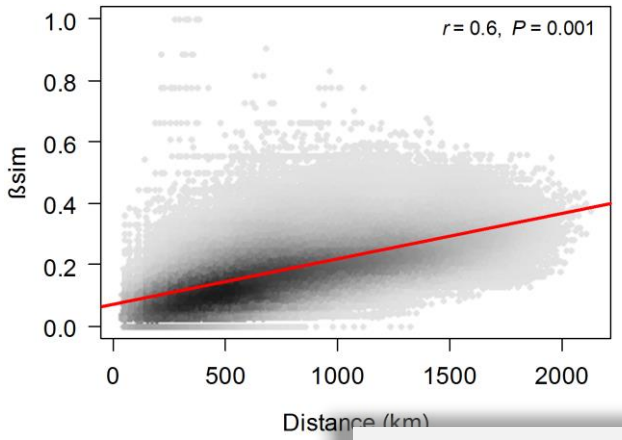
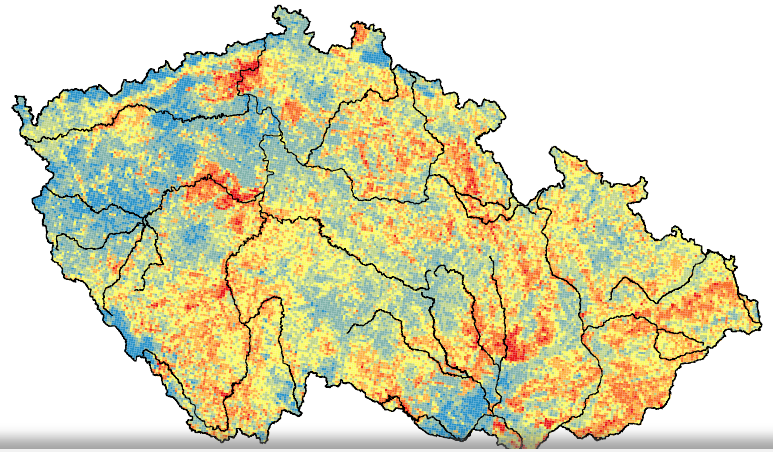
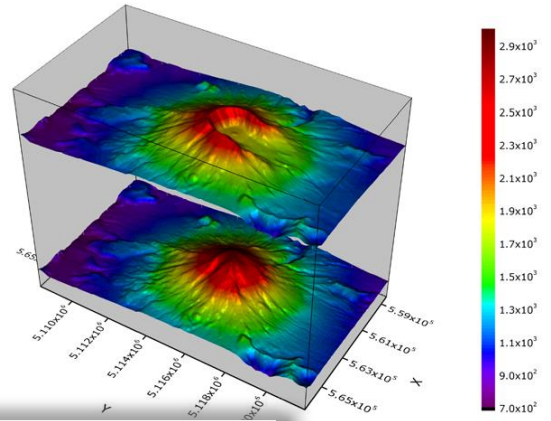
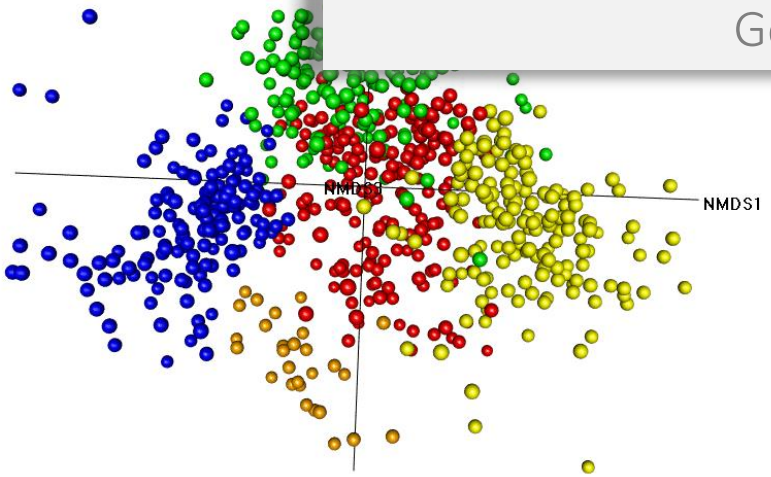


