

Συσχέτιση (Correlation)

Ο όρος συσχέτιση (correlation) αναφέρεται στον βαθμό (ένταση) με τον οποίον συσχετίζονται (συμπεταβάζονται) δύο ή περισσότερες μεταβλητές (απλή ή πολλαπλή συσχέτιση).

Η συσχέτιση μεταξύ δύο μεταβλητών (X, Y) χαρακτηρίζεται γραμμική (linear) όταν στο διάγραμμα διασποράς τα σημεία (x_i, y_i) όλων των διαθέσιμων παρατηρήσεων τείνουν να συγκεντρώνονται γύρω από μία ευθεία γραμμή (Σχήμα 3.2) και μη-γραμμική (non-linear) όταν τα σημεία (x_i, y_i) τείνουν να συγκεντρώνονται γύρω από μία καμπύλη (Σχήμα 3.3). Θα ασχοληθούμε μόνο με την απλή γραμμική συσχέτιση. Η απλή συσχέτιση ασχολείται με τον βαθμό με τον οποίο τα σημεία (x_i, y_i) συγκεντρώνονται γύρω από μία ευθεία γραμμή χωρίς όμως να προσδιορίζει ποια είναι ακριβώς αυτή η ευθεία γραμμή, δηλαδή η εξίσωση της ευθείας που διέρχεται μέσα από το νέφος των σημείων. Με το θέμα αυτό ασχολείται η παλινδρόμηση (regression) με την οποία θα ασχοληθούμε παρακάτω. Στη πράξη οι δύο τεχνικές εφαρμόζονται συνήθως σε συνδυασμό.

Η γραμμική συσχέτιση μεταξύ δύο μεταβλητών, μπορεί να είναι θετική, αρνητική ή οι μεταβλητές να μην συσχετίζονται καθόλου.

Θετική συσχέτιση (positive correlation): Οι δύο μεταβλητές τείνουν να μεταβάλλονται προς την ίδια κατεύθυνση. Τείνουν να αυξάνονται ή να μειώνονται (Σχήμα 3.4α) και το διάγραμμα διασποράς εμφανίζει τη μορφή του Σχήματος 3.2α.

Αρνητική συσχέτιση (negative correlation): Οι δύο μεταβλητές τείνουν να μεταβάλλονται προς αντίθετη κατεύθυνση. Όταν οι τιμές της μίας μεταβλητής αυξάνονται, οι τιμές της άλλης μειώνονται και αντίστροφα (Σχήμα 3.4β). Το διάγραμμα διασποράς παρουσιάζει τη μορφή του Σχήματος 3.2β.

Μηδενική συσχέτιση (zero correlation): Όταν οι μεταβολές των τιμών της μίας μεταβλητής δεν συνδέονται καθόλου με τις μεταβολές των τιμών της άλλης. Τότε, τα σημεία του νέφους είναι πολύ διασκορπισμένα και καλύπτουν σχεδόν όλη την επιφάνεια του διαγράμματος (Σχήμα 3.5).

Η συνδιακύμανση (Covariance) μεταξύ δύο μεταβλητών X, Y (συμβ. $\text{Cov}(X, Y)$) ορίζεται ως εξής:

$$\text{Cov}(X, Y) := E[(X - \mu_X)(Y - \mu_Y)], \text{ όπου } \mu_X = E(X) \text{ και } \mu_Y = E(Y). \text{ Για τον πληθυσμό } \text{Cov}(X, Y) = \frac{\sum_{\text{μεγέθους } N} (X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

Η συνδιακύμανση, βάσει ενός δείγματος μεγέθους n , υπολογίζεται ως εξής:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}, \quad \bar{X} = \frac{\sum_{i=1}^n X_i}{n},$$

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}, (\text{αρερόληπτη εκτίμηση}).$$

-3-

Η συνδιακύρανση είναι ένα στατιστικό μέτρο που περιγράφει συνοπτικά τη γραμμική σχέση δύο μεταβλητών.

$\text{Cov}(X, Y) > 0$: Θετική σχέση X, Y

$\text{Cov}(X, Y) < 0$: αρνητική σχέση X, Y

$\text{Cov}(X, Y) = 0$: ανυπαρξία γραμμικής σχέσης X, Y .

Όπως, όταν δύο μεταβλητές είναι ασυσχέτιστες γραμμικά δεν σημαίνει ότι είναι ανεξάρτητες.

Αν X, Y ανεξάρτητες, $\text{Cov}(X, Y) = (\mu_X - \mu_X)(\mu_Y - \mu_Y) = 0$

χωρίς να ισχύει το αντίστροφο, δηλαδή αν $\text{Cov}(X, Y) = 0$

δεν σημαίνει απαραίτητα ότι X, Y είναι ανεξάρτητες, με την έννοια ότι μπορεί να συνδέονται με κάποια μη γραμμική σχέση.

