

# Komputerowa analiza danych doświadczalnych

Wykład 7  
8.04.2022

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

*Semestr letni 2021/2022*



# Centralne twierdzenie graniczne - przypomnienie

## Sploty

Pobieranie próby,  
estymatory



# 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

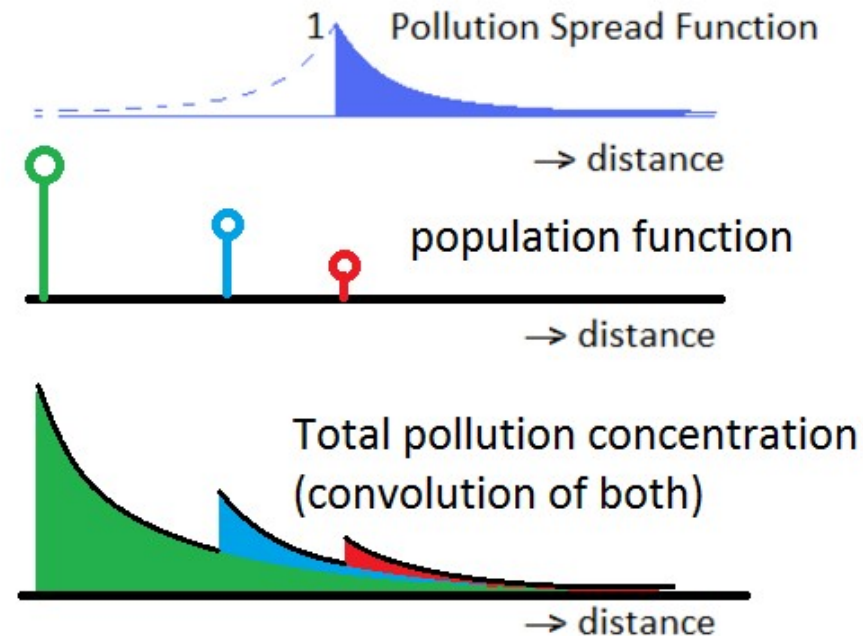


# Sploty

# 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:
  - Mieszkaś 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

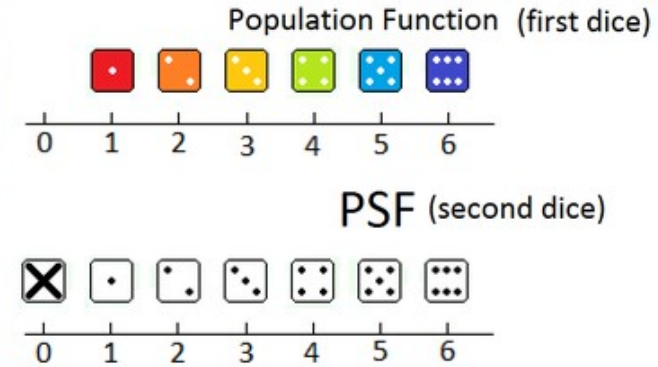


- Jaka jest pełna funkcja opisująca poziom zanieczyszczeń w rzece?
- Jest to **splot** dwóch rozkładów – funkcji populacji oraz funkcji koncentracji odpadów
- Innymi słowy, zastępujemy każdy punkt w funkcji populacji przez funkcję koncentracji przeskalowaną przez wagę funkcji populacji

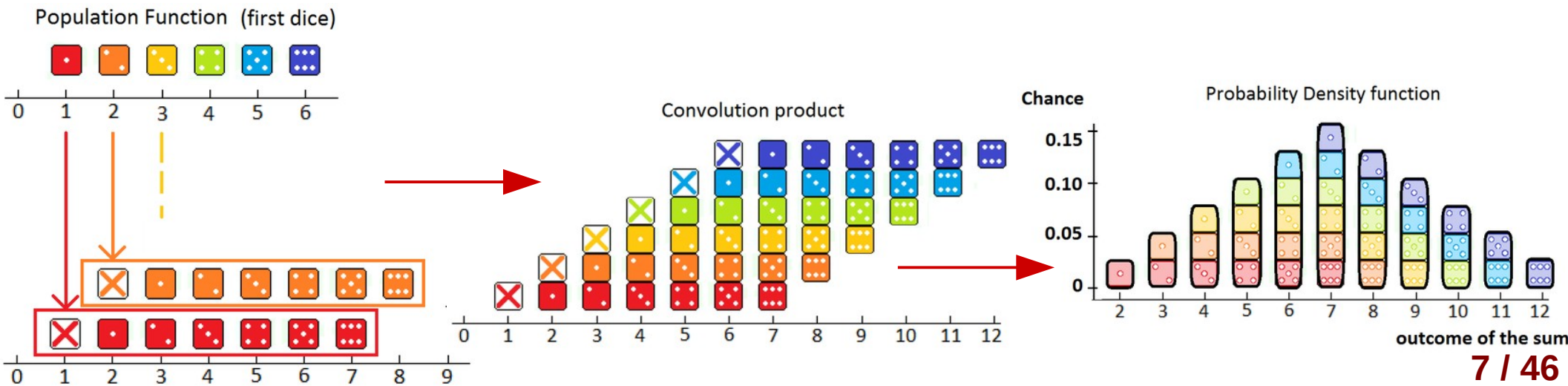
# Suma zmiennych losowych jako splot

<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.



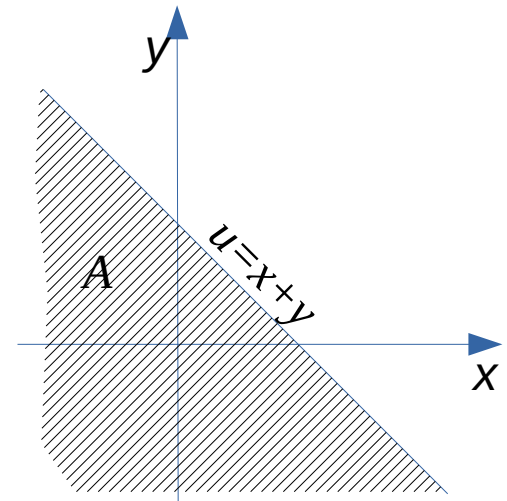
- Drugi rzut kostką oznacza funkcję PSF – jak bardzo dana miejscowość zanieczyszcza rzekę, i znowu 16,7% zanieczyszczeń ląduje 1 km dalej od miasta, 16.7% 2 km dalej od miasta, itp.
- Jak policzyć pełną funkcję zanieczyszczeń? Podmieniamy funkcję populacji poprzez funkcję zanieczyszczeń, dla każdego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

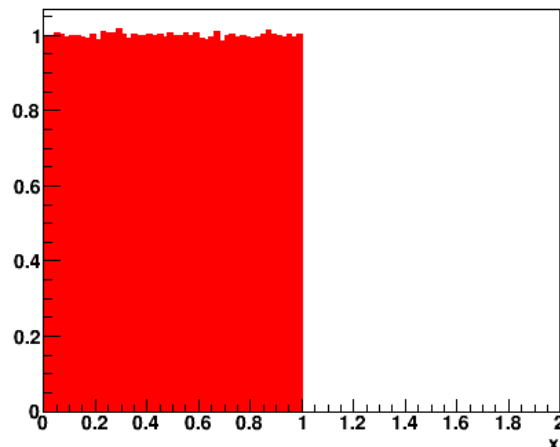
$$f(u) = \frac{dF(u)}{du} = \int_{-\infty}^{\infty} f_x(x) f_y(u-x) dx = \int_{-\infty}^{\infty} f_y(y) f_x(u-y) dy \equiv (f_x * f_y)(u)$$

- 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:

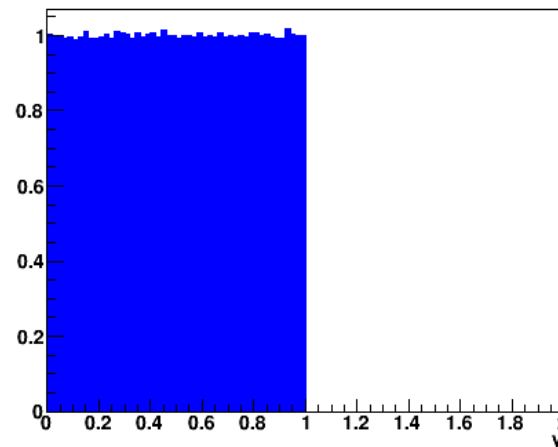
$$f_x(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

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

Rozkład jednostajny



Rozkład jednostajny



# Suma zmiennych losowych jako splot

- Splot dwóch rozkładów jednorodnych:

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

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

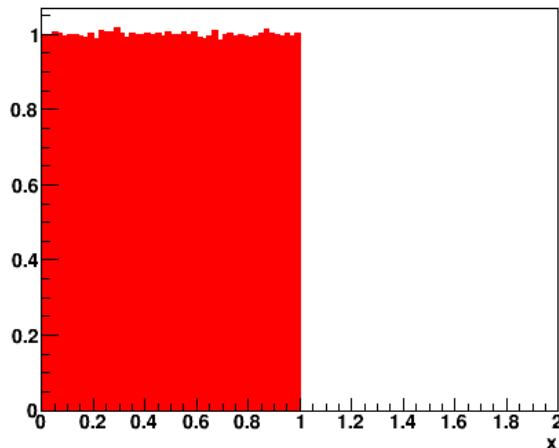
$$f(u) = \int_0^1 f_x(x) f_y(u-x) dx = \int_0^1 1 \cdot f_y(u-x) dx \stackrel{\substack{v=u-x \\ dv=-dx}}{\Rightarrow} f(u) = - \int_u^{u-1} f_y(v) dv = \int_{u-1}^u f_y(v) dv$$

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

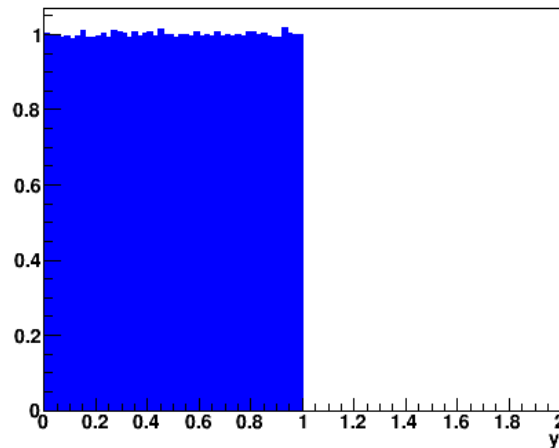
$$(a) \quad 0 \leq u < 1: f_1(u) = \int_0^u f_y(v) dv = \int_0^u 1 dv = u$$

