Lista 2

11. Niech $\{X_i\}_{i=1}^{\infty}$ będzie ciągiem niezależnych zmiennych losowych (zdefiniowanych na wspólnej przestrzeni probabilistycznej Ω), o zadanym poniżej rozkładzie:

    $X_i \in \{-i; 0; i\}$ gdzie $\Pr\{X_i = -i\} = \Pr\{X_i = i\} = \frac{1}{2i \log(i+1)}$.

    Pokaż, że:

    a) $\{X_i\}_{i=1}^{\infty}$ spełnia słabe prawo wielkich liczb

    b) $\{X_i\}_{i=1}^{\infty}$ nie spełnia mocnego prawa wielkich liczb (podpowiedź: użyj drugiego lematu Borela-Cantelli'ego)

12. Indeks Giniego G ciągłej nieujemnej zmiennej losowej X (która m. in. może opisywać rozkład majątku w społeczeństwie) jest zdefiniowana jako podwojone pole pomiędzy linią $y = x$ dla $x \in [0,1]$ i tak zwanej krzywej Lorenza $L_X(x)$, $x \in [0,1]$. Krzywa Lorenza zdefiniowana jest jako:

    $L_X(x) = \frac{1}{\mathbb{E}X} \int_0^{F^{-1}(x)} z\varphi(z)$ gdzie $\varphi(z)$ jest gęstością X, zaś F^{-1} jest odwrotnością funkcji dystrybuanty F zmiennej losowej X.

    a) oblicz indeks Giniego dla zmiennej losowej o ciągłym rozkładzie jednostajnym na odcinku $[0,a]$, $a > 0$

    b) oblicz indeks Giniego dla zmiennej losowej o ciągłym rozkładzie jednostajnym na odcinku $[a,b]$ $b > a > 0$ jako funkcji a i b

    c) oblicz indeks Giniego dla zmiennej losowej o rozkładzie wykładniczym o gęstości $\varphi(z) = \lambda e^{-\lambda z}$ dla $z \geqslant 0$ i 0 w przeciwnym wypadku.

    d) pokaż że krzywa Lorenza zawsze jest ograniczona linią $y = x$ (dla $x \in [0,1]$)

13. Niech $X_1, ..., X_n$ będzie ciągiem niezależnych zmiennych losowych o funkcji gęstości p_θ danej przez $p_\theta(x) = \theta^2 x \exp(-\theta x)$ dla $x \geqslant 0$, gdzie $\theta > 0$ jest nieznanym parametrem.

    a) Wyznacz estymator maksymalnej wiarogodności dla θ.

    b) Oblicz informację Fishera i_θ dla θ i oblicz plug-in estymator dla i_θ

    c) Wyznacz przybliżony przedział ufności dla θ na poziomie ufności $1-\alpha$ na podstawie centralnego twierdzenia granicznego dla estymatora maksymalnej wiarogodności parametru θ.

14. Niech $X_1, ..., X_n$ będzie próbą z rozkładu prawdopodobieństwa z gęstością

    $p_\theta(x) = \theta(1+x)^{-(1+\theta)}$ dla $x > 0$ oraz 0 w przeciwnym wypadku, gdzie $\theta > 1$ jest nieznanym parametrem.

    a) Wyznacz metodę estymatora momentów dla parametru θ.

    b) Użyj metodę delta i centralne twierdzenie graniczne, by wyznaczyć asymptotyczny przedział ufności na poziomie ufności $1-\alpha$ dla tego estymatora.

15. Niech $\Omega = \{\omega = \{\omega_i\}_{i=1}^{\infty} : \omega_i \in \{-1,+1\}\}$ będzie zbiorem nieskończonych ciągów o wyrazach $-1,+1$ wyposażonym w miarę produktową, to znaczy że zdarzenia $A_i^{\pm} = \{\omega : \omega_i = \pm 1\}$ są niezależne dla różnych i. Niech $S_n(\omega) = \sum_{i=1}^{n} i\omega_i$. Ciąg $\{S_n(\omega)\}_{i \geqslant 1}$ nazywamy spacerem losowym.

    a) Przy użyciu centralnego twierdzenia granicznego oszacować prawdopodobieństwo dla dużych n, że $|S_n(\omega)| \geqslant n^\alpha$, $\alpha \in (0,1)$

    b) Niech B będzie zdarzeniem $\{\text{dla nieskończenie wielu } n : |S_n(\omega)| \geqslant n^\alpha\}$. Znajdź najmniejszą liczbę α taką że $\Pr\{B\} = 0$. (Podpowiedź: użyj lematu Borela-Cantellego)

16. Niech $X_1, ..., X_n$ będzie ciągiem niezależnych zmiennych losowych z nieznanego ustalonego rozkładu.

   a) Użyj nierówności Dvoretzky'ego-Kiefera-Wolfowitza do zaprojektowania testu dla hipotezy $H_0 : X_i \sim Exp\,(1)$

   b) Niech $n = 6$, a wartości próby równe: $X_1 = 1; X_2 = 4; X_3 = 0.5; X_4 = 2; X_5 = 0,2$ oraz $X_6 = 3$. Czy odrzucisz czy też zaakceptujesz H_0 jeśli błąd pierwszego rodzaju powinien być równy najwyżej 0.05?

