

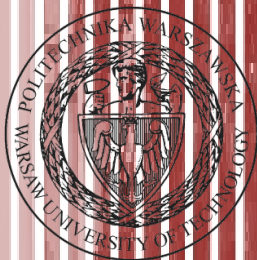


# Komputerowa analiza danych doświadczalnych

Wykład 7  
5.04.2019

dr inż. Łukasz Graczykowski  
[lukasz.graczykowski@pw.edu.pl](mailto:lukasz.graczykowski@pw.edu.pl)

*Semestr letni 2018/2019*



# Wielowymiarowy rozkład Gaussa - przypomnienie

## Centralne twierdzenie graniczne - przypomnienie

### Sploty

### Pobieranie próby, estymatory

# Wielowymiarowy rozkład Gaussa

- Gęstość prawdopodobieństwa wielowymiarowego rozkładu normalnego:  $\mathbf{X} = (X_1, X_2, \dots, X_n)$

$$\phi(\mathbf{x}) = k \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{a})^T B (\mathbf{x} - \mathbf{a})\right) \quad k = \left(\frac{\det B}{(2\pi)^n}\right)^{\frac{1}{2}}$$

- gdzie  $\mathbf{a}$  jest  $n$ -wymiarowym wektorem wartości oczekiwanych

$$E(\mathbf{X}) = \mathbf{a}$$

- Natomiast  $B$  jest dodatnio określoną macierzą symetryczną o wymiarze  $n \times n$  o następującej definicji:

$$C = E((\mathbf{X} - \mathbf{a})(\mathbf{X} - \mathbf{a})^T) = B^{-1}$$

– gdzie  $C$  jest macierzą kowariancji zmiennych losowych  $\mathbf{X}$

- Dla dwóch zmiennych losowych:  $\mathbf{X} = (X_1, X_2)$

$$C = B^{-1} = \begin{pmatrix} \sigma_1^2 & \text{cov}(X_1, X_2) \\ \text{cov}(X_1, X_2) & \sigma_2^2 \end{pmatrix}$$

# Wielowymiarowy rozkład Gaussa

- Jeżeli na moment uznamy, że zmienne losowe  $X_1$  i  $X_2$  są niezależne:

$$B_0 = \begin{pmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{pmatrix}$$

- Wstawiając  $B_0$  do ogólnego wzoru otrzymamy łączną gęstość dwóch niezależnych zmiennych losowych jako iloczyn dwóch rozkładów Gaussa w 1D, znajomy wzór:

$$\phi(x_1, x_2) = k \exp\left(-\frac{1}{2} \frac{(x_1 - a_1)^2}{\sigma_1^2}\right) \exp\left(-\frac{1}{2} \frac{(x_2 - a_2)^2}{\sigma_2^2}\right), \quad k = \frac{1}{2\pi\sigma_1\sigma_2}$$

- **Gdy zmienne losowe nie są niezależne (niezerowe kowariancje), musimy stosować wzór ogólny (poprzedni slajd)!**

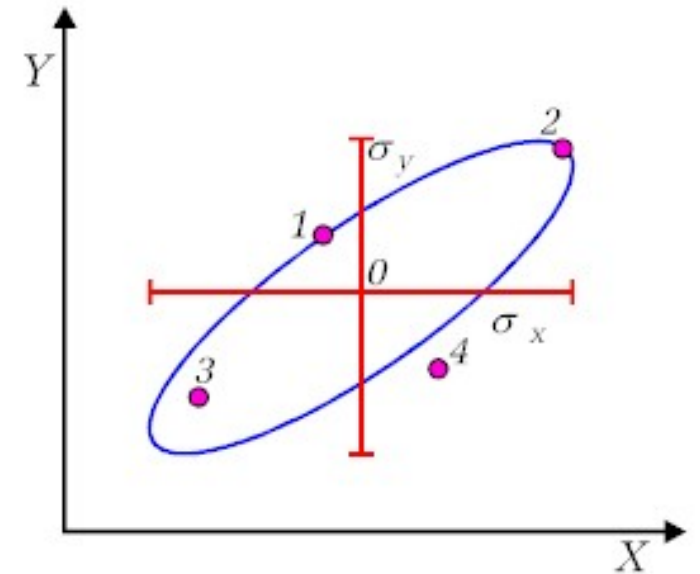
# Elipsa kowariancji

- Przekroje poziome funkcji gęstości prawdopodobieństwa rozkładu Gaussa mają kształt elipsy zwanej elipsą kowariancji:
  - elipsa kowariancji zależy od wartości oczekiwanych oraz odchyleń standardowych i kowariancji
  - elipsa kowariancji wyznacza obszar stałego prawdopodobieństwa
- Dla rozkładu 2D równanie elipsy (**elipsy kowariancji**) o środku w  $(a_1, a_2)$ , której osie główne tworzą kąt  $\alpha$  z osiami głównymi  $x_1, x_2$ :

$$\frac{(x_1 - a_1)^2}{\sigma_1^2} - 2\rho \frac{x_1 - a_1}{\sigma_1} \frac{x_2 - a_2}{\sigma_2} + \frac{(x_2 - a_2)^2}{\sigma_2^2} = 1 - \rho^2$$

$$\operatorname{tg} 2\alpha = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}$$

- Rysunek po prawej:
  - punkty 1 i 2 mają takie samo prawd.
  - prawd. punktu 3 jest większe niż 4



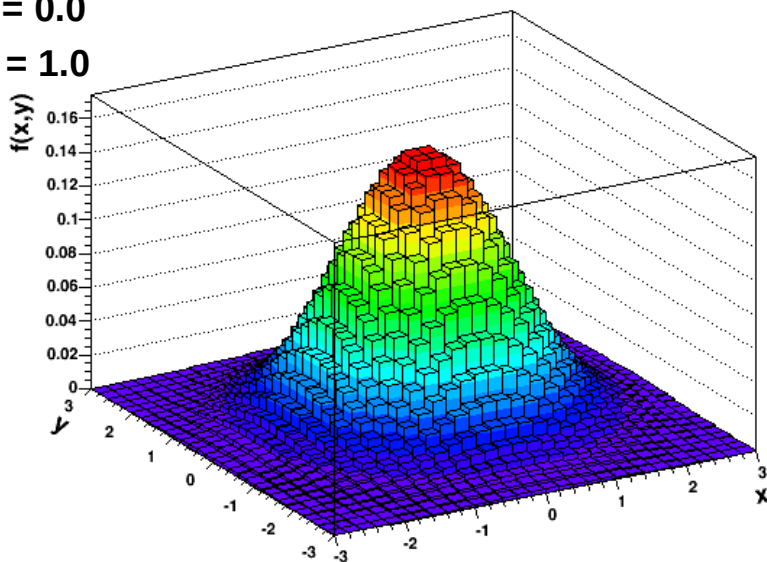


# Elipsa kowariancji

$$\text{cov}(X_1, X_2) = 0.0$$

$$a_1 = a_2 = 0.0$$

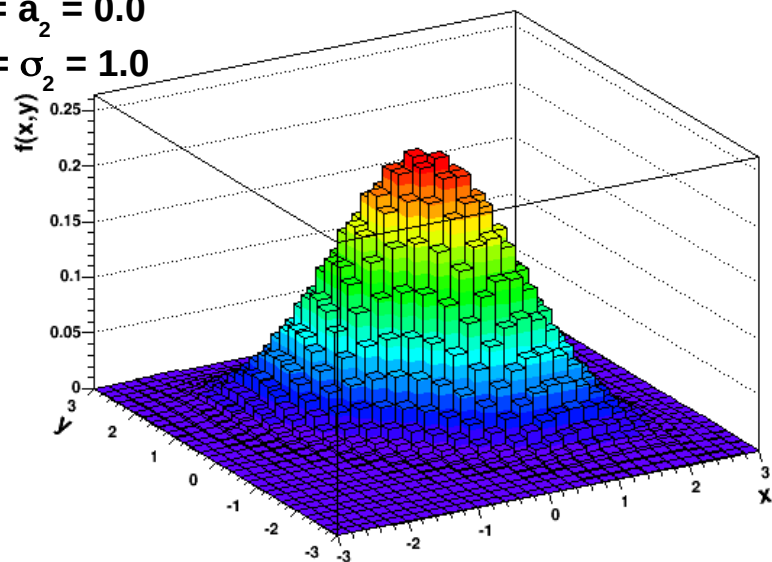
$$\sigma_1 = \sigma_2 = 1.0$$



$$\text{cov}(X_1, X_2) = 0.75$$

$$a_1 = a_2 = 0.0$$

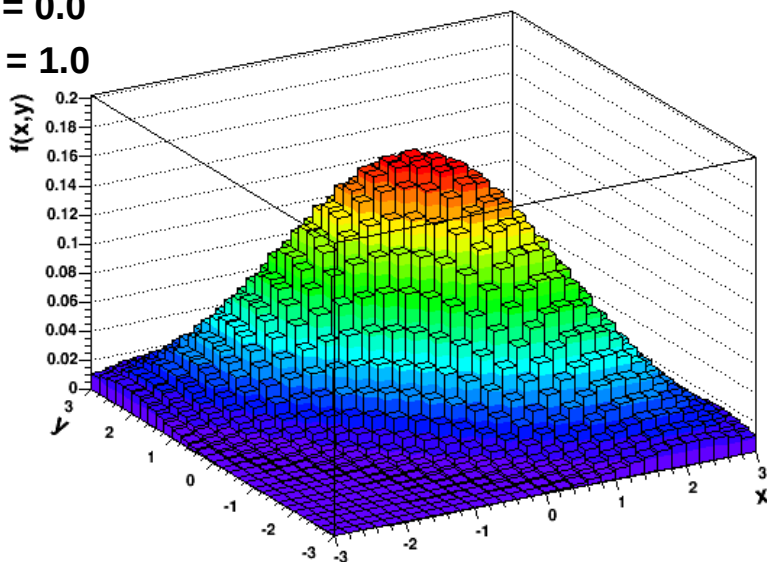
$$\sigma_1 = \sigma_2 = 1.0$$



$$\text{cov}(X_1, X_2) = 0.5$$

$$a_1 = a_2 = 0.0$$

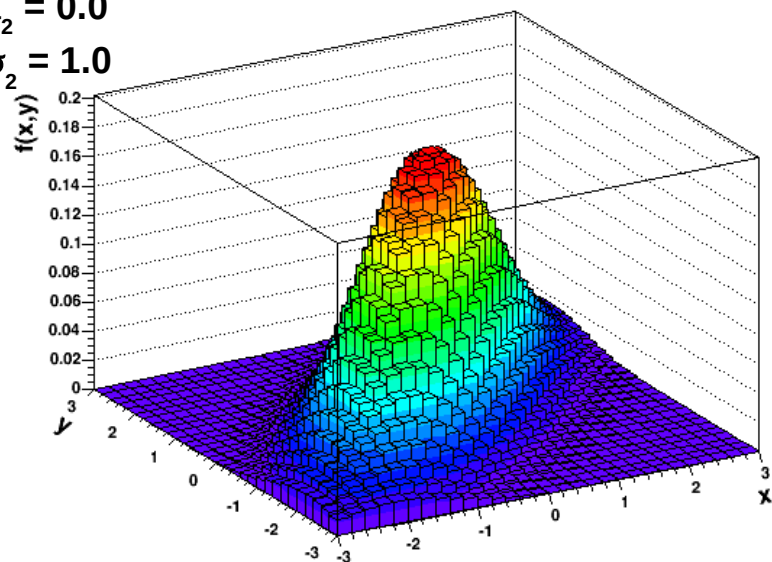
$$\sigma_1 = \sigma_2 = 1.0$$



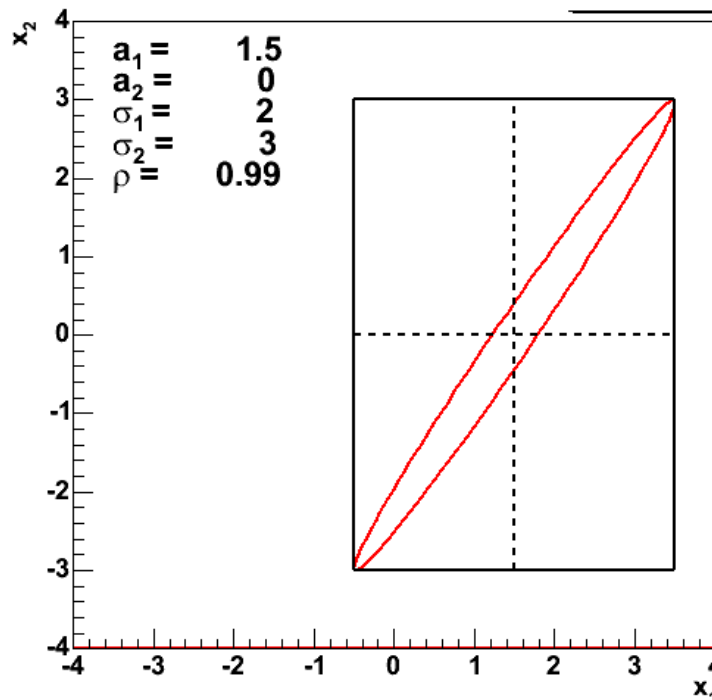
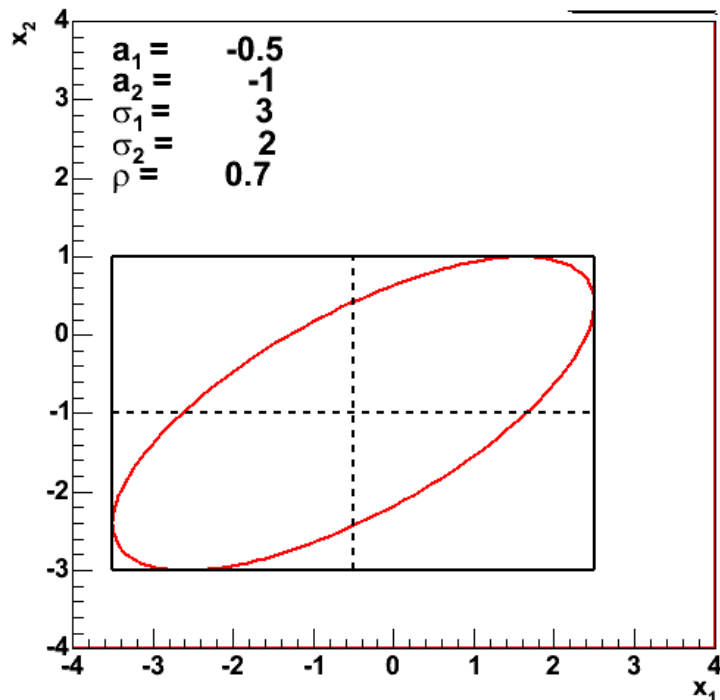
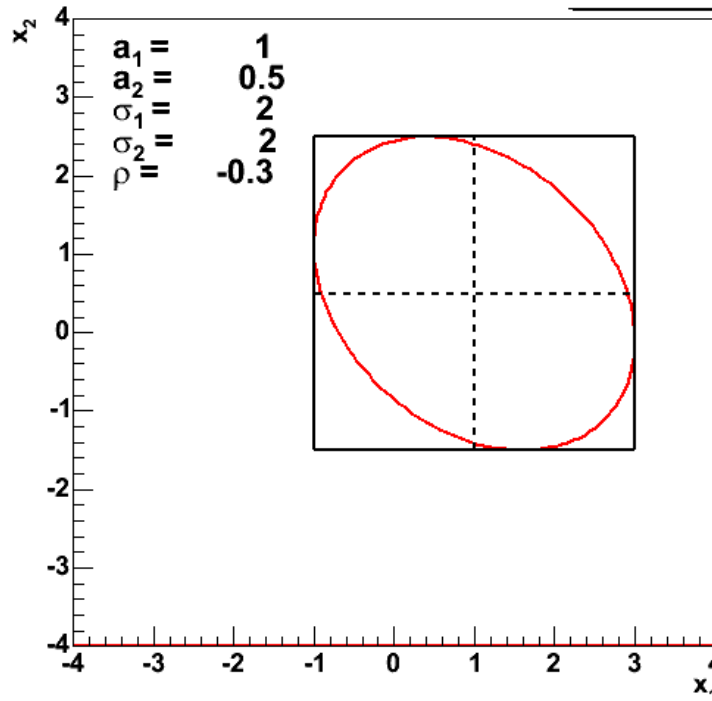
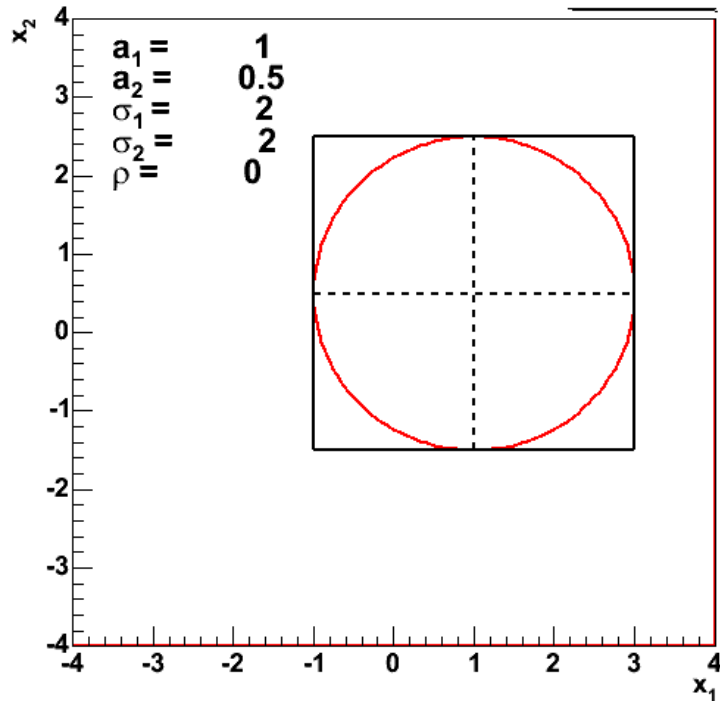
$$\text{cov}(X_1, X_2) = -0.5$$

