

Poprawki i transformacje

Doświadczalnictwo

Autor Wanda Olech
Katedra Genetyki i Ochrony Zwierząt

Transformowanie

Przekształcenie liniowe elementów próby
w celu zbliżenia rozkładu do rozkładu
normalnego

Rodzaje transformacji:

Logarytmiczna (cechy o rozkładzie
geometrycznym)

Blissa (cechy wyrażane w procentach)

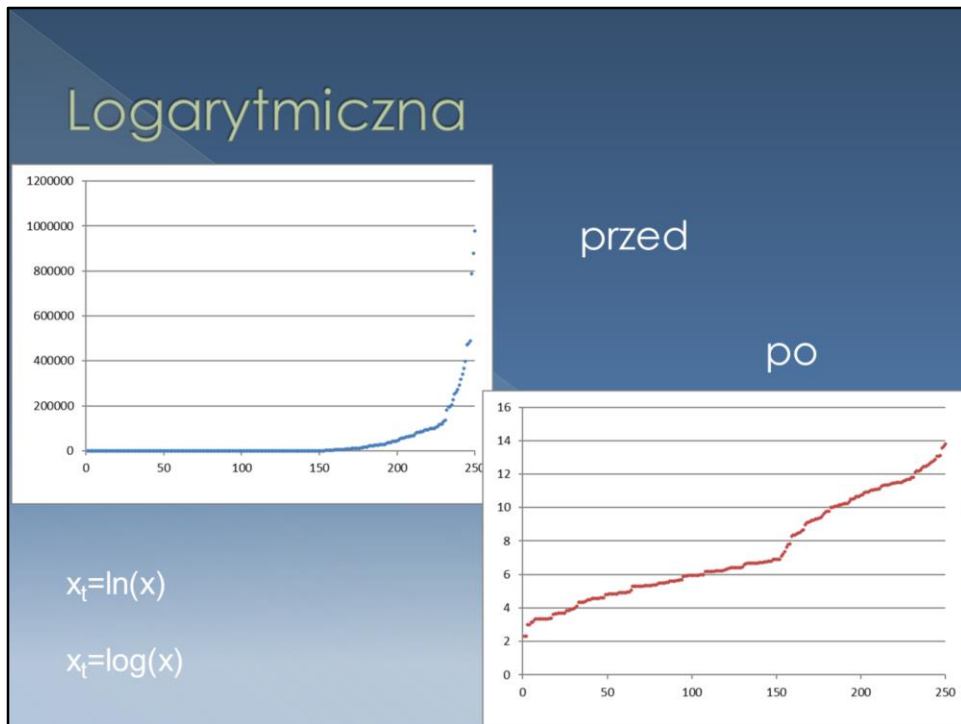
probitowa (cechy skokowe)

Bardzo często się zdarza, że nawet z powodów praktycznych, chcemy liczby przekształcić. Możemy dodawać liczbę, odejmować, mnożyć czy dzielić przez liczbę, obliczać potęgę liczby, logarytmować, obliczać wartość funkcji trygonometrycznej. Możemy też łączyć ze sobą takie przekształcenia. Przekształcenie może być liniowe lub nie, co w tym wypadku oznacza, coś innego niż funkcja liniowa. Liniowe przekształcenie to takie które nie zmienia kolejności liczb. Mamy trzy liczby : 2; -5; 4 – zapisane w kolejności malejącej to 4; 2; -5. Jeśli dodamy do każdej z nich liczbę 10 to uzyskamy: 14; 12; 5. Kolejność nie została zmieniona zatem przekształcenie jest liniowe. Podobnie jak pomnożymy przez 3, to w wyniku mamy 12; 6; -15. Ale jeśli podniesiemy do kwadratu to uzyskamy 16; 4; 25 i mamy zmienioną kolejność czyli takie przekształcenie nie spełnia definicji liniowego.

Drugim warunkiem postawionym transformacji to dążenie (zbliżanie się) do rozkładu normalnego. Jeżeli obydwa warunki są spełnione to przekształcenie nazywamy transformacją, inaczej transformujemy elementy próby.

