

Testowanie hipotez statystycznych.

Krzysztof Topolski

Bioinformatyka

Wrocław, 9 kwietnia 2019

Temat.

Test zgodności χ^2 Pearsona.

Test zgodności χ^2 dla hipotezy złożonej.

Omówiliśmy przypadek, w którym sformułowana hipoteza H_0 określała całkowicie rozkład zmiennej losowej. W wielu zagadnieniach praktycznych stawia się również hipotezy, które nie określają całkowicie rozkładu. Możemy mieć na przykład do czynienia z hipotezą, która orzeka, że zmienna losowa X ma rozkład dyskretny o znanej postaci ogólnej $\{p_i(\theta), i = 1, \dots, k\}$, a parametr θ pozostaje nieokreślony.

W tej sytuacji nie możemy bezpośrednio skorzystać ze statystyki χ^2_P gdyż nie możemy obliczyć, $E_i = E_i(\theta) = np_i(\theta)$, teoretycznych oczekiwanych liczebności klas. Zależą one bowiem od wartości nieznanego parametru θ . Okazuje się, że jeśli nieznaną parametr θ jest oszacowany za pomocą metody największej wiarygodności, to statystyka

$$\chi^2_P = \sum_{i=1}^k \frac{(O_i - E_i(\hat{\theta}))^2}{E_i(\hat{\theta})},$$

gdzie $\hat{\theta}$ oznacza estymator największej wiarygodności parametru θ , ma nadal asymptotyczny rozkład χ^2 ale już nie z $k - 1$ stopniami swobody ale z $k - 2$ stopniami swobody.

Test zgodności χ^2 dla hipotezy złożonej cd.

Przykład

W diploidalnej populacji z losowym kojarzeniem częstość allelu **A** jest równa θ . Jeśli populacja znajduje się w równowadze Hardyego-Weinberga to częstości genotypów **AA**, **Aa** i **aa** są równe odpowiednio θ^2 , $2\theta(1 - \theta)$ i $(1 - \theta)^2$.

Założmy, że w populacji stwierdzono 110 osobników o genotypie **AA**, 235 osobników o genotypie **Aa** oraz 155 osobników o genotypie **aa**. Czy istnieje podstawa do odrzucenia hipotezy o znajdowaniu się tej populacji w warunkach równowagi Hardyego-Wainberga.

Przykład cd.

W tym przypadku hipoteza zerowa ma postać

$$H_0: p_1(\theta) = \theta^2, p_2(\theta) = 2\theta(1 - \theta), p_3(\theta) = (1 - \theta)^2.$$

Estymator parametru θ wyznaczony metodą największej wiarygodności jest postaci

$$\hat{\theta} = \frac{2n_1 + n_2}{2n},$$

gdzie $n = n_1 + n_2 + n_3$.

Ponieważ

Genotyp	AA	Aa	aa
n_i	110	235	155

więc

$$\hat{\theta} = 0,455.$$

Przykład cd.

W tym przypadku hipoteza zerowa ma postać

$$H_0 : p_1(\theta) = \theta^2, p_2(\theta) = 2\theta(1 - \theta), p_3(\theta) = (1 - \theta)^2.$$

Estymator parametru θ wyznaczony metodą największej wiarygodności jest postaci

$$\hat{\theta} = \frac{2n_1 + n_2}{2n},$$

gdzie $n = n_1 + n_2 + n_3$.

Ponieważ

Genotyp	AA	Aa	aa
n_i	110	235	155

więc

$$\hat{\theta} = 0,455.$$

Przykład cd.

W tym przypadku hipoteza zerowa ma postać

$$H_0 : p_1(\theta) = \theta^2, p_2(\theta) = 2\theta(1 - \theta), p_3(\theta) = (1 - \theta)^2.$$

Estymator parametru θ wyznaczony metodą największej wiarygodności jest postaci

$$\hat{\theta} = \frac{2n_1 + n_2}{2n},$$

gdzie $n = n_1 + n_2 + n_3$.

