Bootcamp Data Science
Zajęcia 3

Przemysław Spurek

Szereg rozdzielczy

Zajmiemy się teraz sytuacją, w której nie dysponujemy pełną informacją na temat wartości cechy dla poszczególnych elementów populacji, a jedynie danymi zapisanymi w postaci szeregu rozdzielczego. Także i w tym przypadku można mówić o tendencji centralnej, rozrzucie oraz dystrybuancie empirycznej.

Szereg rozdzielczy

Rozważmy zatem cechę X w skali porządkowej, dla której mamy dany szereg rozdzielczy, którego klasy wyznaczają punkty:

$$a_0 < a_1 < \cdots < a_k.$$

Niech:

$$n_1, \ldots, n_k$$

będą licznościami tych klas oraz niech:

$$y_i = \frac{a_{i-1} + a_i}{2} \text{ dla } i = 1, \ldots, k.$$

Szereg rozdzielczy

Definicja

Wartość średnią dla szeregu rozdzielczego cechy X określamy wzorem:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i y_i.$$

Szereg rozdzielczy

Definicja

Wartość średnią dla szeregu rozdzielczego cechy X określamy wzorem:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i y_i.$$

Jak widać, określona w powyższej definicji wielkość jest sumą składników odpowiadających poszczególnym klasom, przy czym wielkość każdego składnika zależy od położenia danej klasy oraz jej liczności. Zauważmy, że w skrajnym przypadku, to znaczy, że każdy przedział zawiera dokładnie jedną wartość cechy, będącą jego środkiem. Definicja powyższa pokrywa się z definicją średniej z danych surowych. Oczywiście, w ogólnym przypadku średnia obliczona z danych surowych będzie się różnić (na ogół niewiele) od średniej wyznaczonej dla szeregu rozdzielczego.

Szereg rozdzielczy

Definicja

Średnim błędem dla szeregu rozdzielczego cechy X nazywamy liczbę:

$$b = \frac{1}{n} \sum_{i=1}^{k} n_i |y_i - \bar{x}|,$$

wariancją - liczbę:

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{k} n_i (y_i - \bar{x})^2,$$

zaś odchyleniem standardowym - liczbę:

$$s_n = \sqrt{s_n^2}.$$

Dane kategoryczne

- W próbce 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	*Total*
Males	43	9	*52*
Females	44	4	*48*
Total	*87*	*13*	*100*

Frequency Tables

- 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.

Test chi kwadrat

Podczas pracy z danymi kategoryzującymi 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

Załóżmy, że zaobserwowaliśmy częstości wystąpień o_i podczas gdy oczekiwaliśmy częstotliwości (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/bootcamp/blob/master/statistics/D08_Z01.ipynb

Przykład

Załóżmy, 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

Test chi kwadrat

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/bootcamp/blob/master/statistics/D08_Z02.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 kategoryczne 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/bootcamp/blob/master/statistics/D08_Z03.ipynb

Zadanie

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

	Right handed	Left handed	*Total*
Males	43	9	*52*
Females	44	4	*48*
Total	*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/bootcamp/blob/master/statistics/D08_Z04.ipynb`

Chi-Square Contingency Test

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

$$H_0 \text{ Zmienna losowa X jest rozkładem symetrycznym,}$$

$$H_1 \text{ Zmienna losowa X nie jest rozkładem symetrycznym.}$$

Wiemy, że w przypadku rozkładu symetrycznego średnia arytmetyczna $\bar{x}$ i mediana powinny być prawie takie same. Więc prostym sposobem sprawdzenie 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.

$$V = \frac{(n_- - n/2)^2}{n/2} + \frac{(n_+ - n/2)^2}{n/2} \sim \chi^2_1.$$

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/bootcamp/blob/master/statistics/D08_Z05.ipynb`

A Lady Tasting Tea

R. A. Fisher był jednym z założycieli współczesnej statystyki. Jednym z jego pierwszych eksperymentów, a być może najbardziej znanym, było "A Lady Tasting Tea", w którym sprawdzał hipotezę, że pewna kobieta potrafi rozpoznać czy mleko zostało dolane do herbaty, czy herbata do mleka.

