



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΕΙΣΑΓΩΓΙΚΗ ΚΑΤΕΥΘΥΝΣΗ: ΜΑΘΗΜΑΤΙΚΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΜΑΘΗΜΑΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΣΤΙΣ ΦΥΣΙΚΕΣ ΕΠΙΣΤΗΜΕΣ ΚΑΙ ΣΤΙΣ
ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ

Σημειώσεις για το σεμινάριο

ΠΙΘΑΝΟΤΗΤΩΝ-ΣΤΑΤΙΣΤΙΚΗΣ

Θεοδόσης Δημητράκος

e-mail: dimitheo@aegean.gr

Σάμος
2013

Α. ΜΕΡΟΣ

ΣΤΟΙΧΕΙΑ ΠΕΡΙΓΡΑΦΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Τα δεδομένα μίας στατιστικής έρευνας δίνουν ένα πλήθος στοιχείων που αφορούν τον πληθυσμό (από τον οποίον λαμβάνουμε ένα δείγμα) που μας ενδιαφέρει να εξετάσουμε. Τα στοιχεία αυτά οργανώνονται αρχικά σε μορφή πινάκων με τέτοιο τρόπο ώστε να μπορεί κάποιος ερευνητής με μία απλή ανάγνωση να σχηματίζει μία πρώτη ιδέα για το δείγμα (ή τον πληθυσμό). Για πιο αποτελεσματική παρουσίαση γίνεται χρήση είτε γραφικών είτε αριθμητικών μεθόδων. Έστω η μεταβλητή X που λαμβάνει τις ακόλουθες τιμές 7,11,23,35,12,13,11.

Τα παραπάνω δεδομένα καλούνται *αταξινόμητα δεδομένα* και η διαγραμματική παρουσίαση τους γίνεται με τους ακόλουθους τρόπους:

(α) διάγραμμα σημείων ή ραβδόγραμμα

(β) διάγραμμα μίσχου – φύλλου ή κυκλικό διάγραμμα

(γ) ιστόγραμμα ή πολυγωνική γραμμή.

Παράδειγμα: Για το μάθημα της Στατιστικής δίνονται οι βαθμολογίες (με άριστα το 100) δύο ομάδων 10 φοιτητών:

(α) Ομάδα 1: 66,74,78,80,78,81,85,88,85,85.

(β) Ομάδα 2: 65,75,75,66,75,75,80,81,83,85.

Να γίνει το διάγραμμα σημείων για τις δύο παραπάνω ομάδες.

Όλα τα σχήματα που θα χρειαστούν για την πληρότητα των παρόντων σημειώσεων δίνονται σε ξεχωριστό αρχείο μορφής pdf.

Το *διάγραμμα σημείων* ή *σημειόγραμμα* απεικονίζεται στο σχήμα της χειρόγραφης φωτοτυπίας. Το διάγραμμα μίσχου–φύλλου είναι ένας περιγραφικός τρόπος παρουσίασης μίας κατανομής συχνοτήτων και χρησιμεύει στην κατασκευή του ιστογράμματος.

Παράδειγμα: Να γίνει το διάγραμμα μίσχου – φύλλου (φυλογράφημα) των παραπάνω βαθμολογιών στη Στατιστική.

		(65-69)	6+	6 5 6
6	6 5 6	ή	(70-74)	7- 4
7	4 8 8 5 5 5 5		(75-79)	7+ 8 8 5 5 5 5
8	0 1 5 8 5 5 0 1 3 5		(80-84)	8- 0 1 0 1 3
			(85-89)	8+ 5 8 5 5 5

Να κατασκευαστούν διαστήματα τάξεων (ή κλάσεων)

Τάξεις	Συχνότητα f_i της τάξης i	Σχετικές συχνότητες f_i/n
65-69	3	$3/20 = 0.15$
70-74	1	$1/20 = 0.05$
75-79	6	$6/20 = 0.30$
80-84	5	$5/20 = 0.25$
85-89	5	$5/20 = 0.25$
	$\sum f_i = n = 20$	$\sum f_i/n = 1$

Εύρος (πλάτος) τάξεων $\delta = 5$ ($\delta = 74 - 69 = 75 - 70 = 79 - 74$) δηλαδή αφαιρούμε δύο συνεχόμενα κάτω ή άνω όρια δύο διαδοχικών τάξεων. Η κατανομή συχνοτήτων καλείται *διακριτή ή ασυνεχής*. Για να την μετατρέψουμε σε συνεχή προσθαφαιρούμε στα όρια των τάξεων το ± 0.5 . Έτσι, έχουμε το ακόλουθο πίνακα:

Τάξεις X_{i-}	f_i	Αθροιστική F_i συχνότητα	Αθροιστική σχετική συχνότητα
64,5-69,5	3	κάτω από 69,5:3	κάτω από 69,5:3/20=0,1
69,5-74,5	1	κάτω από 74,5:4	κάτω από 74,5:4/20= 0,2
74,5-79,5	6	κάτω από 79,5:10	κάτω από 79,5:10/20=0,5
79,5-84,5	5	κάτω από 84,5:15	κάτω από 84,5:15/20=0,75
84,5-89,5	5	κάτω από 89,5:20	κάτω από 89,5:20/20=1.

Να γίνει το ιστόγραμμα και η πολυγωνική γραμμή.

Ιστόγραμμα (Histogram): Είναι η γραφική απεικόνιση μίας κατανομής με μορφή συνεχόμενων ορθογωνίων παραλληλογράμμων.

Πολυγωνική γραμμή: είναι μία γραμμή που ενώνει τα κέντρα των ορθογωνίων στηλών του ιστογράμματος και η οποία επεκτείνεται στα άκρα κατά $\delta/2$.

Με το σύμβολο που χρησιμοποιούμε στον οριζόντιο άξονα και φαίνεται στο σχήμα της φωτοτυπίας, θεωρούμε ότι συμπίεζουμε τμήματα των αξόνων, χωρίς ενδιαφέρον. Το εμβαδόν της περιοχής κάτω από την πολυγωνική γραμμή είναι πάντοτε μικρότερο από το εμβαδόν του ιστογράμματος. Με μεγάλο αριθμό υποδιαστημάτων και πολύ μικρό εύρος, προεκτείνουμε τα άκρα της πολυγωνικής γραμμής μέχρι τα άκρα του ιστογράμματος (ή μέχρι $\delta/2$ από το ανώτερο και το κατώτερο άκρο (όριο) των τάξεων). Συνεπώς, το εμβαδόν κάτω από την πολυγωνική γραμμή είναι με μεγάλη προσέγγιση ίσο με το εμβαδόν του ιστογράμματος. Η πολυγωνική γραμμή στο ιστόγραμμα έχει τετμημένη τα άνω άκρα των τάξεων και τεταγμένες τις αθροιστικές συχνότητες. Η πολυγωνική γραμμή επεκτείνεται μέχρι το κάτω άκρο του πρώτου υποδιαστήματος και καλείται *δεξιόστροφη πολυγωνική γραμμή αθροιστικών συχνοτήτων*.

Η αριστερόστροφη πολυγωνική γραμμή αθροιστικών συχνοτήτων έχει τετμημένες τα κάτω άκρα των υποδιαστημάτων και τεταγμένες τις αθροιστικές συχνότητες που προκύπτουν ως ίσες προς το κατώτερο όριο ή μεγαλύτερες από το κατώτερο όριο κάθε τάξης.

ΠΙΝΑΚΑΣ ΓΙΑ ΑΡΙΣΤΕΡΟΣΤΡΟΦΕΣ ΣΥΧΝΟΤΗΤΕΣ

Τάξεις X_i	f_i	Αθροιστική συχνότητα F_i
64,5-69,5	3	$\geq 64,5 : 20$
69,5-74,5	1	$\geq 69,5 : 17$
74,5-79,5	6	$\geq 74,5 : 16$
79,5-84,5	5	$\geq 79,5 : 10$
84,5 -89,5	5	$\geq 84,5 : 5$

Η πολυγωνική γραμμή επεκτείνεται μέχρι το άνω άκρο του τελευταίου υποδιαστήματος.

Παράδειγμα: Επιλέγουμε 20 οικογένειες τις οποίες ταξινομούμε ανά επάγγελμα πατέρα, ετήσιο μισθό σε χιλιάδες ευρώ, αριθμό παιδιών.

Ε: εργάτης , Ο: οδηγός , ΔΥ: Δημόσιος Υπάλληλος, Δ: δάσκαλος , Ι: Ιερέας.

Τα δεδομένα δίνονται παρακάτω:

Οικογένεια (i)	Επάγγελμα	(χιλιάδες ευρώ)	
		Ετήσιος μισθός	# παιδιών
1	E	70	0
2	O	75	1
3	E	80	0
4	ΔΥ	70	2
5	ΔΥ	80	2
6	ΔΥ	50	2
7	Δ	90	3
8	I	100	2
9	O	60	4
10	E	60	1
11	Δ	70	1
12	E	60	2
13	E	80	3
14	ΔΥ	70	4
15	I	90	1
16	Δ	100	2
17	E	90	2
18	ΔΥ	65	2
19	Δ	75	2
14	ΔΥ	70	4

15	I	90	1
16	Δ	100	2
17	E	90	2
18	ΔΥ	65	2
19	Δ	75	2
12	E	60	2
13	E	80	3

ΔΥ	6	0.3
Δ	4	0.2
I	2	0.1
E	6	0.3
O	2	0.1
	$\Sigma f_i = 20$	$\Sigma f'_i = 1$

ΠΙΝΑΚΑΣ ΣΥΧΝΟΤΗΤΩΝ ΓΙΑ ΤΟΝ ΕΤΗΣΙΟ ΜΙΣΘΟ

Χιλιάδες ευρώ

Μισθός X_i		f_i	f'_i	F_i	F'_i
50	I	1	0.05	1	0.05
60	III	3	0.15	4	0.2
65	I	1	0.05	5	0.25
70	IIII	4	0.2	9	0.45
75	II	2	0.1	11	0.55
80	IIII	4	0.2	15	0.75
90	III	3	0.15	18	0.9
100	II	2	0.1	20	1

ΠΙΝΑΚΑΣ ΣΥΧΝΟΤΗΤΩΝ ΓΙΑ ΤΟΝ ΑΡΙΘΜΟ ΠΑΙΔΙΩΝ

Αριθμός παιδιών X_i	f_i	f_i	f'_i	F_i	F'_i
0	II	2	0.1	2	0.1
1	IIII	4	0.2	6	0.3
2	IIII IIII	10	0.5	16	0.8
3	II	2	0.1	18	0.9
4	II	2	0.1	20	1

$$\Sigma f_i = 20$$

Να γίνει το ραβδόγραμμα της κατανομής όσον αφορά το επάγγελμα του πατέρα.

Ραβδόγραμμα: Ορθογώνια παραλληλόγραμμα στο ύψος των συχνοτήτων ή των σχετικών συχνοτήτων που απέχουν μεταξύ τους ισομήκη διαστήματα. Είναι κατακόρυφα ή οριζόντια.

Να γίνει το κυκλικό διάγραμμα των συχνοτήτων των επαγγελμάτων.

Κυκλικό διάγραμμα: Κύκλοι που διαιρούνται σε τομείς ώστε το εμβαδόν κάθε τομέα να εκφράζει την αναλογία της τιμής της αντίστοιχης κατηγορίας στο σύνολο. Χαράσσουμε κύκλο με οποιαδήποτε ακτίνα.

Για την κατανομή του αριθμού των παιδιών να γίνει το κυκλικό διάγραμμα, το διάγραμμα συχνοτήτων, το πολύγωνο συχνοτήτων και το ιστόγραμμα.

ΠΑΡΑΔΕΙΓΜΑ: Δίνεται η συγκέντρωση (mgr/cm^3) ενός συγκεκριμένου τύπου αέρος σε δείγμα που λάβαμε σε 57 πόλεις των ΗΠΑ. Να κατασκευαστεί κατανομή συχνοτήτων, το ιστόγραμμα, το ιστόγραμμα αθροιστικών συχνοτήτων, το αθροιστικό διάγραμμα, η πολυγωνική γραμμή, το φυλλογράφημα και το αντίστοιχο ιστόγραμμα.

68	63	42	27	30	36	28	32	79	27
22	23	24	25	24	65	43	25	74	51
36	42	28	31	28	25	45	12	57	51
12	32	49	38	42	27	31	50	38	21
16	24	69	47	23	22	43	27	49	48
23	12	19	46	30	49	49			

Λύση: $X_{\max} = 79$ $X_{\min} = 12 \Rightarrow R = 79 - 12 = 67$

θεωρώ πλάτος – εύρος διαστημάτων $\delta=10$ και έχω $R = X_{\max} - X_{\min} = 79 - 12 = 67$ $n: = \neq \tau\acute{\alpha}\xi\epsilon\omega\nu$

$$K = \frac{R}{\delta} = \frac{67}{10} = 6,7 \approx 7 \text{ διαστήματα τάξεων.}$$

Μπορούμε να εκτιμήσουμε το πλήθος των τάξεων και με τον προσεγγιστικό τύπο του Sturges.

$$n=57$$

$$K=1+3,322 \log n = 1+3,322 \log 57 \\ = 1+3,322 \cdot 1,755 \cong 6,83 \cong 7 \text{ κλάσεις}$$

$$\text{Το εύρος είναι: } \delta = \frac{R}{K} = \frac{67}{7} \cong 9.5 \cong 10$$

<u>Τάξεις</u>	X_i κέντρο <u>τάξης</u>	f_i	f'_i	F_i	F'_i
9,5-19,5	14,5	5	0,088	5	0.088
19,5-29,5	24,5	19	0,333	24	0.421
29,5-39,5	34,5	10	0,175	34	0.596

39,5-49,5	44,5	13	0,228	47	0.824
49,5-59,5	54,5	4	0,070	51	0.895
59,5-69,5	64,5	4	0,070	55	0.965
69,5-79,5	74,5	2	0,035	57	1

Φυλλογράφημα

1	22269
2	122334445557777888
3	0011226688
4	2223335678999
5	0117
6	3589
7	49

Ιστόγραμμα από το Φυλλογράφημα

Άσκηση. Δίνεται η παρακάτω κατανομή.

Τάξεις	f_i	Να γίνει ιστόγραμμα συχνοτήτων, πολυγωνική γραμμή, αθροιστικό διάγραμμα, ιστόγραμμα αθροιστικών συχνοτήτων.
0-10	68	
10-20	24	
20-30	12	
30-40	8	
40-50	12	
50-60	24	
60-70	68	

ΜΕΤΡΑ ΚΕΝΤΡΙΚΗΣ ΤΑΣΗΣ ΚΑΙ ΚΕΝΤΡΙΚΗΣ ΘΕΣΗΣ

Μέση τιμή ή αριθμητικός μέσος $\bar{X} = \frac{\sum_i X_i}{n}$ η πιο αντιπροσωπευτική τιμή της μεταβλητής που εξετάζουμε. Μετριέται σε μονάδες της μεταβλητής X_i . Μετράει την κεντρική τάση των τιμών της κατανομής. Ισχύει ότι $\sum_i (X_i - \bar{X}) = 0$.

Διάμεσος (M) είναι η τιμή που διαχωρίζει το σύνολο των δεδομένων στο 50% αφού πρώτα τοποθετήσουμε τις τιμές κατά αύξουσα τάξη μεγέθους. Η διάμεσος M είναι η τιμή κάτω από την οποία βρίσκεται το 50% των τιμών της μεταβλητής και πάνω από την οποία βρίσκεται το υπόλοιπο 50%. Μετριέται σε μονάδες της μεταβλητής και είναι παράμετρος κεντρικής θέσης. Η θέση της διαμέσου προσδιορίζεται όπως φαίνεται στα ακόλουθα παραδείγματα.

Θέση M: $\frac{n+1}{2}$ π.χ. $X_i : 3,1,7,11,8$

Διάταξη: 1,3,7,8,11

$$\text{Θέση M: } \frac{n+1}{2} = \frac{5+1}{2} = 3$$

Συνεπώς η διάμεσος είναι η 3^η τιμή. Άρα, M=7.

Αν έχουμε $X_i : 1,3,7,8$ τότε, Θέση M: $\frac{n+1}{2} = \frac{4+1}{2} = \frac{5}{2} = 2.5$.

Σε αυτήν την περίπτωση η θέση της διαμέσου είναι το ημίαθροισμα των δύο «μεσαίων» παρατηρήσεων, συνεπώς, θα έχουμε: $M = \frac{x_{(2)} + x_{(3)}}{2} = \frac{3+7}{2} = 5$.

Το 1^ο Τεταρτημόριο (first quartile) είναι η τιμή κάτω από την οποία βρίσκεται το 25% των τιμών της μεταβλητής. Η θέση του πρώτου τεταρτημορίου είναι $Q_1 : \frac{n+1}{4}$.

Το 3^ο Τεταρτημόριο είναι η τιμή κάτω από την οποία βρίσκεται το 75% των τιμών της μεταβλητής.

Η θέση του τρίτου τεταρτημορίου είναι $Q_3 : \frac{3(n+1)}{4}$. Γενικά, η θέση του δεκατημορίου είναι $\alpha\% (n+1)$.

Για παράδειγμα, $\alpha = 0.1, \frac{n+1}{10}, \alpha = 0.2, \frac{2(n+1)}{10}$.

Κορυφή ή Επικρατούσα τιμή (T_0 ή M_0) είναι η τιμή με τη μεγαλύτερη συχνότητα.

Ένα ερώτημα το οποίο συχνά απασχολεί τους ερευνητές όταν θέλουν να αναλύσουν μία σειρά δεδομένων είναι το ακόλουθο:

ΕΡΩΤΗΜΑ: Έχουμε μία σειρά δεδομένων και ζητείται να υπολογίσουμε ένα μέτρο κεντρικής τάσης ή θέσης. Ποιο θα επιλέξουμε από τα μέτρα που γνωρίζουμε ως κατάλληλο;

Η απάντηση σε αυτό το ερώτημα δεν είναι μοναδική αλλά εξαρτάται από την εξέταση διάφορων σημαντικών στοιχείων που πρέπει να ληφθούν υπόψη.

(α) Η ύπαρξη ακραίων τιμών (ακραίων υψηλών ή ακραίων χαμηλών) στο δείγμα τιμών της μεταβλητής. Η διάμεσος δεν επηρεάζεται από την ύπαρξη ακραίων τιμών ενώ ο αριθμητικός (δειγματικός) μέσος επηρεάζεται. Άρα, όταν η κατανομή έχει ακραίες τιμές πιο αντιπροσωπευτικό μέτρο μοιάζει να είναι η διάμεσος. Για παράδειγμα, έστω μία υπό-μελέτη μεταβλητή που λαμβάνει τις ακόλουθες τιμές: 17,13,26,12,92. Η διάμεσος είναι ίση με 17 ενώ ο μέσος είναι ίσος με 32 μία τιμή που σίγουρα είναι «επηρεασμένη» από την ύπαρξη της ακραίας υψηλής τιμής του δείγματος (92).

