

STANDARDIZACE DAT

CÍL STANDARDIZACE

- Odstranit měřítko z originálních dat
 - Jednotky apod.
- Srovnat variabilitu mezi proměnnými
- Implicitně se používá při výpočtu korelace

STANDARDIZACE PROMĚNNÝCH

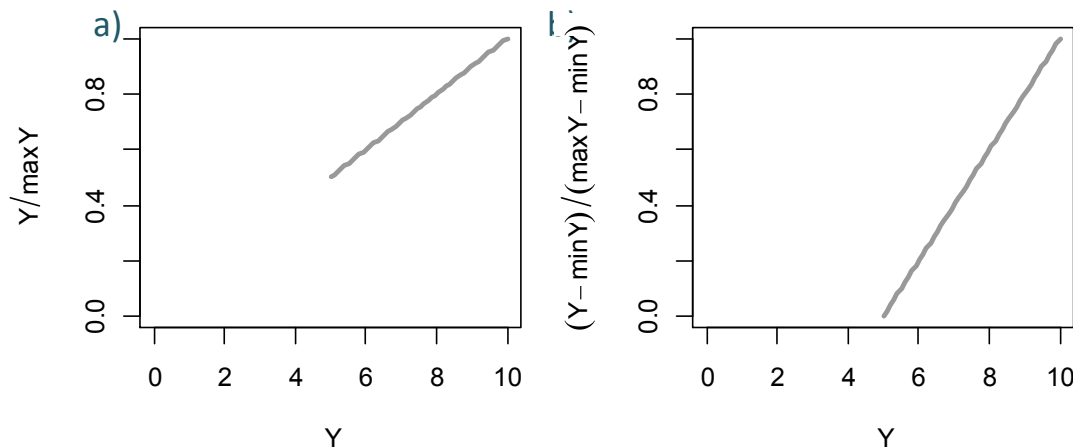
Centrování (*centring*)

- výsledná proměnná má průměr roven nule

Standardizace v úzkém slova smyslu = normalizace

- dává vzniknout bezrozměrným Z-skóre
- výsledná proměnná má průměr roven nule a směrodatnou odchylku rovnu jedné
- „synchronizuje“ proměnné měřené v různých jednotkách a na různých stupnicích
- Umožňuje kombinovat třeba průměrnou teplotu a úhrn srážek
- Implicitně se vždy provádí v regresi, korelaci, v ordinacích pak pro prediktory

Změna rozsahu hodnot (*ranging*)



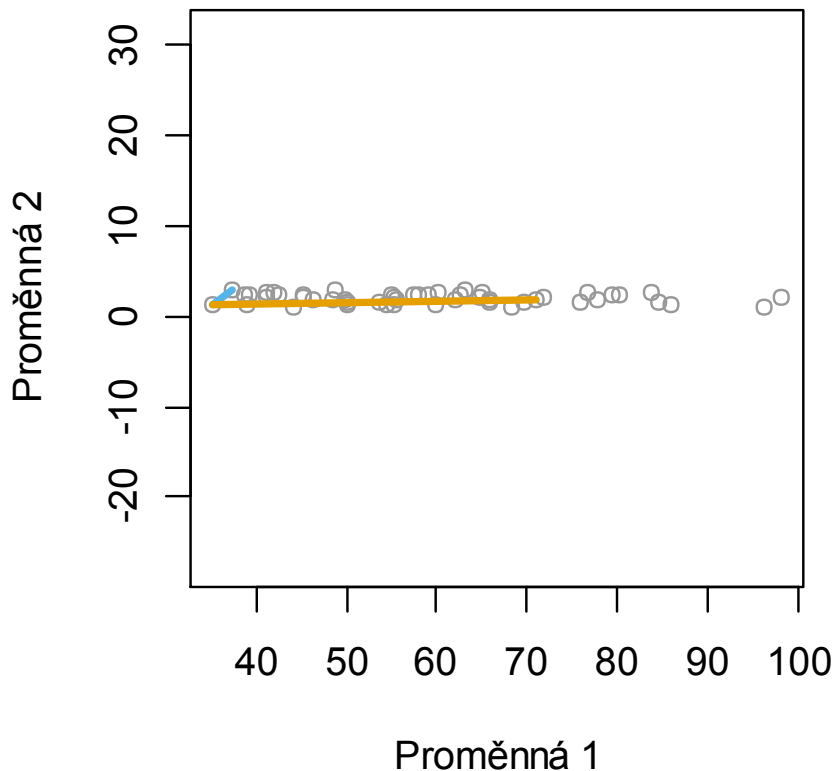
- výsledná proměnná je v rozsahu [0, 1]

