Zadania z Analizy danych

zestaw 7

Zadanie 1 (Metoda najmniejszych kwadratów)

Wykorzystaj metodę najmniejszych kwadratów do znalezienia liczby zawartych zakładów gry Lotto na podstawie liczby wygranych poszczególnych stopni. Wylicz tą metodą odchylenie standardowe liczby zawartych zakładów. Wykorzystaj dane z zadań dotyczących tego zagadnienia z zestawu zadań nr 3. Porównaj wyniki uzyskane obiema metodami.

Dane z losowania z dnia 07-05-2016

Stopień	Liczba wygranych k_i	Oszacowanie liczby zakładów
6	3	41 951 448
5	644	34 905 444
4	31 128	32 124 096
3	571 008	32 547 456

Dane z losowania z dnia 10-05-2016

Stopień	Liczba wygranych
szóstka	0
piątka	42
czwórka	2 312
trójka	46 767

Rozwiązanie

Liczby wygranych poszczególnych stopni k_i są dla nas estymatorami wartości oczekiwanych wygranych poszczególnych stopni np_i. Wariancje liczb wygranych są różne $\sigma^2(k_i) = np_i(1-p_i)$, dlatego musimy minimalizować funkcję o następującej postaci

$$f(n) = \sum_{i=3}^{6} \frac{(np_i - k_i)^2}{\sigma^2(k_i)} = \sum_{i=3}^{6} \frac{(np_i - k_i)^2}{np_i(1-p_i)}.$$

Warunkiem istnienia minimum funkcji jest zerowanie się jej pochodnej

$$\frac{df}{dn} = 0.$$

Policzmy zatem pochodną funkcji $f(n)$.

$$\frac{df}{dn} = \sum_{i=3}^{6} \frac{2p_i(np_i - k_i)np_i(1-p_i) - (np_i - k_i)^2 p_i(1-p_i)}{\left(np_i(1-p_i)\right)^2} =$$

$$\sum_{i=3}^{6} \frac{2p_i(np_i - k_i) - \frac{(np_i - k_i)^2}{n}}{np_i(1-p_i)} = \sum_{i=3}^{6} \frac{2p_i^2 n - 2p_i k_i - np_i^2 + 2p_i k_i - \frac{k_i^2}{n}}{np_i(1-p_i)} =$$

$$\sum_{i=3}^{6} \frac{p_i^2 n^2 - k_i^2}{n^2 p_i(1-p_i)} = \sum_{i=3}^{6} \frac{p_i^2}{p_i(1-p_i)} - \frac{1}{n^2} \sum_{i=3}^{6} \frac{k_i^2}{p_i(1-p_i)} = 0.$$

Z warunku

$$\sum_{i=3}^{6} \frac{p_i^2}{p_i(1-p_i)} - \frac{1}{n^2} \sum_{i=3}^{6} \frac{k_i^2}{p_i(1-p_i)} = 0$$

Znajdujemy

$$n = \sqrt{\frac{\sum_{i=3}^{6} \frac{k_i^2}{p_i(1-p_i)}}{\sum_{i=3}^{6} \frac{p_i^2}{p_i(1-p_i)}}}.$$

Wariancja tej zmiennej losowej wynosi

$$\sigma^2(n) = \sum_{i=3}^{6} \left(\frac{\partial n}{\partial k_i}\right)^2 \sigma^2(k_i) = \frac{1}{\sum_{i=3}^{6} \frac{p_i^2}{p_i(1-p_i)}} \sum_{i=3}^{6} \left(\frac{\frac{2k_i}{p_i(1-p_i)}}{2\sqrt{\sum_{i=3}^{6} \frac{k_i^2}{p_i(1-p_i)}}} \right)^2 np_i(1-p_i) =$$

$$\frac{n}{\sum_{i=3}^{6} \frac{p_i^2}{p_i(1-p_i)}} \cdot \frac{1}{\sum_{i=3}^{6} \frac{k_i^2}{p_i(1-p_i)}} \sum_{i=3}^{6} \frac{k_i^2}{p_i(1-p_i)} = \frac{n}{\sum_{i=3}^{6} \frac{p_i}{1-p_i}}.$$