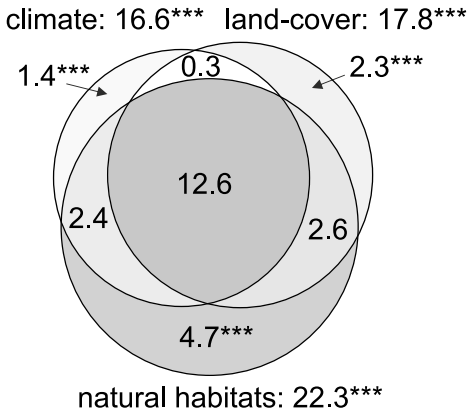




Metody fyzické geografie 3:
Biogeografie & ekologie

Jan Divíšek
Geografický ústav & Ústav botaniky a zoologie





Explorativní analýza, transformace a standardizace dat

Data v biogeografii a ekologii

- **Vysvětlovaná proměnná (*Dependent variable*(*s*))**
  - Distribuce druhů (přítomnost/nepřítomnost), abundance, složení společenstva, vlastnosti druhů atp.
  - společenstvo je typicky sledováno na určité ploše (v případě rostlin a některých málo mobilních živočichů) nebo např. inventarizací jedinců (např. ulovených v pastech v případě mobilních živočichů)
  - složení živého společenstva je popsáno přítomností jednotlivých druhů daného typu organismů, na jedné ploše (v jedné pasti) se většinou vyskytuje více než jeden druh
- **Vysvětlující proměnná (*Explanatory variable*(*s*))**
  - Environmentální faktory, vzdálenosti, fylogenetická podobnost atp.
  - Prostředí je popisováno jednou nebo více proměnnými, o kterých se předpokládá, že ovlivňují studovaný typ organismů
- **Jednorozměrná data (*univariate data*)**
  - pouze jedna proměnná, např. počet druhů
- **Vícerozměrná data (*multivariate data*)**
  - matice dat (data matrix), např. lokality × druhy

Typy proměnných

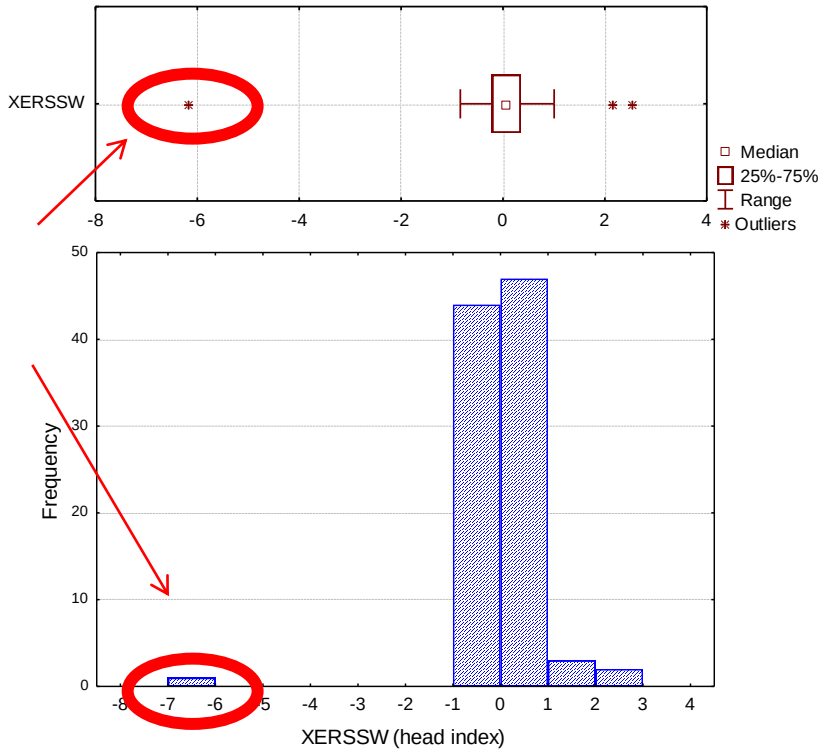
Typ proměnné	Příklady
binární (dvoustavový, presence-absence)	přítomnost nebo absence druhu
mnohostavový	
neseřazený	geologický substrát
seřazený	
semikvantitativní (ordinální)	stupnice pokryvností druhy
kvantitativní (měření)	
diskontinuální (počty, diskrétní)	počet jedinců
kontinuální	teplota, hloubka půdy

Legendre & Legendre (1998)

Explorační analýza dat (exploratory data analysis, EDA)

