Základní analýza dat

literatura:

Hendl, J. 2006: Přehled statistických metod zpracování dat. Analýza a metaanalýza dat. Praha: Portál.

Macháček, J. 2001: Studie k velkomoravské keramice. Metody, analýzy a syntézy, modely. Brno.

Neustupný, E. 1986: Nástin archeologické metody, Archeologické rozhledy 38, 525–549.

► *Úvod*

Deskripce

- v archeologii se deskripce obvykle realizuje ve formě matice, jejíž řádky odpovídají objektům (entity) a sloupce proměnným (znaky, deskriptory, vlastnosti entit). Prvky matice pak nějakým způsobem charakterizují vztah mezi objektem a deskriptorem. Taková matice reprezentuje deskriptivní systém, který je obecně definován množinou objektů, množinou deskriptorů a konkrétním zobrazením, daným prvky matice. Je zřejmé, že jednomu kontextu může odpovídat více než jeden deskriptivní systém.

Proměnné (znaky, deskriptory, vlastnosti entit)

- pomocí proměnných popisujeme zkoumaný soubor. Například u sídliště z období neolitu mohou být proměnné: délka, šířka a hloubka jámy, typ výplně, funkce jámy, datování, celkový počet střepů, počet střepů LBK a počet střepů STK.

Typy proměnných

- závislé a nezávislé (viz. výše) + rušivé proměnné (v archeologii jde např. o proměnné týkající se prostoru a času – zjištěné shodné znaky mezi archeologickými objekty nemusí být známkou jejich např. funkční podobnosti, ale podobnosti chronologické, nebo korelace některých vlastností objektů může vycházet z jejich podobného umístění v rámci sídliště).

Proměnné podle typu použitého měřítka

- nominální (kvalitativní) měřítko, např. pohlaví zemřelého, typ keramického milodaru
  - podtypem je binární (dichotomické) měřítko, tj. rozlišují se pouze 2 třídy, např. přítomnost milodarů v hrobech)
- ordinální měřítko, lze rozlišit řazení podle intenzity nějakého jevu (např. stupeň promíšení výplně jámy: jednolitá, středně promíšená, promíšená)
- intervalové měřítko (scale), tj. vzdálenosti jednotlivých údajů jsou dány jednotkou měření (např. délka hrobové jámy)
  - poměrové měřítko, tj. existuje absolutní nulový bod (např. počet milodarů v hrobu)
- změny (transformace) proměnných jsou standardně možné směrem od vyšších k nižším.

Spolehlivost (reliabilita)

- archeologická data mají obvykle nižší reliabilitu (např. v důsledků nízké kvality terénního výzkumu nebo v důsledků formativních procesů archeologického materiálu, úskalí může být i v nekritickém přejímání výsledků jiných, např. přírodovědných, oborů)
- lze ověřovat výpočtem numerického hodnocení (viz. Hendl 2006, kapitola 7.2.9)

Validita

Validita požaduje, aby procedura měření skutečně měřila to, co předpokládáme, že měří.

- obsahová validita, tj. do jaké míry měření skutečně reprezentuje dané vlastnosti nebo kvality
- kriteriální validita, tj. srovnání s jiným, již ověřeným měřením, srovnání s tzv. zlatým standardem
- konstruktová validita se zabývá teoretickými aspekty měřeného konstruktu (proměnné). Pokud test prokazuje vztahy k těm proměnným, jež podle teorie očekáváme, jde o konvergentní charakter konstruktové validity. Když nemá vztah k proměnným, když tento vztah neočekáváme jde o diskriminační charakter. Důležité je, aby výsledky predikovaly stavy, které podle teorie očekáváme.