Powyższe wydarzenie, które wydaje się banalne, miało ogromny wpływ na historię współczesnej statystyki.
(Box, J. F. (1978). R. A. Fisher: The life of a scientist. New York: Wiley.)

A Lady Tasting Tea

Prawdziwe znaczenie dla nauki płynące z tego eksperymentu są powstałe pytania:

- *Co należy robić w przypadku zmian losowych w temperaturze, słodyczy, i tak dalej?* Idealnie byłoby, aby wszystkie filiżanki herbaty były identyczne, z wyjątkiem kolejności wlewania mleka lub herbaty. Ale nigdy nie można kontrolować wszystkiego. Jeśli nie możemy kontrolować warunków, to najlepsze co możemy zrobić to wprowadzić losowość.
- *Ilu filiżanek należy użyć w teście? W jakiej kolejności powinniśmy je prezentować?* Najważniejszą ideą jest to, że liczba i kolejność filiżanek powinna umożliwić osobie udowodnienia swoich zdolności i wyeliminowanie nadużycia.
- *Jakie wnioski mogą być wyciągnięte z otrzymanego wyniku (pozytywnego i negatywnego)?*

A Lady Tasting Tea

Rzeczywisty scenariusz opisany przez Fishera jako eksperyment "A Lady Tasting Tea" jest następujący:

- Dla każdej filiżanki rejestrujemy kolejność faktycznego wlewania i odpowiedzi pani. Możemy podsumować wynik tabelą:

		Order of actual pouring		
		Tea first	Milk first	*Total*
Lady says	Tea first	a	b	$a+b$
	Milk first	c	d	$c+d$
	Total	$a+c$	$b+d$	n

- n – to całkowita liczba filiżanek herbaty,
- $a+c$ – liczba filiżanek, do których wlewa się herbatę najpierw,
- $a+b$ – liczba filiżanek, dla których kobieta stwierdziła, że wlewa się herbatę najpierw.
- Jeżeli kobieta mogła wyczuć różnicę, to liczby b i c powinny być małe.
- Z drugiej strony, jeśli kobieta zgaduje, to a i d będą takie same.

A Lady Tasting Tea

- Przypuśćmy, że eksperyment wygląda tak, że przygotujemy 8 filiżanek herbaty po cztery każdego typu.
- Kobieta zostaje poinformowana o tej procedurze.
- Załóżmy, że kubki są prezentowane w losowej kolejności.
- Jej zadaniem jest najpierw zidentyfikować cztery pierwszego typu i cztery drugiego.
- W takim przypadku sumy wierszy i kolumn w powyższej tabeli wynoszą 4:

$$a + b = a + c = c + d = b + d = 4.$$

- W konsekwencji, jeżeli znamy jedną z liczb a, b, c, d to pozostałe możemy prosto wyliczyć

$$b = 4 - a, c = 4 - a, d = a.$$

A Lady Tasting Tea

- Możemy przetestować umiejętności pani, poprzez losowanie kolejności filiżanek.
- Jeśli uznamy, że pani nie potrafi rozróżniać wspomnianych sytuacji, to losowe decyzje sprawią, że cztery filiżanki wybrane przez nią jako te z herbatą (najpierw) są równie prawdopodobne? Każda z ośmiu filiżanek
- Jest $\binom{N}{n} = 70$ możliwości (`scipy.misc.comb(8,4,exact=True)`) wybrania 4 z 8 filiżanek.
- Tylko jeden z 70 sposobów prowadzi do prawidłowej klasyfikacji. Więc jeżeli ktoś strzela odpowiedzi ma 1/70 szansy na nie popełnienie błędów.

