Komputerowa analiza danych doświadczalnych

Wykład 2
3.03.2017

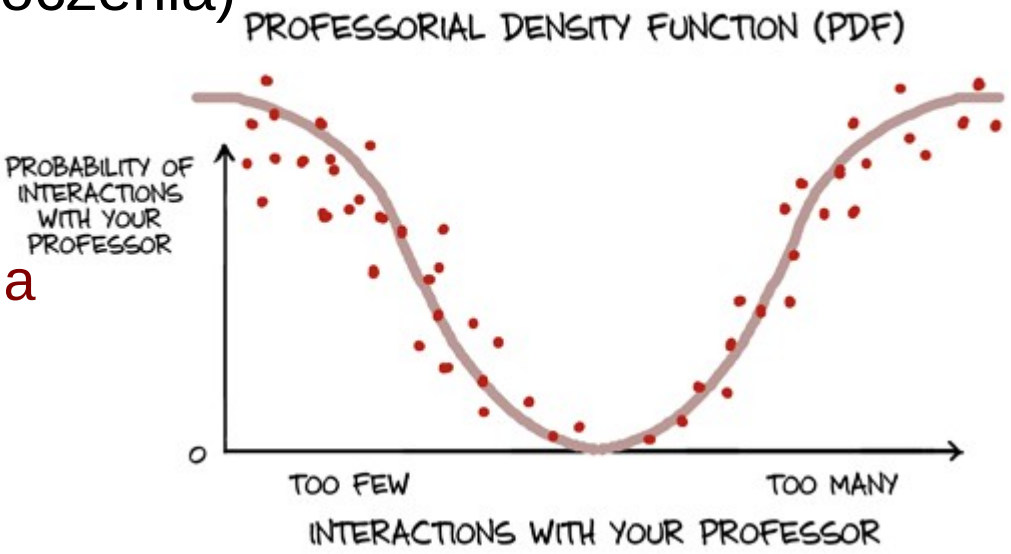
dr inż. Łukasz Graczykowski
lgraczyk@if.pw.edu.pl

Semestr letni 2016/2017

Zmienne losowe, jednowymiarowe rozkłady zmiennych losowych

Pomiar jako zdarzenie losowe

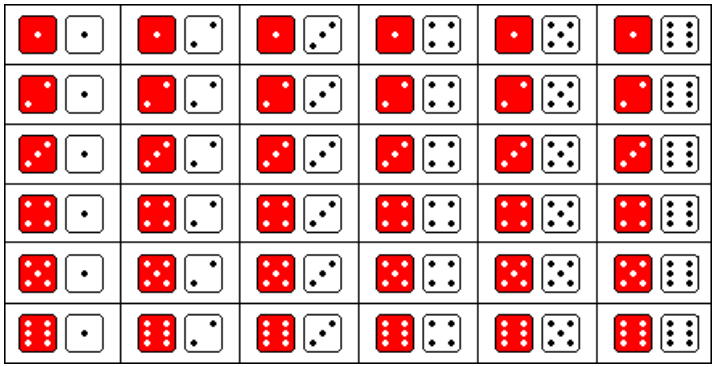
- Wyniki kolejnych pomiarów jakiegoś zjawiska, niezależnie od tego jak bardzo byśmy się starali przestrzegać procedury pomiarowej, będą różne (raz mniejsze, raz większe) – oczywiście zakładając wysoką precyzję przyrządu pomiarowego (patrz Wykład 1)
- Może to wynikać zarówno ze **statystycznego charakteru badanego zjawiska** (np. rozpad promieniotwórczy) jak i **niedokładności przyrządów badawczych** oraz **innych czynników** (np. zmienne warunki otoczenia)
- Z powyższego możemy założyć, że:
  - **pomiar** jest **zdarzeniem losowym**
  - **wynik pomiaru** (realizacja zdarzenia losowego) jest **zmienną losową**
- Uwzględniając powyższe założenia, wnioski na temat pomiaru możemy określać przy pomocy teorii (rachunku) prawdopodobieństwa (patrz Wykład 1)



http://www.phdcomics.com/comics/archive/phd091211s.gif

Typy i rodzaje zmiennych losowych

- **Zmienna losowa** – funkcja przypisująca liczby zdarzeniom elementarnym (np. wynik rzutu 2 kostkami – para liczb)
- Typy zmiennych losowych:
  - jednowymiarowe (dzisiejszy wykład)
  - dwuwymiarowe
  - …
  - n-wymiarowe
- Rodzaje zmiennych losowych
  - dyskretne (lub skokowe)
  - ciągłe
- Oznaczenie: X, Y, …

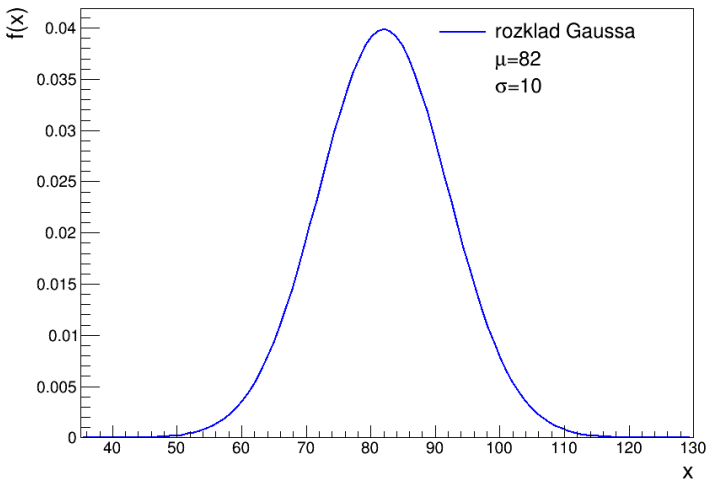


https://mosaicprojects.files.wordpress.com/2013/01/diceposs.gif

Rozkład i dystrybuanta zmiennej losowej

- **Rozkład (gęstość) prawdopodobieństwa** (*ang. probability distribution*, *density*) – funkcja przypisująca zmiennym losowym (np. zmiennej X) prawdopodobieństwo uzyskania danej wartości zmiennej losowej (np. wartości x):

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

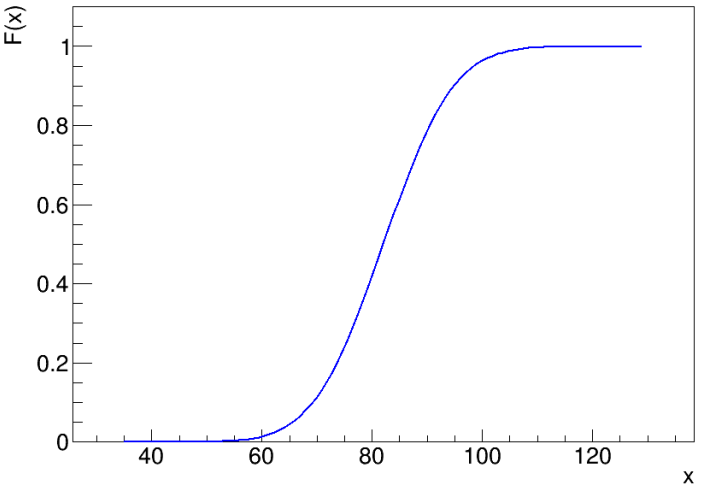


  - rozkład prawdopod. jest unormowany
    - rozkład ciągły: $\int_{-\infty}^{\infty} f(x)\,dx = 1$
    - rozkład dyskretny: $\sum_{i=1}^{\infty} P(X = x_i) = \sum_{i=1}^{\infty} p_i = 1$

- **Dystrybuanta** (*ang. cumulative distribution function*) – funkcja określająca prawdopodobieństwo tego, że zmienna losowa X przyjmie wartość mniejszą bądź równą x:

$$F(x) = P(X \leqslant x) = P((-\infty ; x >_x)$$

  - rozkład ciągły: $F(x) = \int_{-\infty} f(x')\,dx'$
  - rozkład dyskretny: $F(x) = \sum_{i:x_i \leqslant x} P(X = x_i)$