π.χ. $X \in \{-3, -2, -1, 0, 1, 2, 3\}$, $Y = X^2$

$$\text{Cov}(X, Y) = \frac{1}{7} \sum_{i=1}^7 (X_i - \mu_X)(Y_i - \mu_Y) = 0$$

Οι X, Y είναι γραμμικά ασυσχέτιστες (όχι όμως και ανεξάρτητες).

Ιδιότητες: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

$$\text{Cov}(X, X) = \text{Var}(X)$$

αν $Z = a + bX$, $W = c + dY$, $a, b, c, d \in \mathbb{R}$

$$\text{Cov}(Z, W) = bd \text{Cov}(X, Y)$$

$$\text{Cov}(Z, X + Y) = \text{Cov}(Z, X) + \text{Cov}(Z, Y).$$

-4-

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{var}(X-Y) = \text{var}(X) + \text{var}(Y) - 2\text{Cov}(X, Y)$$

Συντελεστής Συσχέτισης (Correlation Coefficient)
Εμφράζει τη συνδιακύρυνση δύο μεταβλητών σε μονάδες των αντίστοιχων τυπικών αποκλίσεων. Ο πληθυσμιακός συντελεστής συσχέτισης (population correlation coefficient) μεταξύ των X, Y συμβολίζεται με ρ_{XY} ή απλά ρ και δίνεται ως εξής: $\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$
$$= \frac{E(X - \mu_X)(Y - \mu_Y)}{\sqrt{E(X - \mu_X)^2} \sqrt{E(Y - \mu_Y)^2}}$$
. Αποδεικνύεται ότι:
$$-1 \leq \rho \leq +1.$$

Ο συντελεστής συσχέτισης εμφράζει τον βαθμό γραμμικής συρρεταβολής δύο μεταβλητών και είναι ορόσηρος με τη συνδιακύρυνση τους, αφού σ_X, σ_Y είναι πάντα θετικές ποσότητες.

Ο δειγματικός συντελεστής συσχέτισης (sample correlation coefficient) αποτελεί ευτίμηση του ρ .

$$r = \hat{\rho} = \frac{\text{Cov}(X, Y)}{\sqrt{s_X^2} \sqrt{s_Y^2}} = \frac{\sum_{i=1}^n (X_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{y})^2}}$$

Επίσης, ισχύει ότι: $-1 \leq r \leq +1$.

$$\text{Επειδή } \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}{n}$$

προκύπτει ο πιο εύχρηστος τύπος:

$$\hat{\rho} = r = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{\sqrt{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2} \sqrt{n \sum_{i=1}^n Y_i^2 - \left(\sum_{i=1}^n Y_i \right)^2}}$$

Θετικές τιμές του ρ (ή του r) σημαίνουν ομόρροπη συρρεταβολή δηλαδή θετική συσχέτιση των μεταβλητών. Τιμή $\rho = +1$ (ή $r = +1$) σημαίνει πλήρη θετική γραμμική συσχέτιση. Αρνητικές τιμές του συντελεστή συσχέτισης σημαίνουν αντίρροπη μεταβολή δηλαδή αρνητική συσχέτιση των μεταβλητών. Τιμή $\rho = -1$ (ή $r = -1$) σημαίνει πλήρη αρνητική γραμμική συσχέτιση. Όσο πλησιέστερα προς το μηδέν είναι η τιμή του συντελεστή συσχέτισης τόσο ασθενέστερη (θετική ή αρνητική) γραμμική σχέση των μεταβλητών. Μηδενική τιμή του συντελεστή συσχέτισης ($\rho = 0$ ή $r = 0$) είναι απόρροια μηδενικής συνδιακύβανσης και σημαίνει απουσία γραμμικής συσχέτισης.

Ιδιότητες: $\rho_{XY} = \rho_{YX}, \rho_{XX} = 1$

$$Z = \alpha + \beta X, W = \gamma + \delta Y, \alpha, \beta, \gamma, \delta \in \mathbb{R}$$

$$\rho_{(\alpha + \beta X)(\gamma + \delta Y)} = \rho_{XY}$$

Ο συντελεστής συσχέτισης είναι ανεξάρτητος από τις μονάδες μέτρησης των δεδομένων.

Έλεγχος σημαντικότητας για τον συντελεστή συσχέτισης ρ .

Η κατανομή του δείγματικού συντελεστή r είναι συρρετρική γύρω από το μηδέν με διακύμανση $\text{var}(r) = \frac{1-r^2}{n-2}$, κάτω από την H_0 .

Αν $\rho=0$ η κατανομή του r εξαρτάται μόνο από το n , και η μεταβλητή $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ (αν $\rho=0$) ακολουθεί

στατιστική συνάρτηση εξέχου t -student με $n-2$ β.ε. Άρα, για τον έλεγχο, χρησιμοποιούνται οι πίνακες της t η οποία ασυμπτωτικά ακολουθεί την κανονική κατανομή.

