Πανεπιστήμιο Αιγαίου

Χωρική Ανάλυση

Ενότητα 8β: Στοιχεία διμεταβλητής περιγραφικής στατιστικής

Κυριακίδης Φαίδων

Τμήμα Γεωγραφίας

Άδειες Χρήσης

Χρηματοδότηση

Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.

Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.

Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Στοιχεία Διμεταβλητής Περιγραφικής Στατιστικής

Φαίδων Κυριακίδης

Καθηγητής

phkyriakidis@geo.aegean.gr

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
Λόφος Πανεπιστημίου, 81100 Μυτιλήνη

ΤΜΗΜΑ ΓΕΩΓΡΑΦΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΓΕΩΓΡΑΦΙΑ ΚΑΙ ΕΦΑΡΜΟΣΜΕΝΗ ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ

Χωρική Ανάλυση

Εισαγωγικές Σημειώσεις

Πολυμεταβλητά δεδομένα

Συχνά, τα δεδομένα αφορούν μετρήσεις πολλαπλών μεταβλητών σε κάθε σταθμό παρατήρησης ή σε κάθε μονάδα δειγματοληψίας. Τίθεται, επομένως, η ανάγκη της **από κοινού** ανάλυσης μετρήσεων πολλών μεταβλητών, με στόχο τη διερεύνηση σχέσεων, αλληλεπιδράσεων, και γενικότερα την ανάπτυξη μοντέλων για πρόβλεψη

Διμεταβλητά δεδομένα

Μια υποπερίπτωση πολυμεταβλητής ανάλυσης είναι και η από κοινού ανάλυση μετρήσεων δύο μεταβλητών, π.χ., θερμοκρασίας και υψομέτρου. Ένα παράδειγμα διμεταβλητής ανάλυσης είναι και αυτό της χωρικής ανάλυσης ενός διδιάστατου σημειακού προτύπου, όπου οι δύο μεταβλητές είναι οι γεωγραφικές συντεταγμένες. Στα παρακάτω, θα μας απασχολήσει η ανάλυση μετρήσεων στο χώρο των μεταβλητών, κι όχι στο γεωγραφικό χώρο. Τα εργαλεία ανάλυσης, όμως, είναι τα ίδια. . .

Στόχοι του μαθήματος αυτού

Παρουσίαση των εννοιών της συνδιακύμανσης (covariance), του συντελεστή συσχέτισης (correlation coefficient), και της ροπής αδράνειας (moment of inertia) ως μέτρα ποσοτικοποίησης της συνάφειας ή διαφοράς μεταξύ ενός συνόλου **ζευγών** μετρήσεων δύο μεταβλητών με συνεχή κλίμακα μέτρησης

Δεδομένα/Μετρήσεις Δύο Μεταβλητών

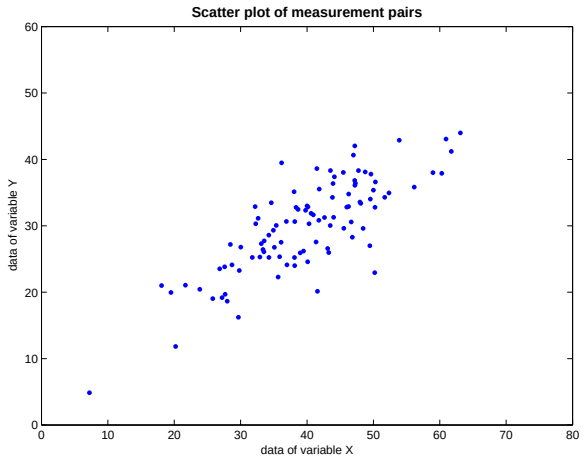
N ζεύγη τιμών $\{(x_i, y_i), i = 1, \ldots, N\}$, όπου κάθε ζεύγος (x_i, y_i) αποτελείται από μια μέτρηση x_i της μεταβλητής X και μια μέτρηση y_i της μεταβλητής Y. Οι N τιμές της μεταβλητής X, και οι N τιμές της μεταβλητής Y μπορούν να αποθηκευτούν σε δύο $(N \times 1)$ διανύσματα: $\mathbf{x} = [x_i, i = 1, \ldots, N]^T$ και $\mathbf{y} = [y_i, i = 1, \ldots, N]^T$, όπου T δηλώνει αναστροφή

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_N \end{bmatrix} \quad \text{και} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix}$$

Διασπορόγραμμα (Scatter Plot) Ζευγών Μετρήσεων

Ορισμός

Η απεικόνιση ζευγών μετρήσεων σε ένα γράφημα, όπου ζεύγος (x_i, y_i) ορίζει ένα σημείο με "συντεταγμένες" x_i και y_i, λέγεται **διασπορόγραμμα** δύο μεταβλητών



Σημείωση

Ο "χώρος" του διασπορογράμματος λέγεται χώρος των μεταβλητών, και οι συντεταγμένες των σημείων που απεικονίζονται στο χώρο αυτό δεν είναι γεωγραφικές, αλλά οι τιμές της κάθε μεταβλητής

Στατιστικά Επιμέρους Μεταβλητών

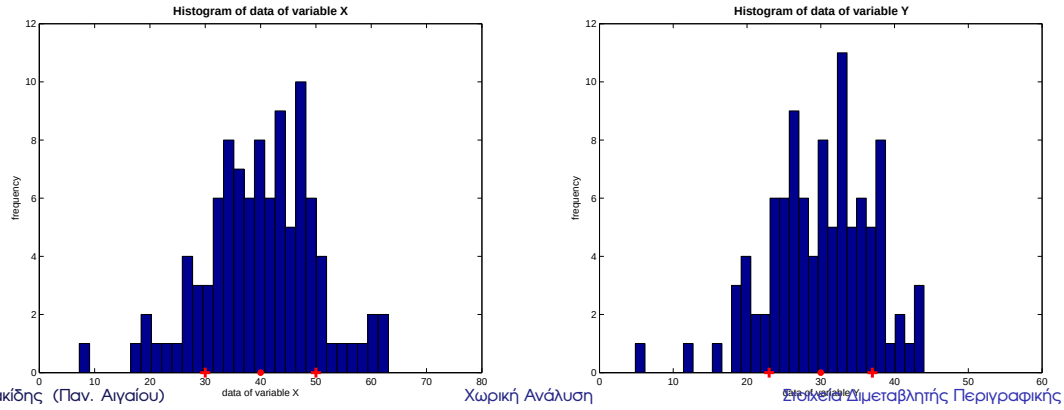
Μέσοι όροι και διακυμάνσεις

Μέσος όρος (mean) m_X και τυπική απόκλιση (standard deviation) s_X των N μετρήσεων της μεταβλητής X, και αντίστοιχα m_Y και s_Y για τις τιμές της μεταβλητής Y

$$m_X = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad s_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[x_i - m_X]^2} \quad \text{και} \quad m_Y = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad s_Y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[y_i - m_Y]^2}$$

