

# Komputerowa analiza danych doświadczalnych

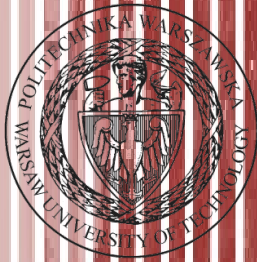
Wykład 3

8.03.2019

dr inż. Łukasz Graczykowski

[lukasz.graczykowski@pw.edu.pl](mailto:lukasz.graczykowski@pw.edu.pl)

*Semestr letni 2018/2019*



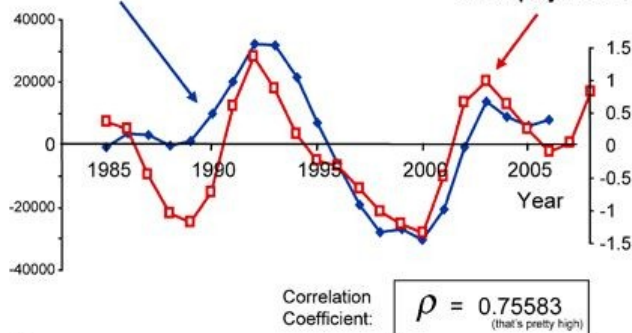
# Dwuwymiarowe rozkłady zmiennych losowych

# 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?

Fluctuations in Grad Student Enrollment (Science & Engineering)

Fluctuations in the Unemployment Rate



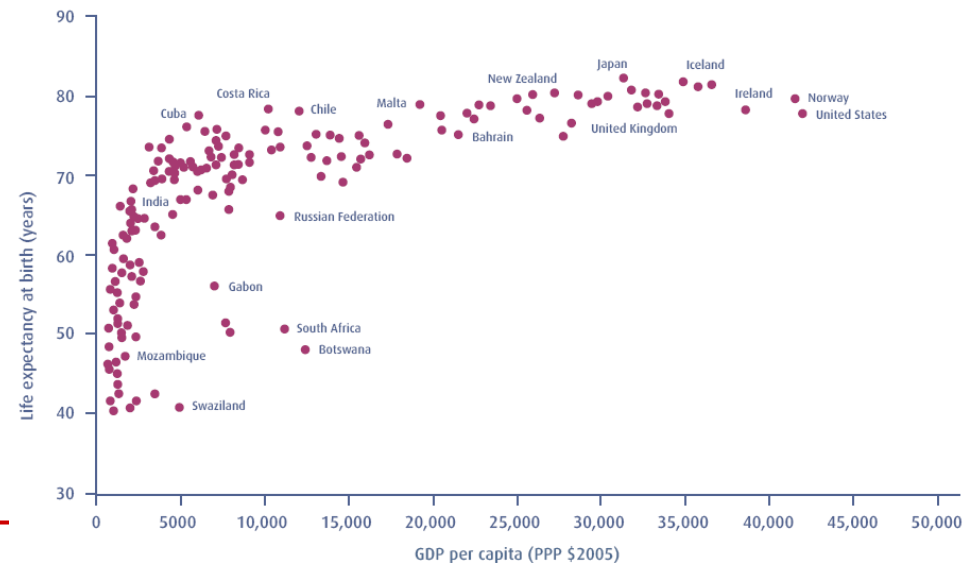
JORGE CHAM © 2008

Guess Who's Coming to Grad School?

Sources: NSF/Bureau of Labor Statistics. Fluctuations obtained by subtracting the mean regression line from the absolute values.

WWW.PHDCOMICS.COM

Figure 8 Life expectancy at birth vs average annual income<sup>16</sup>



# Rozkład i dystrybuanta 2D

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

$$f(x,y) = P(X=x, Y=y)$$

$$p_{ij} = P(X=x_i, Y=y_j)$$

- rozkład prawdopod. jest unormowany

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

- rozkład dyskretny:  $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} p_{ij} = 1$

- Dystrybuanta:**

$$F(x,y) = P(x \leq X, y \leq Y) = \int_{-\infty}^x \int_{-\infty}^y f(x', y') dx' dy'$$

$$F(x) = \sum_{i: x_i \leq x} \sum_{j: y_j \leq Y} p_{ij}$$

- Jeżeli dystrybuanta jest funkcją ciągłą obu zmiennych, to:

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

- Prawdopodobieństwo:**

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y) dx dy = F(b,d) - F(a,b)$$

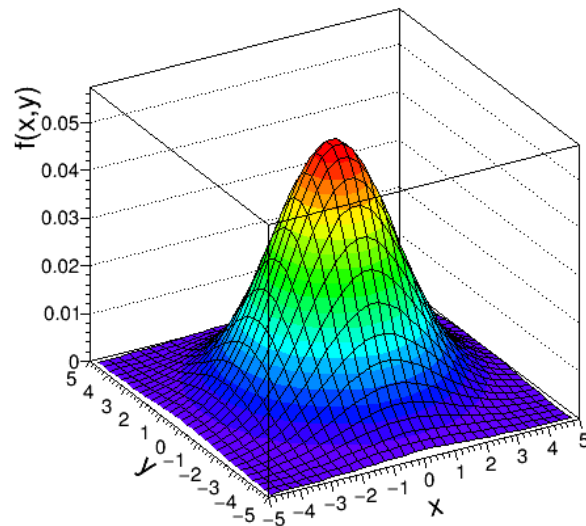
# Zmienna losowa 2D - przykład

- Dwuwymiarowy rozkład Gaussa:

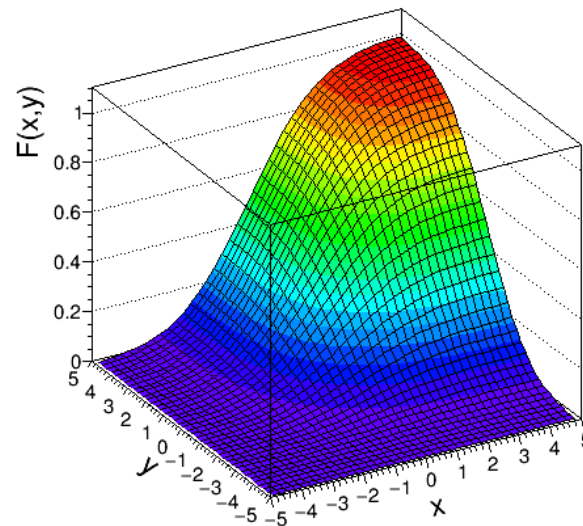
$$f(x, y) = N \cdot \exp \left( - \left( \frac{(x - \hat{x})^2}{2\sigma_x^2} + \frac{(y - \hat{y})^2}{2\sigma_y^2} \right) \right)$$

normalizacja

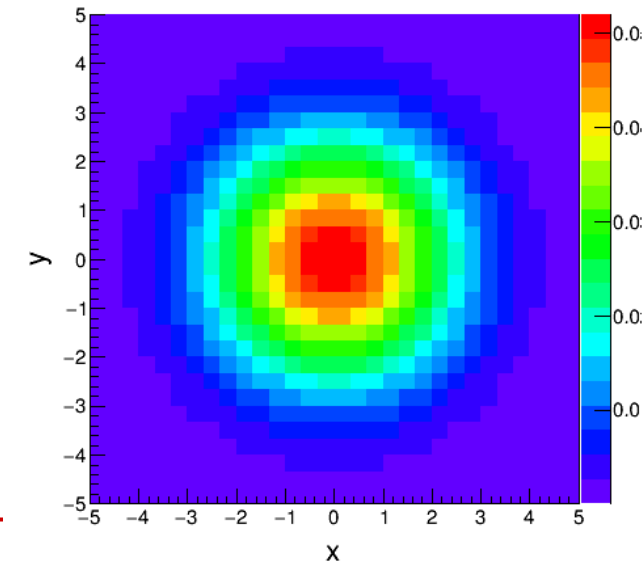
Funkcja gestosci



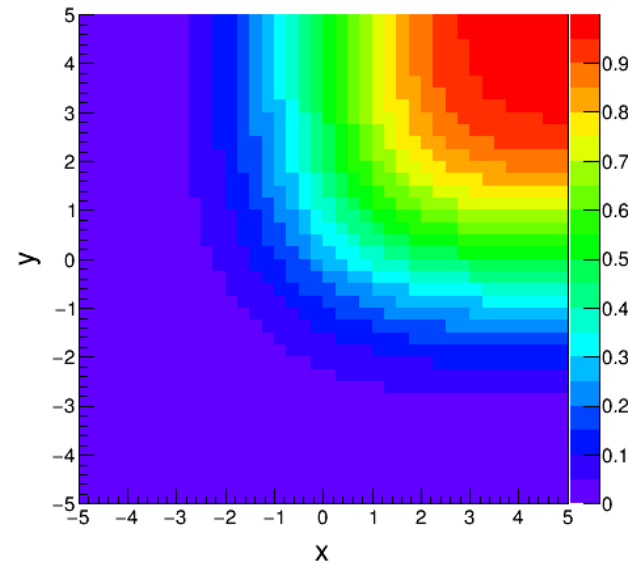
Dystrybuanta



Funkcja gestosci



Dystrybuanta





# 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:**

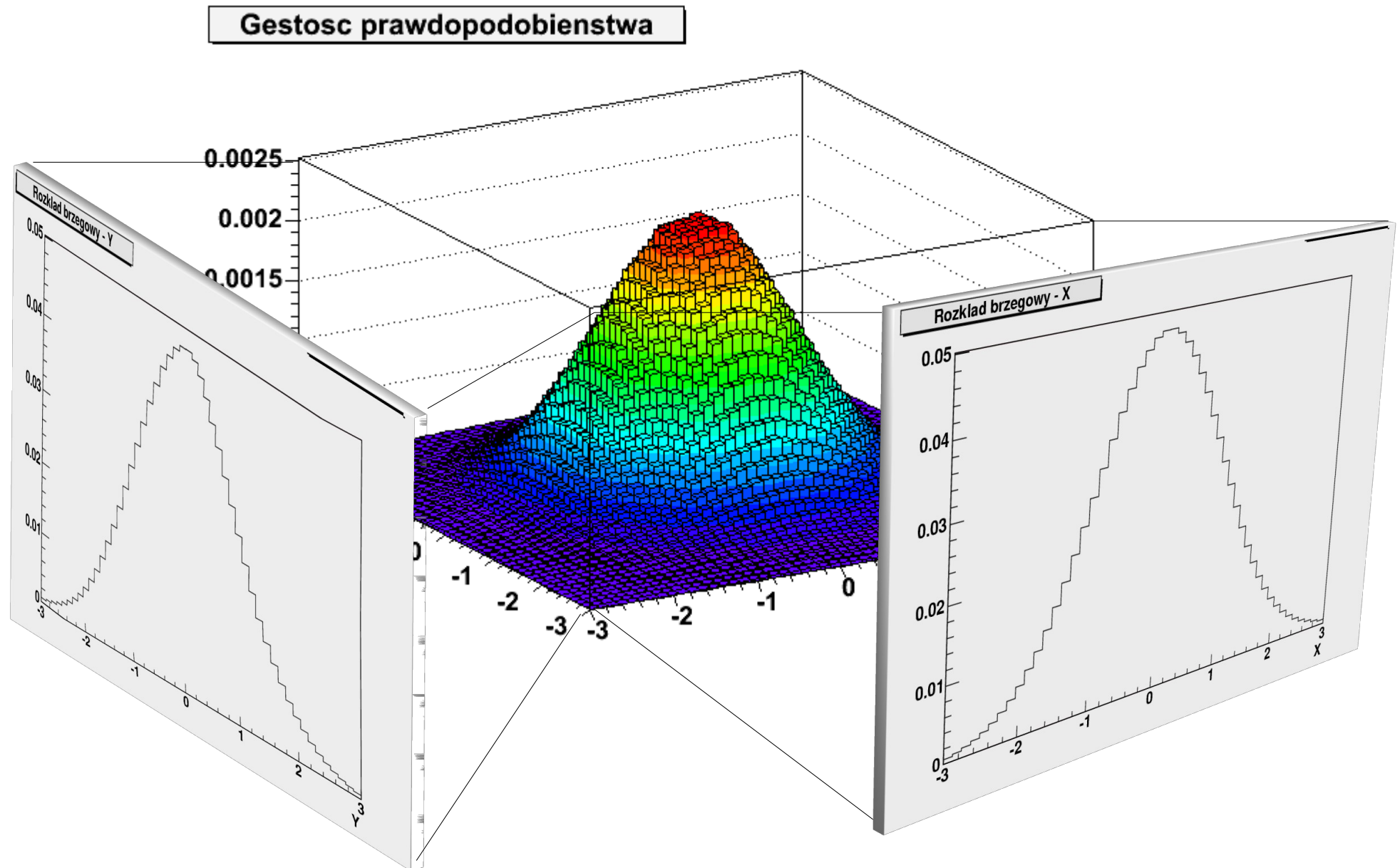
$$P(a \leq X \leq b, -\infty < Y < \infty) = \int_a^b \left[ \int_{-\infty}^{\infty} f(x, y) dy \right] dx = \int_a^b g(x) dx$$
$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \qquad h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