17. Rozważ zbiór 9 punktów w $\mathbb{R}^2$, gdzie każdy punkt jest realizacją pary zmiennych losowych (X, Y). Każdy punkt ma przypisaną binarną wartość $Z \in \{0, 1\}$. Wartości próbek razem z odpowiadającymi im wartościom Z, oznaczone przez $(X, Y \mid Z)$ są następujące:

   $(1, 3 \mid 0)\,; (2, 1 \mid 0)\,; (2, 3 \mid 0)\,; (3, 2 \mid 0)\,; (4, 4 \mid 0)\,; (3, 5 \mid 1)\,; (4, 5 \mid 1)\,; (5, 4 \mid 1)\,; (5, 6 \mid 1)\,.$

   a) Dokonaj regresji logistycznej dla zmiennej klasyfikacji Z jako funkcji (X, Y).

   b) Jak dobrze ta regresja opisuje wartości Z pochodzące z rozważanej próby?

18. Niech M_0 będzie wartością portfela w chwili 0 i niech M_i będzie wartością portfela w roku i. W każdym roku wartość portfela wzrasta o 40% z prawdopodobieństwem $\frac{1}{2}$ lub spada o 20% w stosunku do wartości portfela z poprzedniego roku.

   a) Użyj prawa wielkich liczb, by opisać jak szybko M_n rośnie jako funkcja n dla dużych n.

   b) Użyj lewostronnej nierówności Czebyszewa, by wyestymować prawdopodobieństwo że $M_{10} \leqslant \frac{1}{2} M_0$. (Lewostronna nierówność Czebyszewa oznacza, że dla zmiennej losowej X z wartością oczekiwaną μ i wariancją σ^2 następujące ograniczenie jest spełnione dla każdego $a > 0 : \Pr(X \leqslant \mu - a) \leqslant \frac{\sigma^2}{\sigma^2 + a^2}$.)

19. Niech $\{X_i\}_{i=1}^n$ będzie ciągiem niezależnych zmiennych losowych o tym samym rozkładzie gdzie $X_i \in \{-1, +1\}$ oraz $\Pr\{X_1 = 1\} = \frac{1}{2}$. Niech $S_n = \sum\limits_{i=1}^{n} X_i$.

   a) Użyj centralnego twierdzenia granicznego do oszacowania prawdopodobieństwa, że $S_{1000} < -50$.

   b) Jak duże n jest potrzebne, by prawdopodobieństwo tego, że $S_n > 1000$ było równe co najmniej $\frac{1}{4}$? (Ponownie należy użyć centralnego twierdzenia granicznego).

20. Księgowy chce uprościć swoje obliczenia przez zaokrąglenie wszystkich liczb do najbliższej liczby całkowitej. Jest zainteresowany ile wynosi suma pierwszych 100 błędów zaokrągleń wynikających z jego metody. Aby to przeanalizować, zastosuj poniższy model. Oznaczmy błędy zaokrągleń przez $X_1, ..., X_{100}$, gdzie X_i są niezależnymi zmiennymi losowymi o rozładzie jednostajnym na $(-1/2, 1/2)$.

   a) Oblicz wartość oczekiwaną i wariancję X_1.

   b) Zastosuj centralne twierdzenie graniczne by pokazać, że $\Pr(|\sum\limits_{i=1}^{100} X_i| > 10)$ jest w przybliżeniu równe $2\Pr(N > \sqrt{12})$, gdzie N ma standardowy rozkład normalny.

   c) Znajdź przedział $(a, b) \subset \mathbb{R}$ taki że $\Pr\left\{\sum\limits_{i=1}^{100} X_i \in (a, b)\right\}$ to w przybliżeniu 0.9.

21. Niech X będzie zmienną losową o zadanej dystrybuancie $F_X(x) = 1 - x^{-a}$ dla $x \geqslant 1$ i $F_X(x) = 0$ dla $x < 1$ gdzie $a > 2$. Parametr a jest dla nas nieznany. Załóżmy, że zaobserwowaliśmy 4 niezależne realizacje zmiennej losowej X, które mają wartości $2, 6, 3, 12$. Użyj metody największej wiarogodności do znalezienia estymatora $\tilde{a}$ prawdziwej wartości parametru a. Oblicz wartość oczekiwaną i wariancję X dla znalezionej wartości $\tilde{a}$ oraz oblicz wartość oczekiwaną i wariancję, gdyby a było znane i równe 3.

