Transformácia a štandardizácia dát

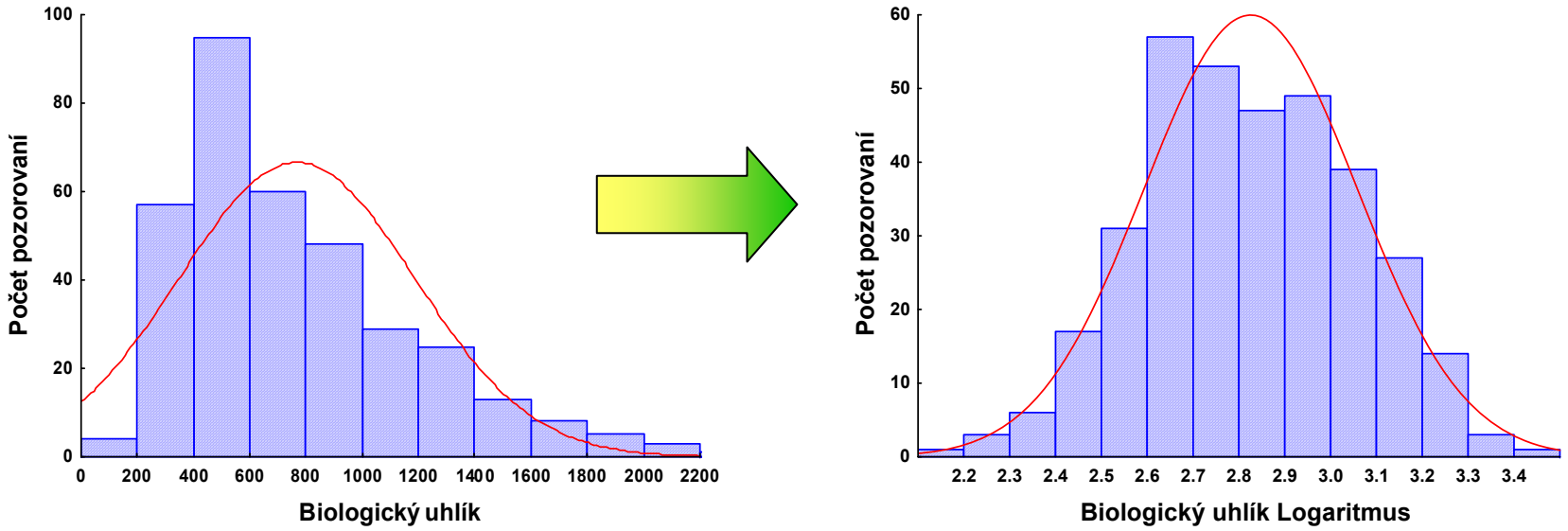
Danka Némethová

Podzim 2008

MU IBA Inštitút bioštatistiky a analýz, Masarykova univerzita

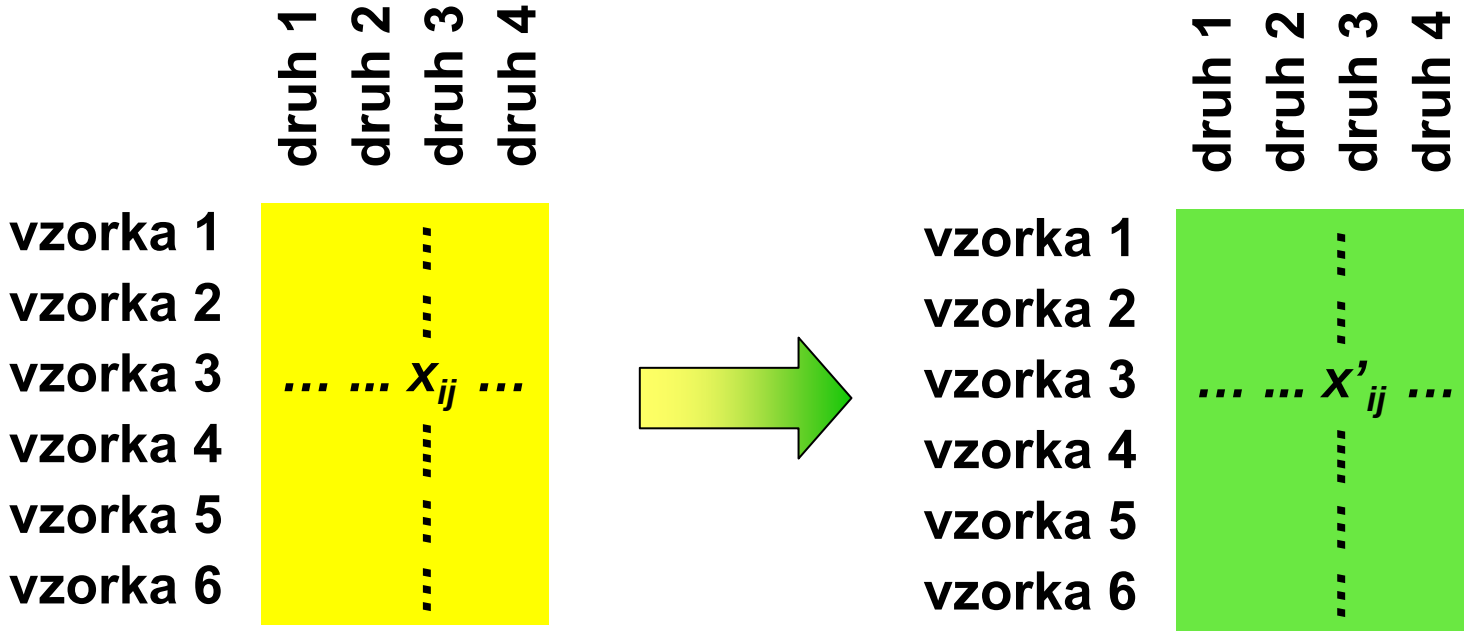
Úvod

- Niektoré mnohorozmerné metódy nevyžadujú **normálne rozdelenie dát**, prípadne sú dostatočne robustné vo vzťahu k odchýlkam od normálneho rozdelenia dát (napr. zhluková analýza).
- Iné metódy mnohorozmerné normálne rozdelenie dát vyžadujú (napr. diskriminačná analýza).
- **Transformáciou** sa dá niekedy rozdelenie dát priblížiť k normálnemu rozdeleniu.



Transformácia

- Transformácia je možná niekoľkými spôsobmi. K **transformácii** sa používajú **konštanty** a **funkcie nezávislé** na analyzovaných dátach.
- Väčšina transformácií, ktoré sa používajú, sú **nelineárne transformácie**. Tieto transformácie menia štruktúru dát.
