Komputerowa analiza danych doświadczalnych

Wykład 7
9.04.2021

dr inż. Łukasz Graczykowski
lukasz.graczykowski@pw.edu.pl

Semestr letni 2020/2021

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:

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

  - rozkład normalny będzie mieć też zmienna:

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

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

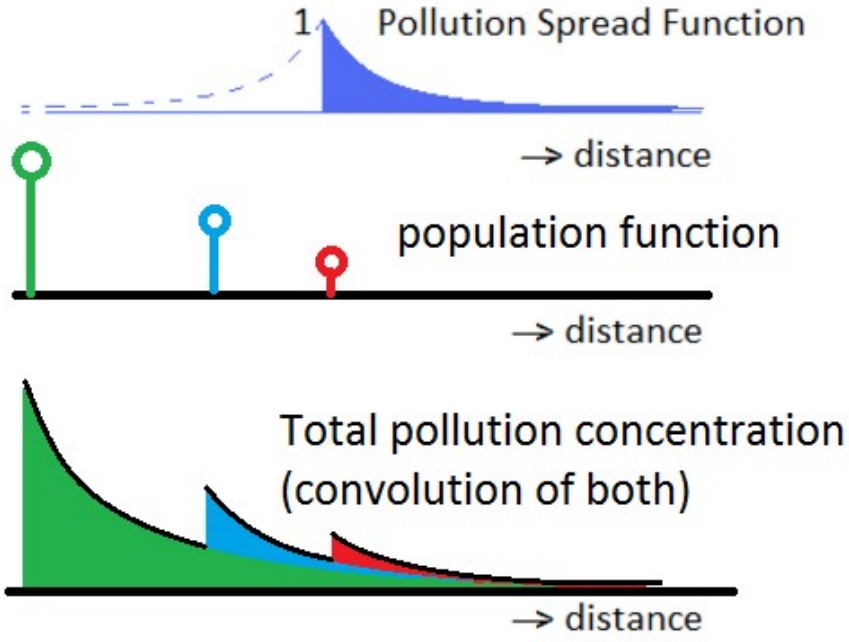

WYDZIAŁ FIZYKI
POLITECHNIKA WARSZAWSKA
POLITECHNIKA WARSZAWSKA
WARSAW UNIVERSITY OF TECHNOLOGY

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

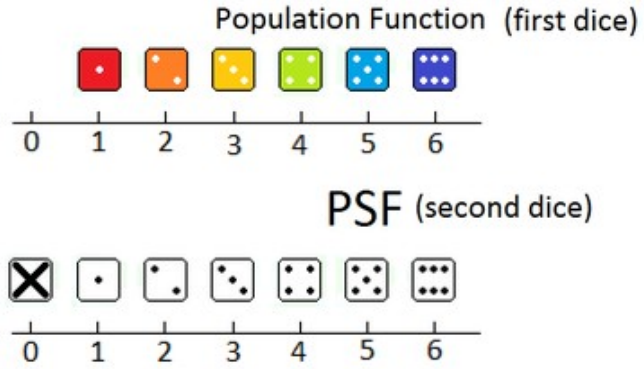


- Jaka jest pełna funkcja opisująca poziom zanieczyszczeń w rzece?
- Jest to **splot** dwóch rozkładów – funkcji populacji oraz funkcji koncentracji odpadów
- Innymi słowy, zastępujemy każdy punkt w funkcji populacji przez funkcję koncentracji przeskalowaną przez wagę funkcji populacji

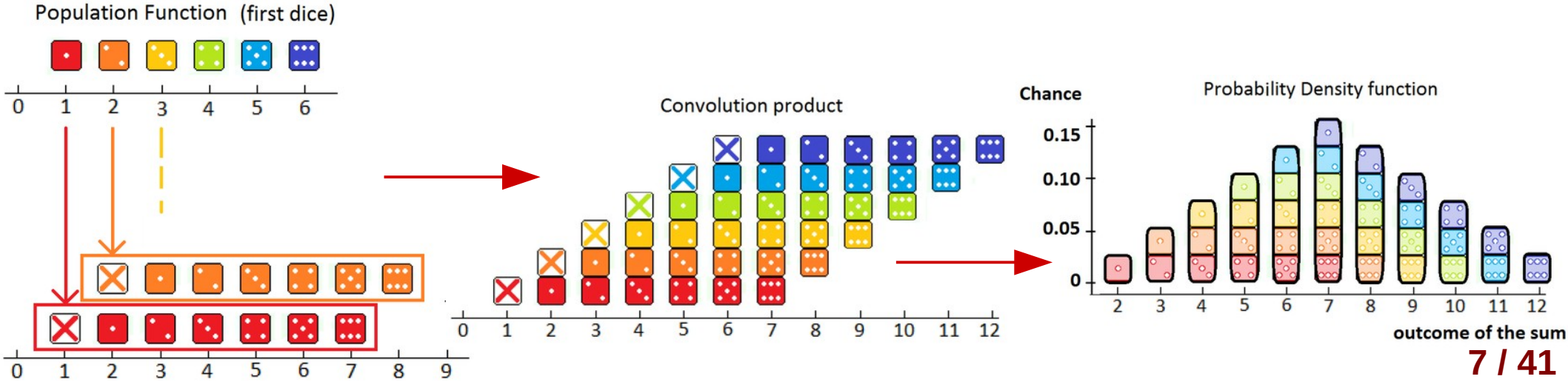
Suma zmiennych losowych jako splot

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% popopulacji 2 km, itd.



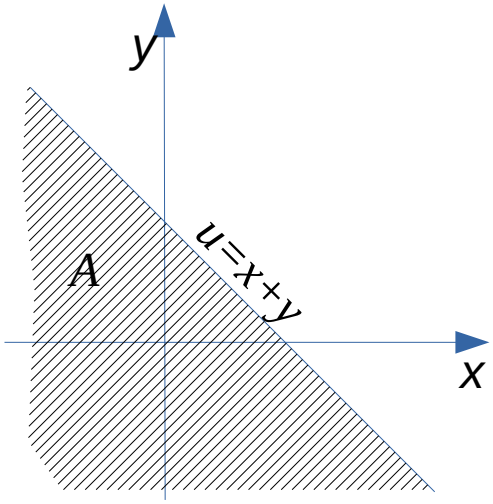
- Drugi rzut kostką oznacza funkcję PSF – jak bardzo dana miejscowość zanieszyszcza rzekę, i znowu 16,7% zanieczyszczeń ląduje 1 km dalej od miasta, 16.7% 2 km dalej od miasta, itp.
- Jak policzyć pełną funkcję zanieczyszczeń? Podmieniamy funkcję populacji poprzez funkcję zanieczyszczeń, dla każdego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

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



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

Zgodnie z definicją dystrybuanty:

$$F(u) = P(U \leqslant u) = P((-\infty; u>)$$

Suma zmiennych losowych jako splot

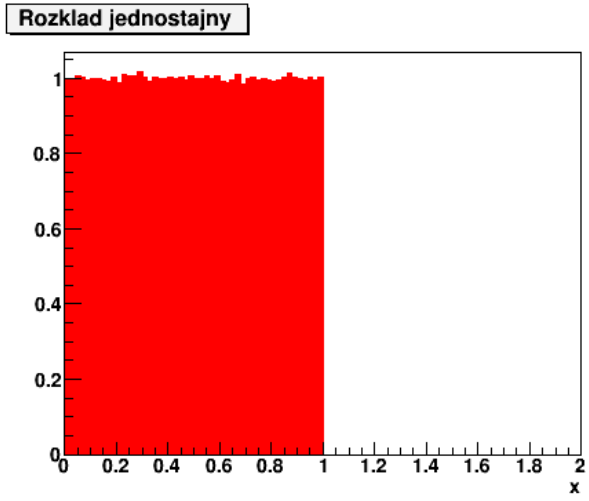
- Z dystrybuanty wyznaczamy funkcję gęstości zmiennej U:

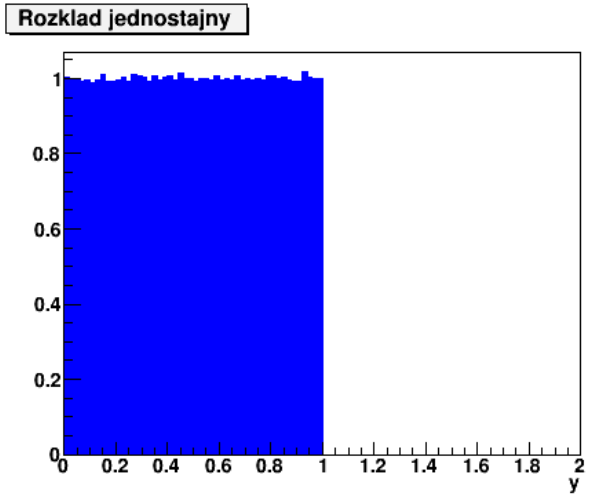
$$f(u)=\frac{dF(u)}{du}=\int_{-\infty}^{\infty} f_x(x) f_y(u-x)\,dx=\int_{-\infty}^{\infty} f_y(y) f_x(u-y)\,dy\equiv(f_x * f_y)(u)$$

- 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:

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

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





Suma zmiennych losowych jako splot

- Splot dwóch rozkładów jednorodnych:

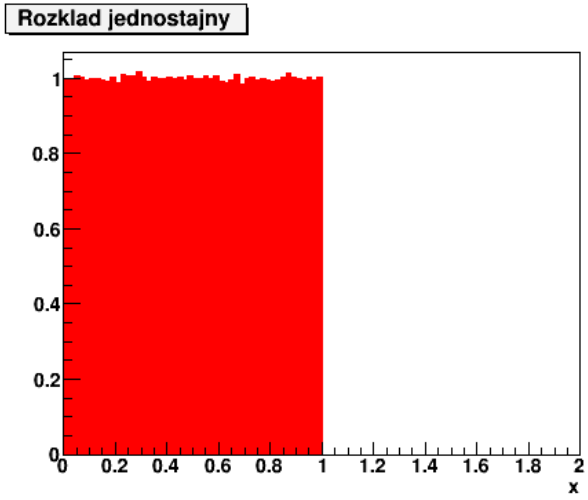
$$f_x(x)=\begin{cases}1, & 0\leqslant x<1\\ 0, & \text{w przeciwnym razie}\end{cases}\qquad f_y(y)=\begin{cases}1, & 0\leqslant y<1\\ 0, & \text{w przeciwnym razie}\end{cases}$$

