STATYSTYKA

Zmienna losowa dwuwymiarowa

Definicja

Zmienna losowa dwuwymiarowa jest połączeniem dwóch zmiennych losowych jednowymiarowych określonych na tym samym zbiorze Ω.

Funkcja rozkładu prawdopodobieństwa (gęstości) takiej zmiennej jest dwuargumentową funkcją zdefiniowaną następująco:

$$f(x,y) = p_{ij} = P((X = x_i) \wedge (Y = y_j))$$

dla zmiennej losowej skokowej

$$f(x,y) = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{P(x < X < x + \Delta x, y < Y < y + \Delta y)}{\Delta x \, \Delta y}$$

dla zmiennej losowej ciągłej

Rozkłady brzegowe

czyli rozkłady każdej zmiennej są określone przez poniższe funkcje jednoargumentowe:

$$f(x) = P(X = x_i) = p_i = \sum_j p_{ij}$$

dla zmiennej losowej skokowej

$$f(y) = P(Y = y_j) = p_j = \sum_i p_{ij}$$

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

dla zmiennej losowej ciągłej

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

Zmienne losowe dwuwymiarowe

X\Y	**2**	**3**	**4**	**5**
3	0,12	0,09	0,04	0
4	0,06	0,08	0,12	0,10
5	0,04	0,05	0,10	0,20

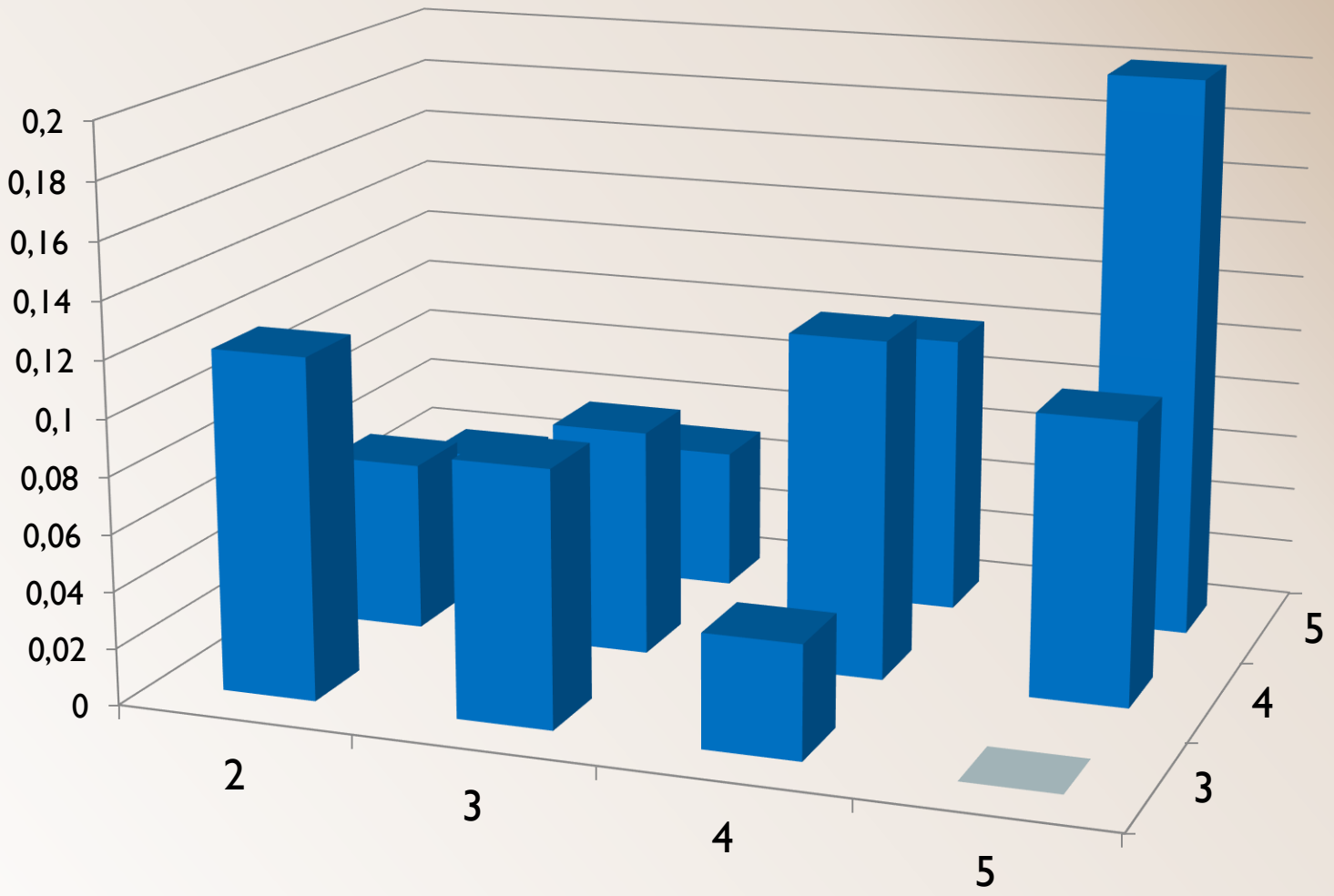
Rozkład brzegowy zmiennej losowej X:

