

Test zgodności χ^2 -Pearsona

Niech będzie dana próba $X_1, \dots, X_n$ pobrana z rozkładu o dystrybuancie F . Testami zgodności nazywamy testy dotyczące hipotez, które mają postać

$$H : F = F_0,$$

gdzie F_0 jest zadaną dystrybuantą.

Przy testach zgodności alternatywą jest na ogół nieparametryczna hipoteza

$$H_1 : F \neq F_0.$$

Jednym z najbardziej rozpowszechnionych testów zgodności jest test χ^2 -Pearsona. Załóżmy, że wyniki próby zostały zgrupowane w r rozłącznych klasach o licznosciach $n_1, \dots, n_r$, gdzie $n_1 + \dots + n_r = n$. Niech p_i oznacza teoretyczne prawdopodobieństwo (przy prawdziwej hipotezie H), że obserwowana zmienna losowa przyjmie wartość z i -tej klasy. Jako miarę rozbieżności pomiędzy wynikami teoretycznymi a wynikami z doświadczenia używamy w tym teście statystyki

$$\chi_p^2 = \sum_{i=1}^r \frac{(n_i - np_i)^2}{np_i}, \quad (1)$$

która dla dużych (przy założeniu prawdziwości hipotezy H) ma w przybliżeniu rozkład χ^2 o $r-1$ stopniach swobody. Wówczas zbiór krytyczny na poziomie istotności α ma postać

$$W = [\chi_{r-1}^2(1-\alpha), \infty), \quad (2)$$

gdzie $\chi_{r-1}^2(1-\alpha) = F^{-1}(1-\alpha)$ dla dystrybuanty rozkładu chi-kwadrat o $r-1$ stopniach swobody.

Zadanie 1. Korzystając z generatora rozkładu Bernoullego $B(5, 0,5)$ wylosowano próbę o licznosci $n=100$. (Generator $B(n, p)$ zwraca liczbę sukcesów w n próbach przy prawdopodobieństwie sukcesu w jednej próbie wynoszącym p). Wyniki są zebrane w poniższej tabeli

Liczba sukcesów, k	0	1	2	3	4	5
Liczba obserwacji	3	16	36	32	11	2

Zweryfikować na poziomie istotności $\alpha = 0,05$ hipotezę, że rozkładem generowanym przez generator jest faktycznie rozkład $B(5, 1/2)$.

Zadanie 2. Rzucono 120 razy kostką do gry i otrzymano wyniki zebrane w tabeli poniżej

Liczba oczek, i	1	2	3	4	5	6
Liczba obserwacji, n_i	14	15	27	13	22	29

Czy są podstawy, aby wątpić w prawidłowość kostki?

Zadanie 3. Załóżmy, że zmienne losowe powiązane są liniowo $Y = aX + b$. W wyniku eksperymentu otrzymujemy wartości, które jednak obarczone są błędami (ε_i).

$$y_i = ax_i + b + \varepsilon_i, \quad i = 1, \dots, n. \quad (3)$$

Oznaczmy przez $\hat{a}$ i $\hat{b}$ estymatory najmniejszych kwadratów współczynników a, b , tzn. rozwiązanie następującego problemu minimalizacji

$$\min \left\{ \sum_{i=1}^n (y_i - ax_i - b)^2 : a, b \in \mathbb{R} \right\} \quad (4)$$

Wyznaczyć współczynniki $\hat{a}$ i $\hat{b}$.

Zadanie 4. Wyznaczyć współczynniki regresji kwadratowej dla następujących danych

x_i	-2	-1,5	0	0,7	0,9	1,2	1,3	2
y_i	-4,1	-2,8	-0,28	0,4	0,62	0,7	0,79	1,05

Wykorzystać Excel do rozwiązania liniowego układu trzech równań z trzema niewiadomymi.

Zadanie 5. W pewnym okresie zmieniano stawki celne na import samochodów. Zarejestrowano następujące procentowe udziały (w stosunku do ogólnej liczby) kupionych samochodów pochodzących z importu:

Stawka cła, x_i	0%	5%	8%	10%	15%	20%
Kupno samochodów z importu, Y_i	32%	25%	20%	17%	16%	11%

- (i) Wyznaczyć metodą najmniejszych kwadratów prostą regresję zmiennej losowej Y względem x .
- (ii) Wyznaczyć współczynnik R^2 i porównać z wynikiem, który daje Excel.
Wsk. Wzór na R^2 (zwany *współczynnikiem determinacji*) jest następujący:

$$R^2 = 1 - \frac{S_{rez}}{S_{tot}}, \quad S_{rez} = \sum_i (y_i - \hat{y}_i)^2, \quad S_{tot} = \sum_i (y_i - \bar{y})^2,$$

$$\hat{y}_i - \text{wartości dopasowane}, \quad \bar{y} = \frac{1}{n} \sum_i y_i.$$

Zadanie 6. Za pomocą kalkulatora lub generatora liczb pseudolosowych w Ms Excel wygenerować 50 liczb losowych z przedziału $[0, 1)$, a następnie sprawdzić czy są losowe. Generator powinien tworzyć liczby losowe o rozkładzie jednostajnym na tym przedziale. Zastosować test zgodności chi-kwadrat Pearsona. W tym celu zaokrąglić wyniki do czterech cyfr, przygotować tabelkę w której będzie zliczona liczba cyfr w wygenerowanym ciągu:

cyfra	0	1	2	3	4	5	6	7	8	9
n_i	?	?	?	?	?	?	?	?	?	?

Teraz testujemy hipotezę H_0 , że rozkład cyfr w ciągu jest jednostajny na poziomie istotności $\alpha = 0,05$ lub $\alpha = 0,01$. Skorzystać ze wzoru (1).

Zadanie 7. W tabelce jest podana liczba godzin/tydzień samodzielnej pracy ucznia oraz średnia ocen. Obliczyć współczynnik korelacji Pearsona i zinterpretować wynik.

Liczba godzin	1	2	3	4	5	6	7	8	10	16
Średnia ocen	3,4	3,6	3,5	3,8	4,0	4,1	4,2	4,3	4,3	4,8