Έλεγχουμε: $H_0: \rho=0$ vs $H_1: \rho \neq 0$

Από τους πίνακες, λαμβάνουμε την πιθανότητα:

$$P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha.$$

Εάν $|t| > t_{\alpha/2, n}$, όπου $n = n-2$ τότε απορρίπτουμε την H_0 και εάν $|t| < t_{\alpha/2, n}$ ο ευτεληθείς συντελεστής δεν είναι στατιστικά σημαντικός.

Αν $\rho \neq 0$. Η κατανομή του r δεν είναι συρρετρική και δεν μπορεί να χρησιμοποιηθεί ο έλεγχος t .

Η συνάρτηση $z_r = \frac{l}{2} \ln \frac{1+r}{1-r}$ ακολουθεί προσεγγιστικά την κανονική κατανομή με μέσο $E(z_r) = z_p = \frac{l}{2} \ln \frac{1+p}{1-p}$ και διακύμανση $\text{var}(z_r) = \sigma_r^2 = \frac{l}{n-3}$. Άρα, η στατιστική σημαντικότητα του r μπορεί να ελεγχθεί μέσω των πινάκων της τυπικής κανονικής κατανομής.

π.χ.1 Αν ο συνεχιστής συσχέτισης μεταξύ του ύψους και του βάρους **10** ατόμων ήταν $r = 0,836$, η στατιστική

$$t = \frac{0,836}{\sqrt{\frac{1-0,836^2}{10-2}}} = 4,309 \quad \text{Για } v = 10-2 = 8$$
$$\alpha/2 = 0,025$$
$$t = 2,306$$

άρα ο πιθανομετρικός συνεχιστής ρ είναι στατιστικά σημαντικός (απορρίπτουμε την H_0).

π.χ.2 Από δείγμα $n=19$ παρατηρήσεων επιβεβαιώθηκε ότι $r=0,762$. Διαφέρει σημαντικά αυτή η τιμή από την υποθετική τιμή του πιθανομετρικού συνεχιστή συσχέτισης $\rho=0,5$. Ελέγχουμε $H_0: \rho=0,5$
 $H_1: \rho \neq 0,5$

$$z_r = \frac{l}{2} \ln \frac{1+r}{1-r} = l$$

$$z_p = \frac{l}{2} \ln \frac{1+p}{1-p} = \frac{l}{2} \ln \frac{1+0,5}{1-0,5} = 0,549$$

$$\sigma_r = \sqrt{\frac{l}{n-3}} = 0,25 \quad \text{και } z = \frac{z_r - z_p}{\sigma_r} =$$
$$= \frac{l - 0,549}{0,25} = 4,804$$

Για $\alpha = 0.05$ για τεστ δύο ουρών, $z_{\alpha=0.05} = 1.96$.

Είναι $z = 1.804 < z_{\alpha} = 1.96$ δεν απορρίπτουμε την H_0 .

ΠΑΡΑΔΕΙΓΜΑΤΑ ΣΤΟ SPSS

- ① Δεδομένα ————— Ετήσιο εισόδημα (income)
Αποταμίευση (savings)
Ηλικία (Age)
Οικογενειακή κατάσταση (Marital)
74 άτομα.
Data-Set-10

Αναμένουμε άτομα με υψηλό εισόδημα τείνουν να αποταμιεύουν περισσότερα χρήματα.

Κάνουμε ένα διάγραμμα διασποράς Income Savings
Y X

Επιλέγουμε: Graphs → Legacy Dialogs

→ Scatter/Dot → Define
Simple Scatter

Y Axis → income

X Axis → Savings

Στο Titles μπορούμε να βάλουμε τίτλους στο γράφημα

Πατάμε OK.

Η σχέση εισοδήματος και αποταμίευσης είναι δυνατή να διαφοροποιείται ανάλογα με την οικογενειακή κατάσταση των ατόμων.

Θέτουμε: Set Markers by → Marital

και επιλέγουμε το Scatter να εμφανίσει και το Marital Status.

Variable View $\rightarrow$ Values (click) βλέπουμε τις επιμέλειες: -9-

για τη μεταβλητή Marital: 1: Single, 3: Widowed
2: Married, 4: divorced.

Μπορούμε να θεωρήσουμε λογικό ότι τα άτομα μεγαλύτερης ηλικίας είναι δυνατόν να διαθέτουν υψηλότερο (πιθανόν) εισόδημα και μεγαλύτερη αποταμίευση.

Επιλέγουμε Graphs $\rightarrow$ Legacy Dialogs

$\rightarrow$ Scatter/Pot

Προβλεπών

$\rightarrow$ Overlay Scatter

4 διαγράμματα

$\rightarrow$ Define

διασποράς

Savings-Income

Y-X

Pair Y X
Savings Income

ξεχωριστά για
κάθε marital
status.