Externí evidence

- validace je prováděna na základě tzv. externí evidence, provádí se prostřednictvím dat, která nebyla součástí deskriptivní matice, z níž byly vyhledávány formální struktury. Může se jednat např. o pohlaví osob pohřbených v jednotlivých hrobech či prostorové vztahy (např. vertikální či horizontální stratigrafie).
- často např. na základě prostorové GIS analýzy

Organizace dat a jejich kontrola, scházející údaje.

kódování

scházející údaje:

- odstranit nereliabilní proměnné nebo data
- imputace průměrných hodnot

► *Grafický a číselný popis rozložení dat*

- zde se zaměříme na jednorozměrný popis a analýzu proměnných, tj. každou proměnnou hodnotíme zvlášť)
- při rozhodování, co budeme pomocí dat počítat nebo jak je budeme zobrazovat, mohou hrát roli čtyři aspekty účelu analýzy:
  1. Explorace: v datech hledáme zajímavé konfigurace a vztahy („data mining“)
     - začínáme se zkoumáním jednotlivých proměnných a teprve pak analyzujeme jejich vztahy
     - začínáme zobrazovat data pomocí grafů, pak přidáme numerické charakteristiky specifických aspektů dat
  2. Kontrola dat: grafické metody slouží i ke kontrole dat, mohou odhalit chyby v zápisu i v měření
  3. Odhadování: platí spíše pro přírodní zákonitosti a sociologii
  4. Komunikace: je zapotřebí data zobrazit tak, aby se jejich důležité vlastnosti efektivně zprostředkovaly příjemci informací. Záleží také na typu sdělení, tj. do textu dáváme tabulky, pro přednášky používáme grafy.

Zobrazení nominálních a ordinálních proměnných

- zobrazení závisí na počtu a typu kategorií, při malém počtu entit je možné některé kategorie znaku sloučit
- volba počtu intervalů se nejčastěji provádí pomocí *Sturgesova pravidla*, které doporučuje volit optimální počet intervalů podle vzorce: $\mathbf{k = 1 + 3{,}3 \log_{10}(n)}$, kde n je počet různých hodnot znaku, jež máme k dispozici. Lze také využít následující tabulku:

počet různých hodnot znaku	optimální počet intervalů daný Sturgesovým pravidlem
1	1
2	2
3-5	3
6-11	4
12-22	5
23-45	6
46-90	7
91-181	8
182-362	9
363-724	10
725-1148	11
1149-2896	12
2897-5792	13
5793-11585	14
11586-23171	15
23172-46341	16
46342-92681	17
...	...

- jako zobrazovací prostředek se používají tabulky s procenty, koláčové a sloupcové grafy

Zobrazení kvantitativních (intervalových či „scale“) dat

- soubor si můžeme představit jako *n*-tici reálných čísel, v níž se jednotlivé prvky mohou opakovat. Například {2; 8; 9; 10; 1; 0; 5} je statistický soubor o 7 prvcích ($n = 7$)
- základním numerickým zobrazením je tabulka četností, relativních četností a kumulativních četností
- i - pořadové číslo (index řádku tabulky)
  x_i - pozorovaná hodnota
  n_i - počet hodnot xi - absolutní četnosti
  f_i - relativní četnost
  N_i - kumulativní absolutní četnost
  F_i - kumulativní relativní četnost

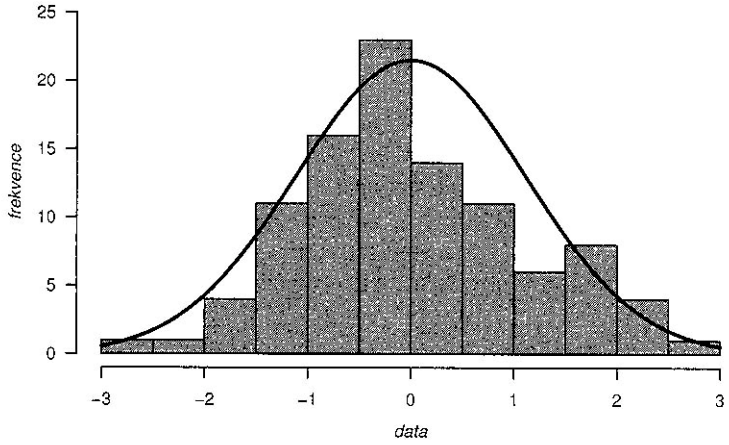
  - četnost (v excelu funkce ČETNOSTI) je počet naměřených hodnot v souboru
  - relativní četnost (procentuální zastoupení počtu naměřených hodnot v souboru): $f_i = 100 \times n_i / n$

příklad v tabulce:

	vlnice	šroubovice	klikatka	rýhy	SUMA
n_i	5	11	7	3	26
$f_i = 100 \times n_i / n$	19,23	42,31	26,92	11,54	100

