Průzkumová analýza dat

B02907 Informační a komunikační technologie

Lubomír Štěpánek,
Ústav biofyziky a informatiky
1. LF UK

Průzkumová analýza dat

- Exploratory Data Analysis
- prozkoumání souboru dat, především *graficky*, vyčíslení hlavních ukazatelů polohy, variability, tvaru

Co vše by měla zahrnout

- kontrola konceptu „tidy data“
- určení typu dat každé proměnné
- nepřítomnost některých pozorování (sparse data)
- přítomnost odlehlých a extrémních hodnot
- pravděpodobnostní rozdělení každé proměnné
- charakteristiky polohy, variability, (tvaru) každé proměnné
- trendy závislostí mezi proměnnými

Co vše by měla zahrnout

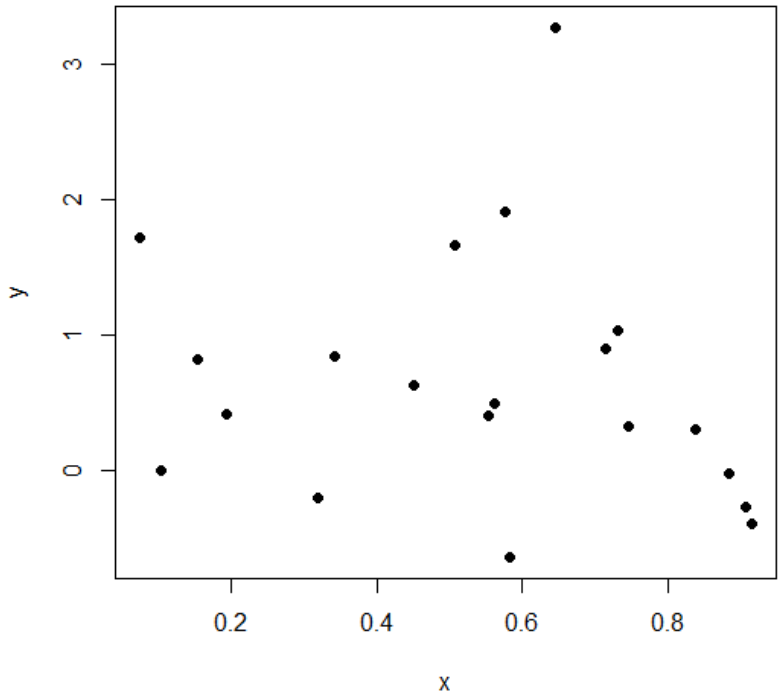
- ideálně vše numericky i graficky!

Grafy („diagramy“)

- přepsání kvantifikovaných údajů (číselných nebo „seřazených“) do soustavy geometrických obrazců
- oproti tabulce zjednodušující, ale přehlednější
- obvykle dvourozměrné – dvě stupnice na kolmých osách (stupnice *přímočaré X křivočaré, rovnoměrné X nerovnoměrné*)
- vzácně polární souřadnice
- dbáme na název grafu, popis os, vhodné měřítko, jednotky, počet desetinných míst

Bodový graf (scatterplot)

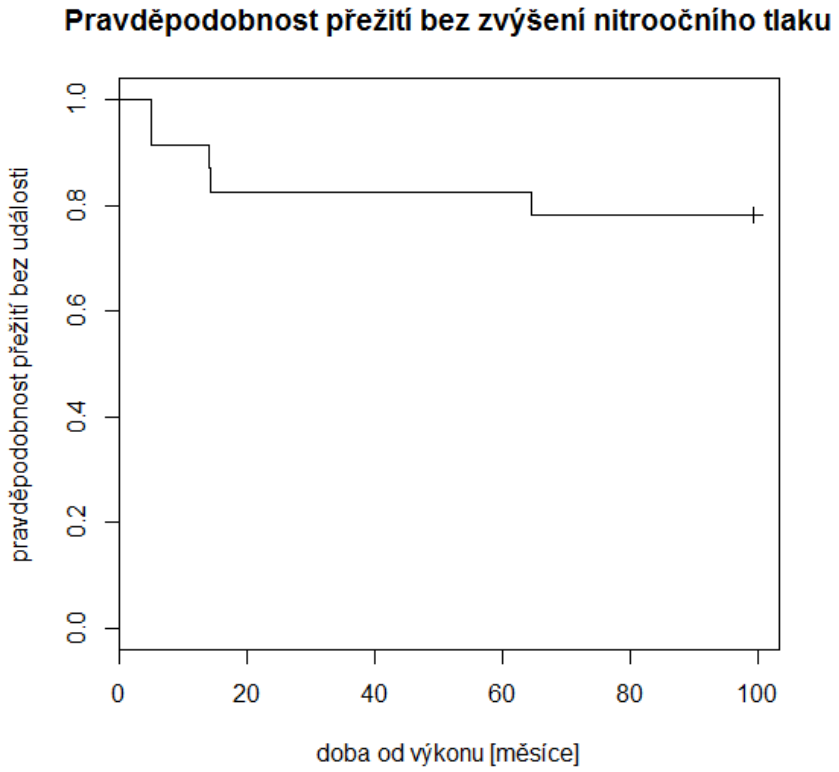
- u každého prvku známe dva znaky, nanášíme do pravoúhlé soustavy souřadnic
- vhodný pro zkoumání závislosti dvou znaků (korelace, regrese)