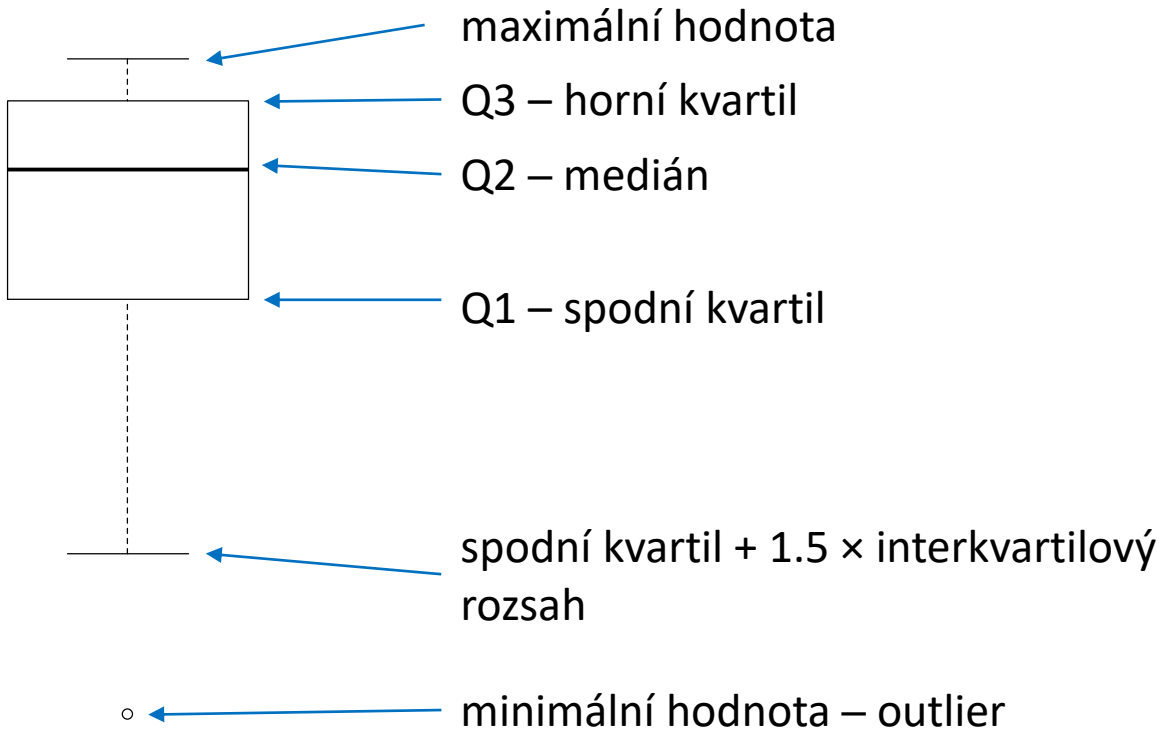
- průzkum dat – kontrola a čištění
  - chyby (*errors*)
    - někdy se chovají jako odlehlé body, je třeba zkontrolovat původní záznam a případně data z analýzy odstranit
  - chybějící data *(missing data,* NA)
    - možnosti jejich nahrazení (interpolace, model)
    - vyloučení proměnné nebo vzorku který má hodně chybějících hodnot
  - odlehlé body (*outliers*)
    - jejich detekce (*outlier analysis*)
- hledání hypotéz, které stojí za to testovat
- **grafická EDA** slouží k
  - odhalení odlehlých bodů (*outlier analysis*)
  - distribuce dat (normalita) a nutnost transformace

potenciálně chybná hodnota



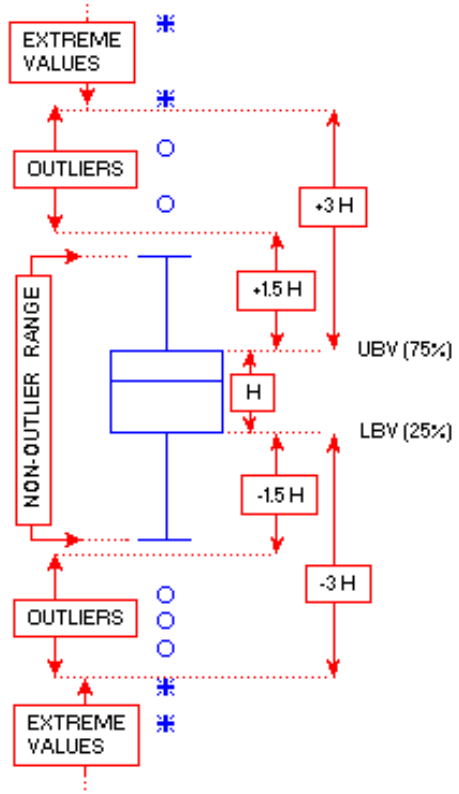
Krabicové grafy (boxplots)

Klasický boxplot
(střední hodnota = medián)