  - kumulativní četnost je postupný součet četností od nejnižší naměřené hodnoty k nejvyšší
- grafické zobrazení vytváří geometrický obraz dat. Nejznámější způsob zobrazení hodnot jedné proměnné se nazývá histogram, kdy osa *X* odpovídá hodnotám proměnné a osa *Y* absolutním nebo relativním četnostem.
- při popisování a analýze toho, co graf zobrazuje, si všímáme nejdříve základní tvarové konfigurace a pak odchylek od tohoto tvaru. Hodnotíme:

  - zhuštění, tj. kde se nalézá místo nebo místa nejvyšší četnosti hodnot
  - shluky, tj. zda existuje jeden nebo více shluků dat v grafu
  - mezery, tj. zda jsou v grafu intervaly nebo oblasti bez hodnot
  - odlehlé hodnoty, tj. jsou-li v grafu údaje podstatně rozdílné od zbytku dat
  - tvar rozdělení, tj. zda lze jednoduše popsat tvar rozdělení dat
- tvar histogramu se porovnává s ideální křivkou, jež se nazývá *hustota*. Nejčastějším typem *hustoty* je tzv. gaussovská křivka nebo-li normální křivka. Jde o symetrickou křivku zvonovitého tvaru. Data s tímto rozdělením se nazývají normálně rozdělená data.



Míry centrální tendence

Statistické zpracování dat pomocí tabulek a grafů usnadňuje jejich vizuální analýzu a celkové posouzení datové konfigurace. Pro další zpracování však potřebujeme data vhodně kondenzovat. Proto se počítají různé číselné charakteristiky – popisné statistiky, které zachycují různé aspekty dat. Jedná se především o charakteristiky centrální tendence a rozptýlenosti, ale i o další charakteristiky jako šikmost nebo špičatost rozdělení dat.

Míry centrální tendence se snaží charakterizovat typickou hodnotu dat a nejznámější z nich jsou aritmetický průměr, medián a modus.

- aritmetický průměr (značí se M nebo x s čárkou nad, v excelu PRŮMĚR) je definovaný jako součet všech naměřených údajů vydělený jejich počtem. Aritmetický průměr je optimální charakteristikou typické hodnoty pro množiny dat s následujícími vlastnostmi: součet odchylek měření od průměru se rovná nule. Jeho nevýhodou je velká citlivost vůči odlehlým hodnotám.
- medián (označuje se Me, v excelu funkce MEDIAN) je prostřední hodnota souboru, tj. dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny. Např. pro řadu hodnot {0; 3; 4; 5; 8; 9; 15} je průměr 6,28 a modus 5.
- medián je na rozdíl od průměru málo citlivý vůči odlehlým hodnotám. Domnívám se, že pro archeologii je median minimálně stejně důležitý jako průměr.
- modus (označuje se Mo, v excelu funkce MODE) nebo modální hodnota je hodnota, která se v datech vyskytuje nejčastěji. Modus nalézá uplatnění především u kategoriálních dat. Z histogramu lze modus rozlišit podle nejvyššího vrcholku dat. Pokud je vrcholků více, uvádí se všechny.

Míra rozptýlenosti

Náhodně proměnlivé údaje nestačí charakterizovat jenom střední hodnotou. Data se stejnou střední hodnotou mohou mít různou rozptýlenost

- variační rozpětí (značí se R) se počítá jako rozdíl maximální a minimální hodnoty: $R = x_{max} - x_{min}$ Nevýhodou variačního rozpětí je velká citlivost vůči odlehlým hodnotám.

- rozptyl a směrodatná odchylka spolu souvisejí. Oběma je společná vlastnost, že na rozdíl od variačního rozpětí využívají při výpočtu všechny údaje a obě se vztahují k aritmetickému průměru, tj. měří rozptýlenost dat kolem aritmetického průměru dat. Nevýhodou je, že dávají velkou váhu extrémním hodnotám.
- rozptyl (značí se s^2, v excelu funkce VAR) je definován jako průměrná kvadratická odchylka měření od aritmetického průměru
- směrodatná odchylka (značí se s, v excelu funkce SMODCH) je odmocnina z rozptylu a vrací míru rozptýlenosti do původních dat. Směrodatná odchylka měří rozptýlenost kolem průměrů a má se používat jenom tehdy, když průměr je vhodný jako míra střední hodnoty. Směrodatná odchylka je silně ovlivněna extrémními hodnotami. Jestliže je rozdělení dat silně zešikmené, směrodatná odchylka neposkytuje dobrou informaci o rozptýlenosti dat – v takovém případě je lepší použít kvantilové míry.
- variační koeficient (značí se VK) se používá pokud chceme porovnat rozptýlenost dat skupin měření stejné proměnné s různým průměrem. Vypočítá se jako podíl směrodatné odchylka a aritmetického průměru: VK = s/M. Výsledek se uvádí jako desetinné číslo nebo v procentech.

