Komputerowa analiza danych doświadczalnych

Wykład 5
26.03.2019

Anna Chmiel na postawie materiałów od Łukasza Graczykowskigo

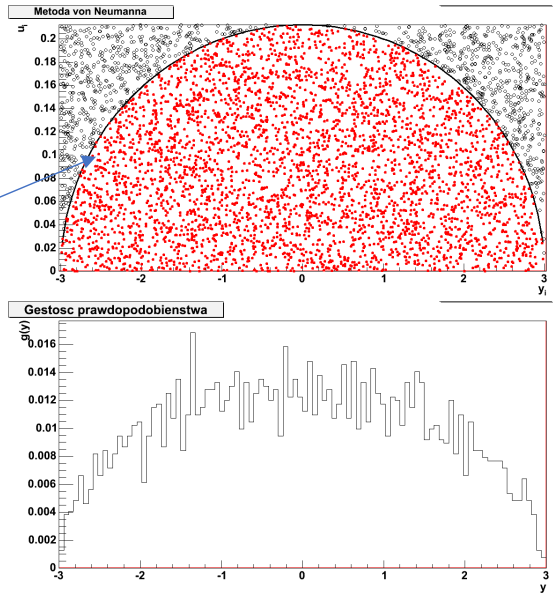
Metody Monte Carlo

Najważniejsze rozkłady prawdopodobieństwa

Metoda akceptacji-odrzuceń von Neumanna

Metoda von Neumanna - definicje

- Geometryczny opis metody von Neumanna:
  - chcemy wygenerować liczby losowe opisane gęstością $g(y)$ w przedziale <a,b>
  - rozważamy krzywą: $u = g(y)$ oraz funkcję stałą: $u = d,\ d \leq g_{max}$
  - losujemy z rozkładu jednorodnego liczby (y_i, u_i), które spełniają warunki: $a \leq y_i \leq b, 0 \leq u_i \leq d$
  - jak łatwo zauważyć, nasze punkty układają się w prostokącie na płaszczyźnie
  - odrzucamy wszystkie punkty spełniające nierówność: $u_i \geq g(y_i)$
  - pozostają jedynie punkty położone pod krzywą: $u = g(y)$
  - zaakceptowane wartości y_i podlegają rozkładowi $g(y)$



Metoda von Neumanna z funkcją pom.

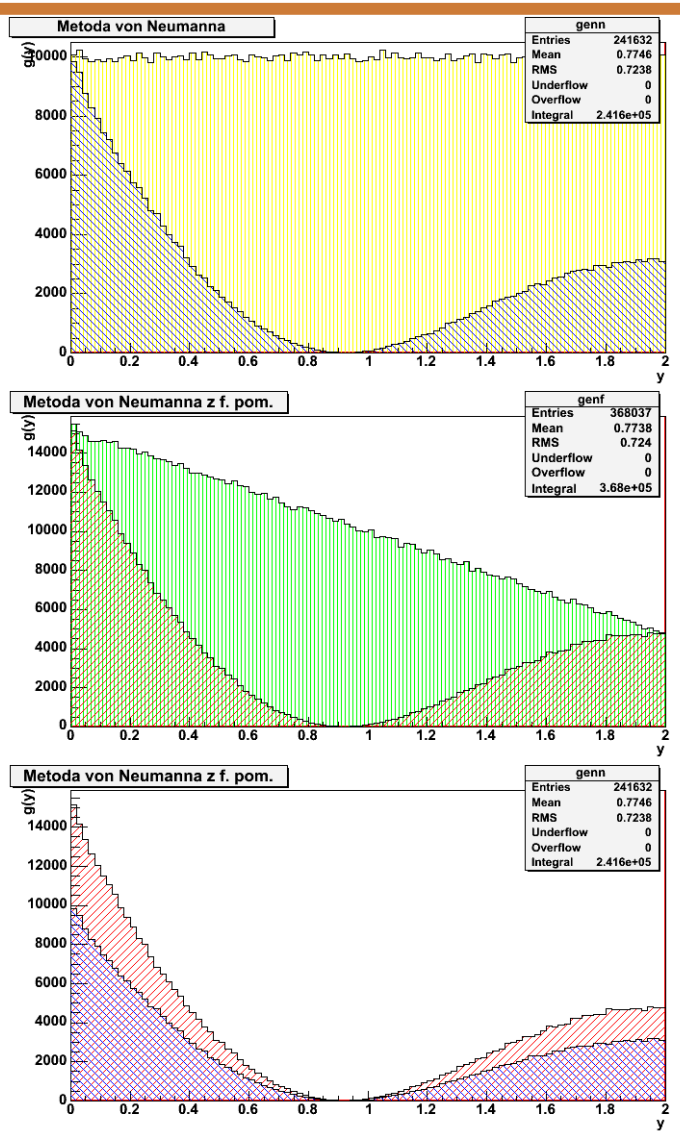
Wydajność metody von Neumanna można poprawić, jeśli odpowiednio zawężimy obszar losowania:

- wprowadzamy funkcję pomocniczą $s(y)$, z ktorej "łatwo" wygenerować zmienne losowe (np. metodą odwrotnej dystrybuanty), i która spełnia warunek:

$$g(y) \leq c \cdot s(y), a < y < b$$

- generujemy liczbę losową y_i z rozkładu $s(y)$ na przedziale $a < y_i < b$ oraz liczbę u_i z rozkładu jednorodnego na przedziale $0 < u_i < 1$
  - odrzucamy liczbę y_i, jeżeli:

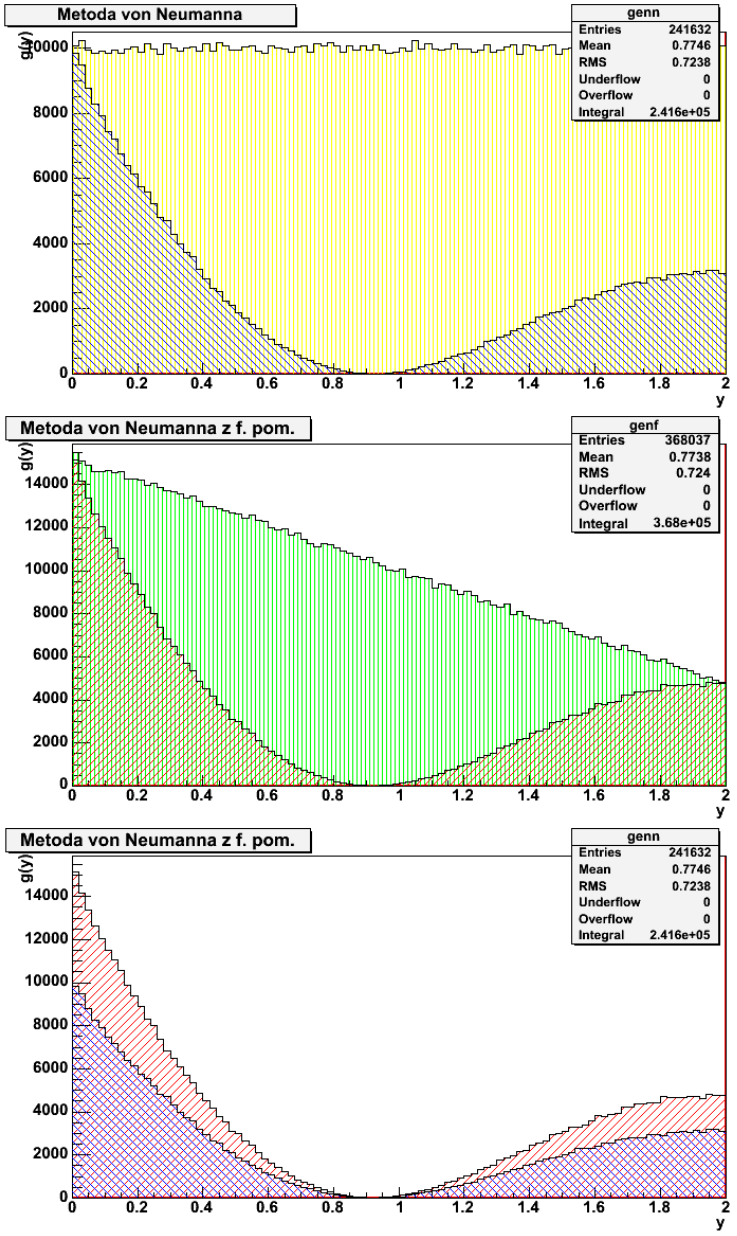
$$u_i \geq \frac{g(y_i)}{c \cdot s(y_i)}$$



Metoda von Neumanna z funkcją pomocniczą

wydajność metody:

$$E = \frac{\int g(y)dy}{c\int s(y)dy}$$

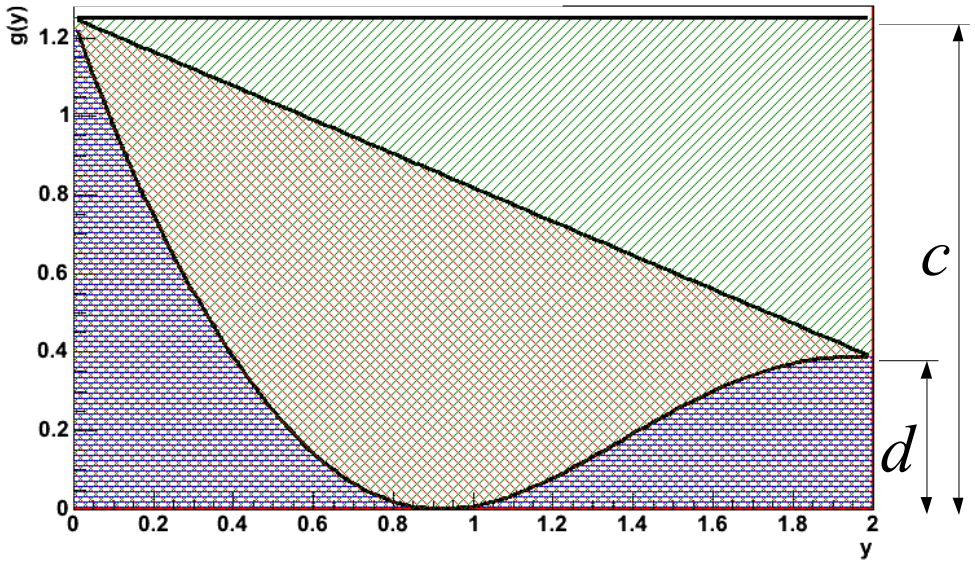


Metoda von Neumanna z funkcją pom.

- Rozważmy funkcję gęstości postaci:

$$g(y) = \cos(\pi y)/(\pi y + 1) + 1/4,\ 0 < y < 2$$

- Funkcja ta, w przedziale od 0 do 2, ma dwa maksima: $g(0) = c$, $g(2) = d$
- W zwykłej metodzie von Neumanna wybieramy prostą: $u_{max} = c$
- Tutaj możemy łatwo wybrać funkcję pomocniczą $s(y)$ jako prostą przechodzącą przez punkty $(0, c)$ i $(2, d)$



- Aby otrzymać wzór $s(y)$ rozważamy układ równań:

$$c = a \cdot 0 + b$$

$$d = a \cdot 2 + b$$

- Z czego wzór na $s(y)$:

$$s(y) = \frac{d - c}{2} y + c$$

- **Jak otrzymać wartość losową z tego rozkładu?**

Metoda von Neumanna z funkcją pom.

