Asymetrické rozdělení, transformace dat

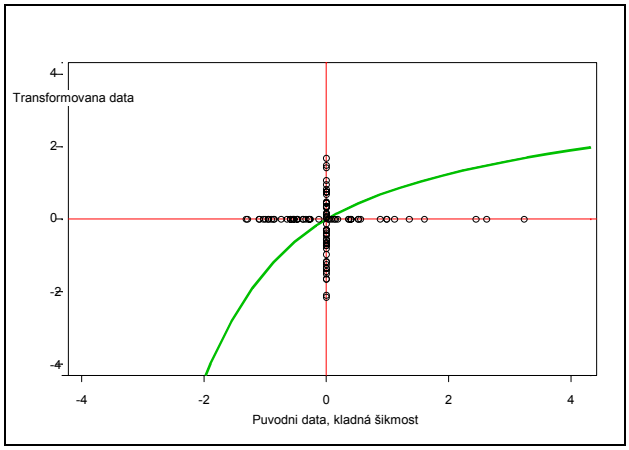
Asymetrie rozdělení je běžným jevem při měření kvantit blízkých mezi detekce přístroje, některých velmi malých veličin (stopové koncentrace, znečištění, úroveň hluku), životnost strojů, velikosti malých částic, některých fyzikálních veličin determinovaných mezními vlastnostmi materiálu jako pevnost, tvrdost, a podobně. Při vyhodnocení nelze v takovém případě použít postupů založených na normálním rozdělení, jako aritmetický průměr, pravidlo 3 sigma, Shewhartovy regulační diagramy, nebo metoda nejmenších čtverců.

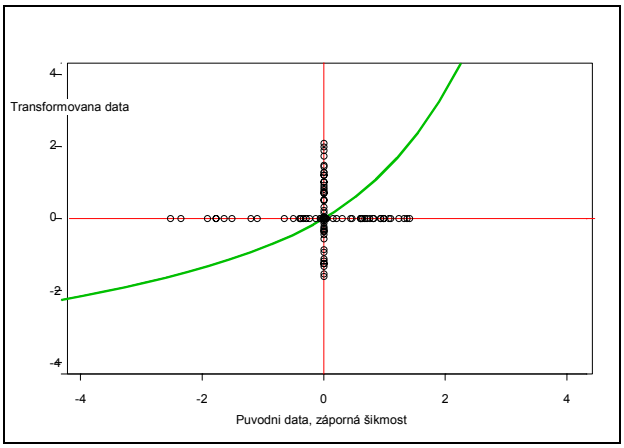
Poměrně jednoduchá technika nelineární transformace umožní i pro asymetricky rozdělená data použít klasických metod. Cílem transformace je nalézt funkci *x*' původních hodnot *x*, která zajistí minimální šikmost, případně maximální věrohodnost transformovaných dat vzhledem k normálnímu rozdělení. Takovou funkcí může být například exponenciála:

$$x' = \frac{\log(r.x+1)}{r} \quad pro\ r<0\ a\ x<0\ nebo\ r>0\ a\ x>0$$

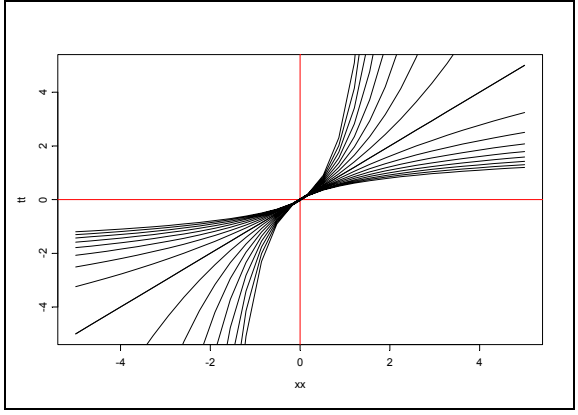
$$x' = -\frac{\exp(-r.x)-1}{r} \quad pro\ r<0\ a\ x>0\ nebo\ r>0\ a\ x<0$$

která se aplikuje na standardizovaná data: $x_i^s = \frac{x_i - \bar{x}}{s_x}$. Parametr r se zvolí tak, aby se rozdělení transformovaných dat co nejvíce blížilo symetrickému, resp. normálnímu. Jako kritérium se volí šikmost nebo věrohodnost. Pro data s kladnou, případně zápornou šikmostí je tvar transformační funkce znázorněn na dvou následujících obrázcích.

