# Komputerowa analiza danych doświadczalnych

Wykład 2
5.03.2021

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

*Semestr letni 2020/2021*

# Prowadzący przedmiot




- **Wykład:**
  - dr inż. Łukasz Graczykowski
    Zakład Fizyki Jądrowej
    pok. 117D, Gmach Fizyki
    lukasz.graczykowski@pw.edu.pl
- **Laboratoria:**
  - dr inż. Łukasz Graczykowski – ROOT
  - dr inż. Anna Chmiel – Matlab
  - mgr inż. Paweł Szymański – ROOT
  - mgr inż. Maria Stefaniak – ROOT
- **Strona przedmiotu:**
  http://www.if.pw.edu.pl/~lgraczyk/wiki/index.php/KADD_2020/2021
- **Konsultacje:** MS Teams

# Sprawy organizacyjne

- **Organizacja zajęć:**
  - wykłady: 1h w tygodniu (15h w semestrze)
  - laboratoria: 2h w tygodniu (30h w semestrze)
- **Laboratoria:**
  - 15 zajęć: 1 wstępne, 11 punktowanych, 2 kolokwia, 1 dodatkowe
  - prowadzone w środowisku ROOT lub Matlab
  - na zajęciach wyjaśnienie nowego zadania oraz indywidualne oddawanie rozwiązań z poprzednich zajęć
  - czas na wykonanie zadania – **48h** (z wyłączeniem weekendów/świąt)
- **Wykłady:**
  - piątki 12:15-13:00, MS Teams

# Warunki zaliczenia (1)

- **Laboratorium:**
  - 11 punktowanych zadań o zróżnicowanym stopniu trudności (**0-5 pkt**)
  - dopuszczenie do wykonania zadania może być warunkowane zaliczeniem kolokwium wstępnego (“wejściówki”)
  - w trakcie pisania programu można korzystać z **własnych** programów i zasobów Internetu
  - program **w pełni** dokończony w domu: **+1 pkt**, ale **max 4 pkt**

# Warunki zaliczenia (2)

- **Nieobecności na laboratorium:**
  - nieobecność nieusprawiedliwiona: **0 pkt.**
  - nieobecność usprawiedliwiona: **max 4 pkt.**
    konieczność zrealizowania materiału we własnym zakresie i przedstawienia na najpóźniej 2 tygodnie od powrotu (na zajęciach lub konsultacjach)
  - maksymalna liczba nieobecności nieuspr.: **2** (w rzypadku dłuższej nieobecności usprawiedliwionej, np. choroba – warunki zaliczenia ustalane będą indywidualnie)

# Warunki zaliczenia (3)

- **Kolokwia na laboratorium:**
  - punktaja: **0-30 pkt** za **każde** kolokwium
  - w trakcie semestru przewidziane są **2 kolokwia**
  - napisanie **3 programów/makr** z materiału zrealizowanego na zajęciach
  - każdy program/makro punktowany **0-10 pkt.**
- **Kolokwium na wykładzie:**
  - punktacja: **0-35 pkt.**
  - kolokwium z wiedzy (pisemne) na wykładzie
- **Zaliczenie kolokwiów: >50% pkt.**
- **Poprawa kolokwium z lab.:**
  - możliwa jednokrotnie, wynik poprawy **zastępuje** wynik regularny
  - tylko na ostatnich (15) zajęciach, **max 24 pkt.** (8 pkt. za zadanie)

# Warunki zaliczenia (4)

- **Punktacja:**
  - maksymalna ilość punktów: **150**
    - laboratoria: **11*5 = 55**
    - kolokwia (lab.): **2*30 = 60**
    - kolokwium (wykł.): **1*35 = 35**
- **Zaliczenie (procent sumy punktów):**
  - >50% - 3 (75,5 pkt. – 90,0 pkt.)
  - >60% - 3,5 (90,5 pkt. – 105,0 pkt.)
  - >70% - 4 (105,0 pkt. – 120,0 pkt.)
  - >80% - 4,5 (120,5 pkt. – 135,0 pkt.)
  - >90% - 5 (135,5 pkt. – 150,0 pkt.)
- **Uwaga! Do zaliczenia przedmiotu konieczne jest zaliczenie (>50% punktów) wszystkich kolokwiów**

# Literatura


1. **S. Brand, “Analiza danych: metody statystyczne i obliczeniowe”, PWN, Warszawa (1998)**
2. R. Nowak, “Statystyka dla fizyków”, PWN, Warszawa (2002)
3. W.T.Eadie, D.Drijard, F.E.James, M.Ross, B.Sadoulet, “Metody statystyczne w fizyce doświadczalnej”, PWN, Warszawa (1989)
4. A.Plucińska, E.Pluciński, “Elementy probabilistyki”, PWN, Warszawa (1979)
5. Przykładowe programy i dokumentacja środowiska ROOT i Matlab

# Program wykładu

1. Pomiary w eksperymentach fizycznych (przypomnienie z rachunku niepewności).
2. Zmienne losowe i ich rozkłady (1D, 2D, nD).
3. Elementy metody Monte Carlo, generacja liczb pseudolosowych za pomocą komputera.
4. Podstawowe rozkłady statystyczne (dyskretne i ciągłe; centralne twierdzenie graniczne).
5. Pomiar jako pobieranie próby. Estymatory.
6. Metoda największej wiarygodności.
7. Weryfikacja hipotez statystycznych (m. in. test $\chi^2$)
8. Metoda najmniejszych kwadratów (przypadek liniowy, wielomianowy, ...)
9. Zagadnienie minimalizacji i optymalizacji.

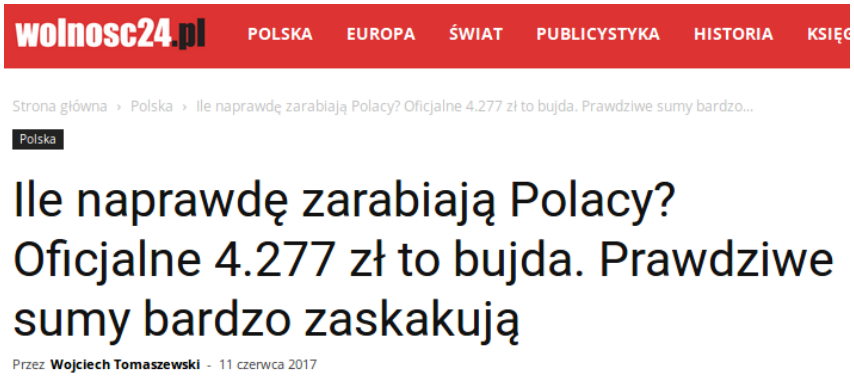
# Po co nam to wszystko?

# Po co nam to wszystko?

Czy te nagłówki i komentarze zawierają słuszne zarzuty?

Publikacje z 2017

https://superbiz.se.pl/wiadomosci/szokujace-dane-gus-ile-naprawde-zarabiaja-polacy-aa-TC9b-tWr2-FRHU.html


https://finanse.wp.pl/polacy-czesto-zarabiaja-mniej-niz-pokazuja-dane-to-wina-statystyki-6192223382357633a



https://gospodarka.dziennik.pl/praca/artykuly/566899,zarabiamy-jak-nigdy-rekordowe-place-w-polsce.html

# Po co nam to wszystko?

Czy my żyjemy w innej rzeczystości?

SUPER BIZ.pl | WIADOMOŚCI BIZ | FIRMA | PRAWO PRZEDSIĘBIORCY | TECHNOLOGIE

SuperBiz / wiadomości / Szokujące dane G REKLAMA aprawdę zarabiają Polacy?

## Szokujące dane GUS! Ile naprawdę zarabiają Polacy?

24.11.2017, godz. 12:40

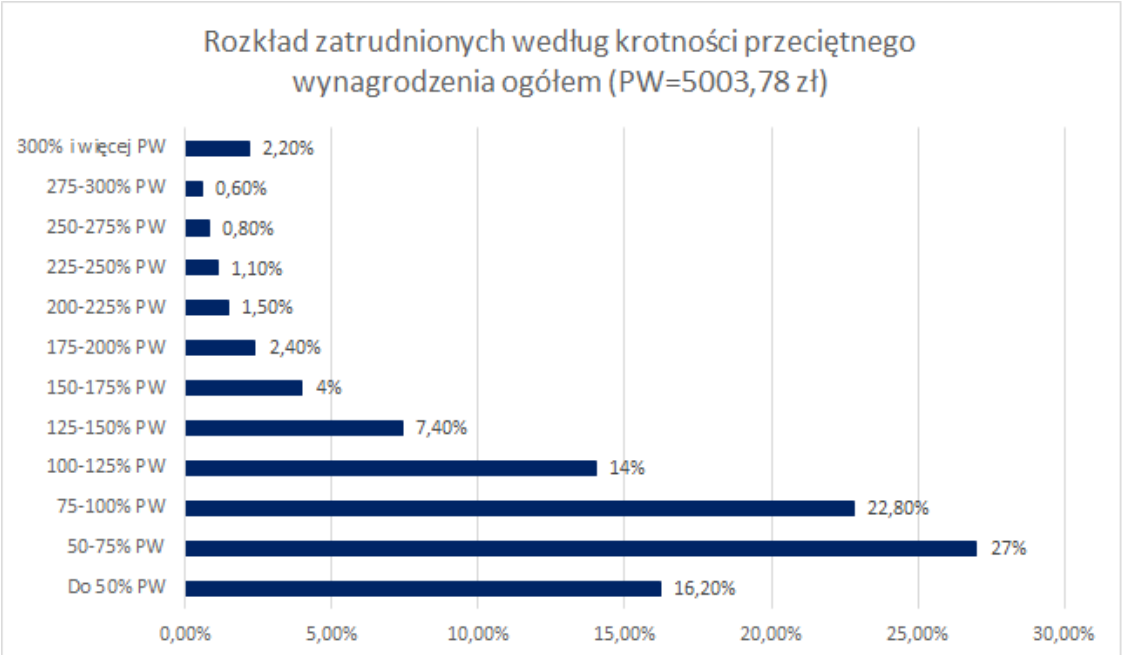
Lubię to! 56 Tweetnij G+ nk Fajne! 0


**Ponad 4,5 tys. złotych wyniosło w październiku przeciętne miesięczne wynagrodzenie brutto w sektorze przedsiębiorstw według Głównego Urzędu Statystycznego. Jednak najnowsze badanie, pokazujące strukturę wynagrodzeń według zawodów w październiku 2016 roku, ukazuje zgoła inną rzeczywistość. Wynika z nich, że najczęstsza pensja w Polsce wyniosła nieco ponad 2 tys. zł brutto, czyli ok. 1,5 tys. zł „na rękę".**

Publikacja z 2017

# Po co nam to wszystko?

Czy my żyjemy w innej rzeczystości?




GUS 2019 za rok 2018

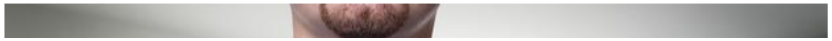
SUPER BIZ.pl | WIADOMOŚCI BIZ | FIRMA | PRAWO PRZEDSIĘBIORCY | TECHNOLOGIE

SuperBiz / wiadomości / Szokujące dane G… aprawdę zarabiają Polacy?

## Szokujące dane GUS! Ile naprawdę zarabiają Polacy?

24.11.2017, godz. 12:40

Lubię to! 56 · Tweetnij · G+ · nk Fajne! 0



**Robert Sosnowiecki**
@Wad_emecum

Follow

Replying to @GUS_STAT

dzlaczego akurat takie wartości dzielące przedziały (i skąd te wartości), a nie np prosty rozkład na decyle po 10% w każdym???

1:23 AM - 23 Nov 2017

**2** Likes

1 · 2


**Ponad 4,5 tys. złotych wyniosło w październiku przeciętne miesięczne wynagrodzenie brutto w sektorze przedsiębiorstw według Głównego Urzędu Statystycznego. Jednak najnowsze badanie, pokazujące strukturę wynagrodzeń według zawodów w październiku 2016 roku, ukazuje zgoła inną rzeczywistość. Wynika z nich, że najczęstsza pensja w Polsce wyniosła nieco ponad 2 tys. zł brutto, czyli ok. 1,5 tys. zł „na rękę".**

Publikacja z 2017

# Po co nam to wszystko?

https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5474/5/6/1/struktura_wynagrodzen_wedlug_zawodow_w_pazdzierniku_2018.pdf

- W październiku 2018 r. przeciętne miesięczne wynagrodzenie ogółem brutto dla jednostek, o liczbie pracujących powyżej 9 osób, wyniosło 5003,78 zł.
- Przeciętne godzinowe wynagrodzenie ogółem brutto wyniosło 27,79 zł.
- Połowa zatrudnionych pracowników otrzymała wynagrodzenie ogółem brutto do 4094,98 zł (mediana = decyl piąty = wynagrodzenie środkowe).
- 10% najniżej zarabiających pracowników otrzymało wynagrodzenie ogółem brutto co najwyżej w wysokości 2224,17 zł (decyl pierwszy).
- 10% najwyżej zarabiających pracowników otrzymało wynagrodzenie ogółem brutto co najmniej w wysokości 8239,84 zł (decyl dziewiąty).
- Miesięczne wynagrodzenie brutto w wysokości 60 tys. zł i więcej otrzymało niewiele ponad 0,04% zatrudnionych, natomiast: od 50 tys. zł – 0,06%, od 40 tys. zł – 0,12%, od 30 tys. zł – 0,30%, od 20 tys. zł – 0,94%, od 10 tys. zł – 6,23%.
- 6,1% ogółu zatrudnionych otrzymało miesięczne wynagrodzenie brutto co najmniej równe dwukrotnemu przeciętnemu wynagrodzeniu miesięcznemu ogółem brutto (tzn. ≥ 10007,56 zł).
- Pracownicy otrzymujący miesięczne wynagrodzenie ogółem brutto mniejsze lub równe przeciętnemu wynagrodzeniu miesięcznemu ogółem brutto (tzn. ≤ 5003,78 zł) stanowili niemalże 2/3 ogółu pracowników (66,0%).
- 16,2% zatrudnionych zarabiało co najwyżej 50% przeciętnego wynagrodzenia miesięcznego brutto, czyli nie więcej niż 2501,89 zł.
- Co 13-ty zatrudniony (7,6% ogółu) otrzymał miesięczne wynagrodzenie brutto co najwyżej równe wysokości minimalnego wynagrodzenia za pracę, tj. 2100,00 zł, ustalanego na podstawie przepisów o minimalnym wynagrodzeniu za pracę.