x_i	**3**	**4**	**5**
$P(X=x_i)$	0,25	0,36	0,39

Rozkład brzegowy zmiennej losowej Y:

y_j	**2**	**3**	**4**	**5**
$P(Y=y_j)$	0,22	0,22	0,26	0,30

Zmienne losowe dwuwymiarowe



Rozkłady warunkowe

Rozkład warunkowy zmiennej Y (pod warunkiem, że X = x_i) określimy wzorem:

$$P(Y = y_j / X = x_i) = \frac{p_{ij}}{p_i}$$

dla zmiennej losowej skokowej,

$$f(y/x) = \frac{f(x,y)}{f(x)}$$

dla zmiennej losowej ciągłej.

Rozkłady warunkowe

Rozkład zmiennej losowej XY

$x_i \setminus y_j$	3	4	5	p_i
1	0,08	0,03	0,08	0,19
2	0,10	0,10	0,30	0,50
4	0,15	0,06	0,10	0,31
p_j	0,33	0,19	0,48	

Rozkłady warunkowe Y/X

$x_i \setminus y_j$	3	4	5	suma
1	0,42	0,16	0,42	1,00
2	0,20	0,20	0,60	1,00
4	0,48	0,19	0,32	1,00

Rozkłady warunkowe X/Y

$x_i \setminus y_j$	3	4	5	
1	0,24	0,16	0,17	
2	0,30	0,53	0,63	
4	0,45	0,32	0,21	
suma	1,00	1,00	1,00	

Parametry

Parametrami dwuwymiarowej zmiennej losowej są wartości oczekiwane obydwu zmiennych (EX i EY), ich wariancje (D^2X i D^2Y) obliczane z funkcji rozkładów brzegowych.

Miarą charakteryzującą współzmienność X i Y jest kowariancja (CXY) :

$$CXY = \sum_{i,j}(x_i - EX)\cdot(y_j - EY)\cdot p_{ij} = \sum_{i,j} x_i y_j p_{ij} - EX \cdot EY$$

dla zmiennej losowej skokowej

$$CXY = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x - EX)\cdot(y - EY)\cdot f(x,y)dxdy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy \cdot f(x,y)dxdy - EX \cdot EY$$

dla zmiennej losowej ciągłej

Rodzaje zależności

Jeśli zmienne są niezależne to kowariancja między nimi ma wartość zero.
Odwrotne twierdzenie jest nieprawdziwe, tzn. jeśli kowariancja jest równa zero, to nie oznacza, że zmienne są niezależne .

