Zjazd 7

Maciej Rosoł

O mnie

- Studia inżynierskie na kierunku Inżynieria Biomedyczna PW 2015-2018
- Studia magisterskie na kierunku Elektronika (elektronika i informatyka medyczna) PW 2018-2020
- Studia w Szkole Doktorskiej na kierunku Inżynieria Biomedyczna PW 2020-obecnie
- Analytic Specialist w OASIS Diagnostics S.A.
- 2 DAN w Aikido

Agenda

- Chi-kwadrat
  - Test zgodności
  - Test niezależności
- Analiza wariancji – ANOVA
- Analiza post-hoc
- Dane wielowymiarowe

Szereg rozdzielczy

Szereg rozdzielczy tworzy się przez uszeregowanie danych według wzrastającej lub malejącej wartości cechy i podzielenie powstałego szeregu na rozłączne podzbiory zwane grupami.

Szereg rozdzielczy

Szereg rozdzielczy tworzy się przez uszeregowanie danych według wzrastającej lub malejącej wartości cechy i podzielenie powstałego szeregu na rozłączne podzbiory zwane grupami.

$$a_0 < a_1 < \cdots < a_k$$

Niech: $n_1, \ldots, n_k$ będą licznościami tych klas oraz niech:

$$y_i = \frac{a_{i-1} + a_i}{2}$$

dla i = 1, . . . , k będą środkami przedziałów klasowych. Wtedy wartość średnia takiego szeregu wyraża się wzorem:

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{k} n_i y_i$$

Szereg rozdzielczy

Szereg rozdzielczy tworzy się przez uszeregowanie danych według wzrastającej lub malejącej wartości cechy i podzielenie powstałego szeregu na rozłączne podzbiory zwane grupami.

Załóżmy, że naszą cechą jest wiek. Wtedy możemy utworzyć następujące przykładowe przedziały:

$$0 < 20 < 30 < 40 < 50 < 60$$

Otrzymujemy wtedy k=5 grup.

Dane kategoryczne

- W próbce liczba danych należących do określonej grupy nazywana jest częstotliwością/częstością wystąpień, więc analiza danych kategorycznych jest analizą częstotliwości/częstości.
- Kiedy porównuje się dwie lub więcej grup, to dane są często prezentowane w formie **Frequency Tables**. Na przykład w poniższej tabeli podana jest liczba osób praworęcznych i leworęcznych w zależności od płci.

	Right handed	Left handed	Total
Males	43	8	51
Females	44	5	49
Total	87	13	

Dane kategoryczne

- Podczas pracy z danymi kategoryzującymi dokładne wartości obserwacji nie są zbyt użyteczne w testach statystycznych, ponieważ kategorie takie jak „mężczyźni”/„kobiety”, „zdrowy”/„chory”, „leworęczny”/„praworęczny” i inne nie mają znaczenia matematycznego.
- Testy dotyczące zmiennych kategorycznych opierają się na liczbie zmiennych, zamiast rzeczywistej wartości samych zmiennych.

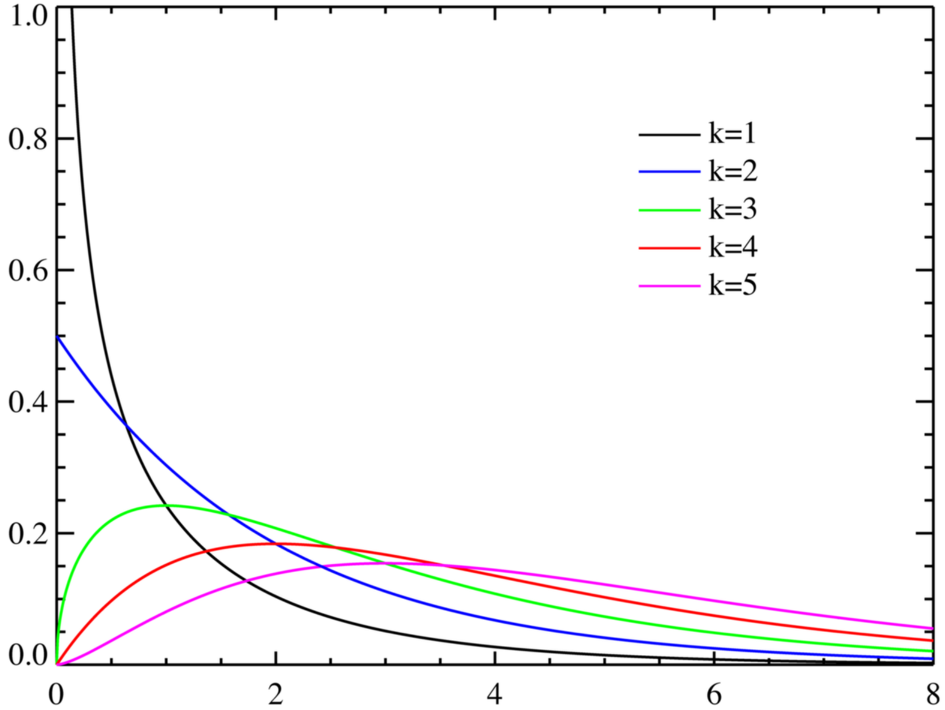
Dane kategoryczne

- Teraz będziemy zakładać, że dane są podane w zestawie kategorii i mamy ich częstości wystąpień (całkowita liczba próbek w każdej kategorii).
- Wiele testów dla takich danych opiera się na analizie odchylenia od wartości oczekiwanej.
- Ponieważ rozkład chi kwadrat charakteryzuje zmienność danych (innymi słowy, ich odchylenie od wartości średniej), wiele z tych testów odnosi się do tego rozkładu i nazywane są testami chi kwadrat.

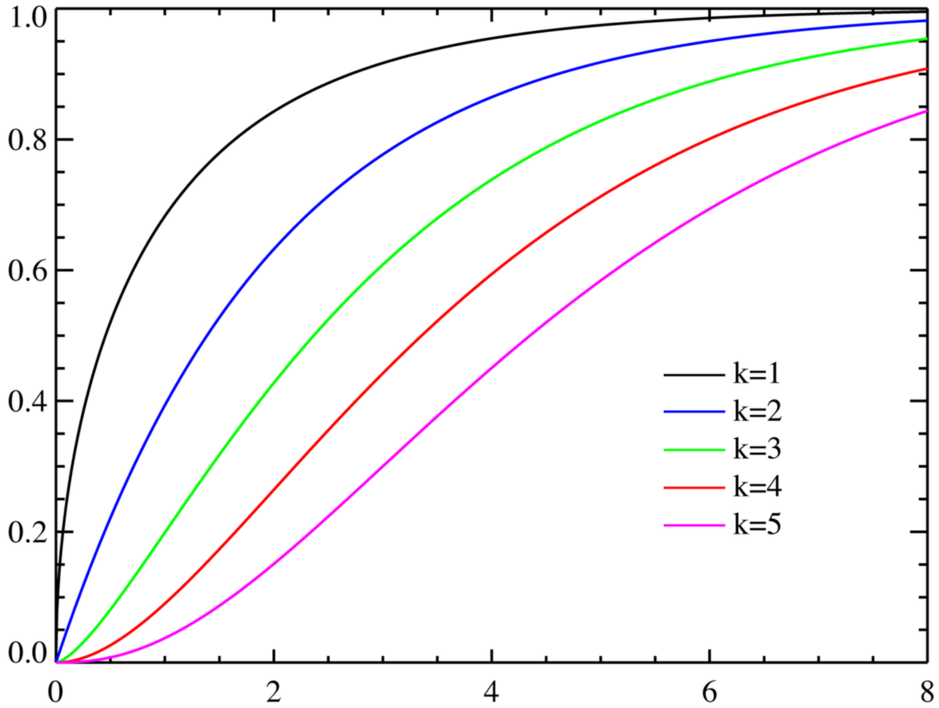
Rozkład chi-kwadrat

Jeżeli zmienna losowa Y ma rozkład chi kwadrat o k stopniach swobody:

$$Y \sim \chi_k^2$$



Gęstość prawdopodobieństwa



Dystrybuanta

https://pl.wikipedia.org/wiki/Rozk%C5%82ad_chi_kwadrat

Test zgodności chi kwadrat

- W przypadku testu t-Studenta weryfikowaliśmy hipotezę czy średnia próbki różni się od oczekiwanej średniej populacji.
- **Test zgodności chi kwadrat** (chi square goodness of fit) jest analogicznym testem dla zmiennych kategorycznych: testuje, czy rozkład przykładowych danych kategorycznych odpowiada oczekiwanemu rozkładowi.

Test zgodności chi kwadrat

Załóżmy, że zaobserwowaliśmy częstości wystąpień o_i podczas gdy oczekiwaliśmy częstotliwości (teoretycznych) e_i.

