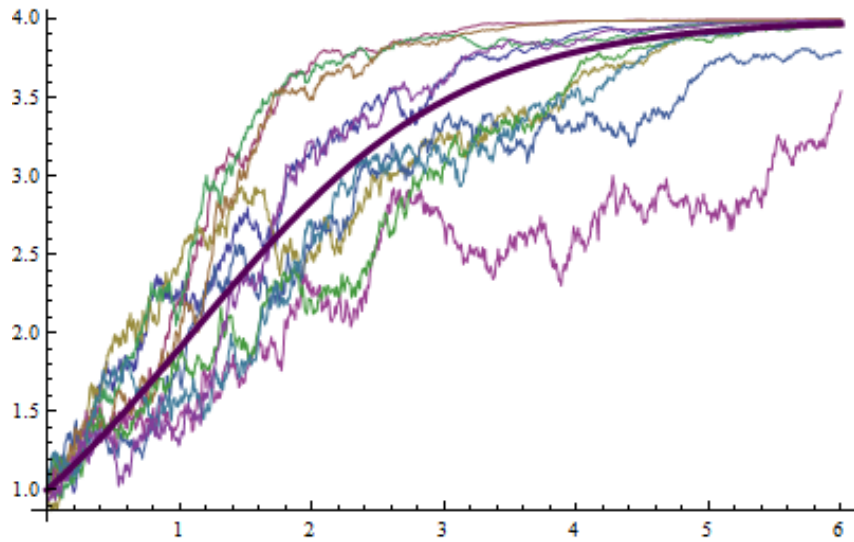
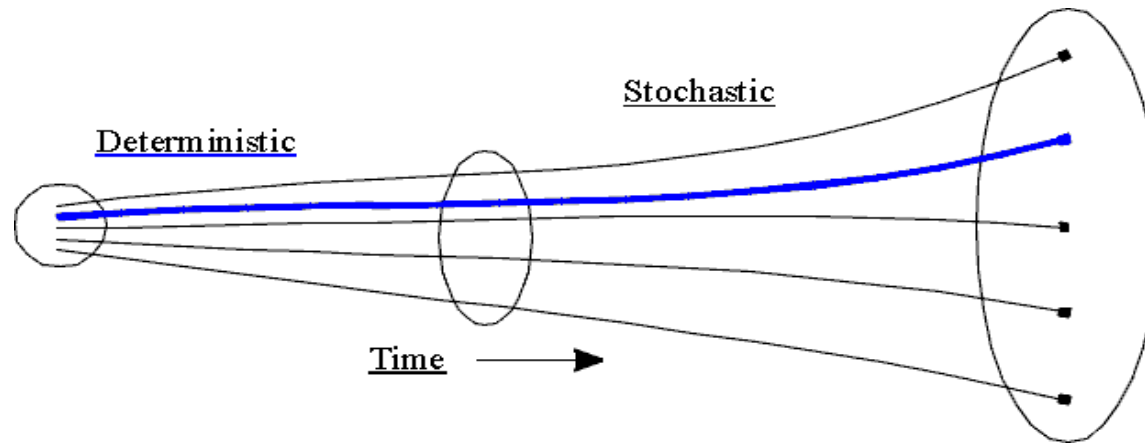




MODELOWANIE STOCHASTYCZNE CZĘŚĆ I

Biomatematyka
Dr Wioleta Drobik-
Czwaro

MODELE STOCHASTYCZNE, A DETERMINISTYCZNE



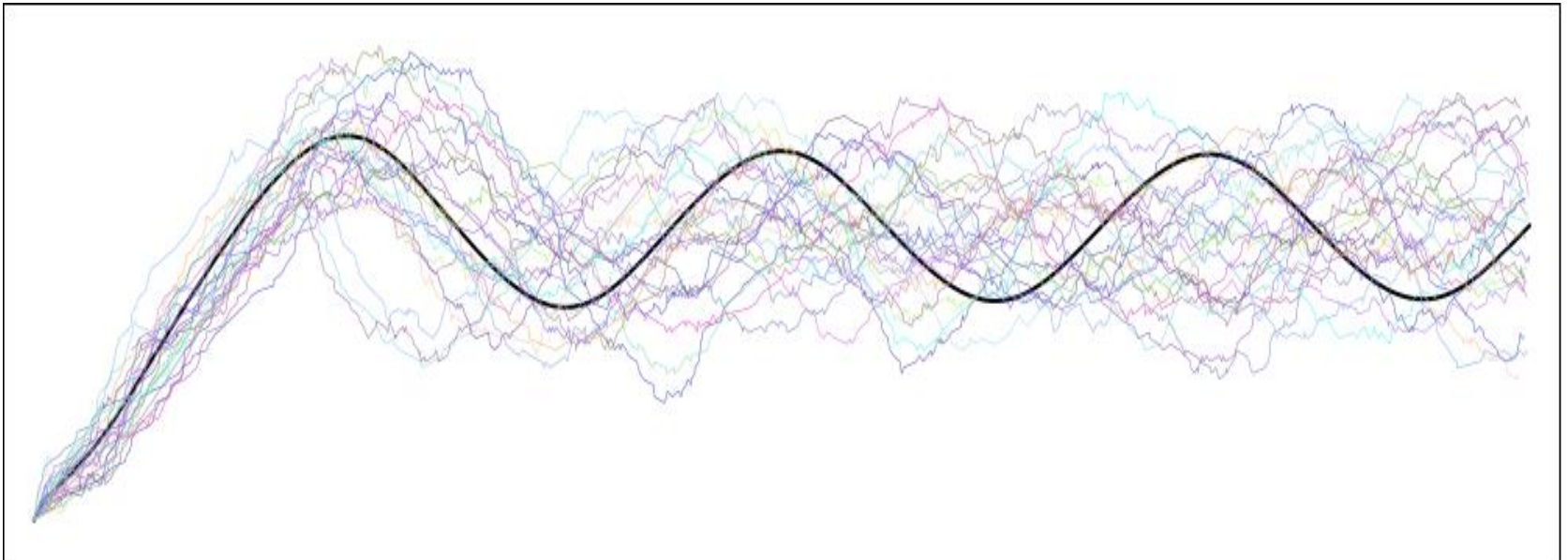
Brak możliwości uwzględnienia czynników losowych w modelu deterministycznym

Znany jest jedynie rozkład prawdopodobieństwa możliwych stanów końcowych

Każdemu wynikowi towarzyszy odpowiednia szansa jego realizacji – poszukujemy najbardziej prawdopodobnego rezultatu

SYMULACJA STOCHASTYCZNA

Symulacja stochastyczna to **eksperyment** przeprowadzany w komputerze, polegający na wielokrotnej symulacji działania analizowanego procesu oraz rejestracji szukanych charakterystyk wynikowych



PRAWDOPODOBIEŃSTWO

Rachunek prawdopodobieństwa – dziedzina matematyki zajmująca się badaniem zjawisk losowych (nie możemy ich przewidzieć z całkowitą pewnością).

