Komputerowa analiza danych doświadczalnych

Wykład 7
8.04.2016

dr inż. Łukasz Graczykowski
lgraczyk@if.pw.edu.pl

Semestr letni 2015/2016

Centralne twierdzenie graniczne - przypomnienie

Sploty

Pobieranie próby, estymatory

Centralne twierdzenie graniczne

Centralne twierdzenie graniczne

- Dlaczego rozkład normalny jest tak ważny w rachunku prawdopodobieństwa i statystyce?
- Mówi o tym **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:

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

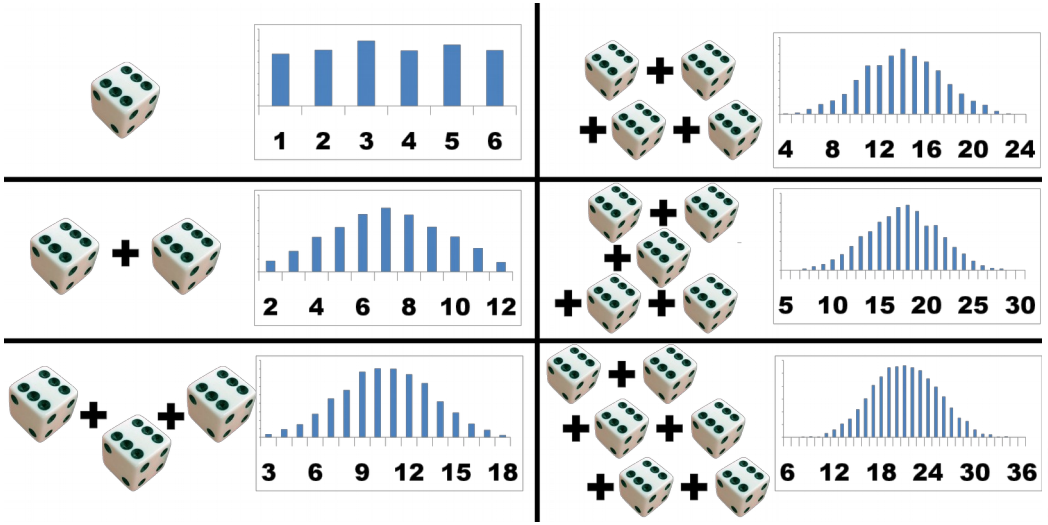
  - ponadto, zmienna $\xi = \frac{1}{n} X = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i$ ma rozkład normalny z:

