Základy statistiky

Statistika kolem nás
Základní pojmy statistické analýzy dat

Adéla Vrtková, Martina Litschmannová

adela.vrtkova@vsb.cz, martina.litschmannova@vsb.cz

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY | KATEDRA APLIKOVANÉ MATEMATIKY

Statistika je...

Statistika je...

- odvětví matematiky zabývající se sběrem, analýzou, interpretací a prezentací dat (Merriam-Webster.com)

Statistika je...

- odvětví matematiky zabývající se sběrem, analýzou, interpretací a prezentací dat (Merriam-Webster.com)
- věda, která se snaží zkoumat reálná data a s pomocí teorie pravděpodobnosti se tato data snaží popisovat (Matematika.cz)

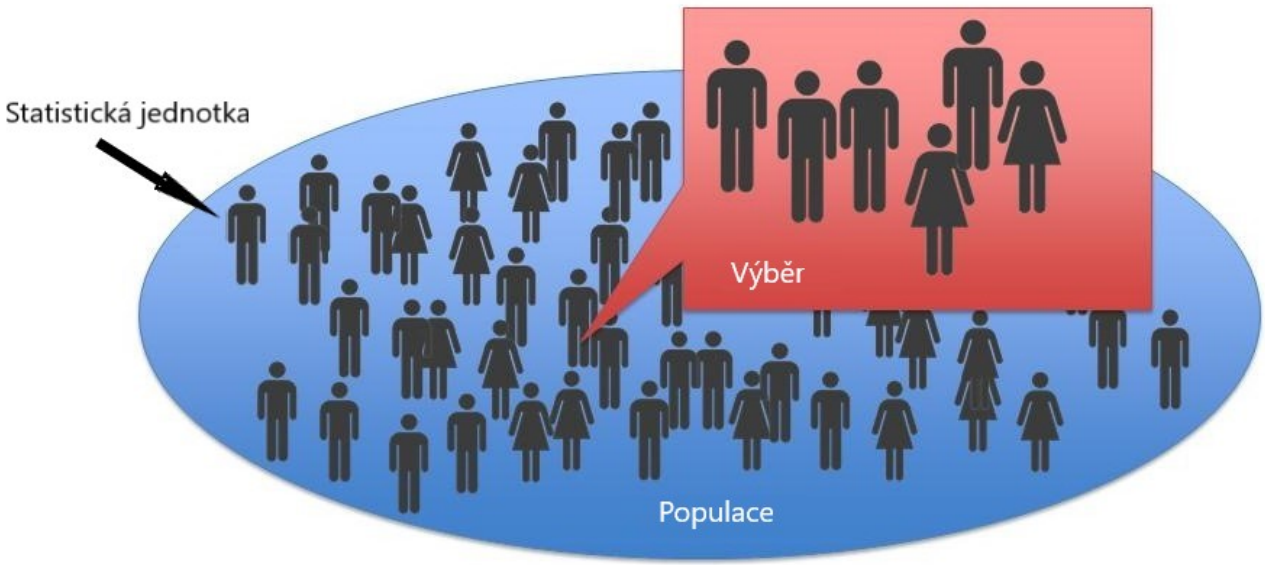
Statistika je...

- odvětví matematiky zabývající se sběrem, analýzou, interpretací a prezentací dat (Merriam-Webster.com)
- věda, která se snaží zkoumat reálná data a s pomocí teorie pravděpodobnosti se tato data snaží popisovat (Matematika.cz)
- přesný součet nepřesných čísel (Poznámka k lidové definici statistiky, časopis Vesmír 5/1998)