Są trzy sytuacje w których zastosowanie transformacji jest bardzo potrzebne a same transformacje nazwane tutaj są stosowane. Wtedy gdy cecha ma rozkład geometryczny, lub wyrażana jest w procentach i rozkłada się odpowiednio na skali procentowej, oraz cechy skokowe, które z definicji nie mogą mieć rozkładu normalnego.

A po co robić transformację próby. Jeżeli jej celem jest zbliżenie się do rozkładu normalnego to zaczynamy mieć w swojej gestii narzędzia dla cech taki rozkład mających. A ponieważ narzędzia dla cech o innych rozkładach są ograniczone i często bardzo słabe to może warto.



Cechy mające rozkład geometryczny – to liczba bakterii w jednostce powierzchni czy objętości, liczba elementów komórkowych i wiele innych. Czasami cecha w całej populacji takiego rozkładu nie ma ale w losowo pobranej próbie mamy bardzo duży rozrzut danych. Jak poznać, że mamy do czynienia z cechą o rozkładzie geometrycznym na następnym slajdzie. Transformacja takich cech polega na obliczeniu logarytmu o dowolnej podstawie, czasami może to dodatkowo złożenie sumy i/lub mnożenie przez liczbę. Na wykresach mamy dane próby przed i po obliczeniu logarytmu. Zmienia się rozrzut wartości, liczby poprzednio odbiegające dziesiątki razy różnią się kilkukrotnie. I o to chodzi w przekształceniu logarytmicznym, aby zmniejszyć skalę zakresu wartości.

Logarytmiczna

średnia	55	223	640	16534	205492
odchylenie	31	75	184	13217	211112
Wskaźnik zmienności	0,566	0,337	0,287	0,799	1,027

średnia	3,84	5,35	6,42	9,26	11,87
odchylenie	0,62	0,34	0,29	1,10	0,80
Wskaźnik zmienności	0,163	0,063	0,045	0,119	0,067

Na tym obrazie jest przykład próby zawierającej wartości wymagające transformacji logarytmicznej. Próba została podzielona na pięć części o jednakowej liczebności w ten sposób, że w pierwszej części znalazły się liczby do kwantyla 20%, potem kolejna partia do kwantyla 40% itd. W każdej części obliczono średnią arytmetyczną, odchylenie standardowe i wskaźnik zmienności. Na górze dane pierwotne. Można zauważyć bardzo różniące się od siebie wartości odchyleń standardowych a na wet wskaźników zmienności. Proporcja pomiędzy ostatnim a pierwszym odchyleniem standardowym wynosi około 70 tysięcy – jedno jest większe tyle razy od drugiego. Takie różnice w parametrach zmienności na sztucznie podzielonej próbie są ewidentnym symptomem geometrycznego rozkładu cechy. Po obliczeniu logarytmu te różnice są bardzo wyraźnie zmniejszone.

Pytanie: jeśli mamy próbę dla cechy która z założenia ma rozkład normalny np. masa ciała, jakie są różnice między parametrami zmienności po podzieleniu próba na kilka części. Proszę sprawdzić, ale proszę pamiętać że próba powinna być liczna, co najmniej kilkadziesiąt danych.

Logarytmiczna

x	log(x)
10	1
100	2
10000	4
100000	5
100	2
10000	4
1000	3
1000	3

To tabela przypominające obliczane logarytmu dziesiętnego. Proszę zauważyć, że logarytm średniej geometrycznej danych jest tożsamy ze średnią arytmetyczną logarytmów. Proszę sprawdzić.

Zatem po transformacji nasze parametry do interpretacji jeśli potrzebne mogą być odwrócone – czyli obliczymy anty logarytm średniej – ale wtedy mamy średnią geometryczną. Antylogarytm to podnoszenie do potęgi równej średniej podstawy logarytmu.

Jeśli ktoś wykonuje prace np. dotyczące liczby elementów komórkowych i do analizy danych musi transformować, to potem bardzo często chce odtworzyć wartość aby interpretacja była bardziej zrozumiała. Trzeba pamiętać o tej zasadzie.

I jeszcze jedno – jeżeli cecha ma rozkład geometryczny to musimy poddać ją transformacji. Średnia arytmetyczna dla takiej cechy bez transformacji nie ma żadnego znaczenia. Tylko średnia geometryczna, czyli odwrócenie średniej arytmetycznej logarytmów.