Prawdopodobieństwo zdarzenia A

$$P(A) = \frac{|A|}{|\Omega|}$$

Moc zbioru A (A – Liczba zdarzeń sprzyjających)
Moc zbioru Ω (Ω – Liczba wszystkich możliwych zdarzeń)

Zbiór zdarzeń elementarnych (Ω) – zbiór zawierający wszystkie możliwe zdarzenia (sytuacje wykluczające się nawzajem)

- Zbiór zdarzeń elementarnych dla dwukrotnego rzutu monetą:

$$\Omega_2 = \{(O, R), (O, O), (R, R), (R, O)\}$$

WŁASNOŚCI PRAWDOPODOBIEŃSTWA - PRZYPOMNIENIE

- $0 \leq P(A) \leq 1$ – zawsze w granicach od 0 do 1
- $P(\Omega) = 1$ – zdarzenia pewnego jest równe 1
- $P(\emptyset) = 0$ – zdarzenia niemożliwego jest równe 0
- $P(A') = 1 - P(A)$ – prawdopodobieństwa zdarzenia przeciwnego $P(A')$
- **Iloczynem zdarzeń** ($A \cap B$) nazywamy zbiór wszystkich zdarzeń elementarnych przestrzeni Ω , które sprzyjają zdarzeniu A i zdarzeniu B
- **Sumą zdarzeń** ($A \cup B$) nazywamy zbiór wszystkich zdarzeń elementarnych przestrzeni Ω , które sprzyjają co najmniej jednemu ze zdarzeń A, B

ZDARZENIA LOSOWE

Zdarzenia losowe rozłączne (albo **wykluczające się**) to para zdarzeń losowych A,B których część wspólna jest zdarzeniem niemożliwym

Zdarzenia losowe niezależne to zdarzenia należące do tej samej przestrzeni zdarzeń, dla których spełniony jest warunek:

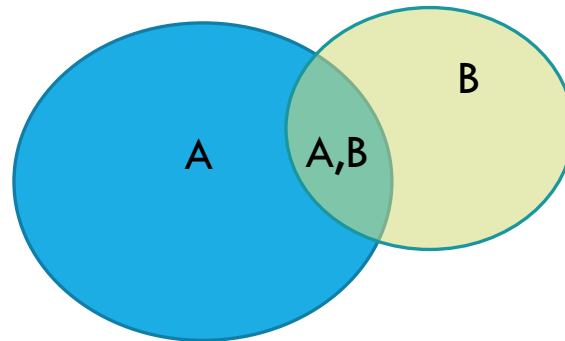
$$P(A \cap B) = P(A) * P(B)$$

- Zależność taką obserwujemy, gdy prawdopodobieństwo zdarzenia A jest takie samo bez względu na to czy zaszło zdarzenie B czy nie i odwrotnie.

PRAWDOPODOBIEŃSTWO WARUNKOWE

Prawdopodobieństwo warunkowe –
prawdopodobieństwo zajścia zdarzenia A pod
warunkiem, że zajdzie zdarzenie B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



gdzie $P(B) > 0$

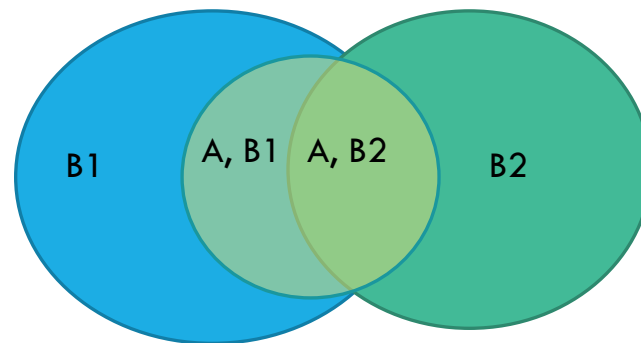
PRAWDOPODOBIEŃSTWO CAŁKOWITE

Prawdopodobieństwo zajścia zdarzenia A, posiadając informację o prawdopodobieństwie tego, że:

- A zaszło pod warunkiem zajścia B₁
- A zaszło pod warunkiem zajścia zdarzenia przeciwnego do B₁ (B₂)

$$P(A) = P(A \cap B_1) \cup (A \cap B_2) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$$

- B₁ i B₂ wykluczają się



PRAWDOPODOBIEŃSTWO CAŁKOWITE

Jeżeli zdarzenia $B_1, B_2, \dots, B_n$ są parami rozłączne oraz mają prawdopodobieństwa dodatnie, które sumują się do jedynki, to dla dowolnego zdarzenia A zachodzi wzór:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n)$$

TWIERDZENIE BAYESA

Twierdzenie – Jeżeli zdarzenia $B_1, B_2, \dots, B_n$ wykluczają się parami i mają prawdopodobieństwa dodatnie, to dla każdego zdarzenia A zawartego w sumie zdarzeń

$$B_1 \cup B_2 \dots \cup B_n$$

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n)}$$

Wzór uproszczony dla pojedynczego zdarzenia B :

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

TWIERDZENIE BAYESA

Jak wyprowadzono twierdzenie Bayesa?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A|B)P(B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(B|A)P(A)$$



$$P(B|A)P(A) = P(A|B)P(B)$$



$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

WNIOSKOWANIE BAYESOWSKIE A KLASYCZNE

Sprawdzamy czy moneta jest symetryczna:

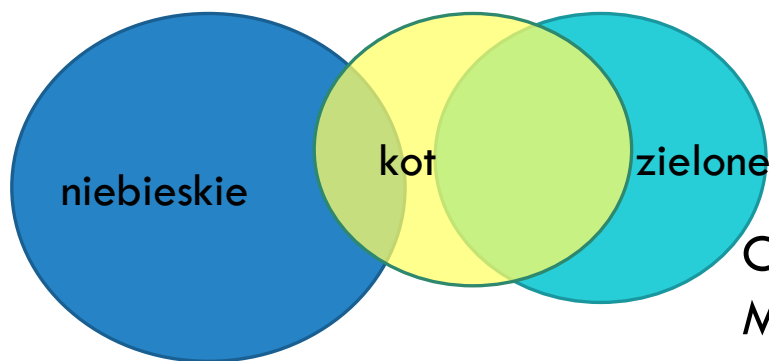
- **Podejście klasyczne:** powtarzamy wielokrotnie eksperyment aby obliczyć częstość empiryczną zdarzenia, co daje nam oszacowanie prawdopodobieństwa
 - wielokrotnie rzucamy monetą, oszacowane prawdopodobieństwo wyrzucenia reszki wynosi 50%
- **Podejście bayesowskie:** Szacujemy w jakim stopniu ufamy, że moneta jest symetryczna

WNIOSKOWANIE BAYESOWSKIE

WZÓR UPROSZCZONY

Przykład:

- $P(\text{zielone}) = P(Z) = 0.16$ – odsetek studentów o zielonych oczach
- $P(\text{kot}) = P(K) = 0.1269$ – odsetek studentów posiadających kota
- $P(\text{kot} | \text{zielone}) = P(K | Z) = 0.5$ – odsetek studentów o zielonych oczach posiadających kota
- Jakie jest prawdopodobieństwo, że student ma zielone oczy jeżeli posiada kota?



$$P(Z|K) = \frac{P(K|Z)P(Z)}{P(K)} = \frac{0,5 * 0,16}{0,1269} = 0,63$$

Co zyskałismy?