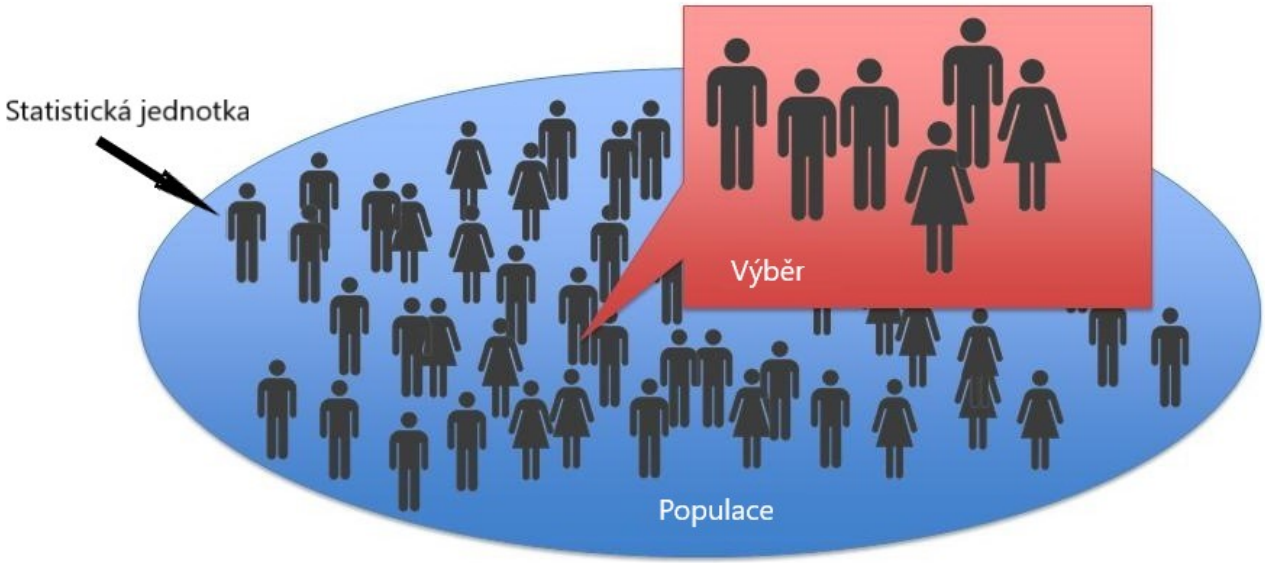
Statistika je...

- odvětví matematiky zabývající se **sběrem, analýzou, interpretací a prezentací dat** (Merriam-Webster.com)
- věda, která se snaží zkoumat reálná data a s pomocí teorie pravděpodobnosti se tato data snaží popisovat (Matematika.cz)
- přesný součet nepřesných čísel (Poznámka k lidové definici statistiky, časopis Vesmír 5/1998)

Populace a výběr



Populace a výběr



- **populace** - všechny statistické jednotky
- **výběr** - podmnožina populace

Není výběr jako výběr...

Představte si, že chcete provést průzkum spokojenosti studentů VŠB-TUO se službami knihovny. Asi není možné se zeptat úplně všech studentů univerzity, resp. těch, kteří služby knihovny využívají. Jak tedy provést výběr studentů do průzkumu?

Není výběr jako výběr...

Představte si, že chcete provést průzkum spokojenosti studentů VŠB-TUO se službami knihovny. Asi není možné se zeptat úplně všech studentů univerzity, resp. těch, kteří služby knihovny využívají. Jak tedy provést výběr studentů do průzkumu?

- Způsoby provedení výběru se zabývá **teorie výběrového šetření** (Steven K. Thompson: Sampling).

Není výběr jako výběr...

Představte si, že chcete provést průzkum spokojenosti studentů VŠB-TUO se službami knihovny. Asi není možné se zeptat úplně všech studentů univerzity, resp. těch, kteří služby knihovny využívají. Jak tedy provést výběr studentů do průzkumu?

- Způsoby provedení výběru se zabývá **teorie výběrového šetření** (Steven K. Thompson: Sampling).
- Ta úzce souvisí s **metodami sběru dat**.

Není výběr jako výběr...

Představte si, že chcete provést průzkum spokojenosti studentů VŠB-TUO se službami knihovny. Asi není možné se zeptat úplně všech studentů univerzity, resp. těch, kteří služby knihovny využívají. Jak tedy provést výběr studentů do průzkumu?

- Způsoby provedení výběru se zabývá **teorie výběrového šetření** (Steven K. Thompson: Sampling).
- Ta úzce souvisí s **metodami sběru dat**.
- Cílem je získání **reprezentativního výběru**, což je takový výběr, který má stejné vlastnosti (strukturu) jako populace.

Pravděpodobnostní výběry

Prostý náhodný výběr (bez vracení / s vracením)

- los z osudí, generátor pseudonáhodných čísel

Systematický výběr

- výběr z populace podle pevně zvoleného kroku - každou třetí jednotku, každou desátou, apod.