$$a_1 = a_2 = 0.0$$

$$\sigma_1 = \sigma_2 = 1.0$$

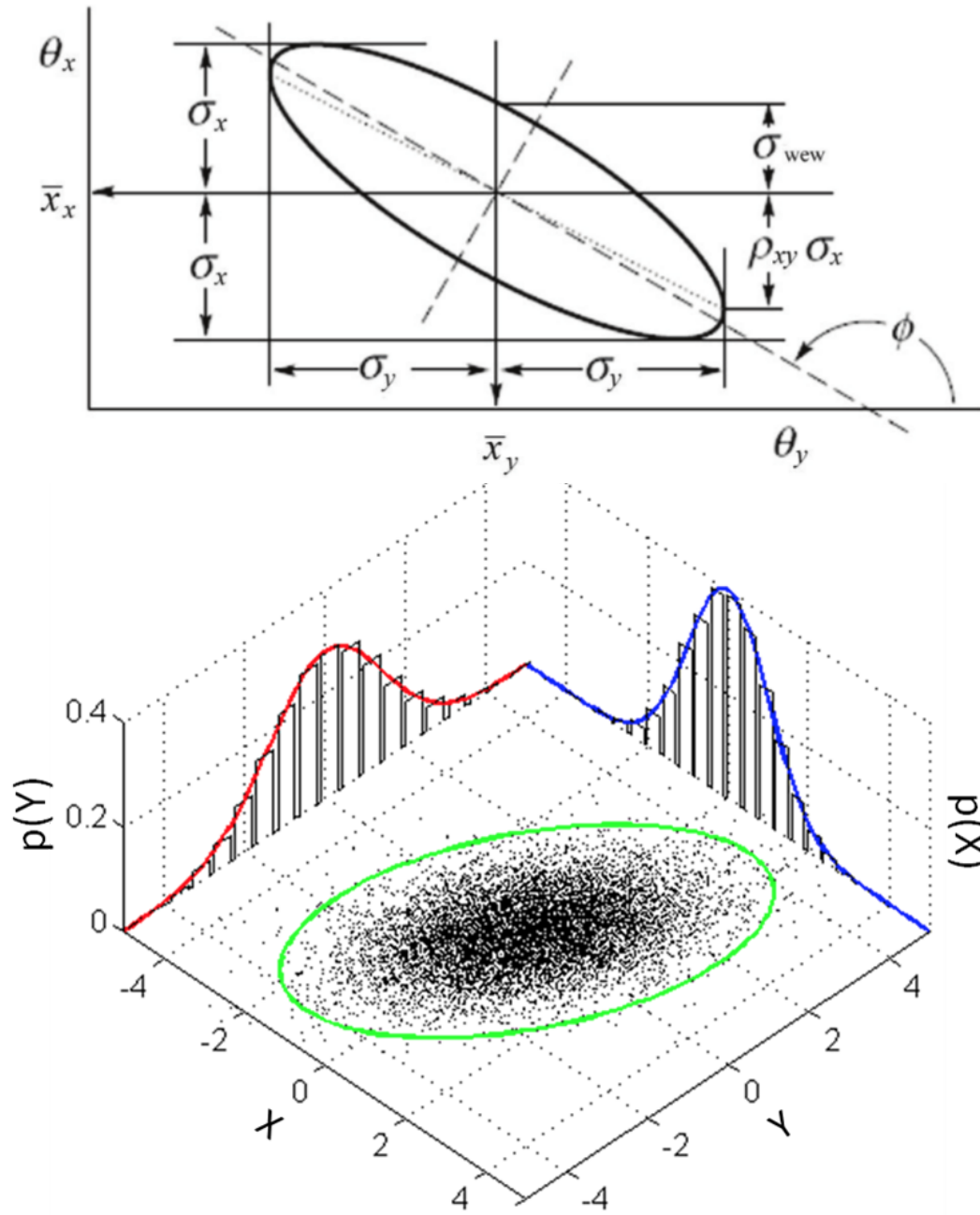


# Elipsa kowariancji



- Korelacja wydłuża i obraca elipsę
- Rozmiar elipsy zależy od wariancji
- Elipsa kowariancji zawiera **pełną** informację o macierzy kowariancji (w przypadku 2D)
- W 3D – elipsoida kowariancji
- W nD – hiperelipsoida kowariancji

# Elipsa kowariancji



- Każda elipsa kowariancji określa obszar prawdopodobieństwa analogicznie jak w przypadku 1D:

$$P(|X - a| \leq \sigma) = 68,3 \%$$

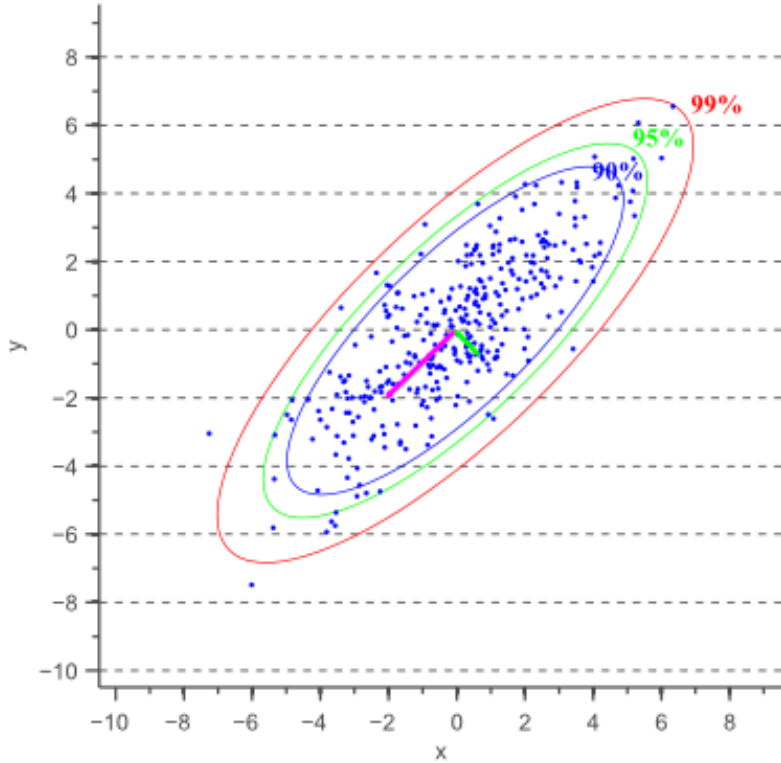
- Wartość prawdopodobieństwa wewnątrz elipsy zależy od ilości wymiarów, w 2D (dla elipsy  $1\sigma$ ):

$$P = 39,3 \%$$

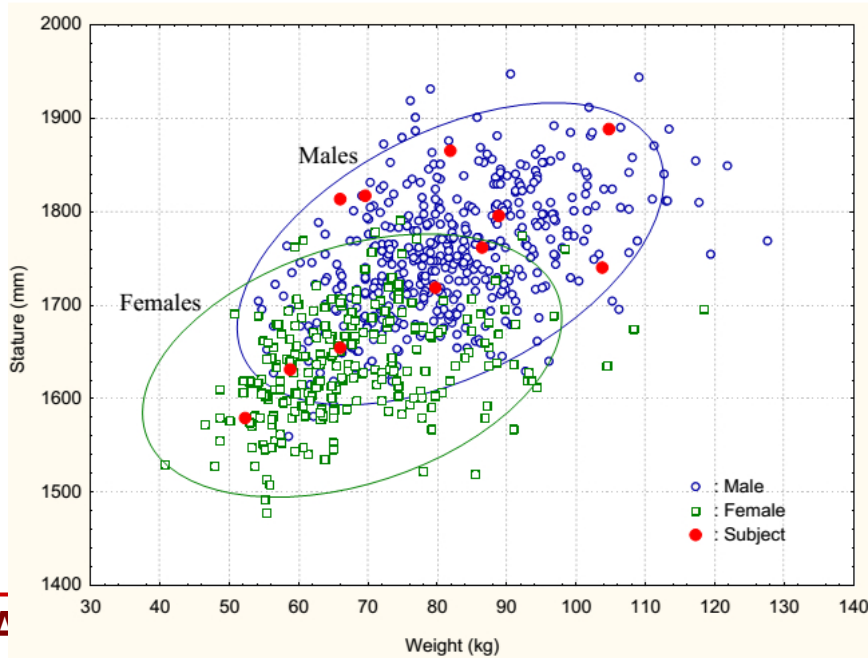
- Inne linie stałego prawdopodobieństwa (elipsy) wyznaczają inne wartości prawdopodobieństwa



# Elipsa kowariancji – wykorzystanie

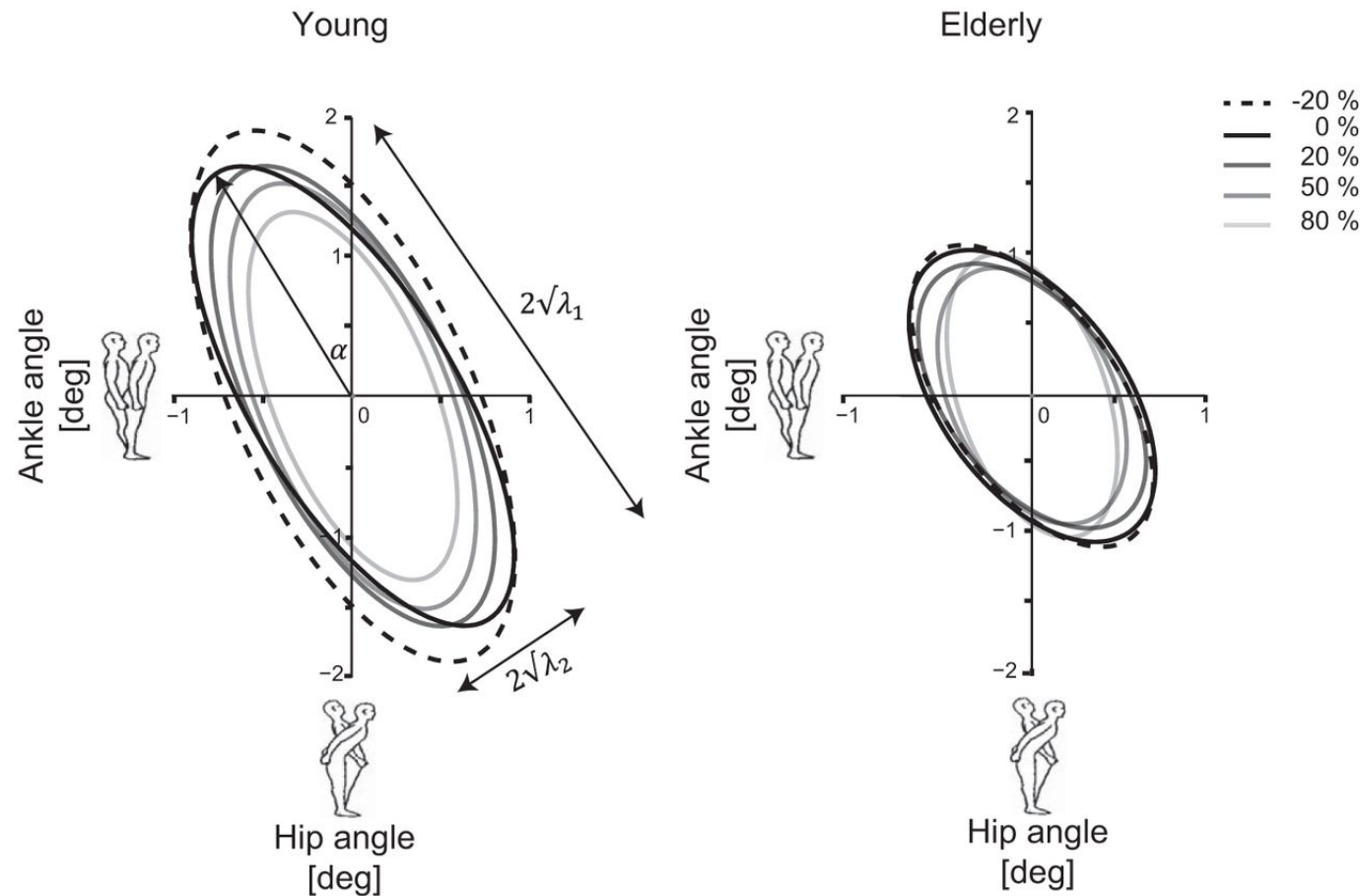


- Elipsy stałego prawdopodobieństwa mają ścisłe powiązanie z przedziałami ufności (o nich w przyszłości)
- Np. najczęściej określa się elipsę zawierającą prawdopodobieństwo 95% z wyników danych
- Przykład – korelacja wzrostu (*stature*) - wagi (*weight*) człowieka
- Analizy tego typu (dwóch lub więcej zmiennych jednocześnie) nazywamy analizą (statystyką) wielowymiarową (*multivariate analysis, statistics*)