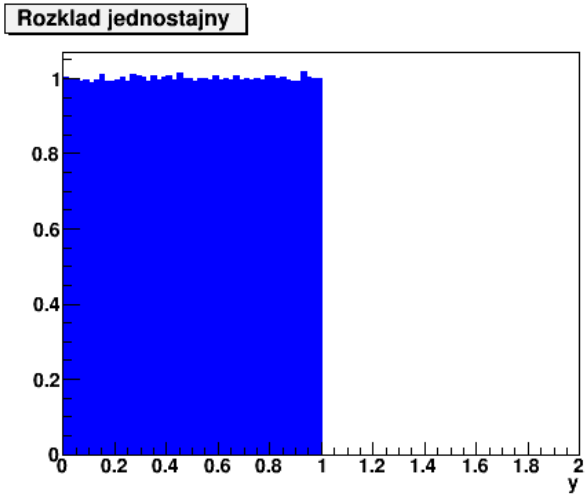
$$f(u)=\int_0^1 f_x(x)f_y(u-x)\,dx=\int_0^1 1\cdot f_y(u-x)\begin{matrix}v=u-x\\ dv=-dx\end{matrix}\Rightarrow \quad f(u)=-\int_u^{u-1} f_y(v)\,dv=\int_{u-1}^{u} f_y(v)\,dv$$

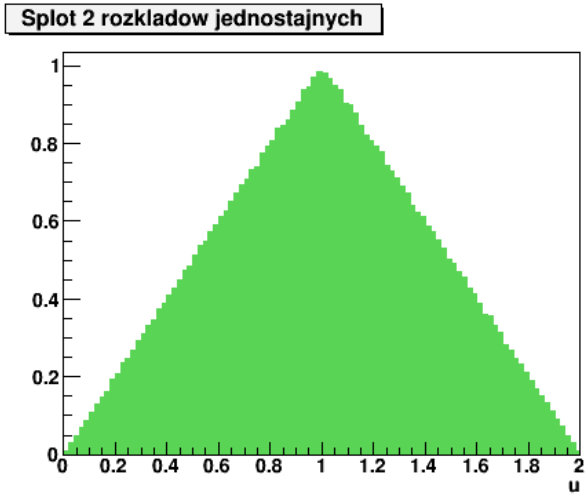
- Zmienna u zmienia się od 0 do 2, zatem rozważmy 2 przypadki:

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

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







Suma zmiennych losowych jako splot

- Rozpatrzmy przypadek splotu dwóch rozkładów jednorodnych:

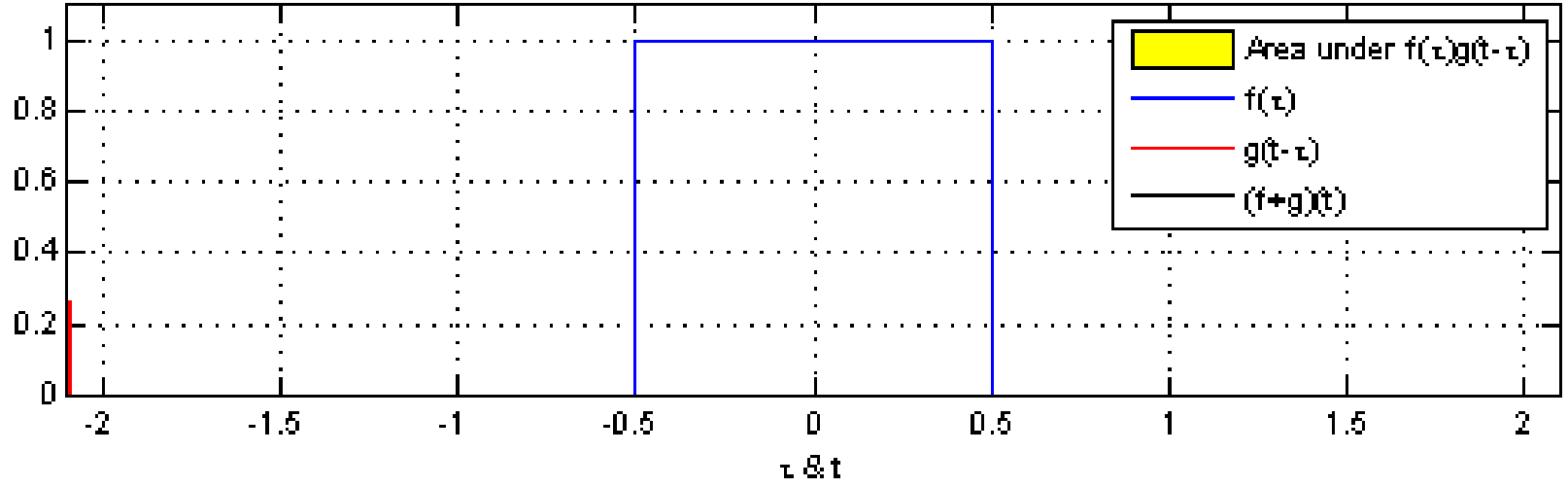
$$f_x(x)=\begin{cases}1, & 0\leqslant x<1\\ 0, & \text{w przeciwnym razie}\end{cases}\qquad f_y(y)=\begin{cases}1, & 0\leqslant y<1\\ 0, & \text{w przeciwnym razie}\end{cases}$$

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

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

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

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



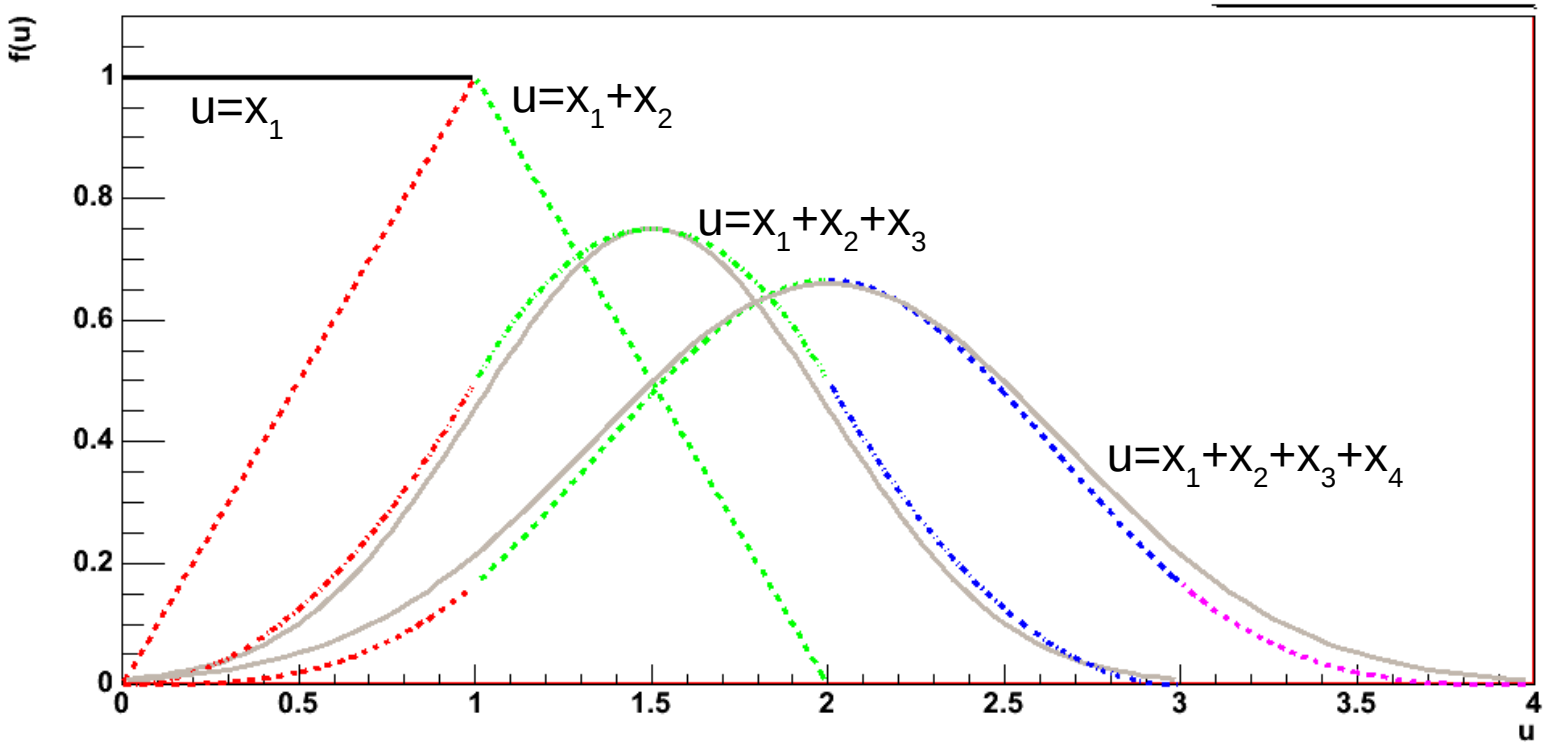
https://en.wikipedia.org/wiki/Convolution#/media/File:Convolution_of_box_signal_with_itself2.gif

Suma zmiennych losowych jako splot

- Analogicznie będzie z sumą trzech zmiennych losowych:

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

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



Sploty z rozkładem normalnym

- Przykład: Mierzymy 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:

