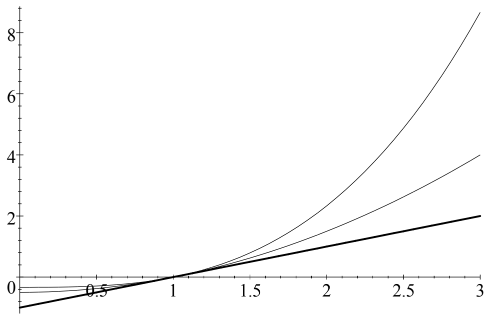
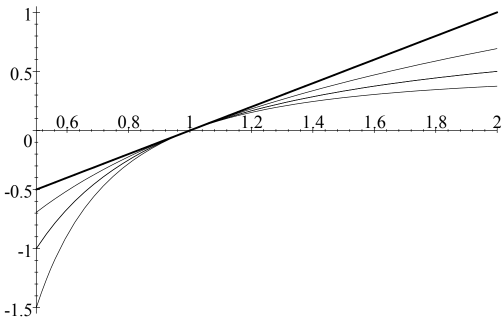
Rodzina przekształceń Boxa-Coxa (dla $x > 0$)

$$h_p(x) = \begin{cases} \frac{x^p-1}{p} & p \neq 0 \\ \ln x & p = 0 \end{cases}$$



Box-Cox dla p=1,2,3



Box-Cox dla p=1,0,-1,-2

$$h_p'(x) = x^{p-1}$$

Zadanie. *Znaleźć p takie, że* $\gamma(\alpha, h_p(x)) \simeq 0$ *dla wszystkich* $\alpha < 0.5$

Niech $x_{(n+1-k)} > x_{(k)}$ dla $k \leq \frac{n+1}{2}$
Dla $\alpha_k = \frac{2k-1}{2n+1}$ kwantyl $q(\alpha_k) = x_{(k)}$
Dla n nieparzystych

$$\gamma(\alpha_k, h_p(x)) = \begin{cases} \frac{x^p_{(n+1-k)}+x^p_{(k)}-2*x^p_{(r)}}{x^p_{(n+1-k)}-x^p_{(k)}} & p \neq 0 \\ \frac{\ln(x_{(n+1-k)})+\ln(x_{(k)})-2*\ln(x_{(r)})}{\ln(x_{(n+1-k)})-\ln(x_{(k)})} & p = 0 \end{cases}, r = \frac{n+1}{2}$$

Dla n parzystych, dane uzupełniamy o dodatkowy wyraz, równy $\widetilde{x} = \frac{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}}{2}$.
Niech

$$u_k^- \stackrel{df}{=} \frac{\widetilde{x}}{x_{(k)}} > 1, u_k^+ \stackrel{df}{=} \frac{x_{(n+1-k)}}{\widetilde{x}} > 1,$$

Wtedy dla ustalonego k przyjmując na chwilę dla uproszczenia oznaczenie $u_k^- = u_-, u_k^+ = u_+$

$$f(p) \stackrel{df}{=} \gamma(\alpha_k, h_p(x)) = \begin{cases} \frac{u_+^p+u_-^{-p}-2}{u_+^p-u_-^{-p}} & p \neq 0 \\ \frac{\ln u_+ - \ln u_-}{\ln u_+ + \ln u_-} & p = 0 \end{cases}$$

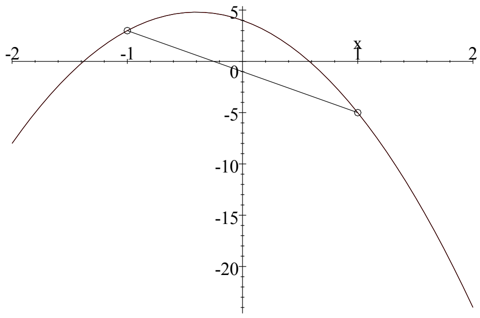
oraz $f(p)$ jest ciągłą funkcją p.
Zadanie sprowadza się do rozwiązania równania $f(p) = 0$.
Zadanie to ma rozwiązanie, gdyż

$$\lim_{p\to\infty} f(p) = \lim_{p\to\infty} \frac{u_+^p - 2}{u_+^p} = 1,$$

$$\lim_{p\to-\infty} f(p) = \lim_{p\to\infty} \frac{u_-^{-p} - 2}{-u_-^{-p}} = -1$$