(β) Θα πρέπει να εξετάσουμε την ευαισθησία στο σχήμα της κατανομής. Όταν η κατανομή έχει τιμές που ισαπέχουν άνω και κάτω περί (γύρω) από την κεντρική τιμή έχουμε συμμετρική κατανομή. Όταν έχω ακραία υψηλές τιμές η κατανομή παρουσιάζει δεξιά ή θετική ασυμμετρία (ή δεξιά στρεβλότητα). Τότε ισχύει ότι: $T_0 < M < \bar{X}$. Όταν έχουμε ακραία χαμηλές τιμές τότε η κατανομή παρουσιάζει αριστερή ασυμμετρία ή αρνητική ασυμμετρία (ή αριστερή στρεβλότητα). Τότε ισχύει ότι: $\bar{X} < M < T_0$. Όταν έχουμε δεξιά ή αριστερή ασυμμετρία τότε προτιμότερο μέτρο μοιάζει να είναι η διάμεσος (από τον αριθμητικό μέσο). Όταν έχουμε περίπου (σχεδόν) συμμετρική κατανομή προτιμούμε το δειγματικό μέσο $\bar{X}$ από τη διάμεσο M. Για παράδειγμα, έστω $X_i : 0,1,0,3,6,7,3,0,1,4$. Να υπολογιστούν τα $\bar{X}$, M, T_0 .

$T_0 = 0$ (μονοκόρυφη κατανομή). Διατάσσουμε τις τιμές της μεταβλητής και έχουμε:

$X_i : 0,0,0,1,1,3,3,4,6,7$.

Η θέση της διαμέσου είναι $\frac{n+1}{2} = 5,5$. Άρα, $M = \frac{X_{(5)} + X_{(6)}}{2} = \frac{1+3}{2} = 2 \Rightarrow M = 2$.

$$\bar{X} = \frac{0+1+0+3+6+7+3+0+1+4}{10} = 2.5.$$

Προτιμότερο μέτρο εδώ φαίνεται να είναι ο αριθμητικός μέσος ($\bar{X} = 2.5$).

(γ) Όταν θέλουμε να εξάγουμε συμπερασματολογία για τον πληθυσμό από το οποίο λαμβάνουμε το δείγμα, υπολογίζοντας τις παραμέτρους του δείγματος, προσπαθώντας να βγάλουμε συμπεράσματα για τις αντίστοιχες παραμέτρους του πληθυσμού, τότε σχεδόν πάντα υπολογίζουμε τον αριθμητικό μέσο $\bar{X}$ αντί της διαμέσου M .

(δ) Αν μας ενδιαφέρει να αναπτύξουμε θεωρητικές ιδιότητες για τα παραπάνω μέτρα, τότε ο αριθμητικός μέσος $\bar{X}$ είναι ο πιο κατάλληλος στη θεωρητική ανάπτυξη της στατιστικής συμπερασματολογίας σε σχέση με το M , γιατί επηρεάζεται από όλες τις τιμές της μεταβλητής.

Στην περίπτωση που τα δεδομένα μας είναι ταξινομημένα σε κλάσεις (ή ομάδες), οι υπολογισμοί των διαφόρων μέτρων θέσης διαφοροποιούνται και προσαρμόζονται στη νέα μορφή (δομή) των δεδομένων.

Για τον αριθμητικό μέσο έχουμε: $\bar{X} = \frac{\sum_i f_i X_i^K}{\sum_i f_i}$, όπου, X_i^K είναι η κεντρική τιμή της τάξης (διαστήματος)

i. Η θέση της διαμέσου M καθορίζεται από την τάξη που έχει αθροιστική συχνότητα $\geq \frac{n}{2}$. Ισχύει ο

ακόλουθος τύπος: $M = L_M + \frac{\delta\left(\frac{n}{2} - F_{M-1}\right)}{f_M}$, όπου L_M είναι το κάτω άκρο της τάξης που περιέχει τη

διάμεσο, F_{M-1} είναι η αθροιστική συχνότητα της προηγούμενης τάξης από αυτήν που περιέχει τη διάμεσο και f_M είναι η συχνότητα της τάξης που περιέχει τη διάμεσο. Σε πλήρη αντιστοιχία με τη διάμεσο, έχουμε τους τύπους για το πρώτο και για το τρίτο τεταρτημόριο που δίνονται παρακάτω:

$$1^\circ \text{ Τεταρτημόριο: } Q_1 = L_{Q_1} + \frac{\delta\left(\frac{n}{4} - F_{Q_1-1}\right)}{f_{Q_1}} \text{ και } 3^\circ \text{ Τεταρτημόριο: } Q_3 = L_{Q_3} + \frac{\delta\left(\frac{3n}{4} - F_{Q_3-1}\right)}{f_{Q_3}}.$$

Η επικρατούσα τιμή (ή τύπος ή κορυφή ή σημείο μέγιστης συχνότητας (T_0 ή M_0)) δίνεται από τον ακόλουθο τύπο: $T_0 = M_0 = L_{T_0} + \delta \frac{\Delta_1}{\Delta_1 + \Delta_2}$, όπου η θέση της κορυφής είναι η τάξη με την μεγαλύτερη

συχνότητα και Δ_1 είναι η διαφορά συχνοτήτων της τάξης με τη μεγαλύτερη συχνότητα με την προηγούμενη τάξη και Δ_2 είναι η διαφορά συχνοτήτων της τάξης με τη μεγαλύτερη συχνότητα με την επόμενη τάξη.

Παράδειγμα: Δίνεται η κατανομή συχνοτήτων 65 εργαζομένων σε ένα εργοστάσιο (σε χιλιάδες δραχμές). Να βρεθούν $\bar{X}, M, Q_1, Q_3, T_0$.

Τάξεις (χιλ.δρχ)	f_i	X_i	συνεχείς	F_i	X_i^K	$f_i X_i^K$
		τάξεις				
30-39	8	29,5-39,5		8	34,5	276
40-49	10	39,5-49,5		18	44,5	445
50-59	16	49,5-59,5		34	54,5	.
60-69	14	59,5-69,5		48	64,5	.
70-79	10	69,5-79,5		58	74,5	.
80-89	5	79,5-89,5		63	84,5	.
90-99	2	89,5-99,5		65	84,5	.

Λύση:

$$\bar{X} = \frac{\sum f_i X_i^K}{\sum f_i} = \frac{3852,5}{65} = 59,27.$$

Αρα, ο μέσος μισθός στο εργοστάσιο των 65 εργαζομένων είναι 59,27 χιλιάδες δραχμές.

Η θέση της διαμέσου είναι η τάξη με αθροιστική συχνότητα $\geq \frac{n}{2}$. Αρα είναι η 3^η τάξη.

Θέση $M \geq \frac{n}{2}$ 3^η τάξη

$$M = 49,5 + \frac{10}{16} \left(\frac{65}{2} - 18 \right) = 58,57 \text{ χιλ. δρχ.}$$

$$Q_1 = L_{Q_1} + \frac{\delta}{f_{Q_1}} \left(\frac{n}{4} - F_{Q_{1-1}} \right) = 39,5 + \frac{10}{10} (16,25 - 8) = 47,75 \text{ χιλ.δρχ}$$

$$\text{Τάξη } Q_1 : \geq \frac{n}{4}$$

2^η

Θέση Q_1

$$\geq \frac{n}{4} = \frac{65}{4} = 16,25 \text{ (2^η τάξη)}$$

$$Q_3 = 69,5 + \frac{10}{10} (48,75 - 48)$$

Θέση Q_3

$$\geq \frac{3n}{4} = \frac{3 \cdot 65}{4} = 48,75 \text{ (5^η τάξη)}$$

$$T_0 = M_0 = L_{T_0} + \delta \frac{\Delta_1}{\Delta_1 + \Delta_2} = 49,5 + 10 \frac{16 - 10}{(16 - 10) + (16 - 14)} = 57 \text{ χιλ. δρχ.}$$

$$\bar{X} = 59,25 \text{ χιλ. δρχ.}$$

$$M = 58,57 \text{ χιλ. δρχ.}$$

$$T_0 = 57 \text{ χιλ.δρχ.}$$

$$\bar{X} > M > T_0$$

έχω δεξιά ή θετική ασυμμετρία

Για να καθορίσουμε τη μορφή και τη θέση μίας κατανομής συχνότητας χρειάζεται να εξετάσουμε:

- (α) Ένα μέτρο κεντρικής θέσης ή τάσης.
- (β) Ένα μέτρο διασποράς (ή μεταβλητότητας).
- (γ) Ένα μέτρο ασυμμετρίας.
- (δ) Ένα μέτρο κύρτωσης.

ΜΕΤΡΑ ΔΙΑΣΠΟΡΑΣ (ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ)

Εκφράζουν τις αποκλίσεις των τιμών μίας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης.

(α) Το εύρος των τιμών μίας μεταβλητής. Ισχύει ότι $R = X_{\max} - X_{\min}$. Όσο πιο μικρό είναι το εύρος, τόσο πιο ομοιόμορφα φαίνεται να κατανέμονται οι τιμές της υπό-μελέτη μεταβλητής γύρω από τον αριθμητικό μέσο.

(β) Η ενδοτεταρτημοριακή απόκλιση είναι η διαφορά πρώτου και τρίτου τεταρτημορίου $Q_3 - Q_1$.

(γ) Η ημί-ενδοτεταρτημοριακή απόκλιση είναι το ήμισυ της ενδοτεταρτημοριακής απόκλισης δηλαδή $\frac{Q_3 - Q_1}{2}$.

(δ) Η μέση απόλυτη απόκλιση δίνεται από τους ακόλουθους τύπους ανάλογα με το εάν έχουμε αταξινόμητα ή ταξινομημένα δεδομένα, αντίστοιχα.

$$MA = \frac{\sum_i |X_i - \bar{X}|}{n} \quad \text{ή} \quad MA = \frac{\sum_i f_i |X_i^K - \bar{X}|}{\sum_i f_i}.$$

Η μέση απόλυτη απόκλιση υπολογίζεται απλά αλλά δεν είναι μέτρο που μπορεί να χρησιμοποιηθεί για θεωρητική συμπερασματολογία. Μετριέται στις ίδιες μονάδες μέτρησης της μεταβλητής.

(ε) Η δειγματική διακύμανση (ή δειγματική διασπορά) S^2 δίνεται από τους ακόλουθους τύπους ανάλογα με το εάν έχουμε αταξινόμητα ή ταξινομημένα δεδομένα, αντίστοιχα.

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} \quad \text{ή} \quad S^2 = \frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n-1}$$
$$\text{ή} \quad S^2 = \frac{1}{n-1} \sum f_i (X_i^K - \bar{X})^2 = \frac{1}{n-1} \left[\sum f_i X_i^{K^2} - \frac{(\sum f_i X_i^K)^2}{n} \right]$$

Η δειγματική διακύμανση S^2 είναι το πιο διαδεδομένο μέτρο διασποράς. Όταν οι τιμές της X_i δεν διαφέρουν πολύ από τη μέση τιμή τότε η διασπορά είναι μικρή και αντίστροφα.

(ζ) Η δειγματική τυπική απόκλιση είναι η θετική τετραγωνική ρίζα της δειγματικής διακύμανσης $S = \sqrt{S^2}$. Για παράδειγμα αν $S = 3.2$, τότε μπορούμε να πούμε ότι οι τιμές της μεταβλητής απέχουν περίπου ± 3.2 μονάδες από το δειγματικό μέσο $\bar{X}$.

Όταν η κατανομή είναι (σχεδόν) συμμετρική, η ζώνη $\bar{X} \pm S$ καλύπτει το 68% περίπου των τιμών της X_i , η ζώνη $\bar{X} \pm 2S$ καλύπτει περίπου το 95% των τιμών της X_i και η ζώνη $\bar{X} \pm 3S$ καλύπτει το 99% περίπου των τιμών της X_i .

Παράδειγμα

Δίνεται η παρακάτω κατανομή. Αν κάθε τιμή της μεταβλητής X_i πολλαπλασιαστεί με μια σταθερά $a > 0$ ή προστεθεί μία σταθερά a ποιες θα είναι οι νέες τιμές των $\bar{X}$ και S^2 ;

Κλάσεις	f_i	X_i^k
10-12	10	11
12-14	20	13
14-16	30	15
16-18	80	17= X_0
18-20	30	19
20-22	20	21
22-24	10	23
$\sum f_i = n = 200$		

Λύση:

(α) Αν $X_i^k = aX_i$

$$S^{*2} = \frac{\sum f_i (X_i^* - \bar{X}^*)^2}{n} = \frac{\sum f_i (aX_i - a\bar{X})^2}{n} = a^2 \frac{\sum f_i (X_i - \bar{X})^2}{n} = a^2 S^2$$

$$S^* = |a|S \text{ και } \bar{X}^* = a\bar{X}$$

Αν κάθε τιμή της X_i πολλαπλασιαστεί με μία σταθερά a η διακύμανση πολλαπλασιάζεται με a^2 η τυπική απόκλιση με $|a|$ και ο αριθμητικός μέσος με a .

(β) $X_i = a + X_i$

$$\bar{X}^* = \frac{\sum f_i X_i^*}{n} = \frac{\sum f_i (X_i + a)}{n} = \frac{\sum f_i X_i}{n} + a \frac{\sum f_i}{n} = \bar{X} + a$$

$$S^{*2} = \frac{\sum f_i (X_i^* - \bar{X}^*)^2}{n} = \frac{\sum f_i (X_i + a - \bar{X} - a)^2}{n} = S^2$$

Αν προσθέσουμε μία σταθερά a σε κάθε τιμή της μεταβλητής, ο δειγματικός μέσος $\bar{X}$ αυξάνεται κατά a , ενώ η δειγματική διακύμανση παραμένει αμετάβλητη

ΘΗΚΟΓΡΑΜΜΑ (Box – Plot)

Το θηκόγραμμα είναι ένα σχήμα στο οποίο συμπεριλαμβάνουμε βασικά περιγραφικά μετρά μίας υπό-μελέτη μεταβλητής. Το κυρίως «σώμα» του αποτελείται από ένα τετράπλευρο όπου στις κάθετες πλευρές τοποθετούμε το πρώτο τεταρτημόριο Q_1 , το τρίτο τεταρτημόριο Q_3 καθώς επίσης και τη διάμεσο M . Σημειώνουμε επίσης δύο οριακές τιμές, την κάτω οριακή και την άνω οριακή που δίνονται από τις ακόλουθες σχέσεις:

$$X_A = Q_1 - 1.5(Q_3 - Q_1) \text{ κάτω οριακή}$$

$$X_B = Q_3 + 1.5(Q_3 - Q_1) \text{ άνω οριακή}$$

Αν υπάρχουν έκτροπες ή ακραίες τιμές για την υπό-μελέτη μεταβλητή X_i αυτές συμβολίζονται με * ή •. Το θηκόγραμμα στο κυρίως σώμα του μας δίνει το κεντρικό διάστημα των παρατηρήσεων μίας μεταβλητής δηλαδή εκεί που βρίσκεται το 50% των παρατηρήσεων. Οι γραμμές στα δύο άκρα του θηκογράμματος και η θέση της διαμέσου μας δίνουν πληροφορίες σχετικές με τη συμμετρία της κατανομής, δηλαδή αν έχουμε θετική ασυμμετρία ή αρνητική ασυμμετρία ή αν μπορούμε να πούμε ότι έχουμε μία (σχεδόν) συμμετρική κατανομή. Μία χαρακτηριστική χρησιμότητα του θηκογράμματος είναι η δυνατότητα που μας παρέχει να μπορούμε να συγκρίνουμε δύο ή περισσότερους πληθυσμούς (ή δείγματα).

Παράδειγμα: Έστω ότι οι τιμές μίας υπό-μελέτη μεταβλητής δίνουν τα ακόλουθα στοιχεία:

$$Q_1 = 9.5$$

$$Q_3 = 12$$

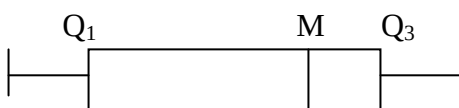
$$M = 11$$

$$X_{\text{ακραιο}} = 18,3$$

$$X_A = Q_1 - 1.5(Q_3 - Q_1) = 9.5 - 1.5(12 - 9.5) = 5.75$$

$$X_B = Q_3 + 1.5(Q_3 - Q_1) = 12 + 1.5(12 - 9.5) = 15.75$$

Με βάση αυτά τα στοιχεία υπολογίζουμε την κάτω οριακή και την άνω οριακή τιμή της μεταβλητής. Αν $M - Q_1 \approx Q_3 - M$ ή $Q_1 - X_{\min} = X_{\max} - Q_3$, η διάμεσος θα βρίσκεται στο μέσο ή περίπου στο μέσο του κυρίως σώματος του θηκογράμματος και η κατανομή των τιμών της μεταβλητής μπορεί να θεωρηθεί σχεδόν συμμετρική. Αν $X_{\max} - Q_3 > Q_1 - X_{\min}$ ή $Q_3 - M > M - Q_1$ τότε η μέση έκταση είναι μεγαλύτερη της διαμέσου M άρα υπάρχουν έκτροπες μεγάλες τιμές. Σ' αυτήν την περίπτωση η κατανομή παρουσιάζει δεξιά ή θετική ασυμμετρία. Αν $X_{\max} - Q_3 < Q_1 - X_{\min}$ ή $Q_3 - M < M - Q_1$ τότε η μέση έκταση είναι μικρότερη της διαμέσου M και η κατανομή παρουσιάζει ακραίες χαμηλές τιμές, όπως φαίνεται στο παρακάτω σχήμα.



Άσκηση: Δίνονται οι παρακάτω τιμές: 7,18,12,17,29,18,4,27,30,2,4,10,21,5,8

Να κατασκευαστεί το Θηκόγραμμα. Τι παρατηρείτε;

ΜΕΤΡΑ ΣΧΕΤΙΚΗΣ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ

Το πιο χαρακτηριστικό μέτρο σχετικής μεταβλητότητας είναι ο συντελεστής μεταβλητότητας που ορίζεται ως εξής: $CV = \frac{S}{\bar{X}}$. Ο συντελεστής μεταβλητότητας εξετάζει τη μεταβλητότητα μίας κατανομής σε σχετικούς όρους και είναι μέτρο ανεξάρτητο των μονάδων μέτρησης της μεταβλητής.

$$\text{π.χ. } \bar{X} = 26$$

$$S = 16.74$$

$$CV = \frac{S}{\bar{X}} = \frac{16.74}{26} = 0.64 \text{ ή } 64\%.$$

Με τον συντελεστή μεταβλητότητας μπορούμε να συγκρίνουμε διάφορες κατανομές μεταξύ τους στις οποίες οι υπό-μελέτη μεταβλητές έχουν διαφορετικές μονάδες μέτρησης και να συγκρίνουμε τη σχετική μεταβλητότητά τους.

