
# EDA a popisná statistika

# EDA a popisná statistika

- **E**xploratory **D**ata **A**nalysis
  - Průzkumová analýza dat
  - Základní numerické charakteristiky dat
  - Základní grafické charakteristiky dat

- Základní popis dat s pomocí numerických a grafických metod a jednoduchých modelů

# Účel průzkumové analýzy

- Detekce chyb a odlehlých měření
  - Zpřesnění vzorkování, eliminace odchylek pro přesnější popis
- Kontrola předpokladů pro následné statistické zpracování
  - Jaké máme typy dat? Jaké je jejich rozdělení? První hypotézy.
- Zkoumání vzorů v datech
  - Existují nějaké vztahy mezi proměnnými v datech?
  - Mají mnou zkoumaná data nějakou vnitřní strukturu?
  - Lze popsat data jednoduchým modelem?

# Typy dat

- Nominální
- Ordinální
- Intervalová
- Poměrová

- Kvalitativní X Kvantitativní

# Typy EDA

- Jednorozměrná
  - Zkoumáme pouze jednu vlastnost proměnné
  - Frekvenční tabulky, histogramy, distribuční fce …
- Dvourozměrná
  - Zkoumáme dvě vlastnosti proměnné a jejich vztah
  - Korelace, lineární regrese, scatterplot, …
- Vícerozměrná
  - Zkoumáme více než dvě vlastnosti proměnné znaků a jejich vztahy
  - Metody vícerozměrné statistiky – MDS, PCA, FA, …

