Aplikovaná statistika v R

Filip Děchtěrenko

Matematicko-fyzikální fakulta

filip.dechterenko@gmail.com

15.5.2014

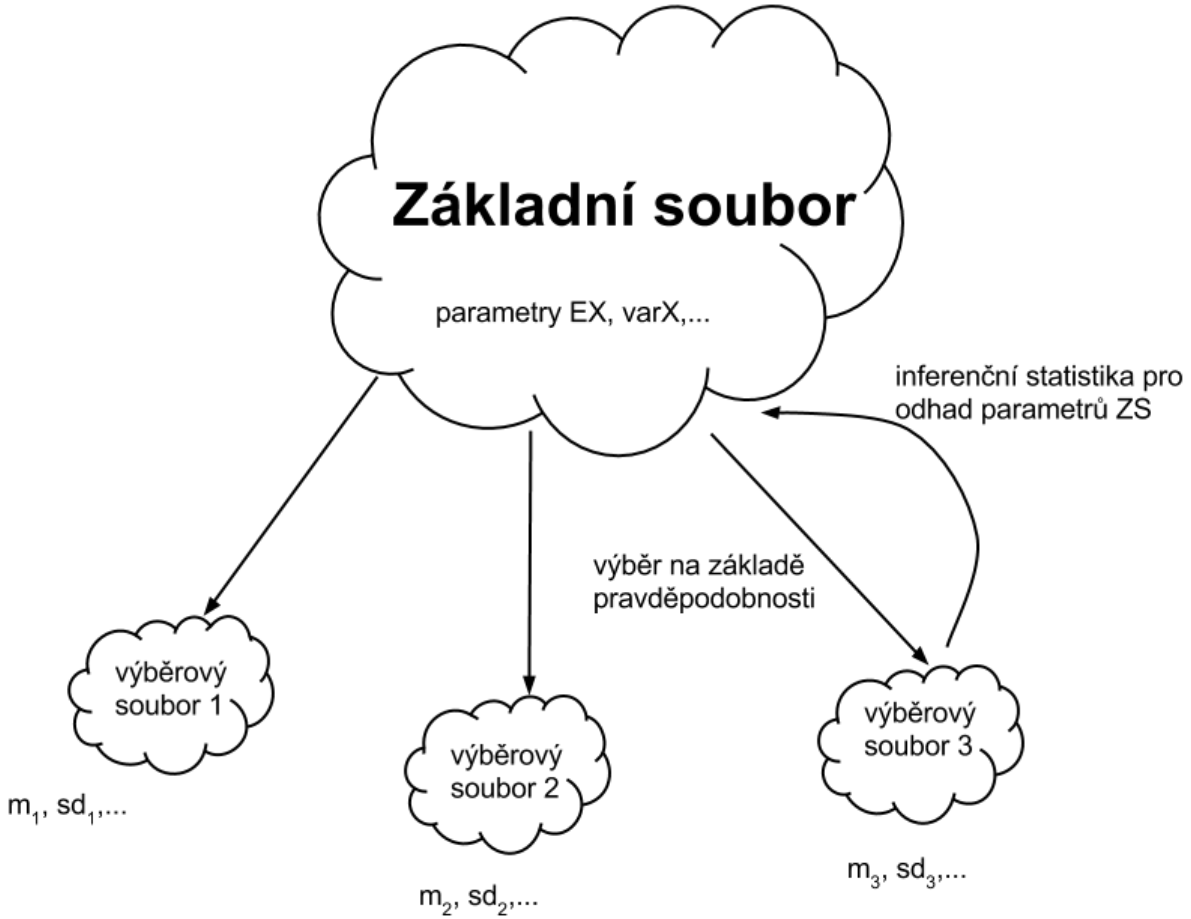
Co bude náplní našich setkání?