- a) relativní škála (poměry mezi hodnotami zachované),
- b) obecná proměnná

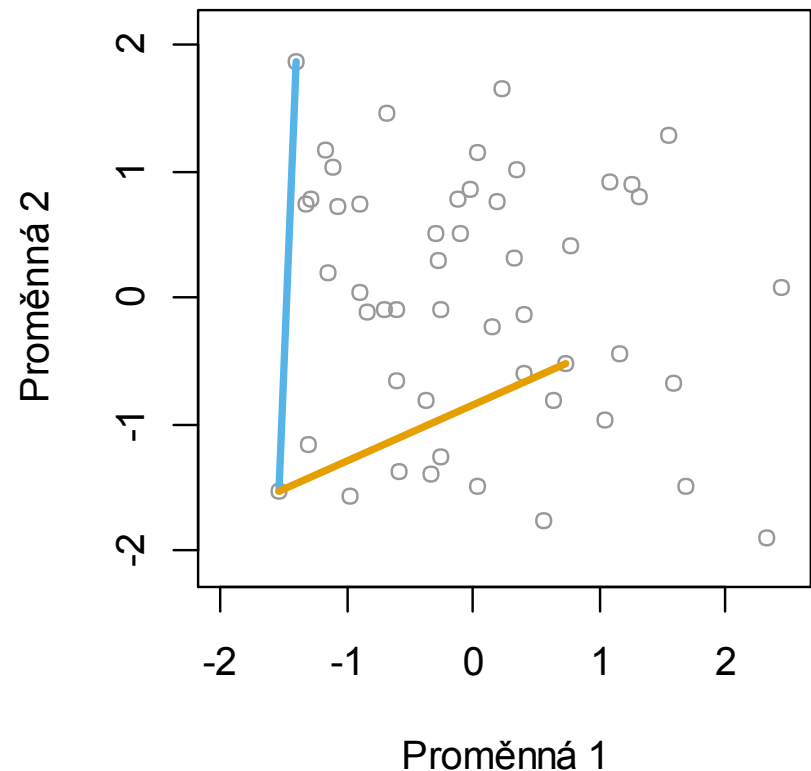
STANDARDIZACE PROMĚNNÝCH

- Odstraní rozdíly v jednotkách
- Bez standardizace vzdálenosti mezi vzorky ovládnou proměnné s velkou variací
- po standardizaci mají všechny proměnné varianci shodnou

Před standardizací



Po standardizaci



STANDARDIZACE DRUHOVÉ MATICE

○ standardizace po druzích (*standardization by species*), tj. po proměnných

- dává stejnou váhu všem druhům – zvýší váhu vzácných druhů a sníží váhu hojných
- ne vždy smysluplná (pokud se druh vyskytuje vzácně v jednom snímku, standardizace po druzích dá tomuto snímku velkou váhu – bude velmi odlišný od ostatních)
- Nutná při analýze proměnných prostředí (odstraní se rozdíly v magnitudě a rozptylu proměnných); proto se v přímé ordinaci prediktory standardizují implicitně
- Nutná při analýze morfometrických parametrů

	sp1	sp2	sp3
vzorek 1	1	3	4
vzorek 2	2	6	8
vzorek 3	10	30	40
průměr	4.333	13	17.33
sd	4.933	14.8	19.73



Odečtení průměru

	sp1	sp2	sp3
vzorek 1	-3.33	-10	-13.33
vzorek 2	-2.33	-7	-9.333
vzorek 3	5.667	17	22.667



Vydělení směrodatnou odchylkou

	sp1	sp2	sp3
vzorek 1	-0.68	-0.68	-0.68
vzorek 2	-0.47	-0.47	-0.47
vzorek 3	1.149	1.149	1.149

STANDARDIZACE DRUHOVÉ MATICE

○ standardizace po vzorcích (*standardization by samples*)

