

Πιθανότητες & Στατιστική

Συσχέτιση

Δημήτρης Μητσούδης

Ας θεωρήσουμε δύο τ.μ. X, Y και n ζεύγη παρατηρήσεων

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

από τ.δ. μεγέθους n .

Αναζητούμε μέτρα τα οποία μπορούν να εκφράσουν και να ποσοτικοποιήσουν την πιθανή συσχέτιση των δύο μεταβλητών.

Ο πιο απλός τρόπος για να πάρουμε μια ιδέα για το αν και πώς δυο μεταβλητές συσχετίζονται, είναι να κατασκευάσουμε το λεγόμενο **διάγραμμα διασποράς** (scatter plot).

Δηλαδή, να αναπαραστήσουμε τα ζεύγη των παρατηρήσεων (x_i, y_i) ως σημεία σε ένα καρτεσιανό σύστημα αξόνων όπου ο οριζόντιος άξονας είναι ο άξονας των x και ο κατακόρυφος των y .

Εισαγωγή

Συχνά ενδιαφερόμαστε να μελετήσουμε τη σχέση μεταξύ δύο ποσοτικών μεταβλητών ή πώς η μια επηρεάζει την άλλη. Π.χ.:

- Ποια είναι η σχέση μεταξύ του ύψους των φοιτητών ενός τμήματος με το βάρος τους;
- Πώς σχετίζεται η επίδοση των μαθητών στην Γ' λυκείου με την επίδοση στις εισαγωγικές εξετάσεις του πανεπιστημίου;
- Πώς σχετίζεται ο μισθός με τα έτη σπουδών των εργαζομένων;

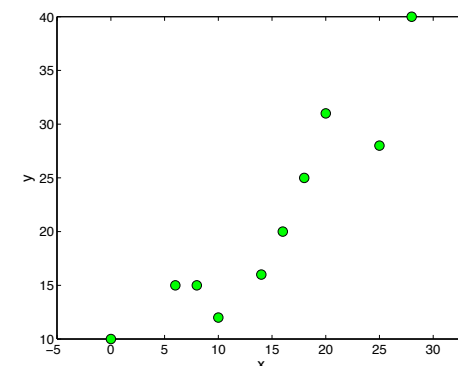
Παρατηρήστε ότι στα παραπάνω παραδείγματα ο ερευνητής δεν προκαθορίζει/ελέγχει τις τιμές καμιάς από τις δύο μεταβλητές.

Προσοχή: Δεν ασχολούμαστε με περιπτώσεις όπως, π.χ., πώς σχετίζεται ο αριθμός των ανοιχτών ταμείων ενός σούπερ μάρκετ (τον οποίο καθορίζει ο διευθυντής του καταστήματος) με τον χρόνο αναμονής των πελατών στα ταμεία;

Παράδειγμα

Ας θεωρήσουμε τα δεδομένα:

x	6	20	0	14	25	16	28	18	10	8
y	15	31	10	16	28	20	40	25	12	15



Από το διάγραμμα διασποράς παίρνουμε μια πρώτη εικόνα για το πώς σχετίζονται αυτές οι μεταβλητές μεταξύ τους: Μικρές τιμές της μεταβλητής X σχετίζονται με μικρές τιμές της Y , ενώ μεγάλες τιμές της X σχετίζονται με μεγάλες τιμές της Y .

Φυσικά, μπορούμε εύκολα να υπολογίσουμε τις ποσότητες:

$$\bar{X} = 14.5, \quad \bar{Y} = 21.2, \quad S_X = 8.7082, \quad S_Y = 9.5778,$$

οι οποίες όμως μας δίνουν πληροφορίες για την κάθε μεταβλητή ξεχωριστά και όχι για τη μεταξύ τους σχέση.

Στη συνέχεια, θα δούμε ένα κλασσικό μέτρο συσχέτισης.

Επομένως

(Συντελεστής γραμμικής συσχέτισης του Pearson για το δείγμα)

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{X} \cdot \bar{Y}}{\sqrt{\sum_{i=1}^n x_i^2 - n \cdot \bar{X}^2} \sqrt{\sum_{i=1}^n y_i^2 - n \cdot \bar{Y}^2}}.$$

Εντελώς ανάλογα ορίζεται ο **πληθυσμιακός** συντελεστής γραμμικής συσχέτισης του Pearson, τον οποίο συμβολίζουμε με ρ .

