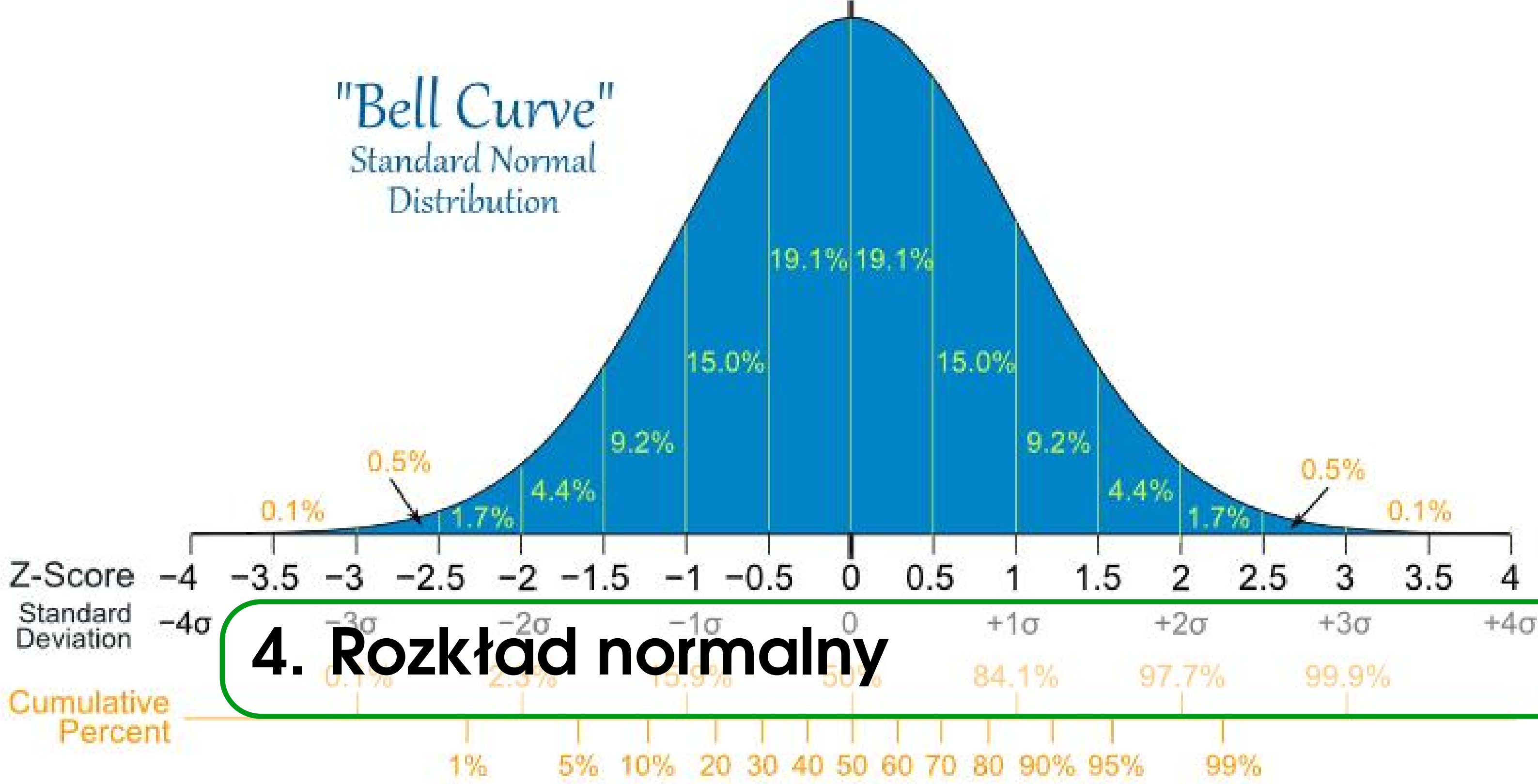



# 4. Rozkład normalny

## 4.1 Statystyka matematyczna

Na poprzednich laboratoriach poznaliśmy statystykę opisową, licząc różne statystyki dla próbki (lub dla populacji jeżeli mieliśmy o niej *wszystkie* dane). Jakiekolwiek wnioski do których dochodziliśmy dotyczyły wyłącznie badanego zbioru danych bez ich generalizowania na wszystkie możliwe dane z danego procesu czy badania. Aby móc wyciągać takie wnioski musimy zacząć jakoś modelować dane których nie widzimy. Naturalnym aparatem do takiego modelowania jest rachunek prawdopodobieństwa, dzięki któremu możemy zamodelować nasze dane np. możemy powiedzieć że dane pochodzą z jakiegoś rozkładu i przy pomocy estymacji parametrów tego rozkładu możemy zdobyć informację o całej populacji. Tym właśnie zajmuje się statystyka matematyczna.

**Definicja 4.1 — Statystyka matematyczna.** Statystyka matematyczna zajmuje się opisywaniem i analizą zjawisk masowych przy użyciu metod rachunku prawdopodobieństwa [3]

### 4.1.1 Rozkład Bernoulliego – powtórka pojęć o zmiennych losowych

Zmienna losowa to funkcja o wartościach, która jest określona na zbiorze zdarzeń elementarnych. Intuicyjnie możemy pojmować zmienną losową jako wartość liczbowa zależna od przypadku. Dla przykładu: wysokość losowo wybranego człowieka jest zmienną losową.

Rozkład prawdopodobieństwa zmiennej losowej jest wyrażony poprzez funkcję prawdopodobieństwa $P(X = x)$ dla zmiennej dyskretnej lub przez funkcję gęstości prawdopodobieństwa (ang. *PDF, probability density function*) $f(x)$ dla zmiennej ciągłej. Dodatkowo, ze zmiennymi losowymi skojarzone jest również pojęcie dystrybuanty (ang. *CDF, cumulative distribution function*) $F(x)$ czyli skumulowanego prawdopodobieństwa od strony $-\infty$. Zauważ, że $F(x) = P(X \leq x)$.