- **Metodą odwrotnej dystrybuanty!**
- Liczymy dystrybuantę:

$$S(y) = \frac{d-c}{4} y^2 + cy$$

- Oraz jej funkcję odwrotną:

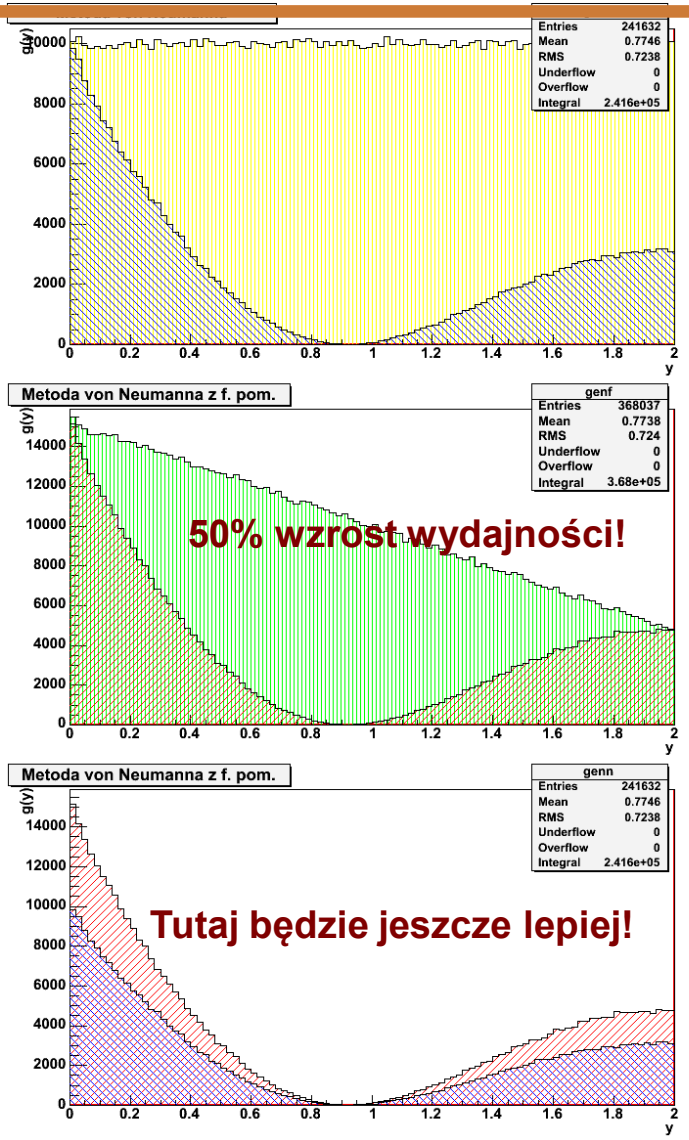
$$y = S^{-1}(x) = 2\frac{c^2 - \sqrt{xc(d-c)+(d-c)^2}}{c(c-d)}$$

- Losujemy wartość z rozkładu jednorodnego w granicach:

$$S(0) = 0,\ S(2) = d + c$$

- I wstawiamy ją do wzoru na odwrotną dystrybuantę by otrzymać y_i z rozkł. $s(y)$
- Losujemy pomocniczą wartość u_i z rozkładu jednorodnego $0 < u_i < 1$
- Sprawdzamy warunek akceptacji y_i:

$$u_i < \frac{g(y_i)}{s(y_i)}$$



Całkowanie metodą Monte Carlo

- Jak już zauważylismy, pole powierzchni pod rozpatrywaną krzywą w stosunku do pola prostokąta, z którego losujemy dwie liczby pseudolosowe, ma się (w przybliżeniu) do siebie tak jak liczba par zaakceptowanych do odrzuconych:

$$\frac{\int_a^b g(y)\,dy}{(b-a)d} \approx \frac{N_{accept}}{N_{all}}$$

- Co pozwala na przybliżone obliczenie wartości całki oznaczonej:

$$\int_a^b g(y)\,dy \approx \frac{N_{accept}}{N_{all}}(b-a)d$$

- W ten sposób można obliczyć **dowolną** całkę oznaczoną poprzez prostą generację dwóch liczb z rozkładu jednorodnego. W wersji n-wymiarowej oczywiście możemy to zrobić dla dowolnej liczby zmiennych losowych (i obliczać całki wielowymiarowe)
- Względna dokładność obliczenia całki:

$$\frac{\Delta I}{I} = \frac{1}{\sqrt{N_{wszystkie}}}$$

Całkowanie metodą Monte Carlo - przykład

- Najpopularniejszy przypadek to wykorzystanie metody Monte Carlo do obliczenia wartości liczby π
- W tym celu rozpatrzmy ćwiartkę okręgu o jednostkowym promieniu. Funkcja opisująca tę ćwiartkę to:

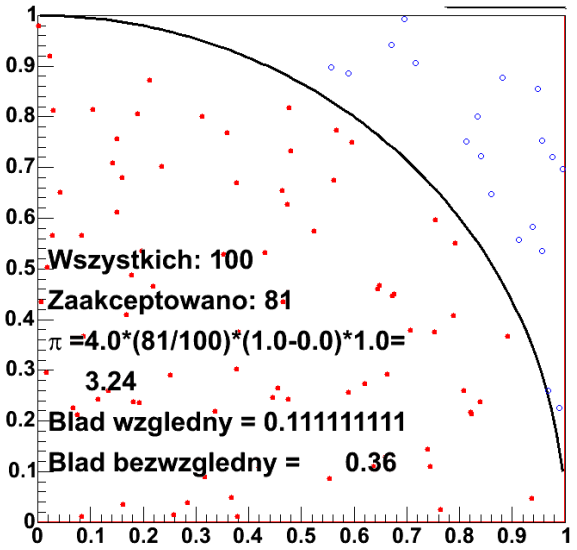
$$g(y) = \sqrt{(R^2 - y^2)}; 0 \leq \quad y \leq 1;$$

- Pole ćwiartki jednostkowego okręgu to:

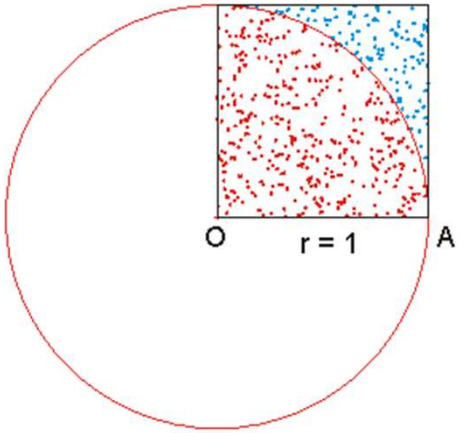
$$I = \int_0^1 g(y)dy = \pi/4 \Rightarrow \pi = 4I$$

- Wartość całki obliczamy metodą Monte Carlo:

$$I \approx \frac{N_{accept}}{N_{all}}(b-a)d$$



wszystko przypomina rzucanie lotkami (darts)



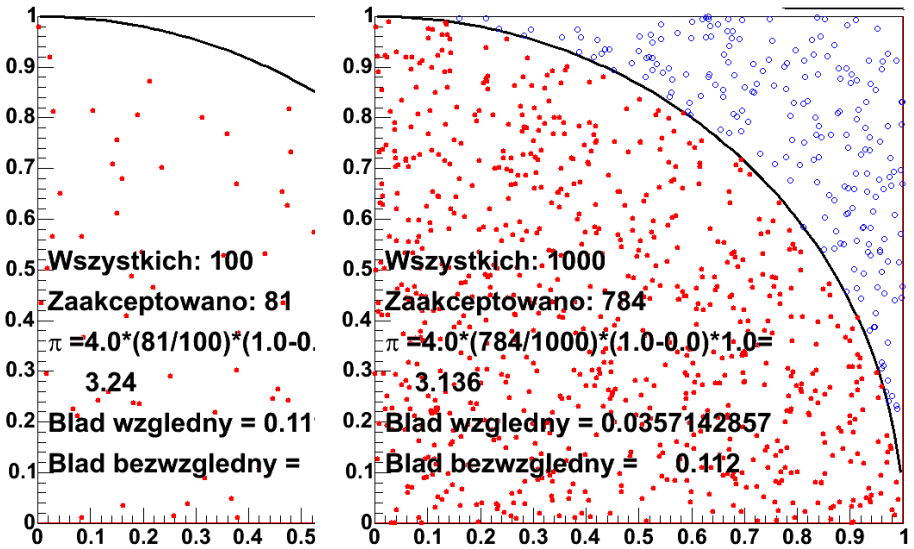
Całkowanie metodą Monte Carlo - przykład

- Najpopularniejszy przypadek to wykorzystanie metody Monte Carlo do obliczenia wartości π
- W tym celu rozpatrzmy ćwiartkę okręgu o jednostkowym promieniu. Funkcja opisująca tę ćwiartkę to:

$$g(y) = \sqrt{(R^2 - y^2)}; 0 \leq \quad y \leq 1;$$

$$I = \int_0^1 g(y)dy = \pi/4 \Rightarrow \pi = 4 \cdot I$$

- Pole ćwiartki jednostkowego okręgu to:
- Wartość całki obliczamy metodą Monte Carlo:



$$I \approx \frac{N_{accept}}{N_{all}}(b-a)d$$

Całkowanie metodą Monte Carlo - przykład

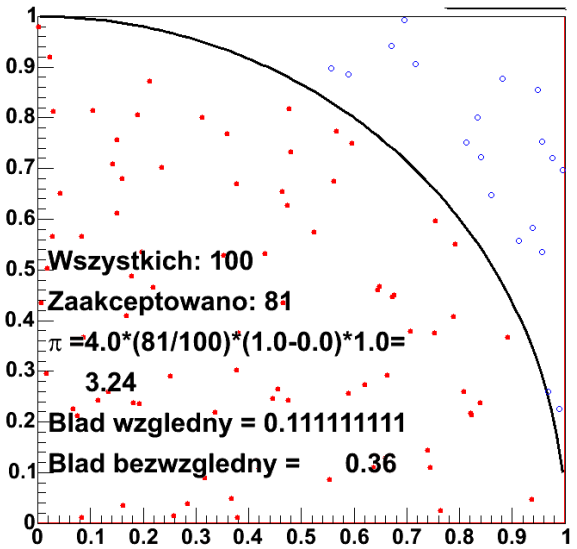
- Najpopularniejszy przypadek to wykorzystanie metody Monte Carlo do obliczenia wartości π
- W tym celu rozpatrzmy ćwiartkę okręgu o jednostkowym promieniu. Funkcja opisująca tę ćwiartkę to:

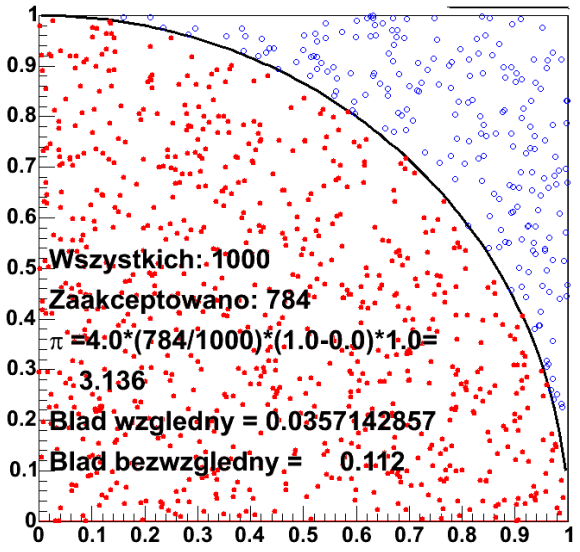
$$g(y) = \sqrt{(R^2 - y^2)}; 0 \leq \quad y \leq 1;$$

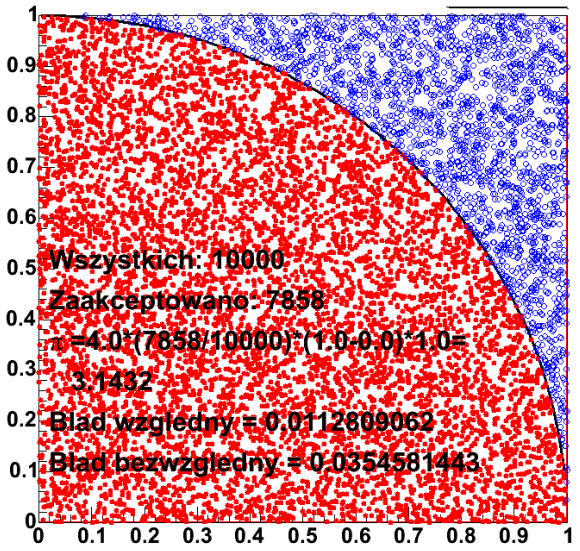
$$I = \int_0^1 g(y)dy = \pi/4 \Rightarrow \pi = 4 \cdot I$$