Σημείωση: Πολλές φορές χρησιμοποιείται $N-1$, αντί N, στον παρονομαστή του τύπου της τυπικής απόκλισης

Ιστογράμματα συχνότητας

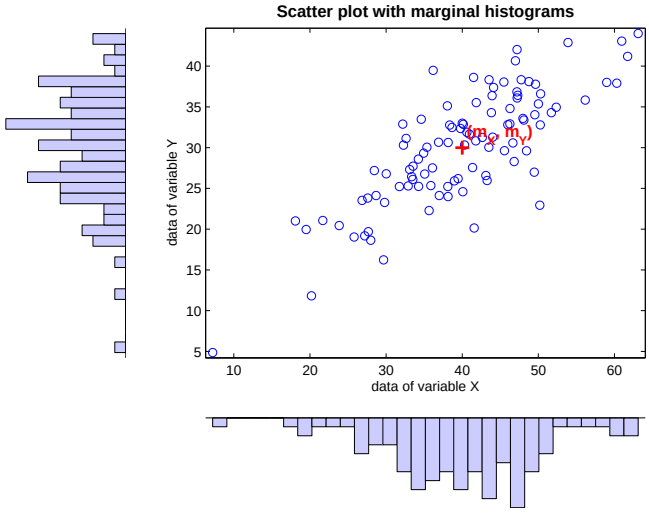


Περιθώρια (Marginal) Ιστογράμματα

Διασπορόγραμμα με περιθώρια ιστογράμματα συχνοτήτων

Τα ιστογράμματα συχνότητας των μετρήσεων x_i και y_i των επιμέρους μεταβλητων X και Y, αποτελούν περιθώρια ιστογράμματα του διασπορογράμματος των N ζευγών μετρήσεων $\{(x_i, y_i), i = 1, \ldots, N\}$. Το κέντρο του διασπορογράμματος είναι ένα σημείο με "συντεταγμένες" (m_X, m_Y)

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix}$$

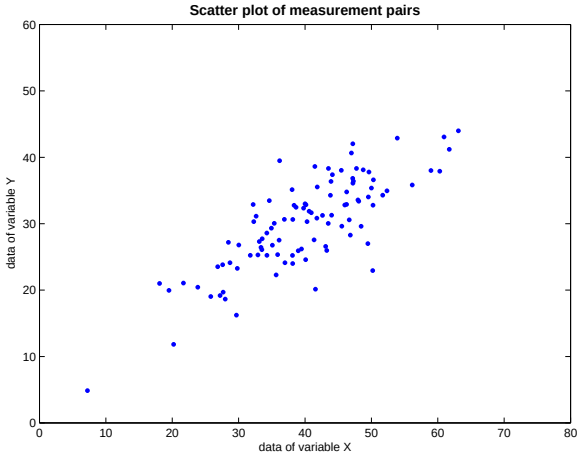


$$[x_1, \cdots, x_i, \cdots, x_N]$$

Προς Ένα Μέτρο Συνάφειας Μεταξύ των Μετρήσεων

Στόχος

Η ποσοτικοποίηση της συμμεταβολής ή συνάφειας μεταξύ των μετρήσεων δύο μεταβλητών, δηλαδή η διερεύνηση του κατά πόσο υψηλές τιμές της μεταβλητής X τείνουν να αντιστοιχούν με υψηλές τιμές της μεταβλητής Y, ή το αντίστροφο



Σημείωση

Ένα μέτρο συνάφειας δε θα πρέπει να επηρεάζεται από τη θέση του διασπορογράμματος, δηλαδή από το "κέντρο" του – συνεπώς χρησιμοποιούνται οι αποκλίσεις των μετρήσεων από τους αντίστοιχους μέσους όρους

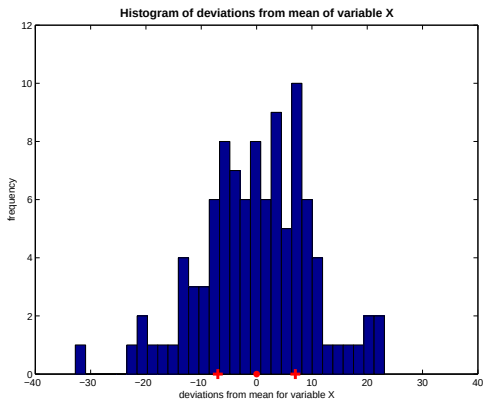
Στατιστικά Επιμέρους Αποκλίσεων

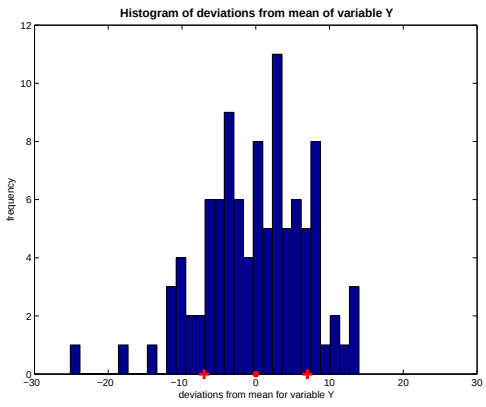
Μέσοι όροι και διακυμάνσεις

Η απόκλιση της μέτρησης x_i από το μέσο όρο m_X της μεταβλητής X ορίζεται ως: $x_i - m_X$. Οι N αποκλίσεις της μεταβλητής X έχουν μέσο όρο $m_{(X-m_X)} = 0$ και τυπική απόκλιση $s_{(X-m_X)} = s_X$, και αντίστοιχα για τις αποκλίσεις της μεταβλητής Y

$$m_{(X-m_X)} = \frac{1}{N}\sum_{i=1}^{N} x_i - m_X = 0, \quad s_{(X-m_X)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(x_i - m_X) - 0]^2} = s_X$$

Ιστογράμματα συχνότητας αποκλίσεων



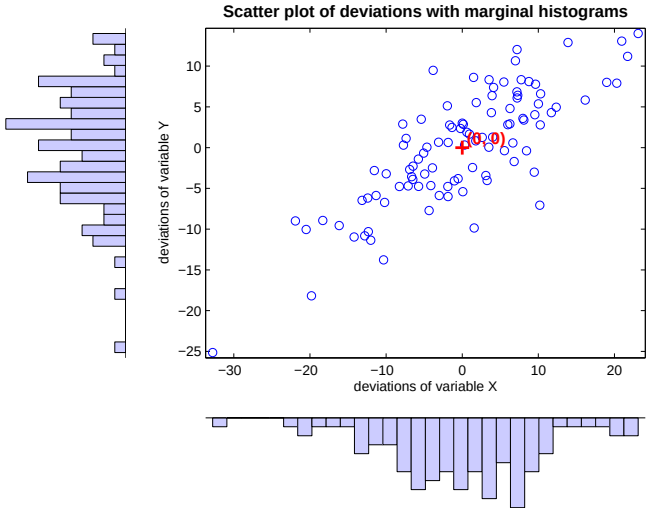


Περιθώρια Ιστογράμματα Αποκλίσεων

Διασπορόγραμμα με περιθώρια ιστογράμματα

Τα ιστογράμματα συχνότητας των αποκλίσεων $x_i - m_X$ και $y_i - m_Y$ των επιμέρους μεταβλητων X και Y, αποτελούν περιθώρια ιστογράμματα του διασπορογράμματος των N ζευγών αποκλίσεων $\{(x_i - m_X, y_i - m_Y), i = 1, \ldots, N\}$. Το κέντρο του διασπορογράμματος είναι ένα σημείο με "συντεταγμένες" $(0, 0)$