Własności dystrybuanty

- Własności dystrybuanty:
  - funkcja niemalejąca

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$

  - jeżeli dystrybuanta F(x) jest ciągła oraz ma 1-szą pochodną:

$$F'(x) = \frac{dF(x)}{dx} = f(x)$$

  - prawdopodobieństwa:

$$P(x \leqslant a) = \int_{-\infty}^{a} f(x)\,dx = F(a)$$

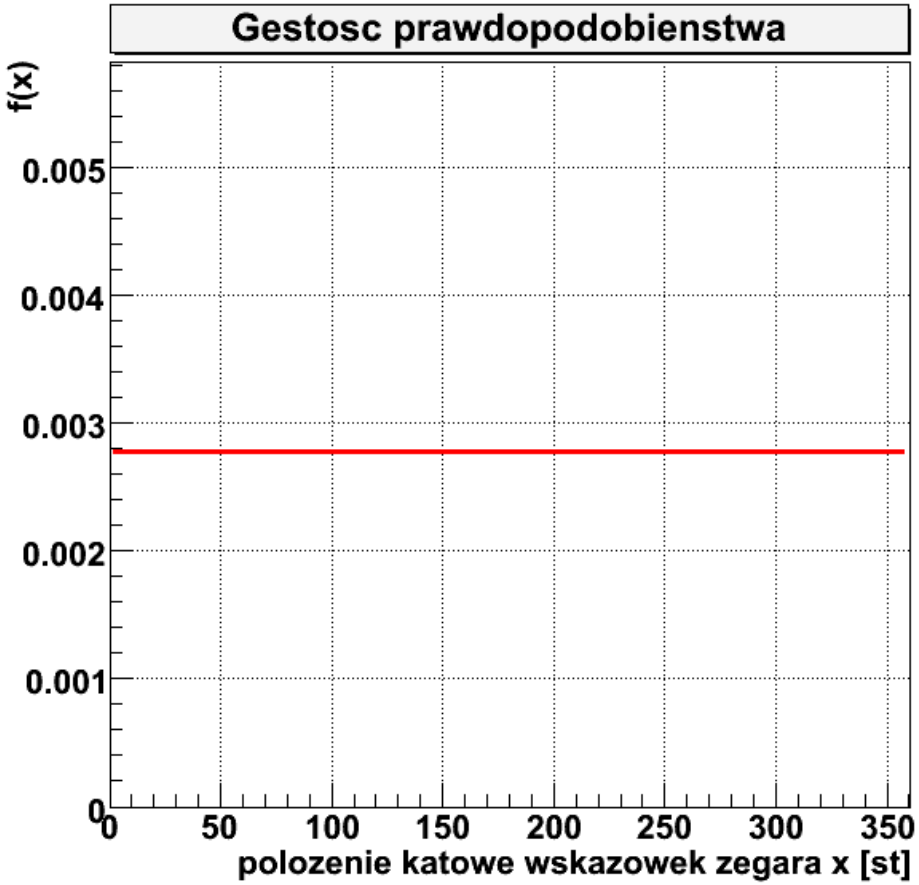
$$P(a \leqslant x \leqslant b) = \int_{a}^{b} f(x)\,dx = F(b) - F(a)$$

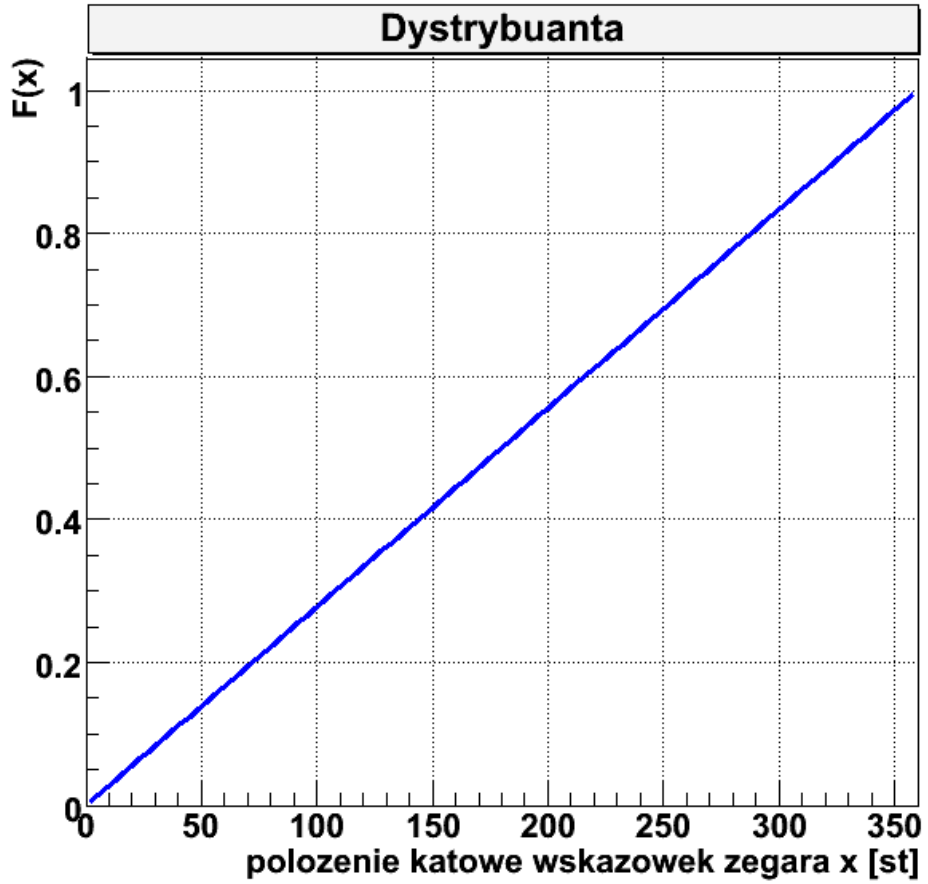
Rozkład i dystrybuanta - przykłady

- Rozkład położenia kątowego wskazówki zegara (rozkład jednorodny, lub jednostajny) – zmienna losowa ciągła:

$$f(x)=\frac{1}{360};\ x\in\langle 0;360\rangle \qquad F(x)=0;\ x<0 \qquad F(x)=1;\ x>360$$

$$f(x)=0;\ x\in\mathbb{R}\setminus\langle 0;360\rangle \qquad F(x)=\int_0^x f(x')\,dx'=\frac{1}{360}x,\ x\in\langle 0;360\rangle$$





Rozkład i dystrybuanta - przykłady

- Rozkład normalny (rozkład Gaussa) – zmienna losowa ciągła:

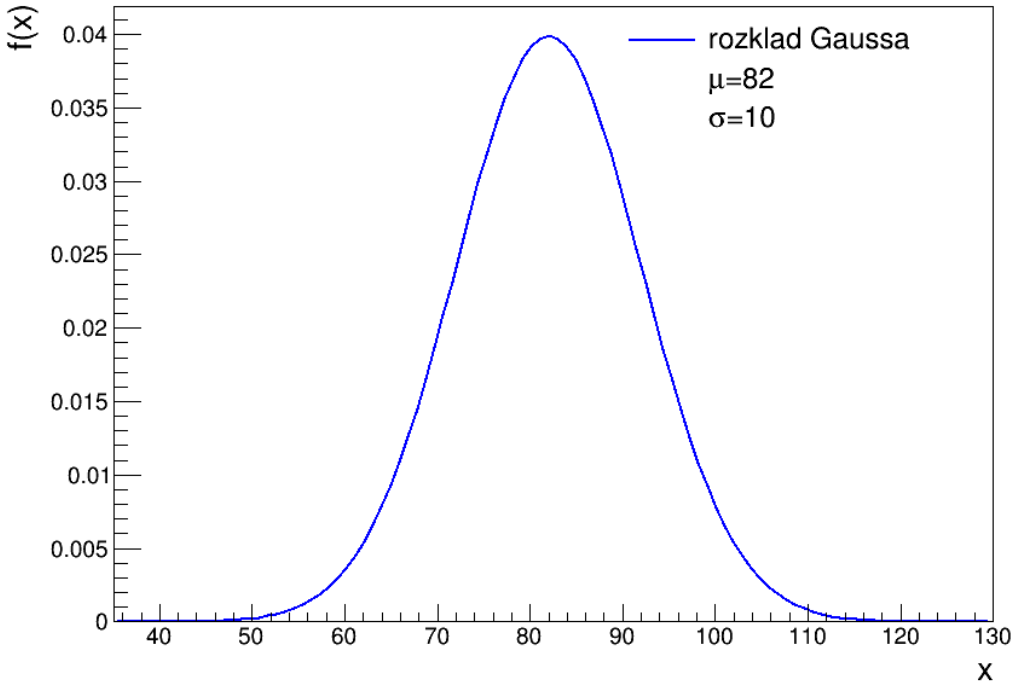
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), x\in\mathbb{R}$$

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(\frac{(x'-\mu)^2}{2\sigma^2}\right)dx'$$

