

Elementarna statystyka

Wnioskowanie dla dwóch zmiennych kategoryjnych

(Inference for two-way tables)

Alexander Bendikov

Uniwersytet Wrocławski

25 maja 2016

Porównywanie kilku proporcji

Przykład: Osobom uzależnionym od kokainy podawano leki mające ułatwić im rzucenie nałogu. W trwającym 3 lata badaniu porównano skuteczność leków opartych na desipraminie oraz na licie, oraz porównano ich skuteczność z placebo. Wyniki badania zamieszczone są w tabeli:

Grupa	Lek	Ilość łącznie	Ilość sukcesów	Proporcja
1	Desipramina	24	14	0,583
2	Lit	24	6	0,250
3	Placebo	24	4	0,167
	Łącznie	72	24	

Porównywanie kilku proporcji

Przykład: Osobom uzależnionym od kokainy podawano leki mające ułatwić im rzucenie nałogu. W trwającym 3 lata badaniu porównano skuteczność leków opartych na desipraminie oraz na litie, oraz porównano ich skuteczność z placebo. Wyniki badania zamieszczone są w tabeli:

Grupa	Lek	Ilość łącznie	Ilość sukcesów	Proporcja
1	Desipramina	24	14	0,583
2	Lit	24	6	0,250
3	Placebo	24	4	0,167
	Łącznie	72	24	

Czy powyższe dane dostarczają dowodu na to, że wśród wszystkich kokainistów proporcje sukcesu w rzucaniu nałogu są różne w zależności od stosowanej terapii?

Hipotezy: $H_0 : p_1 = p_2 = p_3$, H_a : nie wszystkie p_1, p_2, p_3 są równe.

Nasz cel: Chcielibyśmy dokonać wielu porównań jednocześnie z jakąś globalną miarą ufności.

Model: Mamy 3 różne populacje i 3 różne próby. Przy założeniu, że $H_0 : p_1 = p_2 = p_3 (= p)$ jest prawdziwa możemy je potraktować jako próby z tej samej populacji.

(1) W tej sytuacji przybliżamy p przez proporcję połączonych prób $\hat{p}$:

$$\hat{p} = \frac{24}{72} = \frac{1}{3}, \quad \text{oraz} \quad \hat{q} = \frac{48}{72} = \frac{2}{3}.$$

(2) **Oczekiwane obserwacje:** Jeżeli przyjmiemy, że H_0 jest prawdziwa, to oczekiwana ilość sukcesów (wyleczeń) wśród 24 pacjentów w każdej grupie będzie dana przez dwumianową zmienną losową, o parametrach p i $q = (1-p)$. Parametry te przybliżamy przez $\hat{p}$ i $\hat{q}$

	Wyleczenia:			
	obserwowane		oczekiwane	
	Tak	Nie	Tak	Nie
Desipramina	14	10	8	16
Lit	6	18	8	16
Placebo	4	20	8	16

	Wyleczenia:			
	obserwowane		oczekiwane	
	Tak	Nie	Tak	Nie
Desipramina	14	10	8	16
Lit	6	18	8	16
Placebo	4	20	8	16

(3) Statystyka Chi-kwadrat:

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{liczba obserwowana} - \text{liczba oczekiwana})^2}{\text{liczba oczekiwana}} \\
 &= \frac{(14 - 8)^2}{8} + \frac{(10 - 16)^2}{16} + \dots \\
 &= 4,500 + 2,250 + 0,500 + 0,250 + 2,000 + 1,000 \\
 &= 10,500.
 \end{aligned}$$

- (4) p -wartość: rozkład χ^2 zależy od parametru df , liczby stopni swobody. Parametr ten jest równy iloczynowi ilości wierszy pomniejszonej o 1 i ilości kolumn pomniejszonej o 1 w tabeli, w tym przypadku $(2 - 1)(3 - 1) = 2$. Aby otrzymać p -wartość zaglądamy do tabeli rozkładów.

p	0,01	0,005
χ^*	9,21	10,60

Widzimy, że wartość χ^2 10,500 wpada pomiędzy dwie wartości krytyczne 9,21 i 10,60. W takim razie

$$0,005 < p < 0,01.$$

Możemy też skorzystać z komputera i wyliczyć dokładnie, $p = 1 - 0,9948 = 0,0052$.