Blissa

Cechy wyrażone w procentach, jeśli ponad 80% wartości znajduje się na początku lub na końcu skali procentowej, tj. od 0 do 20 lub od 80 do 100%

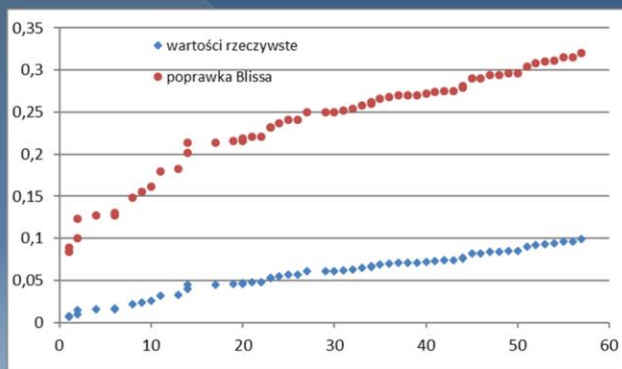
$$x_t = \arcsin \sqrt{x}$$

Cechy wyrażane w procentach, np. odsetek śmiertelności czy przeżywalności, czy inne cechy progowe podlegają transformacji Blissa. Są to cechy, których większość wartości to wartości w przedziale 0-20% lub przedziale 80-100%. Ta większość to około 4/5 wartości jeśli jest na dole lub na górze skali procentowej przekształcenie musi być stosowane.

Transformacja Blissa to podstawienie do wzoru jak na slajdzie.

Pytanie czy to przekształcenie jest liniowe? Proszę sprawdzić na przykładowych danych.

Blissa



średnia	0,0585	0,2367
odchylenie	0,0255	0,0625
V	0,4364	0,2639

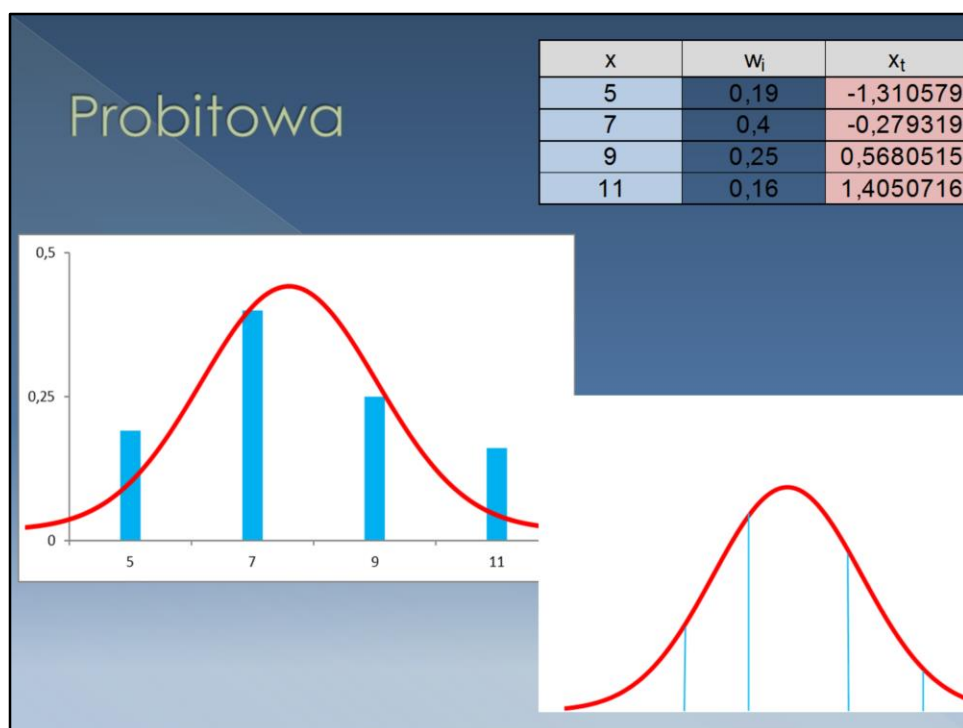
Tu pokazany jest efekt przekształcenia Blissa. Liczby są uporządkowane według swojej wartości. Efektem przekształcenia jest rozciągnięcie skali, o czym świadczą parametry. Po przekształceniu średnia jest większa podobnie jak odchylenie standardowe. Takie przekształcenie „rozciąga” ściśnięte na końcu skali procentowej dane i pozwala na zbliżenie do rozkładu normalnego.