Kvantilové míry

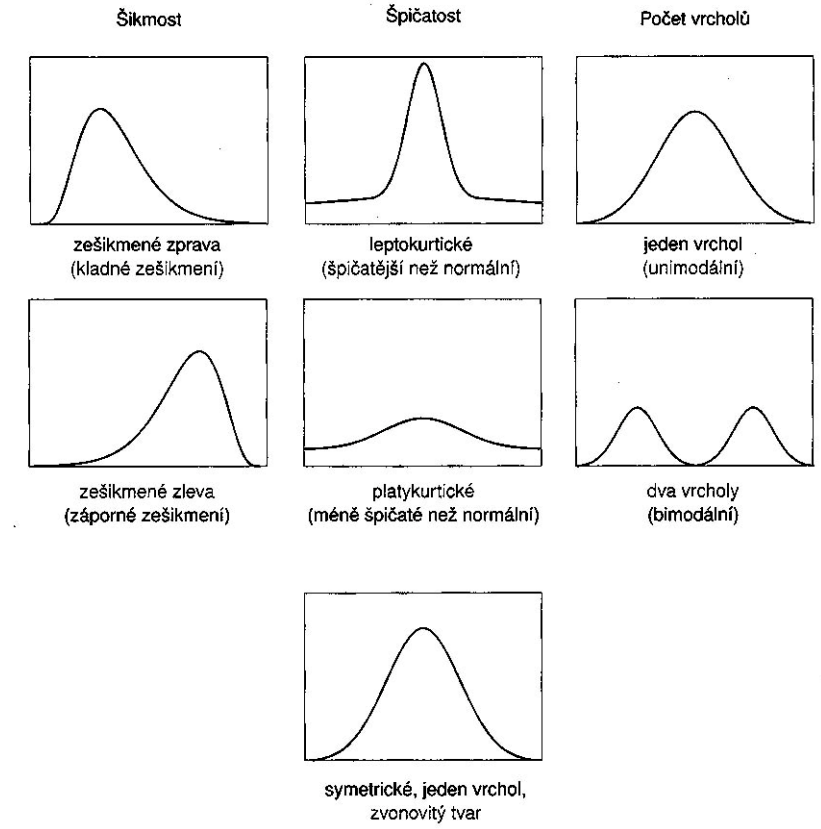
- empirický kvantil je hodnota, pod níž leží definovaná část údajů. Parametr kvantilu (značí se q) je z intervalu hodnot $0 < q < 1$. Často se hladiny q uvádějí v procentech a v tomto případě se nalezené hodnoty označují jako percentily. Např. 25% percentil je rovný kvantilu o hladině 0,25.
- ačkoliv lze z dat vypočítat mnoho různých empirických kvantilů, některé z nich se používají pravidelně. Slouží k popisu jednotlivých částí rozdělení dat a vypočítávají se z nich také míry rozptýlenosti. Jsou to percentily s hladinou 25%, 50% a 75%, které se označují jako kvartily a označují se:
    - Q_I je první neboli dolní kvartil (q = 25%)
    - Q_{II} je druhý kvartil neboli medián (q = 50%)
    - Q_{III} je třetí neboli horní kvartil (q = 75%)
- interkvartilové rozpětí $Q = Q_{III} - Q_I$ je charakteristikou rozptýlenosti, jež se standardně používá k popisu tvaru dat, když se z nějakého důvodu nechceme opřít o průměrové charakteristiky, jako je aritmetický průměr nebo směrodatná odchylka. V intervalu Q_{III} až Q_I se nachází 50% údajů. Interkvartilové rozpětí není na rozdíl od směrodatné odchylky tak citlivé vůči odlehlým hodnotám.