Stratifikovaný výběr

- rozklad populace na disjunktní oblasti (strata) dle určitého kritéria; v každé oblasti pak prostý náhodný výběr

Skupinkový výběr

- populace rozdělena do homogenních skupin, prostý náhodný nebo systematický výběr několika skupin, sledování celých vybraných skupin

Vícestupňový výběr

- analogie skupinkového výběru, u vybraných skupin podskupiny atd.

Výběr s nestejnými pravděpodobnostmi

- statistické jednotky mají různou pravděpodobnost zahrnutí do výběru

Nepravděpodobnostní výběry

Kvótní výběr

- snaha naplnit předem dané kvóty (věkové kategorie, pohlaví apod.), nutná znalost struktury populace

Snowball technika

- výběr malého souboru jednotek splňující kritéria, skrze ně získání kontaktů na další

Účelový výběr

- založen zcela na úsudku výzkumníka, nelze očekávat zobecnitelné výsledky

Příležitostný výběr

- „všichni, kteří jsou po ruce“

Samovýběr

- účast na základě rozhodnutí statistické jednotky

Metody sběru dat

- rozhovory
- dotazníková šetření a ankety
- pozorování
- ohniskové skupiny (focus groups)
- případové studie
- sekundární data, analýza dokumentů
- experimenty, měření

Metody sběru dat

- rozhovory
- **dotazníková šetření a ankety**
- pozorování
- ohniskové skupiny (focus groups)
- případové studie
- sekundární data, analýza dokumentů
- **experimenty, měření**