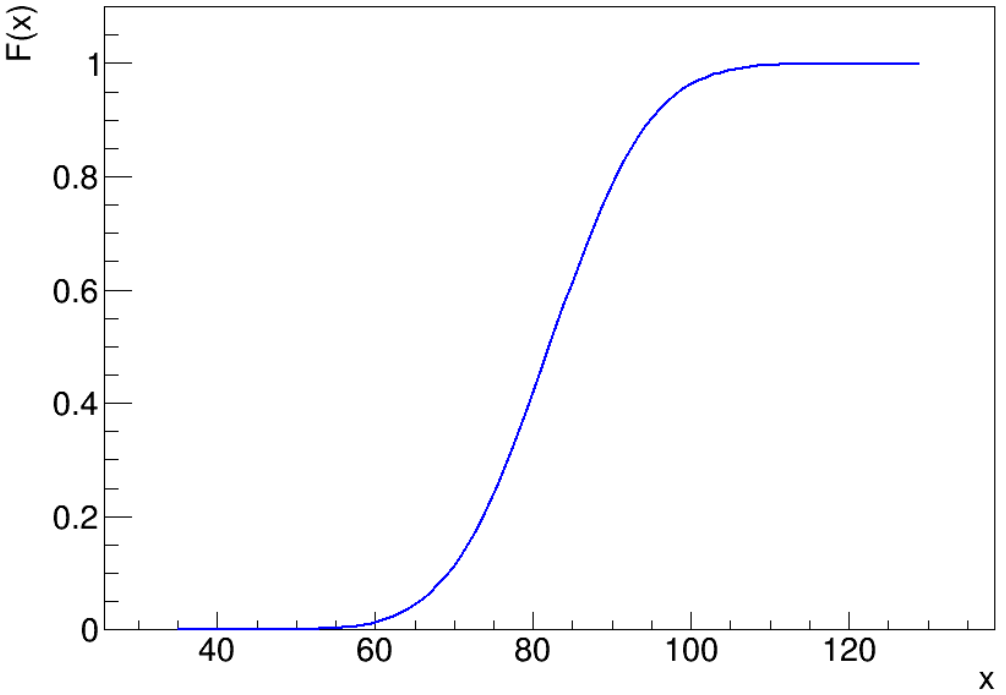
$$\lim_{x\to-\infty} F(x)=0 \qquad \lim_{x\to\infty} F(x)=1$$

Dystrybuanta rozkładu normalnego **nie ma** postaci analitycznej

Rozkład (funkcja gęstości)



Dystrybuanta

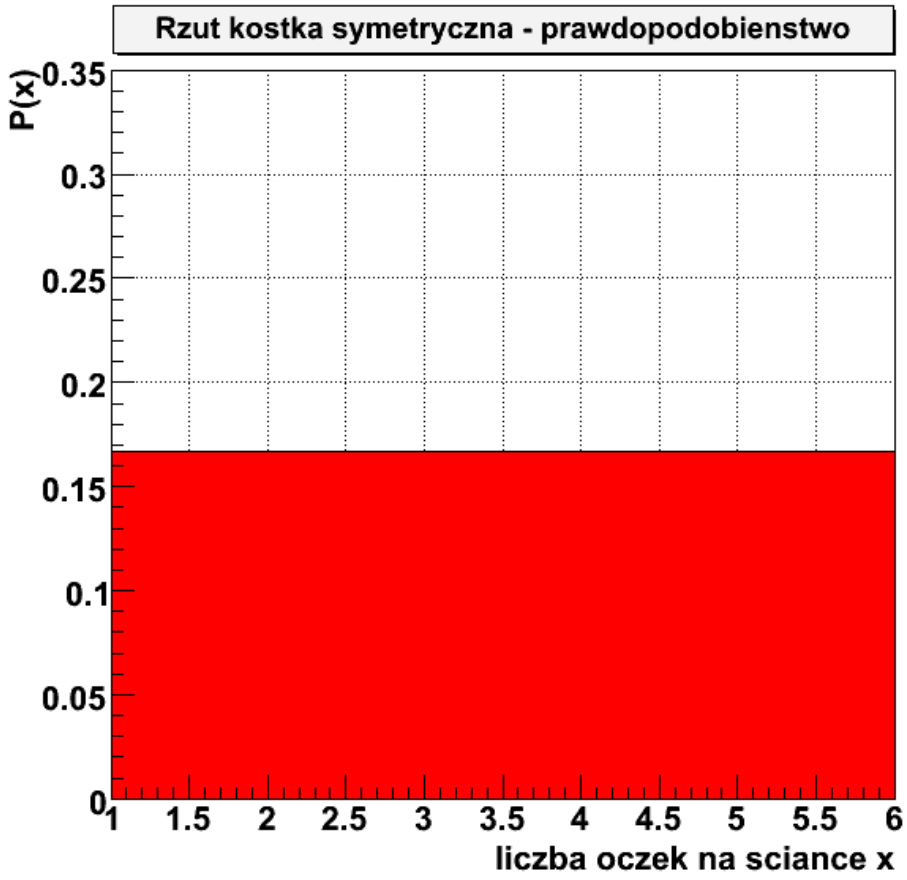


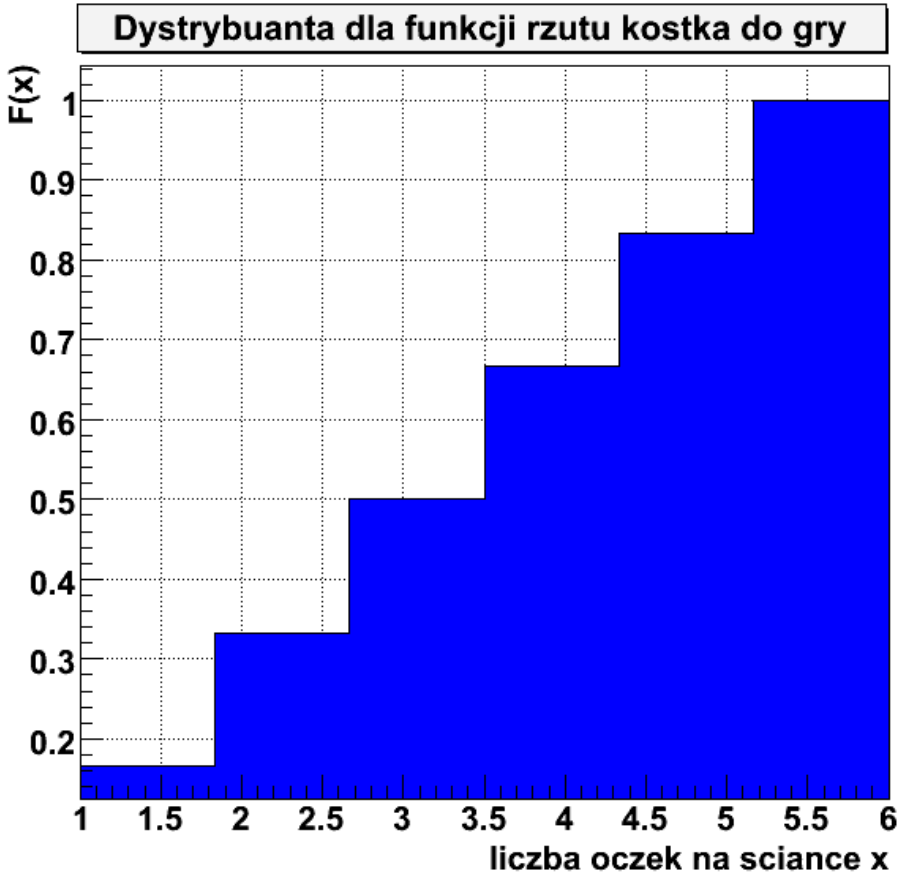
Rozkład i dystrybuanta - przykłady

- Rzut kostką – zmienna losowa dyskretna:

$$P(X=x_i)=P(x_i)=\frac{1}{6}, i=\{1,2,3,4,5,6\}$$

$$F(x_i)=\frac{1}{6}i, i=\{1,2,3,4,5,6\}$$





Funkcje zmiennej losowej, wartość oczekiwana

- Jeżeli Y jest funkcją zmiennej losowej X, to Y również jest zmienną losową (ze swoim rozkładem i dystrybuantą):

$$Y = H(X)$$

- **Wartość oczekiwana (średnia, przeciętna)** (*ang. mean value*) – suma wszystkich możliwych wartości x_i zmiennej X, przemnożonych przez ich prawdopodobieństwa:

$$E(X) \equiv \mu \equiv \hat{x} \equiv \bar{x} = \sum_{i=1}^{n} x_i P(X = x_i) = \sum_{i=1}^{n} x_i p_i$$

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

- Wartość oczekiwana zmiennej Y:

$$E(Y) = E\left(\left(H(X)\right)\right) = \sum_{i=1}^{n} H(x_i) P(X = x_i)$$

- Dla zmiennych losowych typu ciągłego:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx \qquad E(Y) = E\left(H(X)\right) = \int_{-\infty}^{\infty} H(x) f(x)\, dx$$

