MODELOWANIE RZECZYWISTOŚCI

Daniel Wójcik

Instytut Biologii Doświadczalnej PAN
Szkoła Wyższa Psychologii Społecznej

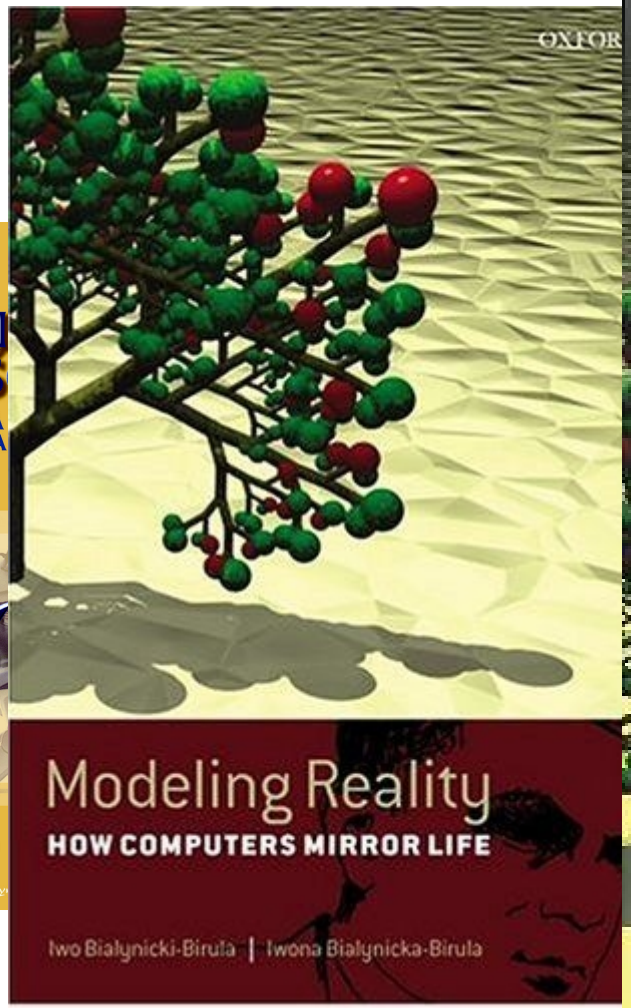
d.wojcik@nencki.gov.pl
dwojcik@swps.edu.pl