- **Dystrybuanty brzegowe:**

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

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

# Rozkłady (gęstości) brzegowe - przykład



# Niezależność zmiennych

- **Prawdopodobieństwo warunkowe:**

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

- Dla zdarzeń niezależnych:

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

- Analogicznie dla **zmiennych losowych niezależnych**:

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

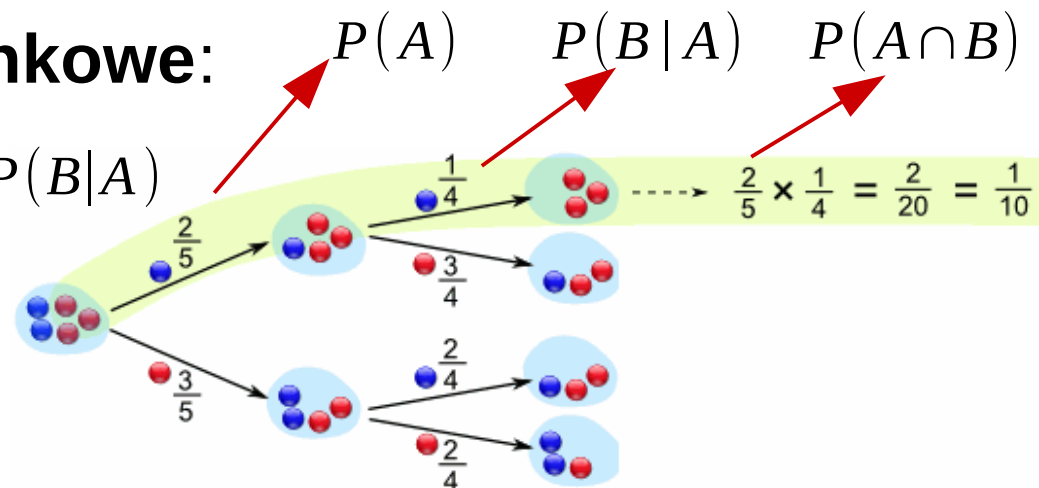
- Warunkowa gęstość prawdopodobieństwa:

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

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

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

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





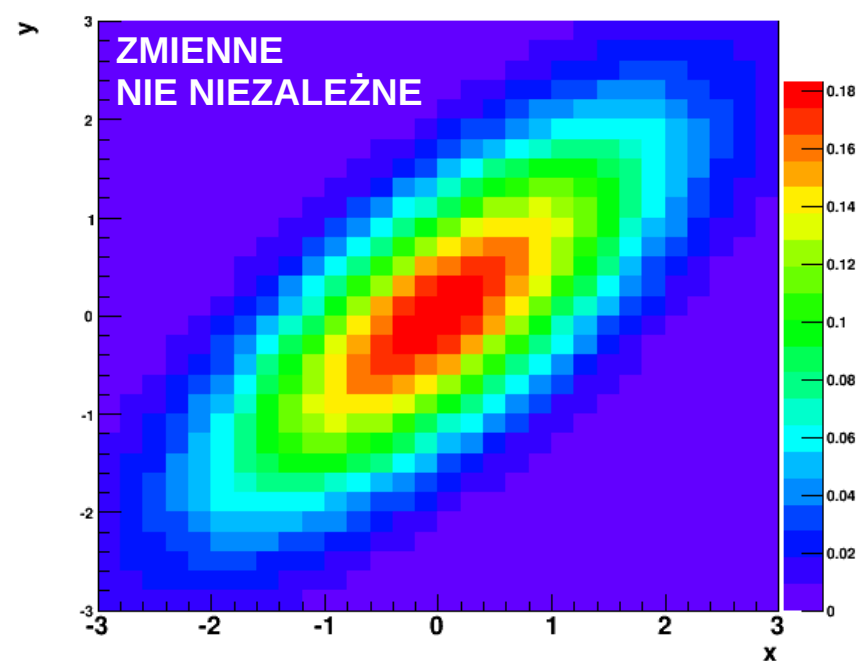
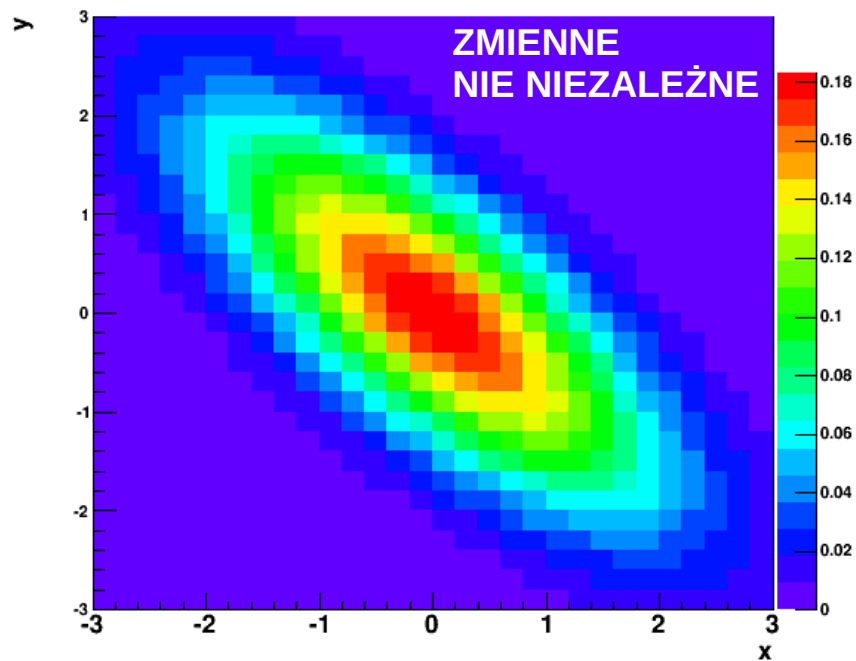
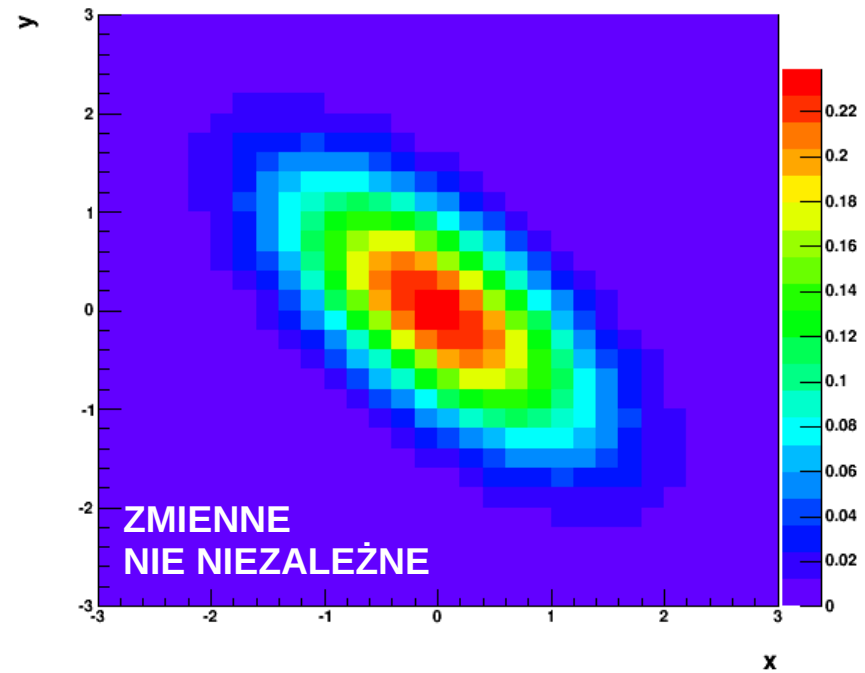
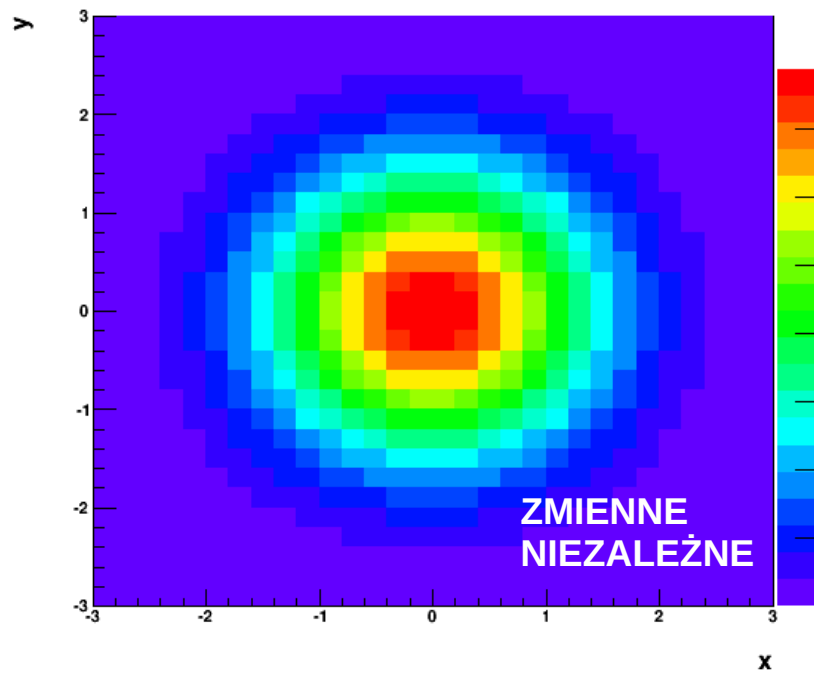
# Niezależność zmiennych

- 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)

# Niezależność zmiennych



# Wartość oczekiwana, wariancja, momenty

- Wartość oczekiwana

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

$$E(XY) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} x_i y_j p_{ij}$$

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

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

- Wariancja:

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

- 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  $b$ :

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

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

# Wartość oczekiwana, wariancja, momenty

- Momenty o specjalnym znaczeniu:

$$\mu_{00} = \lambda_{00} = 1$$

$$\mu_{10} = \mu_{01} = 0$$

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

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

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

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

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

- Wariancja dla zmiennej  $aX + bY$ :

$$\sigma^2(a \cdot X + b \cdot Y) = E(((a \cdot X + b \cdot Y) - E(a \cdot X + b \cdot Y)))^2 = a^2 \sigma^2(X) + b^2 \sigma^2(Y) + 2ab \cdot \text{cov}(X, Y)$$

