Πανεπιστήμιο Κρήτης - Τμήμα Επιστήμης Υπολογιστών

HY-317: Εφαρμοσμένες Στοχαστικές Διαδικασίες-Εαρινό Εξάμηνο 2016
Διδάσκων: Π. Τσακαλίδης

Συνδιασπορά - Συσχέτιση Τυχαίων Μεταβλητών

Επιμέλεια: Κωνσταντίνα Φωτιάδου

1. Συνδιασπορά Τυχαίων Μεταβλητών:

Η συνδιασπορά δύο τυχαίων μεταβλητών X και X, συμβολίζεται ως: $cov(X, Y)$, και ορίζεται ως:

$$cov(X, Y) = E[(X - E[X]) \cdot (Y - E[Y])]$$

Επίσης, ένας εναλλακτικός ορισμός της συνδιασποράς είναι ο ακόλουθος:

$$cov(X, Y) = E[X \cdot Y] - E[X] \cdot E[Y]$$

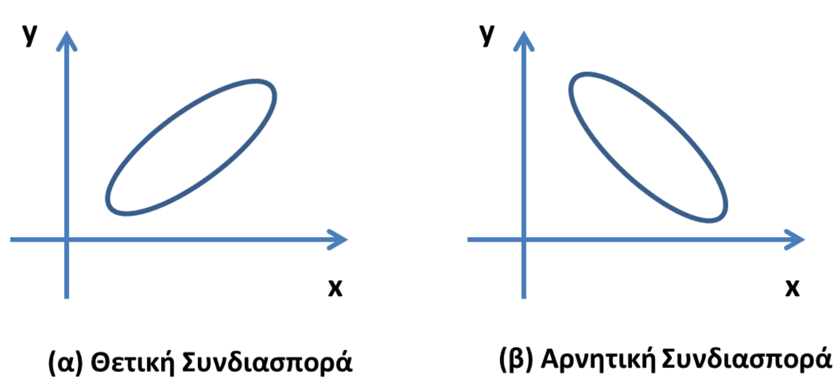
Απόδειξη:

$$cov(X, Y) = E\Big[(X - E[X]) \cdot (Y - E[Y])\Big] = E\Big[X \cdot Y - X \cdot E[Y] - Y \cdot E[X] + E[X] \cdot E[Y]\Big] = E[X \cdot Y] - E[X] \cdot E[Y] - E[Y] \cdot E[X] + E[X] \cdot E[Y] = E[X \cdot Y] - E[X] \cdot E[Y]$$

1.1. Παρατηρήσεις

- Το πρόσημο της συνδιασποράς περιέχει ένα σημαντικό δείκτη της σχέσης των τυχαίων μεταβλητών X και Y.

- Μια θετική ή αρνητική συνδιασπορά μας δείχνει ότι οι τιμές των $X - E[X]$ και $Y - E[Y]$ τείνουν να έχουν ίδιο ή αντίθετο πρόσημο σ' ενα συγκεκριμένο μοναδικό πείραμα. Οι παρακάτω γραφικές παραστάσεις απεικονίζουν δύο γραφικά παραδείγματα θετικής και αρνητικής διασποράς, όπου οι τυχαίες μεταβλητές X και Y είναι ομοιόμορφα κατανεμημένες πάνω σε ελλείψεις.



Σχήμα 1: Παραδείγματα Θετικής και Αρνητικής Συνδιασποράς

Χρήσιμες Ιδιότητες Συνδιασποράς

1. $cov(X,Y) = cov(Y,X)$
2. $cov(X,X) = var(X)$
3. $cov(X+Y,Z) = cov(X,Z) + cov(Y,Z)$
4. $cov(c \cdot X, Y) = c \cdot cov(X,Y)$

1.2. Ασυσχέτιστες Τυχαίες Μεταβλητές

- Όταν $cov(X,Y) = 0$, οι τυχαίες μεταβλητές X και Y είναι ασυσχέτιστες.
- Αν δύο τυχαίες μεταβλητές X και Y είναι ανεξάρτητες, τότε θα είναι και ασυσχέτιστες.