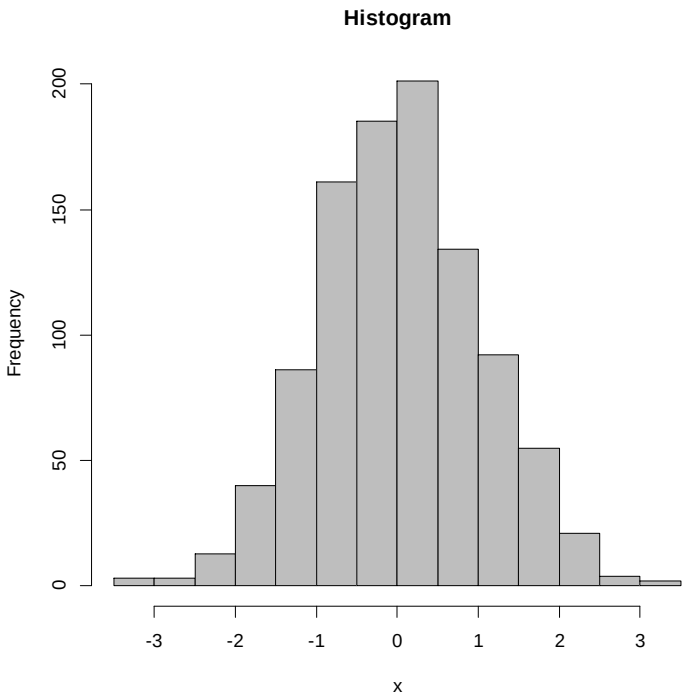



```
boxplot()
```

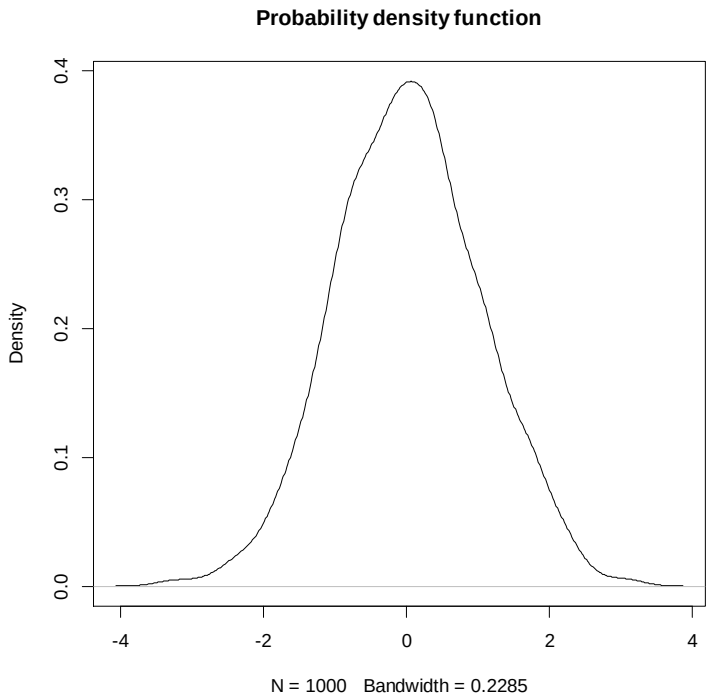
Definice odlehlých bodů a extrémů (STATISTICA)



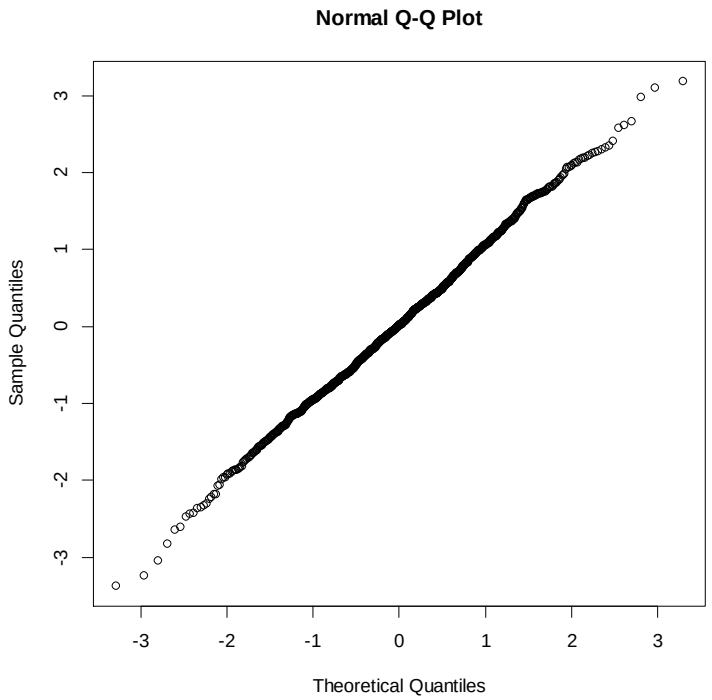
Histograms, PDF plots & Q-Q plots



`hist()`



`density()`



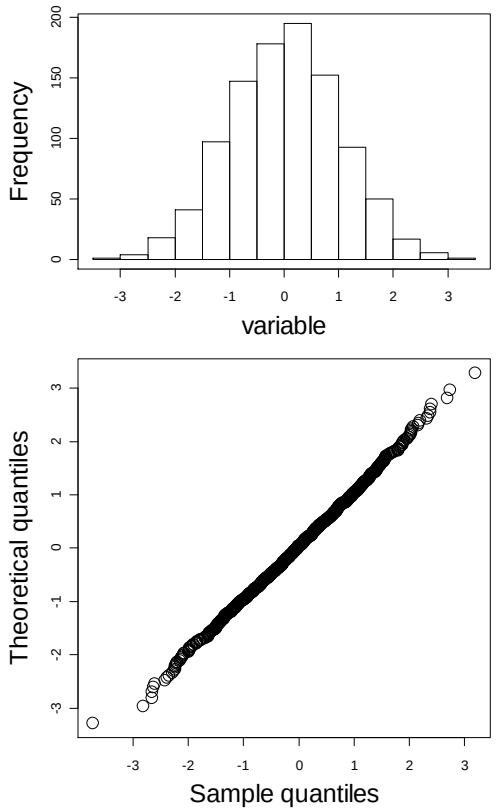
`qqnorm()`

Testování normality dat:

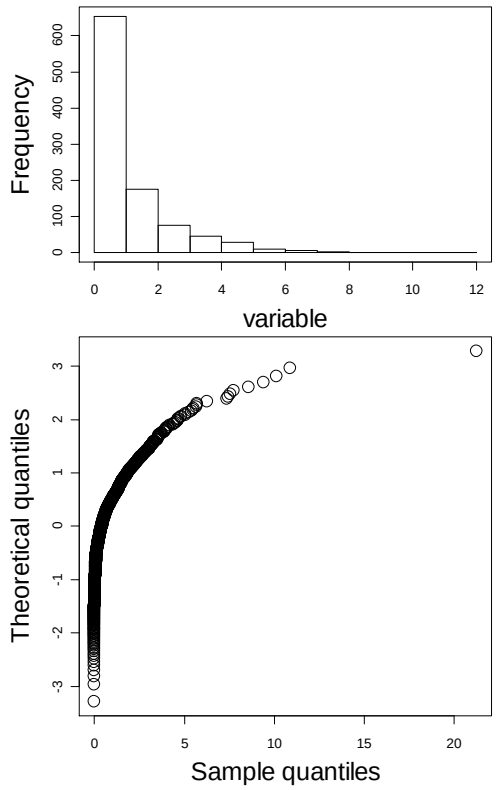
- Shapiro-Wilkův test: `shapiro.test()`
- Kolmogorovův-Smirnovův test: `ks.test()`

Mají data normální rozložení?

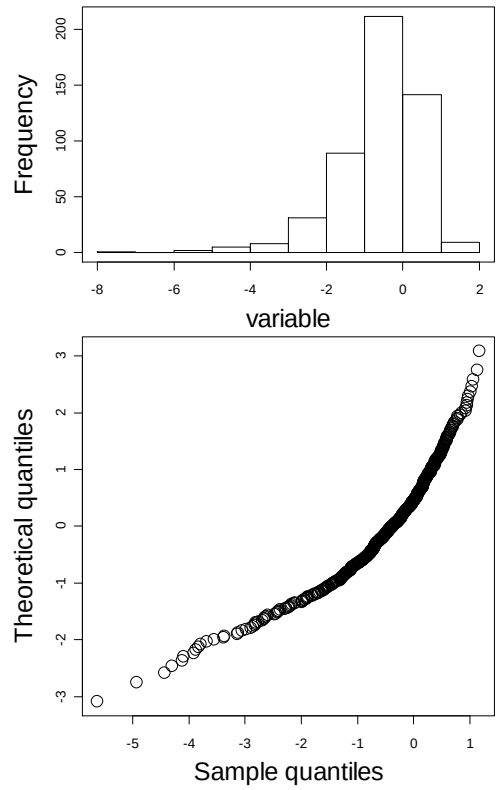
normální rozdělení (*symetrical*)



pozitivně (doprava) sešikmené (*right skewed*)



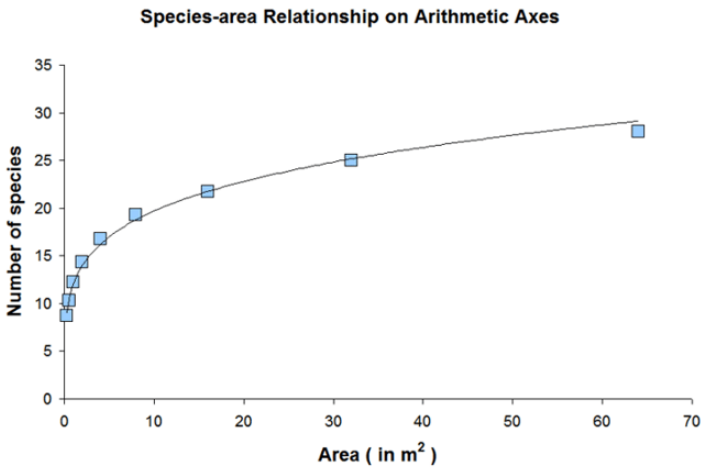
negativně (doleva) sešikmené (*left skewed*)

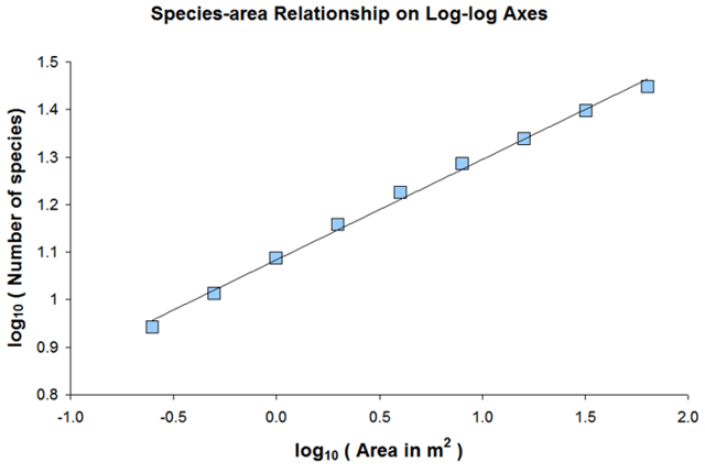


ekologická data jsou často zešikmená pozitivně (doprava), protože jsou omezená nulou na začátku