- 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

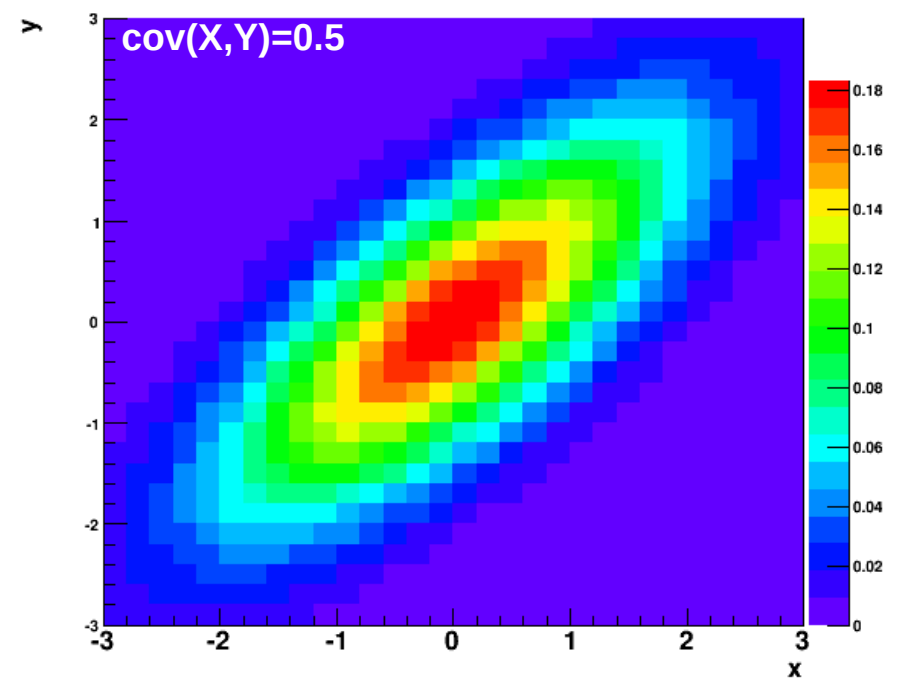
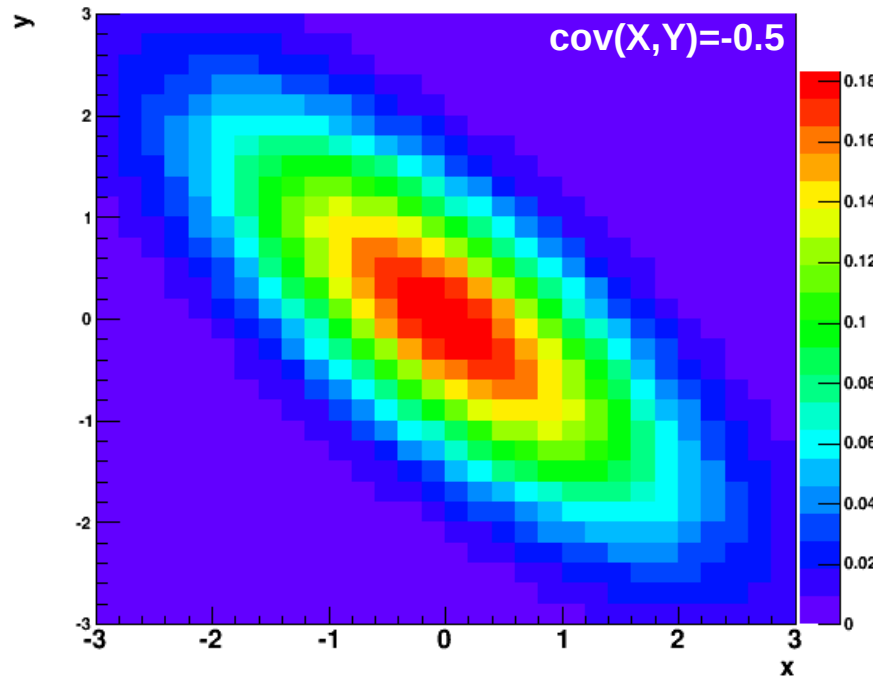
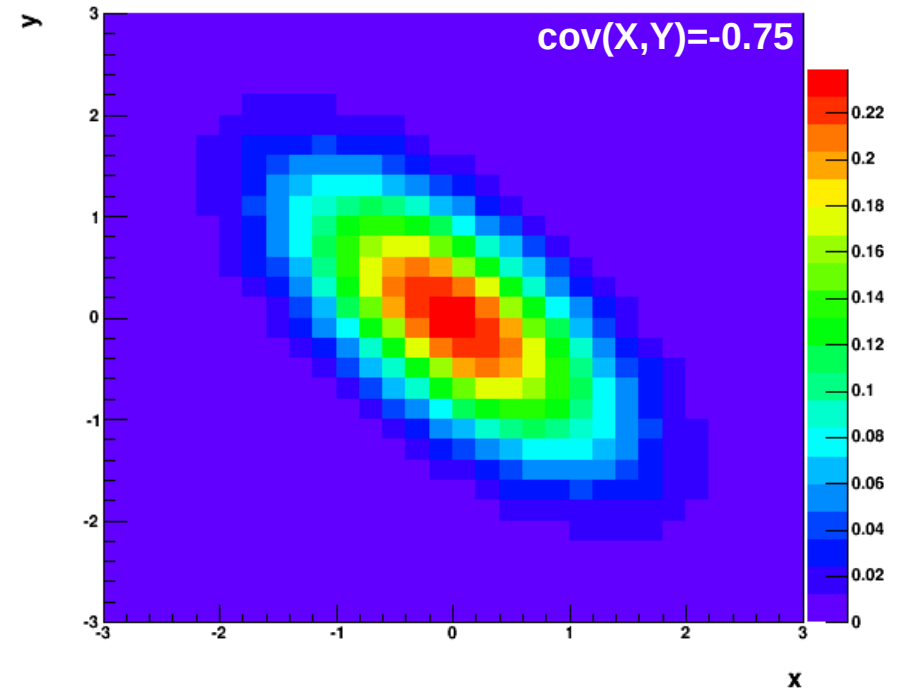
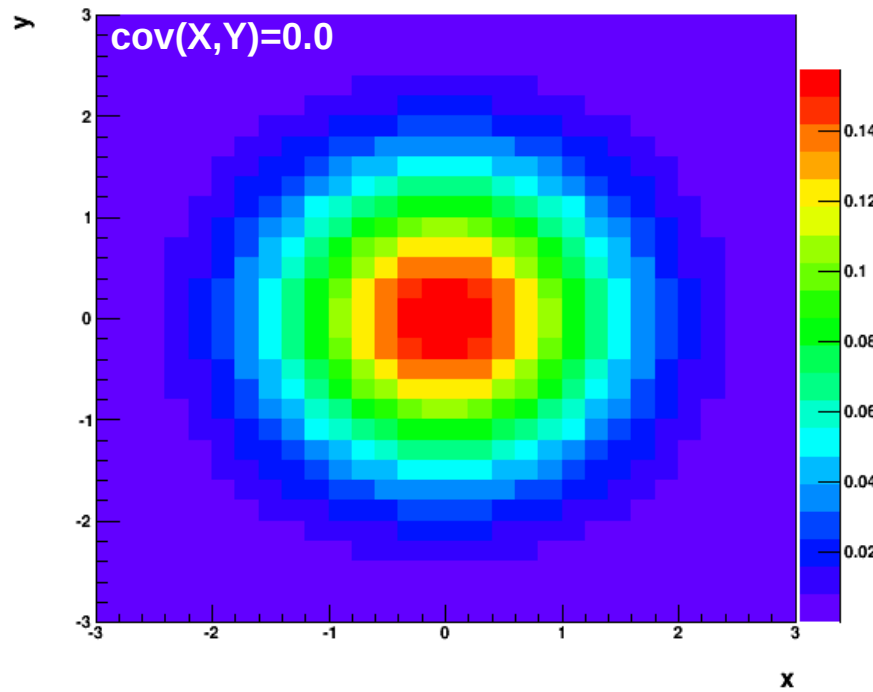
- **Kowariancja**  $cov(X,Y)$  nie ma odpowiednika w przypadku rozkładów 1D
- Określa ona wspólne zmiany zmiennych  $X$  i  $Y$
- Z definicji kowariancji wynika, że:
  - Jeśli dodatnia, wartości  $Y$  na ogół rosną przy wzroście  $X$
  - jest ujemna, wartości  $Y$  na ogół maleją przy wzroście  $X$
  - jeżeli nie ma zależności między  $X$  i  $Y$ , wówczas kowariancja wynosi 0

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

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

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

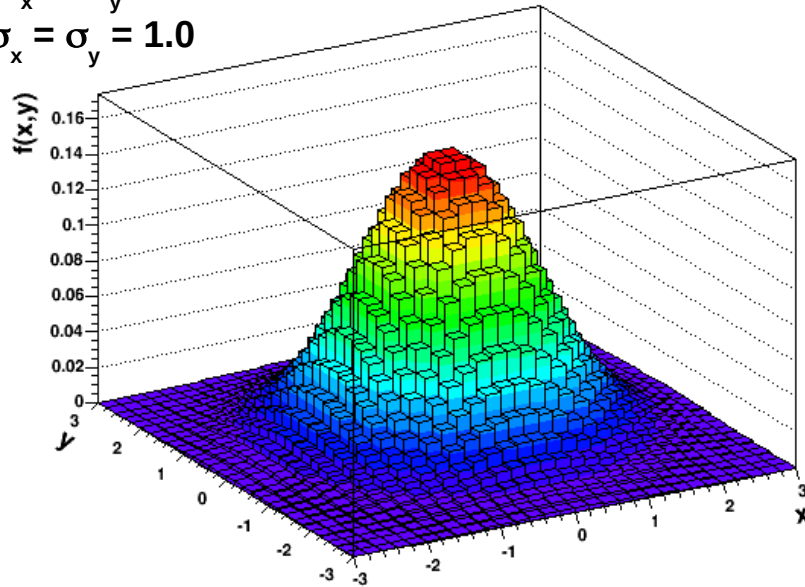
# Kowariancja



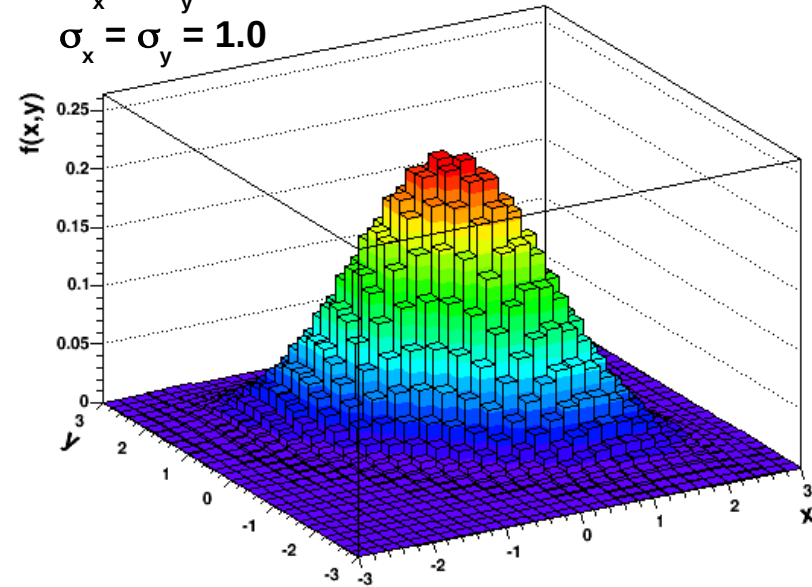


# Kowariancja

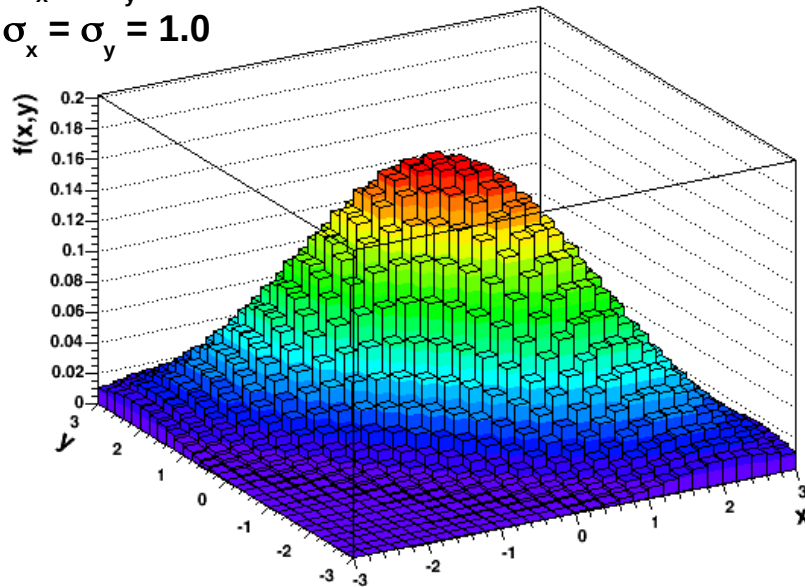
$$\begin{aligned}\text{cov}(X,Y) &= 0.0 \\ m_x &= m_y = 0.0 \\ \sigma_x &= \sigma_y = 1.0\end{aligned}$$



