Komputerowa analiza danych doświadczalnych

Wykład 3
11.03.2016

dr inż. Łukasz Graczykowski

lgraczyk@if.pw.edu.pl

Wykłady z poprzednich lat
(dr inż. H. Zbroszczyk):

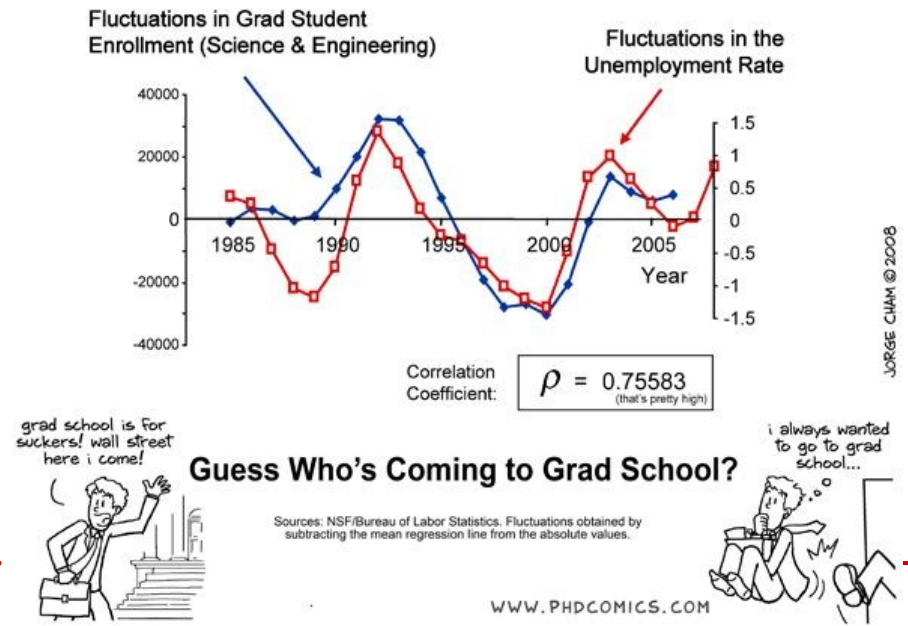
http://www.if.pw.edu.pl/~gos/students/kadd/

Semestr letni 2015/2016

Dwuwymiarowe rozkłady zmiennych losowych

Jednoczesne pomiary dwóch wielkości

- W przypadku znakomitej większości pomiarów jednocześnie mierzymy dwie i więcej wielkości fizycznych (np. napięcie I natężenie prądu w obwodzie)
- Analogicznie w badaniach społecznych, możemy badać jednocześnie kilka cech populacji (np. zamożność i długość życia)
- **Kluczowe pytanie** – czy (i jeśli tak to jaka) jest zależność między tymi wielkościami? Jak jedna wielkość wpływa na drugą? Innymi słowy: jakie są korelacje między tymi zmiennymi?



Rozkład i dystrybuanta 2D

- W wyniku jednokrotnego pomiaru otrzymujemy dwie liczby:
  - (x,y), które są wartościami zmiennej losowej X, oraz Y
- **Rozkład prawdopodobieństwa**:

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

  - rozkład prawdopod. jest unormowany
    - rozkład ciągły: $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy=1$
    - rozkład dyskretny: $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} p_{ij}=1$
- **Dystrybuanta**:

$$F(x,y)=P(x\leqslant X,y\leqslant Y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(x',y')\,dx'dy' \qquad F(x)=\sum_{i:x_i\leqslant X}\sum_{j:y_j\leqslant Y} p_{ij}$$

- Jeżeli dystrybuanta jest funkcją ciągłą obu zmiennych, to:

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

- Prawdopodobieństwo:

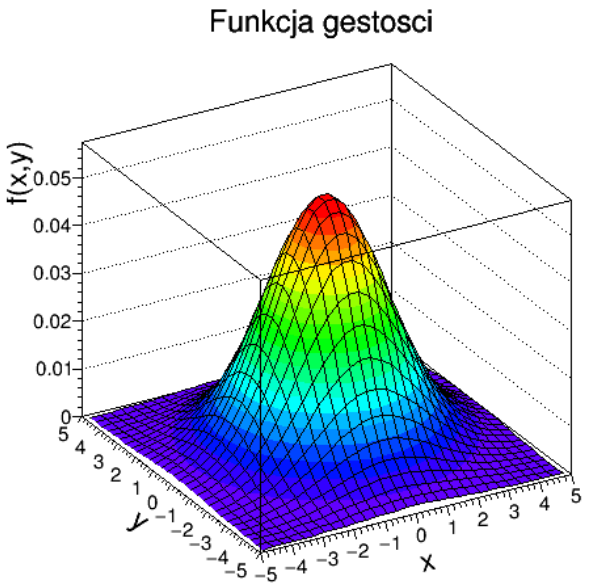
$$P(a\leqslant X\leqslant b,c\leqslant Y\leqslant d)=\int_{a}^{b}\int_{c}^{d} f(x,y)\,dxdy=F(b,d)-F(a,b)$$

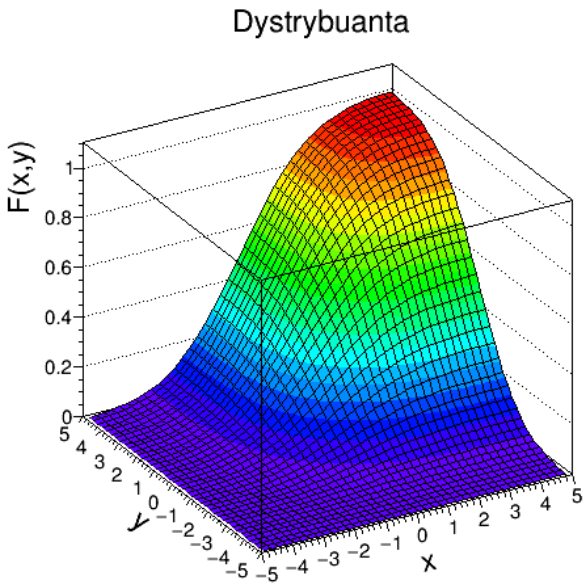
Zmienna losowa 2D - przykład

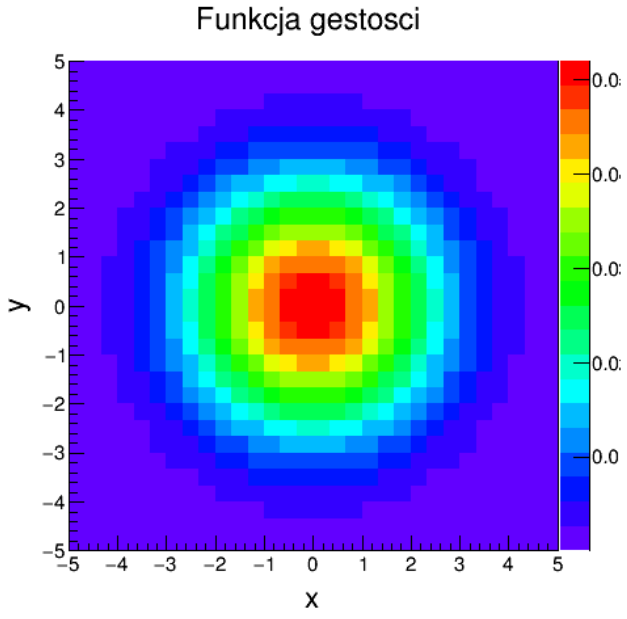
- Dwuwymiarowy rozkład Gaussa:

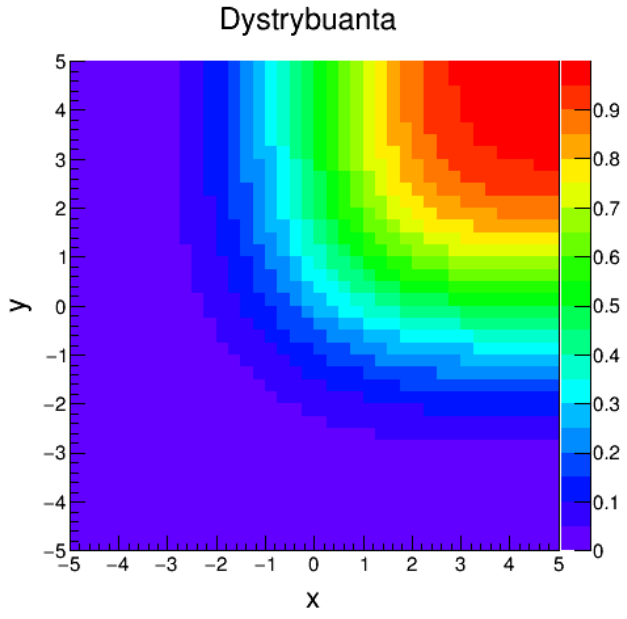
$$f(x,y) = N \cdot \exp\left(-\left(\frac{(x-\hat{x})^2}{2\sigma_x^2} + \frac{(y-\hat{y})^2}{2\sigma_y^2}\right)\right)$$

normalizacja









Rozkłady (gęstości) brzegowe

- Częsty problem doświadczalny:
  - mamy wynik pomiaru – zmienne losowe X, oraz Y, ale interesuje nas tylko zależność od X dla dowolnego Y
  - **przykład:** gęstość prawdopodobieństwa zgonów wywołanych pewnymi chorobami zakaźnymi jest funkcją czasu i położenia geograficznego; w badaniach potrzebujemy zająć się tylko zależnością czasową
- **Brzegowa gęstość prawdopodobieństwa**:

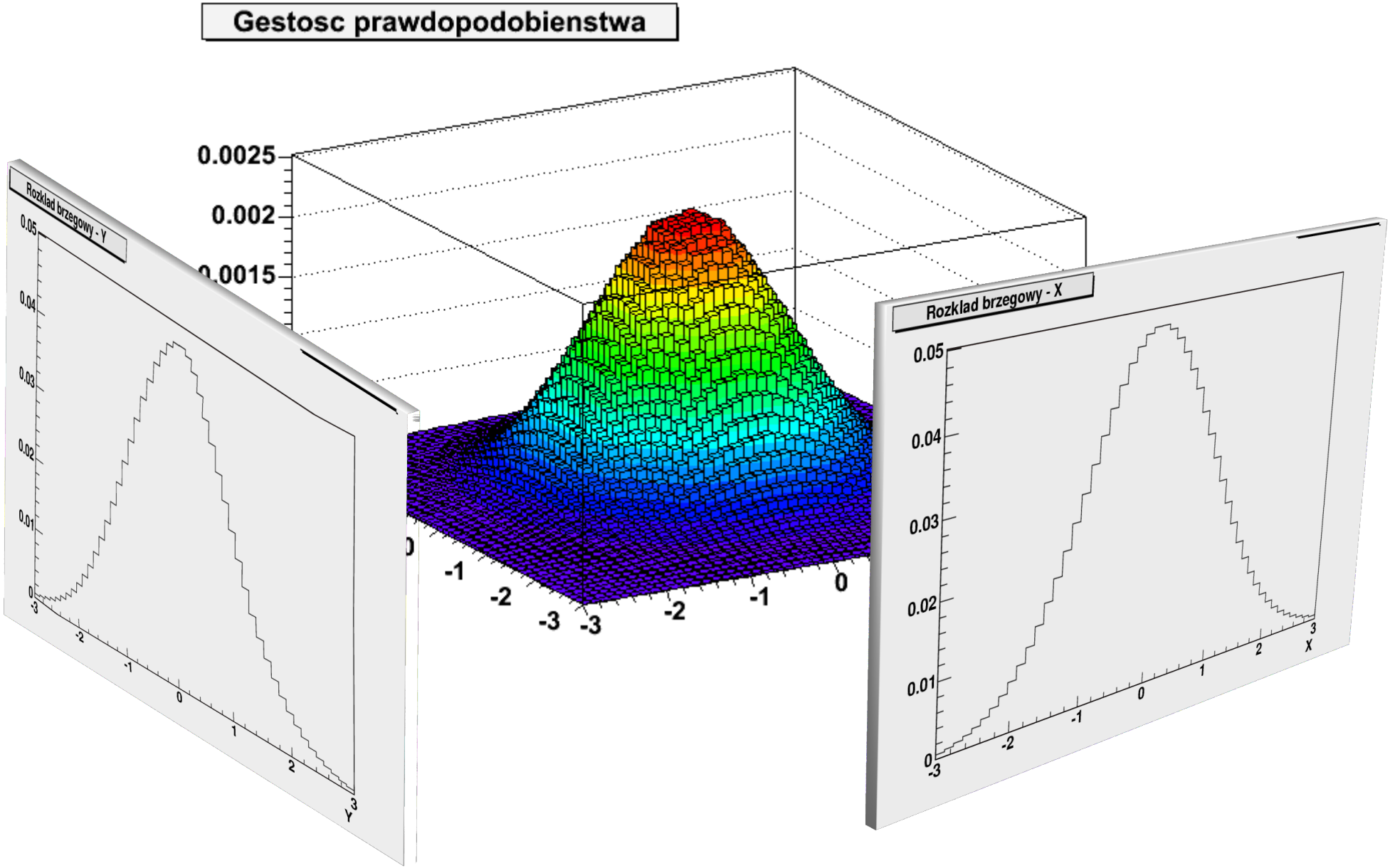
$$P(a \leqslant X \leqslant b, -\infty < Y < \infty) = \int_a^b \left[\int_{-\infty}^{\infty} f(x,y)\,dy \right] dx = \int_a^b g(x)\,dx$$

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

- **Dystrybuanty brzegowe**:

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

Rozkłady (gęstości) brzegowe - przykład



Niezależnosć zmiennych

- **Prawdopodobieństwo warunkowe**:

$$P(B|A)=\frac{P(A\cap B)}{P(B)}\Rightarrow P(A\cap B)=P(B)P(B|A)$$

- Dla zdarzeń niezależnych:

$$P(B|A)=P(B)\Rightarrow P(A\cap B)=P(A)P(B)$$

- Analogicznie dla **zmiennych losowych niezależnych**:

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

- Warunkowa gęstość prawdopodobieństwa:

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

- Prawdopodobieństwo warunkowe zmiennej losowej Y przy znanej wartości zmiennej losowej X:

$$P(y\leqslant Y\leqslant y+dy|x\leqslant X\leqslant x+dx)=f(y|x)dy$$

- Rozkłady brzegowe: $h(y)=\int_{-\infty}^{\infty} f(x|y)g(x)\,dx$ $\qquad$ $g(x)=\int_{-\infty}^{\infty} f(x|y)g(y)\,dy$

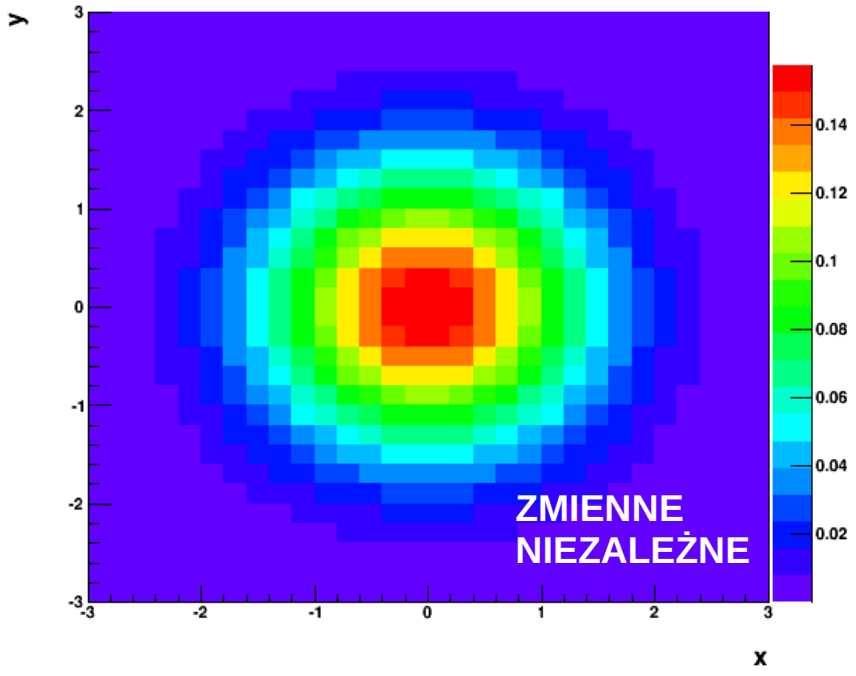
Niezależnosć zmiennych

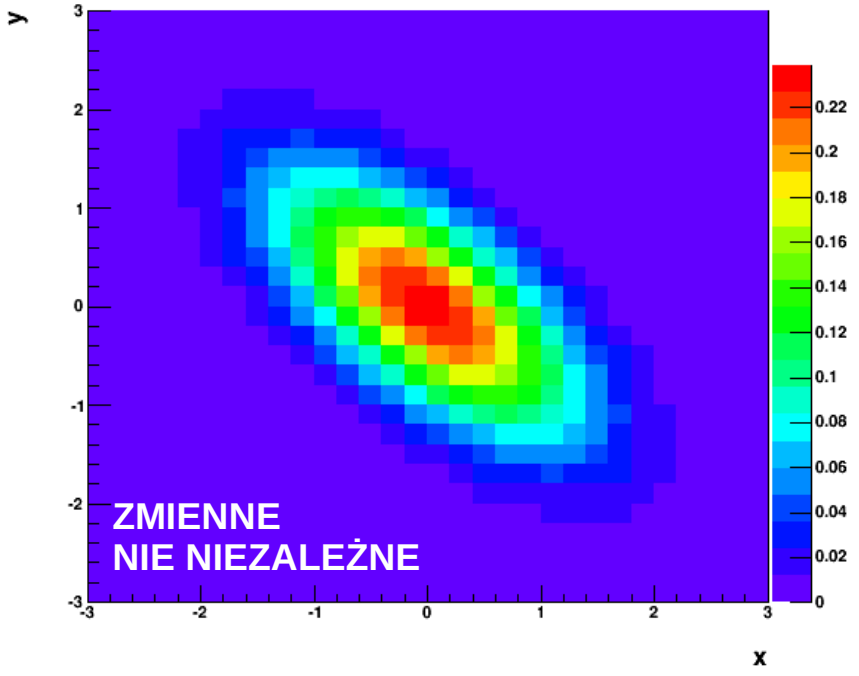
- Dla zmiennych niezależnych otrzymamy:

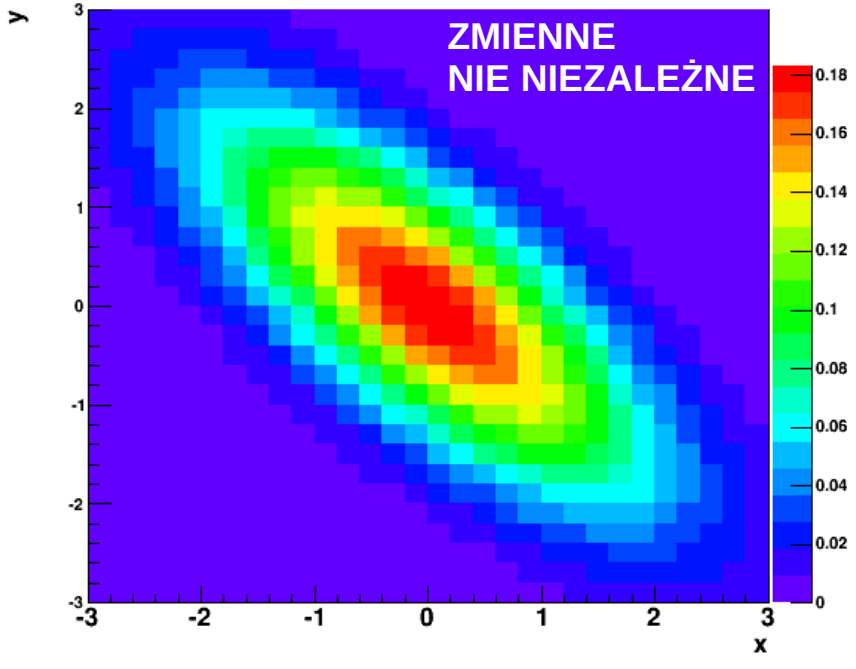
$$f(y|x)=\frac{f(x,y)}{g(x)}=\frac{g(x)h(y)}{g(x)}=h(y)$$

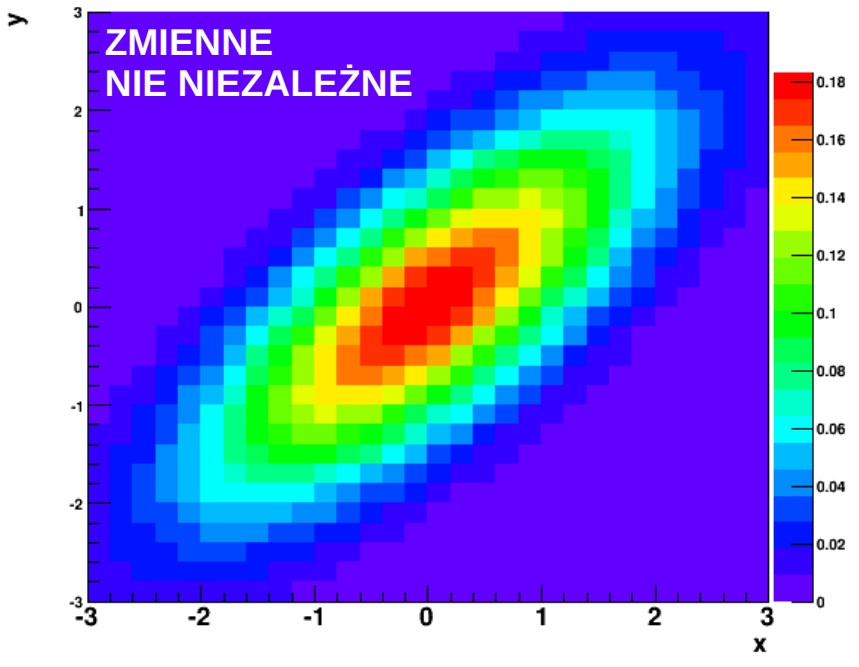
- Wynik ten pokazuje łatwy do przewidzenia fakt – jakikolwiek warunek narzucony na jedną zmienną nie może wpłynąć na rozkład drugiej zmiennej (jeśli są niezależne)

Niezależnosć zmiennych









Wartość oczekiwana, wariancja, momenty

- Wartość oczekiwana

$$E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dxdy \qquad E(XY)=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i y_j p_{ij}$$

- Jeżeli mamy funkcję $H(x,y)$, to wartość oczekiwana:

$$E(H(X,Y))=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)f(x,y)\,dxdy$$

- Wariancja:

$$\sigma^2(X,Y)=E\left(\left[E(X,Y)-E(X,Y)\right]^2\right) \qquad \sigma^2(H(X,Y))=E\left(\left[E(H(X,Y))-E(H(X,Y))\right]^2\right)$$