Ponieważ

Genotyp	AA	Aa	aa
n_i	110	235	155

więc

$$\hat{\theta} = 0,455.$$

Statystyka testowa jest więc równa

$$\chi_P^2 = \frac{(110 - 103,51)^2}{103,51} + \frac{(235 - 247,97)^2}{247,97} + \frac{(155 - 148,51)^2}{148,51} = 1,369.$$

Dla tej wartości statystyki testowej p -wartość jest równa 0,24, czyli nie ma podstaw do odrzucenia hipotezy zerowej.

Statystyka testowa jest więc równa

$$\chi_P^2 = \frac{(110 - 103,51)^2}{103,51} + \frac{(235 - 247,97)^2}{247,97} + \frac{(155 - 148,51)^2}{148,51} = 1,369.$$

Dla tej wartości statystyki testowej p -wartość jest równa 0,24, czyli nie ma podstaw do odrzucenia hipotezy zerowej.

Testowanie hipotezy o niezależności

Za pomocą testu χ^2 możemy przetestować hipotezę o niezależności dwóch zmiennych losowych. Niech (X, Y) będzie dwuwymiarowym wektorem losowym o rozkładzie dyskretnym

$$r_{ij} = P\{X = x_i, Y = y_j\}, \quad i = 1, \dots, a, \quad j = 1, \dots, b.$$

Rozkład $\{r_{ij}\}$ możemy wygodnie zapisać w tablicy

X/Y	y ₁	y ₂	...	y _b
x ₁	r ₁₁	r ₁₂	...	r _{1b}
x ₂	r ₂₁	r ₂₂	...	r _{2b}
⋮	...	...	...	...
x _a	r _{a1}	r _{a2}	...	r _{ab}

Testowanie hipotezy o niezależności

Za pomocą testu χ^2 możemy przetestować hipotezę o niezależności dwóch zmiennych losowych. Niech (X, Y) będzie dwuwymiarowym wektorem losowym o rozkładzie dyskretnym

$$r_{ij} = P\{X = x_i, Y = y_j\}, \quad i = 1, \dots, a, \quad j = 1, \dots, b.$$

Rozkład $\{r_{ij}\}$ możemy wygodnie zapisać w tablicy

X/Y	y₁	y₂	...	y_b
x₁	r_{11}	r_{12}	$\dots$	r_{1b}
x₂	r_{21}	r_{22}	$\dots$	r_{2b}
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$
x_a	r_{a1}	r_{a2}	$\dots$	r_{ab}

Testowanie hipotezy o niezależności

Stawiamy hipotezę zerową H_0 , że zmienne losowe X i Y są niezależne.

$$H_0 : P \{(X, Y) = (x_i, y_j)\} = P \{X = x_i\} P \{Y = y_j\}, \text{ dla wszystkich } i, j$$

Jeśli wprowadzimy oznaczenia

$$p_i = P\{X = x_i\} \quad \text{oraz} \quad q_j = P\{Y = y_j\}$$

to możemy hipotezę H_0 zapisać w postaci

$$H_0 : r_{ij} = p_i q_j, \text{ dla wszystkich } i, j.$$

Testowanie hipotezy o niezależności

Hipoteza zerowa jest hipotezą złożoną. Nie określamy ani prawdopodobieństw r_{ij} ani prawdopodobieństw p_i oraz q_j . Orzekamy jedynie, że prawdopodobieństwo r_{ij} jest równe iloczynowi pewnych liczb p_i oraz q_j , które będziemy traktować jako parametry. Ponieważ

$$p_1 + p_2 + \dots + p_a = 1 \quad \text{oraz} \quad q_1 + q_2 + \dots + q_b = 1$$

więc mamy $a + b - 2$ niezależnych parametrów, które musimy estymować na podstawie danych.