ΜΕΤΡΑ ΑΣΥΜΜΕΤΡΙΑΣ

Το πιο γνωστό μέτρο ασυμμετρίας είναι ο συντελεστής ασυμμετρίας κατά Pearson, S_p . Ο συντελεστής μπορεί να εκφραστεί είτε συναρτήσει της διαμέσου M είτε συναρτήσει της επικρατούσας τιμής M_0 . Οι

σχετικοί τύποι είναι: $S_p(M_0) = \frac{\bar{X} - M_0}{S}$, $S_p(M) = \frac{3(\bar{X} - M)}{S}$. Αν $-0.3 < S_p < 0.3$, τότε η κατανομή

μπορεί να θεωρηθεί σχεδόν συμμετρική. Αν $S_p \approx 0$ ή ισοδύναμα, $\bar{X} - M_0 \approx 3(\bar{X} - M)$, τότε η κατανομή θεωρείται συμμετρική. Ένας άλλος συντελεστής ασυμμετρίας που επίσης χρησιμοποιείται συχνά και υπολογίζεται συναρτήσει της διαμέσου, του πρώτου και του τρίτου τεταρτημορίου είναι ο συντελεστής ασυμμετρίας του Bowley, S_b . Ο συντελεστής ασυμμετρίας του Bowley δίνεται από τον ακόλουθο τύπο:

$S_b = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$. Αν $S_p, S_b > 0$, τότε η κατανομή παρουσιάζει θετική ασυμμετρία, ενώ αν $S_p, S_b < 0$,

τότε η κατανομή παρουσιάζει αρνητική ασυμμετρία. Εναλλακτικά, ανάλογα με το εάν έχουμε αταξινόμητα ή ταξινομημένα δεδομένα, μπορούμε να υπολογίζουμε τον συντελεστή ασυμμετρίας β_1 χρησιμοποιώντας τους ακόλουθους τύπους

$$\beta_1 = \frac{[n^{-1} \sum (X_i - \bar{X})^3]^2}{s^6}, \text{ ή } \beta_1 = \frac{[n^{-1} \sum f_i (X_i - \bar{X})^3]^2}{s^6}.$$

όπου, το πρόσημο του αριθμητή (προφανώς χωρίς να υψωθεί στο τετράγωνο!!) χαρακτηρίζει την ασυμμετρία της κατανομής. Αν ο αριθμητής είναι θετικός, τότε έχουμε θετική ασυμμετρία, διαφορετικά έχουμε αρνητική ασυμμετρία. Αν $\beta_1 \approx 0$, τότε η κατανομή μπορεί να θεωρηθεί σχεδόν συμμετρική.

ΜΕΤΡΟ ΚΥΡΤΩΣΗΣ

Το πιο διαδεδομένο μέτρο κύρτωσης είναι ο συντελεστής κύρτωσης κατά Pearson β_2 ο οποίος, ανάλογα με το εάν έχουμε αταξινόμητα ή ταξινομημένα δεδομένα, μπορεί να υπολογιστεί χρησιμοποιώντας τους ακόλουθους τύπους:

$$\beta_2 = \frac{n^{-1} \sum (X_i - \bar{X})^4}{s^4}, \text{ ή } \beta_2 = \frac{n^{-1} \sum f_i (X_i - \bar{X})^4}{s^4}.$$

Αν $\beta_2 > 3$, τότε η κατανομή χαρακτηρίζεται λεπτόκυρτη. Αν $\beta_2 < 3$, τότε η κατανομή χαρακτηρίζεται πλατύκυρτη και εάν $\beta_2 \approx 3$, τότε η κατανομή χαρακτηρίζεται μεσόκυρτη (για παράδειγμα μεσόκυρτη είναι η κανονική κατανομή).

ΑΣΚΗΣΕΙΣ

ΑΣΚΗΣΗ 1. Δίνονται τα περσινά κέρδη και οι ζημίες 10 επιχειρήσεων σε εκατομμύρια ευρώ:
3, 4, 6, -2, -3, 1, 9, -6, 4, 8.

Να κατασκευαστεί η κατανομή συχνοτήτων με ίσα υποδιαστήματα εύρους $\delta=4$ και αρχή του πρώτου υποδιαστήματος το -6 και να υπολογιστεί η διακύμανση και η διάμεσος των κερδών.

Λύση.

n=10	$\delta=4$	# τάξεων =4	$r \rightarrow$ κέντρο τάξεων		
Τάξεις		Συχνότητα f_i	X_i	X_i^2	F_i :αθροιστική συχνότητα
[-6,-2)		2	-4	16	2
[-2,2)		2	0	0	4
[2,6)		3	4	16	7
[6,10)		3	8	64	10
				96	

$$\sum_{r=1}^4 f_i = 10$$

$$\bar{X} = \frac{\sum_{r=1}^4 f_r X_r}{n} = \frac{-4 \cdot 2 + 0 \cdot 2 + 4 \cdot 3 + 8 \cdot 3}{10} = 2.8.$$

$$\sigma^2 = \frac{\sum_{r=1}^4 f_r X_r^2}{n} - \bar{X}^2 = \frac{96}{10} - 7.84 = 1.76.$$

$$M = L_M + \frac{\delta \left(\frac{n}{2} - F_{M-1} \right)}{f_M} = 2 + 4 \frac{5-4}{3} = 3.33.$$

$$\theta \acute{\epsilon} \sigma \eta \text{ } M: \geq \frac{n}{2} = 5 \quad (\acute{\alpha} \rho \alpha \text{ } \eta \text{ } \tau \acute{\alpha} \xi \eta \text{ } \text{που περιέχει τη διάμεσο είναι η τρίτη}).$$

ΑΣΚΗΣΗ 2. Σε κατανομή λογαριασμών της ΔΕΗ ενός πληθυσμού 100 νοικοκυριών σε ευρώ υπολογίσθηκαν :

$\sum f_i X_i^K = 6.700$, $\sum f_i X_i^{2K} = 457000$ και η διάμεσος $M=67$. Αν διαπιστωθεί ότι πληρούνται οι κατάλληλες συνθήκες, να βρεθεί το ποσοστό των λογαριασμών που είναι κάτω από 58 ευρώ.

Λύση.

$$\text{'Εχουμε } \bar{X} = \frac{\sum f_i X_i^K}{n} = \frac{6700}{100} = 67 = M. \quad n=100.$$

Επειδή $\bar{X} = M$ έχουμε συμμετρική κατανομή. Σε μια συμμετρική κατανομή (π.χ. κανονική κατανομή), ισχύει ότι:

(α) το 68.27% των τιμών του χαρακτηριστικού περιλαμβάνεται στο διάστημα μεταξύ $\bar{X} - \sigma$, $\bar{X} + \sigma$

(β) το 95.45% των τιμών του χαρακτηριστικού περιλαμβάνονται στο διάστημα μεταξύ $\bar{X} - 2\sigma$, $\bar{X} + 2\sigma$ και

(γ) το 99.73% των τιμών του χαρακτηριστικού περιλαμβάνονται στο διάστημα μεταξύ $\bar{X} - 3\sigma$, $\bar{X} + 3\sigma$.

$$\sigma^2 = \frac{\sum f_i X_i^{2K}}{n} - \bar{X}^2 = \frac{457000}{100} - 67^2 = 4570 - 4489 = 81 \Rightarrow \sigma = \sqrt{81} = 9$$

$$\bar{X} \pm \sigma = 67 \pm 9. \quad \text{Άρα, το } \frac{100 - 68.27}{2} = 15,86 \% \text{ των λογαριασμών είναι κάτω από 58€}.$$

ΑΣΚΗΣΗ 3. Δίνονται τα κέρδη και οι ζημιές 9 επιχειρήσεων για το έτος 2000 σε εκατομμύρια ευρώ: 4, 8, -5, 2, -2, 50, 3, 6, 5. Να βρεθούν ο αριθμητικός μέσος και η διάμεσος.

Λύση.

$$\bar{X} = \frac{4+8-5+2-2+50+3+6+5}{9} = \frac{71}{9} = 7,888$$

Διατάσσουμε τις τιμές: -5, -2, 2, 3, 4, 5, 6, 8, 50

$$\text{Θέση διαμέσου: } \frac{n+1}{2} = \frac{10}{2} = 5 \text{ (5}^{\text{η}} \text{ παρατήρηση).}$$

Άρα η διάμεσος είναι $M=4$.

ΑΣΚΗΣΗ 4. Σε συμμετρική κατανομή δαπανών βρέθηκε διάμεσος $M=160$ χιλιάδες ευρώ και συντελεστής μεταβλητότητας $CV=12.5\%$. Ποιο ποσοστό των δαπανών είναι πάνω από 180.000 ευρώ;

$$\text{Λύση. Είναι } CV = \frac{\sigma}{\bar{X}} 100\% \Rightarrow 12.5 = \frac{\sigma}{160} 100 \Rightarrow \underline{\sigma=20}.$$

Η κατανομή είναι συμμετρική άρα $\bar{X} = M$. $\bar{X} + \sigma = 160 + 20 = 180$, $P(\bar{X} - \sigma < X < \bar{X} + \sigma) = 68.27\%$.

Το ποσοστό των τιμών που είναι πάνω από 180 είναι $50\% - \frac{68.27\%}{2} = 15,865\%$ (περίπου 16%).

ΑΣΚΗΣΗ 5. Συμμετρική κατανομή μισθών έχει μέσο $\bar{X} = 1000$ ευρώ.

(α) Να βρεθεί η τυπική απόκλιση σ , αν γνωρίζουμε ότι κάτω από 900 ευρώ και πάνω από 1100 ευρώ παίρνει συνολικά το 31.73% των εργαζομένων.

(β) Αν όλοι αυτοί οι μισθοί αυξηθούν κατά 15% και δοθεί ένα επίδομα 50 ευρώ, να βρεθεί ο συντελεστής μεταβλητότητας CV των νέων μισθών.

Λύση.

$$(\alpha) \frac{31.73\%}{2} = 15.86\% \quad \chi: \text{μισθοί εργαζομένων}$$

$$31.73\% + 68.27\% = 100\% \quad \bar{X} = 1000$$

$$\left. \begin{array}{l} P\left(X > \underset{\substack{|| \\ 1100}}{\bar{X} + 100} \text{ και } X < \underset{\substack{|| \\ 900}}{\bar{X} - 100}\right) = 31.73\% \Rightarrow P\left(\underset{\substack{|| \\ \bar{X} - \sigma}}{900} < X < \underset{\substack{|| \\ \bar{X} + \sigma}}{1100}\right) = 68.27\% \end{array} \right\} \Rightarrow \underline{\sigma=100}$$

$$\begin{aligned} (\beta) \quad \bar{Y} &= \frac{\sum Y_i}{n} = \bar{X}' & \bar{X}' &= \bar{Y} = \bar{X} + 0.15\bar{X} + 50 \\ & & &= 1.15\bar{X} + 50 \\ & & &= 1.15 \cdot 1000 + 50 \\ & & &= 1200 \end{aligned}$$

$$\sigma_Y^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n} = \frac{\sum_i (1.15x_i + 5\theta - 1.15\bar{x} - 5\theta)^2}{n} =$$

$$= \frac{\sum_i 1.15^2 (X_i - \bar{X})^2}{n} = \frac{1.15^2 \sum_i (X_i - \bar{X})^2}{n} =$$

$$= 1.15^2 \sigma_X^2 \Rightarrow \underline{\sigma_Y = 1.15 \sigma_X = 1.15 \cdot 100 = 115}$$

$$CV = \frac{\sigma_Y}{\bar{Y}} 100 = \frac{115}{1200} 100 = 9.58 \Rightarrow \underline{CV = 9.58}$$

ΑΣΚΗΣΗ 7. Δίνεται η κατανομή των επιχειρήσεων ενός βιομηχανικού κλάδου ανάλογα με την αξία των επενδύσεων που πραγματοποιήθηκαν.

Αξία επενδύσεων	f_i	$X_i^K \rightarrow$ κέντρο τάξης	$f_i X_i^K$	$F_i:$ Αθροιστική συχνότητα	$f_i X_i^{K^2}$
10-<20	2	15	30	2	450
20-30	11	25	275	13	6875
30-40	30	35	1050	43	36750
40-50	9	45	405	52	18225
50-60	3	55	165	55	9075
	55		1925		71375

(α) Να υπολογίσετε τον αριθμητικό μέσο, τη διάμεσο, τα τεταρτημόρια, την τεταρτημοριακή απόκλιση και τον συντελεστή μεταβλητότητας.

(β) Υπολογίστε τον συντελεστή ασυμμετρίας του Pearson. Καθορίστε το είδος και την ένταση της ασυμμετρίας.

Λύση.

$$(α) \bar{X} = \frac{\sum_i f_i X_i^K}{n} = \frac{1925}{55} = 35 \quad n=55$$

$$\text{Θέση διαμέσου} \geq \frac{n}{2} : 27.5 \quad M = L_M + \delta \left(\frac{\frac{n}{2} - F_{M-1}}{f_n} \right) = 30 + 10 \frac{27.5 - 13}{30} = 34.83$$

$$Q_1 : \frac{n}{4} = \frac{55}{4} = 13.75 \quad Q_1 = L_{Q_1} + \delta \left(\frac{\frac{n}{4} - F_{Q_1-1}}{f_{Q_1}} \right) = 30 + 10 \frac{13.75 - 13}{30} = 30.25$$

$$Q_3 : \frac{3n}{4} = 41,25$$

$$Q_3 = L_{Q_3} + \delta \left(\frac{3n/4 - F_{Q_3-1}}{f_{Q_3}} \right)$$

$$Q = \frac{Q_3 - Q_1}{2} = \frac{39,42 - 30,2}{2} = 4,585$$

$$= 30 + 10 \left(\frac{41,25 - 13}{30} \right) = 39,42$$

$$\sigma^2 = \frac{\sum_i f_i X_i^{K^2}}{n} - \bar{X}^2 = \frac{71375}{55} - 35^2 = 72,73$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{72,73} \cong 8.53$$

$$CV = \frac{\sigma}{\bar{X}} \cdot 100 = \frac{8,53}{35} \cdot 100 = 24,37\%$$

(β)

$$\bar{X} > M \Rightarrow 35 > 34,83 \text{ (θετικά ασύμμετρη μικρής έντασης)}$$

$$S_p = \frac{3(\bar{X} - M)}{\sigma} = \frac{3(35 - 34,83)}{8.53} = 0.0598 > 0 .$$

Η κατανομή των τιμών παρουσιάζει μικρής έντασης θετική ασυμμετρία (οι τιμές συγκεντρώνονται αριστερά).

Άσκηση 7. Δίνεται η κατανομή 40 διαφορετικών μοντέλων αυτοκινήτων (σε χιλιάδες ευρώ). Να υπολογιστούν (α) η μέση τιμή και η τυπική απόκλιση των τιμών.
(β) η διάμεσος και η τεταρτημοριακή απόκλιση.

Λύση. (α)

ΤΙΜΗ χιλ. €	# ΜΟΝΤΕΛΩΝ f_i	X_i^K	$f_i X_i^K$	$f_i (X_i - \bar{X})^2$	<u>n=40</u>
10-15	5	12.5	62.5	23.63	
15-20	20	17.5	350	70,4	
20-25	10	22.5	225	97,6	
25-30	5	27.5	<u>137,5</u>	<u>330,05</u>	
			775	521,68	

$$\bar{X} = \frac{\sum_i f_i X_i^K}{n} = \frac{775}{40} = 19,375$$

$$\sigma^2 = \frac{\sum_i f_i (X_i - \bar{X})^2}{n} = \frac{521,68}{40} = 13,042$$

$$\sigma = \sqrt{\sigma^2} = 3,61$$

(β)

θέση διαμέσου $M \geq \frac{n}{2} = 20$ (τάξη διαμέσου 2^η)

$$M = L_M + \delta \frac{\left(\frac{n}{2} - F_{M-1}\right)}{f_M} = 15 + 5 \cdot \frac{(20-5)}{20} = 15 + 5 \cdot \frac{15}{20} = 18,75$$

$$Q_1 \geq \frac{n}{4} = 10 \quad Q_1 = L_{Q_1} + \delta \left(\frac{\frac{n}{4} - F_{Q_1-1}}{f_{Q_1}} \right) = 15 + 5 \cdot \frac{(10-5)}{20} = 15 + \frac{25}{20} = 16,25$$

2^η τάξη

$$Q_3 \geq \frac{3n}{4} \Rightarrow Q_3 = L_{Q_3} + \delta \left(\frac{\frac{3n}{4} - F_{Q_3-1}}{f_{Q_3}} \right) = 20 + 5 \cdot \frac{(30-25)}{10} = 22,5$$

3^η τάξη

$$Q_3 - Q_1 = 6,25.$$

Άσκηση 8. Δίνεται ο πίνακας κατανομής ετησίων μισθών 40 τραπεζικών υπαλλήλων (σε χιλιάδες ευρώ). Να υπολογιστεί ο συντελεστής ασυμμετρίας κατά Pearson και η τεταρτημοριακή απόκλιση. Μπορούμε να χαρακτηρίσουμε την κατανομή αυτή συμμετρική; Δικαιολογήστε την απάντησή σας.

Μισθός	# Υπαλλήλων f_i	X_i^K	F_i	$f_i X_i^K$	$f_i (X_i - \bar{X})^2$
10- <20	6	15	6	90	2460,37
20-30	8	25	14	200	840,5
30-40	12	35	26	420	0,75
40-50	7	45	33	315	665,44
50-60	7	55	40	385	2730,44
				1410	6697,5

Λύση.

$$\bar{X} = \frac{\sum_i f_i X_i^K}{n} = \frac{1410}{40} = 35,25$$

$$\sigma^2 = \frac{6697,5}{40} = 167,44 \Rightarrow \sigma = \sqrt{\sigma^2} = 12,94$$

$$\begin{aligned} M &:= \frac{n}{2} = 20 & M &= L_M + \delta \left(\frac{\frac{n}{2} - F_{M-1}}{f_M} \right) \\ & & &= 30 + 10 \frac{(20 - 14)}{12} \\ & & &= 30 + 10 \frac{6}{12} = 35 \end{aligned}$$

$$S_p = \frac{3(\bar{X} - M)}{\sigma} = \frac{3(35,25 - 35)}{12,94} = 0,057 > 0 \quad (\text{ελαφρώς θετικά ασύμμετρη κατανομή}).$$

Μπορεί να χαρακτηριστεί συμμετρική γιατί $-0,3 < S_p < 0,3$.

Άσκηση. Παρακάτω δίνεται ένα δείγμα από τα αποτελέσματα μιας εξέτασης στην οποία έχουν λάβει μέρος 1.5 εκατ. φοιτητές. Η μέγιστη βαθμολογία στην εξέταση αυτή είναι 2.400 μονάδες.

(α) Να κατασκευάσετε τον πίνακα συχνοτήτων και το αντίστοιχο ιστόγραμμα. Ξεκινήστε την πρώτη κλάση από το 800 και χρησιμοποιήστε εύρος κλάσης 200 μονάδες.

(β) Σχολιάστε το σχήμα της κατανομής.