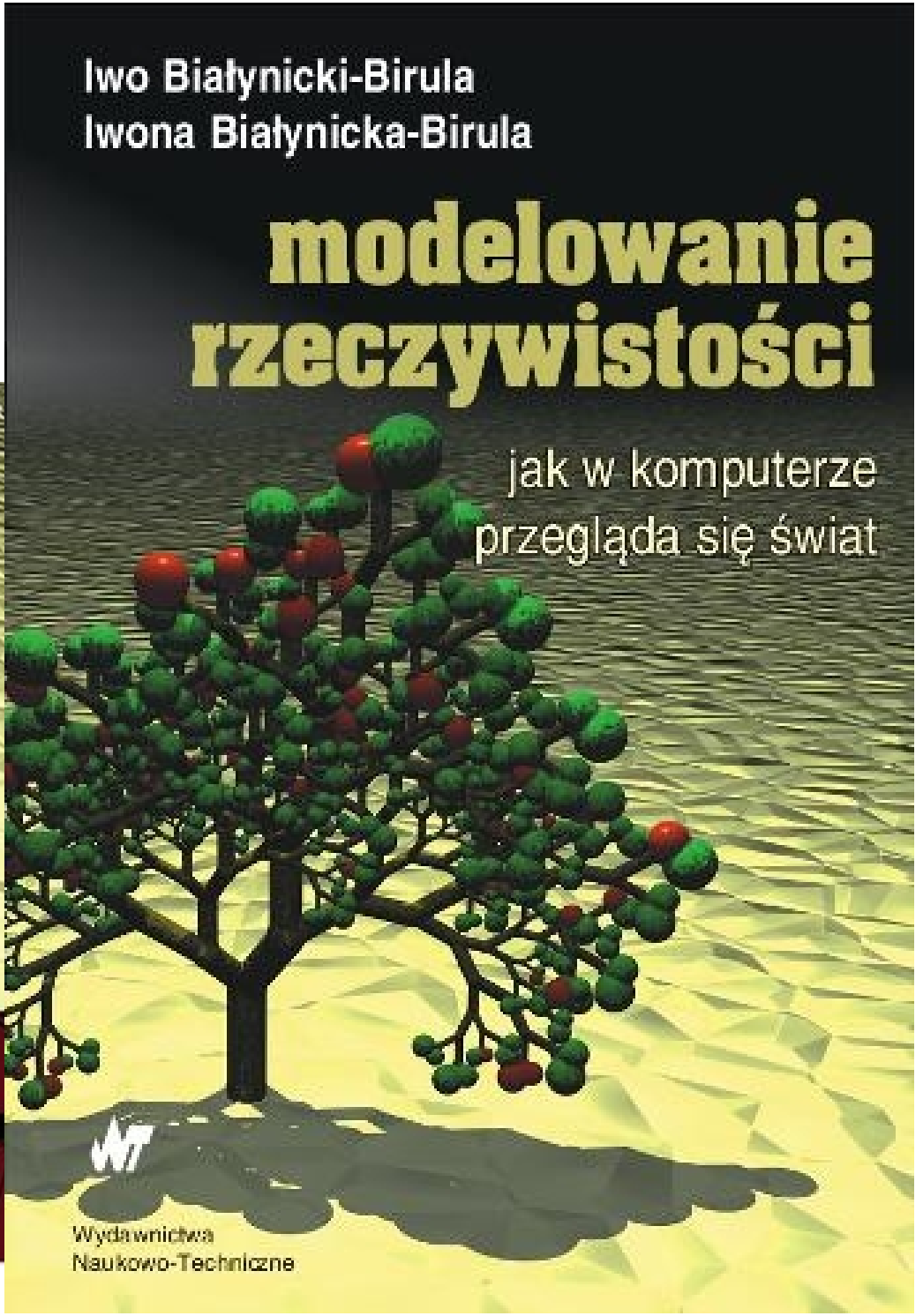
tel. 022 5892 424

http://www.neuroinf.pl/Members/danek/swps/

Podręcznik

Iwo Białynicki-Birula
Iwona Białynicka-Birula





Ciągi liczb losowych

- Czy ciąg 0,1,0,1,0,1,... jest losowy?
- Jaki ciąg uznamy za „typowy przykład ciągu losowego”?
- Napisz ciąg zer i jedynek o długości 128 znaków
- Policz ile jest w tym ciągu podciągów złożonych z trzech, czterech i pięciu jedynek
- Porównaj z wynikami wygenerowanymi przez program **Bernoulli**

Prawdopodobieństwo wylosowania ciągów samych jedynek

- Aby uzyskać samotną jedynkę w binarnym ciągu trzeba wyrzucić po kolei 0,1,0 – prawdopodobieństwo ½ * ½ * ½ = 1/8
- Aby uzyskać n jedynek w binarnym ciągu trzeba wyrzucić po kolei 0,1, ... ,1,0 – to daje prawdopodobieństwo

$$\frac{1}{2} * \underbrace{\frac{1}{2} * \ldots * \frac{1}{2}}_{n\ razy} * \frac{1}{2} * \frac{1}{2} = \frac{1}{2}^{(n+2)}$$

- Zatem prawdopodobieństwo wyrzucenia 6,7,8 jedynek pod rząd to odpowiednio 1/256, 1/512 i 1/1024

Prawdopodobieństwo wylosowania ciągów samych jedynek

Jeżeli prawdopodobieństwo wyrzucenia 8 jedynek pod rząd to 1/1024, ile razy trzeba rzucić monetą, żeby średnio zaobserwować jeden ciąg 8 jedynek w serii?

Ludzkie zachowanie przy zgadywaniu liczb losowych

- Alphone Chapanis, twórca ergonomiki, zauważył, że większość ludzi proszona o wypisanie ciągu liczb losowych starała się unikać powtarzania tej samej liczby trzy razy pod rząd (1953).

- Wiele osób uważało też niektóre liczby za „bardziej losowe” od innych.

Błędne rozumienie przypadkowości

- Jakie jest prawdopodobieństwo tego, że w grupie 30 osób znajdą się dwie urodzone tego samego dnia?

Błędne rozumienie przypadkowości

- Paradoks urodzinowy: prawdopodobieństwo, że w grupie 30 osób znajdą się dwie urodzone tego samego dnia wynosi ponad 70%, chociaż ludziom często wydaje się znacznie mniejsze.
- Takie efekty psychologiczne wykorzystuje się w konstrukcji schematów loteryjnych, żeby wytworzyć u ludzi przekonanie łatwej wygranej.