Zmienna losowa, która może przyjmować jedynie wartości binarne (0 i 1): 1 z pewnym prawdopodobieństwem $p$, a 0 z prawdopodobieństwem $= 1 - p$ nazywamy zmienną o rozkładzie Bernouliego lub o rozkładzie dwupunktowym[1]. Prawdopodobieństwo $p$ często nazywamy prawdopodobieństwem sukcesu, a cały rozkład często zapisujemy jako $Bernoulli(p)$.

Studentów informatyki chyba nie trzeba przekonywać o tym ile rzeczy można zamodelować zmiennymi binarnymi, więc nie dziwi fakt, że jest to jeden z bardzo znanych rozkładów prawdopodobieństwa. Ponadto, ze względu na jego prostotę, posłuży nam do przypomnienia podstawowych pojęć o zmiennych losowych.

Wartość oczekiwana (wartość średnia) zmiennej losowej to wartość określająca spodziewany wynik doświadczenia losowego.

■ **Przykład 4.1** Policzmy wartość oczekiwaną zmiennej $X \sim Bernoulli(p)$.

$$\mathbb{E}[X] = \sum_x x \cdot P(X = x) = 1 \cdot p + 0 \cdot (1 - p) = p$$

■

Zauważmy, że intuicyjnie wartość oczekiwana jest tym samym co średnia arytmetyczna. Wyobraźmy sobie ciąg realizacji zmiennej losowej $X \sim Bernoulli(p)$, może on wyglądać np. tak $0, 0, 1, 0, 1, 0, 0, 1, 1, 0$. Średnia arytmetyczna takiego ciągu będzie wynosiła „liczba jedynek przez liczbę wszystkich elementów" czyli de facto częstotliwość występowania jedynek. Częstotliwość ta (przy dużej próbie) może być interpretowana jako prawdopodobieństwo otrzymania jedynki czyli $p$.