Mając dodatkowe informacje (posiadanie kota) możemy z większym prawdopodobieństwem (0,16 vs 0,63) stwierdzić czy losowo wybrany student ma zielone oczy

WNIOSKOWANIE BAYESOWSKIE

$$P(H|D) = \frac{P(D|H) P(H)}{P(D|H) P(H) + P(D|fH) P(fH)}$$

$P(H|D)$ – prawdopodobieństwo warunkowe - hipoteza jest prawdziwa (H), jeżeli tak mówią dane (D)

$P(D|H)$ – prawdopodobieństwo warunkowe – dane (D) będą potwierdzały hipotezę (H), jeżeli będzie ona prawdziwa

$P(H)$ – prawdopodobieństwo, że hipoteza jest prawdziwa

$P(fH)$ – prawdopodobieństwo, że hipoteza jest fałszywa

$P(D|fH)$ – prawdopodobieństwo, że dowody potwierdzają hipotezę, ale jest ona fałszywa (fD)

CZUŁOŚĆ I SWOISTOŚĆ

Tabela klasyfikacji^a

Obserwowane			Przewidywane				
			surv		Procent poprawnych klasyfikacji		
			nie przeżył	przeżył			
Krok 1	surv	nie przeżył	TN	42	FP	24	63,6
		przeżył	FN	10	TP	44	81,5
Procent ogółem							71,7

a. Punktem podziału jest ,500

Trafnie przewidziany sukces i porażka

Swoistość
(Specyficzność)

Czułość

Dokładność
(ACC)

Procent przewidzianych wartości ogółem
(sukcesów i porażek)

$$\text{Czułość} = \frac{TP}{TP + FN} = \frac{44}{44 + 10} = 0.815$$

$$\text{Specyficzność} = \frac{TN}{TN + FP} = \frac{42}{42 + 24} = 0.636$$

$$\text{Dokładność} = \frac{TN + TP}{TN + FP + TP + FN} = \frac{42 + 44}{42 + 24 + 44 + 10} = 0.717$$

WNIOSKOWANIE BAYESOWSKIE W DIAGNOSTYCE MEDYCZNEJ

Test diagnostyczny ma czułość 99% - wykrywa 99% przypadków choroby – Prawdopodobieństwo, że test wypadnie pozytywnie jeżeli osoba jest chora.

Ten sam test ma swoistość 99% - wypada negatywnie w 99% przypadków jeżeli dana osoba jest zdrowa

Wiemy, że tylko 5 % badanych choruje

Jakie jest prawdopodobieństwo, że osoba jest chora jeżeli test wypadł pozytywnie?

WNIOSKOWANIE BAYESOWSKIE W DIAGNOSTYCE MEDYCZNEJ

$$P(C|+) = \frac{P(+|C)P(C)}{P(+)} = \frac{P(+|C)P(C)}{P(+|C)P(C) + P(+|Z)P(Z)}$$
$$= \frac{0.99 * 0.05}{0.99 * 0.05 + 0.01 * 0.95} = \frac{0.0495}{0.0495 + 0.0095} = 0.839$$

C chory

Z zdrowy

+ wynik pozytywny testu

– wynik negatywny testu

$P(C) = 0,05$ – odsetek chorych wśród badanych

$P(Z) = 1 - 0,05 = 0.95$ – odsetek zdrowych wśród badanych

$P(+|C) = 0,99$ – skuteczność testu przy badaniu osoby chorej, odsetek testów prawdziwie pozytywnych (czułość)

$P(-|Z) = 0,99$ – skuteczność testu przy badaniu osoby zdrowej, odsetek testów prawdziwie negatywnych (swoistość)

Możemy obliczyć $P(+|Z) = 1 - P(-|Z) = 0,01$ – jest to odsetek testów fałszywie pozytywnych

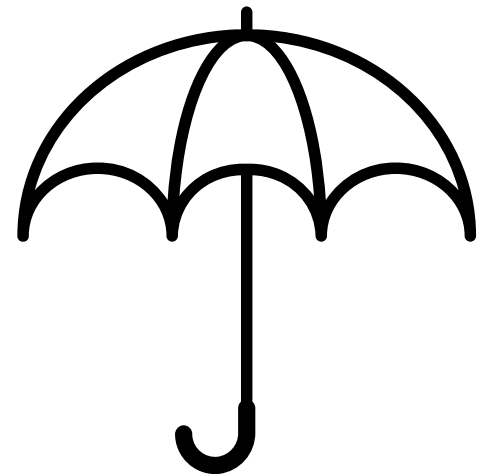
Sprawdź: <https://seeing-theory.brown.edu/bayesian-inference/index.html#section1>

WNIOSKOWANIE BAYESOWSKIE

Zaczynamy z zestawem hipotez opartych na naszej dotychczasowej wiedzy – mamy wstępne przesłanki aby wierzyć, które są prawdziwe a które nie

Zbieramy dane i weryfikujemy nasze założenia – jeżeli dane wspierają hipotezę nasza „wiara” w nią rośnie, jeżeli jest przeciwnie – słabnie.

Podajrzewam, że będzie dzisiaj padać, czy jeżeli zobaczę kogoś z parasolem na ulicy moja wiara w tą hipotezę się zwiększy?



WNIOSKOWANIE BAYESOWSKIE

Czy moneta jest symetryczna?

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$P(D)$ - Prawdopodobieństwo zbioru przesłanek (danych)

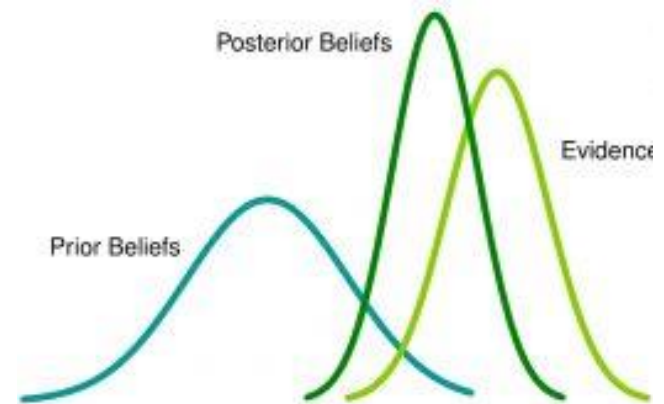
$P(D|\theta)$ - prawdopodobieństwo zaobserwowania wyników zakładając rozkład dla parametru (hipotezę)

$P(\theta)$ - prawdopodobieństwo a priori, zakładany rozkład dla parametru (hipotezy)

$P(\theta|D)$ - prawdopodobieństwo (stopień wiarygodności) a posteriori dla parametru (hipotezy)

θ - parametr wyznaczający symetryczność monety

D – Dane, wynik doświadczenia
np. 4 reszki na n rzutów



Sprawdź: <https://seeing-theory.brown.edu/bayesian-inference/index.html#section1>

WIARYGODNOŚĆ

Funkcja wiarygodności

$$L(\theta|D) = P(D|\theta)$$

Wiarygodność rozkładu pod warunkiem, że mamy określone dane

Rozkład dopasujemy do danych

Sprawdź:

<https://seeing-theory.brown.edu/bayesian-inference/index.html#section1>

Prawdopodobieństwo obserwacji danych pod warunkiem określonego rozkładu

Jak bardzo prawdopodobne są nasze dane zakładając dany rozkład?

TWIERDZENIE BAYESA I ŚLEDZTWO

Dr. Watson: “How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?”

Powiedźmy to inaczej:

How often have I said to you that when $p(D | \theta_i) = 0$ for all $i \neq j$, then, no matter how small the prior $p(\theta_j) > 0$ is, the posterior $p(\theta_j | D)$ must equal one



GENEROWANIE LICZB PSEUDOLOSOWYCH

Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.