prostý náhodný výběr

systematický výběr

stratifikovaný výběr

skupinkový výběr

vícestupňový výběr

výběr s nestejnými p-stmi

kvótní výběr

snowball technika

účelový výběr

příležitostný výběr

samovýběr

rozhovory

dotazníková šetření

pozorování

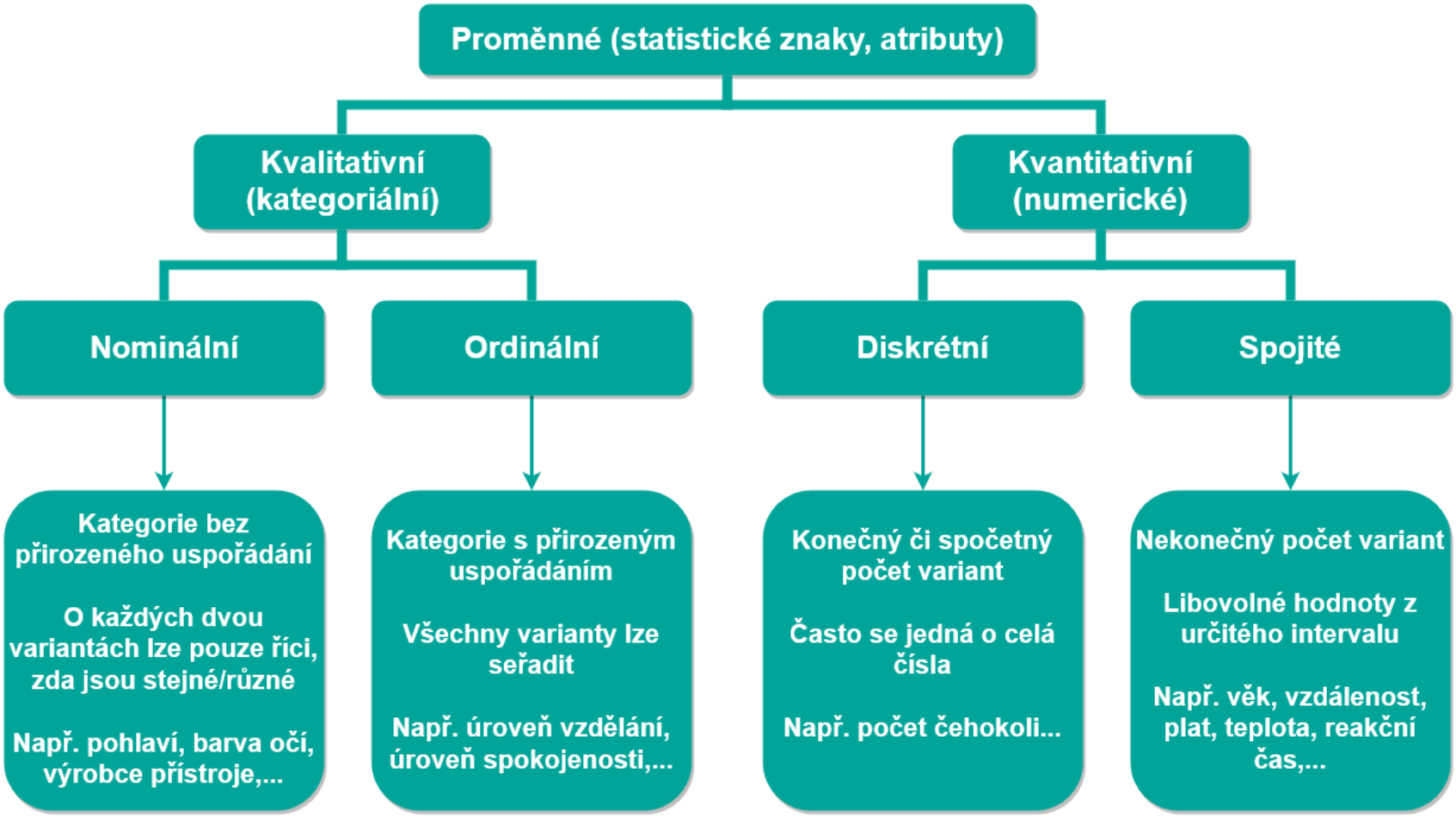
ohniskové skupiny

případové studie

sekundární data

experimenty, měření

Co sbíráme?


Proměnné (statistické znaky, atributy)
Kvalitativní (kategoriální)
Kvantitativní (numerické)
Nominální
Ordinální
Diskrétní
Spojité
Kategorie bez přirozeného uspořádání
O každých dvou variantách lze pouze říci, zda jsou stejné/různé
Např. pohlaví, barva očí, výrobce přístroje,...
Kategorie s přirozeným uspořádáním
Všechny varianty lze seřadit
Např. úroveň vzdělání, úroveň spokojenosti,...
Konečný či spočetný počet variant
Často se jedná o celá čísla
Např. počet čehokoli...
Nekonečný počet variant
Libovolné hodnoty z určitého intervalu
Např. věk, vzdálenost, plat, teplota, reakční čas,...

Datová matice - jak údaje zapisovat

Provedli jsme výběr statistických jednotek, sebrali jsme data, jak je ale zapsat?

Datová matice - jak údaje zapisovat

Provedli jsme výběr statistických jednotek, sebrali jsme data, jak je ale zapsat?

- 1 řádek = 1 statistická jednotka (člověk, měření,...)
- 1 sloupec = 1 proměnná

Datová matice - jak údaje zapisovat

Provedli jsme výběr statistických jednotek, sebrali jsme data, jak je ale zapsat?

- 1 řádek = 1 statistická jednotka (člověk, měření,...)
- 1 sloupec = 1 proměnná

	Proměnná 1	...	Proměnná p
1	x_{11}	...	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	...	x_{np}

Datová matice - jak údaje zapisovat

- 1 řádek = 1 statistická jednotka (člověk, měření,...)
- 1 sloupec = 1 proměnná

ID	věk	hmotnost_před (kg)	hmotnost_po (kg)	výška (cm)	pohlaví
1	23	73,4	70,0	181	M
2	22	79,2	78,6	189	M
3	22	71,9	69,9	167	Ž
4	26	66,3	63,8	165	Ž
5	24	80,3	75,6	170	Ž
⋮	⋮	⋮	⋮	⋮	⋮

Ostatní formáty

- jakákoli jiná struktura dat
- pro tyto jiné struktury budeme používat termín *tabulka* nebo *datová tabulka*

muži				ženy			
věk	hmotnost_před (kg)	hmotnost_po (kg)	výška (cm)	věk	hmotnost_před (kg)	hmotnost_po (kg)	výška (cm)
23	83,4	80,0	175	25	73,4	70,0	165
22	99,2	88,6	182	23	69,2	68,6	169
22	81,9	79,9	190	21	71,9	69,9	172
26	89,3	83,8	182	27	69,3	63,8	160
24	88,3	83,6	180	20	78,3	73,6	170
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