- (4) p -wartość: rozkład χ^2 zależy od parametru df , liczby stopni swobody. Parametr ten jest równy iloczynowi ilości wierszy pomniejszonej o 1 i ilości kolumn pomniejszonej o 1 w tabeli, w tym przypadku $(2 - 1)(3 - 1) = 2$. Aby otrzymać p -wartość zaglądamy do tabeli rozkładów.

p	0,01	0,005
χ^*	9,21	10,60

Widzimy, że wartość χ^2 10,500 wpada pomiędzy dwie wartości krytyczne 9,21 i 10,60. W takim razie

$$0,005 < p < 0,01.$$

Możemy też skorzystać z komputera i wyliczyć dokładnie, $p = 1 - 0,9948 = 0,0052$.

- (5) Wniosek: Różnice w proporcjach sukcesu pomiędzy grupami są statystycznie istotne na poziomie $\alpha = 0,01$. Odrzucamy hipotezę H_0 na poziomie istotności 1%. Dane stanowią mocne zaprzeczenie tezie $p_1 = p_2 = p_3$.

Uwaga: Zauważmy, że w obecnym kontekście nie można sformułować jednostronnych hipotez H_a . Hipoteza H_a jest po prostu zaprzeczeniem hipotezy H_0 .

Uwaga: Zauważmy, że w obecnym kontekście nie można sformułować jednostronnych hipotez H_a . Hipoteza H_a jest po prostu zaprzeczeniem hipotezy H_0 .

Przykład: W pewnej firmie przeprowadzono badanie nad związkiem pomiędzy stanem cywilnym pracowników (mężczyzn) a poziomem ich zatrudnienia (zaszeregowaniem). W tej firmie zaszeregowanie miało 4 poziomy. W badaniu uwzględniono 8235 pracowników a otrzymane dane są w tabeli.

Poziom	Stan cywilny			
	Singiel	Żonaty	Rozw.	Wdowiec
1	58	874	15	8
2	222	3927	70	20
3	50	2396	34	10
4	7	533	7	4

Czy te dane wskazują na statystycznie istotny związek pomiędzy dwiema zmiennymi kategorijnymi: X - stanem cywilnym i Y - poziomem zatrudnienia?

- (1) Hipotezy: H_0 : nie ma związku, H_a : jest związek.
- (2) Ilości oczekiwane: W każdej komórce tabeli możemy dopisać ilość oczekiwaną przy założeniu prawdziwości H_0 :

$$\text{ilość oczekiwana}_{i,j} = \frac{\text{suma wiersza } i \times \text{suma kolumny } j}{\text{suma całej tabeli}}$$

Otrzymujemy tabelę:

Poziom	Stan cywilny								Σ
	Sing.		Żon.		Rozw.		Wdow.		
1	58	39,08	874	896,44	15	14,61	8	4,87	955
2	222	173,47	3927	3979,03	70	64,86	20	21,62	4239
3	50	101,90	2396	2337,30	34	38,10	10	12,70	2490
4	7	22,55	533	517,21	7	8,43	4	2,81	551
Σ	337		7330		126		42		8235

(3) Obliczamy statystykę Chi-kwadrat:

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{obserwowane} - \text{oczekiwane})^2}{\text{oczekiwane}} \\ &= 9,158 + 0,562 + 0,010 + 2,011 + \\ &\quad + 13,575 + 0,681 + 0,407 + 0,121 + \\ &\quad + 26,432 + 1,474 + 0,441 + 0,574 + \\ &\quad + 10,722 + 0,482 + 0,243 + 0,504 \\ &= 67,397\end{aligned}$$

(4) p -wartość: $df = (4 - 1)(4 - 1) = 9$, zaglądamy do odpowiedniej tabeli dla $df = 9$:

p	0,005	0,000...
χ^*	29,67	∞

Ostatnia wypisana wartość krytyczna wynosi 29,67 z prawdopodobieństwem prawego ogona 0,005. Nasza wartość $\chi^2 = 69,3970 \gg 29,67$. W związku z tym p -wartość $< 0,005$. p -wartość obliczona komputerowo wynosi 0,0000 (!).

- (5) Wniosek: dane wykazują statystycznie istotny związek pomiędzy stanem cywilnym pracowników i ich poziomem zaszeregowania. Hipotezę H_0 należy odrzucić.

- (5) Wniosek: dane wykazują statystycznie istotny związek pomiędzy stanem cywilnym pracowników i ich poziomem zaszeregowania. Hipotezę H_0 należy odrzucić.

Uwaga: Testu χ^2 możemy używać jeżeli nie więcej niż 20% liczb oczekiwanych ≤ 5 i wszystkie ≥ 1 .