Statystyczna Analiza Danych – laboratorium

Test χ^2

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 4
25/26 marca 2021

Test χ^2 zgodności

- Mamy próbę prostą $X_1, \ldots, X_n$ pobraną z rozkładu o dystrybuancie F.
- Testem zgodności nazywamy testy dla hipotez postaci: $H_0 : F = F_0$, gdzie F_0 jest zadaną dystrybuantą
- Hipoteza alternatywna $H_1 : F \neq F_0$
- Uwaga: test χ^2 z powodu badania kwadratu różnic traktujemy jak test jednostronny!
- Załóżmy, że wartości z próby są zgrupowane w d rozłącznych klasach o licznościach $c_1, c_2, \ldots, c_m$. Wszystkich obserwacji jest N
- p_i – teoretyczne p-stwo przy prawdziwej H_0, że obserwowana zmienna losowa przyjmie wartość z i-tej klasy

$$k = \sum_{i=1}^{d} \frac{(c_i - Np_i)^2}{Np_i} \rightarrow \chi^2(d-1)$$

Zadanie przykładowe

- Informatyk miał napisać sampler do losowania próby reprezentatywnej pod względem daty urodzenia z pewnej populacji. Dane dotyczące populacji (częstotliwość urodzin danego dnia) otrzymał w formie tabeli.
- Naszym zadaniem jest weryfikacja, czy napisany sampler działa poprawnie. Wykorzystamy test zgodności chi2.
- Zadanie polega na policzeniu statystyki testowej, znalezieniu właściwej wartości krytycznej i zbudowaniu obszaru krytycznego oraz weryfikacji hipotezy (podanie konkluzji)
- Jaka jest H_0? Jaka jest H_1?

Liczenie statystyki testowej

- **Co mamy** – generujemy próbę z samplera (wartości c_i)
- **Co wiemy o zjawisku** – wracamy do pliku, który zawierał częstotliwość urodzin. Obliczamy prawdopodobieństwa pojawienia się dla każdej daty p_i (liczba urodzin / suma urodzin z pliku)
- **Co powinniśmy mieć, jeśli H_0 prawdziwe** – Przemnażamy Np_i – N to liczba obserwacji wygenerowanych przez sampler dla każdej z dat
- Teraz wystarczy wstawić do wzoru (iterujemy się po kolei po d datach):

$$k = \sum_{i=1}^{d} \frac{(co\ mamy - co\ powinnismy)^2}{co\ powinnismy}$$

Weryfikacja

- Na podstawie obszaru krytycznego – potrzebujemy kwantyl rzędu $1 - \alpha$ rozkładu chi2 o (d-1) stopniach swobody
- Na podstawie p-value: użyć `pchisq(k, d-1)`

Mały przykład numeryczny

- Załóżmy dla uproszenia, że rok składa się z 5 dat
- Sampler wygenerował próbę wielkości $N = 100$: [35,12,23,7,23]
- W pliku dla populacji są następujące częstotliwości urodzin: [360, 130, 256, 62, 192]. Populacja liczy 1000 osób
- Załóżmy poziom istotności $\alpha = 0.05$.

Mały przykład numeryczny

co mamy c_i	co wiemy o zjawisku p_i	co powinniśmy Np_i
35	$360/1000 = 0.36$	$100 * 0.36 = 36$
12	$130/1000 = 0.13$	$100 * 0.13 = 13$
23	$256/1000 = 0.256$	$100 * 0.256 = 25.6$
7	$62/1000 = 0.062$	$100 * 0.062 = 6.2$
23	$192/1000 = 0.192$	$100 * 0.192 = 19.2$

$$\chi^2 = \frac{(35-36)^2}{36} + \frac{(12-13)^2}{13} + \frac{(23-25.6)^2}{25.6} + \frac{(7-6.2)^2}{6.2} + \frac{(23-19.2)^2}{19.2} = 1.22$$

p-value: $[1 - \chi^2(1.22, 4)] = 0.87 > 0.05$ – brak podstaw do odrzucenia H_0
Obszar krytyczny: $[\chi^2_{0.95}(4) = 9.49, +\infty)$ – wartość stat. testowej do niego nie wpada, nie ma podstaw do odrzucenia H_0