H_0 - dane **są zgodne** z rozkładem teoretycznym

H_1 - dane **nie są zgodne** z rozkładem teoretycznym

Statystyka testowa przyjmuje postać:

$$V = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{k-1}$$

i ma rozkład chi kwadrat z k-1 stopniami swobody (k to liczba grup).

Notebook – D08_Z01

Notebook – Chi2_zgodnosc1

Notebook – D08_Z02

Test niezależności chi kwadrat

Niezależność jest kluczową koncepcją prawdopodobieństwa opisującą sytuację, w której wiedza o wartościach jednej zmiennej nie mówi nic o wartości innej. Na przykład:

- Miesiąc urodzenia prawdopodobnie nie mówi nic na temat tego jakiej przeglądarki internetowej ktoś używa.
- Więc spodziewamy się, że miesiąc narodzin i preferencje odnośnie przeglądarki będą niezależne.
- Z drugiej strony, miesiąc urodzenia może być związany z wynikami sportowymi w danym roczniku u dzieci (nie być niezależne).

Test niezależności chi kwadrat

Załóżmy, że zaobserwowaliśmy częstości wystąpień n_{ij} podczas gdy oczekiwaliśmy częstotliwości (teoretycznych) e_{ij}.

H_0 - zmienne **są niezależne**

H_1 - zmienne **są zależne**

Statystyka testowa przyjmuje postać:

$$V = \sum_{i=1}^{k}\sum_{j=1}^{s}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(k-1)(s-1)}$$

i ma rozkład chi kwadrat z (k-1)(s-1) stopniami swobody (k to liczba grup pierwszej cechy, s to liczba grup drugiej cechy).

Test niezależności chi kwadrat

Test niezależności jest powszechnie używany do określenia, czy zmienne, takie jak: edukacja, poglądy polityczne i inne preferencje różnią się w zależności od czynników demograficznych, takich jak: płeć, rasa i religia.

Wartości zaobserwowane

	S1	S2	Total
K1	n11	n12	n11+n12
K2	n21	n22	n21+n22
Total	n11+n21	n12+n22	N

Wartości teoretyczne (jeżeli cechy są niezależne)

	S1	S2	Total
K1	(n11+n12)(n11+n21)/N	(n11+n12)(n12+n22)/N	n11+n12
K2	(n21+n22)(n11+n21)/N	(N12+n22)(n21+n22)/N	n21+n22
Total	n11+n21	n12+n22	N

Notebook – D08_Z03

Notebook – D08_Z04

Notebook – Chi2_niezaleznosc

ANOVA

Pomysł analizy wariancji (ANOVA) polega na podzieleniu wariancji na:

- wariancję między grupami (variance between groups),
- wariancję wewnątrz grup (variance within groups),

Na podstawie tych dwóch wartości weryfikujemy hipotezę:

$$H_0 \text{ - } \mu_0 = \mu_1 = \cdots = \mu_n = \mu$$

Względem hipotezy alternatywnej:

$$H_1 \text{ - } \mu_i \neq \mu_j \text{ gdzie } \mathrm{i} \neq j$$

ANOVA

Na przykład, jeżeli porównujemy:

- grupę bez leczenia,
- grupę z leczeniem A,
- grupę z leczeniem B,

wykonujemy jednoczynnikową analizę wariancji (ANOVA), czasami zwaną jednokierunkową. Jeżeli wykonamy taki sam test na mężczyznach i kobietach, to mamy dwuczynnikową analizę wariancji (ANOVA). Względem płci oraz typu leczenia.

ANOVA

Jednoczynnikowa ANOVA zakłada, że wszystkie próbki pochodzą z rozkładu normalnego o tej samej wariancji. Założenie równej wariancji można sprawdzić przy użyciu **testu Levene** (jeżeli p-wartość otrzymana z testu jest większa od założonej wartości krytycznej wtedy założenie jest spełnione)

ANOVA

Aby wykonać analizę wariancji wyliczamy najpierw sum of squares (SS):

$$SS_{Error} = \sum_{j}^{k} \sum_{i}^{n_j} (Y_{ij} - \bar{Y}_j)^2$$

$$SS_{Treatment} = n \sum_{j}^{k} (\bar{Y}_j - \bar{Y})^2$$

Gdzie:

$\bar{Y}$ - średnia wartość cechy

$\bar{Y}_j$ - średnia wartość cechy dla danej klasy

k – liczba grup

n_j - liczba próbek w danej grupie

ANOVA

Następnie wyliczamy stopnie swobody:

$$df_{groups} = n_{groups} - 1$$

$$df_{residuals} = n_{data} - n_{groups}$$

Średnie kwadraty (mean squares-MS), to SS podzielone przez odpowiednie stopnie swobody.

$$MS_{Treatment} = \frac{SS_{Treatment}}{df_{groups}}$$

$$MS_{Error} = \frac{SS_{Error}}{df_{residuals}}$$

ANOVA

Wartość statystki testowej ma postać:

$$F = \frac{MS_{Treatment}}{MS_{Error}} = \frac{SS_{Treatment}/(n_{groups} - 1)}{SS_{Error}/(n_{Total} - n_{groups})}$$

F ma rozkład F Snedecora z df_{groups} i $df_{residuals}$ stopniami swobody.

(w przypadku dwóch grup test t-Studenta prowadzi do dokładnie takiego samego wyniku)

Notebook – D07_Z11

Notebook – ANOVA

Post hoc

- Zerowa hipoteza w jednoczynnikowej analizie wariancji mówi, że wszystkie średnie są takie same. Więc jeżeli odrzucimy hipotezę zerową, to nie mamy żadnej informacji.

Post hoc

- Zerowa hipoteza w jednoczynnikowej analizie wariancji mówi, że wszystkie średnie są takie same. Więc jeżeli odrzucimy hipotezę zerową, to nie mamy żadnej informacji.
- Często nie interesuje nas czy wszystkie próbki są takie same, ale chcielibyśmy też wiedzieć, dla których par próbek takie podobieństwo nie zachodzi.

Post hoc

- Zerowa hipoteza w jednoczynnikowej analizie wariancji mówi, że wszystkie średnie są takie same. Więc jeżeli odrzucimy hipotezę zerową, to nie mamy żadnej informacji.
- Często nie interesuje nas czy wszystkie próbki są takie same, ale chcielibyśmy też wiedzieć, dla których par próbek takie podobieństwo nie zachodzi.
- Analiza takich zależności nazywana jest porównaniami **post hoc** lub testami post hoc.

Post hoc

Do analizy post hoc można wykorzystać następujące testy (uszeregowane od najbardziej do najmniej konserwatywnego):

- test Scheffégo
- test Tukeya
- test Newmana i Keulsa
- test Duncana
- test Najmniejszych Istotnych Różnic (NIR)

Post hoc

Trzej łucznicy - Patryk, Jacek i Aleksander biorą udział w konkursie strzeleckim. Pierścienie na tarczy mają wartości punktacji od 1 do 10 (10 to najwyższy wynik). Każdy uczestnik strzela 6 razy, zdobywając punkty:

Patryk - 5, 4, 4, 3, 9, 4

Jacek - 4, 8, 7, 5, 1, 5

Aleksander - 9, 9, 8, 10, 4, 10

Na podstawie powyższych wyników chcielibyśmy wiedzieć, kto jest najlepszym łucznikiem.

Post hoc

Wyniki testu Tukey pokazują średnią różnicę, przedziały ufności i to, czy należy odrzucić hipotezę zerową dla każdej pary grup na danym poziomie istotności.

```
from statsmodels.stats.multicomp import (pairwise_tukeyhsd, MultiComparison)
multiComp = MultiComparison(data['Score'], data['Archer'])
hsd = multiComp.tukeyhsd()
print((multiComp.tukeyhsd().summary()))
```