Własności wartości oczekiwanej:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Jeżeli zmienne są niezależne to $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Jeżeli zmienne są niezależne to $\mathbb{E}[g(X) \cdot h(Y)] = \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)]$

Jeżeli $Y = g(X)$ to $\mathbb{E}[Y] = \mathbb{E}[g(X)]$

Tą ostatnią własność nazywamy czasami „regułą leniwego statystyka" (ang. *The Rule of the Lazy Statistician*), gdyż pozwala nam na obliczenie wartości oczekiwanej funkcji zmiennej losowej bez konieczności wyznaczenia rozkładu prawdopodobieństwa jej wartości.

Wariancja zmiennej losowej to miara rozrzutu jej wartości i jest intuicyjnym odpowiednikiem poznanej na zeszłych zajęciach wariancji w populacji.

**Problem 4.1** Policz wariancję zmiennej $X \sim Bernoulli(p)$.

Własności wariancji:

$$\mathbb{D}^2[aX + b] = a^2\,\mathbb{D}^2[X]$$

Jeżeli zmienne są niezależne to $\mathbb{D}^2[X + Y] = \mathbb{D}^2[X] + \mathbb{D}^2[Y]$

[1] Dokładniej: rozkład Bernouliego jest zdefiniowany dla zmiennej zero-jedynkowej, a rozkład dwupunktowy dla zmiennej przyjmującej dwie różne, dowolne wartości

## 4.2 Rozkład normalny

Bardzo ważnym rozkładem prawdopodobieństwa jest tzw. rozkład normalny, który opisuje wiele wartości występujących w naturze jak np. wysokość człowieka, grubość pni w lesie, liczba samochodów na autostradzie w ciągu godziny, szybkość samochodu na autostradzie, czas lotu pomiędzy danymi miastami, porowatość gleby z danej próbki, szybkość turbiny wodnej, wyniki badań na spostrzegawczość dzieci przedszkolnych i wiele, wiele innych... Rozkład normalny pojawia się także w analizach danych z portali społecznościowych np. logarytm czasu pomiędzy napisaniem maila czy komentarza, dodaniem nowego znajomego czy nawet zdobyciem nowego followers'a ma rozkład normalny[2].

**Definicja 4.2 — Rozkład normalny.** Ciągła zmienna losowa X ma rozkład normalny (rozkład Gaussa, krzywa de Moivre'a) z parametrami $\mu \in \mathbb{R}$ i $\sigma > 0$, oznaczany przez[a] $X \sim N(\mu, \sigma)$ jeżeli jego funkcja gęstości jest zdefiniowana przez[b]

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

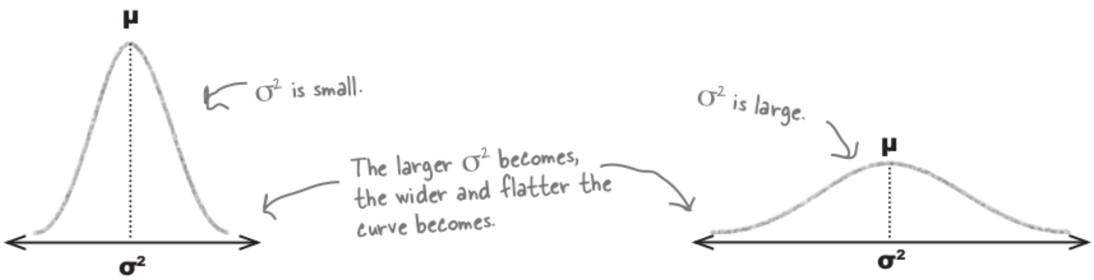
Dodatkowo można pokazać, że $\mathbb{E}[X] = \mu$ i $\mathbb{D}^2[X] = \sigma^2$.

