1 Laboratorium 3. Estymacja dystybuanty i gęstości rozkładu.

1. Wygeneruj próbę rozmiaru $n = 100$ z rozkładu normalnego $N(0,1)$. Na jednym rysunku umieść wykresy dystrybuanty Φ tego rozkładu i dystrybuanty empirycznej $\widehat{F}_n$, odpowiadającej tej próbie. Powtórz tę analizę dla próby z rozkładu wykładniczego z parametrem $\lambda = 1$.

2. **Pasmo ufności dla dystrybuanty:** Niech $X_1, \ldots, X_n$ będzie próbą losową z populacji o dystrybuancie F. Dla ustalonego $\alpha \in (0,1)$ i ustalonego $n \geq 1$ zdefiniujmy $\varepsilon_n = \sqrt{\dfrac{\ln(2/\alpha)}{2n}}$ oraz

   $$L(x) = \max\{\widehat{F}_n(x) - \varepsilon_n, 0\},\ U(x) = \min\{\widehat{F}_n(x) + \varepsilon_n, 1\},\ x \in \mathbb{R}.$$

   Z nierówności Dvoretzky'ego-Kiefera-Wolfowitza, przytocznej na wykładzie, można wywnioskować, że dla każdego $n \in \mathbb{N}$ i każdej dystrybuanty F spełniony jest warunek

   $$\Pr(\, L(x) \leq F(x) \leq U(x) \text{ dla każdego } x \in \mathbb{R}\,) \geq 1 - \alpha.$$

   To oznacza, że $[\,L(x), U(x)\,]$ jest tzw. pasmem ufności dla F na poziomie ufności $1 - \alpha$.

   Badania symulacyjne. Aby sprawdzić jak działa to pasmo wygeneruj $M = 1000$ razy próbę losową rozmiaru $n = 100$ z rozkładu o dystrybuancie F. Dla każdej z tych M prób skonstruuj pasmo ufności dla F na poziomie ufności $1 - \alpha$. W ilu procentach przypadków wykres dystrybuanty F leży pomiędzy wykresami funkcji L i U? Przyjmij $\alpha = 0.05$, $F = \Phi$ oraz $F =$ dystrybuanta rozkładu wykładniczego z parametrem $\lambda = 1$.

3. Wygeneruj próbę rozmiaru $n = 500$ z rozkładu normalnego $N(0,1)$.

   Na jednym rysunku umieść wykres gęstości rozkładu $N(0,1)$ oraz wykresy kilku estymatorów jądrowych, z jądrem gaussowskim, otrzymane dla różnych szerokości pasma h_n. Jak zmiana szerokości pasma wpływa na gładkość wykresu?

   Dodaj do rysunku wykres kolejnego estymatora jądrowego z szerokością pasma wybraną za pomocą **reguły kciuka Silvermana** (ang. Silverman's rule of thumb)*:

   $$h_n = 0.9 \cdot \min\left\{s, \frac{\mathrm{IQR}}{1.34}\right\} n^{-1/5}.$$

   Symbole s i IQR oznaczają odchylenie standardowe w próbie i rozstęp międzykwartylowy w próbie.

   * Ta metoda wyboru szerokości pasma działa dobrze, gdy estymowana gęstość nie różni się zbytnio od gęstości rozkładu normalnego.

4. Wygeneruj próbę* $x_1, x_2, \ldots, x_{500}$ z mieszanki rozkładów normalnych

   $$0.4 \cdot N(0,1) + 0.4 \cdot N(2,1) + 0.2 \cdot N(4, 2^2).$$

Na jednym rysunku umieść wykresy gęstość rozkładu tej mieszanki, histogramu i estymatora jądrowego z jądrem gaussowskim. Wybierz liczbę klas histogramu za pomocą reguły Freedmana-Diaconisa i szerokość pasma estymatora jądrowego za pomocą reguły kciuka Silvermana. Który z tych dwóch estymatorów wydaje się lepszy?

*Aby otrzymać taką próbę, najpierw wygeneruj próbę losową $u_1, \ldots, u_{500}$ z rozkładu jednostajnego na przedziale $(0, 1)$. Następnie, dla każdego $i = 1, \ldots, 500$,

(a) wygeneruj y_i, takie że

$$y_i \stackrel{D}{=} \begin{cases} N(0, 1^2), & \text{gdy} \quad u_i \in [0, 4/10), \\ N(2, 1^2), & \text{gdy} \quad u_i \in [4/10, 8/10), \\ N(4, 2^2), & \text{gdy} \quad u_i \in [8/10, 1); \end{cases}$$

(b) przyjmij $x_i = y_i$.