- Jeżeli: $H(x,y)=a\cdot X+b\cdot Y$ wówczas: $E(a\cdot X+b\cdot Y)=a\cdot E(X)+b\cdot E(Y)$
- Moment zwykły rzędu l i m względem zmiennych X i Y:

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

- Ogólniej – moment rzędu l i m względem punktów a i b:

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

- Momenty centralne: $\mu_{lm}=E\left((X-\lambda_{10})^l(Y-\lambda_{01})^m\right)$

Wartość oczekiwana, wariancja, momenty

- Momenty o specjalnym znaczeniu:

$$\mu_{00}=\lambda_{00}=1$$
$$\mu_{10}=\mu_{01}=0$$
$$\lambda_{10}=E(x)=\hat{x}$$
$$\lambda_{01}=E(y)=\hat{y}$$
$$\mu_{11}=E((X-\hat{x})(Y-\hat{y}))=cov(X,Y)$$
$$\mu_{20}=E\left((X-\hat{x})^2\right)=\sigma^2(X)$$
$$\mu_{02}=E\left((Y-\hat{y})^2\right)=\sigma^2(Y)$$

- Wariancja dla zmiennej $aX+bY$:

$$\sigma^2(a\cdot X+b\cdot Y)=E\left((a\cdot X+b\cdot Y)-E(a\cdot X+b\cdot Y)\right)=a^2\sigma^2(X)+b^2\sigma^2(Y)+2ab\cdot cov(X,Y)$$

- Jeżeli założymy funkcję H w postaci iloczynu: $H(X,Y)=X\cdot Y$

$$E(X\cdot Y)=E(X)E(Y)$$

- Wielkości $E(X)$, $E(Y)$, $\sigma(X)$, $\sigma(Y)$ są podobne jak w 1D

Kowariancja

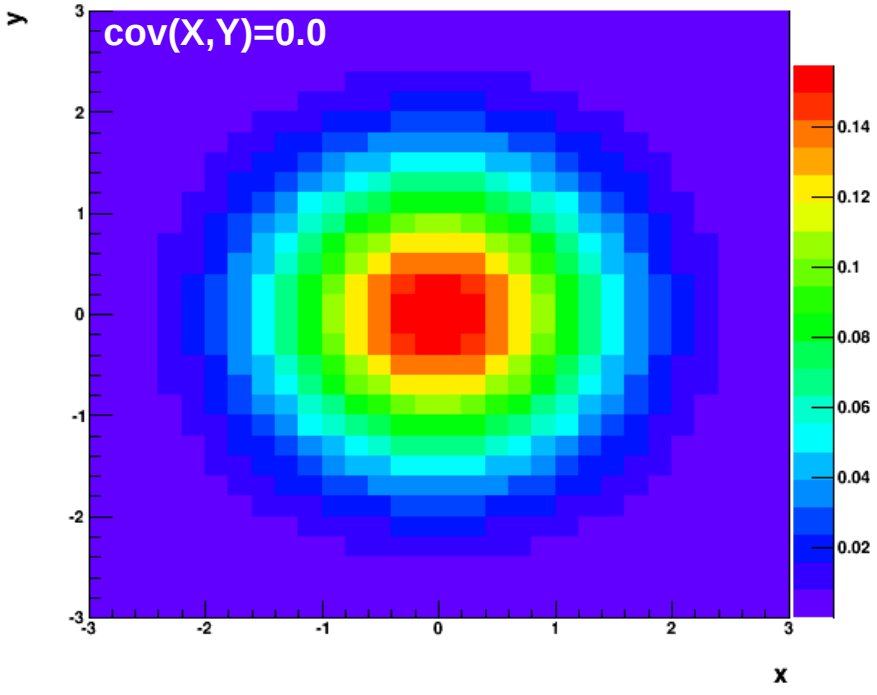
- **Kowariancja** $cov(X,Y)$ nie ma odpowiednika w przypadku rozkładów 1D
- Z definicji kowariancji wynika, że:
  - jest dodatnia, gdy: $x>\hat{x}$ oraz $y>\hat{y}$
  - jest ujemna, gdy: $x>\hat{x}$ oraz $y<\hat{y}$
  - jeżeli nie ma zależności między x i y nie ma zależności, wówczas kowariancja wynosi 0

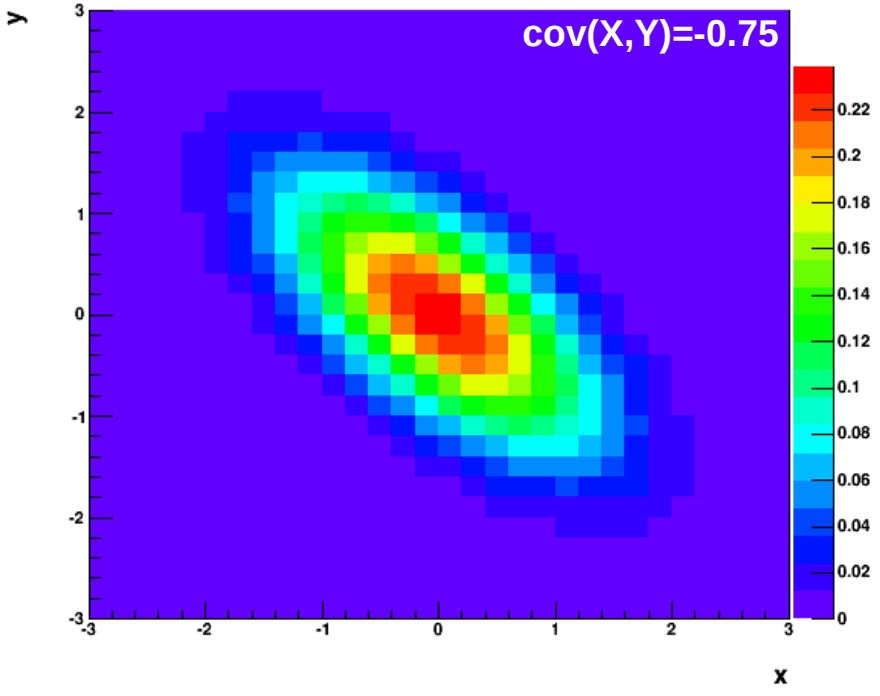
$$cov(X,Y)=\mu_{11}=E((X-E(X))\cdot(Y-E(Y)))=\boxed{E(X\cdot Y)-E(X)\cdot E(Y)}$$

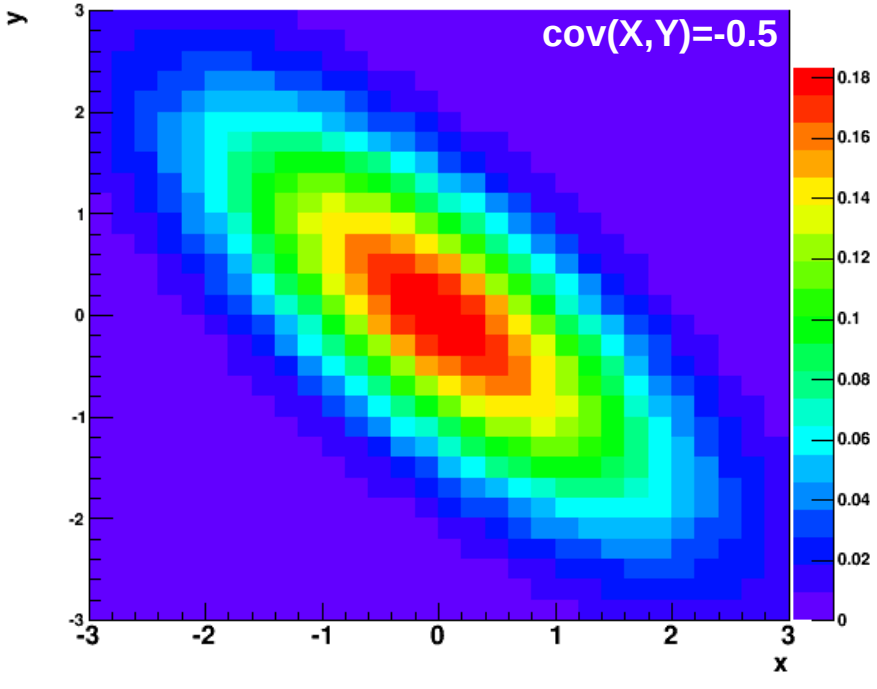
- **Interpretacja**: jeżeli między zmiennymi X i Y nie istnieje żadna korelacja liniowa i istnieją ich wartości oczekiwane, to kowariancja przyjmuje wartość 0. Czyli:

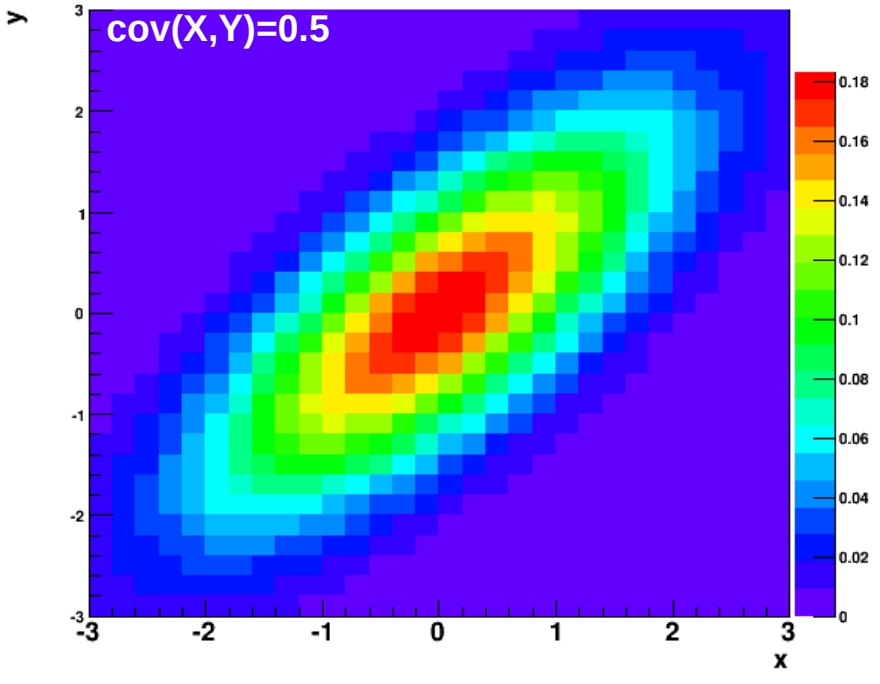
$$cov(X,Y)=0\Rightarrow E(X\cdot Y)=E(X)\cdot E(Y)$$