Momenty

- Jeżeli zdefiniujemy funkcję postaci:

$$Y = H(X) = (X - c)^l$$

- to jej wartości średnie a_l są **momentami rzędu *l* względem *c***:

$$a_l = E\left((X-c)^l\right) = \int_{-\infty}^{\infty} (x-c)^l f(x)\,dx \qquad m_l = E\left(X^l\right) = \int_{-\infty}^{\infty} x^l f(x)\,dx$$ - moment zwykły

- Jeżeli to $c = \hat{x}$, to momenty nazywane są **momentami centralnymi**:

$$\mu_l = E\left((X-\hat{x})^l\right)$$

- Łatwo pokazać, że:

$$\mu_0 = E\left((X-\hat{x})^0\right) = \int_{-\infty}^{\infty} (x-\hat{x})^0 f(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\mu_1 = E\left((X-\hat{x})^1\right) = \int_{-\infty}^{\infty} (x-\hat{x}) f(x)\,dx = \int_{-\infty}^{\infty} x f(x)\,dx - \hat{x}\int_{-\infty}^{\infty} f(x)\,dx = \hat{x} - \hat{x} = 0$$

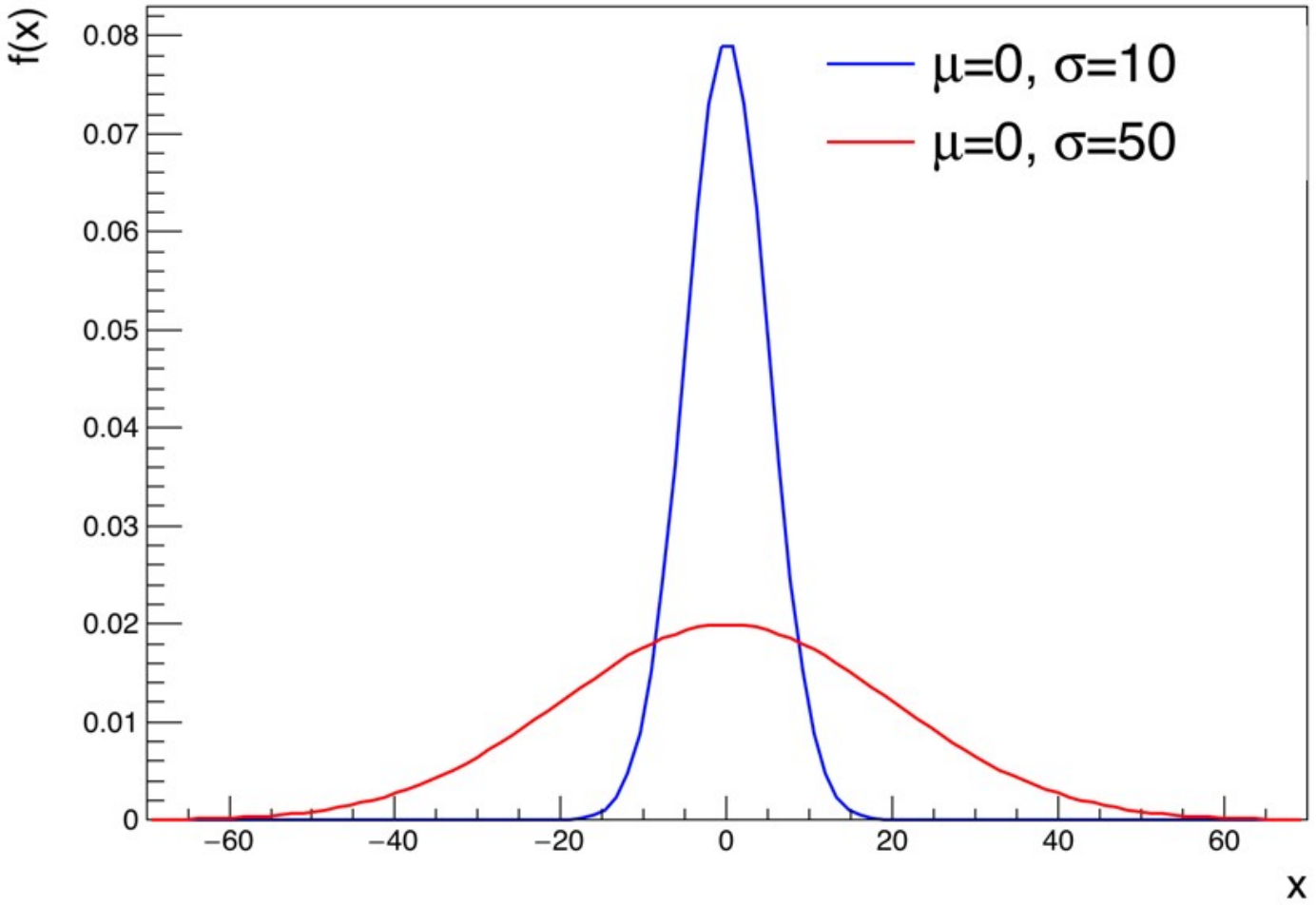
- Najniższy moment, który niesie informacje o odchyleniu zmiennej losowej X od swojej wartości średniej nazywany jest **wariancją** (*ang. variance*):

$$\mu_2 \equiv \sigma^2(X) \equiv var(X) \equiv E\left((X-\hat{x})^2\right) = \int_{-\infty}^{\infty} (x-\hat{x})^2 f(x)\,dx$$

  - jeżeli wariancja jest mała, to wyniki leżą blisko wartości oczekiwanej, jeśli duża, to wyniki są bardziej rozproszone

Rozkład normalny o dużej i małej wariancji

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



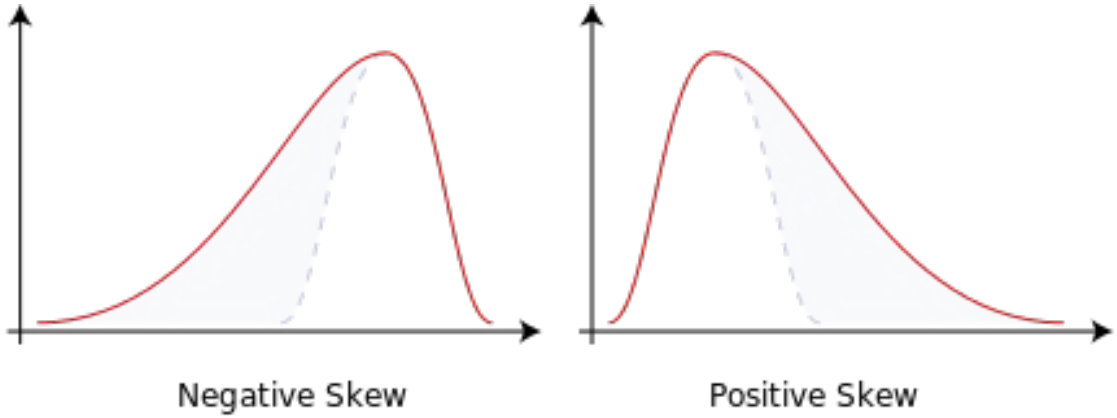
Momenty wyższych rzędów

- Dodatnia wartość pierwiastka z wariancji nazywana jest **odchyleniem standardowym** (*ang. standard deviation*) lub **dyspersją**:

$$\sigma \equiv \sigma(X) = \sqrt{\sigma^2(X)}$$

  - odchylenie standardowe określa niepewność pomiaru (patrz Wykład 1)