# Metody

**Numerické**

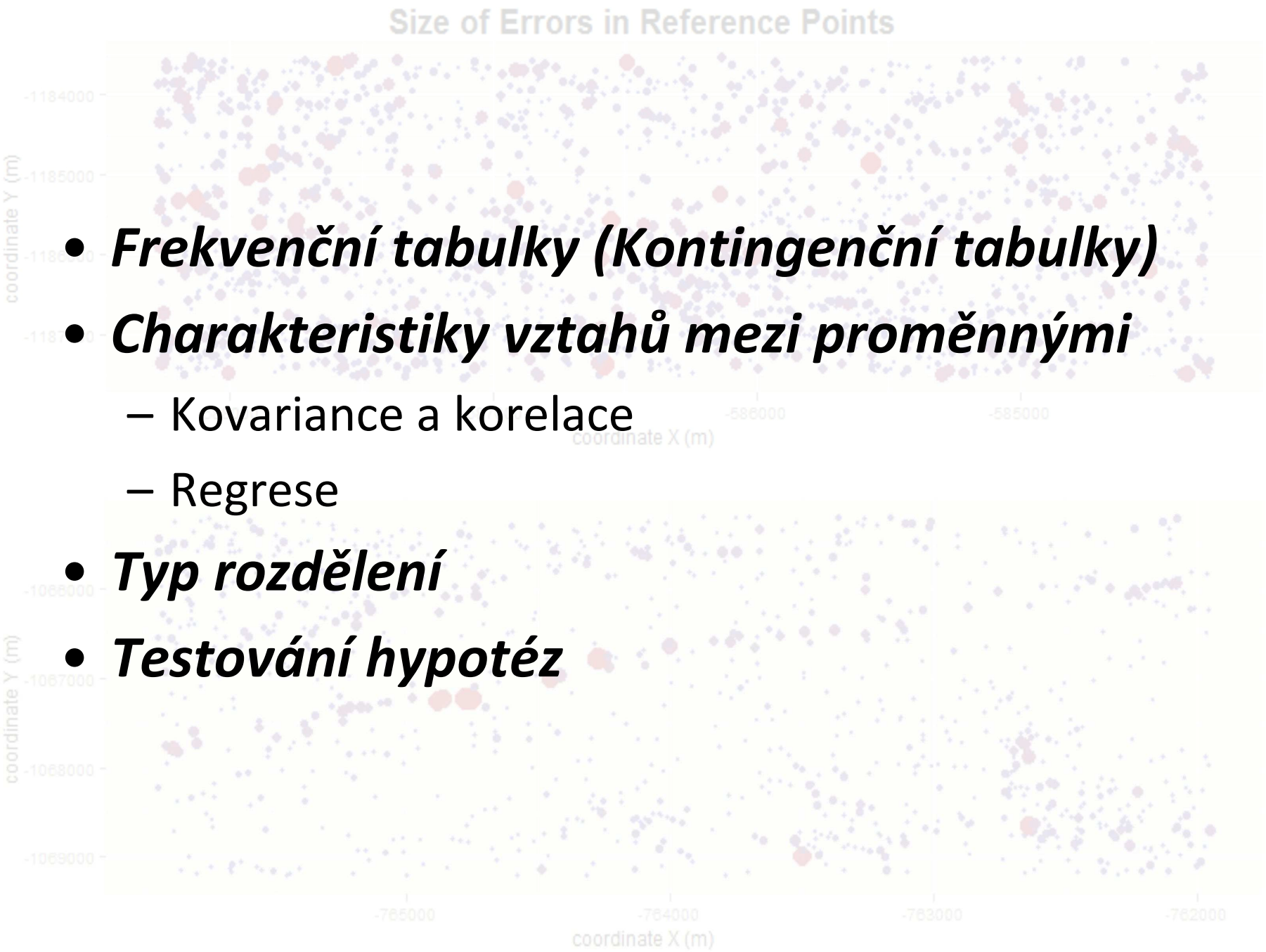
X

**Grafické**

- Popisné
- Robustní
  - Neparametrické
  - Odhad lokálního chování
  - Snaha zachovat původní data – bez transformací

# Numerické charakteristiky

- **Míry polohy**
  - Střední hodnoty
- **Charakteristiky (míry) variability**
  - Rozptyl, směrodatná odchylka, IQR
- **Míry tvaru (rozdělení)**
  - Šikmost, Špičatost




- ***Frekvenční tabulky (Kontingenční tabulky)***
- ***Charakteristiky vztahů mezi proměnnými***
  - Kovariance a korelace
  - Regrese
- ***Typ rozdělení***
- ***Testování hypotéz***

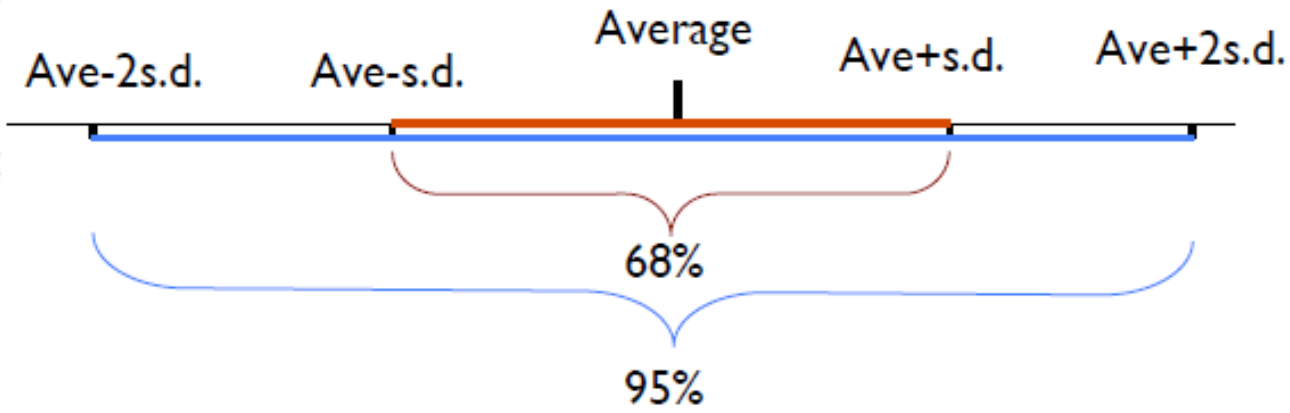
# Míry polohy

- Určují střed kolem kterého jsou data rozptýlena, dále hraniční a významné honoty
- Střední hodnoty
  - Modus
  - Medián
  - Aritmetický průměr
  - Harmonický, geometrický, … průměr
- Minimum, maximum
- Kvantily
  - Kvartily, decily, percentily

# Charakteristiky variability

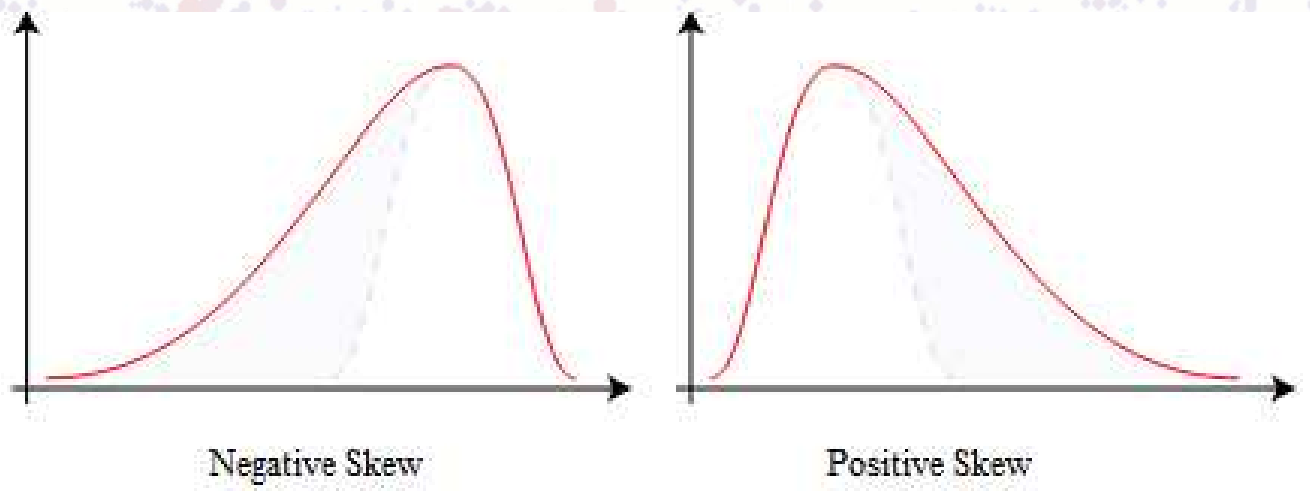
- Popisují rozptýlení dat kolem střední hodnoty

