# Explorační analýza s využitím RkWardu aneb začínáme se softwarem R

Martina Litschmannová, Adéla Vrtková

# Podklady pro seminář

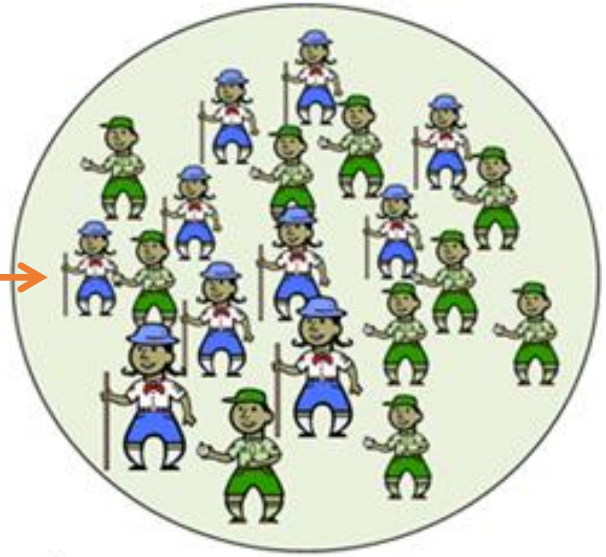
- aku.docx - http://am-nas.vsb.cz/lit40/STA_seminar/2016/aku.docx
- aku.xlsx - http://am-nas.vsb.cz/lit40/STA_seminar/2016/aku.xlsx
- aku.csv - http://am-nas.vsb.cz/lit40/STA_seminar/2016/aku.csv
- seminar_EDA_v_R.R
  - http://am-nas.vsb.cz/lit40/STA_seminar/2016/seminar_EDA_v_R.R