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

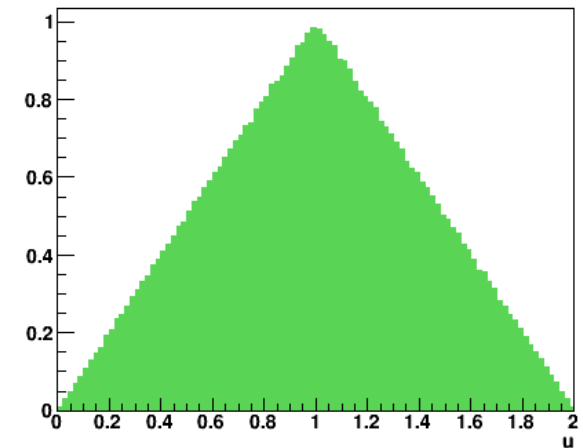
Rozkład jednostajny



Rozkład jednostajny



Splot 2 rozkładów jednostajnych



# Suma zmiennych losowych jako splot

- Rozpatrzmy przypadek splotu dwóch rozkładów jednorodnych:

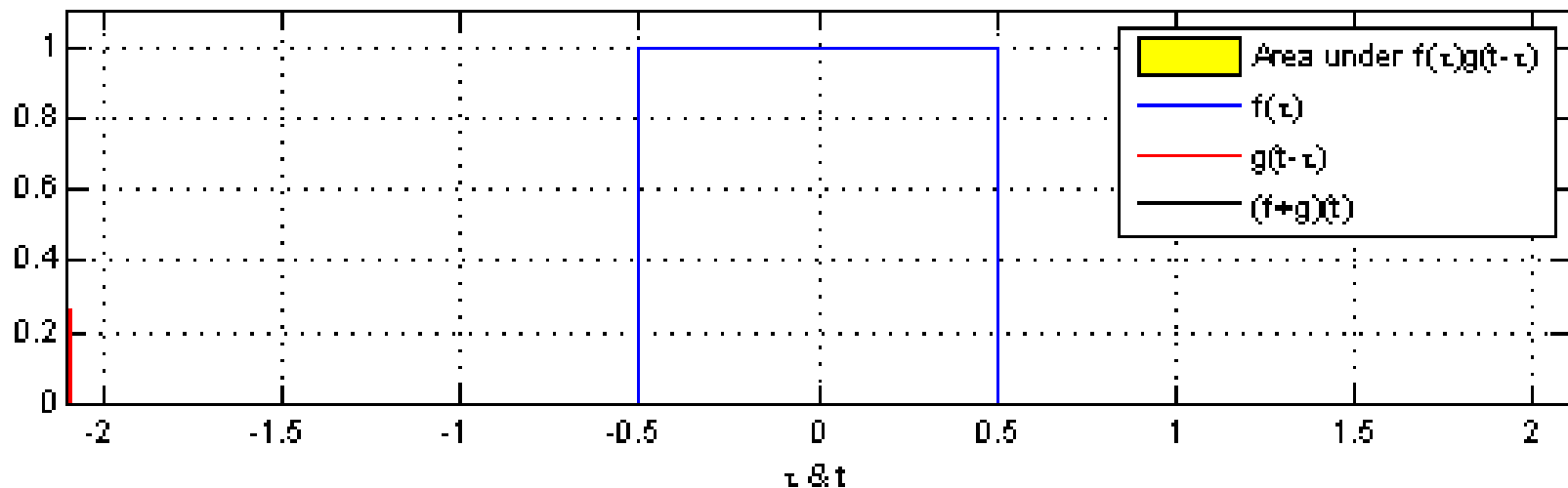
$$f_x(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{w przeciwnym razie} \end{cases} \quad f_y(y) = \begin{cases} 1, & 0 \leq y < 1 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

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

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

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

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



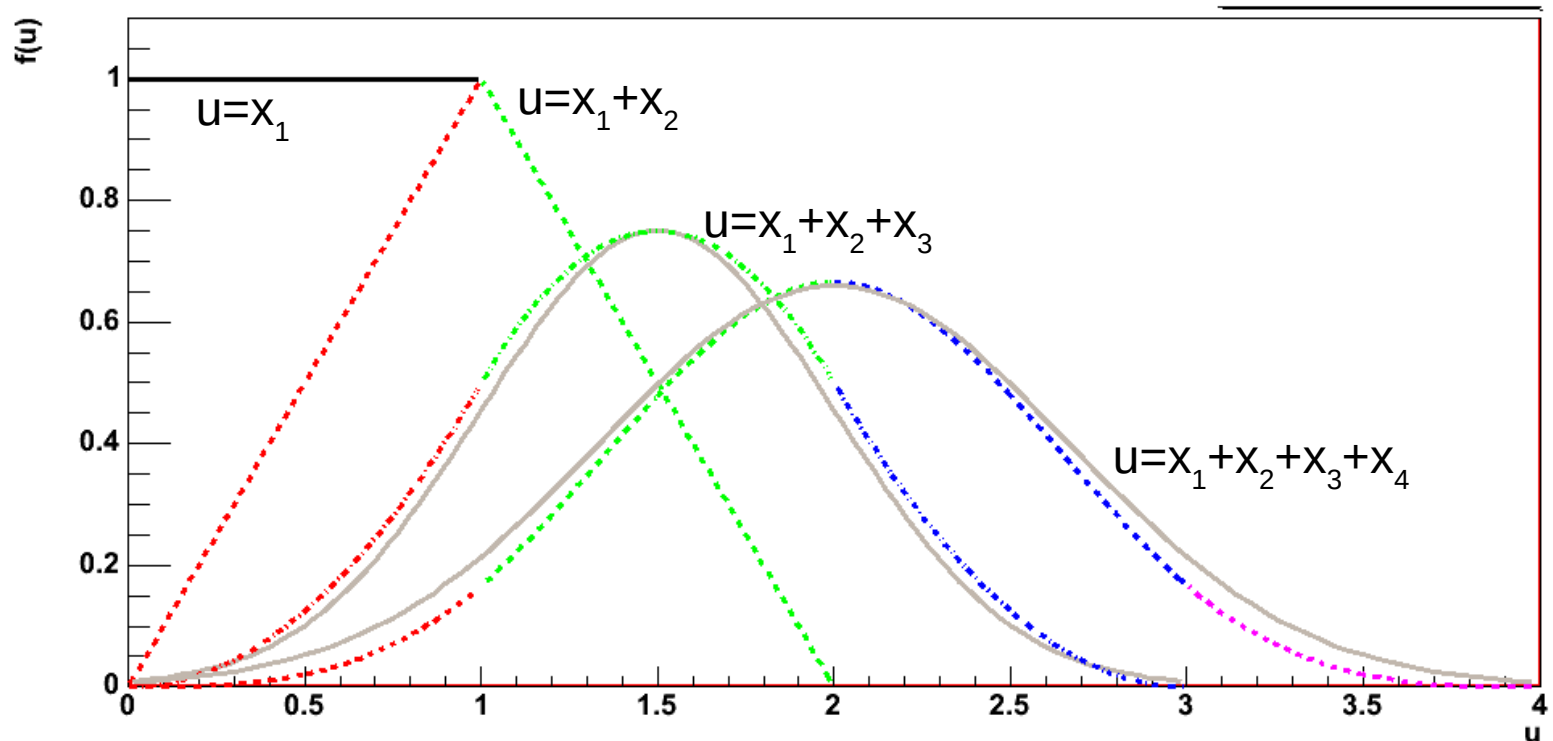
[https://en.wikipedia.org/wiki/Convolution#/media/File:Convolution\\_of\\_box\\_signal\\_with\\_itself2.gif](https://en.wikipedia.org/wiki/Convolution#/media/File:Convolution_of_box_signal_with_itself2.gif)

# 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:

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

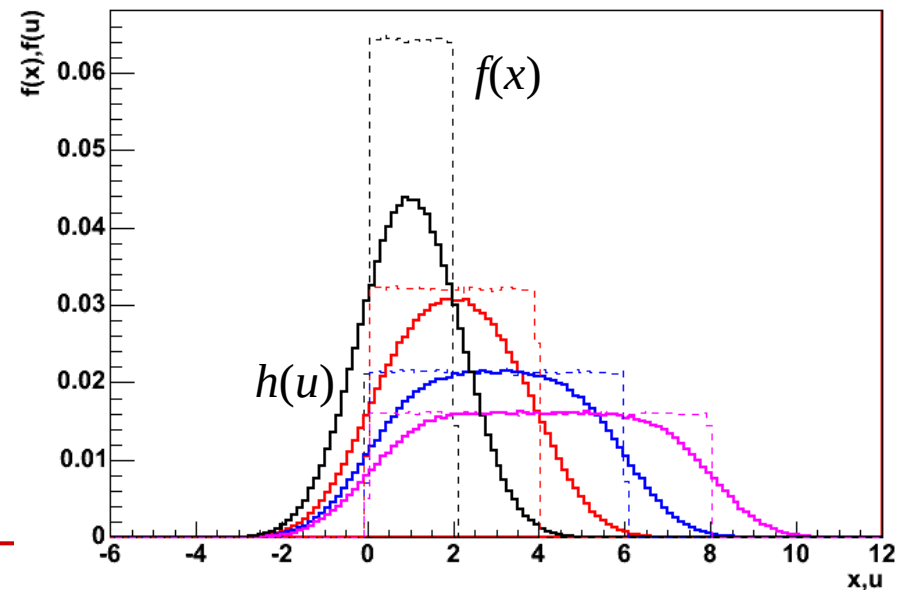
$$f(x) = 0; x \in \mathbb{R} \setminus \langle a, b \rangle$$

- 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:

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

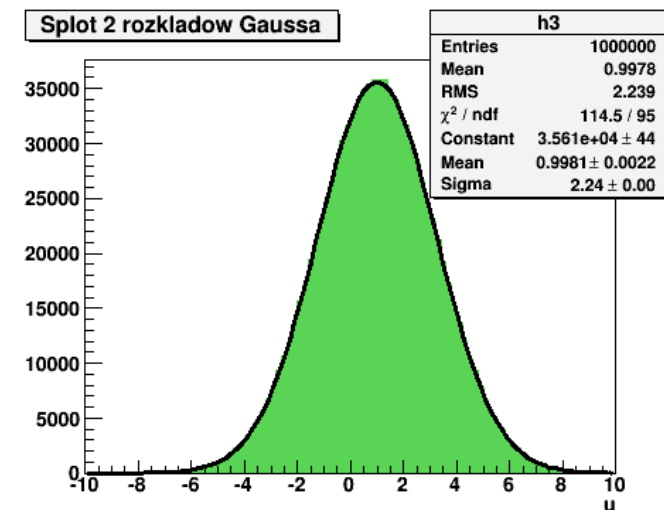
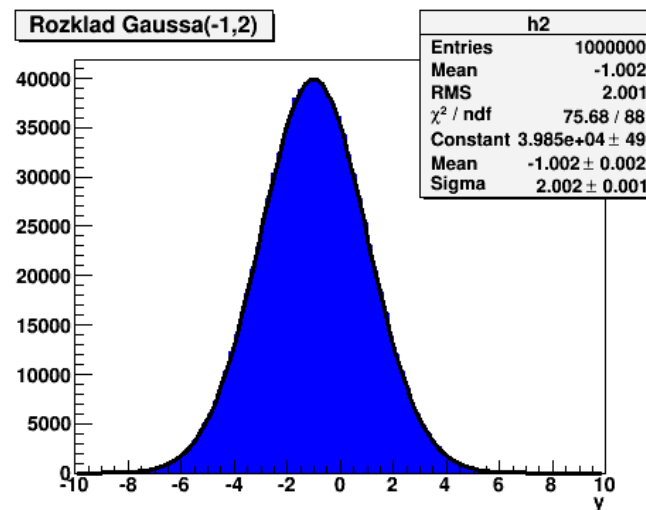
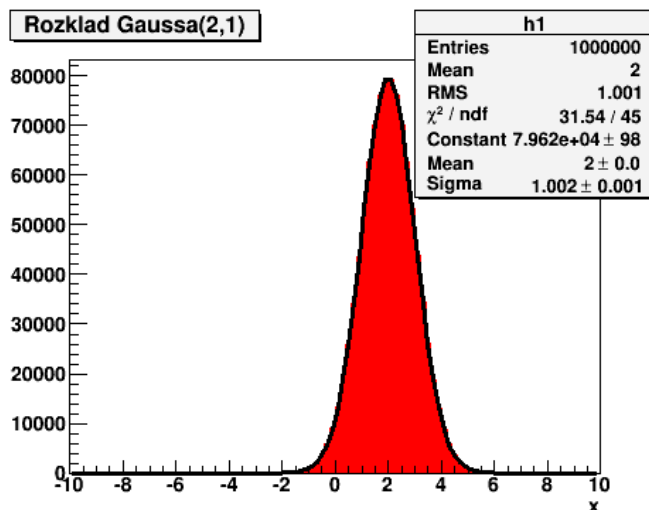


# Sploty z rozkładem normalnym – przykład 2

- 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:

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

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



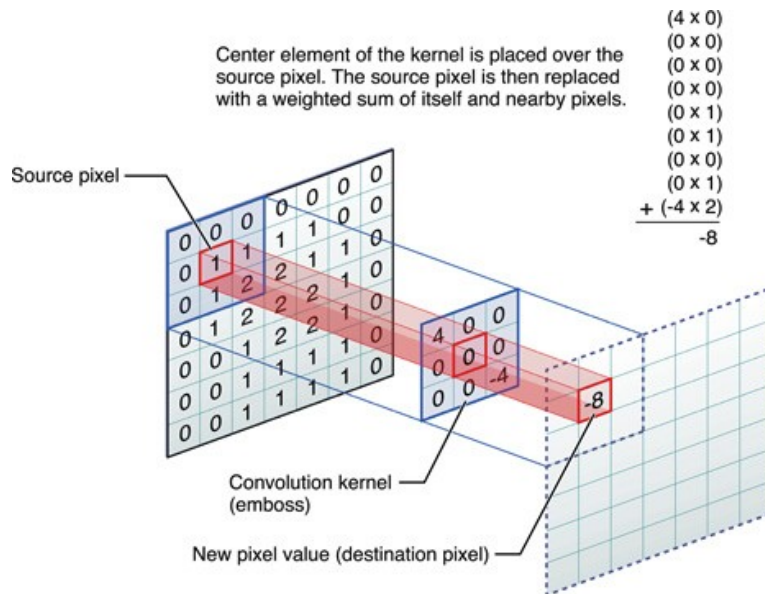
# Zastosowanie splotów

- 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 komput.

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



Original



Emboss



# 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.

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

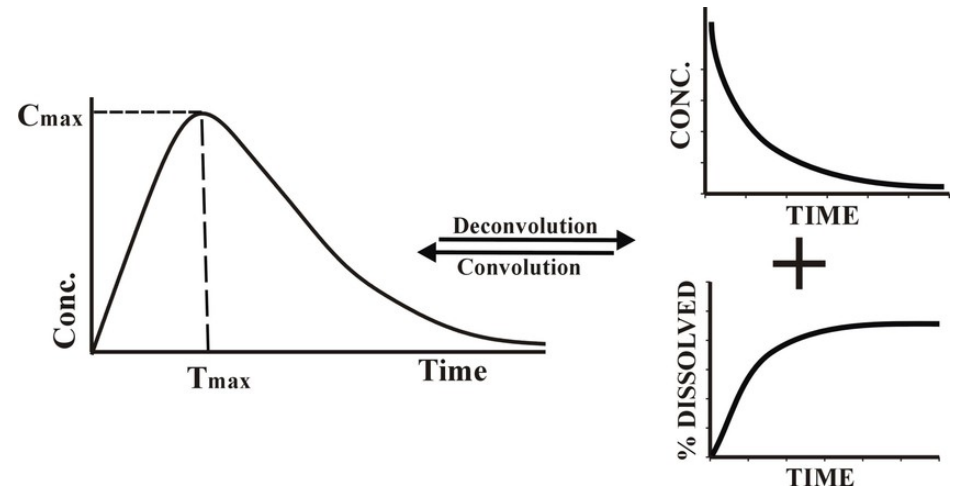
$C_s$  is the concentration time profile resulting from instantaneous absorption of a unit amount of drug which is typically absorbed from bolus IV injection or reference oral solution data

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

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

$u$  is the variable of integration

[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)



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

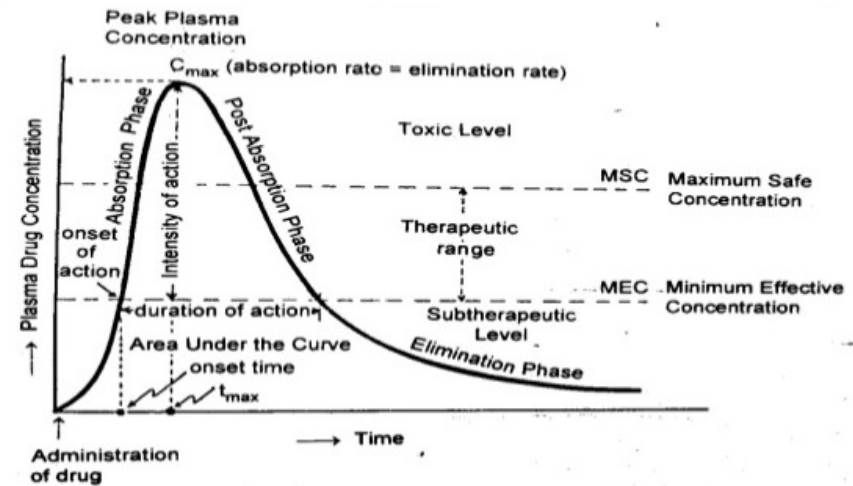
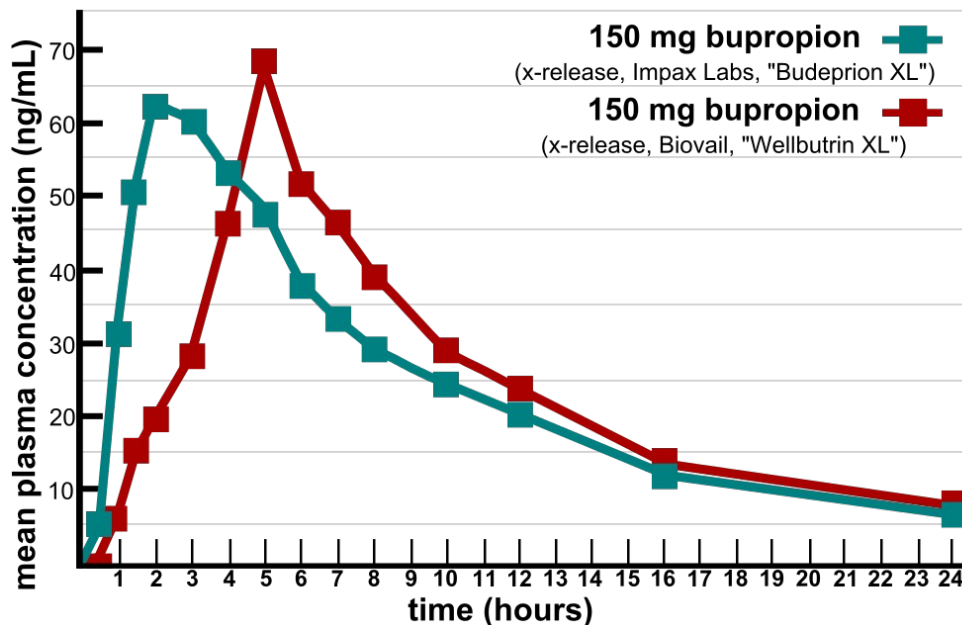


Fig. 9.1 A typical plasma concentration-time profile showing pharmacokinetic and pharmacodynamic parameters, obtained after oral administration of single dose of a drug.

