Nejčastější chyby v explorační analýze

Obecně doporučuju přečíst přednášku 5:

Výběrová šetření, Exploratorní analýza

http://homel.vsb.cz/~lit40/STA1/Materialy/IO.pptx

Použití nesprávných charakteristik pro daný typ proměnné

Nominální proměnná: např. barva auta (červená, zelená, modrá,…) nemá smysl uspořádání např. podle velikosti => používáme pouze absolutní a relativní četnost, modus, výsečový graf, histogram. Není možné použít kumulativní charakteristiky.

Podle jakého kritéria bychom barvy seřazovali?

Tabulka rozdělení četnosti barev projíždějících aut				
Hodnoty	Absolutní četnosti	Relativní četnosti	Kumulativní absolutní četnosti	Kumulativní relativní četnosti
červená	8	0,40	8	0,40
černá	4	0,20	12	0,60
modrá	5	0,25	17	0,85
bílá	3	0,15	20	1,00
Celkem:	20	1,00	20	1,00

Použití nesprávných charakteristik pro daný typ proměnné

Ordinální proměnná: např. známky ve škole (výborně, chvalitebně, dobře,…) má smysl uspořádání např. od nejlepší k nejhorší => používáme absolutní a relativní četnost + kumulativní charakteristiky, modus , výsečový graf, histogram, Lorenzova křivka, Paretův graf

Tabulka rozdělení četnosti výsledné známky ze statistiky				
Hodnoty	Absolutní četnosti	Relativní četnosti	Kumulativní absolutní četnosti	Kumulativní relativní četnosti
dobře	5	0,25	5	0,25
výborně	4	0,20	9	0,45
velmi dobře	8	0,40	17	0,85
nevyhověl	3	0,15	20	1,00
Celkem:	20	1,00	20	1,00

Hodnoty nejsou seřazeny podle daného kritéria

Použití nesprávných charakteristik pro daný typ proměnné

Ordinální proměnná: např. známky ve škole (výborně, chvalitebně, dobře,…) má smysl uspořádání např. od nejlepší k nejhorší => používáme absolutní a relativní četnost + kumulativní charakteristiky, modus , výsečový graf, histogram, Lorenzova křivka, Paretův graf

Tabulka rozdělení četnosti výsledné známky ze statistiky				
Hodnoty	Absolutní četnosti	Relativní četnosti	Kumulativní absolutní četnosti	Kumulativní relativní četnosti
výborně	4	0,20	4	0,20
velmi dobře	8	0,40	12	0,60
dobře	5	0,25	17	0,85
nevyhověl	3	0,15	20	1,00
Celkem:	20	1,00	20	1,00

17 studentů, tj. 85 % u zkoušky ze statistiky prospělo

Použití nesprávných charakteristik pro daný typ proměnné

Numerická proměnná – není možné použít četnosti u numerické proměnné – teoreticky se každá hodnota v souboru může od všech ostatních lišit, potom bychom dostali tabulku se všemi hodnotami a u nich četnost 1 (na co by taková analýza byla?)

Můžeme roztřídit do kategorií a ty poté analyzovat.

Takhle ne

Tabulka rozdělení četnosti výšky studentů 4. ročníku		
Hodnoty [cm]	Absolutní četnosti	Relativní četnosti
168	1	0,029
169	1	0,029
170	2	0,057
171	1	0,029
172	1	0,029
… o 17 řádků dále		
199	2	0,057
200	1	0,029
201	1	0,029
celkem	34	1

Použití nesprávných charakteristik pro daný typ proměnné

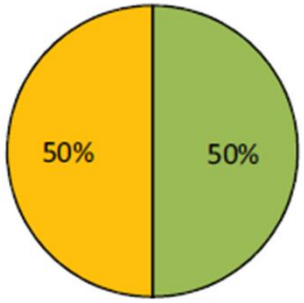
Numerická proměnná – není možné použít četnosti u numerické proměnné – teoreticky se každá hodnota v souboru může od všech ostatních lišit, potom bychom dostali tabulku se všemi hodnotami a u nich četnost 1 (na co by taková analýza byla?)