# Jak provést statistické šetření?

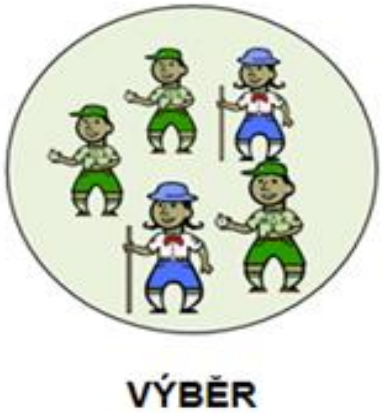
**úplné šetření**

**výběrové šetření**

statistická jednotka



POPULACE = ZÁKLADNÍ SOUB(



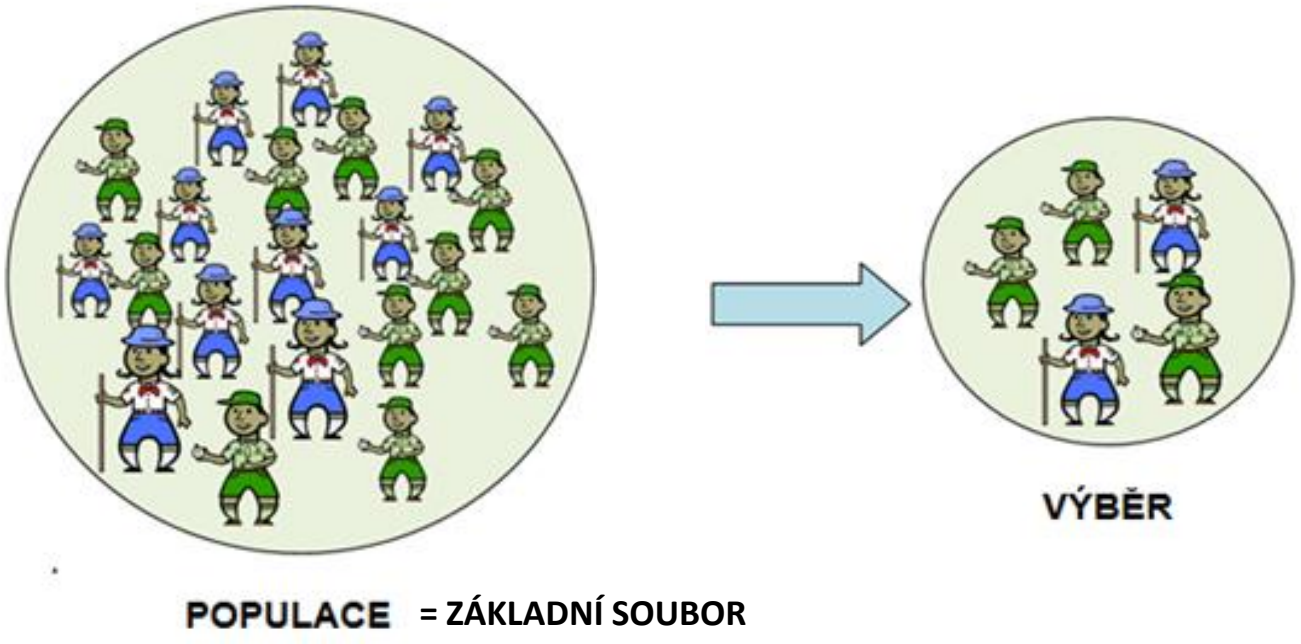
VÝBĚR

REPREZENTATIVNÍ výběr

statistické znaky – údaje, které u statistických znaků sledujeme (např. váha, výška, IQ, …)

# Máme data – a co dál?

Explorační (popisná) analýza




Explorační (popisná) analýza

Statistická indukce

# Software R

- volně šiřitelná verze systému S (komerční verzí systému S je program S-plus)

- je možno zdarma stáhnout a dále distribuovat (při dodržení podmínek GNU GPL licence) z libovolného zrcadla dostupného přes http://www.R-project.org

# RkWard

- interaktivní rozhraní k softwaru R s možností modifikace
- vlastní modifikace RkWardu je ke stažení na

  http://am-nas.vsb.cz/lit40/RkWard/RKWard.zip (KDE/bin/rkward.exe)

# RkWard

POZOR!

Naše modifikace RkWardu obsahuje větší počet naprogramovaných funkcí. V případě použití příkazu, který vymaže paměť programu, dojde k jejich odstranění a tím i omezení funkčnosti interaktivního rozhraní!

# Formáty datového souboru

| kapacita akumulátoru po 5 cyklech (mAh) | | | |
|---|---|---|---|
| Výrobce A | Výrobce B | Výrobce C | Výrobce D |
| 1946.5 | 2006.5 | 1881.8 | 1806.9 |
| 1963.5 | 1991.5 | 1890.4 | 1788.1 |
| 1934.3 | 1988.8 | 1865.7 | 1775,0 |
| 1934.8 | 1975.4 | 1880.7 | 1805.4 |
| 1939.9 | 1998.4 | 1861.1 | 1775.7 |
| 1925.9 | 2012.3 | 1887.3 | 1807.3 |
| 2023,0 | 1995.4 | 1922,0 | 1789.9 |

Seznam proměnných, tabulka

| ID | kapacity akumulátorů po 5 cyklech (mAh) | Výrobce |
|---|---|---|
| 1 | 1946,5 | A |
| 2 | 1963,5 | A |
| 3 | 1934,3 | B |
| 4 | 1934,8 | C |
| 5 | 1939,9 | D |
| 6 | 1925,9 | D |
| 7 | 2023 | B |
| 8 | 1952,5 | A |

Standardní datový formát

# Formáty datového souboru

| kapacita akumulátoru po 5 cyklech (mAh) | | | |
|---|---|---|---|
| Výrobce A | Výrobce B | Výrobce C | Výrobce D |
| 1946.5 | 2006.5 | 1881.8 | 1806.9 |
| 1963.5 | 1991.5 | 1890.4 | 1788.1 |
| 1934.3 | 1988.8 | 1865.7 | 1775,0 |
| 1934.8 | 1975.4 | 1880.7 | 1805.4 |
| 1939.9 | 1998.4 | 1861.1 | 1775.7 |
| 1925.9 | 2012.3 | | 1807.3 |
| 2023,0 | 1995.4 | | 1789.9 |

Seznam proměnných, tabulka

| ID | kapacity akumulátorů po 5 cyklech (mAh) | Výrobce |
|---|---|---|
| 1 | 1946,5 | A |
| 2 | 1963,5 | A |
| 3 | 2006.5 | B |
| 4 | 1991.5 | B |
| 5 | 1806.9 | D |
| 6 | | C |
| 7 | | C |
| 8 | 1925.9 | A |

Standardní datový formát

# Formáty datového souboru

| kapacita akumulátoru po 5 cyklech (mAh) | | | |
|---|---|---|---|
| Výrobce A | Výrobce B | Výrobce C | Výrobce D |
| 1946.5 | 2006.5 | 1881.8 | 1806.9 |
| 1963.5 | 1991.5 | 1890.4 | 1788.1 |
| 1934.3 | 1988.8 | 1865.7 | 1775,0 |
| 1934.8 | 1975.4 | 1880.7 | 1805.4 |
| 1939.9 | 1998.4 | 1861.1 | 1775.7 |
| 1925.9 | 2012.3 | NA | 1807.3 |
| 2023,0 | 1995.4 | NA | 1789.9 |

