Komputerowa Analiza Danych Doświadczalnych

Prowadząca:
dr inż. Hanna Zbroszczyk

e-mail: *gos@if.pw.edu.pl*
Tel: +48 22 234 58 51

Konsultacje:
Poniedziałek: 10.15-11.00
Środa: 11.15-12.00

www: http://www.if.pw.edu.pl/~gos/students/kadd

Politechnika Warszawska
Wydział Fizyki
Pok. 117b (wejście przez 115)

Zmienne losowe dwuwymiarowe

Zmienne losowe dwuwymiarowe – f(x,y), F(x,y)

Mamy 2 zmienne losowe *x* oraz *y*, które spełniają warunek: $x < \mathrm{x}$ oraz $y < \mathrm{y}$

Dystrybuanta: $F(x,y) = P(x < \mathrm{x},\ y < \mathrm{y})$

Jeśli dystrybuanta jest funkcją ciągłych zmiennych x oraz y:

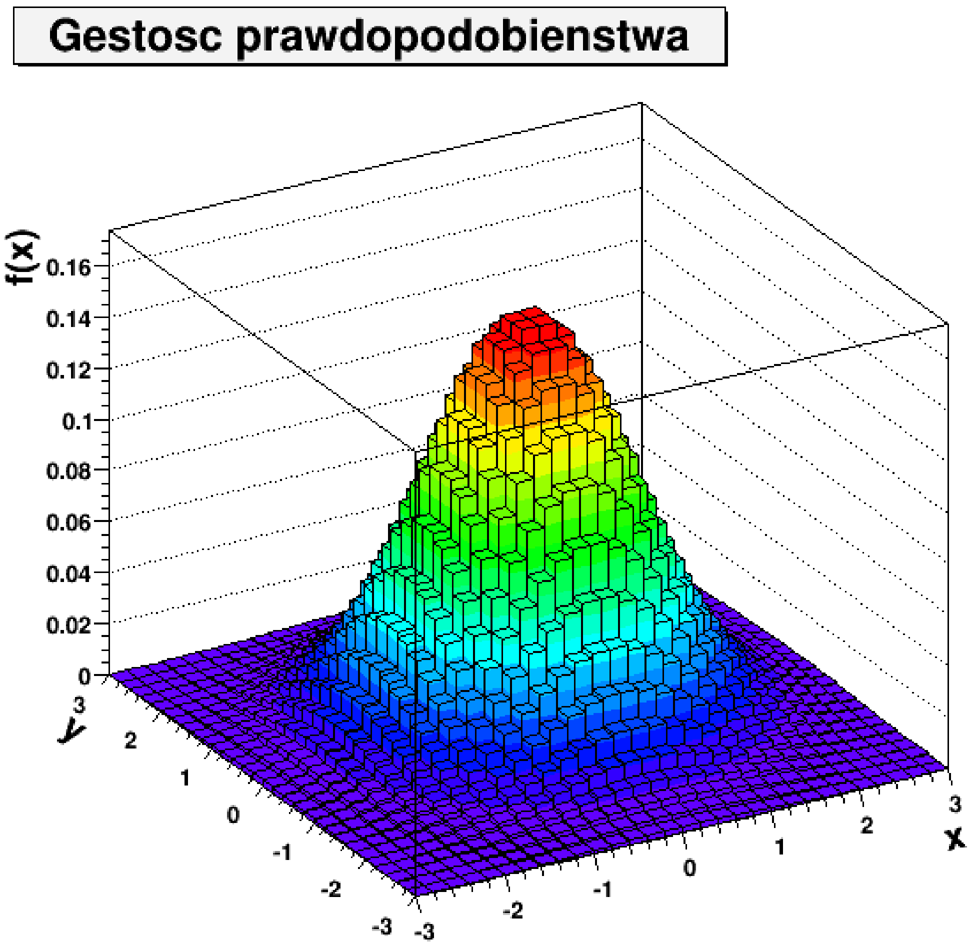
Łączna gęstość prawdopodobieństwa zmiennych losowych *x*, *y*:

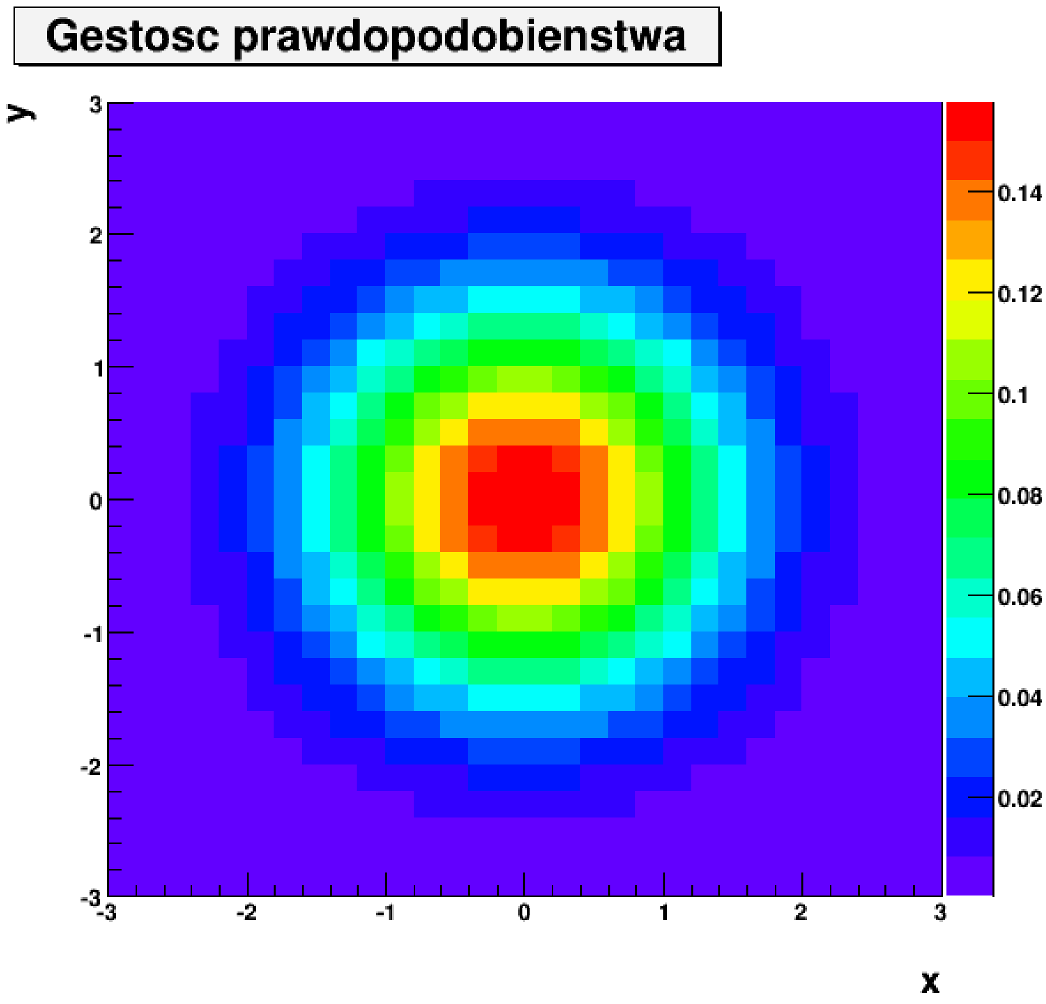
$$f(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y)$$

Prawdopodobieństwo:

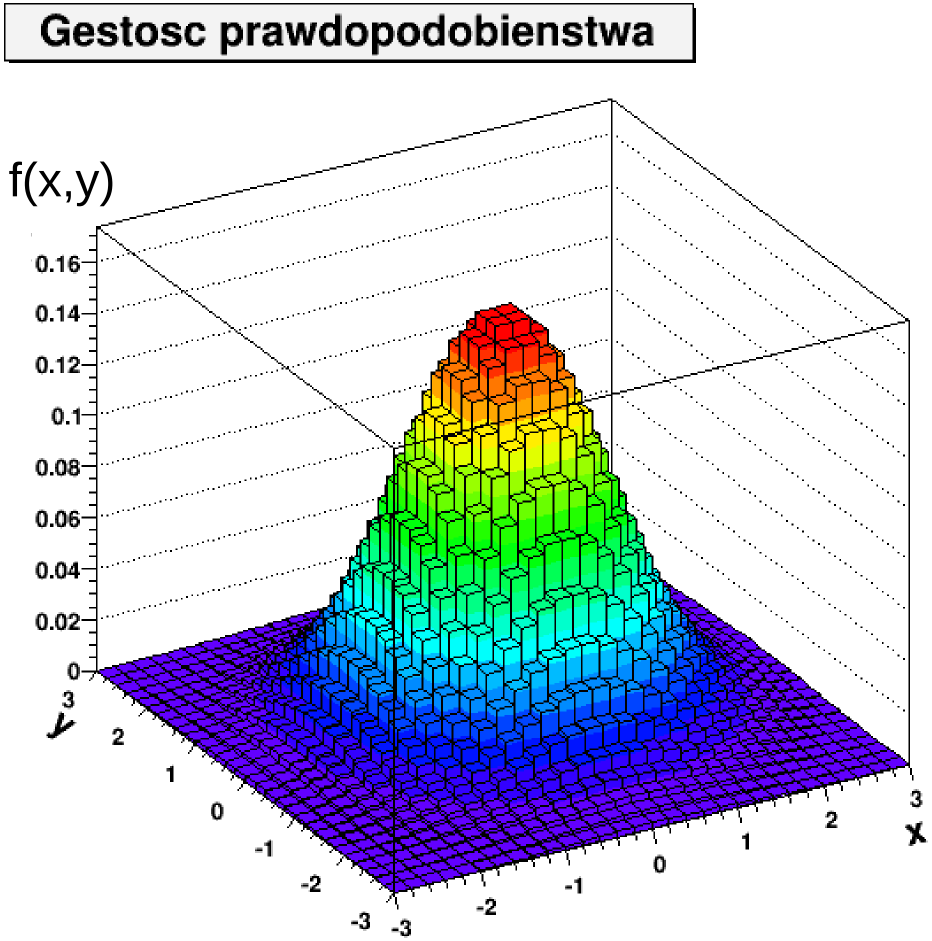
$$P(a \leqslant x < b, c \leqslant y < d) = \int_a^b \left[\int_c^d f(x,y)\,dy\right]dx$$

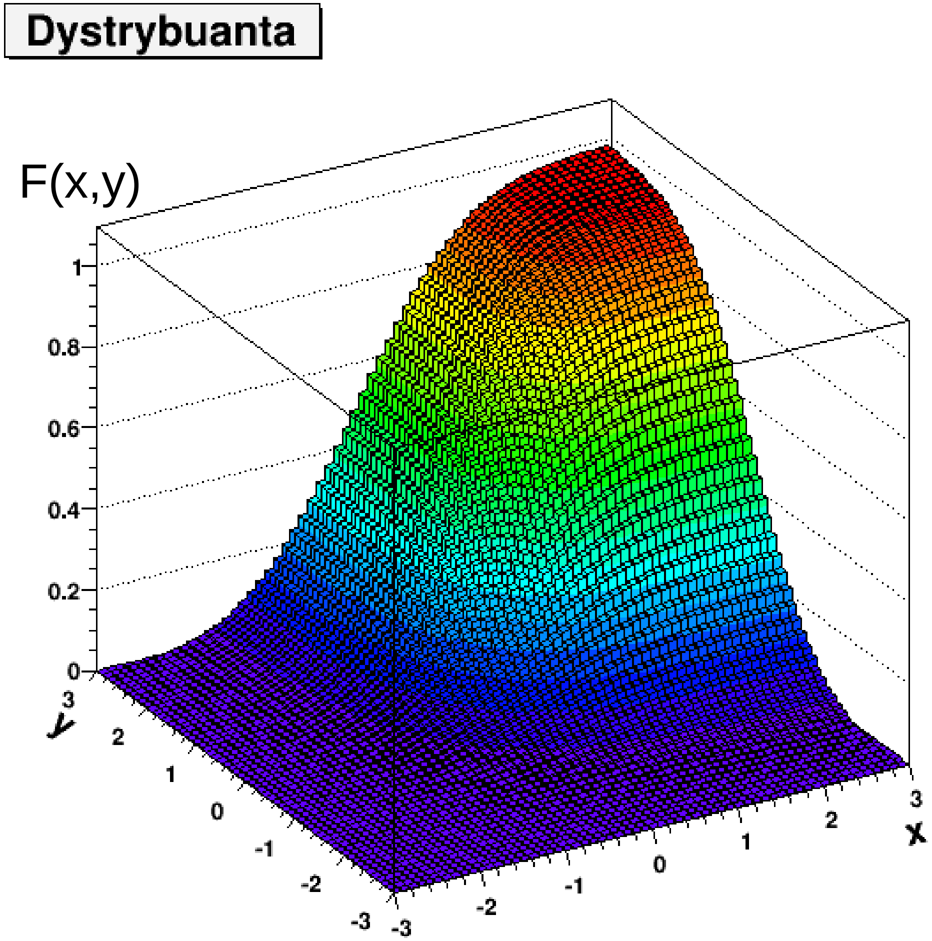
Zmienne losowe dwuwymiarowe – f(x,y)





Zmienne losowe dwuwymiarowe – f(x,y), F(x,y)





Zmienne losowe dwuwymiarowe – gęstości brzegowe

Mierzone są 2 zmienne losowe: *x*, *y*.

Możliwe jest doświadczalne wyznaczenie dystrybuanty F(x,y).

Jaka jest zależność zmiennej *x* dla ustalonego *y*?

$$P(a \leqslant x < b, -\infty < y < \infty) = \int_a^b \left[\int_{-\infty}^{\infty} f(x,y)\,dy\right] dx = \int_a^b g(x)\,dx$$

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

g(x) – **brzegowa gęstość prawdopodobieństwa** dla zmiennej *x*.

Analogicznie:

$$P(-\infty < x < \infty, c < y < d) = \int_c^d \left[\int_{-\infty}^{\infty} f(x,y)\,dx\right] dy = \int_c^d h(y)\,dy$$