Συντελεστής γραμμικής συσχέτισης του Pearson

Ο **δειγματικός συντελεστής γραμμικής συσχέτισης του Pearson** συμβολίζεται με r και ορίζεται ως:

$$r = \frac{S_{XY}}{S_X \cdot S_Y},$$

όπου

$$S_{XY} = \text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}).$$

S_X, S_Y είναι οι δειγματικές τυπικές αποκλίσεις των X, Y , αντίστοιχα.

Παράδειγμα

Ας υπολογίσουμε τον συντελεστή γραμμικής συσχέτισης για τα δεδομένα:

x	6	20	0	14	25	16	28	18	10	8
y	15	31	10	16	28	20	40	25	12	15

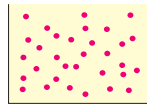
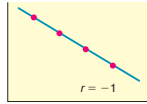
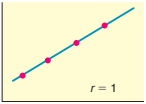
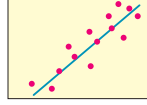
Εύκολα βρίσκουμε ότι:

$$\bar{X} = 14.5, \quad \bar{Y} = 21.2$$

Κατασκευάζουμε τον πίνακα:

x_i	y_i	$x_i - \bar{X}$	$y_i - \bar{Y}$	$(x_i - \bar{X})^2$	$(y_i - \bar{Y})^2$	$(x_i - \bar{X})(y_i - \bar{Y})$
6	15	-8.5	-6.2	72.25	38.44	52.7
20	31	5.5	9.8	30.25	96.04	53.9
0	10	-14.5	-11.2	210.25	125.44	162.4
14	16	-0.5	-5.2	0.25	27.04	2.6
25	28	10.5	6.8	110.25	46.24	71.4
16	20	1.5	-1.2	2.25	1.44	-1.8
28	40	13.5	18.8	182.25	353.44	253.8
18	25	3.5	3.8	12.25	14.44	13.3
10	12	-4.5	-9.2	20.25	84.64	41.4
8	15	-6.5	-6.2	42.25	38.44	40.3
				682.50	825.60	690.00

Συντελεστής γραμμικής συσχέτισης

If r is	Then	The Scatter Diagram Might Look Something Like	
0	There is no linear relation among the points of the scatter diagram.		
1 or -1	There is a perfect linear relation between x and y values; all points lie on the least-squares line.		
Between 0 and 1 ($0 < r < 1$)	The x and y values have a <i>positive correlation</i> . By this, we mean that <i>large</i> x values are associated with <i>large</i> y values, and <i>small</i> x values are associated with <i>small</i> y values.		As we go from left to right, the least-squares line goes <i>up</i> .
Between -1 and 0 ($-1 < r < 0$)	The x and y values have a <i>negative correlation</i> . By this, we mean that <i>large</i> x values are associated with <i>small</i> y values, and <i>small</i> x values are associated with <i>large</i> y values.		As we go from left to right, the least-squares line goes <i>down</i> .

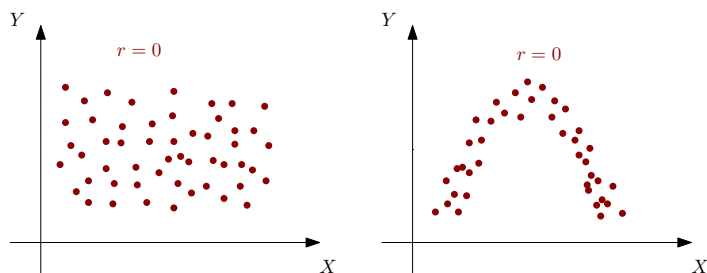
(reproduced from Brase and Brase, Understandable statistics: Concepts and methods)

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{Y})^2}} = \frac{690}{\sqrt{682.5 \cdot 825.6}} = 0.9192.$$

Ιδιότητες του συντελεστή γραμμικής συσχέτισης

- ο συντελεστής γραμμικής συσχέτισης r είναι ένα μέτρο της **γραμμικής** σχέσης μεταξύ δύο μεταβλητών.
- Ο συντελεστής συσχέτισης είναι καθαρός αριθμός (δεν έχει μονάδες μέτρησης).
- $-1 \leq r \leq 1$
- r κοντά στο $+1$ υποδεικνύει **ισχυρή γραμμική θετική συσχέτιση** μεταξύ των μεταβλητών.
- r κοντά στο -1 υποδεικνύει **ισχυρή γραμμική αρνητική συσχέτιση** μεταξύ των μεταβλητών.
- r κοντά στο 0 υποδεικνύει ότι δεν υπάρχει **γραμμική συσχέτιση** μεταξύ των μεταβλητών.