Seznam proměnných, tabulka

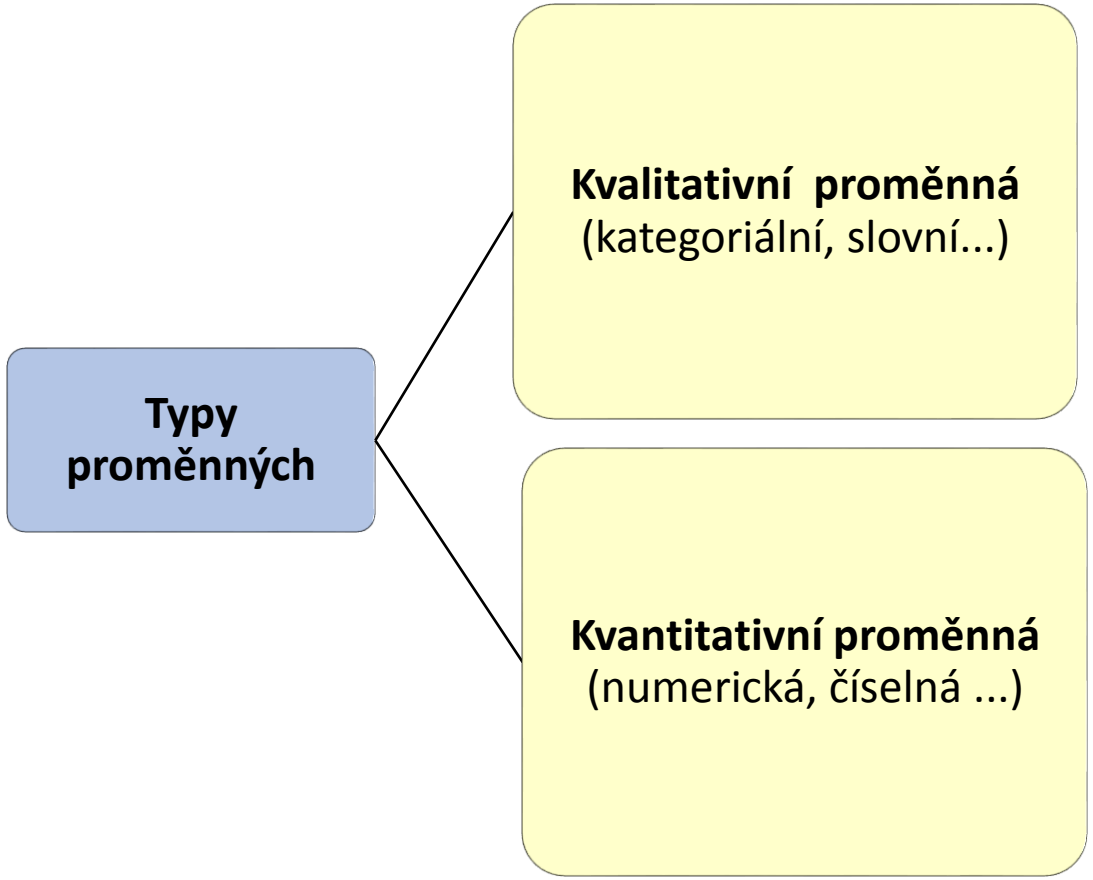
| ID | kapacity akumulátorů po 5 cyklech (mAh) | Výrobce |
|---|---|---|
| 1 | 1946,5 | A |
| 2 | 1963,5 | A |
| 3 | 2006.5 | B |
| 4 | 1991.5 | B |
| 5 | 1806.9 | D |
| 6 | NA | C |
| 7 | NA | C |
| 8 | 1925.9 | A |

Standardní datový formát

# Explorační analýza

Grafická prezentace a uspořádání dat do názornější formy a jejich popis několika málo hodnotami, které by obsahovaly co největší množství informací obsažených v původním souboru.

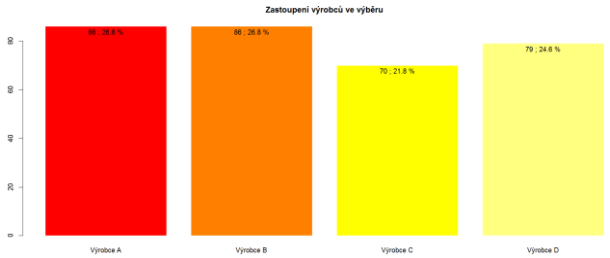
# Typy proměnných



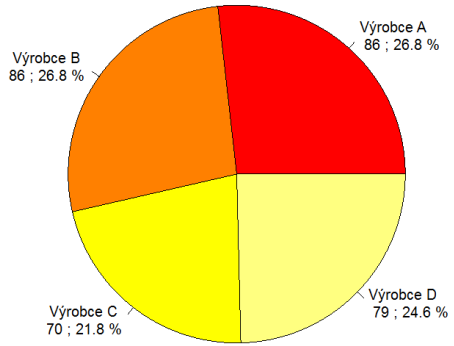
# Explorační analýza kvalitativní proměnné