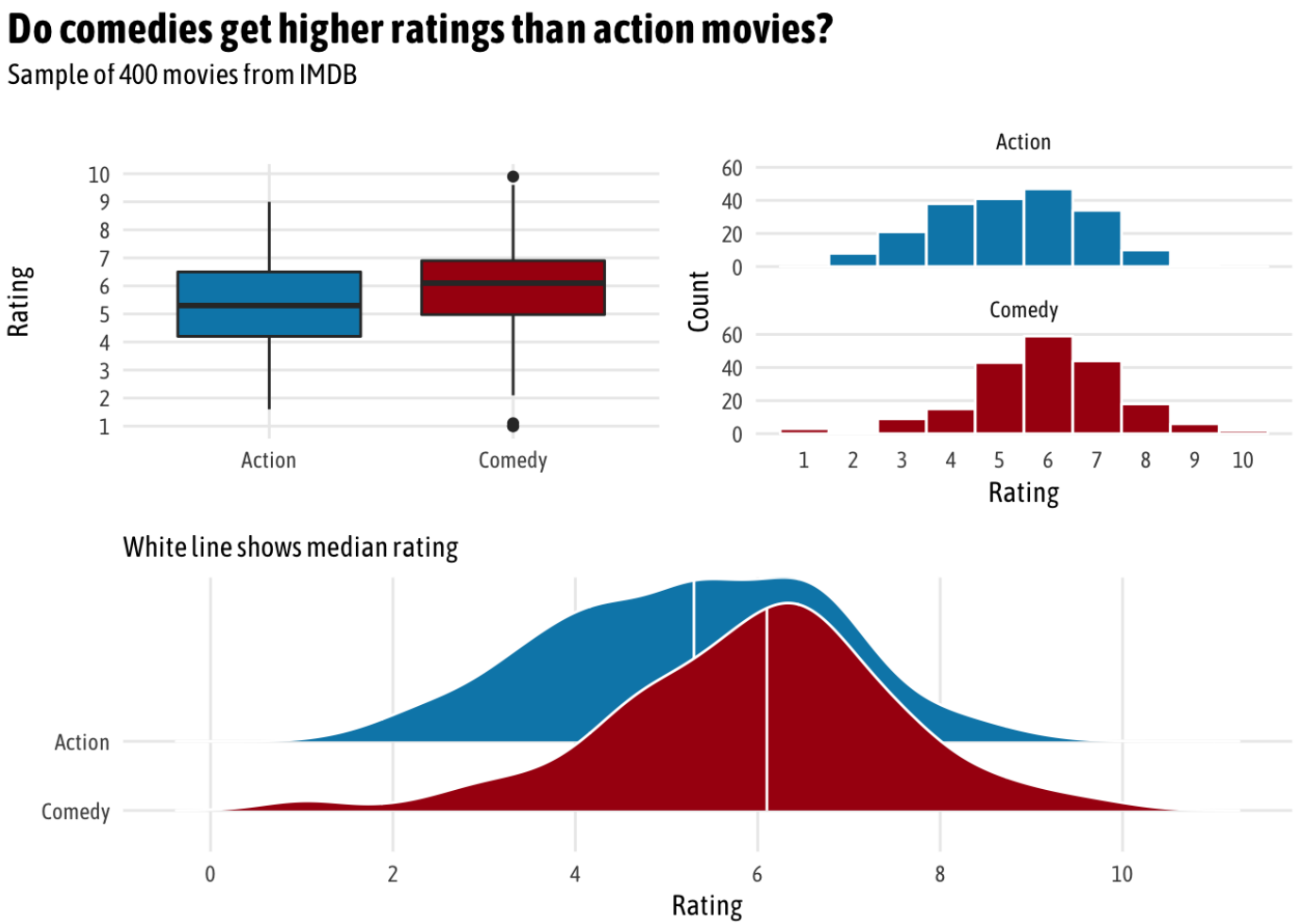
Máme data, a co dál?

Konečně máme sesbíraná a správně zapsaná data. Co s nimi?