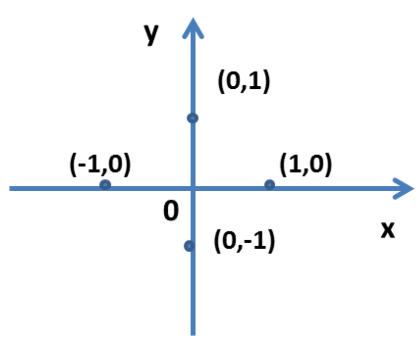
  Απόδειξη:
  Έχουμε ότι:

  $$\begin{aligned} cov(X,Y) = E[(X-E[X]) \cdot (Y-E[Y])] &= E[(X-E[X])] \cdot E[(Y-E[Y])] \\ &= (E[X]-E[X]) \cdot (E[Y]-E[Y]) = 0 \end{aligned}$$

  Οπότε οι τ.μ X και Y είναι ασυσχέτιστες.

- **Προσοχή!** Το αντίστροφο δεν ισχύει πάντα.
  Εάν δύο τυχαίες μεταβλητές είναι ασυσχέτιστες, δεν είναι απαραίτητα ανεξάρτητες!

Παράδειγμα 1.2.1
Έστω ζεύγος τυχαίων μεταβλητών (X,Y) που παίρνει τιμές (1,0), (0,1), (-1,0), (0,-1), κάθε μία με πιθανότητα $p = 1/4$. Εξετάζουμε αν οι τ.μ X και Y είναι ασυσχέτιστες και ανεξάρτητες.



Σχήμα 2: Γραφική παράσταση της κατανομής των τ.μ X και Y του παραδείγματος 1.2.

Στο παραπάνω σχήμα παριστάνεται η κατανομή των τυχαίων μεταβλητών X και Y.
Όπως παρατηρούμε, είναι συμμετρικές ως προς την αρχή των αξόνων.
Για τις μέσες τιμές των τ.μ X και Y ισχύει: $E[X] = E[Y] = 0$.
Επιπλέον, για κάθε ζεύγος (x,y), είτε $x = 0$, είτε $y = 0$. Οπότε συνεπάγεται ότι το γινόμενο $x \cdot y = 0$.
Επομένως, η συνδιασπορά των τ.μ X και Y γίνεται:

$$cov(X,Y) = E[(X-E[X]) \cdot (Y-E[Y])] = E[X \cdot Y] - E[X] \cdot E[Y] = 0$$

Άρα, οι τ.μ X και Y είναι ασυσχέτιστες.

Όμως, δεν είναι ανεξάρτητες! Για παράδειγμα, μια μη-μηδενική τιμή της X προσδιορίζει την τιμή της Y στο μηδεν.

1.3. Συντελεστής Συσχέτισης Τυχαίων Μεταβλητών:

Συντελεστή συσχέτισης δύο τυχαίων μεταβλητών X και Y, συμβολίζουμε την ποσότητα: $\rho(X,Y)$ και ορίζουμε ως εξής:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}}$$

1.3.1. Παρατηρήσεις

- Θεωρείται μια κανονικοποιημένη μορφή της συνδιασποράς των τυχαίων μεταβλητών X και Y.
- Το ρ παίρνει τιμές μεταξύ -1 και 1. Δηλαδή: $\rho \in [-1, 1]$.
- Όταν $\rho > 0$, οι τιμές $X - E[X]$ και $Y - E[Y]$ τείνουν να έχουν το ίδιο πρόσημο.
- Όταν $\rho < 0$, οι τιμές $X - E[X]$ και $Y - E[Y]$ τείνουν να έχουν αντίθετο πρόσημο.
- Το μέγεθος του $|\rho|$ παρέχει ένα κανονικοποιημένο μέτρο του πόσο αυτό αληθεύει.
- Όσο πιο κοντά βρίσκεται η τιμή του $|\rho|$ στο 1, τόσο πιο έντονη γίνεται αυτή η τάση.
- Υποθέτοντας ότι οι X, Y έχουν θετικές διασπορές, αποδεικνύεται ότι $\rho = 1$ ή $\rho = -1$, αν και μόνο αν υπάρχει μια θετική (ή αντίστοιχα αρνητική) σταθερά c ώστε:

$$Y - E[Y] = c \cdot \Big(X - E[X]\Big), \;\; \forall (x,y)$$

Παράδειγμα 1.3.1
Θεωρείστε n ανεξάρτητες ρίψεις ενός νομίσματος με πιθανότητα να έρθει κορώνα ίση με p. Έστω X και Y ο αριθμός των κορώνων και των γραμμάτων, αντίστοιχα.
Να μελετηθεί η συσχέτιση των X και Y.

Λύση
Για κάθε ζευγάρι (x, y) έχουμε ότι: $x + y = n$. Επειδή οι τυχαίες μεταβλητές X και Y είναι διωνυμικές με παραμέτρους p και $1 - p$, αντίστοιχα, έχουμε:

$$E[X] + E[Y] = n \cdot p + n \cdot (1-p) = np + n - np = n$$

Επίσης, $\forall (x, y)$ έχουμε:

$$X - E[X] = -\Big(Y - E[Y]\Big)$$

Για να υπολογίσουμε το συντελεστή συσχέτισης των τυχαίων μεταβλητών X και Y, θα πρέπει πρώτα να βρούμε τη συνδιασπορά των X και Y.

$$cov(X,Y) = E\Big[(X - E[X]) \cdot (Y - E[y])\Big] = -E\Big[(X - E[X]) \cdot (X - E[X])\Big] = -E\Big[(X - E[X])^2\Big] = -var(X)$$

Επομένως, έχουμε:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}} = \frac{-var(X)}{\sqrt{var(X) \cdot var(X)}} = \frac{-var(X)}{\sqrt{var^2(X)}} = -1$$

2. Συσχέτιση τυχαίων μεταβλητών (Correlation)

Ορισμός: Η συσχέτιση δύο τυχαίων μεταβλητών X και Y ορίζεται ως:

$$E[X \cdot Y] = \int_{-\infty}^{+\infty} x \cdot y \cdot f_{X,Y}(x,y) dxdy$$

2.1. Παρατηρήσεις

- Δύο τυχαίες μεταβλητές X και Y λέγονται ορθογώνιες, αν η συσχέτισή τους είναι μηδενική, δηλαδή αν: $E[X \cdot Y] = 0$.
- Προκύπτει ότι αν οι τυχαίες μεταβλητές X και Y είναι στατιστικά ασυσχέτιστες, τότε οι τυχαίες μεταβλητές $X - \mu_X$ και $Y - \mu_Y$, είναι ορθογώνιες.

2.2. Θεωρήματα Συσχέτισης τυχαίων μεταβλητών

Θεώρημα 1.
Αν οι τυχαίες μεταβλητές X και Y είναι στατιστικά ασυσχέτιστες, τότε η διασπορά του αθροίσματος είναι ίση με το άθροισμα των διασπορών των X και Y.

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

Απόδειξη
Θεωρούμε την τυχαία μεταβλητή $Z = X + Y \Leftrightarrow \mu_Z = \mu_X + \mu_Y$.
Η διασπορά της τ.μ Z θα είναι:

$$\sigma_Z^2 = E\Big[(Z - \mu_Z)^2\Big] = E\Big[\Big((X - \mu_X) + (Y - \mu_Y)^2\Big)\Big] = E\Big[(X - \mu_X)^2\Big] + 2E\Big[X - \mu_X\Big] \cdot E\Big[Y - \mu_Y\Big] + E\Big[(Y - \mu_Y)^2\Big] = \sigma_X^2 + \sigma_Y^2 + cov(X,Y) = \sigma_X^2 + \sigma_Y^2, \text{ εφόσον X,Y: ασυσχέτιστες}$$