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

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



# Pobieranie próby

# Pobieranie próby

- **Próba** (*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

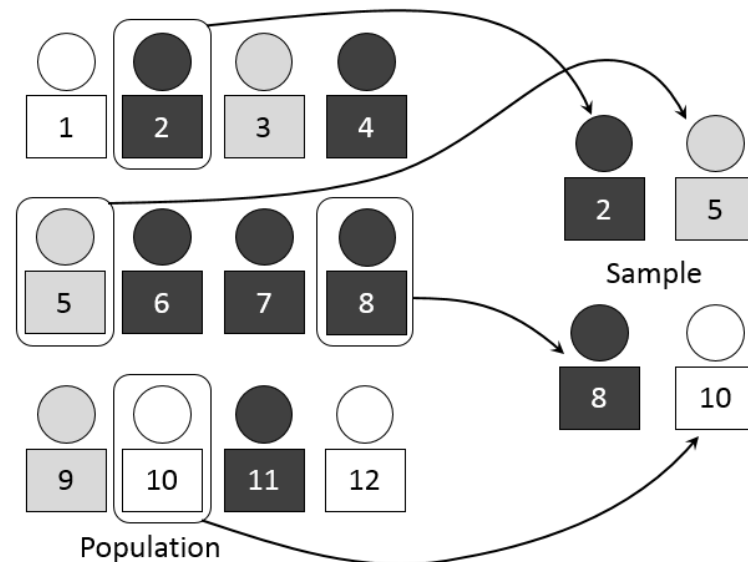
- Właściwości próby losowej:
  - wybór każdego elementu próby jest niezależny
  - każdy element próby jest zmienną losową o takim samym rozkładzie prawdopodobieństwa jak populacja

$$g_1(X_1) = g_2(X_2) = \dots = g_n(X_n) = f(X)$$

czyli

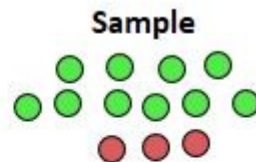
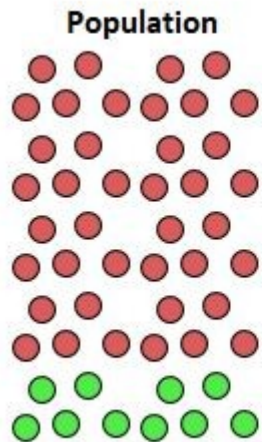
$$E(X_1) = E(X_2) = \dots = E(X_n) = E(X)$$

$$\sigma^2(X_1) = \sigma^2(X_2) = \dots = \sigma^2(X_n) = \sigma^2(X)$$



# Pobieranie próby

- 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



EGESIF\_16-0014-00  
20/01//2017



KOMISJA EUROPEJSKA  
DYREKCJE GENERALNE  
POLITYKA REGIONALNA I MIEJSKA  
ZATRUDNIENIE, SPRAWY SPOŁECZNE I RÓWNOŚĆ SZANS  
GOSPODARKA MORSKA

Wytyczne dotyczące metod doboru próby dla instytucji  
audytowych

Okresy programowania 2007–2013 i 2014–2020

304 strony!

[https://ec.europa.eu/regional\\_policy/sources/docgener/informat/2014/guidance\\_sampling\\_method\\_pl.pdf](https://ec.europa.eu/regional_policy/sources/docgener/informat/2014/guidance_sampling_method_pl.pdf)

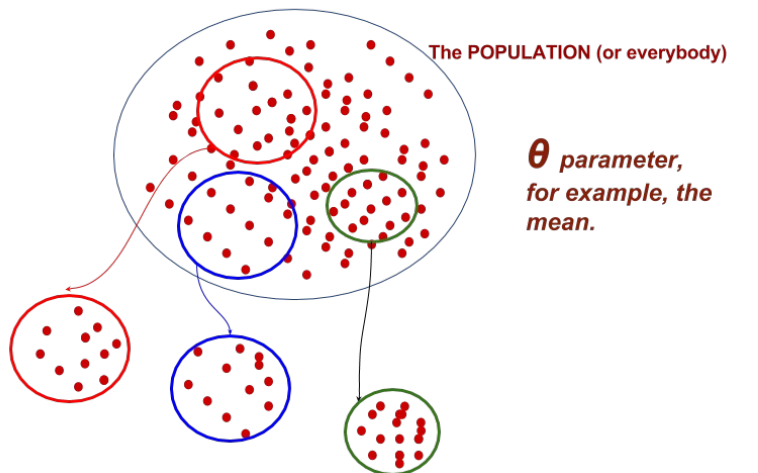
# Estymatory

- Każda funkcja elementów próby nazywana jest **statystyką**

$$S = S(X_1, X_2, \dots, X_n)$$

- Poszukujemy takich statystyk (**estymacja**), które pozwolą nam na podstawie próby wnioskować o własnościach (parametrach) całej populacji
- **Estymator** to statystyka, która pozwala nam na wnioskowanie o zadanym parametrze rozkładu populacji

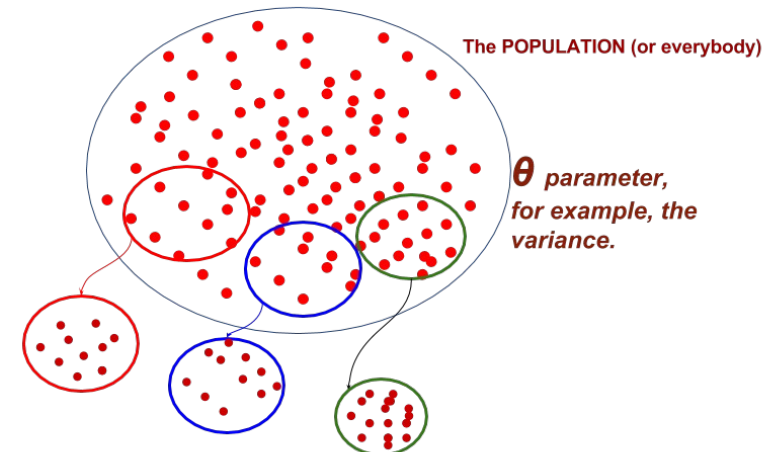
## UN - BIASED ESTIMATOR:



The sample means  $\bar{\theta}$  will reflect  $\theta$  so that the  $E[\bar{\theta}]$ , or expectation (mean) of the sample means is the population mean:

$$\text{Bias} = E[\bar{\theta}] - \theta = 0$$

## BIASED ESTIMATOR:



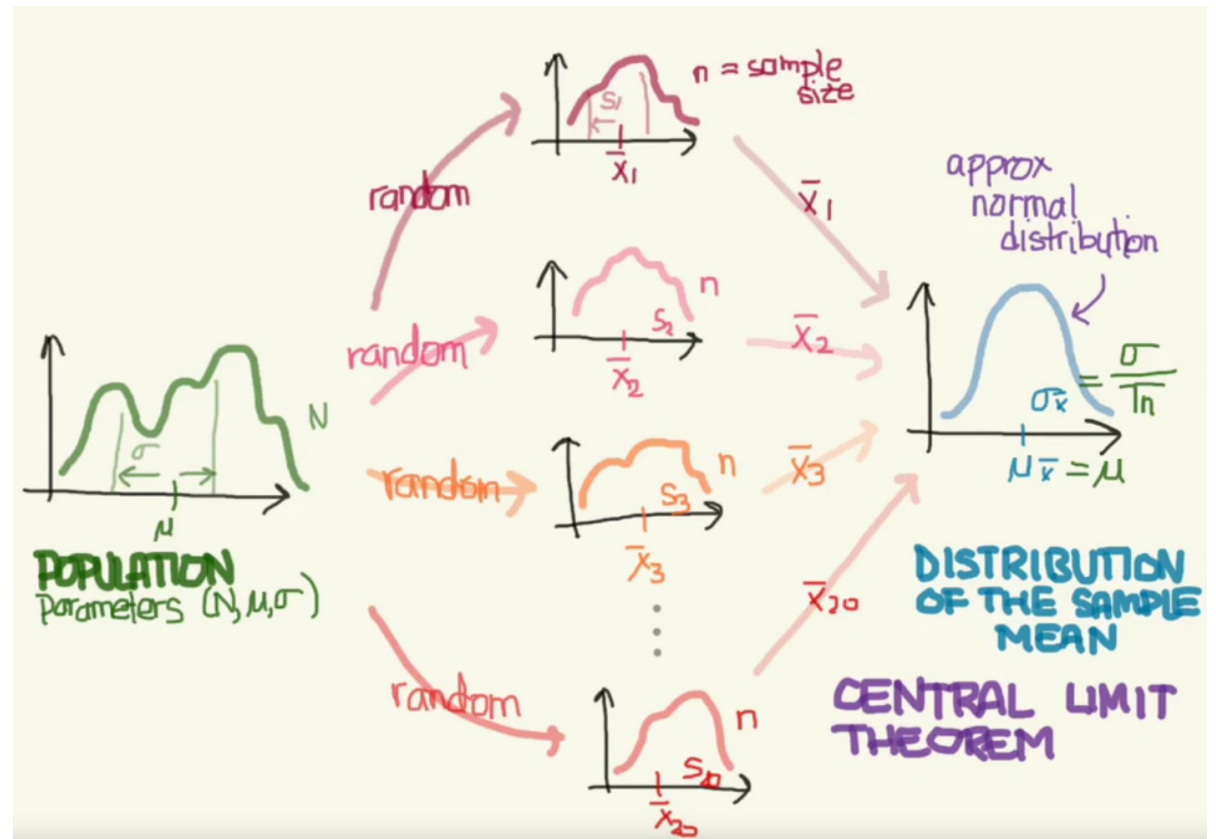
The sample variances  $\bar{\theta}$  will NOT reflect  $\theta$  accurately, so that the  $E[\bar{\theta}]$ , or expectation (mean) of the sample variance is different from the population variance:

$$\text{Bias} = E[\bar{\theta}] - \theta \neq 0$$

# Średnia z próby

- Podstawową przykładową statystyką jest średnia z próby losowej:
  - 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  $\mathbf{X}=(X_1, \dots, X_n)$ ), 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:

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

# Estymator wartości oczekiwanej

## Populacja

- opisana funkcją gęstości:

$$f(x) = P(X=x)$$

- posiada **wartość oczekiwaną**:

$$E(X) = \hat{x} = \int_{-\infty}^{\infty} x f(x) dx$$

- wartość oczekiwana rozkładu to **jedna liczba**
  - nie jest zmienną losową
  - chcemy ją zmierzyć doświadczalnie
- np. dla rozkł. Gaussa:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \mu$$