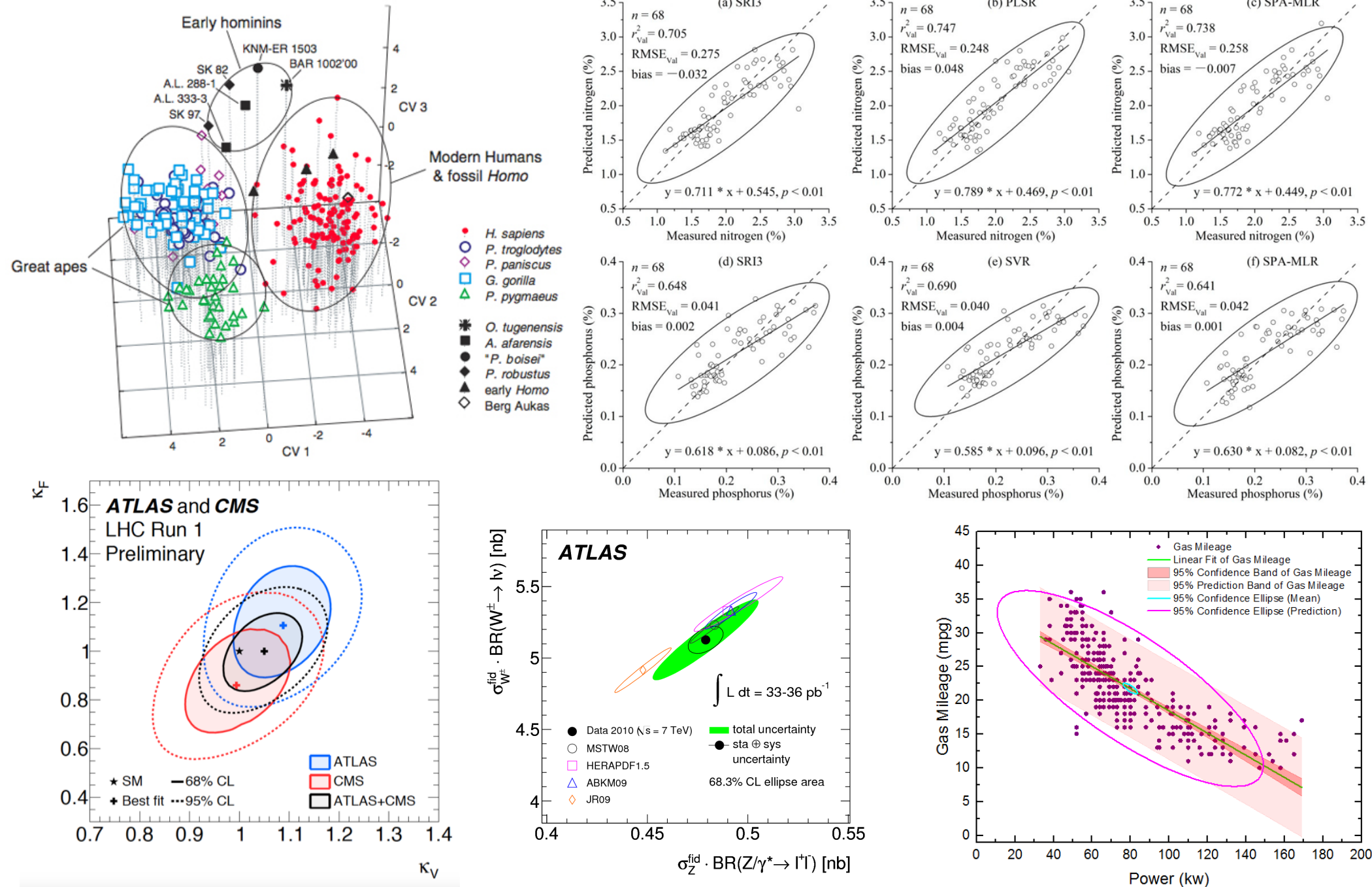


# Elipsa kowariancji – wykorzystanie

- Zależność kąta zgięcia w kostce od kąta zgięcia w biodrze u młodszych i starszych osób



# Elipsa kowariancji – wykorzystanie





# Centralne twierdzenie graniczne

# Centralne twierdzenie graniczne

- **Centralne twierdzenie graniczne** (*ang. central limit theorem*) – jedno z najważniejszych twierdzeń rachunku prawdopodobieństwa:
  - jeżeli zmienne losowe  $X_i$  są zmiennymi niezależnymi o jednakowych wartościach średnich  $a$  i odchyleniach standardowych  $b$ , to **rozkład normalny** ma zmienna:

$$\xi = \frac{1}{n} \sum_{i=1}^n X_i = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i \quad E(\xi) = a, \quad \sigma^2(\xi) = b^2/n$$

- rozkład normalny będzie mieć też zmienna:

$$X = \lim_{n \rightarrow \infty} \sum_{i=1}^n X_i \quad E(X) = na, \quad \sigma^2(X) = nb^2$$

- Innymi słowy – mając  $n$  niezależnych zmiennych o jednakowym (**dowolnym!**) rozkładzie, to ich suma dla dużych  $n$  zbiega do rozkładu normalnego





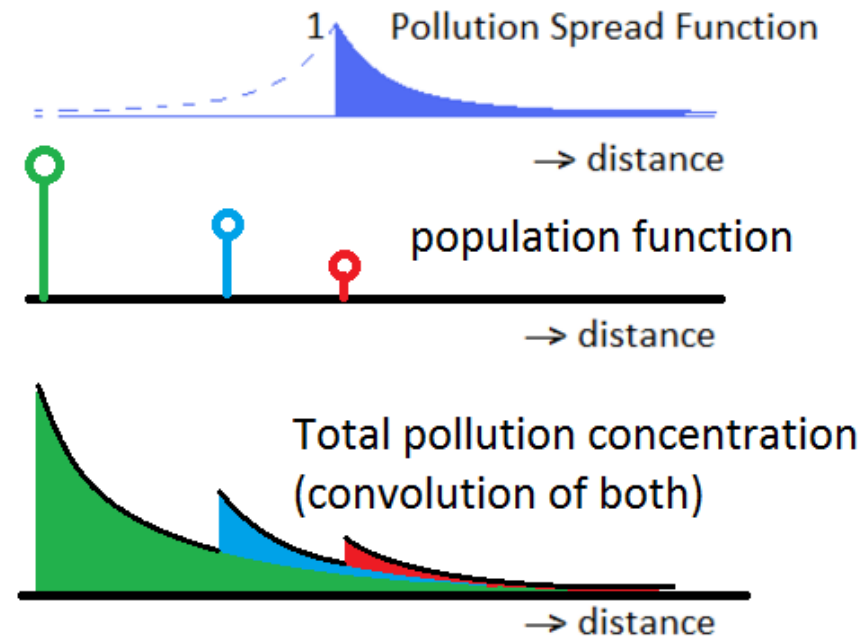
# Sploty



# Suma zmiennych losowych jako splot

<https://www.quora.com/The-density-function-of-the-sum-of-two-random-variables-is-the-convolution-of-their-respective-densities-What-is-the-intuition-behind-this>

- Wyobraźmy sobie taką sytuację:
  - Mieszkasz w wiosce obok rzeki
  - Mieszkańcy wioski wrzucają do rzeki odpady biologiczne
  - Koncentracja odpadów w funkcji odległości od miejsca zrzutu (*Pollution Spread Function, PSF*) jest zależna od ich rozkładu przez mikroorganizmy w rzece
  - Ilość wrzucanych odpadów zależy od populacji miejscowości na rzece

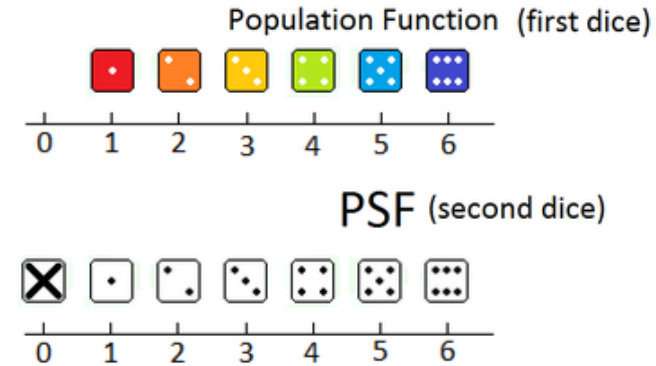


- Jaka jest pełna funkcja opisująca poziom zanieczyszczeń w rzece?
- Jest to **splot** dwóch rozkładów – funkcji populacji oraz funkcji koncentracji odpadów
- Innymi słowy, zastępujemy każdy punkt w funkcji populacji przez funkcję koncentracji przeskalowaną przez wagę funkcji populacji

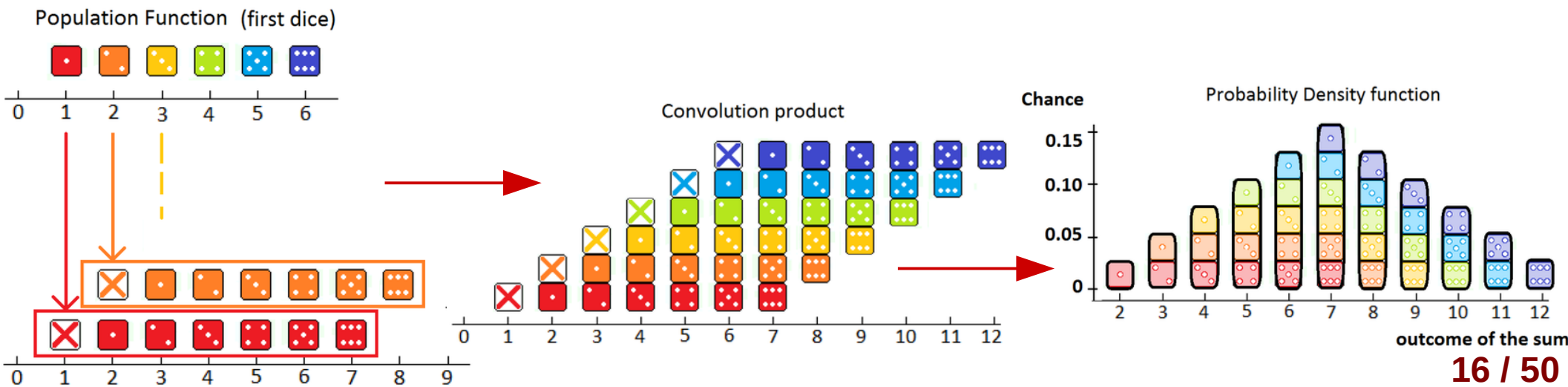
# Suma zmiennych losowych jako splot

<https://www.quora.com/The-density-function-of-the-sum-of-two-random-variables-is-the-convolution-of-their-respective-densities-What-is-the-intuition-behind-this>

- Zamieńmy teraz sytuację na kości do gry
- Pierwszy rzut kostką to funkcja populacji, 16,7% populacji mieszka 1 km w dół rzeki 16,7% populacji 2 km, itd.



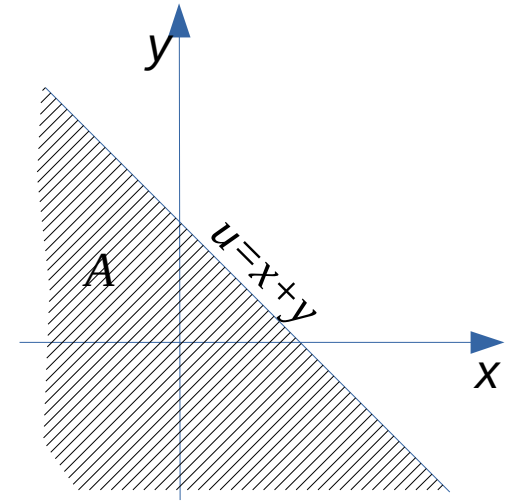
- Drugi rzut kostką oznacza funkcję PSF – jak bardzo dana miejscowość zanieczyszcza rzekę, i znowu 16,7% zanieczyszczeń ląduje 1 km dalej od miasta, 16.7% 2 km dalej od miasta, itp.
- Jak policzyć pełną funkcję zanieczyszczeń? Podmieniamy funkcję populacji poprzez funkcję zanieczyszczeń, dla każdego miasta



# Suma zmiennych losowych jako splot

- Rozważmy zmienną losową:  $U = X + Y$
- Zakładamy niezależność zmiennych:  $f(x, y) = f_x(x)f_y(y)$
- Wtedy dystrybuanta zmiennej  $U$ :

$$\begin{aligned} F(u) &= P(U \leq u) = P(X + Y \leq u) = \\ &= \iint_A f_x(x)f_y(y) dx dy \\ &= \int_{-\infty}^{\infty} f_x(x) dx \int_{-\infty}^{u-x} f_y(y) dy \\ &= \int_{-\infty}^{\infty} f_y(y) dy \int_{-\infty}^{u-y} f_x(x) dx \end{aligned}$$



Pole powierzchni A wyznacza taki obszar prawdopodobieństwa, że wartości  $u$  zmiennej losowej  $U = X + Y$  spełniają warunek:  $U \leq u$

Zgodnie z definicją dystrybuanty:

$$F(u) = P(U \leq u) = P((-\infty; u])$$

# Suma zmiennych losowych jako splot

- Z dystrybuanty wyznaczamy funkcję gęstości zmiennej  $U$ :

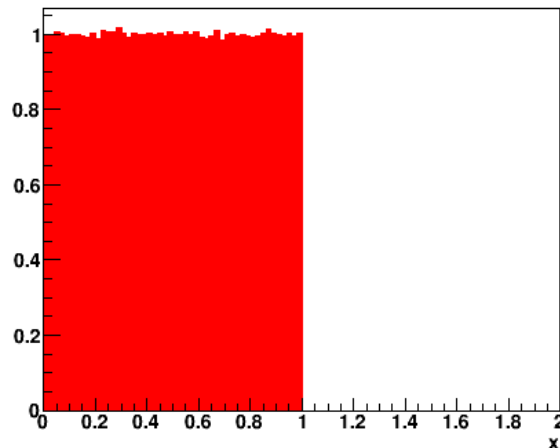
$$f(u) = \frac{dF(u)}{du} = \int_{-\infty}^{\infty} f_x(x) f_y(u-x) dx = \int_{-\infty}^{\infty} f_y(y) f_x(u-y) dy \equiv (f_x * f_y)(u)$$

- Funkcja  $f(u)$  tak zdefiniowana jest **splotem** funkcji  $f_x(x)$  i  $f_y(y)$
- Powyższy wzór jest prawdziwy również wówczas, jeżeli zmienne  $X$  i  $Y$  są zdefiniowane tylko w pewnym ograniczonym obszarze (wtedy ustalamy odpowiednie – węższe i skończone granice całkowania)
- Rozpatrzmy przypadek splotu dwóch rozkładów jednorodnych:

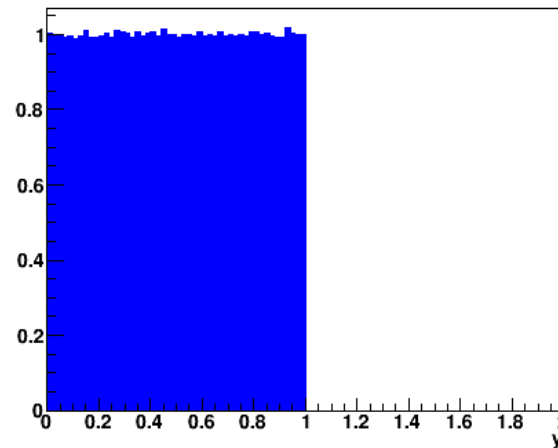
$$f_x(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

$$f_y(y) = \begin{cases} 1, & 0 \leq y < 1 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

Rozkład jednostajny



Rozkład jednostajny



# Suma zmiennych losowych jako splot

- Splot dwóch rozkładów jednorodnych:

$$f_x(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

$$f_y(y) = \begin{cases} 1, & 0 \leq y < 1 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

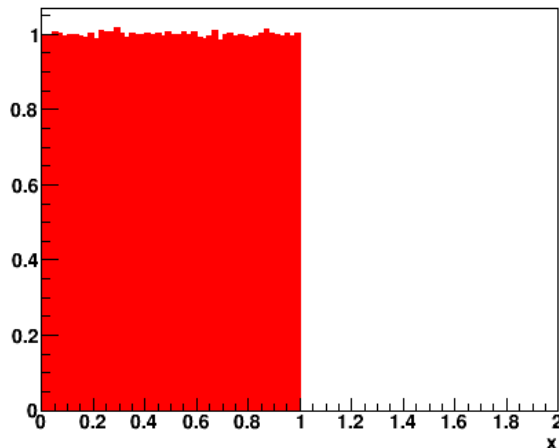
$$f(u) = \int_0^1 f_x(x) f_y(u-x) dx = \int_0^1 1 \cdot f_y(u-x) dx \quad \begin{matrix} v=u-x \\ dv=-dx \end{matrix} \Rightarrow f(u) = - \int_u^{u-1} f_y(v) dv = \int_{u-1}^u f_y(v) dv$$

- Zmienna  $u$  zmienia się od 0 do 2, zatem rozważmy 2 przypadki:

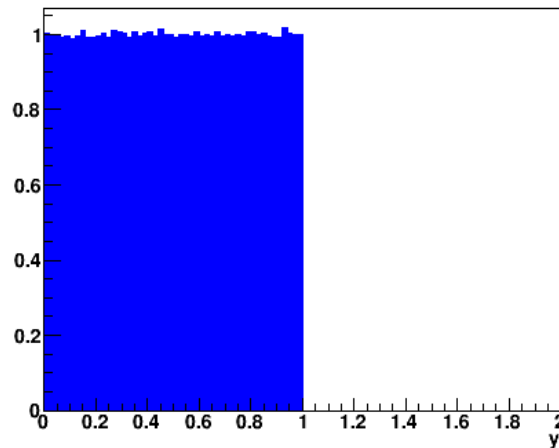
$$(a) \quad 0 \leq u < 1: f_1(u) = \int_0^u f_y(v) dv = \int_0^u 1 dv = u$$

$$(b) \quad 1 \leq u < 2: f_2(u) = \int_{u-1}^1 f_y(v) dv = \int_{u-1}^1 1 dv = 2 - u$$

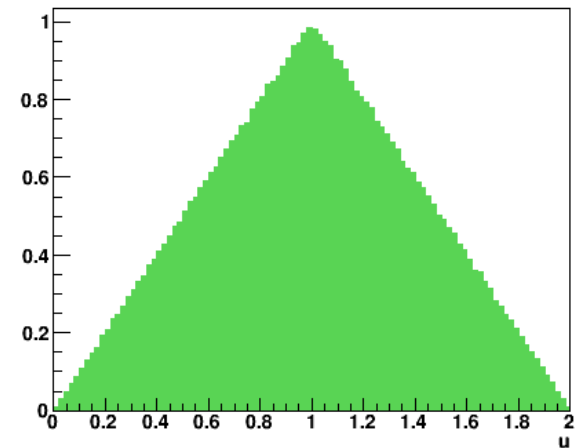
Rozkład jednostajny



Rozkład jednostajny



Splot 2 rozkładów jednostajnych



# Suma zmiennych losowych jako splot

- Rozpatrzmy przypadek splotu dwóch rozkładów jednorodnych:

$$f_x(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

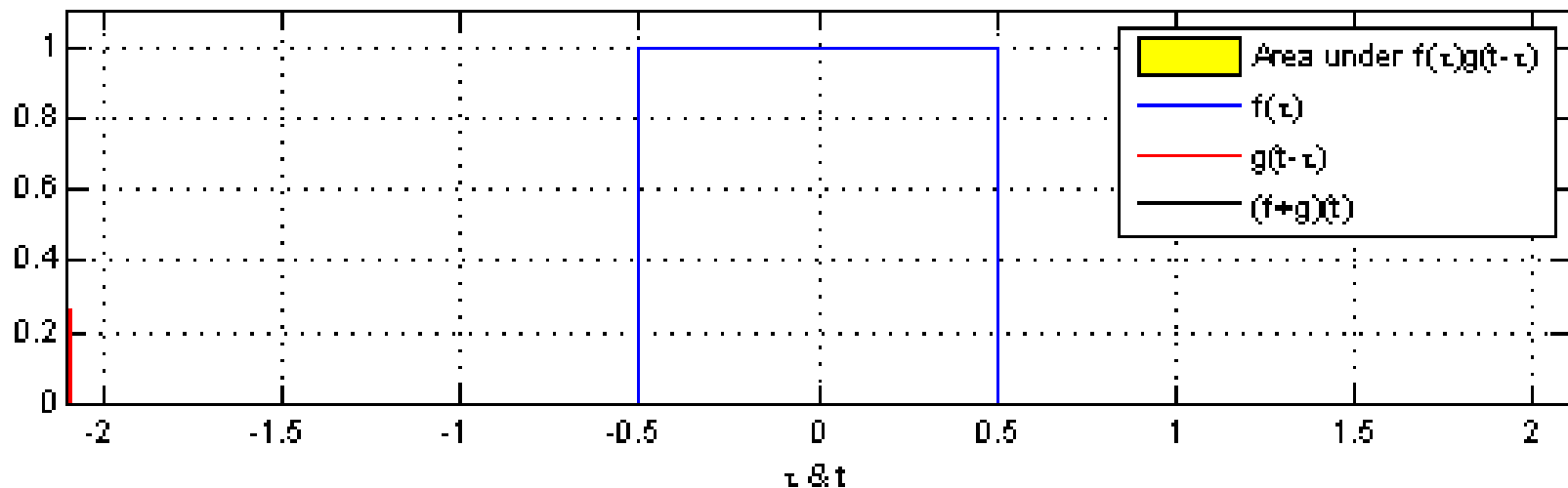
$$f_y(y) = \begin{cases} 1, & 0 \leq y < 1 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

$$f(u) = \int_0^1 f_x(x) f_y(u-x) dx = \int_0^1 f_y(u-x) dx \quad \begin{matrix} v = u-x \\ dv = -dx \end{matrix} \Rightarrow f(u) = - \int_u^{u-1} f_y(v) dv = \int_{u-1}^u f_y(v) dv$$

- Zmienna  $u$  zmienia się od 0 do 2, zatem rozważmy 2 przypadki:

$$(a) \quad 0 \leq u < 1: f_1(u) = \int_0^u f_y(v) dv = \int_0^u 1 dv = u$$

$$(b) \quad 1 \leq u < 2: f_2(u) = \int_{u-1}^1 f_y(v) dv = \int_{u-1}^1 1 dv = 2 - u$$



[https://en.wikipedia.org/wiki/Convolution#/media/File:Convolution\\_of\\_box\\_signal\\_with\\_itself2.gif](https://en.wikipedia.org/wiki/Convolution#/media/File:Convolution_of_box_signal_with_itself2.gif)

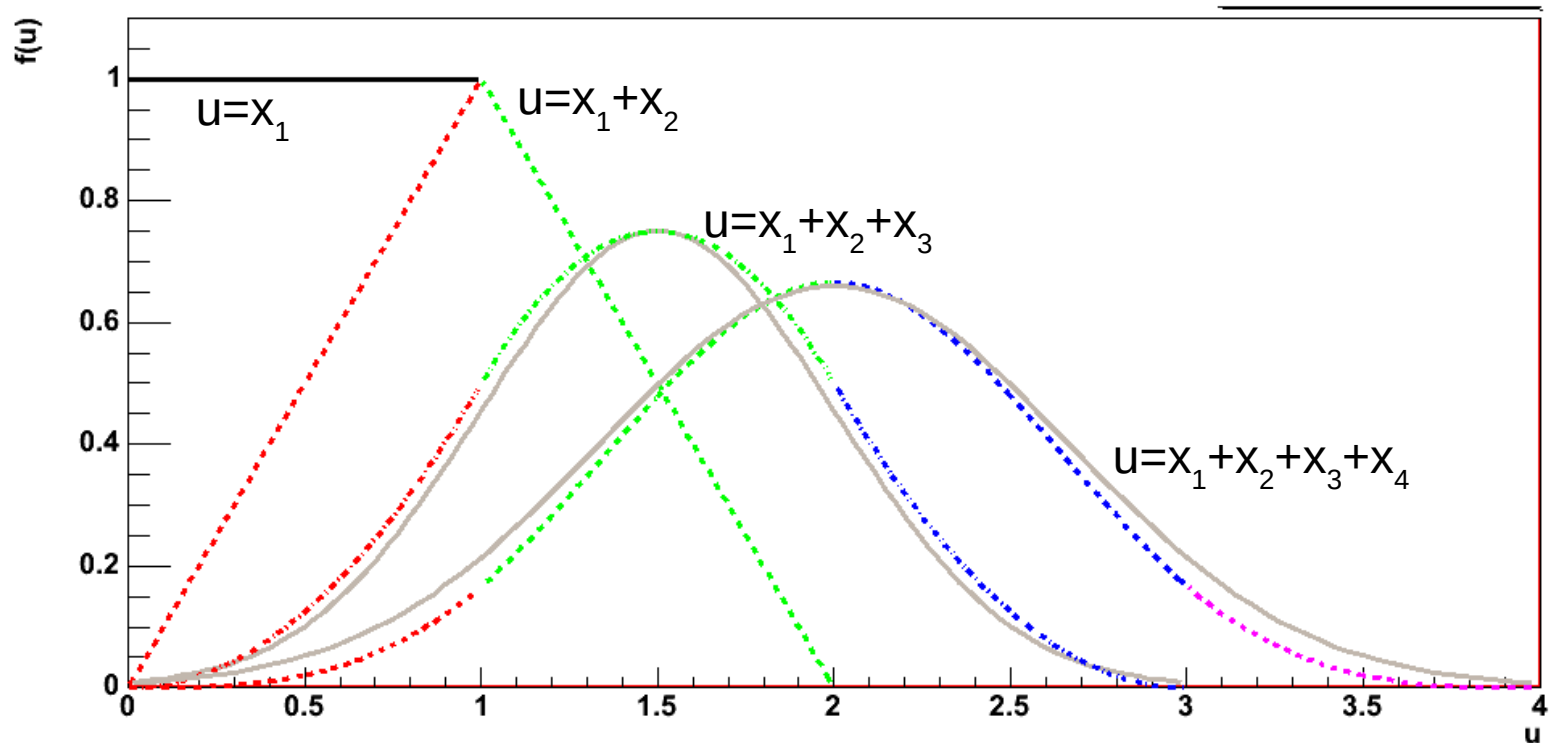


# Suma zmiennych losowych jako splot

- Analogicznie będzie z sumą trzech zmiennych losowych:

$$f(u) = \begin{cases} 1/2 u^2, & 0 \leq u < 1 \\ 1/2 (-2u^2 + 6u - 3), & 1 \leq u < 2 \\ 1/2 (u-3)^2, & 2 \leq u < 3 \end{cases}$$

- Zgodnie z CTG – im więcej rozkładów w splocie, tym bardziej rozkład sumy przypomina rozkład Gaussa:



# Sploty z rozkładem normalnym

- Przykład: Mierzmy zmienną  $X$  opisaną gęstością prawdopodobieństwa  $f_x(x)$ . Pomiar obarczony jest niepewnością  $Y$  mającą rozkład normalny. Wynik jest zatem sumą zmiennych losowych:  $U = X + Y$
- Gęstość prawdopodobieństwa zmiennej  $U$  wynosi wtedy:

$$f(u) = \int_{-\infty}^{\infty} f_x(x) f_y(u-x) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} f_x(x) \exp\left(-\frac{(u-x)^2}{2\sigma^2}\right) dx$$

- Problem: eksperymentalnie otrzymujemy funkcję  $f(u)$ , ale tak naprawdę interesuje nas  $f_x(x)$ . Jak ją wyznaczyć?
  - w ogólnym przypadku jest to niemożliwe
  - można tego dokonać dla pewnej ograniczonej klasy funkcji  $f(u)$
  - najczęściej posługujemy się tutaj metodami Monte Carlo

# Sploty z rozkładem normalnym – przykład 1

- Przykład: Splot rozkładu jednostajnego z rozkładem normalnym (o średniej równej 0)
- W tym przypadku możliwe jest rozwiązanie analityczne. Korzystamy ze wzorów:

$$f(x) = \frac{1}{b-a}; x \in \langle a, b \rangle \quad g(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} \quad h(u) = \int_{-\infty}^{\infty} f(x)g(u-x)dx$$

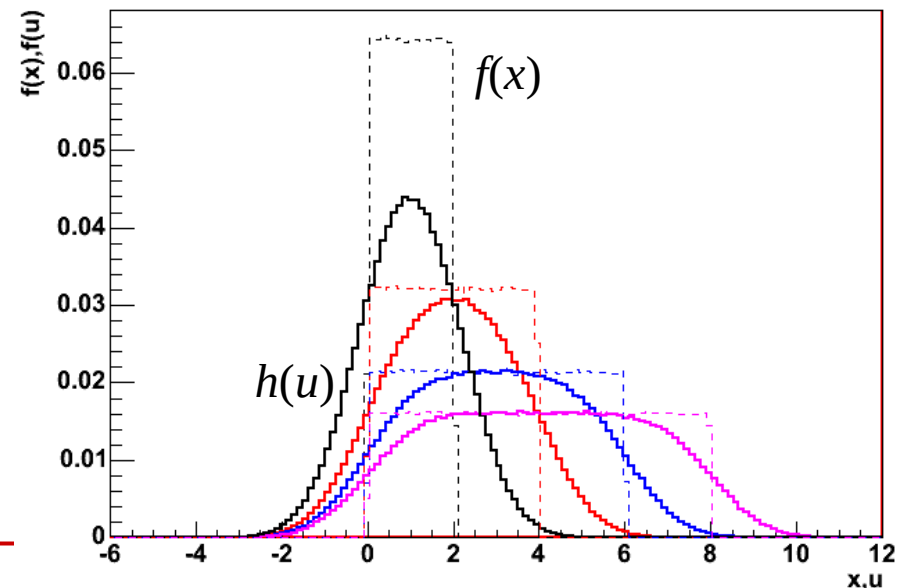
$$f(x) = 0; x \in \mathbb{R} \setminus \langle a, b \rangle$$