$$\begin{bmatrix} y_1 - m_Y \\ \vdots \\ y_i - m_Y \\ \vdots \\ y_N - m_Y \end{bmatrix}$$



$$[x_1 - m_X, \cdots, x_i - m_X, \cdots, x_N - m_X]$$

Συνδιακύμανση

Ορισμός

Συνδιακύμανση s_{XY} δύο μεταβλητών X και Y ονομάζεται η αναμενόμενη τιμή (μέσος όρος) των γινομένων των αποκλίσεων των τιμών της κάθε μεταβλητής από τον αντίστοιχο μέσο όρο

$$s_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - m_X)(y_i - m_Y)$$

Υπολογίζεται και ως $s_{XY} = \frac{1}{N} \sum_{i=1}^{N} x_i y_i - m_X m_Y$, πολλές φορές με $N - 1$ στον παρονομαστή

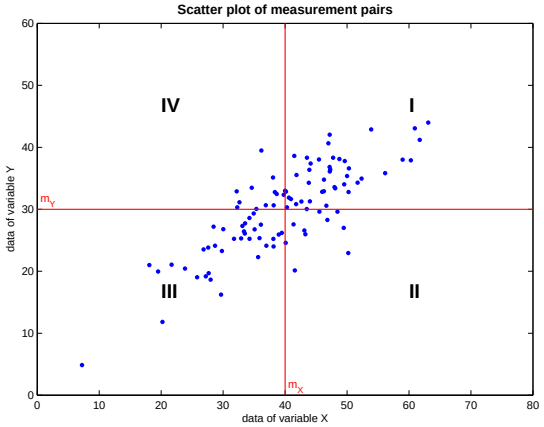
Σημειώσεις

- Η συνδιακύμανση έχει μονάδες μέτρησης αυτές του γινομένου των μονάδων μέτρησης των μεταβλητών. Π.χ., η συνδιακύμανση θερμοκρασίας και υψομέτρου θα μπορούσε να έχει μονάδες (βαθμούς Κελσίου $\times$ μέτρα). Γι' αυτό το λόγο η έννοια της συνδιακύμανσης είναι δυσνόητη, και αντ' αυτής χρησιμοποιείται η έννοια του συντελεστή συσχέτισης που δεν έχει μονάδες μέτρησης
- Η συνδιακύμανση αποτελεί ένα μέτρο συνάφειας μεταξύ N ζευγών μετρήσεων. Η τιμή $s_{XY} = 0$ υποδηλώνει έλλειψη συνάφειας, ενώ $s_{XY} > 0$ και $s_{XY} < 0$ υποδηλώνουν θετική και αρνητική συνάφεια, αντίστοιχα

Συνδιακύμανση (2)

Ερμηνεία

Ο συνδιακύμανση του παρακάτω διασπορογράμματος είναι θετική, γιατί περισσότερα ζεύγη μετρήσεων x_i, y_i εμπίπτουν στους τομείς I και III, απ' ότι στους τομείς II και IV. Στους τομείς I και III, ισχύει $(x_i - m_X) \cdot (y_i - m_Y) > 0$, γιατί οι αντίστοιχες αποκλίσεις $x_i - m_X, y_i - m_Y$ έχουν θετικό ή αρνητικό πρόσημο. Τέλος, οι αποκλίσεις στους τομείς I και III είναι κατ' απόλυτη τιμή μεγαλύτερες από τις αποκλίσεις στους τομείς II και IV

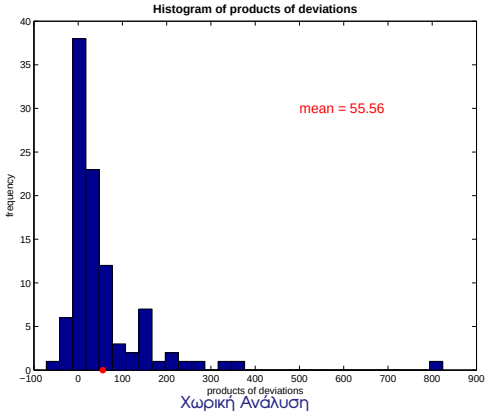


Τομέας I: $x_i > m_X$ και $y_i > m_Y$, συνεπώς $(x_i - m_X) \cdot (y_i - m_Y) > 0$
Τομέας II: $x_i > m_X$ και $y_i < m_Y$, συνεπώς $(x_i - m_X) \cdot (y_i - m_Y) < 0$
Τομέας III: $x_i < m_X$ και $y_i < m_Y$, συνεπώς $(x_i - m_X) \cdot (y_i - m_Y) > 0$
Τομέας IV: $x_i < m_X$ και $y_i > m_Y$, συνεπώς $(x_i - m_X) \cdot (y_i - m_Y) < 0$

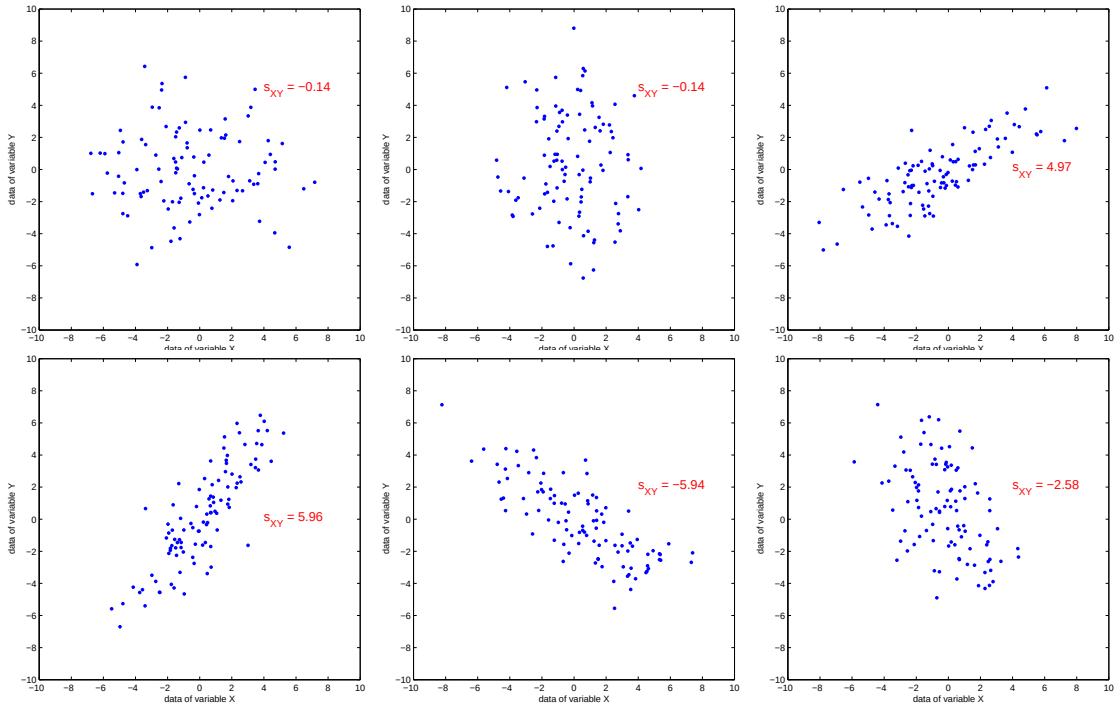
Συνδιακύμανση (3)