- Seznámíme se základními metodami analýzy dat
- Vyzkoušíme si práci v jazyce R

Rozdělení statistiky

- Statistiku rozdělujeme na deskriptivní statistiku a inferenční statistiku
- **Deskriptivní statistika** se zabývá popisem vzorku
- **Inferenční statistika** se zabývá základní populací (pomocí výběru)
- **Základní soubor** (population) je množina všech jevů, kterými se zabýváme
- **Výběrový soubor** (sample) je podmnožina základního souboru
- Kdybychom měli k dispozici celý základní soubor, nemusíme dělat žádnou statistiku

Rozdělení statistiky v obrázku



Obrázek: Vztah mezi základním a výběrovým souborem

Deskriptivní statistika

- Slouží ke zjednodušenému popisu vzorku - typicky ho chceme popsat několika čísly
- Můžeme se zabývat jednou proměnnou nebo více zaráz. Pro jednu proměnnou se typicky zabýváme:
  - **Míry středu** popisují, kde přibližně leží prostředek proměnné
  - **Míry variability** popisují, jak moc se proměnná pohybuje kolem tohoto středu
- **Grafy** nám zobrazí přehledně celý vzorek

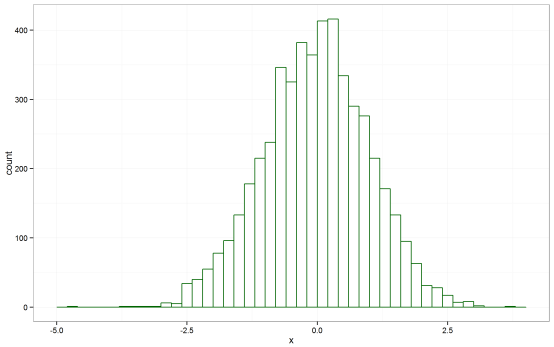
Míry středu

- **Průměr** - určí nám průměrnou hodnotu proměnné
- **Medián** - určí nám prostřední hodnotu proměnné
- **Modus** - určí nám nejčastější hodnotu proměnné
- Označují se někdy jako triple M
- Průměr se používá při parametrických testech, medián při neparametrických testech
- Kromě mediánu se používají i jiná rozdělení dat. Konkrétně percentily, kvantily a kvartily.
- Dolní kvartil je číslo větší než 25% dat, horní než 75% dat

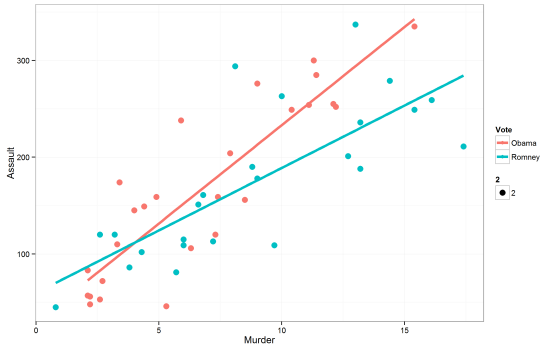
Míry variability

- **Rozpětí** - rozdíl největší a nejmenší hodnoty
- **Mezikvartilové rozpětí** - rozdíl horního a dolního kvartilu
- **Mezikvartilová odchylka** - polovina mezikvartilového rozpětí (odchylka od mediánu)
- **Rozptyl** - Celková míra variability
- **Směrodatná odchylka** - Průměrná míra variability
- Směrodatná odchylka (rozptyl) se používá při parametrických testech, mezikvartilová odchylka (rozpětí) při neparametrických testech