GUS 2019 za rok 2018

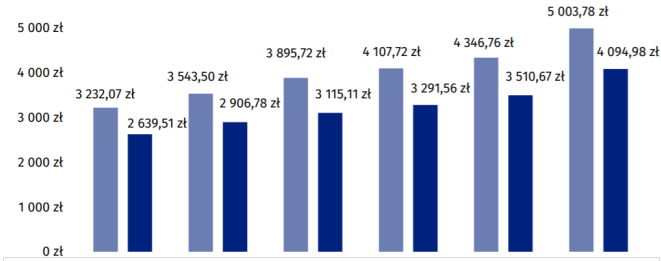
GUS

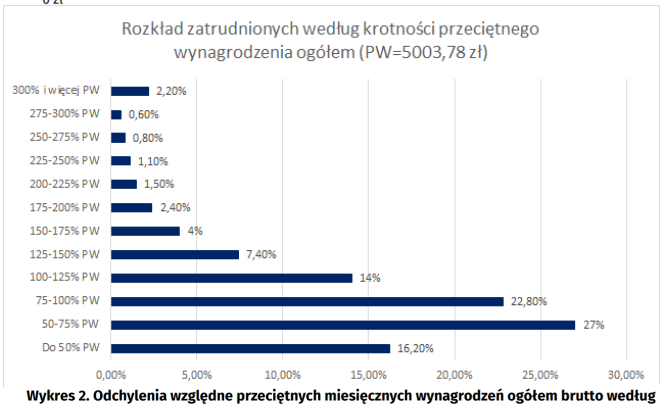
## Struktura wynagrodzeń według zawodów w październiku 2018 roku

**4094,98 zł**
Mediana wynagrodzeń miesięcznych

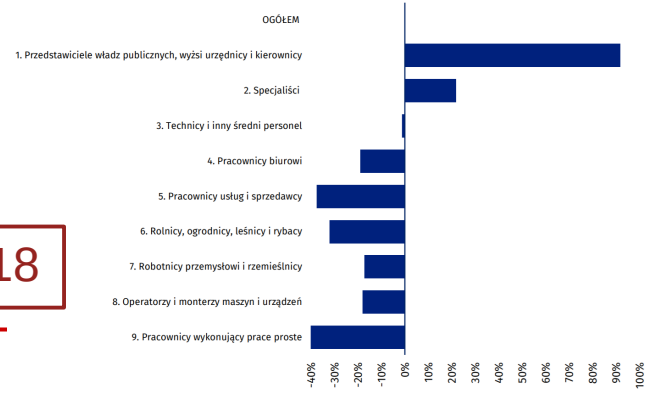
**W październiku 2018 r. środkowe miesięczne wynagrodzenie ogółem[1] brutto dla osób zatrudnionych w jednostkach o liczbie pracujących powyżej 9 osób[2] wyniosło 4094,98 zł. W porównaniu z październikiem 2016 r. mediana była większa o 584,31 zł, tj. o 16,6%.**

**Wykres 1. Przeciętne i środkowe wynagrodzenia ogółem brutto (w zł) za październik**




**Wykres 2. Odchylenia względne przeciętnych miesięcznych wynagrodzeń ogółem brutto według „wielkich" grup zawodów od przeciętnego wynagrodzenia ogółem brutto w październiku 2018 r.**



# Przykłady na czasie

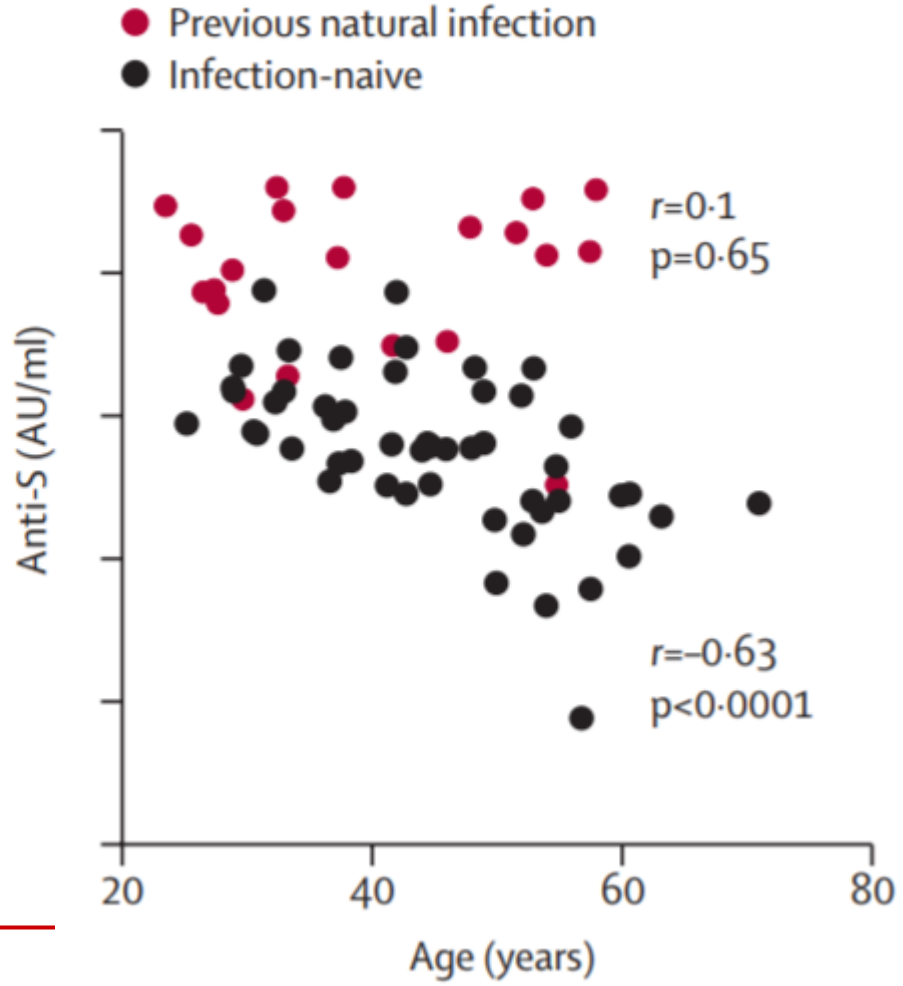
https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00502-X/fulltext

M. Prandecki *et al.*, Lancet, S0140-6736 (2021) 00502-X

THE LANCET

## Effect of previous SARS-CoV-2 infection on humoral and T-cell responses to single-dose BNT162b2 vaccine

Infection-naive individuals showed an inverse correlation between post-vaccination anti-S titre and age (figure B), with individuals older than 50 years generating a significantly weaker serological response than those younger than 50 years (median 230·1 AU/mL *vs* 888·9 AU/mL, p<0·0001; figure A). This correlation was not seen in the group with previous natural infection (figure B).




INDEPENDENT

**Macron claims Oxford-AstraZeneca Covid vaccine 'quasi-ineffective' on older people**

French president criticises UK's rollout strategy amid row over EU delay

Peter Stubley | Friday 29 January 2021 21:36 | 38 comments

ROBERT KOCH INSTITUT

Tabelle 11: Effektivitätsdaten von AZD1222

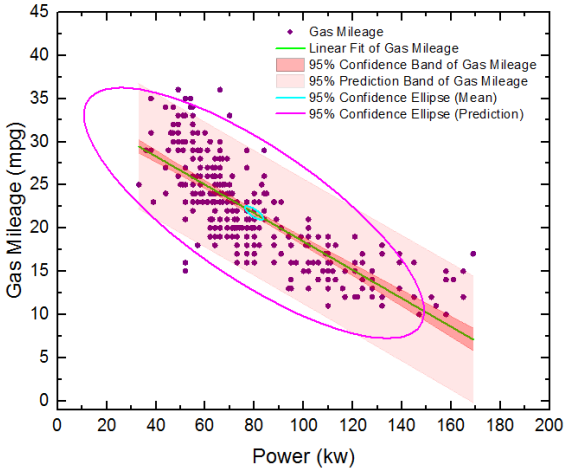
| Endpunkt | | AZD1222 (n/N) | Kontrollen (n/N) | Impfeffektivität (%) | 95% Konfidenzintervall |
|---|---|---|---|---|---|
| | **COVID-19** | | | | |
| Alle | | 30/5.807 | 101/5.829 | 70,4 | 54,8-80,6 |
| 18-64 Jahre | | 29/5.466 | 100/5.510 | 71,1 | 56,3-80,9 |
| ≥65 Jahre | | 1/341 | 1/319 | 6,3 | -1405-94,2 |

35

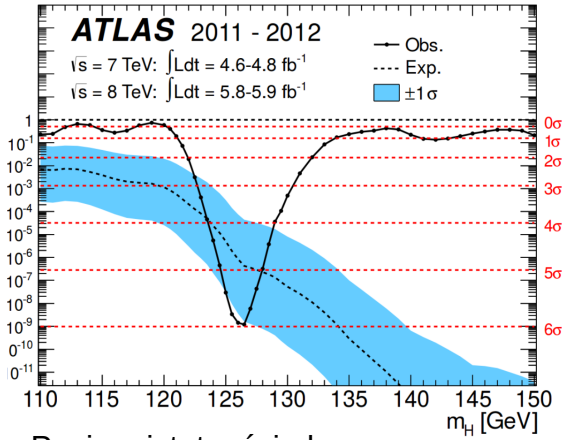
Living Guideline der STIKO zur COVID-19-Impfung und wissenschaftliche Begründung (26.01.2021)

# Po co nam to wszystko?

1. Poprawne opracowanie danych jest niezbędne niemal we wszystkich badaniach empirycznych.
2. Metody analizy danych są w zasadzie bardzo zbliżone niezależnie od dziedziny (fizyka, elektronika, dane medyczne, bankowe, psychologia, itp.)
3. Umiejętność czytania i <u>rozumienia</u> publikacji naukowych.
4. Umiejętność przedstawiania danych i wyników ich analizy (**wykresy**).

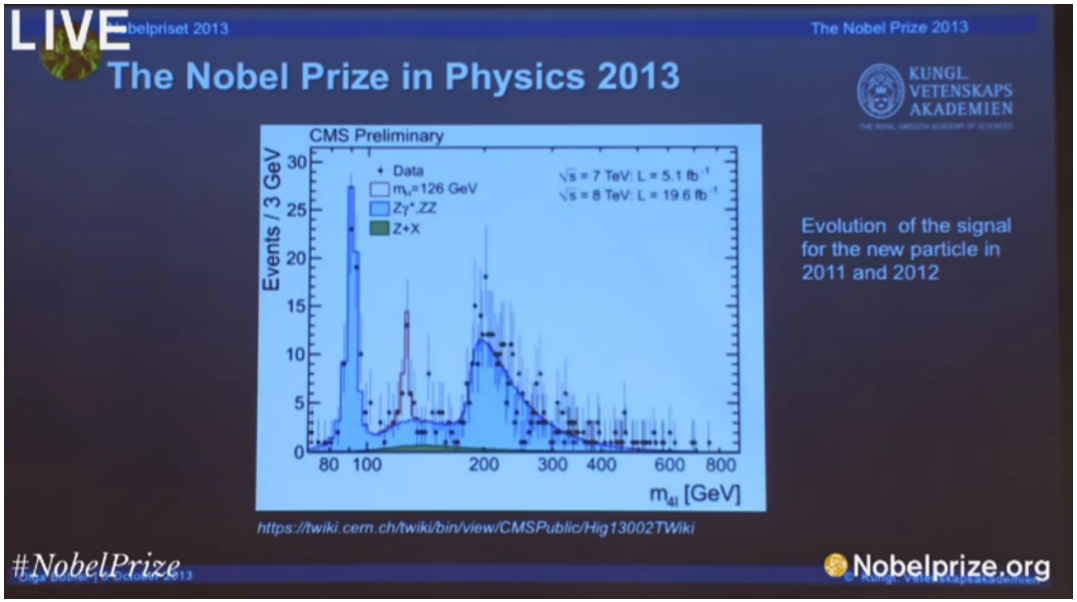



Phys.Lett. B716 (2012) 1-29



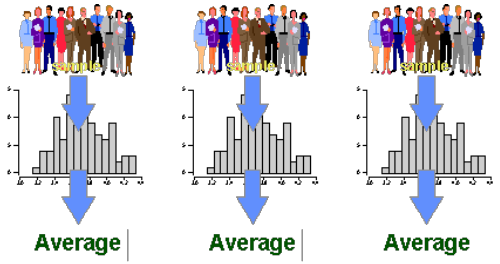

Poziom istotności obserwowanego sygnału bozonu Higgsa.

http://www.nobelprize.org/mediaplayer/index.php?id=1954



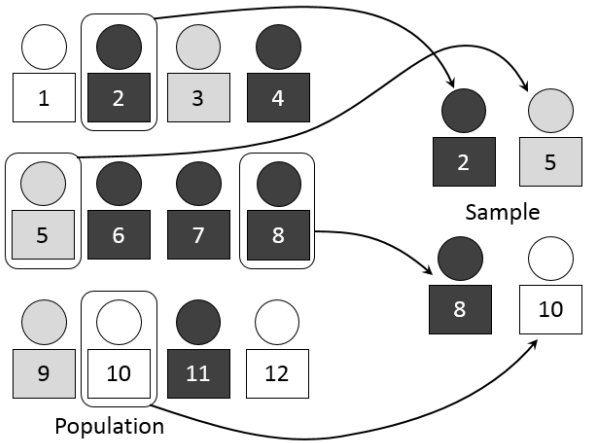
# Statystyczna analiza danych

- Statystyczna analiza danych (to o czym będziemy się uczyć):
  - traktujemy pomiar jako pewien element zbioru wszystkich możliwych pomiarów (pewnej cechy **populacji** o danym rozkładzie prawdopodobieństwa – najczęściej nieznamym)
  - na podstawie skończonej liczby pomiarów, obserwacji (**próby losowej**, podzbioru populacji), która ma swój rozkład prawdopodobieństwa (znany z pomiarów czy obserwacji), próbujemy dowiedzieć się czegoś (czyli **estymować**) na temat parametrów rozkładu całej populacji
  - innymi słowy, na podstawie próby losowej (pomiarów, obserwacji) stawiamy hipotezy i wyciągamy wnioski dotyczące interesującej nas cechy całej populacji




https://www.proprofs.com/quiz-school/story.php?title=3dq-sampling-distributions

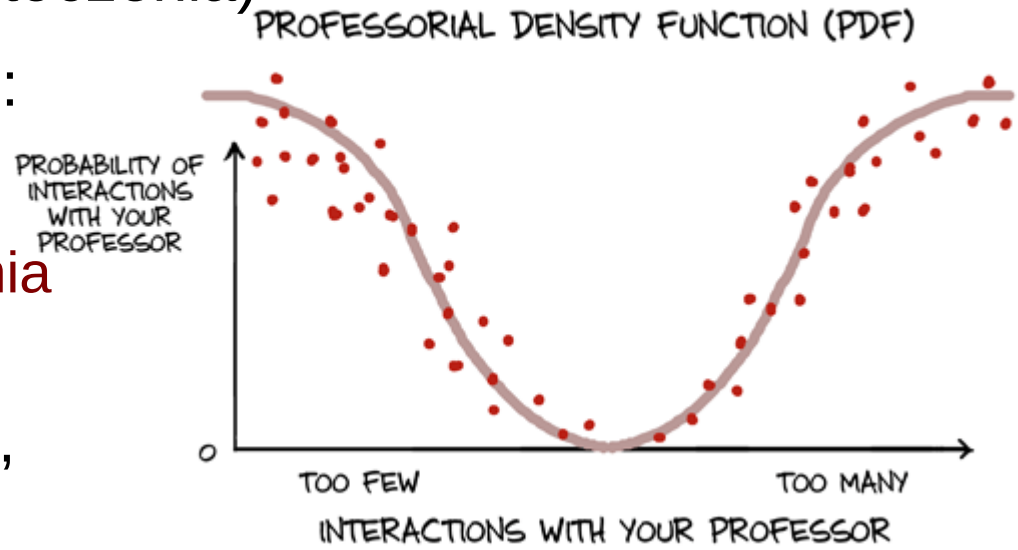
https://en.wikipedia.org/wiki/Sample_(statistics)