Σε κάθε ζεύγος μετρήσεων x_i, y_i αντιστοιχεί ένα ζεύγος αποκλίσεων $(x_i - m_X), (y_i - m_Y)$ και ένα γινόμενο $(x_i - m_X) \cdot (y_i - m_Y)$. Ο μέσος όρος των N γινομένων ονομάζεται συνδιακύμανση των N ζευγών μετρήσεων των δύο μεταβλητών

$$\begin{bmatrix} x_1 - m_X \\ \vdots \\ x_i - m_X \\ \vdots \\ x_N - m_X \end{bmatrix} \odot \begin{bmatrix} y_1 - m_Y \\ \vdots \\ y_i - m_Y \\ \vdots \\ y_N - m_Y \end{bmatrix} = \begin{bmatrix} (x_1 - m_X) \cdot (y_1 - m_Y) \\ \vdots \\ (x_i - m_X) \cdot (y_i - m_Y) \\ \vdots \\ (x_N - m_X) \cdot (y_N - m_Y) \end{bmatrix}$$



Παραδείγματα Διασπορογραμμάτων



Πρόβλημα: Η τιμή της συνδιακύμανσης επηρεάζεται από τις μονάδες μέτρησης των επιμέρους μεταβλητών που συνθέτουν το διασπορόγραμμα. Συνεπώς χρησιμοποιούνται οι τυποποιημένες αποκλίσεις των μετρήσεων από τους αντίστοιχους μέσους όρους

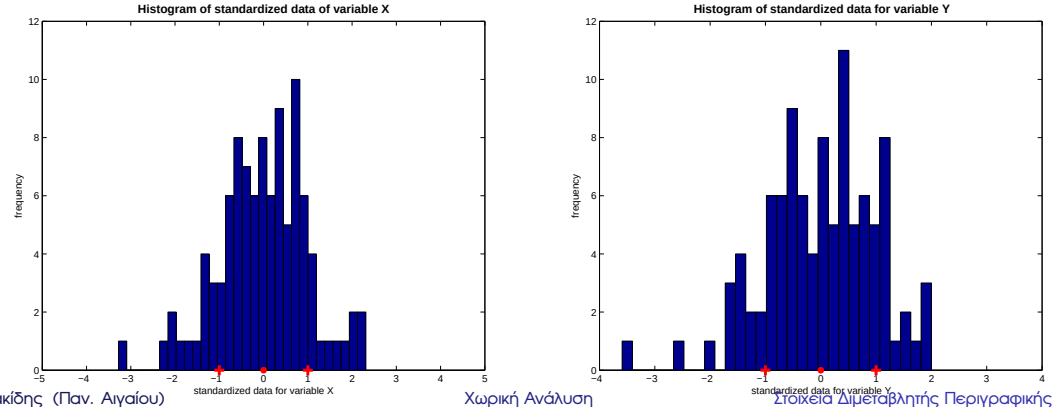
Στατιστικά Επιμέρους Τυποποιημένων Αποκλίσεων

Μέσοι όροι και διακυμάνσεις

Η τυποποιημένη απόκλιση της μέτρησης x_i από το μέσο όρο m_X της μεταβλητής X ορίζεται ως: $(x_i - m_X)/s_X$. Οι N τυποποιημένες αποκλίσεις της μεταβλητής X έχουν μέσο όρο $m_{(X-m_X)/s_X} = 0$ και τυπική απόκλιση $s_{(X-m_X)/s_X} = 1$, και αντίστοιχα για τις αποκλίσεις της μεταβλητής Y

$$m_{(X-m_X)/s_X} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - m_X)}{s_X} = 0, \quad s_{(X-m_X)/s_X} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\frac{(x_i - m_X)}{s_X} - 0]^2} = 1$$

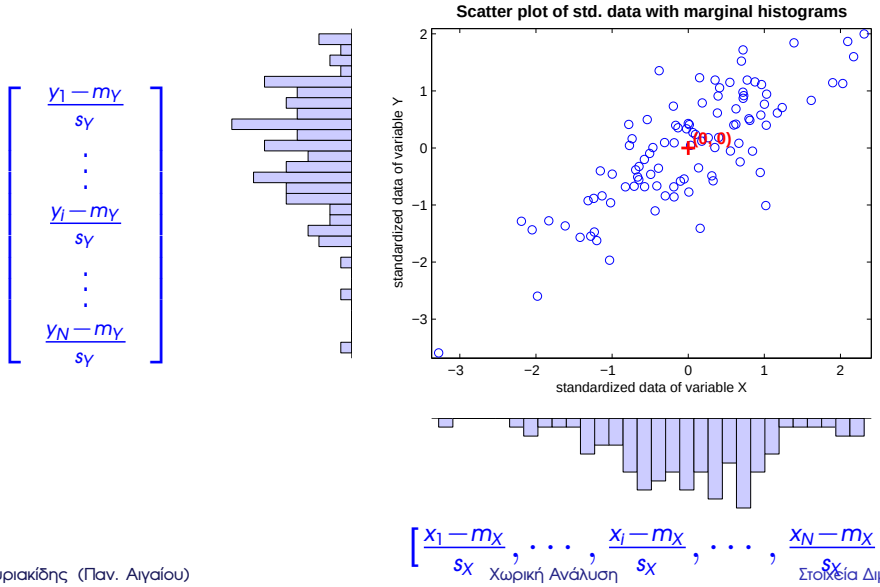
Ιστογράμματα συχνότητας τυποποιημένων αποκλίσεων



Περιθώρια Ιστογράμματα Τυποποιημένων Αποκλίσεων

Διασπορόγραμμα με περιθώρια ιστογράμματα

Τα ιστογράμματα συχνότητας των τυποποιημένων αποκλίσεων $\frac{x_i - m_X}{s_X}$ και $\frac{y_i - m_Y}{s_Y}$ των επιμέρους μεταβλητών X και Y, αποτελούν περιθώρια ιστογράμματα του διασπορογράμματος των N ζευγών αποκλίσεων $\{(\frac{x_i - m_X}{s_X}, \frac{y_i - m_Y}{s_Y}), i = 1, \ldots, N\}$. Το κέντρο του διασπορογράμματος είναι ένα σημείο με "συντεταγμένες" $(0, 0)$. **Προσοχή:** Οι "συντεταγμένες" του διασπορογράμματος έχουν αλλάξει για να ληφθεί υπ' όψην η διαφορετική αρχική διασπορά s_X και s_Y των x- και y-τιμών