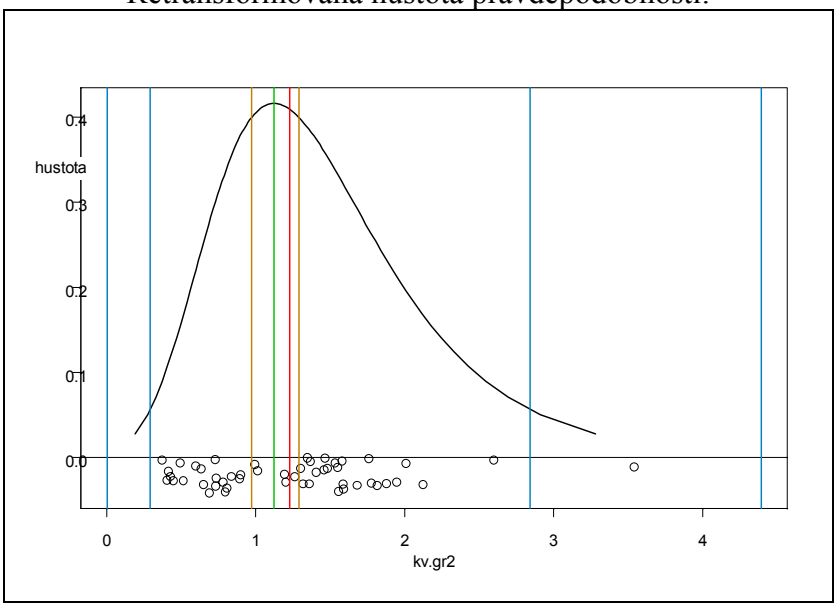




Průběhy transformační funkce pro různá *r*, *r* = 0 odpovídá přímka *y* = *x*

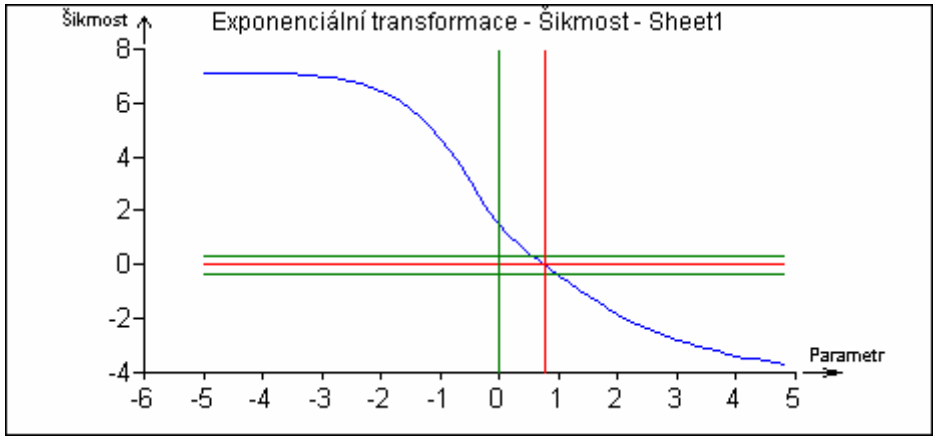


Retransformovaná hustota pravděpodobnosti:



Prostý průměr, retransformovaný průměr s intervalem spolehlivosti, retransformované kvantily: ±2σ, ±3σ (2.5%, 0.135%)

Pro určení optimálního parametru transformace *r* lze využít podmínky nulové šikmosti. Leží-li hodnota *r*, která zajišťuje nulovou šikmost mimo interval spolehlivosti nulové šikmosti, lze považovat transformaci za odůvodněnou (viz graf).