# Zmienne losowe, jednowymiarowe rozkłady zmiennych losowych

# Pomiar jako zdarzenie losowe

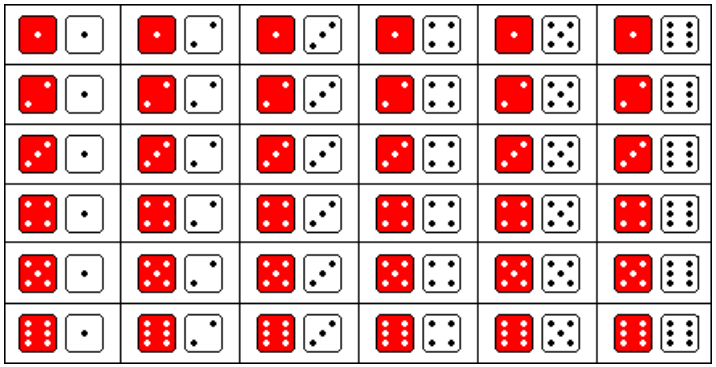
- Wyniki kolejnych pomiarów jakiegoś zjawiska, niezależnie od tego jak bardzo byśmy się starali przestrzegać procedury pomiarowej, będą różne (raz mniejsze, raz większe) – oczywiście zakładając wysoką precyzję przyrządu pomiarowego (patrz Wykład 1)
- Może to wynikać zarówno ze **statystycznego charakteru 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



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

# Typy i rodzaje zmiennych losowych

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

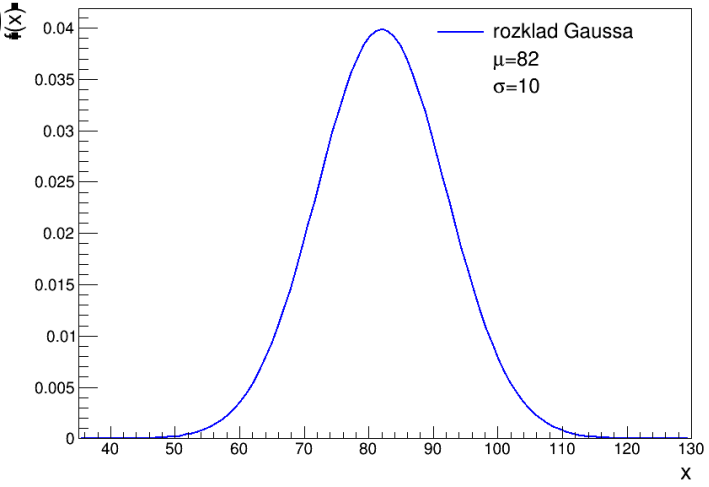


https://mosaicprojects.files.wordpress.com/2013/01/diceposs.gif

# Rozkład i dystrybuanta zmiennej losowej

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

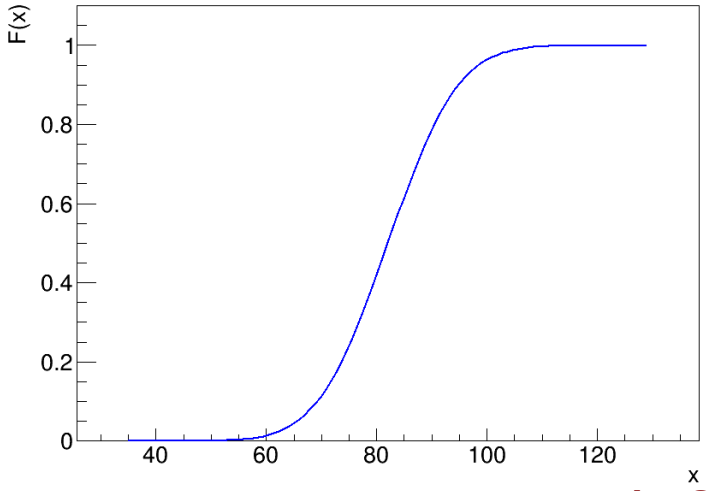
$$f(x)=P(X=x)$$




  - rozkład prawdopod. jest <u>unormowany</u>
    - 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 $x$:

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




  - rozkład ciągły: $F(x)=\int_{-\infty}^{x} f(x')\,dx'$
  - 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:

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

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

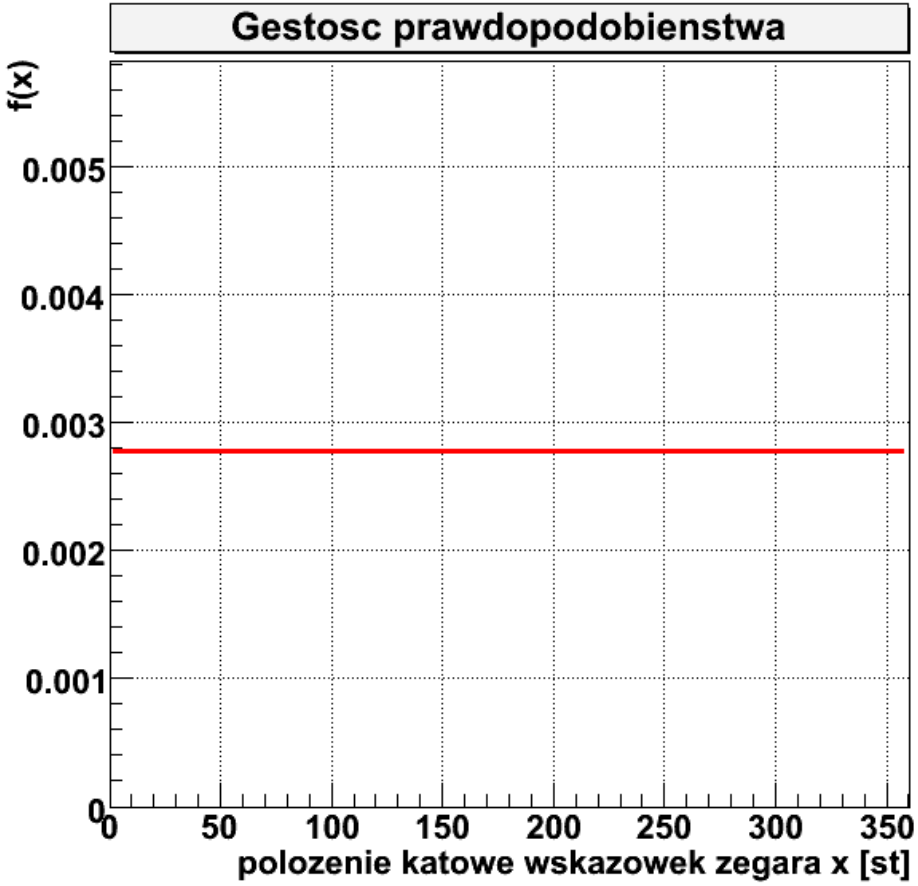
# Rozkład i dystrybuanta - przykłady

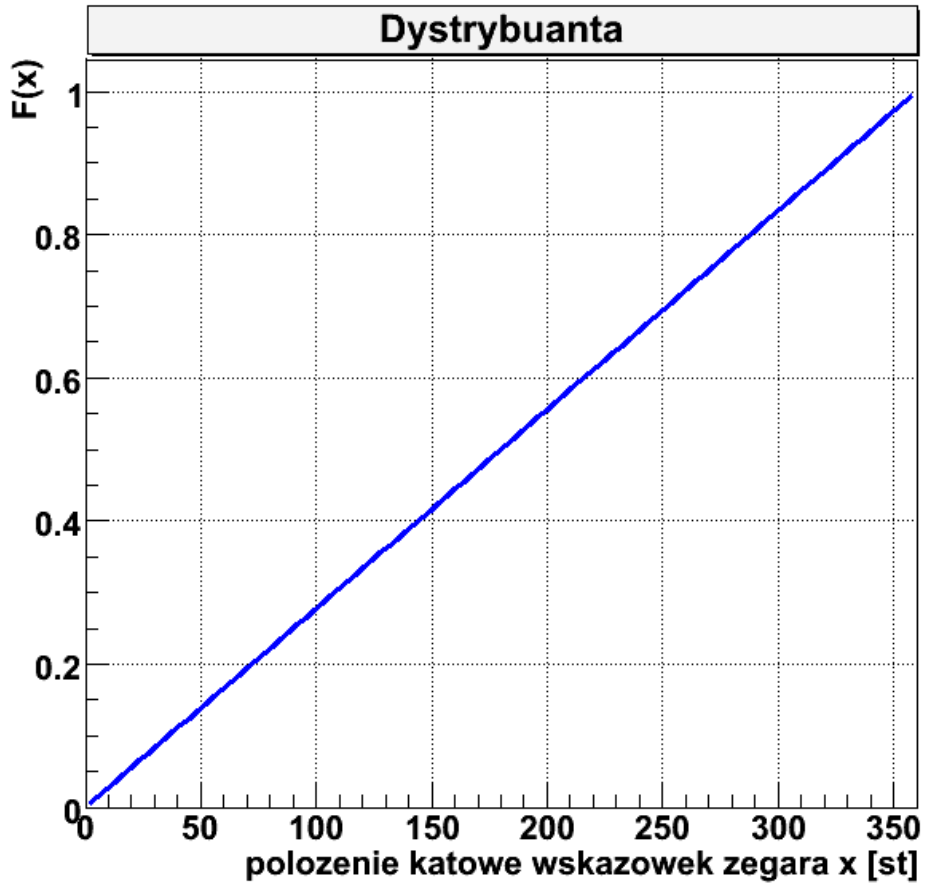
- Rozkład położenia kątowego wskazówki zegara (rozkład jednorodny, lub jednostajny) – zmienna losowa ciągła:

$$f(x)=\frac{1}{360};\ x\in\langle 0;360\rangle \qquad f(x)=0;\ x\in\mathbb{R}\setminus\langle 0;360\rangle$$

$$F(x)=0;\ x<0 \qquad F(x)=1;\ x>360$$

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



# Rozkład i dystrybuanta - przykłady

- Rozkład <u>normalny</u> (rozkład Gaussa) – zmienna losowa ciągła:

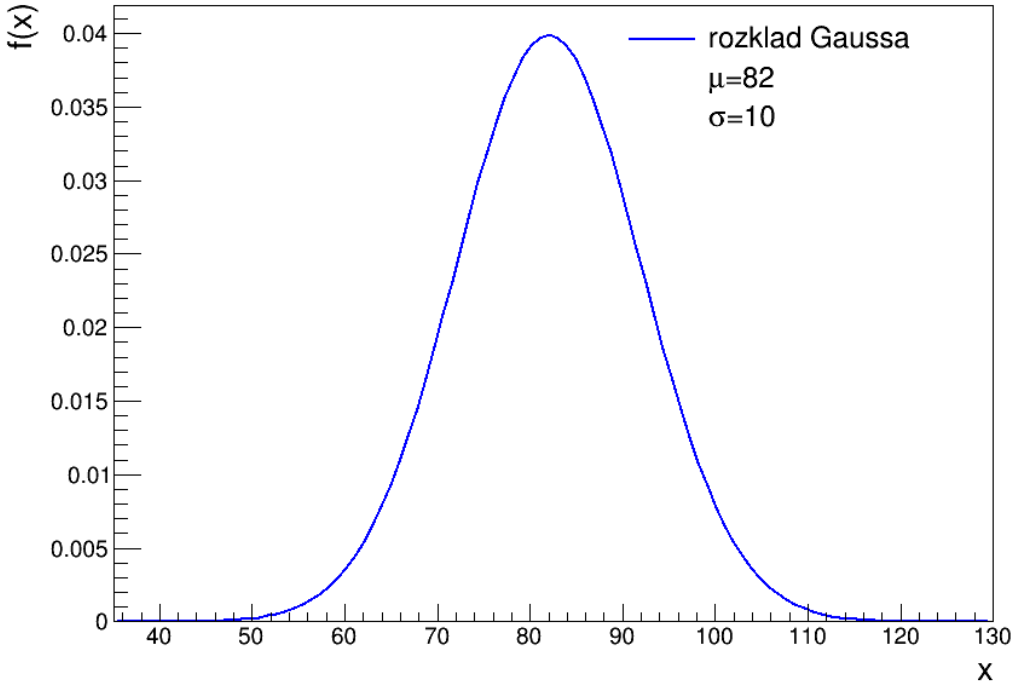
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), x\in\mathbb{R}$$

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

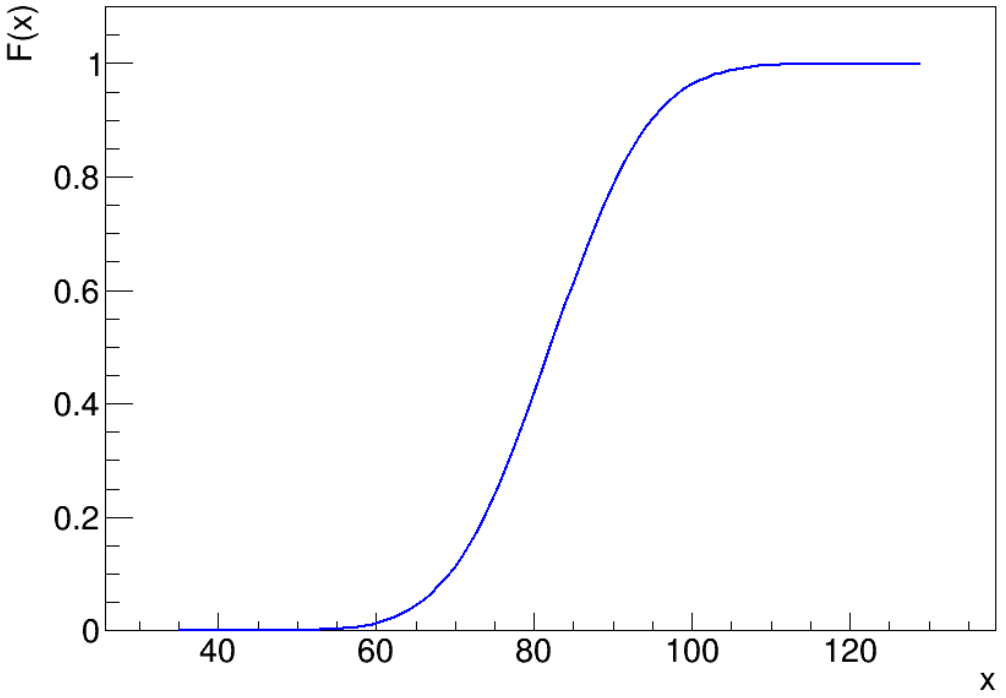
$$\lim_{x\to-\infty} F(x)=0 \qquad \lim_{x\to\infty} F(x)=1$$

Dystrybuanta rozkładu normalnego **nie ma** postaci analitycznej

Rozkład (funkcja gęstości)