- Pole ćwiartki jednostkowego okręgu to:
- Wartość całki obliczamy metodą Monte Carlo:

$$I \approx \frac{N_{accept}}{N_{all}}(b-a)d$$







Całkowanie metodą Monte Carlo - przykład

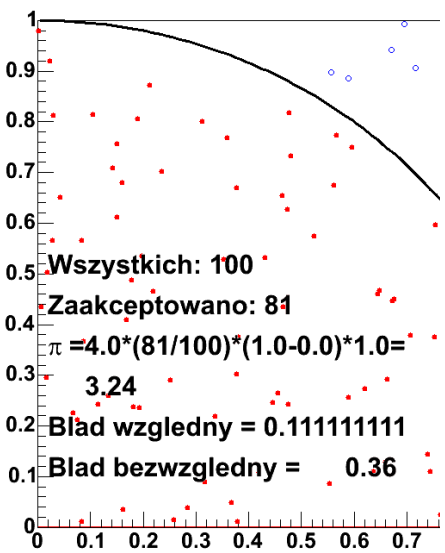
- Najpopularniejszy przypadek to wykorzystanie metody Monte Carlo do obliczenia wartości π
- W tym celu rozpatrzmy ćwiartkę okręgu o jednostkowym promieniu. Funkcja opisująca tę ćwiartkę to:

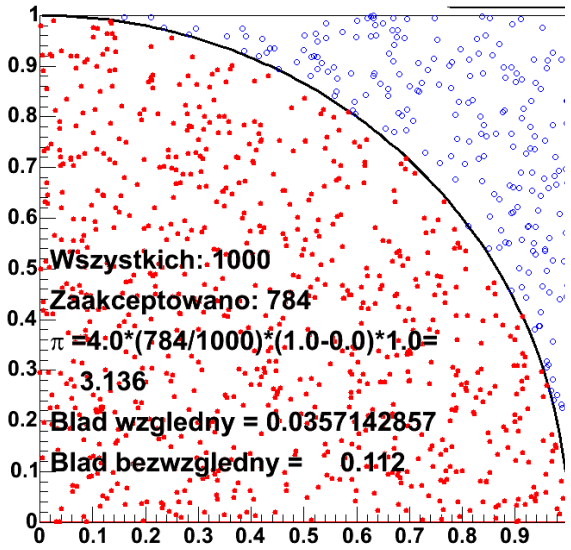
$$g(y) = \sqrt{(R^2 - y^2)}; 0 \leq \quad y \leq 1;$$

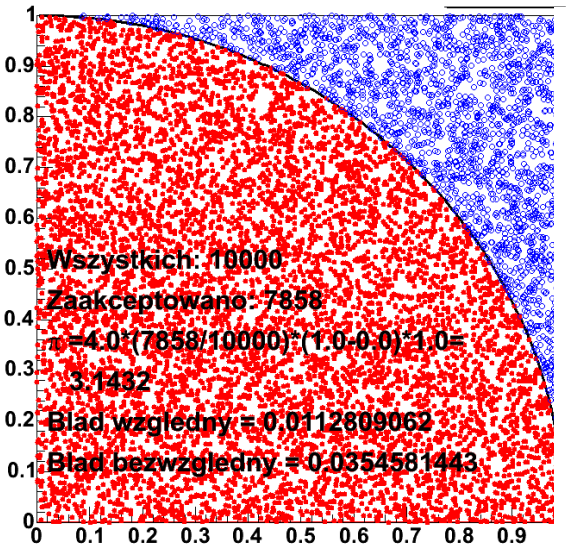
$$I = \int_0^1 g(y)dy = \pi / 4 \Rightarrow \pi = 4 \cdot I$$

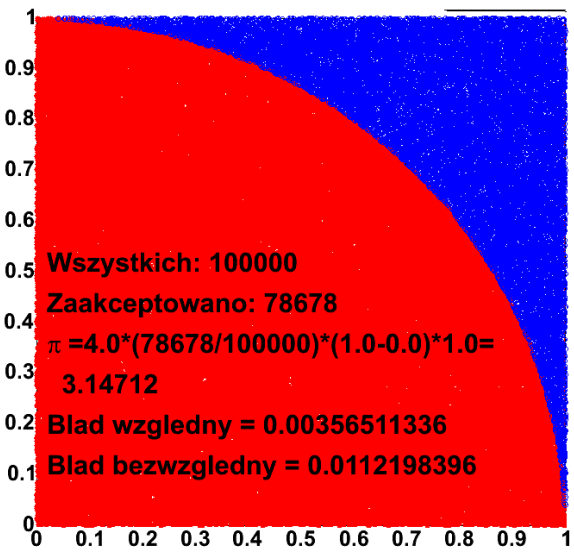
- Pole ćwiartki jednostkowego okręgu to:
- Wartość całki obliczamy metodą Monte Carlo:

$$I \approx \frac{N_{accept}}{N_{all}} (b - a)d$$









Całkowanie metodą Monte Carlo - przykład

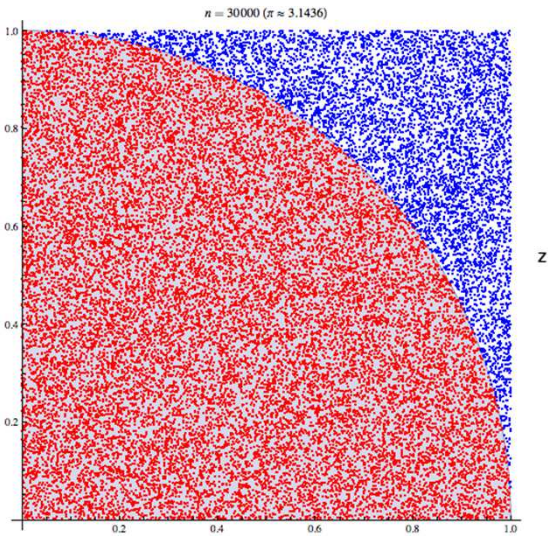
- Najpopularniejszy przypadek to wykorzystanie metody Monte Carlo do obliczenia wartości π
- W tym celu rozpatrzmy ćwiartkę okręgu o jednostkowym promieniu. Funkcja opisująca tę ćwiartkę to:

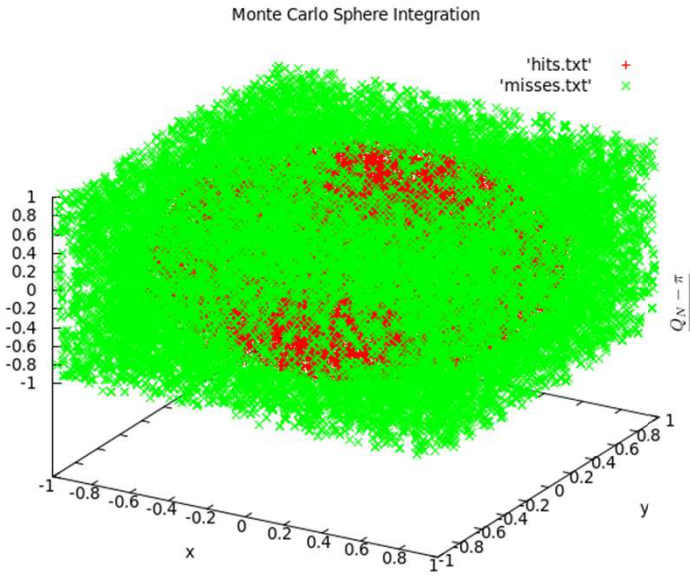
$$g(y) = \sqrt{(R^2 - y^2)}; 0 \leq \quad y \leq 1;$$

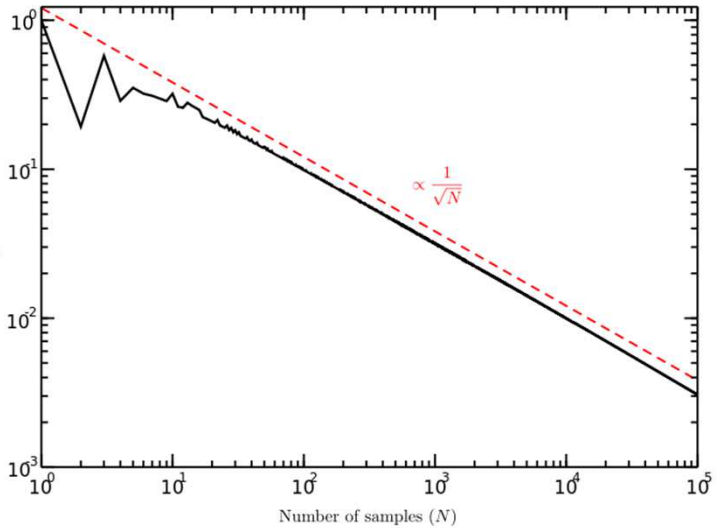
$$I = \int_0^1 g(y)dy = \pi / 4 \Rightarrow \pi = 4 \cdot I$$

- Pole ćwiartki jednostkowego okręgu to:
- Wartość całki obliczamy metodą Monte Carlo:

$$I \approx \frac{N_{accept}}{N_{all}}(b - a)d$$







Generacja liczb o rozkładzie normalnym

- Jak pamiętamy, **rozkład normalny nie ma analitycznej formy dystrybuanty**
- Do generowania liczb z rozkładu normalnego o $\hat{x}=0,\ \sigma=1$ (standardowego) służy **metoda Box'a-Muller'a**

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

transformacja dowolnego rozkł. norm. do standardowego

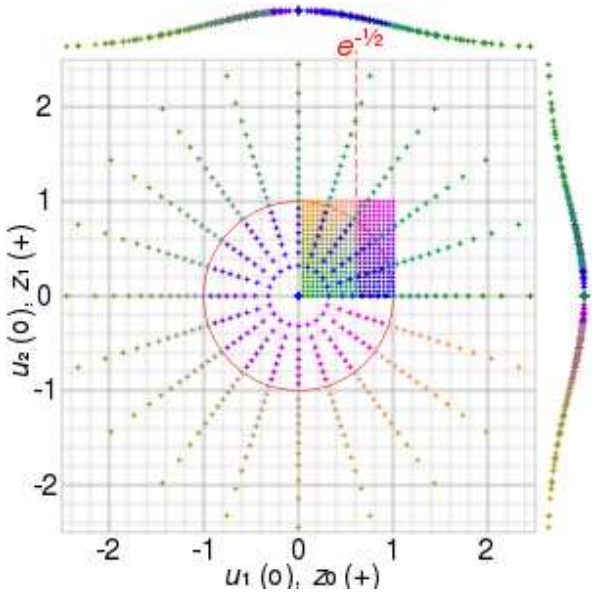
- Generujemy parę liczb (u_1,u_2) z rozkładów jednorodnych (0,1) i dokonujemy zamiany zmiennych:

$$v_1=2u_1-1 \qquad v_2=2u_2-1$$

- Obliczamy: $s=v_1^2+v_2^2$
- Gdy $s\geq 1$ odrzucamy parę
- Otrzymujemy dwie liczby pseudolosowe

opisane rozkładem normalnym standardowym:

$$x_1=v_1\sqrt{-(2/s)\ln s} \qquad x_2=v_2\sqrt{-(2/s)\ln s}$$



Najważniejsze rozkłady prawdopodobieństwa

Rozkład dwumianowy

- W Polsce znany również jako rozkład Bernoulliego (*ang. binomial distribution*) – w innych krajach może oznaczać inny rozkład
- Rozważmy proste doświadczenie – rzut monetą:
  - w wyniku rzutu możemy otrzymać dwa wykluczające się wyniki
  - zatem przestrzeń zdarzeń elementarnych: $E = A + \bar{A}$
  - możemy zdefiniować prawdopodobieństwa:

$$P(A) = p$$

$$P(\bar{A}) = 1 - p = q$$

- Wynik doświadczenia może być zmienną losową X_i, która przybiera wartość 1 lub 0 w zależności od tego, czy zaszło zdarzenie A lub $\bar{A}$
- Jeśli powtórzymy wielokrotnie doświadczenie, to otrzymamy rozkład zmiennej losowej $X = X_1 + X_2 + \ldots . X_n$