http://new.euromise.org/czech/tajne/ucebnice/html/html/img18.gif

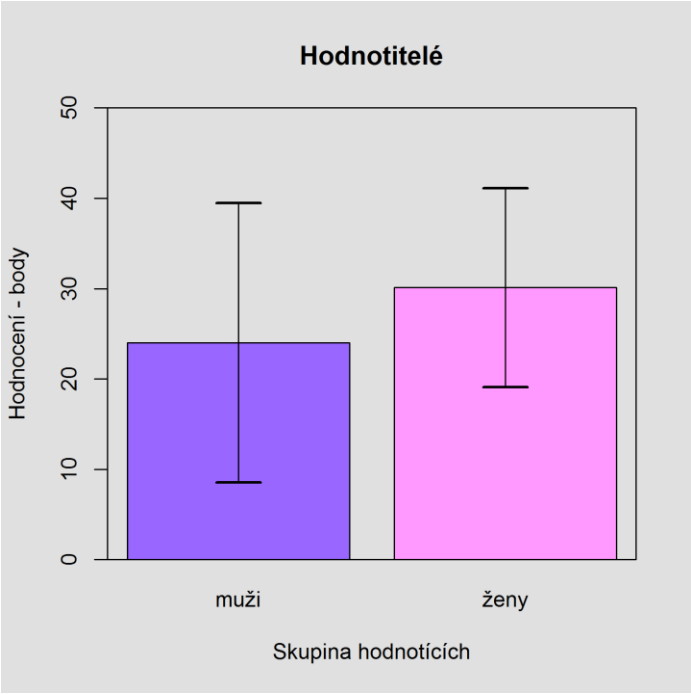
Spojnicový graf (line chart)

- proložíme-li bodovým grafem křivku
- často průběh časové řady či naznačení trendu spojitého znaku (např. Kaplan-Meierova křivka)



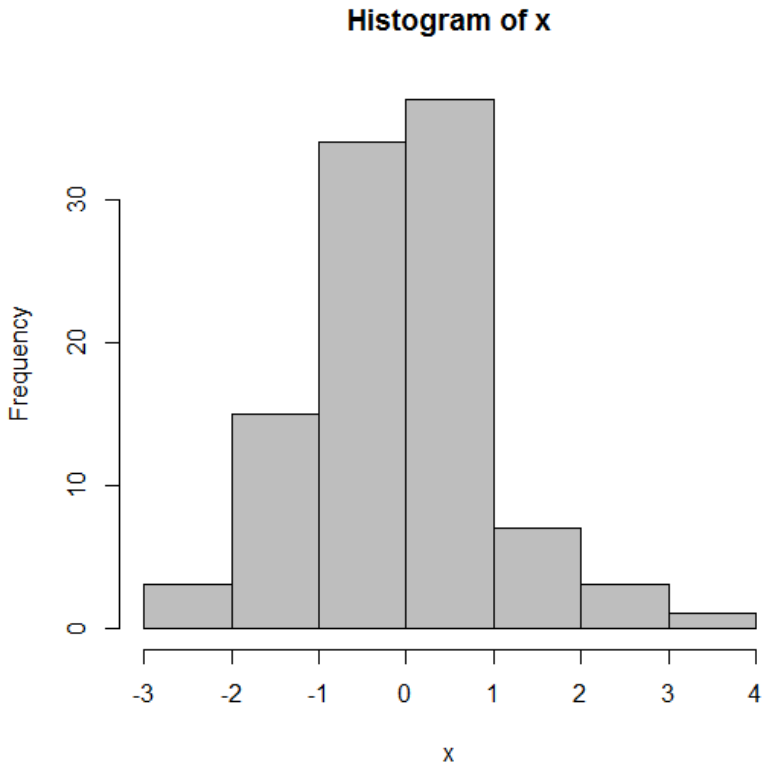
Sloupcový graf (barplot)

