TRANSFORMACE DAT

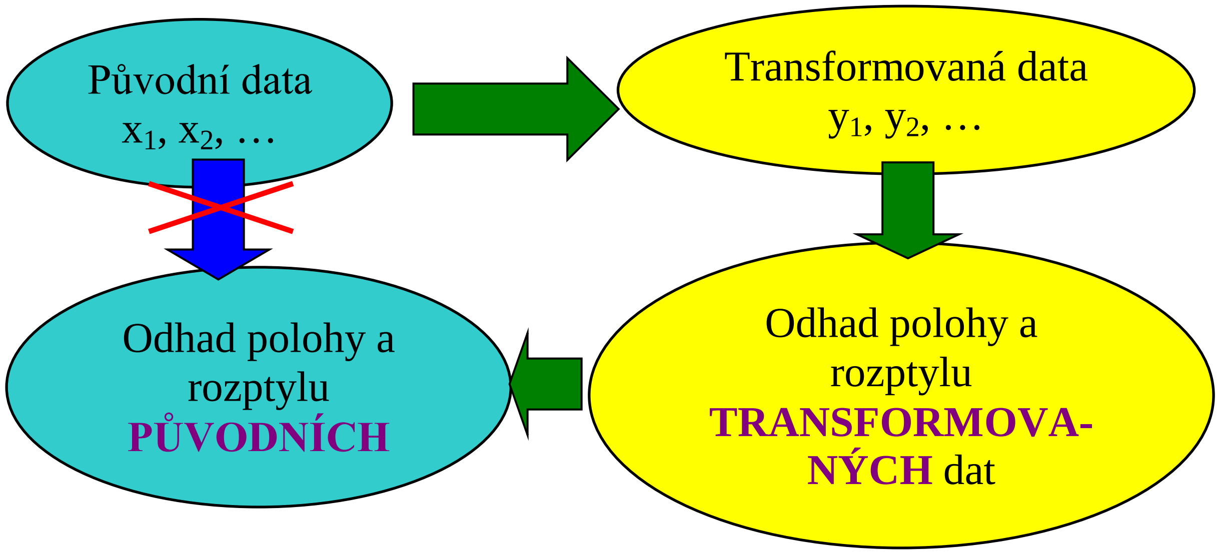
Pokud v EDA zjistíme, že rozdělení se příliš odlišuje od normálního (asymetrie, odlehlé body, nehomogenita), vzniká problém, jak data vůbec vyhodnotit. Často lze použít pro vyhodnocení dat jejich transformaci.

Je vyhledána vhodná transformace, která zajistí největší přiblížení normalitě, tato transformace se provede, vypočte se průměr a jeho IS. Vypočtené údaje se přepočítají do původních souřadnic.

- <u>Důvody pro transformaci:</u> vede ke stabilizaci rozptylu, zesymetričtění rozdělení, je robustní k odlehlým bodům.
- **Obvykle nastává jedna z následujících 3 situací:**

  1. Zamítnuta normalita v ZP, nejsou OB $\Rightarrow$ TRANSFORMACE
  2. Přijata/zamítnuta normalita v ZP, nalezeny OB, které nelze vyloučit $\Rightarrow$ TRANSFORMACE
  3. Přijata/zamítnuta normalita v ZP, nalezeny OB, které lze vyloučit $\Rightarrow$ VYLOUČENÍ OB.

PRINCIP:



(PROSTÁ) MOCNINNÁ TRANSFORMACE

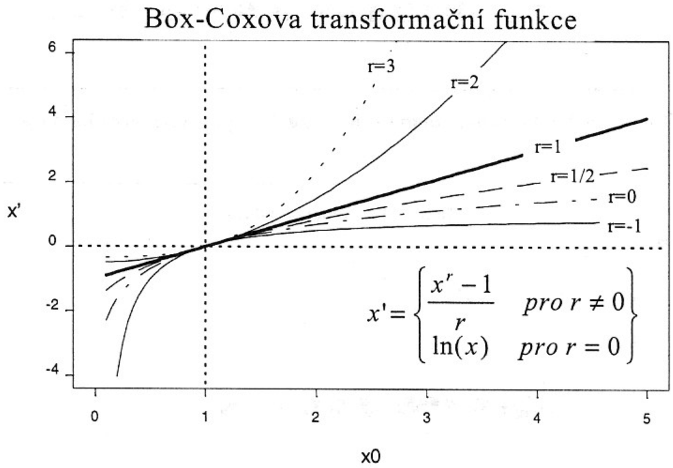
$$y = g(x)\begin{cases} x^{\lambda} & \lambda > 0 \\ -x^{-\lambda} & \lambda < 0 \\ \ln x & \lambda = 0 \end{cases}$$

- Symetrizující transformace.
- Optimální odhad se hledá minimalizací asymetrie (klasického či robustního koeficientu šikmosti).
- Pro $\lambda = 0$ jde o logaritmickou transformaci.