Transformace dat

- mění relativní vzdálenosti mezi jednotlivými hodnotami a tím i tvar jejich distribuce
- **Proč data transformovat?**
  - parametrické testy jsou založené na předpokladu, že data mají nějaké určité (často normální) rozdělení
  - protože lineární vztahy se dají popsat přímkou a lépe se interpretují než vztahy nelineární
  - škála měření je arbitrární a nemusí odpovídat ekologickému významu proměnné (používáme desítkovou soustavu)





https://en.wikipedia.org/wiki/Species-area_curve

Transformace dat

- Na co si dát při transformaci pozor?
  - aby transformace rozložení dat ještě nezhoršila a nevytvořila nové odlehlé body
  - abychom při komentování výsledků používali netransformované hodnoty proměnných

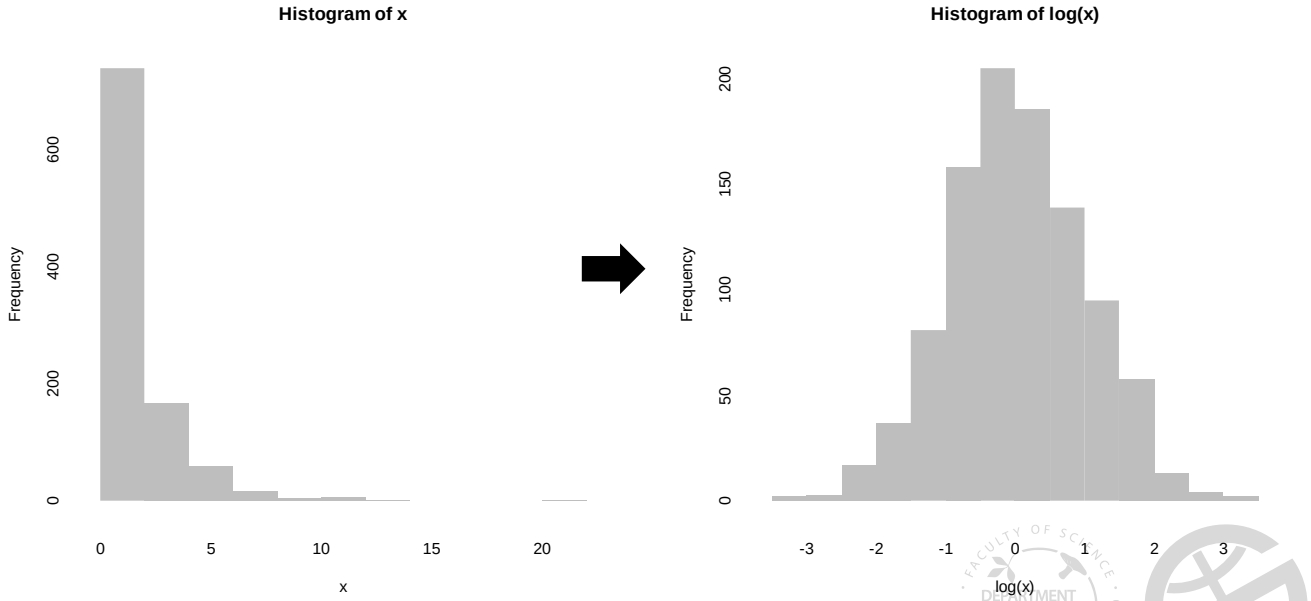
- Typy transformace
  - lineární
    - přičtení konstanty nebo vynásobení konstantou
    - nemění výsledky statistického testování nulových hypotéz
    - např. převod teploty měřené ve stupních Celsia na stupně Fahrenheita
  - nelineární
    - log transformace, odmocninová transformace atd.
    - může změnit výsledky statistického testování

Typy transformací

- **Logaritmická transformace (*log transformation*)**
  - pro data s výrazně pozitivně (doprava) šikmou distribucí (*right skewed*), u kterých existuje vztah mezi průměrem a směrodatnou odchylkou (lognormální rozložení)

$$Y' = \log(Y)$$ případně $$Y' = \log(aY + c)$$

- na základě logaritmu nezáleží (10, 2, e)
- konstanta ***a*** = 1; pokud je Y z intervalu <0;1>, potom $a > 1$
- konstanta ***c*** se přidává, pokud proměnná Y obsahuje nuly
- ***c*** může být např. 1, nebo arbitrárně zvolené malé číslo (0,001)
- na konstantě ***c*** může záležet výsledek analýz (ANOVA), a proto je dobré vybírat takové číslo, aby transformovaná proměnná byla co nejvíce symetrická