| **TABULKA ROZDĚLENÍ ČETNOSTI** | | |
|---|---|---|
| **Varianty** $x_i$ | **Absolutní četnosti** $n_i$ | **Relativní četnosti** $p_i$ |
| $x_1$ | $n_1$ | $p_1 = n_1/n$ |
| $x_2$ | $n_2$ | $p_2 = n_2/n$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $x_k$ | $n_k$ | $p_k = n_k/n$ |
| **Celkem:** | $n_1 + n_2 + \cdots + n_k = n$ | 1 |

Sloupcový graf (Bar Chart)




Výsečový graf (Pie Chart)



# Explorační analýza kvantitativní proměnné

## Míry polohy

- Aritmetický průměr (pozor na citlivost průměru na výskyt odlehlých pozorování)
- Kvantily (např. kvartily)

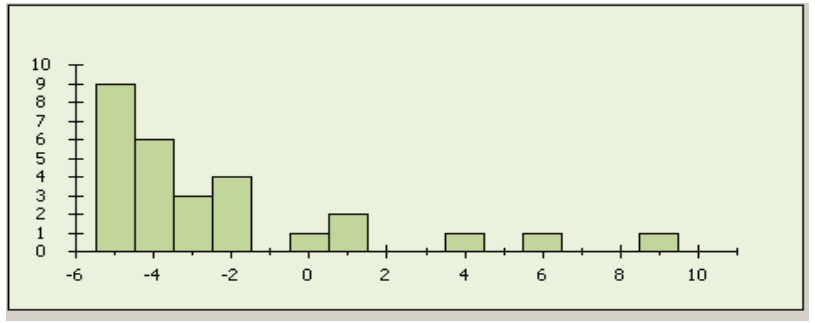
## Míry variability

- Rozptyl
- Směrodatná odchylka
- Variační koeficient (var. koef. větší než 50% naznačuje silnou heterogenitu dat)

# Jakou představu o variabilitě nám dává směrodatná odchylka?

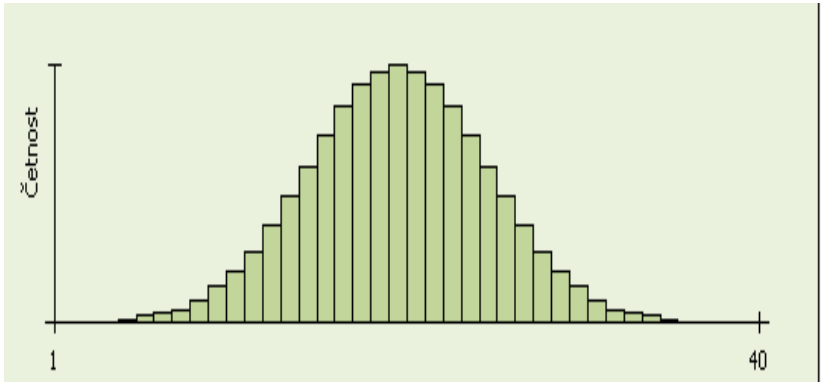
Čebyševova nerovnost: $\forall k > 0: P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}$

| $k$ | $P(\mu - k\sigma < X < \mu + k\sigma)$ |
|---|---|
| 1 | >0 |
| 2 | >0,75 |
| 3 | >0,89 |




Empirické pravidlo 3 sigma

| $k$ | $P(\mu - k\sigma < X < \mu + k\sigma)$ |
|---|---|
| 1 | 0,682 |
| 2 | 0,954 |
| 3 | 0,998 |