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

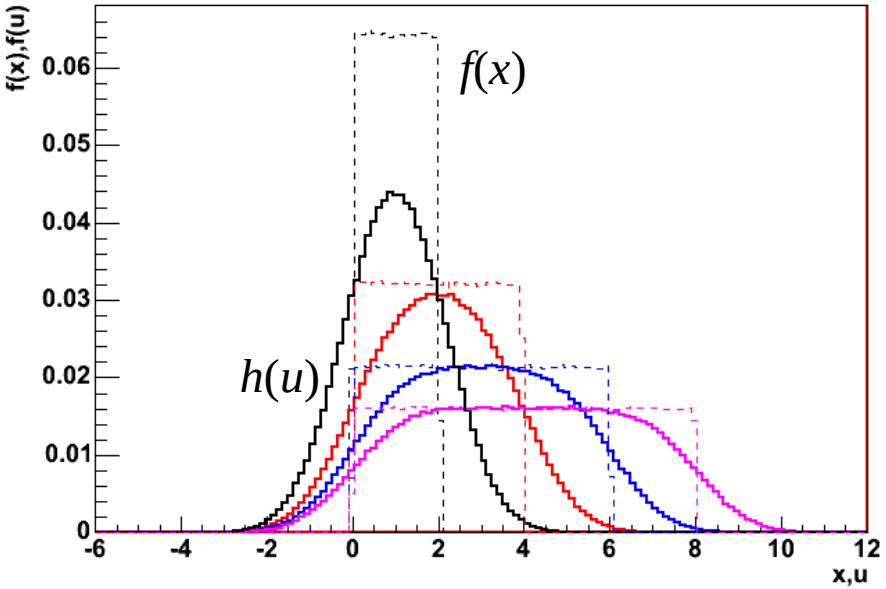
$$f(x)=0;\ x\in\mathbb{R}\setminus\langle a,b\rangle$$

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

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

- Z uwzględnieniem stablicowanej dystrybuanty rozkładu normalnego:

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

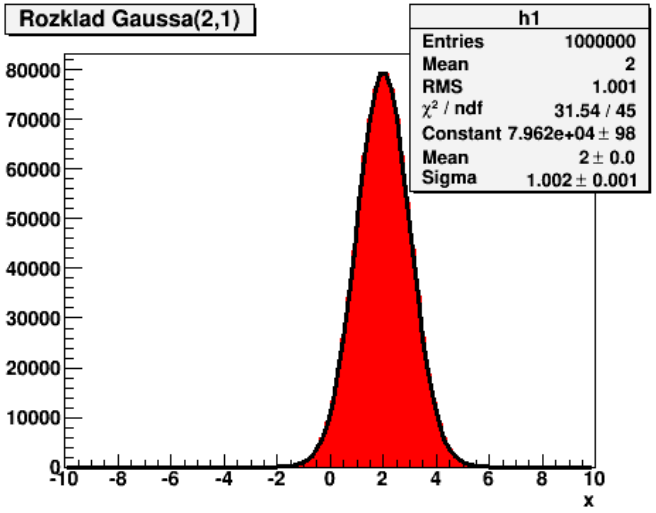


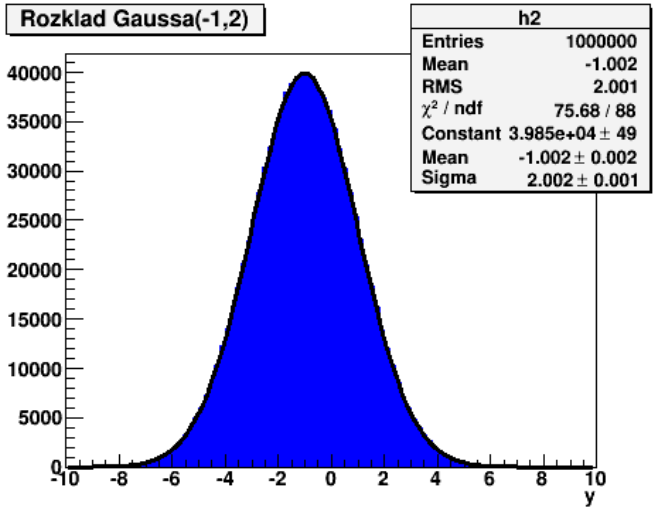
Sploty z rozkładem normalnym – przykład 2

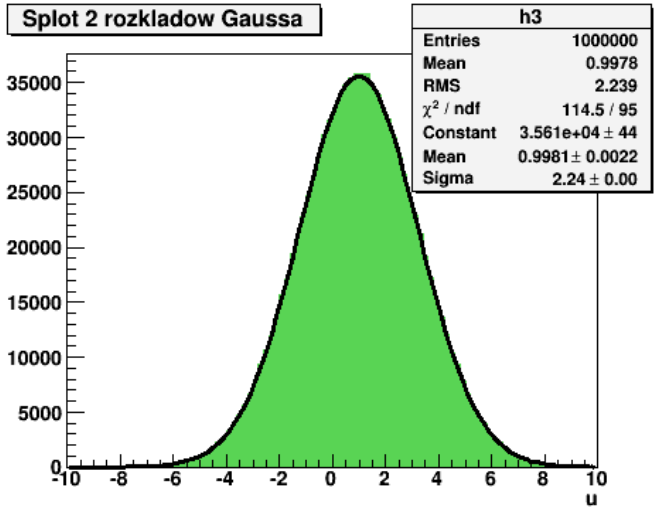
- Przykład: Splot dwóch rozkładów normalnych – dodawanie niepewności “w kwadracie”
- Splot dwóch rozkładów normalnych o wartościach średnich równych 0 i warniancjach σ_x, σ_y ma postać rozkładu normalnego:

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

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



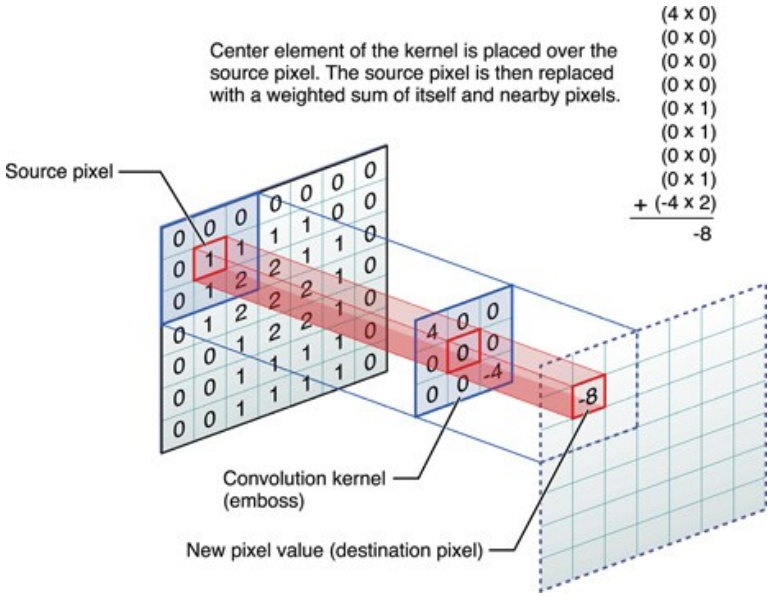




Zastosowanie splotów

- 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

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

Zastosowanie splotów

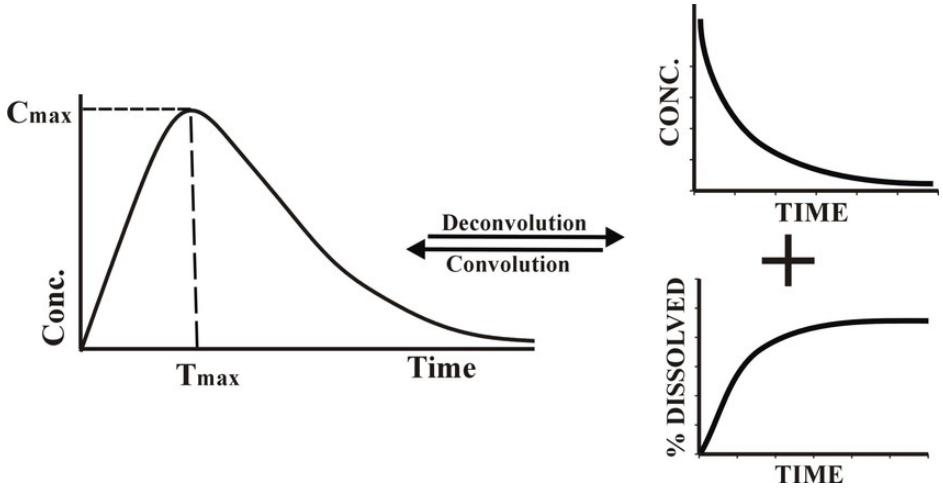
- Bardzo ważnym zastosowaniem splotów są badania farmakokinetyczne leków – koncentracja leku w osoczu krwi w czasie jest splotem funkcji absorpcji leku oraz jego eliminacji

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

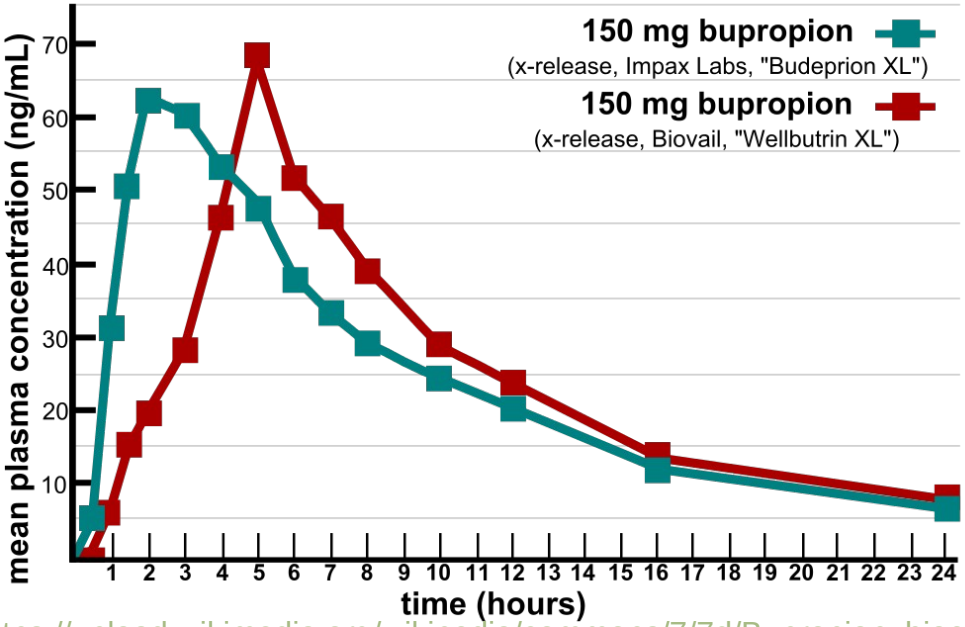
- The absorption rate r_{abs} that results isn plasma concentration c(t) may be estimated by solving following eq.