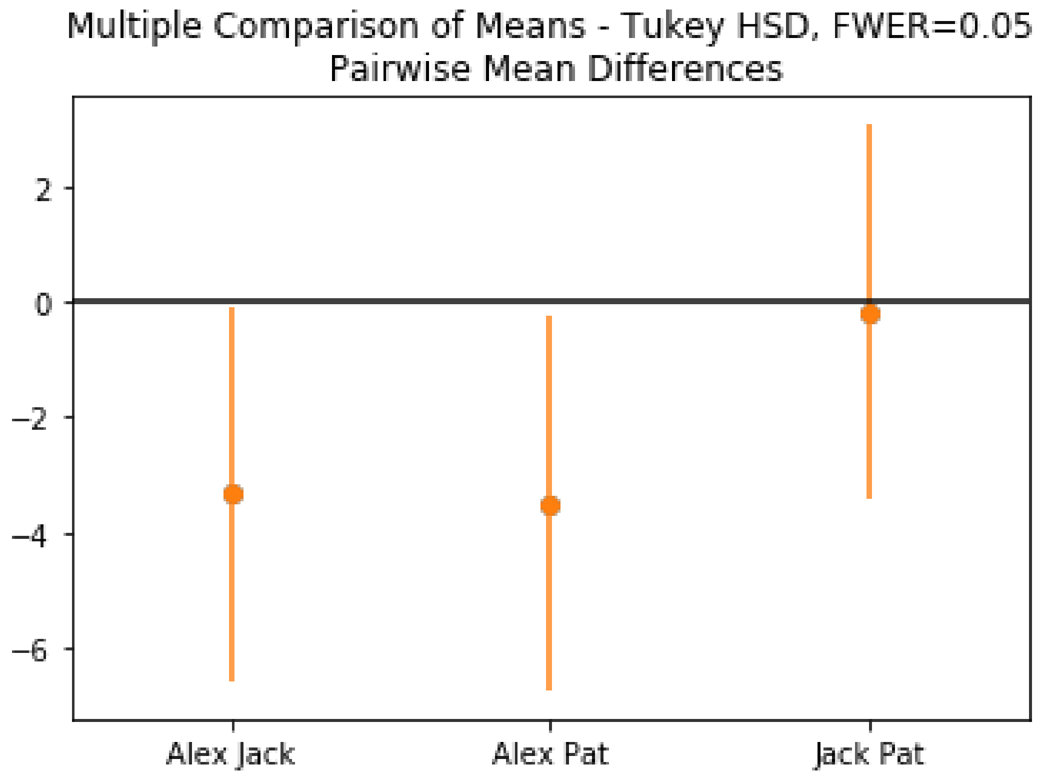
```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
  Alex   Jack  -3.3333 0.0435 -6.5755 -0.0911   True
  Alex    Pat     -3.5 0.0337 -6.7422 -0.2578   True
  Jack    Pat  -0.1667    0.9 -3.4089  3.0755  False
----------------------------------------------------
```

Post hoc

W tym przypadku test sugeruje odrzucenie hipotezy o równości średnich dla par:

- Aleksander Jacek
- Aleksander Patryk

To sugeruje, że wyniki Aleksandra stanowczo różnią się od innych grup. Wizualizacja 95% przedziałów ufności wzmacnia wyniki w sposób wizualny.



Notebook – D07_Z12

Notebook – D07_Z14

Notebook – Post_hoc

Dane wielowymiarowe

Zmienna losowa **dwuwymiarowa** to wektor (X, Y), którego składowe X, Y są zmiennymi losowymi.

Łącznym **rozkładem prawdopodobieństwa** (lub rozkładem łącznym) pary zmiennych losowych (X, Y) określonych na tej samej przestrzeni zdarzeń elementarnych nazywamy przyporządkowanie

$$A \ \rightarrow P((X,Y)\epsilon A)$$

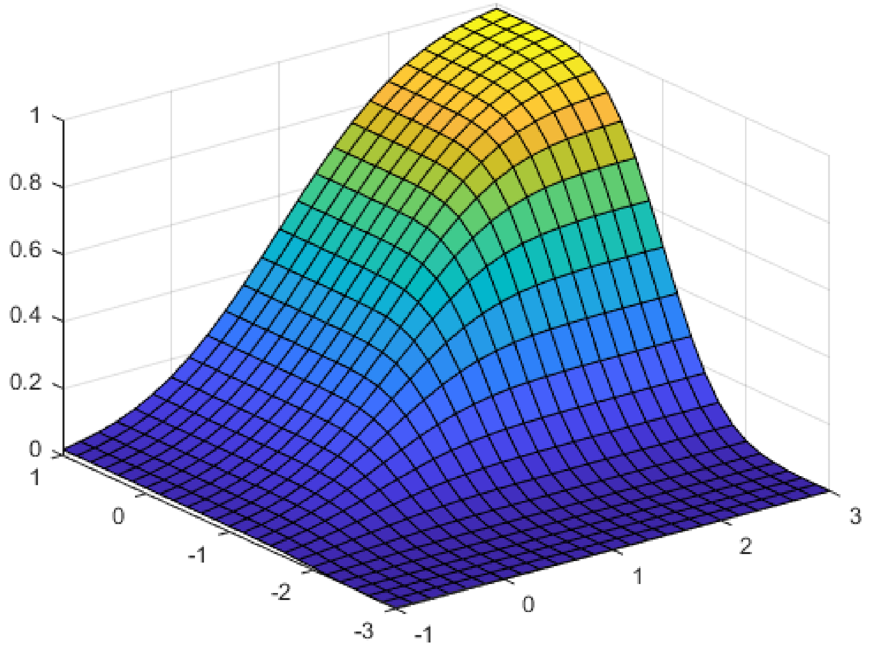
gdzie A - dowolny podzbiór zbioru par wartości zmiennych X, Y

Dane wielowymiarowe

Dystrybuantą zmiennej losowej (X,Y) nazywamy funkcję:

$$F(x, y) = P(X \leq x, Y \leq y)$$

Gdzie $-\infty < x < \infty, -\infty < y < \infty$



https://www.mathworks.com/help/stats/multivariate-normal-distribution.html

Dane wielowymiarowe

Funkcją **prawdopodobieństwa łącznego** dwuwymiarowej zmiennej losowej **dyskretnej** nazywamy funkcję:

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

Funkcja prawdopodobieństwa f oraz jej związek z dystrybuantą dla danych dyskretnych:

- $f(x, y) \geq 0$, dla dowolnej pary wartości (x,y)
- $\sum_x \sum_y f(x, y) = 1$
- $P\big((X, Y) \in A\big) = \sum_{(x,y)\in A} f(x, y)$
- $F(x, y) = \sum_{s\leq x} \sum_{t\leq y} f(s, t)$

Dane wielowymiarowe

W każdym z dwóch etapów teleturnieju można otrzymać 0, 1, lub 2 punkty. Niech zmienne losowe X, Y oznaczają odpowiednio liczby punktów uzyskane w etapie I i II przez losowo wybranego uczestnika. Funkcję prawdopodobieństwa łącznego określa tabela:

y x	0	1	2
0	0.5	0.05	0.01
1	0.2	0.1	0.06
2	0.02	0.03	?

Prawdopodobieństwo $P(X = x, Y = y)$ podane jest na przecięciu wiersza $X = x$ i $Y = y$, na przykład $P(X = 1, Y = 0) = 0.2$

Dane wielowymiarowe

y / x	0	1	2
0	0.5	0.05	0.01
1	0.2	0.1	0.06
2	0.02	0.03	?

Znajdźmy:

$a)\ f(2,2) = P(X = 2, Y = 2)$

$b)\ P(Y = 2)$

$c)\ F(1,1)$

Dane wielowymiarowe

a) $f(2,2) = P(X = 2, Y = 2) = 1 - (0.5 + 0.05 + 0.01 + 0.2 + 0.1 + 0.06 + 0.02 + 0.03) = 1 - 0.97 = 0.03$

y x	0	1	2
0	0.5	0.05	0.01
1	0.2	0.1	0.06
2	0.02	0.03	0.03

Dane wielowymiarowe

$a)$ $f(2,2) = P(X = 2, Y = 2) = 1 - (0.5 + 0.05 + 0.01 + 0.2 + 0.1 + 0.06 + 0.02 + 0.03) = 1 - 0.97 = 0.03$

$b)$ $P(Y = 2) = 0.01 + 0.06 + 0.03 = 0.1$

y x	0	1	2
0	0.5	0.05	0.01
1	0.2	0.1	0.06
2	0.02	0.03	0.03

Dane wielowymiarowe

a) $f(2{,}2) = P(X = 2, Y = 2) = 1 - (0.5 + 0.05 + 0.01 + 0.2 + 0.1 + 0.06 + 0.02 + 0.03) = 1 - 0.97 = 0.03$

b) $P(Y = 2) = 0.01 + 0.06 + 0.03 = 0.1$

c) $F(1{,}1) = P(X \le 1, Y \le 1) = 0.5 + 0.2 + 0.05 + 0.1 = 0.85$

y x	0	1	2
0	0.5	0.05	0.01
1	0.2	0.1	0.06
2	0.02	0.03	0.03

Dane wielowymiarowe

Dwuwymiarowa zmienna losowa (X, Y) nazywana jest ciągłą zmienną losową (krócej - zmienną ciągłą), jeżeli jej łączny rozkład prawdopodobieństwa określony jest przez **funkcję gęstości łącznej** (łączną gęstość prawdopodobieństwa) taką, że:

- $f(x, y) \geq 0$, dla dowolnej pary wartości (x, y)
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$
- $P\big((X, Y) \in A\big) = \iint_A \ f(x, y) dx dy$

Dane wielowymiarowe

Dwuwymiarowa zmienna losowa ma gęstość łączną postaci:

$$f(x, y) = \begin{cases} Cx^2 & \text{gdy } 0 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

dla pewnej stałej C. Znajdź wartość tej stałej C.