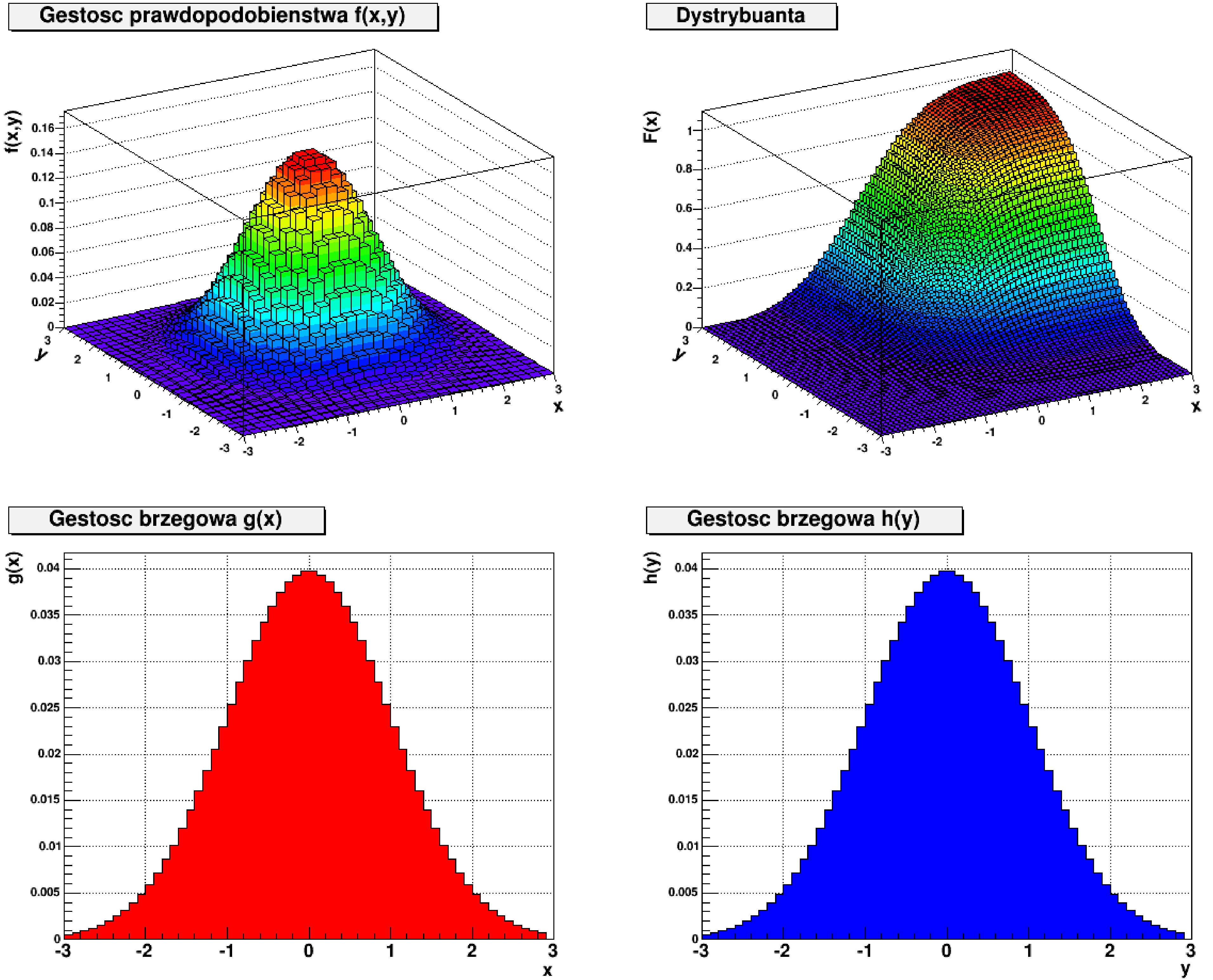
$$h(y)\,dy = \int_{-\infty}^{\infty} f(x,y)\,dx$$

h(y) – **brzegowa gęstość prawdopodobieństwa** dla zmiennej *y*.

Dystrybuanty brzegowe.

$$F_X(x) = \lim_{y \to \infty} F(x,y)$$

$$F_Y(x) = \lim_{x \to \infty} F(x,y)$$



Gestosc prawdopodobienstwa f(x,y)
f(x,y)
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
y
3
2
1
0
-1
-2
-3
x
-3
-2
-1
0
1
2
3
Dystrybuanta
F(x)
1
0.8
0.6
0.4
0.2
0
y
3
2
1
0
-1
-2
-3
x
-3
-2
-1
0
1
2
3
Gestosc brzegowa g(x)
g(x)
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
-3
-2
-1
0
1
2
3
x
Gestosc brzegowa h(y)
h(y)
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
-3
-2
-1
0
1
2
3
y

Zmienne losowe dwuwymiarowe – niezależność zmiennych

Niezależność zmiennych losowych:

$$f(x,y)=g(x)h(y)$$

Prawdopodobieństwo warunkowe dla zmiennej *y* przy znanej wartości *x*:

$$P(y \leqslant \mathit{y} < y+dy | x \leqslant \mathit{x} < x+dx)$$

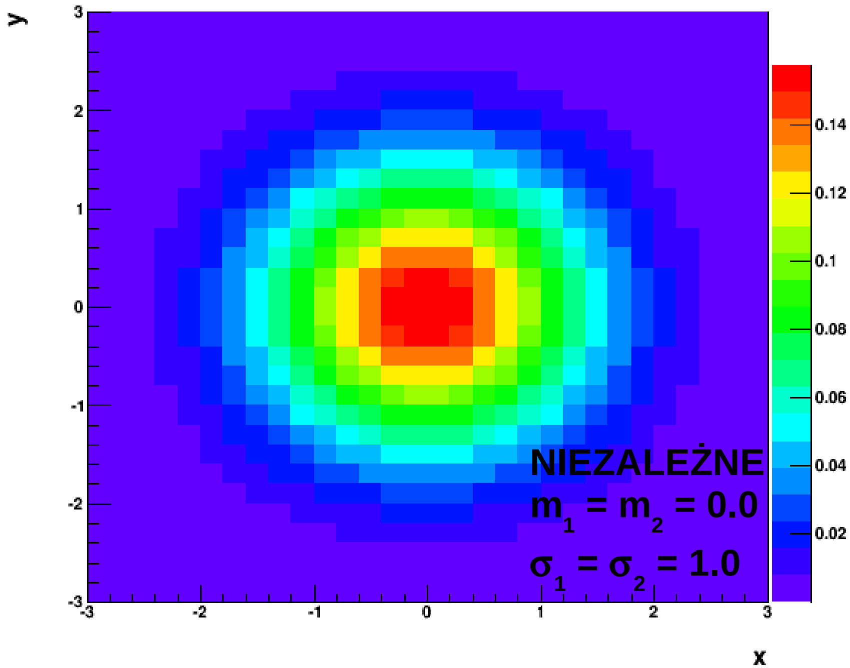
Gęstość prawdopodobieństwa warunkowego:

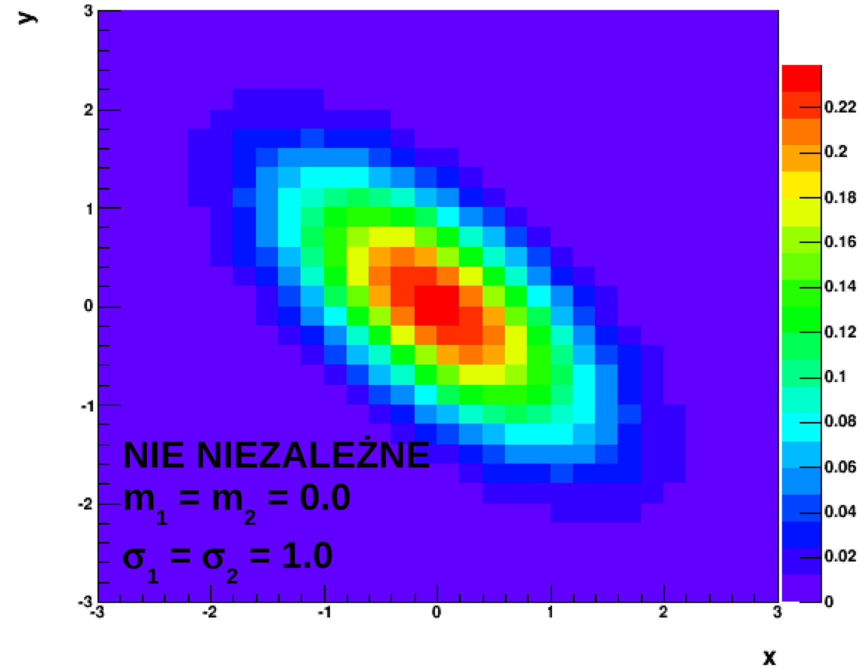
$$f(x|y)=\frac{f(x,y)}{g(x)}$$

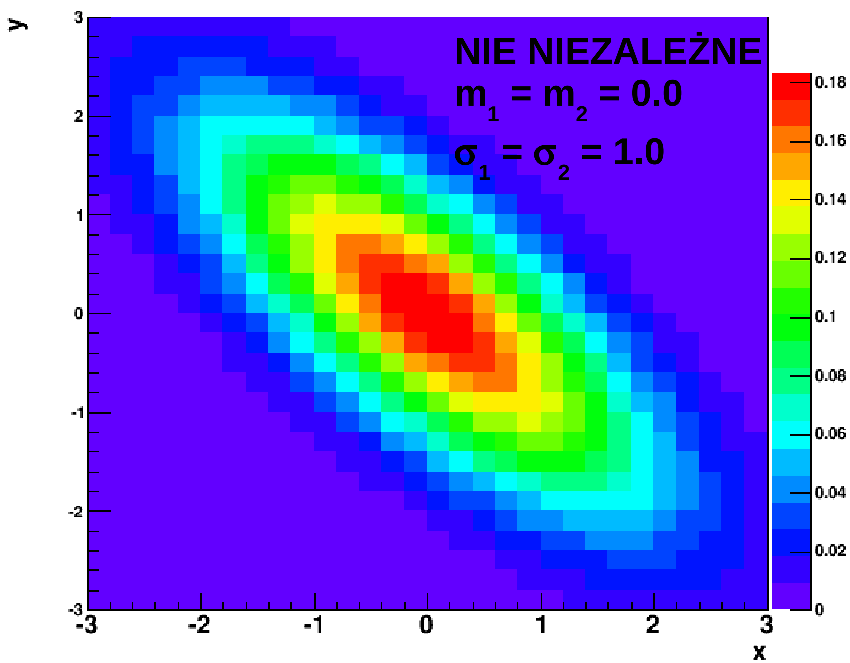
Prawdopodobieństwo warunkowe:

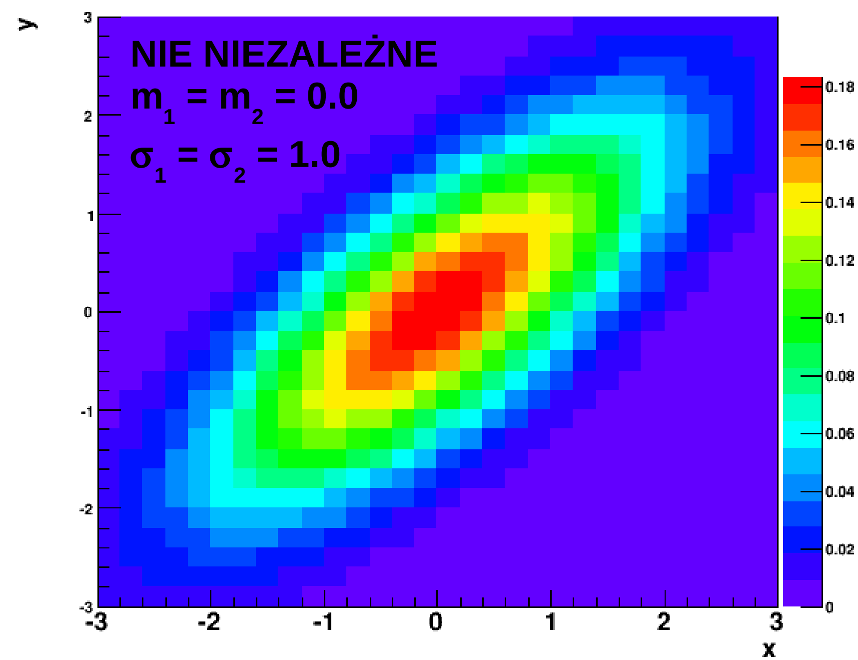
$$P(y \leqslant \mathit{y} < y+dy | x \leqslant \mathit{x} < x+dx)=f(x|y)dy$$

Zmienne losowe dwuwymiarowe – niezależność zmiennych









Gęstość prawdopodobieństwa warunkowego

Prawdopodobieństwo całkowite:

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

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

Dla zmiennych niezależnych:

$$f(y|x)=\frac{f(x,y)}{g(x)}=\frac{g(x)h(y)}{g(x)}=h(y)$$

$$f(x|y)=\frac{f(x,y)}{h(y)}=\frac{g(x)h(y)}{h(y)}=g(x)$$

Warunek narzucony na jedną zmienną nie może mieć wpływu na drugą zmienną.

Wartości oczekiwane rozkładu dwuwymiarowego

Wartość oczekiwana funkcji H(*x*,*y*) zmiennych losowych *x*, *y*:

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

Wariancja:

$$\sigma^2\{H(x,y)\}=E\{[H(x,y)-E(H(x,y))]^2\}$$

Jeśli:

$$H(x,y)=ax+by$$

To:

$$E(ax+by)=aE(x)+bE(y)$$

Wartości oczekiwane rozkładu dwuwymiarowego - momenty

Wartości oczekiwane dla funkcji iloczynowych:

$$H(x, y) = x^l y^m \qquad l, m - \text{liczny całkowite nieujemne}$$

to **momentami rzędu l oraz m względem *x* oraz *y***.

$$\lambda_{lm} = E(x^l y^m)$$

Uogólniając:

$$H(x, y) = (x-a)^l (y-b)^m$$

Wartości oczekiwane:

$$\alpha_{lm} = E\{(x-a)^l (y-b)^m\}$$

To **momenty rzędu l oraz m względem punktów a oraz b**.

Momenty rozkładów dwuwymiarowych, kowariancja

Najważniejsze momenty:

$$\mu_{lm} = E\{(x-\lambda_{10})^l (y-\lambda_{01})^m\}$$

$$\mu_{00} = 1$$

$$\mu_{11} = 0$$

$$\lambda_{10} = E(x) = \hat{x}$$

$$\lambda_{01} = E(y) = \hat{y}$$

$$\mu_{11} = E\{(x-\hat{x})(y-\hat{y})\} = \mathrm{cov}(x, y)$$

$$\mu_{20} = E\{(x-\hat{x})^2\} = \sigma^2(x)$$

$$\mu_{02} = E\{(y-\hat{y})^2\} = \sigma^2(y)$$

Wariancja, wartość oczekiwana rozkładu dwuwymiarowego

Wariancja rozkładu a*x* + b*y*:

$$\sigma^2(a x+b y)=E\{[(a x+b y)-E(a x+b y)]^2\}=E\{[a(x-\hat{x})+b(y-\hat{y})]^2\}$$

$$\sigma^2(a x+b y)=E\{a^2(x-\hat{x})^2+b^2(y-\hat{y})+2ab(x-\hat{x})(y-\hat{y})\}$$

$$\sigma^2(a x+b y)=a^2\sigma^2(x)+b^2\sigma^2(y)+2\,ab\,\mathrm{cov}(x,y)$$

Kiedy funkcja H(*x*,*y*) jest postaci iloczynowej:

$$H(x,y)=x\,y$$

W celu wyznaczenia wartości oczekiwanej, zakładamy niezależność zmiennych *x* oraz *y*:

$$E(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,g(x)\,h(y)\,dx\,dy=\left(\int_{-\infty}^{\infty} x\,g(x)\,dx\right)\left(\int_{-\infty}^{\infty} y\,h(y)\,dy\right)$$

Czyli zachodzi związek:

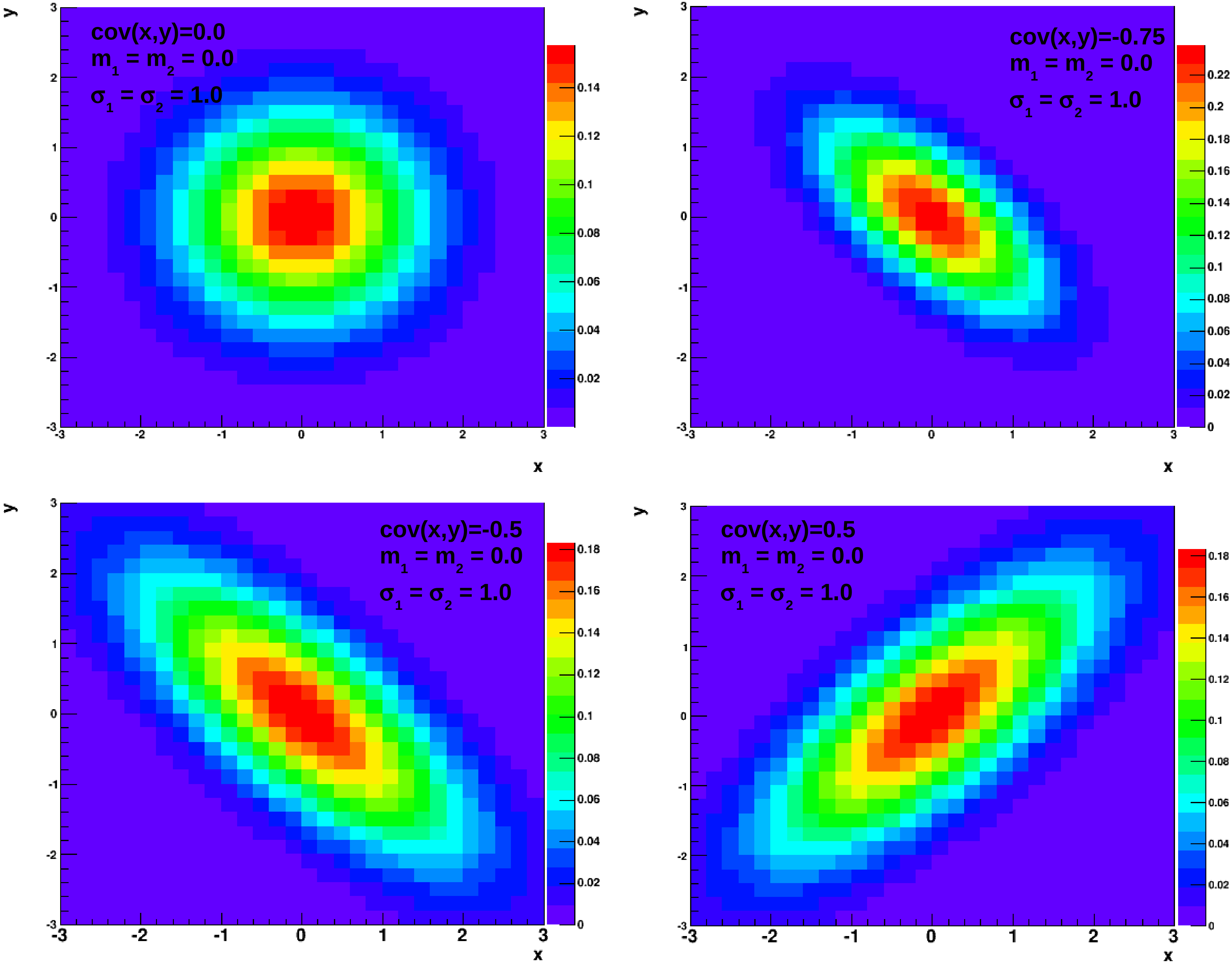
$$E(x,y)=E(x)\,E(y)$$

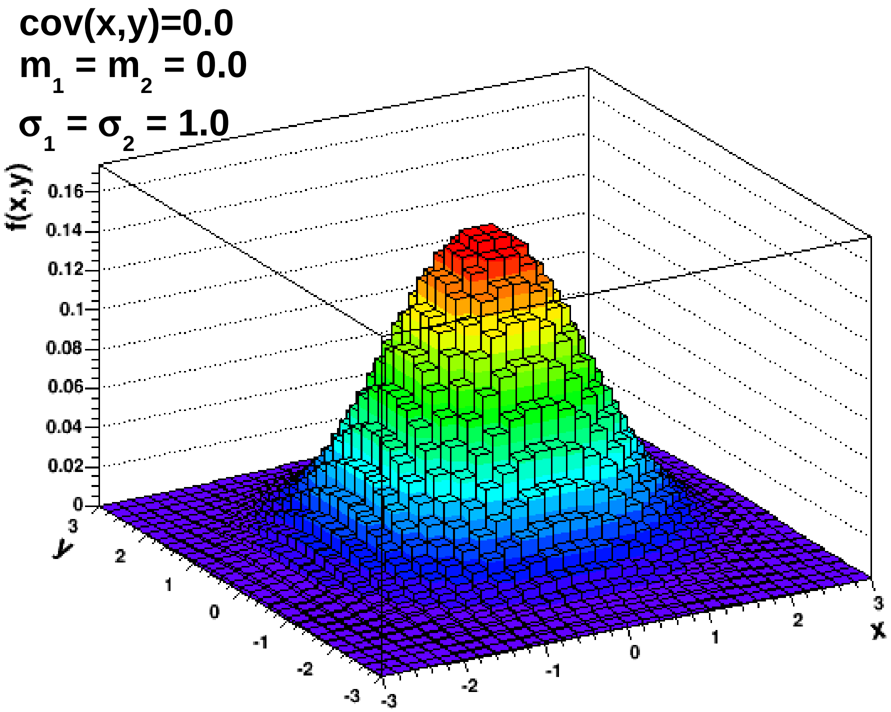
Charakterystyki rozkładu dwuwymiarowego

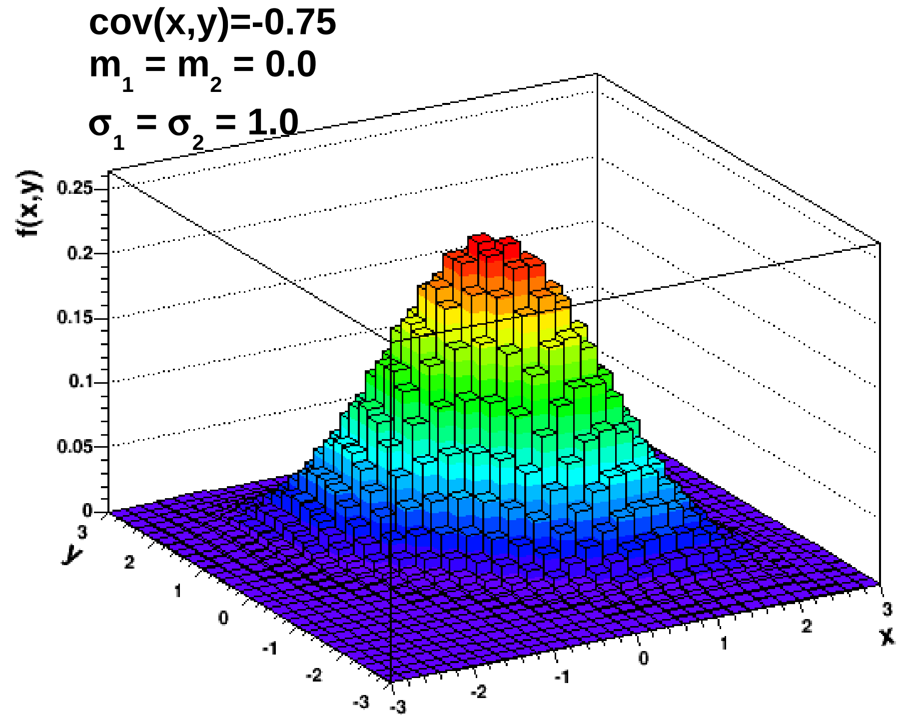
Wielkości: E(x), E(*y*), $\sigma^2(x)$, $\sigma^2(y)$ są podobne do analogicznych wielkości zdefiniowanych w przypadku jednej zmiennej.