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

C_δ 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.slideshare.net/jaspreetguraya/in-vitro-in-vivo-correlation-ivivc



https://upload.wikimedia.org/wikipedia/commons/7/7d/Bupropion_bioequivalency_comparison.svg

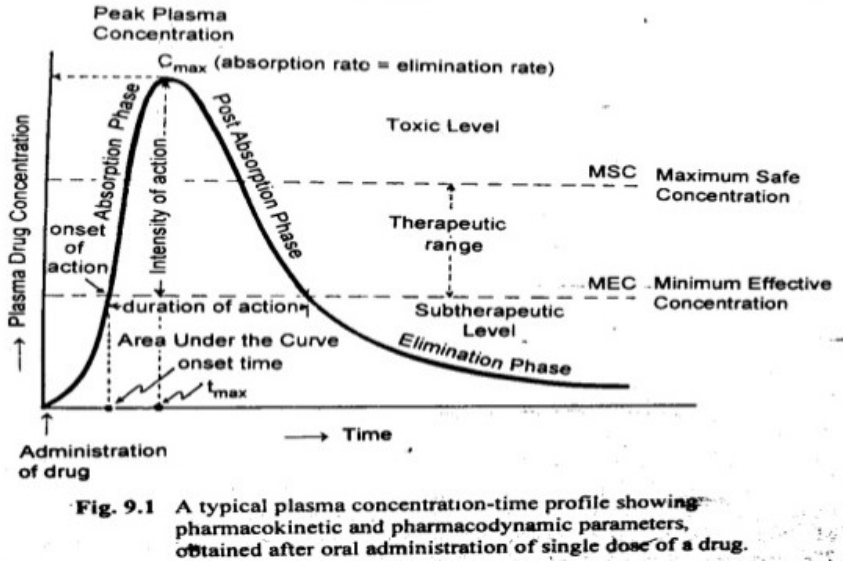


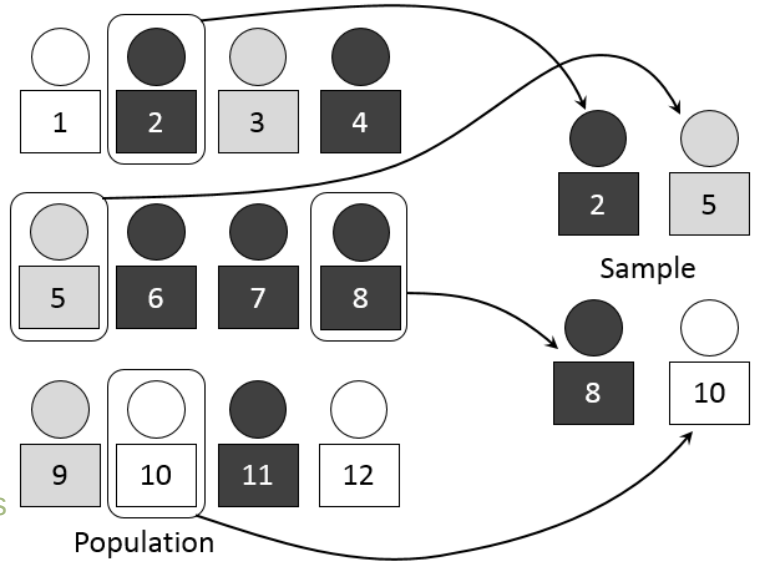
Fig. 9.1 A typical plasma concentration-time profile showing pharmacokinetic and pharmacodynamic parameters, obtained after oral administration of single dose of a drug.

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

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:

$$\begin{aligned} &1.\,próba: && X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)} \\ &\vdots \\ &j\text{-ta próba}: && X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)} \\ &\vdots \\ &l\text{-ta próba}: && X_1^{(l)}, X_2^{(l)}, \ldots, X_n^{(l)} \end{aligned}$$

Oczywiście, to co pobraliśmy jest też zmienną losową!
Elementy próby losowej (wartości zmiennej X), są zmienną losową

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

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

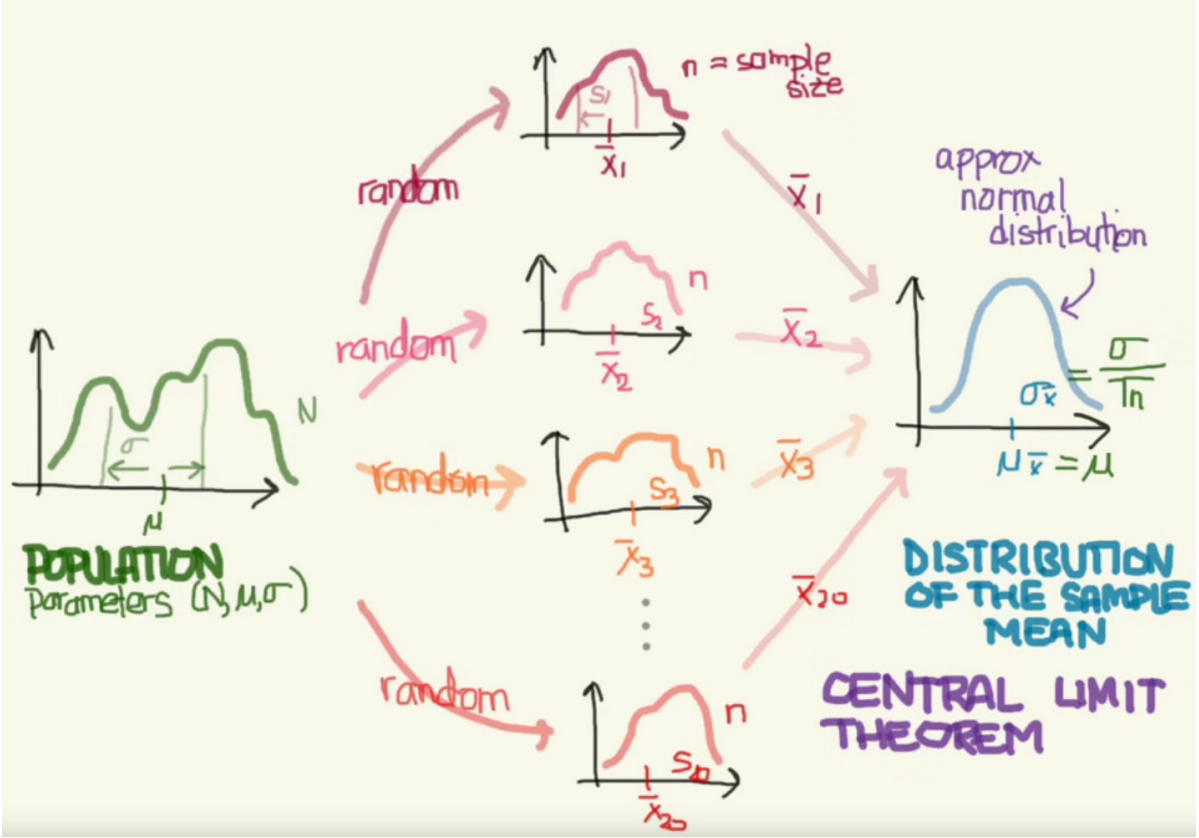
Pobieranie próby

- Aby można było mówić o losowym pobieraniu próby:
  - zmienne X_i muszą być niezależne, czyli: $g(\mathbf{x}) = g_1(x_1) g_1(x_2) \ldots g_n(x_n)$
  - poszczególne rozkłady muszą być jednakowe i identyczne z rozkładem gęstości populacji: $g_1(x_1) = g_2(x_2) = \ldots = g_n(x_n) = f(x)$
- Należy podkreslić, 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:

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



https://upload.wikimedia.org/wikipedia/commons/1/17/Empirical_CDF.png

CTG - powtórzenie

- Przykład:
  - wyobraźmy sobie, że szacujemy wzrost w **populacji** ośmioletnich dzieci w Polsce. Rozkład populacji ma parametry: μ, σ
  - wybieramy losowo 100 8-latków i liczymy średnią wartość z próby losowej – $\overline{X}_1$
  - nasz kolega wykonuje analogiczne doświadczenie → dostaje inny wynik – $\overline{X}_2$
  - zaczynamy więc pracować razem, znowu wybieramy 100 8-latków i dostajemy trzeci wynik –$\overline{X}_3$
  - 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 $\boldsymbol{X}=(X_1,\ldots,X_n)$), 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 λ wyznaczamy na podstawie próby – mierząc skończoną ilość razy ilość rozpadów w czasie → wynik nigdy nie będzie dokłady, bo próba jest skończona, mamy problem **estymacji parametrów**
  - poszukiwana wielkość uzyskiwana jest funkcją elementów próby (**statystyką**) i jest nazywana **estymatorem**: $S = S(X_1, X_2, \ldots, X_n)$
  - estymator jest **nieobciążony**, jeżeli niezależnie od liczebności próby jego wartość oczekiwana jest równa wartości estymowanego parametru:

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

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

$$\lim_{n \to \infty} \sigma\left(S(X_1, X_2, \ldots, X_n)\right) = 0$$

Estymatory – wartość oczekiwana

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

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

- Wniosek: **wartość średnia (arytmetyczna) z próby to estymator <u>nieobciążony</u> wartości oczekiwanej** zmiennej X w populacji