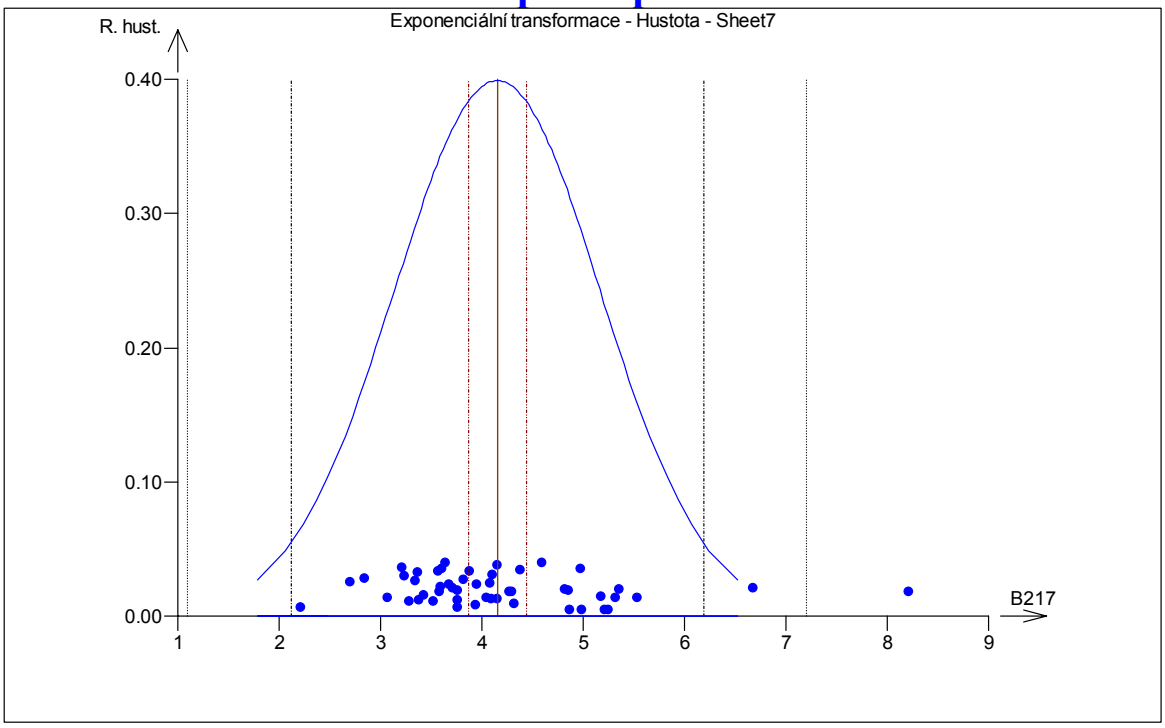
Literatura:

[1] M. Meloun, J. Militký: Chemometrics for Analytical Chemistry: Part 1, Ellis Horwood, 1992
[2] Shewhart, W.A.: Statistical Method from the Viewpoint of Quality Control, Dover Pubns, 1987
[3] Ryan, P. Statistical Methods for Quality Improvement, J. Wiley, 1994
[4] Montgomery D. C.: Introduction to Statistical Quality Control, Chapman and Hall, 1990
[5] Mittag, Rinne: Statistical Methods for Quality Assurance, Chapman and Hall, 1993
[6] Becker, Chambers, Wilks: The New S Language, Chapman and Hall, 1996
[7] Box G. E. P., Cox D. R.: An analysis of transformations. Journal of the Royal Statistical Society, Series B 26(2) 1964: 211-243
[8] M. Meloun, J. Militký: Statsitické zpracování experimentálních dat, VIP, Praha 1998

Příklad:

Data: Stanovený fibrilogen v krvi.
Úkolem je odhadnout důležité populační kvantily (0.1%, 1%, 5%, 25%, 75%, 95%, 99%,99.9%) na základě statistického modelu. Jako statistický model jsou srovnány normální rozdělení a transformované normální rozdělení (exponenciální transformace).