[a] W niektórych książkach stosuje się oznaczenie $X \sim N(\mu, \sigma^2)$

[b] Komentarz dla intuicji: to co stoi przed $e$ pełni tylko funkcję normalizacyjną - pole pod krzywą musi wynosić 1. Cały kształt jest zawarty w $exp(-\frac{(x-\mu)^2}{2\sigma^2})$

Najważniejsze własności rozkładu normalnego:

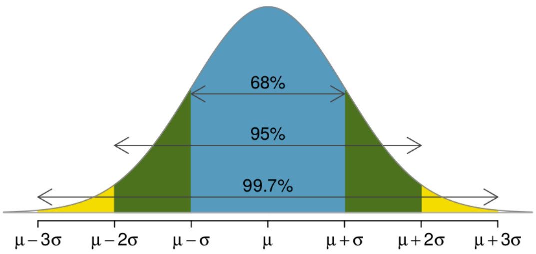
- Rozkład normalny jest unimodalny oraz symetryczny. Największa gęstość prawdopodobieństwa jest dookoła wartości średniej, im dalej od niej tym gęstość jest niższa...
- Parametry rozkładu normalnego mają bardzo prostą interpretację: $\mu$ jest modą krzywej gęstości oraz medianą i średnią arytmetyczną zmiennej. Z kolei $\sigma$ jest odchyleniem standardowym tej zmiennej, którego zwiększenie powoduje jedynie spłaszczenie centralnego wybrzuszenia wykresu oraz pogrubienie jego ogonów.




- Reguła 3 sigma lub reguła 68-95-99.7, która mówi o tym, że 68,26% wszystkich danych leży w odległości jednego odchylenia standardowego od średniej, 95,44% w odległości dwóch odchyleń i 99.74% w odległości trzech odchyleń[3]. Podsumowując: prawie wszystkie obserwacje leżą w obszarze $\mu \pm 3\sigma$.
  **Problem 4.2** Na poprzednich zajęciach poznaliśmy nierówność Czebyszewa. Ile obserwacji wg. tej nierówności leży dalej od średniej niż $3\sigma$? Skąd się bierze różnica

[2] badania m.in. na danych z portalu Reddit.com [4]

[3] Warto zauważyć, że 99% wartości jest już w obszarze $x \pm 2,58\sigma$




Rysunek 4.1: Zasada 68-95-99.7 dla rozkładu normalnego [2]

z regułą 68-95-99.7?

> ! Zasadę 68-95-99.7 można użyć do przybliżenia odchylenia standardowego z prawdziwych danych, jeżeli nie mielibyśmy możliwości jego policzenia. Jeżeli wiemy, że dane mają w przybliżeniu rozkład normalny oraz znamy ich minimum, średnią oraz maksimum to możemy tak dopasować wariancję, żeby $\bar{x} - 3S \approx x_{min}$ i $\bar{x} + 3S \approx x_{max}$. Może to też służyć do zgrubnego sprawdzenia czy rozkład w danych nie ma zbyt wiele obserwacji odstających, aby uważać go za rozkład normalny.

- Ustandaryzowana zmienna losowa transformacją Z ($Z = \frac{X-\mu}{\sigma}$) ma rozkład normalny o $\mu = 0$ i $\sigma = 1$, który można znaleźć w tablicach statystycznych. Jest to o tyle ważna własność, że wynika z niej że każdy rozkład normalny ma bardzo zbliżoną charakterystykę (jeśli go odpowiednio spłaszczyć i zwęzić – nie wszystkie rozkłady mają taką własność).

> ! Znormalizowanie *dowolnej* zmiennej losowej transformacją Z nie spowoduje, że będzie ona miała rozkład normalny. Natomiast, owszem, zmienna ta będzie miała wartość oczekiwaną 0 i odchylenie standardowe 1.