Wyróżniamy dwa rodzaje niezależności

- <u>Stochastyczną</u> (rozkłady warunkowe jednej zmiennej na każdym poziomie drugiej są takie same) inaczej

$$f(x, y) = f(x) \cdot f(y)$$

- <u>Korelacyjną</u> (wartości oczekiwane warunkowych rozkładów są jednakowe) $EX/Y_1=EX/Y_2=.....$

Niezależność stochastyczna

X\Y	0	2	3	p_i
1	0,08	0,20	0,12	0,40
3	0,12	0,30	0,18	0,60
p_j	0,20	0,50	0,30	1,00

$$EX = 1 \cdot 0{,}4 + 3 \cdot 0{,}6 = 0{,}4 + 1{,}8 = 2{,}2$$

$$EY = 0 \cdot 0{,}2 + 2 \cdot 0{,}5 + 3 \cdot 0{,}3 = 0 + 1 + 0{,}9 = 1{,}9$$

$$CXY = (1 \cdot 0 \cdot 0{,}08 + 1 \cdot 2 \cdot 0{,}2 + 1 \cdot 3 \cdot 0{,}12 + 3 \cdot 0 \cdot 0{,}12 + 3 \cdot 2 \cdot 0{,}3 + 3 \cdot 3 \cdot 0{,}18) - 2{,}2 \cdot 1{,}9 =$$

$$= (0 + 0{,}4 + 0{,}36 + 0 + 1{,}8 + 1{,}62) - 2{,}2 \cdot 1{,}9 = 4{,}18 - 4{,}18 = 0$$

Niezależność stochastyczna

X\Y	0	2	3	suma
1	0,08/0,40 =**0,20**	0,20/0,40 =**0,50**	0,12/0,40 =**0,30**	1
3	0,12/0,60 =**0.20**	0,30/0,60 =**0,50**	0,18/0,60 =**0,30**	1

EY/X=1 = 1,9
EY/X=3 = 1.9

X\Y	0	2	3	
1	0,08/0,20 =**0,40**	0,20/0,50 =**0,40**	0,12/0,30 =**0,40**	
3	0,12/0,20 =**0.60**	0,30/0,50 =**0,60**	0,18/0,30 =**0,60**	
suma	1	1	1	

EX/Y=0 = 2,2
EX/Y=2 = 2,2
EX/Y=3 = 2,2

Np.
P(X=1,Y=3)=
0,4*0,3=0,12

Korelacja

Kowariancja jest miarą zależności korelacyjnej (typu liniowego) zmiennych losowych i może przyjmować wartości z przedziału $\langle -DX \cdot DY \; ; \; +DX \cdot DY \rangle$. Znak kowariancji informuje o kierunku zależności.

Ponieważ CXY jest wielkością mianowaną, zależną od jednostek X i Y, to jako miarę współzależności liniowej dwóch zmiennych stosuje się niemianowany współczynnik korelacji liniowej:

$$\rho_{XY} = \frac{CXY}{DX \cdot DY}$$

Współczynnik korelacji liniowej przyjmuje wartości od –1 do 1. Znak informuje o kierunku zależności, a wartość o jej sile. Gdy $\rho_{XY} = 0$ to zmienne są niezależne.

Niezależność korelacyjna

X\Y	0	2	4	p_i
1	0,2	0,1	0,2	0,5
3	0,1	0,3	0,1	0,5
p_j	0,3	0,4	0,3	1

$$EX = 1 \cdot 0{,}5 + 3 \cdot 0{,}5 = 0{,}5 + 1{,}5 = 2{,}0$$

$$EY = 0 \cdot 0{,}3 + 2 \cdot 0{,}4 + 4 \cdot 0{,}3 = 0{,}8 + 1{,}2 = 2{,}0$$

$$CXY = (1 \cdot 0 \cdot 0{,}2 + 1 \cdot 2 \cdot 0{,}1 + 1 \cdot 4 \cdot 0{,}2 + 3 \cdot 0 \cdot 0{,}1 + 3 \cdot 2 \cdot 0{,}3 + 3 \cdot 4 \cdot 0{,}1) - 2{,}0 \cdot 2{,}0 =$$

$$= (0{,}0 + 0{,}2 + 0{,}8 + 0{,}0 + 1{,}8 + 1{,}2) - 4{,}0 = 4{,}0 - 4{,}0 = 0$$

Niezależność korelacyjna

X\Y	0	2	4	suma
1	0,2/0,5 **=0,4**	0,1/0,5 **=0,2**	0,2/0,5 **=0,4**	1
3	0,1/0,5 **=0,2**	0,3/0,5 **=0,6**	0,1/0,5 **=0,2**	1
pj	0,3	0,4	0,3	

EY/X=1 = 2
EY/X=3 = 2

Np.
P(X=1,Y=4)≠0,5*0,3

Zależność jednej zmiennej od drugiej

Można rozpatrywać funkcyjną zależność jednej zmiennej od drugiej. Wtedy jedna ze zmiennych jest zmienną niezależną (argument funkcji), a druga – zmienną zależną (wartością funkcji). Przykładem takich funkcji są regresja.