- mediánová absolutní odchylka je míra rozptýlenosti, která – podobně jako interkvartilové rozpětí – není citlivá k odlehlým hodnotám. Spočítá se jako medián z absolutních hodnot odchylek jednotlivých měřené od mediánu a označuje se jako MAD (median absolute deviation). Vzorec: $MAD = Me\ \{\,|\,x_i - Me\,|\,\}$

Míry špičatosti a šikmosti

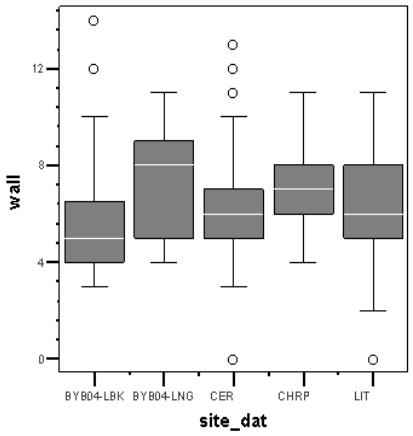
- tyto charakteristiky slouží k jemnějšímu popisu specifických stránek dat. Hodnotí se pomocí nich také to, jak se rozdělení dat podobá normální Gaussově křivce. Nejčastěji se využívají tzv. centrální momenty třetího a čtvrtého stupně. Centrální moment *k*-tého stupně m_k je obecně definován vzorcem: $m_k = \text{suma}\ (x_i - x_{\text{průměr}})^k / n$
- šikmost S_1 (v excelu funkce SKEW) měří zešikmenost, resp. nesymetrii dat a vypočítá se pomocí druhého a třetího momentu podle vzorce: $S_1 = (m_3 / m_2^{3/2})$

- $S_1 = 0$ platí přibližně pro rozdělení přibližně symetrické, $S_1 > 0$ pro rozdělení s prodlouženým pravým koncem, naopak $S_1 < 0$ pro rozdělení s prodlouženým levým koncem.
- koeficient špičatosti S_2 (v excelu funkce KURT) měří odchylku špičatosti zkoumaného rozdělení od normálního rozdělení: $S_2 = (m_4 / m_2^2) - 3$
- takto vypočítaná špičatost má pro normální rozdělení hodnotu 0. Symetrická rozdělení mohou mít stejný rozptyl, ale odlišnou špičatost. Plošší křivky ($S_2 > 0$) nazýváme platykurtické, špičatější křivky ($S_2 < 0$) leptokurtické.



Popis dat pomocí pěti hodnot a krabicový graf s anténami

- vhodným způsobem k popisu jak centrální tendence dat, tak jejich rozptýlenosti je uvedení mediánu jako míry střední hodnoty, kvartilů a nejmenší a největší hodnoty (minima a maxima hodnot).
- těchto pěti hodnot se využívá k sestrojení tzv. krabicového grafu s anténami (box-plot). Používá se pro znázornění jedné množiny dat, ale ještě častěji pro porovnání několika skupin dat. Dovoluje u souborů posoudit a porovnat jak centrální tendence dat, tak i jejich rozptýlenost. Navíc pomocí tohoto grafu posuzujeme i zešikmení a přítomnost odlehlých hodnot (outliers).
- krabička krabicového grafu obsahuje 50% dat a je rozdělena na dvě části mediánem. Dolní hrana krabičky je určena dolním (prvním) kvartilem a horní hrana třetím kvartilem. Pokud je medián blízko jedné z horizontálních hran krabičky, rozdělení dat je zešikmené v opačném směru.



Příklad popisu 5 souborů dat [1]

jemná keramika – síla stěny v mm									
	celkem nádob	minimum	Q_I	medián	průměr	Q_{III}	maximum	rozptyl Q_I - Q_{III}	rozptyl úplný
BYB04-LBK	59	3	4	5	5,68	7	14	3	11
BYB04-LNG	14	4	5	8	7,21	9	11	4	7
CER	315	3	4	6	6,30	7	13	3	10
CHRP	37	4	6	7	7,24	8	11	2	7
LIT	42	2	5	6	6,45	8	11	3	9

[1] J. Macháček uvádí ještě **směrodatnou odchylku** (např. 8,23%) a **variační koeficient** např. (67,35%).