A jak wyglądałby ten wykres gdyby dane były w przedziale 80-100%. Może warto wygenerować takie dane i przekształcić.

Probitowa

Cechy skokowe – transformacja polega na zmianie wartości cechy na zgodną z rozkładem normalnym standaryzowanym

Ostatnia transformacja dotyczy cech skokowych – szczególnie takich które mają niewiele powtarzających się wartości. Najmniej możemy mieć dwie wartości – czyli z reguły liczbami 0 oraz 1, ale wartości może być znacznie więcej. Transformacja nosi nazwę probitowej.



Ta cecha ma cztery wartości i przy pomocy prostokątów narysowany jest jej rozkład. Na wykresie powinny być linie albo punkty aby nie zapomnieć o rodzaju cechy, ale taki obraz jest czytelniejszy. Każda z wartości ma swoją częstość empiryczną, czyli zajmuje fragment rozkładu. Jeżeli podzielimy rozkład normalny na takie fragmenty, o wielkości 0,19; 0,40; 0,25; 0,16. To będzie to mniej więcej tak wyglądać jak na obrazie z prawej. Teraz wystarczy obliczyć średnią wartości dla każdego fragmentu rozkładu normalnego standaryzowanego i przekształcenie gotowe.

Nie dam Państwu teraz szczegółowego rozwiązania – będzie w kolejnym wykładzie. Postaram się pokazać to obliczanie w miarę przejrzystość w wykładzie o cechach progowych.

Wyniki są przedstawione w tabeli – czyli wartość przed i po transformacji.

Poprawki

Przekształcenie liniowe danych w celu usunięcia czynnika

Innym przekształceniem (ale również liniowym) są poprawki, czyli takie przekształcenie, które zmierza do usunięcia wpływu czynnika.

Jeśli średnie w każdej grupie będą takie same to nie będzie różnic między grupami czyli nie będzie możliwości odrzucenia hipotezy o równości wielu średnich – hipotezy w analizie wariancji. Czynników może być kilka i czasem warto pozbyć się jednego z nich, aby „wyekstrahować” pozostałe.

Mamy poprawki addytywne (dodajemy lub odejmujemy liczbę) oraz mnożnikowe (mnożymy lub dzielimy przez liczbę).

Poprawki addytywne

B1	B2	B3	B4	B5
1	3	6	6	3
2	7	5	8	3
3	5	7	9	2
4	6	8	2	1
3	9	4	4	3
6	7	6	7	4
7	8	5	6	5
8	9	3	5	6
4,25	6,75	5,50	5,88	3,38

$$x_{p_{ij}} = x_{ij} + (\bar{x}_w - \bar{x}_i)$$

B1	B2	B3	B4	B5
2,25	1,75	6,00	5,63	5,13
3,25	5,75	5,00	7,63	5,13
4,25	3,75	7,00	8,63	4,13
5,25	4,75	8,00	1,63	3,13
4,25	7,75	4,00	3,63	5,13
7,25	5,75	6,00	6,63	6,13
8,25	6,75	5,00	5,63	7,13
9,25	7,75	3,00	4,63	8,13
5,5	5,5	5,5	5,5	5,5

odchylenie	2,5	2,1	1,6	2,2	1,6
standardowe	2,5	2,1	1,6	2,2	1,6

Przykład poprawiania – poprawki addytywne. Jeden czynnik podzielił populację na pięć subpopulacji, wybraliśmy z każdej reprezentatywną próbę i mamy nasze dane po lewej. Jeśli jedna z grup stanie się swoistym wzorcem to poprawki polegają na dodaniu liczby do danych z innej grupy aby średnia była taka sama. Wybrana została grupa B3 i do niej się dopasowujemy, jej średnia wynosi 5,5. Grupa B1 ma średnia niższą o 1,25 i taka liczba musi być dodana do wszystkich wartości tej grupy. Po poprawieniu wszędzie jest taka sama średnia – nie ma różnic.