Regresją I rodzaju jest funkcja przyporządkowująca wartościom zmiennej niezależnej warunkowe wartości oczekiwane zmiennej zależnej. Jest to funkcja postaci:

$m(x) = E(Y/X=x_i)$ zależność zmiennej Y od X

$m(y) = E(X/Y=y_i)$ zależność zmiennej X od Y

Zależność jednej zmiennej od drugiej

Regresja II rodzaju – funkcja przeprowadzona według metody najmniejszych kwadratów – najczęściej funkcja liniowa postaci:

$$g(y) = \beta_{XY} \cdot y + \alpha_{XY}$$ opisująca zależność Y od X

$$g(x) = \beta_{YX} \cdot x + \alpha_{YX}$$ opisująca zależność X od Y

Współczynnik β_{YX} nazywany jest współczynnikiem regresji liniowej i wyraża wielkość zmiany zmiennej zależnej (Y) przy wzroście zmiennej niezależnej (X) o jednostkę.
Współczynnik α_{YX} nosi nazwę stałej regresji.
Wykresem tej funkcji jest linia prosta.

Współczynniki i stałe regresji obliczane są według poniższych wzorów:

$$\beta_{YX} = \frac{CXY}{D^2 X} \qquad \alpha_{YX} = EY - \beta_{YX} \cdot EX$$

$$\beta_{XY} = \frac{CXY}{D^2 Y} \qquad \alpha_{XY} = EX - \beta_{XY} \cdot EY$$

Współczynnik determinacji

Miarą informującą jaką część wariancji zmiennej zależnej można wyjaśnić przy pomocy danej funkcji regresji jest współczynnik determinacji.

Dla funkcji regresji I rodzaju współczynnik ten oznaczany jest przez ε^2_{YX} oraz ε^2_{XY} (zależnie od kierunku zależności),

a dla liniowej regresji II rodzaju przez $\rho^2_{XY.}$

Dla regresji II rodzaju współczynnik determinacji jest taki sam dla każdego kierunku zależności :

$$\rho^2_{XY} = (\rho_{XY})^2 = \beta_{XY} \cdot \beta_{YX}$$

Współczynnik determinacji

Suma wariancji warunkowych wartości oczekiwanych i wartości oczekiwanej wariancji warunkowych rozkładów dla każdego kierunku zależności jest równa wariancji zmiennej zależnej, czyli:

$$D^2Y = D^2(E(Y/X)) + E(D^2(Y/X))$$

Lub

$$D^2X = D^2(E(X/Y)) + E(D^2(X/Y))$$

Współczynnik determinacji

Współczynniki te obliczane są według wzorów:

$$\varepsilon_{YX}^2 = \frac{D^2\left(E\left(Y/X\right)\right)}{D^2Y} = \frac{D^2Y - E\left(D^2\left(Y/X\right)\right)}{D^2Y}$$

$$\varepsilon_{XY}^2 = \frac{D^2\left(E\left(X/Y\right)\right)}{D^2X} = \frac{D^2X - E\left(D^2\left(X/Y\right)\right)}{D^2X}$$

gdzie $D^2(E(Y/X))$ i $D^2(E(X/Y))$ to wariancje warunkowych wartości oczekiwanych, a $E(D^2(Y/X))$ i $E(D^2(X/Y))$ są wartościami oczekiwanymi wariancji rozkładów warunkowych.

$$D^2\left(E\left(Y/X\right)\right) = \sum_i \left(E\left(Y/X = x_i\right) - EY\right)^2 \cdot P(X = x_i) = \sum_i \left(E\left(Y/X = x_i\right)\right)^2 \cdot P(X = x_i) - (EY)^2$$

$$E\left(D^2\left(Y/X\right)\right) = \sum_i D^2\left(Y/X = x_i\right) \cdot P(X = x_i)$$