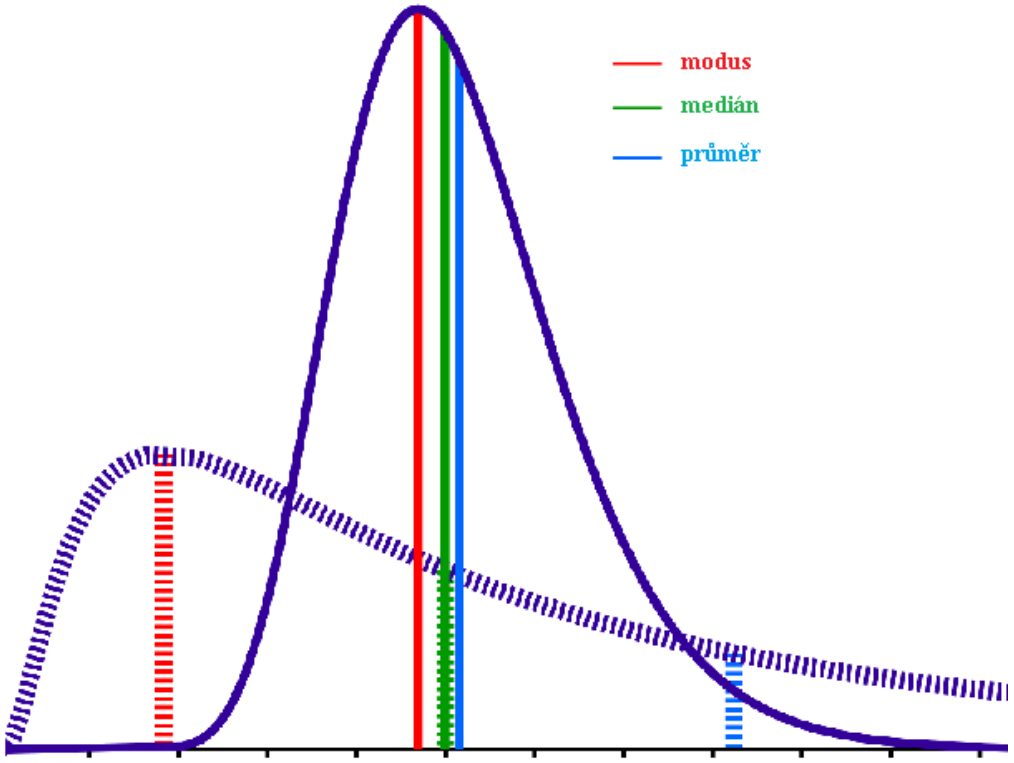


# Explorační analýza kvantitativní proměnné

## Míry šikmosti a špičatosti

- šikmost
- špičatost

pozitivně zešikmená data




Jsou-li data výběrem z normálního rozdělení, pak výběrová šikmost i výběrová špičatost obvykle leží v intervalu $\langle -2; 2 \rangle$.

# Odlehlá pozorování

- ty hodnoty proměnné, které se mimořádně liší od ostatních hodnot a tím ovlivňují např. vypovídací hodnotu průměru.

Jak postupovat v případě, že v datech identifikujeme odlehlá pozorování?

- V případě, že odlehlost pozorování je způsobena:
  - hrubými chybami, překlepy, prokazatelným selháním lidí či techniky ...
  - důsledky poruch, chybného měření, technologických chyb ...

tzn., známe-li příčinu odlehlosti a předpokládáme-li, že již nenastane, jsme oprávněni tato pozorování vyloučit z dalšího zpracování.

- V ostatních případech je nutno zvážit, zda se vyloučením odlehlých pozorování nepřipravíme o důležité informace o jevech vyskytujících se s nízkou četností.