Błędne rozumienie przypadkowości

- Badania nad ludzkim rozumieniem przypadkowości są wykorzystywane do automatycznego wykrywania oszustw.
- Amerykański Urząd Skarbowy (IRS) nie jest w stanie sprawdzić wszystkich formularzy podatkowych. Najpierw sprawdza te, w których testy statystyczne sugerują nieprawidłowości

Prawo Benforda

- Najważniejszą rolę w wykrywaniu oszustw odgrywa prawo Bedforda:

- Weźmy dowolny rzeczywisty zbiór danych dotyczący jakiejś wielkości, np. liczba ludności krajów, powierzchnie krajów, długości rzek, statystyki sportowe, itd. i wybierzmy z niego jedną liczbę.

- Prawdopodobieństwo, że jej pierwszą cyfrą będzie 1 wynosi ok. 30%.

- Prawdopodobieństwo wystąpienia kolejnych cyfr wynosi:

$$P(d) = \log\left(1 + \frac{1}{d}\right)$$

Prawo Benforda

- Program **Benford**

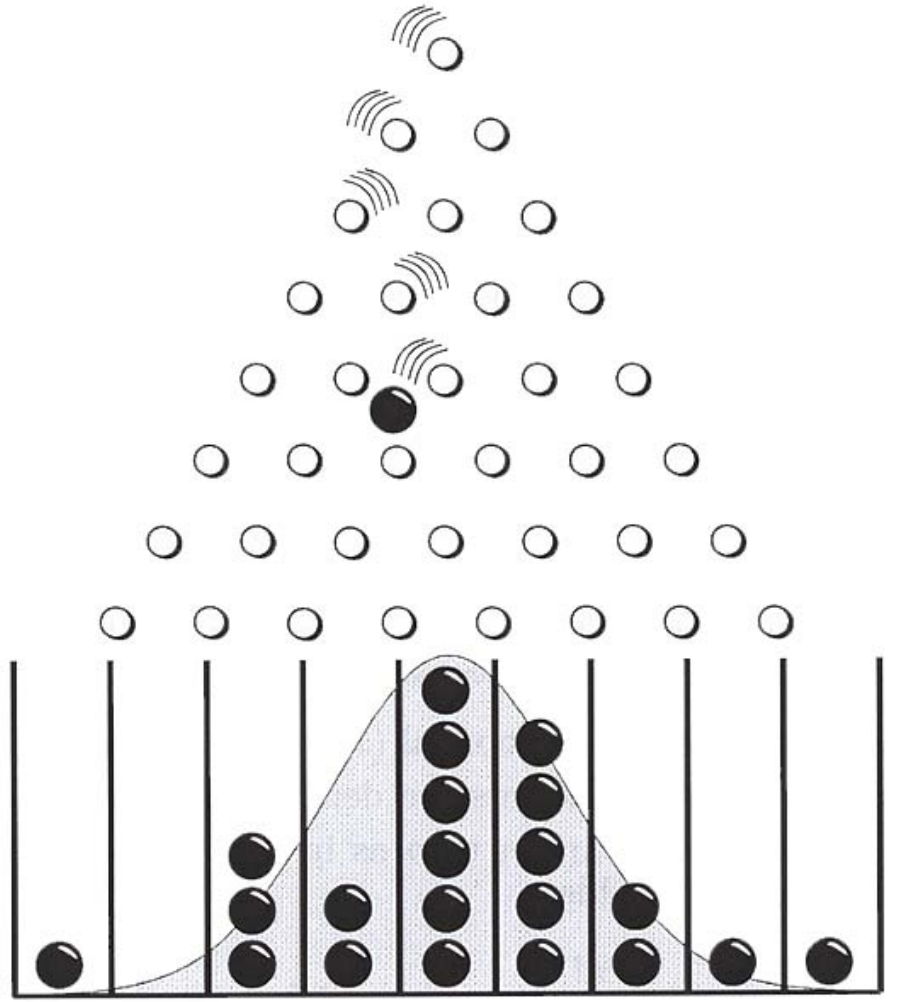
- Prawo to jest nieintuicyjne – większość ludzi spodziewa się, że każda cyfra powinna być równie prawdopodobna.
- Wynika ono z faktu, że jeżeli istnieje prawo dotyczące rozkładu wszystkich cyfr w rzeczywistych zbiorach danych, to musi być ono niezmiennicze ze względu na skalowanie (zmianę jednostek).
- Jedyny taki rozkład to rozkład logarytmiczny dany w prawie Benforda.