Kowariancja



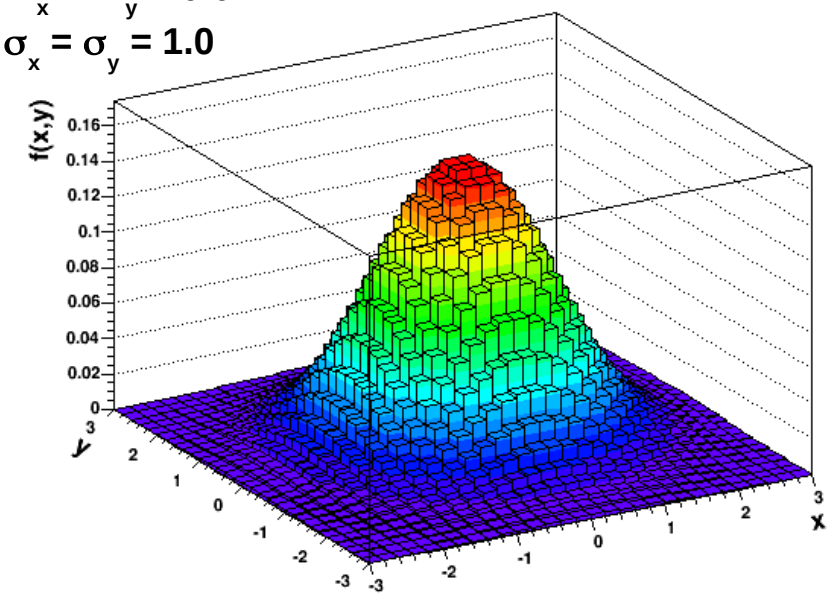




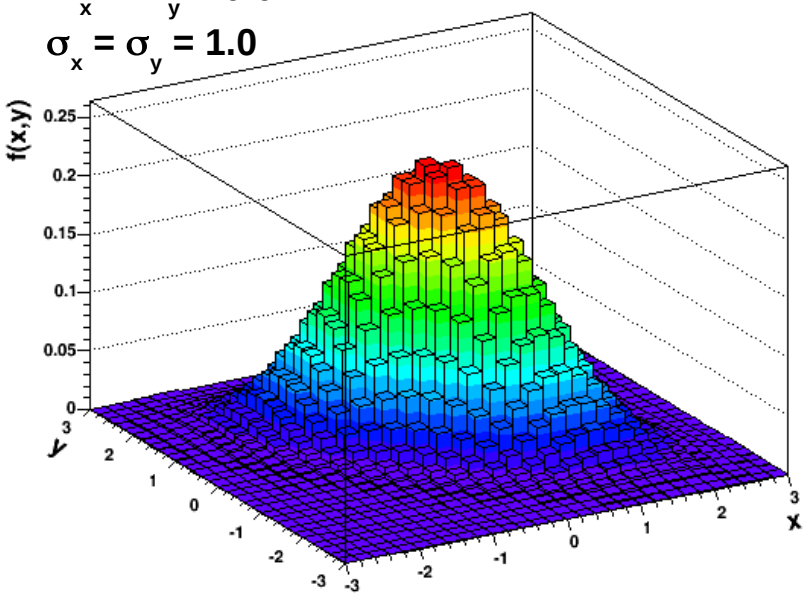


Kowariancja

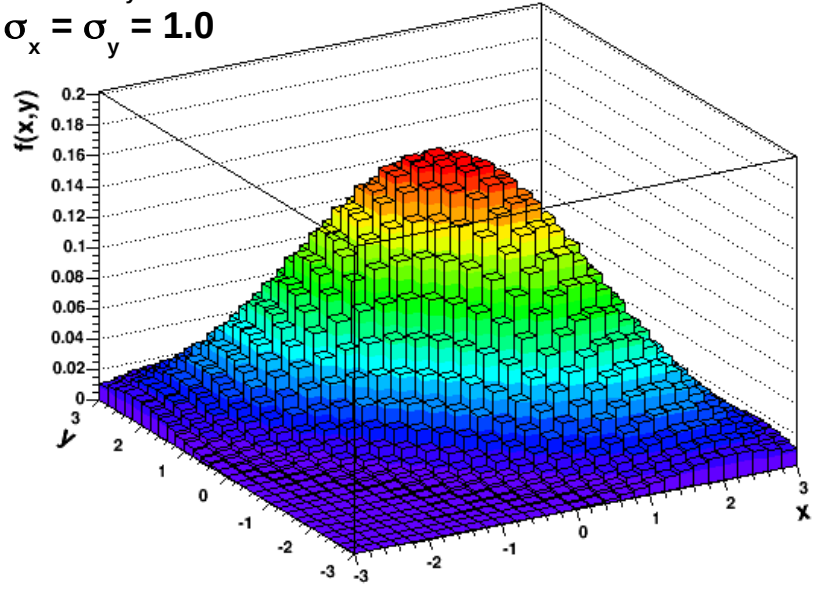
cov(X,Y)=0.0
$m_x = m_y = 0.0$
$\sigma_x = \sigma_y = 1.0$



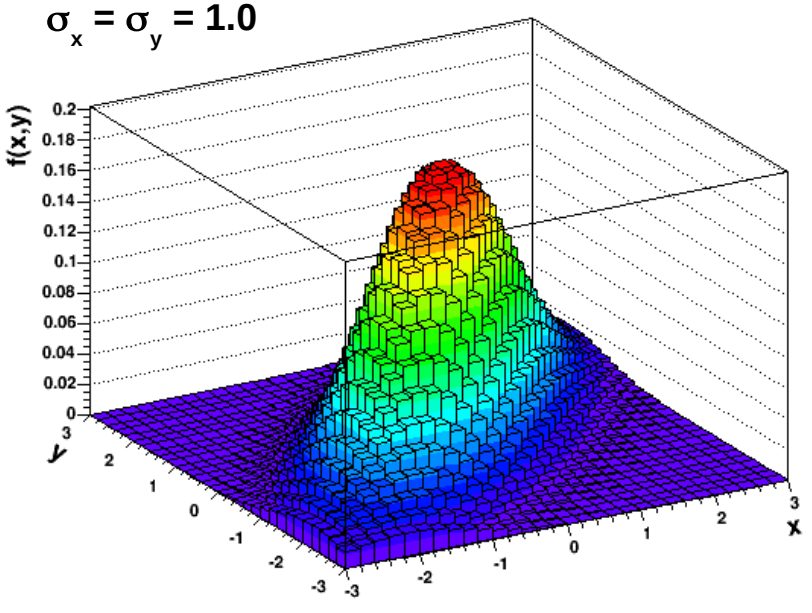
cov(X,Y)=-0.75
$m_x = m_y = 0.0$
$\sigma_x = \sigma_y = 1.0$



cov(X,Y)=-0.5
$m_x = m_y = 0.0$
$\sigma_x = \sigma_y = 1.0$



cov(X,Y)=0.5
$m_x = m_y = 0.0$
$\sigma_x = \sigma_y = 1.0$

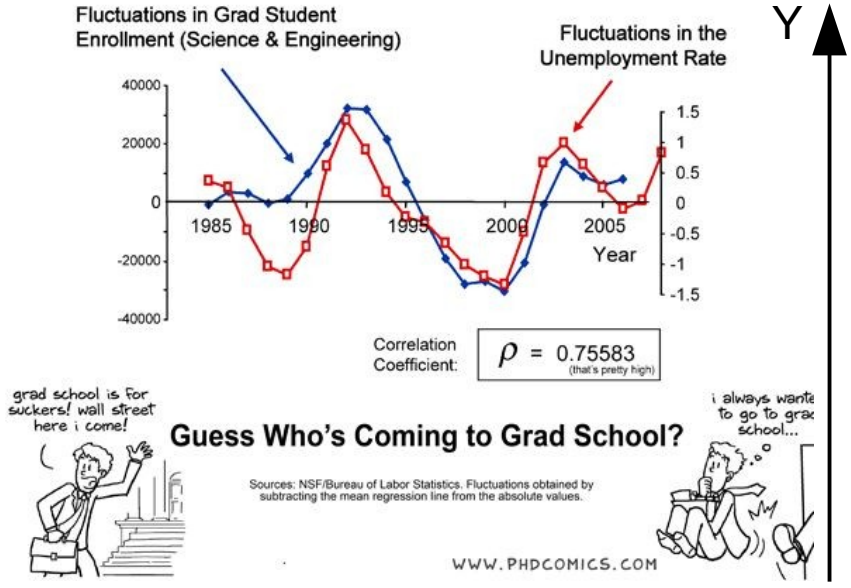


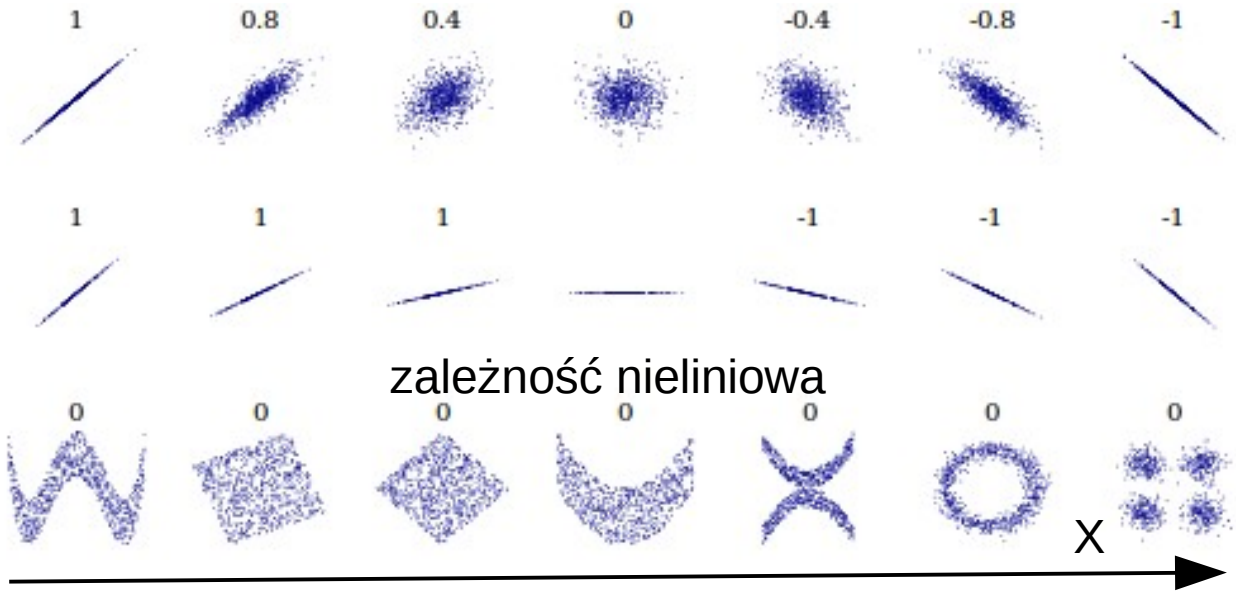
Współczynnik korelacji

- **Współczynnik korelacji (Pearsona):**

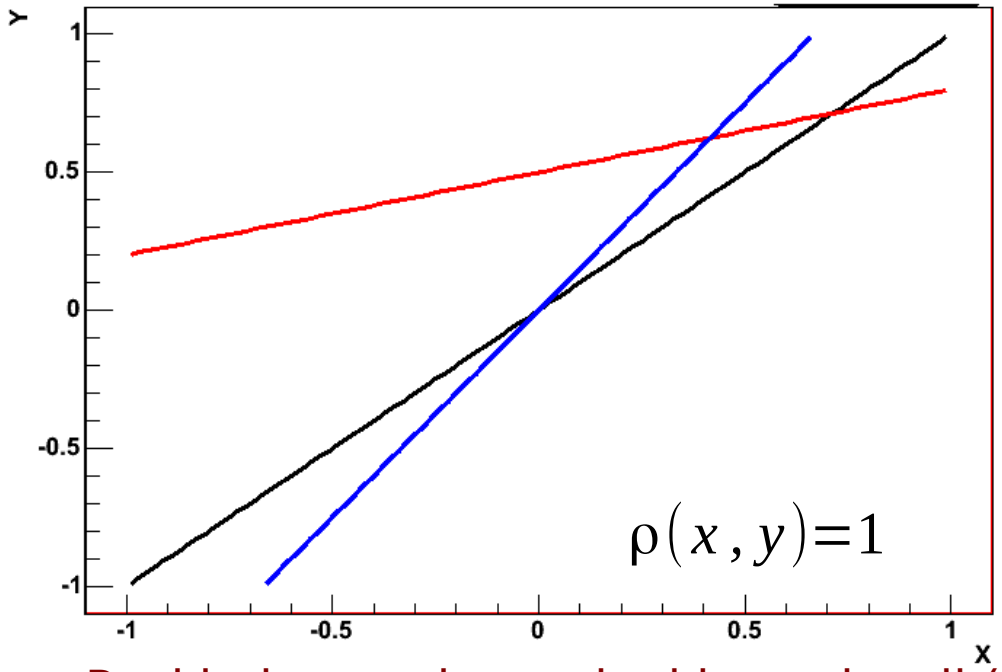
$$\rho(X,Y)=\frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