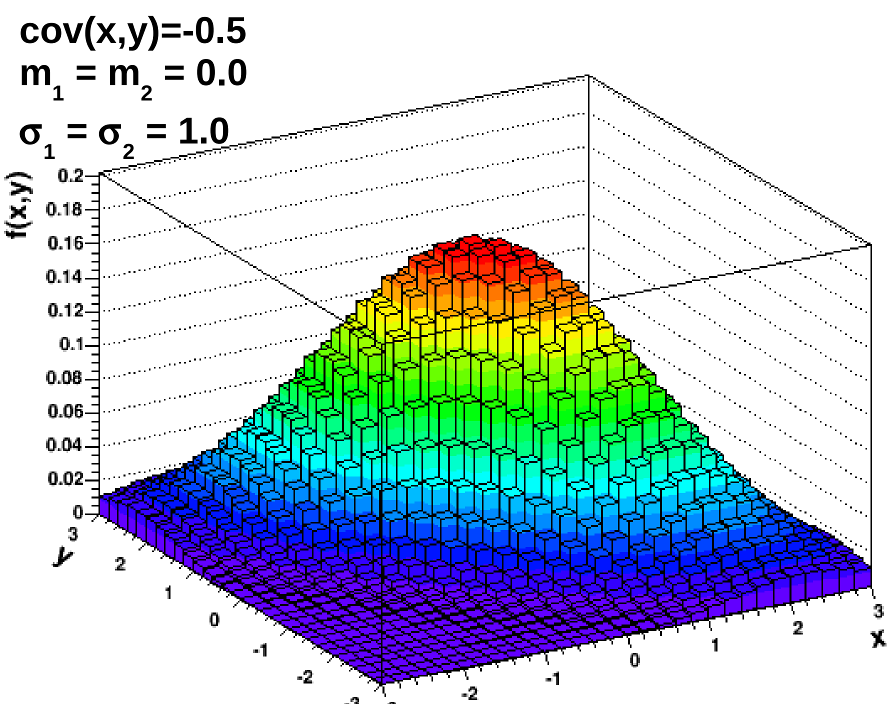
- Kowariancja cov(*x*,*y*) nie ma odpowiednika w dziedzinie funkcji jednej zmiennej.
- Kowariancja cov(*x*,*y*) jest dodatnia, kiedy: $x > \hat{x}$ oraz $y > \hat{y}$
- Kowariancja cov(*x*,*y*) jest ujemna, kiedy: $x > \hat{x}$ oraz $y < \hat{y}$
- Jeśli znajomość wartości *x* nie daje żadnej informacji o wartości *y*, kowariancja jest równa zeru.

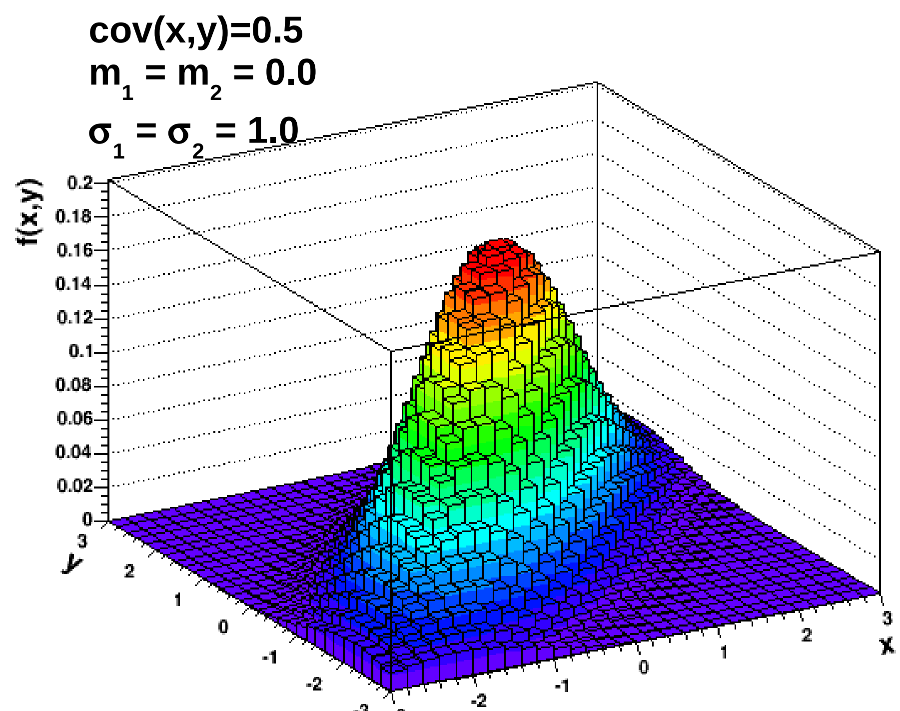
$$\sigma^2(\mathrm{a}\,x+\mathrm{b}\,y)=\mathrm{E}\{[(\mathrm{a}\,x+\mathrm{b}\,y)-\mathrm{E}(\mathrm{a}\,x+\mathrm{b}\,y)]^2\}=\mathrm{E}\{[\mathrm{a}(x-\hat{x})+\mathrm{b}(y-\hat{y})]^2\}$$


cov(x,y)=0.0
m_1 = m_2 = 0.0
σ_1 = σ_2 = 1.0
cov(x,y)=-0.75
m_1 = m_2 = 0.0
σ_1 = σ_2 = 1.0
cov(x,y)=-0.5
m_1 = m_2 = 0.0
σ_1 = σ_2 = 1.0
cov(x,y)=0.5
m_1 = m_2 = 0.0
σ_1 = σ_2 = 1.0
x
y


cov(x,y)=0.0
$m_1 = m_2 = 0.0$
$\sigma_1 = \sigma_2 = 1.0$
f(x,y)
x
y


cov(x,y)=-0.75
$m_1 = m_2 = 0.0$
$\sigma_1 = \sigma_2 = 1.0$
f(x,y)
x
y


cov(x,y)=-0.5
$m_1 = m_2 = 0.0$
$\sigma_1 = \sigma_2 = 1.0$
f(x,y)
x
y


cov(x,y)=0.5
$m_1 = m_2 = 0.0$
$\sigma_1 = \sigma_2 = 1.0$
f(x,y)
x
y

Współczynnik korelacji

Współczynnik korelacji:

$$\rho(x,y)=\frac{\operatorname{cov}(x,y)}{\sigma(x)\sigma(y)}$$

Kowariancja i współczynnik korelacji to miary współzależności x oraz *y*.

Dowiedziemy, że $-1\leqslant\rho(x,y)\leqslant 1$

Wprowadzamy zmienne losowe zredukowane: $\mathrm{u}=\frac{x-\hat{\mathrm{x}}}{\sigma(x)}$ $\mathrm{v}=\frac{y-\hat{\mathrm{y}}}{\sigma(y)}$

Wariancja: $\sigma^2(u+v)=\sigma^2(u)+\sigma^2(v)+2\rho(u,v)\sigma(u)\sigma(v)$

Ponieważ: $\sigma^2(u)=\sigma^2(v)=1$

$$\sigma^2(u+v)=2(1+\rho(u,v))$$

$$\sigma^2(u-v)=2(1-\rho(u,v))$$

Współczynnik korelacji

Wariancja jest nieujemna oraz $\rho(u,v)=\rho(x,y)$ oraz $-1\leqslant\rho(x,y)\leqslant 1$

Zbadajmy wartości krańcowe: $\rho(u,v)=\pm 1$

Kiedy $\rho(u,v)=1$ wariancja znika $\sigma(u-v)=0$ zmienna losowa $(u-v)=\mathrm{const}$

$$(u-v)=\frac{x-\hat{x}}{\sigma(x)}-\frac{y-\hat{y}}{\sigma(y)}=\mathrm{const}$$

Równanie jest spełnione, kiedy:

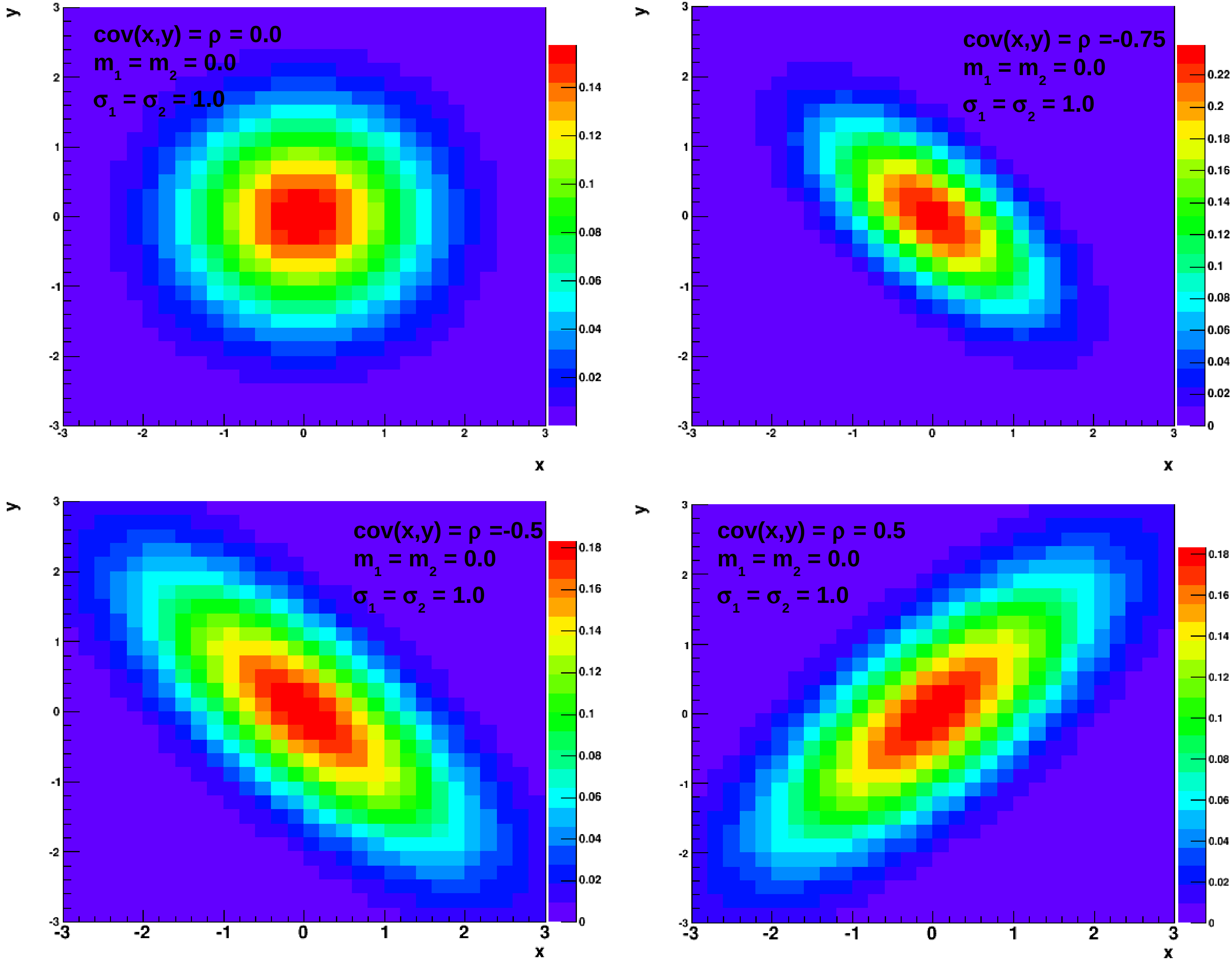
$$y=\mathrm{a}+\mathrm{b}\,x$$

Współczynnik a jest dowolną liczbą rzeczywistą, b jest liczbą rzeczywistą dodatnią,

Współczynnik korelacji $\rho(x,y)=+1$

Współczynnik a jest dowolną liczbą rzeczywistą, b jest liczbą rzeczywistą ujemną,