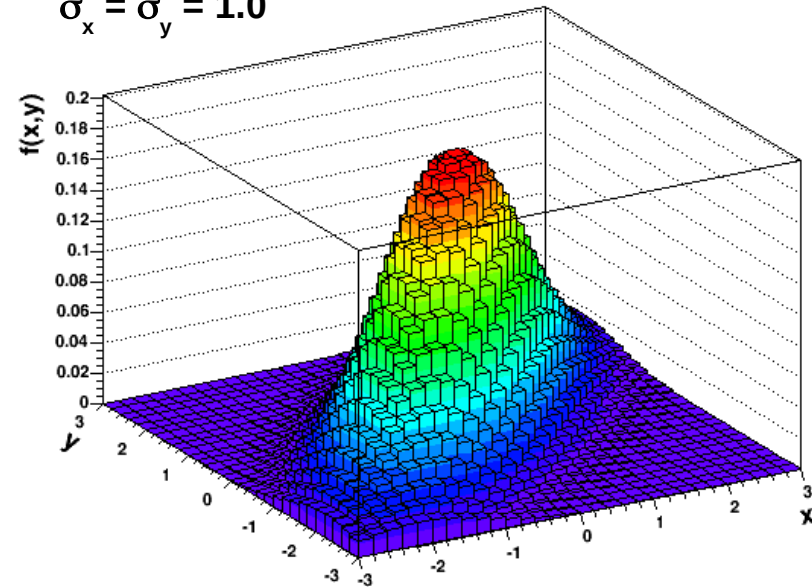
$$\begin{aligned}\text{cov}(X,Y) &= -0.75 \\ m_x &= m_y = 0.0 \\ \sigma_x &= \sigma_y = 1.0\end{aligned}$$



$$\begin{aligned}\text{cov}(X,Y) &= -0.5 \\ m_x &= m_y = 0.0 \\ \sigma_x &= \sigma_y = 1.0\end{aligned}$$



$$\begin{aligned}\text{cov}(X,Y) &= 0.5 \\ m_x &= m_y = 0.0 \\ \sigma_x &= \sigma_y = 1.0\end{aligned}$$

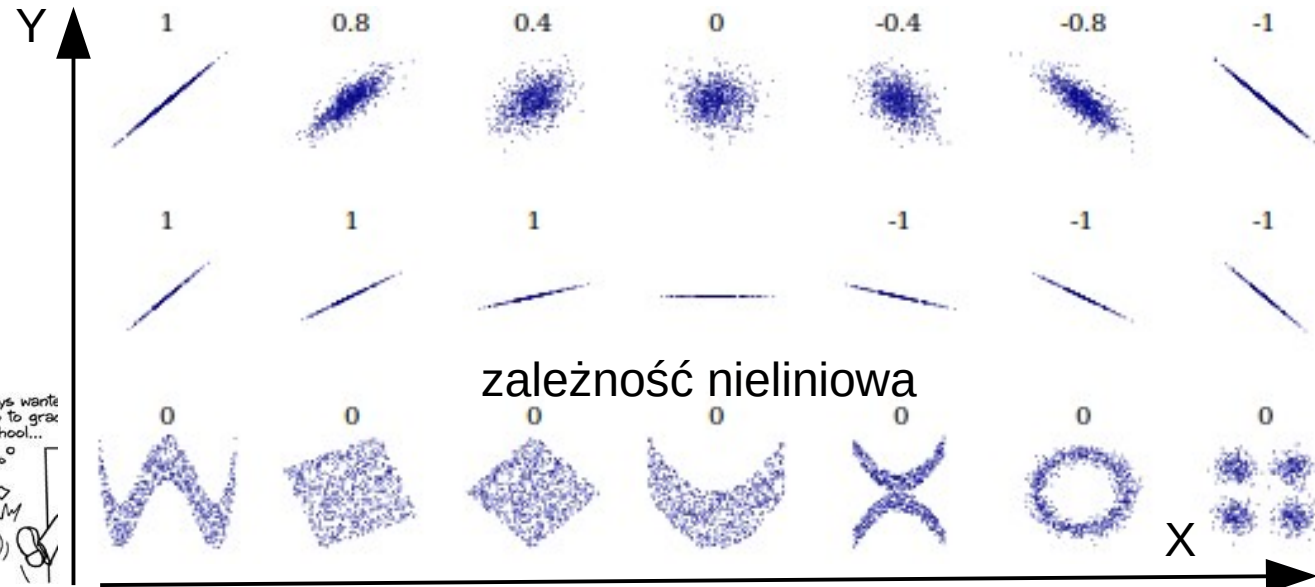
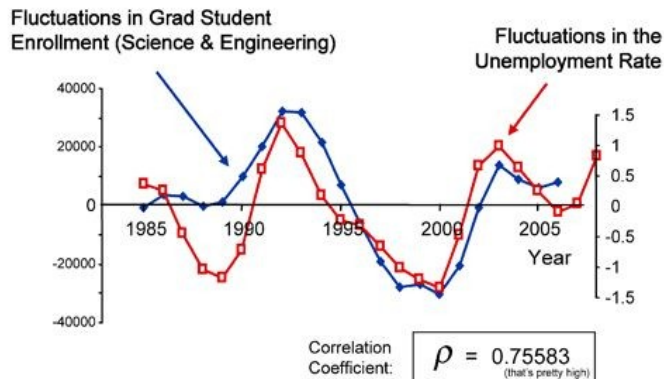


# Współczynnik korelacji

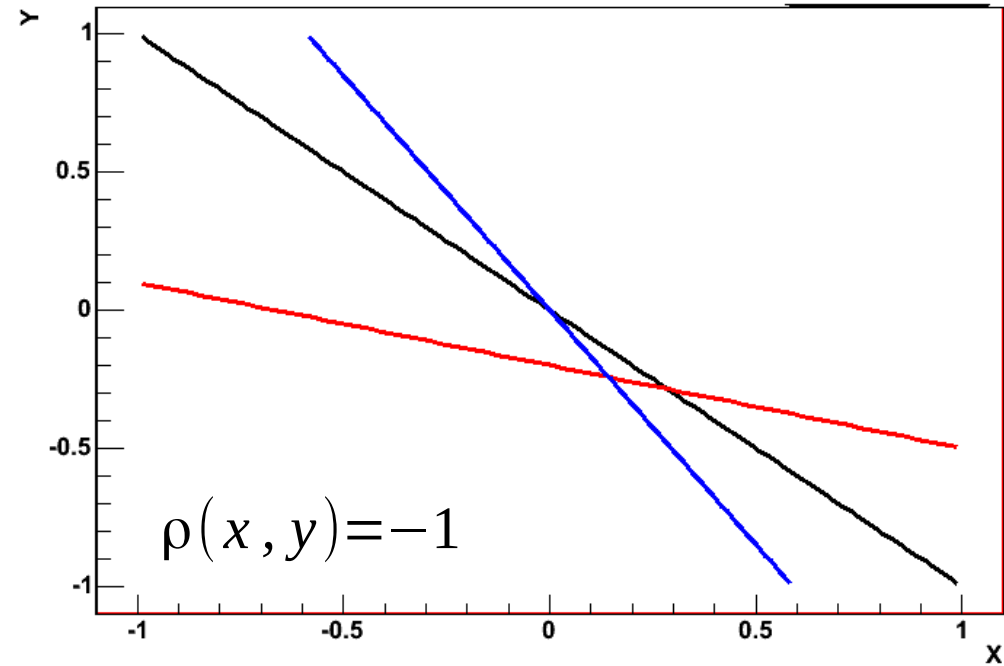
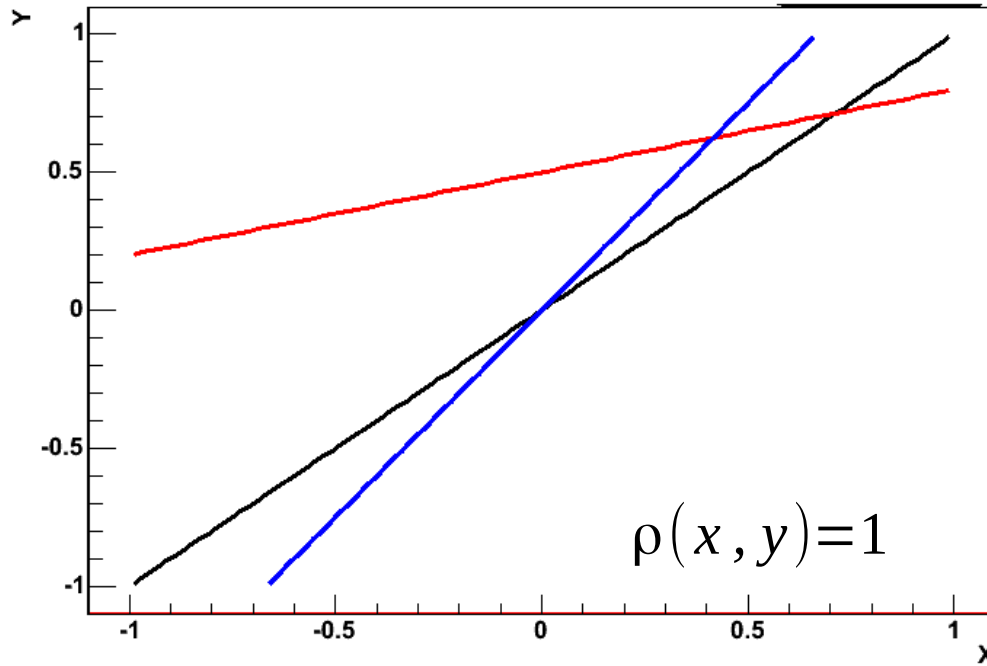
- **Współczynnik korelacji (Pearsona):**

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

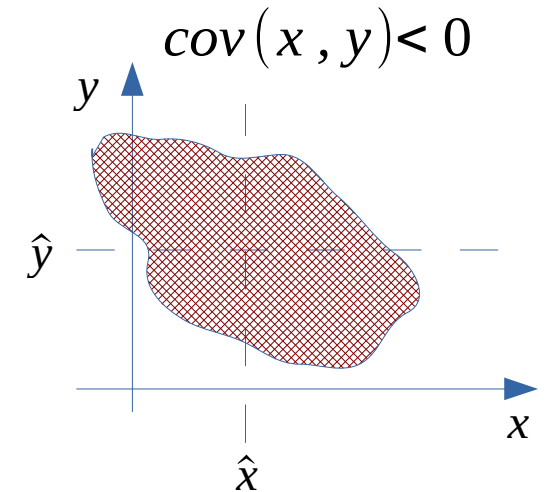
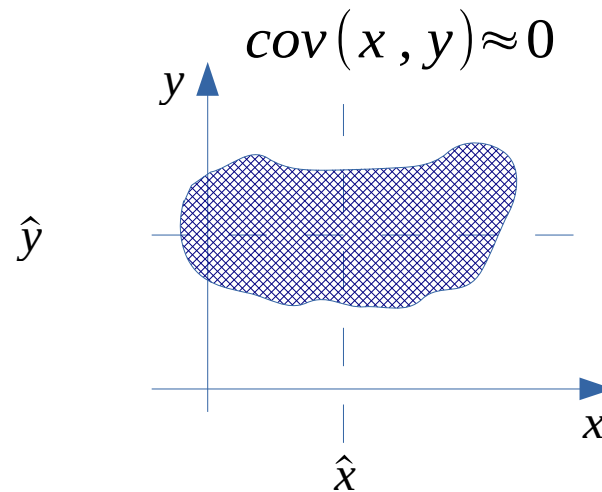
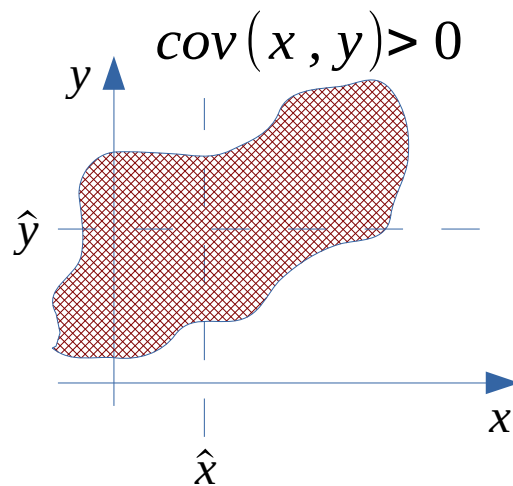
- Kowariancja i współczynnik korelacji to miary zależności liniowej  $X$  oraz  $Y$ :
- Można wykazać (patrz podręcznik Brandt), że:  $-1 \leq \rho(X, Y) \leq 1$ 
  - **współczynnik korelacji jest zatem wielkością znormalizowaną**
- Współczynnik korelacji ocenia **jedynie liniową zależność** (niekoniecznie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