(γ) Τι παρατηρήσεις μπορείτε αν κάνετε χρησιμοποιώντας τις πληροφορίες τόσο του πίνακα συχνοτήτων όσο και του γραφήματος;

Τα δεδομένα είναι:

1996	1525	1355	1645	1780	1275
2135	1280	1060	1585	1650	1560
1150	1485	1990	1590	1880	1420
1755	1375	1475	1680	1440	1260
1730	1490	1560	940	1390	1175

B. ΜΕΡΟΣ

ΔΙΑΝΥΣΜΑΤΙΚΕΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ

ΕΙΣΑΓΩΓΗ

Σε αυτό το μέρος θα κάνουμε μία εισαγωγή στη θεωρία των διανυσματικών τυχαίων μεταβλητών. Θα μελετήσουμε τις διακριτές και τις συνεχείς διανυσματικές τυχαίες μεταβλητές και θα ορίσουμε σε κάθε περίπτωση τις περιθώριες συναρτήσεις κατανομών τους. Επιπλέον θα μελετήσουμε την ανεξαρτησία δύο τυχαίων μεταβλητών και την κατανομή του αθροίσματος ανεξάρτητων τυχαίων μεταβλητών. Οι έννοιες της συνδιακύμανσης και του συντελεστή συσχέτισης δύο τυχαίων μεταβλητών παρουσιάζονται με τη χρήση κατάλληλων σχετικών παραδειγμάτων. Χρησιμοποιώντας μία κατάλληλη σχέση παρουσιάζονται σχετικά παραδείγματα υπολογισμού κατανομών μετασχηματισμένων τυχαίων μεταβλητών. Ορίζονται σημαντικές έννοιες από τη θεωρία των δεσμευμένων κατανομών όπως, μεταξύ άλλων, οι ακόλουθες: δεσμευμένη αθροιστική συνάρτηση κατανομής, δεσμευμένη συνάρτηση πιθανότητας, δεσμευμένη συνάρτηση πυκνότητας και δεσμευμένη μέση τιμή.

ΔΙΑΝΥΣΜΑΤΙΚΕΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ

Έστω οι τυχαίες μεταβλητές $X_1, \dots, X_n$ που ορίζονται στον ίδιο πιθανοθεωρητικό χώρο $(\Omega, \mathfrak{F}, P)$. Για κάθε $i = 1, \dots, n$ οι τυχαίες μεταβλητές X_i είναι τέτοιες ώστε $X_i : \Omega \rightarrow \mathfrak{R}$. Πολλές φορές μας ενδιαφέρει να μελετήσουμε τις τυχαίες μεταβλητές $X_1, \dots, X_n$ σε συνδυασμό για να διερευνήσουμε τυχόν συσχετίσεις τους. Με άλλα λόγια, μας ενδιαφέρει το διάνυσμα των τυχαίων μεταβλητών $\tilde{X} = (X_1, \dots, X_n)$. Για παράδειγμα, έστω ότι η τυχαία μεταβλητή X_1 αναπαριστά το βάρος, η τυχαία μεταβλητή X_2 αναπαριστά το ύψος, η τυχαία μεταβλητή X_3 αναπαριστά την ηλικία και η τυχαία μεταβλητή X_4 αναπαριστά την πίεση ενός ασθενούς. Μας ενδιαφέρει να διερευνήσουμε τις τυχόν συσχετίσεις αυτών των μεταβλητών. Μελετούμε λοιπόν το διάνυσμα $\tilde{X} = (X_1, X_2, X_3, X_4)$. Δίνουμε τον ακόλουθο ορισμό.

Ορισμός. Ένα διάνυσμα $\tilde{X} = (X_1, \dots, X_n)$ που αποτελείται από n τυχαίες μεταβλητές ορισμένες στον ίδιο πιθανοθεωρητικό χώρο $(\Omega, \mathfrak{F}, P)$ καλείται **n -διάστατη τυχαία μεταβλητή ή διανυσματική τυχαία μεταβλητή (δ.τ.μ.)** (n -dimensional random variable or random vector). Για το διάνυσμα $\tilde{X}$ ισχύει ότι $\tilde{X} : \omega \in \Omega \rightarrow \tilde{X}(\omega) = (X_1(\omega), \dots, X_n(\omega)) \in \mathfrak{R}^n$.

Κατά τη ρίψη δύο ζαριών μας ενδιαφέρει συνήθως τόσο η ένδειξη του πρώτου όσο και του δεύτερου ζαριού. Σε μία κλινική μελέτη μας ενδιαφέρει ο τρόπος με τον οποίον η ηλικία ενός ασθενούς επηρεάζει το χρόνο αντίδρασής του σε ένα συγκεκριμένο φάρμακο. Σε αυτές τις περιπτώσεις εμφανίζονται δύο

τυχαίες μεταβλητές. Ιδιαίτερο ενδιαφέρον παρουσιάζει ο προσδιορισμός της συμπεριφοράς της μιας τυχαίας μεταβλητής σε σχέση με τη συμπεριφορά της άλλης. Στη συνέχεια του παρόντος κεφαλαίου, θα αναπτύξουμε τη θεωρία των διανυσματικών τυχαίων μεταβλητών στην περίπτωση κατά την οποία $n = 2$, δηλαδή όταν $\tilde{X} = (X, Y)$.

Ορισμός. Η **συνάρτηση κατανομής** $F : \mathcal{R}^2 \rightarrow [0, 1]$ της διανυσματικής διδιάστατης τυχαίας μεταβλητής (X, Y) ορίζεται ως εξής: $F(x, y) = P[X \leq x, Y \leq y]$, $-\infty < x, y < \infty$ και καλείται **από κοινού (αθροιστική) συνάρτηση κατανομής** (joint distribution function) της διδιάστατης διανυσματικής τυχαίας μεταβλητής (X, Y) .

Η συνάρτηση κατανομής F_X της τυχαίας μεταβλητής X μπορεί να ληφθεί από τη συνάρτηση F της διδιάστατης δ.τ.μ. (X, Y) ως εξής:

$$F_X(x) = P[X \leq x] = P[X \leq x, Y < \infty] = P[\lim_{y \rightarrow \infty} \{X \leq x, Y \leq y\}] = \lim_{y \rightarrow \infty} P[X \leq x, Y \leq y] = \lim_{y \rightarrow \infty} F(x, y).$$

Στην τέταρτη ισότητα η εναλλαγή της θέσης του ορίου και της πιθανότητας είναι επιτρεπτή. Εντελώς ανάλογα προκύπτει ότι $F_Y(y) = \lim_{x \rightarrow \infty} F(x, y)$.

Οι συναρτήσεις F_X και F_Y καλούνται **περιθώριες συναρτήσεις κατανομής** (marginal distribution functions) των τυχαίων μεταβλητών X και Y .

Παράδειγμα 1. Θα υπολογίσουμε την πιθανότητα $P[X > x, Y > y]$ συναρτήσει των περιθωρίων συναρτήσεων κατανομών των τυχαίων μεταβλητών X , Y και της από κοινού αθροιστικής συνάρτησης κατανομής της διδιάστατης δ.τ.μ. (X, Y) . Διαδοχικά έχουμε

$$P[X > x, Y > y] = 1 - P[\{X > x, Y > y\}^c] = 1 - P[\{X > x\}^c \cup \{Y > y\}^c] = 1 - P[\{X \leq x\} \cup \{Y \leq y\}]$$

$= 1 - [P\{X \leq x\} + P\{Y \leq y\} - P\{X \leq x, Y \leq y\}] = 1 - F_X(x) - F_Y(y) + F(x, y)$. Η δεύτερη ισότητα είναι συνέπεια του Νόμου De Morgan για την τομή δύο ενδεχομένων και η τέταρτη ισότητα είναι συνέπεια του προσθετικού νόμου.

ΔΙΑΚΡΙΤΕΣ ΔΙΑΝΥΣΜΑΤΙΚΕΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ

Αν οι τυχαίες μεταβλητές X και Y είναι διακριτές τότε ορίζουμε την **από κοινού συνάρτηση πιθανότητας (σ.π.)** (joint probability function) $p(x, y)$ της διδιάστατης δ.τ.μ. (X, Y) , όπου $x \in S_X$ και $y \in S_Y$, ως εξής: $p(x, y) = P[X = x, Y = y]$.

Απαραίτητη συνθήκη για να είναι η p μία από κοινού σ.π. για τη δ.τ.μ. (X, Y) είναι η εξής:

$$\sum_{x \in S_X} \sum_{y \in S_Y} p(x, y) = 1.$$

Η συνάρτηση πιθανότητας p_X της τυχαίας μεταβλητής X μπορεί να εκφραστεί συναρτήσει της p ως εξής:

$$p_X(x) = P[X = x] = \sum_{y \in S_Y} P[X = x, Y = y] = \sum_{y \in S_Y} p(x, y), \quad x \in S_X.$$

Ομοίως, η συνάρτηση πιθανότητας p_Y της τυχαίας μεταβλητής Y μπορεί να εκφραστεί συναρτήσει της p ως εξής: $p_Y(y) = P[Y = y] = \sum_{x \in S_X} P[X = x, Y = y] = \sum_{x \in S_X} p(x, y), \quad y \in S_Y.$

Παράδειγμα 2. Έστω η διακριτή δ.τ.μ. (X, Y) με σύνολο δυνατών τιμών (φορέα) $S = \{1, 2\} \times \{1, 2, 3\}$ και σ.π. $p(x, y) = P[X = x, Y = y] = \frac{1}{45}(x + 3y), \quad (x, y) \in S$. (α) Δείξτε ότι η p είναι πράγματι μία από κοινού σ.π. (β) Βρείτε τις σ.π. των τυχαίων μεταβλητών X και Y .

Λύση. (α) Είναι $p(x, y) \geq 0, \quad (x, y) \in S$. Επιπλέον $\sum_{x=1}^2 \sum_{y=1}^3 p(x, y) = \frac{1}{45} \sum_{x=1}^2 \sum_{y=1}^3 (x + 3y) = 1$. Επομένως, η p είναι πράγματι μία από κοινού σ.π.

$$(\beta) \text{ Είναι } p_X(x) = P[X = x] = \sum_{y=1}^3 p(x, y) = \frac{x+6}{15} = \begin{cases} \frac{7}{15}, & x = 1 \\ \frac{8}{15}, & x = 2 \end{cases}.$$

$$p_Y(y) = P[Y = y] = \sum_{x=1}^2 p(x, y) = \frac{1+2y}{15} = \begin{cases} \frac{3}{15}, & y = 1 \\ \frac{5}{15}, & y = 2 \\ \frac{7}{15}, & y = 3 \end{cases}.$$

Μπορούμε να συνοψίσουμε τις παραπάνω πιθανότητες στον ακόλουθο πίνακα:

y	1	2	3	$p_X(x)$
x				
1	$\frac{4}{45}$	$\frac{7}{45}$	$\frac{10}{45}$	$\frac{7}{15}$
2	$\frac{5}{45}$	$\frac{8}{45}$	$\frac{11}{45}$	$\frac{8}{15}$
$p_Y(y)$	$\frac{3}{15}$	$\frac{5}{15}$	$\frac{7}{15}$	

Παρατηρούμε ότι η συνάρτηση πιθανότητας p_X της τυχαίας μεταβλητής X μπορεί να ληφθεί αν υπολογίσουμε τα αθροίσματα ανά γραμμή του πίνακα ενώ η συνάρτηση πιθανότητας p_Y της τυχαίας μεταβλητής Y μπορεί να ληφθεί αν υπολογίσουμε τα αθροίσματα ανά στήλη του πίνακα. Επειδή οι σ.π. p_X και p_Y των τυχαίων μεταβλητών X και Y εμφανίζονται στα περιθώρια του παραπάνω πίνακα καλούνται **περιθώριες συναρτήσεις πιθανότητας** (marginal probability functions).

Παράδειγμα 3. Έστω η διδιάστατη διακριτή δ.τ.μ. (X, Y) με συνάρτηση πιθανότητας $p(x, y) = \frac{8}{3^x 5^y}$, $x = 1, 2, \dots$ και $y = 1, 2, \dots$ (α) Να βρεθούν οι περιθώριες συναρτήσεις πιθανότητας των τυχαίων μεταβλητών X και Y . (β) Να υπολογιστεί η πιθανότητα $P(X \geq 2, Y \geq 2)$.

$$\text{Λύση. (α) Είναι } p_X(x) = \sum_{y=1}^{\infty} p(x, y) = \sum_{y=1}^{\infty} \frac{8}{3^x 5^y} = \frac{8}{3^x 5} \sum_{y=1}^{\infty} \left(\frac{1}{5}\right)^{y-1} = \frac{8}{5 \cdot 3^x} \frac{1}{\left(1 - \frac{1}{5}\right)} = \frac{2}{3^x}, x = 1, 2, \dots$$

$$p_Y(y) = \sum_{x=1}^{\infty} p(x, y) = \sum_{x=1}^{\infty} \frac{8}{3^x 5^y} = \frac{8}{5^y 3} \sum_{x=1}^{\infty} \left(\frac{1}{3}\right)^{x-1} = \frac{4}{5^y}, y = 1, 2, \dots$$

(β) Έστω τα ενδεχόμενα $A = \{X < 2\}$ και $B = \{Y < 2\}$.

$$\text{Είναι } P(X \geq 2, Y \geq 2) = 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A \cap B)$$

$$= 1 - P(X \leq 1) - P(Y \leq 1) + P(X \leq 1, Y \leq 1) = 1 - F_X(1) - F_Y(1) + F(1, 1),$$

όπου, $F_X(x)$, $F_Y(y)$ είναι οι περιθώριες αθροιστικές συναρτήσεις κατανομών των τυχαίων μεταβλητών X, Y και $F(x, y)$ είναι η από κοινού αθροιστική συνάρτηση κατανομής της διδιάστατης δ.τ.μ. (X, Y) .

$$F_X(1) = P(X \leq 1) = P(X = 1) = p_X(1) = \frac{2}{3}, F_Y(1) = P(Y \leq 1) = P(Y = 1) = p_Y(1) = \frac{4}{5},$$

$$F(1,1) = P(X \leq 1, Y \leq 1) = P(X = 1, Y = 1) = p(1,1) = \frac{8}{15}.$$

$$\text{Άρα, } P(X \geq 2, Y \geq 2) = 1 - \frac{2}{3} - \frac{4}{5} + \frac{8}{15} = \frac{1}{15}.$$

ΣΥΝΕΧΕΙΣ ΔΙΑΝΥΣΜΑΤΙΚΕΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ

Έστω X, Y δύο συνεχείς τυχαίες μεταβλητές οι οποίες είναι ορισμένες στον ίδιο δειγματικό χώρο. Τότε η διδιάστατη διανυσματική τυχαία μεταβλητή (X, Y) είναι **συνεχής** αν υπάρχει μία μη-αρνητική συνάρτηση δύο μεταβλητών $f: \mathfrak{R} \times \mathfrak{R} \rightarrow [0, \infty)$, $f(x, y)$, $x, y \in \mathfrak{R}$ τέτοια ώστε, για κάθε περιοχή $C \subseteq \mathfrak{R} \times \mathfrak{R}$, η οποία μπορεί να γραφεί μέσω ορθογωνίων με χρήση πεπερασμένου ή απείρως αριθμήσιμου πλήθους πράξεων (τομή, ένωση, συμπλήρωμα), ισχύει ότι $P[(X, Y) \in C] = \iint_{(x,y) \in C} f(x, y) dx dy$. Η

συνάρτηση $f(x, y)$ καλείται **από κοινού συνάρτηση πυκνότητας (σ.π.)** (joint probability density function) της διδιάστατης δ.τ.μ. (X, Y) . Αν A και B είναι υποσύνολα του συνόλου των πραγματικών αριθμών, τότε αν ορίσουμε $C = \{(x, y) : x \in A, y \in B\} = A \times B$, έχουμε

$$P[(X, Y) \in C] = P[X \in A, Y \in B] = \int_B \int_A f(x, y) dx dy.$$

Αν $A = (-\infty, a]$ και $B = (-\infty, b]$, έχουμε

$$P\{X \in (-\infty, a], Y \in (-\infty, b]\} = P[X \leq a, Y \leq b] = F(a, b) = \int_{-\infty}^b \int_{-\infty}^a f(x, y) dx dy.$$

Η συνάρτηση F καλείται **από κοινού (αθροιστική) συνάρτηση κατανομής** της διδιάστατης δ.τ.μ. (X, Y) . Αν παραγωγίσουμε ως προς a και b έχουμε $\frac{\partial^2}{\partial a \partial b} F(a, b) = f(a, b)$. Η τελευταία σχέση συνδέει την από κοινού σ.π. και την από κοινού συνάρτηση κατανομής της διδιάστατης δ.τ.μ. (X, Y) . Απαραίτητη συνθήκη για να είναι η $f(x, y)$ μία από κοινού σ.π. για τη διδιάστατη δ.τ.μ. (X, Y) είναι

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dy dx = 1.$$

Αν η διδιάστατη δ.τ.μ. (X, Y) είναι συνεχής, τότε οι τυχαίες μεταβλητές X και Y είναι συνεχείς (εναλλακτικά καλούνται **από κοινού συνεχείς** (joint continuous)) και οι συναρτήσεις πυκνότητας τους f_X και f_Y , αντίστοιχα, μπορούν να ληφθούν ως εξής:

$$P\{X \in A\} = P\{X \in A, Y \in (-\infty, \infty)\} = \int_A \int_{-\infty}^{\infty} f(x, y) dy dx = \int_A f_X(x) dx, \text{ όπου } f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \text{ είναι η σ.π.}$$

της τυχαίας μεταβλητής X . Ομοίως, η σ.π. της τυχαίας μεταβλητής Y , δίνεται από τη σχέση

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx. \text{ Για τις περιθώριες συναρτήσεις κατανομών } F_X(x), F_Y(y) \text{ ισχύουν οι εκφράσεις:}$$

$$F_X(x) = \lim_{y \rightarrow \infty} F(x, y) \text{ και } F_Y(y) = \lim_{x \rightarrow \infty} F(x, y). \text{ Από τις δύο τελευταίες σχέσεις μπορούμε να}$$

υπολογίσουμε απευθείας τις περιθώριες συναρτήσεις κατανομών των τυχαίων μεταβλητών X και Y από την από κοινού συνάρτηση κατανομής της διδιάστατης δ.τ.μ. (X, Y) .

$$\text{Ισχύει ότι: } F_X(x) = P(X \leq x) = P(X \leq x, -\infty < Y < \infty) = \int_{-\infty}^x \left[\int_{-\infty}^{\infty} f(s, y) dy \right] ds = \int_{-\infty}^x f_X(s) ds.$$

$$\text{Όμοια, } F_Y(y) = P(Y \leq y) = P(Y \leq y, -\infty < X < \infty) = \int_{-\infty}^y \left[\int_{-\infty}^{\infty} f(x, t) dx \right] dt = \int_{-\infty}^y f_Y(t) dt.$$

Παράδειγμα 4. Η από κοινού σ.π. της δ.τ.μ. (X, Y) δίνεται από τον τύπο

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y}, & x, y \in (0, \infty) \\ 0, & x, y \notin (0, \infty) \end{cases}. \text{ Να υπολογιστούν οι πιθανότητες (α) } P[X > 1, Y < 1], \text{ (β) } P[X < Y] \text{ και}$$

$$(\gamma) P[X < a].$$

$$\text{Λύση. (α) Είναι } P[X > 1, Y < 1] = \int_0^1 \int_1^{\infty} 2e^{-x}e^{-2y} dx dy = \int_0^1 2e^{-2y} \left[-e^{-x} \right]_1^{\infty} dy = e^{-1} \int_0^1 2e^{-2y} dy = e^{-1}(1 - e^{-2}).$$

(β)

$$P[X < Y] = \iint_{\{(x, y): x < y\}} 2e^{-x}e^{-2y} dx dy = \int_0^{\infty} \int_0^y 2e^{-x}e^{-2y} dx dy = \int_0^{\infty} 2e^{-2y}(1 - e^{-y}) dy = \int_0^{\infty} 2e^{-2y} dy - \int_0^{\infty} 2e^{-3y} dy = 1 - \frac{2}{3} = \frac{1}{3}.$$

$$(\gamma) P[X < a] = \int_0^a \int_0^{\infty} 2e^{-2y}e^{-x} dy dx = \int_0^a e^{-x} dx = 1 - e^{-a}.$$

Παράδειγμα 5. Η από κοινού σ.π. της δ.τ.μ. (X, Y) δίνεται από τον τύπο $f(x, y) = \begin{cases} e^{-(x+y)}, & x, y \in (0, \infty) \\ 0, & x, y \notin (0, \infty) \end{cases}.$

Βρείτε την σ.π. της τυχαίας μεταβλητής X/Y .