Tabele wielodzielcze

- Test niezależności χ^2 pozwala zbadać, czy pomiędzy dwoma zmiennymi dyskretnymi istnieje zależność
- Do jego przeprowadzenia standardowo korzysta się z tabeli wielodzielczej
- Tabela wielodzielcza pozwala zbadać rozkład obserwacji ze względu na dwie cechy jednocześnie
- Poziomy jednej z cech opisywane są przez kolumny, drugiej przez wiersze

Przykład tabeli wielodzielczej

- Badamy istnienie związku pomiędzy płcią (wiersze), a liczbą wypalanych papierosów dziennie (kolumny).
- Wiersz i kolumna z sumami nie są konieczne, ale przydają się przy obliczeniach

Płeć	Liczba wypalonych papierosów			Suma
	< 10	10-20	> 20	
Kobieta	14	20	6	40
Mężczyzna	12	30	18	60
Suma	26	50	24	100

Test niezależności χ^2

- Porównuje się częstości zaobserwowane z częstościami oczekiwanymi, przy założeniu prawdziwości hipotezy zerowej
- H_0 – zmienne są niezależne; H_1 – istnieje związek pomiędzy zmiennymi
- Częstości oczekiwane:

$$E_{ij} = \frac{\sum_{j=1}^{k} n_j \sum_{i=1}^{w} n_i}{\sum_{i=1}^{w} \sum_{j=1}^{k} n_{ij}} = \frac{suma\ wiersza * suma\ kolumny}{suma\ calkowita}$$

k – liczba kolumn; w – liczba wierszy

- Statystyka testowa:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \rightarrow \chi^2((k-1)(w-1))$$

O_{ij} – obserwowana częstość komórki,

Zadanie przykładowe

- Chcemy sprawdzić, czy liczba wypalanych papierosów dziennie jest niezależna od płci (nie ma związku pomiędzy płcią respondenta a liczbą wypalanych papierosów dziennie)

Liczenie statystyki testowej

- **Co mamy** – tabela, w której kolumny reprezentują przedziały dla wypalanych papierosów (np. mało/średnio/dużo), a w wierszach jest liczba osób, które zadeklarowany taki przedział w zależności od płci
- **Co powinniśmy mieć, jeśli H_0 prawdziwe** – gdyby wyniki były niezależne od płci respondenta, to w każdej komórce tabeli widzielibyśmy wartość $\frac{suma\ \ wiersza * suma\ \ kolumny}{suma\ \ calkowita}$
- Teraz wystarczy wstawić do wzoru (iterujemy się po kolei po komórkach tabeli):

$$k = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{(co\ \ mamy - co\ \ powinnismy)^2}{co\ \ powinnismy}$$

Weryfikacja

- Na podstawie obszaru krytycznego – potrzebujemy kwantyl rzędu $1 - \alpha$ rozkładu chi2 o (w-1)(k-1) stopniach swobody
- Na podstawie p-value: użyć `pchisq(k, (w-1)(k-1))`

Mały przykład numeryczny

- Sprawdźmy, czy pomiędzy płcią a paleniem papierosów występuje zależność
- Przyjmijmy poziom istotności $\alpha = 0.05$

Płeć	Liczba wypalonych papierosów			Suma
	< 10	10-20	> 20	
Kobieta	14	20	6	40
Mężczyzna	12	30	18	60
Suma	26	50	24	100

Mały przykład numeryczny

- Co mamy

Płeć	Liczba wypalonych papierosów			Suma
	< 10	10-20	> 20	
K	14	20	6	40
M	12	30	18	60
Suma	26	50	24	100

- Co powinniśmy mieć jeśli H_0 prawdziwe

Płeć	Liczba wypalonych papierosów			Suma
	< 10	10-20	> 20	
K	40*26/100 = 10.4	40*50/100 = 20	40*24/100 = 9.6	40
M	60*26/100 = 15.6	60*50/100 = 30	60*24/100 = 14.4	60
Suma	26	50	24	100

$$\chi^2 = \frac{(14-10.4)^2}{10.4} + \frac{(20-20)^2}{20} + \frac{(6-9.6)^2}{9.6} + \frac{(12-15.6)^2}{15.6} + \frac{(30-30)^2}{30} + \frac{(18-14.4)^2}{14.4} = 4.33$$

p-value: $[1 - \chi^2(4.33, (4-1)*(2-1) = 2)] = 0.12 > 0.05$ – brak podstaw do odrzucenia H_0
Obszar krytyczny: $[\chi^2_{0.95}(2) = 5.99, +\infty)$ – wartość stat. testowej do niego nie wpada, nie ma podstaw do odrzucenia H_0