

1 Wykład 9

Test jednorodności

Przykład 1.1 Obserwowano dużą liczbę pacjentów. Części z nich podawano aspirynę, części placebo. Obserwowano wystąpienia ataku serca u pacjentów:

	Atak serca	Bez ataku	suma
aspiryna	104	10933	11037
placebo	189	10845	11034
suma	293	21778	22071

Interesuje nas odpowiedź na pytanie, czy reakcja pacjentów na aspirynę różni się istotnie od reakcji na placebo. ■

Mamy dane, pochodzące z obserwacji n obiektów, podzielone na w grup o ustalonych z góry frakcjach $q_1, q_2, \dots, q_w$ ($\sum_{i=1}^w q_i = 1, q_i > 0$) przedstawione w postaci prostokątnej tabeli:

$$X = \begin{array}{c|cccc} & B_1 & B_2 & \cdots & B_k \\ \hline G_1 & x_{11} & x_{12} & \cdots & x_{1k} \\ G_2 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_w & x_{w1} & x_{w2} & \cdots & x_{wk} \end{array}$$

gdzie B_j są pewnymi zdarzeniami, x_{ij} jest liczbą obserwacji, w których zdarzenie B_j zaszło w grupie G_i . Zdarzenia B_j stanowią układ kompletny¹, opisujący rozkład wartości cechy opisowej B . W naszym przykładzie pacjentów pogrupowano według zastosowanej terapii, cecha B – objawy choroby (wystąpienie lub brak ataku serca). Czy pogrupowanie ma wpływ na rozkład prawdopodobieństwa wartości cechy B ? Pytanie to można sformułować w postaci hipotezy konkurencyjnej H_1 : "pogrupowanie ma wpływ na rozkład prawdopodobieństwa wartości cechy B " i hipotezy zerowej H_0 : "pogrupowanie nie ma wpływu na rozkład prawdopodobieństwa wartości cechy B ".

Matematycznie hipotezę zerową można sformułować w postaci równości:

$$H_0 : p_{1j}/q_1 = p_{2j}/q_2 = \dots = p_{wj}/q_w \stackrel{\text{def}}{=} \mu_j \text{ dla każdego } j = 1, 2, \dots, k$$

gdzie

$$p_{ij} = P(\text{w grupie } G_i \text{ zajdzie zdarzenie } B_j)$$

Oczywiście hipoteza konkurencyjna jest negacją hipotezy zerowej. Macierz obserwacji X ma rozkład wielomianowy

$$P(X) = n! \prod_{i=1}^w \prod_{j=1}^k \frac{p_{ij}^{x_{ij}}}{x_{ij}!}, \quad n = \sum_{i=1}^w \sum_{j=1}^k x_{ij}$$

¹ tzn są rozłączne i ich suma jest zdarzeniem pewnym

Logarytm wiarygodności hipotezy konkurencyjnej wynosi

$$\log L_x(H_1) = \sup \left\{ \log \left(n! \prod_{i=1}^w \prod_{j=1}^k \frac{p_{ij}^{x_{ij}}}{x_{ij}!} \right) : 0 \leq p_{ij}, \forall_i \sum_{j=1}^k p_{ij} = q_i \right\}$$

Wprowadźmy pomocniczo parametry $\lambda_{ij} \stackrel{\text{def}}{=} \frac{p_{ij}}{q_i}$. Warunek

$$\forall_i \sum_{j=1}^k p_{ij} = q_i$$

jest równoważny warunkowi

$$\forall_i \sum_{j=1}^k \lambda_{ij} = 1$$

Podobnie, jak przy wyprowadzaniu wzoru na statystykę χ^2 Pearsona supremum to osiągnięte jest dla $\widehat{p}_{ij} = \widehat{\lambda}_{ij} = \frac{x_{ij}q_i}{x_{i.}}$, gdzie $x_{i.} = \sum_{j=1}^k x_{ij} = nq_i$. Stąd wynika, że $\widehat{p}_{ij} = \frac{x_{ij}}{n}$ oraz, że

$$\log L_x(H_1) = \log \left(\frac{n!}{\prod_{i=1}^w \prod_{j=1}^k x_{ij}!} \right) + \sum_{i=1}^w \sum_{j=1}^k x_{ij} \log \left(\frac{x_{ij}}{n} \right)$$