Dane wielowymiarowe

Ponieważ f jest gęstością to:

$$1 = \int_0^1 \int_0^1 Cx^2 dxdy = \int_0^1 \left(\int_0^1 Cx^2 dx \right) dy =$$
$$= C \int_0^1 [\tfrac{1}{3} x^3]_0^1 dy = C \int_0^1 \tfrac{1}{3} dy = \tfrac{c}{3}.$$

Czyli:

$$1 = \frac{C}{3}$$

$$C = 3.$$

Dane wielowymiarowe

Rozkładem brzegowym pary (X, Y) nazywamy rozkład prawdopodobieństwa zmiennej losowej X lub zmiennej losowej Y:

a) dla dyskretnych zmiennych X, Y, brzegowe funkcje prawdopodobieństwa są postaci:

$$f_X(x) = P(X = x) = \sum_y f(x, y)\,, f_Y(y) = P(Y = y) = \sum_x f(x, y)$$

b) dla ciągłych zmiennych X, Y, brzegowe gęstości są postaci:

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

Dane wielowymiarowe

Dwuwymiarowa zmienna losowa (X, Y) ma gęstość:

$$f(x,y) = \begin{cases} \frac{3}{8}(x-y)^2 & \text{gdy } -1 \leq x \leq 1,\ -1 \leq y \leq 1 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Znajdziemy gęstość zmiennej losowej X.

Dane wielowymiarowe

Dwuwymiarowa zmienna losowa (X, Y) ma gęstość:

$$f(x,y) = \begin{cases} \frac{3}{8}(x-y)^2 & \text{gdy } -1 \leq x \leq 1,\ -1 \leq y \leq 1 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Znajdziemy gęstość zmiennej losowej X.

Wyznaczmy gęstość zmiennej losowej X:

$$f_X(x) = \int_{\infty}^{\infty} f(x,y)dy = \int_{-1}^{1} \frac{3}{8}(x-y)^2 dy =$$

$$= \frac{3}{8} \int_{-1}^{1} (x^2 - 2xy + y^2)dy = \frac{3}{8}[x^2 y - xy^2 + \frac{1}{3}y^3]_{-1}^{1} = \frac{1}{4}(3x^2 + 1).$$

Dane wielowymiarowe

Uzupełnij tabelę:

y / x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	?
1	0.2	0.1	0.06	?
2	0.02	0.03	0.03	?
$f_Y(y)$	?	?	?	

Dane wielowymiarowe

Uzupełnij tabelę:

y / x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	0.56
1	0.2	0.1	0.06	0.36
2	0.02	0.03	0.03	0.08
$f_Y(y)$	0.72	0.18	0.1	

$f_Y(0)$ = 0.5 + 0.2 + 0.02 = 0.72
$f_Y(1)$ = 0.05 + 0.1 + 0.03 = 0.18
$f_Y(2)$ = 0.01 + 0.06 + 0.03 = 0.1

$f_X(0)$ = 0.5 + 0.05 + 0.01 = 0.56
$f_X(1)$ = 0.2 + 0.1 + 0.06 = 0.36
$f_X(2)$ = 0.02 + 0.03 + 0.03 = 0.08.

Dane wielowymiarowe

Rozkład zmiennej losowej (X|Y = y) nazywamy **rozkładem warunkowym** zmiennej losowej X przy ustalonej wartości zmiennej losowej Y.

Mówimy, że funkcja p(X|Y = y): R → [0, 1] jest **warunkową funkcją prawdopodobieństwa** zmiennej losowej X pod warunkiem, że Y = y jeśli dla każdego x ∈ R :

$$p(X|Y = y)(x) = P(X = x|Y = y)$$

Dane wielowymiarowe

Warunkowa funkcja prawdopodobieństwa (dla danych **dyskretnych**) zmiennej losowej X pod warunkiem, że Y = y dana jest wzorem:

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

O ile $p_Y(y) > 0$

Warunkowa gęstość prawdopodobieństwa (dla danych **ciągłych**) zmiennej losowej X pod warunkiem, że Y = y dana jest wzorem:

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

O ile $f_Y(y) > 0$

Dane wielowymiarowe

Załóżmy, że chcemy policzyć prawdopodobieństwo, że zmienna ciągła X jest z przedziału <a,b> pod warunkiem, że zmienna losowa Y jest równa y. W takim wypadku prawdopodobieństwo to będzie wyrażało się wzorem:

$$P(X \in < a, b > | Y = y) = \int_a^b f(X|Y = y)(x)dx$$

Dane wielowymiarowe

Niech (X, Y) będzie parą zmiennych losowych o rozkładzie określonym przez funkcję f (x, y) będącą funkcją prawdopodobieństwa łącznego lub gęstością. Zmienne losowe X, Y są **niezależne**, jeżeli:

$$f(x, y) = f_X(x) f_Y(y)$$

Dla **wszystkich** wartości x i y.

Zmienne X i Y, które nie są niezależne, nazywamy **zależnymi** zmiennymi losowymi.

Dane wielowymiarowe

Czy liczby punktów uzyskane w I i II etapie teleturnieju przez losowo wybranego uczestnika są niezależnymi zmiennymi losowymi?

y / x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	0.56
1	0.2	0.1	0.06	0.36
2	0.02	0.03	0.03	0.08
$f_Y(y)$	0.72	0.18	0.1	

Dane wielowymiarowe

Czy liczby punktów uzyskane w I i II etapie teleturnieju przez losowo wybranego uczestnika są niezależnymi zmiennymi losowymi?

y / x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	0.56
1	0.2	0.1	0.06	0.36
2	0.02	0.03	0.03	0.08
$f_Y(y)$	0.72	0.18	0.1	

Musimy sprawdzić:

$$f(x,y) = f_X(x) f_Y(y)$$

Wystarczy znaleźć jeden przykład dla którego powyższa nierówność nie zachodzi.

Mamy:

$$0.5 = f(0,0) \neq f_X(0) f_Y(0) = 0.56 * 0.72 = 0.4032$$

Dane wielowymiarowe

Czy X, Y są niezależnymi zmiennymi losowymi, jeśli ich łączna gęstość ma postać:

$$f(x,y) = \begin{cases} \frac{3}{8}(x-y)^2 & \text{gdy } -1 \le x \le 1,\ -1 \le y \le 1 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Musimy sprawdzić:

$$f(x,y) \stackrel{?}{=} f_X(x) f_Y(y).$$

Dane wielowymiarowe

Mamy:

$$f_Y(y) = \tfrac{3}{8} \int\limits_{-1}^{1} (x^2 - 2xy + y^2) dx = \tfrac{3}{8} (\tfrac{x^3}{3} - x^2 y + y^2 x)|_{-1}^{1} =$$
$$= \tfrac{3}{8} [(\tfrac{1}{3} - y + y^2) - (-\tfrac{1}{3} - y - y^2)] = \tfrac{3}{8}(\tfrac{2}{3} + 2y^2) =$$
$$= \tfrac{1}{4} + \tfrac{3}{4} y^2 = \tfrac{1}{4}(3y^2 + 1)$$

Ponieważ funkcja f jest symetryczna ze względu na parametry x i y, to:

$$f_X(x) = \tfrac{3}{8} \int\limits_{-1}^{1} (x^2 - 2xy + y^2) dy = \tfrac{1}{4}(3x^2 + 1).$$

Dane wielowymiarowe

Policzmy:

$$f_X(x)f_Y(y) = \tfrac{1}{4}(3x^2+1)\tfrac{1}{4}(3y^2+1) = \tfrac{1}{16}(3x^2+1)(3y^2+1).$$

Wystarczy znaleźć jeden przykład, dla którego powyższa nierówność nie zachodzi. Sprawdźmy więc $x = 0$ i $y = 0$. Mamy:

$$f(x,y) = 0$$

oraz

$$f_X(x)f_Y(y) = \frac{1}{16}.$$

Zmienne losowe X oraz Y są zależnymi ponieważ:

$$f(x,y) \neq f_X(x)f_Y(y).$$

Notebook – Dane wielowymiarowe

Dane wielowymiarowe

Wartością oczekiwaną (średnią) zmiennej losowej g(X,Y) nazywamy:

$$E\big(g(X,Y)\big) = \sum_{x}\sum_{y} g(x,y)f(x,y)$$

gdy X i Y są dyskretne, natomiast

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

gdy X i Y są ciągłe.

Jeżeli zmienne X i Y są **niezależne** to:

$$E(X,Y) = EX * EY$$

Dane wielowymiarowe

Co jeżeli dane są zależne?

Czy można jakoś tą zależność określić liczbowo?