Panel by
Rows Marital status
[marital]

Όπου, Y X
age income } $\Rightarrow$ 4 διαγράμματα
διασποράς
ξεχωριστά για
κάθε marital status.

② Δεδομένα αρτηριακής πίεσης και βασικών αιματο-
λογικών εξετάσεων 24 υπερτασικών ανδρών. Τα στοιχεία
αναφέρονται στη συστολική πίεση (systol), τη διαστολική
πίεση (diastol), το σάκχαρο (gluco), τη χοληστερόλη
(cholest). Επιθυμούμε να διερευνήσουμε αν υπάρχει
συσχέτιση μεταξύ των 4 αυτών μεταβλητών.

Data-Set-11

Analyze $\rightarrow$ Correlate $\rightarrow$ Bivariate

Variables (v Pearson) Two-tailed
v flag significant correlations

- ✓ Means and standard deviations
- ✓ Cross-product deviations and Covariances

Παράε OK.

Ο πίνακας Correlations

εμφανίζει τις συνδιακυρνώσεις
και τους συντελεστές συσχέτισης
για όλες τις μεταβλητές.

Κάνει και τα σχετιικά τεστ σημαντικότητας
για τους συντελεστές συσχέτισης
αν είναι στατιστικά σημαντικοί και
δίνει τα αποτελέσματα τονίζοντας
με * και ** τα στατιστικά σημαντικά
σχετικά αποτελέσματα για $\alpha = 0.05 (*)$
και $\alpha = 0.01 (**)$.

Συντελεστής συσχέτισης του Spearman (Spearman's
rank correlation coefficient) r_s , είναι ο συντελεστής
συσχέτισης του Pearson αλλά υπολογιζόμενος με βάση
τις τάξεις μεγέθους των παρατηρήσεων και όχι τις
παρατηρήσεις:

$$r_s = \frac{\sum_{i=1}^n [R(X_i) - \overline{R(X)}][R(Y_i) - \overline{R(Y)}]}{\left(\sum_{i=1}^n [R(X_i) - \overline{R(X)}]^2 \right)^{1/2} \left(\sum_{i=1}^n [R(Y_i) - \overline{R(Y)}]^2 \right)^{1/2}}$$

$$\text{όπου } \overline{R(X)} = \frac{1}{n} \sum_{i=1}^n R(X_i), \quad \overline{R(Y)} = \frac{1}{n} \sum_{i=1}^n R(Y_i).$$

-11-

Ισχύει επίσης ότι $-1 \leq r_s \leq +1$. Το κριτήριο ελέγχου για την υπόθεση: $H_0: \rho_s = 0$ vs $H_1: \rho_s \neq 0$ είναι της ίδιας μορφής με τον έλεγχο του συντελεστή συσχέτισης του Pearson, όπου ρ_s είναι ο αντίστοιχος πληθυσμιακός συντελεστής συσχέτισης, εφόσον τα ζεύγη των σχετιζόμενων θέσεων (x_{ri}, y_{ri}) , $i=1, \dots, n$ αποτελούν τυχαία δείγματα από τον αντίστοιχο πληθυσμό.

Ισχύει ότι η στατιστική συνάρτηση ελέγχου που χρησιμοποιείται είναι η ίδια όπως στον έλεγχο Pearson, αντί για r χρησιμοποιείται r_s : $t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$, κάτω από την H_0 . Το παραπάνω στατιστικό δείχρατος ακολουθεί την κατανομή t -student με $n-2$ β.ε.

Ο συντελεστής συσχέτισης του Spearman είναι ένας μη-παραμετρικός συντελεστής και είναι πολύ λιγότερο ευαίσθητος στην ύπαρξη ακραίων δειγματικών τιμών. 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, αφού στηρίζεται στις σχετικές θέσεις των τιμών.

Συντελεστής συσχέτισης του Kendall (Kendall's tau) και αυτός βασίζεται στις σχετικές θέσεις των τιμών των μεταβλητών X και Y .

Κατά τον υπολογισμό του οι τιμές των x_i και y_i , $i=1, \dots, n$ των δύο μεταβλητών διατάσσονται και έστω x_{ri}, y_{ri} , $i=1, \dots, n$ οι σχετικές θέσεις τους κατά τη διάταξη. Υπολογίζονται οι διαφορές $y_{ri} - x_{ri}$ για $i=1, \dots, n-1$

-12-

και $j > i$ και αθροίζονται οι περιπτώσεις κατά τις οποίες οι διαφορές αυτές είναι θετικές καθώς και οι περιπτώσεις κατά τις οποίες είναι αρνητικές. Τα δύο αθροίσματα συμβολίζονται με n_c (θετικές διαφορές) και n_d (αρνητικές διαφορές). Ο συντελεστής συσχέτισης του Kendall ορίζεται ως εξής:
$$\tau = \frac{n_c - n_d}{\frac{n(n-1)}{2}}$$

Αν όλες οι διαφορές είναι θετικές $\tau = +1$ ενώ αν όλες οι διαφορές είναι αρνητικές $\tau = -1$.