- Αν $r = 0$, τότε **δεν συνεπάγεται ότι δεν υπάρχει συσχέτιση μεταξύ των μεταβλητών**, αλλά ότι δεν υπάρχει **γραμμική συσχέτιση** μεταξύ των μεταβλητών.



- Ο συντελεστής γραμμικής συσχέτισης r χρησιμοποιείται ως μια εκτιμήτρια του πληθυσμιακού συντελεστή γραμμικής συσχέτισης ρ , **μόνον** όταν τα ζεύγη $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ προέρχονται από **τυχαία δειγματοληψία**. Εν γένει, δεν χρησιμοποιείται σε περιπτώσεις πειραματικών δεδομένων, όπου οι τιμές της μιας μεταβλητής ελέγχονται/καθορίζονται από τον ερευνητή.

- **Συσχέτιση $\nRightarrow$ αιτιότητα**

Όταν σε μια δειγματοληψία δύο μεταβλητές X και Y βρίσκονται συσχετισμένες αυτό σημαίνει απλά και μόνον ότι οι μεταβλητές αυτές συνδέονται με κάποια σχέση.

Δεν συνεπάγεται όμως, κατ' ανάγκη, **αιτιότητα**, δηλαδή ότι υπάρχει μεταξύ τους κάποια **σχέση αιτίου-αποτελέσματος**.

Π.χ., αν υπολογίσουμε το συντελεστή συσχέτισης μεταξύ του **μήκους του ποδιού** και της **ικανότητας ανάγνωσης** για τους μαθητές του δημοτικού σίγουρα θα προκύψει μια μεγάλη τιμή.

Φυσικά, αυτό δεν σημαίνει ότι μεγάλο πόδι $\Rightarrow$ μεγαλύτερη ικανότητα ανάγνωσης!

Εδώ υπάρχει ένας τρίτος παράγοντας, που **προφανώς είναι η ηλικία**, με τον οποίο συνδέονται και τα δύο χαρακτηριστικά.

Έλεγχος συσχέτισης

Ο συντελεστής συσχέτισης (r) που είδαμε έχει υπολογιστεί βάσει των παρατηρήσεων του δείγματος. Ο συντελεστής συσχέτισης για τον πληθυσμό (ρ) είναι μια άγνωστη παράμετρος την οποία μπορούμε να εκτιμήσουμε με τη βοήθεια του r .

Συχνά μας ενδιαφέρει να ελέγξουμε σε ε.σ. α κατά πόσο οι δύο τ.μ. X και Y είναι ασυσχέτιστες ή όχι, δηλαδή πραγματοποιούμε τον έλεγχο:

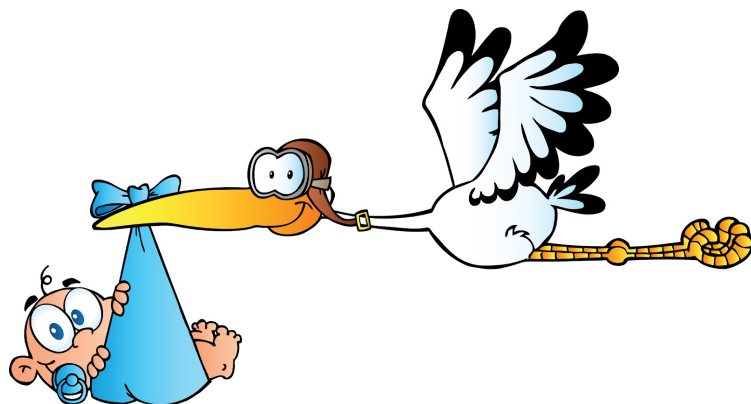
$$H_0: \rho = 0 \text{ έναντι της } H_1: \rho \neq 0$$

Αποδεικνύεται ότι κάτω από την μηδενική υπόθεση η τ.μ.

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim \text{κατανομή } t \text{ (Student) με } n-2 \text{ βαθμούς ελευθερίας.}$$

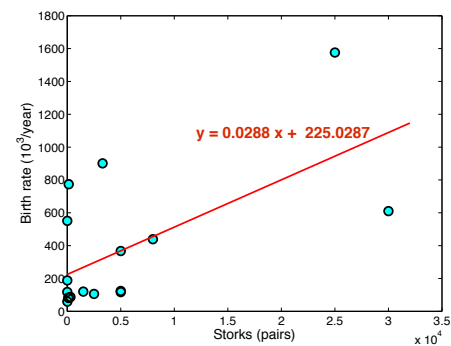
Υπολογίζουμε λοιπόν την τιμή του στατιστικού ελέγχου T και τότε η p -value του αμφίπλευρου ελέγχου είναι δύο φορές η πιθανότητα δεξιά του T με βάση την t_{n-2} . Τέλος, **απορρίπτουμε την H_0 αν $p\text{-value} < \alpha$** .