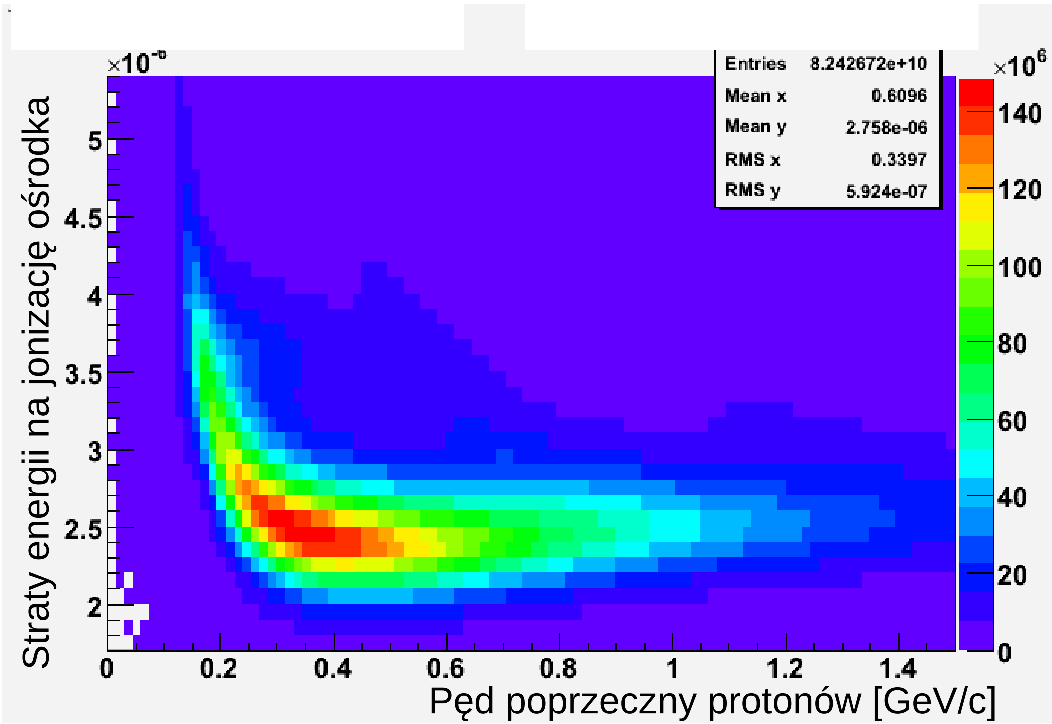
Współczynnik korelacji $\rho(x,y)=-1$

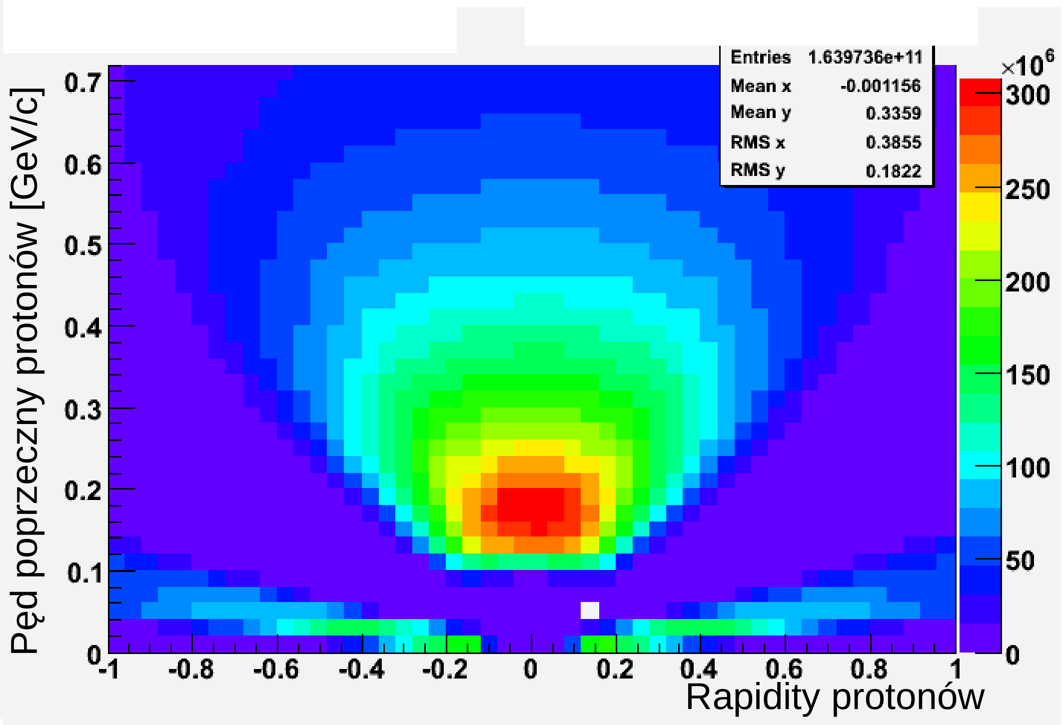

cov(x,y) = ρ = 0.0
m1 = m2 = 0.0
σ1 = σ2 = 1.0
cov(x,y) = ρ =-0.75
m1 = m2 = 0.0
σ1 = σ2 = 1.0
cov(x,y) = ρ =-0.5
m1 = m2 = 0.0
σ1 = σ2 = 1.0
cov(x,y) = ρ = 0.5
m1 = m2 = 0.0
σ1 = σ2 = 1.0
x
y

Rozkłady zmiennej losowej dwuwymiarowej

Dla dwóch zmiennych niezależnych kowariancja (i współczynnik korelacji) wynosi 0.

$$\mathrm{cov}(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\hat{x})(y-\hat{y})\,g(x)\,h(y)\,dx\,dy=\left(\int_{-\infty}^{\infty}(x-\hat{x})\,g(x)\,dx\right)\left(\int_{-\infty}^{\infty}(y-\hat{y})\,h(y)\,dy\right)=0$$





$$Y=\frac{1}{2}\ln\frac{E+p_L}{E-p_L}$$

Zmienne losowe N-wymiarowe

Rozkłady wielu zmiennych losowych – dystrybuanta, gęstość prawdopodobieństwa

Dystrybuanta dla n zmiennych losowych x_1, x_2, ..., x_n

$$F(x_1, x_2, \ldots, x_n) = P(x_1 < x_1, x_2 < x_2, \ldots, x_n < x_n)$$

Jeśli dystrybuanta jest funkcją ciągłą oraz posiada pochodne cząstkowe względem x_i,

Łączna gęstość prawdopodobieństwa:

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F(x_1, x_2, \ldots, x_n)$$

Gęstość rozkładu brzegowego:

$$g_r(x_r) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_{r-1} dx_{r+1} \ldots dx_n$$

To gęstość prawdopodobieństwa zmiennej x_r.

Rozkłady wielu zmiennych losowych – ciąg dalszy

$H(x_1, x_2, ..., x_n)$ - dowolna funkcja n zmiennych

Wartość oczekiwana:

$$E\{H(x_1, x_2, \ldots, x_n)\} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} H(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n) dx_1 \ldots dx_n$$

Jeśli $H(x) = x_r$ to:

$$E\{x_r\} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_r f(x_1, x_2, \ldots, x_n) dx_1 \ldots dx_n$$

$$E\{x_r\} = \int_{-\infty}^{\infty} x_r g_r(x_r) dx_r$$

Niezależność zmiennych losowych:

$$f(x_1, x_2, \ldots, x_n) = g_1(x_1) g_2(x_2) \ldots g_n(x_n)$$

Łączna gęstość rozkładu brzegowego dla dowolnych l spośród n zmiennych ($l<n$):

$$g(x_1, x_2, \ldots x_l) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_{l+1} \ldots dx_n$$

Zmienne x_1, x_2, ..., x_l są niezależne, kiedy:

$$g(x_1, x_2, \ldots x_l) = g_1(x_1) g_2(x_2) \ldots g_l(x_l)$$

Rozkłady wielu zmiennych losowych - momenty

Momenty rzędu l_1, l_2, ..., l_n to wartości oczekiwane funkcji:

$$H = x_1^{l_1} x_2^{l_2} \dots x_n^{l_n}$$

Oznaczamy je przez:

$$\lambda_{l_1 l_2 \dots l_n} = E(x_1^{l_1} x_2^{l_2} \dots x_n^{l_n})$$

W szczególności:

$$\lambda_{100\dots0} = E(x_1) = \hat{x}_1$$

$$\lambda_{010\dots0} = E(x_2) = \hat{x}_2$$

$$\lambda_{000\dots1} = E(x_n) = \hat{x}_n$$

Momenty (momenty centralne) względem wartości średnich:

$$\mu_{l_1 l_2 \dots l_n} = E\{(x_1 - \hat{x}_1)^{l_1} (x_2 - \hat{x}_2)^{l_2} \dots (x_n - \hat{x}_n)^{l_n}\}$$

Rozkłady wielu zmiennych losowych – momenty c.d.

Wariancje:

$$\mu_{200\ldots0}=\mathrm{E}\{(x_1-\hat{x}_1)^2\}=\sigma^2(x_1)$$

$$\mu_{020\ldots0}=\mathrm{E}\{(x_2-\hat{x}_2)^2\}=\sigma^2(x_2)$$

$$\mu_{000\ldots2}=\mathrm{E}\{(x_n-\hat{x}_n)^2\}=\sigma^2(x_n)$$

Moment centralny dla $l_i = l_j = 1$ oraz $l_k = 0$ (i, k, j są różne), **kowariancja** między x_i, x_j

$$c_{ij}=\mathrm{cov}(x_i, x_j)=\mathrm{E}\{(x_i-\hat{x}_i)(x_j-\hat{x}_j)\}$$

Notacja wektorowa dla zmiennych losowych n-wymiarowych

Reprezentacja n zmiennych x_1, x_2, ..., x_n jako składowych wektora **x** w przestrzeni n-wymiarowej.

Dystrybuanta: $F=F(\mathbf{x})$

Gęstość prawdopodobieństwa: $f(\mathbf{x})=\frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F(\mathbf{x})$

Wartość oczekiwana: $E\{H(\mathbf{x})\}=\int H(x) f(x) dx$

Macierz kowariancji: $$C=\begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{bmatrix}$$

Elementy c_{ij} to współczynniki kowariancji, elementy diagonalne macierzy, to $c_{ii}=\sigma^2(x_i)$

Notacja wektorowa dla zmiennych losowych n-wymiarowych

Macierz kowariancji to macierz symetryczna: $c_{ij} = c_{ji}$

Wartości oczekiwane: $E(\mathbf{x}) = \hat{\mathbf{x}}$

Element macierzy kowariancji: $c_{ij} = E\{(x_i - \hat{x}_i)(x_j - \hat{x}_j)^T\}$

Jest to wartość średnia elementu o wskaźnikach *ij* iloczynu diadycznego wektorów:

$(\boldsymbol{x_i} - \hat{\mathbf{x}}_{\mathbf{i}})(\boldsymbol{x_i} - \hat{\mathbf{x}}_{\mathbf{i}})^T$

$$x^T = (x_1 \quad x_2 \quad \dots \quad x_n)$$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

Macierz kowariancji: $C = E\{(\boldsymbol{x} - \hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}})^T\}$