Η τ χρησιμοποιείται για τον έλεγχο: $H_0: \rho_\tau = 0$ vs $H_1: \rho_\tau \neq 0$. Κάτω από την H_0 , η $Z = \frac{n_c - n_d}{\sqrt{\frac{n(n-1)(2n+5)}{18}}}$

ακολουθεί την τυπική κανονική κατανομή. Είναι ένας μη-παραμετρικός συντελεστής συσχέτισης με τα πλεονεκτήματα και τα μειονεκτήματα που έχει αντίστοιχα και ο συντελεστής Spearman. Δηλαδή, πλεονέκτημα: Υπολογίζεται όχι μόνο για ποσοτικές αλλά και για διατεταγμένες μεταβλητές. Μειονέκτημα: Δεν υπολογίζεται από τις αρχικές τιμές των μεταβλητών αλλά από τις σχετικές θέσεις που έχουν αυτές οι τιμές στην εσωτερική διάταξή τους.

Για το π.χ. 2: Analyze $\rightarrow$ Correlate $\rightarrow$ Bivariate

✓ Kendall's tau

✓ Spearman Two-Tailed

✓ Flag significant correlations

εμφανίζει τα αποτελέσματα των Non-parametric Correlations.

Μερική Συσχέτιση (Partial Correlation): Ευφράζει -13-
τη γραμμική σχέση μεταξύ δύο ποσοτικών μεταβλητών
X και Y όταν από αυτήν αφαιρεθούν οι γραμμικές επι-
δράσεις μίας ή περισσότερων άλλων μεταβλητών.

Ο συντελεστής μερικής συσχέτισης των X και Y απομα-
κρύνοντας τις επιδράσεις της Z δίνεται ως εξής:

$$r_{XY,Z} = \frac{(r_{XY} - r_{XZ}r_{YZ})}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}}, \text{ όπου}$$

r_{XY} : συντελεστής συσχέτισης της X με την Y
και r_{XZ}, r_{YZ} : συντελεστής συσχέτισης της X με την Z
και της Y με την Z, αντίστοιχα.

Στο π.χ. 3 Analyze $\rightarrow$ Correlate $\rightarrow$ Partial

Data-Set-12 $\rightarrow$ Variables Height
Weight

Height-Weight-Age Controlling for age

Pearson CC: 0.972 partial coefficient

Partial CC(age): 0.878 (μικρότερο από 0.972)

(still significant; $\hat{\alpha} = 0.021$).



Αποδεικνύεται ότι:

$$r = \hat{\beta} \frac{s_x}{s_y} \quad (7.17)$$

Έτσι:

- i) Αν $r \approx 0$, τότε $\hat{\beta} \approx 0$ που σημαίνει ότι δεν υπάρχει γραμμική σχέση των x και y , δεν αποκλείεται όμως η ύπαρξη άλλων σχέσεων ή συσχετίσεων.
- ii) Αν $r \approx 1$ τότε $\hat{\beta} > 0$ και υπάρχει θετική γραμμική εξάρτηση των x και y .
- iii) Αν $r \approx -1$ τότε $\hat{\beta} < 0$ και υπάρχει αρνητική γραμμική εξάρτηση των x και y .