Prosta, przechodząca przez punkty $(-1, f(-1))$ i $(1, f(1))$ ma równanie

$$\frac{f(1)-f(-1)}{2}(p-1)+f(1)$$



Z metody siecznych przybliżone rozwiązanie równania

$$f(p)=0$$

jest rozwiązaniem równania

$$\frac{f(1)-f(-1)}{2}(p-1)+f(1)=0$$

skąd

$$\begin{aligned}
p &\simeq \frac{f(-1)+f(1)}{f(-1)-f(1)} = \\
&= \frac{\frac{u_+^{-1}+u_--2}{u_+^{-1}-u_-}+\frac{u_++u_-^{-1}-2}{u_+-u_-^{-1}}}{\frac{u_+^{-1}+u_--2}{u_+^{-1}-u_-}-\frac{u_++u_-^{-1}-2}{u_+-u_-^{-1}}} = \frac{\frac{1+u_+u_--2u_+}{1-u_-u_+}+\frac{u_+u_-+1-2u_-}{u_+u_--1}}{\frac{1+u_+u_--2u_+}{1-u_-u_+}-\frac{u_+u_-+1-2u_-}{u_+u_--1}} \\
&= \frac{(1+u_+u_--2u_+)-(u_+u_-+1-2u_-)}{(1+u_+u_--2u_+)+(u_+u_-+1-2u_-)} \\
&= \frac{2(u_--u_+)}{2(1+u_+u_--u_+-u_-)} \\
&= \frac{1}{u_+-1}-\frac{1}{u_--1}
\end{aligned}$$

Przykład 1.

k	1	2	3	4	5	6	7
$x_{(k)}$	1	4	9	16	25	36	49

Dane są kwadratami kolejnych liczb naturalnych.

k	1	2	3
u_-	$\frac{16}{1}=16$	$\frac{16}{4}=4$	$\frac{16}{9}=1.78$
u_+	$\frac{49}{16}=3.06$	$\frac{36}{16}=2.25$	$\frac{25}{16}=1.56$
p	$\frac{1}{3.06-1}-\frac{1}{16-1}=0.42$	$\frac{1}{2.25-1}-\frac{1}{4-1}=0.47$	$\frac{1}{1.56-1}-\frac{1}{1.78-1}=0.50$

Wartości p są w pobliżu 1/2. Pierwiastek symetryzuje te dane - co było jasne od początku.

Przykład 2.

Odległości planet od Słońca w jednostkach astronomicznych (*odległość Ziemi*=1)

1	2	3	4	5	6	7	8	9
0.4	0.7	1.0	1.5	5.2	9.5	19.2	30.0	39.4

	1	2	3	4	5
u_-	13.0	7.43	5.20	3.47	
u_+	7.58	5.77	3.69	1.83	
p	0.07	0.05	0.13	0.80	

Mediana wartości p, równa 0.10 jest najlepszym oszacowaniem p. Wybierając najbliższą wartosc całkowitą otrzymać można skalę, symetryzującą dane o planetach. Jest to skala logarytmiczna. ■

Poniższy diagram ilustruje tezę, że gdy dane dadzą się zsymetryzować, to najlepszą miarą typowości jest mediana

$$\begin{array}{ccc} x & \xrightarrow{h_p} & y \\ \downarrow & & \downarrow \\ \widetilde{x} & \xleftarrow{h_p^{-1}} & \widetilde{y} \end{array}$$