Testy zgodności

Martyna Kobielnik

Spis treści

1 Testy zgodności w Statistice

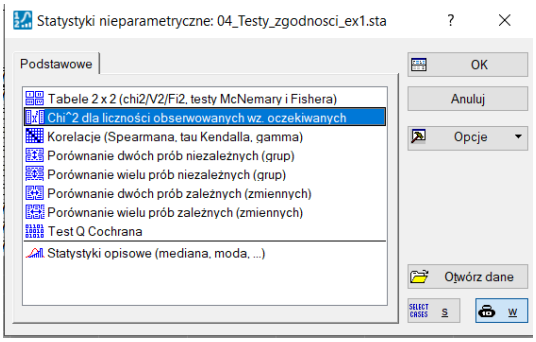
Przy pomocy modułu *Nieparametryczne* dostępnego w zakładce *Statystyka* można wykonać test χ^2 dla dowolnego rozkładu teoretycznego. Potrzebny nam będzie jedynie szereg rozdzielczy i liczebności oczekiwane. Zweryfikujemy hipotezę $H_0 : F(x) = F_0(x)$, gdzie F_0 jest dystrybuantą rozkładu normalnego $N(5, 2)$ przeciwko hipotezie $H_1 : F(x) \neq F_0(x)$. W arkuszu przedstawionym na rys. 1a (*02_ Testy_ zgodnosci_ ex1.sta*) zawarty jest pewien szereg rozdzielczy. Ostatnia kolumna zawiera liczebności oczekiwane, wyznaczone z rozkładu normalnego $N(5, 2)$ wg wzoru

$$np_i = n\left(F_0(a_{i+1}) - F_0(a_i)\right),$$

gdzie a_i i a_{i+1} są lewą i prawą granicą i–tej klasy.

	1 lewy	2 środek	3 prawy	4 liczność empiryczna	5 **liczność oczekiwana**
1	0,5	1	1,5	1	1,11338737
2	1,5	2	2,5	3	2,62362467
3	2,5	3	3,5	5	4,83910315
4	3,5	4	4,5	6	6,98665288
5	4,5	5	5,5	7	7,89650605
6	5,5	6	6,5	7	6,98665288
7	6,5	7	7,5	5	4,83910315
8	7,5	8	8,5	3	2,62362467
9	8,5	9	9,5	2	1,11338737
10	9,5	10	10,5	1	0,369788377

(a) Szereg rozdzielczy



(b) Statystyki nieparametryczne

Rysunek 1: Test χ^2

W oknie z wyborem testów nieparametrycznych przedstawionym na rys. 1b wybieramy *Chi^ 2 dla liczności obserwowanych wz. oczekiwanych*. Wybieramy zmienne zawierające liczebności empiryczne i teoretyczne. W wyniku otrzymamy arkusz przedstawiony na rys. 2. W górnej części

podana jest wartość statystyki testowej oraz wartość p. W tym przypadku nie mamy podstaw do odrzucenia hipotezy mówiącej o tym, że rozkład badanej cechy jest rozkładem normalnym ze średnią $\mu = 5$ i odchyleniem standardowym $\sigma = 2$.

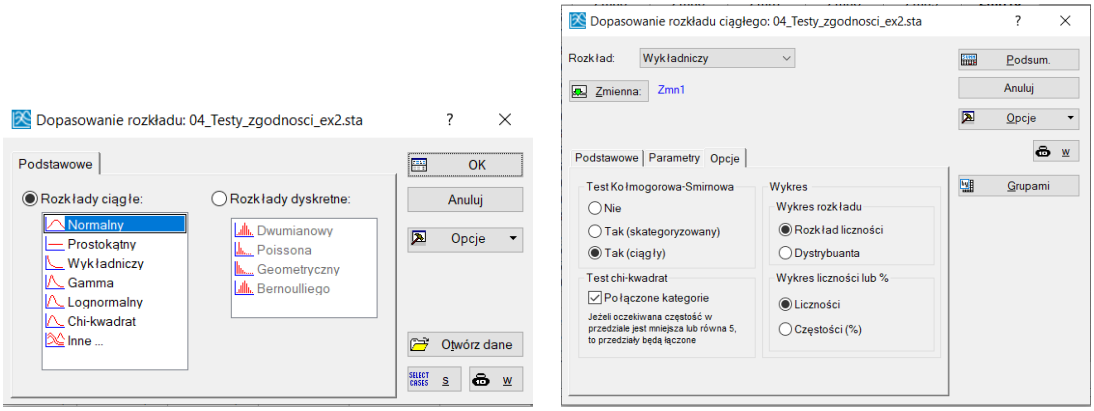
Liczności obserwowane i oczekiwane (04_Testy_zgodnosci_ex1.sta)
Chi kwadrat= 2,151441 df = 9 p = ,988840
UWAGA: Różne sumy oczekiwanych i obserwowanych

