Komputerowa Analiza Danych Doświadczalnych

Prowadząca:
dr inż. Hanna Zbroszczyk

e-mail: *gos@if.pw.edu.pl*
tel: +48 22 234 58 51
konsultacje: poniedziałek, 10-11, środa: 11-12

www: http://www.if.pw.edu.pl/~gos/students/kadd

Politechnika Warszawska
Wydział Fizyki
Pok. 117b (wejście przez 115)

WIELOWYMIAROWY ROZKŁAD NORMALNY

Wielowymiarowy rozkład normalny

Wektor **zmiennych losowych**: $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$

Gęstość prawdopodobieństwa rozkładu normalnego:

$$\phi(\boldsymbol{x}) = k \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{a})^T B(\boldsymbol{x}-\boldsymbol{a})\right\} = k \exp\left\{-\frac{1}{2} g(\boldsymbol{x})\right\}$$

a to n-wymiarowy wektor wartości oczekiwanych.

Można dowieść, że $E\left\{(\boldsymbol{x}-\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{a})^T\right\} B = I$

czyli: $C = B^{-1}$ C jest macierzą kowariancji.

Zmienne 2- wymiarowych:

$$C = B^{-1} = \begin{bmatrix} \sigma_1^2 & cov(x_1, x_2) \\ cov(x_1, x_2) & \sigma_2^2 \end{bmatrix}$$

$$B = \frac{1}{\sigma_1^2 \sigma_2^2 - cov(x_1, x_2)^2} \begin{bmatrix} \sigma_2^2 & -cov(x_1, x_2) \\ -cov(x_1, x_2) & \sigma_1^2 \end{bmatrix}$$

Wielowymiarowy rozkład normalny

Dla znikających kowariancji: $B_0 = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}$

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

Zmienne losowe zredukowane: $u_i = \frac{x_i - a_i}{\sigma_i}$

Współczynnik korelacji: $\rho = cov\,\frac{(x_1, x_2)}{\sigma_1\sigma_2} = cov\,(u_1, u_2)$

$$\phi(u_1, u_2) = k \exp\left\{-\frac{1}{2}\boldsymbol{u}^T B\,\boldsymbol{u}\right\} = k \exp\left\{-\frac{1}{2} g(\boldsymbol{u})\right\}$$

$$B = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

Wielowymiarowy rozkład normalny

Równanie elipsy o środku w a_1 i a_2: $$\frac{(x_1-a_1)^2}{\sigma_1^2}-2\rho\frac{x_1-a_1}{\sigma_1}\frac{x_2-a_2}{\sigma_2}+\frac{(x_2-a_2)^2}{\sigma_2^2}=1-\rho^2$$

Osie główne elipsy tworzą kąt α z osiami x_1 oraz x_2.

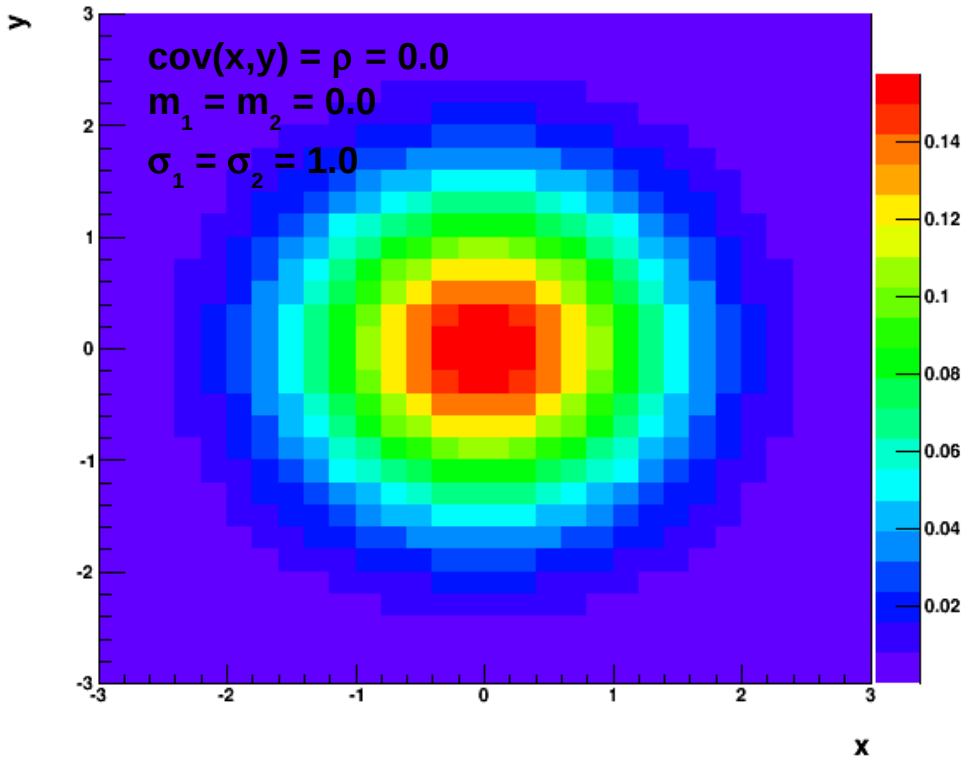
$$tg\,2\alpha=\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2-\sigma_2^2}$$

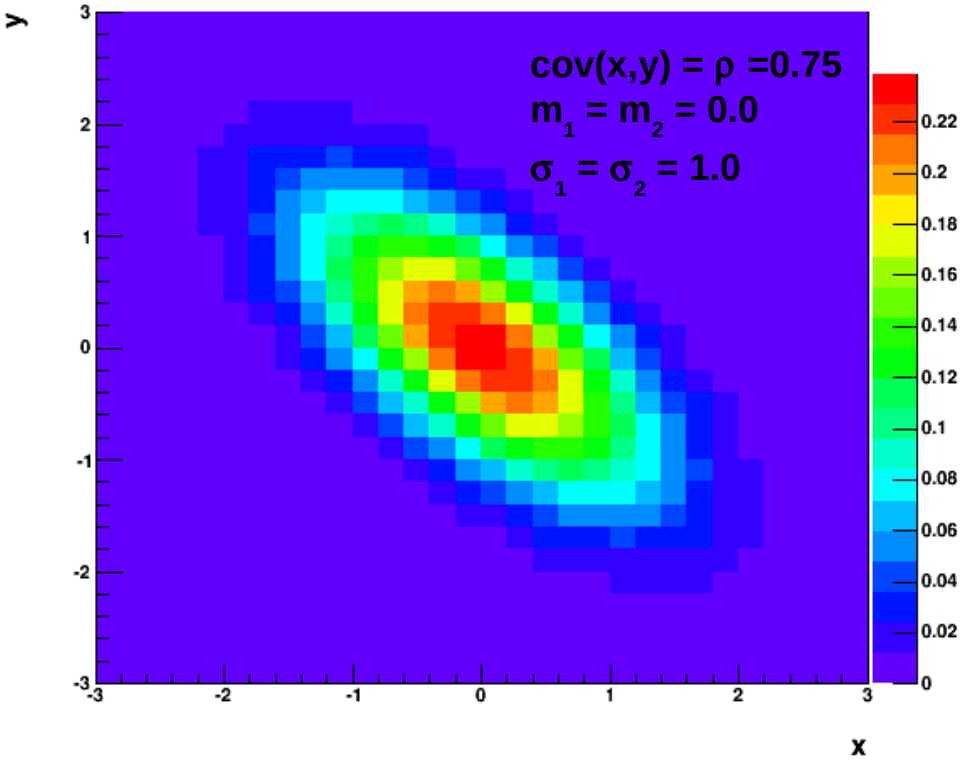
Półosie elipsy:

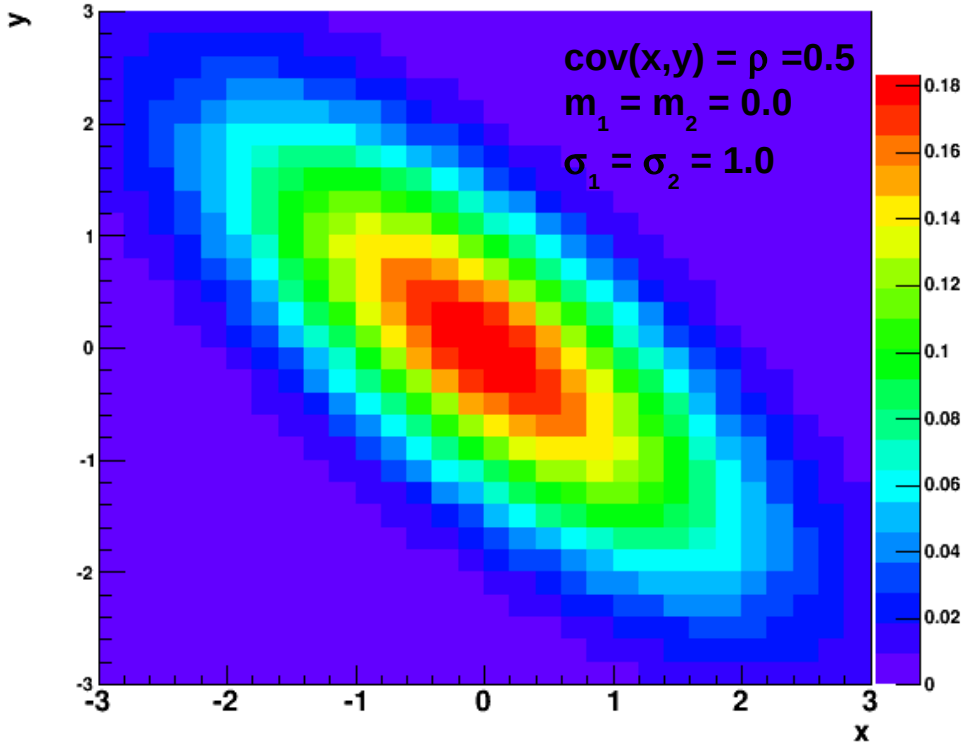
$$p_1^2=\frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_2^2\cos^2\alpha-2\rho\sigma_1\sigma_2\sin\alpha\cos\alpha+\sigma_1^2\sin^2\alpha}$$

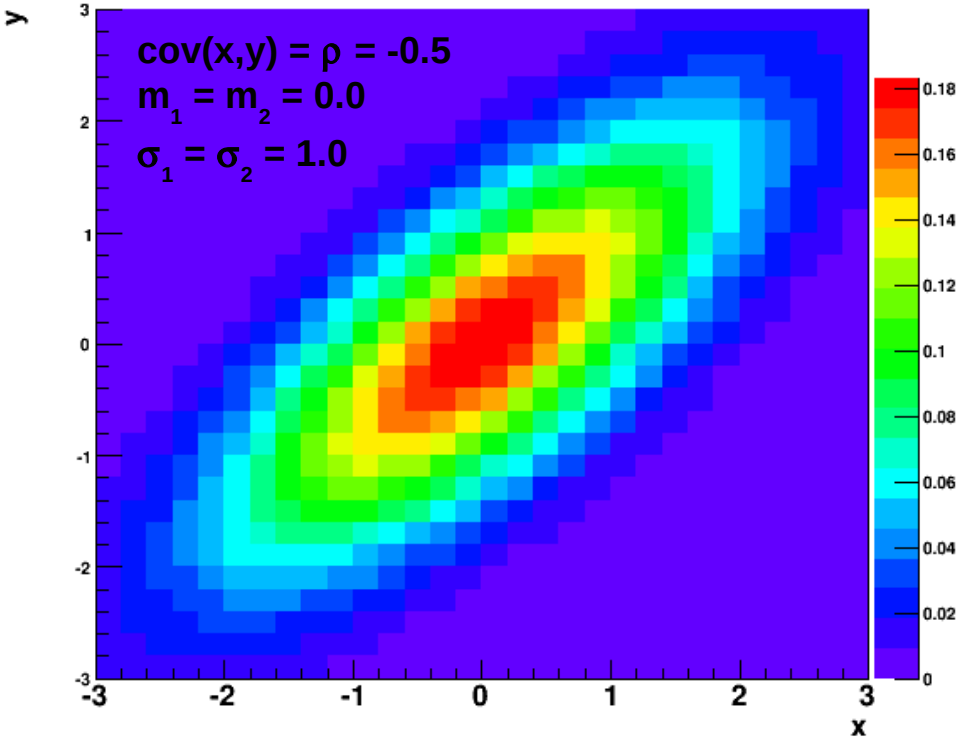
$$p_2^2=\frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_2^2\sin^2\alpha-2\rho\sigma_1\sigma_2\sin\alpha\cos\alpha+\sigma_1^2\cos^2\alpha}$$

Elipsa o takich parametrach to **elipsa kowariancji** (dla 2D rozkładu normalnego).