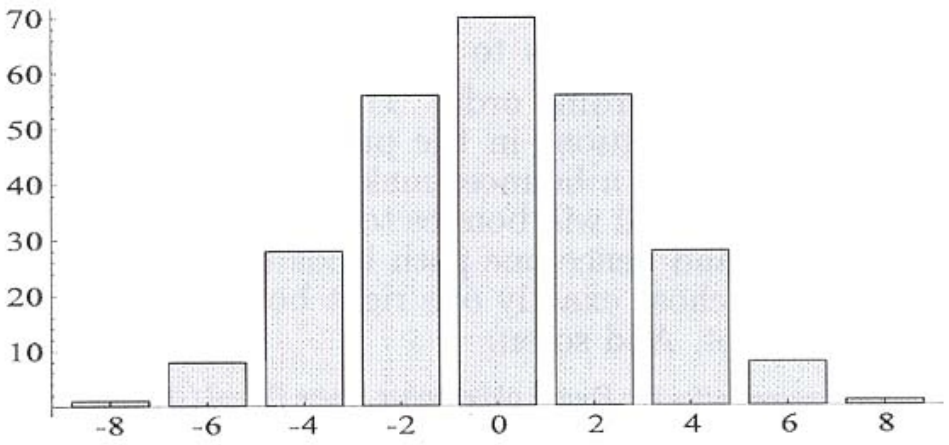
Prawo Benforda

- Prawo Benforda stosowano wielokrotnie do wykrywania nadużyć i błędów, zarówno ludzkich jak i maszynowych (komputerowych).
- Przy jego pomocy nie można wykryć jednego błędu w dużym zbiorze danych, ale można stwierdzić niewłaściwą procedurę generacji lub przetwarzania danych.

Deska Galtona

Pochylona deska z wbitymi gwoździami ułożonymi w trójkąt. Można jej użyć do wizualizacji wielokrotnego rzucania monetą





Deska Galtona

- Jeżeli prawdopodobieństwo skoku w prawo lub w lewo na każdym gwoździu jest takie samo, to prawdopodobieństwo rozkładu na ostatnim poziomie dane jest przez trójkąt Pascala
- Program **Galton**

n \ s	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
0										1									
1									1		1								
2								1		2		1							
3							1		3		3		1						
4						1		4		6		4		1					
5					1		5		10		10		5		1				
6				1		6		15		20		15		6		1			
7			1		7		21		35		35		21		7		1		
8		1		8		28		56		70		56		28		8		1	
9	1		9		36		84		126		126		84		36		9		1

Trójkąt Pascala

- Prawdopodobieństwo dojścia do najdalszego końca wynosi $1/2^n$. Łatwo to zgadnąć, bo istnieje tylko jedna droga, która tam prowadzi.
- Prawdopodobieństwo dojścia do pozycji s przy n rzutach (odbiciach) jest równe sumie prawdopodobieństw uzyskania s-1 w n-1 rzutach i odbicia w prawo, powiększonej o prawdopodobieństwo uzyskania s+1 w n-1 rzutach i odbicia w lewo.
- Tą regułę opisuje trójkąt Pascala.

Trójkąt Pascala

Trójkąt Pascala powstaje przy obliczaniu n-tej potęgi dwumianu. Każdy współczynnik w trójkącie Pascala równy jest liczbie dróg jakimi można do niego dojść

n \ s	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
0										1									
1									1		1								
2								1		2		1							
3							1		3		3		1						
4						1		4		6		4		1					
5					1		5		10		10		5		1				
6				1		6		15		20		15		6		1			
7			1		7		21		35		35		21		7		1		
8		1		8		28		56		70		56		28		8		1	
9	1		9		36		84		126		126		84		36		9		1

Współczynniki dwumianu

- Liczby w n-tym wierszu trójkąta Pascala, zwane współczynnikami dwumianowymi, oznaczane są symbolem Newtona i dane są wzorem:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Zatem prawdopodobieństwa trafienia do odpowiedniej przegródki wynoszą

$$p(n,k) = \frac{1}{2^n}\frac{n!}{k!(n-k)!}$$

- Zachodzi

$$\sum_{k=0}^{n} p(n,k) = 1$$

Błądzenie przypadkowe

- Ruch punktu na prostej lub w przestrzeni o dowolnym wymiarze polegający na wykonywaniu losowych kroków o stałej długości w jednym z kilku wybranych kierunków.
- Deska Galtona jest równoważna błądzeniu przypadkowemu na prostej.
- Błądzenie przypadkowe jest modelem ruchu cząstki Browna