Ορισμός 7.7

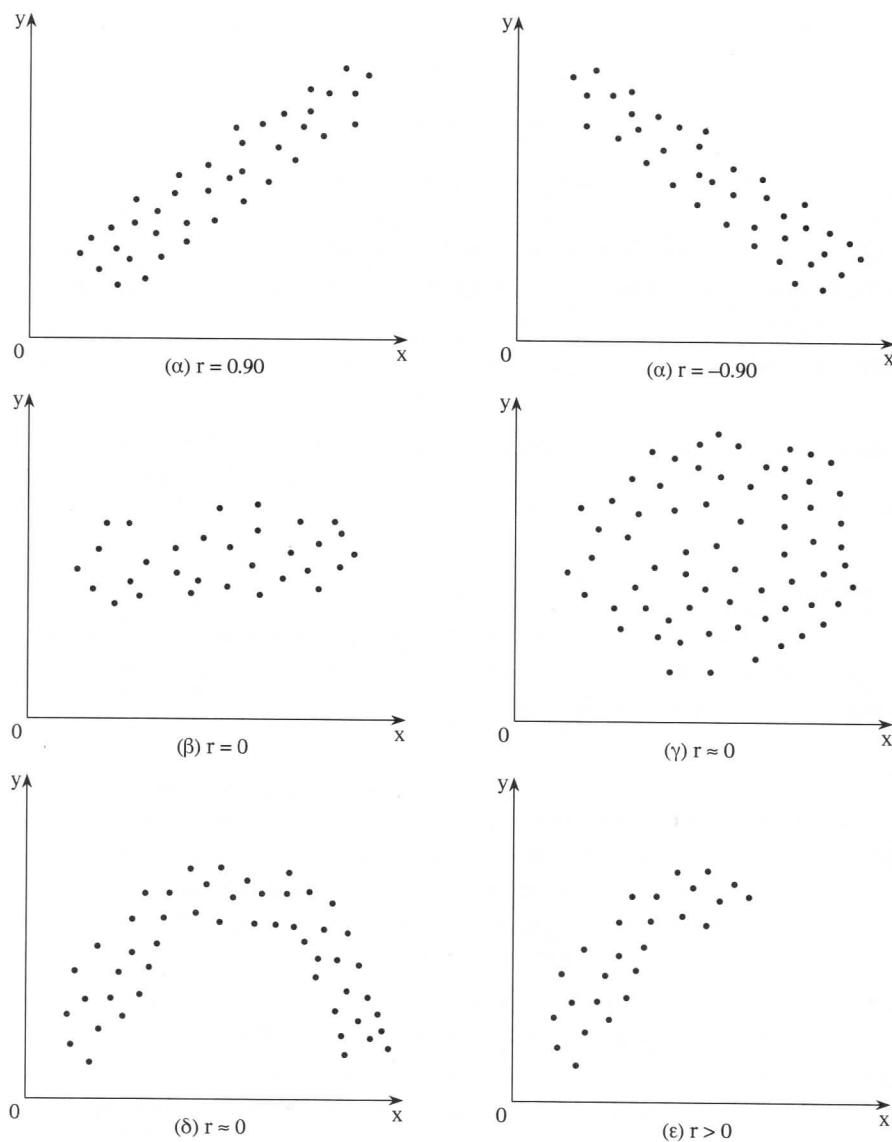
Η ποσότητα r^2 καλείται **συντελεστής προσδιορισμού** και η ποσότητα $1-r^2$ παριστάνει το ποσοστό της συνολικής μεταβλητότητας που οφείλεται στο συνολικό σφάλμα.

Στην πράξη πρώτα κάνουμε τη γραφική παράσταση του σμήνους των σημείων (διάγραμμα διασποράς). Έτσι έχουμε εποπτική εικόνα του προβλήματος. Ο συντελεστής συσχέτισης δεν υπολογίζεται παρά μόνον όταν το σμήνος έχει σχήμα επίμηκης κεκλιμένης έλλειψης ή πλατυσμένου J.

Το σχήμα 7.6 δίνει το διάγραμμα διασποράς σε σχέση με τον συντελεστή συσχέτισης r και διαπιστώνονται τα εξής:

- Στο (α) υπάρχει γραμμική εξάρτηση η οποία είναι μεγαλύτερη, όσο η εκκεντρότητα της έλλειψης πλησιάζει τη μονάδα.
- Στο (β) αν και το σμήνος των σημείων έχει σχήμα έλλειψης, ο κύριος άξονας είναι παράλληλος προς τον άξονα των x οπότε $\beta=0$ δηλ. τα x και y είναι γραμμικά ανεξάρτητα.
- Στο (γ) δεν υπάρχει σχέση που να συνδέει τα x και y .
- Στο (δ) δεν υπάρχει γραμμική σχέση που να συνδέει τα x και y , όμως το σμήνος των σημείων είναι προσανατολισμένο κατά μήκος μιας καμπύλης που μπορεί να είναι πολυώνυμο, ή να είναι της μορφής $y = ce^{ax+\beta}$ ή $y = \log(ax+\beta)$.
- Στο (ε) το σμήνος έχει σχήμα J και μπορούμε να πούμε ότι η εξάρτηση είναι σχεδόν γραμμική.

(7.16)



Σχήμα 7.6

Σημείωση Σ' αυτό το σημείο θα πρέπει να τονισθεί ότι χρειάζεται μεγάλη προσοχή στην επιλογή των τυχαίων μεταβλητών στην εφαρμογή των μεθόδων διότι μπορεί να καταλήξουμε σε λάθος συμπεράσματα. π.χ. Τα τελευταία χρόνια παρατηρείται στην Ευρώπη αύξηση των τροχαίων ατυχημάτων και μείωση του αριθμού των γεννήσεων. Τα δύο αυτά γεγονότα είναι ανεξάρτητα μεταξύ τους αν όμως υπολογισθεί ο συντελεστής συσχέτισής τους βάσει των αριθμητικών δεδομένων αυτός μπορεί να είναι κοντά στο -1 χωρίς αυτό να σημαί-