- Możemy obliczyć wariancję wartości średniej:

$$\sigma^2(\bar{X}) = E\left\{\bar{X} - E(\bar{X})\right\}^2 = E\left\{\left(\frac{x_1 + x_2 + \ldots + x_n}{n} - \hat{x}\right)^2\right\}$$

$$= \frac{1}{n^2} E\left\{\left[(X_1 - \hat{x}) + (X_2 - \hat{x}) + \ldots + (X_n - \hat{x})\right]^2\right\}$$

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

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

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

Estymatory - wariancja

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

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

- Wartość oczekiwana tej wielkości:

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

$$=\frac{1}{n}E\left\{\sum_{i=1}^{n}(X_i-\hat{x})^2+\sum_{i=1}^{n}(\hat{x}-\bar{X})^2+2\sum_{i=1}^{n}(X_i-\hat{x})(\hat{x}-\bar{X})\right\}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\{E\left((X_i-\hat{x})^2\right)-E\left((\bar{X}-\hat{x})^2\right)\right\}=\frac{1}{n}\left\{n\sigma^2(X)-n\left(\frac{1}{n}\sigma^2(X)\right)\right\}$$

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

- 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:

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

- Otrzymyjemy **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 <u>z próby</u>**:

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

- Zaś odpowiadające odchylenie standardowe (**niepewność średniej <u>z próby</u>**):

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

Estymatory - wariancja

- **Estymator wariancji wartości średniej <u>z próby</u>**:

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

- Zaś odpowiadające odchylenie standardowe (**estymator niepewność średniej <u>z próby</u>**):

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

- 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**):

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

  - wariancji populacji – wariancja z próby (aproksymowana):

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

  - 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

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

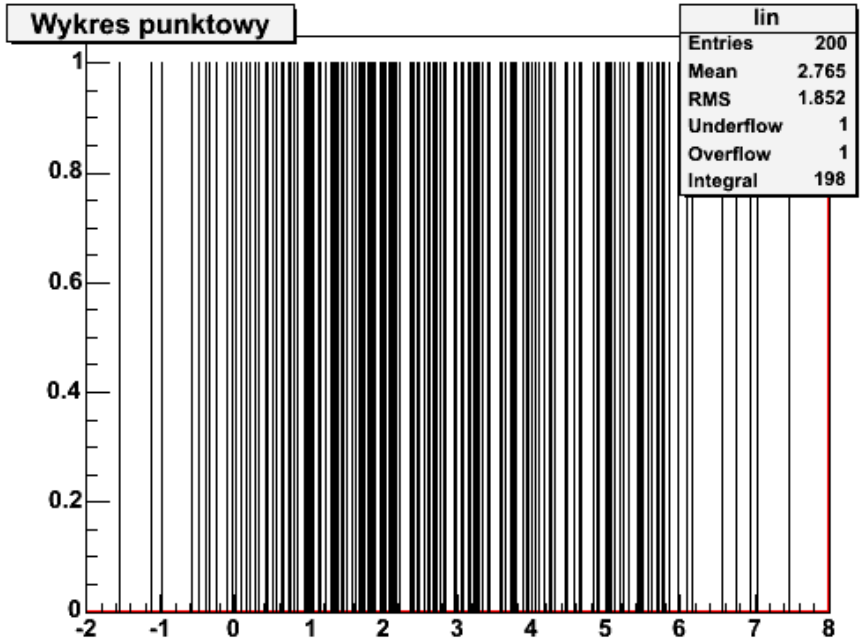
  - dalej możemy wyznaczać np. wariancję wariancji aproksymowanego estymatora wariancji próby i tak dalej (w nieskończoność)...

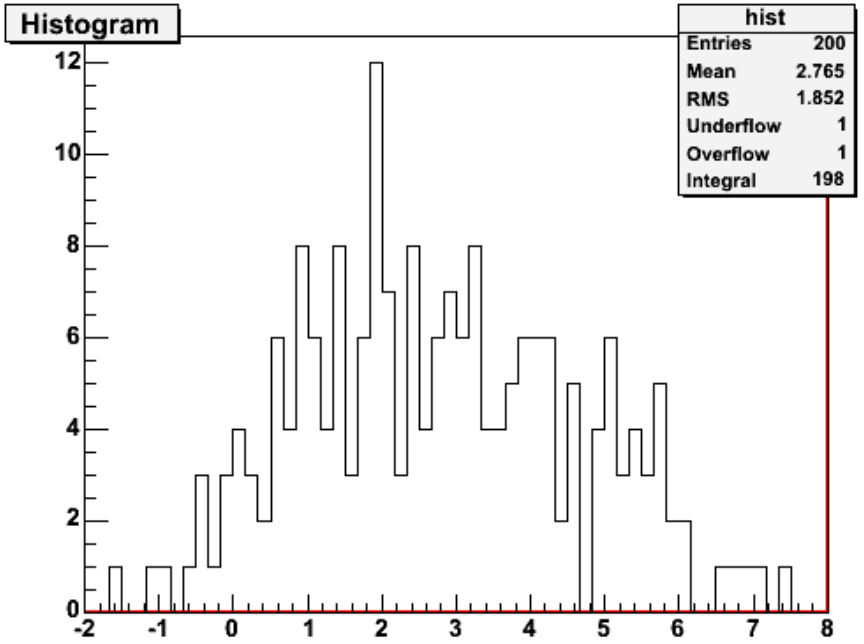
Graficzne przedstawienie próby

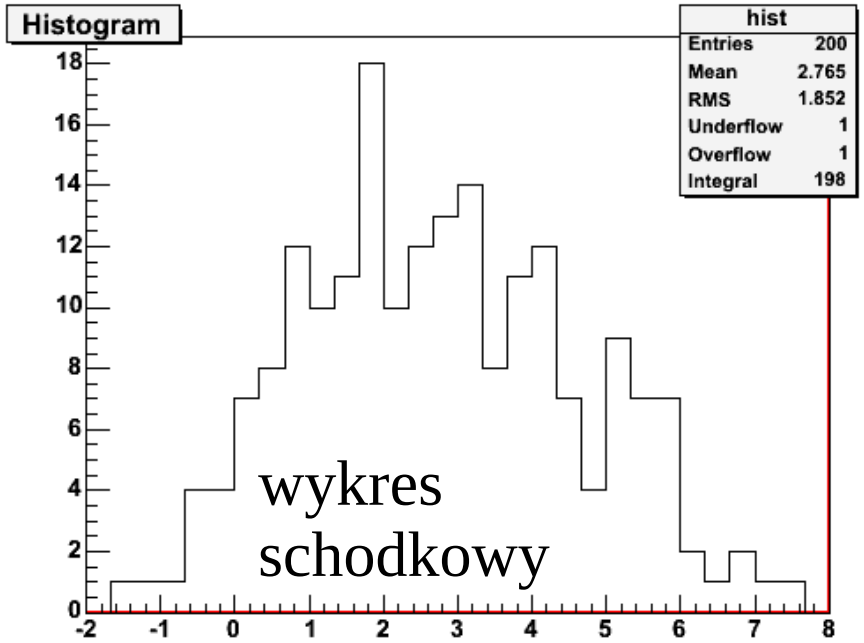
Graficzne przedstawienie próby

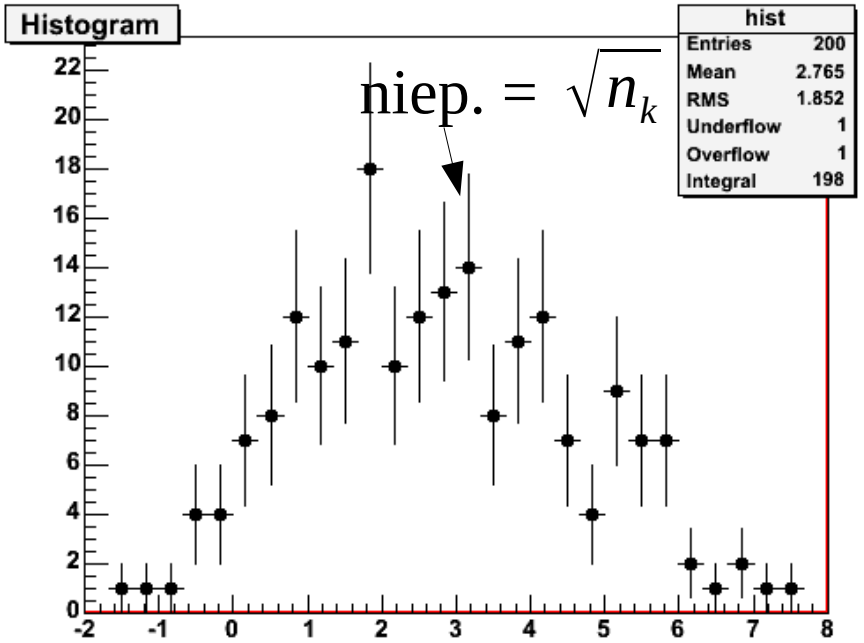
- Rozważmy próbę: $x_1, x_2, \ldots, x_n$, która zależy od jednej zmiennej losowej X
- Możemy tę próbę przedstawić jako wykres 1D – punkty na osi x – jednowymiarowy wykres punktowy
  - **wada:** co w przypadku, gdy mamy dwa takie same pomiary?
- Z reguły stosujemy zatem wykres 2D, zwany **histogramem**: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