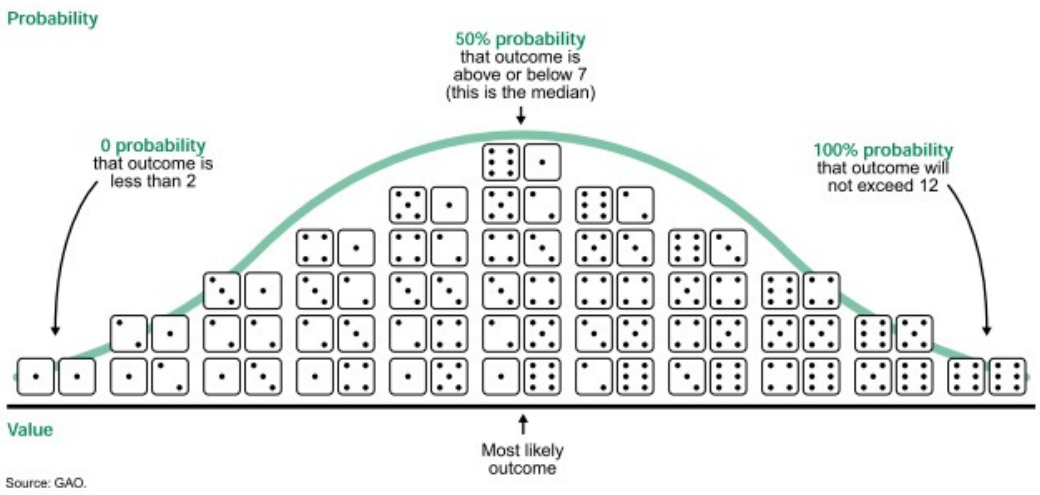
$$E(\xi) = a,\ \sigma^2(\xi) = b^2/n$$

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

Centralne twierdzenie graniczne - przykład 1

- Wyobraźmy sobie eksperyment polegający na rzucie kostką (kostkami) i obserwowaniu całkowitej liczby oczek:
  - kolejne rzuty kostką (kostkami) są niezależne
  - jeśli rzucamy kostką jednokrotnie (albo 1 kostką), to prawdopodobieństwo uzyskania danej wartości jest jednakowe
  - jeśli rzucamy kostką dwukrotnie (albo 2 kostkami), to prawdopodobieństwo uzyskania sumy oczek nie jest już jednakowe
  - jeśli rzucimy kostką n-krotnie (n-kostkami) → rozkład normalny





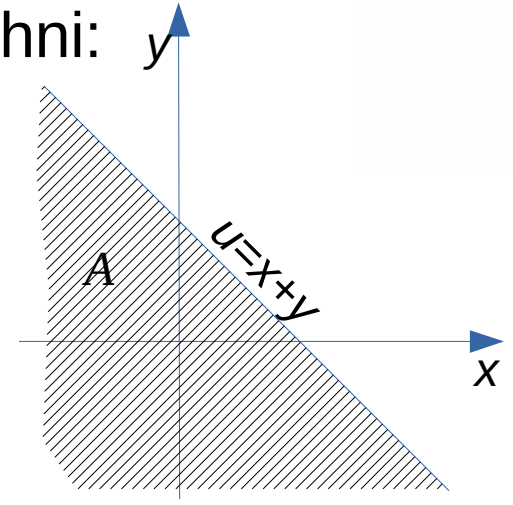

WYDZIAŁ FIZYKI
POLITECHNIKA WARSZAWSKA
POLITECHNIKA WARSZAWSKA
WARSAW UNIVERSITY OF TECHNOLOGY

Sploty

Suma zmiennych losowych jako splot

- W doświadczeniach eksperymentalnych bardzo często mamy do czynienia z sumą dwóch zmiennych losowych
- Na przykład – rozpad cząstek nietrwałych opisnany jest pewnym kątem rozpadu, wynikającym ze statystycznego charakteru zjawiska fizycznego, zaś niepewność jego pomiaru z niedokładnośni przyrządu. Obserwowany rozkład jest **splotem** dwóch rozkładów
- Rozważmy zmienną losową: $U = X + Y$
- Zakładamy niezależność zmiennych: $f(x,y) = f_x(x) f_y(y)$
- Wtedy dystrybuanta zmiennej U: może być wyznaczona jako pole powierzchni:

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



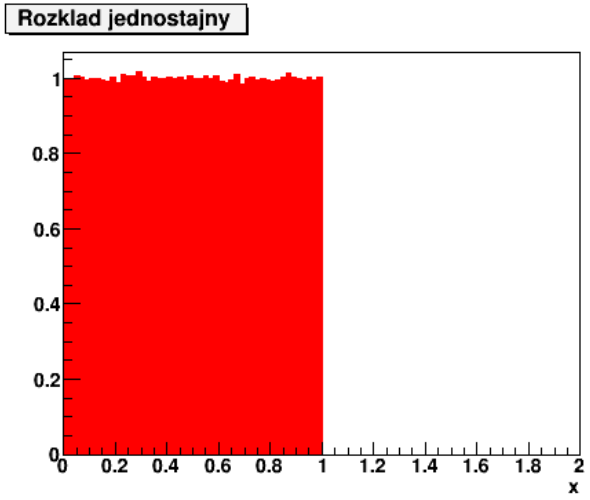
Suma zmiennych losowych jako splot

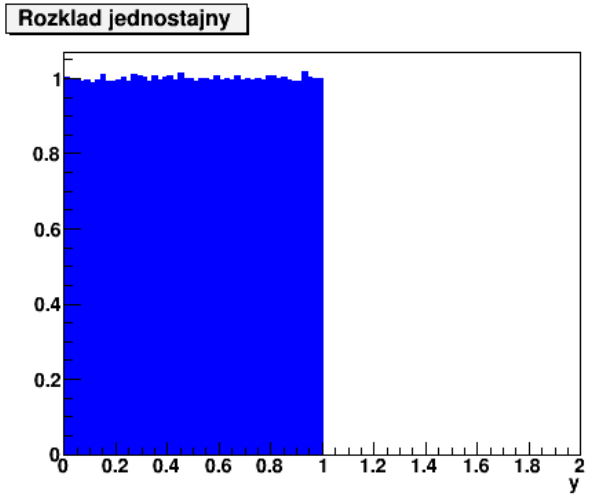
- Z dystrybuanty wyznaczamy funkcję gęstości zmiennej U:

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

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

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





$$f(u)=\int_0^1 f_x(x) f_y(u-x)\,dx=$$

$$=\int_0^1 f_y(u-x)$$

Suma zmiennych losowych jako splot

- Rozpatrzmy przypadek splotu dwóch rozkładów jednorodnych:

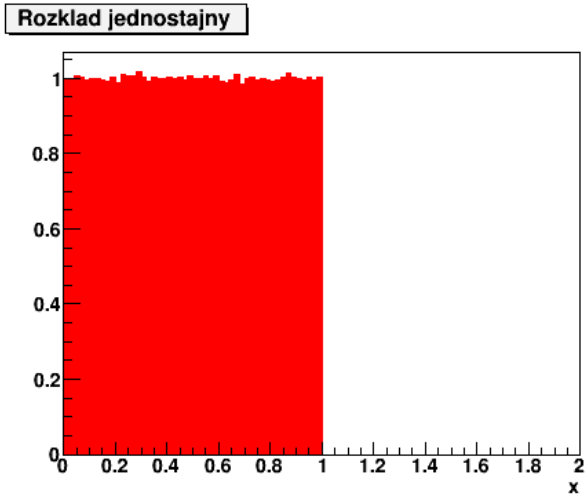
$$f_x(x)=\begin{cases}1, & 0\leqslant x<1\\ 0, & \text{w przeciwnym razie}\end{cases}\qquad f_y(y)=\begin{cases}1, & 0\leqslant y<1\\ 0, & \text{w przeciwnym razie}\end{cases}$$

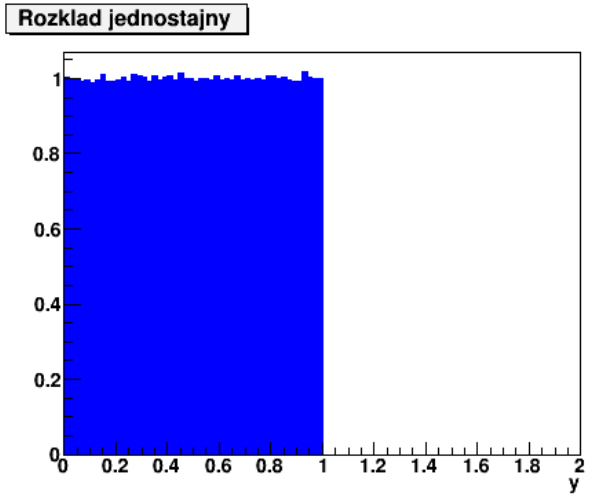
$$f(u)=\int_0^1 f_x(x)f_y(u-x)\,dx=\int_0^1 f_y(u-x)\quad \begin{matrix}v=u-x\\ dv=-dx\end{matrix}\Rightarrow\quad f(u)=-\int_u^{u-1} f_y(v)\,dv=\int_{u-1}^{u} f_y(v)\,dv$$