Rozkład zmiennej losowej XY

$x_i \setminus y_j$	3	4	5	p_i
1	0,08	0,03	0,08	0,19
2	0,10	0,10	0,30	0,50
4	0,15	0,06	0,10	0,31
p_j	0,33	0,19	0,48	

Prawdopodobieństwa warunkowe P(Y/X)

$x_i \setminus y_j$	3	4	5	suma
1	0,42	0,16	0,42	1,00
2	0,20	0,20	0,60	1,00
4	0,48	0,19	0,32	1,00

E(Y/X)	D^2(Y/X)
4,00	0,84
4,40	0,64
3,84	0,78

EY	D^2Y
4,15	0,7875

0,0656

0,7219

Prawdopodobieństwa warunkowe P(X/Y)

$x_i \setminus y_j$	3	4	5	
1	0,24	0,16	0,17	
2	0,30	0,53	0,63	
4	0,45	0,32	0,21	
suma	1,00	1,00	1,00	

E(X/Y)	2,67	2,47	2,25
D^2(X/Y)	1,62	1,20	0,94

EX	2,43
D^2X	1,25

0,0344

1,2107

COVXY	-0,164
ρ^2	0,0276

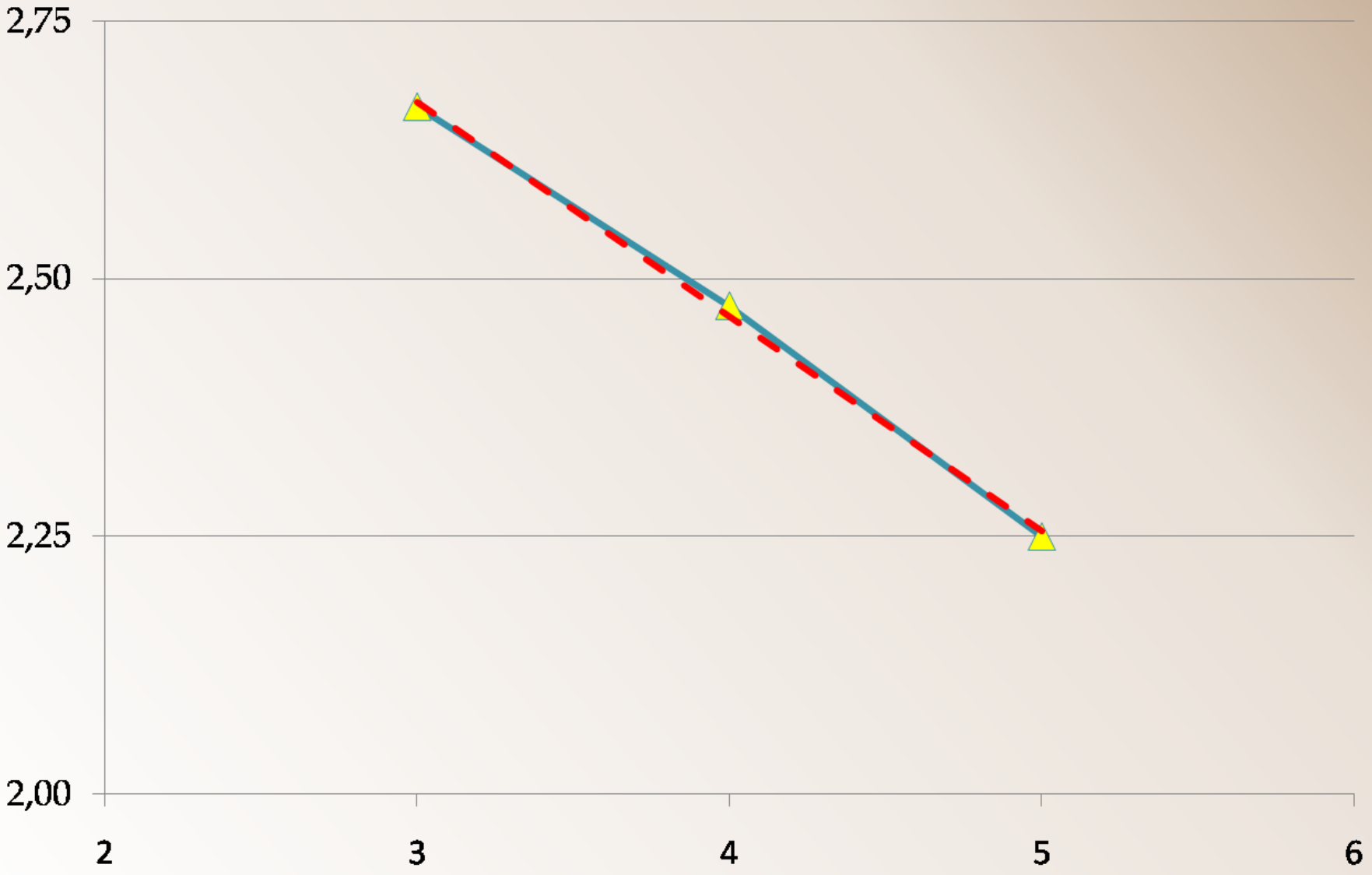
ε^2y/x

0,08326

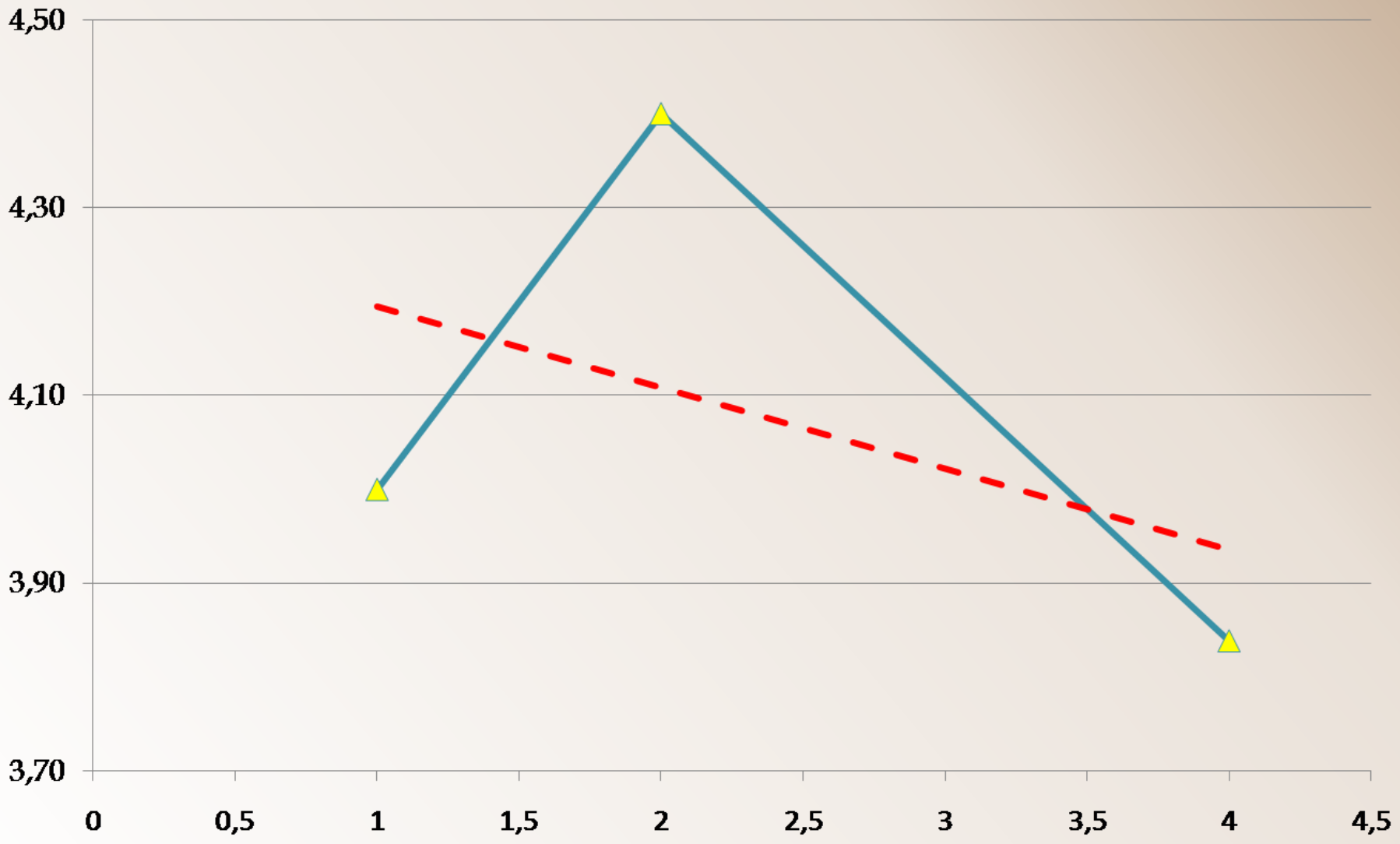
ε^2x/y

0,02763

Zależność X od Y



Zależność Y od X



Tworzenie zmiennych losowych złożonych poprzez liniowe kombinacje zmiennych

Zmienne losowe można ze sobą łączyć, tzn. dodawać, odejmować, mnożyć przez liczby.