Warto zwrócić uwagę, że addytywna poprawka nie zmienia odległości między danymi i odchylenie standardowe pozostaje niezmienione.

Poprawki mnożnikowe

B1	B2	B3	B4	B5
1	3	6	6	3
2	7	5	8	3
3	5	7	9	2
4	6	8	2	1
3	9	4	4	3
6	7	6	7	4
7	8	5	6	5
8	9	3	5	6
4,25	6,75	5,50	5,88	3,38

$$x_{pij} = x_{ij} \cdot (\bar{x}_w / \bar{x}_i)$$

B1	B2	B3	B4	B5
1,29	2,44	6	5,62	4,89
2,59	5,7	5	7,49	4,89
3,88	4,07	7	8,43	3,26
5,18	4,89	8	1,87	1,63
3,88	7,33	4	3,74	4,89
7,76	5,7	6	6,55	6,52
9,06	6,52	5	5,62	8,15
10,4	7,33	3	4,68	9,78
5,5	5,5	5,5	5,5	5,5

odchylenie	2,5	2,1	1,6	2,2	1,6
standardowe	3,2	1,7	1,6	2,1	2,6

Przykład poprawiania – poprawki mnożnikowe. Jeden czynnik podzielił populację na pięć subpopulacji, wybraliśmy z każdej reprezentatywną próbę i mamy nasze dane po lewej. Jeśli jedna z grup stanie się swoistym wzorcem to poprawki polegają na mnożeniu przez liczbę danych z innej grupy aby średnia była taka sama. Wybrana została grupa B3 i do niej się dopasowujemy, jej średnia wynosi 5,5. Grupa B1 ma średnią niższą zatem wartości te grupy muszą zostać pomnożone przez iloraz 5,5/4,25 czyli liczbę większą od jeden równą ok. 1,29. Po poprawieniu wszędzie jest taka sama średnia – nie ma różnic.

Warto zwrócić uwagę, że mnożnikowa poprawka zmienia odległości między danymi i odchylenie standardowe zmienia się zgodnie z wielkością poprawki dla grupy.

Poprawki

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	201	5,16		
m.poz.B	4	57,3	14,3	3,49	0,0168
błąd	35	144	4,11		

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	144	3,69		
m.poz.B	4	0	0	0	1,0000
błąd	35	144	4,11		

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	188	4,83		
m.poz.B	4	0	0	0	1,0000
błąd	35	188	5,38		

Tutaj są analizy wariancji przed poprawieniem i po zastosowaniu poprawki addytywnej i mnożnikowej.

Czy są Państwo w stanie powiedzieć po której poprawce są tabele analiz wariancji?

Układ wieloczynnikowy - poprawki addytywne

	B1	B2	B3	B4	B5
A1	1	3	6	6	3
	2	7	5	8	3
	3	5	7	9	2
	4	6	8	2	1
A2	3	9	4	4	3
	6	7	6	7	4
	7	8	5	6	5
	8	9	3	5	6

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	201	5,16		
m.poz.A	1	14,4	14,4	5,27	0,0289
m.poz.B	4	57,3	14,3	5,25	0,0025
interakcja	4	47,4	11,8	4,33	0,0070
błąd	30	82	2,73		

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	144	3,69		
m.poz.A	1	14,4	14,4	5,27	0,0289
m.poz.B	4	0	0	0	1,0000
interakcja	4	47,3	11,8	4,33	0,0070
błąd	30	82	2,73		

Poprawianie to usuwanie czynnika – tutaj w modelu dwuczynnikowym wykonujemy poprawkę addytywną dla jednego z czynników. Pozbywamy się tego czynnika ale pozostaje interakcja z nim, zatem nie powinno się usuwać czynnika, jeżeli interakcja jest istotna !!!!