# Kowariancja



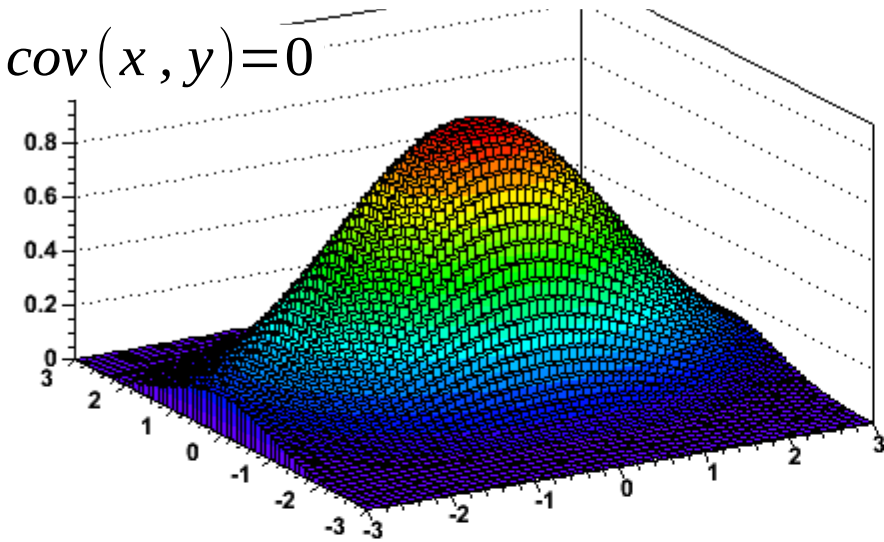
Rozkłady o maksymalnej kowariancji (wsp. korelacji)



# Korelacja a niezależność zmiennych

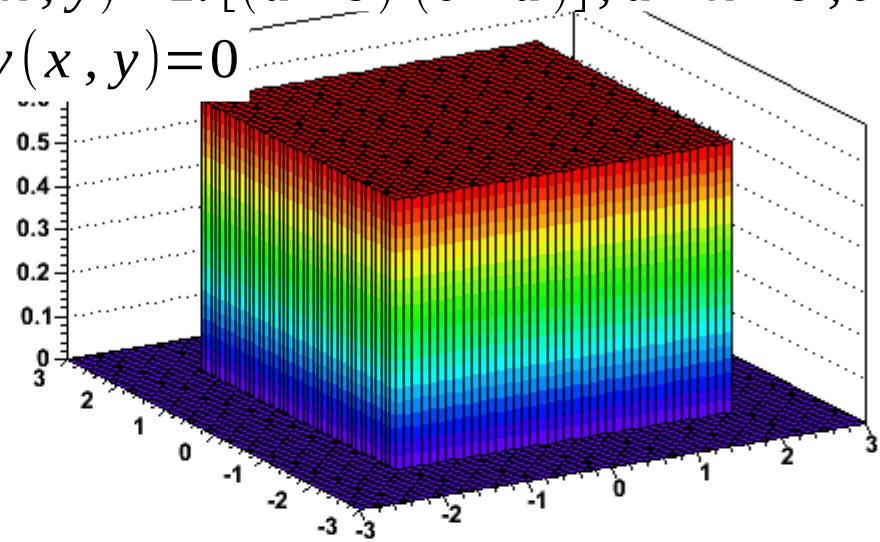
$$f(x, y) = \text{gaus}(x) \cdot \text{gaus}(y)$$

$$\text{cov}(x, y) = 0$$



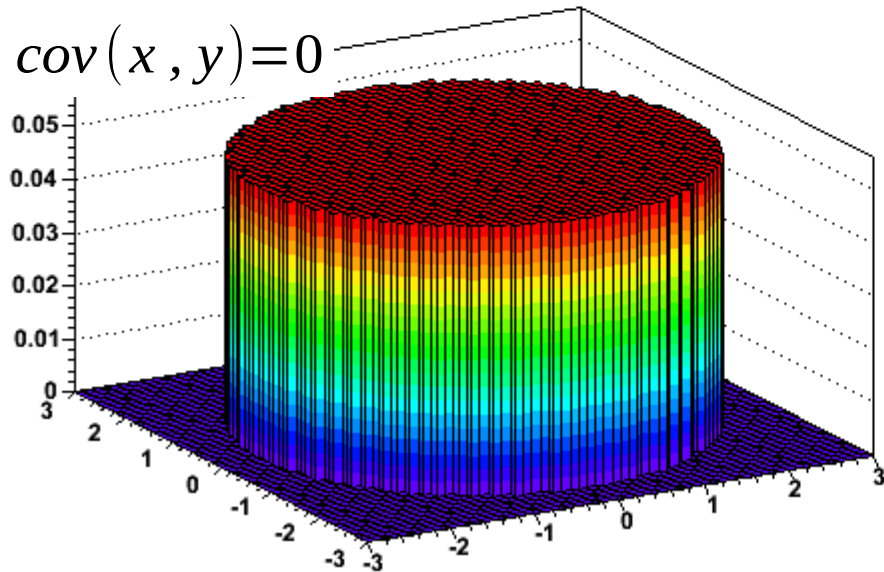
$$f(x, y) = 1/[(a-b) \cdot (c-d)]; a < x < b, c < y < d$$

$$\text{cov}(x, y) = 0$$



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

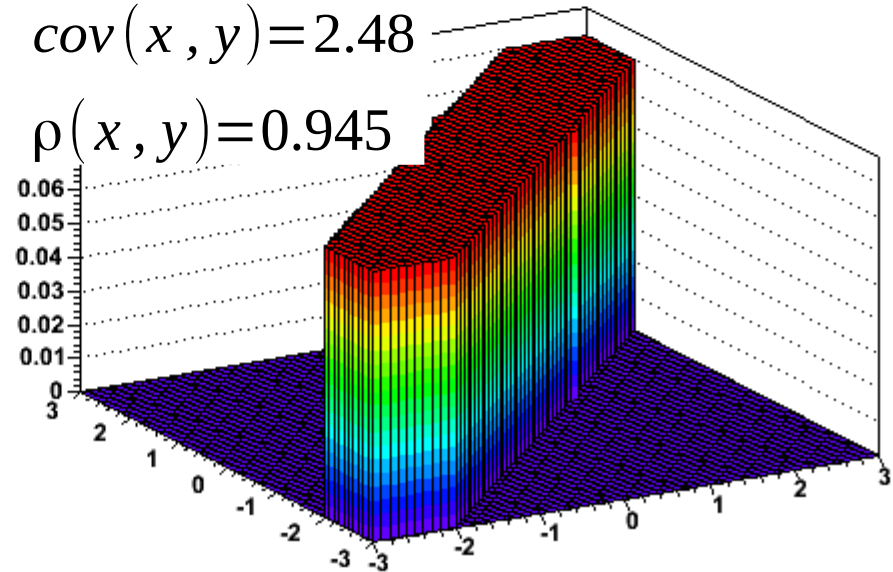
$$\text{cov}(x, y) = 0$$



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

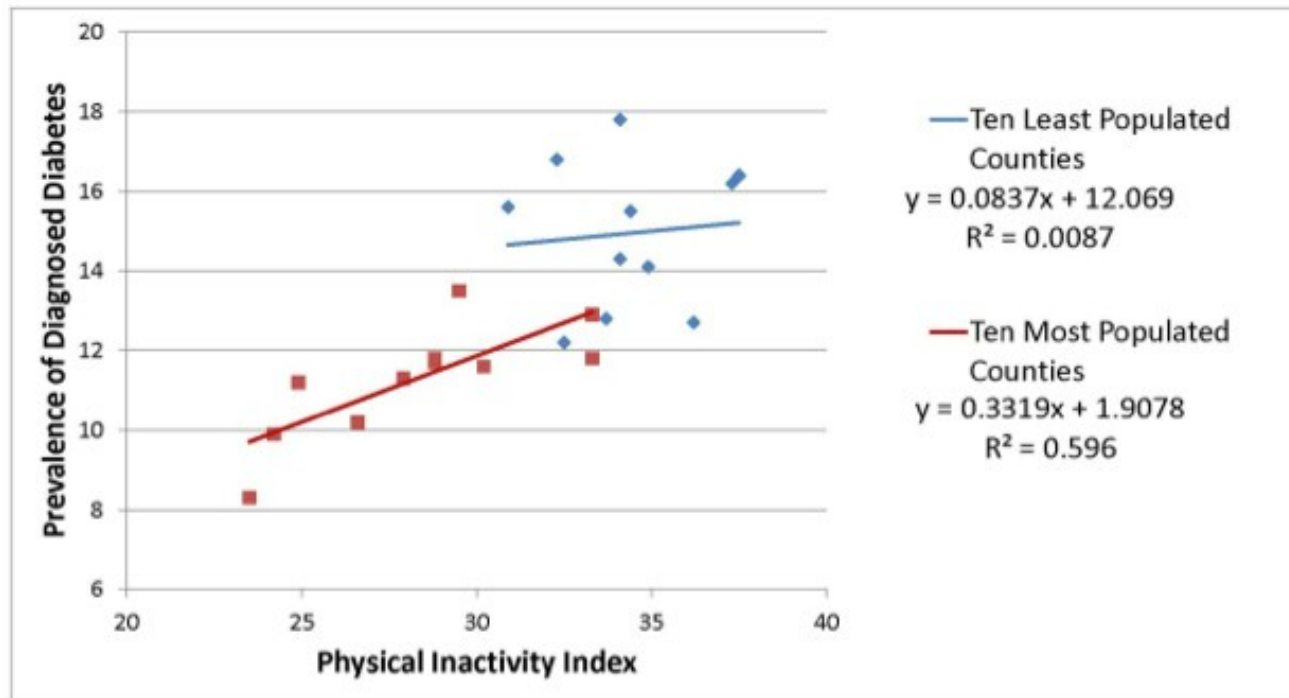
$$\text{cov}(x, y) = 2.48$$

$$\rho(x, y) = 0.945$$



# Przykład – aktywność fiz. a cukrzyca

- Im mniejszy Physical Inactivity Index – tym większy wysiłek fizyczny
- W hrabstwach USA o małej populacji inne czynniki również mogą wpływać na cukrzycę
- Wysoka korelacja między aktywnością fizyczną a cukrzycą w krajach o wysokiej populacji (np. spowodowana miejskim trybem życia)



# Rozkład i dystrybuanta N-wymiarowa

- Mamy  $N$  zmiennych losowych:

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

- Rozkład prawdopodobieństwa:**

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

- rozkład prawdopod. jest unormowany

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

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

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

- Jeżeli dystrybuanta jest funkcją ciągłą, to:

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

- Gęstość brzegowa zmiennej  $x_r$ :

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



# Wartość oczek., 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  $H(x_1, x_2, \dots, x_N)$ :

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

- Wariancja zmiennej  $X_r$ :

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

- 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  $x_j$ :

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

# Momenty

- Przez analogię definiujemy również momenty (zwykłe):

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

- W szczególności, wartości oczekiwane:

$$\lambda_{100\dots 0} = E(X_1) = \hat{x}_1$$

$$\lambda_{010\dots 0} = E(X_2) = \hat{x}_2$$

$$\lambda_{000\dots N} = E(X_N) = \hat{x}_N$$

- Momenty centralne:

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

- W szczególności, wariancje:

$$\mu_{200\dots 0} = E((X_1 - \hat{x}_1)^2) = \sigma^2(X_1)$$

$$\mu_{020\dots 0} = E((X_2 - \hat{x}_2)^2) = \sigma^2(X_2)$$

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

Kowariancja:

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

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



# Zapis wektorowy

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

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}$$

- Funkcja gęstości:

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

- Dystrybuanta:

$$F(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$$

- Jeżeli istnieją pierwsze pochodne:

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

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

$$E(H(\mathbf{X})) = \int H(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$



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:

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

- 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:  $E(\mathbf{X}) = \hat{\mathbf{x}}$