Λύση. Αρχικά θα υπολογίσουμε τη συνάρτηση κατανομής της τυχαίας μεταβλητής X/Y . Για $a > 0$, έχουμε διαδοχικά

$$F_{X/Y}(a) = P\left[X/Y \leq a\right] = \iint_{\substack{x/y \leq a \\ y > 0}} e^{-(x+y)} dx dy = \int_0^{\infty} \int_0^{ay} e^{-(x+y)} dx dy = \int_0^{\infty} (1 - e^{-ay}) e^{-y} dy = \left[-e^{-y} + \frac{e^{-(a+1)y}}{(a+1)} \right]_0^{\infty} = 1 - \frac{1}{(a+1)}.$$

Αν παραγωγίσουμε ως προς a , λαμβάνουμε τη σ.π. της τυχαίας μεταβλητής X/Y , η οποία δίνεται από τον τύπο: $F'_{X/Y}(a) = f_{X/Y}(a) = (a+1)^{-2}$, $a > 0$.

Παράδειγμα 6. Η από κοινού συνάρτηση πυκνότητας μιας διδιάστατης διανυσματικής τυχαίας μεταβλητής (X, Y) δίνεται από τον τύπο $f(x, y) = \frac{3x(x+y)}{5}$, $0 < x < 1$, $0 < y < 2$ και $f(x, y) = 0$, αλλού.

(α) Να υπολογιστεί η πιθανότητα $P\left(0 < X < \frac{1}{2}, 1 < Y < 2\right)$.

(β) Να βρεθούν οι περιθώριες συναρτήσεις πυκνότητας $f_X(x)$, $f_Y(y)$.

(γ) Να βρεθεί η από κοινού συνάρτηση κατανομής $F(x, y)$.

(δ) Να βρεθούν οι περιθώριες συναρτήσεις κατανομών $F_X(x)$, $F_Y(y)$.

Λύση. (α) $P\left(0 < X < \frac{1}{2}, 1 < Y < 2\right) = \int_0^{\frac{1}{2}} \int_1^2 \frac{3x(x+y)}{5} dy dx = \int_0^{\frac{1}{2}} \frac{3x}{5} \left[\frac{y^2}{2} + xy \right]_1^2 dx = \frac{11}{80}.$

(β) $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \frac{3x}{5} \int_0^2 (x+y) dy = \frac{6x(x+1)}{5}$, $0 < x < 1$.

$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^1 \frac{3x(x+y)}{5} dx = \frac{1}{5} + \frac{3y}{10}$, $0 < y < 2$.

(γ) $F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y \frac{3t(t+s)}{5} ds dt = \frac{yx^3}{5} + \frac{3y^2x^2}{20}$, $0 < x < 1$, $0 < y < 2$.

(δ) $F_X(x) = \int_{-\infty}^x f_X(s) ds = \int_0^x \frac{6s(s+1)}{5} ds = \frac{6}{5} x^2 \left(\frac{x}{3} + \frac{1}{2} \right)$, $0 < x < 1$.

$F_Y(y) = \int_{-\infty}^y f_Y(s) ds = \int_0^y \left(\frac{1}{5} + \frac{3s}{10} \right) ds = \frac{1}{5} y + \frac{3}{20} y^2$, $0 < y < 2$.

Παράδειγμα 7. Η από κοινού αθροιστική συνάρτηση κατανομής των χρόνων ζωής X, Y δύο λαμπτήρων A, B σε χιλιάδες ώρες δίνεται από τον τύπο: $F(x, y) = (1 - e^{-4x^2})(1 - e^{-3y})$, $x, y > 0$ και $F(x, y) = 0$, αλλού.

(α) Να υπολογιστούν οι περιθώριες συναρτήσεις κατανομών των τυχαίων μεταβλητών X και Y .

(β) Ποια είναι η πιθανότητα (i) και οι δύο λαμπτήρες να ζήσουν περισσότερο από 500 ώρες (ii) τουλάχιστον ένας από τους δύο λαμπτήρες να ζήσει περισσότερο από 500 ώρες;

(γ) Να βρεθούν οι περιθώριες συναρτήσεις πυκνότητας των τυχαίων μεταβλητών X και Y . Ποια είναι η από κοινού συνάρτηση πυκνότητας της διδιάστατης διανυσματικής τυχαίας μεταβλητής (X, Y) ;

Λύση. (α) $F_X(x) = \lim_{y \rightarrow \infty} F(x, y) = (1 - e^{-4x^2}) \lim_{y \rightarrow \infty} (1 - e^{-3y}) = 1 - e^{-4x^2}$, $x > 0$.

$F_Y(y) = \lim_{x \rightarrow \infty} F(x, y) = (1 - e^{-3y}) \lim_{x \rightarrow \infty} (1 - e^{-4x^2}) = 1 - e^{-3y}$, $y > 0$.

(β) (i) Έστω τα ενδεχόμενα A : ο λαμπτήρας A ζει λιγότερο από 500 ώρες δηλαδή $X < \frac{1}{2}$ και B : ο λαμπτήρας B ζει λιγότερο από 500 ώρες δηλαδή $Y < \frac{1}{2}$. Ζητάμε την πιθανότητα

$$P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A \cap B).$$

$$\text{Είναι } P(A) = P(X < \frac{1}{2}) = F_X(\frac{1}{2}) = 1 - e^{-4(\frac{1}{2})^2} = 1 - e^{-1} \cong 63\%.$$

$$P(B) = P(Y < \frac{1}{2}) = F_Y(\frac{1}{2}) = 1 - e^{-3 \cdot \frac{1}{2}} = 1 - e^{-\frac{3}{2}}.$$

$$P(A \cap B) = P(X < \frac{1}{2}, Y < \frac{1}{2}) = F(\frac{1}{2}, \frac{1}{2}) = (1 - e^{-1})(1 - e^{-\frac{3}{2}}).$$

(ii) Ζητάμε την πιθανότητα $P(A^c \cup B^c) = P((A \cap B)^c) = 1 - P(A \cap B)$.

$$(γ) f_X(x) = F'_X(x) = (1 - e^{-4x^2})' = 8xe^{-4x^2}, x > 0.$$

$$f_Y(y) = F'_Y(y) = (1 - e^{-3y})' = 3e^{-3y}, y > 0.$$

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y) = 24xe^{-3y}e^{-4x^2}.$$

ΑΝΕΞΑΡΤΗΤΕΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ

Θεωρούμε ένα πείραμα στο οποίο ρίχνουμε ένα νόμισμα και ένα ζάρι. Διαισθητικά πιστεύουμε ότι όποιο κι αν είναι το αποτέλεσμα της ρίψης ενός νομίσματος δεν θα πρέπει να έχει καμία επίδραση στο αποτέλεσμα της ρίψης του ζαριού και αντίστροφα. Έστω X η τυχαία μεταβλητή που ισούται με 1 ή 0 ανάλογα με το αν το νόμισμα δείχνει κεφαλή ή γράμματα και έστω Y η τυχαία μεταβλητή που παίρνει τις τιμές 1,2,3,4,5 ή 6 ανάλογα με το αν η άνω έδρα του ζαριού φέρει τον αριθμό 1,2,3,4,5 ή 6 αντίστοιχα. Το αποτέλεσμα του διπλού πειράματος περιγράφεται από τη διδιάστατη διακριτή δ.τ.μ. (X,Y) . Το διαισθητικό μας συμπέρασμα ότι τα αποτελέσματα των ρίψεων του νομίσματος και του ζαριού δεν έχουν καμία επίδραση το ένα στο άλλο, μπορεί να διατυπωθεί αυστηρά ως εξής: αν x είναι ένας από τους αριθμούς 0 ή 1 και y είναι ένας από τους αριθμούς 1,2,3,4,5,6, τότε τα ενδεχόμενα $\{X = x\}$ και $\{Y = y\}$ πρέπει να είναι ανεξάρτητα. Δίνουμε τον ακόλουθο ορισμό.

Ορισμός. Οι τυχαίες μεταβλητές X και Y καλούνται **ανεξάρτητες** (independent) αν για οποιαδήποτε υποσύνολα A και B του συνόλου των πραγματικών αριθμών, ισχύει ότι

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\} \quad (1)$$

Ισοδύναμα, οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες όταν και μόνο όταν για οποιαδήποτε σύνολα A και B , τα ενδεχόμενα $E_A = \{X \in A\}$ και $E_B = \{Y \in B\}$ είναι ανεξάρτητα. Για οποιαδήποτε $a, b \in \mathcal{R}$, ισχύει ότι $P\{X \leq a, Y \leq b\} = P\{X \leq a\}P\{Y \leq b\}$ ή ισοδύναμα $F(a,b) = F_X(a)F_Y(b)$, όπου F είναι η από κοινού αθροιστική συνάρτηση κατανομής της διδιάστατης δ.τ.μ. (X,Y) και F_X, F_Y είναι οι συναρτήσεις κατανομών των τυχαίων μεταβλητών X και Y , αντίστοιχα.

Όταν οι X και Y είναι **διακριτές** τυχαίες μεταβλητές, η συνθήκη ανεξαρτησίας (1) είναι ισοδύναμη με τη σχέση

$$p(x,y) = p_X(x)p_Y(y), \quad (2)$$

ή με τη σχέση

$$P[X = x, Y = y] = P[X = x]P[Y = y], \text{ για κάθε } x, y$$

Η ισοδυναμία των σχέσεων (1) και (2) εξηγείται ως εξής: Αν στη σχέση (1), θέσουμε $A = \{x\}$ και $B = \{y\}$, λαμβάνουμε τη (2). Επιπλέον, αν η σχέση (2) ισχύει, τότε για οποιαδήποτε σύνολα A και B , έχουμε

$$P\{X \in A, Y \in B\} = \sum_{y \in B} \sum_{x \in A} P\{X = x, Y = y\} = \sum_{y \in B} \sum_{x \in A} P\{X = x\}P\{Y = y\} = \sum_{y \in B} P\{Y = y\} \sum_{x \in A} P\{X = x\}$$

$$= P\{Y \in B\}P\{X \in A\}.$$

Όταν οι X και Y είναι **συνεχείς** τυχαίες μεταβλητές, η συνθήκη της ανεξαρτησίας (6.1) είναι ισοδύναμη με τη σχέση $f(x, y) = f_X(x)f_Y(y)$, για κάθε x, y .

Διαισθητικά, οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες όταν η γνώση της μίας δεν επηρεάζει την κατανομή της άλλης. Αν δύο τυχαίες μεταβλητές δεν είναι ανεξάρτητες, τότε καλούνται **εξαρτημένες** (dependent).

Παράδειγμα 8. Ένας άνδρας και μία γυναίκα έχουν συμφωνήσει να συναντηθούν σε ένα ζαχαροπλαστείο κάποια χρονική στιγμή μεταξύ 12:00 και 1:00 το μεσημέρι. Υποθέτουμε ότι ο άνδρας φθάνει στο ζαχαροπλαστείο στις $12 + X$ και η γυναίκα φθάνει στις $12 + Y$. Υποθέτουμε επιπλέον ότι οι X και Y είναι ανεξάρτητες τυχαίες μεταβλητές και ότι ακολουθούν την Ομοιόμορφη κατανομή στο διάστημα $(0,1)$. Να υπολογιστεί η πιθανότητα $P\left(|X - Y| > \frac{1}{6}\right)$, δηλαδή η πιθανότητα ο άνδρας να περιμένει τη γυναίκα ή αντίστροφα για περισσότερο από δέκα λεπτά.

Λύση. Έστω $f(x, y)$ η από κοινού συνάρτηση πυκνότητας της διδιάστατης δ.τ.μ. (X, Y) και f_X, f_Y οι σ.π. των τυχαίων μεταβλητών X και Y , αντίστοιχα.

$$\text{Είναι } \left(|X - Y| > \frac{1}{6}\right) = \{X - Y > \frac{1}{6}\} \cup \{X - Y < -\frac{1}{6}\} \text{ και τα ενδεχόμενα } \{X - Y > \frac{1}{6}\}, \{X - Y < -\frac{1}{6}\}$$

είναι ασυμβίβαστα.

$$\text{Άρα, } P\left(|X - Y| > \frac{1}{6}\right) = P\left\{Y + \frac{1}{6} < X\right\} + P\left\{X + \frac{1}{6} < Y\right\} = 2P\left\{X + \frac{1}{6} < Y\right\} = 2 \iint_{x+\frac{1}{6} < y} f(x, y) dx dy$$

$$= 2 \iint_{x+\frac{1}{6} < y} f_X(x) f_Y(y) dx dy = 2 \iint_{x+\frac{1}{6} < y} dx dy = 2 \int_{\frac{1}{6}}^1 \int_0^{y-\frac{1}{6}} dx dy = 2 \int_{\frac{1}{6}}^1 \left(y - \frac{1}{6}\right) dy = 2 \frac{25}{72} = \frac{25}{36}.$$

Η δεύτερη ισότητα είναι συνέπεια της συμμετρίας και η τέταρτη ισότητα είναι συνέπεια της ανεξαρτησίας των τυχαίων μεταβλητών X και Y .

Παράδειγμα 9. Υποθέτουμε ότι ο αριθμός των ατόμων που εισέρχονται κατά τη διάρκεια μιας μέρας σε ένα ταχυδρομείο ακολουθεί την κατανομή Poisson με παράμετρο λ . Δείξτε ότι, αν το κάθε άτομο που εισέρχεται στο ταχυδρομείο είναι άνδρας με πιθανότητα p και γυναίκα με πιθανότητα $1 - p$, τότε ο αριθμός των ανδρών και των γυναικών που μπαίνουν στο ταχυδρομείο κατά τη διάρκεια μιας μέρας είναι

ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κατανομή Poisson με παραμέτρους λp και $\lambda(1-p)$, αντίστοιχα.

Λύση. Έστω X και Y , αντίστοιχα, ο αριθμός των ανδρών και των γυναικών που εισέρχονται στο ταχυδρομείο κατά τη διάρκεια μιας μέρας. Από το Θεώρημα Ολικής Πιθανότητας και αν δεσμευτούμε ως προς την τυχαία μεταβλητή $X+Y$, έχουμε

$$\begin{aligned} P[X=i, Y=j] &= P[X=i, Y=j | X+Y=i+j]P[X+Y=i+j] + P[X=i, Y=j | X+Y \neq i+j]P[X+Y \neq i+j] \\ &= P[X=i, Y=j | X+Y=i+j]P[X+Y=i+j] = \binom{i+j}{i} p^i (1-p)^{i+j-i} e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}. \end{aligned}$$

Η δεύτερη ισότητα προκύπτει διότι προφανώς ισχύει ότι $P[X=i, Y=j | X+Y \neq i+j] = 0$. Η τρίτη ισότητα προκύπτει διότι, εξ υποθέσεως, η τυχαία μεταβλητή $X+Y$ ακολουθεί την κατανομή Poisson με παράμετρο λ . Επιπλέον, δοθέντος ότι $i+j$ άνθρωποι εισέρχονται στο ταχυδρομείο και ο καθένας από αυτούς είναι άνδρας με πιθανότητα p , η πιθανότητα ακριβώς i άτομα να είναι άνδρες (και επομένως

ακριβώς j άτομα να είναι γυναίκες) είναι ίση με $P[X=i, Y=j | X+Y=i+j] = \binom{i+j}{i} p^i (1-p)^{i+j-i}$,

δηλαδή η παραπάνω πιθανότητα είναι διωνυμική με παραμέτρους $i+j$ και p . Συνεπώς,

$$P[X=i, Y=j] = e^{-\lambda} (\lambda p)^i \frac{[\lambda(1-p)]^j}{i!j!}, \quad i, j = 0, 1, \dots$$

Η περιθώρια συνάρτηση κατανομής της τυχαίας μεταβλητής X βρίσκεται ως εξής:

$$P[X=i] = \sum_{j=0}^{\infty} P[X=i, Y=j] = e^{-\lambda} \frac{(\lambda p)^i}{i!} \sum_{j=0}^{\infty} \frac{[\lambda(1-p)]^j}{j!} = e^{-\lambda} \frac{(\lambda p)^i}{i!} e^{\lambda(1-p)} = e^{-\lambda p} \frac{(\lambda p)^i}{i!}, \quad i = 0, 1, \dots$$

Άρα, $X \sim \text{Poisson}(\lambda p)$. Ομοίως, η περιθώρια συνάρτηση κατανομής της τυχαίας μεταβλητής Y είναι

$$P[Y=j] = \sum_{i=0}^{\infty} P[X=i, Y=j] = e^{-\lambda(1-p)} \frac{[\lambda(1-p)]^j}{j!}, \quad j = 0, 1, \dots \quad \text{Άρα, } Y \sim \text{Poisson}[\lambda(1-p)].$$

Παρατηρούμε ότι $P[X=i, Y=j] = P[X=i]P[Y=j]$, $i, j = 0, 1, \dots$. Επομένως, οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες.

Παράδειγμα 10. Η από κοινού συνάρτηση πιθανότητας της διδιάστατης διακριτής τυχαίας μεταβλητής

(X, Y) δίνεται από τον τύπο $f(x, y) = \frac{3}{8} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{2y}$, $x, y = 0, 1, 2, \dots$. Να εξεταστεί αν οι τυχαίες

μεταβλητές X και Y είναι ανεξάρτητες και να υπολογιστεί η πιθανότητα $P(X \geq 2 \cup Y \leq 2)$.

