Ćwiczenia 3

1. Wygeneruj dwie próby 100-elementowe z rozkładu $N(0,1)$ i wykonaj dla nich histogramy. Czy wyglądają tak samo? Dlaczego?

2. **(S)** Wygeneruj 100 liczb z rozkładu $N(100, 10^2)$. Ile procent z nich znajduje się w odległości co najwyżej dwóch odchyleń standardowych od średniej?

3. **(S)** Wykres normalności jest bardzo zgrabną metodą sprawdzania normalności danych. Bardziej prymitywną metodą jest sprawdzenie, czy zachodzi reguła 3 sigm, czyli że około 68% danych jest w odległości odchylenia standardowego od średniej, około 95% w odległości 2 odchyleń no i 99,8% w odległości do trzech odchyleń. Wygeneruj 200 obserwacji z rozkładu $N(0,1)$ i zastosuj do nich opisaną metodę.

4. **(S)** Napisz funkcję, która jako wynik zwraca wykres słupkowy sumy oczek uzyskanych na trzech kostkach k6. Jej jedynym parametrem jest liczba rzutów. Za jej pomocą sprawdź jak zmieniają się wyniki wraz ze wzrostem liczby rzutów.

5. Napisz funkcję, która symuluje liczbę działających żarówek z 500, gdzie każda żarówka ma szansę działania równą 0,99. Używając tej funkcji oszacuj wartość oczekiwaną oraz wariancję zmiennej losowej X, która przyjmuje wartość 1 jeśli żarówka się pali i 0 jeśli się przepaliła. Jakie są wartości teoretyczne tej wartości oczekiwanej i wariancji?

6. **(*)** *Czas oczekiwania n-tego klienta w kolejce do pojedynczego serwera.* Niech klienci oznaczeni $C_0, C_1, \ldots, C_n$ przybywają w chwilach $\tau = 0, \tau_1, \ldots, \tau_n$ w celu skorzystania z serwera. Czasy pomiędzy przybyciami $A_1 = \tau_1 - \tau_0, \ldots, A_n = \tau_n - \tau_{n-1}$ są niezależnymi zmiennymi losowymi o tym samym rozkładzie wykładniczym z parametrem λ_A. Czasy obsługi $S_0, S_1, \ldots, S_n$ są niezależnymi zmiennymi losowymi, które są również niezależne od czasów pomiędzy przybyciami i mają rozkład wykładniczy z parametrem λ_S. Niech W_j oznacza czas oczekiwania klienta C_j. W związku z tym klient C_j opuszcza miejsce o czasie

   $$\tau_j + W_j + S_j.$$

   Jeśli ten czas jest większy od τ_{j+1} to następny klient C_{j+1} musi czekać przez czas:

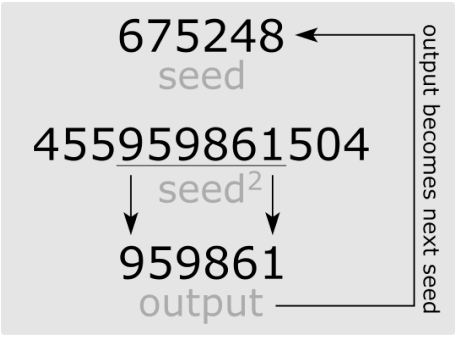
   $$\tau_j + W_j + S_j - \tau_{j+1}.$$

   Mamy zatem następującą relację:

   $$\begin{aligned} W_0 &= 0 \\ W_{j+1} &= \max 0, W_j + S_j - A_{j+1}, \end{aligned}$$

dla $j = 0, 1, \ldots, n-1$. Napisz funkcję, która symuluje zmienną W_n. Użyj jej następnie do wygenerowania 10 000 obserwacji tej zmiennej losowej ($n = 50$, $\lambda_A = 2$, $\lambda_S = 2$). Na tej podstawie wyznacz wartość oczekiwaną i odchylenie standardowe zmiennej losowej W_n.