Graficzne przedstawienie próby

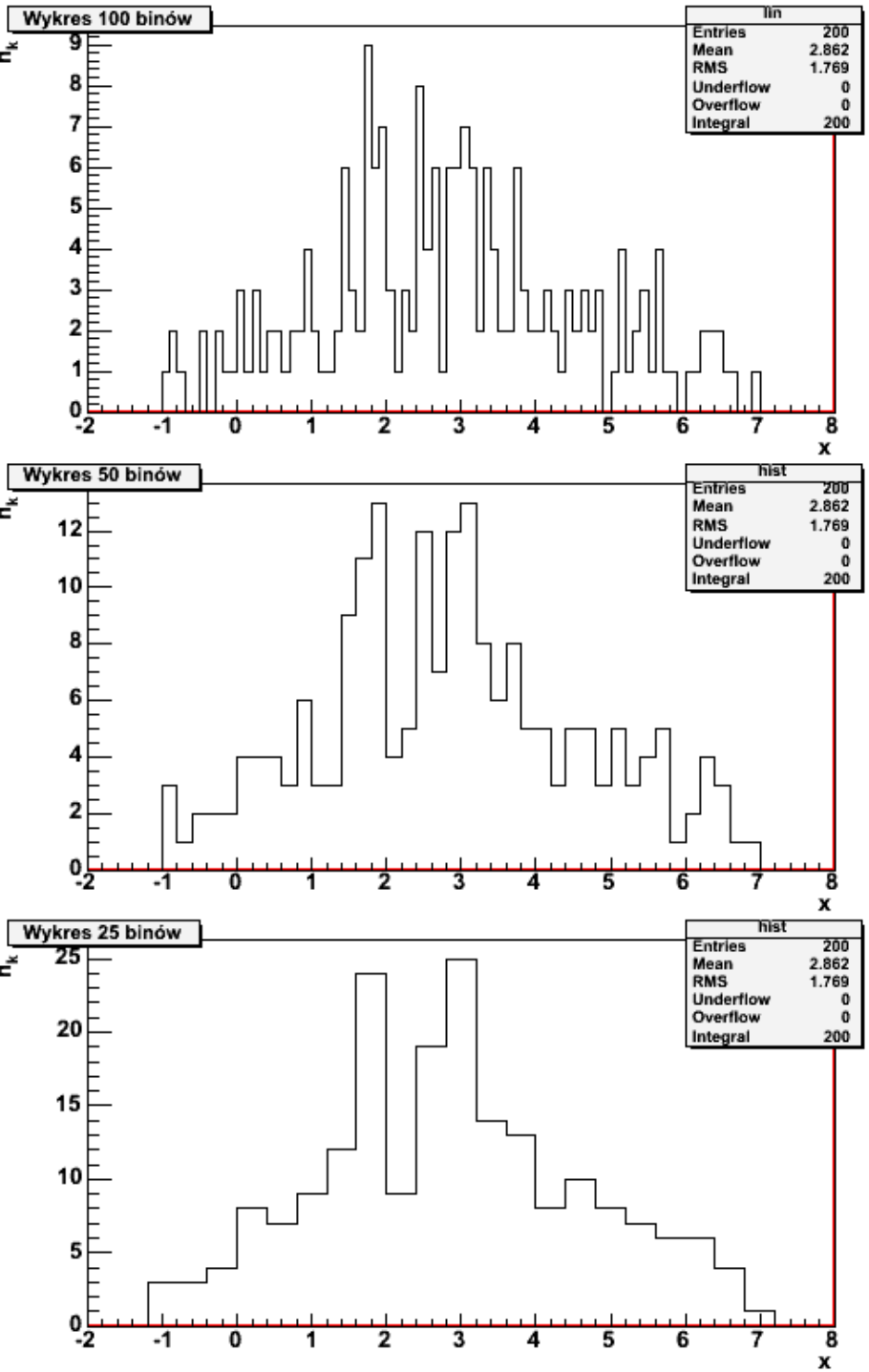








Histogram – szerokość przedziału



- Im więcej przedziałów, tym informacja o próbie jest dokładniejsza
- Większa ilość przedziałów powoduje jednak większe wahania statystyczne *od punktu do punku*
- 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 liczność (liczbę składników) *r*.
- Badamy zachowanie estymatorów charakterystyk rozkładu i estymatorów ich niepewności w funkcji liczności próby *r*

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

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

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

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)

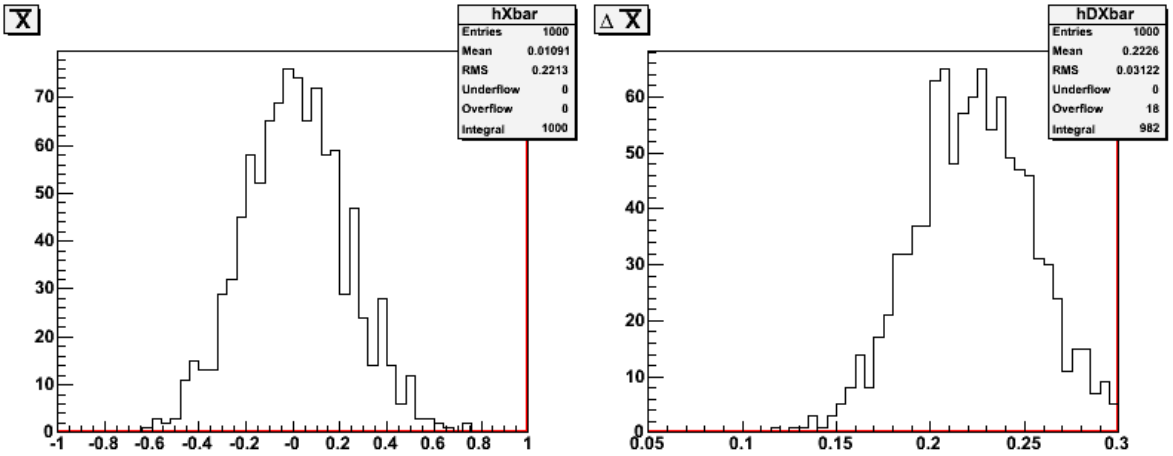
$$\sigma(S(X)) = \Delta S(X) = \frac{S(X)}{\sqrt{2(n-1)}}$$

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

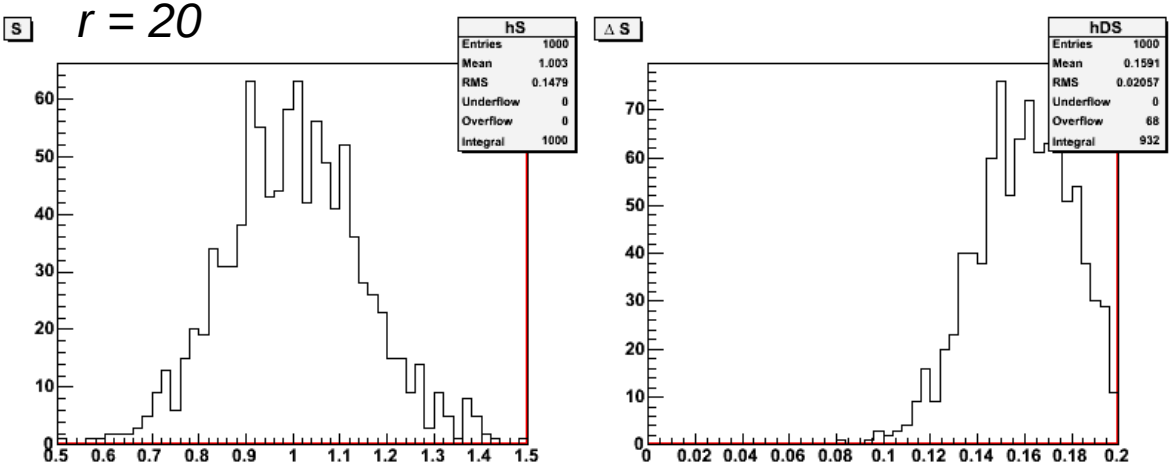
$$\sigma(S^2(X)) = \Delta S^2(X) = S^2(X)\sqrt{\frac{2}{n-1}}$$

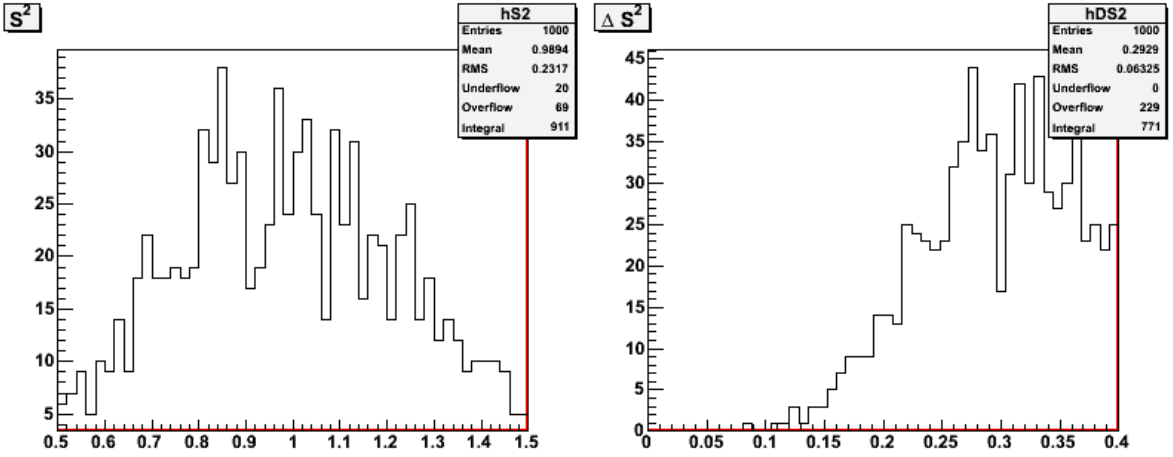
niepewność estymatora wariancji populacji – estymator odch. std. estymatora **wariancji** populacji