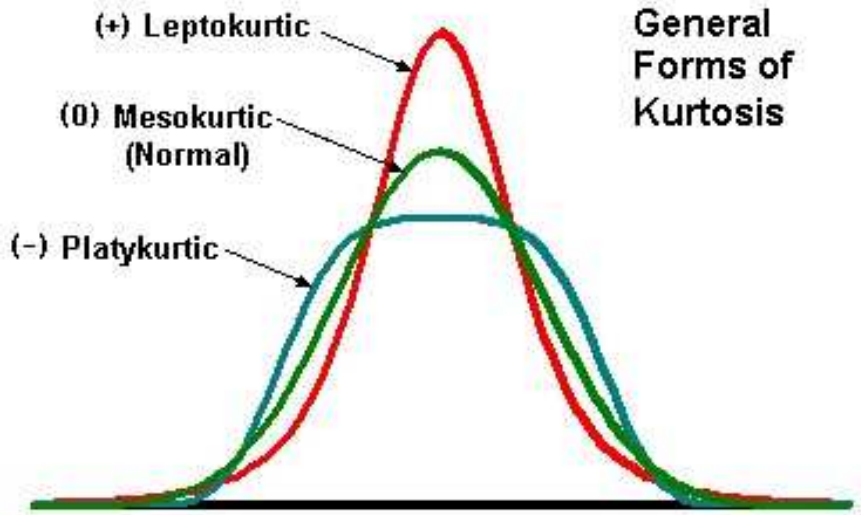
- Směrodatná odchylka
- Rozptyl
- IQR



# Míry tvaru rozdělení

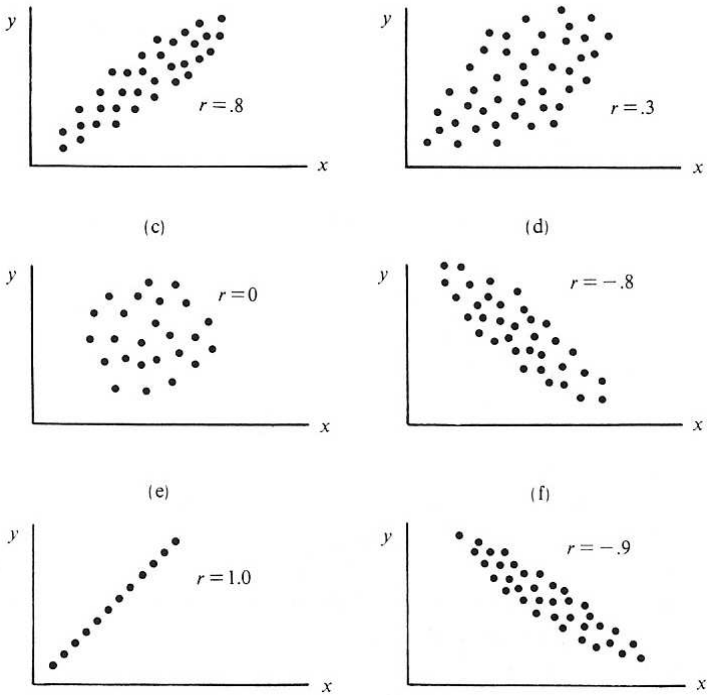
- Popisují tvar rozdělení a relativní polohu konkrétních měření
- ***Šikmost***
  - symetrie rozdělení
  - Je více malých nebo velkých hodnot?
- ***Špičatost***
  - Jsou data kumulována kolem střední hodnoty?