Dane wielowymiarowe

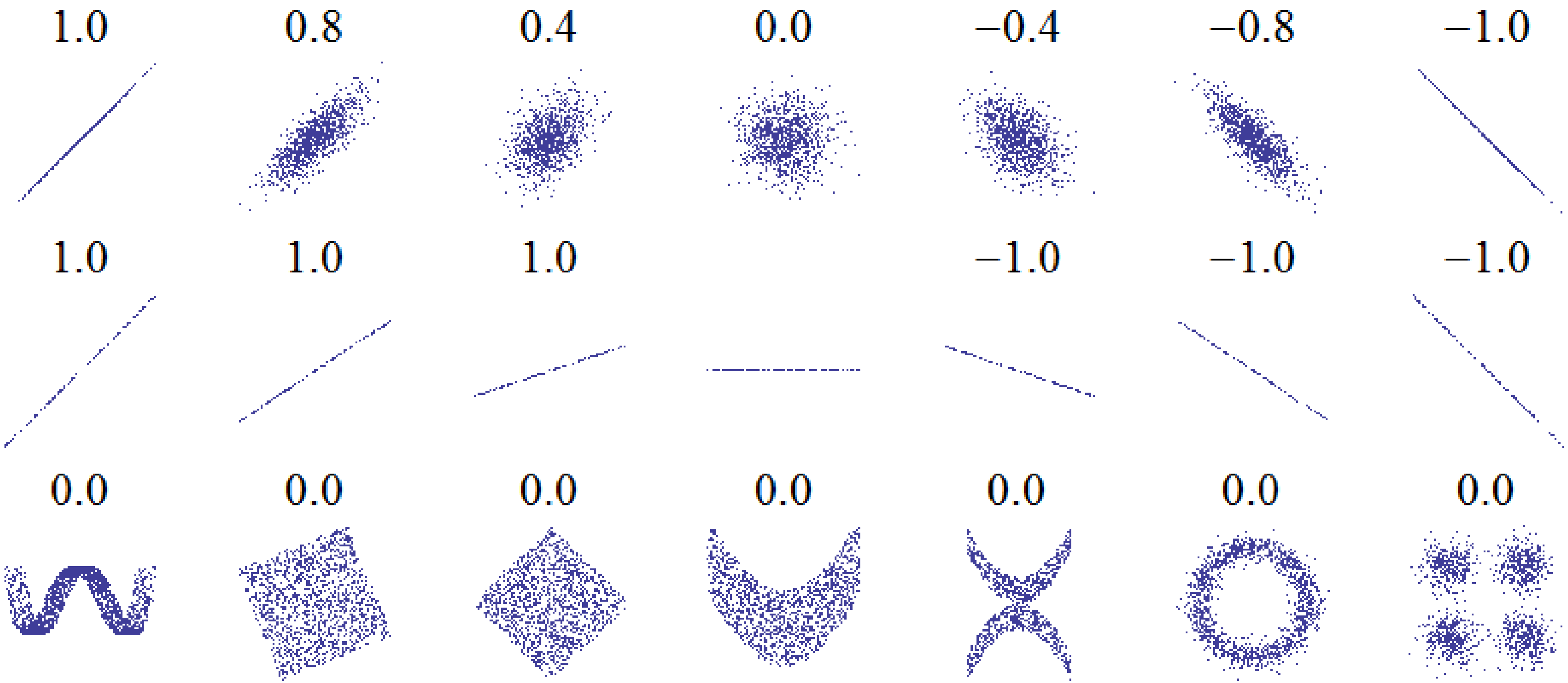
Kowariancja - liczba określająca odchylenie elementów od sytuacji idealnej, w której występuje zależność liniowa pomiędzy zmiennymi losowymi X i Y

$$cov(X,Y) = E[(X - EX)(Y - EY)] = E(XY) - EXEY$$

Współczynnik korelacji liniowej Pearsona – współczynnik określający poziom zależności liniowej między zmiennymi losowymi X i Y, przyjmujący wartość z zakresu <-1,1>

$$r_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Dane wielowymiarowe



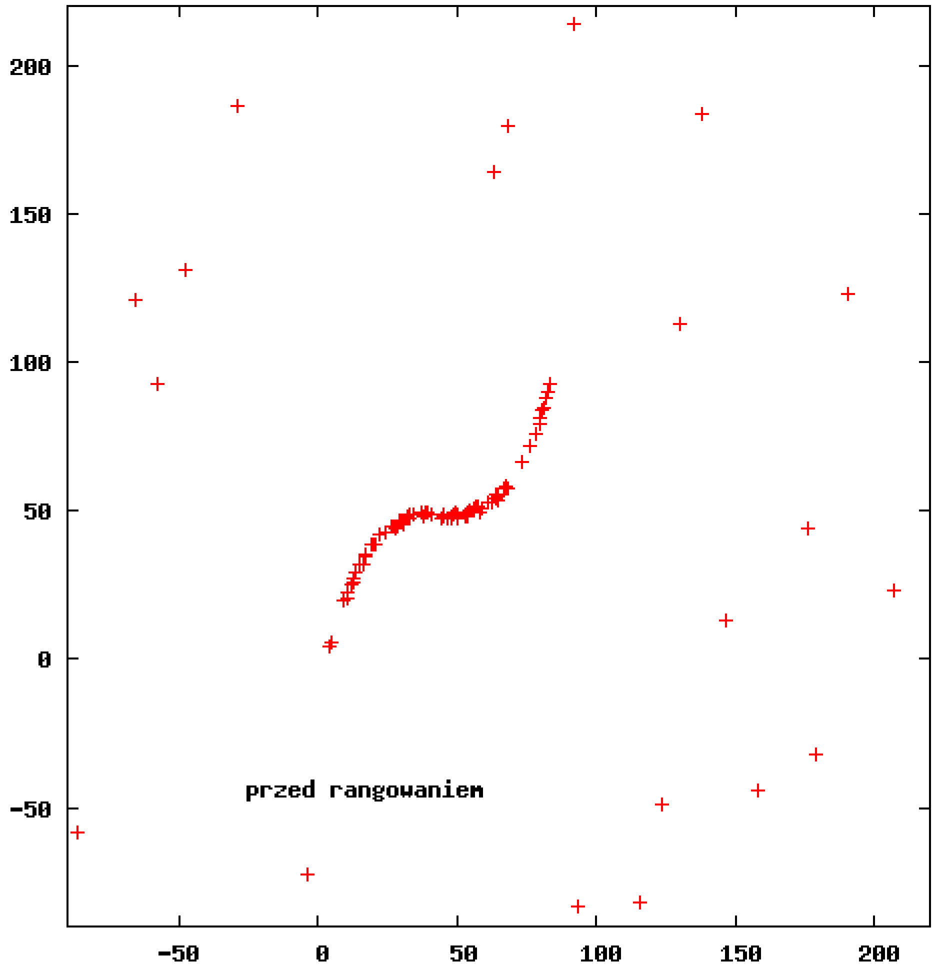
https://pl.wikipedia.org/wiki/Wsp%C3%B3%C5%82czynnik_korelacji_Pearsona

Dane wielowymiarowe

Korelacja rang Spearmana (lub: korelacja rangowa Spearmana, rho Spearmana) jedna z nieparametrycznych miar monotonicznej zależności statystycznej między zmiennymi losowymi.

Korelacja rangowa przyjmuje zawsze wartości z przedziału $[-1, +1]$. Ich interpretacja jest podobna do klasycznego współczynnika korelacji Pearsona, z jednym zastrzeżeniem: w odróżnieniu od współczynnika Pearsona, który mierzy liniową zależność między zmiennymi, a wszelkie inne związki traktuje jak zaburzone zależności liniowej, korelacja rangowa pokazuje dowolną monotoniczną zależność (także nieliniową).