Θεώρημα 2.
Αν οι τυχαίες μεταβλητές X και Y είναι ορθογώνιες τότε:

$$E[(X+Y)^2] = E[X^2] + E[Y^2]$$

Απόδειξη
Έχουμε ότι:

$$E[(X+Y)^2] = E[X^2] + 2E[X \cdot Y] + E[Y^2] = E[X^2] + E[Y^2], \quad \text{εφόσον οι τ.μ X,Y είναι ορθογώνιες}$$

Θεώρημα 3.
Αν οι τυχαίες μεταβλητές X και Y είναι ασυσχέτιστες και η μέση τιμή της μιας τουλάχιστον είναι μηδέν, τότε είναι ορθογώνιες.

Απόδειξη
Εφόσον οι τ.μ X και Y είναι ασυσχέτιστες, θα ισχύει: $E[X \cdot Y] = E[X] \cdot E[Y]$.
Επίσης, είτε $E[X] = 0$ ή $E[Y] = 0 \Rightarrow E[X \cdot Y] = 0 \Rightarrow$ οπότε οι τ.μ X και Y είναι ορθογώνιες.

Ανισότητα Schwarz
Για δύο τυχαίες μεταβλητές X και Y ισχύει η ακόλουθη ανισότητα:

$$\Big(E[X \cdot Y]\Big)^2 \leq E[X^2] \cdot E[Y^2]$$

Απόδειξη
Ορίζουμε την τυχαία μεταβλητή $X - \lambda Y, \ \lambda \in \mathbb{R}$.
Και έστω $P(\lambda) = E\Big[(X - \lambda Y)^2\Big] \ \forall \lambda \in \mathbb{R}$ έχουμε:

$$\begin{aligned} 0 \leq P(\lambda) = E\Big[(X-\lambda Y)^2\Big] &= \lambda^2 \cdot E[Y^2] + E[X^2] - 2 \cdot \lambda \cdot E[X \cdot Y] \quad \Leftrightarrow \\ \lambda^2 \cdot E[Y^2] - 2\lambda \cdot E[X \cdot Y] + E[X^2] &= 0 \quad \Leftrightarrow \end{aligned}$$

Βρίσκουμε την πρώτη παράγωγο ως προς λ:

$$P'(\lambda) = 2\lambda \cdot E[Y^2] - 2E[X \cdot Y]$$

Εξετάζουμε το πρόσημο και της δεύτερης παραγώγου της $P(\lambda)$:

$$P''(\lambda) = 2E[Y^2] > 0$$

Επειδή η δεύτερη παράγωγος είναι θετική, το τριώνυμο έχει τοπικό ελάχιστο. Μηδενίζοντας την πρώτη παράγωγο, βρίσκουμε το ελάχιστο:

$$\begin{aligned} 2\lambda \cdot E[Y^2] - 2E[X \cdot Y] = 0 \quad &\Leftrightarrow \\ \lambda_{min} = \frac{E[X \cdot Y]}{E[Y^2]} \end{aligned}$$

Υπολογίζουμε την αντίστοιχη τιμή του τριωνύμου:

$$\begin{aligned} 0 \leq E[X^2] + \frac{\Big(E[X \cdot Y]\Big)^2}{E[Y^2]} - 2 \cdot \frac{E[X \cdot Y]}{E[Y^2]} \cdot E[X \cdot Y] \quad &\Leftrightarrow \\ 0 \leq E[X^2] - \frac{\Big(E[X \cdot Y]\Big)^2}{E[Y^2]} \quad &\Leftrightarrow \\ \Big(E[X \cdot Y]\Big)^2 \leq E[X^2] \cdot E[Y^2], \quad \text{Αποδείχθηκε} \end{aligned}$$