Negative Skew
Positive Skew


(+) Leptokurtic
(0) Mesokurtic
(Normal)
(-) Platykurtic
General
Forms of
Kurtosis

# Charakteristiky vztahů mezi proměnnými

- Zjištění, popis a kvantifikace vztahu mezi dvěma (a více) proměnnými
- Kovariance
  - Absolutní hodnota
- Korelace
  - Relativní hodnota
  - Pearsonova, Spearmanova, …
- Lineární regrese
  - y = ax + b



# Typy rozdělení a testování hypotéz

- Zjištění konkrétního typu rozdělení
  - Specifické metody pro různé druhy rozdělení
  - Transformace

- Testování hypotéz a předpokladů
  - T-test
  - F-test
  - ANOVA, K-W test, Tukey HSD
  - Wilcoxonův, $\chi^2$
  - Shapiro – Wilkův, K-S test, …

# Frekvenční a kontingenční tabulky

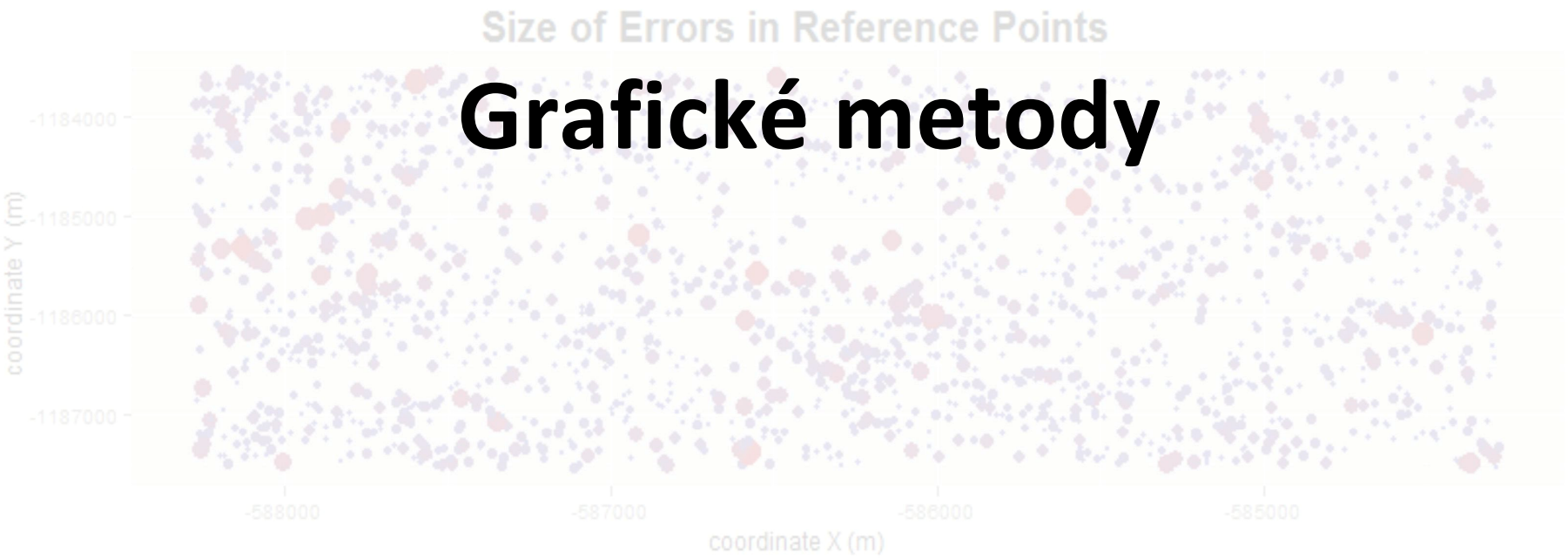
- Frekvenční tabulky
  - Frekvence (počet) výskytu dané kategorie

| Půdní typ | | |
|---|---|---|
| 1 | 2 | 3 |
| 97 | 46 | 12 |

- Kontingenční tabulky
  - Přehledné zobrazení vzájemného výskytu dvou a více znaků jednoho datového souboru

| Půda / LU | Ab | Ah | Am | Bw | Fh | Fw | W | |
|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 21 | 14 | 3 | 0 | 7 | 41 | 90 |
| 2 | 4 | 14 | 7 | 3 | 1 | 0 | 7 | 36 |
| 3 | 0 | 4 | 1 | 0 | 0 | 3 | 2 | 10 |
| | 8 | 39 | 22 | 6 | 1 | 10 | 50 | 136 |