Dane wielowymiarowe



Dane wielowymiarowe

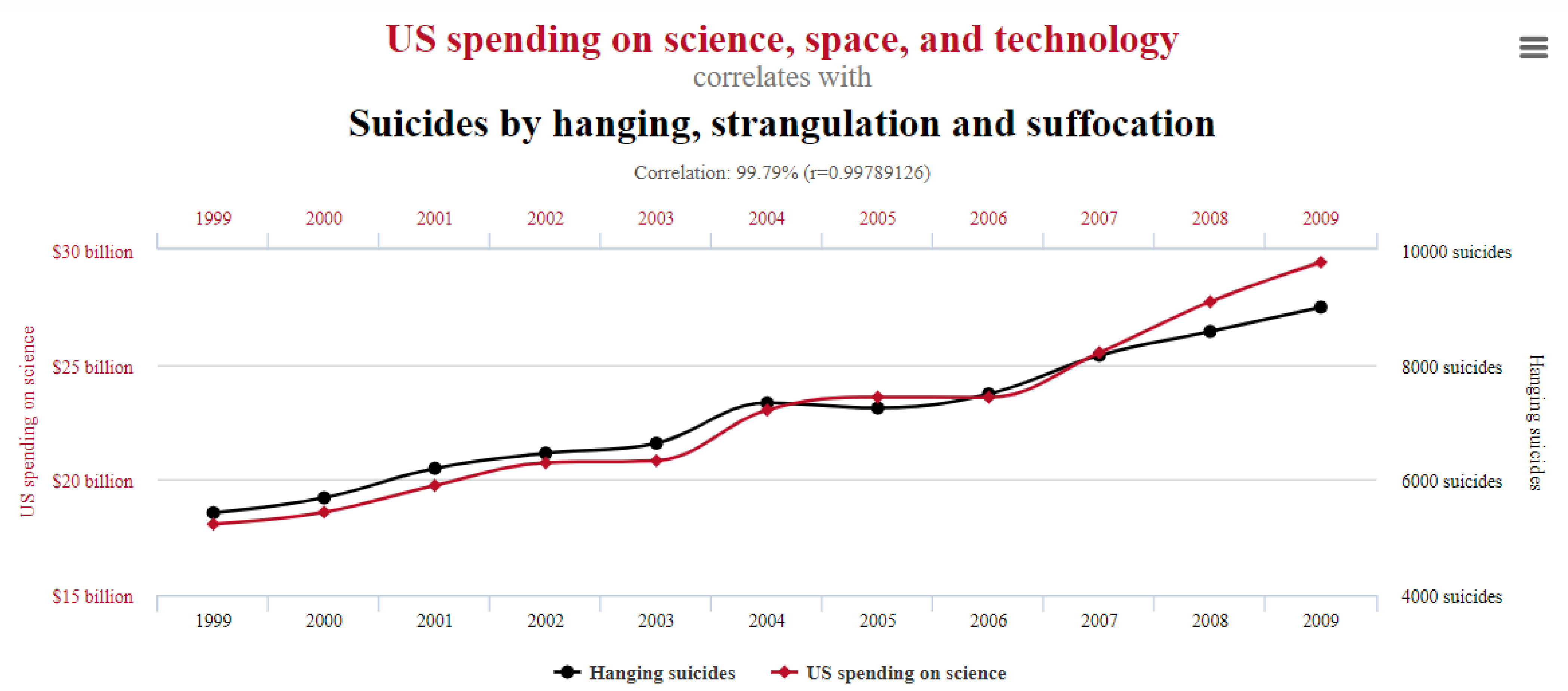
Tau Kendalla - statystyka będąca jedną z miar monotonicznej zależności dwóch zmiennych losowych. Służący w praktyce do opisu korelacji między zmiennymi porządkowymi.

Tau Kendalla przyjmuje wartości od -1 do 1 włącznie. +1 oznacza, że każda ze zmiennych rośnie przy wzroście drugiej. -1 oznacza, że każda maleje przy wzroście drugiej. Tym samym tau Kendalla, podobnie jak korelacja rangowa jest miarą monotonicznej zależności zmiennych losowych.

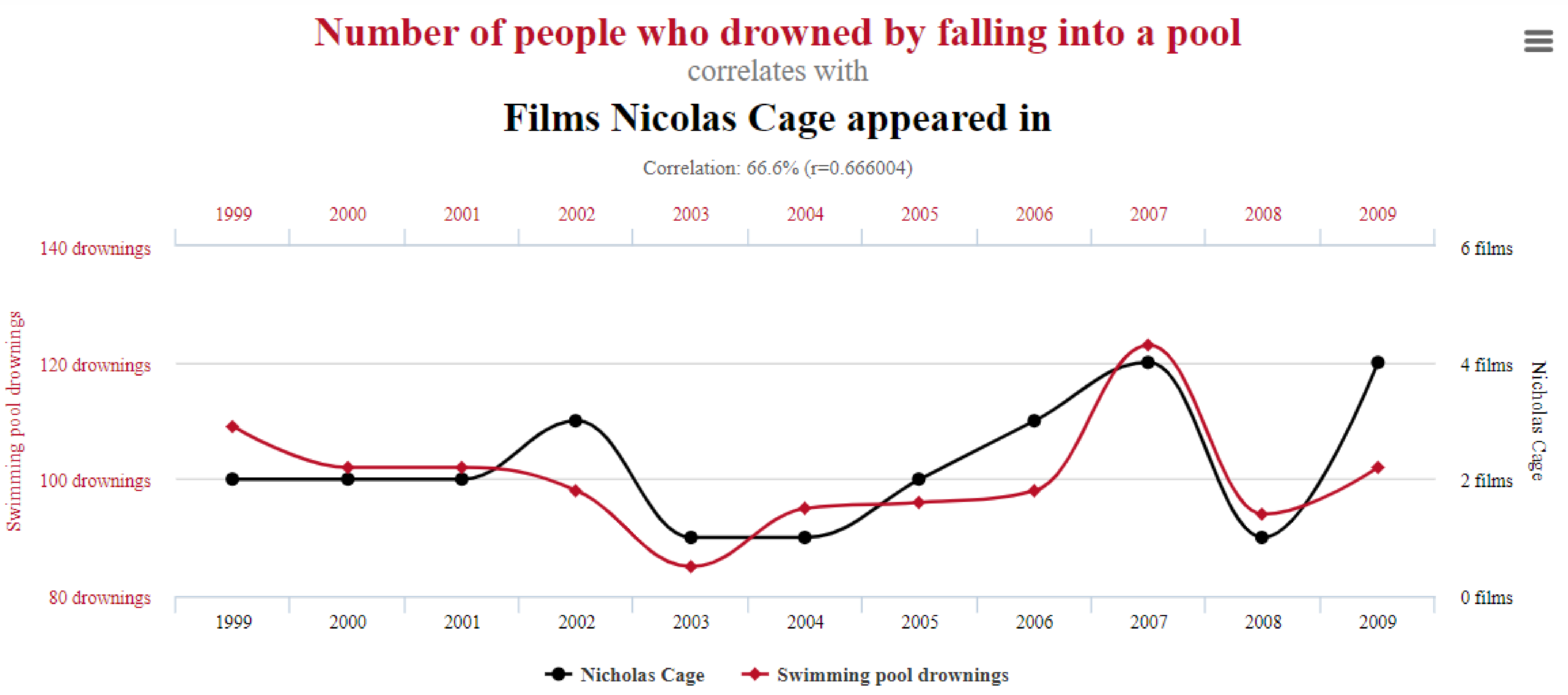
Dane wielowymiarowe

Czy wysoka korelacja musi oznaczać, że występuje zależność lub przyczynowość pomiędzy danymi?