- exploratční analýza dat
- aplikace metod statistické indukce

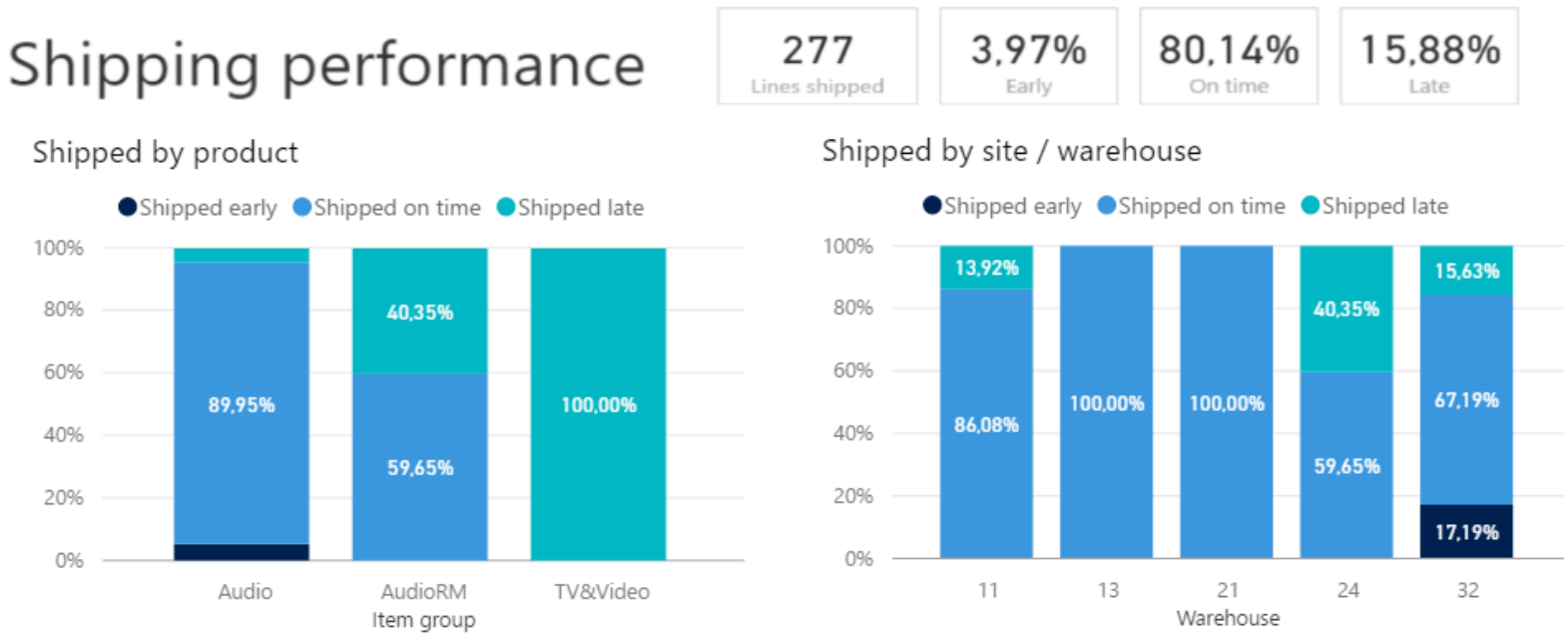
Explorační analýza dat

= popisná statistika, cílí na **prvotní průzkum dat**, nalezení **chyb**, **odlehlých pozorování**, získání přehledu o tom, s jakými daty pracujeme. Díky závěrům explorační analýzy můžeme také **zformulovat statistické hypotézy** nebo dopředu **odhadnout odpověď** na některé předem definované výzkumné otázky. Zároveň je nutnou součástí **analýzy předpokladů** statistických metod.