Rozkład dwumianowy

- Z rachunku prawdopodobieństwa wiemy, że jeżeli przestrzeń zdarzeń elementarnych $E = A_1 + A_2 + \ldots + A_n$ i zdarzenia są niezależne, to:

$$P(A_1 A_2 \ldots A_n) = P(A_1) P(A_2) \ldots P(A_n)$$

- Z tego wynika, że prawdopodobieństwo, że k pierwszych doświadczeń (z n) da wynik zdarzenia A a pozostałe n-k dadzą wynik zdarzenia A, wynosi:

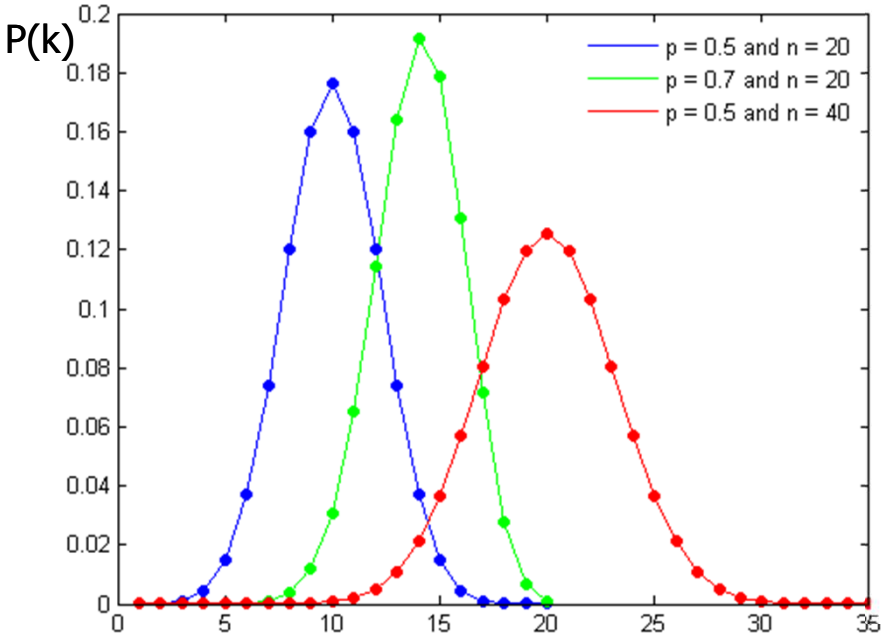
$$P\left(A^k \overline{A}^{n-k}\right) = p^k q^{n-k}$$

- Zgodnie z kombinatoryką, pojawienie się k razy zdarzenia A w n doświadczeniach realizuje się na "n po k" sposobów: różniących się kolejnością zdarzeń A i A
- Prawdopodobieństwo wystąpienia k razy zdarzenia A i n-k razy zdarzenia A w n doświadczeniach, w dowolnej kolejności, wynosi: $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$
- Tak zdefiniowany rozkład nazywamy **rozkładem dwumianowym**

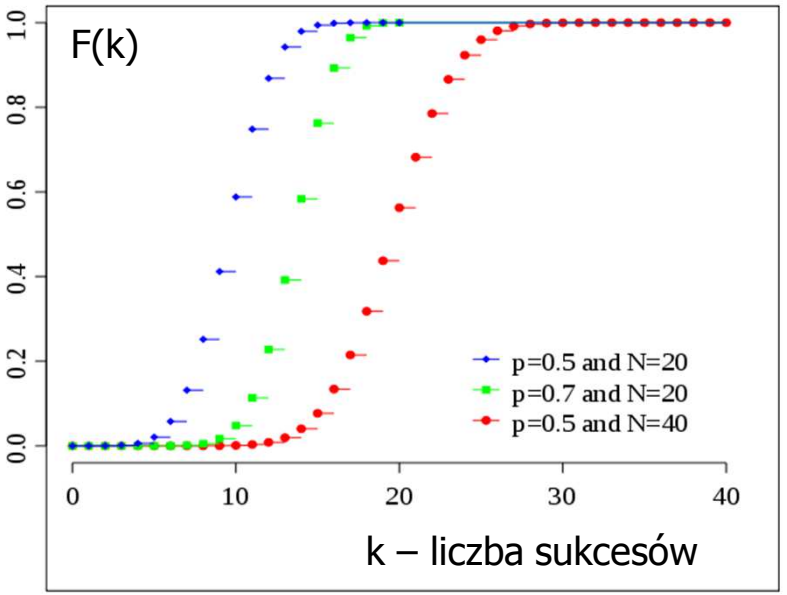
$$P(k) = W_k^n = \binom{n}{k} p^k q^{n-k};\ q = 1 - p$$

Rozkład dwumianowy

Rozkład prawdopodobieństwa



Dystrybuanta



Rozkład dwumianowy

- Policzmy wartość oczekiwaną i wariancję rozkładu dwumianowego
- Dla pojedynczego doświadczenia X_i (zmiennej losowej, która może przyjąć wartość 1 lub 0):

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

$$E(X_i) = 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0)$$

$$\boxed{E(X_i) = 1 \cdot p + 0 \cdot q = p}$$

$$\sigma^2(X_i) = E\left((x_i - p)^2\right) = (1-p)^2 p + (0-p)^2 q = \boxed{pq}$$

- Z własności warotści oczekiwanej:

$$E(X = X_1 + X_{2\ldots} + X_n) = \sum_{i=1}^{n} E(X_i) = np$$

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

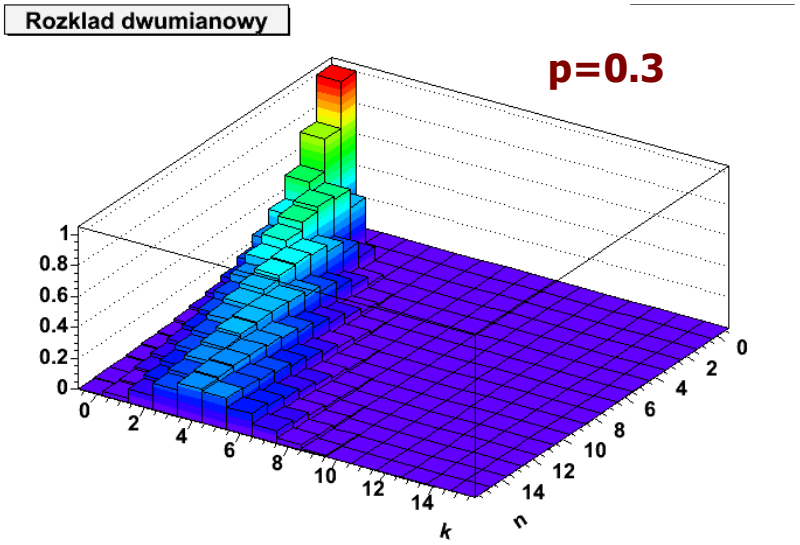
- Zakładając niezależność zmiennych (zerowe kowariancje) otrzymamy z kolei:
- Dla 2 zdarzeń losowych:

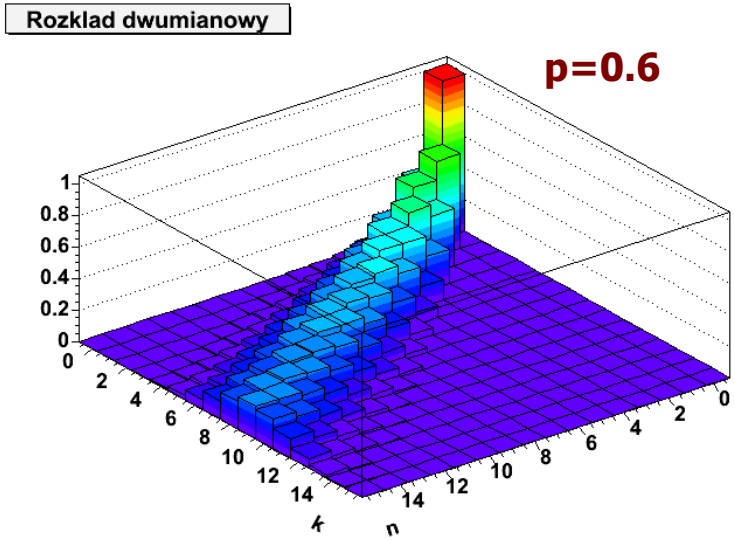
$$\sigma^2(X) = \binom{2}{2} p^2 (2-2p)^2 + \binom{2}{1} pq(1-2p)^2 + \binom{2}{0} q^2 (0-2p)^2 = i$$

$$2p^2(4 - 8p + 4p^2) + 2(p - 2p^2) + (1 - 2p + p^2)4p^2 = 2p(1-p) = 2pq$$

Rozkład dwumianowy - właściwości

Dla różnych n, stałe p

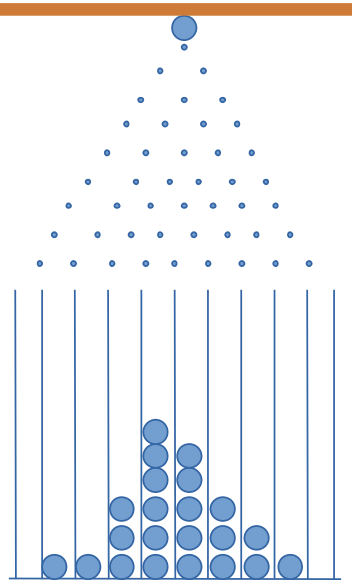




Rozkład dwumianowy – tablica Galtona

- Innym przykładem realizacji rozkładu dwumianowego jest **tablica (deska) Galtona**:
  - mamy n rzędów kołeczków
  - kuleczka może przesunąć się w lewo (z prawdopod. p=0,5) lub w prawo (q=0.5)
  - kuleczka przesunie się k razy w lewo i n-k razy w prawo
  - każde przesunięcie jest niezależne
  - zatem dla jednej konkretnej konfiguracji (drogi) "spadku" kulki prawdopodobieństwo: $p^k q^{n-k}$
  - jeśli mamy różne konfiguracje przesunięć:

$$P(k) = W_k^n = \binom{n}{k} p^k q^{n-k};\ q = 1 - p$$



deska Galtona na Wydziale Fizyki PW
http://www.if.pw.edu.pl/~pluta/pl/tgak.jpg

Rozkład dwumianowy – inne przykłady

$$P(k) = W_k^n = \binom{n}{k} p^k q^{n-k}; q = 1 - p$$

1) n – ilość studentów na 3 roku fizyki
p – prawdopodobieństwo zaliczenia KADD (załóżmy, że p>0.5 :))
k – ilość osób, które przedmiot zaliczyły

1) n – liczba dzieci urodzonych w 2015 roku
p – prawdopodobieństwo, że urodzi się dziewczynka (p=0,5)
k – ilość urodzonych dziewczynek

2) Małe i duże ryby w stawie
n - liczba wszystkich ryb
p - prawdopodobieństwo złowienia dużej ryby
k - liczba dużych ryb

Rozkład Poissona

- Rozważmy rozkład dwumianowy: $P(k) = W_k^n = \binom{n}{k} p^k q^{n-k}\,;\, q = 1 - p$
- dla $n \to \infty$ ale przy stałym $np = \lambda$ rozkład dwumianowy dąży do **rozkładu Poissona** (wyprowadzenie – Brandt):

$$\lim_{n\to\infty} W_k^n = f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

  - normalizacja:

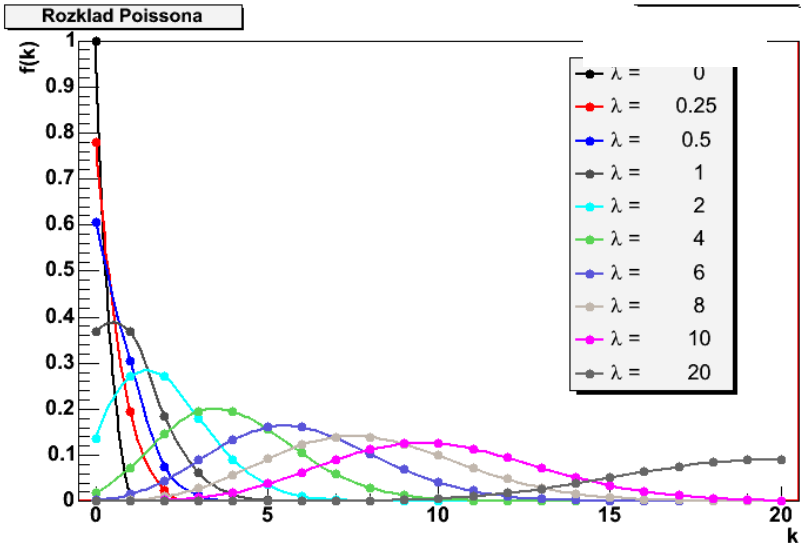
$$\sum_{k=0}^{\infty} f(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right) = e^{-\lambda} e^{\lambda} = 1$$

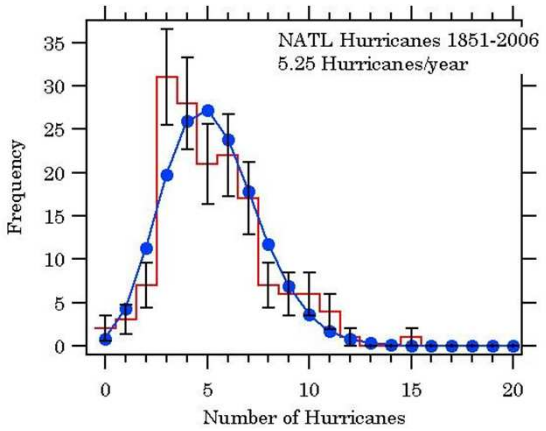
  - wartość oczekiwana: $E(K) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} e^{-\lambda} = \lambda$
  - wariancja: $\sigma^2(K) = E(K^2) - (E(K))^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda$
  - Skosność i wsp. asymetrii:

$$\mu_3 = E\left((k-k)^3\right) = \lambda \qquad \gamma = \frac{\mu_3}{\sigma^3} = \frac{\lambda}{\lambda^{3/2}} = \lambda^{-1/2}$$

Rozkład Poissona - przykłady

- Rozkład Poissona stosujemy wtedy, gdy mamy dużą liczbę niezależnych zdarzeń, z których tylko nieliczne mają interesującą nas własność (duże n, małe p w rozkł. dwumianowym)
- Rozkład Poissona występuje tam, gdzie mamy zjawiska dyskretne, gdy prawdopodobieństwo wystąpienia zjawiska jest stałe w czasie lub przestrzeni:
  - liczba połączeń przychodzących do centrali na minutę
  - liczba mutacji w danym odcinku DNA po ekspozycji na pewną dawkę promieniowania
  - liczbę zabitych każdego roku przez kopnięcie konia w korpusie kawalerii w Prusach (Wikipedia)

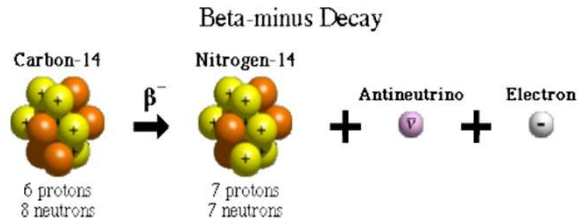


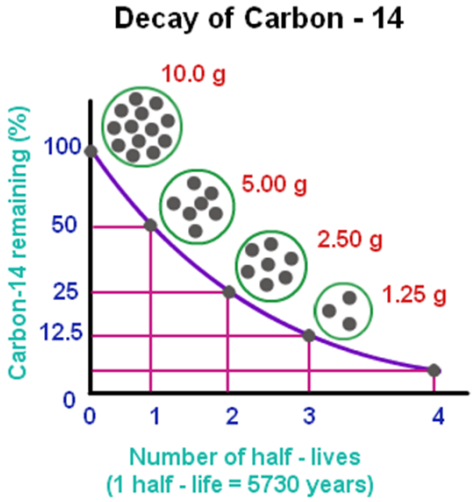


Rozkład Poissona – przykłady

- Mamy jądro promieniotwórcze o czasie życia τ. Obserwujemy je w czasie $T \ll \tau$. Prawdopodobieństwo rozpadu jądra w tym czasie $W \ll 1$. Dzielimy czas T na n przedziałów, prawdopodobieństwo: $p = W/n$.
- Obserwujemy w czasie T źródło zawierające N jąder. Liczba przedziałów czasowych n_k, w których zaobserwowano $k = 0, 1, 2, 3$ itd. rozpadów. Wtedy częstość $h(k) = n_k/n$.
- Doświadczalnie zaobserwowano, że dla $N \to \infty$ i dużych n rozkład $h(k)$ dąży do rozkładu Poissona, co stanowi bezpośredni dowód na niezależność i statystyczny charakter rozpadów promieniotwórczych (badania Rutherforda i Geigera).

Analogicznie – częstość obserwowania k gwiazd w elemencie kąta bryłowego sfery niebieskiej lub k rodzynek w jednostkowym elemencie objętości ciasta





Rozkład potęgowy

Rozkład potęgowy $p(x) = Cx^{-\alpha}$ ***power law distribution***

fat tails tłuste ogony

Cechy rozkładu potęgowego

1. *Scale in variance* $p(ax) = C(ax)^{-\alpha} = a^{-\alpha}p(x) \sim p(x)$
2. Wartości średnia

$$\langle x \rangle = \int_{X\,\min}^{\infty} x p(x)\,dx = C\int_{X\,\min}^{\infty} X^{-\alpha+1}dx$$

$$= \frac{C}{2-\alpha}\left[x^{-\alpha+2}\right]_{X\,\min}^{\infty}.$$

$\alpha \le 2$	∞
$\alpha > 2$	$\langle x \rangle = \frac{\alpha-1}{\alpha-2}x_{\min}$

3. Wariancje

$$\left\langle x^2 \right\rangle = \frac{C}{3-\alpha}\left[x^{-\alpha+3}\right]_{X\,\min}^{\infty}$$

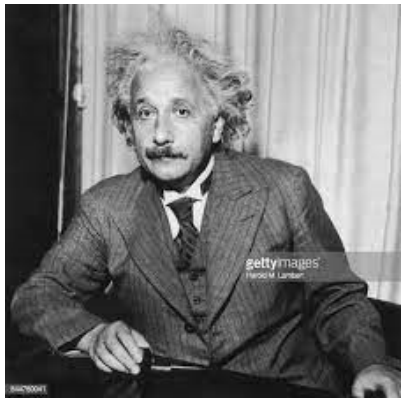
$\alpha \le 3$	∞
$\alpha > 3$	$\left\langle x^2 \right\rangle = \frac{\alpha-1}{\alpha-3}x_{\min}^2$

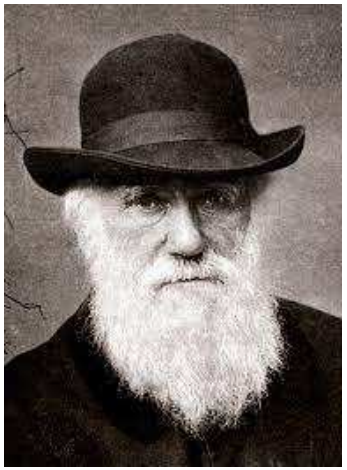
4.Dystrybuanta tez jest funkcja potęgowa

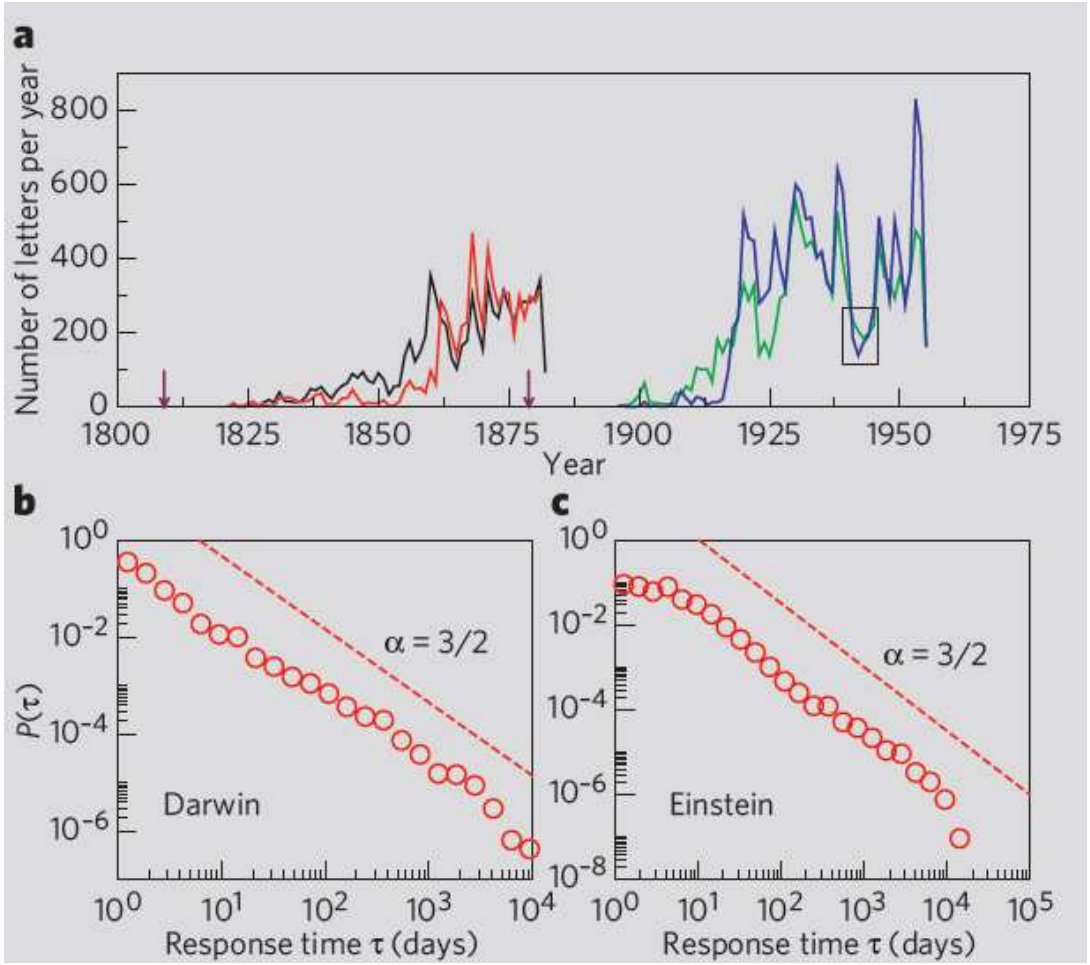
Konsekwencją posiadania nieskończonej średniej lub odchylenia standardowego jest to, że centralne twierdzenie graniczne nie zachowuje się dla takich rozkładów, a zatem średnia i wariancja próbki (która zawsze będzie skończona) nie mogą być używane jako estymatory średniej populacji i wariancji