Dane wielowymiarowe



Dane wielowymiarowe

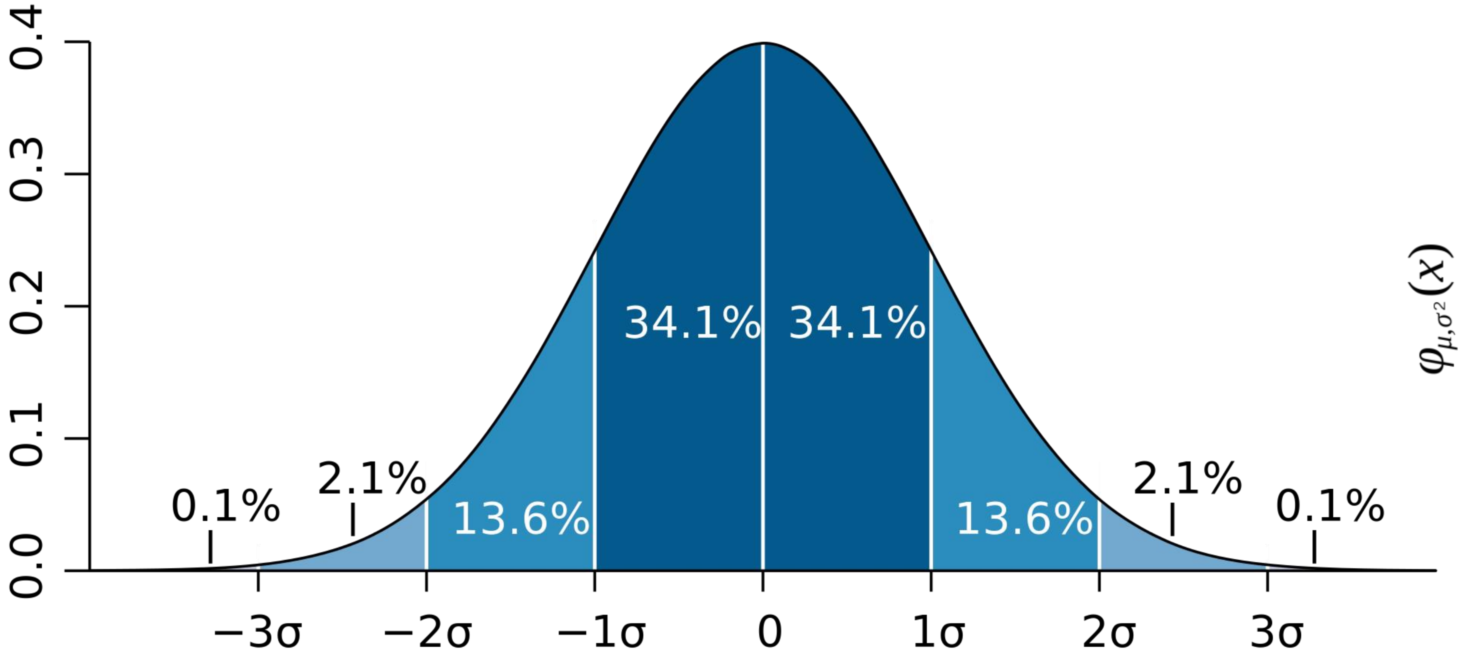


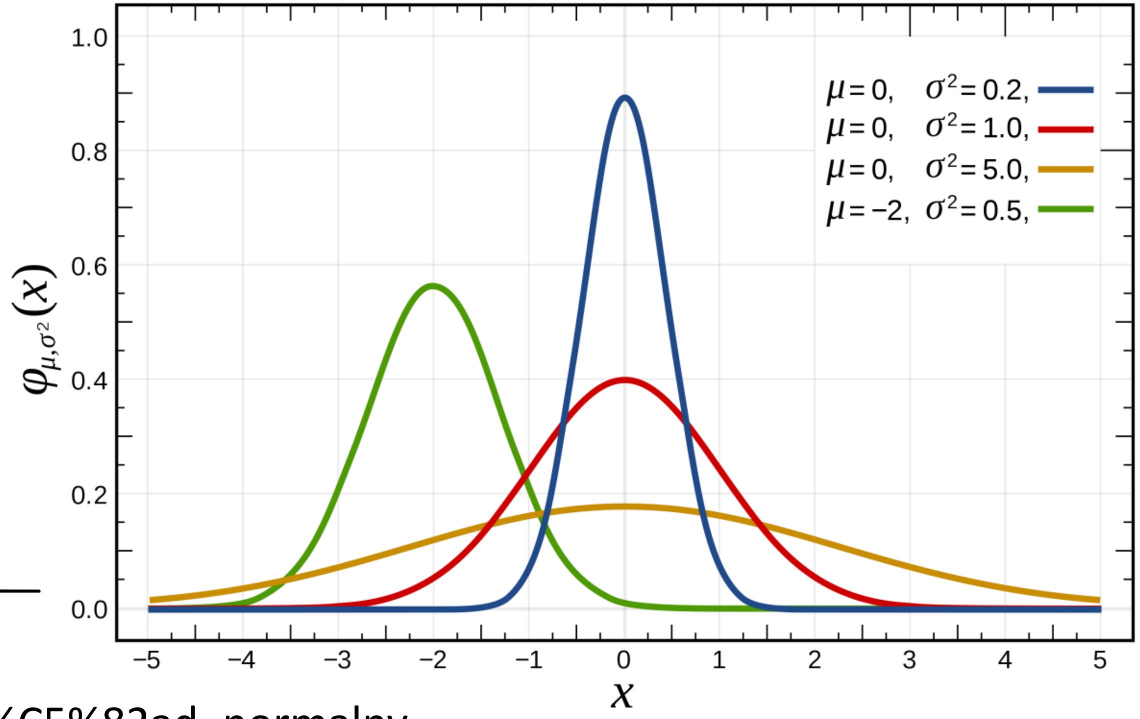
Notebook – D09_Z02

Wielowymiarowy rozkład normalny

Jednowymiarowy rozkład normalny określony jest wzorem

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$





https://pl.wikipedia.org/wiki/Rozk%C5%82ad_normalny

Wielowymiarowy rozkład normalny

Jednowymiarowy rozkład normalny definiowany jest przez średnią i odchylenie standardowe (pierwiastek z wariancji), natomiast wielowymiarowy rozkład normalny definiowany jest przez wektor średnich (zawierający średnią wartość dla każdego wymiaru) oraz macierz kowariancji Σ (uogólnienie pojęcia wariancji na przypadek wielowymiarowy).

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

$\sigma_i^2 - wariancja\ zmiennej\ X_i$

$\sigma_{ij} - kowariancja\ między\ zmiennymi\ X_i\ i\ X_j$

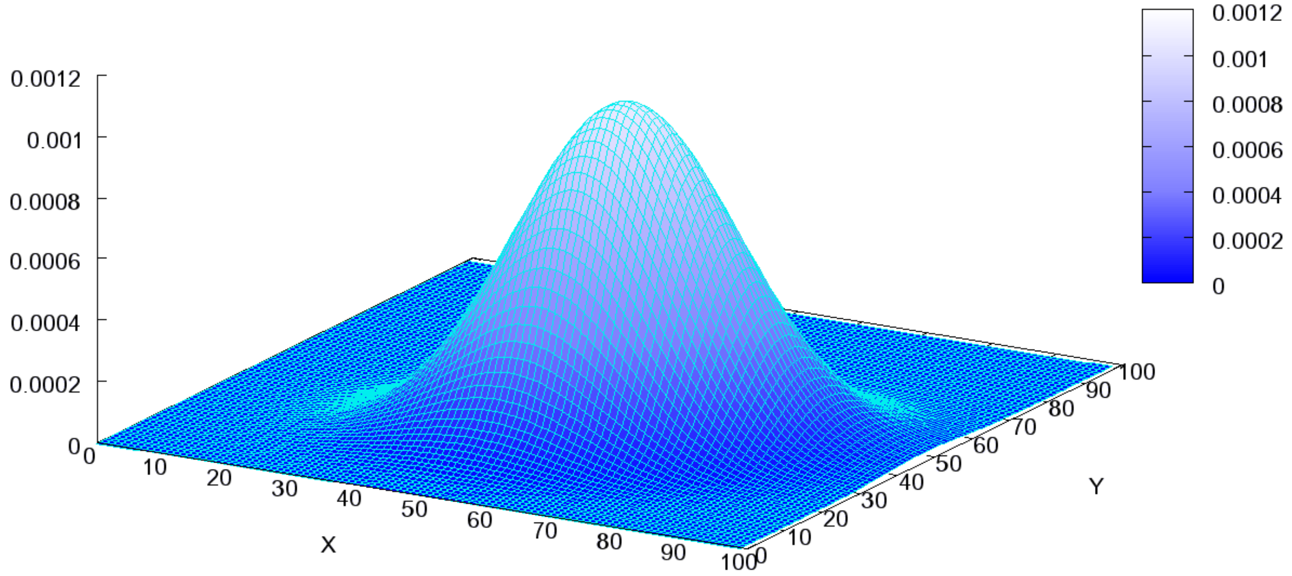
Wielowymiarowy rozkład normalny

N-wymiarowy rozkład normalny dla macierzy kowariancji Σ oraz średniej μ ma gęstość:

$$f_{\boldsymbol{\mu},\Sigma}(X) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\boldsymbol{\mu})^T \Sigma^{-1} (X-\boldsymbol{\mu})\right)$$