A Lady Tasting Tea

- Okazuje się, że jeśli przyjmiemy, że pani nie ma umiejętności rozpoznawania, liczba prawidłowych klasyfikacji herbaty ma rozkład "hipergeometryczny" (`hd = stats.hypergeom (8,4,4 )`).
- Istnieje pięć możliwości: $0, 1, 2, 3, 4$ oraz odpowiadające im prawdopodobieństwa:

Number of correct calls	*Python* command	Probability
0	hd.pmf(0)	1/70
1	hd.pmf(1)	16/70
2	hd.pmf(2)	36/70
3	hd.pmf(3)	16/70
4	hd.pmf(4)	1/70

A Lady Tasting Tea

- Przy powyższych założeniach możemy policzyć p-value hipotezy: *Pani nie potrafi rozpoznać dwóch herbat.*
- Przypomnijmy, że p-value jest prawdopodobieństwem zaobserwowania wyniku ekstremalnego lub bardziej ekstremalnego niż obserwowany przy założeniu hipotezy zerowej.
- W takim przypadku gdy pani odgadnie wszystkie herbaty poprawnie to p-value wynosi $1/70$, a jeśli popełni jeden błąd, to wówczas wartość p-value wynosi $1/70 + 16/70 = 0.24$.

A Lady Tasting Tea

Test opisany w powyższym przykładzie nazywamy Fisher's exact test (dokładnym testem Fishera).

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		
		0	1	*Total*
A	0	a	b	$a + b$
	1	c	d	$c + d$
Total		$a + c$	$b + d$	$N = a + b + c + d$

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/bootcamp/blob/master/statistics/D08_Z06.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	*Total*
Before: present	101	121	*222*
Before: absent	59	33	*92*
Total	*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:

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

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).

McNemar's Test

Używając korekty Yates's otrzymujemy:

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

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

Cochran's Q Test

- Test Cochrana Q jest testem, w którym odpowiedzi mogą przyjmować tylko dwa możliwe wyniki (oznaczone jako 0 i 1).
- Jest to nieparametryczny test statystyczny, który sprawdza, czy k różnych schematów leczenia ma identyczne efekty.

Cochran's Q Test

https://github.com/przem85/bootcamp/blob/master/statistics/D08_Z07.ipynb

Rozważmy przykład, w którym dwanaście osób poproszone zostało o wykonanie trzech zadań. Wynikiem każdego zadania jest sukces lub porażka. Wyniki są kodowane przez 0 dla niepowodzenia i 1 dla sukcesu.

Subject	Task 1	Task 2	Task 3
1	0	1	0
2	1	1	0
3	1	1	1
4	0	0	0
5	1	0	0
6	0	1	1
7	0	0	0
8	1	1	0
9	0	1	0
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0

Czy istnieje różnica w trudności między tymi zadaniami?

Cochran's Q Test

Hipoteza zerowa dla Q-testu mówi, że nie ma różnic między zmiennymi. Jeśli obliczona p-value jest poniżej wybranego poziomu hipoteza zerowa zostaje odrzucona i możemy stwierdzić, że proporcje co najmniej 2 zmiennych są znacząco różne od siebie.

Dla naszych danych statystyka przyjmuje wartość $Q = 8.6667$, a p-value $p = 0.013$. Innymi słowy, co najmniej jedno z trzech zadań jest łatwiejsze lub trudniejsze od pozostałych.

Zadanie

https://github.com/przem85/bootcamp/blob/master/statistics/D08_Z08.ipynb

Zadanie

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

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

Zadanie

https://github.com/przem85/bootcamp/blob/master/statistics/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

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?

Zadanie

https://github.com/przem85/bootcamp/blob/master/statistics/D08_Z10.ipynb

Zadanie (Cz 1.)

W pozwie dotyczącym morderstwa obrona wykorzystuje kwestionariusz do wykazania, że pozwany jest szalony. W wyniku kwestionariusza oskarżony twierdzi, że "nie jest winny z powodu szaleństwa".
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 "szalonym", 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 "szalony" jest uznanych za "szalonych".

Zadanie

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.

Zadanie

https://github.com/przem85/bootcamp/blob/master/statistics/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.

Zadanie

https://github.com/przem85/bootcamp/blob/master/statistics/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 cięższi?
- Czy powyższe dane pochodzą z rozkładów normalnych?

Zadanie

https://github.com/przem85/bootcamp/blob/master/statistics/D08_Z13.ipynb

Zadanie

Pobieraj dane z pliku https://github.com/przem85/statistics/blob/master/D8/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?