- Wtedy, wprowadzając zmienną  $v = (x-u)/\sigma$  otrzymujemy:

$$h(u) = \frac{1}{b-a} \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp\left(-\frac{(u-x)^2}{2\sigma^2}\right) dx = \frac{1}{b-a} \frac{1}{\sqrt{2\pi}} \int_{(a-u)/\sigma}^{(b-u)/\sigma} \exp\left(-\frac{1}{2}v^2\right) dv$$

- Z uwzględnieniem stabilizowanej dystrybuanty rozkładu normalnego:

$$h(u) = \frac{1}{b-a} \left( \Phi_0\left(\frac{b-u}{\sigma}\right) - \Phi_0\left(\frac{a-u}{\sigma}\right) \right)$$

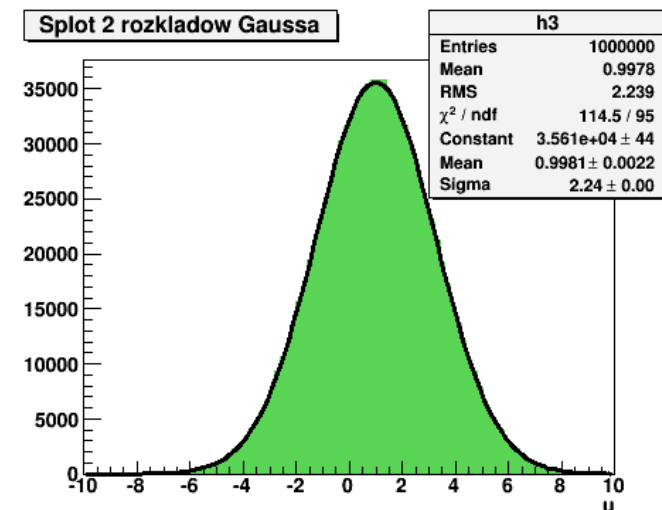
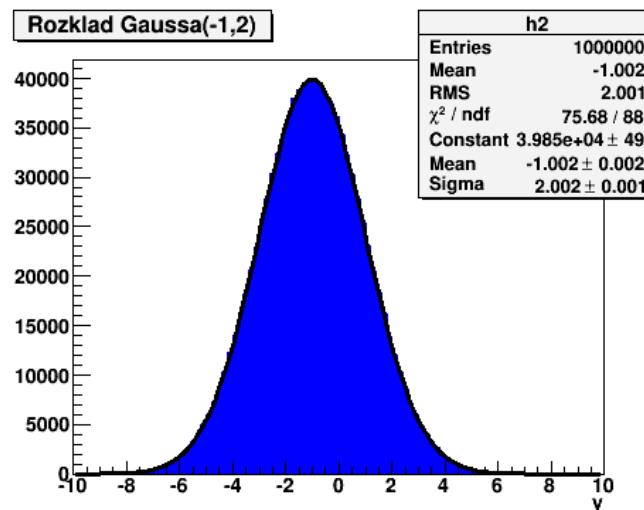
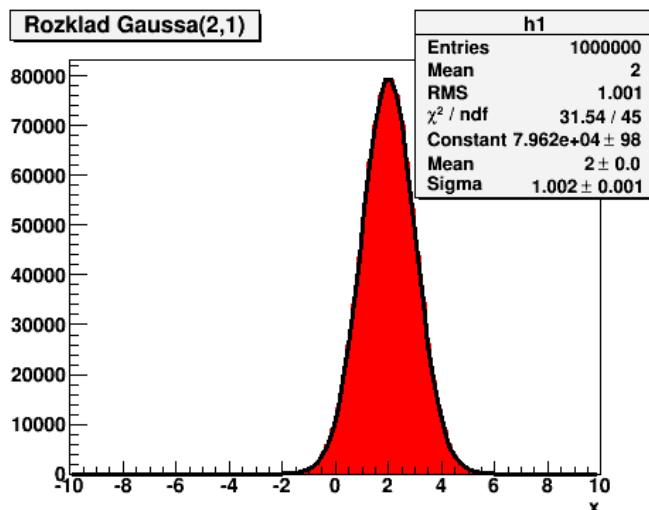


# Sploty z rozkładem normalnym – przykład 2

- Przykład: Splot dwóch rozkładów normalnych – dodawanie niepewności “w kwadracie”
- Splot dwóch rozkładów normalnych o wartościach średnich równych 0 i wariancjach  $\sigma_x$ ,  $\sigma_y$  ma postać rozkładu normalnego:

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-u^2/2\sigma^2\right), \quad \sigma^2 = \sigma_x^2 + \sigma_y^2$$

- Widzimy, że **wariancje się dodają** (odchylenia std. dodają się w kwadracie)
- Jeśli średnie rozkładów różne od 0 – **wartości oczekiwane również się dodają**



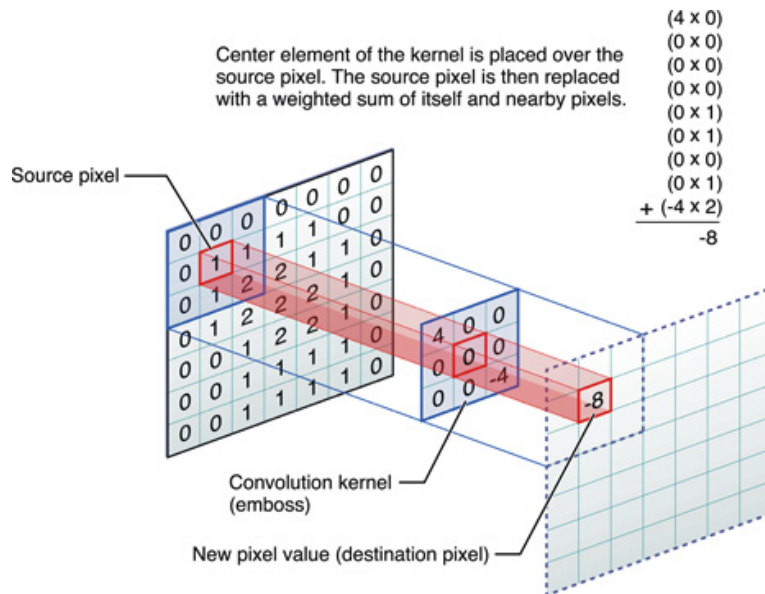
# Zastosowanie splotów

- Cyfrowe przetwarzanie obrazów
- Akustyka
- Muzyka elektroniczna
- W fizyce gdzie się pojawia superpozycja
- W planowaniu radioterapii (rozkłady dawki)



<https://upload.wikimedia.org/wikipedia/en/2/24/Lenna.png>  
Playboy 1972 – standardowy obrazek w grafice komputerowej.

	-2	-1	0
	-1	1	1
	0	1	2



Original



Emboss

<https://developer.apple.com/library/content/documentation/Performance/Conceptual/vImage/ConvolutionOperations/ConvolutionOperations.html>



# Zastosowanie splotów

- Bardzo ważnym zastosowaniem splotów są badania farmakokinetyczne leków – koncentracja leku w osoczu krwi w czasie jest splotem funkcji absorpcji leku oraz jego eliminacji

- The absorption rate  $r_{abs}$  that results in plasma concentration  $c(t)$  may be estimated by solving following eq.

$$c(t) = \int_0^t c_d(t-u)r_{abs}(u)du$$

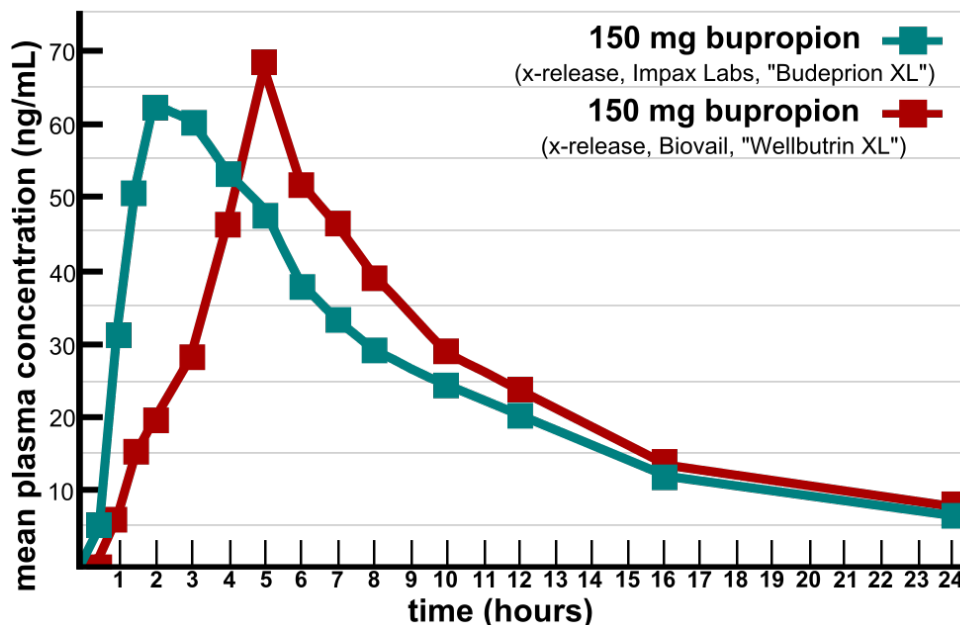
$C_d$  is the concentration time profile resulting from instantaneous absorption of a unit amount of drug which is typically absorbed from bolus IV injection or reference oral solution data

$c(t)$  is plasma conc. versus time profiles of tested formulation

$r_{abs}$  is the input rate of the oral solid dosage form into the body

$u$  is the variable of integration

<https://www.slideshare.net/jaspreetguraya/in-vitro-in-vivo-correlation-ivivc>



[https://upload.wikimedia.org/wikipedia/commons/7/7d/Bupropion\\_bioequivalency\\_comparison.svg](https://upload.wikimedia.org/wikipedia/commons/7/7d/Bupropion_bioequivalency_comparison.svg)

[https://www.researchgate.net/publication/228486042\\_In\\_Vitro-In\\_Vivo\\_Correlation\\_IVIVC\\_and\\_Determining\\_Drug\\_Concentrations\\_in\\_Blood\\_from\\_Dissolution\\_Testing-A\\_Simple\\_and\\_Practical\\_Approach](https://www.researchgate.net/publication/228486042_In_Vitro-In_Vivo_Correlation_IVIVC_and_Determining_Drug_Concentrations_in_Blood_from_Dissolution_Testing-A_Simple_and_Practical_Approach)

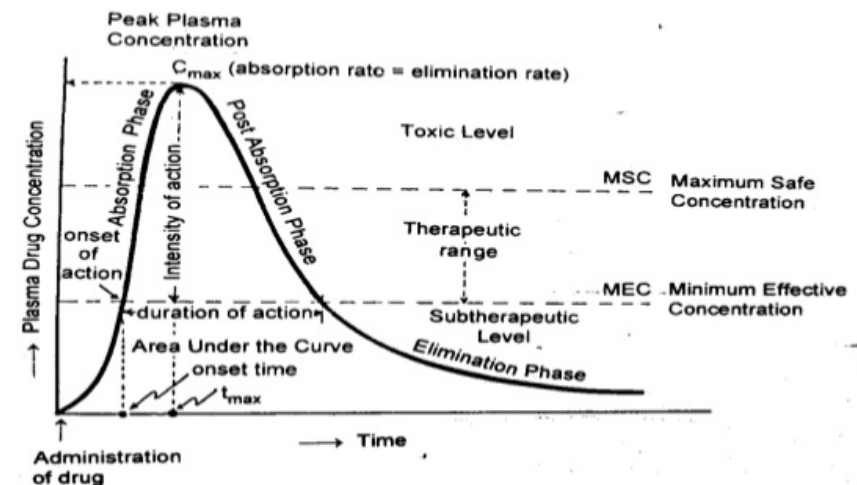
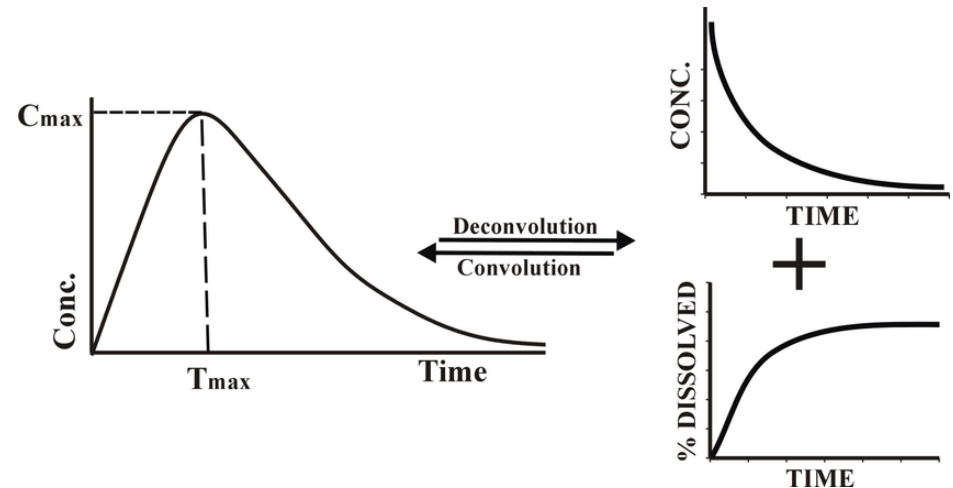


Fig. 9.1 A typical plasma concentration-time profile showing pharmacokinetic and pharmacodynamic parameters, obtained after oral administration of single dose of a drug.