# Jak identifikovat odlehlá pozorování?

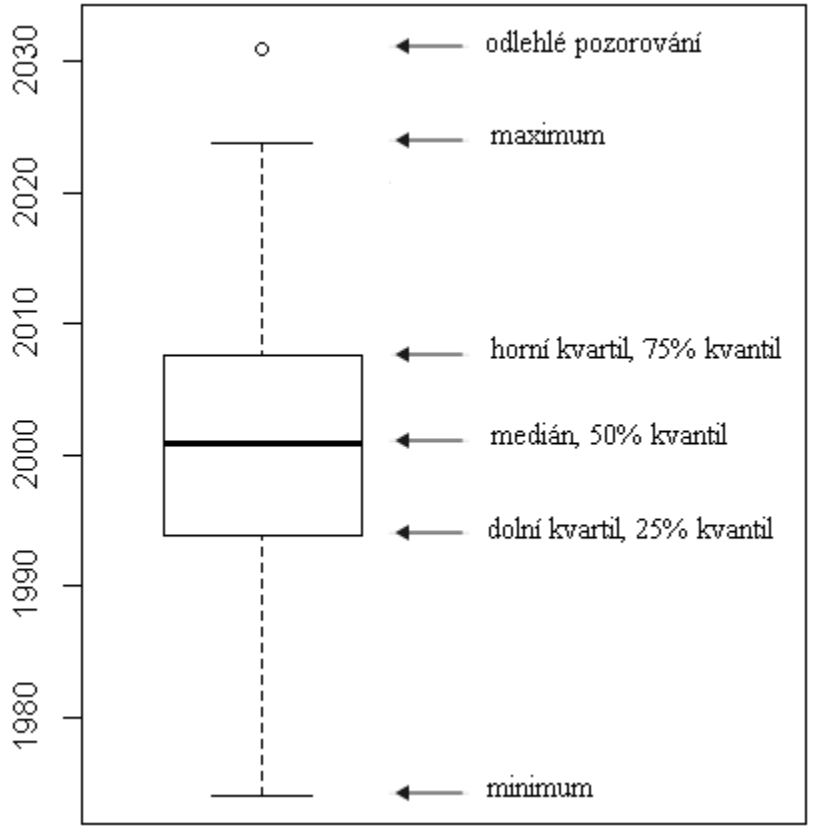
## Metoda vnitřních hradeb

$$\left[\left(x_i < x_{0,25} - 1{,}5IQR\right) \vee \left(x_i > x_{0,75} + 1{,}5IQR\right)\right] \Rightarrow x_i \ \ je\ odlehlým\ \ pozorováním$$

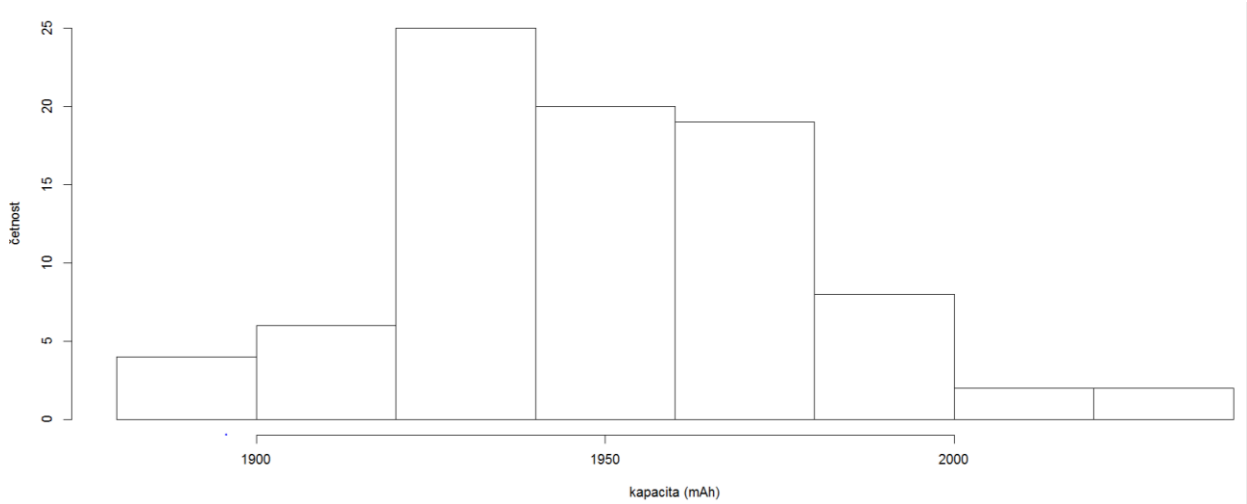
**Dolní mez vnitřních hradeb**

**Horní mez vnitřních hradeb**

# Vizualizace kvantitativní proměnné




Krabicový graf (box plot)




Histogram

# Jak zaokrouhlovat výběrové charakteristiky?

Směrodatnou odchylku jakožto míru nejistoty měření zaokrouhlujeme na jednu, maximálně dvě platné cifry a míry polohy (průměr, kvantily…) zaokrouhlujeme tak, aby nejnižší zapsaný řád odpovídal nejnižšímu zapsanému řádu směrodatné odchylky.

# Chybný zápis číselných charakteristik

| | Délka (m) | Váha (kg) | Teplota ($^0$C) |
|---|---|---|---|
| **Průměr** | 2,26 | 127,6 | 14 567 |
| **Medián** | 2,675 | 117,8 | 13 700 |
| **Směrodatná odchylka** | 0,78 | 23,7 | 1 200<br>(před zaokrouhlením 1235) |
| **Proč je zápis chybný?** | | | |

# Chybný zápis číselných charakteristik

| | Délka (m) | Váha (kg) | Teplota ($^0$C) |
|---|---|---|---|
| **Průměr** | 2,26 | 127,6 | 14 567 |
| **Medián** | 2,675 | 117,8 | 13 700 |
| **Směrodatná odchylka** | 0,78 | 23,7 | 1 200<br>(před zaokrouhlením 1235) |
| **Proč je zápis chybný?** | *Různý počet des. míst.* | | |