Podsumujmy

$$n = \sqrt{\frac{\sum_{i=3}^{6} \frac{k_i^2}{p_i(1-p_i)}}{\sum_{i=3}^{6} \frac{p_i}{1-p_i}}}.$$

$$\sigma(n) = \sqrt{\frac{n}{\sum_{i=3}^{6} \frac{p_i}{1-p_i}}}.$$

Dla danych z dnia 07-05-2016 dostajemy

$$n = 32\,528\,236, \quad \sigma(n) = 41\,546.$$

$$n = 32\,528(42) \cdot 10^3.$$

Przypomnijmy, że w rozwiązaniu z zestawu 3 otrzymaliśmy

$$\tilde{n} = 32\,527(42) \cdot 10^3.$$

Dla danych z dnia 10-05-2016 dostajemy

$$n = 2\,651\,682, \quad \sigma(n) = 11\,862.$$

$$n = 2\,652(12) \cdot 10^3.$$

Przypomnijmy, że w rozwiązaniu z zestawu 3 otrzymaliśmy

$$\tilde{n} = 2\,647(12) \cdot 10^3.$$

Zadanie 2 (Metoda Boxa i Mullera)

Korzystając z metody Boxa i Mullera wygeneruj ciąg 100 liczb o rozkładzie normalnym z parametrami $\mu = 3$ i $\sigma = 0{,}4$. Policz wartość średnią i odchylenie standardowe otrzymanego ciągu i porównaj z wartościami parametrów μ i σ. Zbuduj histogram dla otrzymanego ciągu i porównaj z teoretyczną krzywą Gaussa.

Zadanie 3 (Centralne twierdzenie graniczne)

Powtórz poprzednie zadanie generując liczby według wzoru

$$x = \mu + \sigma \cdot \left(\sum_{i=1}^{12} y_i - 6 \right),$$

gdzie liczby y_i podlegają rozkładowi jednostajnemu na przedziale $[0,1]$.

Zadanie 4

Zastosuj metodę Monte Carlo do wyliczenia liczby π.

a) Wygeneruj dużą liczbę N par liczb x_i, y_i o rozkładzie jednostajnym na przedziale $[0,1]$. Policz liczbę n wygenerowanych par, dla których $x_i^2 + y_i^2 \leq 1$. Wylicz przybliżoną wartość liczby π ze wzoru

$$\pi \approx \frac{4n}{N}.$$

b) Wygeneruj dużą liczbę N liczb x_i o rozkładzie jednostajnym na przedziale $[0,1]$. Wylicz przybliżoną wartość liczby π ze wzoru

$$\pi \approx \frac{4}{N} \sum_{i=1}^{n} \sqrt{1 - x_i^2}.$$

Zadanie 5

Wylicz metodą Monte Carlo całkę potrójną

$$I = \int_{x=0}^{x=1} \int_{y=0}^{y=1} \int_{z=0}^{z=1} \sin^2(x + y + z) dx dy dz.$$

Przybliżenie całki wylicz ze wzoru

$$I \approx \frac{1}{N}\sum_{i=1}^{N} \sin^2(x_i + y_i + z_i).$$

Porównaj wynik otrzymany tą metodą dla $N = 10^k, k = 3,4, \dots ,8$ z wynikiem dokładnym

$$I = \frac{1}{16}(8 + 3\sin 2 - 3\sin 4 + \sin 6) \approx 0{,}79493.$$

Porównaj wyliczony błąd z estymatorem odchylenia standardowego otrzymanej przez nas wartości:

$$S(I) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n} (\sin^2(x_i + y_i + z_i) - T_n(I))^2}.$$