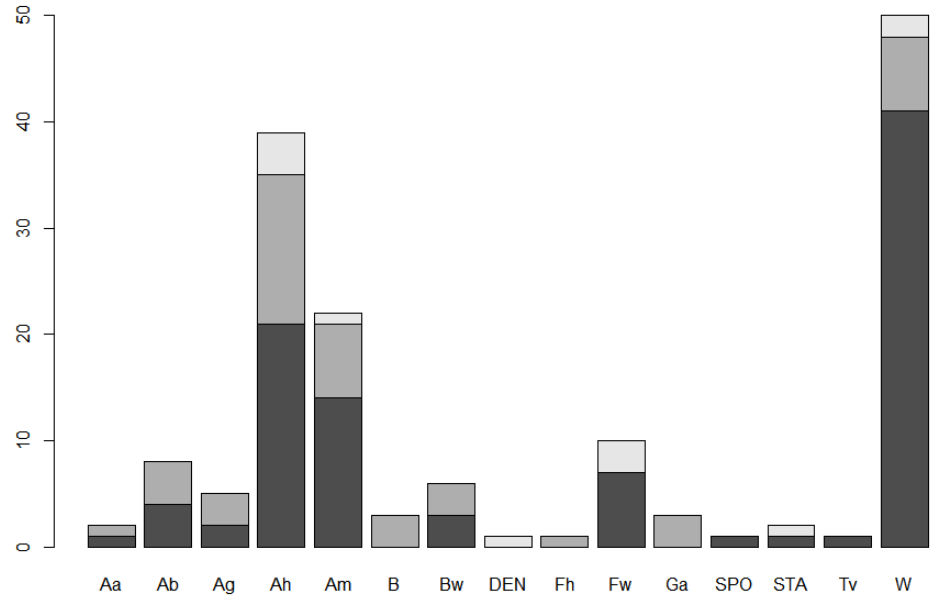
# Grafické metody




# Obrázek (graf) vydá za 1000 slov

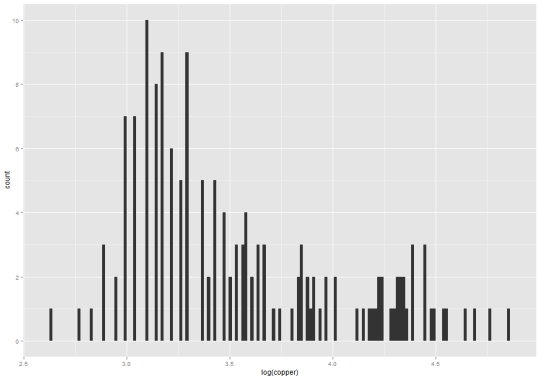
# Sloupcový graf

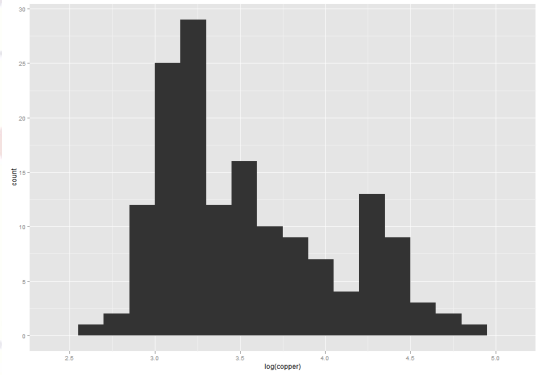
- Zobrazení kategorických dat
  - Frekvenční (kontingenční) tabulky

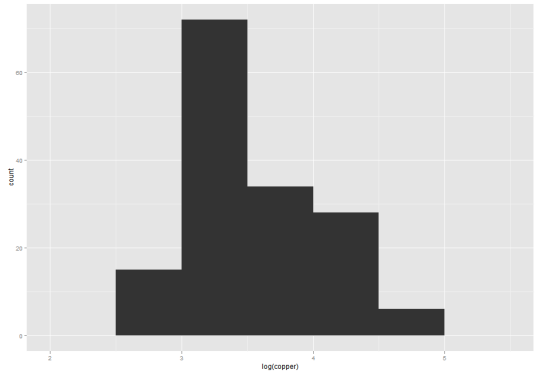


# Histogram

- Zobrazení rozdělení hodnot spojité proměnné
  - Tvar rozdělení, symetrie, rozsah, variabilita, …
  - Pozor na šířku intervalu (sloupce) histogramu
  - Absolutní X relativní hodnoty

