Statystyka i modelowanie w naukach o środowisku

Wykład 11

Transformacja danych

Rozkład χ^2

Normalizacja

Dla rozkładów normalnych opracowano dużą liczbę testów pozwalających na wyciąganie wniosków co do prawdziwości różnych hipotez.

Co robić, gdy założenia o normalności nie są spełnione?

Można zastosować test nieparametryczny lub wykonać transformację danych.

Transformacja danych

Przekształcenie danych mających rozkład inny niż normalny do rozkładu normalnego.

Często stosowane przekształcenia (transformacje) zmiennej x:

Arc sin X

Transformacja Boxa-Coxa $\frac{x^{\lambda}-1}{\lambda}$

logarytmowanie, potęgowanie, pierwiastkowanie itp. $\sqrt{x}$

Arc sin (tzw. transformacja Blissa) – stosujemy zazwyczaj dla danych mających rozkład dwumianowy wyrażonych w procentach, przyjmujących najczęściej wartości w przedziale (0-20% lub 80-100%)

Transformacja Boxa-Coxa – jest to często stosowana transformacja, w przypadku rozkładów asymetrycznych (lewostronnie lub prawostronnie skośnych lub też „uciętych" rozkładów normalnych)

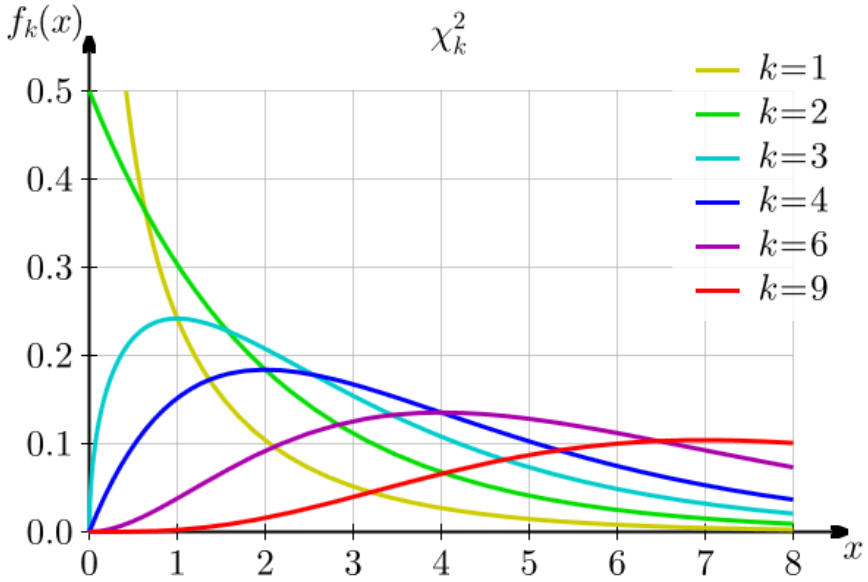
Logarytmowanie – Stosujemy zazwyczaj w przypadku, gdy wraz ze wzrostem wartości średniej zwiększa się wariancja (a tym samym odchylenie standardowe), czyli występuje korelacja między średnią a wariancją. Stosowanie transformacji log(x) może nie być możliwe, np. w takim przypadku jeśli zmienna przyjmuje wartość 0, wtedy można zastosować transformację log(x+1).

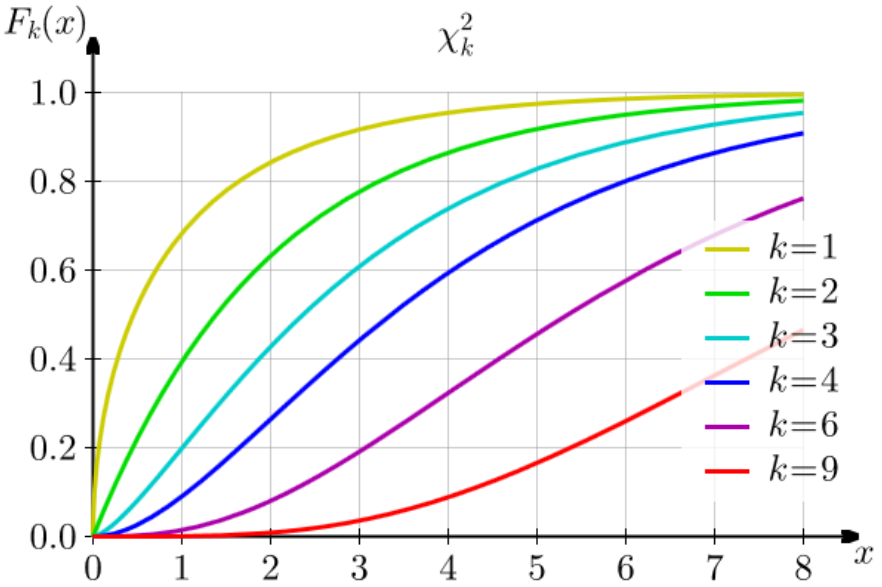
Pierwiastkowanie – stosujemy w przypadku rozkładów zbliżonych do rozkładu Poissona, tzn. w rozkładach prawostronnie skośnych, w których wartość średnia jest zbliżona do wariancji. Podobnie jak w przypadku transformacji log(x) może występować problem, jeśli zmienna przyjmuje wartość 0 (lub wartości ujemne).