Συντελεστής (Γραμμικής) Συσχέτισης

Ορισμός

Συντελεστής (γραμμικής) συσχέτισης r_{XY} δύο μεταβλητών X και Y ονομάζεται η αναμενόμενη τιμή (μέσος όρος) των γινομένων των τυποποιημένων αποκλίσεων των τιμών της κάθε μεταβλητής από τους αντίστοιχους μέσους όρους m_X και m_Y

$$r_{XY} = \frac{1}{N} \sum_{i=1}^{N} (\frac{x_i - m_X}{s_X})(\frac{y_i - m_Y}{s_Y}) = \frac{s_{XY}}{s_X s_Y}$$

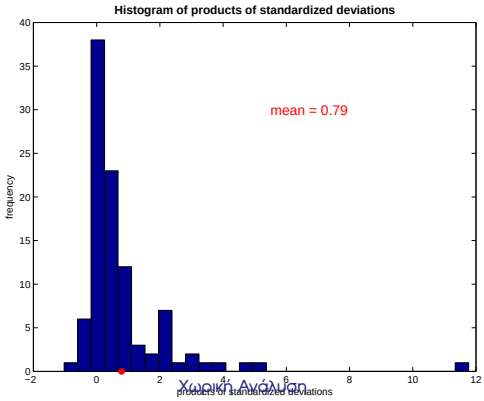
Σημειώσεις

- Ο συντελεστής συσχέτισης αποτελεί ένα μέτρο συνάφειας μεταξύ N ζευγών μετρήσεων, δεν έχει μονάδες μέτρησης και κυμαίνεται στο διάστημα $[-1, 1]$
- Η τιμή $r_{XY} = 0$ υποδηλώνει έλλειψη συσχέτισης, ενώ $r_{XY} > 0$ και $r_{XY} < 0$ αντιστοιχούν σε θετική και αρνητική συσχέτιση. Οι τιμές $r_{XY} = 1$ και $r_{XY} = -1$ υποδηλώνουν τέλεια θετική και τέλεια αρνητική γραμμική συσχέτιση, αντίστοιχα (στις περιπτώσεις αυτές, τα σημεία του διασπορογράμματος "πέφτουν" σε μια ευθεία)
- Η τιμή του συντελεστή συσχέτισης r_{XY} δύο μεταβλητών X και Y δεν επηρεάζεται από το μέσο όρο και τη διακύμναση της κάθε μεταβλητής. Δηλαδή, για δύο νέες μεταβλητές $X' = a + b \cdot X$ και $Y' = c + d \cdot Y$ (γραμμικός μετασχηματισμός των μεταβλητών X και Y), ισχύει: $r_{X'Y'} = r_{XY}$

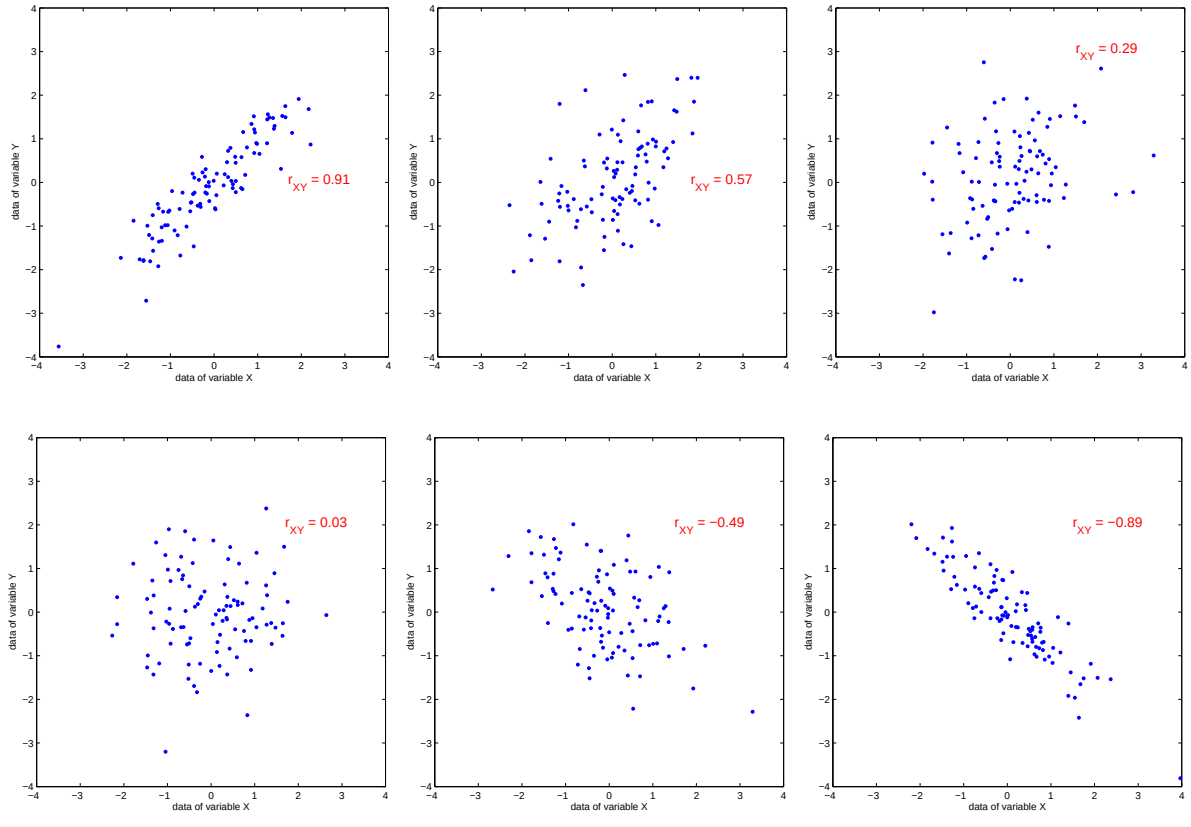
Συντελεστής (Γραμμικής) Συσχέτισης (2)