# Chybný zápis číselných charakteristik

| | Délka (m) | Váha (kg) | Teplota ($^0$C) |
|---|---|---|---|
| **Průměr** | 2,26 | 127,6 | 14 567 |
| **Medián** | 2,675 | 117,8 | 13 700 |
| **Směrodatná odchylka** | 0,78 | 23,7 | 1 200<br>(před zaokrouhlením 1235) |
| **Proč je zápis chybný?** | *Různý počet des. míst.* | *3 platné cifry u směrodatné odchylky.* | |

# Chybný zápis číselných charakteristik

| | Délka (m) | Váha (kg) | Teplota ($^0$C) |
|---|---|---|---|
| **Průměr** | 2,26 | 127,6 | 14 567 |
| **Medián** | 2,675 | 117,8 | 13 700 |
| **Směrodatná odchylka** | 0,78 | 23,7 | 1 200<br>(před zaokrouhlením 1235) |
| **Proč je zápis chybný?** | *Různý počet des. míst.* | *3 platné cifry u směrodatné odchylky.* | *Nejnižší zapsaný řád průměru (jednotky) neodpovídá nejnižšímu zapsanému řádu směrodatné odchylky (stovky)+ směr. odch. není zaokrouhlena nahoru.* |

# Oprava

| | Délka (m) | Váha (kg) | Teplota ($^0$C) |
|---|---|---|---|
| **Průměr** | 2,26 | 127,6 | 14 567 |
| **Medián** | 2,68 | 117,8 | 13 700 |
| **Směrodatná odchylka** | 0,78 | 23,7 | 1 200<br>(před zaokrouhlením 1235) |
| **Proč je zápis chybný?** | | *3 platné cifry u směrodatné odchylky.* | *Nejnižší zapsaný řád průměru (jednotky) neodpovídá nejnižšímu zapsanému řádu směrodatné odchylky (stovky)+ směr. odch. není zaokrouhlena nahoru.* |