Rozkład potęgowy- przykłady

Rozkład potęgowy $p(x) = Cx^{-\alpha}$ ***power law distribution***

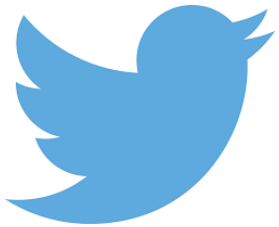


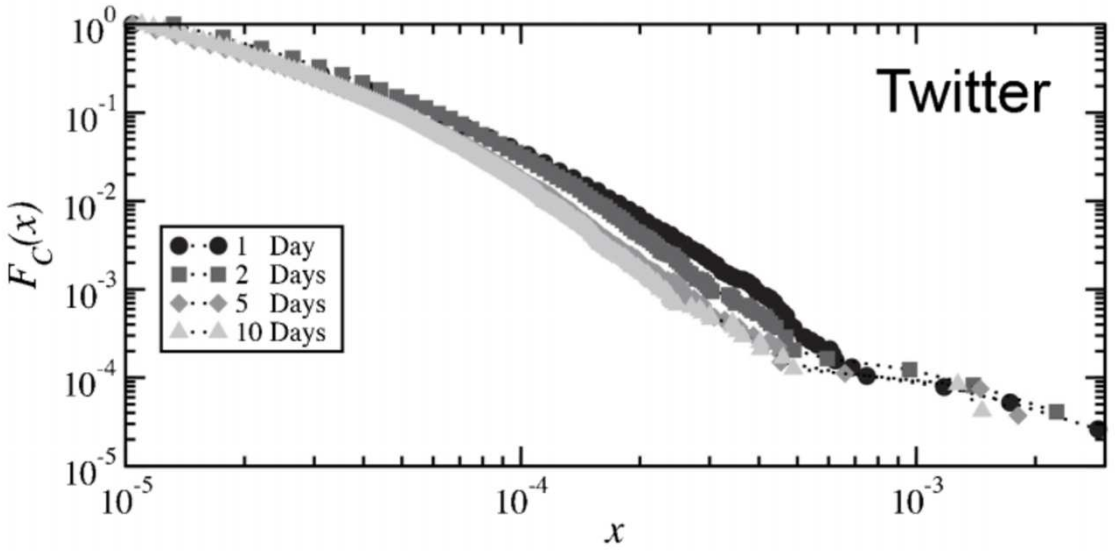


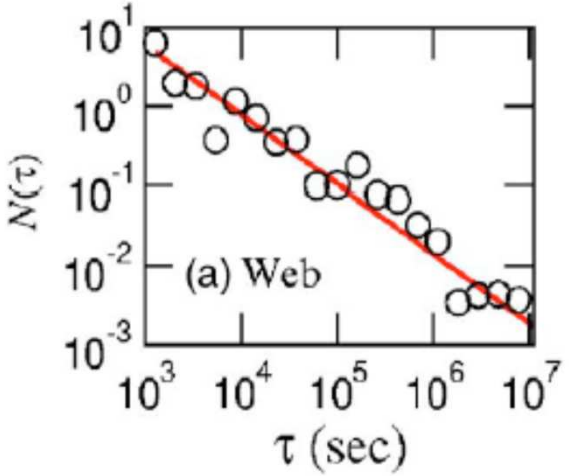


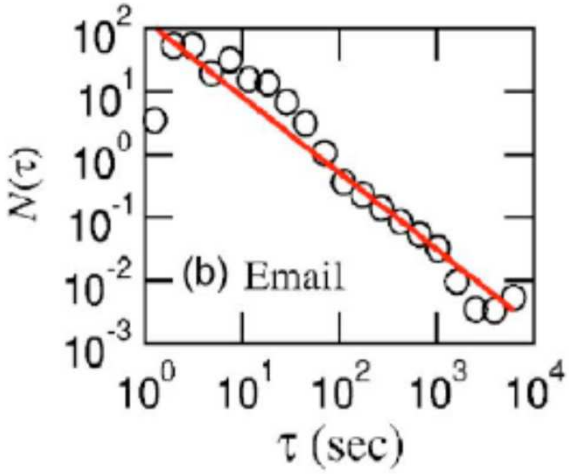
Rozkład potęgowy- przykłady

Rozkład potęgowy $p(x) = Cx^{-\alpha}$ ***power law distribution***









Rozkład potęgowy- przykłady

Astronomia
- Kepler’s trzecie prawo Kryminologia

Fizyka
- Prawo Stefana – Boltzmanna
- Prawa odwrotnych kwadratów grawitacji Newtona i elektrostatyki,
- Samoorganizująca się krytyczność z punktem krytycznym jako atraktorem
- Model siły van der Waalsa
- Siła i potencjał w prostym ruchu harmonicznym
- Korekcja gamma związana z natężeniem światła i napięciem
- Zachowanie w pobliżu przejść fazowych drugiego rzędu z udziałem krytycznych wykładników

Biologia
- Prawo Taylora dotyczące średniej wielkości populacji i wariancji wielkości populacji w ekologii
- Lawiny neuronalne

Meteorologia
- Wielkość ogniw deszczowni
- rozpraszanie energii w cyklonach

Ekonomia
- Wielkość populacji miast w regionie lub sieci miejskiej, prawo Zipfa.
- Podział dochodu w gospodarce rynkowej.

Finanse
- Średnia bezwzględna zmiana logarytmicznych cen średnich

Rozkład potęgowy- przykłady

Nauka ogólna

- Wzrost wykładniczy i obserwacja losowa (lub zabijanie)
- Postęp poprzez wykładniczy wzrost i wykładniczą dyfuzję innowacji
- Różowy szum
- Populacje miast (prawo Gibrata)
- Bibliogramy i częstości występowania słów w tekście (prawo Zipfa)
- Prawo Richardsona dotyczące dotkliwości konfliktów zbrojnych (wojny i terroryzm)
- Zależność między rozmiarem pamięci podręcznej procesora a liczbą braków pamięci podręcznej jest zgodna z prawem mocy braków pamięci podręcznej.

Matematyka

- Fraktale
- Rozkład Pareto i zasada Pareto zwana także „regułą 80–20”
- Dystrybucja Yule – Simon (dyskretna)
- Rozkład t-Studenta (ciągły), którego szczególnym przypadkiem jest rozkład Cauchy’ego
- Prawo Lotki
- Model sieci bez skali

Rozkład wykładniczy

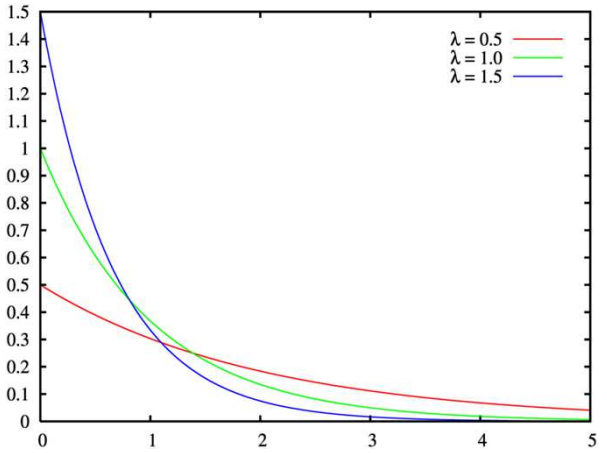
- Gęstość prawdopodobieństwa: $f(x) = \lambda e^{-\lambda x};\ x \geq 0;\ \lambda > 0$

$$f(x) = 0;\ x < 0$$

- Dystrybuanta: $F(x) = 0;\ x < 0$

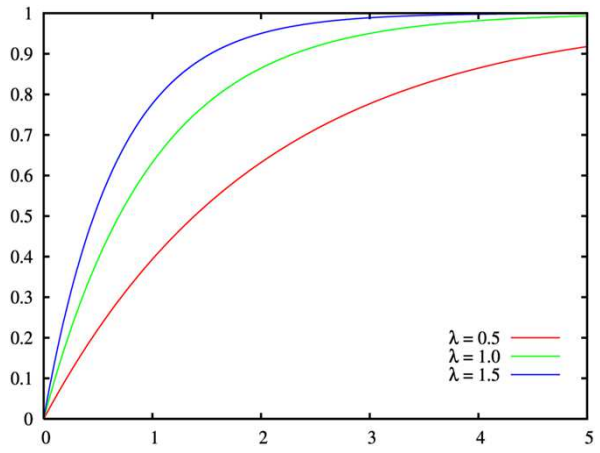
$$F(x) = \int_0^x f(x)dx = \lambda \int_0^x e^{-\lambda x'} dx' = \left[\frac{-\lambda}{\lambda} e^{-\lambda x'} \right]_0^x$$

$$F(x) = 1 - e^{-\lambda x};\ x \geq 0$$



- Wartość oczekiwana:

$$E(x) = \hat{x} = \int_0^\infty x f(x)dx = \lambda \int_0^\infty e^{-\lambda x} x dx = \boxed{\frac{1}{\lambda}}$$



- Wariancja:

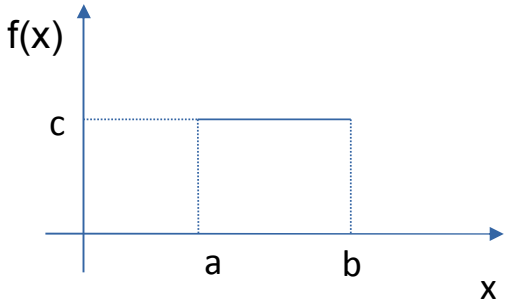
$$E(x^2) = \int_0^\infty x^2 f(x)dx = \frac{2}{\lambda^2}$$

$$\sigma^2(x) = E(x^2) - (E(x))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \boxed{\frac{1}{\lambda^2}}$$

Rozkład jednostajny

- Gęstość prawdopodobieństwa: $f(x)=c;\ x\in\langle a,b\rangle$

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



- Współczynnik (normalizacja) c:

$$\int_{-\infty}^{\infty} f(x)dx = c\int_{a}^{b} dx = c(b-a)=1 \Rightarrow c=\frac{1}{b-a}$$

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

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

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

$$E(X^2)=\frac{1}{b-a}\int_{a}^{b} x^2dx=\frac{(b^3-a^3)}{3(b-a)}=$$

$$=\frac{(b-a)(b^2+ba+a^2)}{3(b-a)}=\frac{b^2+ba+a^2}{3}$$

$$\sigma^2(X)=\frac{b^2+ba+a^2}{3}-\left(\frac{b+a}{2}\right)^2=$$

$$=\frac{b^2+ba+a^2}{3}-\frac{b^2+2ba+a^2}{4}=\boxed{\frac{(b-a)^2}{12}}$$

- Wartość oczekiwana:

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

Rozkład wielomianowy – uogólnienie rozkładu dwumianowego

- Uogólnienie, gdy mamy więcej możliwości niż dwie (sukces i porażka)
- Jeśli przestrzeń zdarzeń elementarnych: $E = A_1 + A_2 + \ldots + A_l$
- Zdarzenia się wzajemnie wykluczają: $P(A_j) = p_j, \quad \sum_{j=1}^{l} p_j = 1$
- To prawdopodobieństwo zajścia k_j razy zdarzenia A_j:

$$P = W^n_{k_1, k_2, \ldots, k_l} = \frac{n!}{\prod_{j=1}^{l} k_j!} \prod_{j=1}^{l} p_j^{k_j}, \quad \sum_{j=1}^{l} k_j = n$$

- Taki rozkład nazywamy **rozkładem wielomianowym**
- Jeśli zdefiniujemy zmienne losowe X_{ij} równe 1, gdy wynikiem i-tego doświadczenia jest zdarzenie A_j, lub równe 0 w przeciwnym razie, oraz

- Wtedy wartość oczekiwana i kowariancja:

$$X_j = \sum_{i=1}^{n} X_{ij}$$

$$E(X_j) = \hat{x}_j = n\, p_j \qquad c_{ij} = np_i(\delta_{ij} - p_j)$$

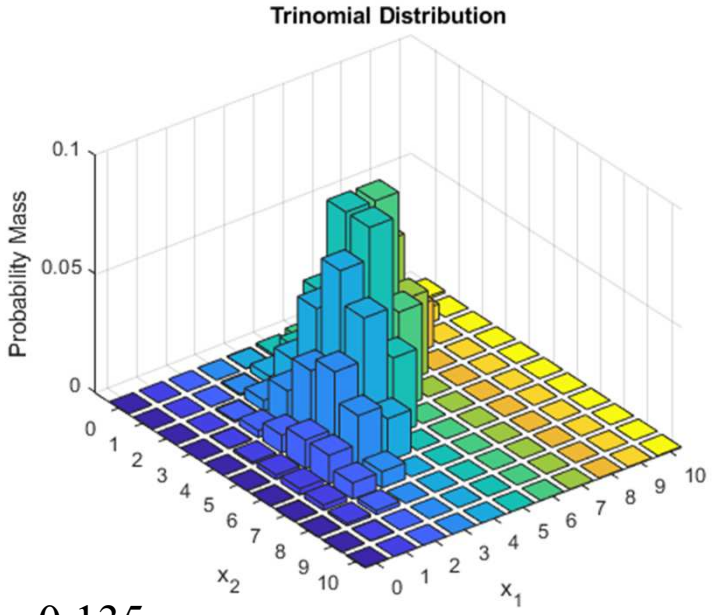
Rozkład wielomianowy – uogólnienie

- Przykład – gra w karty: troje graczy (A, B ,C) rozgrywa serię gier:
  prawdopodobieństwo, że gracz A wygra dowolną grę jest 20%
  prawdopodobieństwo, że gracz B wygra dowolną grę jest 30%
  prawdopodobieństwo, że gracz C wygra dowolną grę jest 50%
- Jeśli rozegrają 6 gier, jakie jest prawdopodobieństwo, że gracz A wygra 1 grę, gracz B wygra 2 gry, a gracz C wygra 3 gry?