Dla hipotezy zerowej

$$\log L_x(H_0) = \sup \left\{ \log \left(n! \prod_{i=1}^w \prod_{j=1}^k \frac{(q_i \mu_j)^{x_{ij}}}{x_{ij}!} \right) : 0 \leq \mu_j, \sum_{j=1}^k \mu_j = 1 \right\}$$

supremum jest osiągnięte dla $\widehat{\mu}_j = x_{.j}/n$, gdzie $x_{.j} = \sum_{i=1}^w x_{ij}$ ² i jest równe

$$\begin{aligned} \log L_x(H_0) &= \log \left(\frac{n!}{\prod_{i=1}^w \prod_{j=1}^k x_{ij}!} \right) + \sum_{i=1}^w \sum_{j=1}^k x_{ij} \log \left(\frac{q_i x_{.j}}{n} \right) = \\ &= \log \left(\frac{n!}{\prod_{i=1}^w \prod_{j=1}^k x_{ij}!} \right) + \sum_{i=1}^w \sum_{j=1}^k x_{ij} \log \left(\frac{x_{ij}}{n^2} \right) \end{aligned}$$

Tak więc

$$2 \log L_x(H_0, H_1) = 2 \sum_{i=1}^w \sum_{j=1}^k x_{ij} \log \left(\frac{x_{ij} n}{x_{i.} x_{.j}} \right)$$

Oznaczmy $e_{ij} \stackrel{\text{def}}{=} \frac{x_{i.} x_{.j}}{n} = n \frac{q_i x_{.j}}{n}$. Liczba ta jest oczekiwaną liczbą zajść zdarzenia B_j w grupie G_i gdy prawdziwa jest hipoteza zerowa. Mamy wtedy

$$2 \log L_x(H_0, H_1) = 2 \sum_{i=1}^w \sum_{j=1}^k x_{ij} \log \left(\frac{x_{ij}}{e_{ij}} \right)$$

²czyli, gdy prawdziwa jest H_0 to estymatorami największej wiarygodności prawdopodobieństw p_{ij} są $\widehat{p}_{ij} = \frac{q_i x_{.j}}{n}$

Podobnie więc jak w już rozpatrywanym przykładzie testu zgodności, zbiór krytyczny będzie miał postać:

$$C = \{x : D(x, e) > k\}$$

gdzie odchylenie w tym przypadku będzie miało postać:

$$D(x, e) = 2 \sum_{i=1}^w \sum_{j=1}^k x_{ij} \log \left(\frac{x_{ij}}{e_{ij}} \right)$$

Gdy hipoteza zerowa jest prawdziwa, odchylenie ma rozkład χ^2 z liczbą stopni swobody równą $w(k-1) - (k-1) = (w-1)(k-1)$. Wynika to z faktu, że hipoteza H_1 ma $w(k-1)$ stopni swobody: wszystkich parametrów λ_{ij} jest wk natomiast równań ograniczających $\forall_i \sum_{j=1}^k \lambda_{ij} = 1$ jest w czyli wolnych parametrów jest $wk - w = w(k-1)$; dla hipotezy H_0 wszystkich parametrów μ_j jest k z jednym ograniczeniem $\sum_{j=1}^k \mu_j = 1$ - wolnych parametrów jest $k-1$.