- pokud je analýza zaměřená na relativní proporce mezi druhy, ne jejich absolutní abundance
- vhodné také v případě, že výsledné abundance závisí na důkladnosti, s jakou sbíráme data (např. při odchytu živočichů doba strávená na ploše, počet pastí nebo vliv špatného počasí na mobilitu živočichů)

Původní hodnoty

	sp1	sp2	sp3	průměr	sd
vzorek 1	1	3	4	2.666	1.528
vzorek 2	2	6	8	5.333	3.055
vzorek 3	10	30	40	26.66	15.28



Odečtení průměru
a podělení výsledku
směrodatnou
odchylkou

Výpočet hodnot v prvním sloupci

Hodnoty standardizované po vzorcích

	sp1	sp2	sp3
vzorek 1	-1.09	0.218	0.873
vzorek 2	-1.09	0.218	0.873
vzorek 3	-1.09	0.218	0.873

DALŠÍ STANDARDIZACE (PŘES VZORKY)

○ *Species profile*

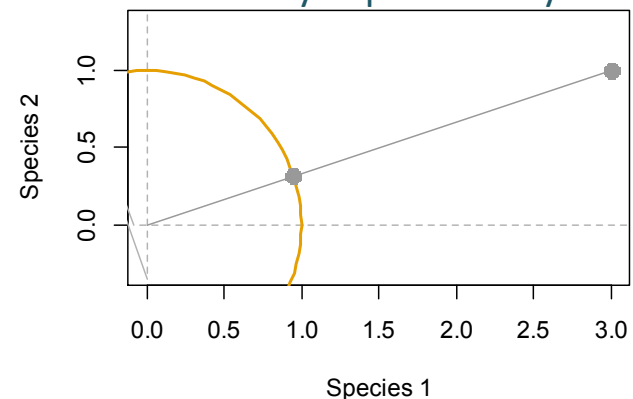
- Abundance druhů jsou vyděleny sumou abundancí v dané vzorku
-> Relativní podíly abundancí

○ Hellingerova standardizace

- modifikovaný species profile, lepší statistické vlastnosti
- Euklidovské vzdálenosti vypočítané na transformovaných datech vedou k Hellingerově vzdálenosti (viz další část)

○ Tětivová standardizace (*chord standardization*)

- Euklidovské vzdálenosti vypočítané na datech standardizovaných přes vzorky vedou k tětivové vzdálenosti



TRANSFORMACE

- matematická funkce, jejíž argumenty nejsou odvozené z dat, na která je transformace aplikovaná
- Změna hodnoty není závislá na ostatních hodnotách proměnné (*data independent*)
- nejčastější důvod je **změnit tvar rozložení** proměnné a zajistit homoskedasticitu
- Mění tvar rozdělení hodnot

STANDARDIZACE

- mění data pomocí statistiky, která je spočtená na datech samotných, např. průměr, součet, rozsah aj.
- **Změna hodnoty závisí na ostatních hodnotách proměnné** (data dependent)
- nejčastější důvod použití je **vyrovnat rozdíly v relativním významu (váze) proměnných**, druhů nebo vzorků
- Nemění tvar rozdělení hodnot

DUMMY VARIABLES, FUZZY CODING

○ Dummy variables

- převod **kvalitativní** (kategoriální) proměnné na sérii **kvantitativních** (binárních, 0/1) proměnných
- pokud má kategoriální proměnná n stavů (kategorií), pro její vyjádření stačí $n-1$ dummy proměnných – obvykle je ale lépe kódovat jako n proměnných kvůli kreslení (např. ordinační diagramy), statistiku to neovlivňuje.
- V R funkce `model.matrix (~factor)` a následná manuální úprava

○ Fuzzy-coding

- Kromě 0/1 používá i mezistupně např. 0.5, nebo 1/3
- součet pro daný vzorek vždy do jedné
- Nelze jednoduše vyjádřit jako faktor o n stupních volnosti

Sample	bahno	písek	vegetace	další proměnné
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	0	1
6	1	0	0

Sample	Substrát	další proměnné
1	bahno
2	písek
3	bahno
4	vegetace
5	vegetace
6	bahno



KÓDOVÁNÍ DAT (*DATA CODING*)

- 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

*) van der Maarel (2007), Table 1

***) Turboveg for Windows 2

METADATA

- zaznamenat veškeré transformace, standardizace, kódování do metadat!