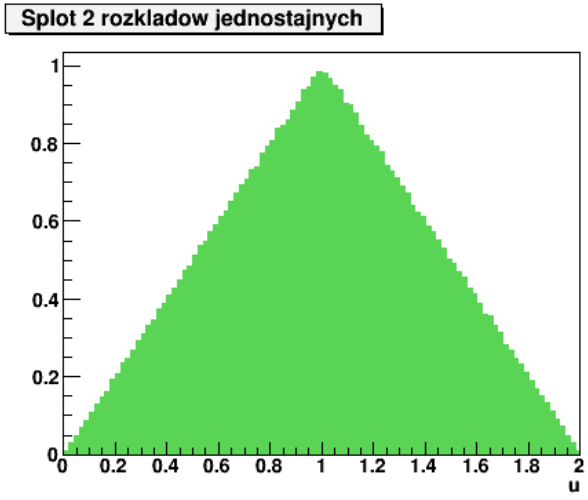
- Zmienna u zmienia się od 0 do 2, zatem rozważmy 2 przypadki:

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

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







Suma zmiennych losowych jako splot

- Rozpatrzmy przypadek splotu dwóch rozkładów jednorodnych:

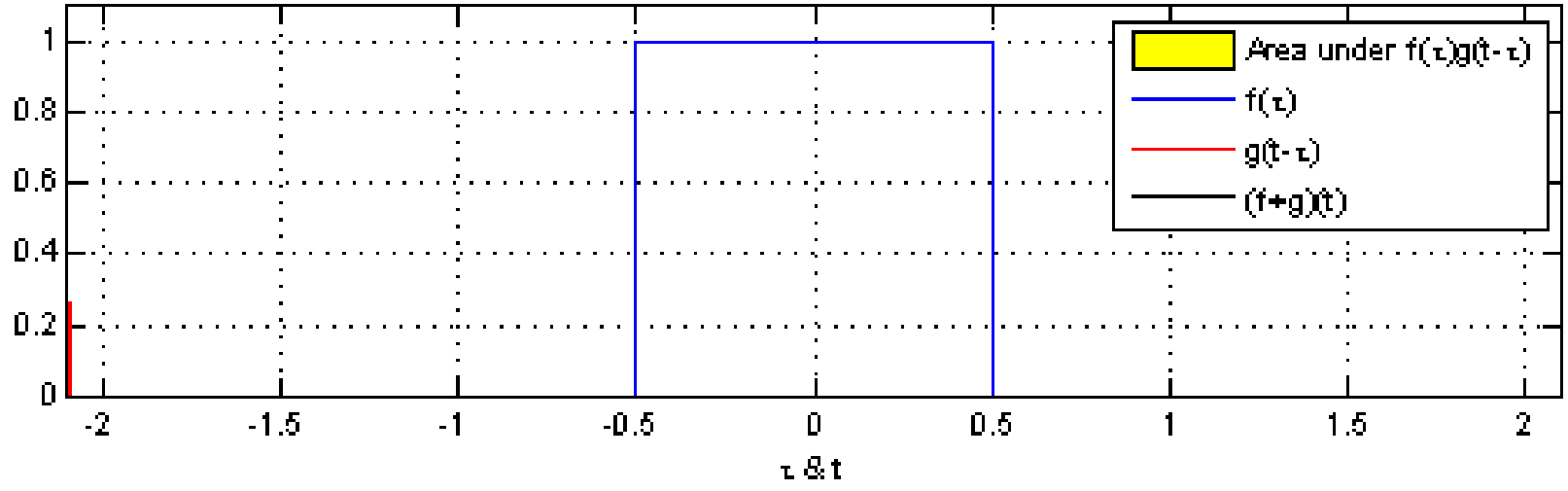
$$f_x(x)=\begin{cases}1, & 0\leqslant x<1\\ 0, & \text{w przeciwnym razie}\end{cases}\qquad f_y(y)=\begin{cases}1, & 0\leqslant y<1\\ 0, & \text{w przeciwnym razie}\end{cases}$$