7. **(*)** *Generowanie liczb pseudolosowych.* Pierwszą znaną metodę generowania liczb pseudolosowych podał John von Neumann w 1949 roku. Jest to metoda środkowego kwadratu (*ang. middle-square method*). Dla zadanego ziarna generatora w metodzie tej podnosimy ziarno do kwadratu, ewentualnie dodajemy na początku zera, tak aby długość otrzymanego wyniku była równa $2n$, gdzie n to liczba cyfr ziarna (powinna być parzysta – dlaczego?). Ze środka liczby wybieramy n cyfr. To jest pierwsza liczba pseudolosowa, która staje się nowym ziarnem. Proces ten kontynuujemy aż do uzyskania zadanej liczebności liczb pseudolosowych. Napisz funkcję, która dla dwóch parametrów (ziarno oraz liczba elementów, które chcemy otrzymać) realizuje ten algorytm.



Rysunek 1: Przykład działania metody środkowego kwadratu (Zadanie 7).

8. **(*)** *Spacer losowy.* Symetryczny prosty spacer losowy, rozpoczynający się w początku układu współrzędnych., jest zdefiniowany następująco. Załóżmy, że $X_1, X_2, \ldots$ to niezależne zmienne losowe o tym samym rozkładzie:

$$\begin{cases} +1 & \text{z prawdopodobieństwem } 1/2 \\ -1 & \text{z prawdopodobieństwem } 1/2 \end{cases}.$$

Zdefiniujmy ciąg $\{S_n\}_{n \geq 0}$:

$$\begin{aligned} S_0 &= 0 \\ S_n &= S_{n-1} + X_n \end{aligned}$$

Ciąg ten jest nazywany symetrycznym prostym spacerem losowym zaczynającym się w początku układu współrzędnych. Napisz funkcję, która symuluje ten proces i zwraca wektor realizacji dla ustalonego n. Funkcja powinna mieć parametr, który pozwala określić czy chcemy aby narysowany był również wykres tego spaceru. Następnie napisz funkcję, która symuluje jedno wystąpienie spaceru, który trwa przez czas n i następnie zwraca długość czasu, przez który spacer odbywa się powyżej osi X (czas jaki spacer spędza powyżej osi X, jest taki sam jak liczba punktów w wektorze $(S_0 + S_1, S_1 + S_2, \ldots, S_{n-1} + S_n)$, które są większe od 0). Użyj tej ostatniej funkcji do oszacowania symulacyjnie ile średnio czasu spędza space losowy nad osią X jeśli jest on długości 1000 (wygeneruj 10 000 procesów).

9. **(S)** Wykonano 50 pomiarów ciepła wydzielanego w pewnym procesie. Otrzymano $\bar{x} = 4{,}8$ oraz $s = 0{,}4$. Zakładając, że pomiary miały rozkład normalny, obliczyć prawdopodobieństwo, że pojedynczy pomiar różniłby się od średniej o 0,8 lub więcej. Ilu wyników większych od tej wartości można się spodziewać? Czy stosując kryterium Chauveneta odrzucimy obserwację o wartości 4? A o wartości 6?

10. **(S)** Przeprowadzono 12 pomiarów aktywności długożyciowego źródła promieniotwórczego otrzymując wyniki: 12, 34, 22, 14, 22, 17, 24, 22, 18, 14, 18, 12. Stosując test Dixona, Grubbsa (poziom istotności 0,05) oraz kryterium Chauveneta oceń czy wynik o wielkości 34 należy uznać za odstający.

11. Zbiór danych `trees` zawiera informacje o średnicy, wysokości i miąższości drewna 31 ściętych czereśni. Sprawdź za pomocą lasów izolacyjnych oraz współczynnika odstawania lokalnego czy znajdują się w nim obserwacje odstające. Wykonaj wykres, gdzie dla każdej obserwacji, na osi X mamy wartości współczynnika odstawania lokalnego, a na osi Y wartości uzyskane za pomocą lasów izolacyjnych. Wielkość punktów ma być proporcjonalna do iloczynu wyznaczonych wskaźników. Opisz najbardziej odstające punkty za pomocą numeru wiersza ze zbioru danych. Czy obie metody wskazują tę samą obserwację jako najbardziej odstającą?

12. Dla zbioru danych `trees` zastąp wszystkie wartości 80 cechy `Height` poprzez NA. Na tak przygotowanym zbiorze wypróbuj różne metody imputacji. Porównaj imputacje z oryginałami, która metoda okazała się najlepsza?