

Bootcamp Data Science

Przemysław Spurek

Dane kategoryczne

- W próbie liczba danych należących do określonej grupy nazywana jest częstotliwością/częstością wystąpień, więc analiza danych kategorycznych jest analizą częstotliwości/częstości.
- Kiedy porównuje się dwie lub więcej grup, to dane są często prezentowane w formie **Frequency Tables**. Na przykład w poniższej tabeli podana jest liczba osób praworęcznych i leworęcznych.

	Right handed	Left handed	<i>Total</i>
Males	43	9	52
Females	44	4	48
<i>Total</i>	87	13	100

- Teraz będziemy zakładać, że dane są podane w zestawie kategorii i mamy ich częstości wystąpień (całkowita liczba próbek w każdej kategorii).
- Wiele testów dla takich danych opiera się na analizie odchylenia od wartości oczekiwanej.
- Ponieważ rozkład chi kwadrat charakteryzuje zmienność danych (innymi słowy, ich odchylenie od wartości średniej), wiele z tych testów odnosi się do tego rozkładu i często nazywane są testami chi kwadrat.

Test chi kwadrat

W przypadku testu t-Studenta weryfikowaliśmy hipotezę czy średnia próbki różni się od oczekiwanej średniej populacji.

Chi-squared goodness-of-fit (chi kwadrat) jest analogicznym testem dla zmiennych kategorycznych: testuje, czy rozkład przykładowych danych kategorycznych odpowiada oczekiwanemu rozkładowi.

Podczas pracy z danymi kategorizującymi dokładne wartości obserwacji nie są zbyt użyteczne w testach statystycznych, ponieważ kategorie takie jak “mężczyźni”, “kobiety” i inne nie mają znaczenia matematycznego.

Testy dotyczące zmiennych kategorycznych opierają się na liczbie zmiennych, zamiast rzeczywistej wartości samych zmiennych.

Test chi kwadrat

Założmy, że zaobserwowaliśmy częstości wystąpień o_i podczas gdy oczekiwaliśmy częstości (teoretycznych) e_i .

Hipoteza zerowa mówi, że wszystkie dane pochodzą z tej samej populacji. Statystyka testowa ma postać:

$$V = \sum_i \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{f-1}$$

ma rozkład chi kwadrat z $f - 1$ stopniami swobody (gdzie f to liczba klas).

Test chi kwadrat

https://github.com/przem85/statistic_4/blob/master/D08_Z01_chi_square_goodness_of_fit.ipynb

Przykład

Założmy, że wyjeżdżasz wędkować z kolegami. Każdego wieczoru losujecie kto ma sprzątać. Ale po zakończeniu podróży wydaje Ci się, że zrobiłeś większość pracy:

You	Peter	Hans	Paul	Mary	Joe
10	6	5	4	5	3

Zadanie

Wykonaj test chi kwadrat w celu sprawdzenia czy dane demograficzne w USA mają ten sam rozkład co dane w Minesocie. Użyj danych z:

```
https://github.com/przem85/statistic\_4/blob/master/D08\_Z02\_chi\_square\_goodness\_of\_fit.ipynb
```


Chi-Square Contingency Test

Niezależność jest kluczową koncepcją prawdopodobieństwa opisującą sytuację, w której wiedza o wartościach jednej zmiennej nie mówi nic o wartości innej.

- Na przykład: miesiąc w którym urodziłeś się prawdopodobnie nie mówi nic na temat tego jakiej przeglądarki internetowej używasz.
- Więc spodziewamy się, że miesiąc narodzin i preferencje odnośnie przeglądarki będą niezależne.
- Z drugiej strony, miesiąc urodzenia może być związany z twoimi wynikami sportowymi w twoim roczniku (nie być niezależne).

Test chi kwadrat niezależności sprawdza, czy dwie zmienne kateryczne są niezależne (Hipoteza H_0 mówi, że zmienne są niezależne). Test niezależności jest powszechnie używany do określenia, czy zmienne, takie jak: edukacja, poglądy polityczne i inne preferencje różnią się w zależności od czynników demograficznych, takich jak: płeć, rasa i religia.

Chi-Square Contingency Test

https://github.com/przem85/statistic_4/blob/master/D08_Z03_chi-square_contingency.ipynb

Zadanie

W przypadku danych z poniższej tabeli sprawdzić czy to, że ktoś jest praworęczny lub leworęczny zależy od płci.

	Right handed	Left handed	<i>Total</i>
Males	43	9	52
Females	44	4	48
<i>Total</i>	87	13	100

Chi-Square Contingency Test

Statystyka testowa V ma rozkład χ^2 gdy:

- dla wszystkich obserwacji, częstości są większe od 1 ($e_i \geq 1$),
- dla co najmniej 80% wszystkich obserwacji, częstości są większe od 5 ($e_i \geq 5$).