Σε κάθε ζεύγος μετρήσεων x_i, y_i αντιστοιχεί ένα ζεύγος τυποποιημένων αποκλίσεων $\frac{(x_i - m_X)}{s_X}, \frac{(y_i - m_Y)}{s_Y}$ και ένα γινόμενο $\frac{(x_i - m_X)}{s_X} \cdot \frac{(y_i - m_Y)}{s_Y}$. Ο μέσος όρος των N γινομένων ονομάζεται συντελεστής συσχέτισης των N ζευγών μετρήσεων των δύο μεταβλητών

$$\begin{bmatrix} \frac{x_1 - m_X}{s_X} \\ \vdots \\ \frac{x_i - m_X}{s_X} \\ \vdots \\ \frac{x_N - m_X}{s_X} \end{bmatrix} \odot \begin{bmatrix} \frac{y_1 - m_Y}{s_Y} \\ \vdots \\ \frac{y_i - m_Y}{s_Y} \\ \vdots \\ \frac{y_N - m_Y}{s_Y} \end{bmatrix} = \begin{bmatrix} \frac{(x_1 - m_X)}{s_X} \cdot \frac{(y_1 - m_Y)}{s_Y} \\ \vdots \\ \frac{(x_i - m_X)}{s_X} \cdot \frac{(y_i - m_Y)}{s_Y} \\ \vdots \\ \frac{(x_N - m_X)}{s_X} \cdot \frac{(y_N - m_Y)}{s_Y} \end{bmatrix}$$



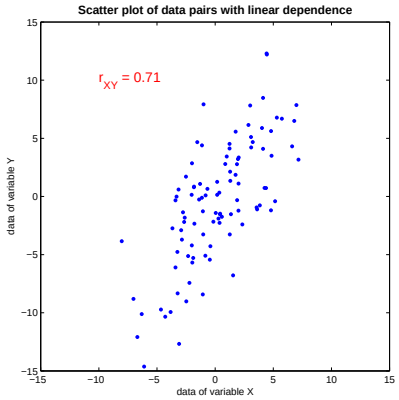
Παραδείγματα Διασπορογραμμάτων

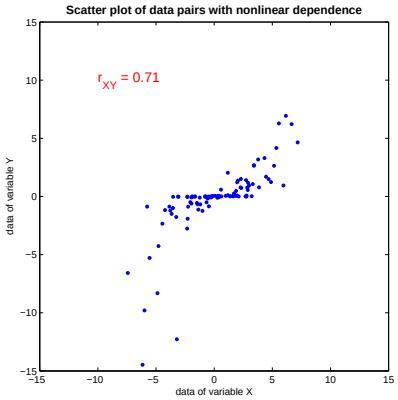


Προβλήματα του Συντελεστή (Γραμμικής) Συσχέτισης

Τι ακριβώς περιγράφει ο συντελεστής συσχέτισης

Ο συντελεστής συσχέτισης r_{XY} μεταξύ N ζευγών μετρήσεων δύο μεταβλητών ποσοτικοποιεί γραμμική συμμεταβολή. Με απλά λόγια, περιμένουμε ένα διασπορόγραμμα με σχετικά υψηλό συντελεστή γραμμικής συσχέτισης (αριστερά) να παρουσιάζεται "ομοιόμορφα" διεσπαρμένο γύρω από μια ευθεία γραμμή που "περνάει" βέλτιστα από τα δεδομένα. Επιπλέον, ο συντελεστής γραμμικής συσχέτισης είναι συνδεδεμένος με την κλίση της ευθείας αυτής





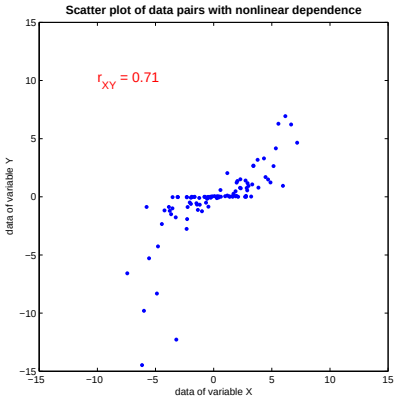
Το πρόβλημα

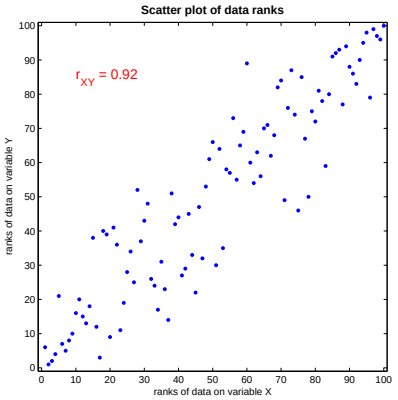
Ο συντελεστής γραμμικής συσχέτισης δεν είναι επαρκής για να περιγράψει ένα διασπορόγραμμα το οποίο εμφανίζει ισχυρή **μη-γραμμική** συμμεταβολή (δεξιά)

Διατάξεις (Ranks) Μετρήσεων

Διαδικασία Υπολογισμού

Κάθε τιμή x_i μετατρέπεται σε έναν ακέραιο r_i^X στο διάστημα $[1, N]$, όπου $r_i^X = 1$ αν η τιμή x_i είναι η μικρότερη στο δείγμα, και $r_i^X = N$ αν η τιμή x_i είναι η μεγαλύτερη στο δείγμα. Αντίστοιχα, υπολογίζονται οι διατάξεις $\{r_i^Y, i = 1, \ldots, N\}$ για τις μετρήσεις της μεταβλητής Y. Όταν μια τιμή επαναμβάνεται στο δείγμα, τότε η αντίστοιχη διατάξη ορίζεται ως ο μέσος όρος των διατάξεων των ιδίων τιμών





Αν οι τιμές της μεταβλητής X και της μεταβλητής Y σχετίζονται μη γραμμικά αλλά **μονοτονικά** (αριστερά), τότε το διασπορόγραμμα των διατάξεων των μετρήσεων (δεξιά) εμφανίζει γραμμική (ή τουλάχιστον λιγότερο μη γραμμική) σχέση

Συντελεστής Συσχέτισης Διατάξεων του Spearman

Διαδικασία Υπολογισμού

- Ο συντελεστής συσχέτισης του Spearman είναι ο συντελεστής συσχέτισης των διατάξεων των μετρήσεων (rank correlation coefficient):

$$r_{XY}^S = \frac{1}{N}\sum_{i=1}^{N}\frac{r_i^X - m_{r_X}}{s_{r_X}}\frac{r_i^Y - m_{r_Y}}{s_{r_Y}}$$

  όπου m_{r_X} και s_{r_X} είναι ο μέσος όρος και η τυπική απόκλιση, αντίστοιχα, των διατάξεων των μετρήσεων της μεταβλητής X
- Όταν δεν υπάρχουν επαναλαμβανόμενες μετρήσεις στο δείγμα, ο συντελεστής r_{XY}^S υπολογίζεται ως:

$$r_{XY}^S = 1 - \frac{6}{N(N^2-1)}\sum_{i=1}^{N}[r_i^X - r_i^Y]^2$$

Χρησιμότητα

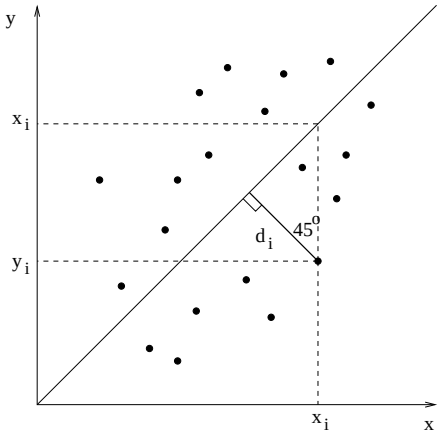
Ο συντελεστής συσχέτισης διατάξεων του Spearman χρησιμοποιείται όταν: (1) η σχέση των μετρήσεων των δύο μεταβλητών X και Y εμφανίζεται ισχυρά μη γραμμική, (2) οι τιμές των μεταβλητών αφορούν τάξεις, κι όχι μετρήσεις συνεχούς κλίμακας, (3) εμφανίζονται ακραίες τιμές στο δείγμα, και (4) οι επιμέρους (περιθώριες) κατανομές δεν προσεγγίζονται ικανοποιητικά από την Κανονική κατανομή