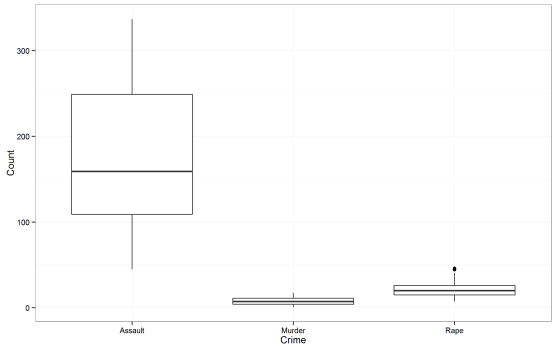
Grafy



Obrázek: Histogram



Obrázek: Scatter plot



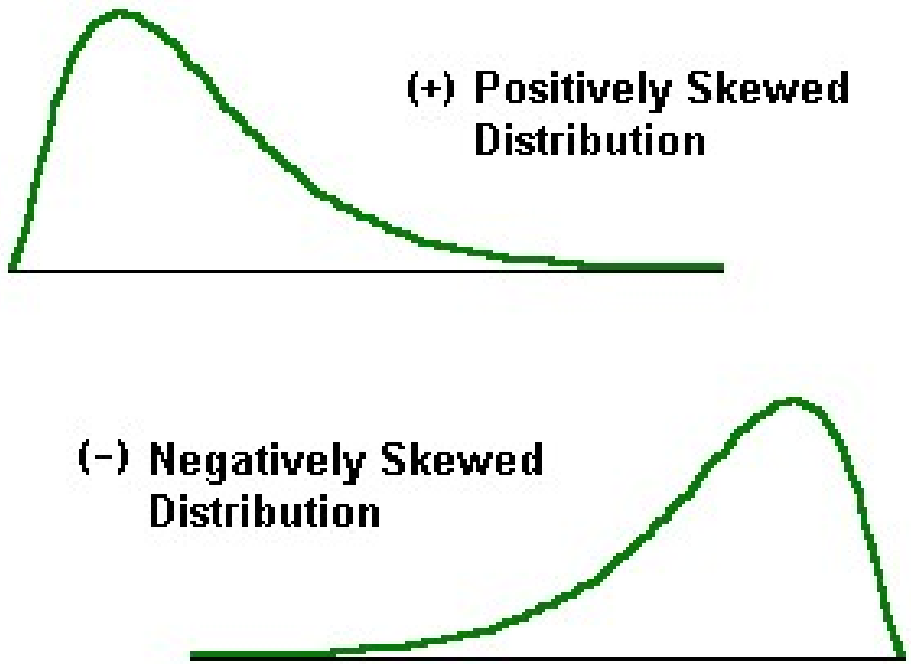
Obrázek: Box plot

Obrázek: Ukázky běžně používaných grafů

Další popisné statistiky

- **Šikmost** - jak moc nahnuté je rozdělení. Kladná šiknost určuje padání doleva, záporná doprava
- **Trimmed mean** - spočítáme průměr bez spodních a horních x%, funkce describe používá spodních a horních 10%
- **Median absolute deviation** - MAD; absolutní odchylka od mediánu
- Tr. mean a MAD jsou robustní míry středu a polohy, používají se v robustních testech
- **Střední chyba průměru** - Standard error of the mean, SEM; určuje kolísání všech možných výběrů ze základní populace (normovaná směrodatná odchylka)

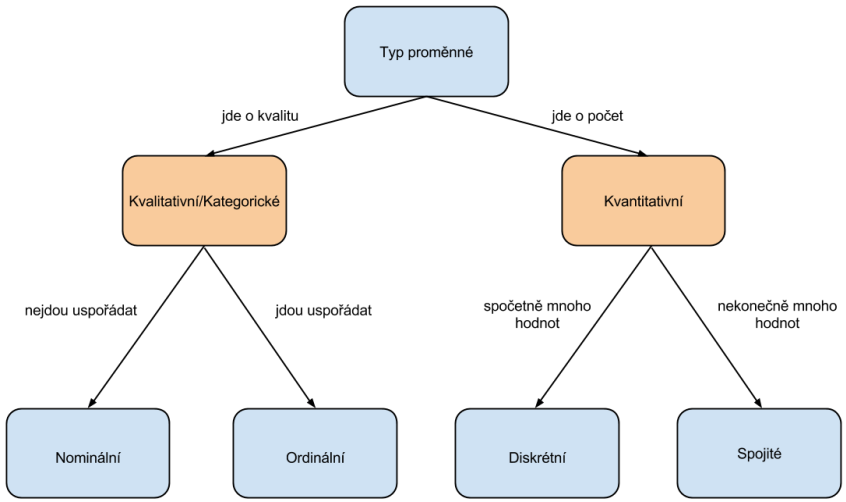
Ukázka šikmosti



Obrázek: Pozitivně a negativně zešikmená data

Druhy proměnných

- Jednotlivé výzkumné proměnné mohou být různých typů. Podle typu porměnných volíme statistický nástroj



Obrázek: Pozitivně a negativně zešikmená data

- Kvantitativní se někdy rozděluje na intervalové a poměrové
- Častokrát není přiřazení proměnných jednoznačné

Příklady proměnných

- Nominální - barva, pohlaví, třídy ve škole
- Ordinální - Likertova škála, známky ve škole
- Kvantitativní - věk, výška

Vyzkoušíme si to prakticky

- Založte si v RStudiu nový projekt
- Stáhněte si soubor z `http://goo.gl/t0iofL`, rozbalte ho do adresáře s projektem a otevřete soubor cviceni1.R

A nyní sami:-)

- Otevřete si soubor cviceni1_test.R a prozkoumejte data, zobrazte vztah mezi proměnnými

Konec cvičení