Możemy też w konstrukcji zbioru krytycznego posłużyć się statystyką Pearsona:

$$\chi^2 = \sum_{i=1}^w \sum_{j=1}^k \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$$

Przykład 1.2 (ciąg dalszy) Obliczmy liczebności oczekiwane, gdyby hipoteza zerowa była prawdziwa:

e_{ij}	Atak serca	Bez ataku	suma
aspiryna	146.52	10890.5	11037
placebo	146.48	10887.5	11034
suma	293	21778	22071

Na przykład, $e_{11} = \frac{293 \cdot 11037}{22071} = 146.52$ Statystyka χ^2 Pearsona będzie miała wartość

χ^2	Atak serca	Bez ataku	suma
aspiryna	$\frac{(104-146.52)^2}{146.52} = 12.339$	$\frac{(10933-10890.5)^2}{10890.5} = 0.166$	12.505
placebo	$\frac{(189-146.48)^2}{146.48} = 12.343$	$\frac{(10845-10887.5)^2}{10887.5} = 0.166$	12.509
suma	24.682	0.332	25.014

Poziom krytyczny dla tego testu można oszacować z tablic statystycznych:

$$P(\chi_1^2 > 25.014) < P(\chi_1^2 > 7.879) = 0.005$$

Prowadzi to do odrzucenia hipotezy zerowej o braku wpływu terapii na zachorowania. Z tablicy χ^2 można odczytać, że:

- grupy terapeutyczne są do siebie podobne - poziomy χ^2 są praktycznie takie same
- różnica grup terapeutycznych ujawnia się wśród chorych, którzy przeszli atak serca ($\chi^2 = 24.682$): gdyby H_0 była prawdziwa, oczekiwana liczba chorych w grupie zażywającej aspirynę, (146.52), byłaby większa od zaobserwowanej (104); w grupie leczonej placebo oczekiwana liczba chorych (146.48), byłaby mniejsza od zaobserwowanej (189).■

W podobny sposób można testować hipotezę o niezależności dwóch kompletnych układów zdarzeń³ $(A_1, A_2, \dots, A_w)$ i $(B_1, B_2, \dots, B_k)$.

Hipotezy w tym przypadku mają postać:

$$H_0 : \forall_{ij} P(A_i \cap B_j) = P(A_i) P(B_j),$$

$$H_0 : \exists_{ij} P(A_i \cap B_j) \neq P(A_i) P(B_j)$$

Tablica danych ma taką samą postać jak w przypadku testu jednorodności:

$$X = \begin{array}{|c|c|c|c|c|} \hline & B_1 & B_2 & \cdots & B_k \\ \hline A_1 & x_{11} & x_{12} & \cdots & x_{1k} \\ \hline A_2 & x_{21} & x_{22} & \cdots & x_{2k} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline A_w & x_{w1} & x_{w2} & \cdots & x_{wk} \\ \hline \end{array}$$

gdzie x_{ij} jest liczbą obserwacji, w których zaszło zdarzenie $A_i \cap B_j$. Zadziwiającym zbiegiem okoliczności test, oparty na ilorazie wiarygodności⁴ jest taki sam jak w przypadku testu jednorodności.

W analogiczny sposób można też sprawdzić, czy dane pochodzą z rozkładu o ustalonej dystrybucji. Obserwacje liczbowe możemy podzielić na grupy obserwacji należących do rozłącznych przedziałów i zapisać, ile obserwacji wpadło do poszczególnych przedziałów. W ten sposób otrzymamy ciąg liczebności obserwowanych. Możemy też, mając daną dystrybucję hipotetyczną obliczyć prawdopodobieństwo p_i , że obserwację wpadły do i -tego przedziału. Liczebności oczekiwane można więc obliczyć ze wzoru $e_i = np_i$, gdzie n jest liczbą obserwacji. W ten sposób można, stosując test zgodności⁵, zweryfikować hipotezę o tym, czy obserwacje mają rozkład o danej dystrybucji⁶.

³tym razem zdarzenia $(A_1, A_2, \dots, A_w)$ są losowe

⁴patrz zadanie 1 z listy 6

⁵w obliczaniu liczby stopni swobody należy uwzględnić, ile parametrów rozkładu było wyznaczanych z próby

⁶patrz zadanie 3 z listy 6