Můžeme roztřídit do kategorií a ty poté analyzovat.

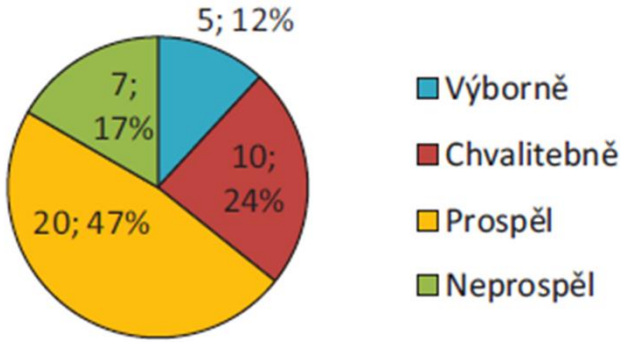
Tabulka rozdělení četnosti výšky studentů 4. ročníku		
Hodnoty [cm]	Absolutní četnosti	Relativní četnosti
160-169	2	0,057
170-179	11	0,314
180-189	10	0,286
190-199	11	0,314
Nad 200	1	0,029
celkem	35	1

Zavádějící grafy

Neuvedení absolutních četností u koláčových grafů (popř. celkový rozsah souboru v blízkosti grafu)

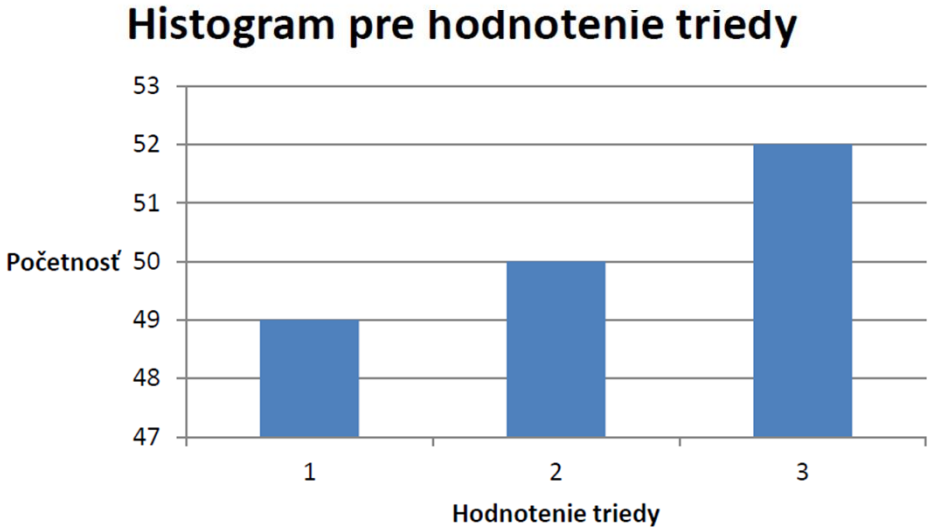


<-NE | ANO ->



Posunuté osy u histogramů:

kdyby osa y začínala na 0, byl by rozdíl ve výškách sloupců zanedbatelný, takhle to vypadá, že se četnosti jednotlivých odnocení výrazně liší



Neokomentované a zbytečné výstupy

Každý výstup (graf, tabulka) by měly být:

- okomentované – co jsme danou analýzou zjistili, jaké jsou hlavní ukazatele, co to v kontextu se zpracovávanými daty a problematikou znamená? (tzn. v tabulce okomentovaná také každá charakteristika)
- opodstatněné – vybereme pouze takový typ výstupu, který dokáže sdělit nejvíce informací
  - velmi častá zbytečnost rozptylogramu (scatterplot) u jednorozměrné analýzy, většinou lépe poslouží histogram
  - velké množství např. koláčových grafů (lepší je použít jeden 100% skládaný pruhový graf)
  - velké množství krabicových grafů pro každou kategorii stejné numerické proměnné zvlášť (lepší je do jednoho grafu vykreslit všechny krabicové grafy – lze lépe pozorovat rozdíly nebo podobnosti mezi jednotlivými kategoriemi)
  - příliš velké grafy (např. 1 koláčový graf nebo 1 histogram přes celou stránku, šetříme papír a barvu do tiskárny)