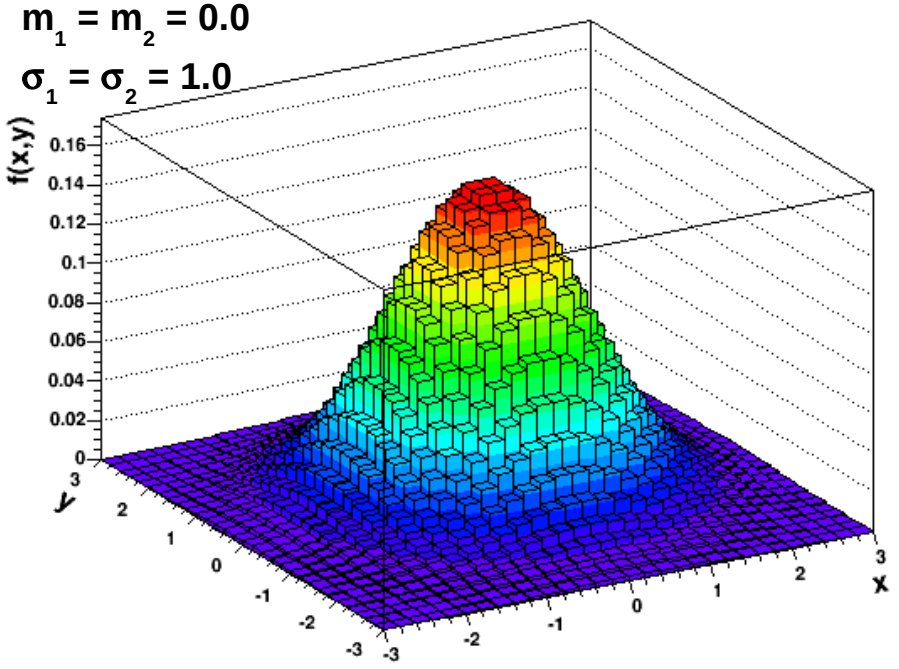

cov(x,y) = ρ = 0.0
m1 = m2 = 0.0
σ1 = σ2 = 1.0
y
x


cov(x,y) = ρ =0.75
m1 = m2 = 0.0
σ1 = σ2 = 1.0
y
x

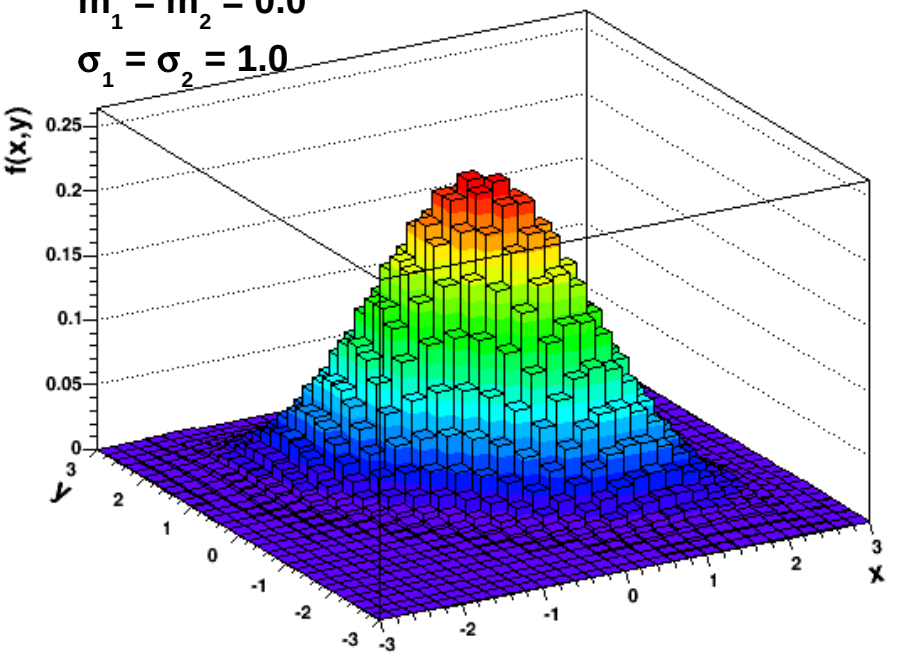

cov(x,y) = ρ =0.5
m1 = m2 = 0.0
σ1 = σ2 = 1.0
y
x


cov(x,y) = ρ = -0.5
m1 = m2 = 0.0
σ1 = σ2 = 1.0
y
x

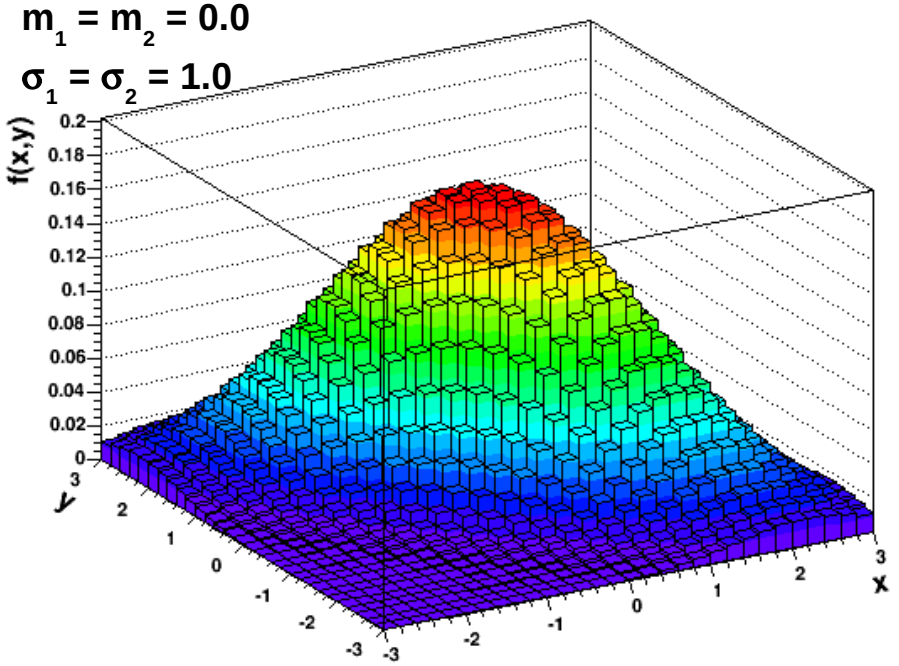
cov(x,y)=0.0
$m_1 = m_2 = 0.0$
$\sigma_1 = \sigma_2 = 1.0$



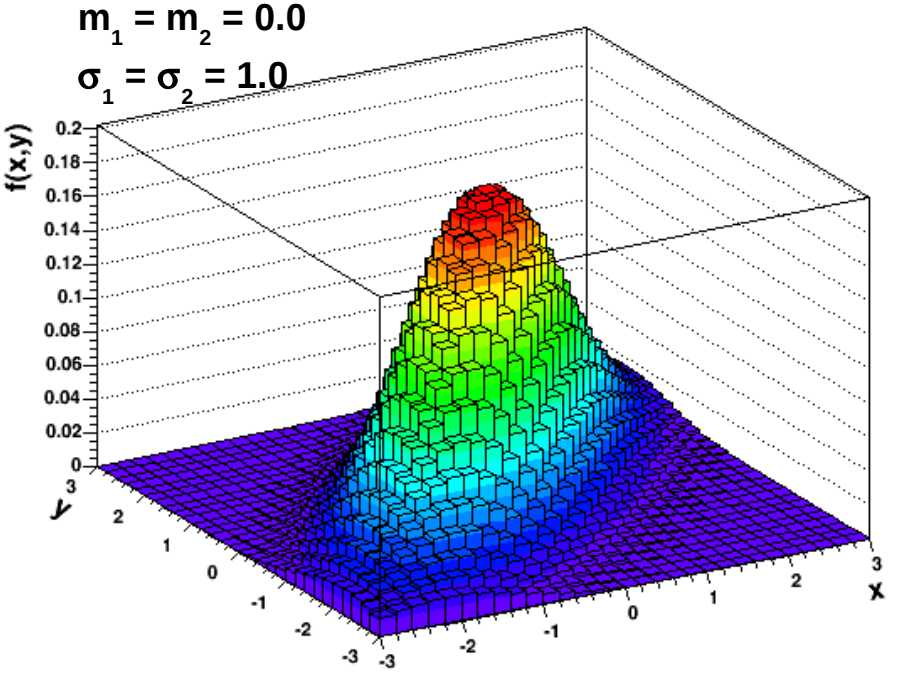
cov(x,y)=0.75
$m_1 = m_2 = 0.0$
$\sigma_1 = \sigma_2 = 1.0$