<https://image.slidesharecdn.com/pharmacokineticmodels-140930004231-phapp01/95/pharmacokinetic-models-8-638.jpg?cb=1412037860>

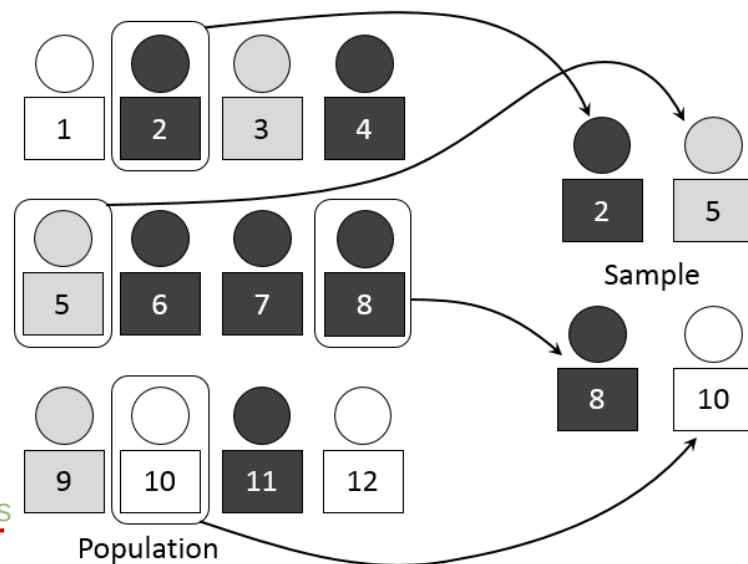




# Pobieranie próby

# Pobieranie próby

- W przypadku pomiarów eksperymentalnych najczęściej nie znamy rozkładu prawdopodobieństwa opisującego dany pomiar (np. parametru rozkładu Poissona w rozpadach promieniotwórczych, czy parametrów rozkładu Gaussa opisującego jakąś populację)
- Te parametry chcemy wyznaczyć doświadczalnie, nie jesteśmy jednak w stanie zebrać nieskończenie wiele pomiarów
- W konsekwencji jesteśmy zmuszeni **przybliżać rozkład gęstości za pomocą rozkładu częstości** (histogramu o skończonej liczbie wejść)
- Próbą** (*ang. sample*) nazywamy zespół doświadczeń wykonywanych w celu określenia kształtu (parametrów) poszukiwanego rozkładu:
  - próba otrzymywana jest poprzez wybór elementów z (często nieskończonego) zbioru wszystkich możliwych doświadczeń (wszystkich możliwych pomiarów), zwanego **populacją generalną**
  - próbę o  $n$  składnikach nazywamy próbą  $n$ -wymiarową



[https://en.wikipedia.org/wiki/Sampling\\_%28statistics%29#/media/File:Simple\\_random\\_sampling.PNG](https://en.wikipedia.org/wiki/Sampling_%28statistics%29#/media/File:Simple_random_sampling.PNG)

# Pobieranie próby

- Cała “sztuka” polega na odpowiednim wybraniu próby z populacji, by aproksymacja rozkładu gęstości była jemu jak najwierniejsza
- Załóżmy, że rozkład zmiennej losowej  $X$  opisywany jest funkcją  $f(x)$  – interesują nas wartości zmiennej  $X$  uzyskane przez poszczególne elementy próby
- Pobieramy  $l$  prób, każda o wymiarze  $n$ , i zaobserwowaliśmy następujące wartości zmiennej  $X$ :

1. próba:  $X_1^{(1)}, X_2^{(1)}, \dots, X_n^{(1)}$

$\vdots$

$j$ -ta próba:  $X_1^{(j)}, X_2^{(j)}, \dots, X_n^{(j)}$

$\vdots$

$l$ -ta próba:  $X_1^{(l)}, X_2^{(l)}, \dots, X_n^{(l)}$

Oczywiście, to co pobraliśmy jest też zmienną losową!

**Elementy próby losowej (wartości zmiennej  $X$ ), są zmienną losową**

- Każdą próbę możemy przedstawić jako wektor ( $n$ -wymiarową zmienną losową):  $\mathbf{X}^{(j)} = (X_1^{(j)}, X_2^{(j)}, \dots, X_n^{(j)})$
- Wektor ma rozkład gęstości prawdopodobieństwa:

$$g(\mathbf{x}) = g(x_1, x_2, \dots, x_n)$$

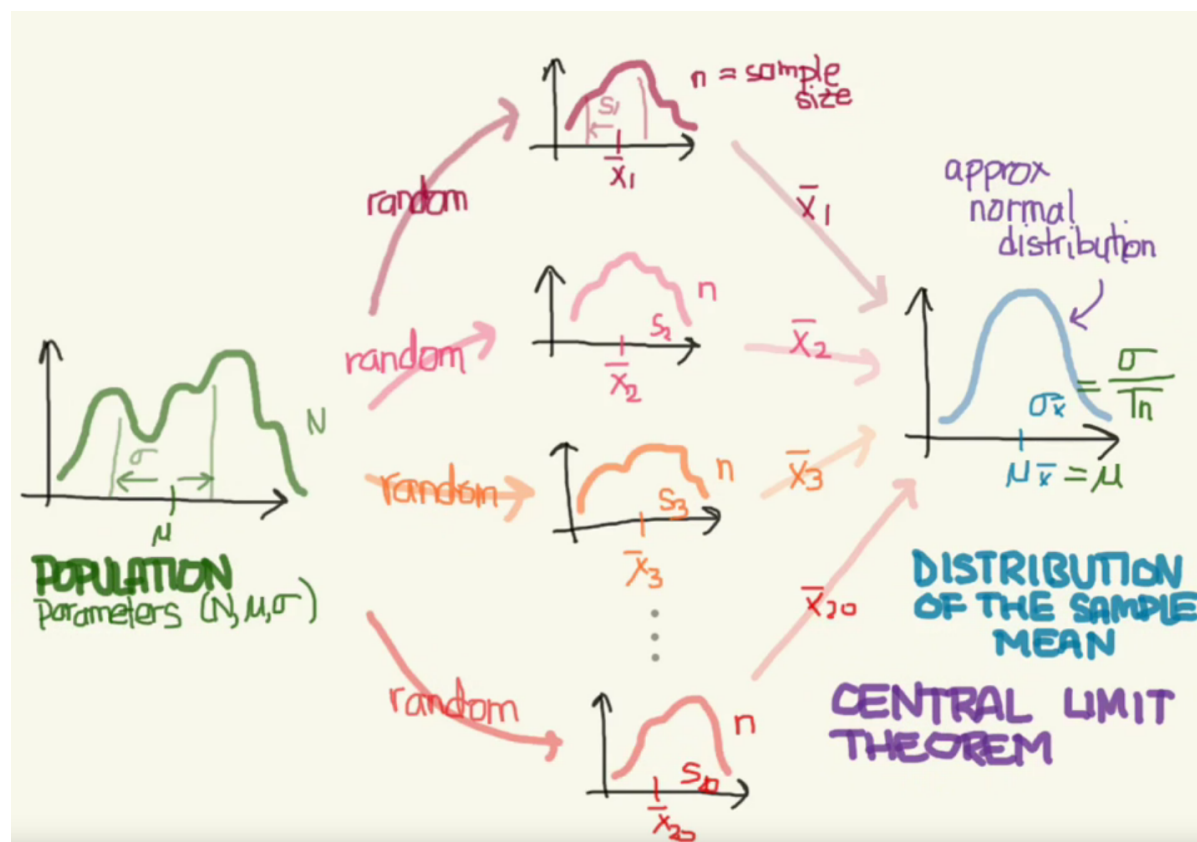
# Pobieranie próby

- Aby można było mówić o losowym pobieraniu próby:
  - **zmienne  $X_i$  muszą być niezależne, czyli:**  $g(\mathbf{x}) = g_1(x_1)g_1(x_2)\dots g_n(x_n)$
  - **poszczególne rozkłady muszą być jednakowe i identyczne z rozkładem gęstości populacji:**  $g_1(x_1) = g_2(x_2) = \dots = g_n(x_n) = f(x)$
- Należy podkreślić, że w rzeczywistym procesie pobierania próby często bardzo trudno jest zapewnić pełną losowość – nie ma tutaj jednej recepty jak to zrobić (należy starać się spełnić powyższe warunki)
  - przykładowo: prowadząc badania kliniczne leków powinniśmy zapewnić w każdym ośrodku próbę losową i kontrolną pacjentów, która jest “taka sama” jak w innych ośrodkach, co bardzo często nie jest możliwe praktycznie

# Pobieranie próby

- Teraz zdefiniujemy pojęcia, które charakteryzują próbę losową:
  - funkcję elementów próby losowej nazywamy **statystyką**
  - najważniejszym przykładem statystyki jest **średnia z próby** (*ang. sample mean*) zdefiniowana jako średnia z elementów próby:

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$$



# CTG - powtórzenie

- Przykład:

- wyobraźmy sobie, że szacujemy wzrost w **populacji** ośmioletnich dzieci w Polsce. Rozkład populacji ma parametry:  $\mu, \sigma$
- wybieramy losowo 100 8-latków i liczymy średnią wartość z próby losowej –  $\bar{X}_1$
- nasz kolega wykonuje analogiczne doświadczenie → dostaje inny wynik –  $\bar{X}_2$
- zaczynamy więc pracować razem, znowu wybieramy 100 8-latków i dostajemy trzeci wynik –  $\bar{X}_3$
- ale przecież jest tylko **jeden prawdziwy** średni wzrost 8-latek w całej populacji!
- ponieważ **średnia z próby jest również zmienną losową**, możemy wykonać wielokrotnie próbę losową i dostać wiele średnich → **otrzymujemy rozkład wartości średniej z próby**
- jeśli mamy dużo prób losowych → **rozkład wartości średniej z prób dąży do rozkładu normalnego (CTG):**  $N(\mu, \sigma/\sqrt{n})$



# Pobieranie próby - przykład

- Przykład – wzrost Polaków
- Niewątpliwie, wzrost Polaków (zmienna losowa  $X$ ) podlega pewnemu rozkładowi  $f(x)$  z dystrybuantą  $F(x)$ 
  - Pomiar wzrostu pojedynczego Polaka daje wartość  $x_1$
  - Losowy wybór tego jednego polaka to zmienna losowa  $X_1$
- Jeżeli stworzymy  $n$ -wymiarową próbę losową, tzn. wybierzemy  $n$  Polaków, to rozkład prawdopodobieństwa wyboru dla każdej z osób (od  $g_1(x_1)$  do  $g_n(x_n)$ ) jest taki sam jak dla całej populacji i równy  $f(x)$
- **Zadaniem estymacji** jest znalezienie takiej statystyki (a więc funkcji określonej na wektorze  $X=(X_1, \dots, X_n)$ ), aby najlepiej przybliżała ona rzeczywistą wartość parametru opisującego rzeczywisty rozkład zmiennej losowej  $X$

# Estymatory

- Typowy problem analizy danych: znamy (np. z prawa fizycznego) ogólną postać gęstości prawdopodobieństwa w danej populacji, należy “jedynie” wyznaczyć parametry tego rozkładu. Przykład:
  - **mierzymy rozpad radioaktywny w czasie:**  $N(t) = N_0(1 - \exp(-\lambda t))$
  - parametr  $\lambda$  wyznaczamy na podstawie próby – mierząc skończoną ilość razy ilość rozpadów w czasie → wynik nigdy nie będzie dokładny, bo próba jest skończona, mamy problem **estymacji parametrów**
  - poszukiwana wielkość uzyskiwana jest funkcją elementów próby (**statystyką**) i jest nazywana **estymatorem**:  $S = S(X_1, X_2, \dots, X_n)$
  - estymator jest **nieobciążony**, jeżeli niezależnie od liczebności próby jego wartość oczekiwana jest równa wartości estymowanego parametru:

$$E(S(X_1, X_2, \dots, X_n)) = \lambda, \text{ dla każdego } n$$

- estymator jest **zgodny**, jeżeli jego wariancja znika:

$$\lim_{n \rightarrow \infty} \sigma(S(X_1, X_2, \dots, X_n)) = 0$$

# Estymatory – wartość oczekiwana

- **Wartość średnia** ze wszystkich elementów próby jest zmienną losową (jest funkcją zmiennych losowych). Jej wartość oczekiwana (tej średniej):

$$E(\bar{X}) = \frac{1}{n} (E(X_1) + E(X_2) + \dots + E(X_n)) = E(X) = \hat{x}, \text{ dla każdego } n$$

- Wniosek: **wartość średnia (arytmetyczna) z próby to estymator nieobciążony wartości oczekiwanej** zmiennej  $X$  w populacji
- Możemy obliczyć wariancję wartości średniej:

$$\begin{aligned} \sigma^2(\bar{X}) &= E\{\bar{X} - E(\bar{X})\}^2 = E\left\{\left(\frac{X_1 + X_2 + \dots + X_n}{n} - \hat{x}\right)^2\right\} \\ &= \frac{1}{n^2} E\{[(X_1 - \hat{x}) + (X_2 - \hat{x}) + \dots + (X_n - \hat{x})]^2\} \end{aligned}$$

- Z uwagi na niezależność zmiennych kowariancje między zmiennymi  $X_i$  znikają, czyli ostatecznie:

$$\sigma^2(\bar{X}) = \frac{1}{n} \sigma^2(X) \quad \lim_{n \rightarrow \infty} \sigma^2(\bar{X}) = 0$$

- Wniosek: **wartość średnia (arytmetyczna) z próby jest również estymatorem zgodnym wartości oczekiwanej**

# Estymatory - wariancja

- Jak pamiętamy z definicji wariancji, nie jest ona zmienną losową
- Możemy wariancję przybliżyć przez średnią arytmetyczną odchyłeń kwadratowych od wartości średniej:

$$S'^2(X) = \frac{1}{n} \left( (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right)$$

- Wartość oczekiwana tej wielkości:

$$\begin{aligned} E(S'^2(X)) &= \frac{1}{n} E \left\{ \sum_{i=1}^n (X_i - \bar{X})^2 \right\} = \frac{1}{n} E \left\{ \sum_{i=1}^n (X_i - \hat{x} + \hat{x} - \bar{X})^2 \right\} \\ &= \frac{1}{n} E \left\{ \sum_{i=1}^n (X_i - \hat{x})^2 + \sum_{i=1}^n (\hat{x} - \bar{X})^2 + 2 \sum_{i=1}^n (X_i - \hat{x})(\hat{x} - \bar{X}) \right\} \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ E((X_i - \hat{x})^2) - E((\bar{X} - \hat{x})^2) \right\} = \frac{1}{n} \left\{ n \sigma^2(X) - n \left( \frac{1}{n} \sigma^2(X) \right) \right\} \\ &= \frac{n-1}{n} \sigma^2(X) \end{aligned}$$

- Widać więc, że  $S'^2$  jest **estymatorem obciążonym** dla wariancji populacji mającym wartość oczekiwaną mniejszą niż  $\sigma^2(X)$

# Estymatory - wariancja

- Możemy jednak nieznacznie zmodyfikować definicję wariancji z próby i wprowadzić estymator:

$$S^2(X) = \frac{1}{n-1} \left( (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \right)$$

- Otrzymujemy **estymator nieobciążony wariancji populacji**
- Jeśli wykorzystamy znany z CTG wzór:  $\sigma^2(\bar{X}) = \frac{1}{n} \sigma^2(X)$
- To otrzymamy **estymator wariancji wartości średniej z próby**:

$$S^2(\bar{X}) = \frac{1}{n} S^2(X) = \frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Zaś odpowiadające odchylenie standardowe (**niepewność średniej z próby**):

$$\Delta \bar{X} = \sqrt{S^2(\bar{X})} = S(\bar{X}) = \frac{1}{\sqrt{n}} S(X) \qquad S = \sqrt{S^2} = \frac{1}{\sqrt{n-1}} \sqrt{\sum_{i=1}^n (X_i - \hat{X})^2}$$

# Estymatory - wariancja

- **Estymator wariancji wartości średniej z próby:**

$$S^2(\bar{X}) = \frac{1}{n} S^2(X) = \frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Zaś odpowiadające odchylenie standardowe (**estymator niepewność średniej z próby**):

$$\Delta \bar{X} = \sqrt{S^2(\bar{X})} = S(\bar{X}) = \frac{1}{\sqrt{n}} S(X)$$

- Jaka jest zaś **niepewność wariancji z próby** (bez wyprowadzenia)?
  - czyli: **estymator wariancji estymatora wariancji wartości średniej z próby?**

$$\Delta S^2 = S^2 \sqrt{\frac{2}{n-1}}$$

- I tak dalej możemy tworzyć kolejne poziomy estymatorów...