Cizojazyčné popisy ve výstupech

Jestliže píšu projekt (domácí úkol, zprávu, závěrečnou práci,…) v určitém jazyce, budou v tom samém jazyce i popisy grafů (Box-and-Whiskers plot → krabicový graf) a os, názvy charakteristik v tabulkách (mean → průměr) a naměřené hodnoty (red → červená)

Jestliže používám Statgraphics, nastavím správný font pro českou diakritiku:

- Kliknout pravým tlačítkem myši na graf → Graphics Option → Top Title (X-Axis, Y-Axis) → Line 1 Fonts (Title Fonts) → Script nastavit na Středoevropské (1 nahoru od Západního)

Nesprávné zaokrouhlení číselných charakteristik

Jak zaokrouhlit číselné charakteristiky:

- směrodatná odchylka směrem nahoru na 2 platné cifry (pozor 2 platné cifry neznamenají na 2 řády)
- průměr, rozptyl, kvantily na stejný řád jako směrodatná odchylka
- min, max – na stejný řád, v jakém provádíme měření
- počet prvků ve výběru – celé číslo
- variační koeficient, šikmost, špičatost – max. na 3 desetinná místa

Př. 1. Zpracování hodnot bílkovinné sérum z dat na cvičení (jednovýb. testy)

Výstup ze Statgraphicsu

```
Summary Statistics for Col_6

Count = 218
Average = 34,4868
Median = 34,5105
Variance = 0,154986
Standard deviation = 0,393683
Minimum = 33,57
Maximum = 35,485
Range = 1,915
Lower quartile = 34,195
Upper quartile = 34,736
Stnd. skewness = -0,170919
Stnd. kurtosis = -1,6797
```

Nesprávné přepsání hodnot do tabulky

Počet	218	Směr. odchylka	0,393683
Průměr	34,4868	Minimum	33,57
Medián	34,5105	Maximum	35,485
Dolní kvartil	34,195	Šikmost	-0,170919
Horní kvartil	34,736	Špičatost	-1,6797

Nesprávné zaokrouhlení číselných charakteristik

Jak zaokrouhlit číselné charakteristiky:

- směrodatná odchylka směrem nahoru na 2 platné cifry (pozor 2 platné cifry neznamenají na 2 řády)
- průměr, rozptyl, kvantily na stejný řád jako směrodatná odchylka
- min, max – na stejný řád, v jakém provádíme měření
- počet prvků ve výběru – celé číslo
- variační koeficient, šikmost, špičatost – max. na 3 desetinná místa

Př. 1. Zpracování hodnot bílkovinné sérum z dat na cvičení (jednovýb. testy)

Výstup ze Statgraphicsu

```
Summary Statistics for Col_6

Count = 218
Average = 34,4868
Median = 34,5105
Variance = 0,154986
Standard deviation = 0,393683
Minimum = 33,57
Maximum = 35,485
Range = 1,915
Lower quartile = 34,195
Upper quartile = 34,736
Stnd. skewness = -0,170919
Stnd. kurtosis = -1,6797
```

Počet	218	Směr. odchylka	0,40
Průměr	34,49	Minimum	33,570
Medián	34,51	Maximum	35,485
Dolní kvartil	34,20	Šikmost	-0,171
Horní kvartil	34,74	Špičatost	-1,680

Nesprávné zaokrouhlení číselných charakteristik

Jak zaokrouhlit číselné charakteristiky:

- směrodatná odchylka směrem nahoru na 2 platné cifry (pozor 2 platné cifry neznamenají na 2 řády)
- průměr, rozptyl, kvantily na stejný řád jako směrodatná odchylka
- min, max – na stejný řád, v jakém provádíme měření
- počet prvků ve výběru – celé číslo
- variační koeficient, šikmost, špičatost – max. na 3 desetinná místa

Př. 2. Zpracování hodnot doba přežití z dat na cvičení (jednovýb. testy)