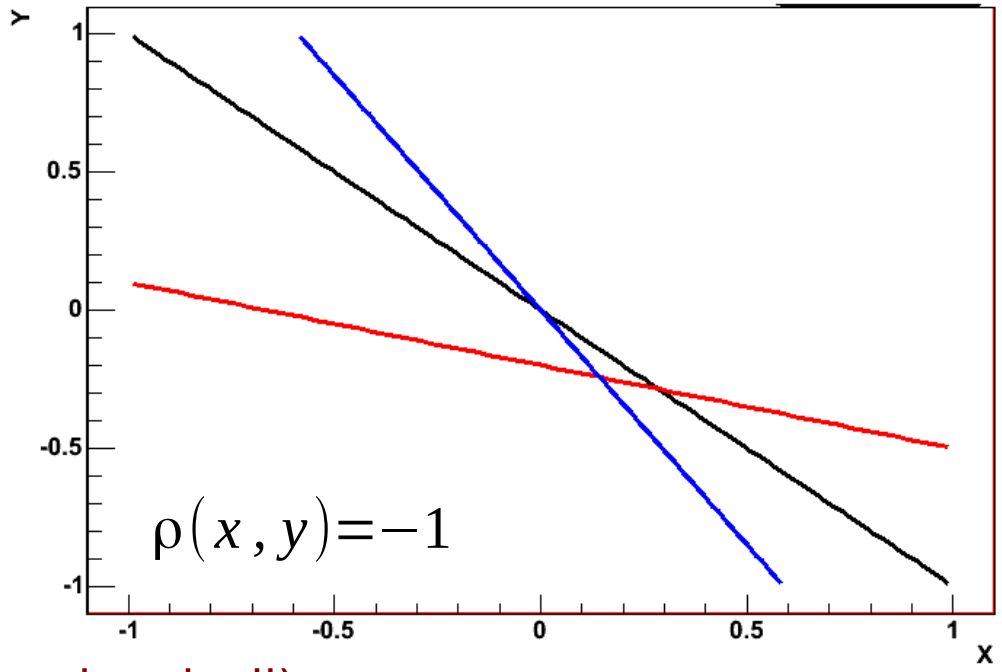
- Kowariancja i współczynnik korelacji to miary zależności liniowej X oraz Y:
- Można wykazać (patrz Brandt), że: $-1 \leqslant \rho(X,Y) \leqslant 1$
  - współczynnik korelacji jest zatem wielkością znormalizowaną
- Współczynnik korelacji ocenia **jedynie liniową zależność (nie jest to zależność przyczyna-skutek!)**



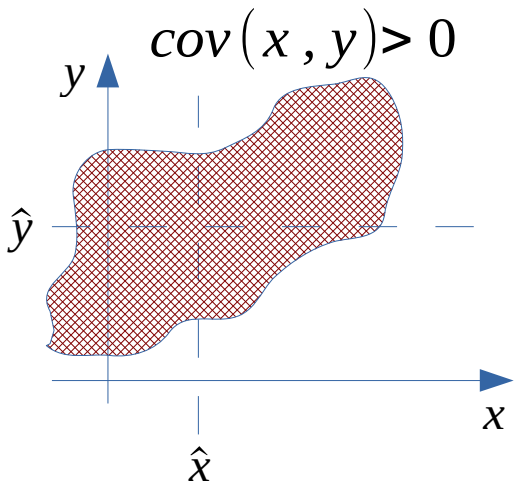


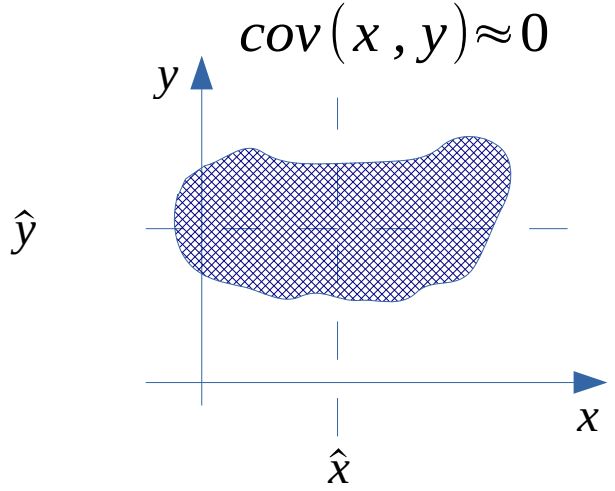
Kowariancja

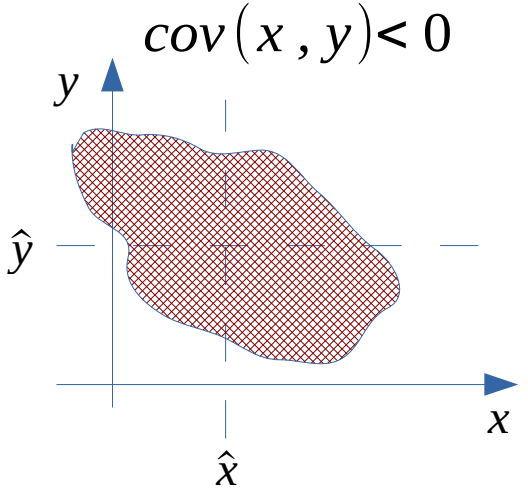




Rozkłady o maksymalnej kowariancji (wsp. korelacji)



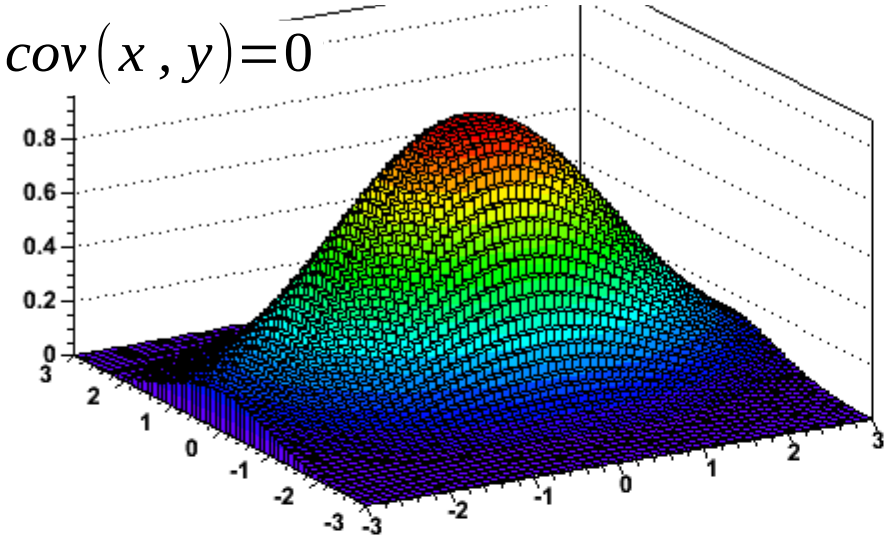




Korelacja a niezależność zmiennych

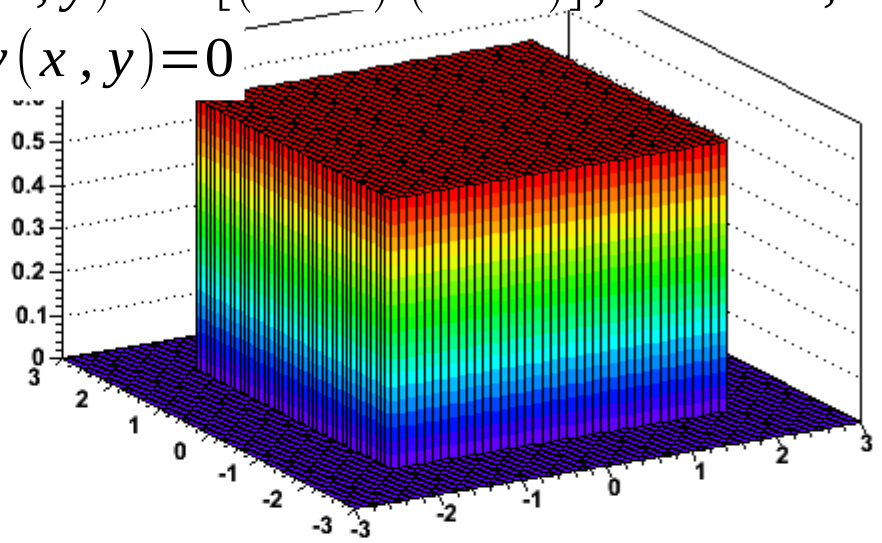
$f(x,y) = gaus(x) \cdot gaus(y)$

$cov(x,y) = 0$



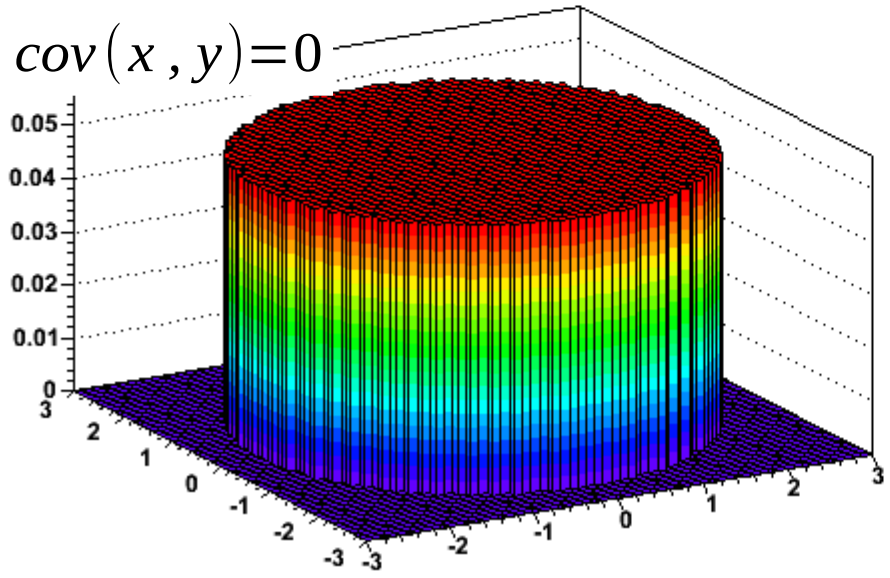
$f(x,y) = 1/[(a-b)\cdot(c-d)]; a < x < b, c < y < d$