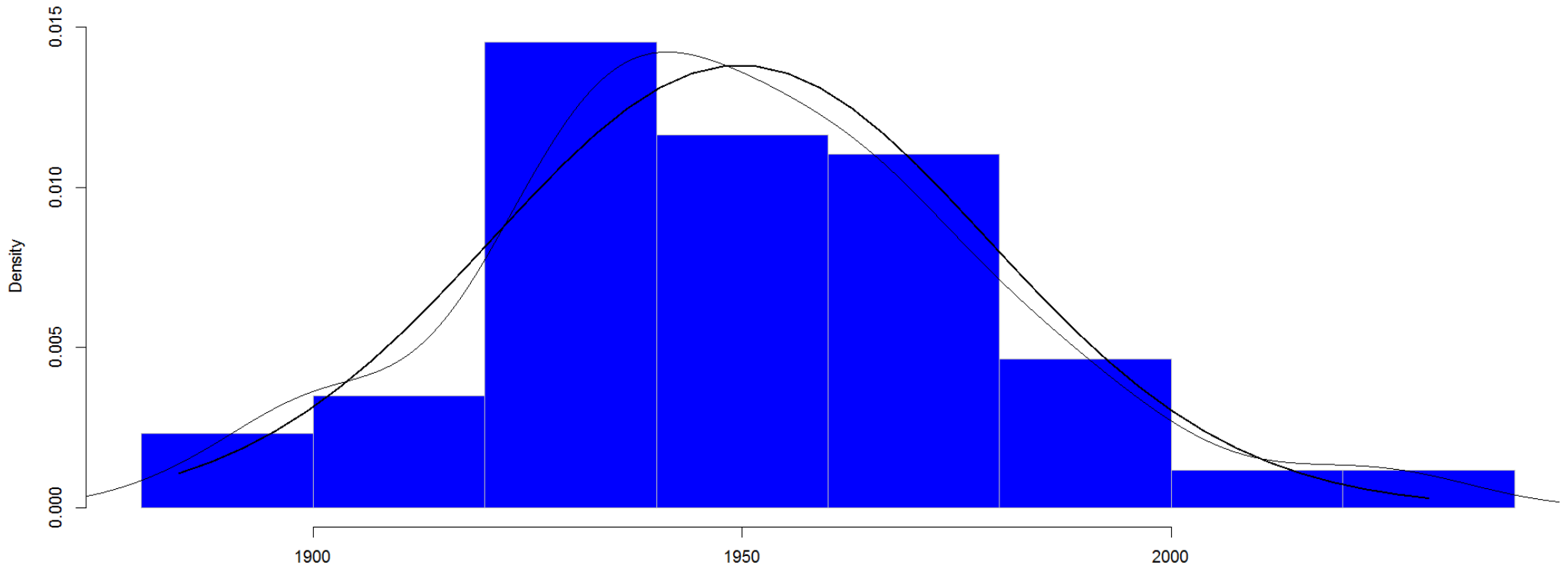
# Oprava

| | Délka (m) | Váha (kg) | Teplota ($^0$C) |
|---|---|---|---|
| **Průměr** | 2,26 | 128 | 14 567 |
| **Medián** | 2,68 | 118 | 13 700 |
| **Směrodatná odchylka** | 0,78 | 24 | 1 200<br>(před zaokrouhlením 1235) |
| **Proč je zápis chybný?** | | | *Nejnižší zapsaný řád průměru (jednotky) neodpovídá nejnižšímu zapsanému řádu směrodatné odchylky (stovky)+ směr. odch. není zaokrouhlena nahoru.* |

# Správný zápis číselných charakteristik

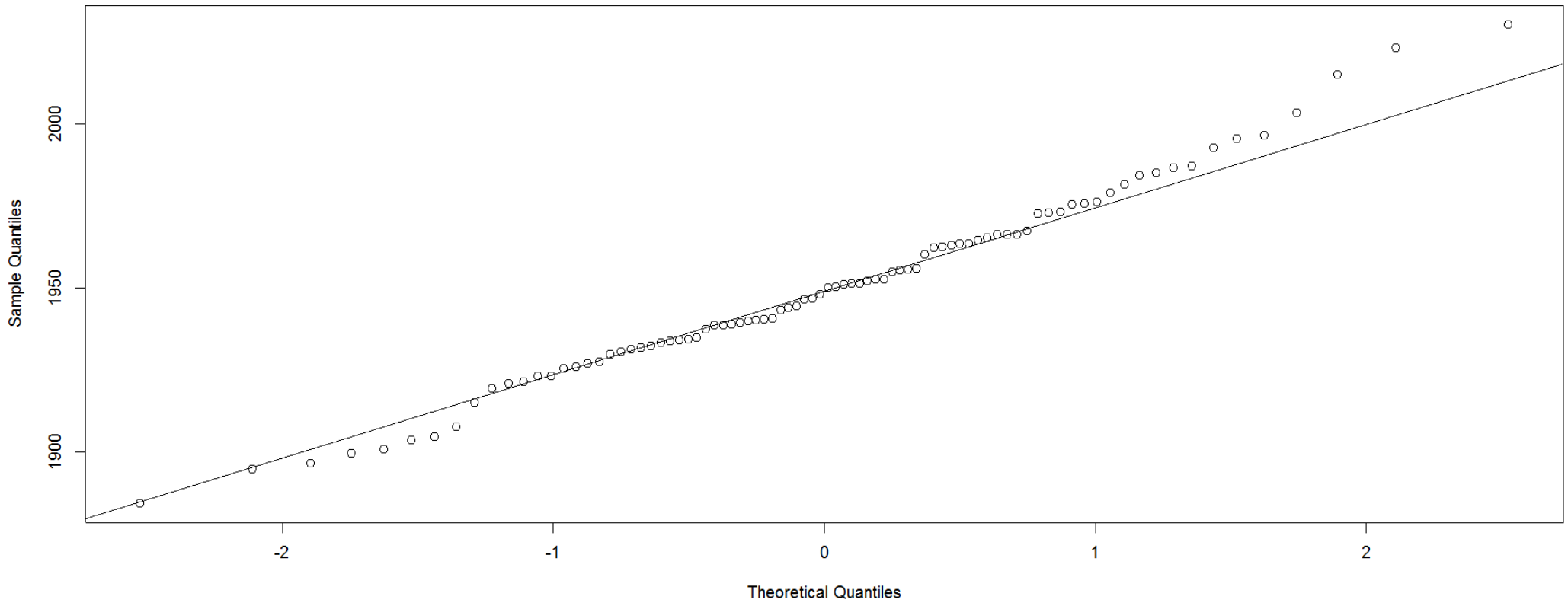
| | Délka (m) | Váha (kg) | Teplota ($^0$C) |
|---|---|---|---|
| **Průměr** | 2,26 | 127,6 | 14 600 |
| **Medián** | 2,675 | 117,8 | 13 700 |
| **Směrodatná odchylka** | 0,78 | 23,7 | 1 300 |

# Vizualizace kvantitativní proměnné




Kombinace histogramu, odhadu hustoty pravděpodobnosti a odpovídající hustoty pravděpodobnosti normálního rozdělení

# Vizualizace kvantitativní proměnné




Q-Q graf