## Próba losowa

- zakładamy, że **średnia arytmetyczna** z elementów próby jest estymatorem wartości oczekiwanej

- **średnia arytmetyczna** jest statystyką:

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

- **jest zmienną losową** (zależy od elementów próby)
- posiada swoją **wartość oczekiwaną** oraz **wariancję**
- oczekujemy, że będzie ona *estymatorem nieobciążonym* i *zgodnym* wartości oczekiwanej populacji:

$$E(\bar{X}) = E(X) = \hat{x}, \text{ dla każdego } n$$

$$\lim_{n \rightarrow \infty} \sigma(\bar{X}) = 0$$

- **jak to sprawdzić?**

# Estymator wartości oczekiwanej

## Próba losowa

- **średnia arytmetyczna** z elem. próby jest statystyką:

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

- **wartość oczekiwana** średniej arytmetycznej:

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

Czyli wykazaliśmy, że estymator jest **nieobciążony**

- **wariancja** średniej arytmetycznej:

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

Musimy pokazać, że to wyrażenie zbiega do 0 dla dużych  $n$

# Estymator wartości oczekiwanej

## Próba losowa

- **wariancja średniej arytmetycznej:**

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

- **ponieważ elementy próby losowej są niezależne:**

$$\text{Cov}(X_i, X_j) = E\{(X_i - \hat{x})(X_j - \hat{x})\} = 0$$

- **to upraszczamy nawias:**

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

- **zatem ostatecznie wariancja średniej arytmetycznej:**

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

- **ponieważ wariancja rozkładu populacji jest jedną liczbą, to:**

$$\lim_{n \rightarrow \infty} \sigma(\bar{X}) = 0$$

Wykazaliśmy, że estymator jest **zgodny**

# Estymator wartości oczekiwanej

## Populacja

- opisana funkcją gęstości:

$$f(x) = P(X = x)$$

- posiada **wartość oczekiwaną**:

$$E(X) = \hat{x} = \int_{-\infty}^{\infty} x f(x) dx$$

- wartość oczekiwana rozkładu to **jedna liczba**
  - nie jest zmienną losową

- np. dla rozkł. Gaussa:

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu$$

## Próba losowa

- **średnia arytmetyczna** z elem. próby jest statystyką:

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

- jest **estymatorem nieobc. i zgodnym** wartości oczekiwanej populacji

- **wartość oczekiwana** średniej arytmetycznej:

$$E(\bar{X}) = E(X) = \hat{x}, \text{ dla każdego } n$$

- **wariancja** (niepewność wyznaczenia) średniej arytmetycznej:

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

Ale... żeby wyznaczyć jej wartość musimy znaleźć estymator wariancji populacji  $\sigma^2(X)$

# 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:

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

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

- 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 musimy policzyć: **estymator wariancji estymatora wariancji wartości średniej z próby?**

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

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



# Estymatory - podsumowanie

- Przykładowe **estymatory nieobciążone**:

- **wartości oczekiwanej populacji** → **średnia arytmetyczna z próby**:

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

- **wariancji populacji** → **średnia odchyleń kwadratowych**:

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

- Wariancje (niepewności) estymatorów:

- **wariancja średniej arytmetycznej**:

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

- **wariancja średniej odchyleń kwadratowych**:

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

- **Uwaga!** Wariancje estymatorów są również estymatorami – możemy więc liczyć np. wariancję wariancji średniej arytmetycznej, itd.

- **Czy te wzory coś przypominają?**

# Przykład 1 – pomiar bezpośredni

- Mierzymy przy pomocy suwmiarki bok pręta  $d$  o przekroju kwadratowym
- Dokładność suwmiarki (niepewność wzorcowania):

<http://www.kreocen.pl/img/p/1057776/1/TESA-Suwmiarka-STANDARD-005-mm.jpg>  
[http://www.drut.com.pl/images/com\\_sobi2/gallery/69/69\\_image\\_1\\_bml.jpg](http://www.drut.com.pl/images/com_sobi2/gallery/69/69_image_1_bml.jpg)



$$\Delta d = 0,1 \text{ mm}$$

- Seria  $n=11$  pomiarów (w mm): 12,5; 12,3; 12,6; 12,5; 12,6; 12,5; 12,4; 12,3; 12,5; 12,4; 12;6

- Wynik (średnia arytmetyczna):

$$\bar{d} = \sum_{i=1}^n d_i = 12,4727 \text{ mm}$$

- Niepewność typu A:

$$u_A(d) = \sqrt{\sum_{i=1}^n \frac{1}{n(n-1)} (d_i - \bar{d})^2} = 0,033278 \text{ mm}$$

- Niepewność typu B:

$$u_B(d) = \frac{\Delta d}{\sqrt{3}} = 0,057735 \text{ mm}$$

- Niepewność całkowita:

$$u(d) = \sqrt{u_A^2(d) + u_B^2(d)} = 0,06639 \text{ mm} \approx 0,066 \text{ mm}$$

- Wynik:  $d = 12,473(66) \text{ mm}$

**WYKŁAD 1**

# 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):

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

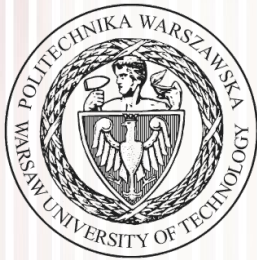
- wariancji wartości średniej z próby (**patrz niepewność typu A**):

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

- 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)...

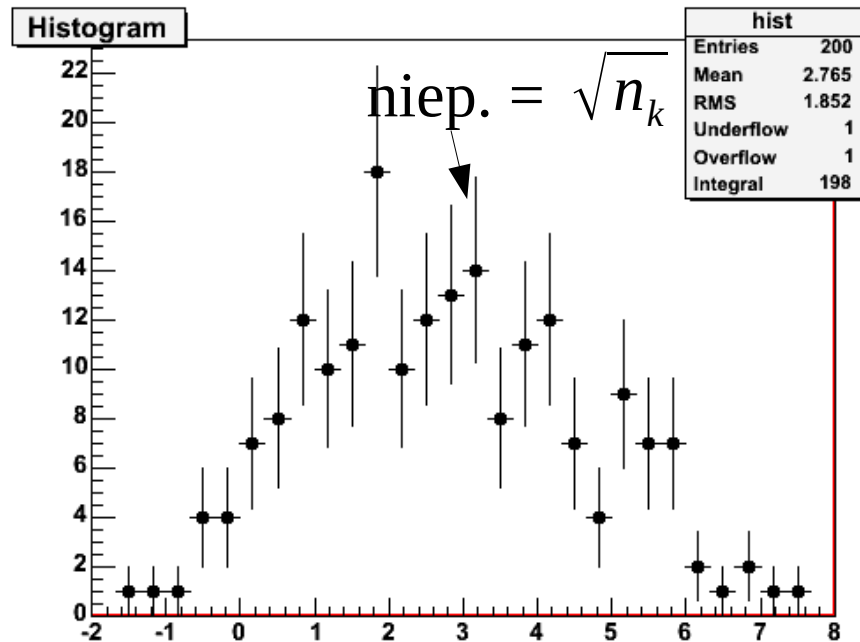
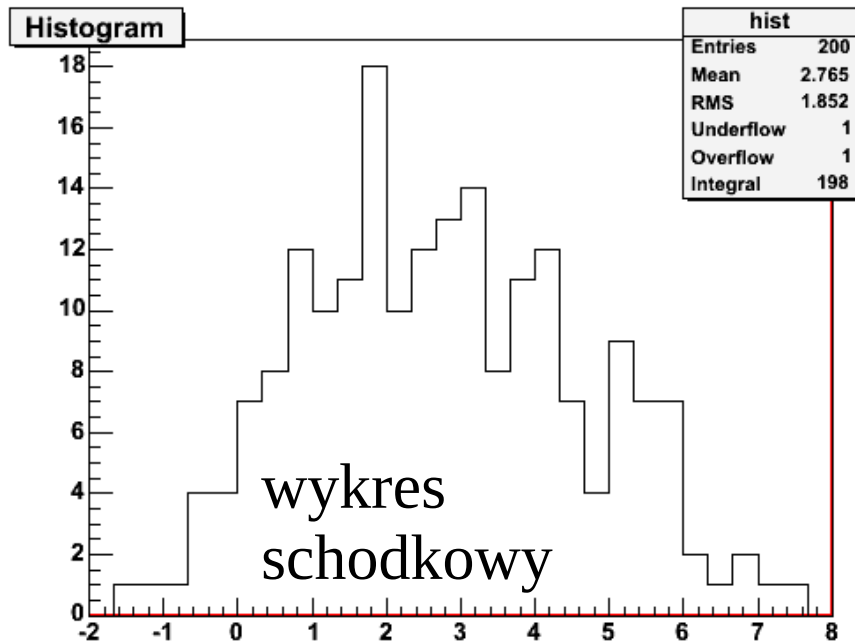
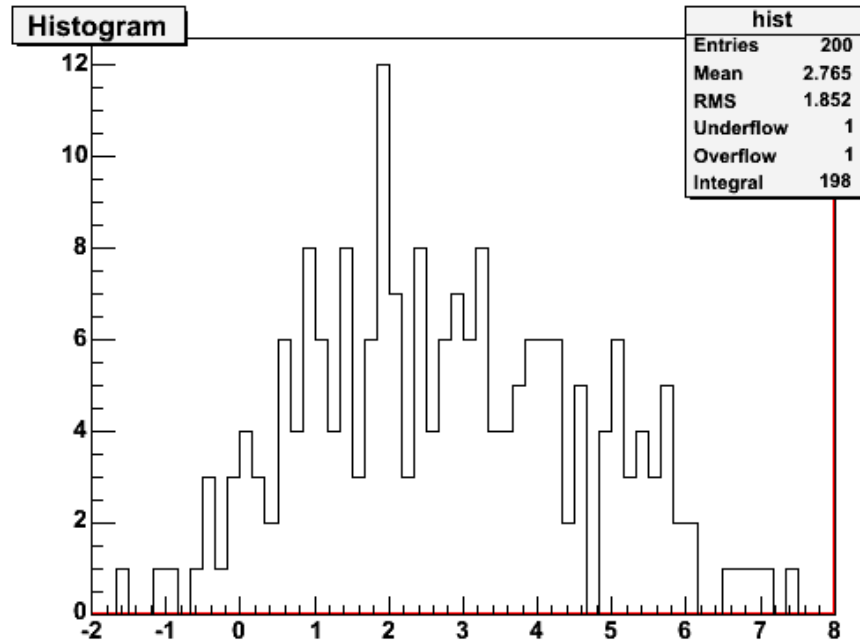
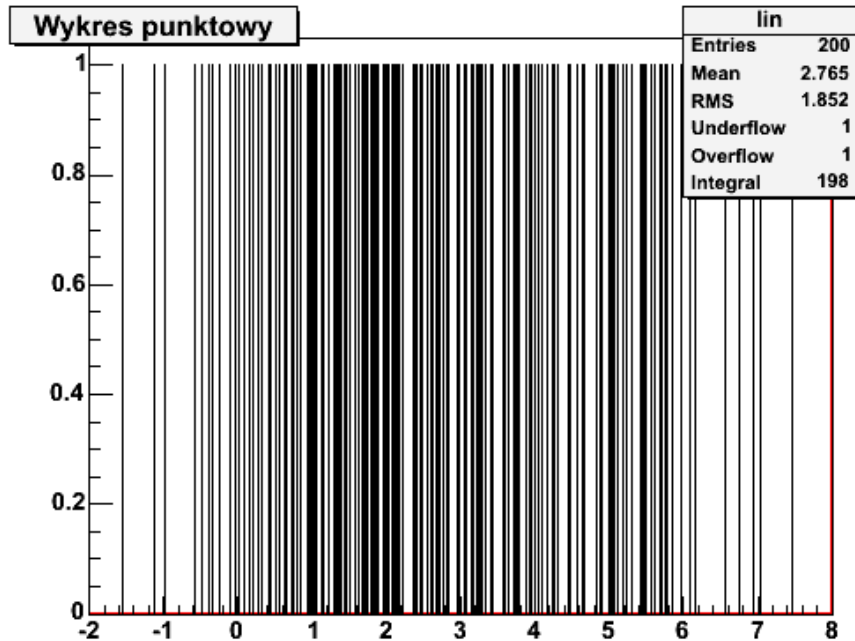