Transformacja danych - problemy

1) Brak możliwości transformowania niektórych rozkładów do rozkładu normalnego, np. nie da się przekształcić zmiennej skokowej do zmiennej ciągłej, tak więc w przypadku jeśli zmienna jest zmienną skokową (dyskretną), która przyjmuje niewielką liczbę wartości (np. 1, 2, 3, 4 i 5) to niemożliwe jest zastosowanie transformacji, tak aby rozkład tej zmiennej był rozkładem normalnym
2) Trudności w interpretacji wyników. Ze względu na to, że po transformacji wartości parametrów (np. wartość średnia) ulegają zmianie, to nie można wnioskować np.. O procentowej różnicy między średnimi na podstawie parametrów obliczonych na zmiennej transformowanej.

Rozkład χ^2

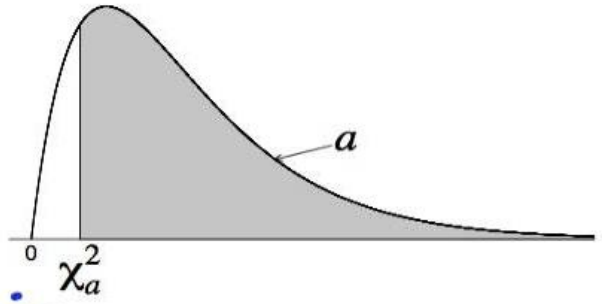




The chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.

http://en.wikipedia.org/wiki/Chi-squared_distribution

Tablica rozkładu χ^2



df	$\chi^2_{0.9995}$	$\chi^2_{0.999}$	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.95}$	$\chi^2_{0.90}$	$\chi^2_{0.85}$	$\chi^2_{0.80}$
1	0.000	0.000	0.000	0.000	0.001	0.004	0.016	0.036	0.064
2	0.001	0.002	0.010	0.020	0.051	0.103	0.211	0.325	0.446
3	0.015	0.024	0.072	0.115	0.216	0.352	0.584	0.798	1.005
4	0.064	0.091	0.207	0.297	0.484	0.711	1.064	1.366	1.649
5	0.158	0.210	0.412	0.554	0.831	1.145	1.610	1.994	2.343
6	0.299	0.381	0.676	0.872	1.237	1.635	2.204	2.661	3.070
7	0.485	0.598	0.989	1.239	1.690	2.167	2.833	3.358	3.822
8	0.710	0.857	1.344	1.646	2.180	2.733	3.490	4.078	4.594
9	0.972	1.152	1.735	2.088	2.700	3.325	4.168	4.817	5.380
10	1.265	1.479	2.156	2.558	3.247	3.940	4.865	5.570	6.179

https://www.youtube.com/watch?v=HwD7ekD5l0g

Test zgodności χ^2

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Test zgodności χ^2 - przykład

Petrograf bada mikroskopowo cienką próbkę skały magmowej. Jego zadaniem jest nadanie badanej skale poprawnej nazwy. Stosując specjalny aparat sprzężony z mikroskopem zlicza 100 kryształów występujących w próbce. Z literatury wiadomo, że w granicie stosunek 4 głównych minerałów ma się do siebie tak jak 4:1:2:3. Czy badana próbka, w której stosunek odpowiednich kryształów jest 35:12:22:31 może być nazwana granitem?

Test zgodności χ^2 - przykład

H_0 – rozkłady są zgodne

	100		
	E	O	$(O-E)^2/E$
4	40	35	0.625
1	10	12	0.4
2	20	22	0.2
3	30	31	0.033333
			$\sum$
			1.258333

χ^2 - ze wzoru wynosi 1,258

p=ROZKŁAD.CHI(1.258;3)=0,739

$p>\alpha$

H_0 przyjmujemy, można powiedzieć, że rozkłady są zgodne

Test zgodności χ^2 - przykład

H_0 – rozkłady są zgodne

Statistica →Nieparametryczne→Chi^2 dla liczności obserwowanych wz. oczekiwanych→ wskazać, które zmienne są oczekiwane (E), a które obserwowane (O)

	Liczności obserwowane i oczekiwane (Arkusz1) Chi kwadrat= 1.258333 df = 3 p = .739050			
ypadek	**obserw. O**	oczekiw. E	obs-ocz	(ob-oc)2 /ocz
1	35.0000	40.0000	-5.00000	0.625000
2	12.0000	10.0000	2.00000	0.400000
3	22.0000	20.0000	2.00000	0.200000
4	31.0000	30.0000	1.00000	0.033333
	100.0000	100.0000	0.00000	1.258333

p=0,739

p>α

H_0 przyjmujemy, można powiedzieć, że rozkłady są zgodne

Test niezależności χ^2 - przykład

Badano strukturę lasu na pięciu stanowiskach. Określono liczbę gatunków występujących we wszystkich trzech piętrach lasu. Sprawdź czy bogactwo gatunkowe tych lasów zależy od piętra.

		stanowisko				
		A	B	C	D	E
warstwa	runo	24	33	52	7	64
	podszyt	22	11	6	16	23
	sklepienie	5	7	6	11	4

https://www.socscistatistics.com/tests/chisquare2/default2.aspx

Test niezależności χ^2 - przykład

https://www.socscistatistics.com/tests/chisquare2/default2.aspx

H_0 zmienne są niezależne lub bogactwo gatunkowe tych lasów nie zależy od piętra

The chi-square statistic is 52.0032. The p-value is < 0.00001. The result is significant at p < .05.

H_0 odrzucamy

zmienne są zależne lub bogactwo gatunkowe tych lasów zależy od piętra