John von Neumann 1951

Komputer nie jest w stanie generować liczb prawdziwie losowych

Generujemy ciągi liczbowe o **z góry ustalonych wartościach**, które imitują ciągi liczb losowych

Liczby takie generowane są w sposób deterministyczny

Punkt startowy to **ziarno** losowania

- Ręcznie ustawione ziarno umożliwia replikację obliczeń i symulacji

GENEROWANIE LICZB PSEUDOLOSOWYCH

Generator liczb pseudolosowych

- generuje deterministycznie „potencjalnie nieskończony” ciąg bitów na podstawie ziarna
- jeśli ziarno jest reprezentowane przez k bitów informacji, to generator może wygenerować n -bitowy ciąg jedynie na 2^k sposobów spośród 2^n możliwych

Przykładowe zastosowanie

- **metoda monte carlo** – wystarczają „słabe” generatory
- kryptografia – wymaga „silnych” generatorów

LOSOWOŚĆ A PSEUDOLOSOWOŚĆ

Student 1:

THHHTHTTTTHTTHTTTTHHTHTTHT
HHHTHTHHTHTTTHHTTTTHTTTHTH
TTHHTTTTTTTHTHHHHHTHTHTH
THTHTHHHHHTHHTTTTHTTHHTH

Student 2:

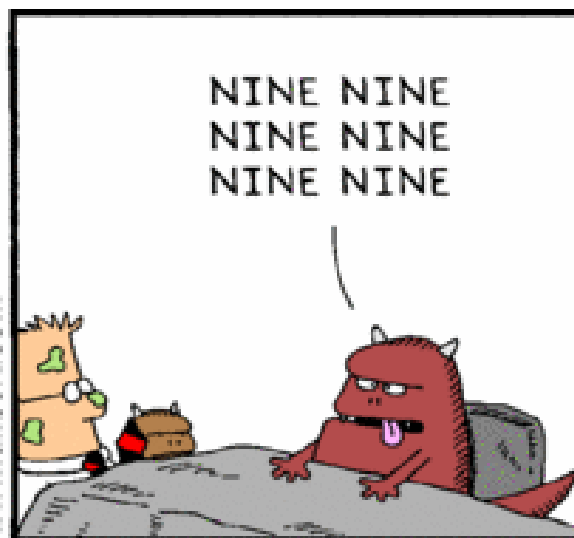
HTTHTTHTHHTTHTHTTTHHTHTT
HTTHHHTTHTTHTHTHTHHTTHTTH
THTHTHTHHHTTHTHTHTHHTHTTT
HTHHHTHTHTHTHHTTHTHTTTHHT

Nauczyciel poprosił dwóch studentów o wykonanie 100 rzutów monetą i zapisanie wyników. Jeden ze studentów podał wymyślone wartości, zamiast wykonać eksperyment. Który?

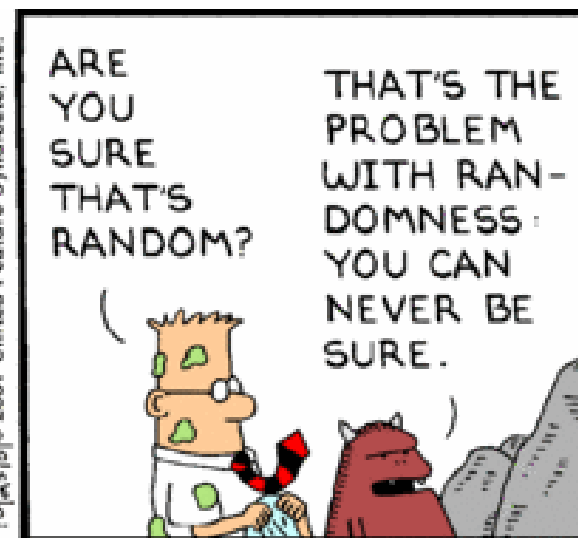
LOSOWOŚĆ A PSEUDOLOSWOŚĆ



www.dilbert.com
scottadams@aol.com



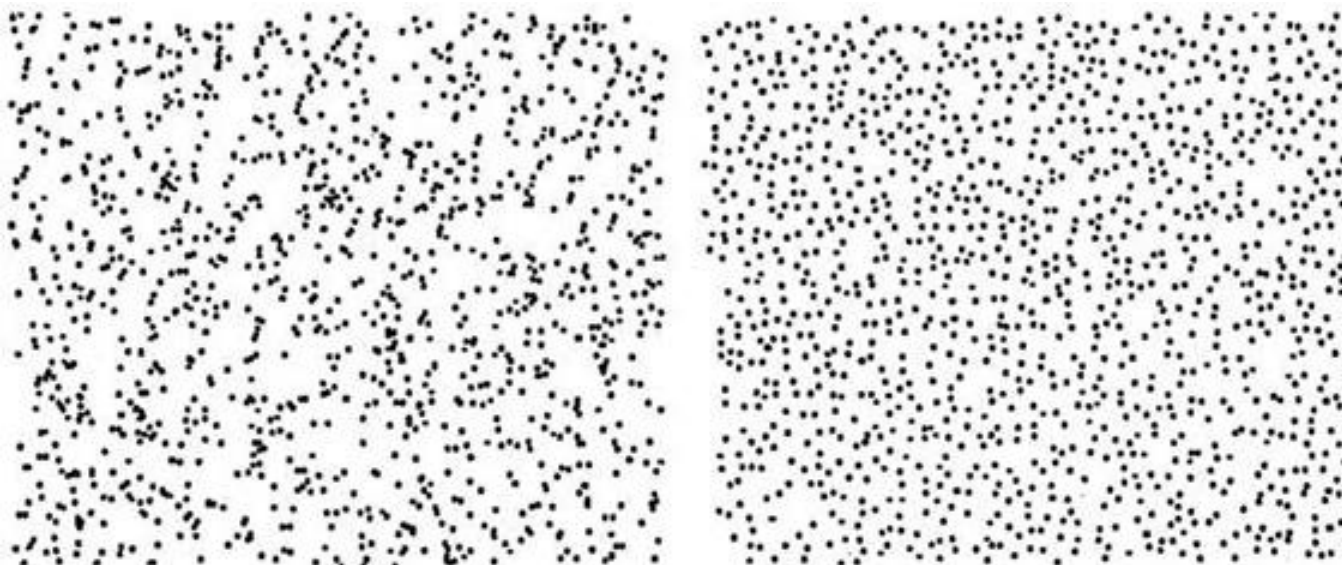
10/15/01 © 2001 United Feature Syndicate, Inc.