Dystrybuanta

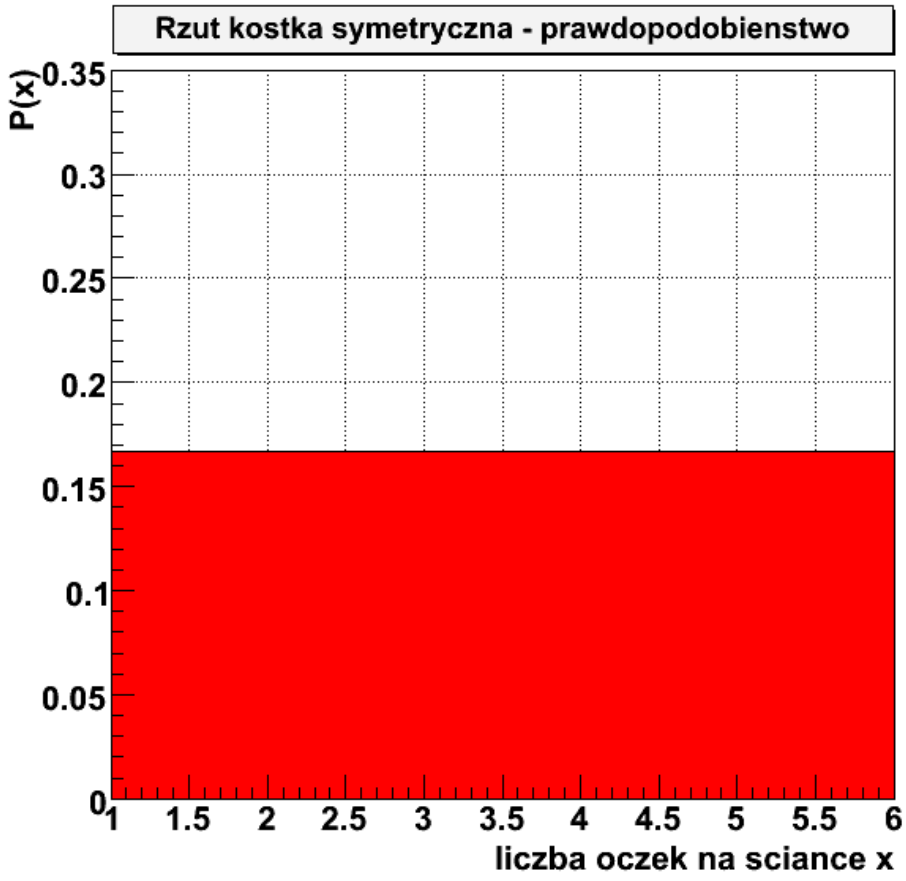


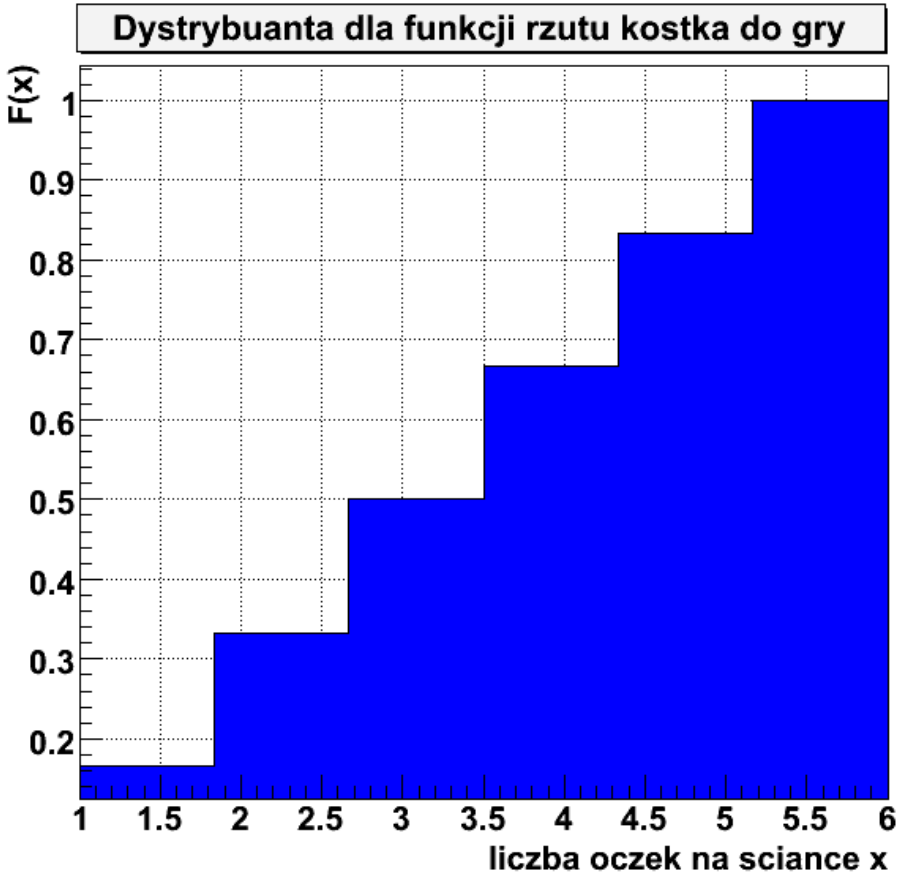
# Rozkład i dystrybuanta - przykłady

- Rzut kostką – zmienna losowa dyskretna:

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

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



# Rozkład i dystrybuanta - przykłady

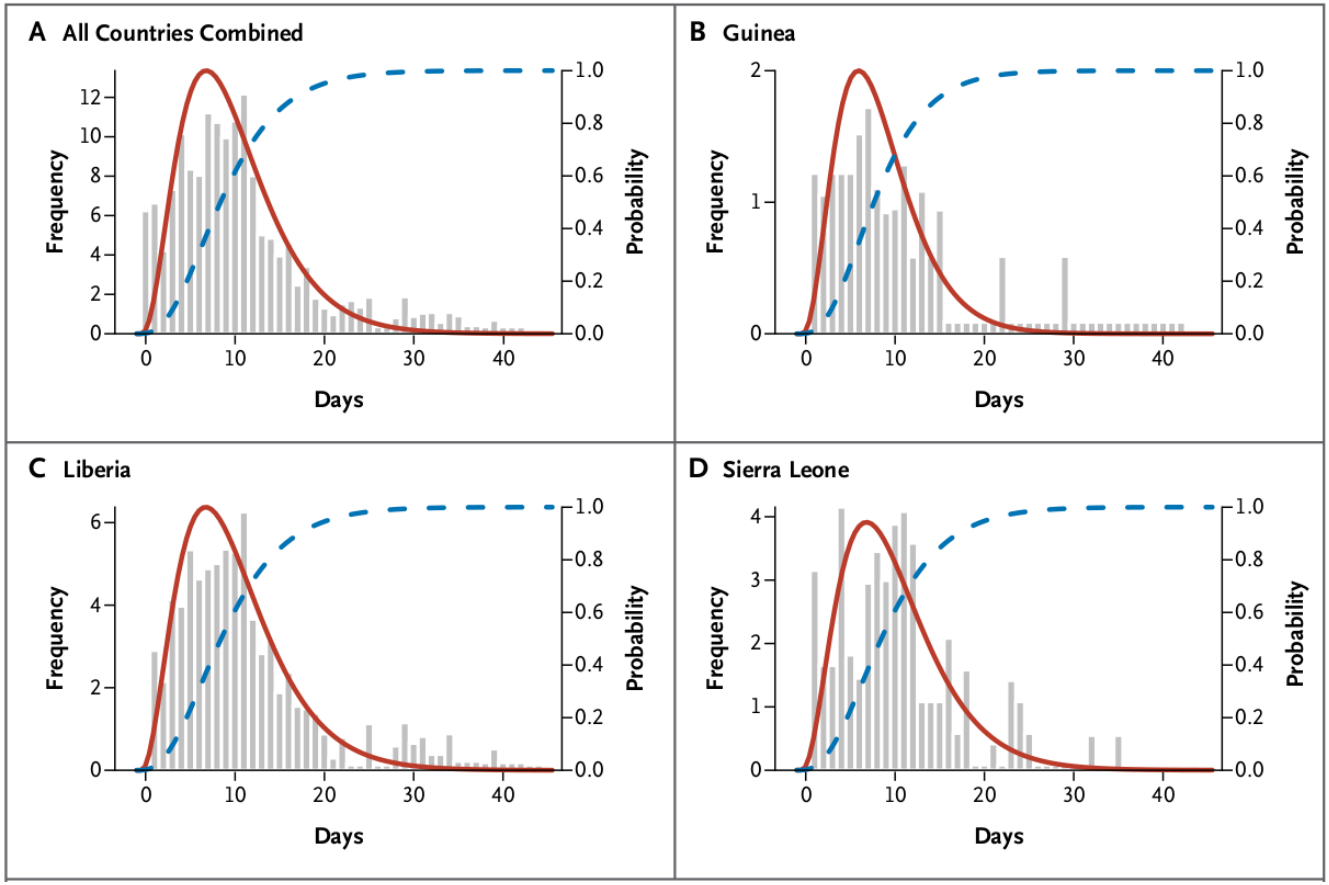
The NEW ENGLAND JOURNAL of MEDICINE

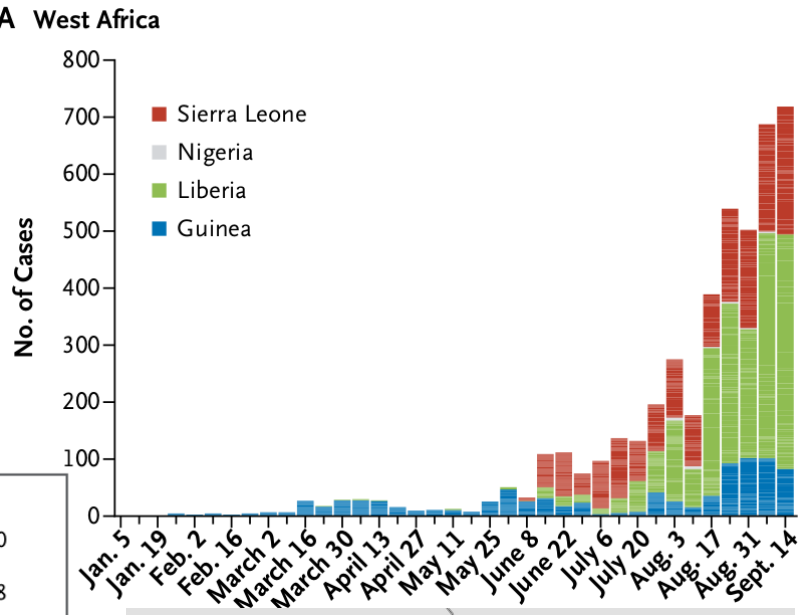
ESTABLISHED IN 1812 OCTOBER 16, 2014 VOL. 371 NO. 16

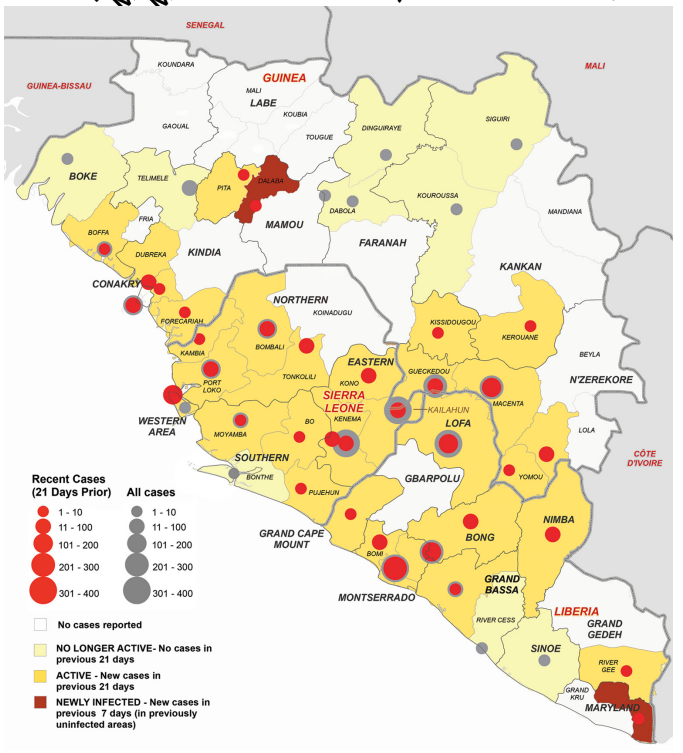
Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections

WHO Ebola Response Team*

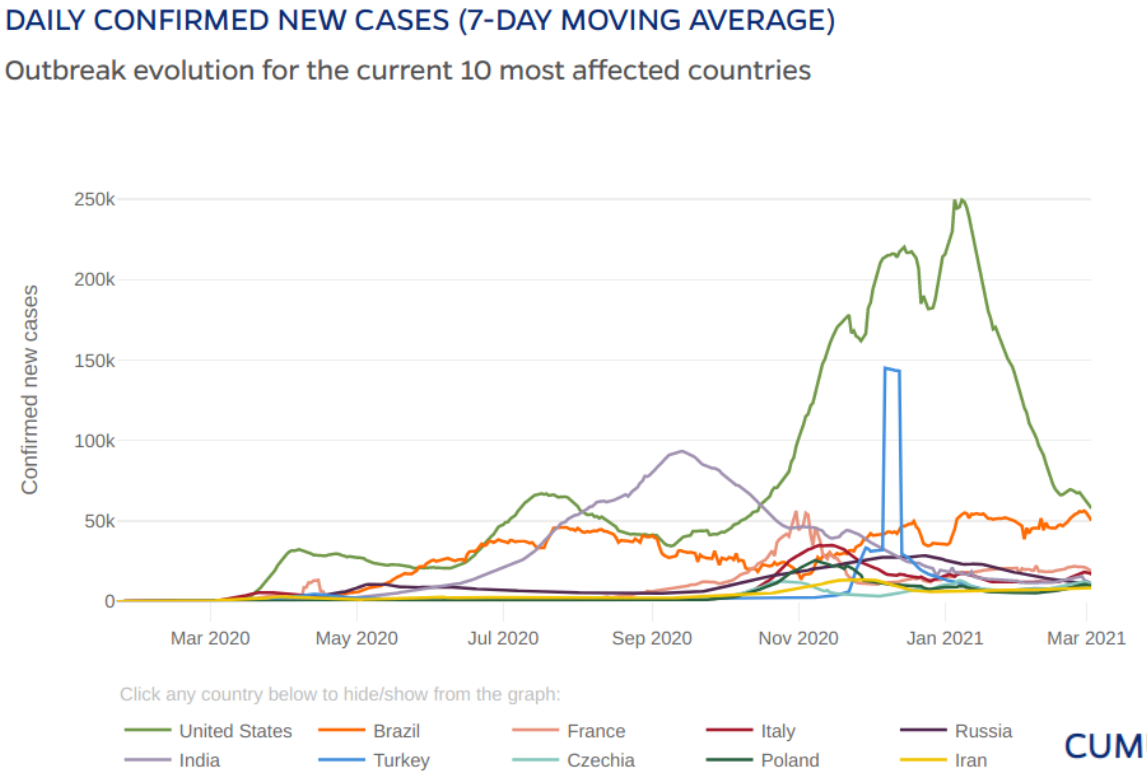
**Ebola**



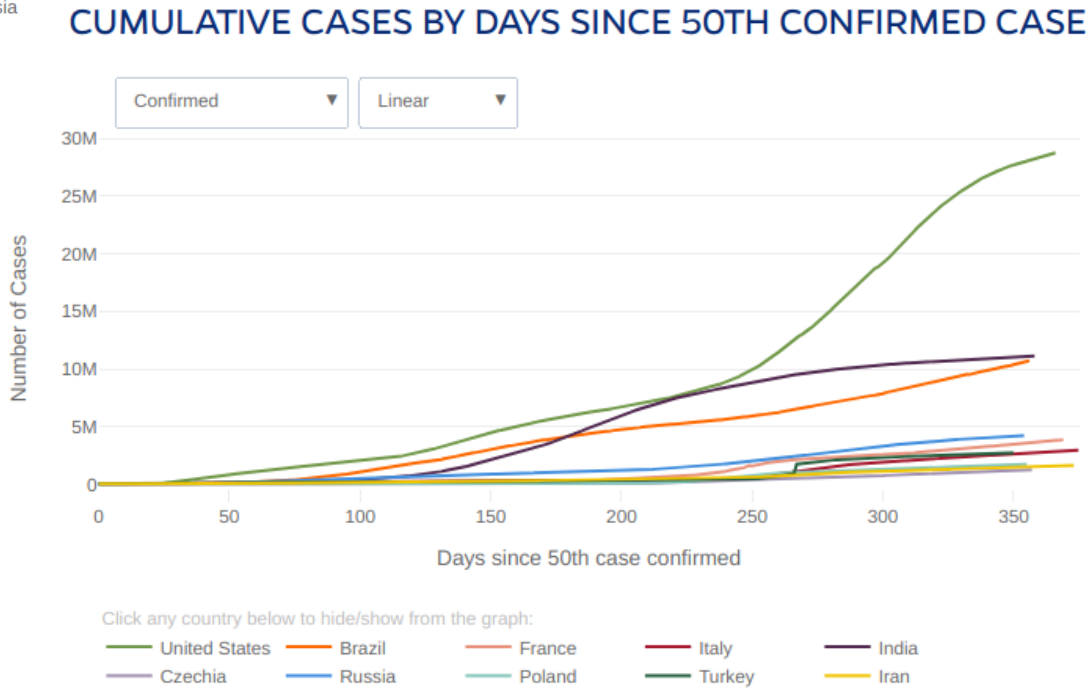
# Rozkład i dystrybuanta - przykłady



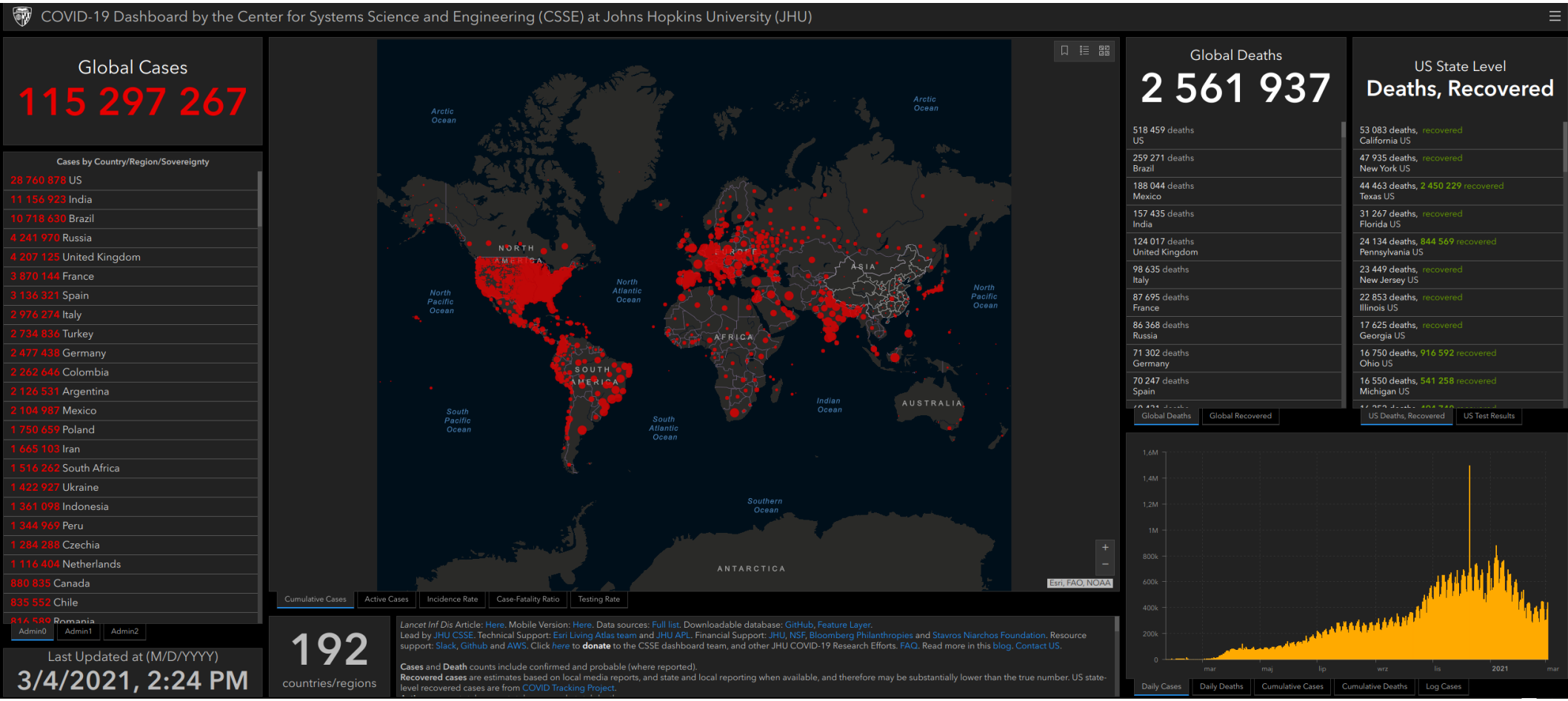

**Koronawirus SARS-CoV-2**

JOHNS HOPKINS UNIVERSITY & MEDICINE | CORONAVIRUS RESOURCE CENTER

https://coronavirus.jhu.edu/



# Rozkład i dystrybuanta - przykłady




https://coronavirus.jhu.edu/map.html

# Funkcje zmiennej losowej, wartość oczek.

- 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:

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

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

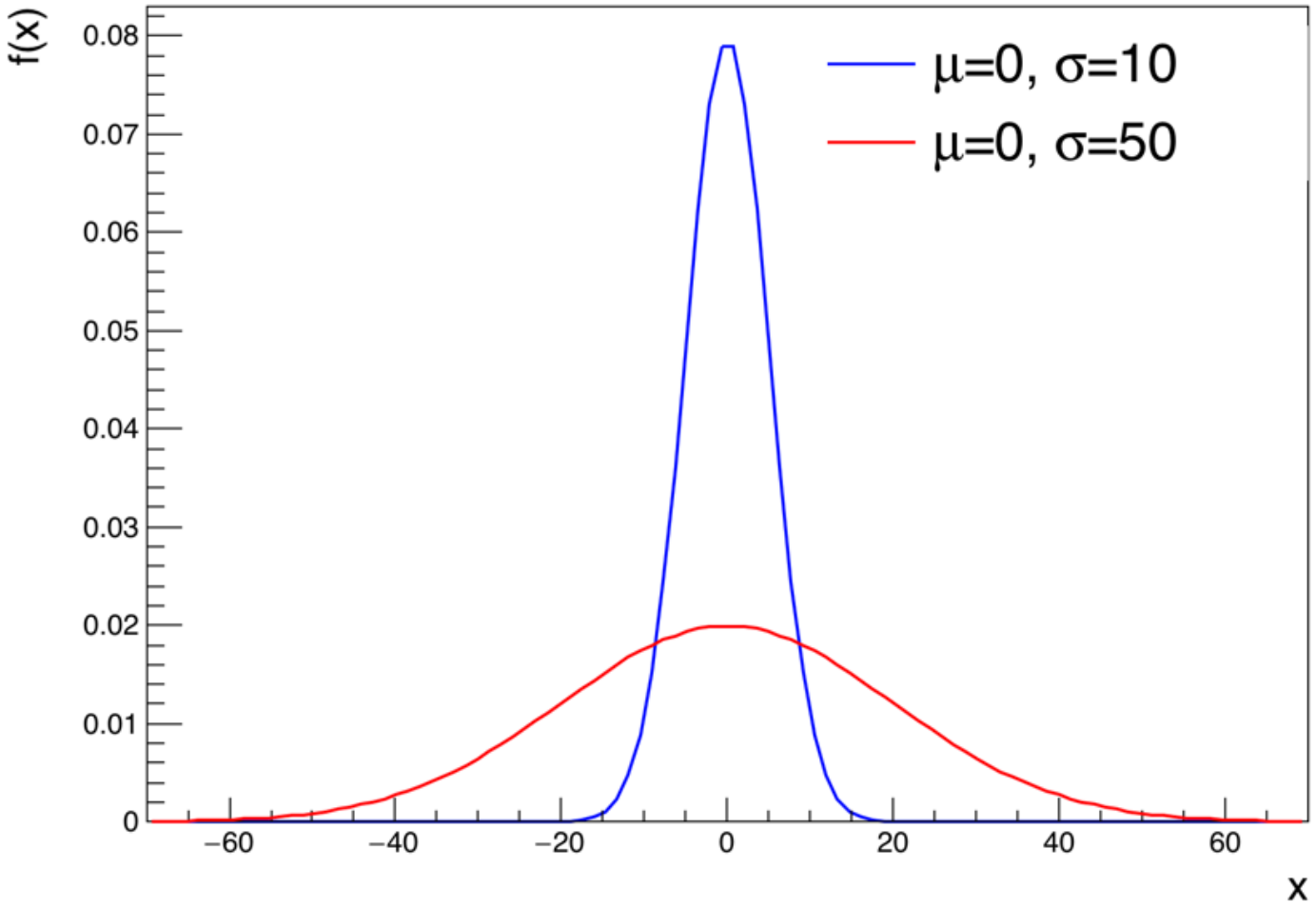
- Najniższy moment, który niesie informacje o odchyleniu zmiennej losowej $X$ od swojej wartości średniej nazywany jest **wariancją** (*ang. variance*):

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

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

# Rozkład normalny o dużej i małej wariancji

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



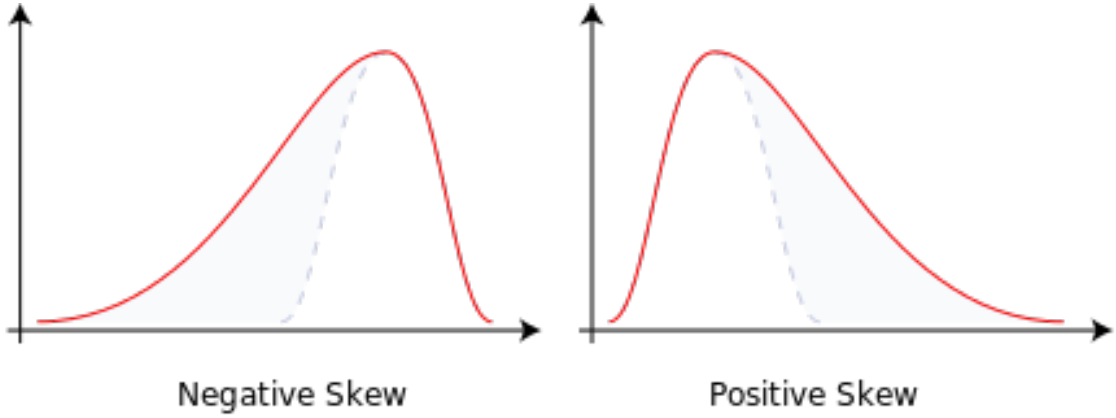
# Momenty wyższych rzędów

- 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

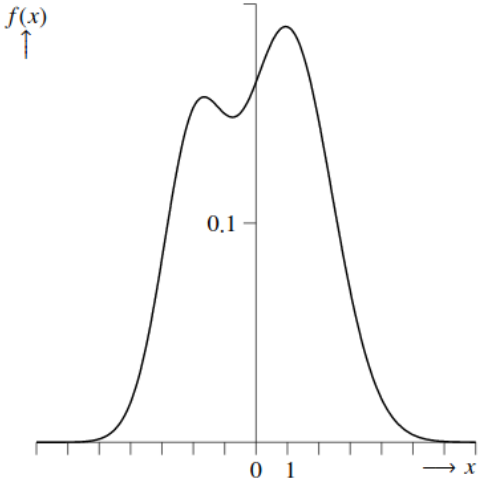
# Momenty wyższych rzędów

- Da się skonstruować rozkład prawdopodobieństwa, który ma zerową skośność i mimo wszystko jest asymetryczny:
  - nastąpi to w **bardzo wąskiej** klasie pryzpadków, gdy istnieje szczególna zależność między nieparzystymi momentami (np. momenty pierwszy i trzeci się zerują) → w rzeczywistym świecie bardzo mało prawdopodobne, by gdzieś się zdarzyła taka sytuacja
- Przykład tutaj:

https://poseidon01.ssrn.com/delivery.php?ID=951089070074087024005127098027090002060050030038051052104082031101114067073122093077023022018005104099039122076115073089115117013032027064082027009007080097112097066059040089117009108084090099020124094127127022007068121119106097077006091001120126087&EXT=pdf

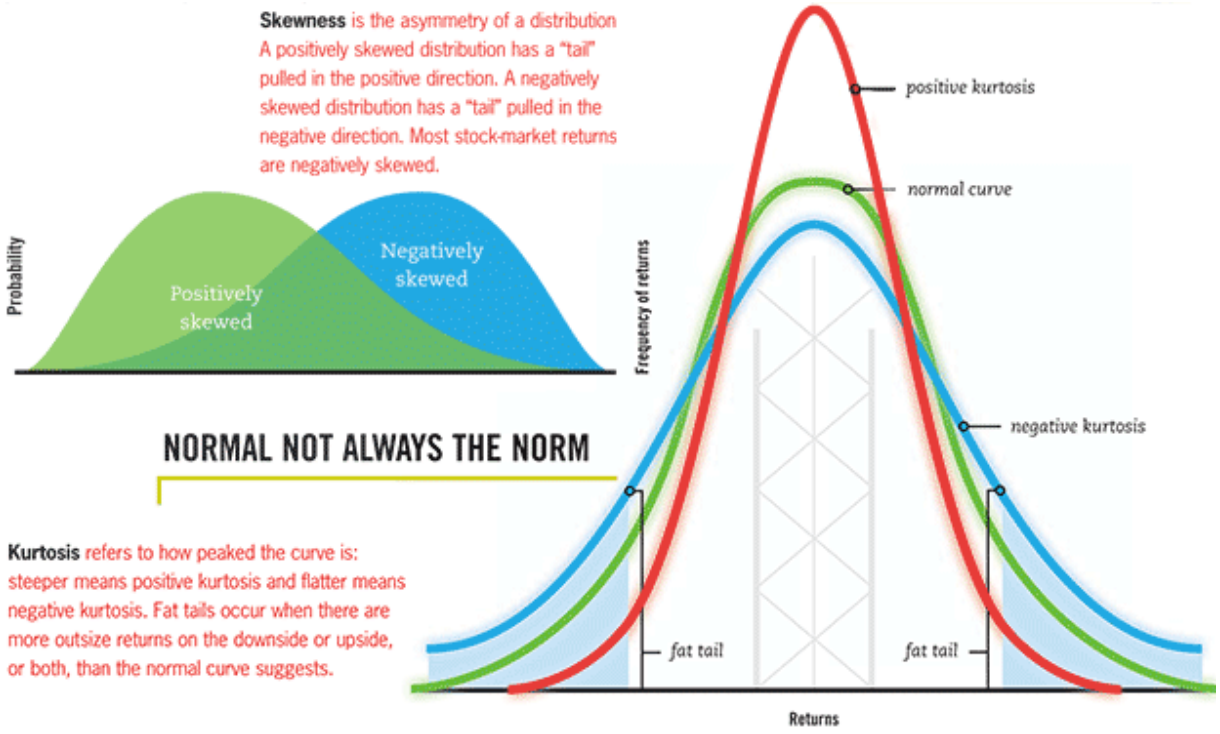
$$f(x) = \sum_{k=1}^{K} \frac{\pi_k}{\sigma_k} \phi\left(\frac{x-\mu_k}{\sigma_k}\right),$$

For example, consider the case $K = 2$, $\pi_1 = 1/3$, $\pi_2 = 2/3$, $\mu_1 = -2$, $\mu_2 = 1$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$. Then $\mathrm{E}(X) = 0$, $\mathrm{E}(X^3) = 0$, $\mathrm{E}(X^5) = 20/3 \neq 0$. Thus, this is an asymmetric distribution with zero skewness. It is depicted in figure 1.