# 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  $(\mathbf{X} - \hat{\mathbf{x}})^T$  i  $(\mathbf{X} - \hat{\mathbf{x}})$

$$\mathbf{x}^T = (x_1, x_2, \dots, x_N)$$

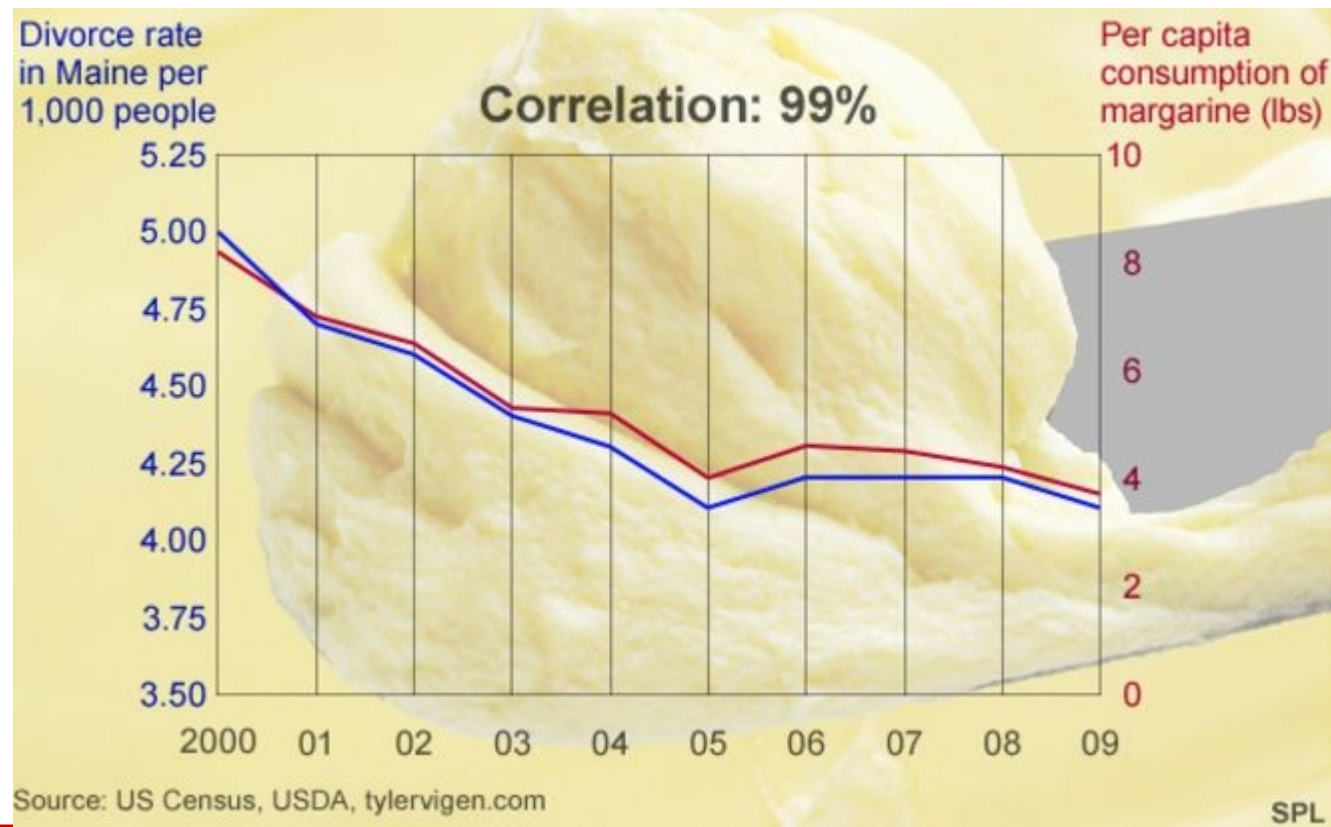
$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

- Wtedy macierz kowariancji możemy zapisać krótko:

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

# 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

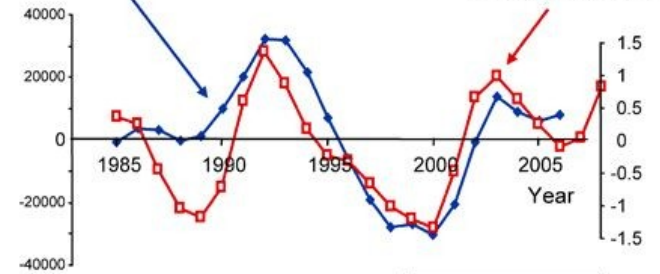


# Uwagi do interpretacji wsp. korelacji

- 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”

Fluctuations in Grad Student Enrollment (Science & Engineering)

Fluctuations in the Unemployment Rate



Correlation Coefficient:  $\rho = 0.75583$   
(that's pretty high)



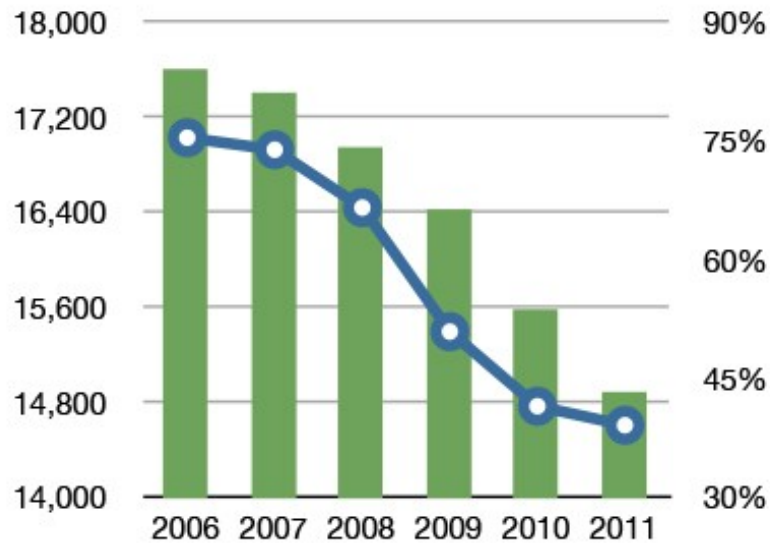
Guess Who's Coming to Grad School?

Sources: NSF/Bureau of Labor Statistics. Fluctuations obtained by subtracting the mean regression line from the absolute values.