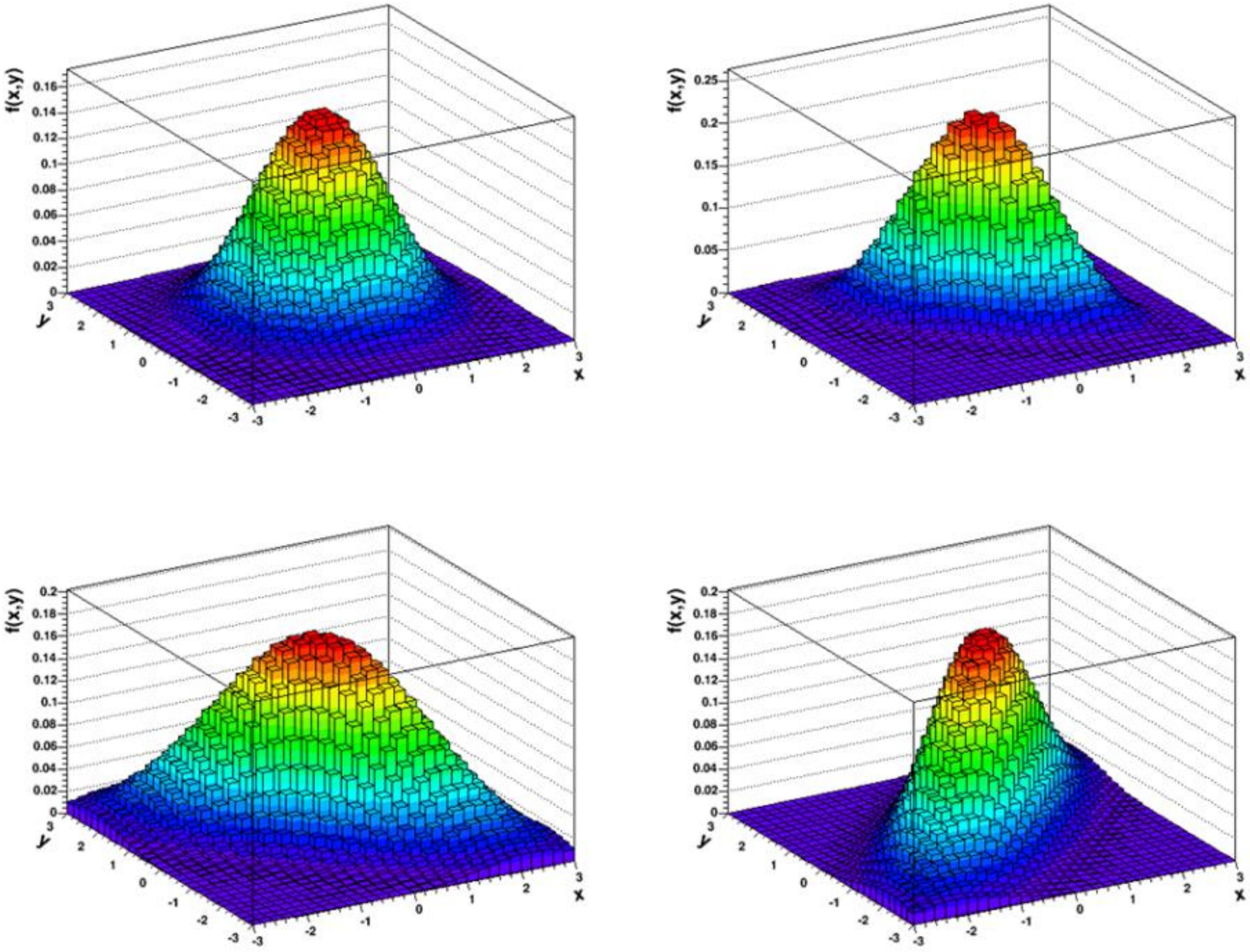
Oznacza się to w skrócie:

$$X \sim N(\boldsymbol{\mu}, \Sigma)$$

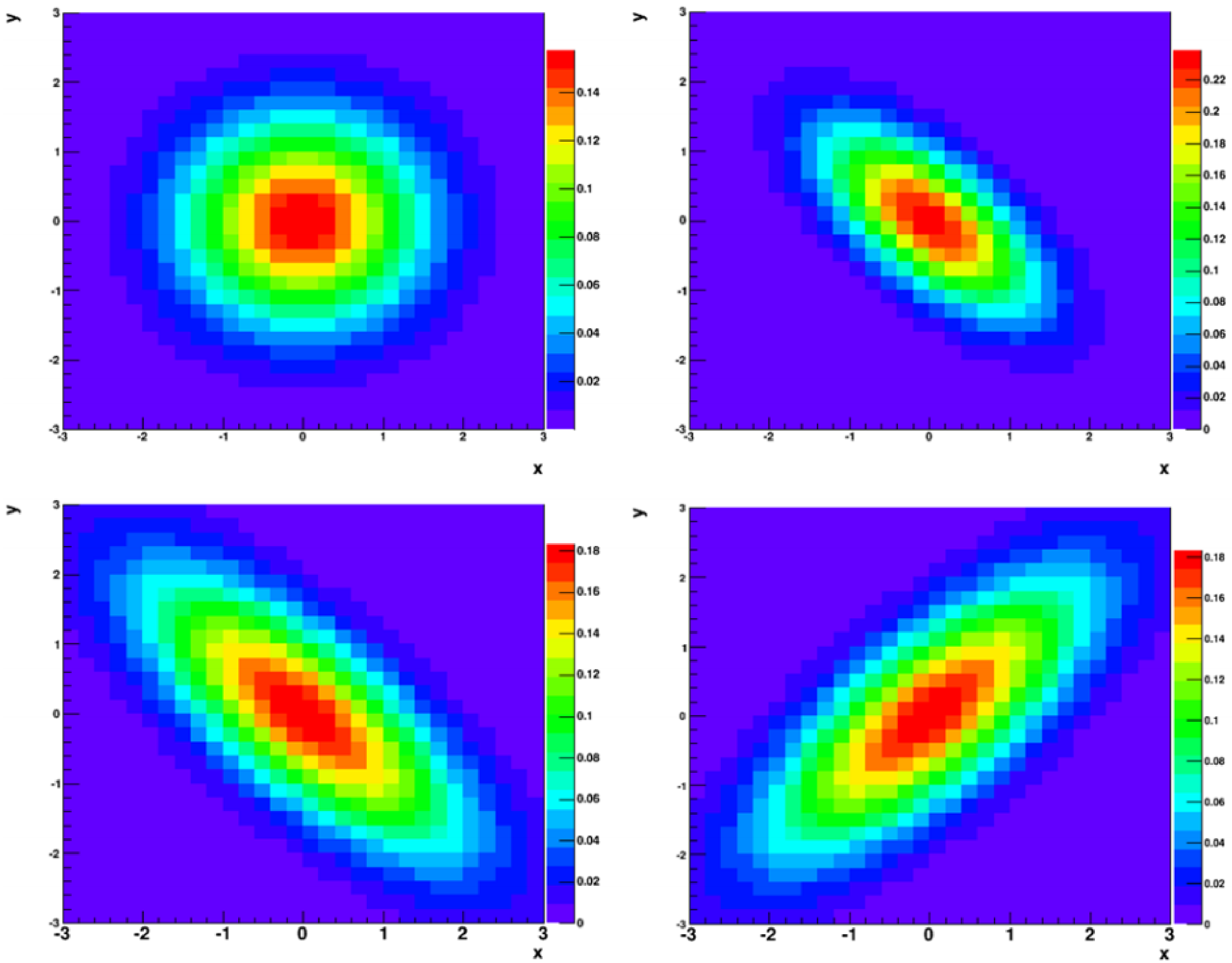


https://pl.wikipedia.org/wiki/Wielowymiarowy_rozk%C5%82ad_normalny

Wielowymiarowy rozkład normalny

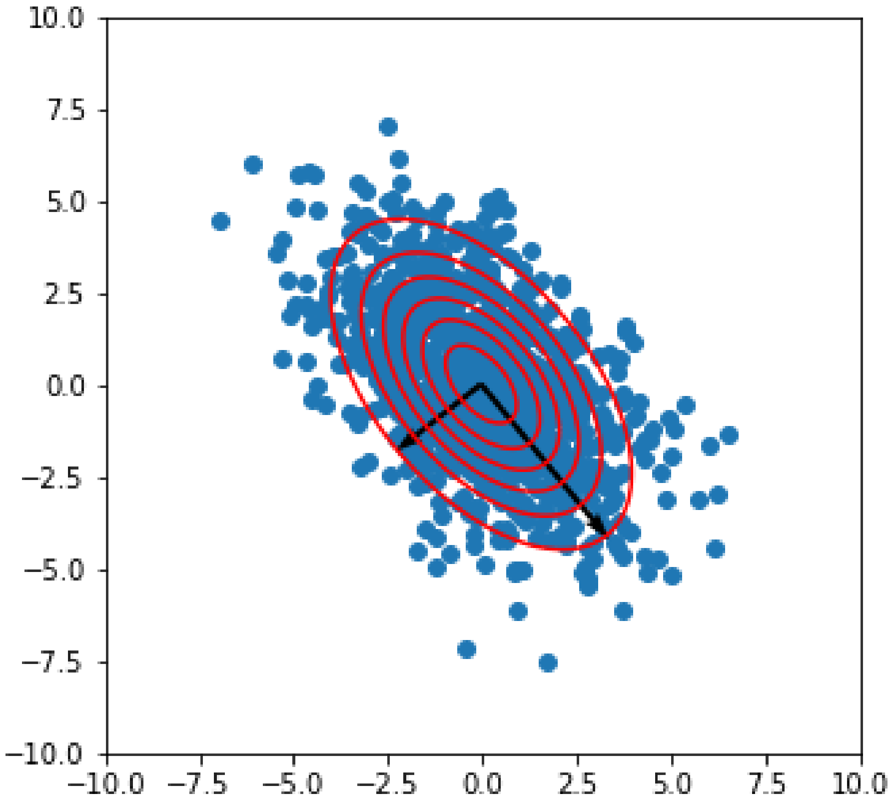


Wielowymiarowy rozkład normalny



Notebook – D09_Z03

Wielowymiarowy rozkład normalny



Notebook – D09_Z04