.

3. Ο δείκτης Kendall W

Χρησιμοποιείται για να καθορισθεί ο βαθμός συμφωνίας μεταξύ δύο ή περισσοτέρων κριτών στην κατάταξη (τακτική σειρά) που δίνουν σε δύο ή περισσότερα πρόσωπα ή αντικείμενα.

$$W = \frac{12 \cdot s^2}{k^2 \cdot \left(\nu^2 - 1\right)}$$

όπου, ν το μέγεθος του δείγματος (το πλήθος των αξιολογούμενων), k ο αριθμός των κριτών και s η τυπική απόκλιση των αθροισμάτων των τακτικών τιμών που αντιστοιχούν σε κάθε αξιολογούμενο.

Παράδειγμα:

Προϊόντα	1ος Κριτής	2ος Κριτής	3ος Κριτής	4ος Κριτής	5ος Κριτής	Άθροισμα (x_i)	x_i^2
Α	5	6	4	5	4	24	576
Β	3	5	1	3	3	15	225
Γ	1	1	2	1	2	7	49
Δ	2	2	3	4	5	16	256
Ε	4	3	5	2	1	15	225
ΣΤ	7	4	6	6	7	30	900
Ζ	6	7	7	8	8	36	1296
Η	8	8	8	7	6	37	1369
						180	**4896**

$$\bar{x} = \frac{180}{8} = 22{,}5$$

$$s^2 = \frac{1}{7} \cdot \left(4.896 - 8 \cdot 22{,}5^2\right) \approx 121$$

$$W = \frac{12 \cdot 121}{5^2 \cdot (8^2 - 1)} = 0{,}92$$

Ιδιότητες-χρήσεις του Δείκτη Kendall W

- Παίρνει τιμές στο κλειστό διάστημα [0 , 1]. Αν συμφωνούν πλήρως οι κριτές είναι $W = 1$, ενώ όταν υπάρχει ριζική διαφωνία, είναι $W = 0$.
- Η περίπτωση ισοβαθμιών αντιμετωπίζεται όπως και στον υπολογισμό του συντελεστή *rho*.
- Μπορούμε να εργασθούμε και με το συντελεστή *rho* υπολογίζοντάς τον μεταξύ όλων των δυνατών ζευγών κατατάξεων (1ος κριτής, 2ος κριτής), (1ος κριτής, 3ος κριτής), … (1ος κριτής, kος κριτής), (2ος κριτής, 3ος κριτής), κ.ο.κ. και παίρνοντας τη μέση τιμή $\overline{rho}$. Ισχύει $\overline{rho} = W$.