WWW.PHDCOMICS.COM



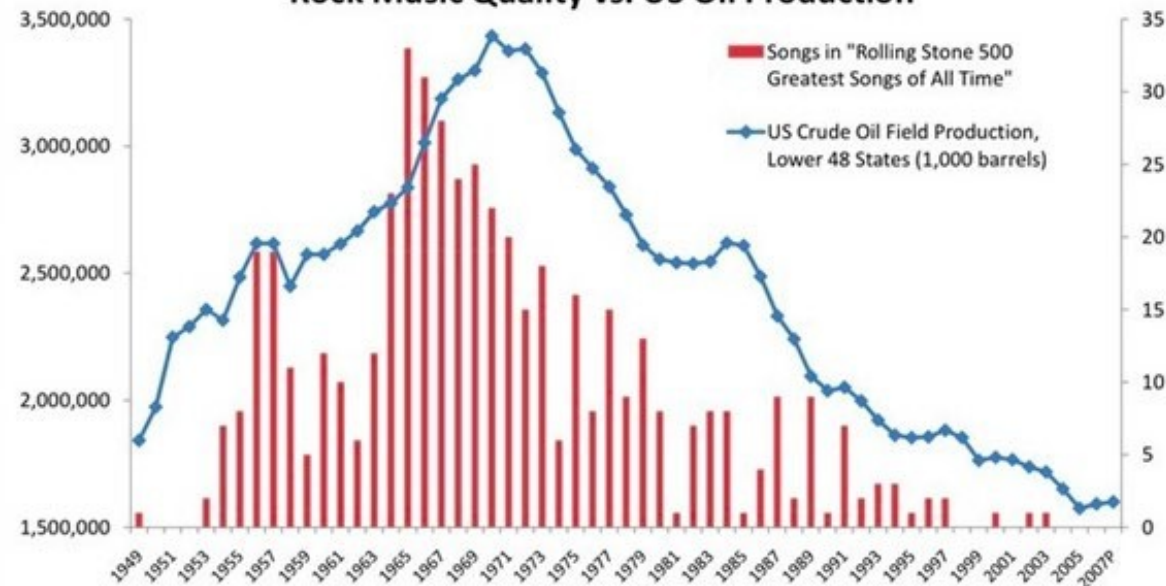
Internet Explorer vs Murder Rate



Murders in US

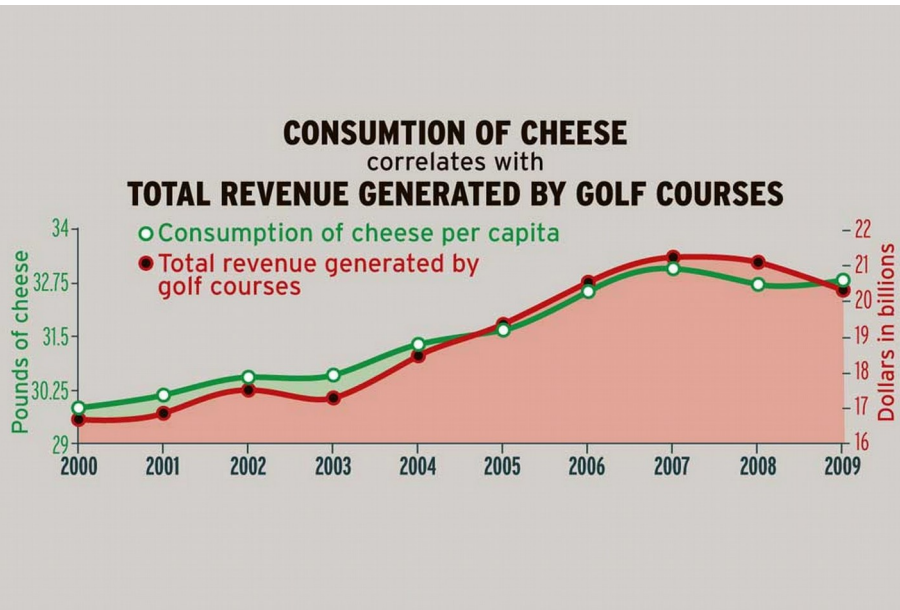
Internet Explorer Market Share

Rock Music Quality vs. US Oil Production

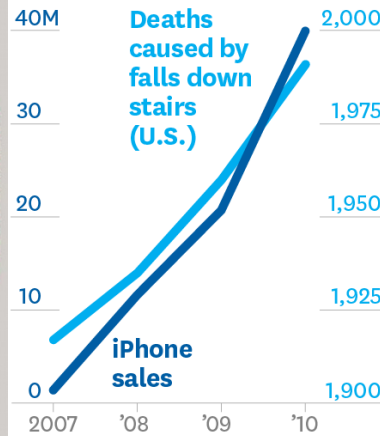




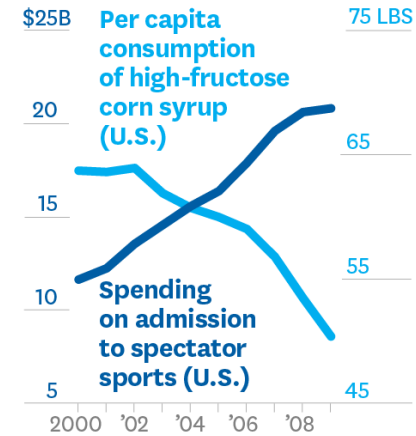
# Uwagi do interpretacji wsp. korelacji



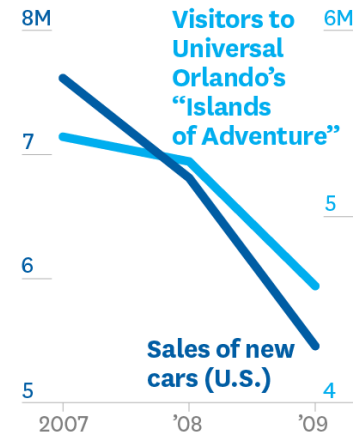
**MORE IPHONES MEANS  
MORE PEOPLE DIE FROM  
FALLING DOWN STAIRS**



**LET'S CHEER ON  
THE TEAM, AND  
WE'LL LOSE WEIGHT**



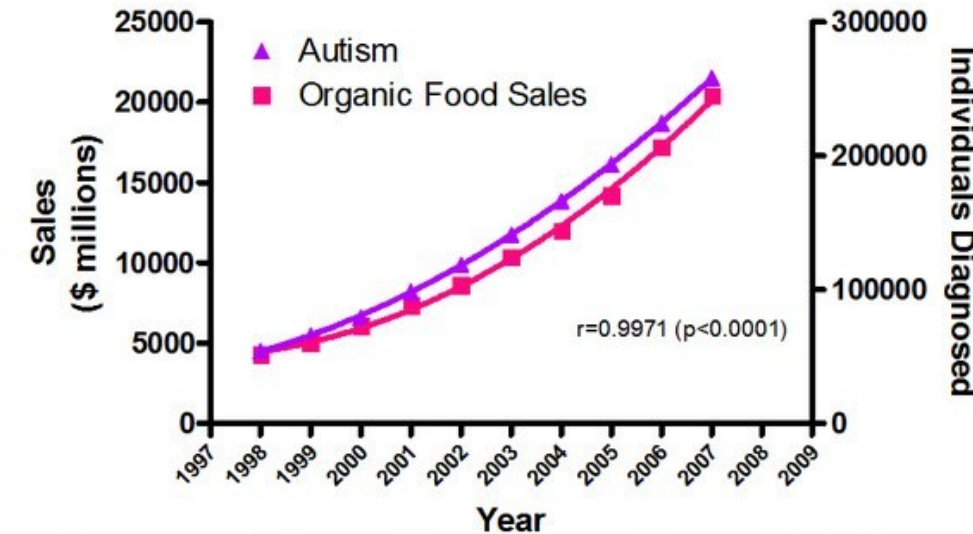
**TO INCREASE AUTO  
SALES, MARKET TRIPS  
TO UNIVERSAL ORLANDO**



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

© HBR.ORG

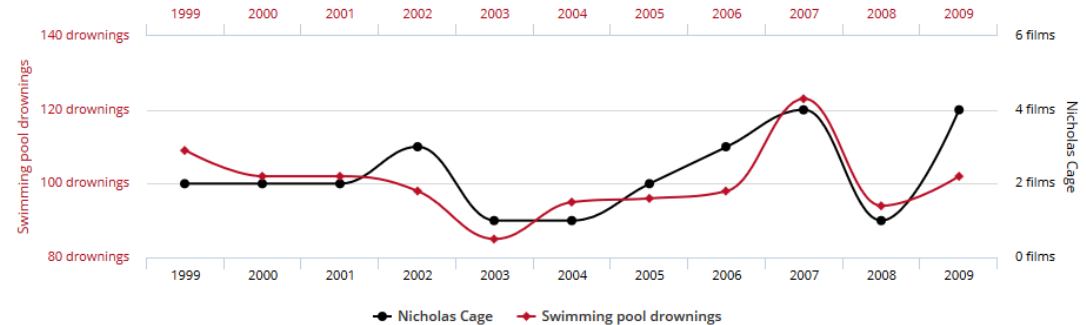
**The real cause of increasing autism prevalence?**



**Number of people who drowned by falling into a pool**

correlates with  
**Films Nicolas Cage appeared in**

Correlation: 66.6% ( $r=0.666004$ ,  $p>0.05$ )



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

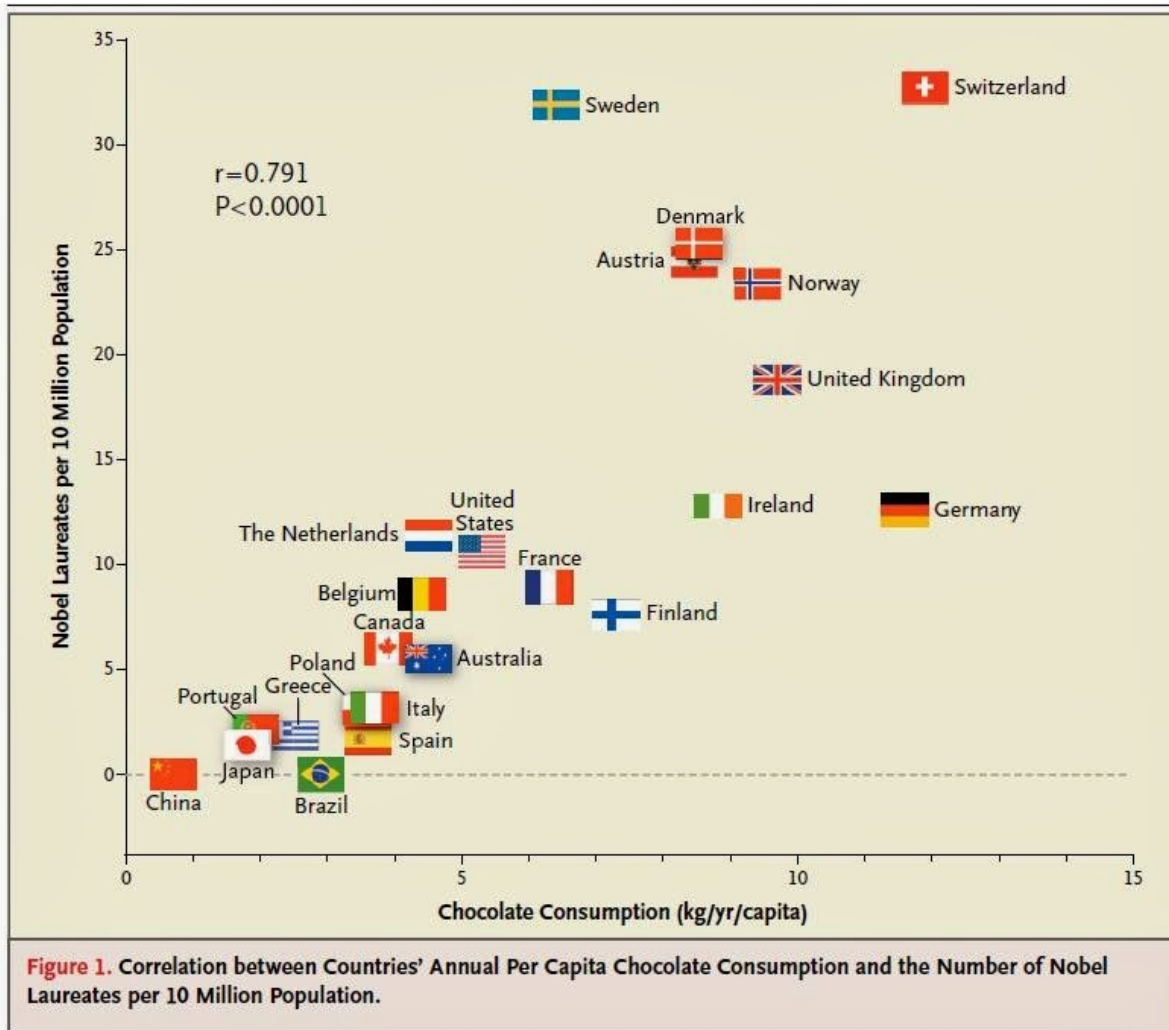
tylervigen.com

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

<http://www.biostat.jhsph.edu/courses/bio621/misc/Chocolate%20consumption%20cognitive%20function%20and%20nobel%20laureates%20%28NEJM%29.pdf>

The New England Journal Medicine 2012; 367:1562-1564, Impact Factor (2015): 59.998!!!! (Nature IF: 38.138)



## CONCLUSIONS

Chocolate consumption enhances cognitive function, which is a sine qua non for winning the Nobel Prize, and it closely correlates with the number of Nobel laureates in each country. It remains to be determined whether the consumption of chocolate is the underlying mechanism for the observed association with improved cognitive function.

Dr. Messerli reports regular daily chocolate consumption, mostly but not exclusively in the form of Lindt's dark varieties.

Disclosure forms provided by the author are available with the full text of this article at NEJM.org.

From St. Luke's–Roosevelt Hospital and Columbia University, New York.

This article was published on October 10, 2012, at NEJM.org.

- Jaki jest “problem” z tym artykułem?



# Uwagi do interpretacji wsp. korelacji

<http://news.nationalpost.com/news/correlation-and-causation-the-good-statistical-lesson-behind-study-associating-drinking-milk-and-nobel-prizes>

“It’s only tongue in cheek,” said Franz Messerli, the New York doctor who did the chocolate analysis. “Those among us who are in medicine and do research, they’ll immediately realize that this basically is a joke.”

## NATIONAL POST

HOME • FINANCIAL POST • NEWS • COMMENT • PERSONAL FINANCE • INVESTING • TECH • SPORTS • ARTS • LIFE • HEALTH • HOUSING  
NEWS CANADA • WORLD • ISRAEL & THE MIDDLE EAST • RELIGION • TORONTO • GRAPHICS • PHOTOS • EDITORS BLOG • NEWS

## NEWS

**TRENDING** Real estate | Lotto Max | Russia | CPC leadership race | Trump | North Korea

### Correlation and causation: The good statistical lesson behind study associating drinking milk and Nobel prizes



JOSEPH BREAN | January 14, 2013 6:56 PM ET  
[More from Joseph Brean](#) | [@JosephBrean](#)



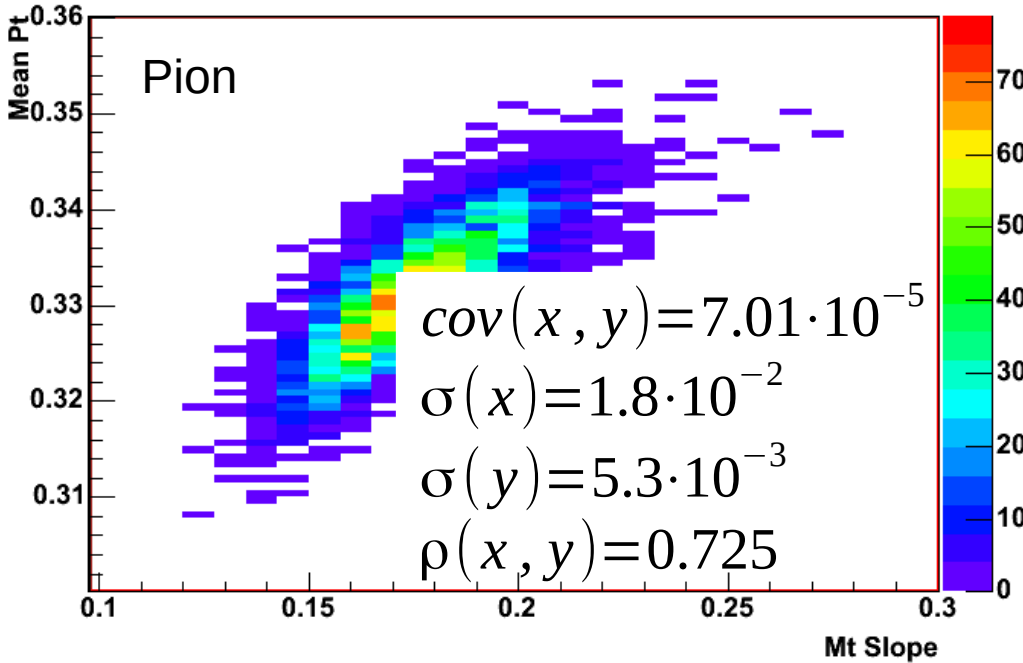
Republish  
Reprint

He compared it to the observation that after the Second World War in Germany, there were distinct drops in both the human birth rate and the stork population, which led to the tantalizing conclusion that these phenomena were related — that there were fewer babies because there were fewer storks to bring them.

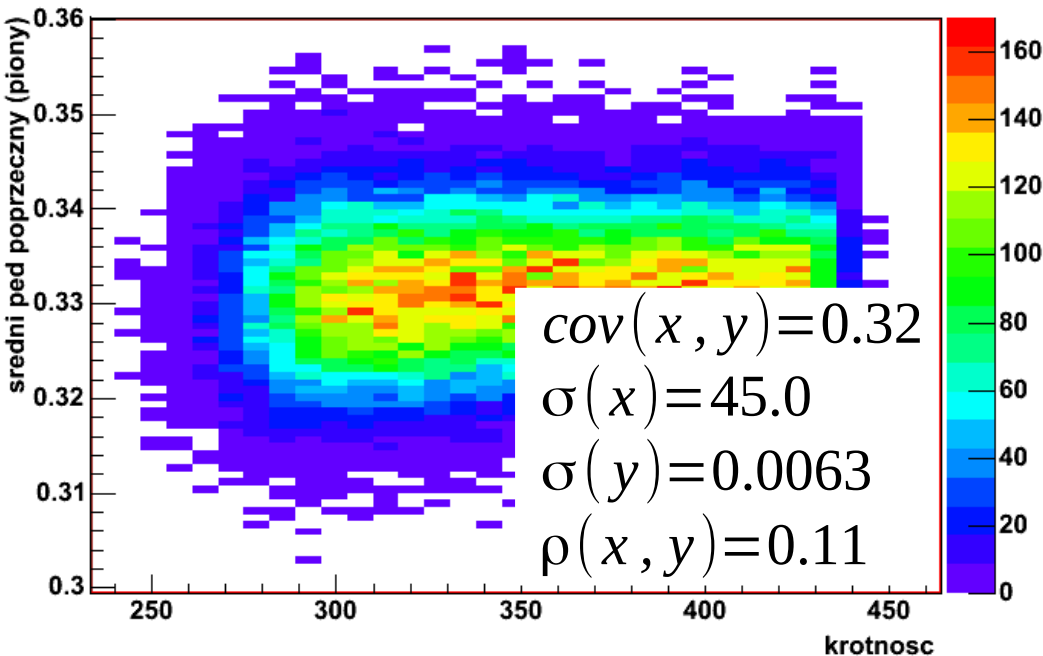
“There was a highly significant correlation,” Dr. Messerli said. “But when you think about this, what happened after the Second World War in Germany was a building boom. Logically, storks like open air, they like farms, but they don’t like housing projects. And these small apartments are not conducive to have six and seven children ... So actually there is a rational reason why storks dwindled and the number of children dwindled at the same time. So here you can find a cause and effect relationship. But I have yet to find one for the chocolate consumption and the Nobel prize winners.”