LOSOWOŚĆ A PSEUDOLOSOWOŚĆ

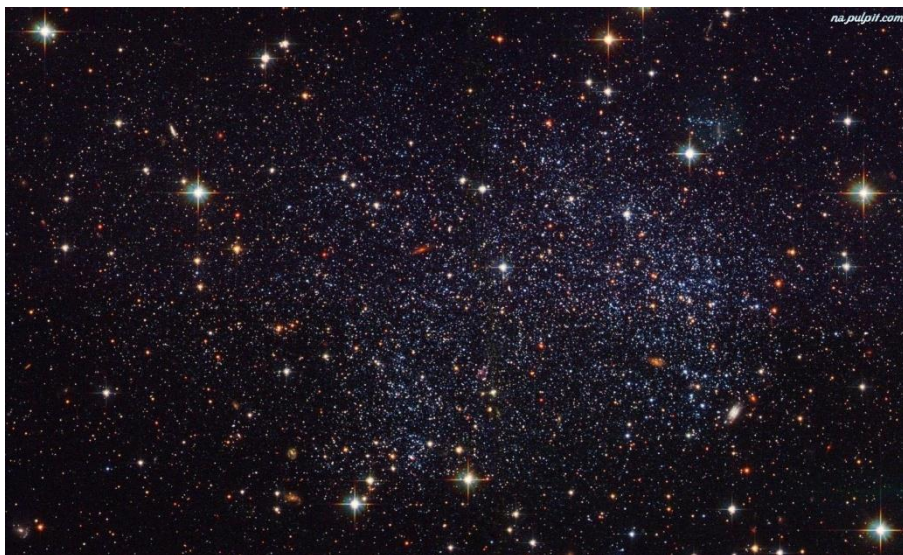
Kiedy liczby są prawdziwie losowe?

- Teoria liczb
- Teoria układów dynamicznych
- Filozofia



Który z tych wykresów przedstawia losowe ułożenie punktów?

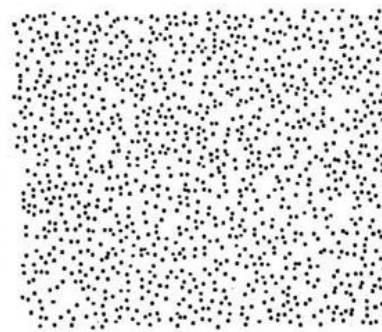
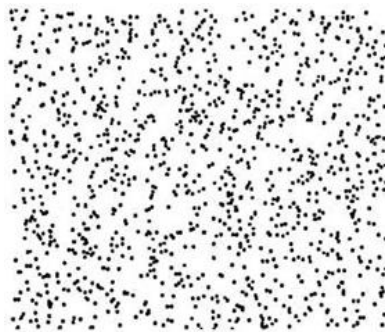
LOSOWOŚĆ A PSEUDOLOSOWOŚĆ



Symulacja układu gwiazd



Jaskinia Waitomo Glowworm i świecące
muchówki *Arachnocampa luminosa*



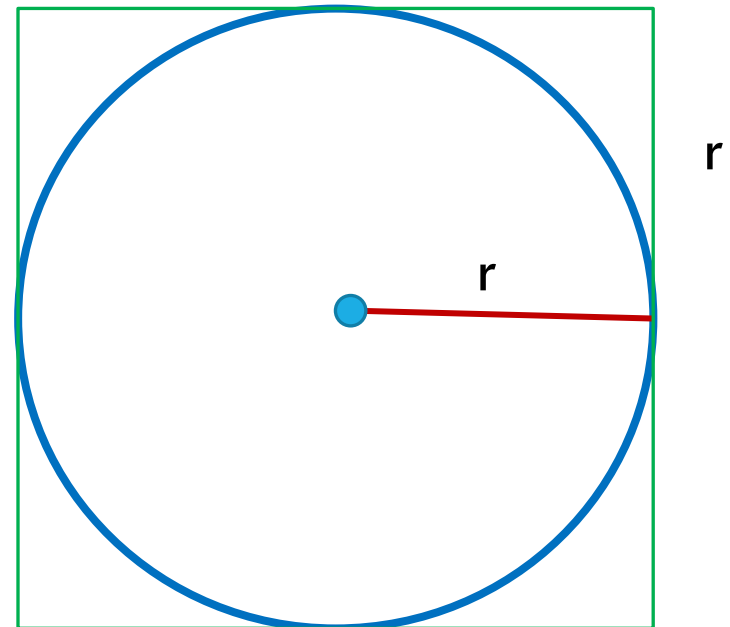
MONTE CARLO

- Za twórcę najczęściej uznaje się Stanisława Ulama
- Są to metody obliczeniowe polegające na wykorzystaniu generatorów liczb pseudolosowych
- Stosowane do modelowania procesów:
 - zbyt złożonych aby można było przewidzieć wynik za pomocą podejścia analitycznego
 - tam gdzie podejście analityczne jest możliwe, ale bardzo skomplikowane
- Metoda Monte Carlo polega na wielokrotnym losowaniu z określonego wcześniej rozkładu

MONTE CARLO - PRZYKŁADY

Obliczanie liczby π :

- Chcemy obliczyć pole koła wpisanego w kwadrat o boku równym $2r$, gdzie r to promień koła. Wzory:
- Pole koła: πr^2
- Pole kwadratu $(2r)^2$: $4r^2$
- $\pi = 4P_{\text{koło}}/P_{\text{kwadrat}}$



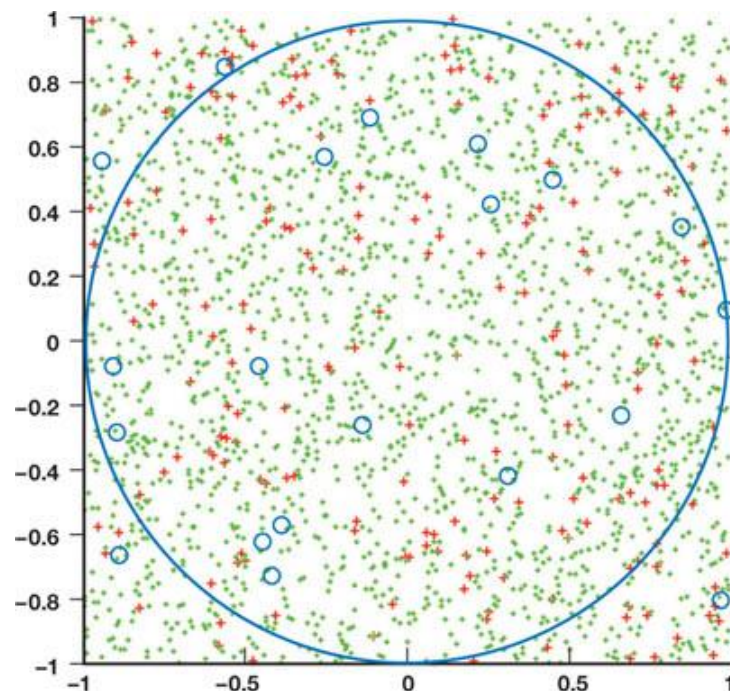
LICZBA π

Symulacja:

Wyznaczamy wewnątrz kwadratu bardzo dużo losowych punktów

Zliczamy te punkty, które znajdują się we wnętrzu koła

Sprawdź: <https://seeing-theory.brown.edu/frequentist-inference/index.html#section1>



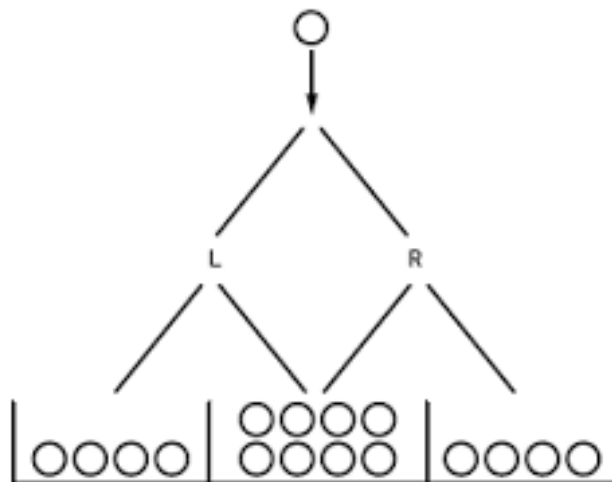
Stosunek liczby punktów zawierających się w kole do wszystkich wylosowanych punktów będzie dążył w nieskończoności do stosunku pola koła do pola kwadratu

$$\pi \approx 4n_{\text{koło}}/n_{\text{kwadrat}}$$

DESKA GALTONA

Deska Galtona to pochylona deska z wbitymi gwoździami ułożonymi w trójkąt, pod którą znajdują się przegródki do których wpadają kulki.