# 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:

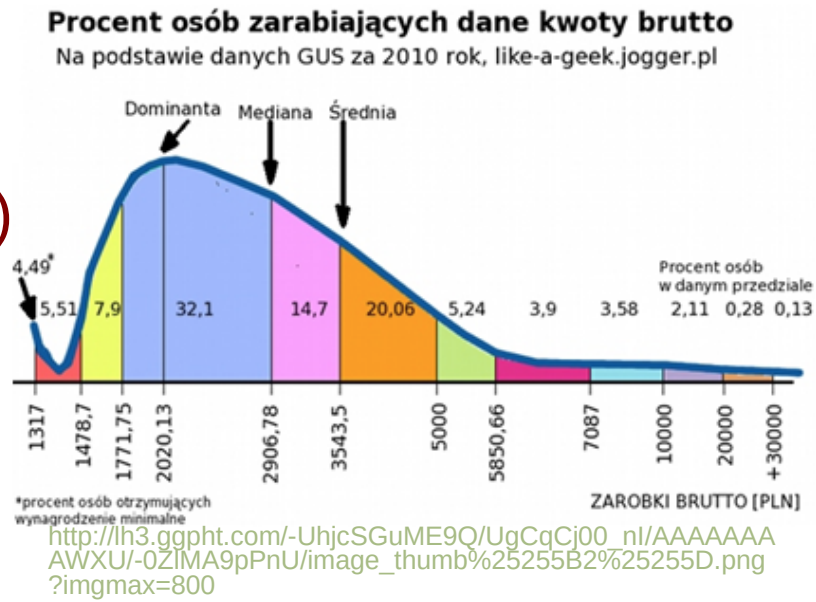
$$\sigma^2(X) = E((X - \hat{x})^2) = E(X^2 - 2X \cdot \hat{x} + \hat{x}^2) = E(X^2) - 2(E(X))^2 + (E(X))^2 = E(X^2) - (E(X))^2$$

- 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., moda, mediana

- **Zmienna standardowa** (o wartości oczekiwanej 0 i odchyleniu 1):
  - rozważmy zmienną losową: $U = \frac{X - \hat{x}}{\sigma(X)}$
  - 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**, **moda**, **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/-0ZIMA9pPnU/image_thumb%25255B2%25255D.png?imgmax=800

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

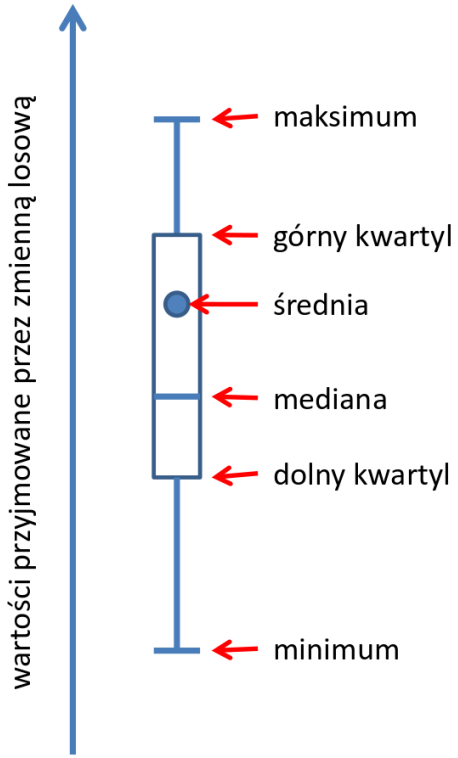
$$F(x_{0,5}) = P(X < x_{0,5}) = 0{,}5$$

# 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 $x_q$

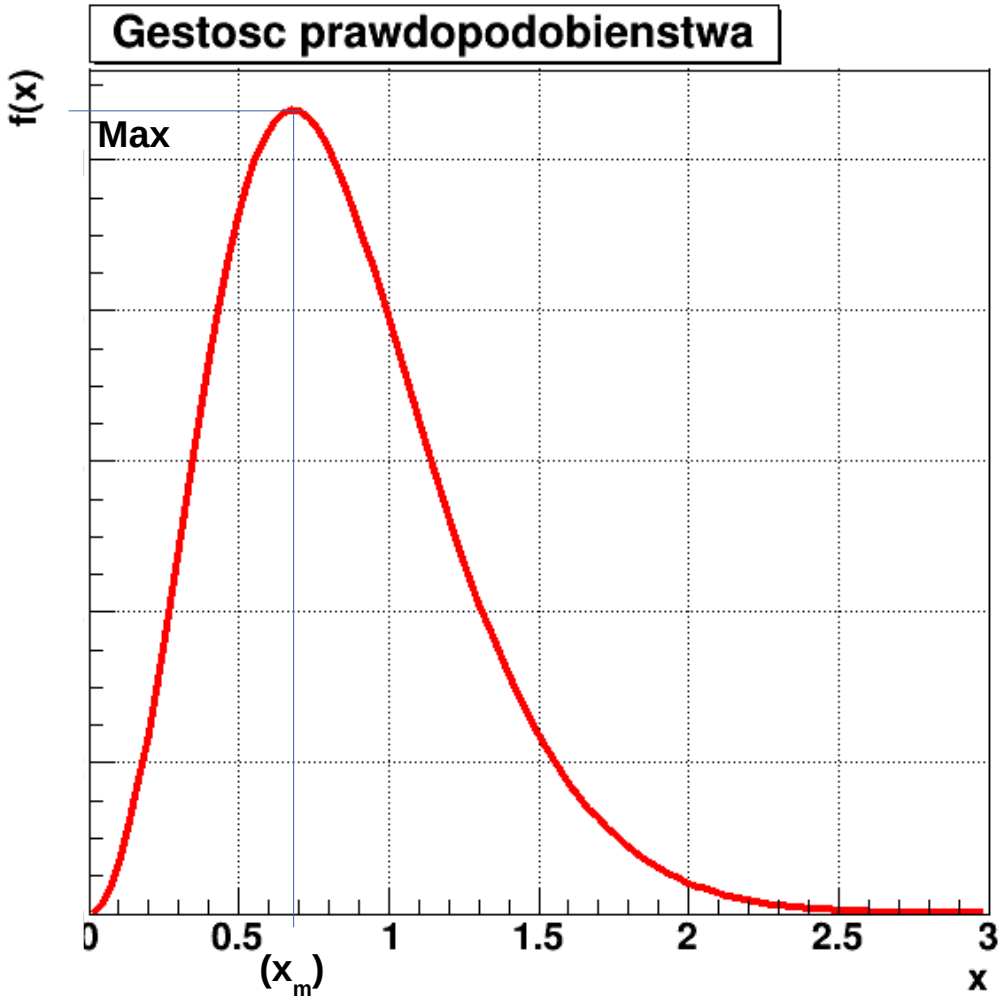
# Kwantyle

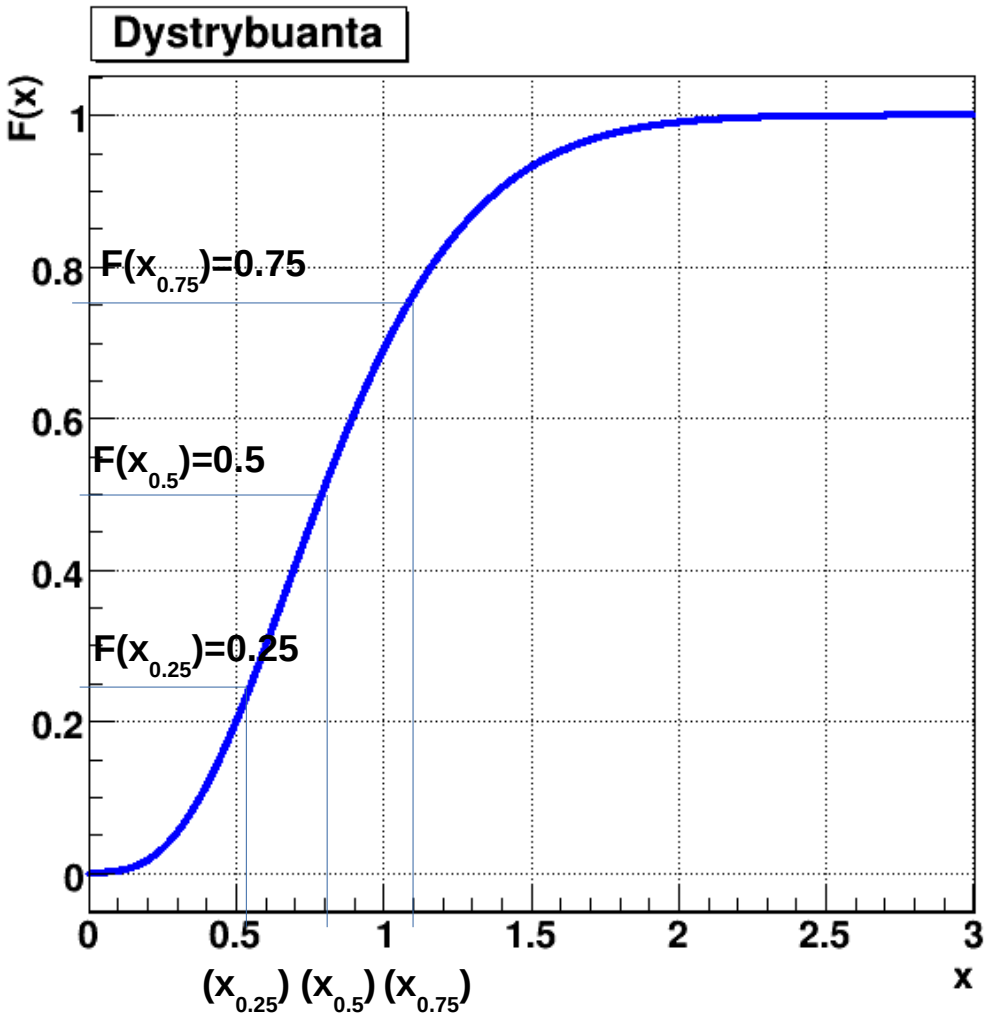
$$F(x_q)=\int_{-\infty}^{x_q} f(x)\,dx=q$$ - kwantyl rzędu q

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

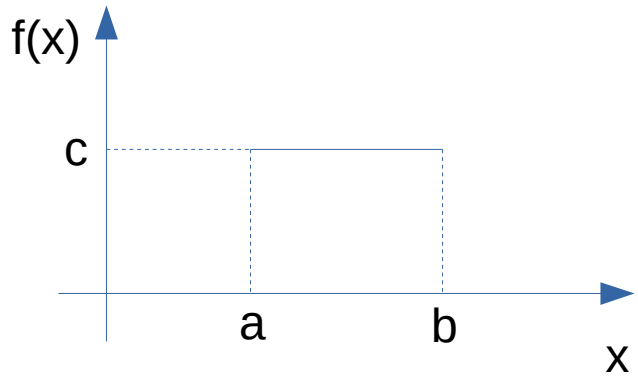


# Przykład 1 – rozkład jednostajny

- 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:

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

# Niepewność typu B - przykład

- **Niepewność typu B** – obliczana na drodze innej niż metoda A. Przykład:

  **Wykład 1**

  - tylko jeden pomiar wielkości mierzonej
  - urządzenie pomiarowe jest mało dokładne (np. mierzymy grubość płytki linijką zamiast śrubą mikrometryczną) – wyniki nie wykazują rozrzutu

- Obliczanie niepewności typu B oparte jest o naukowy osąd eksperymentatora – bierzemy pod uwagę wiedzę o przyrządach (wzorcowanie), badanym materiale, itp.