BOX-COXOVA TRANSFORMACE

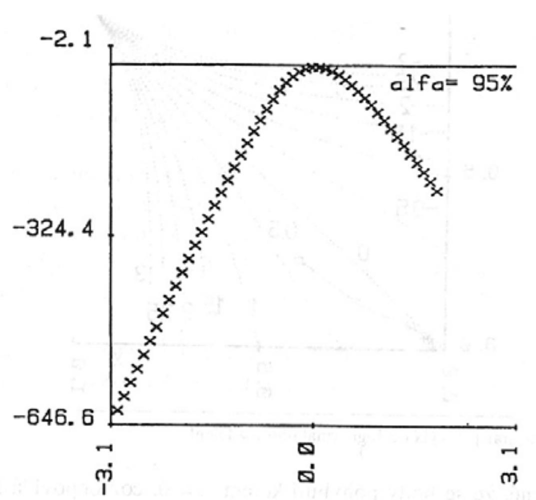
$$y = g(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases}$$

- Přibližuje rozdělení výběru k normálnímu vzhledem k šikmosti a špičatosti.
- Takto definovaná transformace je použitelná pouze pro kladná data. Tvar transformační funkce pro některé parametry uvádí následující obrázek:



Posouzení (statistické) výhodnosti transformace:
- Q-Q graf (pro data před a po transformaci),
- graf logaritmu věrohodnostní funkce (osa y) na parametru λ (osa x).

Maximum = optimální parametr λ. Vodorovná přímka odpovídá 95 % IS maxima věrohodnosti a svislé přímky odpovídají IS odhadu λ $<\lambda_D, \lambda_H>$. Obsahuje-li tento interval +1, není nutné transformovat (transformace není přínosná).



- Čím bude IS širší, tím je transformace méně významná.
- Logaritmus věrohodnostní funkce lnL se používá ve tvaru:

$$\ln L(\lambda) = -\frac{n}{2}\ln s^2(y) + (\lambda - 1)\sum_{i=1}^{n} \ln x_i$$

ZPĚTNÁ TRANSFORMACE (RETRANSFORMACE)

Pokud se podaří nalézt vhodnou transformaci, která vede k přibližné normalitě, lze určit průměr a rozptyl transformovaných dat. Pak je zapotřebí provést zpětnou transformaci na původní proměnnou (data), protože požadujeme odhady parametrů původních dat a ne transformovaných.

Lze použít dva přístupy:

- ***nekorektní (naivní)***: provedení prosté zpětné transformace

$$\bar{x}_R = g^{-1}(\bar{y})$$

- ***korektní***: vychází se z Taylorova rozvoje funkce y = g(x) v okolí $\bar{y}$.

DALŠÍ TRANSFORMACE

- EXPONENCIÁLNÍ – založena na minimalizaci asymetrie.
- MODIFIKOVANÁ MOCNINNÁ TRANSFORMACE – eliminace špičatosti u symetrického rozdělení; nejprve se od dat odečte medián a následně se aplikuje „Box-Coxova transformace“:

$$y = g(x) = SIGN\frac{(|x - \tilde{x}_{0,5}| + 1)^{\lambda} - 1}{\lambda},$$

kde SIGN má význam znaménka, než se aplikuje absolutní hodnota.

- TRANSFORMACE STABILIZUJÍCÍ ROZPTYL – vyžaduje nalezení transformace y = g(x), ve které je $s^2(y)$ konstantní:

$$y = g(x) \approx C\int \frac{dx}{\sqrt{f(x)}} \quad kde\ C = kons.\ ze\ vztahu \quad C = \left[\frac{d\,g(x)}{dx}\right]^2 f(x)$$