- **Lineárne transformácie** (v ekológii napr. násobenie abundancií druhu konštantou) nemenia výsledky analýzy ak sa aplikujú na všetky premenné (druhy). Ak sa však takouto transformáciou upravia hodnoty jedného druhu, dôjde k jeho **váženiu**.

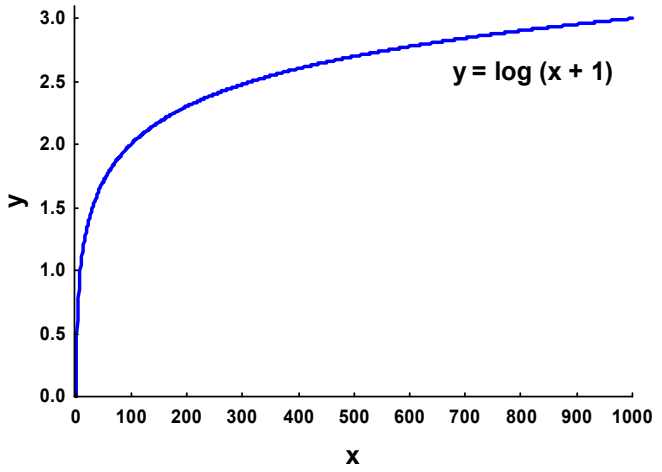


Logaritmická transformácia

$$x'_{ij} = \log_c x_{ij}$$

alebo, ak sú prítomné nuly

$$x'_{ij} = \log_c (x_{ij} + 1)$$



POUŽITIE

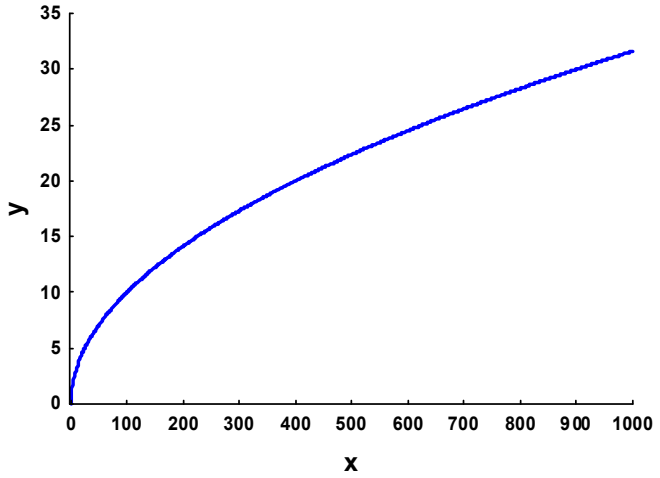
- Na priradenie menšej váhy dominantným druhom, na zvýraznenie kvalitatívnej stránky dát (pri rádových rozdieloch medzi vzorkami príp. druhmi)
- Transformácia hodnôt env. premenných s lognormálnym rozdelením
- U env. premenných na zobrazenie lineárneho vzťahu mnohých druhov napr. k logaritmu obsahu toxickej látky alebo logaritmu obsahu živín