$cov(x,y) = 0$



$f(x,y) = 1/(\pi R^2); \sqrt{x^2 + y^2} < R$

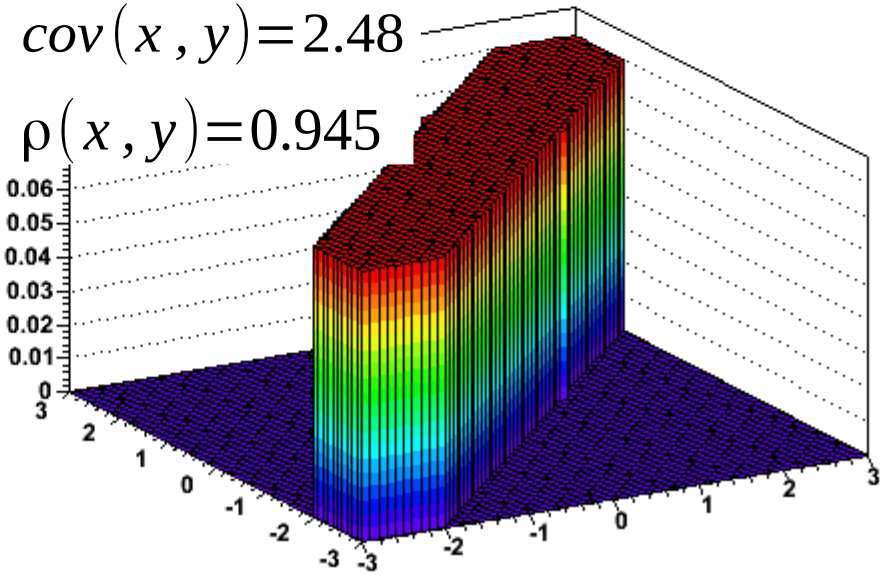
$cov(x,y) = 0$



$f(x,y) = 1/a; |x-y| < a$

$cov(x,y) = 2.48$

$\rho(x,y) = 0.945$



Rozkład i dystrybuanta N-wymiarowa

- Mamy N zmiennych losowych:

  $(X_1, X_2, \ldots, X_N)$

- **Rozkład prawdopodobieństwa**:

$$f(x_1, x_2 \ldots, x_N) = P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$$

  - rozkład prawdopod. jest unormowany

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_N)\, dx_1\, dx_2 \ldots dx_N = 1$$

- **Dystrybuanta**: $F(x_1, x_2, \ldots, x_N) = P(x_1 \leqslant X_1, x_2 \leqslant X_2, \ldots, x_N \leqslant X_N) =$

$$= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_1} \int_{-\infty}^{x_N} f(x'_1, x'_2, \ldots, x'_N)\, dx'_1\, dx'_2 \ldots dx'_N$$

- Jeżeli dystrybuanta jest funkcją ciągłą, to:

$$f(x_1, x_2, \ldots, x_N) = \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \ldots \frac{\partial}{\partial x_N} F(x_1, x_2, \ldots, x_N)$$

- Gęstość brzegowa zmiennej x_r:

$$g_r(x_r) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_N)\, dx_1\, dx_2 \ldots d_{x_{r-1}}\, d_{x_{r+1}} \ldots dx_N$$

Wartość oczekiwana, wariancja, kowariancja

- Wartość oczekiwana zmiennej x_r:

$$E(X_r)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} x_r\, f(x_1,x_2,\dots,x_N)\,dx_1\,dx_2\dots dx_N=\int_{-\infty}^{\infty} x_r\, g_r(x_r)\,dx_r$$

- Dla funkcji $H(x_1,x_2,\dots,x_N)$:

$$E\left(H(x_1,x_2,\dots,x_N)\right)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} H(x_1,x_2,\dots,x_N)\, f(x_1,x_2,\dots,x_N)\,dx_1\,dx_2\dots dx_N$$

- Wariancja zmiennej X_r:

$$\sigma^2(X_r)=E\left(\left(X_r-\hat{x}_r\right)^2\right)$$

- Jeżeli zmienne losowe są niezależne:

$$f(x_1,x_2,\dots,x_N)=g_1(x_1)\,g_2(x_2)\dots g_N(x_N)$$

- Łączny rozkład brzegowy dla dowolnych l spośród N zmiennych:

$$g(x_1,x_2,\dots,x_l)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} f(x_1,x_2,\dots,x_N)\,dx_{l+1}\dots dx_N$$

- Kowariancja pomiędzy zmiennymi x_i i x_j:

$$cov(X_i,X_j)=E\left(\left(X_i-\hat{x}_i\right)\left(X_j-\hat{x}_j\right)\right)$$

Momenty

- Przez analogię definiujemy również momenty (zwykłe):

$$\lambda_{l_1,l_2,\ldots,l_N} = E\left(X_1^{\,l_1} X_2^{\,l_2} \ldots X_N^{\,l_N}\right)$$

- W szczególności, wartości oczekiwane:

$$\lambda_{100\ldots0} = E\left(X_1\right) = \hat{x}_1$$

$$\lambda_{010\ldots0} = E\left(X_2\right) = \hat{x}_2$$

$$\lambda_{000\ldots N} = E\left(X_N\right) = \hat{x}_N$$

- Momenty centralne:

$$\mu_{l_1,l_2,\ldots,l_N} = E\left(\left(X_1-\hat{x}_1\right)^{l_1}\left(X_2-\hat{x}_2\right)^{l_2}\ldots\left(X_N-\hat{x}_N\right)^{l_N}\right)$$

- W szczególności, wariancje:

$$\mu_{200\ldots0} = E\left(\left(X_1-\hat{x}_1\right)^2\right) = \sigma^2\left(X_1\right)$$

$$\mu_{020\ldots0} = E\left(\left(X_2-\hat{x}_2\right)^2\right) = \sigma^2\left(X_2\right)$$

$$\mu_{000\ldots N} = E\left(\left(X_N-\hat{x}_N\right)^2\right) = \sigma^2\left(X_N\right)$$

Kowariancja:

$$l_i = l_j = 1;\ \ \forall\, l_k = 0 \left(i \neq k \neq j\right)$$

$$c_{ij} = cov\left(X_i, X_j\right) = E\left(\left(X_i-\hat{x}_i\right)\left(X_j-\hat{x}_j\right)\right)$$

Zapis wektorowy

- N zmiennych losowych można przedstawić jako N-wymiarowy wektor (**wektor losowy**):

$$\boldsymbol{X}=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} \qquad \boldsymbol{x}=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

- Funkcja gęstości:

$$f(\boldsymbol{x})=P(\boldsymbol{X}=\boldsymbol{x})$$

- Dystrybuanta:

$$F(\boldsymbol{x})=P(\boldsymbol{X}\leqslant\boldsymbol{x})$$

- Jeżeli istnieją pierwsze pochodne:

$$f(\boldsymbol{x})=\frac{\partial^N}{\partial x_1 \partial x_2 \ldots \partial x_N} F(\boldsymbol{X})$$

- Wartość oczekiwana funkcji $H(\boldsymbol{X})$:

$$E\left(H(\boldsymbol{X})\right)=\int H(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

Wektor wartości zmiennych losowych

Macierz kowariancji

- Macierz, której elementy to odpowiednie momenty odpowiadające wariancjom i kowariancjom nazywamy **macierzą kowariancji**:

$$C=\begin{pmatrix} c_{11} & c_{12} & \cdots c_{1N} \\ c_{21} & c_{22} & \cdots c_{2N} \\ \vdots & \vdots & \\ c_{N1} & c_{N2} & \cdots c_{NN} \end{pmatrix}$$

  - elementy c_{ij} dane są wzorem na kowariancję:

$$c_{ij}=cov(X_i,X_j)=E\left(\left(X_i-\hat{x}_i\right)\left(X_j-\hat{x}_j\right)\right)$$

  - elementy diagonalne c_{ii} to wariancje: $c_{ii}=\sigma^2(X_i)$
  - macierz kowariancji jest symetryczna: $c_{ij}=c_{ji}$

- wartość oczekiwana w zapisie wektorowym: $E(\boldsymbol{X})=\hat{\boldsymbol{x}}$

Macierz kowariancji

- Każdy element c_{ij}

$$c_{ij} = E\left(\left(X_i - \hat{x}_i\right)\left(X_j - \hat{x}_j\right)^T\right)$$

macierzy kowariancji możemy interpretować jako wartość średnią elementu o wskaźnikach ij iloczynu diadycznego wektorów $\left(\boldsymbol{X} - \hat{\boldsymbol{x}}\right)^T$ i $\left(\boldsymbol{X} - \hat{\boldsymbol{x}}\right)$

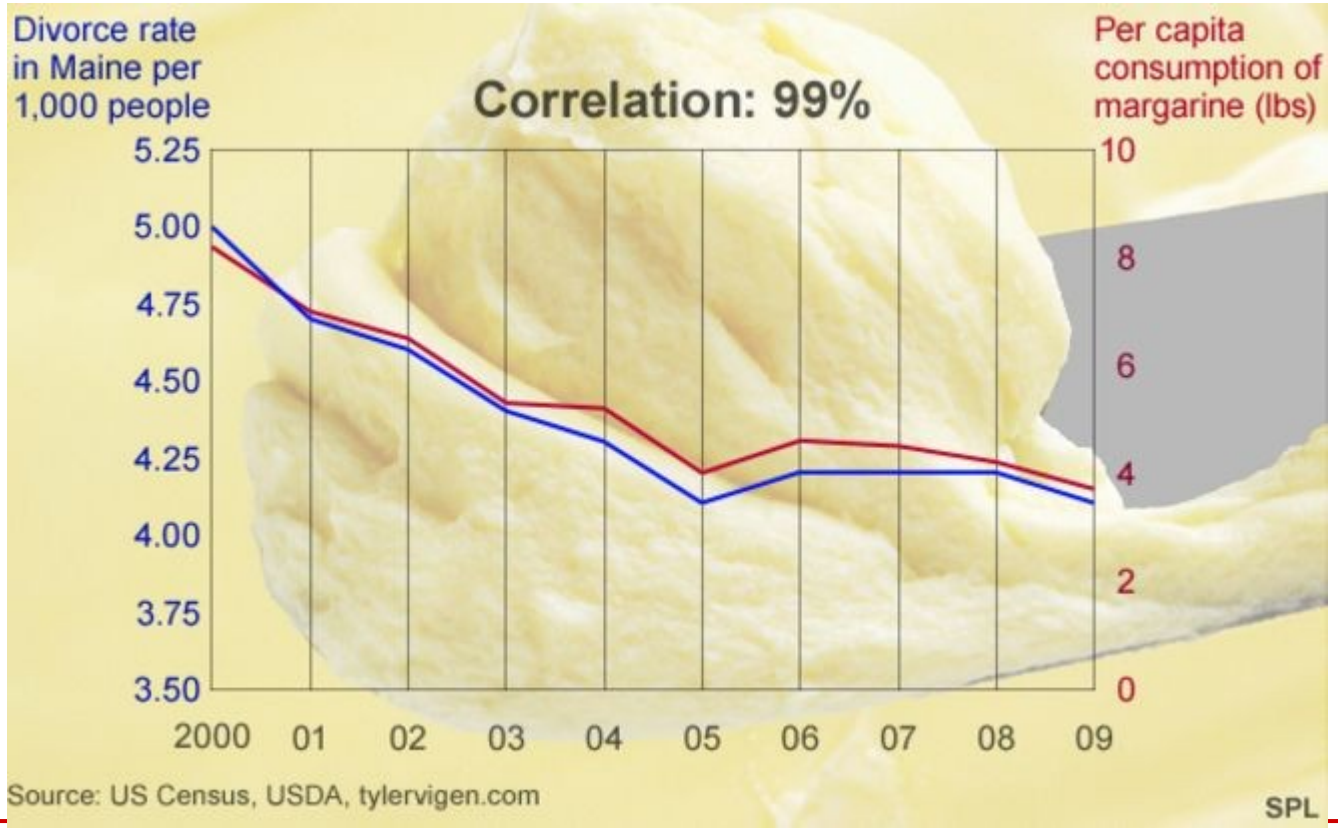
$$\boldsymbol{x}^T = \left(x_1, x_2, \ldots, x_N\right) \qquad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