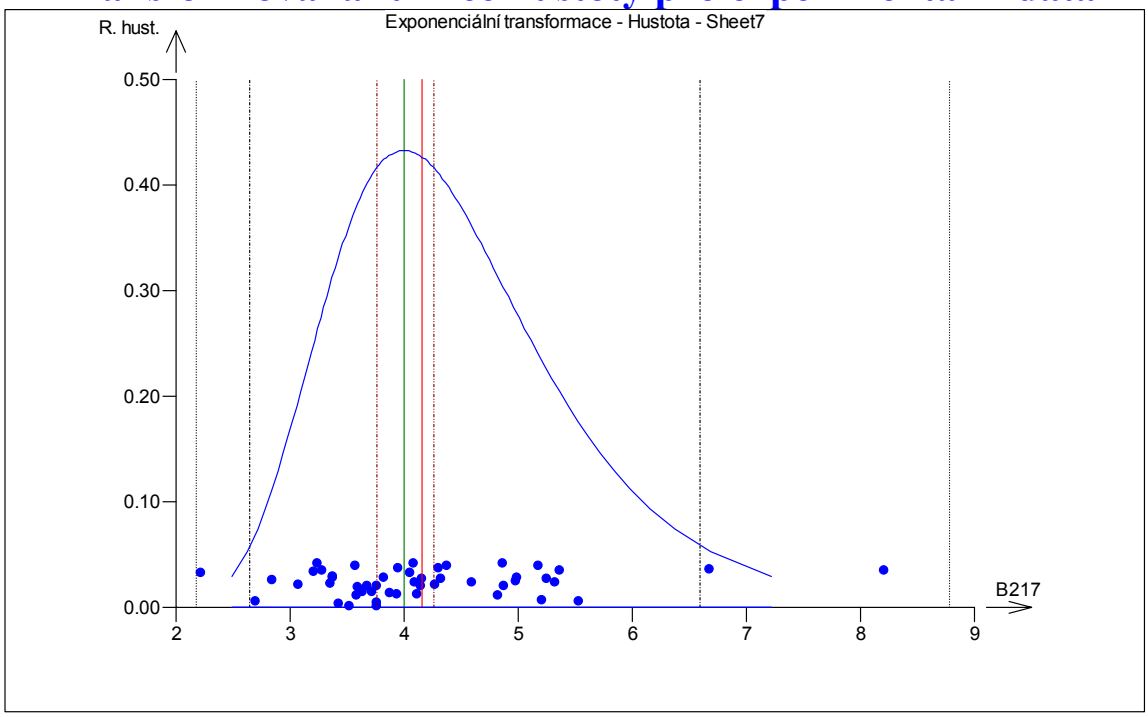
Gaussiva křivka pro experimentální data



Výsledky

Exponenciální transformace dat :			
Název úlohy :	Sheet7		
Optimální parametr :	0.392578125		
Zvolený parametr :	**0 (bez transformace)**		
Oprávněnost transformace :	Ano		
Opravený průměr :	4.1526		
Interval spolehlivosti :			
Spodní :	3.863218426		
Horní :	4.441981574		
Významné opravené kvantily	p	spodní	horní
	50 %	4.1526	4.1526
	25 %	3.465805261	4.839394739
	20 %	3.295624741	5.009575259
	15 %	3.097258551	5.207941449
	12.5 %	2.981264341	5.323935659
	10 %	2.847668574	5.457531426
	7.5 %	2.686806569	5.618393431
	6.25 %	2.590491864	5.714708136
	5 %	2.477738629	5.827461371
	3 %	2.237494296	6.067705704
	2.5 %	2.156879584	6.148320416
	2 %	2.061383703	6.243816297
	1.5 %	1.942919775	6.362280225
	1 %	1.783811615	6.521388385
	0.5 %	1.529778804	6.775421196
	0.25 %	1.294356389	7.010843611
	0.125 %	1.074102532	7.231097468
	0.1 %	1.005991348	7.299208652
	0.03 %	0.6583812751	7.646818725

Transformovaná funkce hustoty pro experimentální data



Výsledky

1. Předpoklad normálního rozdělení			
Název úlohy :	Fibrilogen		
Optimální parametr :	0.392578125		
Zvolený parametr :	0.392578125		
Oprávněnost transformace :	Ano		
Opravený průměr :	3.999691458		
Interval spolehlivosti :			
Spodní :	3.759769975		
Horní :	4.264055091		
Významné opravené kvantily	p	spodní	horní
	50 %	3.999691458	3.999691458
	25 %	3.462737264	4.674233838
	20 %	3.345255906	4.868380354
	15 %	3.21471294	5.110004461
	12.5 %	3.141315718	5.259386738
	10 %	3.059276917	5.439240068
	7.5 %	2.963818629	5.667437958
	6.25 %	2.908301206	5.810442904
	5 %	2.844783104	5.984191979
	3 %	2.714419889	6.378448695
	2.5 %	2.672104293	6.518463476
	2 %	2.622854633	6.689597003
	1.5 %	2.563031375	6.91011922
	1 %	2.48479456	7.221287932
	0.5 %	2.36457602	7.75603929
	0.25 %	2.257930002	8.296600561
	0.125 %	2.161975689	8.844947006
	0.1 %	2.133007363	9.023387657
	0.03 %	1.990003721	10.00446021

Závěr

	O D H A D			
p	**Normální**	**Transformovaný**	**Normální**	**Transformovaný**
25 %	3.4658	3.4627	4.8393	4.6742
5 %	2.4777	2.8447	5.8274	5.9841
1 %	1.7838	2.4847	6.5213	7.2212
0.1 %	1.0059	2.1330	7.2992	9.0233