Τετραγωνική Ημιδιαφορά Ζεύγους Μετρήσεων

"Απόσταση" σημείου από διχοτόμο

Τετραγωνική απόσταση d_i^2 του σημείου με συντεταγμένες (x_i, y_i) από το πλησιέστερο σημείο στην ευθεία των 45°:

$$\cos(45) = \frac{d_i}{|x_i - y_i|} \Rightarrow d_i = \frac{\sqrt{2}}{2}|x_i - y_i| \Rightarrow d_i^2 = \frac{1}{2}(x_i - y_i)^2$$



Με άλλα λόγια

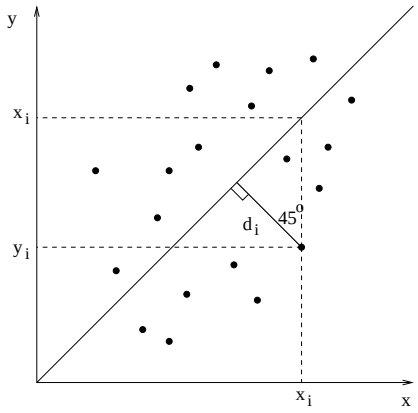
Η τετραγωνική ημιδιαφορά $\frac{1}{2}(x_i - y_i)^2$ μεταξύ των δύο τιμών που απαρτίζουν ένα ζεύγος μετρήσεων, μπορεί να ερμηνευτεί και ως ένα μέτρο της απόστασης του σημείου (x_i, y_i) από τη διχοτόμο. **Προσοχή:** Η διχοτόμος ενός διασπορογράμματος δεν έχει πάντα νόημα...

Ροπή Αδράνειας Διασπορογράμματος Ι

Ορισμός

Μέσος όρος τετραγωνικών αποστάσεων σημείων του διασπορογράμματος από τη διχοτόμο, ή μέσος όρος των N τετραγωνικών ημιδιαφορών των N ζευγών μετρήσεων:

$$g_{XY} = \frac{1}{N}\sum_{i=1}^{N} d_i^2 = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}(x_i - y_i)^2 = \frac{1}{2N}\sum_{i=1}^{N}(x_i - y_i)^2$$



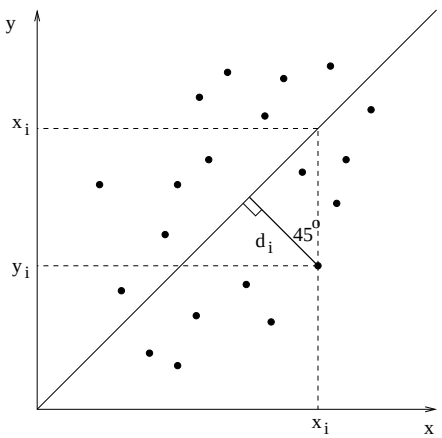
Ερμηνεία

Όσο πιο διαφορετικά είναι μεταξύ τους τα επιμέρους στοιχεία που απαρτίζουν τα N ζεύγη μετρήσεων, δηλαδή όσο πιο διαφορετικές είναι οι x-μετρήσεις από τις αντίστοιχες y-μετρήσεις, τόσο πιο 'άτακτο' είναι ένα διασπορόγραμμα

Ροπή Αδράνειας Διασπορογράμματος ΙΙ

Φυσική ερμηνεία

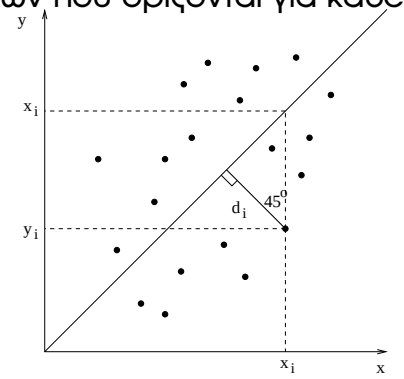
Η ροπή αδράνειας ενός σώματος που βρίσκεται σε μια θέση με συντεταγμένες (x_i, y_i) συνδέεται με τη δύναμη που πρέπει να ασκήσει κανείς στο σώμα αυτό για να το περιστρέψει γύρω από κάποιον άξονα. Η δύναμη που απαιτείται είναι συνάρτηση της μάζας του σώματος και της απόστασής του από τον άξονα περιστροφής. Για σταθερή μάζα, όσο πιο μακριά βρίσκεται το σώμα από τον άξονα περιστροφής, τόσο πιο μεγάλη είναι η δύναμη που πρέπει να ασκηθεί για την περιστοφή του



Ροπή Αδράνειας Διασπορογράμματος III

Στατιστική αναλογία

Στην περίπτωση ενός σημείου (x_i, y_i) ενός διασπορογράμματος, δεν μιλάμε για κάποιο σώμα, άρα δεν υπεισέρχεται η έννοια της μάζας, ή ισοδύναμα μπορούμε να θεωρήσουμε ότι η μάζα είναι σταθερή (έστω 1) για οποιοδήποτε σημείο. Στην περίπτωση ενός διασπορογράμματος με N σημεία, η ροπή αδράνειας g_{XY} του διασπορογράμματος είναι ο μέσος όρος των N ροπών που ορίζονται για κάθε ένα από τα N σημεία



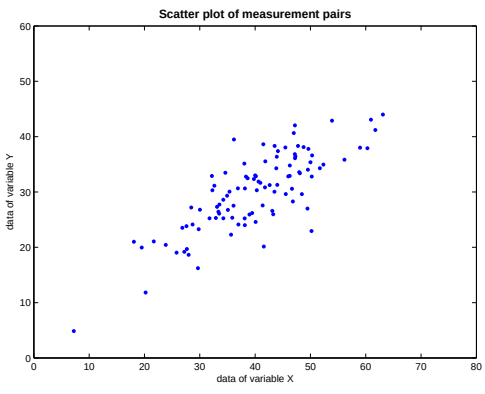
Ελάχιστη τιμή ροπής αδράνειας

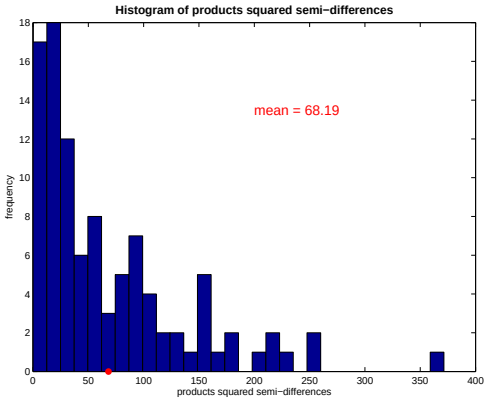
Η ροπή αδράνειας ενός διασπορογράμματος του οποίου τα σημεία "πέφτουν" στη διχοτόμο είναι 0. Δηλαδή, αν $x_i \equiv y_i, \forall i$, τότε $g_{XY} = g_{XX} = g_{YY} = 0$, αφού η μέση διαφορά των τιμών μιας μεταβλητής από τους εαυτούς τους είναι 0