Výstup ze Statgraphicsu

```
Summary Statistics for Col_7

Count = 100
Average = 29,23
Median = 21,0
Variance = 761,391
Standard deviation = 27,5933
Minimum = 3,0
Maximum = 177,0
Range = 174,0
Lower quartile = 11,5
Upper quartile = 37,0
Stnd. skewness = 10,8897
Stnd. kurtosis = 19,4207
```

Nesprávné přepsání hodnot do tabulky

Počet	100	Směr. odchylka	27,5933
Průměr	29,23	Minimum	3,0
Medián	21,0	Maximum	177,0
Dolní kvartil	11,5	Šikmost	10,8897
Horní kvartil	37,0	Špičatost	19,4207

Nesprávné zaokrouhlení číselných charakteristik

Jak zaokrouhlit číselné charakteristiky:

- směrodatná odchylka směrem nahoru na 2 platné cifry (pozor 2 platné cifry neznamenají na 2 řády)
- průměr, rozptyl, kvantily na stejný řád jako směrodatná odchylka
- min, max – na stejný řád, v jakém provádíme měření
- počet prvků ve výběru – celé číslo
- variační koeficient, šikmost, špičatost – max. na 3 desetinná místa

Př. 2. Zpracování hodnot doba přežití z dat na cvičení (jednovýb. testy)

Výstup ze Statgraphicsu

```
Summary Statistics for Col_7

Count = 100
Average = 29,23
Median = 21,0
Variance = 761,391
Standard deviation = 27,5933
Minimum = 3,0
Maximum = 177,0
Range = 174,0
Lower quartile = 11,5
Upper quartile = 37,0
Stnd. skewness = 10,8897
Stnd. kurtosis = 19,4207
```

Počet	100	Směr. odchylka	28
Průměr	29	Minimum	3
Medián	21	Maximum	177
Dolní kvartil	12	Šikmost	10,9
Horní kvartil	37	Špičatost	19,4

Nesprávné zaokrouhlení číselných charakteristik

Jak zaokrouhlit číselné charakteristiky:

- směrodatná odchylka směrem nahoru na 2 platné cifry (pozor 2 platné cifry neznamenají na 2 řády)
- průměr, rozptyl, kvantily na stejný řád jako směrodatná odchylka
- min, max – na stejný řád, v jakém provádíme měření
- počet prvků ve výběru – celé číslo
- variační koeficient, šikmost, špičatost – max. na 3 desetinná místa

Př. 3. Zpracování hodnot osmolalita 8^{00} h z dat na cvičení (dvouvýb. testy)

Výstup ze Statgraphicsu

```
Summary Statistics for Col_8

Count = 16
Average = 577,75
Median = 477,5
Variance = 56777,4
Standard deviation = 238,28
Minimum = 300,0
Maximum = 1300,0
Range = 1000,0
Lower quartile = 456,0
Upper quartile = 616,0
Stnd. skewness = 3,5419
Stnd. kurtosis = 4,376
```

Nesprávné přepsání hodnot do tabulky

Počet	16	Směr. odchylka	238,28
Průměr	577,75	Minimum	300,0
Medián	477,5	Maximum	1300,0
Dolní kvartil	456,0	Šikmost	3,5419
Horní kvartil	616,0	Špičatost	4,376

Nesprávné zaokrouhlení číselných charakteristik

Jak zaokrouhlit číselné charakteristiky:

- směrodatná odchylka směrem nahoru na 2 platné cifry (pozor 2 platné cifry neznamenají na 2 řády)
- průměr, rozptyl, kvantily na stejný řád jako směrodatná odchylka
- min, max – na stejný řád, v jakém provádíme měření
- počet prvků ve výběru – celé číslo
- variační koeficient, šikmost, špičatost – max. na 3 desetinná místa

Př. 3. Zpracování hodnot osmolalita 8^{00} h z dat na cvičení (dvouvýb. testy)