Λύση. Για τη περιθώρια συνάρτηση πιθανότητας της τυχαίας μεταβλητής X έχουμε

$$f_X(x) = \sum_{y=0}^{\infty} \frac{3}{8} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{2y} = \frac{3}{8} \left(\frac{1}{2}\right)^x \sum_{y=0}^{\infty} \left(\frac{1}{4}\right)^y = \frac{3}{8} \left(\frac{1}{2}\right)^x \frac{1}{1 - \frac{1}{4}} = \left(\frac{1}{2}\right)^{x+1}, \quad x = 0, 1, 2, \dots$$

Για τη περιθώρια συνάρτηση πιθανότητας της τυχαίας μεταβλητής Y έχουμε

$$f_Y(y) = \sum_{x=0}^{\infty} \left(\frac{3}{8}\right) \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{2y} = \frac{3}{8} \left(\frac{1}{2}\right)^{2y} \sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^x = \frac{3}{8} \left(\frac{1}{2}\right)^{2y} \frac{1}{1 - \frac{1}{2}} = \frac{3}{4} \left(\frac{1}{2}\right)^{2y}, \quad y = 0, 1, \dots$$

Οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες διότι $f(x, y) = f_X(x) f_Y(y)$, για όλα τα $x, y = 0, 1, \dots$

Χρησιμοποιούμε την ανεξαρτησία των τυχαίων μεταβλητών X και Y και έχουμε

$$P(X \geq 2 \cup Y \leq 2) = P(X \geq 2) + P(Y \leq 2) - P(X \geq 2, Y \leq 2) = 1 - P(X \leq 1) + P(Y \leq 2) - [1 - P(X \leq 1)]P(Y \leq 2)$$

Όμως,

$$P(X \leq 1) = f_X(0) + f_X(1) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4},$$

$$P(Y \leq 2) = f_Y(0) + f_Y(1) + f_Y(2) = \frac{3}{4} + \frac{3}{16} + \frac{3}{64} = \frac{63}{64}.$$

Με αντικατάσταση στον προηγούμενο τύπο έχουμε

$$P(X \geq 2 \cup Y \leq 2) = 1 - \frac{3}{4} + \frac{63}{64} - \left(1 - \frac{3}{4}\right) \frac{63}{64} = \frac{253}{256}.$$

ΚΑΤΑΝΟΜΗ ΤΟΥ ΑΘΡΟΙΣΜΑΤΟΣ ΑΝΕΞΑΡΤΗΤΩΝ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Η μέση τιμή μιας συνάρτησης $g(X, Y)$ της δ.τ.μ. (X, Y) είναι $E[g(X, Y)] = \sum_y \sum_x g(x, y) p(x, y)$, αν η

δ.τ.μ. (X, Y) είναι διακριτή με από κοινού συνάρτηση πιθανότητας $p(x, y)$, ενώ

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy, \quad \text{αν η δ.τ.μ. } (X, Y) \text{ είναι συνεχής με από κοινού συνάρτηση}$$

πυκνότητας $f(x, y)$. Ισχύει η ακόλουθη πρόταση.

Πρόταση. Έστω τυχαίες μεταβλητές X, Y που ορίζονται στον ίδιο πιθανοθεωρητικό χώρο $(\Omega, \mathfrak{F}, P)$. Αν οι X, Y είναι ανεξάρτητες τότε $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$, όπου g και h είναι πραγματικές συναρτήσεις.

Απόδειξη. Θα αποδείξουμε την πρόταση στην περίπτωση κατά την οποία οι τυχαίες μεταβλητές X και Y είναι συνεχείς. Η απόδειξη είναι παρόμοια στην περίπτωση κατά την οποία οι X και Y είναι διακριτές τυχαίες μεταβλητές. Έστω $f(x, y)$ η από κοινού συνάρτηση πυκνότητας της διδιάστατης δ.τ.μ. (X, Y) και f_X, f_Y οι συναρτήσεις πυκνότητας των X, Y . Διαδοχικά, έχουμε

$$E[g(X)h(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f(x, y)dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy = \int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

$$= E[h(Y)]E[g(X)]. \blacksquare$$

Παράδειγμα 11. Αν $X \sim \text{Poisson}(\lambda_1)$ και $Y \sim \text{Poisson}(\lambda_2)$, $\lambda_1, \lambda_2 > 0$ και X, Y είναι ανεξάρτητες τυχαίες μεταβλητές, τότε $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Λύση. Το ενδεχόμενο $\{X + Y = i\}$, $i = 0, 1, \dots$ μπορεί να γραφεί ως η τομή των ξένων μεταξύ τους ενδεχομένων $\{X = k\}$ και $\{Y = i - k\}$, $k = 0, \dots, i$. Διαδοχικά έχουμε

$$P[X + Y = i] = \sum_{k=0}^i P[X = k, Y = i - k] = \sum_{k=0}^i P[X = k]P[Y = i - k] = \sum_{k=0}^i e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{i-k}}{(i-k)!}$$

$$= e^{-(\lambda_1 + \lambda_2)} \sum_{k=0}^i \frac{\lambda_1^k \lambda_2^{i-k}}{k!(i-k)!} = \frac{e^{-(\lambda_1 + \lambda_2)}}{i!} \sum_{k=0}^i \frac{i!}{k!(i-k)!} \lambda_1^k \lambda_2^{i-k} = \frac{e^{-(\lambda_1 + \lambda_2)}}{i!} (\lambda_1 + \lambda_2)^i.$$

Άρα, $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

ΜΕΣΗ ΤΙΜΗ ΚΑΙ ΔΙΑΣΠΟΡΑ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Στο παρόν εδάφιο θα ασχοληθούμε με την εύρεση της μέσης τιμής και της διασποράς του αθροίσματος δυο τυχαίων μεταβλητών. Ισχύει η ακόλουθη πρόταση.

Πρόταση. Έστω δύο τυχαίες μεταβλητές X, Y με πεπερασμένες μέσες τιμές. Τότε η τυχαία μεταβλητή $X + Y$ έχει πεπερασμένη μέση τιμή και επιπλέον ισχύει ότι $E(X + Y) = E(X) + E(Y)$.

Απόδειξη. Έστω ότι η δ.τ.μ. (X, Y) είναι συνεχής και έστω $f(x, y), f_X, f_Y$ οι σ.π. της διδιάστατης δ.τ.μ. (X, Y) και των τυχαίων μεταβλητών X, Y αντίστοιχα. Θεωρούμε τη συνάρτηση $g(X, Y) = X + Y$. Διαδοχικά έχουμε

$$E[g(X, Y)] = E(X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y)f(x, y)dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x, y)dydx + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x, y)dxdy =$$

$= \int_{-\infty}^{\infty} xf_X(x)dx + \int_{-\infty}^{\infty} yf_Y(y)dy = E(X) + E(Y)$. Αν η διδιάστατη δ.τ.μ. (X,Y) είναι διακριτή η απόδειξη γίνεται με παρόμοιο τρόπο. ■

Αν $X_1, \dots, X_r$ είναι τυχαίες μεταβλητές με πεπερασμένες μέσες τιμές, τότε με επαγωγή αποδεικνύεται ότι

$$E\left[\sum_{i=1}^r X_i\right] = \sum_{i=1}^r E[X_i].$$

Η τελευταία σχέση είναι πολύ χρήσιμη για τον υπολογισμό μέσων τιμών.

Παράδειγμα 12. Έστω ότι N άνδρες πετάνε τα καπέλα τους στο πάτωμα ενός δωματίου. Τα καπέλα ανακατεύονται και ο κάθε άνδρας διαλέγει ένα καπέλο κατά τυχαίο τρόπο. Βρείτε τη μέση τιμή του αριθμού των ανδρών που διαλέγουν τα δικά τους καπέλα.

Λύση. Έστω η τυχαία μεταβλητή $X_i, i=1, \dots, N$ τέτοια ώστε $X_i = 1$, αν ο άνδρας i διαλέγει το δικό του καπέλο και $X_i = 0$, αν ο άνδρας i δεν διαλέγει το δικό του καπέλο. Έστω η τυχαία μεταβλητή $X = X_1 + \dots + X_N$. Ζητάμε να βρούμε τη μέση τιμή $E(X)$. Ισχύει ότι $EX_i = P(X_i = 1) = \frac{1}{N}$, για κάθε $i = 1, \dots, N$. Συνεπώς, $E(X) = E(X_1) + \dots + E(X_N) = 1$.

Παράδειγμα 13. Δέκα κυνηγοί περιμένουν να περάσει ένα σμήνος από δέκα κοτσύφια. Όταν περάσει το σμήνος, οι κυνηγοί πυροβολούν ταυτόχρονα αλλά ο καθένας διαλέγει το στόχο του τυχαία και ανεξάρτητα από τους άλλους. Αν ο κάθε κυνηγός πετυχαίνει το στόχο του με πιθανότητα p , να βρεθεί ο αναμενόμενος αριθμός των κοτσυφιών που ξεφεύγουν από τους πυροβολισμούς των κυνηγών.

Λύση. Έστω η τυχαία μεταβλητή $X_i, i=1, \dots, 10$ τέτοια ώστε $X_i = 1$, αν το i -οστό κοτσύφι ξεφεύγει και $X_i = 0$, αν το i -οστό κοτσύφι δεν ξεφεύγει από τους πυροβολισμούς των κυνηγών.

Έστω η τυχαία μεταβλητή $X = X_1 + \dots + X_{10}$. Ζητάμε να βρούμε τη μέση τιμή $E(X)$.

Ισχύει ότι $E(X) = E(X_1 + \dots + X_{10}) = E(X_1) + \dots + E(X_{10})$.

Για κάθε $i = 1, \dots, 10$ έχουμε $E(X_i) = P(X_i = 1)$. Ο κάθε κυνηγός, ανεξάρτητα από τους άλλους, πετυχαίνει το i -οστό κοτσύφι με πιθανότητα $\frac{p}{10}$. Άρα, $P(X_i = 1) = \left(1 - \frac{p}{10}\right)^{10}$. Συνεπώς,

$$E(X) = 10 \left(1 - \frac{p}{10}\right)^{10}.$$

Παράδειγμα 14. Η από κοινού συνάρτηση πιθανότητας μιας διακριτής διδιάστατης τυχαίας μεταβλητής

$$(X, Y) \text{ δίνεται από τον τύπο } p(x, y) = c \left(\frac{x}{y} \right), \quad x = 1, 2, 3, \quad y = 1, 2.$$

(α) Ποια είναι η τιμή της σταθεράς c και ποιες οι περιθώριες συναρτήσεις πιθανότητας των τυχαίων μεταβλητών X και Y .

(β) Να υπολογιστούν οι μέσες τιμές των τυχαίων μεταβλητών X και Y .

Λύση. (α) Πρέπει $\sum_{x=1}^3 \sum_{y=1}^2 c \left(\frac{x}{y} \right) = 1 \Rightarrow \sum_{x=1}^3 c x \left(\frac{3}{2} \right) = 1 \Rightarrow c = \frac{1}{9}.$

$$p_X(x) = \sum_{y=1}^2 p(x, y) = \sum_{y=1}^2 \frac{1}{9} \left(\frac{x}{y} \right) = \frac{x}{6}, \quad x = 1, 2, 3.$$

$$p_Y(y) = \sum_{x=1}^3 p(x, y) = \sum_{x=1}^3 \frac{1}{9} \left(\frac{x}{y} \right) = \frac{6}{9y}, \quad y = 1, 2.$$

$$(β) \quad E[X] = \sum_{x=1}^3 x p_X(x) = \sum_{x=1}^3 \frac{x^2}{6} = \frac{14}{6}, \quad E[Y] = \sum_{y=1}^2 y p_Y(y) = \sum_{y=1}^2 y \frac{6}{9y} = \frac{12}{9}.$$

Παράδειγμα 15. Η από κοινού συνάρτηση πυκνότητας μιας διδιάστατης συνεχούς τυχαίας μεταβλητής

(X, Y) δίνεται από τον τύπο $f(x, y) = 1$, αν $0 < x, y < 1$ και $f(x, y) = 0$, διαφορετικά.

(α) Να υπολογιστούν οι μέσες τιμές των τυχαίων μεταβλητών X, Y και XY .

(β) Να υπολογιστεί η μέση τιμή $E(X^2 + Y^2)$.

(γ) Να υπολογιστούν αρχικά οι πιθανότητες $P(X + Y < 0.5)$, $P(0.5 \leq X + Y < 1)$ και $P(X + Y \geq 1)$ και να βρεθεί στη συνέχεια η μέση τιμή της τυχαίας μεταβλητής Z , όπου $Z = 0$, αν $X + Y < 0.5$, $Z = 1$, αν $0.5 \leq X + Y < 1$ και $Z = 2$, αν $X + Y \geq 1$.

Λύση. (α) Έχουμε $E(X) = \int_0^1 \int_0^1 x dy dx = \int_0^1 x dx = \frac{1}{2}$, $E(Y) = \int_0^1 \int_0^1 y dy dx = \int_0^1 \frac{1}{2} dx = \frac{1}{2}$,

$$E(XY) = \int_0^1 \int_0^1 xy dy dx = \int_0^1 \frac{x}{2} dx = \frac{1}{4}.$$

$$(\beta) E(X^2 + Y^2) = \int_0^1 \int_0^1 (x^2 + y^2) dy dx = \int_0^1 \left(x^2 [y]_0^1 + \frac{1}{3} [y^3]_0^1 \right) dx = \int_0^1 \left(x^2 + \frac{1}{3} \right) dx = \frac{1}{3} [x^3]_0^1 + \frac{1}{3} [x]_0^1 = \frac{2}{3}.$$

$$(\gamma) \text{ Είναι } P(X + Y < 0.5) = \int_0^{0.5} \int_0^{0.5-x} dy dx = \int_0^{0.5} (0.5 - x) dx = \frac{1}{8},$$

$$P(0.5 \leq X + Y < 1) = \int_0^{0.5} \int_{0.5-x}^{1-x} dy dx + \int_{0.5}^1 \int_0^{1-x} dy dx = \int_0^{0.5} 0.5 dx + \int_{0.5}^1 (1-x) dx = \frac{3}{8},$$

$$P(X + Y \geq 1) = 1 - P(X + Y < 1) = 1 - (P(X + Y < 0.5) + P(0.5 \leq X + Y < 1)) = 1 - \frac{1}{8} - \frac{3}{8} = \frac{1}{2}.$$

Για τη μέση τιμή της τυχαίας μεταβλητής Z , η οποία είναι διακριτή τυχαία μεταβλητή, έχουμε

$$E(Z) = \sum_{z=0}^2 zP(Z=z) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{1}{2} = \frac{11}{8}.$$

ΣΥΝΔΙΑΚΥΜΑΝΣΗ ΔΥΟ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Έστω X, Y δύο τυχαίες μεταβλητές. Η διακύμανση μιας τυχαίας μεταβλητής αποτελεί ένα μέτρο της μεταβλητότητάς της. Αν, για παράδειγμα, θεωρήσουμε ένα γραμμικό συνδυασμό $Z = aX + bY$, των τυχαίων μεταβλητών X και Y , όπου $a, b \in \mathbb{R}$ σταθερές, είναι διαισθητικά προφανές ότι η διακύμανση της Z θα πρέπει να επηρεάζεται τόσο από τις διακυμάνσεις των X, Y όσο και από την από κοινού συμπεριφορά των τυχαίων μεταβλητών X, Y . Η ποσότητα που αντικατοπτρίζει την από κοινού συμπεριφορά των X, Y δίνεται στον ακόλουθο ορισμό.

Ορισμός (Συνδιακύμανση). Η **συνδιακύμανση** (covariance) δύο τυχαίων μεταβλητών X και Y , ορίζεται ως εξής: $\text{Cov}(X, Y) = E\{(X - E(X))(Y - E(Y))\}$.

Από τον παραπάνω ορισμό, αναπτύσσοντας το δεξιό μέλος της τελευταίας ισότητας, διαδοχικά έχουμε

$$\begin{aligned} \text{Cov}(X, Y) &= E\{XY - YE(X) - XE(Y) + E(X)E(Y)\} = E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y). \end{aligned}$$

Αν X, Y είναι ανεξάρτητες τυχαίες μεταβλητές, τότε από την Πρόταση 4.1, έχουμε $\text{Cov}(X, Y) = 0$.

Το αντίστροφο δεν ισχύει. Δηλαδή, αν $\text{Cov}(X, Y) = 0$, δεν έπεται ότι οι X, Y είναι ανεξάρτητες. Ένα απλό παράδειγμα δύο εξαρτημένων τυχαίων μεταβλητών X, Y που έχουν μηδενική συνδιακύμανση

μπορεί να ληφθεί αν υποθέσουμε ότι η X είναι τέτοια ώστε $P(X = 0) = P(X = 1) = P(X = -1) = \frac{1}{3}$, και ορίσουμε την Y , έτσι ώστε $Y = 0$, αν $X \neq 0$ και $Y = 1$, αν $X = 0$. Τότε $XY = 0$, και επομένως, $E(XY) = 0$. Επιπλέον, $E(X) = 0$ και $Cov(X, Y) = E(XY) - E(X)E(Y) = 0$. Οι X, Y είναι φανερά εξαρτημένες τυχαίες μεταβλητές.

Ισχύει η ακόλουθη πρόταση.

Πρόταση. Έστω X, Y δύο τυχαίες μεταβλητές με πεπερασμένες ροπές δεύτερης τάξης. Τότε η τυχαία μεταβλητή $X + Y$ έχει πεπερασμένη ροπή δεύτερης τάξης και επομένως έχει πεπερασμένη διασπορά. Επιπλέον, ισχύει ότι $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$.

Απόδειξη. Είναι $Var(X + Y) = E[X + Y - E(X + Y)]^2 = E[(X - E(X)) + (Y - E(Y))]^2$
 $= E(X - E(X))^2 + E(Y - E(Y))^2 + 2E[(X - E(X))(Y - E(Y))] = Var(X) + Var(Y) + 2Cov(X, Y)$. ■

Ισχύει η ακόλουθη πρόταση.

Πρόταση. Αν X, Y, Z τυχαίες μεταβλητές και $a, b \in \mathbb{R}$, ισχύει ότι

$$Cov(aX + bY, Z) = aCov(X, Z) + bCov(Y, Z).$$

Απόδειξη. Είναι

$$Cov(aX + bY, Z) = E[(aX + bY)Z] - E(aX + bY)E(Z) = a[E(XZ) - E(X)E(Z)] + b[E(YZ) - E(Y)E(Z)]$$

$$= aCov(X, Z) + bCov(Y, Z). \quad \blacksquare$$

Παράδειγμα 16. Για το Παράδειγμα 12 υπολογίστε τη διασπορά του αριθμού των ανδρών που διαλέγουν τα δικά τους καπέλα.