# Graficzne przedstawienie próby

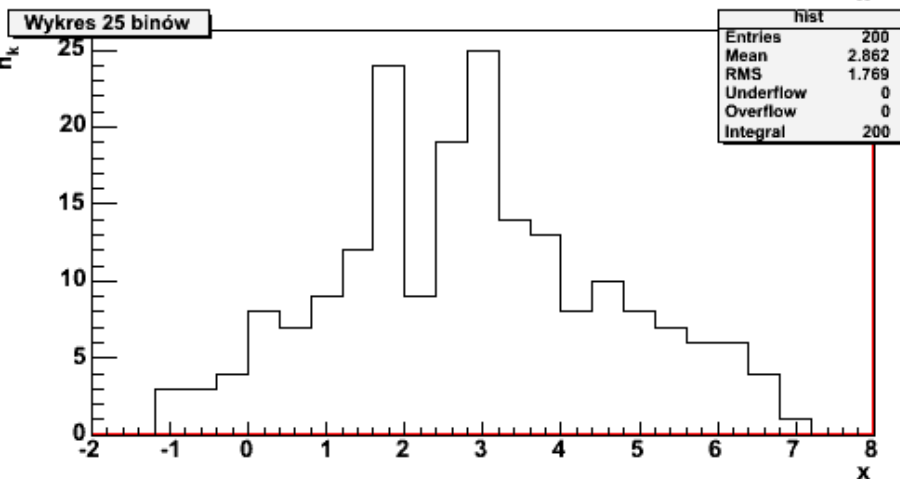
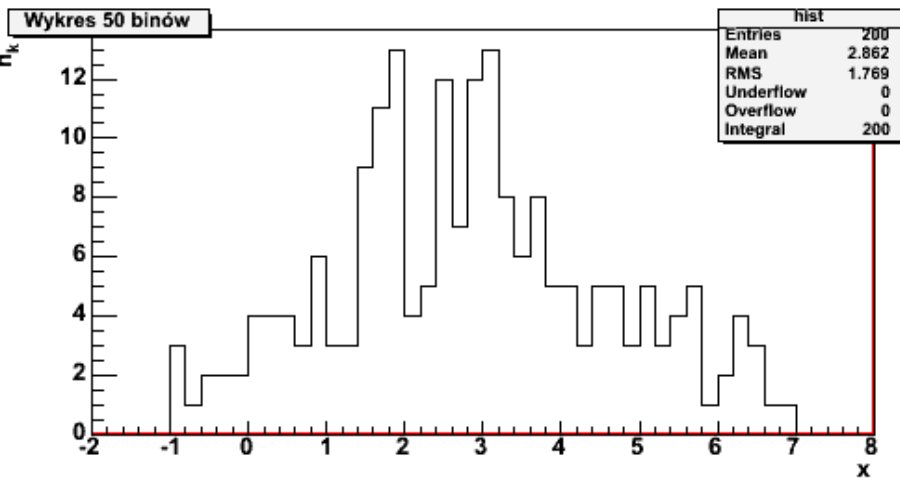
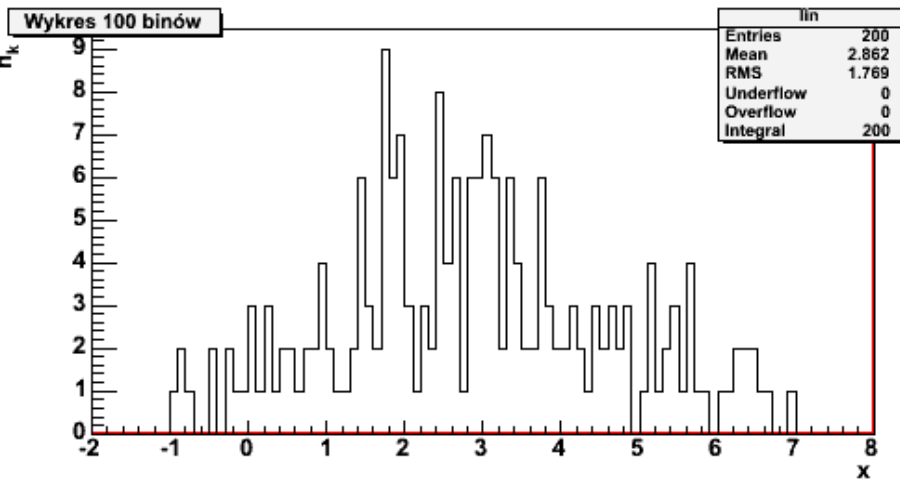
# Graficzne przedstawienie próby

- 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**
  - **wykres częstości służy do przybliżania rozkładu gęstości**

# 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 licznosci 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

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

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)

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

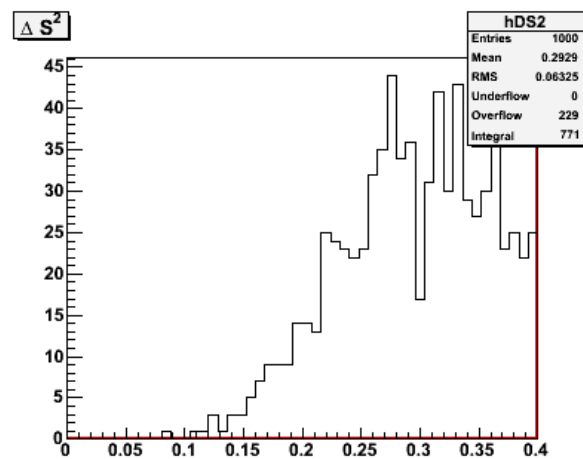
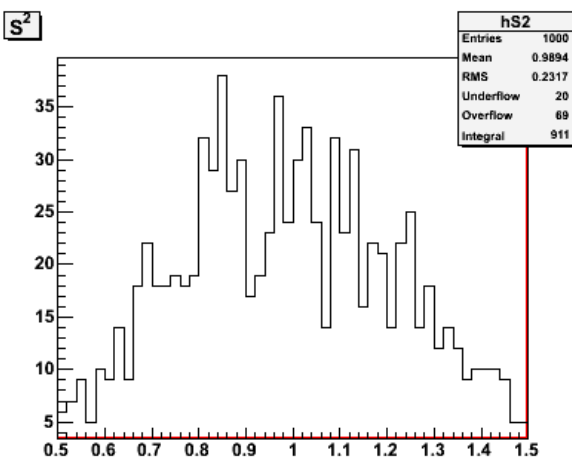
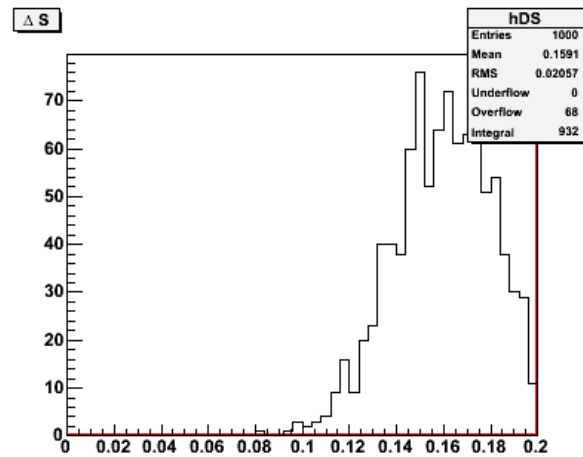
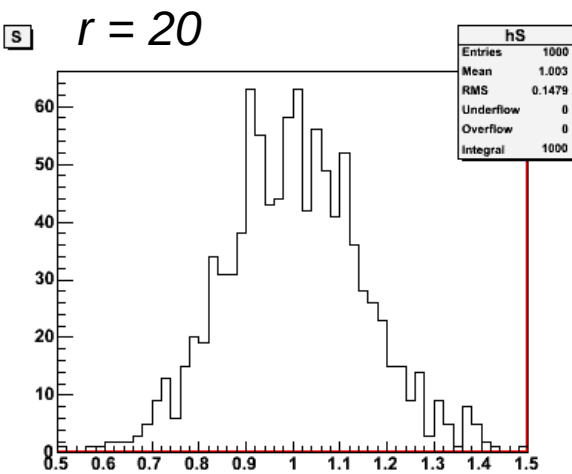
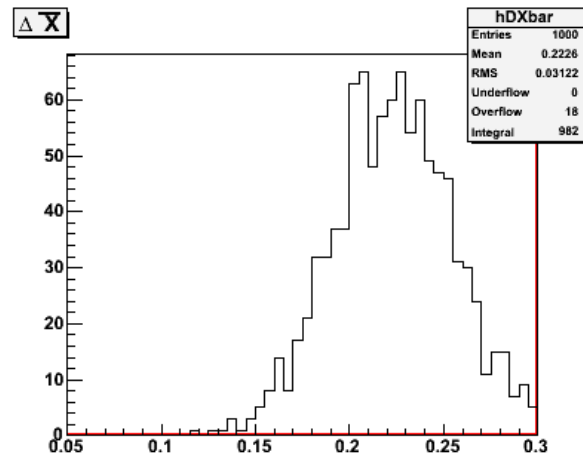
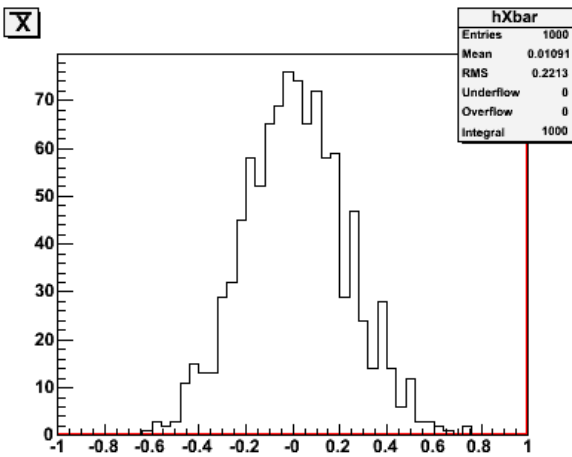
niepewność estymatora odch. std. populacji – estymator odch. std. estymatora **odch. std.** populacji