- Założenie:
  - prawdopodobieństwo uzyskania pomiaru mieszczącego się w przedziale wyznaczonym przez wynik i (znaną) niepewność wzorcowania Δx jest jednakowe

- Efekt:
  - rozkład pomiarów jest **rozkładem jednostajnym**
  - **wynik:** jedna wartość (jeden pomiar)
  - **niepewność:** odchylenie standardowe wartości oczekiwanej

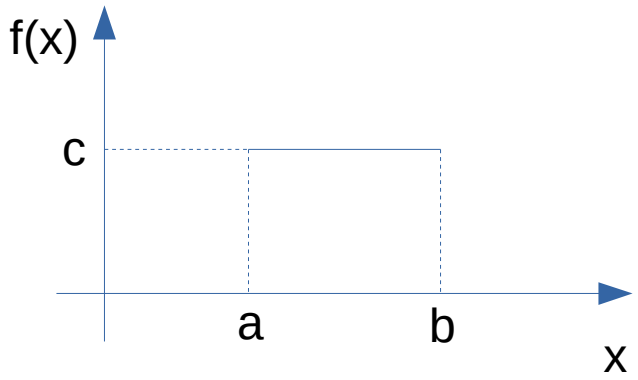
$$u(x)=\frac{\Delta x}{\sqrt{3}}=\sqrt{\frac{(\Delta x)^2}{3}}$$

# Przykład 1 – rozkład jednostajny

- Gęstość prawdopodobieństwa:

$$f(x)=\frac{1}{a-b};\ x\in\langle a,b\rangle$$

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




- Wartość oczekiwana:

$$E(X)=\hat{x}=\frac{b+a}{2}$$

- Wariancja:

$$\sigma^2(X)=\frac{(b-a)^2}{12}$$

- Przewodnik GUM mówi o niepewności typu B tak:

Figure 2 a). Then $x_i$, the expectation or expected value of $X_i$, is the midpoint of the interval, $x_i = (a_- + a_+)/2$, with associated variance

$$u^2(x_i) = (a_+ - a_-)^2/12 \qquad (6)$$

If the difference between the bounds, $a_+ - a_-$, is denoted by $2a$, then Equation (6) becomes

$$u^2(x_i) = a^2/3 \qquad (7)$$

www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

# Przykład 2 – rozkład dwumianowy

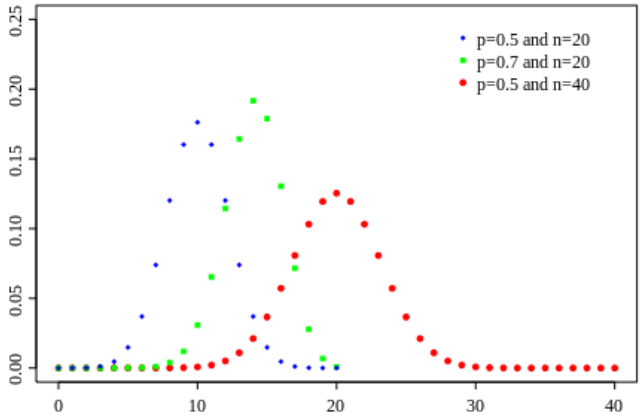
- *ang. binomial distribution*
- Wynik zawsze jedną z dwóch wykluczających się wartości (sukces i porażka)
- Funkcja prawdopodobieństwa:

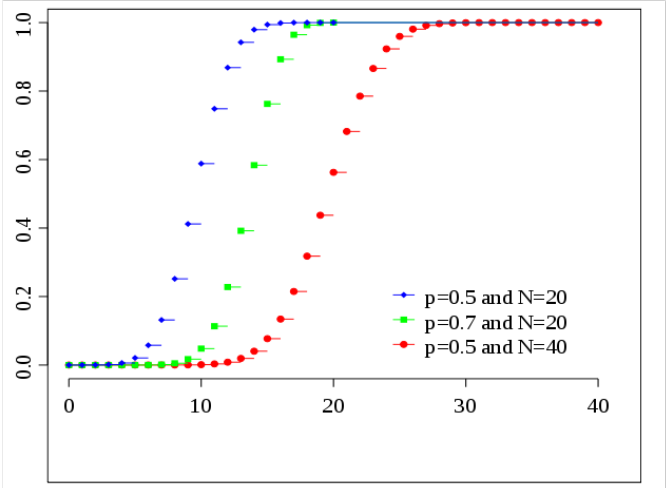
$$p_n(k)=\binom{n}{k}p^k q^{n-k}=\frac{n!}{k!(n-k)!}p^k q^{n-k};\ p\in\langle 0;1\rangle;\ q=1-p$$

  - $k$ 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

Wartość oczekiwana:

$$E(X)=np$$

Wariancja poj. próby $x_i$:

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

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:

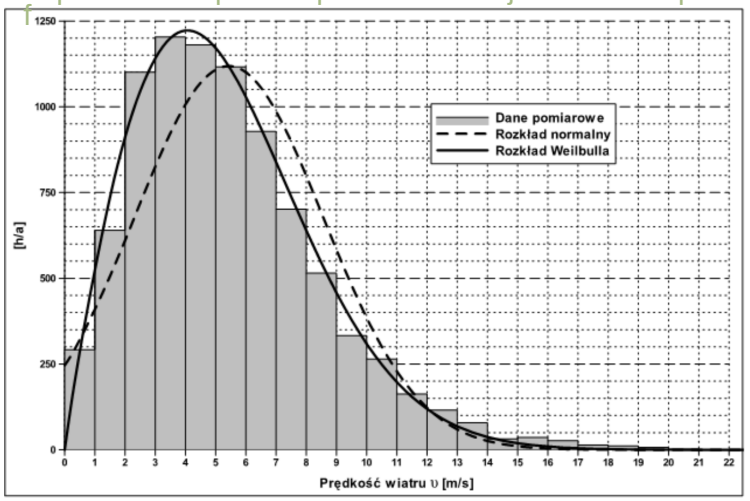
$$\sigma^2(x)=A^2\left[\Gamma\left(1+\frac{2}{k}\right)-\left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right]$$

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

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

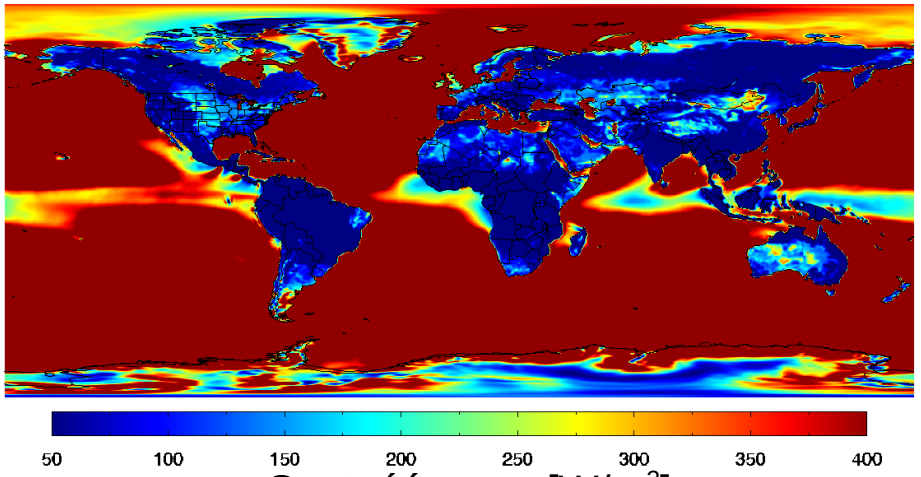
ρ – gęstość powietrza

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

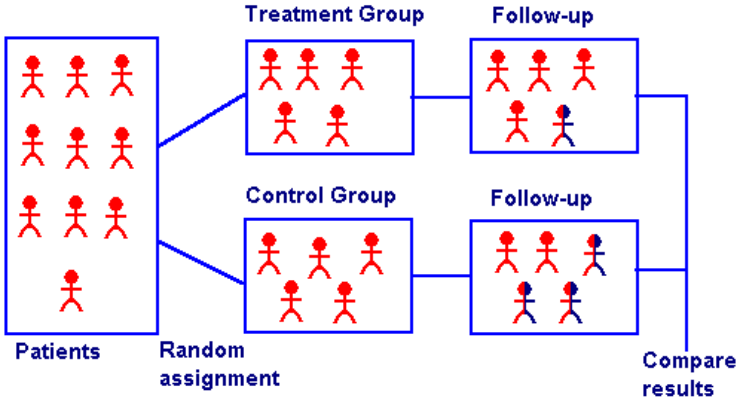



Gęstość mocy [W/m$^2$]

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

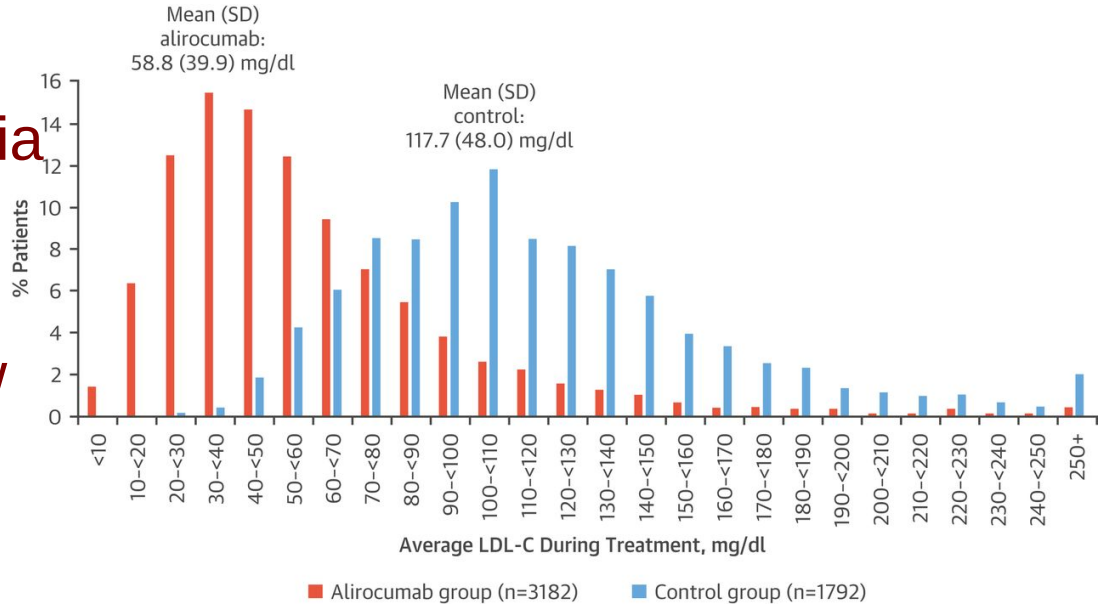
# Rozkłady w praktyce – dane medyczne

- 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)




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

- Przykład 1:
  - 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



# Rozkłady w praktyce – dane medyczne

- Jednym z ważniejszych zastosowań statystyki są **badania medyczne**
- Statystykę wykorzystuje się również do badania **farmakokinetyki** i **bezpieczeństwa** stosowania leków
- Przykład 2:
  - substancja **imatinib** nazwa handlowa **Glivec**
  - lek stosowany w celu leczenia **przewlekłej białaczki szpikowej**
  - Mierzono **trough level** (stężenie leku przed podaniem kolejnej dawki)

http://www.bloodjournal.org/content/111/8/4022

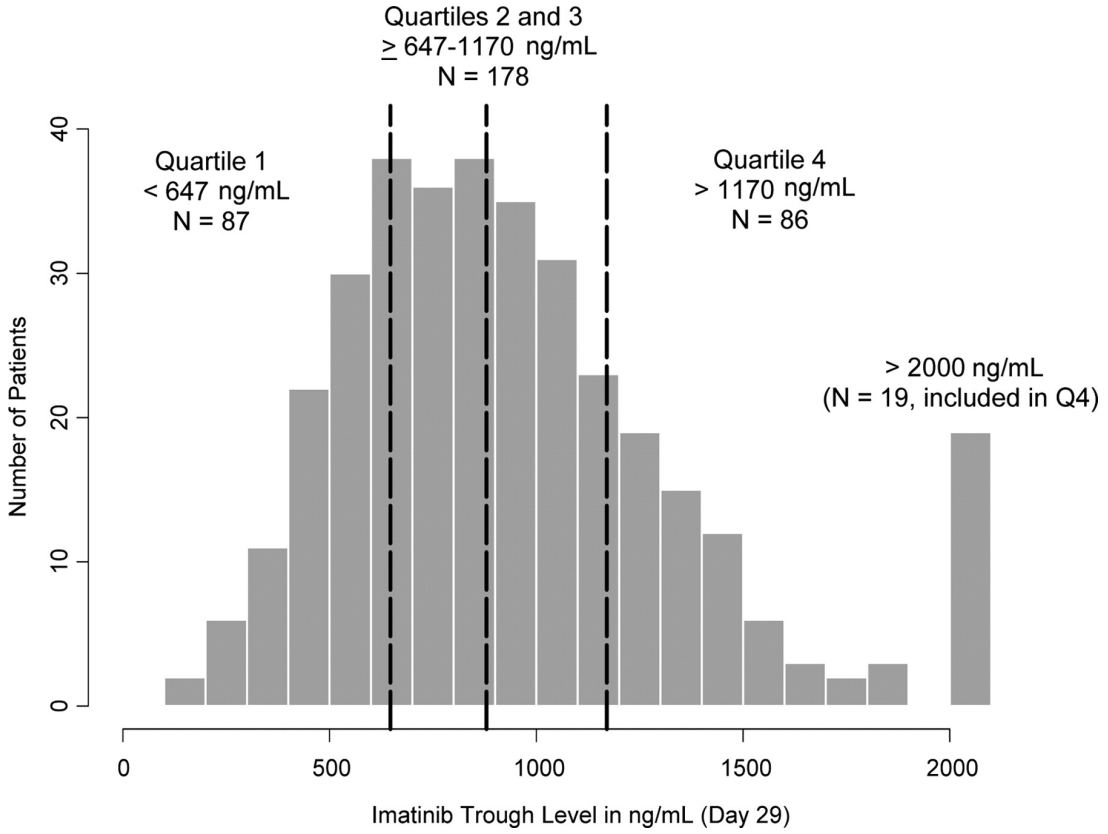



**Figure 1**

**Distribution of imatinib trough levels at 400 mg daily at steady state on day 29 (n = 351).** The vertical dashed lines represent 25th, 50th (median), and 75th percentiles (ie, 647, 879, and 1170 ng/mL, respectively).

# Rozkłady w praktyce – dane medyczne

There also appears to be a trend toward lower EFS with lower imatinib trough levels (**Figure 5**). Estimated EFS rates at 5 years were 78%, 83%, and 89%, for Q1, Q2-Q3, and Q4, respectively. However, no statistically significant difference was observed with the available follow-up (*P* = .16, log-rank test).