Rozkład wielomianowy – uogólnienie

- Przykład – gra w karty: troje graczy (A, B ,C) rozgrywa serię gier:
  prawdopodobieństwo, że gracz A wygra dowolną grę jest 20%
  prawdopodobieństwo, że gracz B wygra dowolną grę jest 30%
  prawdopodobieństwo, że gracz C wygra dowolną grę jest 50%
- Jeśli rozegrają 6 gier, jakie jest prawdopodobieństwo, że gracz A wygra 1 grę, gracz B wygra 2 gry, a gracz C wygra 3 gry?

$$P = W^{n}_{k_1, k_2, \ldots, k_l} = \frac{n!}{\prod_{j=1}^{l} k_j!} \prod_{j=1}^{l} p_j^{k_j}$$



n = 6 – liczba gier
k_1 = 1 – wygrywa gracz A
(ilość sukcesów zdarzenia A_1)
k_2 = 2 – wygrywa gracz B
k_3 = 3 – wygrywa gracz C
p_1 = 0.2 – prawd. wygrania gracza A
p_2 = 0.3 – prawd. wygrania gracza B
p_3 = 0.5 – prawd. wygrania gracza C

$$P(A=1, B=2, C=3) = \frac{6!}{1!2!3!} 0.2^1 \cdot 0.3^2 \cdot 0.5^3 = 0.135$$

Częstość i prawo wielkich liczb

Prawdopodobieństwo a priori to prawdopodobieństwo obliczane przed realizacją doświadczenia losowego, czyli klasyczne prawdopodobieństwo
Prawdopodobieństwa a posteriori, obliczanego na podstawie wyników doświadczenia, czyli **częstości**.

- Definicja prawdopodobieństwa – przeprowadzenie n prób dostatecznie dużo razy (N) umożliwia pomiar prawdopodobieństwa zdarzenia A

$$P(A) = \lim_{N \to \infty} \frac{n}{N}$$

- W rzeczywistości nie zawsze znamy prawodpodobieństw zdarzeń (np. p_j w rozkł. wielomianowym) – wyznaczamy je eksperymentalnie
- **Częstość** wystąpienia zdarzenia A_j w n doświadczeniach będzie określona wzorem:

$$H_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij} = \frac{1}{n} X_j$$

- Częstość jest zmienną losową, dla ktorej (przy n próbach):

$$E(H_j) = h_j = E\left(\frac{X_j}{n}\right) = p_j \qquad \sigma^2(H_j) = \sigma^2\left(\frac{X_j}{n}\right) = \frac{1}{n^2}\sigma^2(X_j) = \frac{1}{n} p_j(1 - p_j) = \frac{1}{n} p_j q_j$$

Częstość i prawo wielkich liczb

- **Częstość** wystąpienia zdarzenia A_j w n doświadczeniach będzie określona wzorem:

$$H_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij} = \frac{1}{n} X_j$$

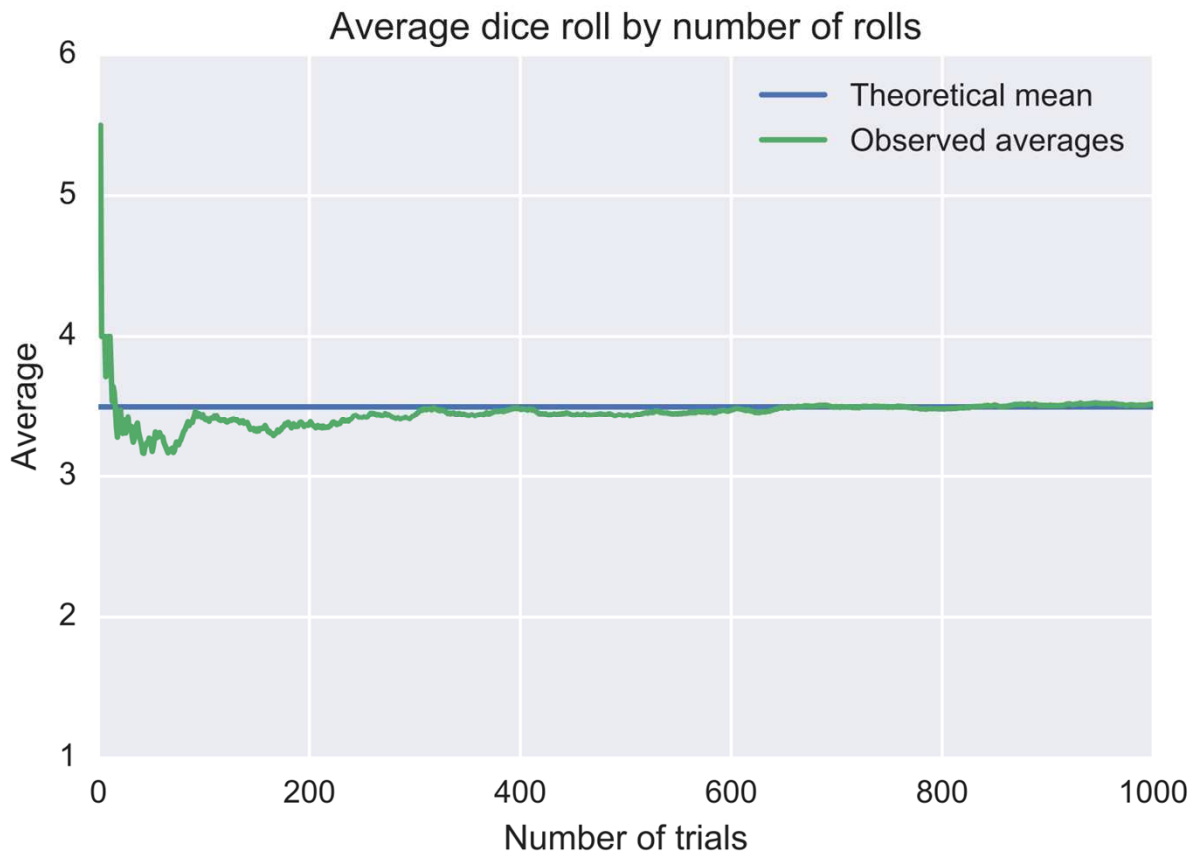
Wartość oczekiwana częstości jest równa jego prawdopodobieństwu.

- Odchylenie standardowe częstości jest mniejsze niż $1/\sqrt{n}$ i może osiągać dowolnie małe wielkości (gdy n→∞). Jest to **prawo wielkich liczb**

- Możemy zatem użyć **częstości jako przybliżonej wartości prawdopodobieństwa** z odpowiednia niepewnością jej wyznaczenia

- **Kwadrat niepewności jest w przybliżeniu odwrotnie proporcjonalny do liczby przeprowadzonych prób** – jest to niepewność statystyczna

Częstość i prawo wielkich liczb



Rozkład hipergeometryczny

- W urnie jest N kul – k białych i N-K czarnych
- W **n** próbach wyciągamy (bez zwracania) k kul białych i n-k=l kul czarnych.

Jakie jest prawdopodobieństwo wyciągnięcia k kul białych?

- Wylosowanie kolejnej kulki zmienia proporcje kul białych do czarnych i wpływa na wynik kolejnego losowania – rozkład dwumianowy nie ma tu zastosowania. Mamy jednak:
  - liczbę możliwości wylosowania n z N kulek: $\binom{N}{n}$
  - prawdopodobieństwo takiego zdarzenia: $1/\binom{N}{n}$
  - możliwość wylosowania k spośród K białych $\binom{K}{k}$
  - i l spośród L czarnych kulek wynoszą: $\binom{L}{l}$
  - prawdopodobieństwo szukane wynosi zatem: $W_k = \dfrac{\binom{K}{k}\binom{L}{l}}{\binom{N}{n}}$
- Analogicznie jak w rozkładzie dwumianowym, definiujemy zmienną losową:

$$X = \sum_{i=1}^{n} X_i$$

Rozkład hipergeometryczny

- Analogicznie jak w rozkładzie dwumianowym, definiujemy zmienną losową:

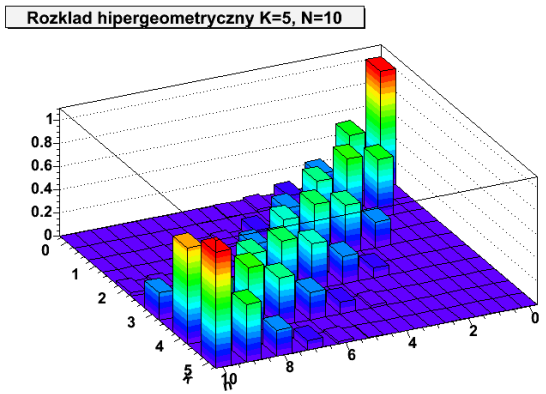
$$X = \sum_{i=1}^{n} X_i$$

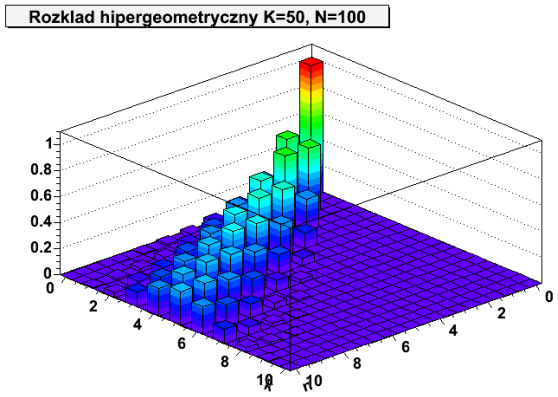
  - X_i przyjmuje wartość 1 dla białych i 0 dla czarnych wylosowanych kul
- Można pokazać, że (Brandt):

$$E(X) = n\frac{K}{N} \qquad \sigma^2(X) = \frac{nK(K-N)(N-n)}{N^2(N-1)}$$

- Dla $n \ll N$ rezultat kolejnego losowania niewiele wpływa na następne wyniki. Wtedy rozkład hipergeometryczny upodabnia się do dwumianowego:

$$p = \frac{K}{N}, \quad q = \frac{N-K}{N}, \quad E(X) = n\frac{K}{N} = np, \quad \sigma^2(X) = \frac{npq(N-n)}{N-1}$$





Rozkład normalny standardowy

- Gęstość prawdopodobieństwa:

$$f(x) \equiv {}_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

  - rozkład o średniej 0 i wariancji 1
- Dystrybuanta nie ma postaci analitycznej (korzystamy z tabel)
- Rozkład jest unormowany:

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$$

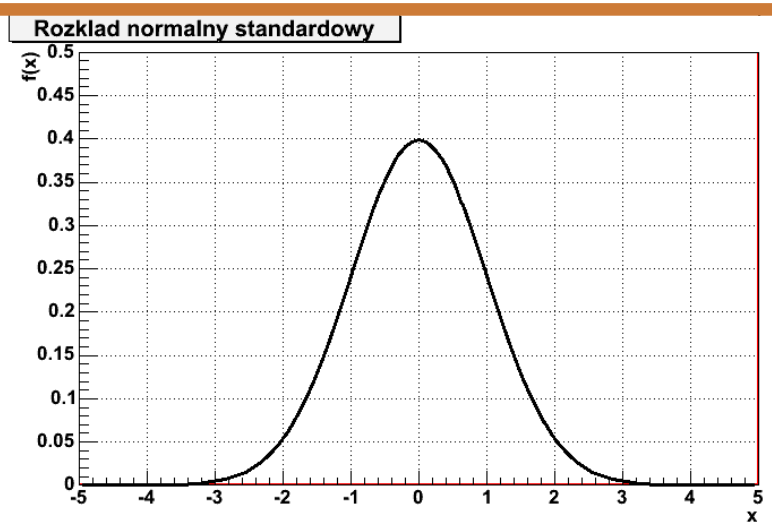
- Jeśli wprowadzimy zmienną:

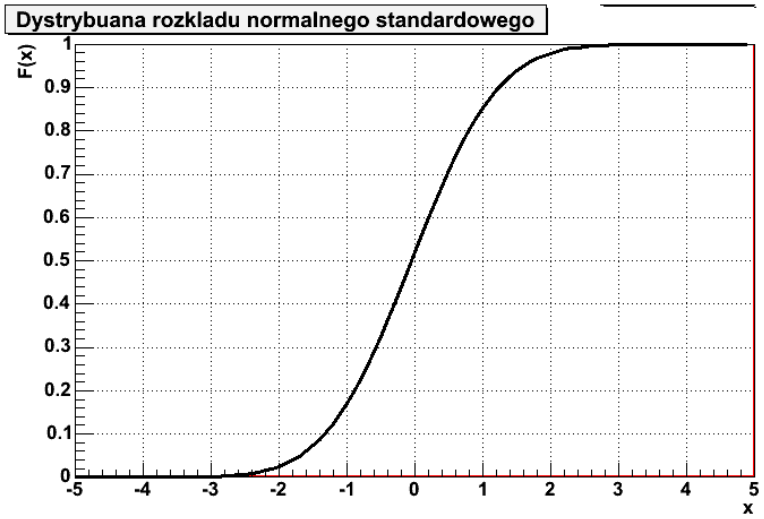
$$Y = (X - a)/b$$

- Otrzymamy rozkład Gaussa:

$$f(y) \equiv (y) = \frac{1}{\sqrt{2\pi b}} e^{-(y-a)^2/2b^2}$$

  - średnia (przesunięcie): $\hat{y} = a$
  - wariancja (szerokość): $\sigma^2(Y) = b$





Rozkład normalny standardowy - własności

- Punkt przegięcia rozkładu:
  - standardowego $x=\pm 1$
  - Gaussa $x=a\pm b$
- Załóżmy, że znamy dystrybuantę:

$$F_0(x)\equiv \Phi_0(x)= P(X\le x)$$

- Ze względu na asymetrię gęstości:

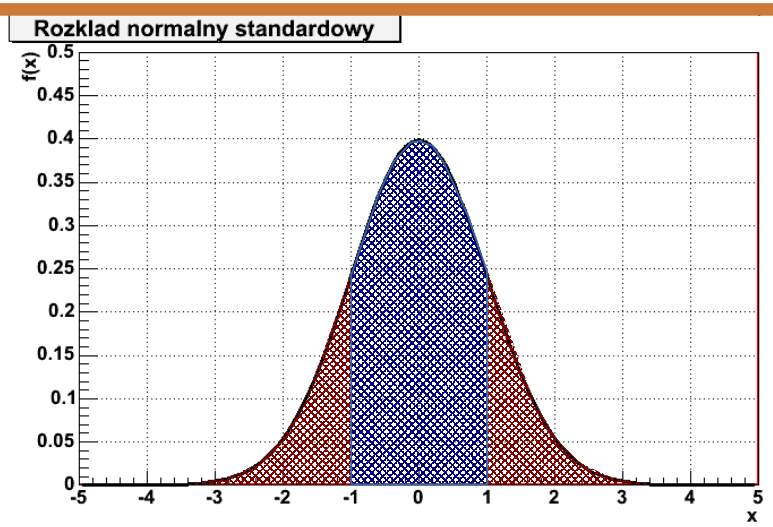
$$P(|X|>x)=2\Phi_0(-|x|)=2(1-{}_0(|x|))$$

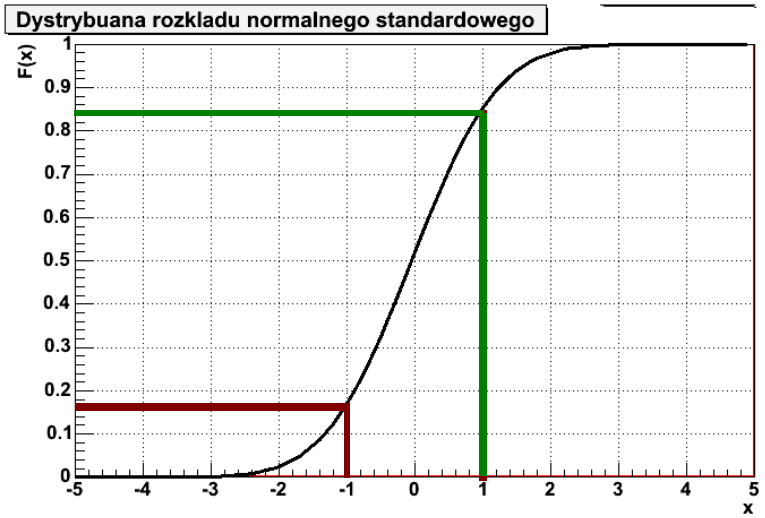
- Analogicznie, wewnątrz przedziału $2x$:

$$P(|X|\ \ x)=2\Phi_0(|x|)-1$$

- Dystrybuantę r. norm. można uogólnić na r. Gaussa:

$$\Phi(y)=\Phi_0\left(\frac{x-a}{b}\right)$$





Rozkład normalny standardowy - własności

- Wtedy szczególnie interesujące jest obliczenie występowania zmiennej los. dla wielokrotności odchylenia standardowego:

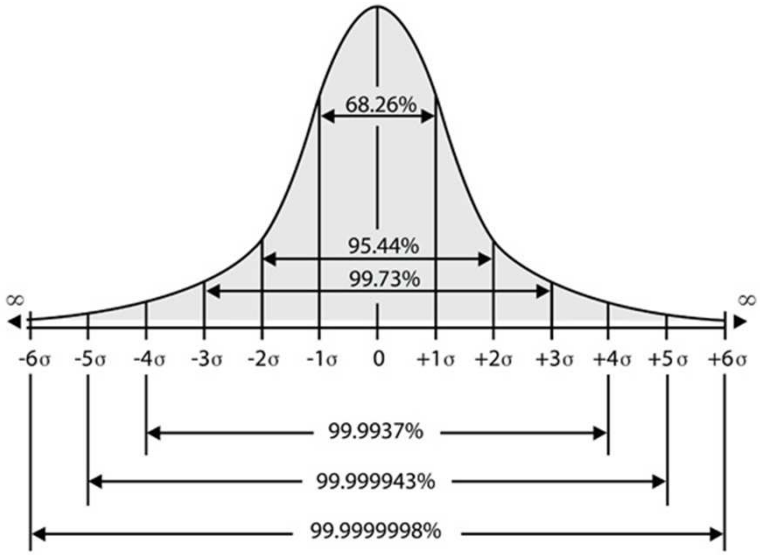
$$P(|Y-a| \ \ n\sigma) = 2\Phi_0\left(\frac{nb}{b}\right) - 1 = 2\Phi_0(n) - 1$$

- Otrzymamy wtedy:

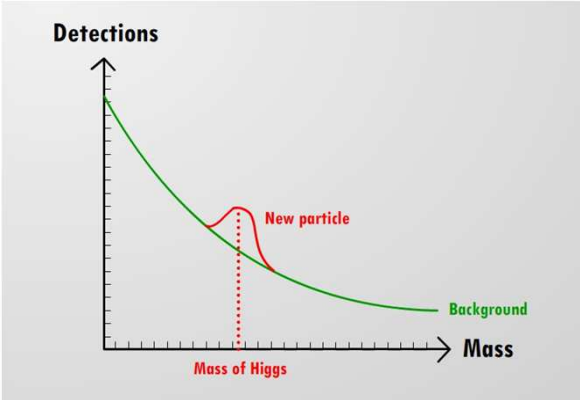
$$P(|Y-a| \ \ \sigma) = 68{,}3\% \qquad P(|Y-a| > \sigma) = 31{,}7\%$$

$$P(|Y-a| 2\sigma) = 95{,}4\% \qquad P(|Y-a| > 2\sigma) = 4{,}6\%$$

$$P(|Y-a| 3\sigma) = 99{,}8\% \qquad P(|Y-a| > 3\sigma) = 0{,}2\%$$



- Z Wykładu 1 pamiętamy, że **współczynnik rozszerzenia** niepewność typu A zwykle jest między 2 a 3 – tu widać dlaczego
- W nauce przez odchylenie standardowe określamy również różnice w obserwowanym sygnale eksperymentalnym w stosunku do sytuacji, gdy efektu fizycznego nie ma

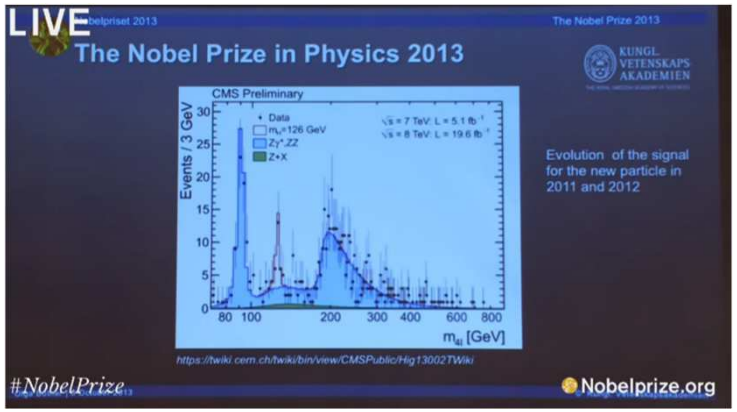


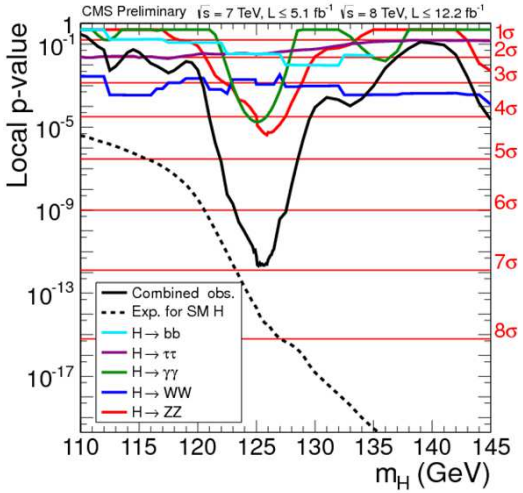
Wielokrotności sigma

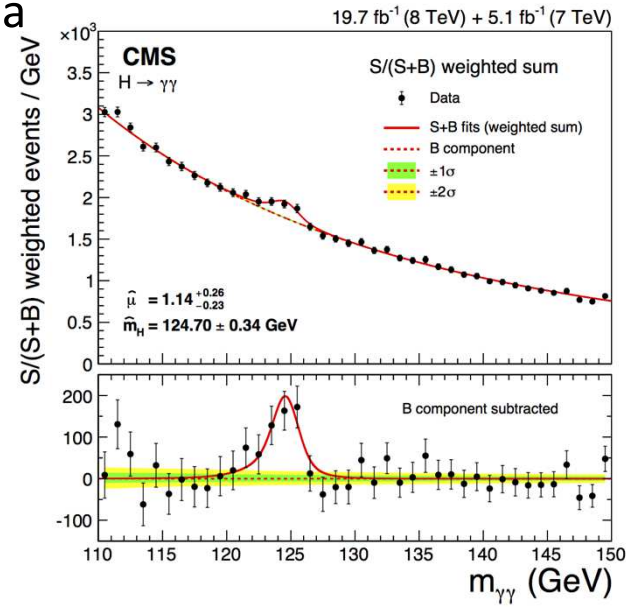
- Idealnym przykładem jest odkrycie bozonu Higgsa
- W fizyce cząstek przyjęło się, że dopiero mając **odchylenie 5σ można mówić o odkryciu**:

$$P(|Y - a| \square 5\sigma) = 99{,}99994\,\%$$

- Różnica na takim poziomie wymagała zebrania dużej ilości danych, stąd potwierdzenie jego istnienia zajęło ponad 3 lata







KONIEC