# Estymatory - wariancja

- Podsumowując zatem **estymatory nieobciążone**:

- wartości oczekiwanej populacji → średnia z próby (**wynik doświadczenia**):

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$$

- wariancji populacji – wariancja z próby (aproksymowana):

$$S^2(X) = \frac{1}{n-1} ((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2)$$

- wariancji wartości średniej z próby (**patrz niepewność typu A**):

$$S^2(\bar{X}) = \frac{1}{n} S^2(X) = \frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2$$

- wariancji (aproksymowanej) wariancji z próby

$$\text{Var}(S^2) = S^4 \left( \frac{2}{n-1} \right)$$

- dalej możemy wyznaczać np. wariancję wariancji aproksymowanego estymatora wariancji próby i tak dalej (w nieskończoność)...

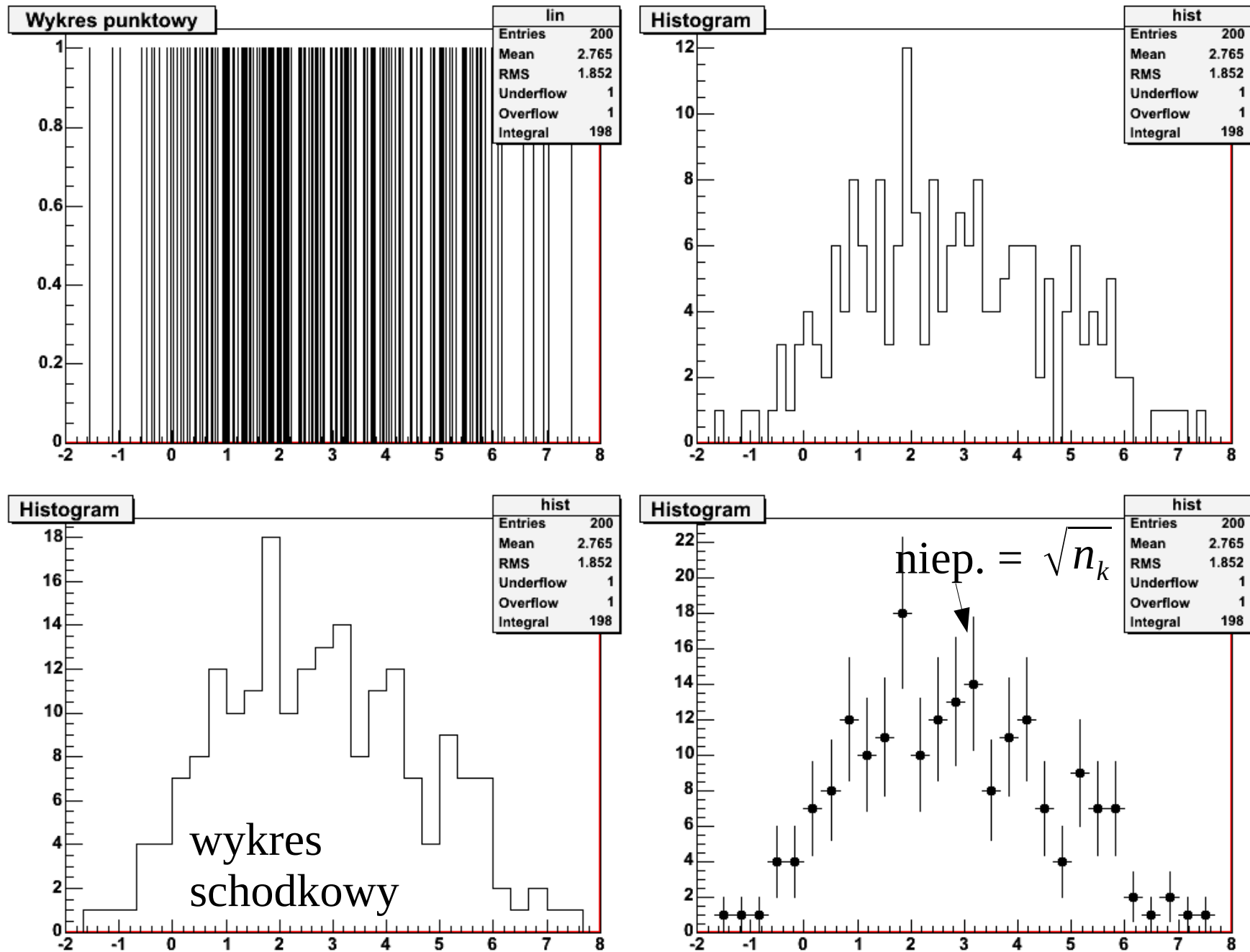


# Graficzne przedstawienie próby

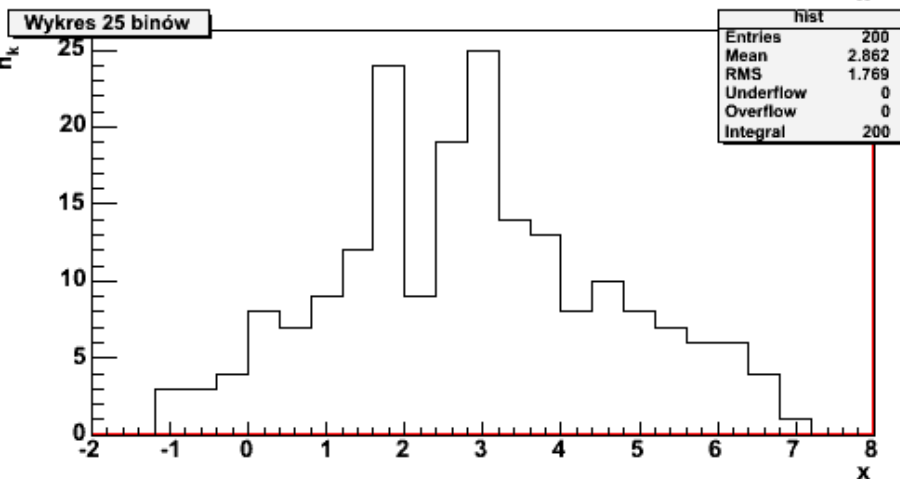
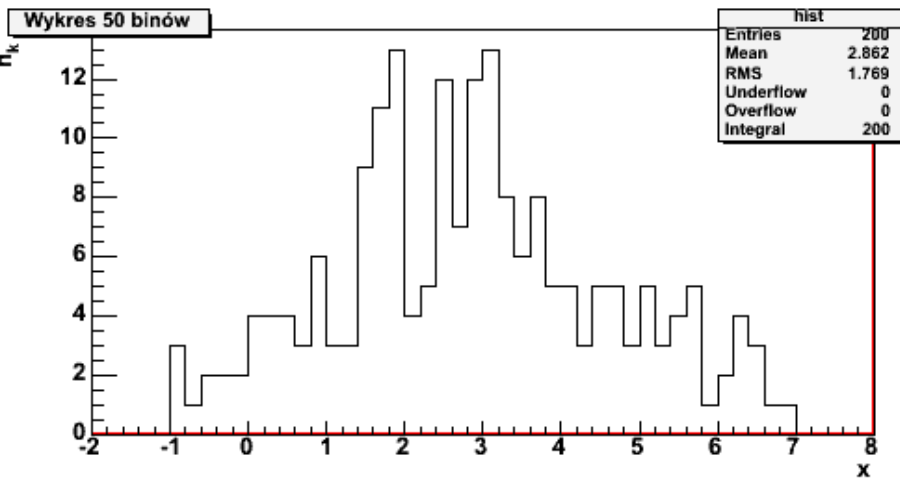
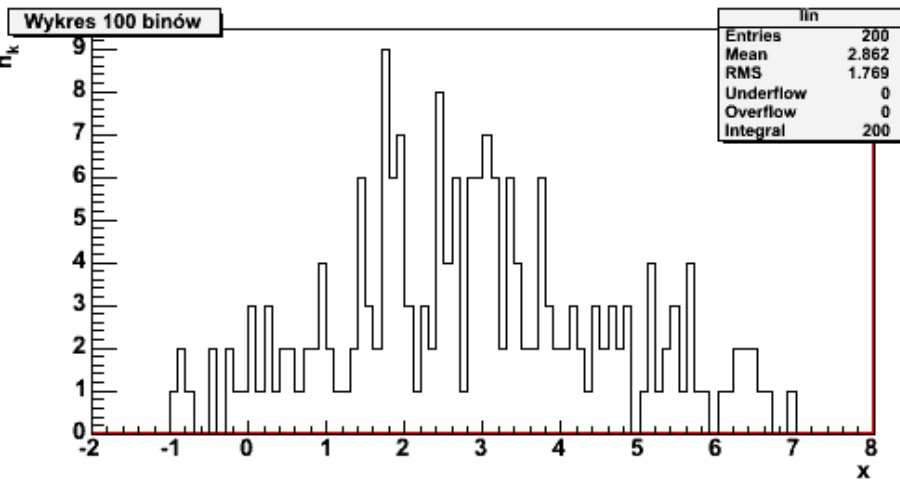
# Graficzne przedstawienie próby

- Rozważmy próbę:  $x_1, x_2, \dots, x_n$ , która zależy od jednej zmiennej losowej  $X$
- Możemy tę próbę przedstawić jako wykres 1D – punkty na osi  $x$  – jednowymiarowy wykres punktowy
  - **wada:** co w przypadku, gdy mamy dwa takie same pomiary?
- Z reguły stosujemy zatem wykres 2D, zwany **histogramem**:
  - dzielimy przedział zmienności  $x$  (lub jego część) na  $r$  **przedziałów** o jednakowej szerokości  $\Delta x$ :  $\xi_1, \xi_2, \dots, \xi_r$
  - środki przedziałów znajdują się w punktach:  $x_1, x_2, \dots, x_r$
  - na osi  $y$  odkładamy liczbę elementów próby przypadającą na dany przedział:  $n_1, n_2, \dots, n_r$
  - tak otrzymany wykres nazywamy **wykresem częstości** lub **histogramem**

# Graficzne przedstawienie próby



# Histogram – szerokość przedziału



- Im więcej przedziałów, tym informacja o próbie jest dokładniejsza
- Większa ilość przedziałów powoduje jednak większe wahania statystyczne *od punktu do punktu*
- Pole pod krzywą schodkową jest proporcjonalne do wielkości próby (jeśli je przeskalujemy przez  $1/n$ , otrzymamy częstość)

# Graficzne przedstawienie próby - przykład

- Badamy “nieznany” rozkład prawdopodobieństwa
- Symulujemy taką sytuację poprzez generację 1000 prób z rozkładu Gaussa o wartości średniej 0 i wariancji 1. Każda próba ma licznosc (liczbę składników)  $r$ .
- Badamy zachowanie estymatorów charakterystyk rozkładu i estymatorów ich niepewności w funkcji liczności próby  $r$

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$$

estymator wartości oczekiwanej populacji  
średnia z próby

$$S(X) = \sqrt{S^2(X)} = \frac{1}{\sqrt{n-1}} \sqrt{\sum (X_i - \bar{X})^2}$$

estymator **odch. std.** populacji

$$S^2(X) = \frac{1}{n-1} \{ (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \}$$

estymator **wariancji** populacji

$$\sigma(\bar{X}) = \Delta \bar{X} = \sqrt{(S^2(\bar{X}))} = S(\bar{X}) = \frac{1}{\sqrt{n}} S(X)$$

niepewność wart. średniej - estymator odch. st. wartości średniej z próby (estymatora wart. oczekiwanej)

$$\sigma(S(X)) = \Delta S(X) = \frac{S(X)}{\sqrt{2(n-1)}}$$

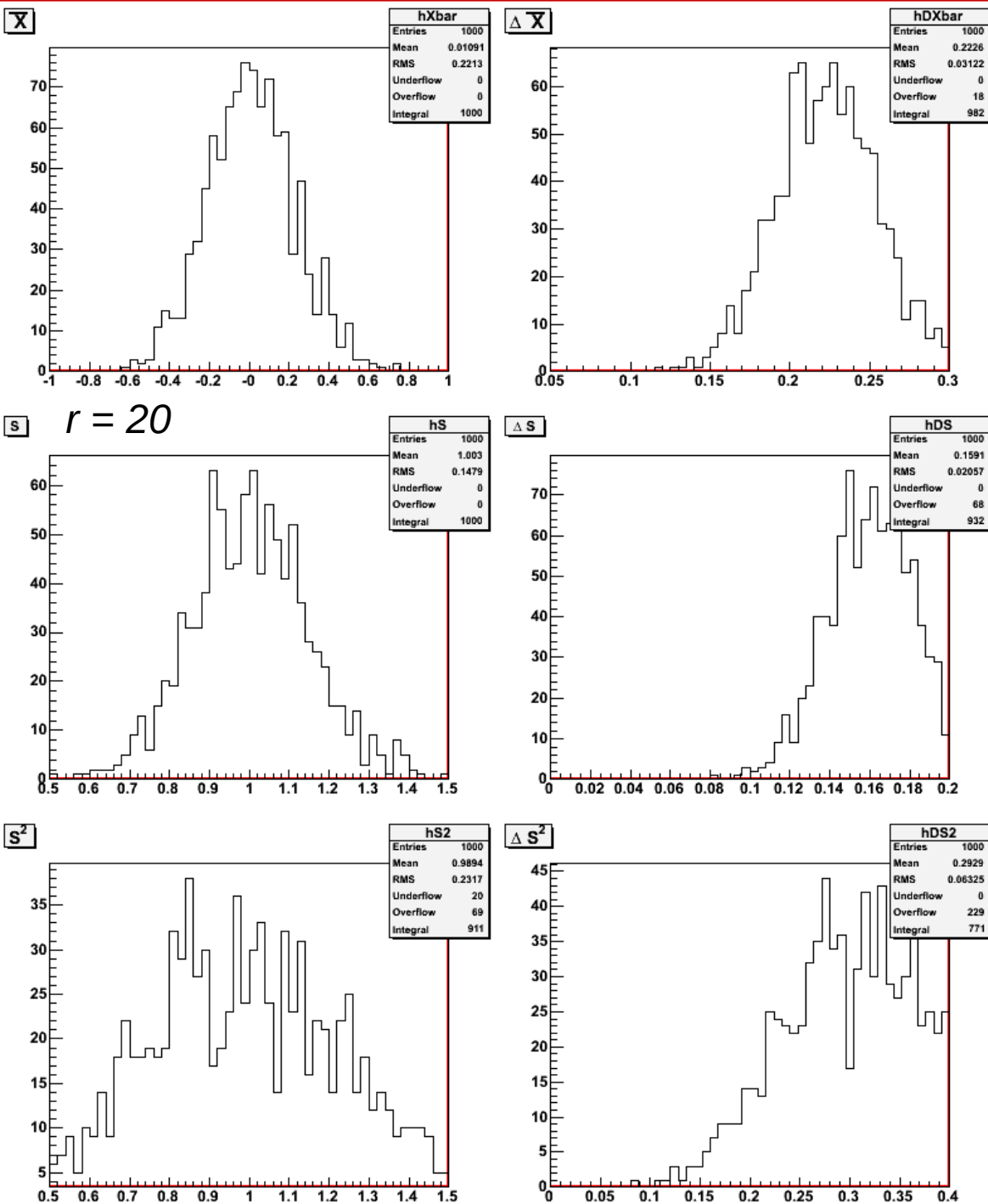
niepewność estymatora odch. std. populacji – estymator odch. std. estymatora **odch. std.** populacji