- Wtedy macierz kowariancji możemy zapisać krótko:

$$C = E\left(\left(\boldsymbol{X} - \hat{\boldsymbol{x}}\right)\left(\boldsymbol{X} - \hat{\boldsymbol{x}}\right)^T\right)$$

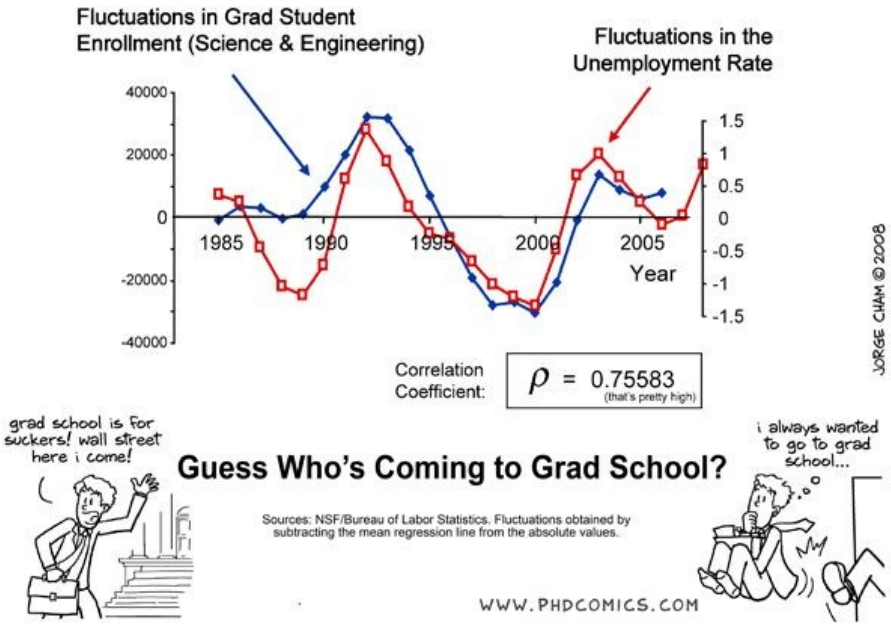
Uwagi do interpretacji wsp. korelacji

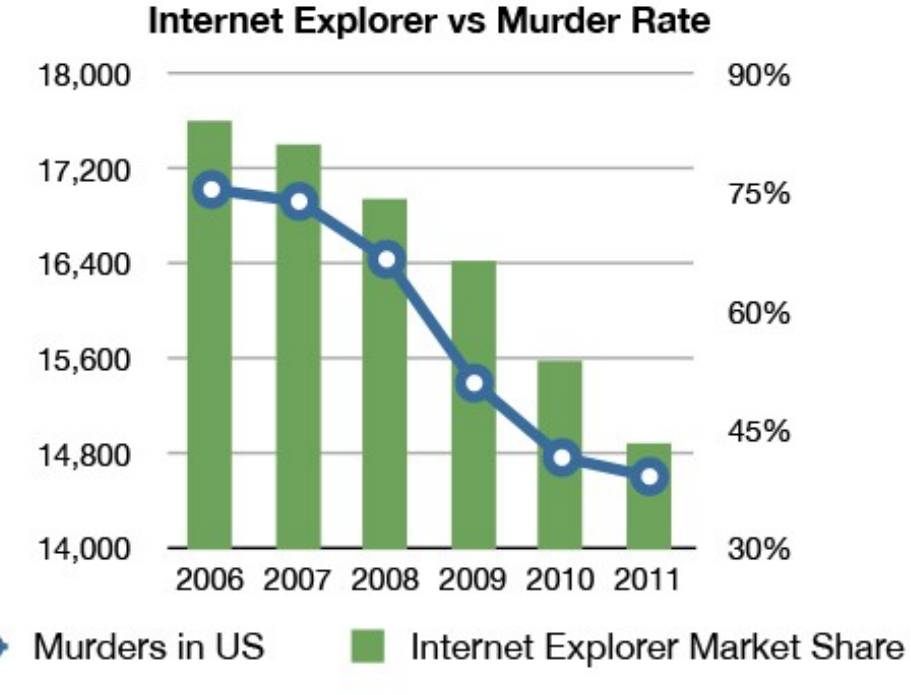
- Zerknijmy na taką zależność:
  - konsumpcja margaryny w czasie i liczba rozwodów w stanie Maine na 1000 osób
  - współczynnik korelacji liniowej wynosi aż 0,99!
  - wynika to z faktu, że obie zmienne są **skorelowane ze wspólną zmienną – czasem** a nie same ze sobą

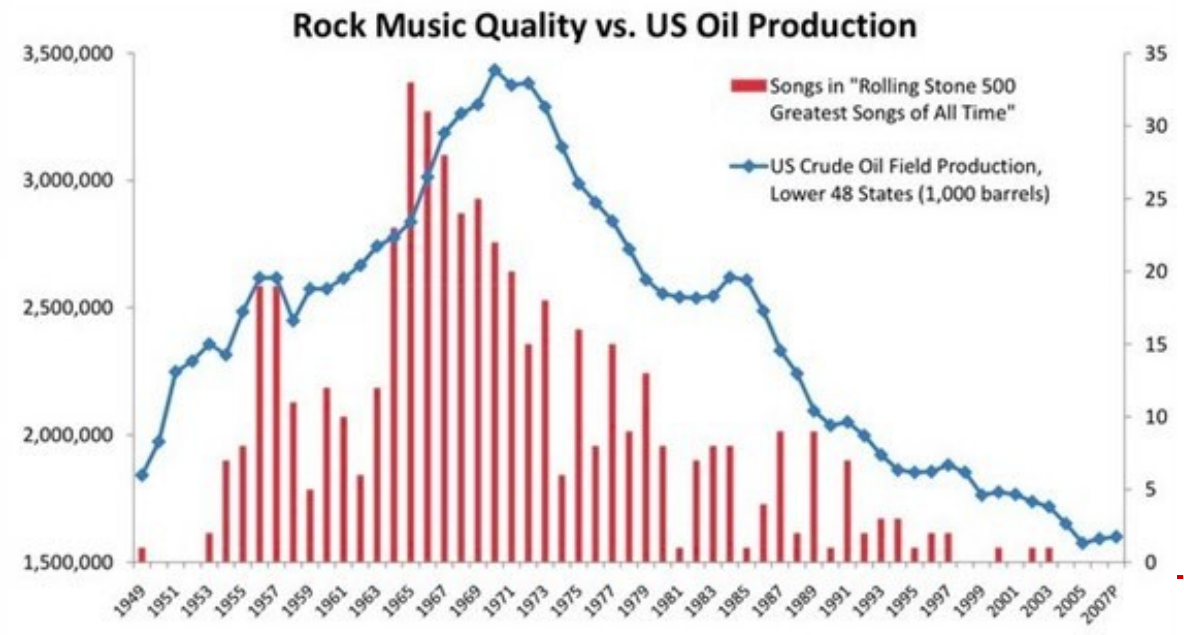


Uwagi do interpretacji wsp. korelacji

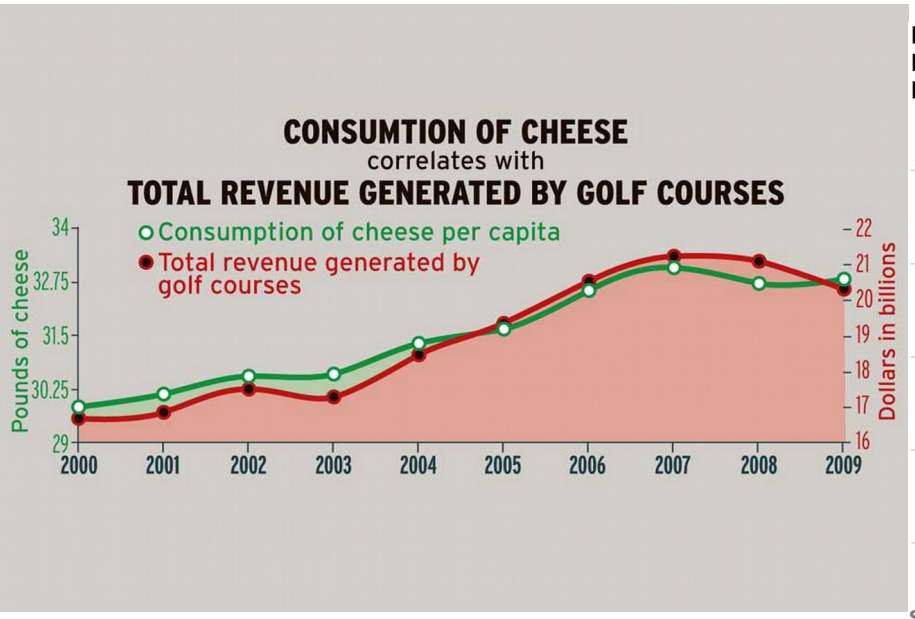
- Podobne przykłady:
- Korelację dwóch wielkości zależnych od czasu można symulować przez **błądzenie losowe** (**random walk**):
  - im dłuższy przedział czasowy tym korelacja się zwiększa
- Wygooglaj “spurious correlations”