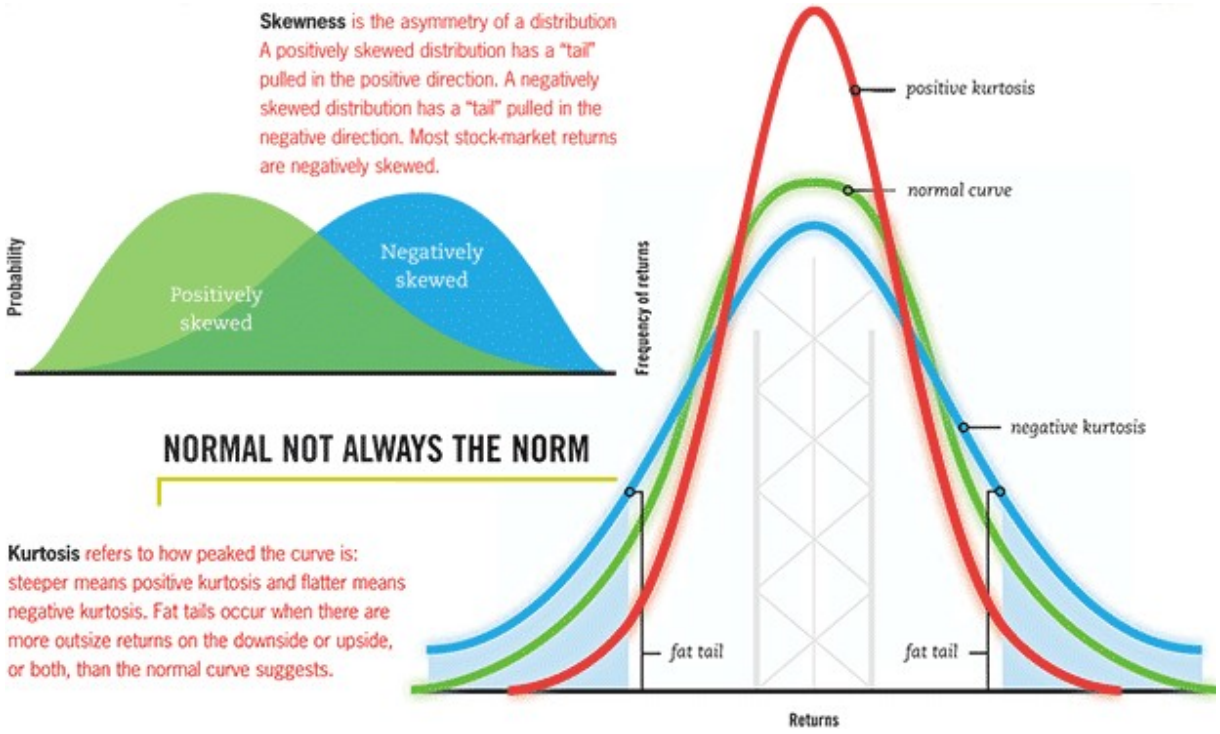
- Trzeci moment centralny nazywany jest **skośnością** lub **współczynnikiem skośności** (*ang. skewness*):
  - najczęściej wprowadza się bezwymiarową wielkość nazywaną **współczynnikiem asymetrii** rozkładu: $\gamma = \frac{\mu_3}{\sigma^3}$
  - dla rozkładów symetrycznych (względem średniej) parametr ten wynosi 0



https://en.wikipedia.org/wiki/Skewness#/media/File:Negative_and_positive_skew_diagrams_(English).svg

Momenty wyższych rzędów

- Czwarty moment centralny nazywany jest **kurtozą** (*ang. kurtosis*):
  - Analogicznie do skośności, najczęściej wprowadza się bezwymiarową wielkość: $K = \frac{\mu_4}{\sigma^4}$
  - ponieważ kurtoza rozkładu normalnego wynosi 3, często kurtozę (zwaną **kurtozą nadmiarową** – *ang. excess kurtosis*) definiuje się odejmując 3 (by dla rozkł. normalnego wynosiła 0): $K = \frac{\mu_4}{\sigma^4} - 3$



http://www.advisor.ca/wp-content/uploads/2012/07/normal-not-always-the-norm.gif

Własności wartości oczekiwanej i wariancji

- Własności wartości oczekiwanej:
  - $E(c \cdot X) = c \cdot E(X); c \in \mathbb{R}$
  - $E(X+Y) = E(X) + E(Y)$
  - $E(X+c) = E(X) + c; c \in \mathbb{R}$
  - $E(c) = c; c \in \mathbb{R}$
- Z czego wynika:
  - $E(a \cdot X + b \cdot Y + c) = a \cdot E(X) + b \cdot E(Y) + c; a, b, c \in \mathbb{R}$
  - $E(X - E(X)) = E(X) - E(E(X)) = E(X) - E(X) = 0$
- Zależność między wariancją a wartością oczekiwaną:

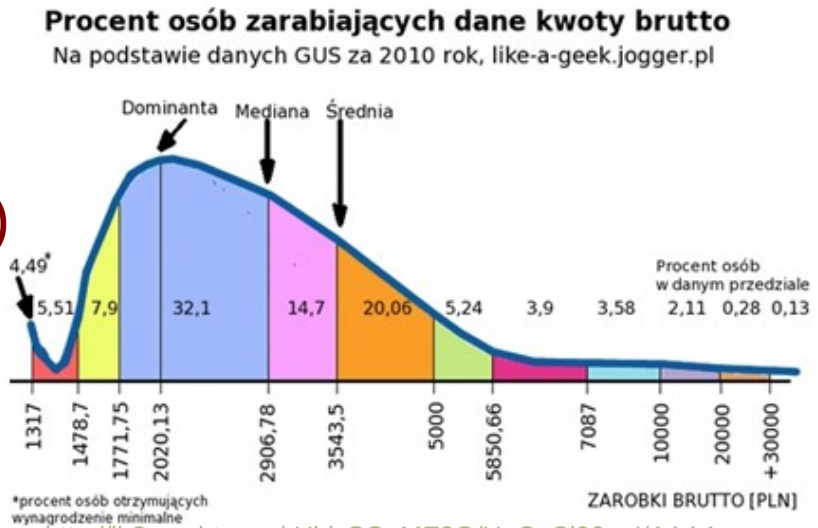
$$\sigma^2(X) = E\left((X - \hat{x})^2\right) = E\left(X^2 - 2X \cdot \hat{x} + \hat{x}^2\right) = E(X^2) - 2(E(X))^2 + (E(X))^2 = E(X^2) - (E(X))^2$$

- Własności wariancji: $\sigma^2(c) = 0; c \in \mathbb{R}$
  - $\sigma^2(c \cdot X) = c^2 \cdot \sigma^2(X); c \in \mathbb{R}$
  - $\sigma^2(X + c) = \sigma^2(X); c \in \mathbb{R}$

Zmienna stand., wartość modalna, mediana

- **Zmienna standardowa** (o wartości oczekiwanej 0 i odchyleniu 1):
  - rozważmy funkcję: $U = \frac{X - \hat{x}}{\sigma(X)}$
  - wartość oczekiwana: $E(U) = \frac{1}{\sigma(X)} E(X - \hat{x}) = \frac{1}{\sigma(X)} (\hat{x} - \hat{x}) = 0$
  - wariancja: $\sigma^2(U) = \frac{1}{\sigma^2(X)} E\{(X - \hat{x})^2\} = \frac{\sigma^2(X)}{\sigma^2(X)} = 1$
- **Wartość modalna**, **dominanta** (ang. *mode*): $P(X = x_{max}) = max$
  - wartość najbardziej prawdopodobna
  - rozkład jednomodalny (1 maksimum)
  - rozkład wielomodalny (wiele maksimów)
  - warunki maksimum:

$$\frac{df(x)}{dx} = 0 \qquad \frac{d^2 f(x)}{dx^2} < 0$$



http://lh3.ggpht.com/-UhjcSGuME9Q/UgCqCj00_nI/AAAAAAAAWXU/-0ZlMA9pPnU/image_thumb%25255B2%25255D.png?imgmax=800

- **Mediana** (*ang. median*):
  - wartość zmiennej losowej, dla której dystrybuanta wynosi 1/2

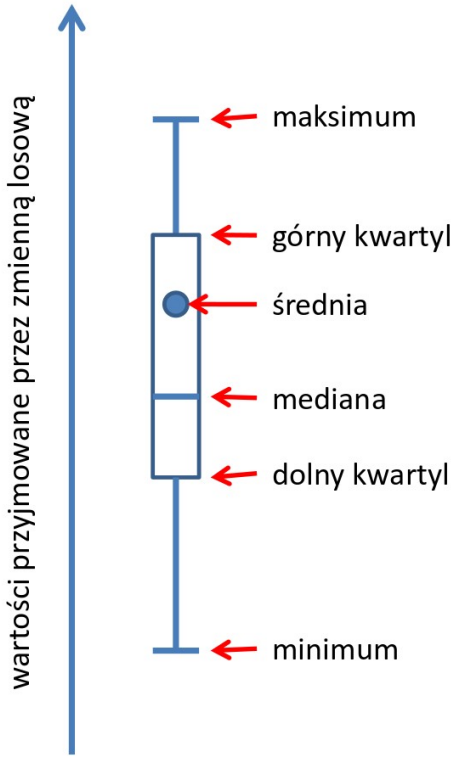
$$F(x_{0,5}) = P(X < x_{0,5}) = 0{,}5$$

Kwantyle

- Mediana dzieli rozkład prawdopodobieństwa na dwa obszary o równym prawdopodobieństwie
- W przypadku rozkładów symetrycznych jednomodalnych wartości: średnia = dominanta = mediana
- Mediana $x_{0,5}$ jest **kwantylem** (*ang. quantile*) rzędu 0,5
- Ogólna definicja **kwantylu rzędu q**, x_q:
  - **kwartyl** dolny $x_{0,25}$
  - kwartyl górny $x_{0,75}$
  - **decyle** $x_{0,1}$, $x_{0,2}$, ..., $x_{0,9}$
  - funkcja $x_q(q)$ jest funkcją odwrotną do dystrybuanty

$$F(x_q) = P(X < x_q) = q$$

$$F(x_q) = \int_{-\infty}^{x_q} f(x)\,dx = q,\ q \in \langle -1;1 \rangle$$



- Kwantyl rzędu q jest taką liczbą x_q, że $q \cdot 100\%$ elementów w danej próbce (populacji) ma wartość pomiaru (badanej cechy) nie większą niż x_q