3. Συνδιασπορά Αθροίσματος Τυχαίων μεταβλητών

Μέσω της έννοιας της συνδιασποράς, μπορούμε να επεκτείνουμε την έννοια της διασποράς του αθροίσματος τ.μ και σε αθροίσματα μη- ανεξάρτητων τ.μ.
Εάν $X_1, X_2, \cdots, X_n$, τ.μ με πεπερασμένες διασπορές, έχουμε:

$$var(X_1 + X_2) = var(X_1) + var(X_2) + 2 \cdot cov(X_1, X_2)$$

Στην γενικευμένη περίπτωση, έχουμε ότι:

$$var(X) = var\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} var(X_i) + 2\sum_{i}^{n-1}\sum_{j=i+1}^{n} cov(X_i, X_j)$$

Απόδειξη
Θέτουμε: $\tilde{X}_i = X_i - E[X_i]$
Έχουμε:

$$\begin{aligned} var\Big(\sum_{i=1}^{n} X_i\Big) = E\Big[(\sum_{i=1}^{n} \tilde{X}_i)^2\Big] = E\Big[\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{X}_i\tilde{X}_j\Big] = \sum_{i=1}^{n}\sum_{j=1}^{n} E[\tilde{X}_i \cdot \tilde{X}_j] &= \\ \sum_{i=1}^{n} E[\tilde{X}_i^2] + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} E[\tilde{X}_i \cdot \tilde{X}_j] &= \\ \sum_{i=1}^{n} var(X_i) + 2\sum_{i}^{n-1}\sum_{j=i+1}^{n} cov(X_i, X_j), \quad \text{Αποδείχθηκε} \end{aligned}$$

Εφαρμογή
Θεωρήστε το πρόβλημα των καπέλων, όπου n άτομα πετούν τα καπέλα τους σ'ενα κουτί και ακολούθως ανασύρουν από ένα καπέλο στην τύχη. Να βρεθεί η διασπορά της τ.μ X, η οποία εκφράζει τον αριθμό των ατόμων που ανασύρουν το δικό τους καπέλο.

Λύση
Έστω ότι $X = X_1 + X_2 + \cdots + X_n$.
Επίσης, έστω X_i: Η τυχαία μεταβλητή που παίρνει την τιμή 1, αν το $i-$οστό άτομο ανασύρει το δικό του καπέλο. Διαφορετικά, παίρνει την τιμή 0.
Η τ.μ $X_i \sim$ Bernoulli, με $p = P(X_i = 1) = \frac{1}{n}$.
Η διασπορά της τ.μ X_i, $var(X_i)$, θα είναι:

$$var(X_i) = \frac{1}{n} \cdot (1 - \frac{1}{n})$$

Για να βρούμε τη διασπορά της τ.μ X, θα πρέπει πρώτα να υπολογίσουμε την συνδιασπορά των τ.μ X_i και X_j.
Έχουμε:

$$\begin{aligned} cov(X_i, X_j) = E\Big[(X_i - E[X_i]) \cdot (X_j - E[X_j])\Big] = E[X_i \cdot X_j] - E[X_i] \cdot E[X_j] &= \\ P(X_i = 1 \text{ και} P(X_j = 1)) - P(X_i = 1) \cdot P(X_j = 1) &= \\ P(X_i = 1) \cdot P(X_j = 1 | X_i = 1) - P(X_i = 1) \cdot P(X_j = 1) &= \\ \frac{1}{n} \cdot \frac{1}{n-1} - \frac{1}{n^2} = \frac{1}{n^2(n-1)} \end{aligned}$$

Επομένως, η διασπορά της τ.μ X υπολογίζεται από τον ακόλουθο τύπο:

$$var(X) = var\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} var(X_i) + 2\sum_{i}^{n-1}\sum_{j=i+1}^{n} cov(X_i, X_j) \quad =$$

$$n \cdot \frac{1}{n} \cdot (1 - \frac{1}{n}) + 2 \cdot \frac{n(n-1)}{2} \cdot \frac{1}{n^2 \cdot (n-1)} = 1$$