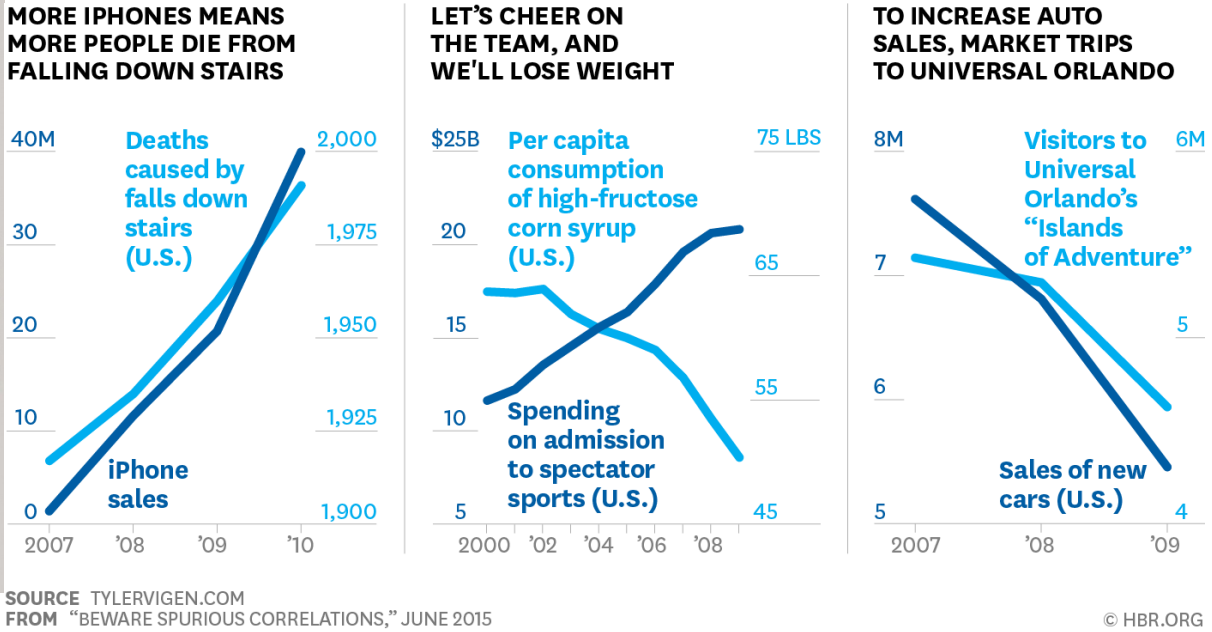




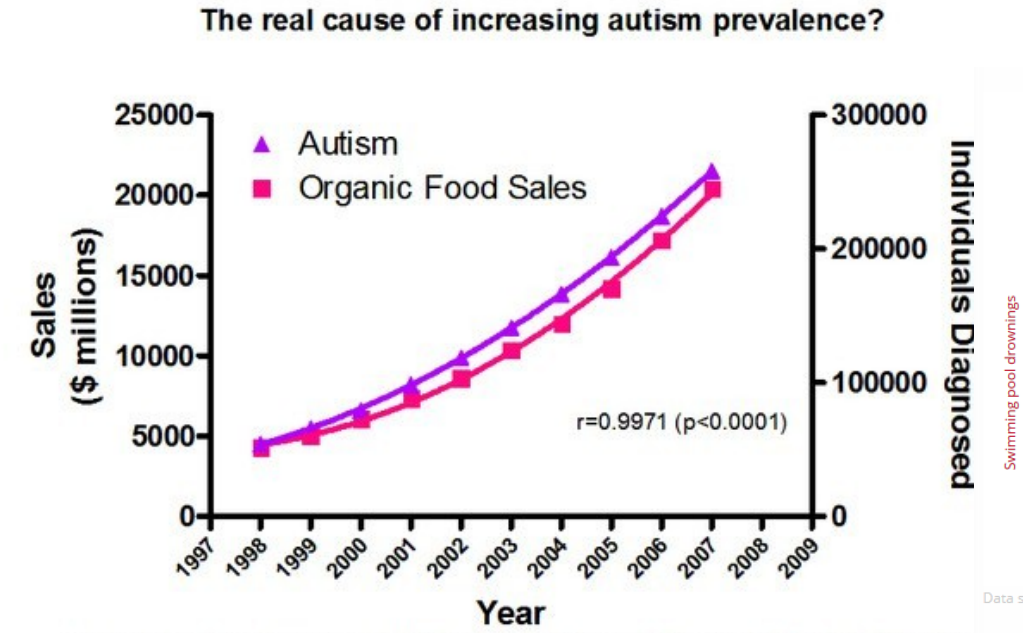


Uwagi do interpretacji wsp. korelacji

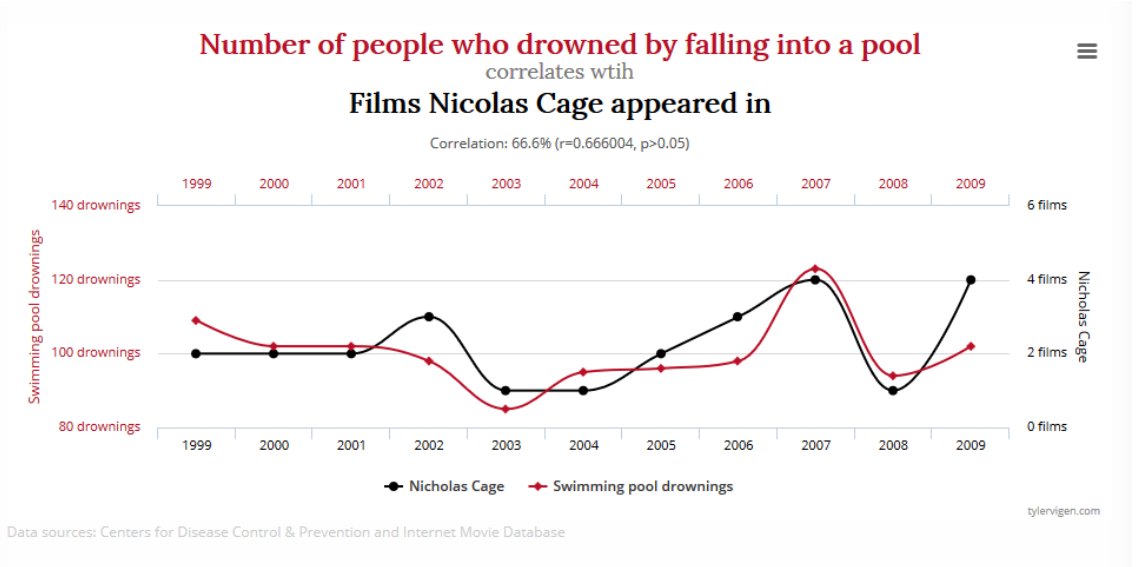




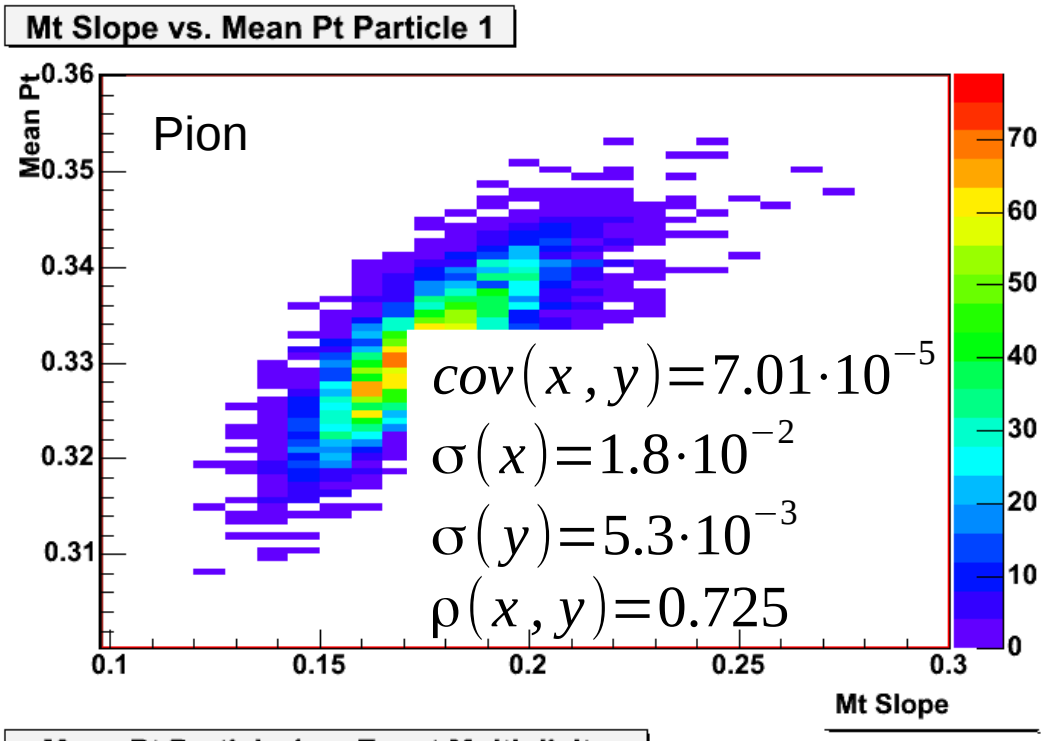
SOURCE TYLERVIGEN.COM
FROM "BEWARE SPURIOUS CORRELATIONS," JUNE 2015

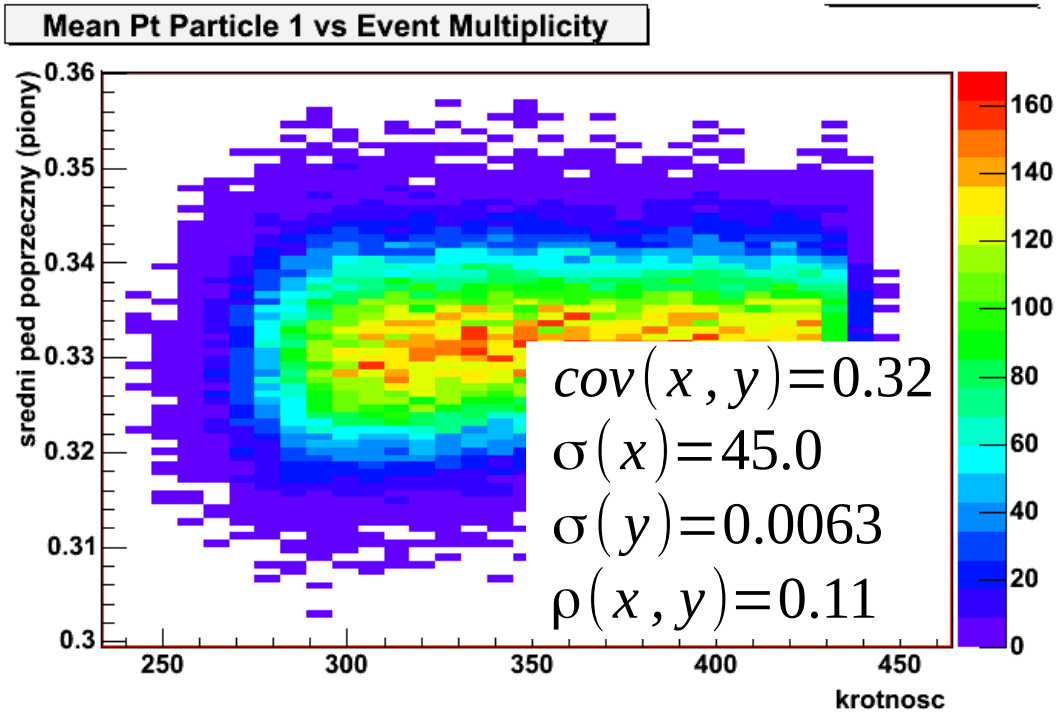


Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

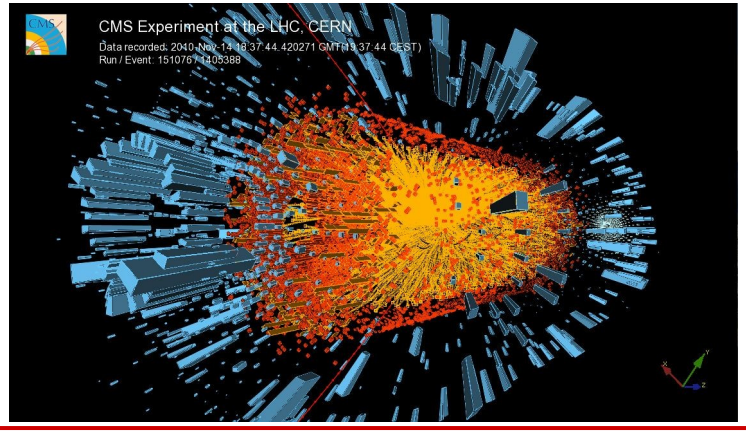


Uwagi do interpretacji wsp. korelacji





- Przykłady rzeczywistych korelacji eksperymentalnych:
  - góra – zależność między średnim pędem cząstki (pionu) a nachyleniem rozkładu pędowego rozkładu pędu
  - dół – praktycznie brak korelacji pomiędzy żrednim pędem cząstki (pionu) a liczbą cząstek produkowanych w zderzeniu



Uwagi do interpretacji wsp. korelacji