Spacery losowe – model dyfuzji

- Błądzenie przypadkowe (spacer losowy) w przestrzeni stanowi model dyfuzji i ruchów Browna – rozprzestrzeniania się cząsteczek w danym środowisku.
- Średnia odległość od punktu początkowego rośnie z czasem t jak $\sqrt{t}$
- Program **Smoluchowski**

Wzór Stirlinga

- Prawdopodobieństwo powrotu do miejsca startu wynosi

$$p(2n,n)=\frac{(2n)!}{2^{2n} n!\, n!}$$

- Dla n = 2,4,6,8,10,12... otrzymujemy 1/2, 3/8, 5/16, 35/128, 63/256, 231/1024...
- Ze wzoru Stirlinga

$$n! \approx \sqrt{2\pi n}\, n^n e^{-n}$$

otrzymujemy przybliżenie:

$$p(2n,n) \approx \frac{\sqrt{2}}{\sqrt{\pi n}}$$

Asymptotyka rozkładu dwumianowego

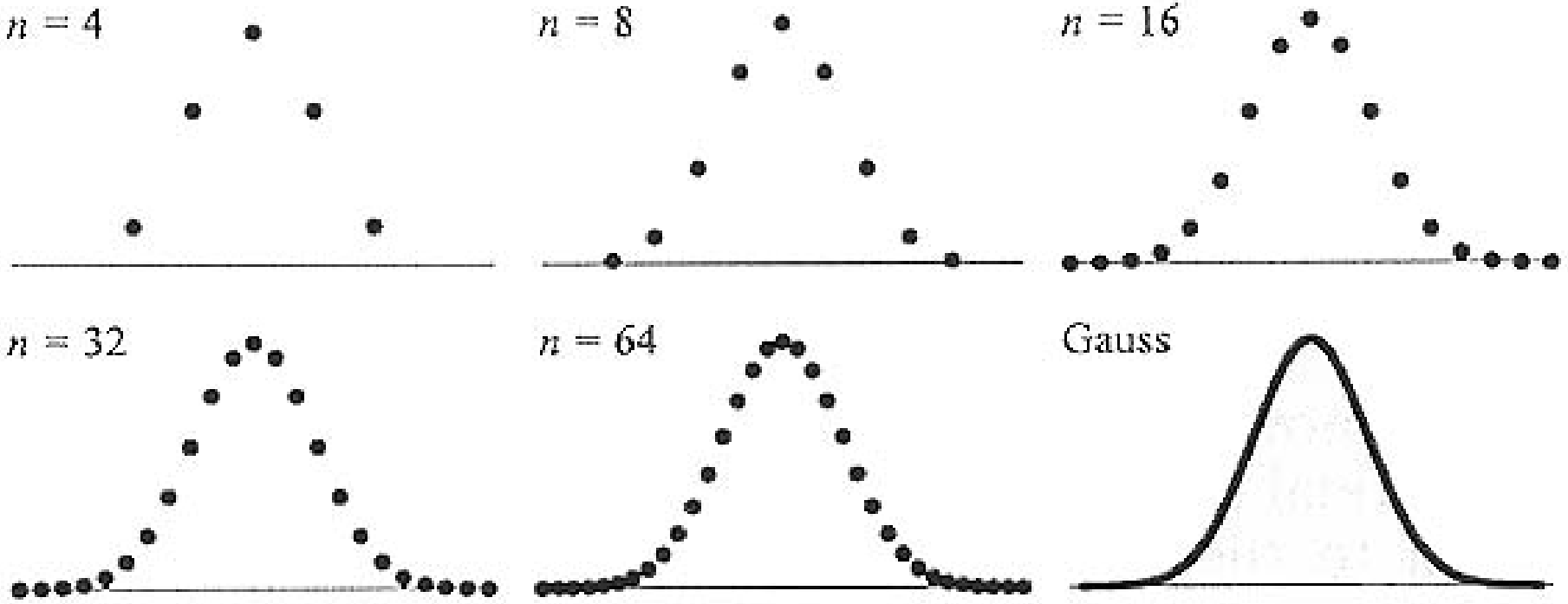
- Widzimy, że prawdopodobieństwo powrotu cząstki do punktu startu maleje do 0
- Zauważmy, że także liczba możliwych wyników rośnie. Żeby badać zachowanie rozkładu dla dużych *n* musimy go znormalizować.
- Mnożąc prawdopodobieństwa przez $\sqrt{n/2}$ i tak samo kurcząc skalę na osi *x* otrzymujemy dla dużych *n* rozkład Gaussa:

$$p(x) = \sqrt{1/\pi}\exp(-x^2)$$

Asymptotyka rozkładu dwumianowego

- Mnożąc prawdopodobieństwa przez $\sqrt{n/2}$ i tak samo kurcząc skalę na osi *x* otrzymujemy dla dużych *n* rozkład Gaussa:

$$p(x) = \sqrt{1/\pi}\exp(-x^2)$$



Rozkład Gaussa

- Zwany też rozkładem normalnym, bardzo często występuje w przyrodzie.
- Jeżeli jakaś własność opisywana jest jedną liczbą, to bardzo często rozkład jej wartości dany jest krzywą Gaussa.
- Na przykład histogram wartości wzrostu ludzi w danej populacji jest dobrze przybliżany rozkładem Gaussa.
- Krzywa normalna opisuje też wagę, rozmiar buta, iloraz inteligencji i wiele innych rozkładów własności organizmów żywych.

Generatory liczb losowych

- Komputery generują liczby pseudolosowe, nie losowe.
- Generacja polega na wielokrotnym stosowaniu funkcji mieszającej do liczby początkowej (zarodka).
- Wybieramy zarodek, x_0. Wyliczamy

  $x_1=f(x_0)$,

  $x_2=f(x_1)$,

  i tak dalej.

Generatory liczb losowych

- Wygenerowane liczby powtarzają się po pewnym czasie. Im dłuższy okres, tym lepszy generator. Im lepiej „potasowane” liczby, tym lepszy generator.

- Popularne generatory liczb losowych w kompilatorach mają często krótki okres, np. 2^{24}.

Generatory liczb losowych

- Popularne generatory liczb losowych w kompilatorach mają często krótki okres, np. 2^{24}.
- Taki generator przy rzutach monetą nigdy nie wygeneruje ciągu identycznych cyfr (na przykład jedynek) dłuższego niż 25.
- W symulacjach i w zastosowaniach potrzebujemy zwykle znacznie lepszych generatorów liczb losowych, takich jak np. Mersenne twister – okres 2^{19937}-1

Inne źródła liczb losowych

- Tabele liczb losowych,
  np. rozwinięcia liczb normalnych
- Pomiary fizyczne odpowiednich układów
- Funkcje mieszające

Do czego potrzebujemy losowych ciągów?

- Do każdego praktycznego zastosowania teorii prawdopodobieństwa:
  - Wybór próbki statystycznej
  - Gry matematyczne (np. negocjacje)
  - Obliczenia metodą Monte Carlo
  - bezpieczny Internet: zakupy, komunikacja

Liczby normalne ponownie

- Liczba normalna – liczba, w której rozwinięciu w danym układzie każdy blok cyfr jest tak samo prawdopodobny jak każdy inny blok tej samej długości
- Rozwijając liczbę normalną w bazie o podstawie 36 możemy generować losowe teksty: 10 cyfr + 26 liter łacińskich, albo 35 polskich liter i spacja.
- Przykład rozwinięcia liczby pi po polsku.

Prawdopodobieństwo wystąpienia dowolnego tekstu

- Trzydziestotomowa encyklopedia „Britannica” zawiera około 30 milionów znaków. Prawdopodobieństwo wystąpienia jej tekstu w przypadkowym tekście wynosi

$$\left(\frac{1}{36}\right)^{30\,000\,000} \approx \left(\frac{1}{10}\right)^{45\,000\,000}$$

- Prawdopodobieństwo wystąpienia krótkich słów, jak „Budda” wynosi

$$\left(\frac{1}{36}\right)^{5} \approx \left(\frac{1}{60\,000\,000}\right)$$

czyli w tekście długości 60 mln znaków takie słowa powinny wystąpić przynajmniej raz

Kodowanie tekstów w rozwinięciach

- Znajdowanie tekstów w rozwinięciach liczb normalnych nie tylko zależy od podstawy ale i od kodowania, tj. od tego co przypiszemy danemu symbolowi w rozwinięciu