Υπολογισμός Ροπής Αδράνειας Διασπορογράμματος

Σε κάθε ζεύγος μετρήσεων x_i, y_i αντιστοιχεί ένα ζεύγος διαφορών $\frac{(x_i - y_i)}{\sqrt{2}}, \frac{(x_i - y_i)}{\sqrt{2}}$ και ένα γινόμενο $\frac{(x_i - y_i)^2}{2}$. Ο μέσος όρος των N γινομένων τετραγωνικών ημιδιαφορών είναι η ροπή αδράνειας g_{XY} του διασπορογράμματος

$$\begin{bmatrix} \frac{x_1 - y_1}{\sqrt{2}} \\ \vdots \\ \frac{x_I - y_I}{\sqrt{2}} \\ \vdots \\ \frac{x_N - y_N}{\sqrt{2}} \end{bmatrix} \odot \begin{bmatrix} \frac{x_1 - y_1}{\sqrt{2}} \\ \vdots \\ \frac{x_I - y_I}{\sqrt{2}} \\ \vdots \\ \frac{x_N - y_N}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{(x_1 - y_I)^2}{2} \\ \vdots \\ \frac{(x_I - y_I)^2}{2} \\ \vdots \\ \frac{(x_N - y_N)^2}{2} \end{bmatrix}$$

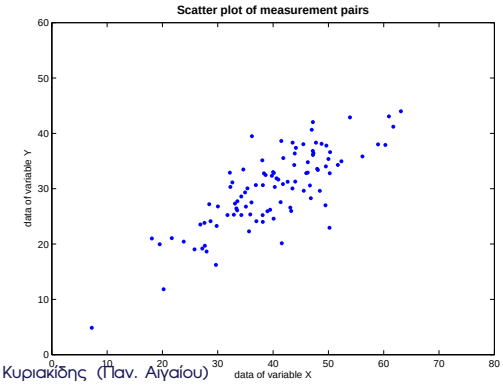


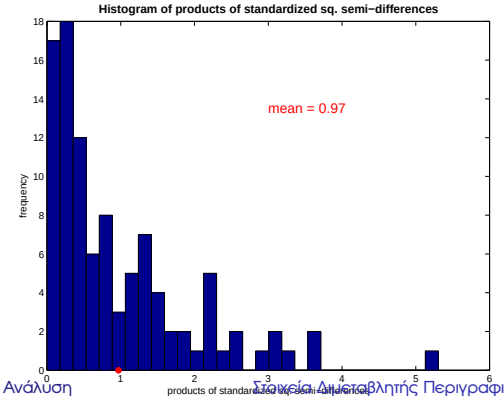


Τυποποιημένη Ροπή Αδράνειας Διασπορογράμματος

Επειδή οι τετραγωνικές ημιδιαφορές έχουν ασαφείς μονάδες, και μπορούν να είναι πολύ μεγάλες, εξετάζουμε τις τυποποιημένες διαφορές $\frac{(x_i-y_i)}{\sqrt{2}s_X}$, $\frac{(x_i-y_i)}{\sqrt{2}s_Y}$, και συνεπώς τα γινόμενα $\frac{(x_i-y_i)^2}{2s_Xs_Y}$. Ο μέσος όρος των N γινομένων τυποποιημένων τετραγωνικών ημιδιαφορών είναι η τυποποιημένη ροπή αδράνειας g^S_{XY} του διασπορογράμματος

$$\begin{bmatrix} \frac{x_1-y_1}{\sqrt{2}s_X} \\ \vdots \\ \frac{x_i-y_i}{\sqrt{2}s_X} \\ \vdots \\ \frac{x_N-y_N}{\sqrt{2}s_X} \end{bmatrix} \odot \begin{bmatrix} \frac{x_1-y_1}{\sqrt{2}s_U} \\ \vdots \\ \frac{x_i-y_i}{\sqrt{2}s_Y} \\ \vdots \\ \frac{x_N-y_N}{\sqrt{2}s_Y} \end{bmatrix} = \begin{bmatrix} \frac{(x_1-y_i)^2}{2s_Xs_Y} \\ \vdots \\ \frac{(x_i-y_i)^2}{2s_Xs_Y} \\ \vdots \\ \frac{(x_N-y_N)^2}{2s_Vs_V} \end{bmatrix}$$





Βασικά Σημεία Διάλεξης

Μονομεταβλητά περιγραφικά στατιστικά δύο μεταβλητών

Ξεχωριστή παρουσίαση και ανάλυση μετρήσεων δύο μεταβλητών: περιθώρια στατιστικά, περιθώρια ιστογράμματα συχνοτήτων, κάθε μεταβλητής...

Διμεταβλητά περιγραφικά στατιστικά

Από κοινού παρουσίαση και ανάλυση συμμεταβολής ζευγών μετρήσεων δύο μεταβλητών: συνδιακύμανση, συντελεστής συσχέτισης (Pearson), συντελεστής συσχέτισης διατάξεων (Spearman), ροπή αδράνειας (moment of inertia). Σχέσεις μεταξύ τους: $r_{XY} = s_{XY}/(s_Xs_Y)$, $s_{XY} = r_{XY}s_Xs_Y$, $g_{XY} = s_X + s_Y - 2s_{XY} + [m_X - m_Y]^2$. Καθένα από τα παραπάνω στατιστικά περιγράφει ένα διασπορόγραμμα ζευγών μετρήσεων: τα πρώτα δύο περιγράφουν τη (μέση) συνάφεια μεταξύ των ζευγών, ενώ το τελευταίο περιγράφει τη (μέση) διαφορά τους

Παρατηρήσεις

Ο συντελεστής συσχέτισης χρησιμοποιείται περισσότερο από τη συνδιακύμανση, διότι η τιμή του δεν επηρεάζεται από τις μονάδες μέτρησης των επιμέρους μεταβλητών. Η ροπή αδράνειας χρησιμοποιείται όταν οι μετρήσεις αφορούν την ίδια μεταβλητή, αλλά αντιστοιχούν σε διαφορετικές χρονικές στιγμές ή σε διαφορετικά όργανα μέτρησης. Οι επιμέρους μέσοι όροι των μεταβλητών δεν υπεισέρχονται στον υπολογισμό της ροπής αδράνειας