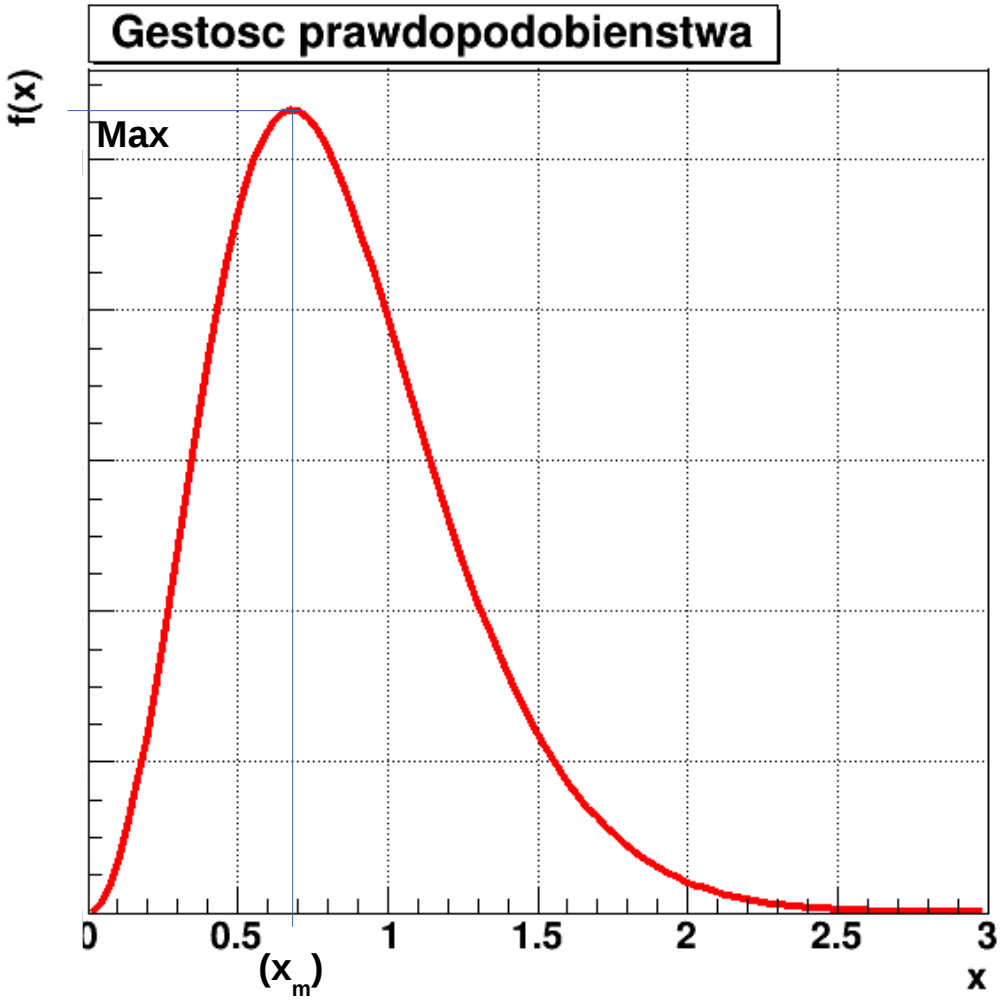
Kwantyle

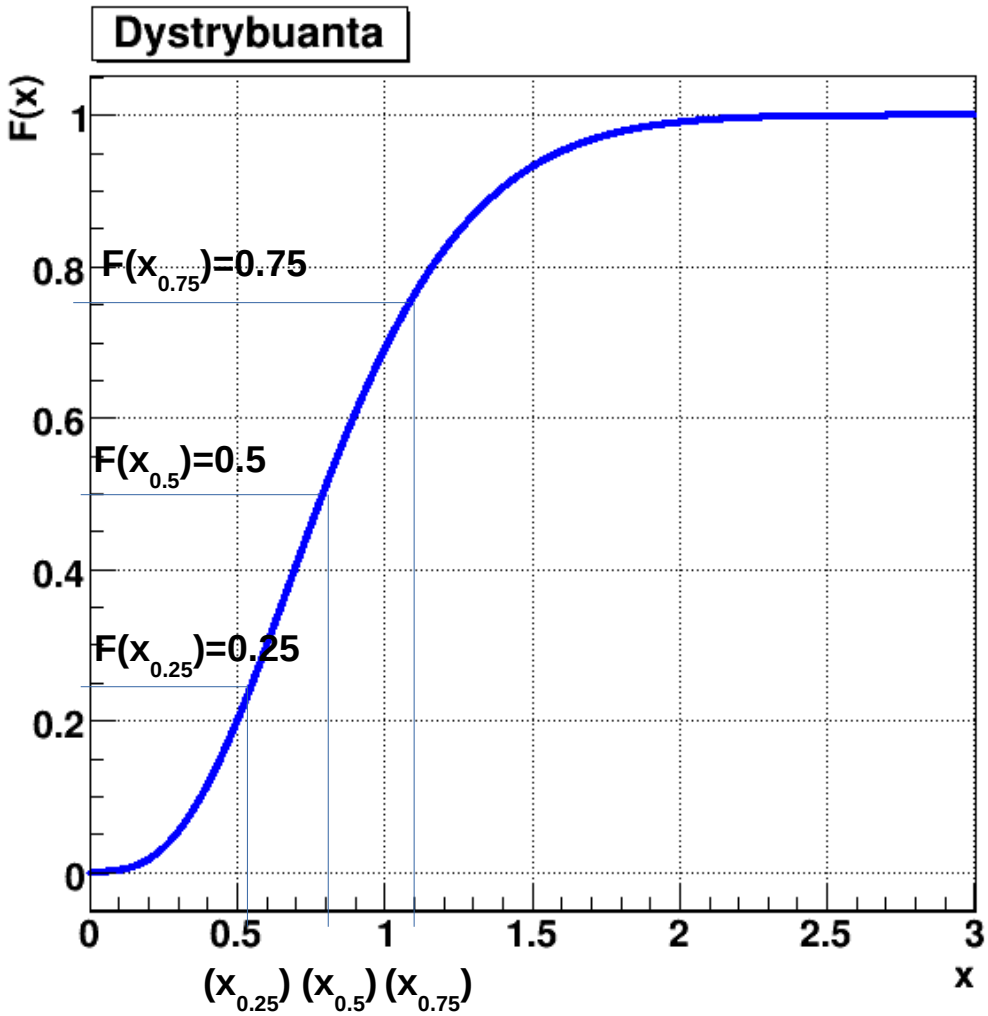
$$F(x_q) = \int_{-\infty}^{x_q} f(x)\,dx = q$$ - kwantyl rzędu q

$$F(x_{0,5}) = \int_{-\infty}^{x_{0,5}} f(x)\,dx = 0{,}5$$ - mediana

$$F(x_{0,25}) = \int_{-\infty}^{x_{0,25}} f(x)\,dx = 0{,}25$$ - kwartyl dolny

$$F(x_{0,75}) = \int_{-\infty}^{x_{0,75}} f(x)\,dx = 0{,}75$$ - kwartyl górny

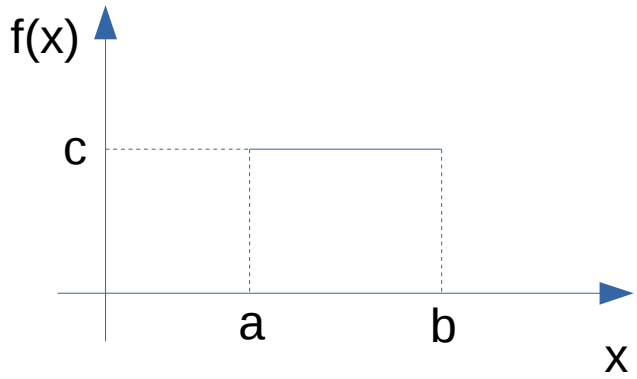




Przykład 1 - rozkład jednostajny

- Gęstość prawdopodobieństwa:

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



- Współczynnik (normalizacja) c:

$$\int_{-\infty}^{\infty} f(x)\,dx=c\int_{a}^{b} dx=c(b-a)=1\Rightarrow c=\frac{1}{b-a}$$
$$f(x)=\frac{1}{b-a}\,;x\in\langle a,b\rangle$$
$$f(x)=0\,;x\in\mathbb{R}\setminus\langle a,b\rangle$$