- na jedné ose třídy znaku (kvalitativního či kvantitativního), na druhé obvykle četnosti tříd
- sloupce svislé či vodorovné
- *error bar graph* – sloupcový graf s *chybovými úsečkami* („fousy“) = násobky směrodatné odchylky výběru



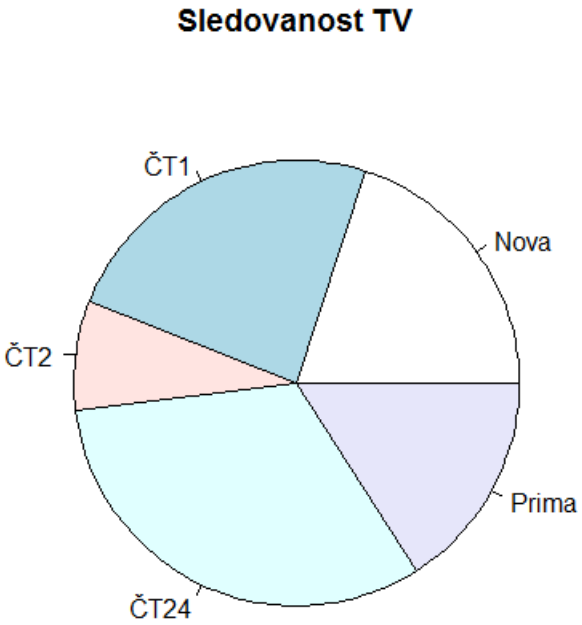
Histogram

- na vodorovné ose třídy kvantitativního spojitého znaku (nemusí být stejně velké), na svislé absolutní či méně častěji relativní četnosti tříd



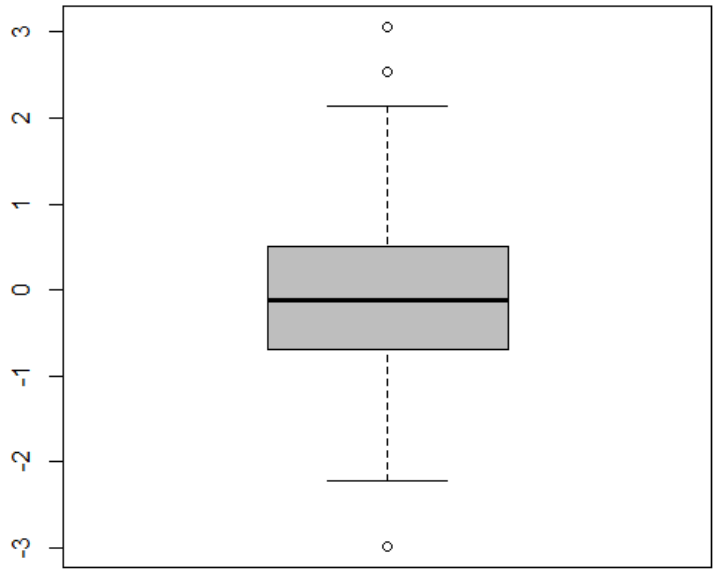
Kruhový graf (pie chart)

- = výsečový, koláčový; vhodný pro zobrazení relativních četností (v %) tříd kvalitativního či kvantitativního znaku



Krabicový graf (boxplot)

- box graph, whisker plot; pro data s nenormálním, negaussovským rozložením („asymetrická“)
- zobrazuje kvartily a medián, odlehlé hodnoty



Odlehlé hodnoty

- v souboru metrických dat mohou existovat tzv. odlehlé a extrémní hodnoty
- x je odlehlá (outlier value), pokud:

$$x < Q_1 - 1{,}5 \cdot (Q_3 - Q_1) \text{ nebo } x > Q_3 + 1{,}5 \cdot (Q_3 - Q_1)$$

- x je extrémní hodnota (extreme value), pokud:

$$x < Q_1 - 3{,}0 \cdot (Q_3 - Q_1) \text{ nebo } x > Q_3 + 3{,}0 \cdot (Q_3 - Q_1)$$

- s extrémními hodnotami dále nepočítáme, odlehlé je dobré přeměřit, případně také vyřadit

Chybějící hodnoty

- missing values
- často jako NA (Not Available)
- nejsou nuly!
- pozorování s chybějícími hodnotami vyřazeno, nebo NAs nahrazeny např. průměrem zbylých hodnot