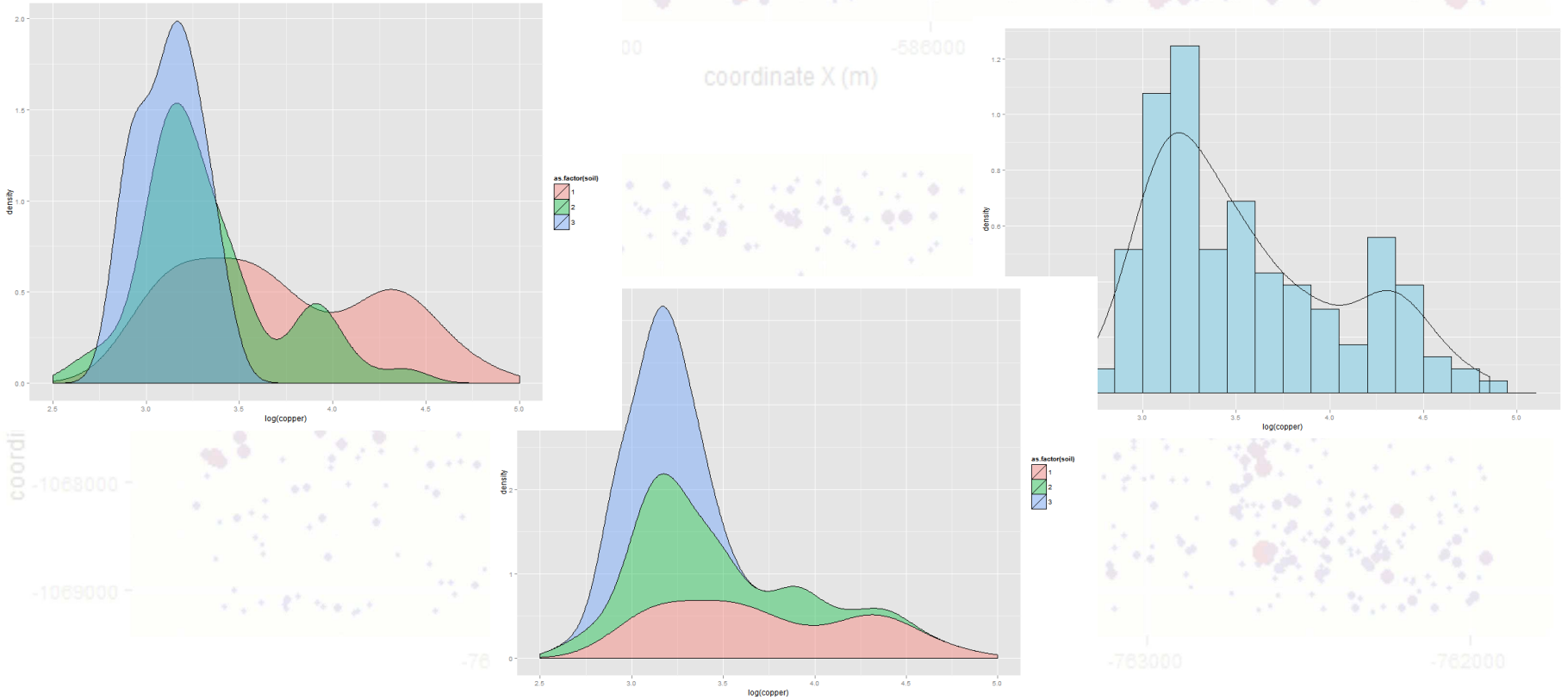


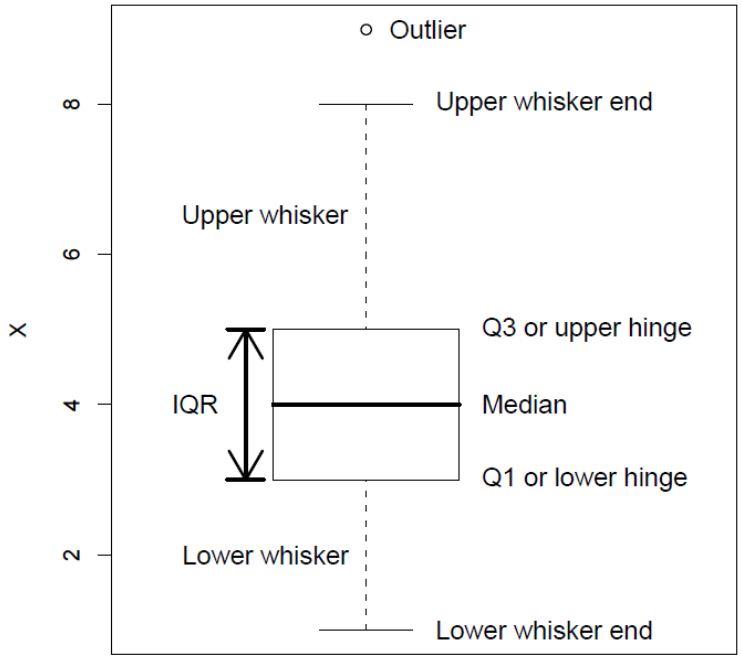
# Distribuční funkce

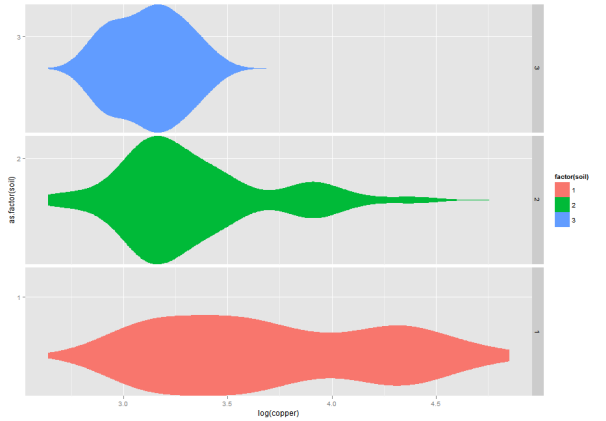
- Zobrazení funkce pravděpodobnosti výskytu dané hodnoty v rámci celé sady