Zdroj: www.andrewheiss.com/blog/2019/01/29/diff-means-half-dozen-ways/

Exploračná analýza dat

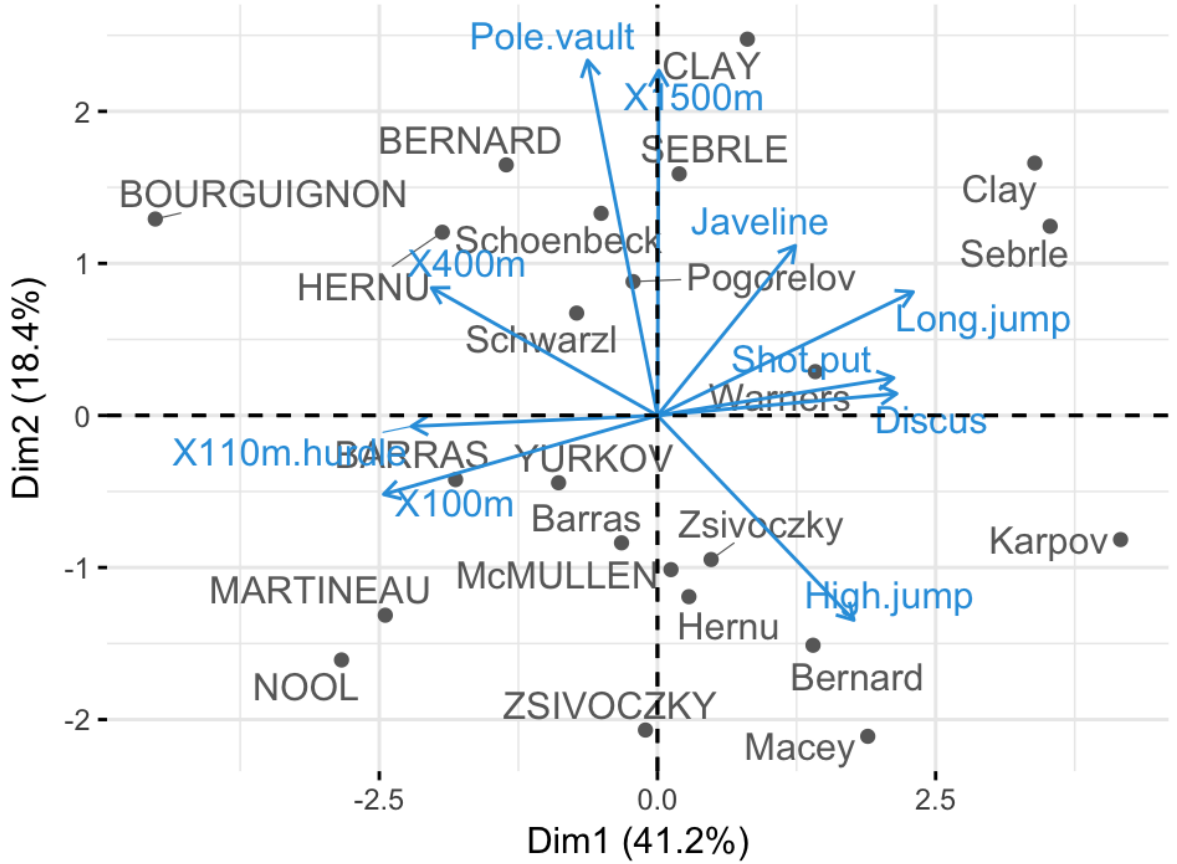


Zdroj: ax-dynamics.com/article/4-more-power-bi-content-packs-for-2017

Aplikace metod statistické indukce

Z těch nejznámějších metod spadajících pod tuto oblast uvedeme např. **testování hypotéz, regresní modely, metodu hlavních komponent, shlukovou analýzu nebo analýzu přežití**. Jedná se prakticky o všechny metody, kdy jsou **na základě analýzy výběru vyvozovány závěry o populaci**.

Metoda hlavních komponent



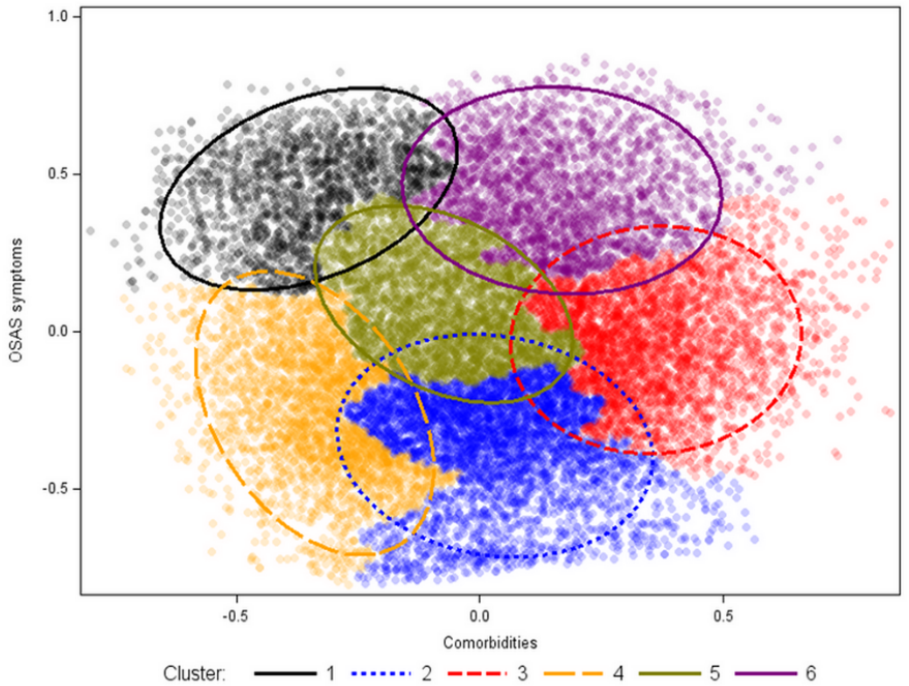
Zdroj: www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