Przypadek	obserw. liczność empiryczna	oczekiw. liczność oczekiwana	obs-ocz	(ob-oc)2 /ocz
C: 1	1,00000	1,11339	-0,113387	0,011547
C: 2	3,00000	2,62362	0,376375	0,053993
C: 3	5,00000	4,83910	0,160897	0,005350
C: 4	6,00000	6,98665	-0,986653	0,139335
C: 5	7,00000	7,89651	-0,896506	0,101782
C: 6	7,00000	6,98665	0,013347	0,000025
C: 7	5,00000	4,83910	0,160897	0,005350
C: 8	3,00000	2,62362	0,376375	0,053993
C: 9	2,00000	1,11339	0,886613	0,706027
C: 10	1,00000	0,36979	0,630212	1,074038
Sum	40,00000	39,39183	0,608169	2,151441

Rysunek 2: Wyniki testu χ^2

Dla wybranych rozkładów, test χ^2 można wykonać w module *Dopasowanie rozkładu* dostępnego w zakładce *Statystyka*. Jest tam dostępny również test Kołmogorowa–Smirnowa. Proces przeprowadzenia testów zgodności zostanie zaprezentowany na przykładzie danych zawartych w pliku *02_ Testy_ zgodnosci_ ex2.sta*.

W pierwszej kolejności, w oknie z rys. 3a, należy wybrać rozkład, który chcemy dopasować do danych. Do wyboru mamy rozkłady ciągłe i dyskretne.



(a) Wybór rozkładu (b) Opcje

Rysunek 3: Dopasowanie rozkładu

Dla przykładu, do zmiennych *Zmn1* i *Zmn2* spróbujemy dopasować rozkład wykładniczy. W zakładce *Parametry* należy wybrać parametry rozkładu, który próbujemy dopasować. Automatycznie pola te zostaną uzupełnione parametrami empirycznymi. W zakładce *Opcje* widocznej na rys. 3b mamy możliwość wyboru, czy w razie za małej liczebności w klasie, powinna ona zostać połączona z sąsiednią. Możemy również zdecydować, czy ma zostać wykonany dodatkowo test Kołmogorowa–Smirnowa. Do wyboru mamy wersję ciągłą (opartą na pojedynczych wartościach z próby) i skategoryzowaną (przed przeprowadzeniem testu dane zostaną podzielone na klasy). W

przypadku dużych zbiorów danych ten wybór może znacząco wpłynąć na czas obliczeń. Wynik analizy dla obu zmiennych przedstawiony został na rys. 4.

Zmienna: Zmn1, Rozkład: Wykładniczy (04_Testy_zgodnosci_ex2.sta)
d Kołmogorowa-Smirnowa 0,37629, p < 0,01
Chi-kwadrat = 70,66148, df = 5 (dopasow.) , p = 0,00000