Λύση. Ισχύει ότι $Var(X) = \sum_{i=1}^N Var(X_i) + 2 \sum_{i=1}^N \sum_{j=i+1}^N Cov(X_i, X_j)$.

$$\text{Όμως, } Var(X_i) = EX_i^2 - (EX_i)^2 = P(X_i = 1) - (EX_i)^2 = \frac{1}{N} - \left(\frac{1}{N}\right)^2 = \frac{N-1}{N^2}, \text{ για } i = 1, \dots, N.$$

Επίσης, για i, j με $i < j$, έχουμε ότι $Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$. Όμως, $X_i X_j = 1$, αν αμφότεροι οι άνδρες i και j διαλέγουν τα δικά τους καπέλα και $X_i X_j = 0$, σε οποιαδήποτε άλλη περίπτωση.

$$\text{Άρα, } E(X_i X_j) = P(X_i = 1, X_j = 1) = P(X_i = 1)P(X_j = 1 | X_i = 1) = \frac{1}{N} \cdot \frac{1}{(N-1)}.$$

$$\text{Συνεπώς, } \text{Cov}(X_i, X_j) = \frac{1}{N(N-1)} - \left(\frac{1}{N}\right)^2 = \frac{1}{N^2(N-1)}.$$

$$\text{Επομένως, } \text{Var}(X) = \frac{N-1}{N} + 2 \frac{N(N-1)}{2} \cdot \frac{1}{N^2(N-1)} = \frac{N-1}{N} + \frac{1}{N} = 1.$$

ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ ΔΥΟ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Έστω X και Y δύο τυχαίες μεταβλητές με πεπερασμένες μη-μηδενικές διασπορές για τις οποίες διαπιστώσαμε ότι $\text{Cov}(X, Y) \neq 0$. Τότε οι X, Y δεν είναι ανεξάρτητες. Μας ενδιαφέρει να αποδώσουμε ποσοτικά το βαθμό εξάρτησης των X και Y με έναν κατάλληλο αριθμό. Η τιμή της συνδιακύμανσης των τυχαίων μεταβλητών X και Y επηρεάζεται σημαντικά από τις μονάδες μέτρησης των X και Y . Για να εξαλείψουμε την επίδραση των μονάδων μέτρησης των X και Y στο βαθμό της εξάρτησής τους δίνουμε τον ακόλουθο ορισμό.

Ορισμός. Ο **συντελεστής συσχέτισης** (correlation coefficient) δύο τυχαίων μεταβλητών X, Y τέτοιων

ώστε $\text{Var}(X) \cdot \text{Var}(Y) > 0$, ορίζεται ως εξής: $\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$ και αποτελεί ένα μέτρο του

βαθμού εξάρτησης μεταξύ τους.

Πρόταση. Ισχύει ότι $-1 \leq \rho(X, Y) \leq 1$.

Απόδειξη. Έστω σ_X^2 και σ_Y^2 οι διασπορές των τυχαίων μεταβλητών X και Y , αντίστοιχα. Ισχύει ότι

$$0 \leq \text{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = \frac{\text{Var}(X)}{\sigma_X^2} + \frac{\text{Var}(Y)}{\sigma_Y^2} + \frac{2\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \Rightarrow 0 \leq 1 + 1 + 2\rho(X, Y) \Rightarrow \rho(X, Y) \geq -1.$$

Επίσης, ισχύει ότι

$$0 \leq \text{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = \frac{\text{Var}(X)}{\sigma_X^2} + \frac{\text{Var}(Y)}{\sigma_Y^2} - \frac{2\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \Rightarrow 0 \leq 1 + 1 - 2\rho(X, Y) \Rightarrow \rho(X, Y) \leq 1.$$

Συνεπώς, $-1 \leq \rho(X, Y) \leq 1$. ■

Ο συντελεστής συσχέτισης δύο τυχαίων μεταβλητών X και Y είναι ένα μέτρο του βαθμού της γραμμικής εξάρτησης των X και Y . Μία τιμή του συντελεστή συσχέτισης κοντά στο +1 ή στο -1 είναι ένδειξη υψηλού βαθμού γραμμικής εξάρτησης μεταξύ των X και Y , ενώ μία τιμή του συντελεστή συσχέτισης κοντά στο 0 είναι ένδειξη ότι δεν υπάρχει γραμμική εξάρτηση. Μία θετική τιμή του

συντελεστή συσχέτισης είναι ένδειξη ότι η Y αυξάνει καθώς η X αυξάνει ενώ μία αρνητική τιμή του είναι ένδειξη ότι η Y μειώνεται καθώς η X αυξάνει. Αν $\rho(X, Y) = 0$, τότε οι τυχαίες μεταβλητές X και Y καλούνται **ασυσχέτιστες** (uncorrelated).

Παράδειγμα 17. Έστω X_1, X_2, X_3 τρεις ανεξάρτητες τυχαίες μεταβλητές που έχουν πεπερασμένες θετικές διασπορές $\sigma_1^2, \sigma_2^2, \sigma_3^2$, αντίστοιχα. Υπολογίστε τον συντελεστή συσχέτισης των $X_1 - X_2$ και $X_2 + X_3$.

Λύση. Ισχύει ότι $\text{Cov}(X_1 - X_2, X_2 + X_3) = E[(X_1 - X_2)(X_2 + X_3)] - E[X_1 - X_2]E[X_2 + X_3]$.

Από την ανεξαρτησία των X_1, X_2, X_3 , μετά από πράξεις, προκύπτει ότι

$$\text{Cov}(X_1 - X_2, X_2 + X_3) = -(E[X_2^2] - (E[X_2])^2) = -\text{Var}(X_2) = -\sigma_2^2.$$

$$\text{Συνεπώς, } \rho(X_1 - X_2, X_2 + X_3) = -\frac{\sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(\sigma_2^2 + \sigma_3^2)}}.$$

Παράδειγμα 18. Ένας αριθμός X επιλέγεται τυχαία στο διάστημα $[0, 1]$. Έστω R η απόκλιση του αριθμού που επιλέχτηκε από ένα σταθερό αριθμό a , όπου $0 \leq a \leq 1$. (α) Να βρεθεί ο συντελεστής συσχέτισης των τυχαίων μεταβλητών X και R . (β) Για ποια τιμή του a οι τυχαίες μεταβλητές X και R είναι ασυσχέτιστες;

Λύση. (α) Η τυχαία μεταβλητή X ακολουθεί την Ομοιόμορφη κατανομή στο διάστημα $[0, 1]$ ενώ για την τυχαία μεταβλητή R έχουμε $R = |a - X|$. Αφού η συνάρτηση πυκνότητας της X είναι $f(x) = 1$, για $0 \leq x \leq 1$, έχουμε

$$E[R] = E[|a - X|] = \int_{-\infty}^{\infty} |a - x| f(x) dx = \int_0^a (a - x) dx + \int_a^1 (x - a) dx = a^2 - a + \frac{1}{2}.$$

$$E[R^2] = E[|a - X|^2] = \int_{-\infty}^{\infty} (a - x)^2 f(x) dx = \int_0^1 (a - x)^2 dx = a^2 - a + \frac{1}{3}.$$

$$\text{Var}[R] = E[R^2] - (E[R])^2 = 2a^3 - a - a^2 + \frac{1}{12}.$$

$$E[XR] = E[X|a - X|] = \int_{-\infty}^{\infty} x|a - x| dx = \int_0^a x(a - x) dx + \int_a^1 x(x - a) dx = \frac{a^3}{3} - \frac{a}{2} + \frac{1}{3}.$$

Επειδή $E[X] = \frac{1}{2}$, $Var[X] = \frac{1}{12}$, προκύπτει, μετά από πράξεις, ότι

$$\rho(X, R) = \frac{E(XR) - E(X)E(R)}{\sqrt{Var(X)}\sqrt{Var(R)}} = \frac{4a^3 - 6a^2 + 1}{\sqrt{12}\sqrt{2a^3 - a^4 - a^2 + \frac{1}{12}}}.$$

(β) Για να είναι οι τυχαίες μεταβλητές X, R ασυσχέτιστες θα πρέπει ο συντελεστής συσχέτισης $\rho(X, R)$ να είναι ίσος με μηδέν ή ισοδύναμα $4a^3 - 6a^2 + 1 = (2a - 1)(2a^2 - 2a - 1) = 0$. Η εξίσωση έχει μοναδική λύση στο διάστημα $[0, 1]$, την $a = \frac{1}{2}$.

Παράδειγμα 19. Η από κοινού συνάρτηση πυκνότητας της διδιάστατης συνεχούς τυχαίας μεταβλητής (X, Y) δίνεται από τον τύπο $f(x, y) = 1$, αν $|x| < y < 1$ και $f(x, y) = 0$, διαφορετικά. Ναδειχθεί ότι, ενώ οι τυχαίες μεταβλητές X και Y είναι ασυσχέτιστες, δεν είναι ανεξάρτητες.

Λύση. Για τις τυχαίες μεταβλητές X και Y έχουμε $R_X = (-1, 1)$ και $R_Y = [0, 1]$.

$$\text{Είναι } f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{|x|}^{\infty} dy = 1 - |x|, \quad -1 < x < 1, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{-y}^y dx = 2y, \quad 0 \leq y < 1.$$

Οι τυχαίες μεταβλητές X, Y δεν είναι ανεξάρτητες διότι δεν ισχύει η σχέση $f(x, y) = f_X(x)f_Y(y)$ για κάθε x και y . Για παράδειγμα, αν πάρουμε $x = y = \frac{1}{2}$, θα έχουμε $1 = f\left(\frac{1}{2}, \frac{1}{2}\right) \neq f_X\left(\frac{1}{2}\right)f_Y\left(\frac{1}{2}\right) = \frac{1}{2}$. Για τις μέσες τιμές των τυχαίων μεταβλητών X, Y και XY έχουμε

$$E(X) = \int_{-1}^1 x(1 - |x|) dx = \int_{-1}^0 x(1 + x) dx + \int_0^1 x(1 - x) dx = \left[\frac{x^2}{2} + \frac{x^3}{3} \right]_{-1}^0 + \left[\frac{x^2}{2} - \frac{x^3}{3} \right]_{-1}^0 = 0,$$

$$E(Y) = \int_0^1 2y^2 dy = \frac{2}{3}, \quad E(XY) = \int_0^1 \int_{-y}^y xy dx dy = \int_0^1 y \left[\frac{x^2}{2} \right]_{-y}^y dy = \int_0^1 y \cdot 0 dy = 0.$$

Επειδή $Cov(X, Y) = E(XY) - E(X)E(Y) = 0 - \frac{2}{3} \cdot \frac{1}{2} = -\frac{1}{3} \neq 0$, έπεται ότι οι τυχαίες μεταβλητές X και Y είναι ασυσχέτιστες.

ΔΕΣΜΕΥΜΕΝΗ ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΤΗΤΑΣ ΚΑΙ ΔΕΣΜΕΥΜΕΝΗ ΣΥΝΑΡΤΗΣΗ ΠΥΚΝΟΤΗΤΑΣ

Έστω (X, Y) μία διδιάστατη διακριτή διανυσματική τυχαία μεταβλητή και $p(x, y)$ η από κοινού συνάρτηση πιθανότητάς της. Για κάθε δυνατή τιμή y της Y τέτοια ώστε $p_Y(y) = P(Y = y) \neq 0$, η

δεσμευμένη συνάρτηση πιθανότητας (δ.σ.π.) (conditional probability function) της X δοθέντος ότι $Y = y$, ορίζεται ως εξής:

$$p_{X|Y}(x|y) := P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x, y)}{p_Y(y)},$$

όπου $p_Y(y)$ είναι η συνάρτηση πιθανότητας της τυχαίας μεταβλητής Y .

Παρατηρούμε ότι, αν οι τυχαίες μεταβλητές X, Y είναι ανεξάρτητες, τότε η δ.σ.π. της X δοθέντος ότι $Y = y$ συμπίπτει με τη (μη δεσμευμένη) συνάρτηση πιθανότητας $p_X(x)$ της X , διότι

$$p_{X|Y}(x|y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x)P(Y = y)}{P(Y = y)} = p_X(x).$$

Η δεύτερη ισότητα είναι συνέπεια της ανεξαρτησίας των τυχαίων μεταβλητών X και Y .

Για κάθε συγκεκριμένη τιμή $y \in S_Y$, όπου S_Y είναι ο φορέας της Y , ορίζεται μία συνάρτηση πιθανότητας $g(x) = p_{X|Y}(x|y)$ στην οποία αντιστοιχεί μία αθροιστική συνάρτηση κατανομής $G(x) = \sum_{t \leq x} g(t)$.

Η συνάρτηση $G(x)$ καλείται **δεσμευμένη αθροιστική συνάρτηση κατανομής** (conditional cumulative distribution function) της X δοθέντος ότι $Y = y$.

Μπορούμε να γράψουμε $G(x) = P_{X|Y}(x|y) = P(X \leq x | Y = y) = \sum_{t \leq x} p_{X|Y}(t|y)$.

Έστω (X, Y) μία διδιάστατη συνεχής διανυσματική τυχαία μεταβλητή και $f(x, y)$ η από κοινού συνάρτηση πυκνότητάς της. Για μία συγκεκριμένη τιμή y τέτοια ώστε $f_Y(y) \neq 0$, η **δεσμευμένη συνάρτηση πυκνότητας (δ.σ.π.)** (conditional density function) της X δοθέντος ότι $Y = y$, ορίζεται ως

εξής: $f_{X|Y}(x|y) := \frac{f(x, y)}{f_Y(y)}$, όπου $f_Y(y)$ είναι η συνάρτηση πυκνότητας της τυχαίας μεταβλητής Y .

Παρατηρούμε ότι, αν οι τυχαίες μεταβλητές X, Y είναι ανεξάρτητες, τότε η δ.σ.π. της X δοθέντος ότι $Y = y$ συμπίπτει με τη (μη δεσμευμένη) συνάρτηση πυκνότητας $f_X(x)$ της X , διότι

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x).$$

Η δεύτερη ισότητα είναι συνέπεια της ανεξαρτησίας των τυχαίων μεταβλητών X και Y .

Για κάθε συγκεκριμένη τιμή $y \in S_Y$ τέτοια ώστε $f_Y(y) \neq 0$ ορίζεται μία συνάρτηση πυκνότητας

$$g(x) = f_{X|Y}(x|y) \text{ στην οποία αντιστοιχεί μία αθροιστική συνάρτηση κατανομής } G(x) = \int_{-\infty}^x g(t)dt.$$

Η συνάρτηση $G(x)$ καλείται **δεσμευμένη αθροιστική συνάρτηση κατανομής** της X δοθέντος ότι

$$Y = y. \text{ Μπορούμε να γράψουμε } G(x) = F_{X|Y}(x|y) = P(X \leq x | Y = y) = \int_{-\infty}^x f_{X|Y}(t|y)dt.$$

Παράδειγμα 20. Υποθέτουμε ότι η από κοινού συνάρτηση πιθανότητας της διδιάστατης διακριτής δ.τ.μ. (X, Y) δίνεται από τις σχέσεις: $p(0,0) = 0.4$, $p(0,1) = 0.2$, $p(1,0) = 0.1$, $p(1,1) = 0.3$. Υπολογίστε τη δεσμευμένη συνάρτηση πιθανότητας της X , δοθέντος ότι $Y = 1$.

Λύση. Ισχύει ότι $p_Y(1) = P(Y = 1) = \sum_{x=0}^1 p(x,1) = 0.2 + 0.3 = 0.5$. Επομένως, $p_{X|Y}(0|1) = \frac{p(0,1)}{p_Y(1)} = \frac{2}{5}$ και

$$p_{X|Y}(1|1) = \frac{p(1,1)}{p_Y(1)} = \frac{3}{5}.$$

Παράδειγμα 21. Έστω ότι $X \sim \text{Poisson}(\lambda_1)$ και $Y \sim \text{Poisson}(\lambda_2)$. Υποθέτουμε ότι οι τυχαίες μεταβλητές X, Y είναι ανεξάρτητες. Αν $Z := X + Y$, δείξτε ότι η δεσμευμένη κατανομή της X

δοθέντος ότι $Z = X + Y = z$ είναι διωνυμική με παραμέτρους z και $p := \frac{\lambda_1}{\lambda_1 + \lambda_2}$,

δηλαδή $X | Z = X + Y = z \sim \text{Bin}(z, p := \frac{\lambda_1}{\lambda_1 + \lambda_2})$ και $p_{X|Z}(x|z) = \binom{z}{x} p^x (1-p)^{z-x}$, $x = 0, 1, \dots, z$.

Λύση. Ισχύει ότι $Z = X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. Διαδοχικά έχουμε ότι

$$\begin{aligned} p_{X|Z}(x|z) &= \frac{P(X = x, Z = z)}{P(Z = z)} = \frac{P(X = x, X + Y = z)}{P(Z = z)} = \frac{P(X = x, Y = z - x)}{P(Z = z)} = \frac{P(X = x)P(Y = z - x)}{P(Z = z)} \\ &= \frac{\frac{e^{-\lambda_1} \lambda_1^x}{x!} \cdot \frac{e^{-\lambda_2} \lambda_2^{z-x}}{(z-x)!}}{\frac{e^{-(\lambda_1 + \lambda_2)} (\lambda_1 + \lambda_2)^z}{z!}} = \binom{z}{x} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^x \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{z-x} = \binom{z}{x} p^x (1-p)^{z-x}. \end{aligned}$$

Η τέταρτη ισότητα είναι συνέπεια της ανεξαρτησίας των τυχαίων μεταβλητών X και Y .

Παράδειγμα 22. Η από κοινού συνάρτηση πυκνότητας της διδιάστατης συνεχούς δ.τ.μ. (X, Y) δίνεται από τον τύπο $f(x, y) = \frac{12}{5}x(2 - x - y)$, αν $0 < x, y < 1$ και $f(x, y) = 0$, διαφορετικά. Υπολογίστε τη δεσμευμένη συνάρτηση πυκνότητας της X δοθέντος ότι $Y = y$, όπου $0 < y < 1$.

Λύση. Για $0 < x < 1$, $0 < y < 1$, έχουμε

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y) dx} = \frac{x(2 - x - y)}{\int_0^1 x(2 - x - y) dx} = \frac{x(2 - x - y)}{\frac{2}{3} - \frac{y}{2}} = \frac{6x(2 - x - y)}{4 - 3y}.$$

ΔΕΣΜΕΥΜΕΝΗ ΜΕΣΗ ΤΙΜΗ

Αν X και Y είναι δύο **διακριτές τυχαίες μεταβλητές**, τότε για κάθε δυνατή τιμή y της Y , τέτοια ώστε $p_Y(y) = P(Y = y) \neq 0$, η **δεσμευμένη μέση τιμή** (conditional expectation) της X , δοθέντος ότι $Y = y$, ορίζεται ως εξής:

$$E[X | Y = y] = \sum_{x \in S_X} xP[X = x | Y = y] = \sum_{x \in S_X} xp_{X|Y}(x|y),$$

όπου S_X είναι το σύνολο των δυνατών τιμών (φορέας) της X και $p_{X|Y}(x|y)$ είναι η δεσμευμένη συνάρτηση πιθανότητας της X , δοθέντος ότι $Y = y$.