Παρά

Ο παρ
και τη
τους.
αρχικό

X

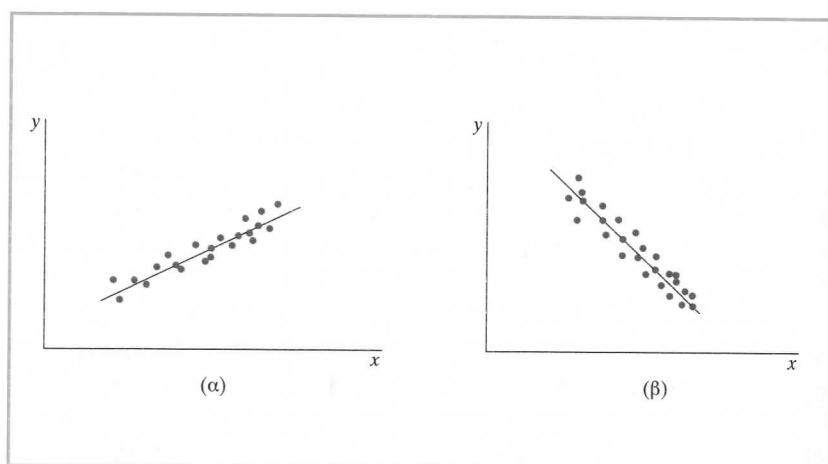
2
3
3
32
35
27
37
24

Να βρεθ

Αύση

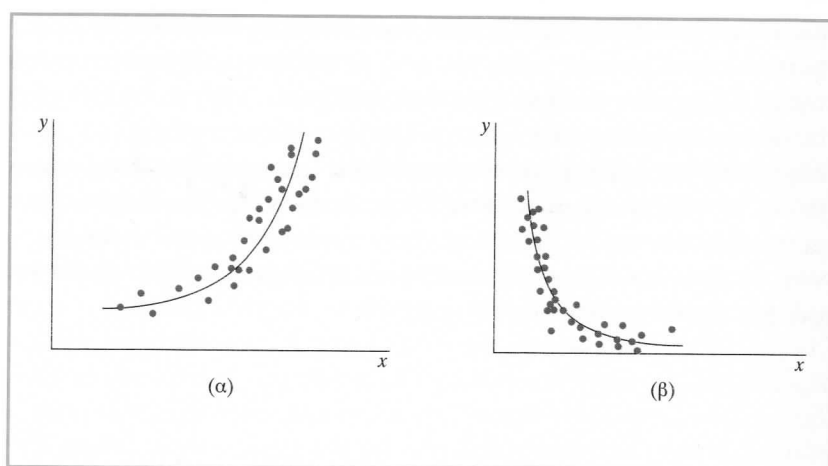
ΣΧΗΜΑ 3.2

Θετική (α) και αρνητική (β) γραμμική συσχέτιση

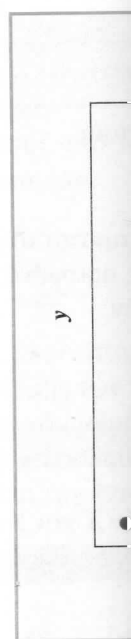
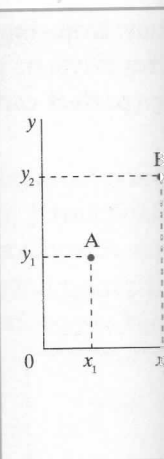