Górna Granica	**Obserw. Liczność**	Skumulow. Obserw.	Procent Obserw.	Skumul. % Obserw.	Oczekiwana Liczność	Skumulow. Oczekiwana	Procent Oczekiwana	Skumul. % Oczekiwana	Obserw. - Oczekiwana
<= 1,00000	0	0	0,00000	0,0000	4,439414	4,43941	8,87883	8,8788	-4,43941
2,00000	0	0	0,00000	0,0000	4,045246	8,48466	8,09049	16,9693	-4,04525
3,00000	1	1	2,00000	2,0000	3,686076	12,17074	7,37215	24,3415	-2,68608
4,00000	1	2	2,00000	4,0000	3,358795	15,52953	6,71759	31,0591	-2,35880
5,00000	0	2	0,00000	4,0000	3,060574	18,59011	6,12115	37,1802	-3,06057
6,00000	1	3	2,00000	6,0000	2,788831	21,37894	5,57766	42,7579	-1,78883
7,00000	4	7	8,00000	14,0000	2,541215	23,92015	5,08243	47,8403	1,45878
8,00000	5	12	10,00000	24,0000	2,315585	26,23574	4,63117	52,4715	2,68441
9,00000	5	17	10,00000	34,0000	2,109988	28,34572	4,21998	56,6914	2,89001
10,00000	5	22	10,00000	44,0000	1,922646	30,26837	3,84529	60,5367	3,07735
11,00000	3	25	6,00000	50,0000	1,751938	32,02031	3,50388	64,0406	1,24806
12,00000	4	29	8,00000	58,0000	1,596386	33,61669	3,19277	67,2334	2,40361
13,00000	6	35	12,00000	70,0000	1,454646	35,07134	2,90929	70,1427	4,54535
14,00000	6	41	12,00000	82,0000	1,325490	36,39683	2,65098	72,7937	4,67451
15,00000	5	46	10,00000	92,0000	1,207802	37,60463	2,41560	75,2093	3,79220
16,00000	1	47	2,00000	94,0000	1,100563	38,70520	2,20113	77,4104	-0,10056
17,00000	1	48	2,00000	96,0000	1,002846	39,70804	2,00569	79,4161	-0,00285
18,00000	2	50	4,00000	100,0000	0,913805	40,62185	1,82761	81,2437	1,08619
<nieskończoność	0	50	0,00000	100,0000	9,378153	50,00000	18,75631	100,0000	-9,37815

(a) Wyniki dla *Zmn1*

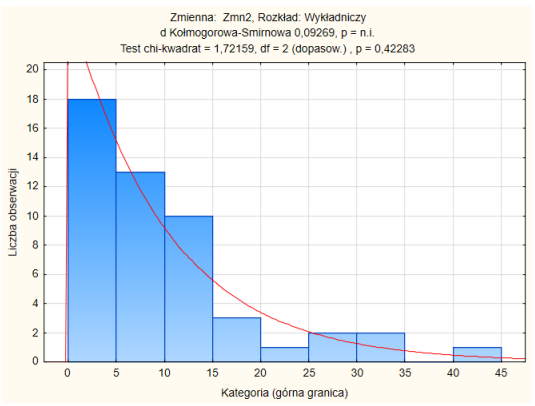
Zmienna: Zmn2, Rozkład: Wykładniczy (04_Testy_zgodnosci_ex2.sta)
d Kołmogorowa-Smirnowa 0,09269, p = n.i.
Chi-kwadrat = 1,72159, df = 2 (dopasow.) , p = 0,42283

Górna Granica	**Obserw. Liczność**	Skumulow. Obserw.	Procent Obserw.	Skumul. % Obserw.	Oczekiwana Liczność	Skumulow. Oczekiwana	Procent Oczekiwana	Skumul. % Oczekiwana	Obserw. - Oczekiwana
<= 5,00000	18	18	36,00000	36,0000	19,60185	19,60185	39,20370	39,2037	-1,60185
10,00000	13	31	26,00000	62,0000	11,91720	31,51905	23,83440	63,0381	1,08280
15,00000	10	41	20,00000	82,0000	7,24522	38,76426	14,49043	77,5285	2,75478
20,00000	3	44	6,00000	88,0000	4,40482	43,16909	8,80965	86,3382	-1,40482
25,00000	1	45	2,00000	90,0000	2,67797	45,84706	5,35594	91,6941	-1,67797
30,00000	2	47	4,00000	94,0000	1,62811	47,47516	3,25621	94,9503	0,37189
35,00000	2	49	4,00000	98,0000	0,98983	48,46499	1,97966	96,9300	1,01017
40,00000	0	49	0,00000	98,0000	0,60178	49,06677	1,20356	98,1335	-0,60178
<nieskończoność	1	50	2,00000	100,0000	0,93323	50,00000	1,86645	100,0000	0,06677