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

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

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

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

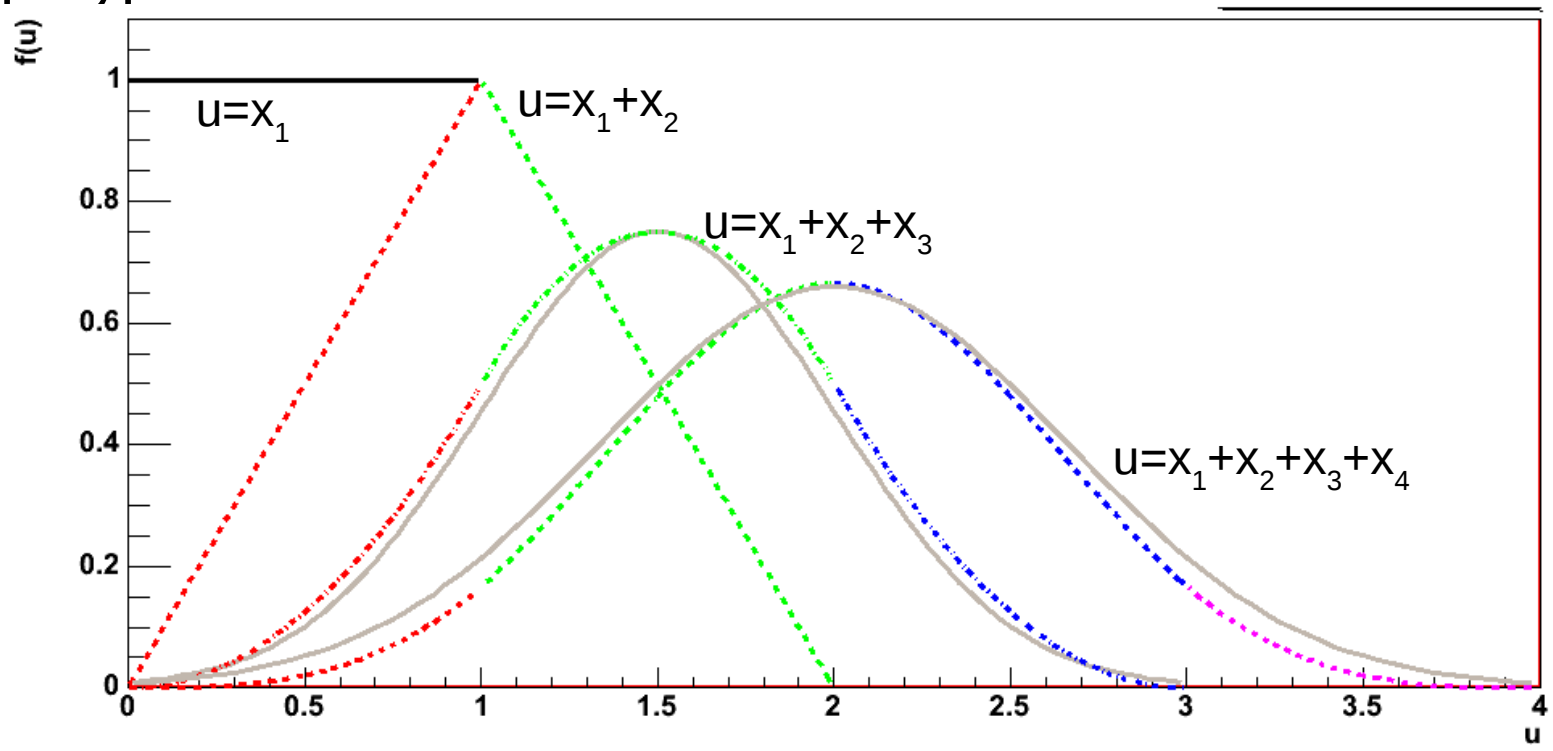


Suma zmiennych losowych jako splot

- 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:

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

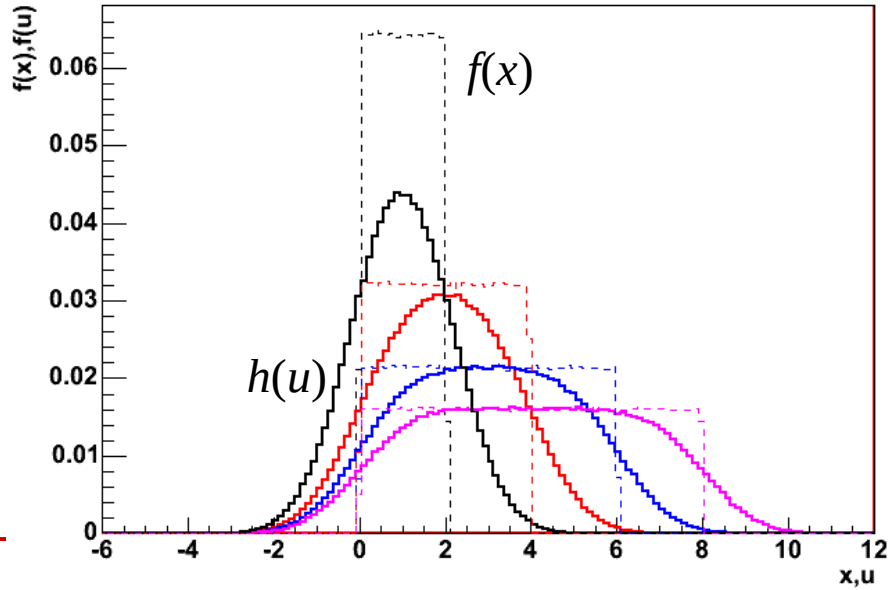
$$f(x)=0;\ x\in\mathbb{R}\setminus\langle a,b\rangle$$

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

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

- Zaś uwzględniając dystrybuantę rozkładu normalnego:

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

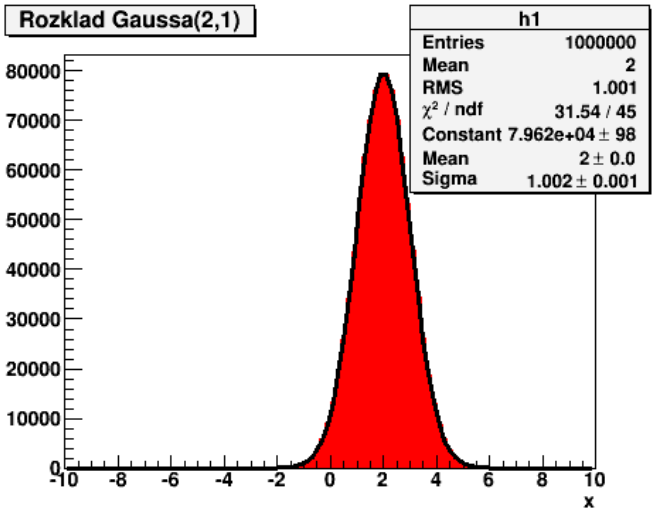


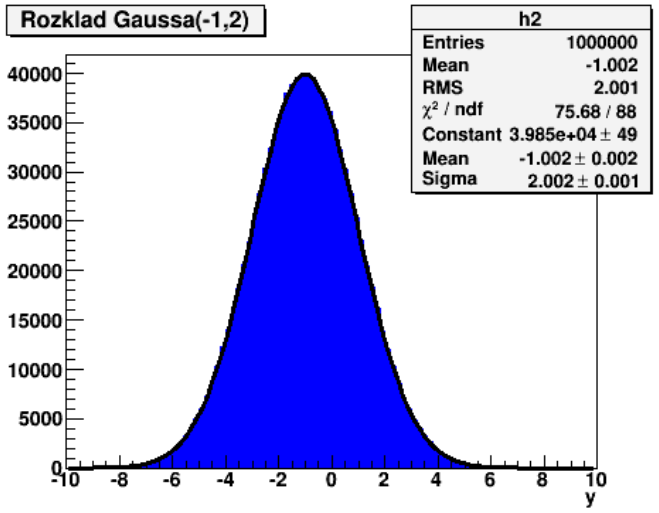
Sploty z rozkładem normalnym - przykład 2