22. Automat do kawy akceptuje monety 1\$ i 2\$ dolarowe, pobiera 1\$ dolar za kawę i wydaje resztę gdy ktoś zapłaci monetą dwudolarową pod warunkiem, że w automacie wciąż były monety jednodolarowe. Automat ma dwa pojemniki: jeden na monety jednodolarowe, a drugi na monety dwudolarowe. Całkowita ilość monet które mieszą się w automacie jest równa $N \gg 1$. Prawdopodobieństwo że osoba płaci za kawę monetą 1\$ wynosi p_1, a prawdopodobieństwo że osoba płaci monetą 2\$ wynosi p_2. Osoby przychodzą po kawę niezależnie i po jednej w danej chwili (czyli można przyjąć że czas jest dyskretny i w każdym kroku czasowym inna osoba próbuje zakupić kawę w automacie). Automat przestaje pracować w przypadku gdy jeden z pojemników (pojemnik na monety 1\$ lub pojemnik na monety 2\$) będzie pełny.

   a) Zaproponuj dobór wielkości N_1 dla pojemnika na monety 1\$ i N_2 dla pojemnika na monety 2\$ jeśli automat do kawy powinien pracować tak długo, jak to tylko możliwe. Znajdź przybliżoną odpowiedź dla dużego N ($N = N_1 + N_2$) znając wartości p_1 oraz p_2. (Podpowiedź: użyj mocnego prawa wielkich liczb.)

23. Niech $\{X_i\}_{i=1}^n$ będzie ciągiem niezależnych zmiennych losowych o tym samym rozkładzie gdzie $X_i \in \{-1, +1\}$ oraz $\Pr\{X_1 = 1\} = \frac{1}{2}$. Niech $S_n = \sum\limits_{i=1}^n X_i$.

   a) Podaj dokładne górne ograniczenie prawdopodobieństwa, że $S_{2500} > 100$ poprzez połączenie centralnego twierdzenia granicznego i twierdzenia Berry'ego-Esseena.

      (Twierdzenie Berry'ego-Esseena: Mamy ciąg niezależnych zmiennych losowych o tym samym rozkładzie $\{Y_i\}_{i=1}^n$ spełniających $\mathbb{E}Y_i = 0$, $Var\left(Y_i\right) = \sigma^2$ oraz $\rho = \mathbb{E}\left(|Y_i|^3\right)$. Niech dalej F_n będzie dystrybuantą $\frac{\sum_{i=1}^n Y_i}{\sigma\sqrt{n}}$ oraz niech Φ będzie dystrybuantą standardowego rozkładu normalnego $\mathcal{N}(0,1)$. Wówczas następujące nierówności są spełnione dla każdego $n \geqslant 1$:

$$\sup_x |F_n(x) - \Phi(x)| \leqslant \frac{1}{2} \cdot \frac{\rho}{\sigma^3\sqrt{n}}.)$$

   b) Podaj inne górne ograniczenie na $\Pr\{S_{2500} > 100\}$ poprzez użycie prawostronnej nierówności Hoeffdinga.

      (Prawostronna nierówność Hoeffdinga:
      Mamy ciąg niezależnych zmiennych losowych o tym samym rozkładzie $\{Y_i\}_{i=1}^n$ spełniających $a \leqslant Y_i \leqslant b$ dla pewnych liczb a i b i dla wszystkich i. Wówczas następujące nierówności są spełnione dla wszystkich $t > 0$ oraz $n \geqslant 1$

$$\Pr\left\{\sum_{i=1}^n Y_i - \mathbb{E}\left(\sum_{i=1}^n Y_i\right) \geqslant t\right\} \leqslant \exp\left(-\frac{2t^2}{nc^2}\right) \tag{1}$$

      gdzie $c := b - a$.)

   c) Podaj jeszcze inne górne ograniczenie na $\Pr\{S_{2500} > 100\}$ poprzez użycie tak zwanego ograniczenia Chernoffa.

      (Ograniczenie Chernoffa:
      Mamy ciąg niezależnych zmiennych losowych o tym samym rozkładzie Bernoulliego $\{Y_i\}_{i=1}^n$, – to jest $Y_i \in \{0; 1\}$ – oraz $\Pr\{Y_i = 1\} = p$. Niech $S_n = \sum\limits_{i=1}^n Y_i$. Wówczas następująca nierówność zachodzi dla wszystkich $n \geqslant 1$ i dla wszystkich $\delta > 0$:

$$\Pr\{S_n \geqslant (1+\delta)\, np\} \leqslant \exp\left(-\frac{\delta^2 np}{2+\delta}\right).)$$

   d) Które z powyższych trzech ograniczeń jest najlepsze?

   e) Które z trzech podejść do otrzymania górnego oszacowania pozwala również obliczyć nietrywialne dolne oszacowanie na $\Pr\{S_{2500} > 100\}$?