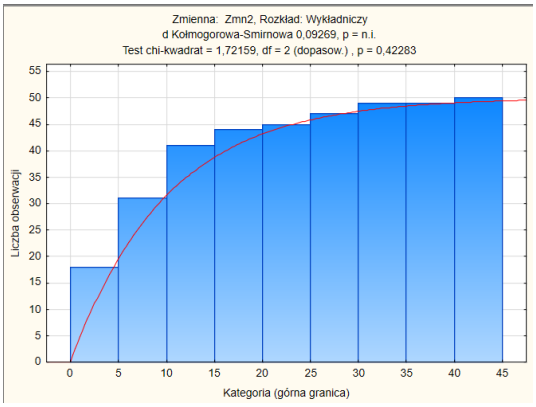
(b) Wyniki dla *Zmn2*

Rysunek 4: Wyniki testów χ^2 i Kołmogorowa–Smirnowa

Na rys. 5 przedstawione są wykresy liczebności i dystrybuanty z dopasowanymi funkcjami teoretycznymi dla *Zmn1*.



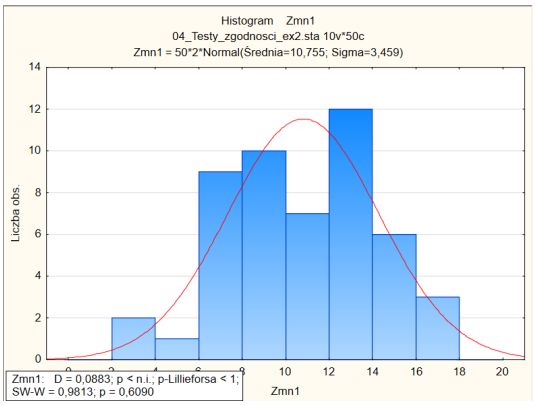
(a) Wykres liczności (histogram)



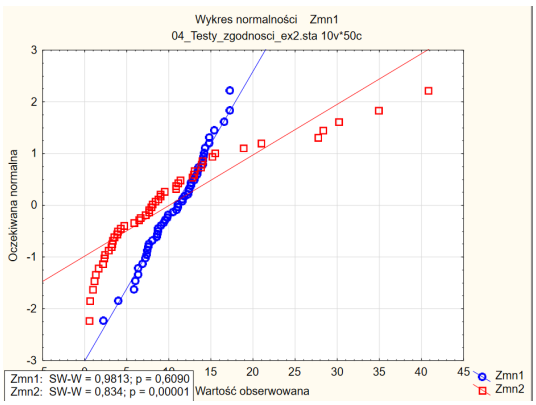
(b) Dystrybuanta

Rysunek 5: Wykresy liczności i dystrybuanty

Testy zgodności można wykonać również przy okazji tworzenia niektórych wykresów. Tworząc histogram, możemy wybrać jeden z rozkładów. Wtedy, krzywa gęstości tego rozkładu zostanie naniesiona na wykres. Jeśli wybrany został rozkład normalny, to dodatkowo możemy wykonać test Kołmogorowa–Smirnowa lub test Shapiro–Wilka. Histogram zawierający wyniki obu testów dla zmiennej *Zmn1* znajduje się na rys. 6a.