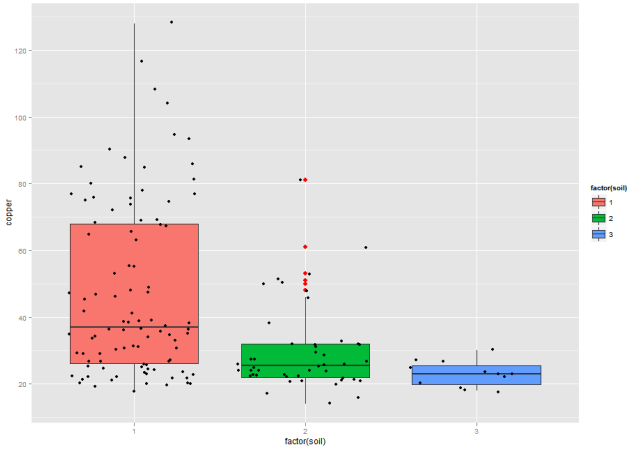


# Boxplot

- Krabicový graf, Box – Whisker plot, …

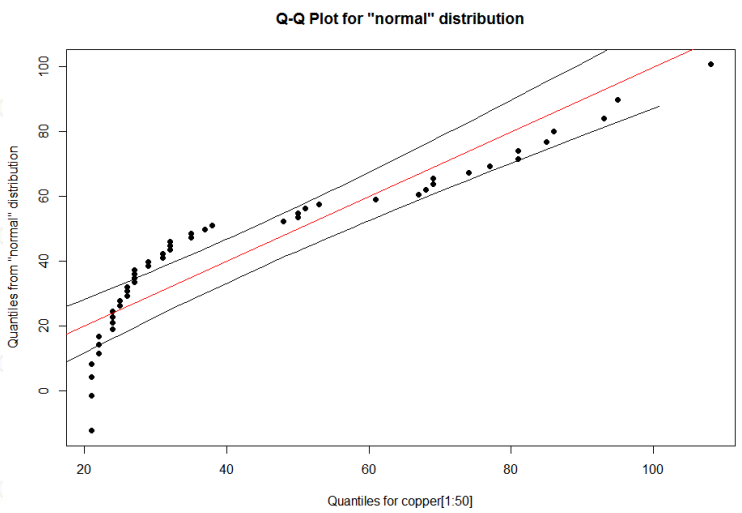


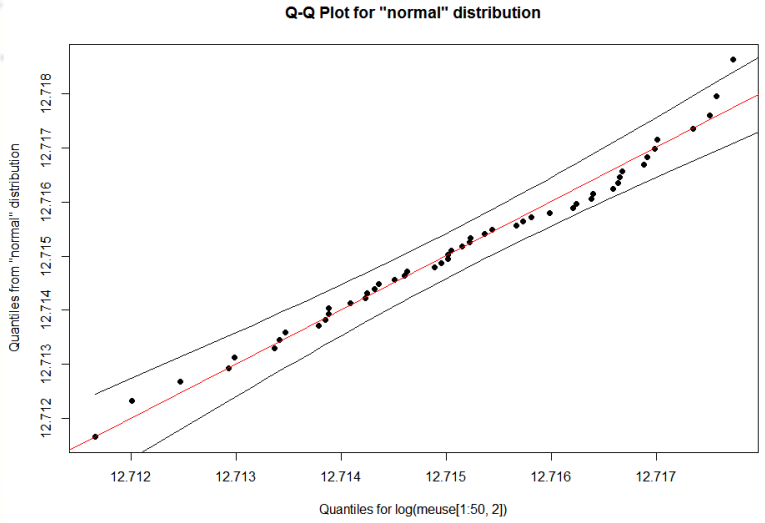



# Q-Q plot

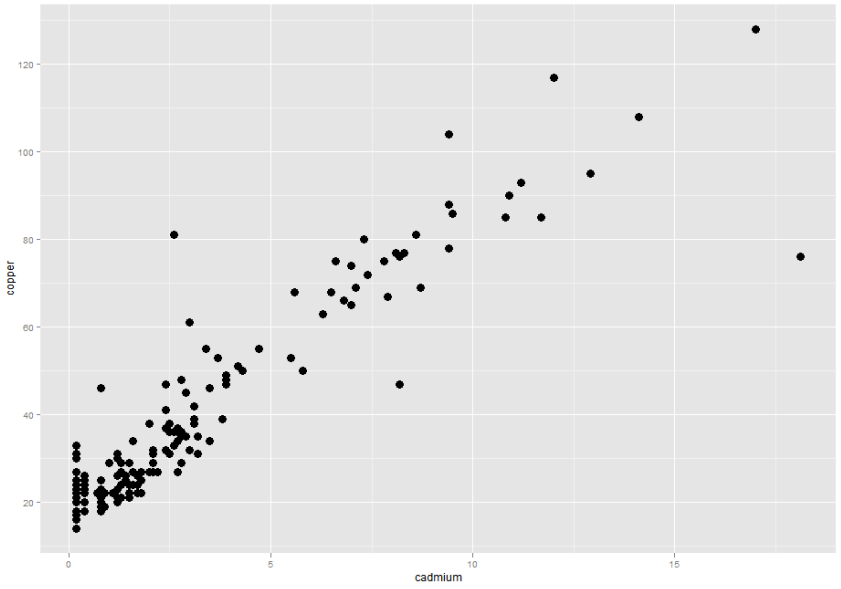
- Srovnání reálných hodnot s teoretickým odhadem hodnot vybraného rozdělení
- Nejčastěji porovnání s normálním rozdělením