- Kulki tworzą histogram rozkładu dwumianowego, który zgodnie z centralnym twierdzeniem granicznym zbliża się do rozkładu normalnego przy bardzo dużej liczbie kulek i przegródek



DESKA GALTONA

Prawdopodobieństwo, że kulka wpadnie do przegródki numer k jest równe

$$0,5^k \times 0,5^{n-k}$$

n – liczba wszystkich przegródek



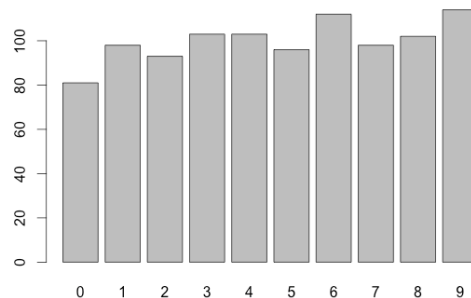
Deska Galtona ilustruje więc sposób powstawania w naturze rozkładu normalnego pod wpływem dużej liczby drobnych losowych odchyleń

Sprawdź: https://www.youtube.com/watch?v=b216a8O_IVU

GENEROWANIE LICZB LOSOWYCH ZE ZNANEGO ROZKŁADU W R

Przykład: Rozkład jednostajny (każda wartość ma identyczne prawdopodobieństwo)

- `runif(5,1,4)` – losujemy pięć liczb z rozkładu jednostajnego (min=1 i max=4)
- `punif(0.5)` – wyznaczamy wartość dystrybuanty rozkładu jednostajnego w punkcie 0.5
- `dunif(0.5)` – wyznaczamy gęstość rozkładu jednostajnego w punkcie 0.5
- `qunif(0.1)` – wyznaczamy wartość kwantyla rzędu 0.1 dla rozkładu jednostajnego



PROCESY STOCHASTYCZNE

Procesem stochastycznym nazywamy zbiór zmiennych losowych $\{X_t\}_{t \in T}$, które przyjmują wartości w przestrzeni i są indeksowane przez zbiór T

- D – pewne zdarzenie losowe – np. pogoda danego dnia
- $X_1, X_2, X_3, \dots, X_i$ – zbiór wszystkich możliwych wyników np. słonecznie, pochmurno, deszczowo
- W trakcie powtarzania doświadczenia N razy w próbie z numerem t (gdzie $t \in N$) dostajemy wynik o numerze X_t (gdzie $X_t \in \{1, 2, 3, \dots, n\}$)

Prawdopodobieństwa różnych wartości W_{x_t} mogą zależeć od:

- niczego - są niezależne (Proces Poissona, Proces Bernoulliego)
- wyników wszystkich poprzednich doświadczeń
- tylko od wyniku poprzedniego doświadczenia (Procesy Markowa)

PROCES BERNOULLEGO

Proces stochastyczny złożony z ciągu niezależnych zmiennych losowych $X_1, X_2, X_3, \dots, X_n$ takich, że:

- Dla każdego i wartość X_i to 0 (porażka) lub 1 (sukces)
- Prawdopodobieństwa osiągnięcia wartości 0 i 1 jest identyczne dla każdego i – **zmienne losowe są niezależne.**

Rzut monetą

- Dwa możliwe wyniki każdej próby: O (orzeł) lub R (reszka)
- Prawdopodobieństwo otrzymania jednego z tych dwóch wyników wynosi 0.5 i jest niezależne od poprzedniego rzutu
- Liczba reszek jakie wyrzucimy w serii rzutów ma rozkład Bernoulliego

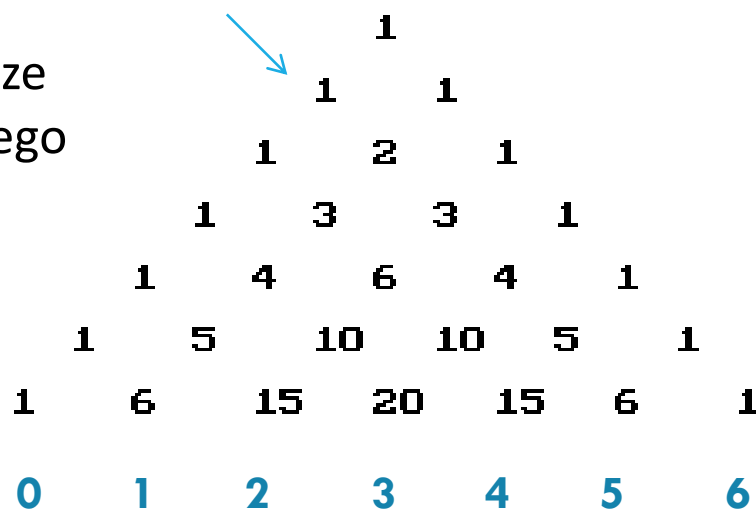
ROZKŁAD BERNOULLIEGO (DWUMIANOWY)

Prawdopodobieństwo zajścia
pojedynczego scenariusza

$$P(X = k) = \binom{n}{k} q^k (1 - q)^{n-k}$$

Liczba „scenariuszy” ze
zbioru n-elementowego

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



A Pascal's triangle for n=6, showing the number of scenarios for each k. The triangle is symmetric and the sum of the numbers in each row is 2^n. A blue arrow points from the 'n' in the formula above to the top row (n=6). Another blue arrow points from the binomial coefficient in the formula to the first column (k=0).

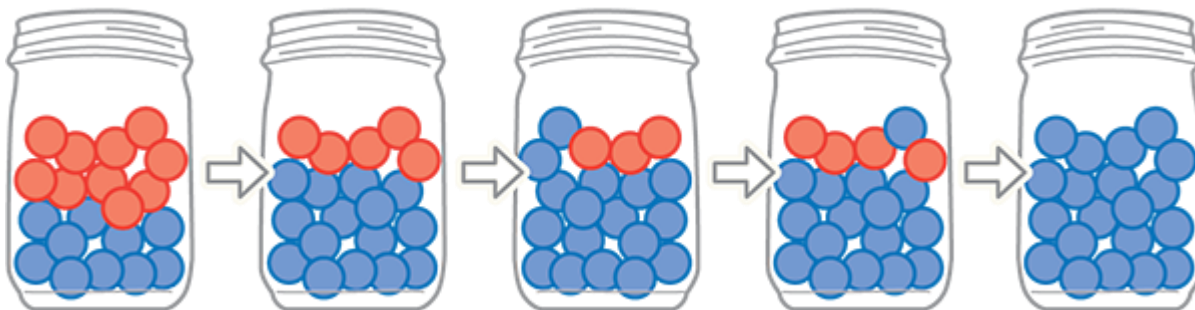
				1			
			1		1		
		1		2		1	
		1	3		3		1
	1		4	6		4	
	1	5		10	10		5
	1	6	15	20	15	6	1
k:	0	1	2	3	4	5	6