Ilość stopni swobody (DOF) dla tabeli $r \times c$ o r wierszach i c kolumnach wynosi:

$$df = (r - 1) \times (c - 1).$$

Ponadto wiemy, że suma oczekiwanych częstości sumuje się do n :

$$\sum_i o_i = n.$$

Chi-Square Contingency Test

Zadanie

Wykonaj test niezależności chi kwadrat w celu sprawdzenia czy preferencje wyborcze zależą od czynnika demograficznego:

```
https://github.com/przem85/statistic\_4/blob/master/D08\_Z04\_chi-square\_contingency.ipynb
```

Chi-Square Contingency Test

Test chi kwadrat można wykorzystać do wygenerowania testu normalności, np.

H_0 Zmienna losowa X jest rozkładem symetrycznym,

H_1 Zmienna losowa X nie jest rozkładem symetrycznym.

Wiemy, że w przypadku rozkładu symetrycznego średnia arytmetyczna \bar{x} i mediana powinny być prawie takie same. Więc prostym sposobem sprawdzenia tej hipotezy byłoby policzenie, ile obserwacji jest mniejszych niż średnia arytmetyczna (n_-), a ile obserwacji jest większych niż średnia arytmetyczna (n_+). Jeśli średnia i mediana są takie same, to 50% obserwacji powinna być mniejsza niż średnia i 50% powinna być większa niż średnia.

$$V = \frac{(n_- - n/2)^2}{n/2} + \frac{(n_+ - n/2)^2}{n/2} \sim \chi_1^2.$$

Fisher's Exact Test

- Jeśli założenie mówiące, że 80% komórek posiada co najmniej 5 elementów nie jest spełnione, to używamy testu Fishera (Fisher's Exact Test).
- Ten test oparty jest na sumach w wierszach i kolumnach.
- Metoda polega na ocenie prawdopodobieństwa związanego ze wszystkimi możliwymi tabelami 2×2 , które mają takie same sumy wierszy i kolumn, co obserwowane dane.
- Hipoteza zerowa mówi, że zmienne wierszy i kolumn są niezależne.
- W większości przypadków dokładny test Fishera jest korzystniejszy od testu chi kwadrat. Ale, aż do pojawienia się potężnych komputerów, nie było on powszechnie stosowany.

Fisher's Exact Test

Zadanie

Wykonaj test Fisher's Exact Test. Użyj danych z:

```
https://github.com/przem85/statistic\_4/blob/master/D08\_Z05\_Fisher\_exact\_test.ipynb
```

McNemar's Test

Chociaż test McNemara wykazuje powierzchowne podobieństwo do testu chi kwadrat 2×2 lub testu prawdopodobieństwa 2×2 Fishera, robi coś całkiem innego.

Poprzednie testy badały związki, które istnieją między komórkami tabeli. W teście McNemara sprawdza się różnicę między proporcjami, które wynikają z sumy marginalnej tabeli

$$p_A = (a + b/N) \text{ oraz } p_B = (a + c)/N.$$

		B		<i>Total</i>
		0	1	
A	0	a	b	$a + b$
	1	c	d	$c + d$
<i>Total</i>		$a + c$	$b + d$	$N = a + b + c + d$

McNemar's Test

- Pytaniem w teście McNemara jest: czy te dwie proporcje p_A i p_B różnią się istotnie?
- Odpowiedź otrzymana musi uwzględniać fakt, że te dwie proporcje nie są niezależne.
- Korelacja p_A i p_B wynika z faktu, że obie wykorzystują wartość w górnej lewej komórce tabeli.
- Test McNemara może być wykorzystany na przykład w badaniach, w których pacjenci pełnią własną kontrolę lub w badaniach “przed i po”.

McNemar's Test

https://github.com/przem85/statistic_4/blob/master/D08_Z06_McNemar_test.ipynb

W poniższym przykładzie będziemy próbować ustalić, czy lek ma wpływ na konkretną chorobę. W tabeli należy podać liczbę pacjentów oraz diagnozę (choroba: obecna lub nieobecna) przed leczeniem podawaną w rzędach oraz diagnozę po leczeniu w kolumnach.

	After: present	After: absent	<i>Total</i>
Before: present	101	121	222
Before: absent	59	33	92
<i>Total</i>	160	154	314

Test wymaga, aby te same pomiary były zawarte w pomiarach przed i po (dopasowane pary).

McNemar's Test

W tym przykładzie zerowa hipoteza mówi o “jednorodności marginalnej”, co oznacza, że leczenie nie daje żadnego efektu. Z powyższych danych statystyką testową McNemara z ciągłą poprawką Yatesa obliczmy:

$$\chi^2 = \frac{(|b - c| - \text{correctionFactor})^2}{b + c},$$

gdzie χ^2 ma rozkład chi kwadrat z 1 stopniem swobody. Dla małych liczb próbek wartość korekty powinna wynosić 0.5 (korekta Yatesa) lub 1.0 (korekta Edwarda).

Używając korekty Yates's otrzymujemy:

$$\chi^2 = \frac{(|121 - 59| - 0.5)^2}{121 + 59} = 21.01.$$

P-value jest mniejsze od 0.05 więc odrzucamy hipotezę zerową i stwierdzamy, że leczenie miało wpływ na badanych pacjentów.

