

Wybrane zagadnienia Matematyki II - Zadanie 3

Testy nieparametryczne

Zaimplementuj testy zgodności (Pearsona χ^2 oraz Kołmogorowa) weryfikujące hipotezę o pochodzeniu danej próbki z rozkładu:

- jednostajnego
- normalnego.

Dane wejściowe program pobiera z pliku tekstowego sformatowanego w identyczny sposób jak w poprzednim zadaniu. Dla każdej zmiennej program powinien wyświetlić wartości odpowiednich statystyk. Dla testu χ^2 należy też podać liczbę stopni swobody, dla testu Kołmogorowa dodatkowo wartość statystyki granicznej D_n .

Test χ^2 Pearsona

Dzielimy próbę $X_1, X_2, \dots, X_n$ na k klas (szereg rozdzielczy). Oznaczmy przez n_i liczbę elementów w i -tej klasie. Statystyka testu jest postaci:

$$\chi_d^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \quad (1)$$

gdzie $n = \sum n_i$, zaś $p_i = P(a_{i-1} < X \leq a_i) = F_0(a_i) - F_0(a_{i-1})$ jest prawdopodobieństwem, że zmienna znajduje się w klasie ograniczonej przedziałem $(a_{i-1}, a_i]$. Przedział krytyczny testu wyznaczamy z tablic kwartyli rozkładu χ^2 $[\chi^2(1 - \alpha, f), \infty)$, gdzie $f = k - 1$ jest liczbą stopni swobody.

Np. dla $k = 10$ wartości krytyczne $\chi^2(1 - \alpha, 9)$ wynoszą:

α	0.05	0.01	0.005
$\chi^2(1 - \alpha)$	16.919	21.666	23.589

Test Kołmogorowa

Statystyka testowa jest postaci:

$$D_n = \sup_x |F_0(x) - S_n(x)| \quad (2)$$

gdzie S_n jest dystrybuantą empiryczną, F_0 - dystrybuantą testowanego rozkładu. W praktyce wyznaczenie tej statystyki sprowadza się do wykonania następujących kroków. Porządkujemy wyniki w rosnącej kolejności $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ i obliczamy

$$d_n^+ = \max_{1 \leq i \leq n} \left| \frac{i}{n} - F_0(x_{(i)}) \right| \quad d_n^- = \max_{1 \leq i \leq n} \left| F_0(x_{(i)}) - \frac{i-1}{n} \right| \quad (3)$$

Następnie wybieramy większą z liczb d_n^+, d_n^- , czyli $d_n = \max(d_n^+, d_n^-)$. Dla danego poziomu istotności α oraz n odczytujemy z tablic wartość krytyczną statystyki Kołmogorowa $d_n(1-\alpha)$. Dla dużej liczby $n \rightarrow \infty$ (w praktyce dla $n > 100$) można posłużyć się rozkładem granicznym statystyki $\sqrt{n}d_n$, dla której wartości krytyczne wynoszą:

α	0.05	0.01	0.005
$D_n(1-\alpha)$	1.224	1.354	1.628

Przybliżona wartość dystrybuanty rozkładu normalnego

Weryfikując hipotezę o pochodzeniu próbki z rozkładu normalnego niezbędna jest znajomość wartości dystrybuanty tego rozkładu $F_0(x) = P(X < x)$ dla każdego x . Jedną z numerycznych metod wyznaczania przybliżonej wartości dystrybuanty rozkładu normalnego jest rozwinięcie w szereg postaci

$$P(X < x) \approx 1 - Z(x) \sum_i^N a_i t^i \quad (4)$$

gdzie

$$Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad t = \frac{1}{1 + px}$$

Wartości odpowiednich stałych dla $N = 5$ znajdują się w tabeli. Dokładność tego przybliżenia wynosi $|\epsilon(x)| < 7.5 * 10^{-8}$ [1].

a_1	a_2	a_3	a_4	a_5	p
0.319381530	-0.356563782	1.781477937	-1.821255978	1.330274429	0.2316419

Literatura

- [1] Abramowitz and Stegun, *Handbook of Mathematical Functions* Dover Publications, Inc, New York, 1965.
- [2] W. Kryszicki, J. Bartos, W. Dyska, K. Królikowska, M. Wasilewski, *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach (cz. 2)*, Wydawnictwo Naukowe PWN, Warszawa 2002.