$$\sigma(S^2(X)) = \Delta S^2(X) = S^2(X) \sqrt{\frac{2}{n-1}}$$

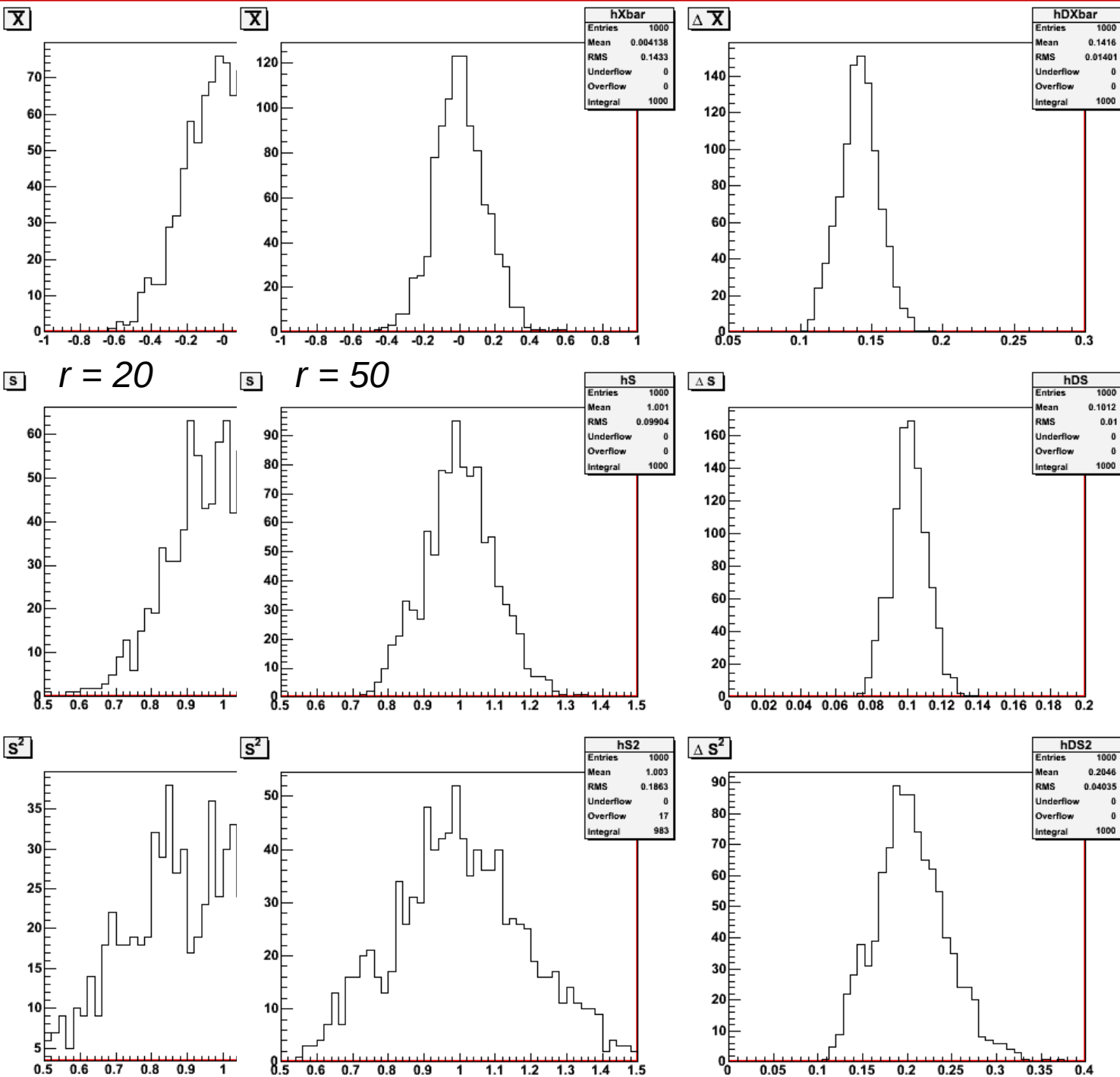
niepewność estymatora wariancji populacji – estymator odch. std. estymatora **wariancji** populacji



# Estymatory - histogramy



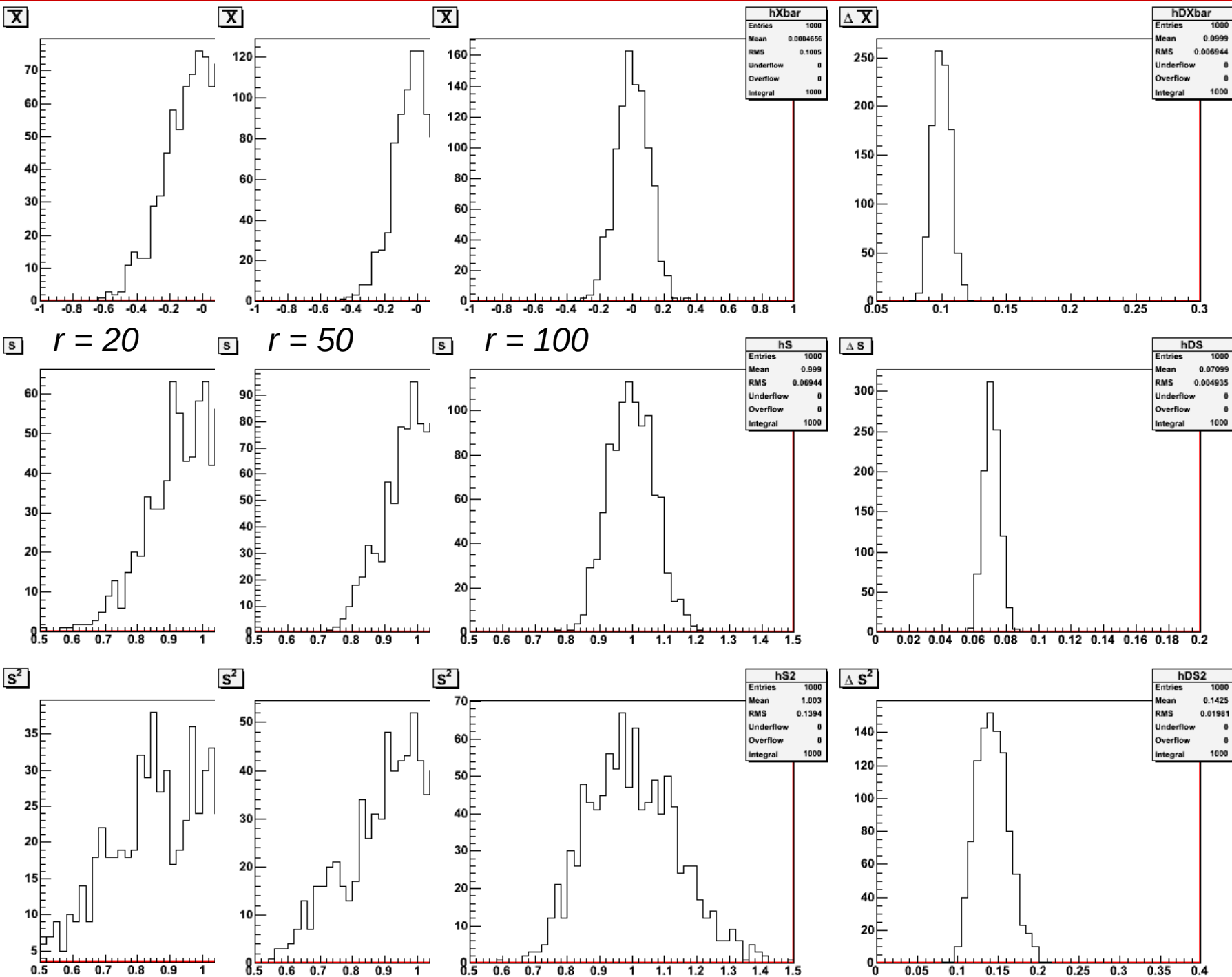
# Estymatory - histogramy



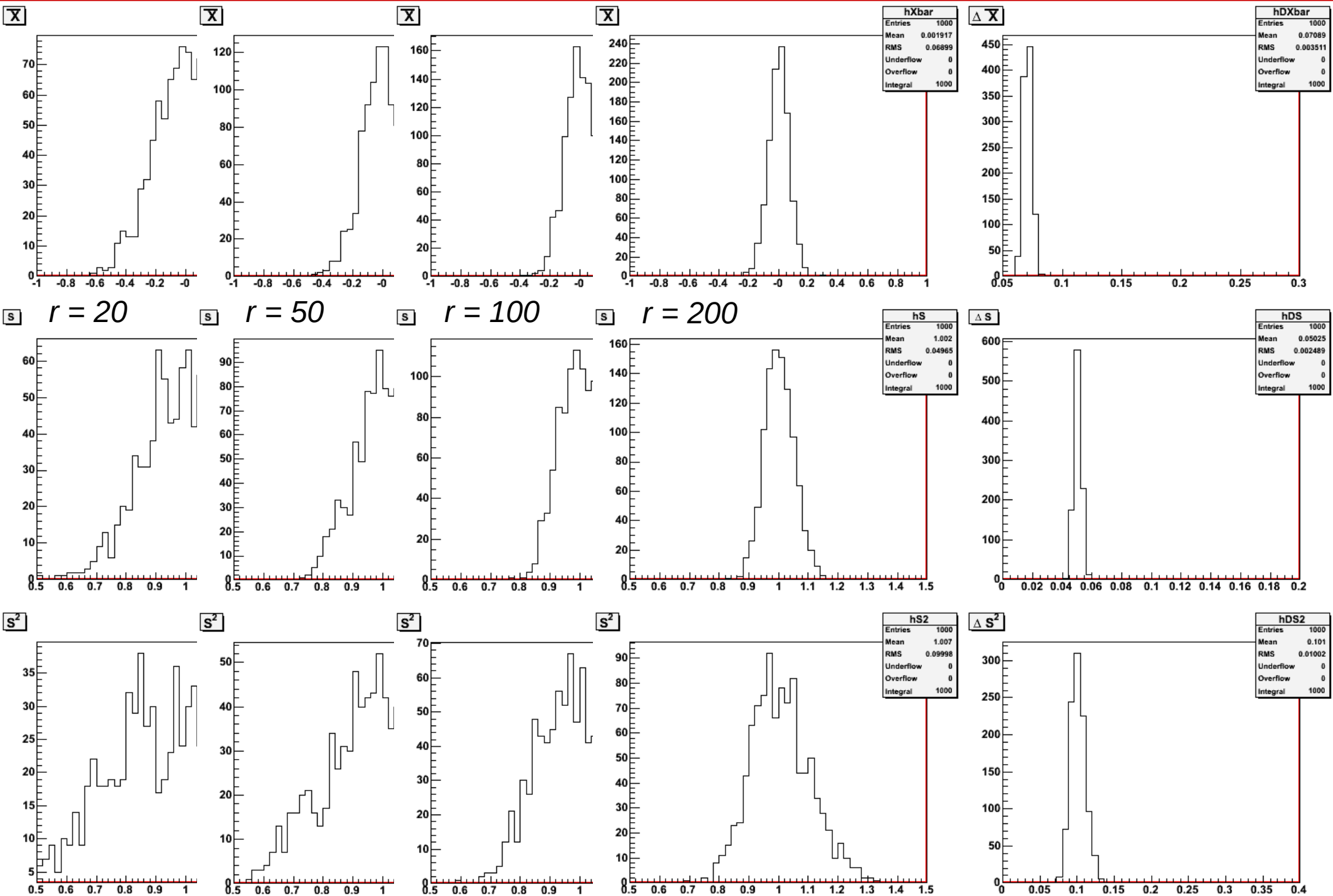


# Wielowymiarowy rozkład Gaussa

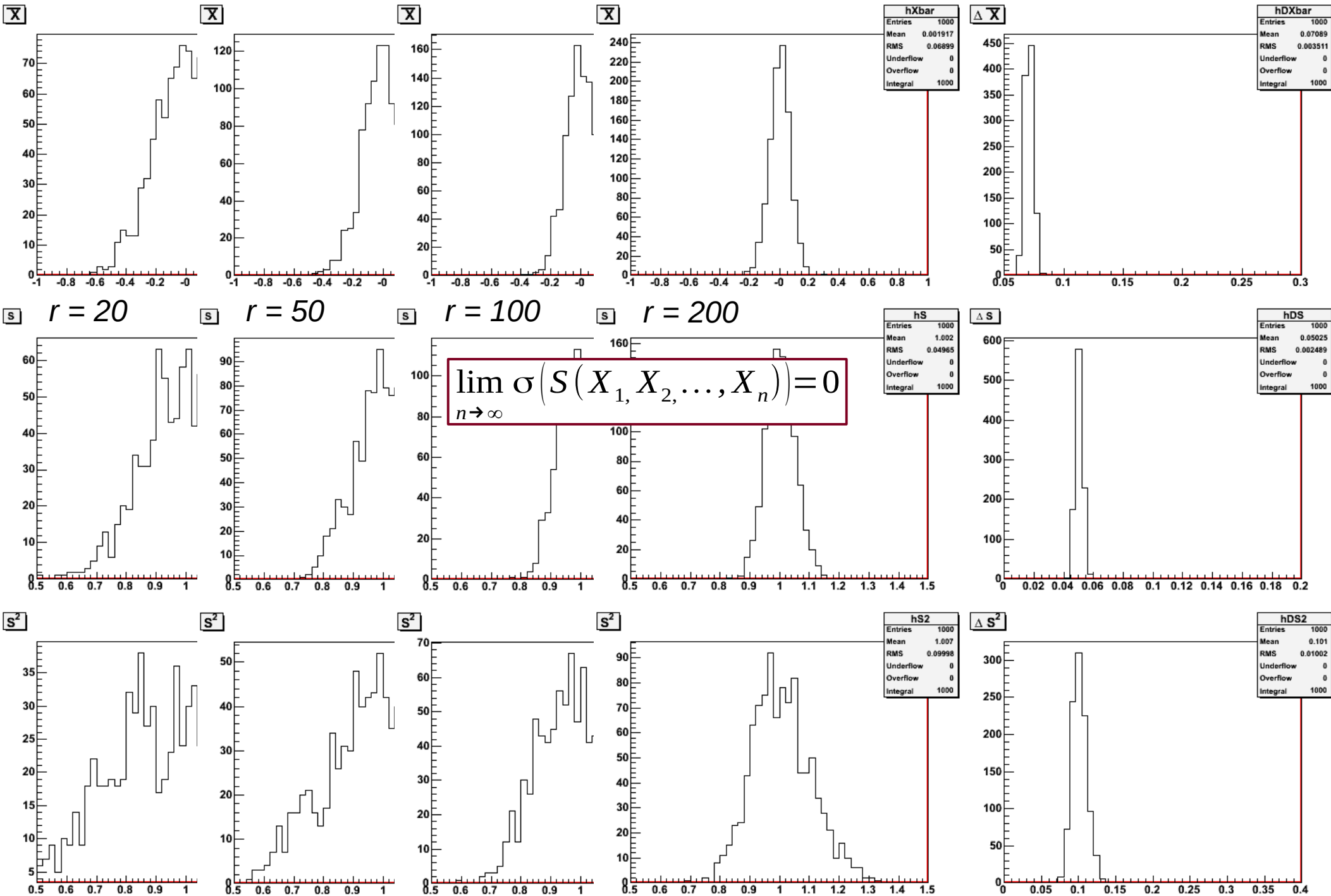
# Estymatory - histogramy



# Estymatory - histogramy



# Estymatory - histogramy







# KONIEC