Z tego powodu rozkład $N(0,1)$ doczekał się specjalnej nazwy: „rozkładu standardowego” lub „standardowego rozkładu normalnego”. Jego funkcję gęstości zwyczajowo oznaczamy przez $\phi(z)$, a jego dystrybuantę przez $\Phi(z)$. Dodatkowo, pamiętając że funkcja gęstości jest symetryczna, dystrybuanta rozkładu normalnego ma następującą własność:

$$\Phi(x) = 1 - \Phi(-x)$$

Warto też pamiętać, że zmienna $Z \sim N(0,1)$ przyjmuje wartości ujemne dla wartości mniejszej od średniej i wartości dodatnie dla wartości większych od średniej.

> ! Zawsze jak pracujesz z tabelą rozkładu normalnego upewnij się co ona dokładnie pokazuje. Czasem są to wartości dystrybuanty $\Phi(z)$, czasem $1 - \Phi(z)$, a zdarza się też $\Phi(z) - 0,5$. Często tablicuje się tylko wartości dodatnie $z$ albo tylko wartości ujemne. Dobrą metodą weryfikacji czy rozumiesz daną tabelę jest próba odczytania z niej reguły 68-95-99.7.

**Problem 4.3** Odczytaj z tablic statystycznych regułę 68-95-99.7.

**Twierdzenie 4.1** Jeżeli zmienne losowe $X_i \sim N(\mu_i, \sigma_i)$ mają rozkład normalny i są niezależne to rozkład ich sumy jest również normalny $\sum_i^n X_i \sim N(\sum_i \mu_i, \sqrt{\sum_i \sigma_i^2})$

**Problem 4.4** Podczas lotu samolotem jednym z kluczowych momentów jest wzbicie się w powietrze, kiedy samolot na krótkim pasie startowym musi osiągnąć tzw. prędkość decyzji i oderwać się od podłoża. Aby było to możliwe samolot musi być wyposażony w mocne silniki, ale także nie bez znaczenia jest waga załadunku, a w przypadku samolotu pasażerskiego waga ludzi, która jak wiemy ma rozkład normalny. Samolot może unieść maksymalnie $5,8t$, średnia waga pasażera to $91,64kg \pm 16,66kg$, całkowita liczba pasażerów to 60. Jakie jest prawdopodobieństwo, że samolot nie będzie mógł się wzbić w powietrze z powodu zbyt dużej wagi?

■ **Przykład 4.2** Zgodnie z próbką pobraną z USDA Food Commodity Intake Database [2] średni wzrost mężczyzny pomiędzy 20 i 62 rokiem życia w USA wynosi $177,8cm$ z odchyleniem standardowym $8,38cm$. Jakie jest prawdopodobieństwo, że losowy dorosły mężczyzna będzie miał wzrost pomiędzy $175,26cm$ a $188cm$?

Dokonajmy standaryzacji obu wartości: $175cm$ i $188cm$

$$z_1 = \frac{x-\mu}{\sigma} = \frac{175-177,8}{8,38} \approx -0,3$$

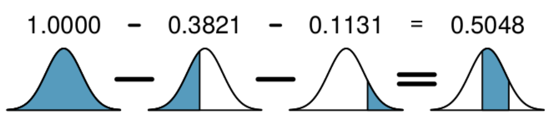
$$z_2 = \frac{x-\mu}{\sigma} = \frac{188-177,8}{8,38} \approx 1.21$$

Dalej zapiszmy szukane prawdopodobieństwo:

$$\begin{aligned}
P(175,26 < X < 188) &= P(-0,3 < Z < 1.21) \\
&= P(Z > -0.3) - P(Z > 1.21) \\
&= 1 - P(Z < -0.3) - P(Z > 1.21) \\
&= 1 - P(Z < -0.3) - (1 - P(Z < 1.21)) \\
&= 1 - P(Z < -0.3) - 1 + P(Z < 1.21) \\
&= -P(Z < -0.3) + P(Z < 1.21)
\end{aligned}$$