http://www.bloodjournal.org/content/111/8/4022

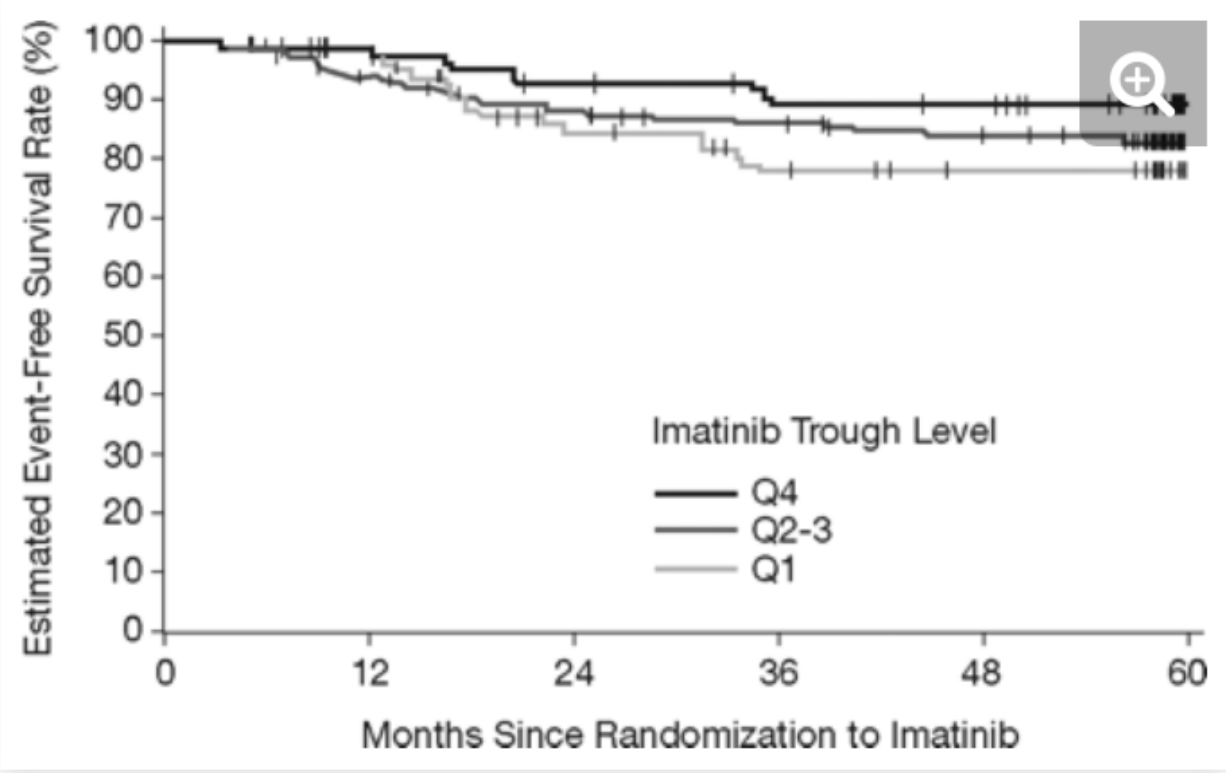



**Figure 5**

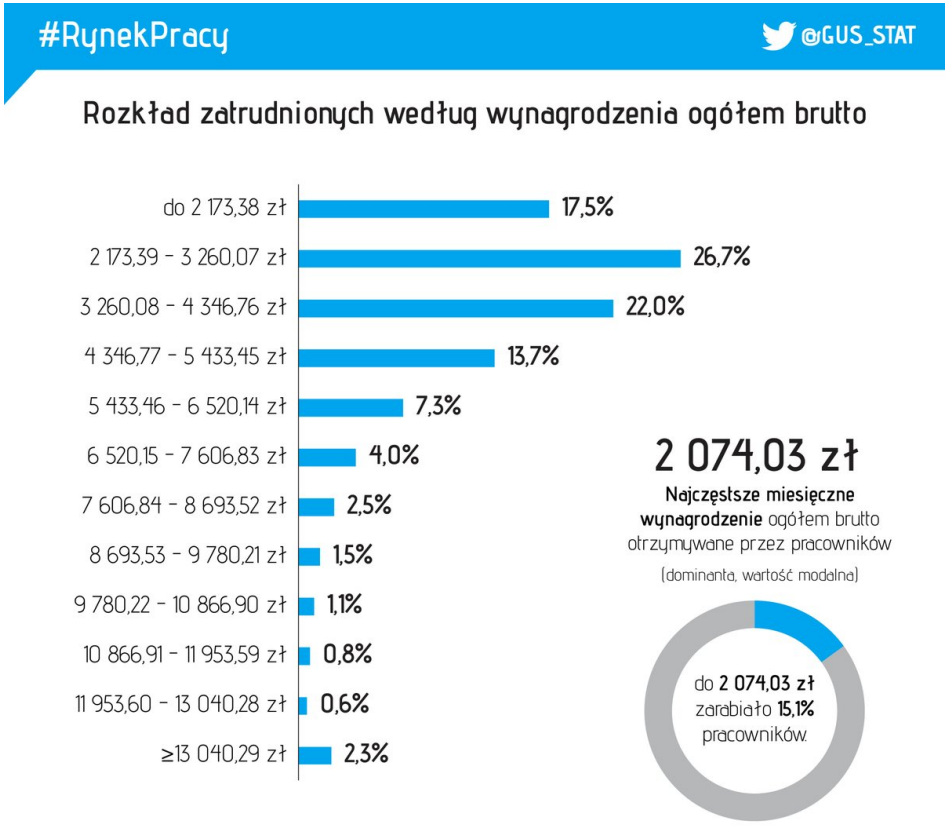
Download figure | Open in new tab | Download powerpoint

**Event-free survival by category of steady-state imatinib trough levels**. The estimated EFS rates at 5 years were 78%, 83%, and 89% in the Q1, Q2-Q3, and Q4 groups, respectively (*P* = .16, log-rank test).

# Rozkłady w praktyce – wynagrodzenia

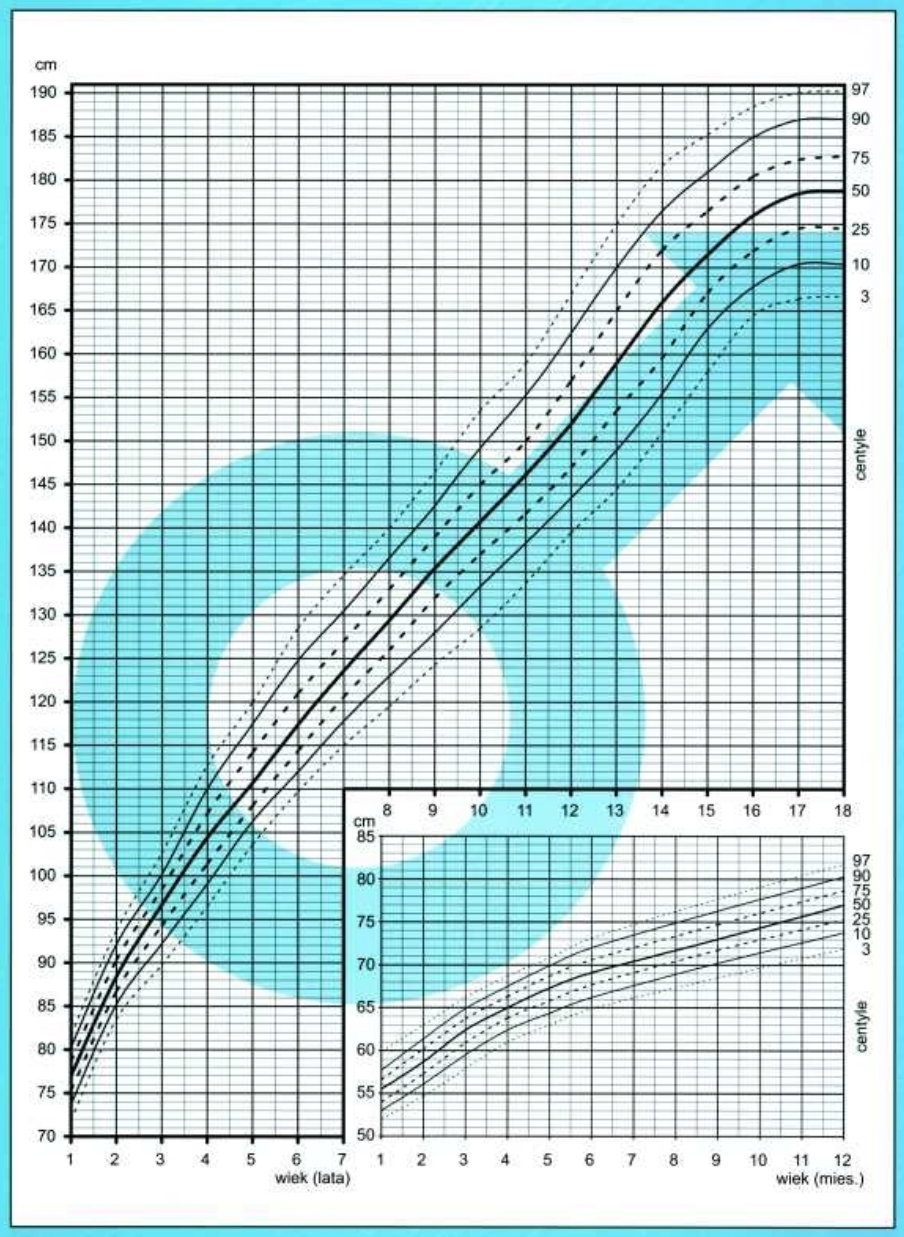
- **Przeciętne miesięczne wynagrodzenie** ogółem brutto w gospodarce narodowej (dla jednostek o liczbie pracujących powyżej 9 osób) w październiku 2016 r. wyniosło **4346,76 zł**.
- **Przeciętne godzinowe wynagrodzenie** ogółem brutto wyniosło **26,37 zł**.
- **Najczęstsze miesięczne wynagrodzenie** ogółem brutto otrzymywane przez pracowników wynosiło **2074,03 zł** (**dominanta**, wartość modalna).
- **Połowa** zatrudnionych pracowników otrzymała wynagrodzenie ogółem brutto do **3510,67 zł** (**mediana** = decyl piąty = wynagrodzenie środkowe).
- **10% najniżej zarabiających** pracowników otrzymało wynagrodzenie ogółem brutto co najwyżej w wysokości **1890,32 zł** (**decyl pierwszy**).
- **10% najwyżej zarabiających** pracowników otrzymało wynagrodzenie ogółem brutto co najmniej w wysokości **7200,00 zł** (**decyl dziewiąty**).

**Dane z 2017 za rok 2016**

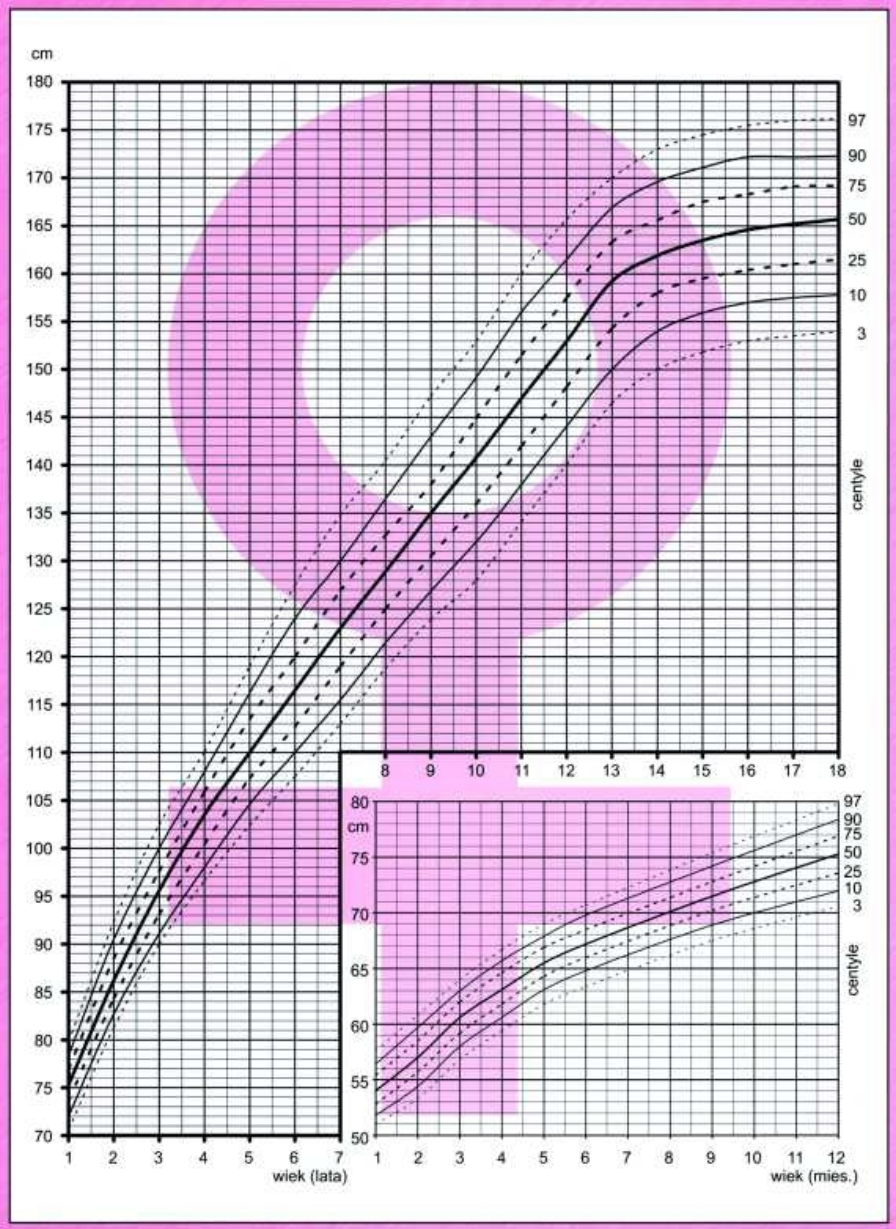



https://stat.gov.pl/obszary-tematyczne/rynek-pracy/pracujacy-zatrudnieni-wynagrodzenia-koszty-pracy/struktura-wynagrodzen-wedlug-zawodow-w-pazdzierniku-2016-r-,5,5.html

# Rozkłady w praktyce – tablice centylowe




chłopcy




dziewczynki

# KONIEC

# Rozkład dwumianowy

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

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

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

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

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

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

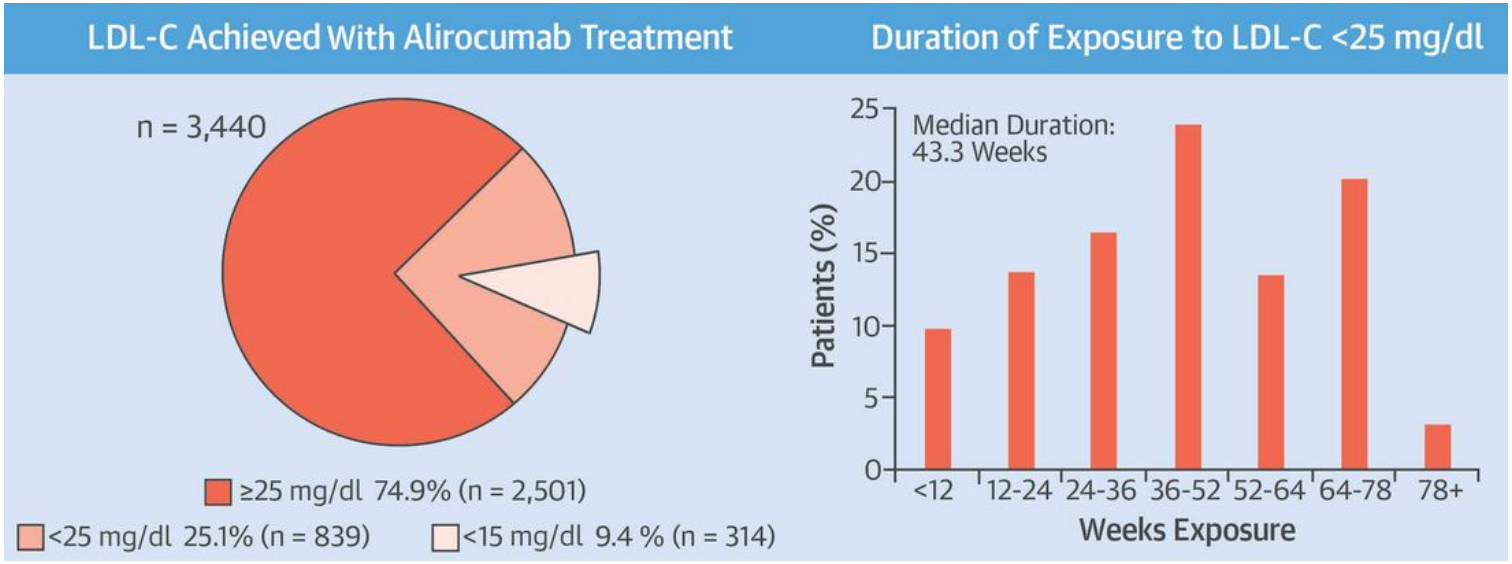
$$\sigma^2(x)=n(n-1)p^2+np-(np)^2=npq$$

**Odchylenie standardowe:**

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

# Rozkłady w praktyce – dane medyczne

- **Mediana** długości przyjmowania leku wyniosła 78 tygodni
- W badaniu alirokumabu sprawdzano 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