Układ wieloczynnikowy - poprawki mnożnikowe

	B1	B2	B3	B4	B5
A1	1	3	6	6	3
	2	7	5	8	3
	3	5	7	9	2
	4	6	8	2	1
A2	3	9	4	4	3
	6	7	6	7	4
	7	8	5	6	5
	8	9	3	5	6

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	201	5,16		
m.poz.A	1	14,4	14,4	5,27	0,0289
m.poz.B	4	57,3	14,3	5,25	0,0025
interakcja	4	47,4	11,8	4,33	0,0070
błąd	30	82	2,73		

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	188	4,83		
m.poz.A	1	25,2	25,2	7,59	0,0099
m.poz.B	4	0	0	0	1,0000
interakcj	4	63,6	15,9	4,79	0,0041
błąd	30	99,6	3,32		

Podobnie co się dzieje z drugim czynnikiem i interakcja po zastosowaniu poprawki mnożnikowej.

Podsumowując tak jak transformacje są potrzebne i powinny być stosowane to poprawki są pewnym archaizmem. Kiedyś wieloczynnikowe układy były zbyt skomplikowane, ale teraz mają Państwo narzędzia które pozwalają na ich analizę.

Zadanie – proszę sprawdzić co się dzieje w układach wieloczynnikowych jeżeli przekształcimy dane wejściowe poprzez poprawianie mające na celu usunięcie czynnika.

Układ wieloczynnikowy - poprawki addytywne

	B1	B2	B3	B4	B5
A1	1	2	2	6	3
	2	4	5	7	3
	3	5	3	9	2
	2	6	4	2	1
A2	3	9	8	11	7
	6	7	6	13	4
	7	8	7	6	8
	8	9	9	10	6

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	346	8,86		
m.poz.A	1	160	160	47,5	0,0000
m.poz.B	4	84,6	21,2	6,28	0,0009
interakcja	4	0	0	0	1,0000
błąd	30	101	3,37		

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	261	6,69		
m.poz.A	1	160	160	47,5	0,0000
m.poz.B	4	0	0	0	1,0000
interakcja	4	0	0	0	1,0000
błąd	30	101	3,37		

Brak interakcji - regresja

Cecha zależna (analizowana)					Cecha niezależna										
B1	B2	B3	B4	B5	B1	B2	B3	B4	B5						
1	3	6	6	3	-1,8	-1,1	0,1	0,2	-0,9						
2	7	5	8	3	-1,4	0,3	-0,2	0,7	-1,2						
3	5	7	9	2	-1,0	-0,3	0,2	0,9	-1,4						
4	6	8	2	1	-0,7	0,1	0,8	-1,5	-1,8						
3	9	4	4	3	-1,0	1,0	-0,8	-0,7	-1,0						
6	7	6	7	4	0,0	0,4	0,0	0,4	-0,1						
7	8	5	6	5	0,3	0,6	-0,3	0,1	-0,4						
8	9	3	5	6	0,7	1,0	-1,1	-0,4	0,1						
											LSS	SKO	ŚKO	F _{emp}	p
										ogólna	39	201	5,16		
										m.poz.B	4	57,3	14,3	2,99	0,0343
										błąd	30	144	4,79		

reszty					
B1	B2	B3	B4	B5	
0,12	0,16	-0,22	-0,50	-0,41	
0,00	0,22	-0,37	0,10	0,44	
-0,12	-0,09	0,50	0,54	0,00	
0,03	-0,22	-0,18	0,28	0,12	
-0,12	0,25	0,31	0,03	-0,12	
0,07	-0,06	0,07	-0,06	-1,74	
0,22	0,38	-0,09	-0,22	0,19	
0,10	0,25	0,16	0,19	-0,22	

	LSS	SKO	ŚKO	F _{emp}	p
ogólna	39	5,32	0,14		
m.poz.B	4	0,51	0,13	0,79	0,5421
błąd	30	4,82	0,16		

I ostatnia rzecz. Wydaje się że poprawka ale to jest całkowicie inny problem. Mamy dwie cechy niezależna i zależna skorelowane ze sobą i ta druga podlega analizie. W modelu konieczne jest uwzględnienie regresji, czy jak w SPSS współzmiennnej. Ale jeśli nie ma możliwości albo jest jakaś inna przyczyna to można przeprowadzić analizę cechy zależnej na resztach z regresji. Tu jest przykład jak zmienia to wyniki. Proszę sprawdzić na swoich przykładach. To taka swoista prosta poprawka jeśli nie ma możliwości innej analizy.