Już w tym miejscu można skorzystać z tabeli zawierającej wartości dystrybuanty, ale zakładając, że mamy dostępne tylko dystrybuanty dla ujemnych wartości $Z$ przekształcamy dalej...

$$\begin{aligned}
P(175,26 < X < 188) &= -P(Z < -0.3) + (1 - P(Z < -1.21)) \\
&= 1 - P(Z < -0.3) - P(Z < -1.21) \\
&= 1 - \Phi(-0,3) - \Phi(-1,21) \\
&= 1 - 0,3821 - 0,1131 = 0,5048
\end{aligned}$$




Prawdopodobieństwo, że losowy dorosły mężczyzna będzie miał pomiędzy $175,26cm$ a $188cm$ wynosi ok. 50%. ■

Przy rozwiązywaniu tego typu zadań często dużo łatwiej jest narysować sobie rozkład normalny i zakolorować szukane pole pod wykresem (jak na rysunku w przykładzie). W ten sposób często unikniesz błędów!

**Problem 4.5** Mamy zmienną losową o rozkładzie N(1,3), prawdopodobieństwo, że $P(X < x)$ wynosi $0,04$. Ile się równa $x$?

**Ćwiczenie 4.1 — Rozkład normalny.** Rozwiąż zadanie pod adresem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/03/cw-1.xls` ■

## 4.3 Rozkłady prawdopodobieństwa w R

R jak każdy pakiet statystyczny posiada funkcje pozwalające na pracę z różnymi rozkładami prawdopodobieństwa. Każdy z rozkładów prawdopodobieństwa posiada w R cztery funkcje:

- dla rozkładów ciągłych funkcja gęstości prawdopodobieństwa (ang. *probability density function, PDF*) lub dla rozkładów dyskretnych funkcja masy prawdopodobieństwa (ang. *probability mass function, PMF*) – każda z tych funkcji posiada prefiks `d`.
- dystrybuanta (ang. *cumulative distribution function, CDF*) posiada prefiks `p`. Przypomnijmy, że używamy jej do odpowiedzi na pytania jak „jakie jest prawdopodobieństwo, że zmienna przyjmie wartość mniejszą niż $x$".
- odwrotna dystrybuanta (ang. *inverse cumulative distribution function*) ma prefiks `q` i służy do znalezienia odpowiedzi na pytanie „Od jakiej liczby $x$% populacji jest mniejsze?".
- generator liczb losowych z danego rozkładu ma prefiks `r`

Znamy już prefiksy odpowiednich rodzajów funkcji, które należy jeszcze połączyć z angielskim skrótem nazwy rozkładu:

| | |
|---|---|
| rozkład dwumianowy | `binom` |
| rozkład $\chi^2$ | `chisq` |
| rozkład normalny | `norm` |
| rozkład F | `f` |
| rozkład t-Studenta | `t` |
| rozkład jednorodny | `unif` |

1. Znajdź wartość prawdopodobieństwa, że w sekwencji $n = 50$ niezależnych zdarzeń Bernoulliego było $k = 5$ sukcesów, jeżeli prawdopodobieństwo sukcesu wynosi $p = 0.05$.

```
dbinom(x=5, size=50, prob=0.05)
```

Nazwy argumentów można pominąć o ile znasz ich kolejność:

```
dbinom(5, 50, 0.05)
```

2. Znajdź prawdopodobieństwo, że uzyskano co najwyżej 5 sukcesów

   ```
   pbinom(5, 50, 0.05)
   ```

   Znajdź to samo prawdopodobieństwo używając funkcji `dbinom()` i funkcji `sum()`.[a]
3. Wylosuj wektor 1000 liczb pochodzących z rozkładu dwumianowego z $n = 100$ i $p = 0.5$.