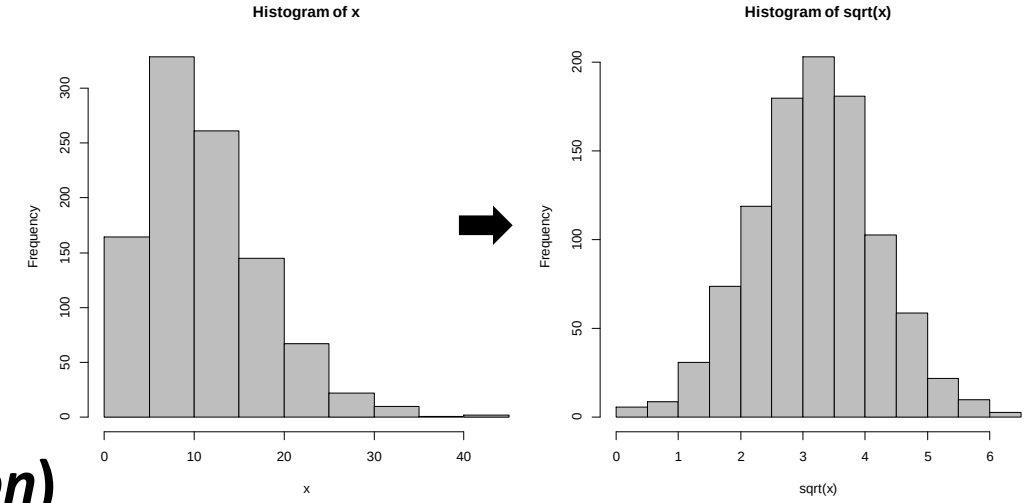


Typy transformací

- **Odmocninová transformace (*square-root transformation*)**
  - vhodná pro mírně doprava zešikmená data (*right skewed*), např. počty druhů (*Poisson distribution*)
  - třetí a vyšší odmocnina je účinnější na více zešikmená data (čtvrtá odmocnina se používá pro abundance druhů s mnoha nulami a několika vysokými hodnotami)

$$Y' = \sqrt{Y}$$ případně $$Y' = \sqrt{Y + c}$$

- konstanta ***c*** se přičítá, pokud soubor obsahuje nuly
- ***c*** může být např. 0,5, nebo 3/8 (0,325)