Odmocninová transformácia

$$x'_{ij} = \sqrt{x_{ij}}$$

znaky nesmú dosahovať nulových hodnôt, preto sa niekedy používa v tvare:

$$x'_{ij} = \sqrt{x_{ij} + 0.5}$$



POUŽITIE

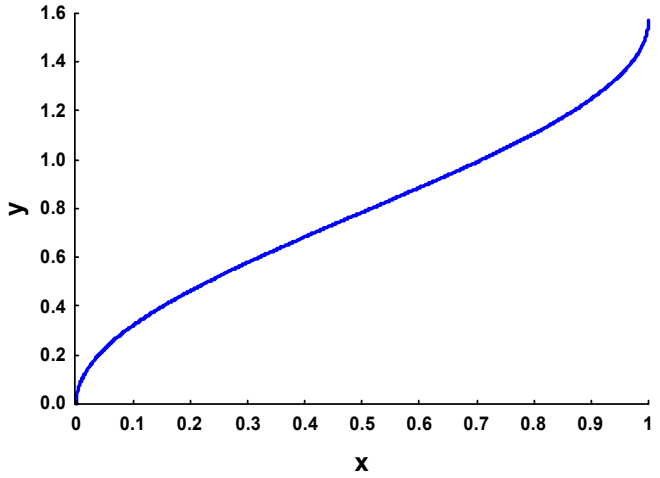
- pred analyzovaním premenných s Poissonovým rozdelením (napr. počet jedincov určitého druhu získaných z jednej pasce za určitú časovú jednotku)
- na priradenie nižšej váhy dominantným druhom

Arcussinová transformácia

Používa sa v kombinácii s odmocninovou transformáciou.

$$x'_{ij} = \arcsin \sqrt{x_{ij}}$$

Predpokladá, že dáta sú merané v intervale <0,1>.

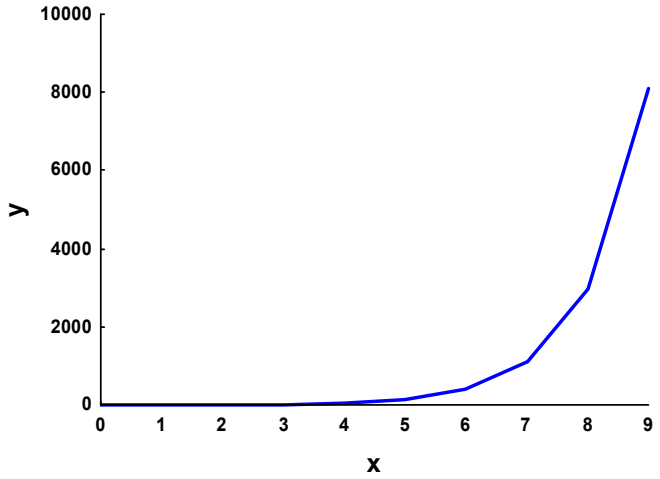


POUŽITIE

- Používa sa na úpravu percentuálnych hodnôt, vyjadrených v intervale <0,1> (napr. pokryvnosti vegetačných druhov).

Exponenciálna transformácia

$$x'_{ij} = a^{x_{ij}}$$



Ak a je reálne číslo väčšie ako 1, sú zvýraznené dominantné druhy.