cov(x,y)=0.5
$m_1 = m_2 = 0.0$
$\sigma_1 = \sigma_2 = 1.0$



cov(x,y)=-0.5
$m_1 = m_2 = 0.0$
$\sigma_1 = \sigma_2 = 1.0$

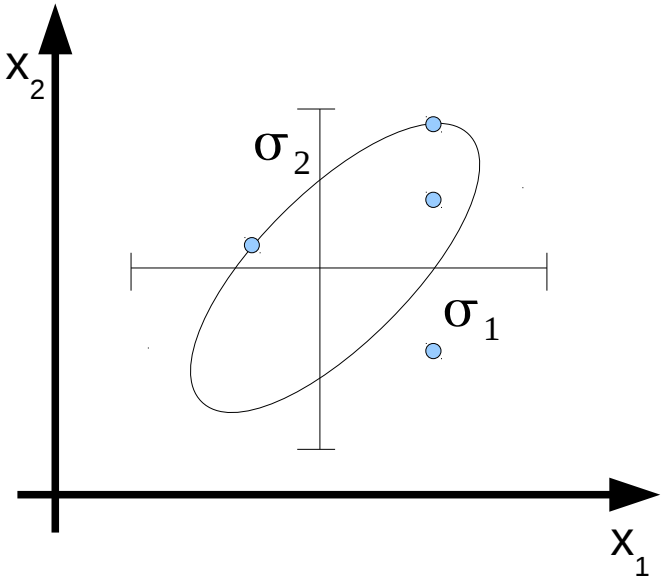


Wielowymiarowy rozkład normalny

Linie stałego prawdopodobieństwa to elipsy (wzajemnie koncentryczne, leżą wewnątrz lub na zewnątrz elipsy kowariancji).

W rozkładzie błędów dla **zmiennej 1D**, (przy obszarze całkowania) $a-\sigma \le x \le a+\sigma$, (niezależnie od wielkości σ) była **stała wartość**.

Dla zmiennej **2D**, należy uwzględnić wartości: σ_1, σ_2, ρ ,
(**nie prostokąt**)



Zmienne **3D** –
elipsoida kowariancji.

Zmiennych **nD** –
hiperelipsoida w przestrzeni n-wymiarowej.

Wielowymiarowy rozkład normalny

Elipsoida kowariancji - hiperpowierzchnia w przestrzeni n-wymiarowej o równaniu g(**x**) = 1.

Dla innych wartości **g(x) = const** - inne elipsoidy (wewnątrz - g<1 lub na zewnątrz - g>1 elipsoidy kowariancji.

Można wykazać, że jeśli **x** ma rozkład normalny, to g(**x**) ma rozkład χ^2 o **n** stopniach swobody.

Prawdopodobieństwo wystąpienia wartości wektora **x** wewnątrz elipsoidy **g = const:**

$$W = \int_0^g f(X^2; n)\, d\, X^2 = P(\frac{n}{2}, \frac{g}{2})$$

P to niepełna funkcja Gamma.

Dla elipsoidy kowariancji:

$$W_n = P(\frac{n}{2}, \frac{1}{2})$$

Dla małych n:

$W_1 = 0.68269$; $W_2 = 0.39347$; $W_3 = 0.19875$;

$W_4 = 0.09020$; $W_5 = 0.03734$; $W_6 = 0.01439$;

Wielowymiarowy rozkład normalny

Aby **wyznaczyć obszary, dla różnych n o tym samym prawdopodobieństwie**:

- ustalamy wartość W

- obliczyć odpowiadającą mu wartość g.

- g staje się kwantylem z prawdopodobieństwem w rozkładzie X^2 o n stopniach swobody. $g = X^2_W(n)$

Elipsoida odpowiadająca wartości g, że zawiera wektor **x** z prawdopodobieństwem W - **elipsoida ufności**.

W= 0.9 → wektor x leży wewnątrz elipsoidy ufności z prawdopodobieństwem 90%.

Wariancje, odchylenia standardowe w przypadku N-wymiarowym.

Prawdopodobieństwo uzyskania wartości zmiennej losowej $\boldsymbol{x}_i$ z przedziału $a_i - \sigma_i < x_i < a_i + \sigma_i$

nie zależy od liczby zmiennych n, (jak w przypadku zmiennej 1D) 68.3%.

Słuszne tylko wtedy, gdy nie występują dodatkowe ograniczenia na przybierane wartości.

SPLOTY ROZKŁADÓW

Sploty rozkładów

Przykład: rozkład kątów emisji cząstek wtórnych w rozpadach cząstek elementarnych.

(do wyznaczenia własności cząstek, np. spinu).

Kąt rozpadu = kąt emisji + błąd pomiaru (2 zmienne losowe, kąt emisji – przypadkowy)

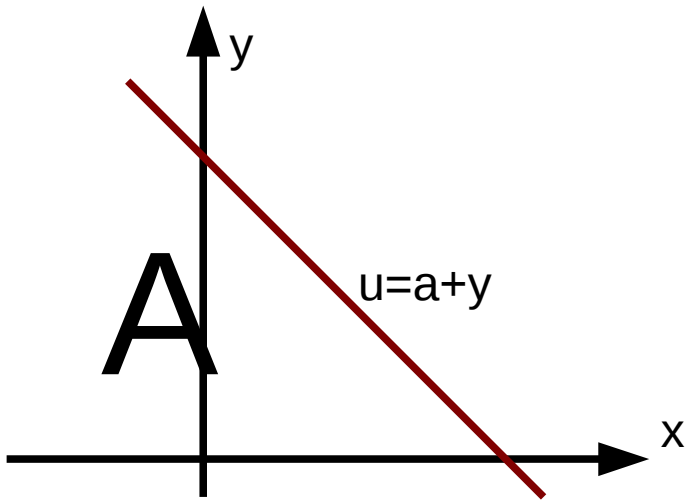
Kąt rozpadu (rozkład) to **splot** (dwóch) innych rozkładów.

Pierwotne wielkości: zmienne x, y, suma: $u = x + y$

Niezależność zmiennych: $f(x, y) = f_x(x)\, f_y(y)$.

Dystrybuanta zmiennej u: $F(u) = P(u < u) = P(x + y < u)$

Możliwe jest jej wyznaczenie poprzez całkowanie całkowanie funkcji f(x, y).



Sploty rozkładów

$$F(u)=\int\int f_x(x)\,f_y(y)\,dx\,dy=\int_{-\infty}^{\infty} f_x(x)\,dx\int_{-\infty}^{u-x} f_y(y)\,dy=\int_{-\infty}^{\infty} f_y(y)\,dy\int_{-\infty}^{u-y} f_x(x)\,dx$$

Gęstość prawdopodobieństwa:

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

Wzór słuszny nawet wtedy, gdy zmienne *x*, *y* określone tylko w pewnym obszarze,

(zmieniają się granice całkowania - zawężają się).