Výstup ze Statgraphicsu

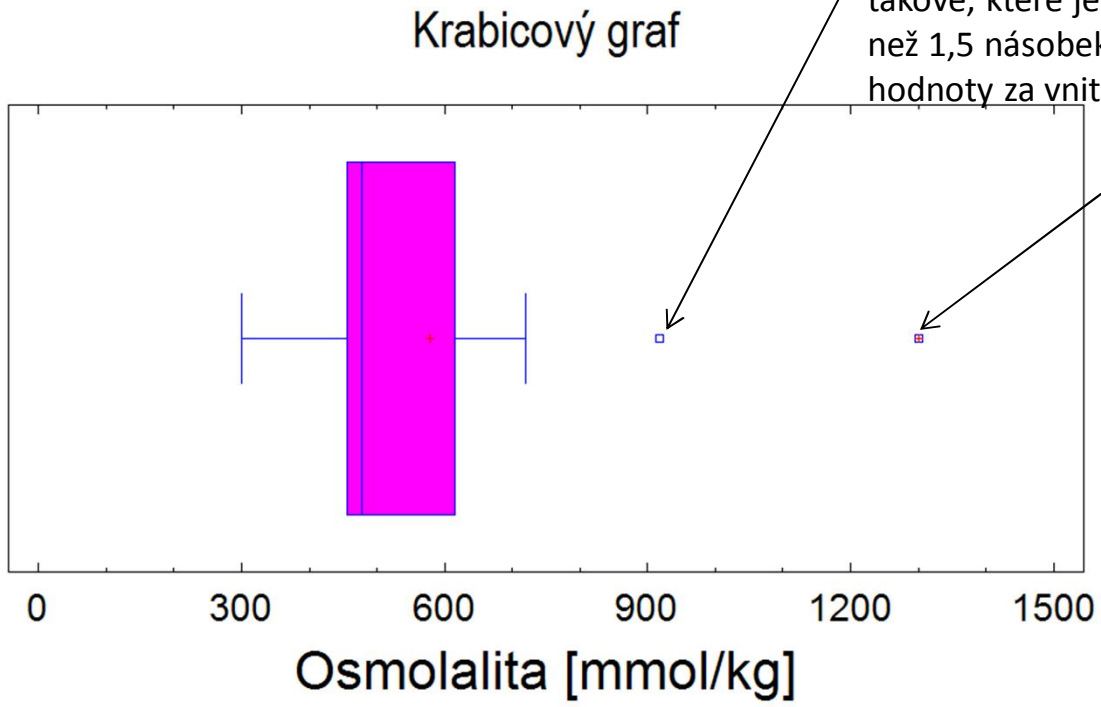
```
Summary Statistics for Col_8

Count = 16
Average = 577,75
Median = 477,5
Variance = 56777,4
Standard deviation = 238,28
Minimum = 300,0
Maximum = 1300,0
Range = 1000,0
Lower quartile = 456,0
Upper quartile = 616,0
Stnd. skewness = 3,5419
Stnd. kurtosis = 4,376
```

Počet	16	Směr. odchylka	240
Průměr	580	Minimum	300
Medián	480	Maximum	1300
Dolní kvartil	460	Šikmost	3,54
Horní kvartil	620	Špičatost	4,38

Odlehlá pozorování

Máme několik kritérií pro posouzení odlehlých pozorování (viz skripta str. 32): vnitřní (resp. vnější) hradby, *z*-souřadnice, $x_{0,5}$-souřadnice. Obecně je *z*-souřadnice méně přísná než vnitřní hradby. Co vykresluje Statgraphics v krabicovém grafu = odlehlé a extrémní pozorování podle vnitřních (resp. vnějších) hradeb.



Odlehlé pozorování (označeno čtverečkem) – takové, které je vzdáleno od horního kvartilu o více než 1,5 násobek interkvartilového rozpětí, tzn. hodnoty za vnitřními hradbami