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

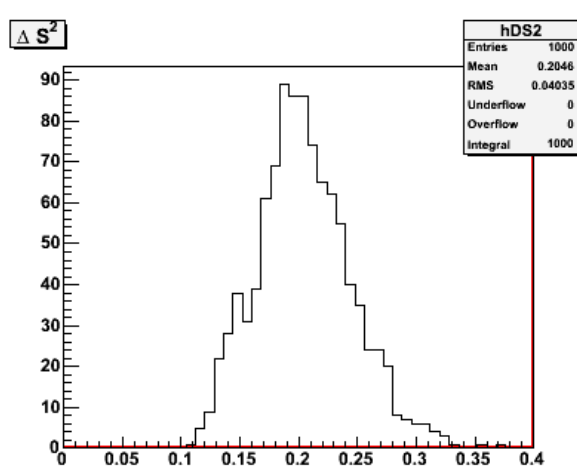
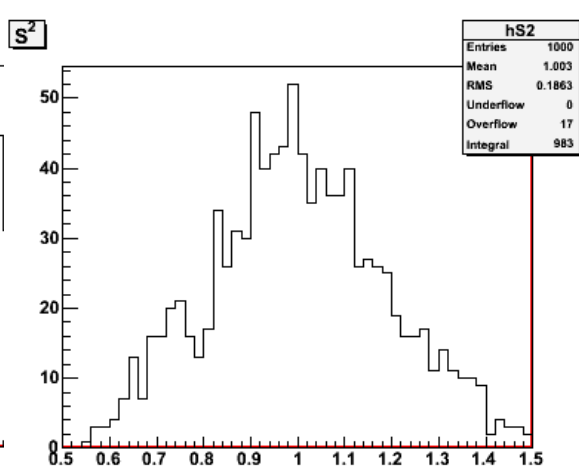
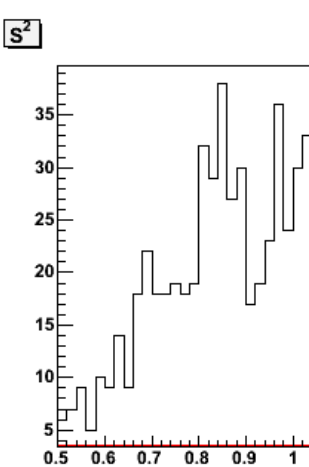
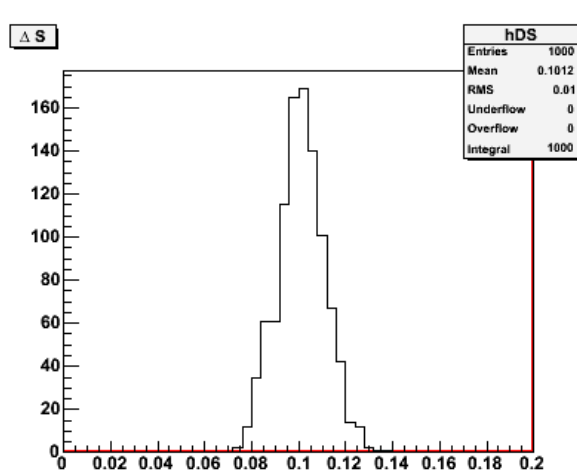
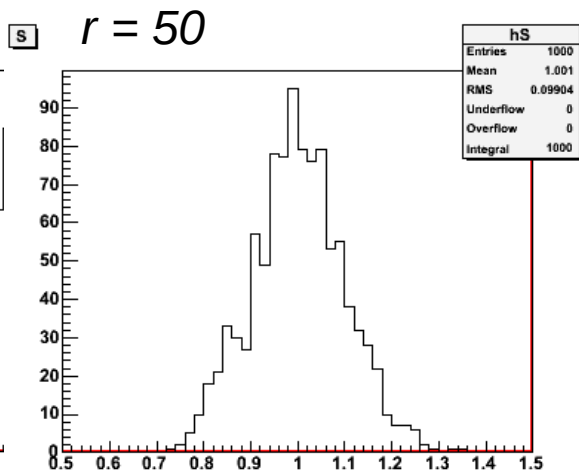
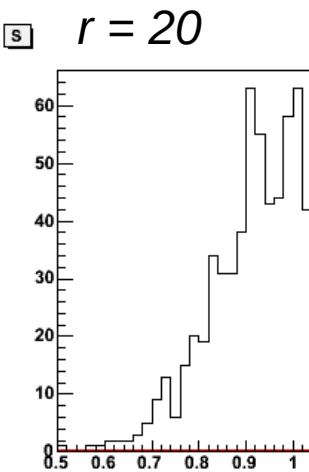
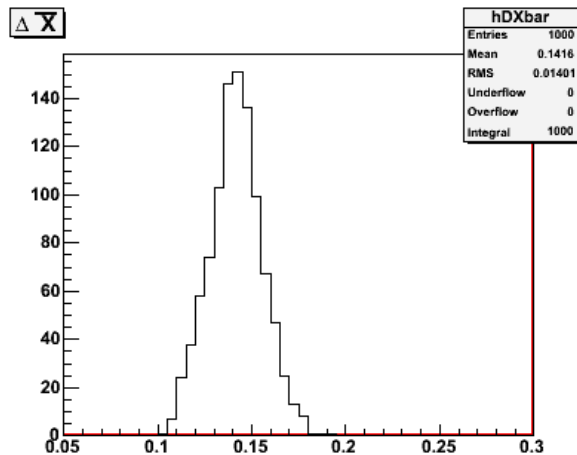
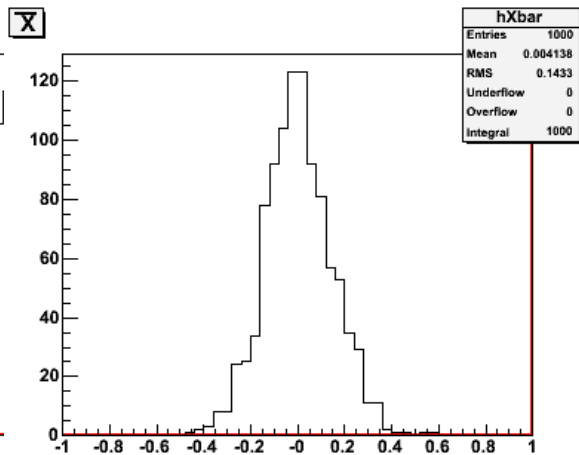
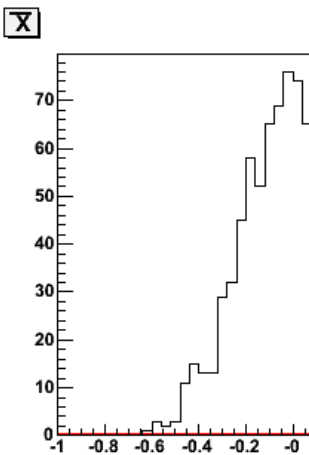
niepewność estymatora wariancji populacji – estymator odch. std. estymatora **wariancji** populacji