Shluková analýza

Obstructive Sleep Apnea: A Cluster Analysis at Time of Diagnosis

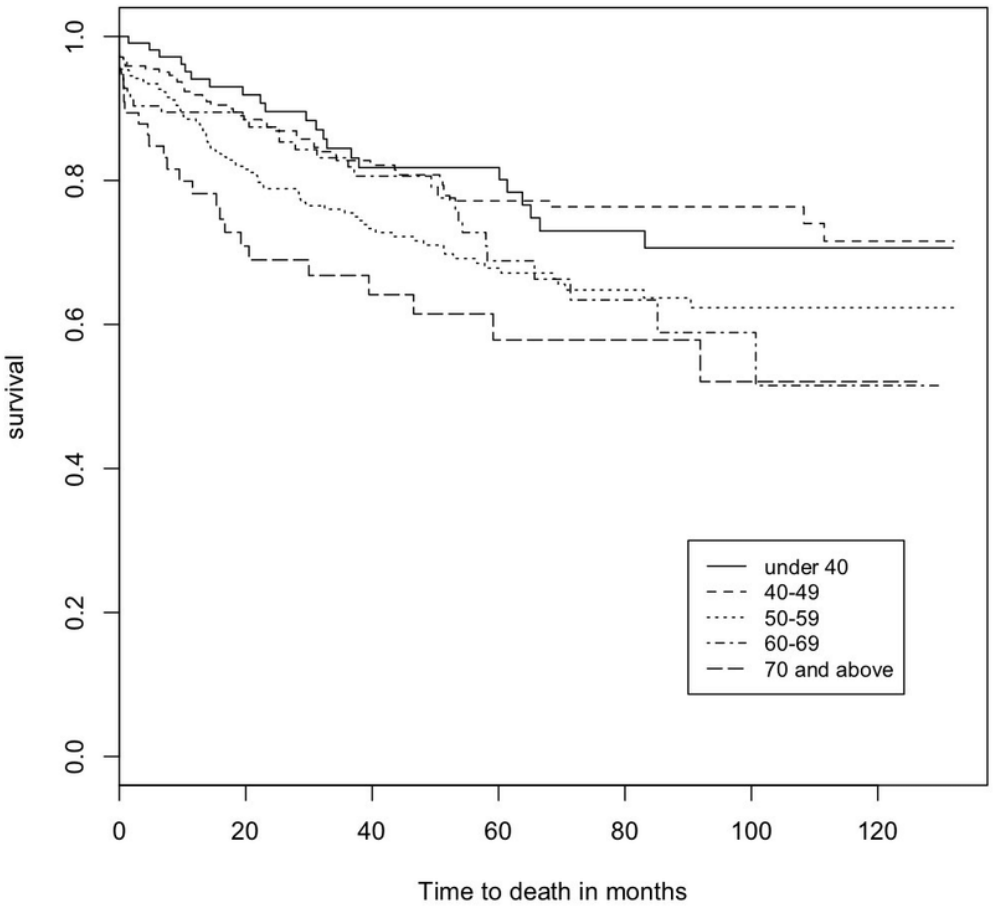
Representation of six clusters after ascending hierarchical clustering analysis.

Axes correspond to individual coordinates for the two main dimensions of the multiple correspondence analysis. Cluster 1: the young symptomatic. Cluster 2: the old obese. Cluster 3: the multi-disease (MD) old obese. Cluster 4: the young snorers. Cluster 5: the drowsy obese. Cluster 6: the MD obese symptomatic.



Zdroj: BAILLY, Sébastien, et al. Obstructive sleep apnea: a cluster analysis at time of diagnosis. PLoS One, 2016, 11.6: e0157318.

Analýza přežití



Zdroj: KATANODA, Kota; MATSUDA, Tomohiro. Five-year relative survival rate of breast cancer in the USA, Europe and Japan. Japanese journal of clinical oncology, 2014, 44.6: 611-611.