   ```
   sample <- rbinom(1000, 100, 0.5)
   ```

   a następnie narysuj ich histogram.
4. * Chcielibyśmy porównać uzyskany histogram z funkcją gęstości rozkładu normalnego $N(50,5)$. Z tego powodu potrzebujemy znormalizować histogram z częstotliwości do prawdopodobieństw. Znajdź odpowiedni przełącznik funkcji `hist` w pomocy.
5. * Aby dodać do histogramu linię przedstawiającą gęstość rozkładu normalnego wykorzystamy funkcję `lines()`, która przyjmuje dwa argumenty: wektor $x$ i wektor $y$. Wygeneruj odpowiednie wektory, a następnie wywołaj funkcję `lines`.
6. * Alternatywnie można narysować krzywą rozkładu normalnego używając funkcji `curve(wyrazenie, add=T)` przyjmującej wyrażenie z argumentem `x` pod które są podstawiane wartości z osi x wykresu.

! Dla chętnych: aby sprawdzić czy pobrana próbka ma rozkład (prawie) normalny możesz wykorzystać specjalny typ wykresu do tego celu: wykres kwantylowy (ang. *Q-Q plot*). W R możesz go skonstruować poprzez wywołanie `qqnorm(sample) qqline(sample)`.

[a]Wskazówka: funkcja `dbinom()` może przyjąć wektor liczb jako pierwszy argument.

**Ćwiczenie 4.2** Oblicz prawdopodobieństwo $P(2 < X < 5)$ jeżeli $X \sim N(7,1)$. ■

**Ćwiczenie 4.3** Narysuj wykres funkcji gęstości standardowego rozkładu normalnego. ■

## 4.4 Symulacje

### 4.4.1 Uniwersalność rozkładu jednorodnego

W komputerze generujemy symulujemy zdarzenia losowe poprzez wykorzystanie tzw. generatorów liczb pseudolosowych. Taki generator to nic innego jak bardzo skomplikowany wzór matematyczny nakładany na jedną lub kilka zmiennych przechowywanych w generatorze. Wzór ten jest deterministyczny tzn. gdy zainicjujemy zmienne generatora w taki sam sposób to zawsze otrzymamy ten sam ciąg liczb pseudolosowych. Z tego powodu ważne jest inicjalizowanie generatora różnymi wartościami początkowymi np. liczbą pobraną z zegara systemowego, choć w niektórych sytuacjach np. przy debugowaniu powtarzalność generacji liczb pseudolosowych dla tej samej wartości początkowej może być bardzo użyteczna.

Jednakże, taki generator zwykle generuje liczby pseudolosowe o rozkładzie jednorodnym z przedziału $[0,1)$. Co zrobić jeśli interesuje nas generowanie liczb z dowolnego rozkładu? Jeżeli interesuje nas inny rozkład jednorodny możemy to zwykle osiągnąć poprzez odpowiednie wymnożenie i dodanie/odjęcie jakieś wartości od tego rozkładu, ale co z rozkładem normalnym, Poissona itd.? Przypomnijmy sobie rozkład jednorodny, a potem poznajmy jego bardzo użyteczną własność, która pozwoli nam na rozwiązanie tego problemu.

**Definicja 4.3 — Rozkład jednorodny.** Rozkład jednorodny (jednostajny, równomierny, prostokątny) to ciągły rozkład prawdopodobieństwa, dla którego gęstość prawdopodobieństwa w przedziale od $a$ do $b$, jest stała i różna od zera, a poza nim równa zeru.

**Problem 4.6 — Własności rozkładu jednostajnego.** Odpowiedz na poniższe pytania:

- Jaką wartość przyjmuje funkcja gęstości prawdopodobieństwa w przedziale od $a$ do $b$?
- Ile wynosi wartość oczekiwana?
- Ile wynosi wariancja?

**Problem 4.7** W jaki sposób możesz wygenerować liczby losowe z rozkładu Bernoulliego z prawdopodobieństwem sukcesu $p$?