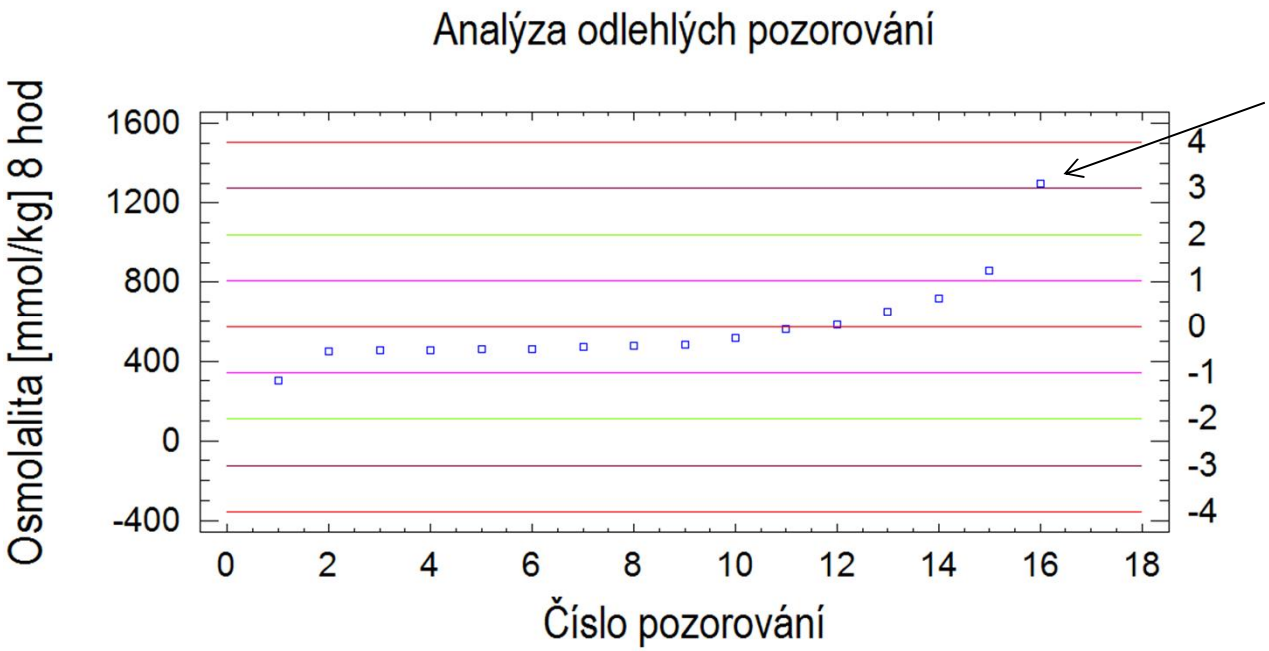
Extrémní pozorování (označeno čtverečkem s křížkem) – takové, které je vzdáleno od horního kvartilu o více než 3 násobek interkvartilového rozpětí, tzn. hodnoty za vnějšími hradbami

Odlehlá pozorování

Musíte se rozhodnout, které kritérium použijete. Pokud označíte nějakou hodnotu jako odlehlé pozorování, musíte zhodnotit důvod přítomnosti v datech (např. chyba v zápise, nebo důsledek chybného měření) a postup, co s ním dále uděláte (viz skripta str. 33).

Jak ve Statgraphicsu určit odlehlá pozorování podle *z*-souřadnice:

Describe → Numeric Data → Outlier Identification



Odlehlé pozorování je taková hodnota, která je vzdálena od průměru o více než ± trojnásobek směrodatné odchylky

Vodorovné čáry znázorňují hodnotu, které je od průměru (vodorovná čára v 0) vzdálena *k* násobek směrodatné odchylky

Data nepříliš vhodná na projekt

Příklad – ukázka semestrálního projektu ze statistiky
http://homel.vsb.cz/~lit40/STA1/Materialy/Strakova_SMAD.pdf

2000	2005	2010
2591	2501	1323
8849	6274	7389
6787	5487	2629
10976	9369	9157
8891	7646	8502
13532	12121	10499
6983	5514	4179
7985	8046	7876
4536	4068	3103
7707	6177	4903

1) Jedná se o data závislá: pro 1 okres máme vždy 3 údaje v jednotlivých letech, tzn. stav dobytka je závislý např. na velikosti okresu (pokud mám malý okres, je v něm ve všech třech letech méně dobytka než ve velikých okresech)-> z pohledu metod týkajících se našeho předmětu bychom měli použít Friedmanův test (namísto nesprávně použité ANOVY)
2) daleko vhodnější je ale použít metody na zpracování časových řad (údaje v několika letech), které nejsou náplní studia našeho předmětu

Doporučení: zvolit jiný projekt.