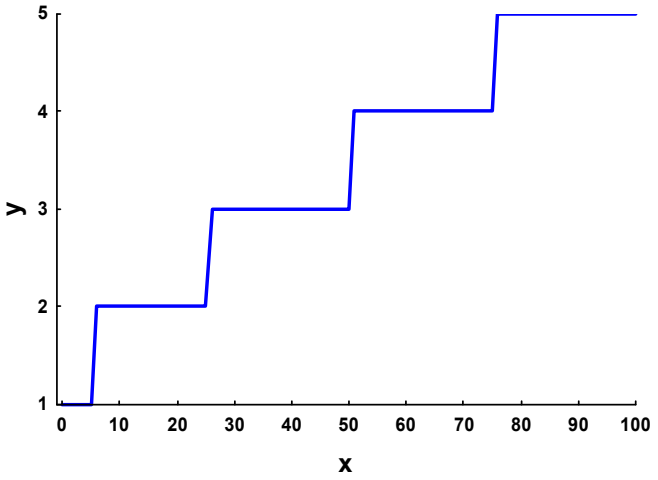
Transformácia na ordinálnu škálu

Abundancie druhov prevedené do tried.

Čím vyššie je číslo triedy, tým vyššia je abundancia.

Typickou transformáciou na ordinálnu škálu je použitie Braun-Blanquetovej stupnice pri kvantifikovaní pokryvnosti vegetácie.

Extrémom je binarizácia – transformácia na prezenciu a absenciu.



$$x'_{ij} = 0 \quad \text{ak} \quad x_{ij} = 0 \qquad\qquad x'_{ij} = 1 \quad \text{ak} \quad x_{ij} > 0$$

- Ak sú k dispozícii spojité dáta, je vhodnejšia akákoľvek iná transformácia.
- Je však výhodné zbierať dáta v teréne na ordinálnej škále (v botanike).

Štandardizácia

- Ku štandardizácii sa používajú štatistiky odvodené z analyzovaného súboru dát (rozpätie, smerodatná odchýlka, priemer, maximum atď.). Znaky sa týmto postupom prevádzajú na rovnaké merítko (čiže prestáva záležať na skutočnom rozmere príslušného znaku).
- Existuje viacero spôsobov štandardizácie dát a dôvody na ich použitie sú rôzne.
- Štandardizácia: použitie určitého štandardu pre všetky premenné (druhy) alebo objekty (vzorky, lokality) pred vypočítaním (ne)podobností alebo pred aplikovaním zhlukovej analýzy.

Štandardizácia na celkovú abundanciu vzorky

- Abundancie druhov vo vzorke sa spočítajú a každá abundancia je vydelená týmto súčtom. Takto sa určia relatívne abundancie (dominancia) druhov.
- Je potrebné používať túto štandardizáciu opatrne ak sú súčty abundancií vo vzorkách veľmi rozdielne, pretože vzácne druhy a objavujú až vo vzorkách s vysokým počtom jedincov.

$$x'_{ij} = \frac{x_{ij}}{\sum_i x_{ij}}$$

Štandardizácia na celkovú abundanciu druhu

- Pre každý druh sú spočítané abundancie cez všetky vzorky a potom sú vydelené celkovou sumou.
- Táto štandardizácia silne nadváži vzácne druhy a podváži bežné druhy. Preto sa táto štandardizácia odporúča len vtedy, ak sa frekvencie druhov v tabuľke veľmi nelíšia.
- Býva používaná v prípadoch, ak sa v zozname druhov vyskytujú rôzne trofické úrovne, pretože vyššie trofické úrovne sú menej zastúpené.

$$x'_{ij} = \frac{x_{ij}}{\sum_j x_{ij}}$$

Štandardizácia na maximum vzorky

- Všetky abundancie druhov sú vydelené maximálnou abundanciou dosiahnutou nejakým druhom vo vzorke.
- Táto štandardizácia je aplikovaná z rovnakého dôvodu ako štandardizácia na celkovú abundanciu vo vzorke.
- Je menej citlivá na počet druhov, ale je potrebné používať ju opatrne v prípadoch ak sú veľké rozdiely vo vyrovnanosti vzoriek.

$$x'_{ij} = \frac{x_{ij}}{\max_i \{x_{ij}\}}$$

Štandardizácia na maximum druhu

- Táto štandardizácia je odporúčaná podobne ako štandardizácia na celkovú abundanciu druhu, ak sú prítomné rôzne trofické úrovne.

$$x'_{ij} = \frac{x_{ij}}{\max_j \{x_{ij}\}}$$

Štandardizácia na jednotkovú dĺžku vektora vzorky

- Vydelením abundancie druhu na vzorke odmocninou sumy štvorcov abundancií sa všetky vektory vzoriek zobrazia na jednotkovej kružnici druhového priestoru.
- Euklidovské vzdialenosti sa touto štandardizáciou redukujú na tetivové vzdialenosti (cord distance).

$$x'_{ij} = \frac{x_{ij}}{\sqrt{\sum_i x_{ij}^2}}$$