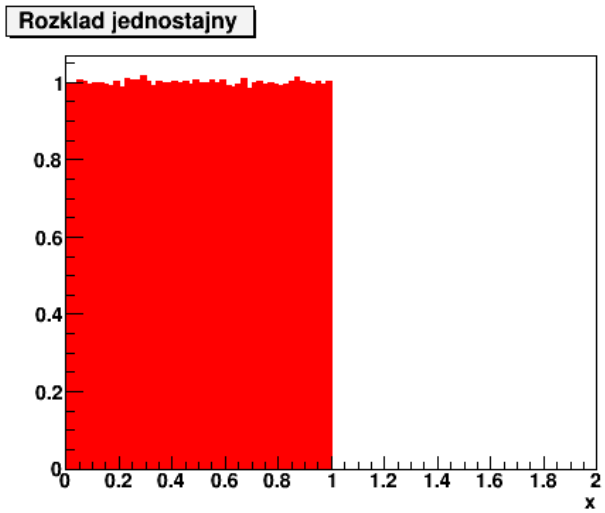
Sploty rozkładów jednostajnych

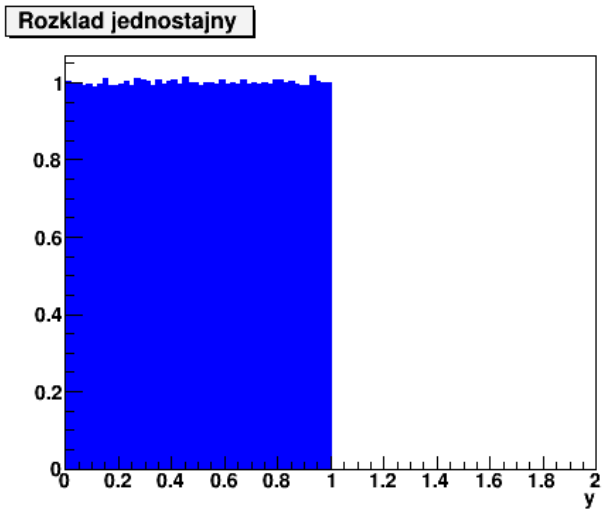
$$f_x(x)=1 \; x\in[0,1] \qquad f_y(y)=1 \; y\in[0,1]$$

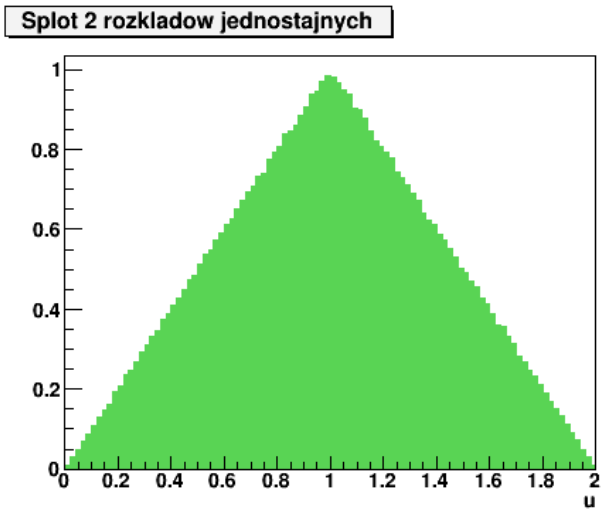
$$f(u)=\int_0^1 f_y(u-x)\,dx \qquad v=u-x \qquad dv=-dx \qquad f(u)=-\int_u^{u-1} f_y(v)\,dv=\int_{u-1}^{u} f_y(v)\,dv$$

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

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







Sploty rozkładów

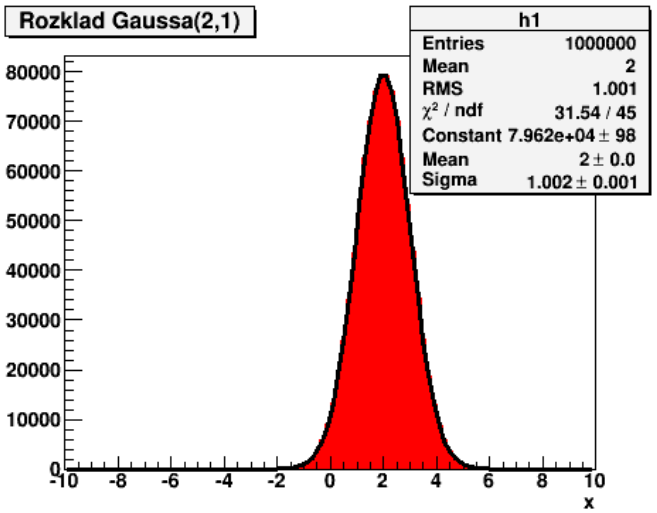
Splot dwóch rozkładów normalnych – dodawanie błędów w kwadratach.

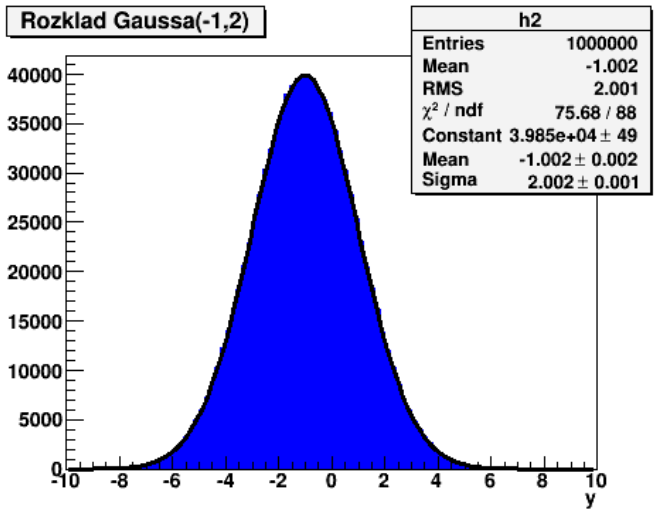
Splot 2 rozkładów normalnych o wartościach oczekiwanych 0 oraz odchyleniach standardowych σ_x, σ_y **jest rozkładem normalnym**:

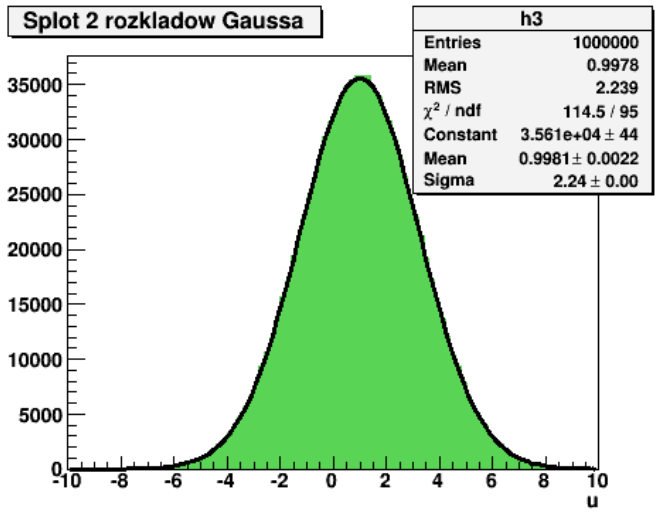
$$f(u) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp \frac{-u^2}{2\sigma^2} \qquad \sigma^2 = \sigma_1^2 + \sigma_2^2$$

Wartości oczekiwane oraz wariancje dodają się

(odchylenia standardowe dodają się w kwadratach).







POBIERANIE PRÓBY

Pobieranie próby

Nieznany rozkład prawdopodobieństwa?

Jego aproksymacja → rozkład częstości (otrzymany doświadczalnie).

Próba - skończony zespół doświadczeń wykonywanych w celu określenia kształtu rozkładu.

Otrzymywana ze zbioru elementów składającego się np. ze zbioru wszystkich możliwych do wyobrażenia doświadczeń i przypadków określonego typu.

Zbiór ten jest nieskończony, → **populacja generalną.**

Próba ma n elementów → próba **n-wymiarowa**.

Rozkład zmiennej losowej *x* określony przez gęstość prawdopodobieństwa f(*x*).

Szukamy wartości zmiennej *x* (uzyskane poprzez l poszczególnych elementów próby).

Pobieranie próby

1. próba: $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$

.

.

j. próba $x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}$

.