- Dystrybuanta:

$$F(x)=0\,;x<a$$
$$F(x)=\frac{1}{b-a}\int_{a}^{x} dx'=\frac{x-a}{b-a}\,;x\in\langle a\,;b\rangle$$
$$F(x)=1\,;x>b$$

Wariancja: $\sigma^2(X)=E(X^2)-(E(X))^2$

$$E(X^2)=\frac{1}{b-a}\int_{a}^{b} x^2\,dx=\frac{(b^3-a^3)}{3(b-a)}=$$
$$=\frac{(b-a)(b^2+ba+a^2)}{3(b-a)}=\frac{b^2+ba+a^2}{3}$$
$$\sigma^2(X)=\frac{b^2+ba+a^2}{3}-\left(\frac{b+a}{2}\right)^2=$$
$$=\frac{b^2+ba+a^2}{3}-\frac{b^2+2ba+a^2}{4}=\boxed{\frac{(b-a)^2}{12}}$$

- Wartość oczekiwana:

$$E(X)=\hat{x}=\frac{1}{b-a}\int_{a}^{b} x\,dx=\frac{1}{2(b-a)}(b^2-a^2)=\frac{(b-a)(b+a)}{2(b-a)}=\boxed{\frac{b+a}{2}}$$

Przykład 2 - rozkład dwumianowy

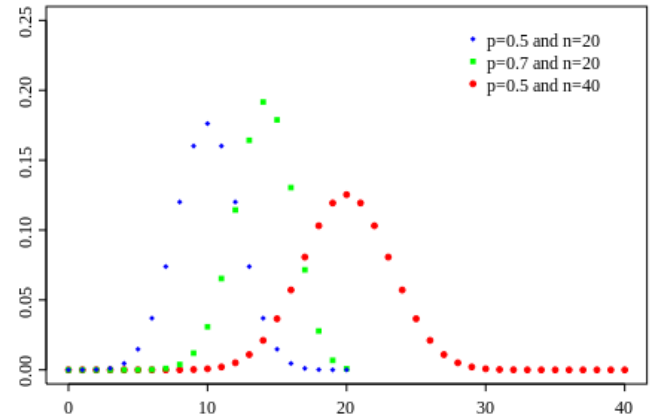
- *ang. binomial distribution*
- Wynik zawsze jedną z dwóch wykluczających się wartości (sukces i porażka)
- Funkcja prawdopodobieństwa:

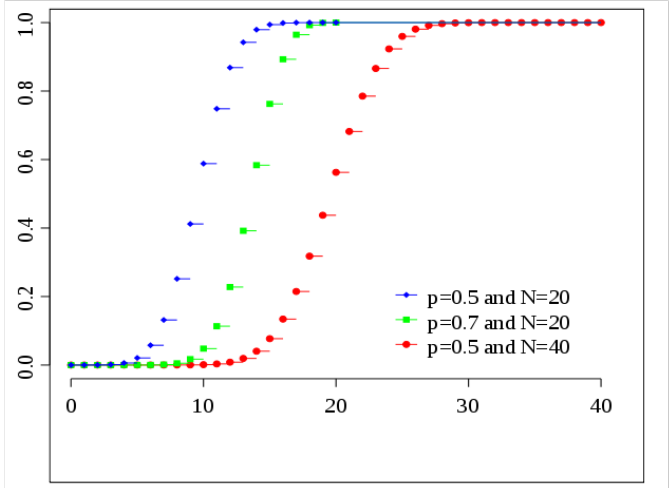
$$p_n(k)=\binom{n}{k}p^k q^{n-k}=\frac{n!}{k!(n-k)!}p^k q^{n-k};\ p\in\langle 0;1\rangle;\ q=1-p$$

  - k sukcesów w n niezależnych próbach przeprowadzonych w identycznych warunkach
  - p – prawdopodobieństwo sukcesu w pojedynczej próbie
  - $q=1-p$ – prawdopodobieństwo porażki w pojedynczej próbie

- Wartość oczekiwana pojedynczej próby x_i:

$$E(x_i)=1\cdot p+0\cdot q=p$$





https://pl.wikipedia.org/wiki/Rozk%C5%82ad_dwumianowy

Wariancja poj. próby x_i:

$$\sigma^2(x_i)=E\left((x_i-p)^2\right)=$$
$$=(1-p)^2\cdot p+(0-p)^2\cdot q=pq$$

Wartość oczekiwana:

$$E(X)=np$$

Wariancja:

$$\sigma^2(X)=npq$$

Przykład 3 - rozkład prędkości wiatru

- Rozkład czestości występowania danej prędkości wiatru opisuje funkcja Weibulla
- Funkcja prawdopodobieństwa:

$$f(v)=\frac{k}{A}\cdot\left(\frac{v}{A}\right)^{k-1}\exp\left[-\left(\frac{v}{A}\right)^{k}\right];v\geq 0$$

  - k, A – parametry rozkładu (otrzymywane z danych dośw.)

- Wartość oczekiwana:

$$E(v)=A\,\Gamma\left(1+\frac{1}{k}\right);\Gamma(z)=\int_0^{\infty}t^{z-1}e^{-t}\,dt$$

- Wariancja:

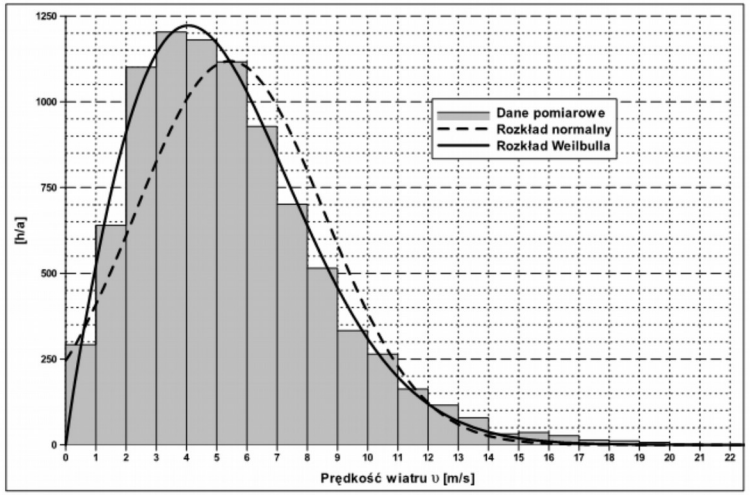
$$\sigma^2(x)=A^2\left[\Gamma\left(1+\frac{2}{k}\right)-\left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right]$$

- Trzeci moment rozkładu prędkości wiatru służy do obliczenia gęstości mocy wiatru:

$$P_w=\frac{1}{2}\rho\int_0^{\infty}v^3 f(v)\,dv$$

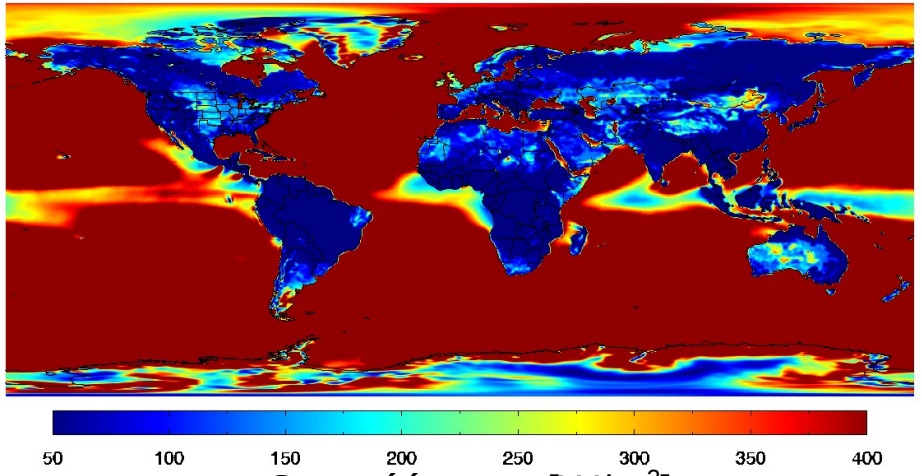
ρ – gęstość powietrza

http://www.ien.pw.edu.pl/EIG/instrukcje/Elektr-EW.pdf



Rys. 2. Histogram prędkości wiatru dla Łeby i zastosowanie rozkładu normalnego oraz Weibulla.