Estymatory - histogramy

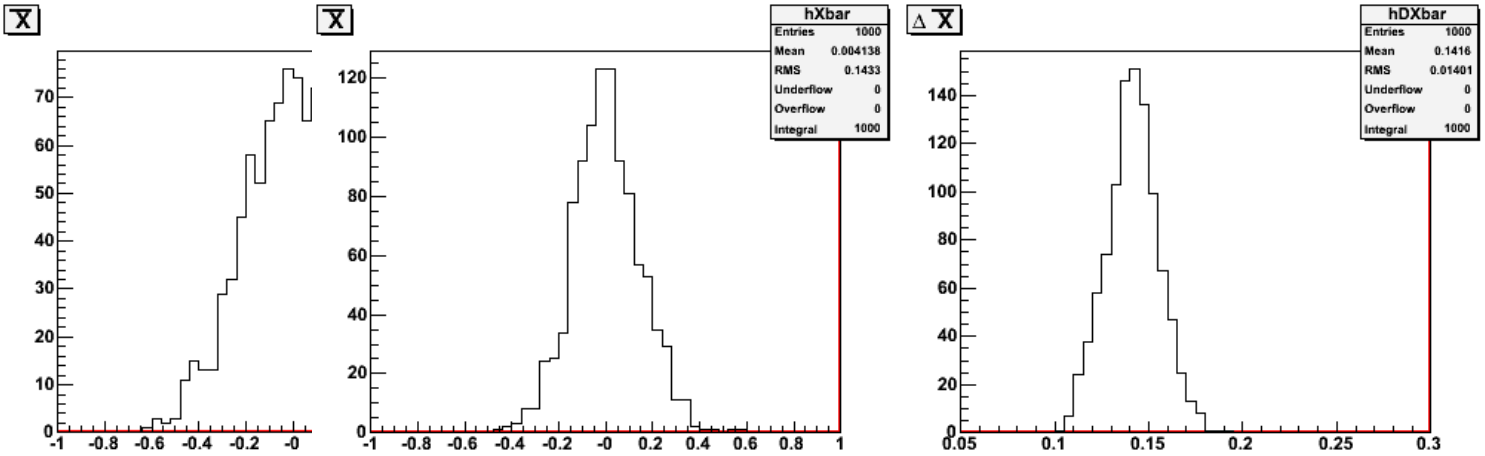


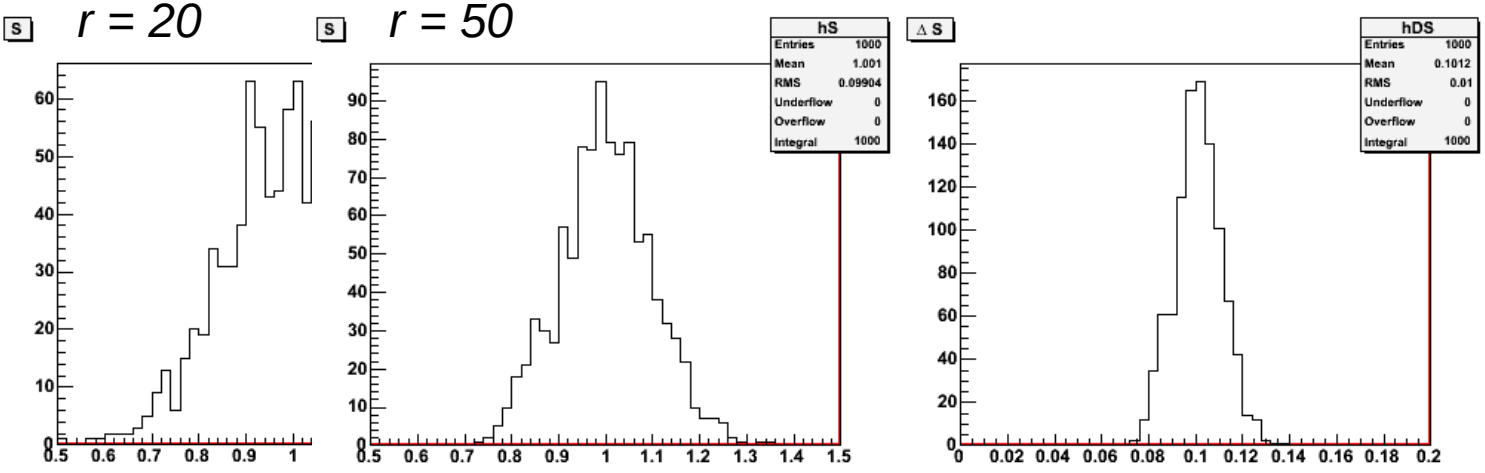
r = 20

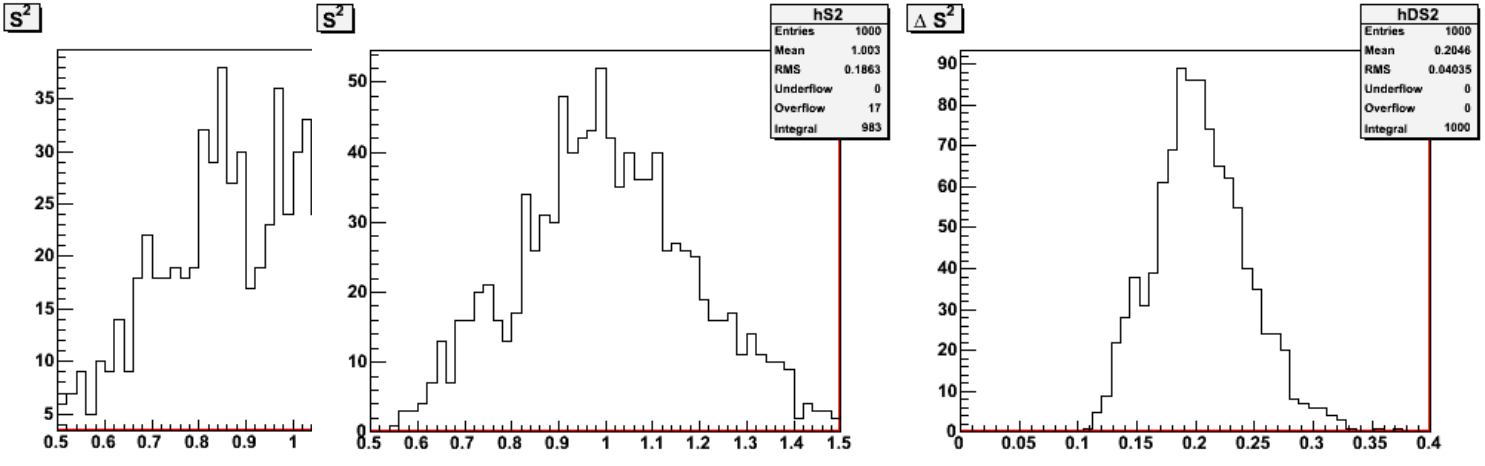




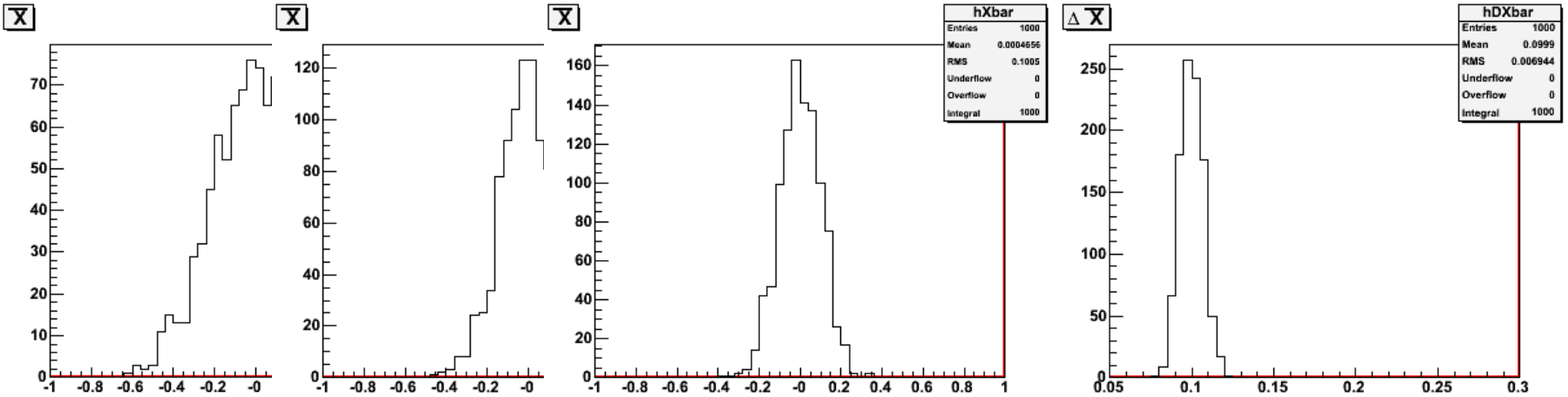
Estymatory - histogramy

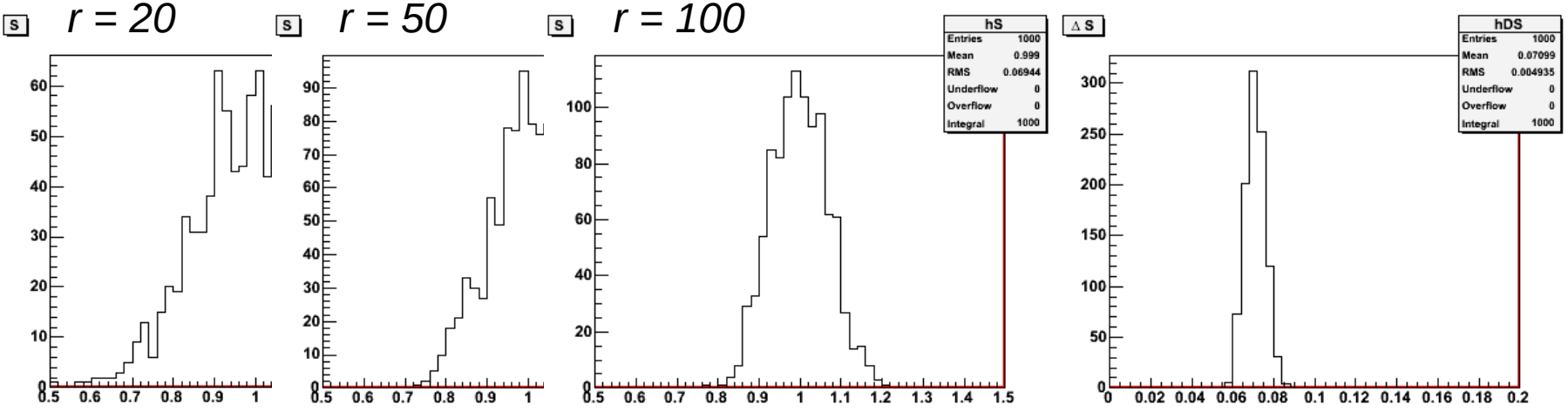


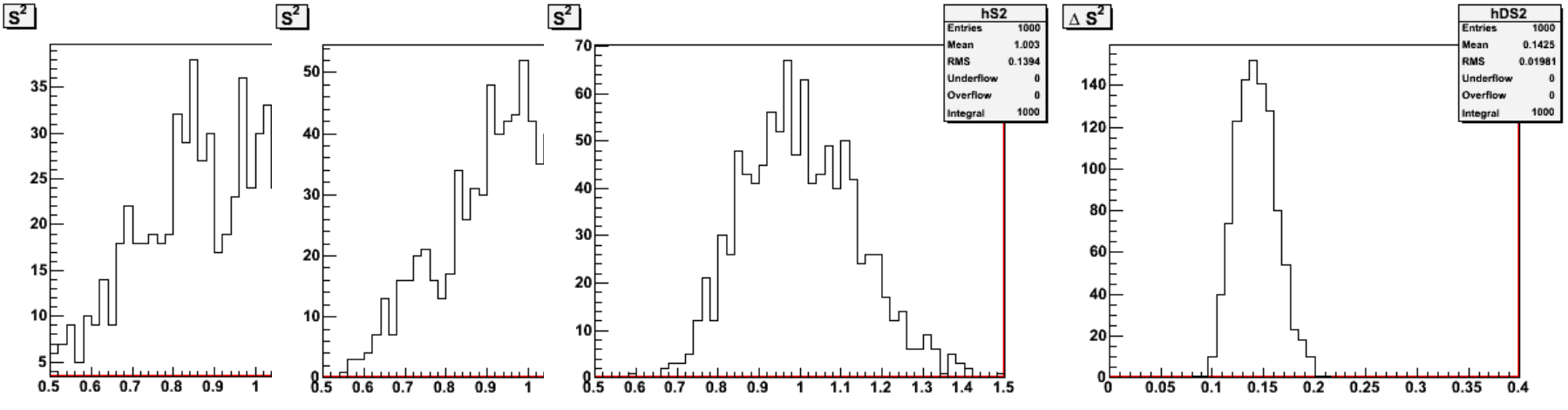