Zadanie

https:

[//github.com/przem85/statistic_4/blob/master/D08_Z08.ipynb](https://github.com/przem85/statistic_4/blob/master/D08_Z08.ipynb)

Zadanie

Badania nad efektami nowego leku na serce doprowadziły do uzyskania następujących danych:

	Heart rate		Total
	Increased	NOT-increased	
Treated	36	14	50
Not treated	30	25	55
Total	66	39	105

- Czy lek ma wpływ na chorobę?
- Co by się stało gdyby dla jednej osoby zmienił się wynik?

	Heart rate		Total
	Increased	NOT-increased	
Treated	36	14	50
Not treated	29	26	55
Total	65	40	105

https:

[//github.com/przem85/statistic_4/blob/master/D08_Z09.ipynb](https://github.com/przem85/statistic_4/blob/master/D08_Z09.ipynb)

Zadanie (Cz 1.)

Miasto Linz chce wiedzieć, czy ludzie chcą zbudować długą plażę wzdłuż Dunaju. Rozmawiają z miejscowymi ludźmi i decydują się zebrać 20 odpowiedzi z każdej z pięciu grup wiekowych:

(<15, 15-30, 30-45, 45-60, > 60)

Kwestionariusz stwierdza: "Rozwój przybrzeżny przyniesie korzyści Linz" i możliwe odpowiedzi są:

1	2	3	4
Strongly agree	Agree	Disagree	Strongly disagree

Zadanie (Cz 2.)

Rada Miasta chce dowiedzieć się, czy wiek ludzi wpływał na odpowiedzi, szczególnie tych, którzy odczuwali negatywnie (tj. “Nie zgadzali się” lub “Zdecydowanie nie zgadzali się”).

Age group (type)	Frequency of negative responses (observed values)
<15	4
15–30	6
30–45	14
45–60	10
>60	16

- Czy te różnice są znaczące?
- Jaki rozkład i z iloma stopniami swobody ma statystyka testowa?

https:
//github.com/przem85/statistic_4/blob/master/D08_Z10.ipynb

Zadanie (Cz 1.)

W pozwie dotyczącym morderstwa obrona wykorzystuje kwestionariusz do wykazania, że pozwany jest niepoczytalny. W wyniku kwestionariusza oskarżony twierdzi, że “nie jest winny z powodu niepoczytalności”. W odpowiedzi, adwokat oskarżyciela chce pokazać, że kwestionariusz nie działa. Zatrudnia doświadczonego neurologa i przedstawia mu 40 pacjentów, dla których 20 ukończyło test z wynikiem “niepoczytalny”, a 20 ze “zdrowy”. Po wykonaniu badań przez neurologa otrzymujemy, że: 19 osób z wynikiem testu “zdrowy” jest zdrowych, ale tylko 6 spośród 20 z wynikiem “niepoczytalny” jest uznanych za “niepoczytalnego”.

Zadanie (Cz 2.)

	Sane by expert	Insane by expert	Total
Sane	19	1	20
Insane	6	14	20
Total	22	18	40

- Czy ten wynik jest znacząco różny od kwestionariusza?
- Czy wynik byłby znacząco różny, gdyby ekspert orzekł, że 20, a nie 19 ludzi z wynikiem “zdrowy” jest naprawdę zdrowych.

`https://github.com/przem85/statistic_4/blob/master/D08_Z11.ipynb`

Zadanie

Dzienne spalanie energii przez 11 zdrowych kobiet wynosi:
[5260., 5470., 5640., 6180., 6390., 6515., 6805., 7515., 7515., 8230.,
8770.] kJ.

Czy ta wartość znacznie różni się od zalecanej wartości 7725?
Wykorzystaj dwa testy.

https:

[//github.com/przem85/statistic_4/blob/master/D08_Z12.ipynb](https://github.com/przem85/statistic_4/blob/master/D08_Z12.ipynb)

Zadanie

W klinice 15 pacjentów (leniwych) waży:

[76, 101, 66, 72, 88, 82, 79, 73, 76, 85, 75, 64, 76, 81, 86.] kg

i 15 pacjentów (sportowców) waży:

[64, 65, 56, 62, 59, 76, 66, 82, 91, 57, 92, 80, 82, 67, 54] kg.

- Czy leniwi pacjenci są znacznie ciężsi?
- Czy powyższe dane pochodzą z rozkładów normalnych?

Zadanie

https:

[//github.com/przem85/statistic_4/blob/master/D08_Z13.ipynb](https://github.com/przem85/statistic_4/blob/master/D08_Z13.ipynb)

Zadanie

Pobieraj dane z pliku `https://github.com/przem85/statistic_4/blob/master/ANOVA4.txt`

Zawiera on dane z eksperymentu na roślinach, które były hodowane w

trzech różnych warunkach wzrostu.

- Wykonaj ANOVA
- Czy trzy grupy są różne?
- Wykonaj analizę post hoc, który z par jest inny?
- Czy porównanie nieparametryczne (Kruskal-Wallis test) prowadzi do innego wyniku?