Mapa gęstości mocy wiatru na wys. 10 m

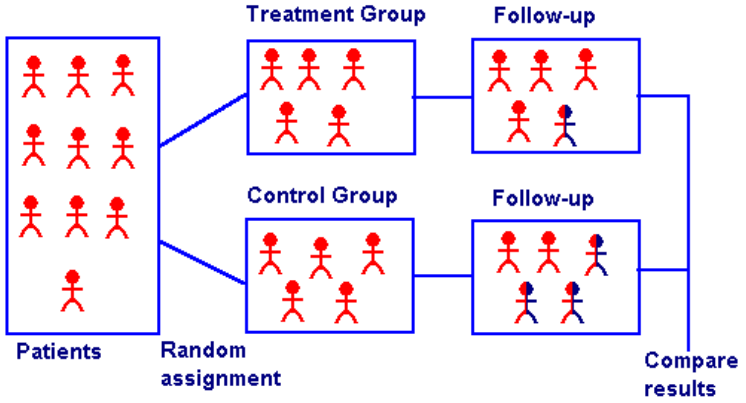


Gęstość mocy [W/m^2]

http://www.renewableenergyst.org/wind.htm

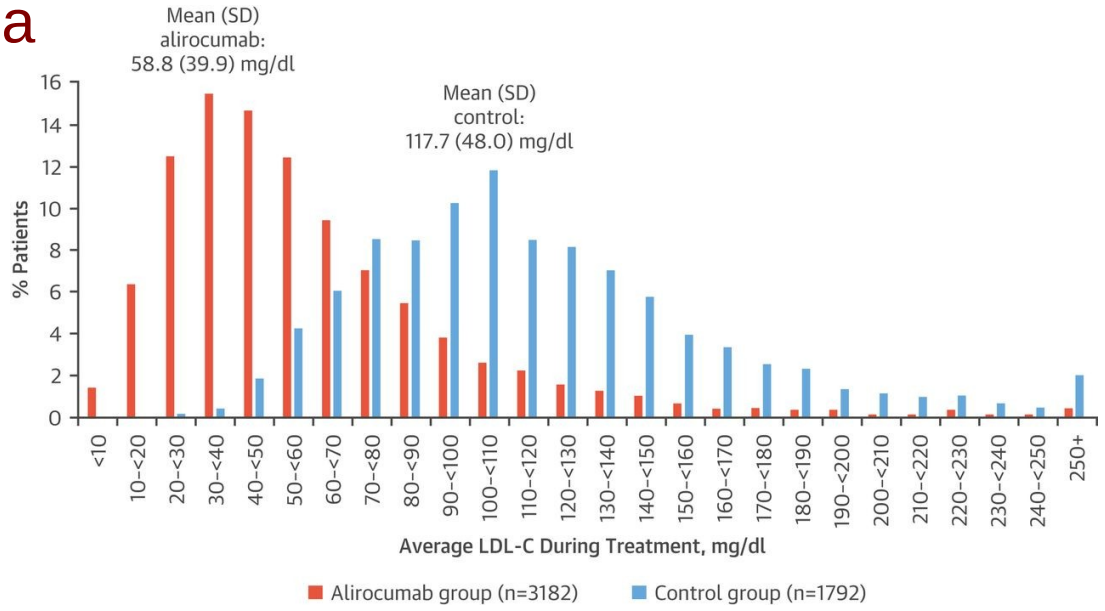
Rozkłady w praktyce - dane medyczne

- Jednym z ważniejszych zastosowań statystyki są **badania medyczne**
- W testach nowych leków wykonuje się badania kliniczne **podwójnie ślepej próby** – **double blinded trial** (ani lekarz ani pacjent nie wiedzą, czy przyjmują lek, czy placebo)



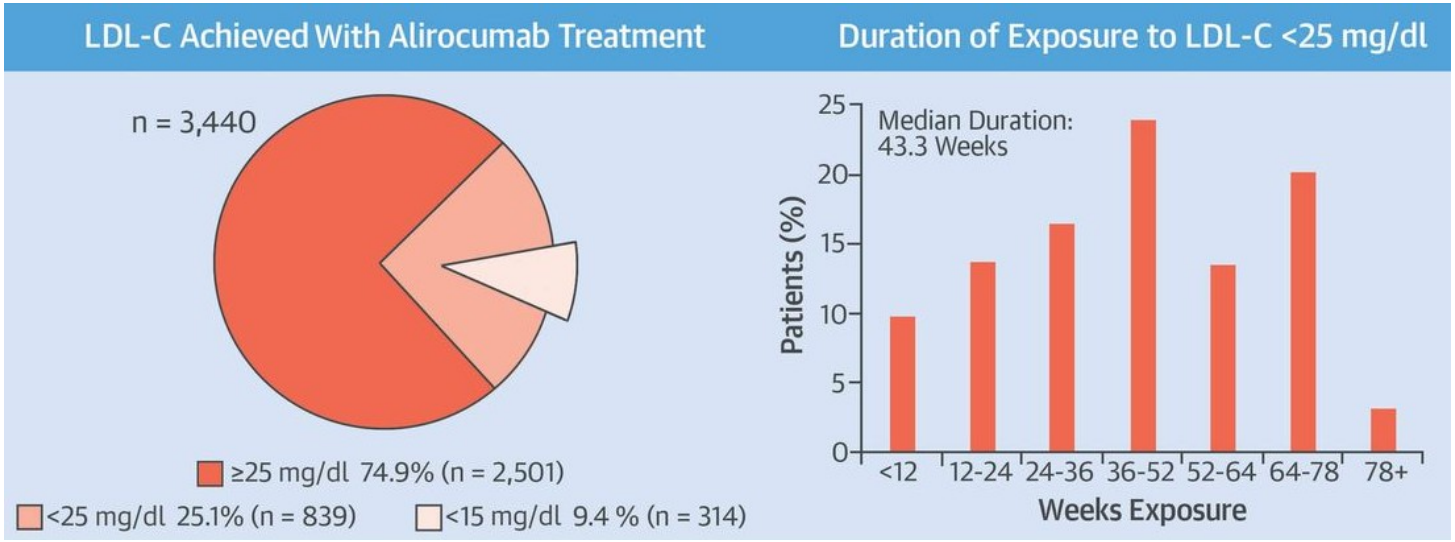
- Przykład:
  - substancja **alirokumab** nazwa handlowa **Praluent**
  - lek stosowany w celu obniżenia dużego stężenia "złego" cholesterolu LDL we krwi
  - średni poziom LDL u pacjetów przyjmujących lek: 58,8 mg/dl
  - średni LDL u pacjentów placebo: 117,7 mg/dl

http://www.onlinejacc.org/content/69/5/471



Rozkłady w praktyce - dane medyczne

- Mediana długości przyjmowania leku wyniosła 78 tygodni
- W badaniu alirokumabu sprawdzanu również pacjentów, którzy uzystaki w trakcie przyjmowania leku stężenie LDL < 25 mg/dl
- W przypadku pacjentów, którzy w przynajmniej dwóch badaniach kontrolnych uzyskali LDL < 25 mg/dl, mediana jego utrzymywania się wynosiła 43,3 tygodnie
- Rozkład utrzymywania się LDL < 25 mg/dl w czasie prezentuje wykres po lewej



http://www.onlinejacc.org/content/69/5/471


WYDZIAŁ FIZYKI
POLITECHNIKA WARSZAWSKA
POLITECHNIKA WARSZAWSKA
WARSAW UNIVERSITY OF TECHNOLOGY

KONIEC

Rozkład dwumianowy

Prawdopodobieństwo: $p_n(k)=\frac{n!}{k!(n-k)!}p^k q^{n-k}$ $p\in[0,1]$ $q=1-p$

Wartość oczekiwana:

$$E(x)=\sum_{k=0}^{n} k\, p_n(k)=\sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k q^{n-k}$$

$$E(x)=np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}q^{n-k}=np(p+q)^{n-1}=np$$

Wariancja:

$$\sigma^2(x)=E(x^2)-(E(x))^2$$

$$E(x^2)=\sum_{k=0}^{n} k^2\frac{n!}{k!(n-k)!}p^k q^{n-k}=\sum_{k=1}^{n}((k-1)+1)k\frac{n!}{k!(n-k)!}p^q q^{n-k}$$

$$E(x^2)=\sum_{k=2}^{n}\frac{n!}{(k-2)!(n-k)!}p^k q^{n-k}+\sum_{k=1}^{n}\frac{n!}{(k-1)!(n-k)!}p^k q^{n-k}$$

$$E(x^2)=n(n-1)p^2\sum_{k=2}^{n}\frac{(n-2)!}{(k-2)!(n-k)!}p^k q^{n-k}+np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}q^{n-k}$$

$$E(x^2)=n(n-1)p^2(p+q)^{n-2}+np(p+q)^{n-1}=n(n-1)p^2+np$$

$$\sigma^2(x)=n(n-1)p^2+np-(np)^2=npq$$

Odchylenie standardowe:

$$\sqrt{\sigma^2(x)}=\sqrt{(npq)}$$