Dwa pozostałe parametry, p_a oraz q_b , określamy za pomocą równości

$$p_a = 1 - \sum_{i=1}^{a-1} p_i \quad \text{oraz} \quad q_b = 1 - \sum_{j=1}^{b-1} q_j.$$

Testowanie hipotezy o niezależności

Estymacji dokonujemy na podstawie n elementowej próby prostej $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. Oznaczmy przez O_{ij} liczbę par (X_k, Y_k) równych (x_i, y_j) . Wygodnie jest zapisać dane w tabeli

X/Y	y₁	y₂	...	y_b	Suma
x₁	O_{11}	O_{12}	...	O_{1b}	$O_{1\cdot}$
x₂	O_{21}	O_{22}	...	O_{2b}	$O_{2\cdot}$
$\vdots$	...	...	...	...	$O_{1\cdot}$
x_a	O_{a1}	O_{a2}	...	O_{ab}	$O_{a\cdot}$
Suma	$O_{\cdot 1}$	$O_{\cdot 2}$	...	$O_{\cdot b}$	n

gdzie

$O_{i\cdot}$ — oznacza sumę liczb w i — tym wierszu,

$O_{\cdot j}$ — oznacza sumę liczb w j — tej kolumnie.

Testowanie hipotezy o niezależności

Estymatorami wyznaczonymi metoda największej wiarygodności nieznanych parametrów $p_1, \dots, p_a$ oraz $q_1, \dots, q_b$ są odpowiednio

$$\hat{p}_i = \frac{O_{i\cdot}}{n}, \quad i = 1, \dots, a,$$

oraz

$$\hat{q}_j = \frac{O_{\cdot j}}{n}, \quad j = 1, \dots, b.$$

Procedurę testowania opieramy na statystyce testowej

$$\chi_P^2 = \sum_{j=1}^b \sum_{i=1}^a \frac{\left(O_{ij} - n \frac{O_{i.}}{n} \frac{O_{.j}}{n}\right)^2}{n \frac{O_{i.}}{n} \frac{O_{.j}}{n}}.$$

Po przekształceniu

$$\chi_P^2 = n \sum_{j=1}^b \sum_{i=1}^a \frac{\left(O_{ij} - \frac{O_{i.} O_{.j}}{n}\right)^2}{O_{i.} O_{.j}}.$$

Statystyka χ_P^2 ma asymptotyczny rozkład chi kwadrat z $(a-1)(b-1)$ stopniami swobody.

Procedurę testowania opieramy na statystyce testowej

$$\chi_P^2 = \sum_{j=1}^b \sum_{i=1}^a \frac{\left(O_{ij} - n \frac{O_{i.}}{n} \frac{O_{.j}}{n}\right)^2}{n \frac{O_{i.}}{n} \frac{O_{.j}}{n}}.$$

Po przekształceniu

$$\chi_P^2 = n \sum_{j=1}^b \sum_{i=1}^a \frac{\left(O_{ij} - \frac{O_{i.} O_{.j}}{n}\right)^2}{O_{i.} O_{.j}}.$$

Statystyka χ_P^2 ma asymptotyczny rozkład chi kwadrat z $(a-1)(b-1)$ stopniami swobody.

Procedurę testowania opieramy na statystyce testowej

$$\chi_P^2 = \sum_{j=1}^b \sum_{i=1}^a \frac{\left(O_{ij} - n \frac{O_{i.}}{n} \frac{O_{.j}}{n}\right)^2}{n \frac{O_{i.}}{n} \frac{O_{.j}}{n}}.$$

Po przekształceniu

$$\chi_P^2 = n \sum_{j=1}^b \sum_{i=1}^a \frac{\left(O_{ij} - \frac{O_{i.} O_{.j}}{n}\right)^2}{O_{i.} O_{.j}}.$$

Statystyka χ_P^2 ma asymptotyczny rozkład chi kwadrat z $(a-1)(b-1)$ stopniami swobody.

Liczba stopni swobody wynika z faktu, że mamy $k = ab$ klatek oraz szacujemy $a + b - 2$ niezależnych parametrów, a więc liczba swobody rozkładu statystyki testowej jest równa

$$\nu = ab - 1 - (a + b - 2) = (a - 1)(b - 1).$$

Liczba stopni swobody wynika z faktu, że mamy $k = ab$ klatek oraz szacujemy $a + b - 2$ niezależnych parametrów, a więc liczba swobody rozkładu statystyki testowej jest równa

$$\nu = ab - 1 - (a + b - 2) = (a - 1)(b - 1).$$

Przykład

W fabryce telewizorów sądzi się, że jakość obrazu wyprodukowanego telewizora nie zależy od jakości dźwięku. Aby sprawdzić to przypuszczenie, zbadano 100 wybranych losowo telewizorów i otrzymano wynik podany w tablicy:

Obraz/Dźwięk	zła	dobra	wysoka	Suma
zła	11	17	6	34
dobra	15	18	7	40
wysoka	8	8	10	26
Suma	34	43	23	100

Przykład

W fabryce telewizorów sądzi się, że jakość obrazu wyprodukowanego telewizora nie zależy od jakości dźwięku. Aby sprawdzić to przypuszczenie, zbadano 100 wybranych losowo telewizorów i otrzymano wynik podany w tablicy:

Obraz/Dźwięk	zła	dobra	wysoka	Suma
zła	11	17	6	34
dobra	15	18	7	40
wysoka	8	8	10	26
Suma	34	43	23	100

Należy postawić i przetestować hipotezę, przyjmując poziom istotności $1 - \alpha = 0,9$.

Hipoteza zerowa orzeka, że jakość dźwięku jest niezależna od jakości obrazu.

Należy postawić i przetestować hipotezę, przyjmując poziom istotności $1 - \alpha = 0,9$.

Hipoteza zerowa orzeka, że jakość dźwięku jest niezależna od jakości obrazu.

Należy postawić i przetestować hipotezę, przyjmując poziom istotności $1 - \alpha = 0,9$.

Hipoteza zerowa orzeka, że jakość dźwięku jest niezależna od jakości obrazu.

Oczekiwana częstość w poszczególnych klatkach.

Jakość obrazu/dźwięku	zła	dobra	wysoka	Suma
zła	11, 56	14, 62	7, 82	34
dobra	13, 60	17, 20	9, 20	40
wysoka	8, 84	11, 18	5, 98	26
Suma	34, 00	43, 00	23, 00	100

Otrzymujemy wartość statystyki $\chi^2_P = 5,23$ dla rozkładu χ^2 z 4 stopniami swobody daje to p –wartość równą 0,26.

Tak więc nie ma podstaw do odrzucenia hipotezy zerowej. Jakość dźwięku jest niezależna od jakości obrazu.

Oczekiwana częstość w poszczególnych klatkach.

Jakość obrazu/dźwięku	zła	dobra	wysoka	Suma
zła	11, 56	14, 62	7, 82	34
dobra	13, 60	17, 20	9, 20	40
wysoka	8, 84	11, 18	5, 98	26
Suma	34, 00	43, 00	23, 00	100

Otrzymujemy wartość statystyki $\chi^2_P = 5,23$ dla rozkładu χ^2 z 4 stopniami swobody daje to p –wartość równą 0,26.

Tak więc nie ma podstaw do odrzucenia hipotezy zerowej. Jakość dźwięku jest niezależna od jakości obrazu.

Oczekiwana częstość w poszczególnych klatkach.

Jakość obrazu/dźwięku	zła	dobra	wysoka	Suma
zła	11, 56	14, 62	7, 82	34
dobra	13, 60	17, 20	9, 20	40
wysoka	8, 84	11, 18	5, 98	26
Suma	34, 00	43, 00	23, 00	100

Otrzymujemy wartość statystyki $\chi^2_P = 5,23$ dla rozkładu χ^2 z 4 stopniami swobody daje to p –wartość równą 0,26.

Tak więc nie ma podstaw do odrzucenia hipotezy zerowej. Jakość dźwięku jest niezależna od jakości obrazu.

Dla danych, które możemy przedstawić w tablicy dwudzielczej

X/Y	0	1	Suma
0	a	b	$a + b$
1	c	d	$c + d$
Suma	$a + c$	$b + d$	n

statystyka χ_P^2 przyjmuje postać

$$\chi_P^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}.$$

Dla danych, które możemy przedstawić w tablicy dwudzielczej

X/Y	0	1	Suma
0	a	b	$a + b$
1	c	d	$c + d$
Suma	$a + c$	$b + d$	n

statystyka χ_P^2 przyjmuje postać

$$\chi_P^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}.$$