**Twierdzenie 4.2 — Uniwersalność rozkładu jednorodnego.** Niech $F$ będzie ciągłą, silnie rosnącą [a] dystrybuantą wtedy $X = F^{-1}(U) \sim F$ jeżeli $U \sim Uniform(0,1)$.

Także jeżeli $X \sim F$ to wtedy $F(X) \sim Uniform(0,1)$.

[a] Zależy nam na tym, aby $F^{-1}$ była dobrze określona. Jeżeli zdefiniujemy $F^{-1}(x) = \inf\{y : F(y) \geq x\}$ to możemy pominąć to wymaganie.

### 4.4.2 Metody Monte Carlo

Metoda Monte Carlo jest stosowana do modelowania matematycznego procesów zbyt złożonych (obliczania całek, łańcuchów procesów statystycznych), aby można było przewidzieć ich wyniki za pomocą podejścia analitycznego. Istotną rolę w metodzie Monte Carlo odgrywa losowanie (wybór przypadkowy) wielkości charakteryzujących proces, przy czym losowanie dokonywane jest zgodnie z rozkładem, który musi być znany. Metoda została opracowana i pierwszy raz zastosowana przez Stanisława Ulama [**?** ], polsko-amerykańskiego matematyka.

**Problem 4.8** W jaki sposób możemy uzyskać oszacowanie liczby $\pi$ używając generatora liczb pseudolosowych?

**Problem 4.9** W jaki sposób możemy przybliżyć $F(x) = P(X \leq x)$ umiejąc wylosować liczby z danego rozkładu?

**Ćwiczenie 4.4 — Metody Monte Carlo.** Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/03/cw-2.xls` i rozwiąż ćwiczenie. ■

# Literatura

## Literatura powtórkowa

Podstawowe statystyki opisowe można znaleźć w rozdziałach 3.1—3.2 darmowej książki „OpenIntro Statistics” [2]. Książka jest dostępna do ściągnięcia na stornie `www.openintro.org/stat/textbook.php`. Z literatury w języku polskim polecam czwarty rozdział książki [1]. W Internecie jest też wiele filmów opisujących proces obliczania prawdopodobieństw rozkładu normalnego np. `https://www.youtube.com/watch?v=4R8xm19DmPM`.

## Literatura dla chętnych

Polecam do obejrzenia film o zastosowaniach rozkładu normalnego: `https://www.youtube.com/watch?v=W5ZiwMczBNY`, a dla tych którzy mają trochę więcej wolnego czasu: `https://www.youtube.com/watch?v=HdkVVAOOmgU`.

# Pytania sprawdzające zrozumienie

**Pytanie 4.1** Przy danej $\mu$ i danym prawdopodobieństwie $P(X > x) = p$ oraz przy użyciu tabeli ze standardowym rozkładem normalnym znajdź odchylenie standardowe zmiennej losowej o rozkładzie normalnym (oraz analogiczne zadania z daną $\sigma$, nieznaną średnią itd.).

**Pytanie 4.2** Dla zmiennej losowej o rozkładzie normalnym wyznacz prawdopodobieństwo $P(a < X < b)$.

# Bibliografia

[1] Amir D. Aczel. *Statystyka w zarządzaniu. Pełny wykład.* PWN, 2005.

[2] D.M. Diez, C.D. Barr, i M. Çetinkaya Rundel. *OpenIntro Statistics: Third Edition*. OpenIntro, Inc., 2015. ISBN 194345003X. URL `openintro.org`.

[3] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, i M. Wasilewski. *Rachunek prawdopodobienstwa i statystyka matematyczna w zadaniach. Część II: Statystyka matematyczna*. Wydawnictwo Naukowe PWN, 2002. ISBN 8301113847.

[4] P. Van Mieghem N. Blenn, C. Doerr. The Overlooked Role of the Lognormal Distribution in Social Networks , 2014. Network Architectures and Services, TU Delft, the Netherlands.