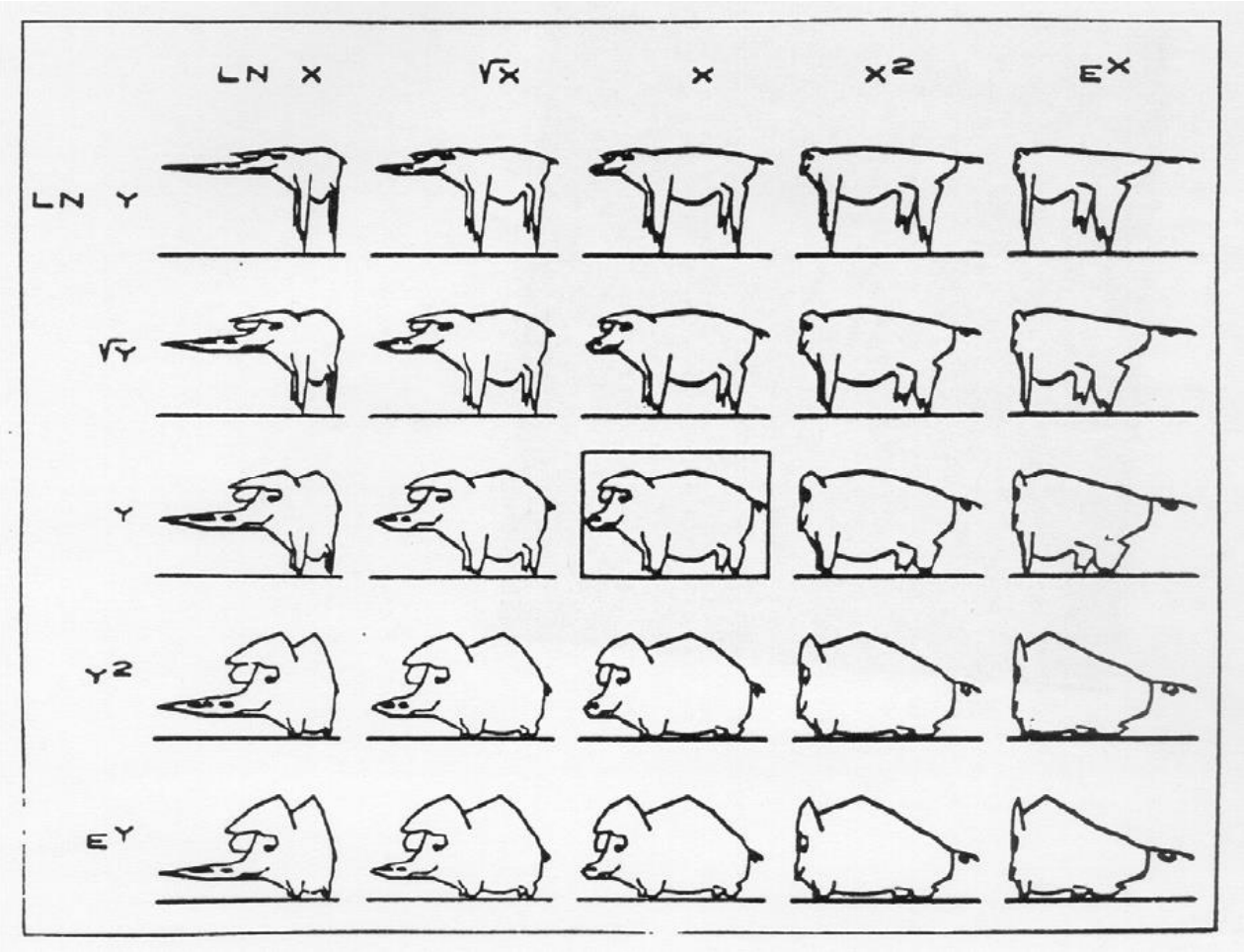


- **Mocninná transformace (*power transformation*)**
  - vhodná pro data negativně (doleva) sešikmená (*left skewed*)

$$Y' = Y^p$$

- pokud $p < 1$ - **odmocninová transformace** (p = 0,5 – druhá odmocnina, p = 0,25 – čtvrtá odmocnina atd.)

Transformace



Münch. Med. Wschr. 124, 1982

Další transformace

- **Transformace pomocí arcsin (*angular transformation*)**
  - vhodná pro procentické hodnoty (a obecně podíly)

$$Y' = \sin Y$$ případně $$Y' = \sin \sqrt{Y}$$

  - použitelná pro hodnoty v intervalu <-1; 1>
  - transformované hodnoty jsou v radiánech

- **Reciproká transformace (*reciprocal transformation*)**
  - vhodná pro poměry (například výška/hmotnost, počet dětí v populaci na počet žen atd.)

$$Y' = 1/Y$$

- **Box-Cox transformace (zobecněná mocninná transformace)**
  - zobecněná parametrická transformace
  - iterativní hledání parametru λ (lambda), pro které je rozdělení transformované proměnné nejblíže normálnímu rozdělení
  - používá se v případě, že nemáme *a priori* představu, jakou transformaci použít

Standardizace dat

- vyrovnává rozdíly v relativním významu (váze) jednotlivých ekologických proměnných (měřené na různých škálách), druhů nebo vzorků
- mění data pomocí statistiky, která je spočtená na datech samotných, např. průměr, součet, rozsah aj. (*data dependent*)
- ve své podstatě je to další typ transformace

Standardizace dat

- **Centrování (*centring*)**
  - výsledná proměnná má průměr roven nule

$$Y'_i = Y_i - \text{průměr}\ (Y)$$

- **Standardizace v úzkém slova smyslu**
  - výsledná proměnná má průměr roven nule a směrodatnou odchylku rovnu jedné
  - „synchronizuje" proměnné měřené v různých jednotkách a na různých stupnicích

$$Y'_i = (Y_i - \text{průměr}\ (Y))/\text{směrodatná odchylka}\ (Y)$$

- **Změna rozsahu hodnot** (*ranging*)
  - výsledná proměnná je v relativních hodnotách nebo v rozsahu [0, 1]

$$Y'_i = Y_i/Y_{max} \quad \text{nebo} \quad Y'_i = (Y_i - Y_{min})/(Y_{max} - Y_{min})$$

Kódování dat

- *Dummy variables*
  - metoda, jak převést **kvalitativní** (kategoriální) proměnnou na **kvantitativní** (binární) proměnné použitelné v analýzách
  - pokud má kategoriální proměnná *n* stavů (hodnot), pro její vyjádření stačí *n*-1 dummy proměnných (jedna z proměnných je vždy lineárně závislá na ostatních)
  - `dummy{dummies}`

hodnoty	dummy proměnné			
	KAMB	LITO	RANK	FLUVI
kambizem	1	0	0	0
litozem	0	1	0	0
ranker	0	0	1	0
fluvizem	0	0	0	1

Kódování dat

- např. nahrazení kódů u alfa-numerických stupnic, např. Braun-Blanquetovy stupnice dominance-abundance

Braun-Blanquetova stupnice:	**r**	**+**	**1**	**2**	**3**	**4**	**5**
ordinální hodnoty:	1	2	3	4	5	6	7
střední hodnoty procent:	1	2	3	13	38	63	88

Literatura

- Legendre, P. & Legendre, L. (2012): Numerical ecology. Third Edition. Elsevier, Amsterdam.
- Borcard, D., Gillet, F. & Legendre, P. (2011): Numerical ecology with R. Springer, New York.