.

l. próba: $x_1^{(l)}, x_2^{(l)}, \ldots, x_n^{(l)}$

Wyniki **grupowane** w n-wymiarowe wektory przestrzeni prób:

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

Rozkład gęstości prawdopodobieństwa:

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

Pobieranie próby

Pobieranie próby losowe:

(a) poszczególne **zmienne losowe** muszą być **niezależne**

$$g(\boldsymbol{x}) = g_1(x_1)\, g_2(x_2) \ldots g_n(x_n)$$

(b) poszczególne **rozkłady** muszą być **jednakowe** i identyczne z rozkładem gęstości dla populacji.

$$g_1(x_1) = g_2(x_2) = \ldots = g_n(x_n)$$

W rzeczywistym procesie pobierania próby niezwykle trudno jest zachować pełną losowość.

Wszystkie n elementów próby porządkujemy w kolejności niemalejącej,

$$W_n(x) = n_x / n$$ **dystrybuanta empiryczna.**

- funkcja schodkowa, zwiększą się o 1/n;
- przybliżenie dystrybuanty,
- dąży do tej dystrybuanty <=> n bardzo duże, (n -> inf).

Pobieranie próby

Statystyka - funkcja elementów.

Wartość średnia z próby (średnia arytmetyczna z próby):

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$$

Przy znanej postaci matematycznej gęstości prawdopodobieństwa dla populacji,

gdy trzeba wyznaczyć z populacji wartość parametru (-ów):

Rozpad promieniotwórczy (liczba rozpadających się jąder w przedziale czasowym (0, t)):

$$N_t = N_0\left(1 - \exp -\lambda t\right)$$

λ wyznaczony na podstawie **skończonej próby** (mierząc skończoną liczbę r. promieniotwórczych).

Wynik **nie** może być **dokładny** (ograniczoną liczebność próby).

Parametry są więc **estymowane**.

Estymowana wartość otrzymywana za pomocą pobierania próby (jest więc **statystyką, estymatorem.**

$$S = S\left(x_{1,}\, x_{2,} \ldots, x_n\right)$$

Pobieranie próby

Estymator nieobciążony: (niezależnie od liczebności próby) jego wartość oczekiwana jest równa wartości estymowanego parametru (n dowolne).

$$E\{S(x_1, x_2, \ldots, x_n)\} = \lambda$$

Estymator zgodny: wariancja znika dla dowolnie dużej próby.

$$\lim_{n \to \infty} \sigma(S) = 0$$

(Czasami podaje się dolną granicę wariancji dla estymatorów.

Jeśli znajdziemy S_0, którego wariancja jest równa tej granicy →

→ to stosowanie go pozwala uzyskać maksymalną efektywność w estymacji parametrów.

To **estymator efektywny** parametru λ.

Pobieranie próby z populacji typu ciągłego

Wartość średnia ze wszystkich elementów próby (średnia arytmetyczna) to zmienną losową.

Jej wartość oczekiwana:

$$E\{\bar{x}\}=\frac{1}{n}\{E(x_1)+E(x_2)+\ldots E(x_n)\}=\hat{x}$$

Jest ona równa wartości oczekiwanej zmiennej *x*.

Ponieważ powyższa równość jest spełniona dla każdego n, to jest to **wartość średnia (średnia arytmetyczna) to estymator nieobciążony dla wartości oczekiwanej** zmiennej *x* w populacji.

Funkcja charakterystyczna wartości średniej:

$$\varphi_{\bar{x}}(t)=\{\varphi_{\frac{x}{n}}(t)\}^n=\{\varphi_x(\frac{t}{n})\}^n$$

Pobieranie próby z populacji typu ciągłego

Wariancja wartości średniej (średniej arytmetycznej):

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

$$\sigma^2(\bar{x}) = \frac{1}{n^2} E\{[(x_1 - \hat{x}) + (x_2 - \hat{x}) + \ldots + (x_n - \hat{x})]^2\}$$

Wszystkie elementy próby niezależne, wszystkie kowariancje

$$E\{(x_i - \hat{x})(x_j - \hat{x})\} = 0 \qquad i \neq j$$

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

Wartość średnia z próby to estymator zgodny wartości oczekiwanej.

Pobieranie próby z populacji typu ciągłego

Wariancja to nie zmienna losowa (nie można wyznaczyć doświadczalnie).

Wyznaczenie **estymatora wariancji**: (w pierwszym przybliżeniu:

średnia arytmetyczna odchyleń kwadratowych od wartości średniej z próby):

$$S'^2 = \frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\}$$

Można dowieść, że:

$$E(S'^2) = \frac{n-1}{n}\sigma^2(x)$$

S'^2 to estymator obciążony dla wariancji.

Estymator nieobciążony wariancji:

$$S^2 = \frac{1}{n-1}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\}$$

Pobieranie próby z populacji typu ciągłego

Estymator wariancji wartości średniej:

$$S^2(\bar{x}) = \frac{1}{n} s^2(x) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Błąd wartości średniej z próby:

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

Wariancja wyrażenia:

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

Błąd wariancji z próby:

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

Estymator odchylenia standardowego próby i jego błąd:

$$s = \sqrt{s^2} = \frac{1}{\sqrt{n-1}} \sqrt{\sum (x_i - \bar{x})^2} \qquad \Delta s = \frac{s}{\sqrt{2(n-1)}}$$

WIZUALIZACJA

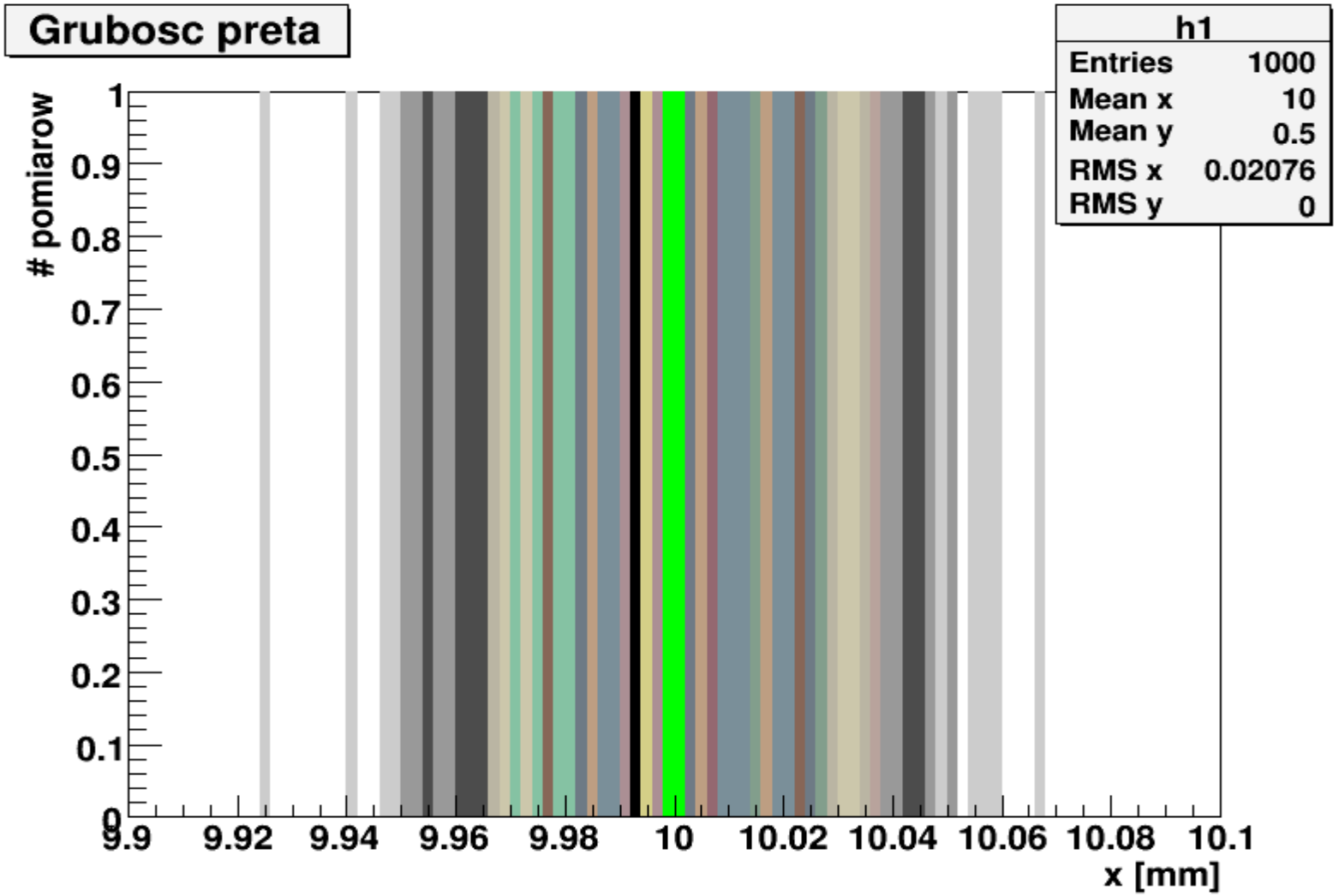
Przedstawianie próby w postaci graficznej