Połączenie zmiennych, w którym występują one zawsze w pierwszej potędze, nazywamy liniowymi.
Znając parametry zmiennych składowych można obliczać parametry zmiennych złożonych.

Jeżeli zmienna V jest liniową kombinacją zmiennych X_1, X_2..., czyli:

$$V = \sum_{i=1}^{k} a_i \cdot X_i$$

to wartość oczekiwana zmiennej złożonej jest złożeniem wartości oczekiwanych według tej samej kombinacji:

$$EV = \sum_{i=1}^{k} a_i EX_i$$

Tworzenie zmiennych losowych złożonych poprzez liniowe kombinacje zmiennych

a wariancja zmiennej złożonej oprócz wariancji zmiennych prostych uwzględnia powiązania między każdą parą zmiennych prostych (kowariancje):

$$D^2V = \sum_{i=1}^{k} a_i^2 \cdot D^2 X_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} 2 \cdot a_i \cdot a_j \cdot CX_i X_j$$

Kowariancję zmiennej złożonej ze zmienną składową X_m obliczamy następująco:

$$CVX_m = \sum_{i=1}^{k} a_i CX_i X_m \quad \text{przy czym} \quad CX_m X_m = D^2 X_m$$

Przykład

Dane są zmienne losowe X oraz Y o parametrach: EX=2, $D^2X=5$, EY=3, $D^2Y=7$, CXY=4.

Zmienne Z i W są kombinacjami liniowymi postaci:

Z = 3X + 2Y

W = 4X – 3Y

$$EZ = 3 \cdot EX + 2 \cdot EY = 3 \cdot 2 + 2 \cdot 3 = 12$$

$$EW = 4 \cdot EX - 3 \cdot EY = 4 \cdot 2 - 3 \cdot 3 = -1$$

$$D^2Z = 3^2 D^2X + 2^2 D^2Y + 2 \cdot 3 \cdot 2 \cdot CXY = 9 \cdot 5 + 4 \cdot 7 + 12 \cdot 4 = 121$$

$$D^2W = 4^2 D^2X + 3^2 D^2Y - 2 \cdot 4 \cdot 3 \cdot CXY = 16 \cdot 5 + 9 \cdot 7 - 24 \cdot 4 = 47$$

$$CZW = 3 \cdot 4 \cdot D^2X + 3 \cdot (-3) \cdot CXY + 2 \cdot 4 \cdot CXY + 2 \cdot (-3) \cdot D^2Y = 12 \cdot 5 - 9 \cdot 4 + 8 \cdot 4 - 6 \cdot 7 = 14$$

kowariancje zmiennej złożonej ze składowymi X i Y:

$$CXZ = 3 \cdot D^2X + 2 \cdot CXY = 3 \cdot 5 + 2 \cdot 4 = 23$$

$$CXW = 4 \cdot D^2X - 3 \cdot CXY = 4 \cdot 5 - 3 \cdot 4 = 8$$

$$CYZ = 3 \cdot CXY + 2 \cdot D^2Y = 3 \cdot 4 + 2 \cdot 7 = 26$$

$$CYW = 4 \cdot CXY - 3 \cdot D^2Y = 4 \cdot 4 - 3 \cdot 7 = -5$$

Ursus arctos