Φέρνουν τα μωρά οι πελαργοί;



R. Matthews, Storks deliver babies ($p = 0.008$), *Teaching Statistics*, **22** (2000), 36–38.

Σχέση μεταξύ πληθυσμού πελαργών και ρυθμού γεννήσεων



Προκύπτει ότι ο συντελεστής γραμμικής συσχέτισης $r = 0.62$

Πληθυσμιακά στοιχεία ανθρώπων και πελαργών για 17 χώρες

Country	Storks (pairs)	Humans (millions)	Birth rate (thousands/year)
Albania	100	3.2	83
Austria	300	7.6	87
Belgium	1	9.9	118
Bulgaria	5000	9.0	117
Denmark	9	5.1	59
France	140	56.0	774
Germany	3300	78.0	901
Greece	2500	10.0	106
Holland	4	15.0	188
Hungary	5000	11.0	124
Italy	5	57.0	551
Poland	30000	38.0	610
Portugal	1500	10.0	120
Romania	5000	23.0	367
Spain	8000	39.0	439
Switzerland	150	6.7	82
Turkey	25000	56.0	1576

Adapted from Matthews, "Storks deliver babies ($p=0.008$)". Data from 1990.

Έλεγχος συσχέτισης

$H_0 : \rho = 0$ (δεν υπάρχει γραμμική συσχέτιση)

$H_1 : \rho \neq 0$

Ε.σ. $\alpha = 1\%$

Στατιστική συνάρτηση ελέγχου:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.62\sqrt{17-2}}{\sqrt{1-0.62^2}} \approx 3.06$$

Με $(n-2) = 15$ β.ε. βρίσκουμε $p\text{-value} = 0.008$

∴ Απορρίπτουμε την H_0 και συμπεραίνουμε «ισχυρή στατιστικά σημαντική σχέση μεταξύ του αριθμού των πελαργών και του αριθμού γεννήσεων»!!!

Προσοχή! Μη συγχέετε συσχέτιση με αιτιότητα

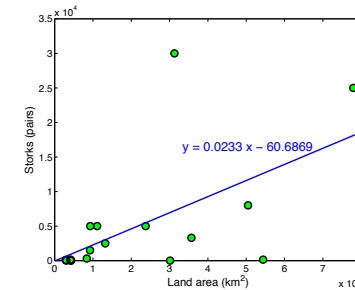


Παρόλα αυτά, ... «όπως και με κάθε άλλη p -value, αυτό δεν σημαίνει ότι η πιθανότητα αυτό να συμβαίνει στην τύχη είναι μόλις $1/125 = 0.008$, και σε καμία περίπτωση **δεν σημαίνει ότι με ποσοστό $124/125 = 99.2\%$ πράγματι οι πελαργοί φέρνουν τα μωρά.**»

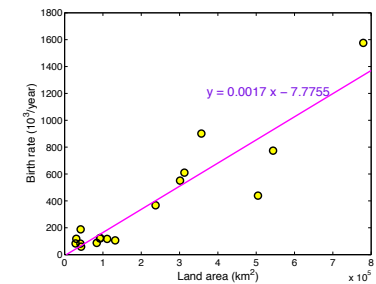
Μια πιθανή εξήγηση: Η ύπαρξη ενός τρίτου παράγοντα με τον οποίο συνδέονται και οι δύο μεταβλητές.

Πιθανός «πιστευτός» υποψήφιος: Η έκταση της κάθε χώρας ...

Σχέση πληθυσμού πελαργών και έκτασης, και αριθμού γεννήσεων και έκτασης



Πληθυσμός πελαργών και έκταση



Αριθμός γεννήσεων και έκταση

Έκταση & πληθυσμιακά στοιχεία ανθρώπων και πελαργών για 17 χώρες

Country	Area (sq. km)	Storks (pairs)	Humans (millions)	Birth rate (thousands/year)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56.0	774
Germany	357,000	3300	78.0	901
Greece	132,000	2500	10.0	106
Holland	41,900	4	15.0	188
Hungary	93,000	5000	11.0	124
Italy	301,280	5	57.0	551
Poland	312,680	30000	38.0	610
Portugal	92,390	1500	10.0	120
Romania	237,500	5000	23.0	367
Spain	504,750	8000	39.0	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25000	56.0	1576