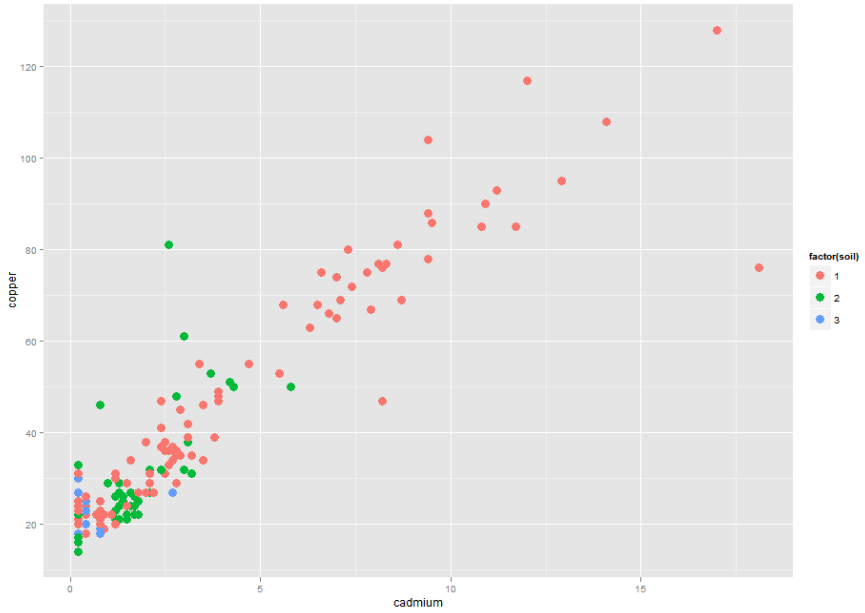


# Scatterplot

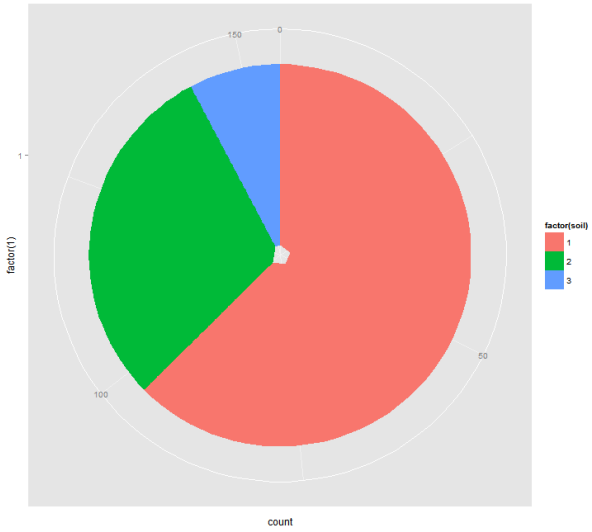
- X-Y bodový graf



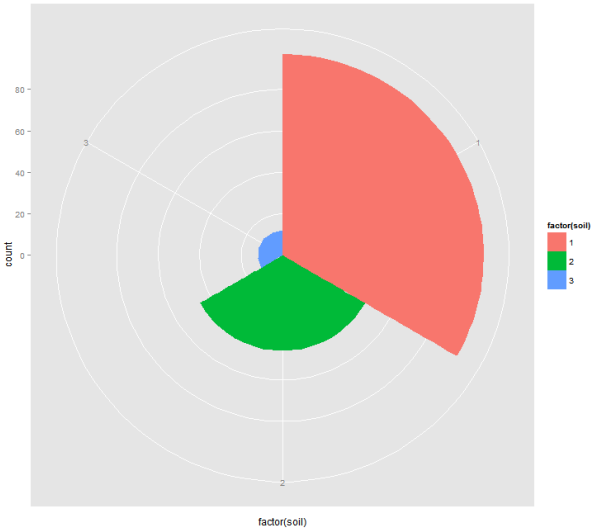



# Kruhový diagram

- Výsečový graf
- nikdy ne „Koláčový graf“




- Špatně interpretovatelný – hlavně u velkého množství skupin
- Příliš mnoho prostoru pro příliš málo informace
- 3D efekty



# Jak udělat špatný graf

- Zobrazte tak málo informací, jak je to jen možné
- Vystavte, co chcete ukázat - rotujte
- Použijte 3D efekty a podivné barvy
- Používejte koláčové grafy
- Zvolte špatné měřítko
- Nepopisujte osy