Αν X και Y είναι δύο **συνεχείς τυχαίες μεταβλητές**, τότε για κάθε δυνατή τιμή y της Y , τέτοια ώστε $f_Y(y) \neq 0$, η **δεσμευμένη μέση τιμή** της X , δοθέντος ότι $Y = y$, ορίζεται ως εξής:

$$E[X | Y = y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y) dx,$$

όπου $f_{X|Y}(x|y)$ είναι η δεσμευμένη συνάρτηση πυκνότητας της X δοθέντος ότι $Y = y$.

Παράδειγμα 23. Η από κοινού συνάρτηση πιθανότητας μιας διδιάστατης διακριτής διανυσματικής τυχαίας μεταβλητής (X, Y) δίνεται από τον τύπο $p(x, y) = \frac{1}{15}(x + y)$, $x = 0, 1, 2$ και $y = 1, 2$. Να υπολογιστούν οι δεσμευμένες μέσες τιμές $E[X | Y = 1]$, $E[Y | X = 1]$ και να συγκριθούν μεταξύ τους.

Λύση. Είναι $E[X | Y = 1] = \sum_{x=0}^2 xp_{X|Y}(x|1)$. Όμως, $p_{X|Y}(x|y) = \frac{p(x, y)}{p_Y(y)}$ και $p_Y(y) = \sum_{x=0}^2 p(x, y) = \frac{1}{5}y + \frac{1}{5}$.

Άρα, $p_{X|Y}(x|y) = \frac{1}{3} \frac{(x+y)}{(y+1)}$ και $p_{X|Y}(x|1) = \frac{p(x,1)}{p_Y(1)} = \frac{x+1}{6}$. Είναι $E[X|Y=1] = \sum_{x=0}^2 x \left(\frac{x+1}{6} \right) = \frac{8}{6}$.

Είναι $E[Y|X=1] = \sum_{y=1}^2 y p_{Y|X}(y|1)$. Όμως, $p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)} = \frac{x+y}{2x+3}$

και $p_X(x) = \sum_{y=1}^2 p(x,y) = \frac{1}{15}(2x+3)$. Άρα, $p_{Y|X}(y|1) = \frac{1+y}{5}$.

Είναι $E[Y|X=1] = \sum_{y=1}^2 y p_{Y|X}(y|1) = \sum_{y=1}^2 \frac{y(1+y)}{5} = \frac{8}{5}$. Επομένως, $E[Y|X=1] > E[X|Y=1]$.

Παράδειγμα 24. Η από κοινού συνάρτηση πυκνότητας της διδιάστατης συνεχούς δ.τ.μ. (X, Y) δίνεται

από τον τύπο $f(x, y) = \frac{e^{-\frac{x}{y}} e^{-y}}{y}$, $0 < x, y < \infty$ και $f(x, y) = 0$, διαφορετικά. Υπολογίστε τη δεσμευμένη μέση τιμή $E(X|Y=y)$, όπου $y > 0$.

Λύση. Αρχικά, θα υπολογίσουμε τη δεσμευμένη συνάρτηση πυκνότητας της X , δοθέντος ότι $Y=y$, όπου $y > 0$. Διαδοχικά έχουμε

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y) dx} = \frac{\frac{e^{-\frac{x}{y}} e^{-y}}{y}}{e^{-y} \int_0^{\infty} \left(\frac{1}{y} \right) e^{-\frac{x}{y}} dx} = \frac{\frac{1}{y} e^{-\frac{x}{y}}}{\int_0^{\infty} \frac{1}{y} e^{-\frac{x}{y}} dx} = \frac{\left(\frac{1}{y} \right) e^{-\frac{x}{y}}}{[-e^{-\frac{x}{y}}]_0^{\infty}} = \frac{1}{y} e^{-\frac{x}{y}}, x > 0.$$

Παρατηρούμε ότι $X|Y=y \sim \text{Εκθετική}\left(\frac{1}{y}\right)$. Επομένως, $E(X|Y=y) = \int_0^{\infty} \frac{x}{y} e^{-\frac{x}{y}} dx = y$.

Παράδειγμα 25. Η από κοινού συνάρτηση πυκνότητας της διδιάστατης συνεχούς διανυσματικής τυχαίας μεταβλητής (X, Y) δίνεται από τον τύπο $f(x, y) = 4xy$, $0 < x, y < 1$ και $f(x, y) = 0$, αλλού.

(α) Να υπολογιστούν οι συναρτήσεις $f_X(x)$, $f_Y(y)$, $f_{X|Y}(x|y)$ και $f_{Y|X}(y|x)$.

(β) Να υπολογιστούν οι ποσότητες $E\left[X|Y = \frac{1}{2}\right]$ και $E\left[Y|X = \frac{1}{3}\right]$.

Λύση. (α) Είναι $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^1 4xy dy = 4x \left[\frac{y^2}{2} \right]_0^1 = 2x, \quad 0 < x < 1.$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^1 4xy dx = 4y \left[\frac{x^2}{2} \right]_0^1 = 2y, \quad 0 < y < 1.$$

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{4xy}{2y} = 2x, \quad 0 < x < 1.$$

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{4xy}{2x} = 2y, \quad 0 < y < 1.$$

(β) Είναι $E\left[X | Y = \frac{1}{2}\right] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y) dx = \int_0^1 2x^2 dx = \frac{2}{3}.$

$$E\left[Y | X = \frac{1}{3}\right] = \int_{-\infty}^{\infty} yf_{Y|X}(y|x) dy = \int_0^1 2y^2 dy = \frac{2}{3}.$$

Παράδειγμα 26. Η από κοινού συνάρτηση πυκνότητας μιας διδιάστατης συνεχούς τυχαίας μεταβλητής (X, Y) δίνεται από τον τύπο $f(x, y) = xe^{-x(y+1)}, \quad x, y > 0$ και $f(x, y) = 0$, διαφορετικά.

(α) Να βρεθούν οι περιθώριες συναρτήσεις πυκνότητας των τυχαίων μεταβλητών X και Y .

(β) Να βρεθούν οι δεσμευμένες συναρτήσεις πυκνότητας $f_{X|Y}(x|y)$ και $f_{Y|X}(y|x)$.

(γ) Να βρεθούν οι δεσμευμένες συναρτήσεις κατανομής $F_{X|Y}(x|y)$ και $F_{Y|X}(y|x)$.

(δ) Να βρεθούν οι δεσμευμένες μέσες τιμές $E(X | Y = y)$ και $E(Y | X = x)$.

Λύση. (α) Για $y > 0$, έχουμε

$$f_X(x) = xe^{-x} \int_0^{\infty} e^{-xy} dy = xe^{-x} \left[-\frac{1}{x} e^{-xy} \right]_0^{\infty} = xe^{-x} \left[-\frac{1}{x} (0 - 1) \right] = e^{-x}, \quad x > 0.$$

Για $x > 0$, έχουμε

$$f_Y(y) = \frac{1}{y+1} \int_0^{\infty} x(y+1)e^{-x(y+1)} dx = \frac{1}{(y+1)^2}, \quad y > 0. \text{ Το τελευταίο ολοκλήρωμα είναι ίσο με τη μέση τιμή}$$

της Εκθετικής κατανομής με παράμετρο $y+1$.

(β) Για $y > 0$, έχουμε $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = x(y+1)^2 e^{-x(y+1)}$, $x > 0$ και για $x > 0$, έχουμε

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = x e^{-xy}, \quad y > 0.$$

(γ) Για $y > 0$, αν $x \leq 0$, τότε $f_{X|Y}(s|y) = 0$, για κάθε $s \leq x$, οπότε $F_{X|Y}(x|y) = 0$, ενώ αν $x > 0$, τότε

$$F_{X|Y}(x|y) = (y+1)^2 \int_0^x s e^{-s(y+1)} ds = 1 - (1+x(y+1))e^{-x(y+1)}.$$

Για τον υπολογισμό της $F_{Y|X}(y|x)$, $x > 0$, παρατηρούμε ότι η δεσμευμένη κατανομή της Y δοθέντος ότι $X = x$, είναι η Εκθετική κατανομή με παράμετρο x . Άρα, $F_{Y|X}(y|x) = 0$, αν $y \leq 0$ και $F_{Y|X}(y|x) = 1 - e^{-xy}$, αν $y > 0$.

(δ) Για τη δεσμευμένη μέση τιμή $E(X|Y=y)$, $y > 0$, έχουμε

$$E(X|Y=y) = (y+1) \int_0^\infty x^2 (y+1) e^{-x(y+1)} dx = (y+1) \frac{2}{(y+1)^2} = \frac{2}{y+1}.$$

Το τελευταίο ολοκλήρωμα είναι ίσο με την ποσότητα $E(Z^2)$, όπου η τυχαία μεταβλητή Z ακολουθεί την Εκθετική κατανομή με παράμετρο $y+1$. Για τη δεσμευμένη μέση τιμή $E(Y|X=x)$, $x > 0$, έχουμε $E(Y|X=x) = \frac{1}{x}$.

Για την εύρεση της δεσμευμένης μέσης τιμής μιας συνάρτησης $g(X)$ της τυχαίας μεταβλητής X , δοθέντος ότι $Y = y$, ισχύουν οι εξής τύποι: $E[g(X)|Y=y] = \sum_x g(x) p_{X|Y}(x|y)$, αν οι X και Y είναι

διακριτές τυχαίες μεταβλητές και $E[g(X)|Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx$, αν οι X και Y είναι συνεχείς τυχαίες μεταβλητές.

Αν θέσουμε $E[X] = \mu$ και θεωρήσουμε τη συνάρτηση $h(x) = (x - \mu)^2$, $x \in \mathfrak{R}$, τότε, από τους παραπάνω τύπους, η **δεσμευμένη διακύμανση** (conditional variance) της X δοθέντος ότι $Y = y$, είναι

$$\text{Var}[X|Y=y] = E[h(X)|Y=y] \quad \text{και} \quad \text{δίνεται από τον τύπο} \\ \text{Var}[X|Y=y] = E[X^2|Y=y] - (E[X|Y=y])^2.$$

Συμβολίζουμε με $E[X | Y]$ τη συνάρτηση της τυχαίας μεταβλητής Y , της οποίας η τιμή όταν $Y = y$ είναι $E[X | Y = y]$. Παρατηρούμε ότι η ποσότητα $E[X | Y]$ είναι επίσης μία τυχαία μεταβλητή. Για τη συνάρτηση $Var(X | Y)$ ισχύει η σχέση $Var(X | Y) = E(X^2 | Y) - (E(X | Y))^2$. Η παρακάτω πρόταση είναι σημαντικότερη και έχει πολλές εφαρμογές.

Πρόταση. Έστω δύο τυχαίες μεταβλητές X, Y . Ισχύει ότι $E[X] = E[E[X | Y]]$. (5.1)

(α) Αν η Y είναι μία διακριτή τυχαία μεταβλητή τότε η σχέση (5.1) γράφεται ως εξής:

$$E[X] = \sum_y E[X | Y = y]P[Y = y].$$

(β) Αν η Y είναι μία συνεχής τυχαία μεταβλητή με συνάρτηση πυκνότητας f_Y τότε η σχέση (5.1)

$$\text{γράφεται ως εξής: } E[X] = \int_{-\infty}^{\infty} E[X | Y = y]f_Y(y)dy.$$

Στη συνέχεια θα αποδείξουμε την παραπάνω πρόταση για την περίπτωση κατά την οποία και οι δύο τυχαίες μεταβλητές X, Y είναι διακριτές.

Απόδειξη. Διαδοχικά, έχουμε

$$\begin{aligned} E[E[X | Y]] &= \sum_y E[X | Y = y]P[Y = y] = \sum_y \left[\sum_x xP[X = x | Y = y] \right] P[Y = y] \\ &= \sum_y \sum_x x \frac{P(X = x, Y = y)}{P(Y = y)} P(Y = y) = \sum_x \sum_y xP(X = x, Y = y) = \sum_x x \sum_y P(X = x, Y = y) \\ &= \sum_x xP(X = x) = E[X]. \blacksquare \end{aligned}$$

Η παραπάνω πρόταση ισχύει και όταν μία από τις τυχαίες μεταβλητές X, Y είναι διακριτή και η άλλη είναι συνεχής.

Παράδειγμα 27. Αν A είναι ένα ενδεχόμενο και Y είναι μία τυχαία μεταβλητή, τότε

$$P(A) = \sum_y P(A | Y = y)P(Y = y), \text{ αν η } Y \text{ είναι μία διακριτή τυχαία μεταβλητή και}$$

$$P(A) = \int_{-\infty}^{\infty} P(A | Y = y)f_Y(y)dy, \text{ αν η } Y \text{ είναι μία συνεχής τυχαία μεταβλητή με σ.π. } f_Y.$$

Λύση. Έστω η τυχαία μεταβλητή X τέτοια ώστε $X = 1$, αν το ενδεχόμενο A συμβαίνει και $X = 0$, αν το ενδεχόμενο A δεν συμβαίνει. Ισχύει ότι $P(A) = E[X]$ και $P(A|Y = y) = E[X|Y = y]$. Εφαρμόζοντας την πρόταση, έχουμε

$$P(A) = E[X] = E[E[X|Y]] = \begin{cases} \sum_y E[X|Y = y]P(Y = y) \\ \int_{-\infty}^{\infty} E[X|Y = y]f_Y(y)dy \end{cases} = \begin{cases} \sum_y P(A|Y = y)P(Y = y) \\ \int_{-\infty}^{\infty} P(A|Y = y)f_Y(y) \end{cases}.$$

Το άνω σκέλος των τελευταίων δύο ισοτήτων αντιστοιχεί στην περίπτωση που η Y είναι διακριτή και το κάτω σκέλος αντιστοιχεί στην περίπτωση που η Y είναι συνεχής με σ.π. f_Y .

Παράδειγμα 28. Οι τυχαίες μεταβλητές $X_1, \dots, X_N$ ακολουθούν τη διωνυμική κατανομή με παραμέτρους m , p και η τυχαία μεταβλητή N είναι ανεξάρτητη των $X_1, \dots, X_N$ και ακολουθεί την κατανομή Poisson με παράμετρο λ . Υπολογίστε τη μέση τιμή $E\left(\sum_{i=1}^N X_i\right)$.

Λύση. Ισχύει ότι $E\left(\sum_{i=1}^N X_i | N = n\right) = E\left(\sum_{i=1}^n X_i | N = n\right) = E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E[X_i] = nmp$.

Η δεύτερη ισότητα είναι συνέπεια της ανεξαρτησίας των τυχαίων μεταβλητών N και $X_i, i = 1, \dots, n$.

Άρα, $E\left(\sum_{i=1}^N X_i | N\right) = mpN$.

Συνεπώς, $E\left(\sum_{i=1}^N X_i\right) = E\left[E\left(\sum_{i=1}^N X_i | N\right)\right] = E(mpN) = mpE(N) = mp\lambda$. Η τελευταία ισότητα προκύπτει

διότι $N \sim \text{Poisson}(\lambda)$.

Παράδειγμα 29. Υποθέτουμε ότι ο αριθμός των ταξιδιωτών που φθάνουν σε ένα σιδηροδρομικό σταθμό μέχρι τη χρονική στιγμή t , ακολουθεί την κατανομή $\text{Poisson}(\lambda t)$, όπου λ είναι ο ρυθμός με τον οποίον οι ταξιδιώτες φθάνουν στο σιδηροδρομικό σταθμό. Αν η χρονική στιγμή άφιξης του τρένου στο σταθμό ακολουθεί την Ομοιόμορφη κατανομή στο διάστημα $(0, T)$ και είναι ανεξάρτητη των αφίξεων των ταξιδιωτών στο σταθμό, ποια είναι η μέση τιμή του αριθμού των ταξιδιωτών που επιβιβάζονται στο τρένο;

Λύση. Έστω η τυχαία μεταβλητή $N(t)$ που αναπαριστά τον αριθμό των αφίξεων των ταξιδιωτών στο σταθμό μέχρι τη χρονική στιγμή t , $t \geq 0$. Έστω ότι η τυχαία μεταβλητή Y αναπαριστά τη χρονική

στιγμή άφιξης του τρένου στο σταθμό. Ισχύει ότι $N(t) \sim \text{Poisson}(\lambda t)$ και $Y \sim \text{Ομοιόμορφη}(0, T)$. Ζητάμε τη μέση τιμή $E[N(Y)]$.

Ισχύει ότι $E[N(Y) | Y = t] = E[N(t) | Y = t] = E[N(t)] = \lambda t$. Η δεύτερη ισότητα είναι συνέπεια της ανεξαρτησίας των τυχαίων μεταβλητών Y και $N(t)$. Η τελευταία ισότητα προκύπτει διότι $N(t) \sim \text{Poisson}(\lambda t)$.

$$\text{Άρα, } E[N(Y) | Y] = \lambda Y. \text{ Συνεπώς, } E[N(Y)] = E[E[N(Y) | Y]] = E(\lambda Y) = \lambda E(Y) = \frac{\lambda T}{2}.$$

Η τελευταία ισότητα προκύπτει διότι $Y \sim \text{Ομοιόμορφη}(0, T)$.

Παράδειγμα 30. Ένας μεταλλωρύχος είναι παγιδευμένος σε ένα ορυχείο το οποίο έχει τρεις πόρτες. Η πρώτη πόρτα οδηγεί σε ένα τούνελ το οποίο τον απελευθερώνει μετά από πορεία τριών ωρών. Η δεύτερη πόρτα οδηγεί σε ένα τούνελ το οποίο τον ξαναφέρει στο ορυχείο μετά από πορεία πέντε ωρών. Η τρίτη πόρτα οδηγεί σε ένα τούνελ το οποίο τον ξαναφέρει στο ορυχείο μετά από πορεία επτά ωρών. Αν υποθέσουμε ότι πάντοτε ο μεταλλωρύχος διαλέγει κάποια από τις τρεις πόρτες με την ίδια πιθανότητα, ποιος είναι ο αναμενόμενος χρόνος μέχρι την απελευθέρωσή του;

Λύση. Έστω Y η τυχαία μεταβλητή που αναπαριστά το χρόνο μέχρι την απελευθέρωση και X η τυχαία μεταβλητή που αναπαριστά την πόρτα που διαλέγει αρχικά ο μεταλλωρύχος. Ζητάμε τη μέση τιμή $E[Y]$.

$$\text{Ισχύει ότι } E[Y] = E[Y | X = 1]P[X = 1] + E[Y | X = 2]P[X = 2] + E[Y | X = 3]P[X = 3]$$

$$= \frac{1}{3}(E[Y | X = 1] + E[Y | X = 2] + E[Y | X = 3]).$$

$$\text{Όμως, εξ υποθέσεως, ισχύει ότι } E[Y | X = 1] = 3, E[Y | X = 2] = 5 + E[Y], E[Y | X = 3] = 7 + E[Y].$$

$$\text{Άρα, } E[Y] = \frac{1}{3}(3 + 5 + E[Y] + 7 + E[Y]) \Rightarrow E[Y] = 15.$$