# Uwagi do interpretacji wsp. korelacji

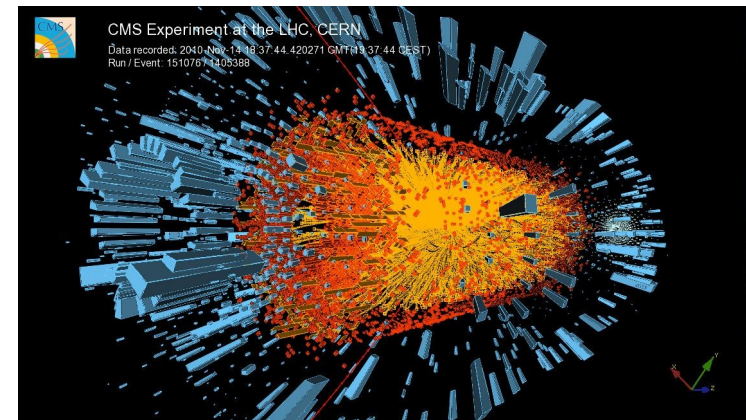
Mt Slope vs. Mean Pt Particle 1



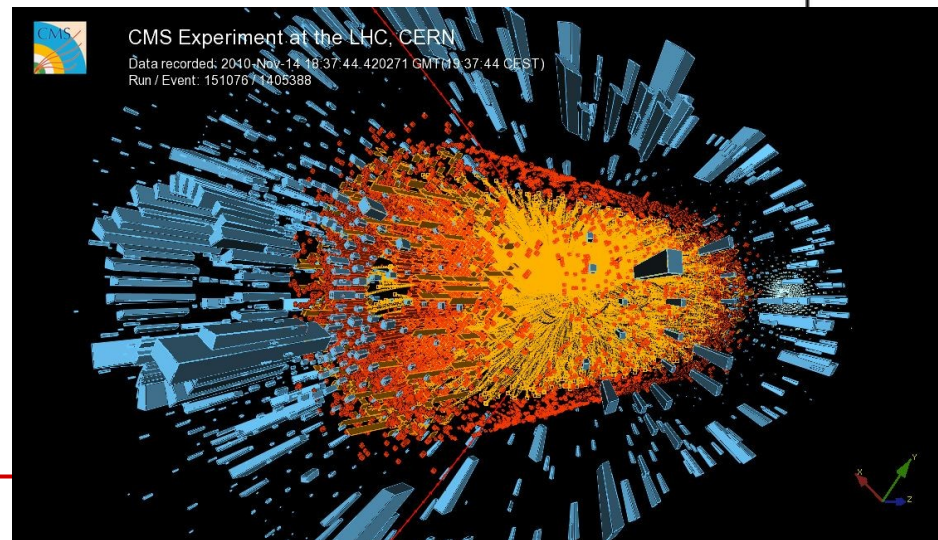
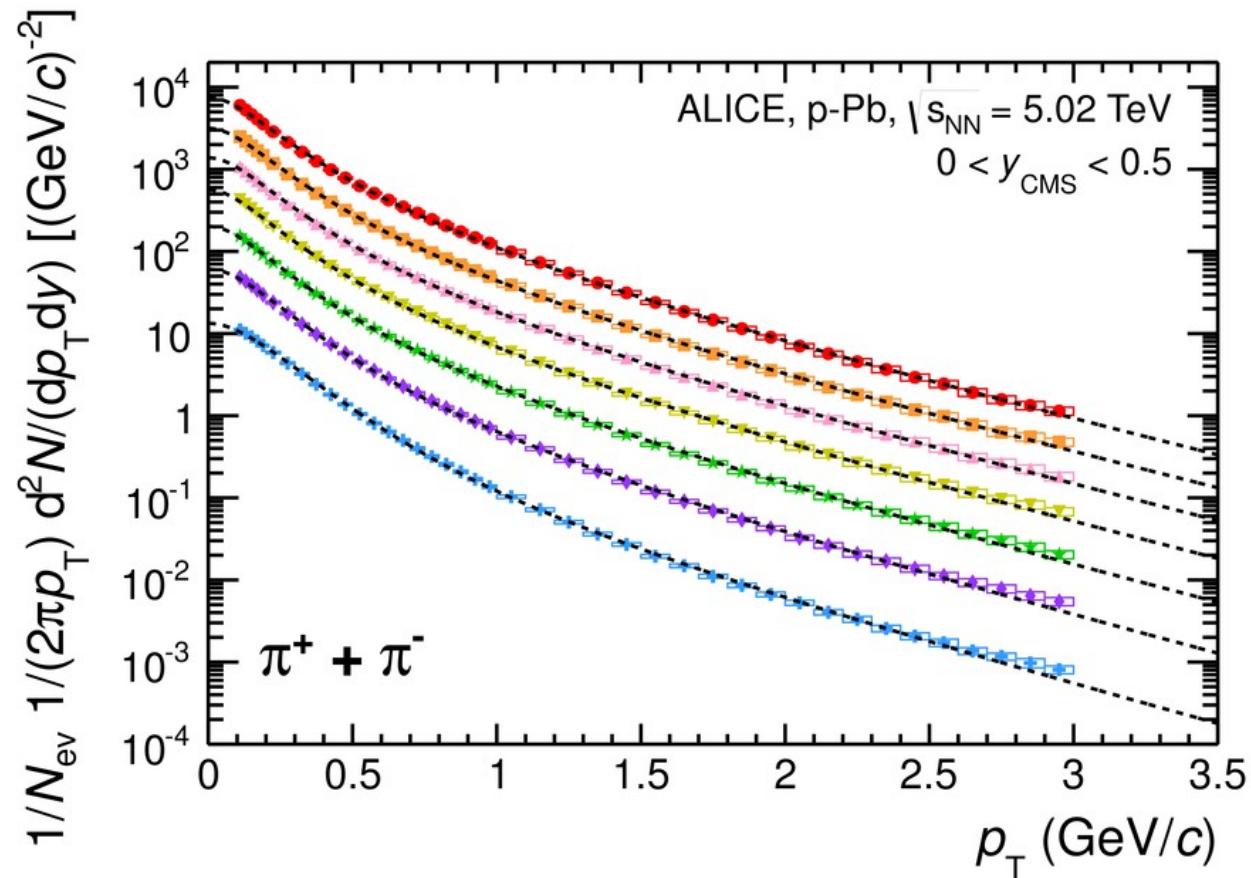
Mean Pt Particle 1 vs Event Multiplicity



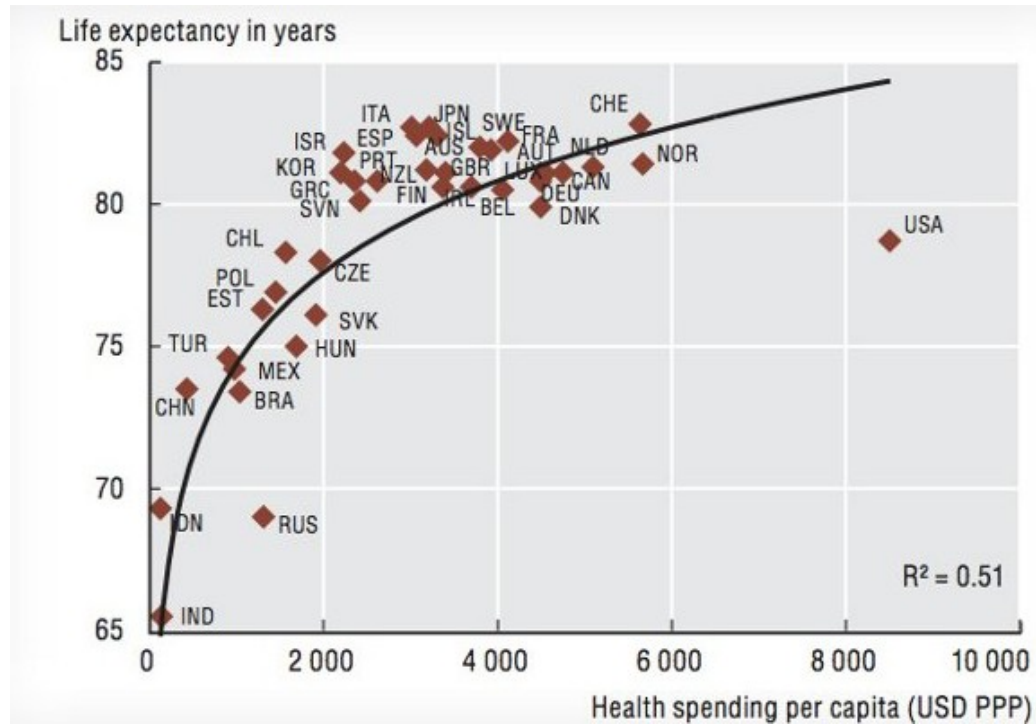
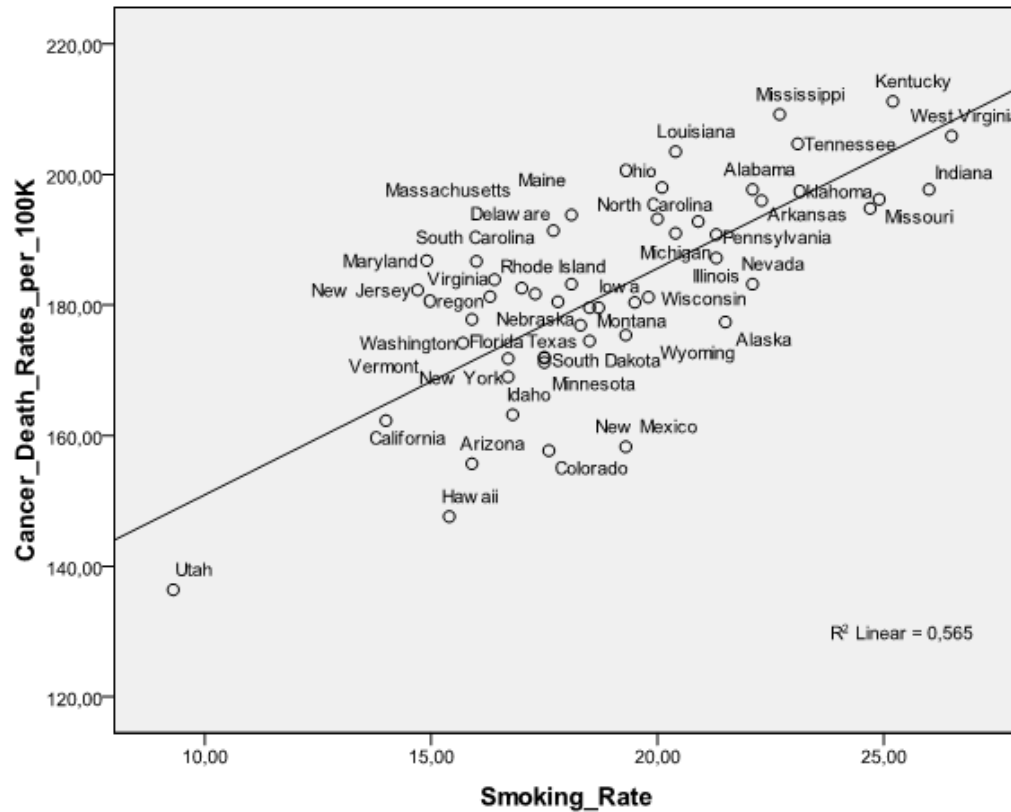
- Przykłady rzeczywistych korelacji eksperymentalnych:
  - góra – zależność między średnim pędem cząstki (pionu) a nachyleniem rozkładu pędowego rozkładu pędu
  - dół – praktycznie brak korelacji pomiędzy średnim pędem cząstki (pionu) a liczbą cząstek produkowanych w zderzeniu



# Uwagi do interpretacji wsp. korelacji



# Uwagi do interpretacji wsp. korelacji



WHO Data