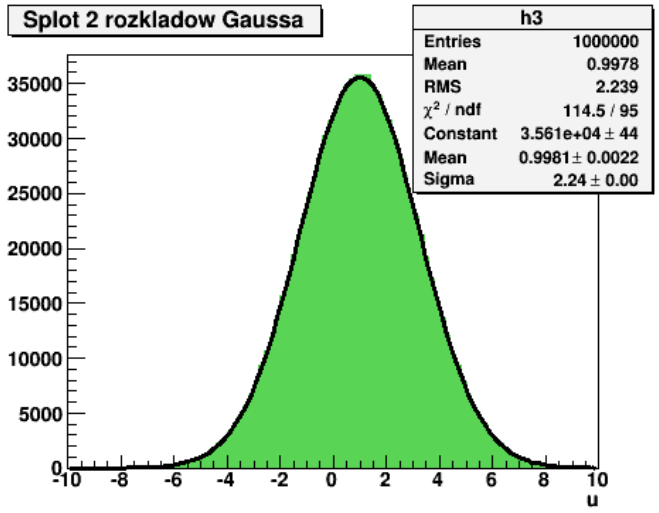
- Przykład: Splot dwóch rozkładów normalnych – dodawanie niepewności “w kwadracie”
- Splot dwóch rozkładów normalnych o wartościach średnich równych 0 i warniancjach σ_x, σ_y ma postać rozkładu normalnego:

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

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



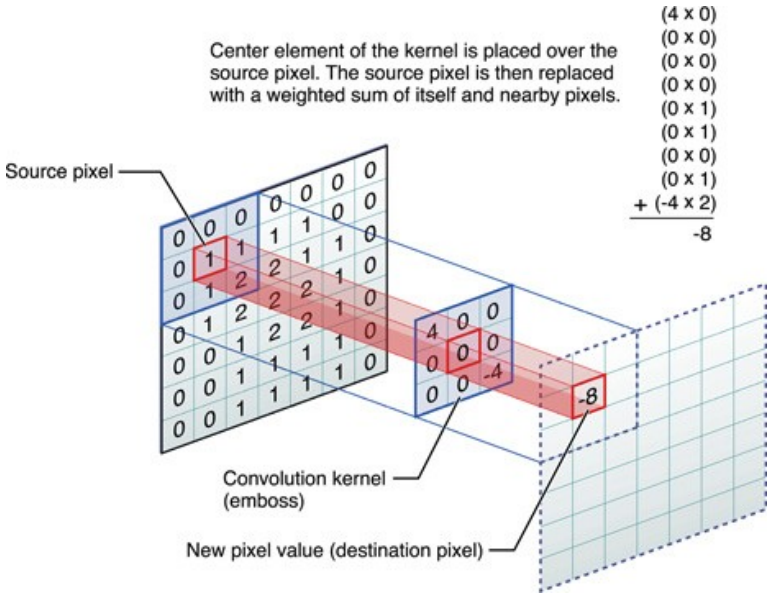




Zastosowanie splotów

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

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



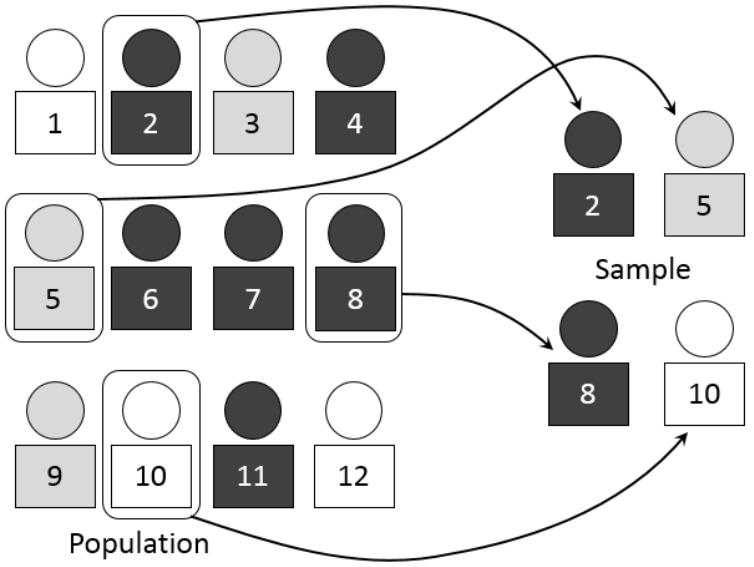
Original

Emboss

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



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:

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

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

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

Pobieranie próby

- Aby można było mówić o losowym pobieraniu próby:
  - zmienne X_i muszą być niezależne, czyli: $g(\mathbf{x}) = g_1(x_1) g_1(x_2) \dots g_n(x_n)$
  - 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 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)
- Teraz zdefiniujemy pojęcia, które charakteryzują próbę losową:
  - załóżmy, że mamy n-elementową próbę i odkładamy wyniki na osi liczb. Przez n_x oznaczmy taką liczbę wartości, które są mniejsze niż pewna stała x, czyli mamy spełnioną definicję dystrybuanty: $X \leqslant x$
  - wielkość $W_n(x) = n_x / n$ nazywamy **dystrybuantą empiryczną**
  - jest to funkcja schodkowa zwiększająca się o $1/n$ dla każdej kolejnej wartości z próby; dla dużych n dąży do dystrybuanty

Pobieranie próby

- Teraz zdefiniujemy pojęcia, które charakteryzują próbę losową:
  - załóżmy, że mamy n-elementową próbę i odkładamy wyniki na osi liczb. Przez n_x oznaczmy taką liczbę wartości, które są mniejsze niż pewna stała x, czyli mamy spełnioną definicję dystrybuanty: $X \leqslant x$
  - wielkość $W_n(x) = n_x / n$ nazywamy **dystrybuantą empiryczną**
  - jest to funkcja schodkowa zwiększająca się o $1/n$ dla każdej kolejnej wartości z próby; dla dużych n dąży do dystrybuanty
  - funkcję elementów próby (czyli zmiennej losowej X) nazywamy **statystyką**
  - najważniejszym przykładem statystyki jest **średnia z próby** (*ang. sample mean*) zdefiniowana jako średnia z elementów próby:

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

Pobieranie próby - przykład

- Przykład – wzrost Polaków
- Niewątpliwie wzrost Polaków (zmienna losowa X) podlega pewnemu rozkładowi $f(x)$ z dysteybuantą $F(x)$
- Pomiar wzrostu pojedynczego Polaka daje wartość x
- Jeżeli stworzymy n-wymiarową próbę losową, tzn. wybierzemy n Polaków, to rozkład prawdopodobieństwa dla każdej z osób (od $g_1(x_1)$ do $g_n(x_n)$) jest taki sam jak dla całej populacji i równy $f(x)$
- Dla każdej tak skonstruowanej próby możemy teraz policzyć jej $W_n(x)$. Oczywiście im większe będzie n, im więcej ludzi weźmiemy do naszej próby, tym rozkład wyliczony z próby będzie bliższy rozkładowi rzeczywiście istniejącemu w populacji
- **Zadaniem estymacji** jest znalezienie takiej statystyki (a więc funkcji określonej na wektorze $\boldsymbol{X}$), 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:

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

- 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:

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

- 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 podstawimy ten wzór do wzoru: $\sigma^2(\bar{X})=\frac{1}{n}\sigma^2(X)$
- To otrzymamy **estymator wariancji wartości średniej**:

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

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

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

- Jaka jest zaś **niepewność wariancji z próby** (bez wyprowadzenia)?

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

- Odchylenie standardowe próby: $S=\sqrt{S^2}=\frac{1}{\sqrt{n-1}}\sqrt{\sum_{i=1}^{n}(X_i-\hat{X})^2}$

Estymatory - wariancja

- Podsumowując zatem **estymatory nieobciążone**:
  - wartości oczekiwanej populacji → średnia z próby (**wynik doświadczenia**):

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

  - 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

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

  - odchylenia standardowego próby:

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

  - dalej możemy wyznaczać np. wariancję odchylenia std. próby...


WYDZIAŁ FIZYKI
POLITECHNIKA WARSZAWSKA
POLITECHNIKA WARSZAWSKA
WARSAW UNIVERSITY OF TECHNOLOGY

KONIEC