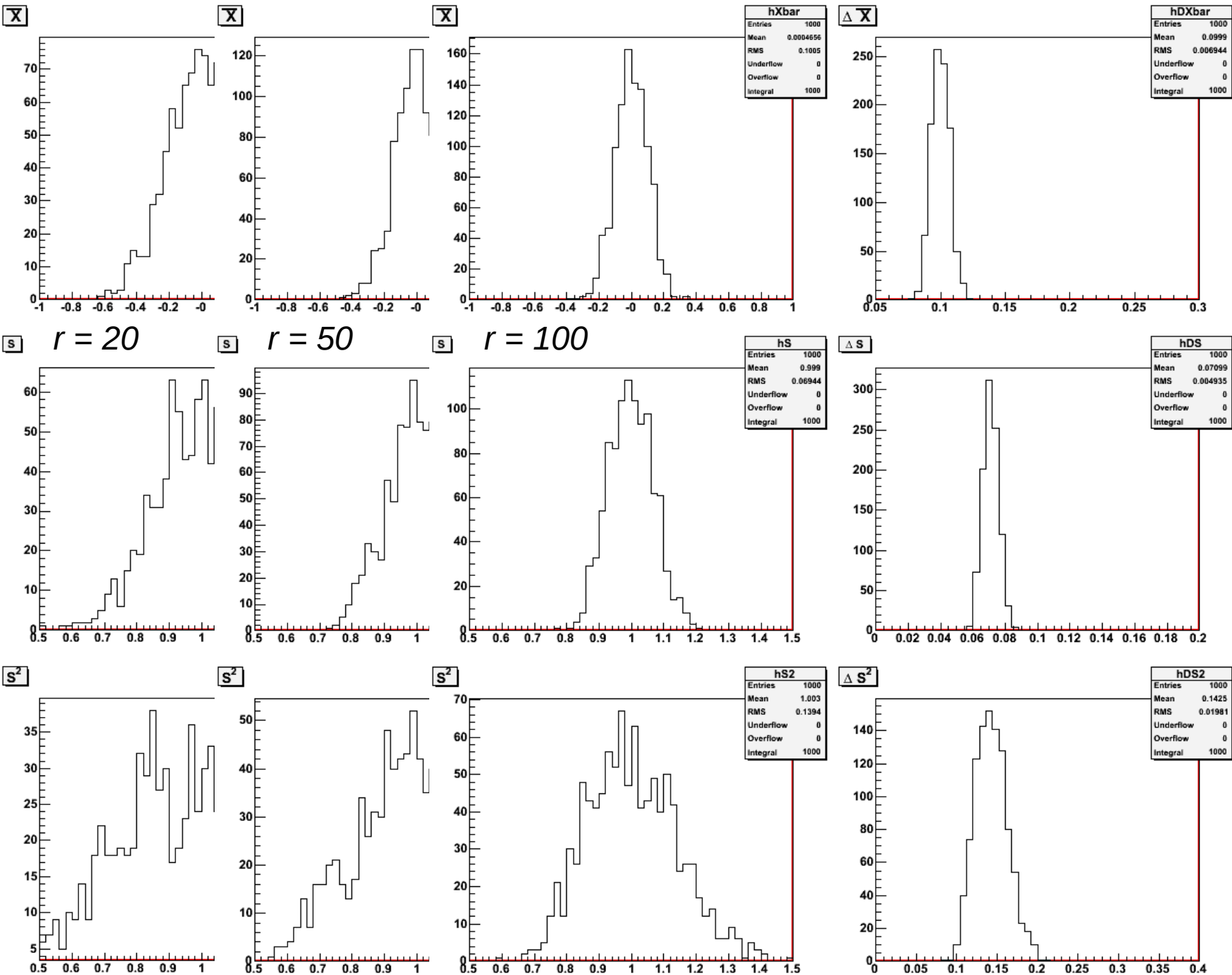
# Estymatory - histogramy



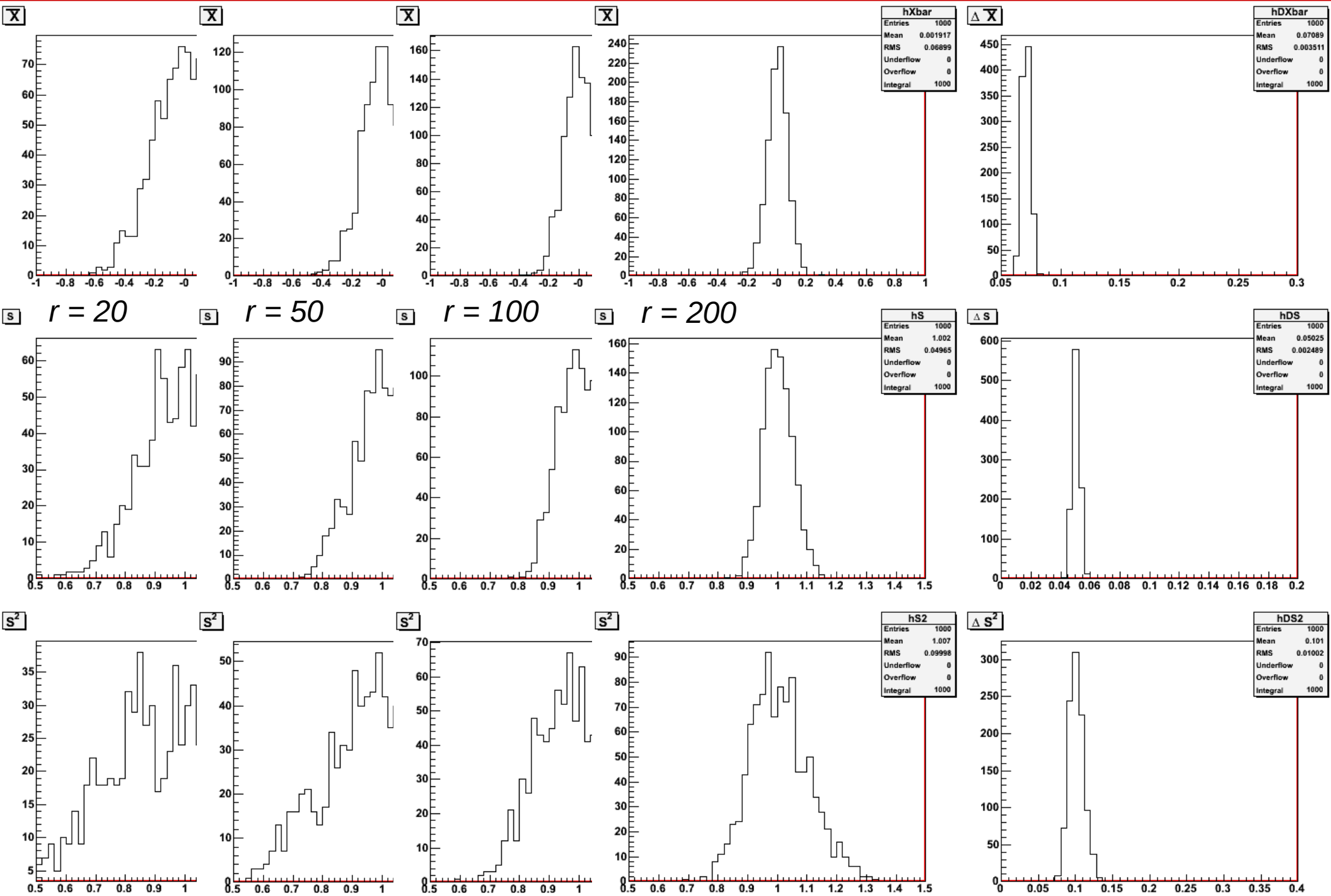
# Estymatory - histogramy



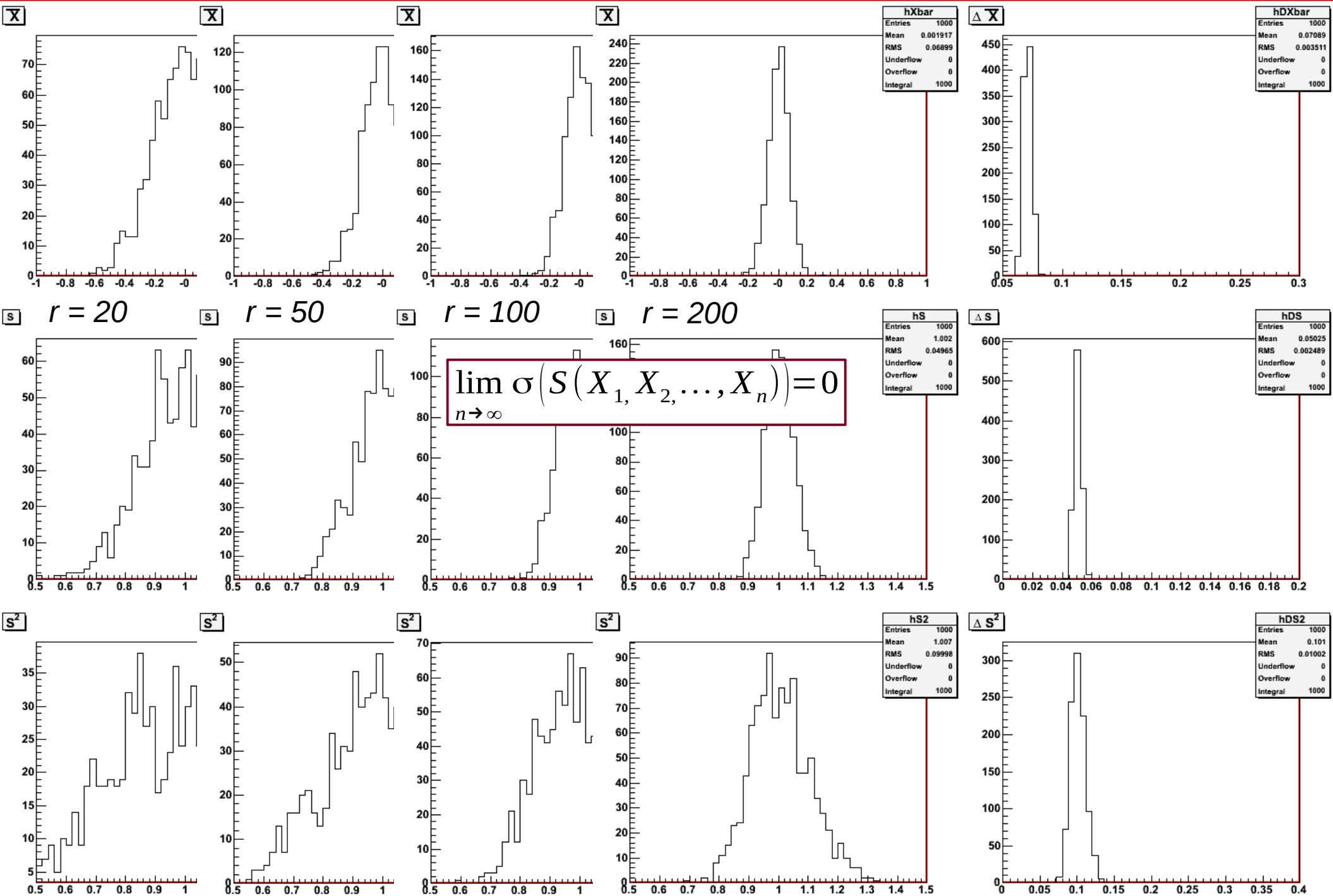
# Estymatory - histogramy



# Estymatory - histogramy



# Estymatory - histogramy





**KONIEC**