Zmierzona została 1000 razy grubość pręta ołowianego.

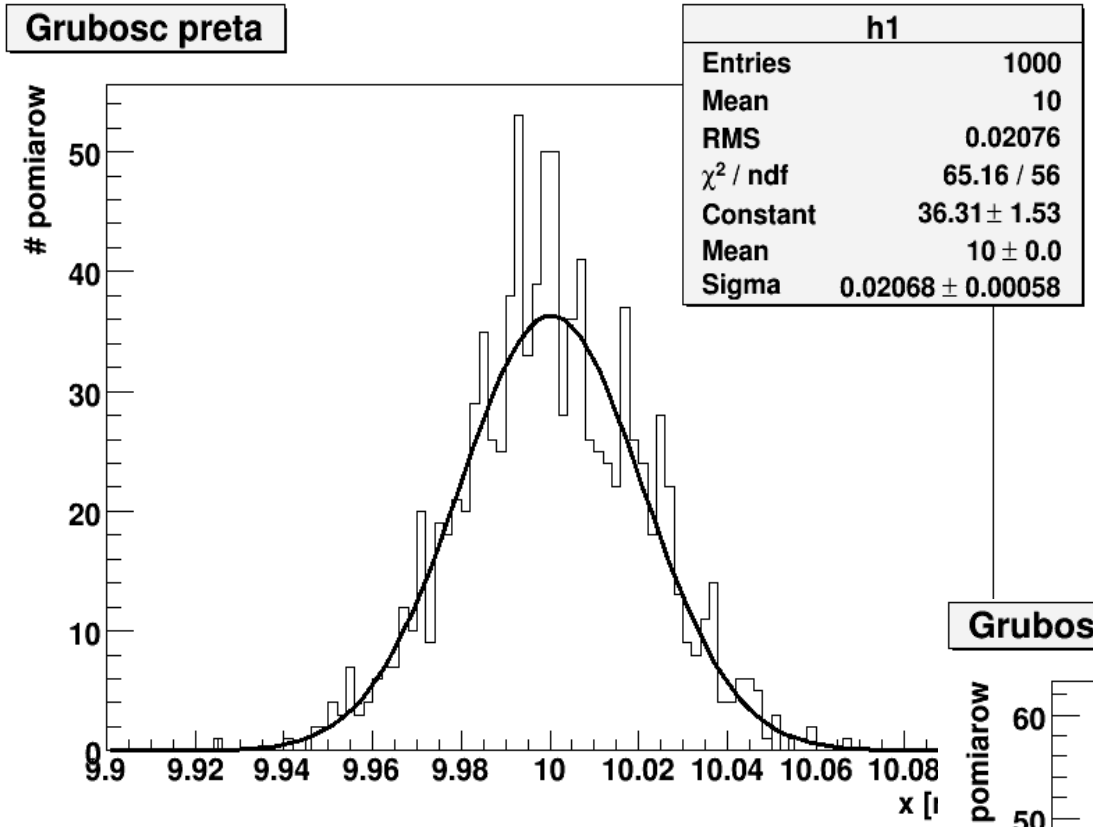
W jaki sposób przedstawić wyniki pomiarów oraz wyznaczyć grubość wraz z błędem?

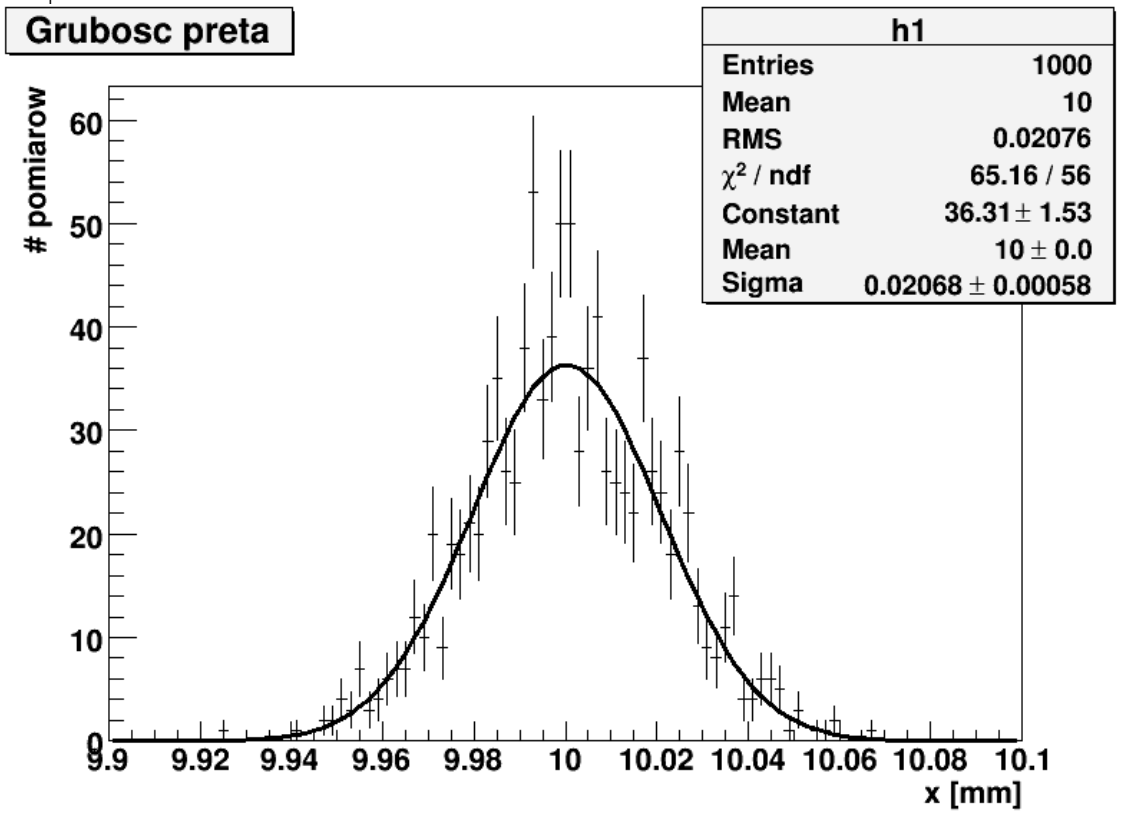
1) Określamy przedział zmienności grubości x