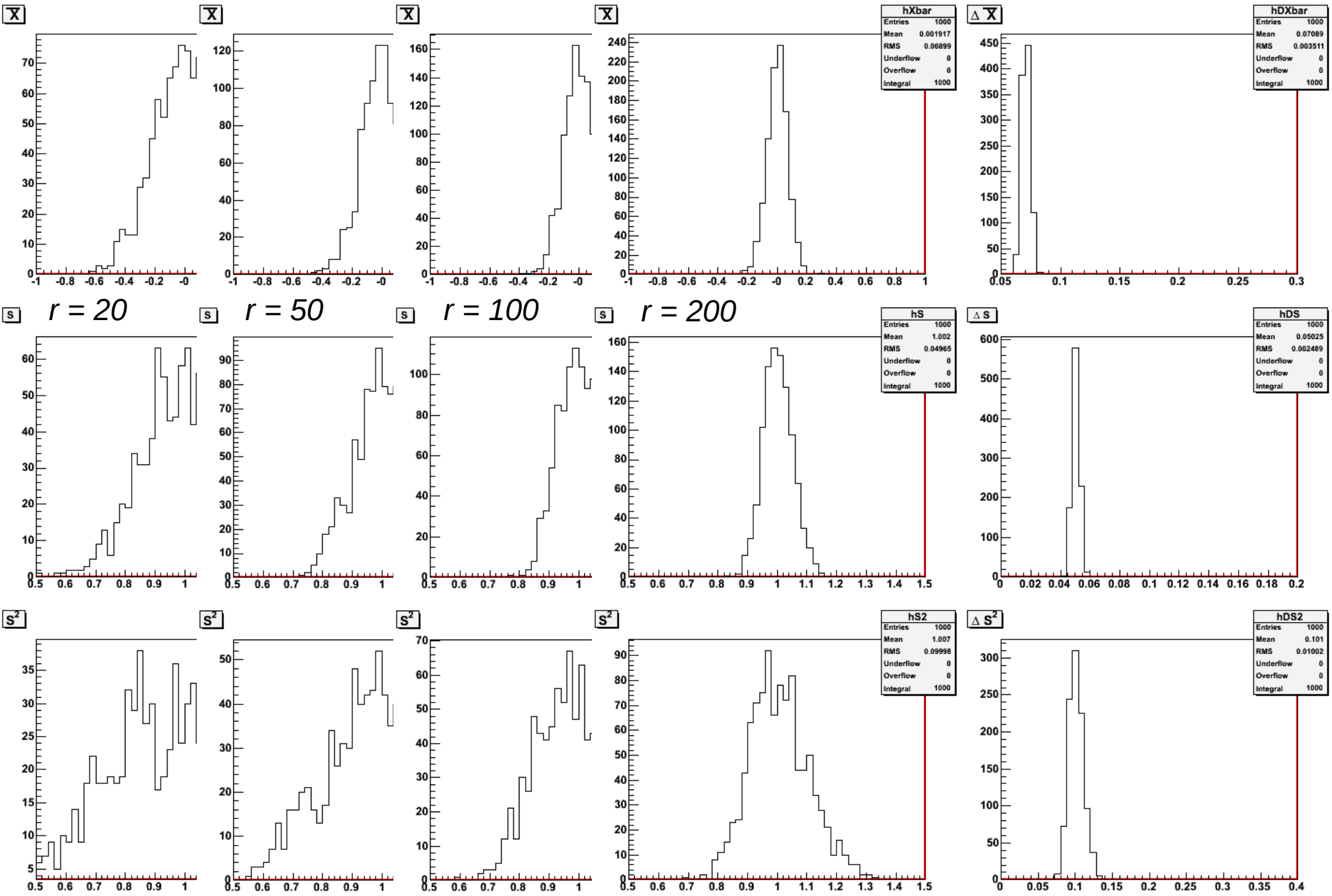
Estymatory - histogramy



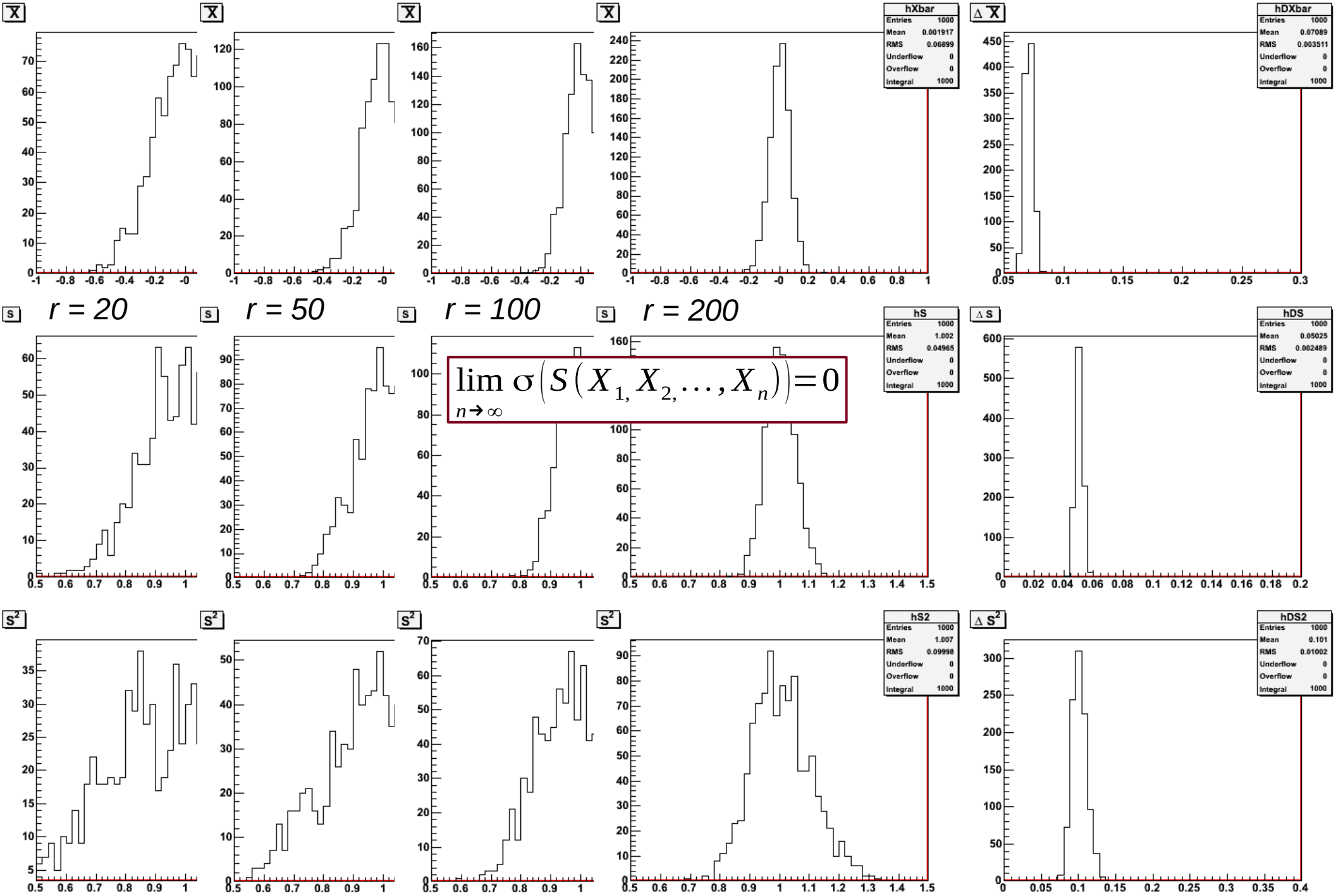




Estymatory - histogramy



Estymatory - histogramy



$$\lim_{n \to \infty} \sigma\left(S\left(X_1, X_2, \ldots, X_n\right)\right) = 0$$


WYDZIAŁ FIZYKI
POLITECHNIKA WARSZAWSKA
POLITECHNIKA WARSZAWSKA
WARSAW UNIVERSITY OF TECHNOLOGY

KONIEC