(a) Histogram



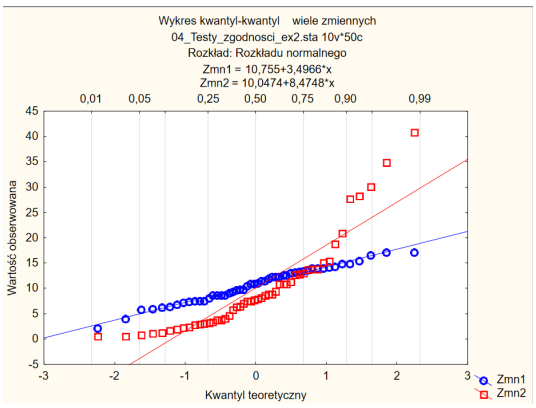
(b) Wykres normalności

Rysunek 6: Histogram i wykres normalności

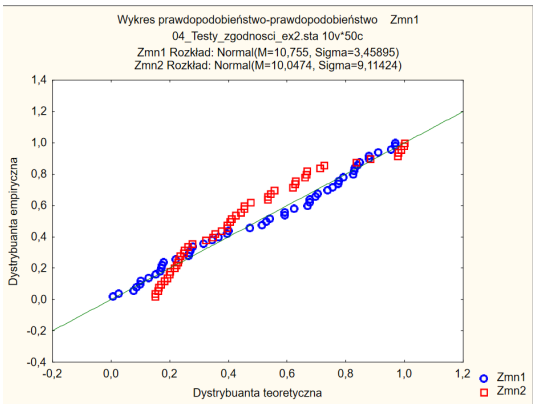
Wybierając *Wykres normalności* możliwe jest wykonanie przy okazji testu Shapiro–Wilka. Taki wykres dla zmiennych *Zmn1* i *Zmn2* przedstawiony jest na rys. 6b. Skala dla wykresu dobrana jest tak, aby rozkładowi normalnemu odpowiadała linia prosta. Próba jest układana w szereg wariacyjny, wartości są rangowane, a następnie wyznacza się oczekiwane wartości dla poszczególnych rang przy założeniu, że rozkład jest normalny. Tak powstałe punkty, jeśli założenie o normalności jest spełnione, powinny tworzyć linię prostą. Im bliżej wykreślonej prostej znajdują się punkty,

tym większe wskazanie na normalność rozkładu. Na wykresie widzimy, że dla zmiennej *Zmn2* odchylenia od prostej są o wiele większe niż dla *Zmn1*. W lewym dolnym rogu widoczne są wyniki testu Shapiro–Wilka, które potwierdzają wcześniejszą obserwację.

Poza wymienionymi wyżej możliwościami, możliwe jest utworzenie wykresów pomocnych przy ocenie, czy próba pochodzi z populacji o określonym rozkładzie. Są to wykres kwantyl-kwantyl i wykres prawdopodobieństwo-prawdopodobieństwo. Oba wykresy przedstawione zostały na rys. 7. Na ich podstawie możemy dokonać wizualnej oceny dopasowania.



(a) Wykres kwantyl-kwantyl



(b) Wykres prawdopodobieństwo-prawdopodobieństwo

Rysunek 7: Wykresy

Wykres kwantyl-kwantyl tworzony jest poprzez wykreślenie punktów, gdzie jedna ze współrzędnych jest kwantylem teoretycznym, a druga, kwantylem empirycznym wyznaczonym na podstawie próby. Jeśli rzeczywiście rozkład empiryczny jest zgodny z teoretycznym, punkty te powinny układać się w linii prostej, która jest prostą regresji (temat regresji liniowej omawiany będzie pod koniec semestru).

Wykres prawdopodobieństwo-prawdopodobieństwo tworzą punkty, których współrzędnymi są $(F_n(x_i), F(x_i))$, gdzie F_n to dystrybuanta empiryczna wyznaczona na podstawie próby n-elementowej, a F to dystrybuanta teoretyczna.

2 Zadania

Raport z zadań należy zapisać w pliku 02_imie_nazwisko.pdf i umieścić w odpowiednim miejscu na PZE. Należy w nim uwzględnić wszystkie wyniki pośrednie wraz z ich krótkim omówieniem.

1. Na poziomie istotności $\alpha = 0.07$ zweryfikuj hipotezę mówiącą o tym, że rozkład zmiennej, z której próba znajduje się w arkuszu 02_Testy_zgodnosci_zad1.sta jest rozkładem wykładniczym. Parametr μ rozkładu oszacuj na podstawie wyników z próby. Pamiętaj o uwzględnieniu wag przypadków.

2. Na podstawie arkusza 02_Testy_zgodnosci_zad2.sta [1] sprawdź, czy pierwsza zmienna może pochodzić z rozkładu jednostajnego a druga z rozkładu geometrycznego. Parametry należy oszacować na podstawie dostępnych prób. Za poziom istotności przyjmij $\alpha = 0.1$. Dla obu zmiennych utwórz histogram z naniesioną gęstością dopasowywanego rozkładu oraz wykres dystrybuanty empirycznej z teoretyczną.

3. Sprawdź, które ze zmiennych w arkuszu 02_Testy_zgodnosci_zad2.sta mogą pochodzić z rozkładu normalnego. Odpowiedź uzasadnij i zwizualizuj przy pomocy odpowiednich wykresów.

[1] Przed przystąpieniem do rozwiązywania zadania należy przejść do zakładki *Dane*, wybrać opcję *Przelicz* i przeliczyć wartości wszystkich zmiennych