2) Określamy na ile przedziałów o jednakowej grubości go podzielimy

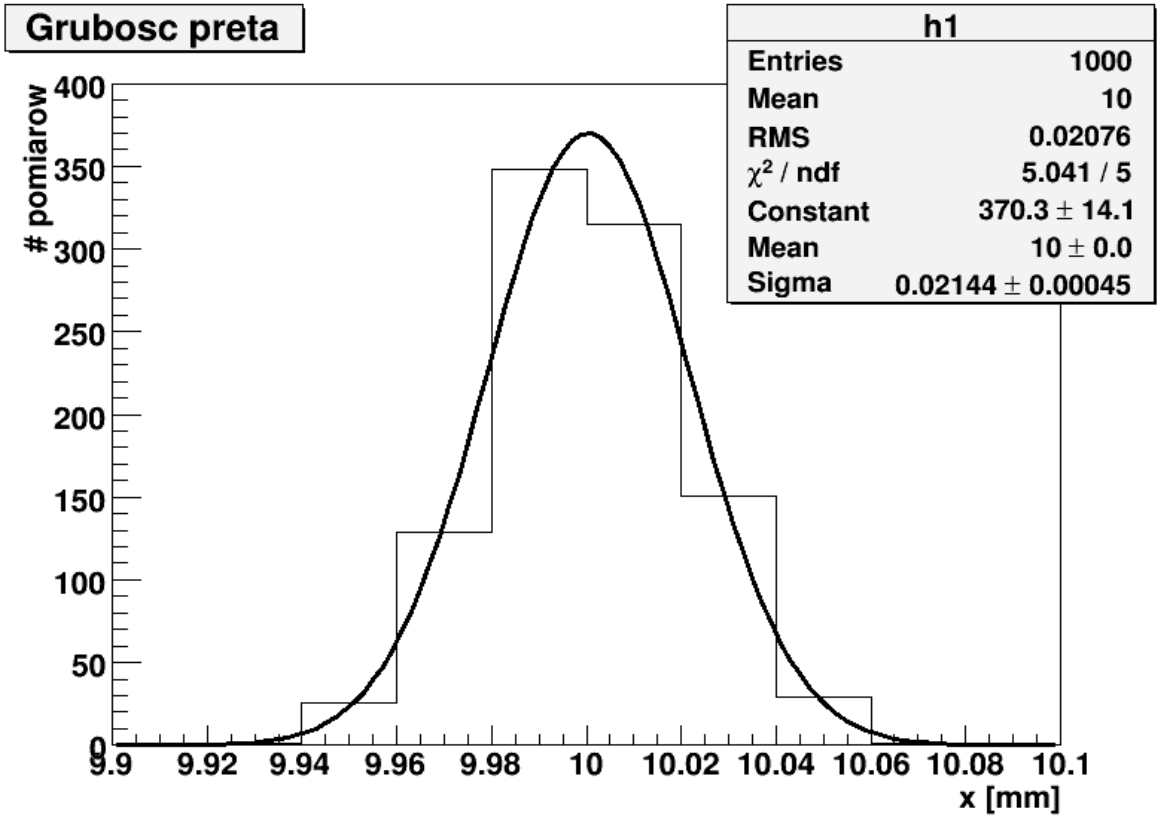


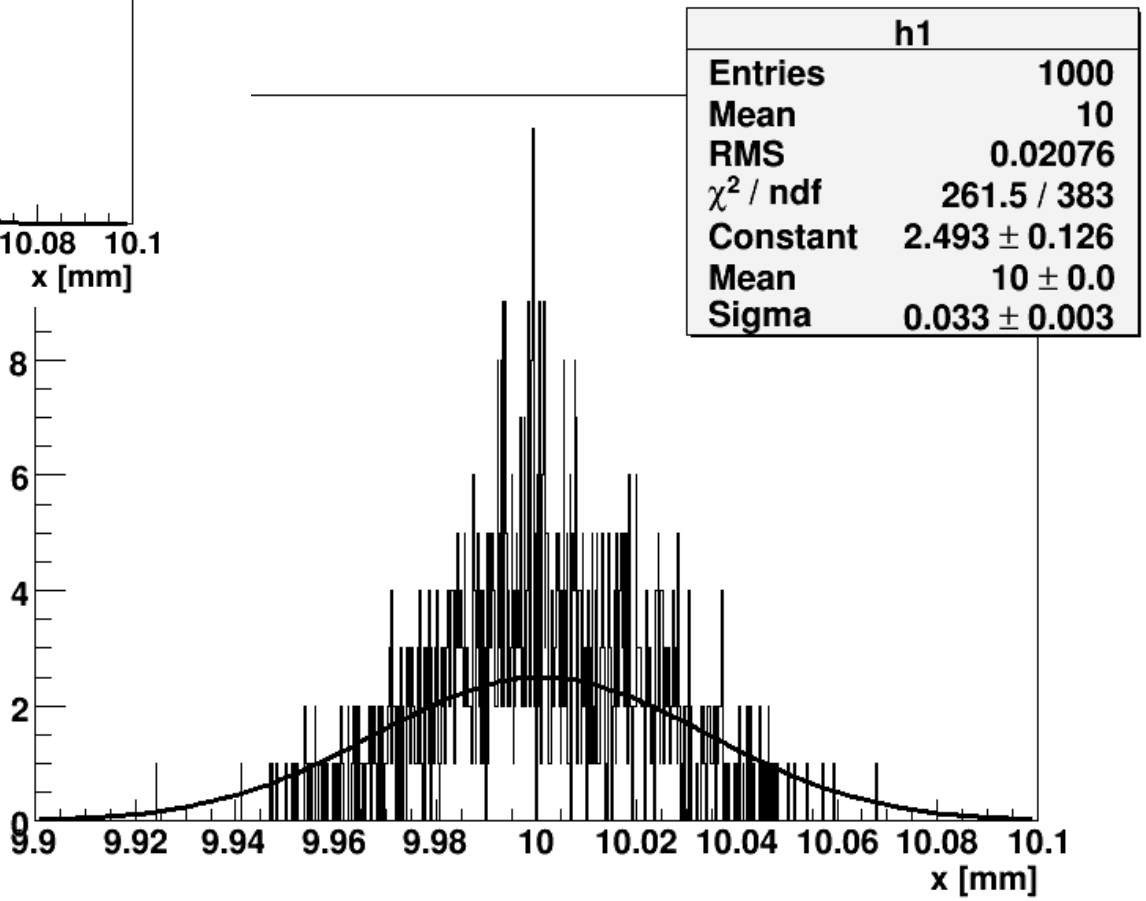
Przedstawianie próby w postaci graficznej





Przedstawianie próby w postaci graficznej





KONIEC WYKŁADU 6