ΣΧΗΜΑ 3.3

Θετική (α) και αρνητική (β) μη γραμμική συσχέτιση

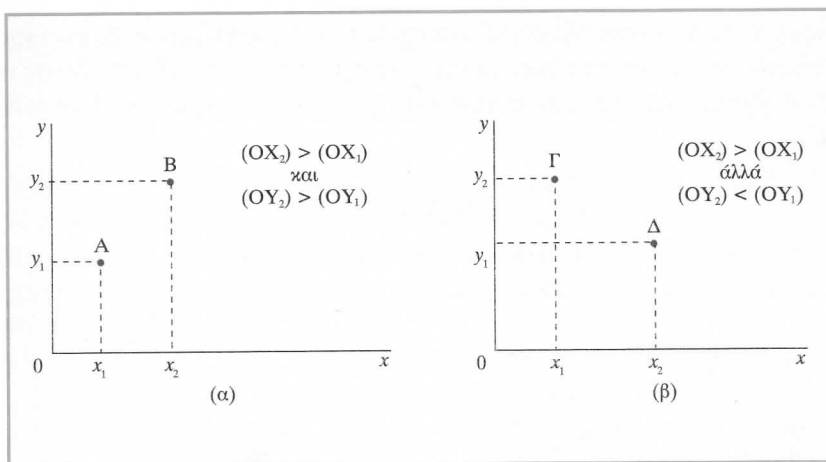


Θε

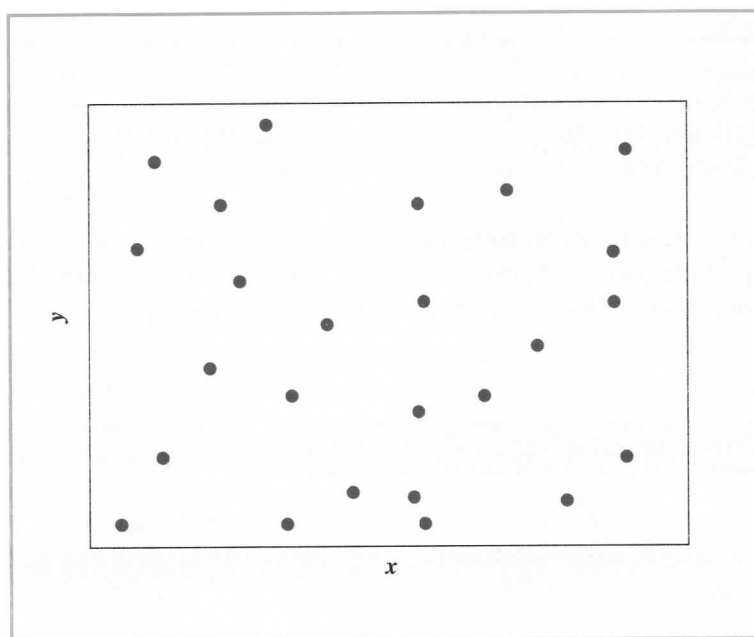


ΣΧΗΜΑ 3.4

Θετική (α) και αρνητική (β) σχέση δεδομένων

**ΣΧΗΜΑ 3.5**

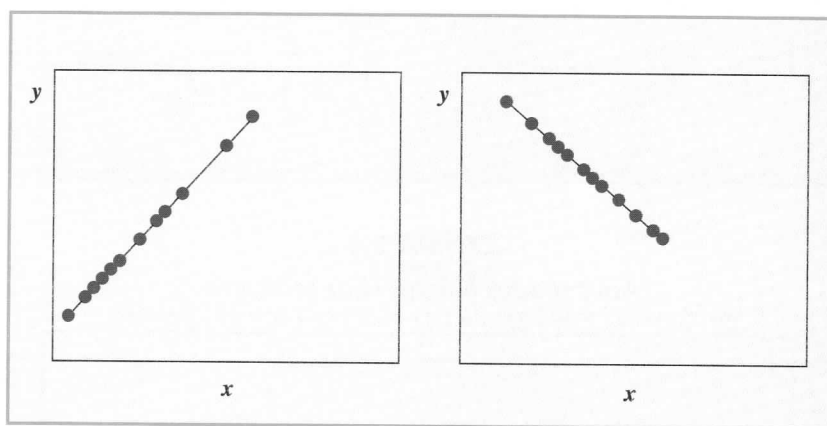
Αυσχέτιστα διμεταβλητά δεδομένα



Γενικά, όσο πιο συγκεντρωμένα είναι τα σημεία γύρω από την ευθεία γραμμή τόσο πιο ισχυρή είναι η γραμμική σχέση που συνδέει τις δύο μεταβλητές. Αντίθετα, όσο περισσότερο διασκορπισμένα εμφανίζονται τα σημεία, τόσο πιο ασθενής είναι η συσχέτιση των μεταβλητών. Στην ακραία περίπτωση που όλα τα σημεία (x_i, y_i) συμβεί να βρίσκονται πάνω σε μία ευθεία γραμμή (Σχήμα 3.6) η **συσχέτιση** καλείται **πλήρης** (**perfect correlation**).

ΣΧΗΜΑ 3.6

Πλήρης Γραμμική Συσχέτιση



(α) Πλήρης Θετική Συσχέτιση

(β) Πλήρης Αρνητική Συσχέτιση

Πριν ασχοληθούμε με τους τρόπους με τους οποίους εκτιμάται το είδος και ο βαθμός γραμμικής συσχέτισης, θα αναφερθούμε σε ορισμένα χρήσιμα στατιστικά μέτρα, τις ροπές διμεταβλητών πληθυσμών.

3.3 Ροπές Διμεταβλητών Πληθυσμών

Έστω $x_1, x_2, \dots, x_N$ και $y_1, y_2, \dots, y_N$ οι τιμές δύο μεταβλητών X και Y σε ένα διμεταβλητό πληθυσμό N και k, l ακέραιοι μη αρνητικοί αριθμοί. Ορίζουμε ως:

- απλή ροπή τάξης k για τη μεταβλητή X και τάξης l για τη μεταβλητή Y

(συμβ
των υ
χα, δη

- κεντρ
τή Y (
ποκλύ
ριθμη
ντίστο

Από τ

$$\nu_{0,0} =$$

$$\nu_{1,0} =$$

$$\nu_{0,1} =$$

ενώ η $\nu_{1,1}$

ταβλητώ

Επίσ

προκύπτ

$$\mu_{0,0} =$$

$$\mu_{2,0} =$$

$$\mu_{0,2} =$$

η δε $\mu_{1,1}$

γκαταν

τω, η συ

μελέτη μ

όμως κα

Παράδειγμα

$X:$

$Y:$