DRYF GENETYCZNY

Zmiany częstości występowania genów/alleli w populacji, które nie wynikają z doboru naturalnego, migracji czy mutacji, ale ze zdarzeń losowych



losowe pobieranie prób o ograniczonej liczebności



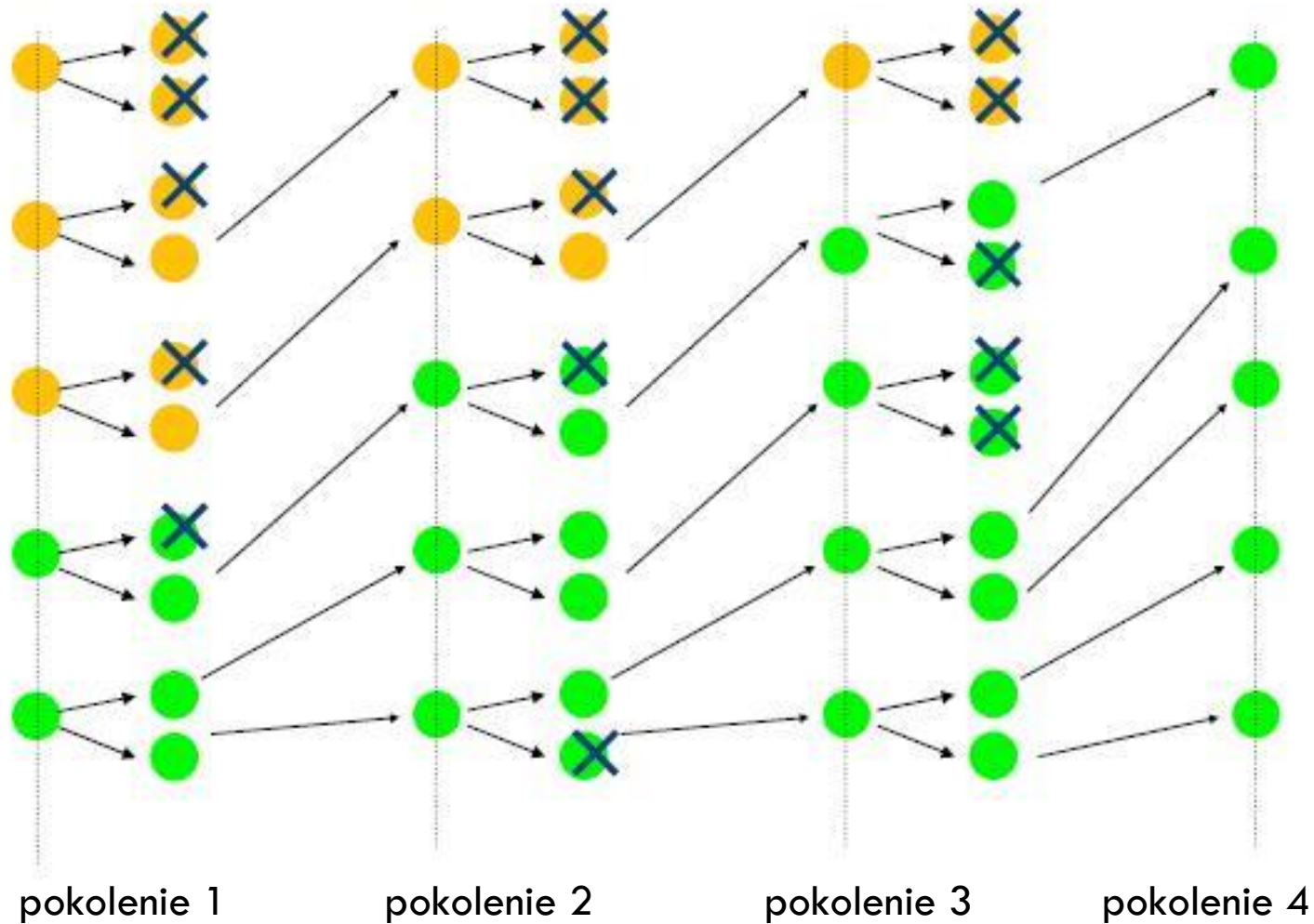
MUTACJE NEUTRALNE

Rozprzestrzenianie mutacji za pomocą dryfu genetycznego

Model Wrighta-Fishera:

- Istnieje **m kopii sekwencji z mutacją** w jednym pokoleniu, chcemy ustalić liczbę kopii sekwencji z mutacją w następnym pokoleniu
- Każda kopia genu w nowej populacji jest dziedziczona z jednej przypadkowo wybranej kopii genu z poprzedniego pokolenia
- Dowolna kopia genu w pokoleniu t zawiera mutację z prawdopodobieństwem równym $\alpha = m/N$ i jest wolna od mutacji z prawdopodobieństwem $1 - \alpha$
- Prawdopodobieństwo, że będzie n kopii zmutowanego genu w kolejnym pokoleniu opisuje **rozkład dwumianowy**

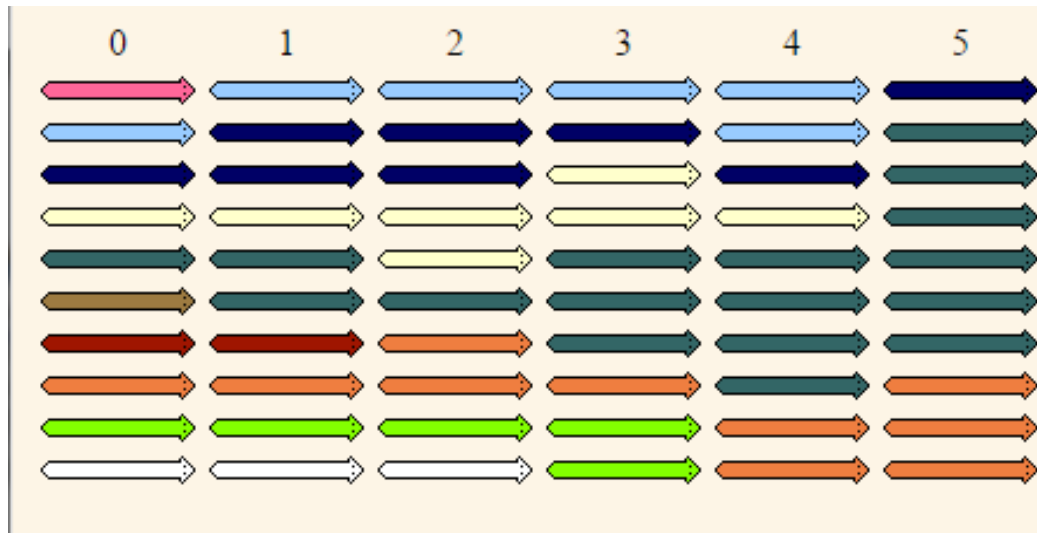
DRYF GENETYCZNY



DRYF GENETYCZNY

Ciekawe symulacje:

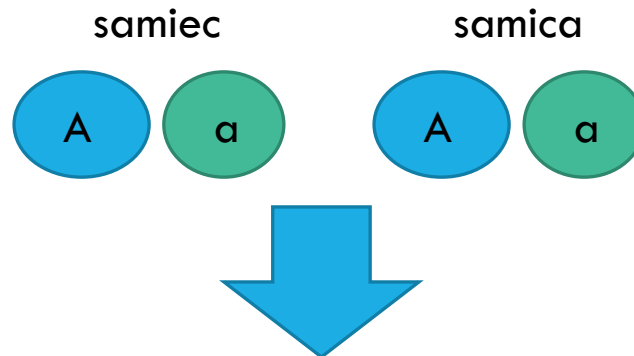
<http://www.biology.arizona.edu/evolution/act/drift/drift.html>



DRYF GENETYCZNY

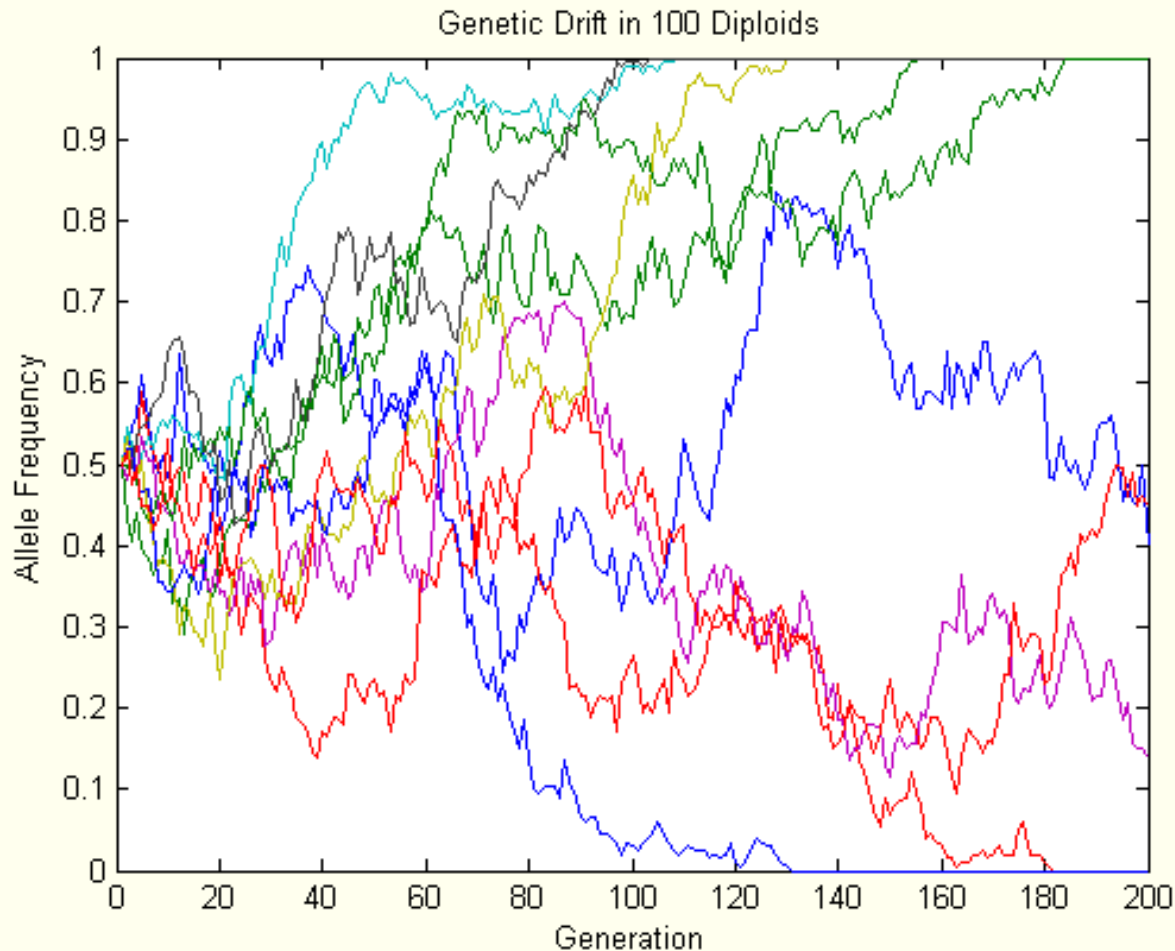
W programie populus: <http://cbs.umn.edu/populus/>

- Wykorzystuje metodę Monte Carlo: losowanie alleli z małej populacji rodzicielskiej i przekazanie ich potomstwu
- Dryf może być zasymulowany max dla 10 dwuallelicznych loci
- Przykład: populacja złożona z dwóch osobników, samca i samicy



Dwa osobniki potomne, Jakie będą ich genotypy?
Z jakim prawdopodobieństwem?

DRYF GENETYCZNY – SYMULACJA

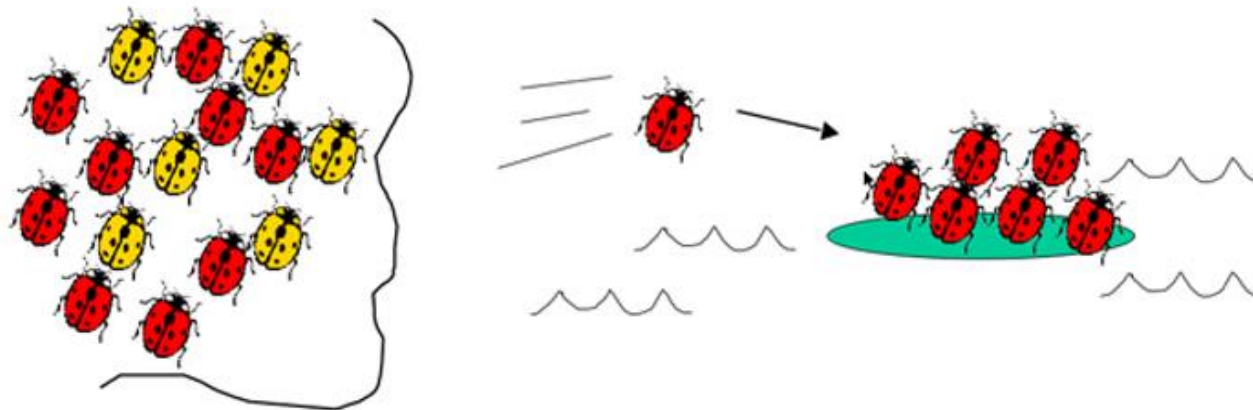


EFEKT ZAŁOŻYCIELA

Odtłączenie się od populacji wyjściowej niewielkiej grupy osobników (o innej frekwencji alleli niż populacja wyjściowa) która migruje lub zostaje przeniesiona na nowy, odizolowany obszar.

- Np. z licznej populacji żyjącej na rozległym terenie kontynentu część osobników przedostaje się na odizolowaną wyspę

Po wielu pokoleniach możemy mieć do czynienia z nową jednostką systematyczną np. podgatunkiem



LITERATURA

Foryś U. 2011. Modelowanie matematyczne w biologii i medycynie.

Higgs P., Atwood T. 2011. Bioinformatyka i ewolucja molekularna. PWN.

Tiurn J. 2006. Wstęp do obliczeniowej biologii molekularnej. Wykłady

<http://www.cs.hunter.cuny.edu/~saad/courses/compbio/lectures/lecture9.pdf>

Polecane: <https://www.youtube.com/channel/UCoxbmeSkLm-NymFjNgeFbCw>