# Pravděpodobnost a statistika

## Úvod do analýzy závislostí

Vilém Vychodil

KMI/PRAS, Přednáška 2

Vytvořeno v rámci projektu 2963/2011 FRVŠ

# Přednáška 2: Přehled

1. Uspořádané výběry:
   - číslicové dendrogramy, krabicové grafy,
   - percentily a hodnoty odvozené z percentilů,
   - zkoumání odlehlých hodnot.
2. Porovnávání hodnot dvou výběrů:
   - grafické metody porovnávání diagramů,
   - metody založené na kvantilech a kk-grafy,
3. Korelační a regresní analýza:
   - základní problémy lineární regresní analýzy,
   - metoda nejmenších čtverců, kovariance, korelační koeficient,
   - vlastnosti regresních přímek a korelačních koeficientů,
   - nelineární regrese.

# Číslicové dendrogramy

**Alternativní způsob grafického znázornění dat ve výběru**
Výběr $x_1, \ldots, x_n$, kde každé číslo $x_i$ obsahuje alespoň dvě cifry.

## Sestrojení číslicového dendrogramu, angl.: *stem-and-leaf display*

1. Zvolíme číselný řád (tisíce, stovky, desítky, ...), pak:
   - **stonek**, angl.: *stem*: hodnoty daného řádu nacházející se ve výběru,
   - **list**, angl.: *leaf*: hodnoty nižšího řádu nacházející se ve výběru.
2. Vypíšeme všechny možné hodnoty stonků do sloupce pod sebe.
3. Za každý stonek napíšeme seznam odpovídajících listů.
4. Obvykle doplňujeme přidaným sloupcem s absolutní četností listů.

**Výhody**

- Graf znázorňuje „rozložení hodnot“ ve výběru (analogicky jako histogram).
- Narozdíl od histogramu jsou data z výběru přímo obsažena v diagramu.

## Příklad (Vstupní data a absolutní a relativní četnosti)

Výběr obsahující skóre $40$ studentů na zkoušce:

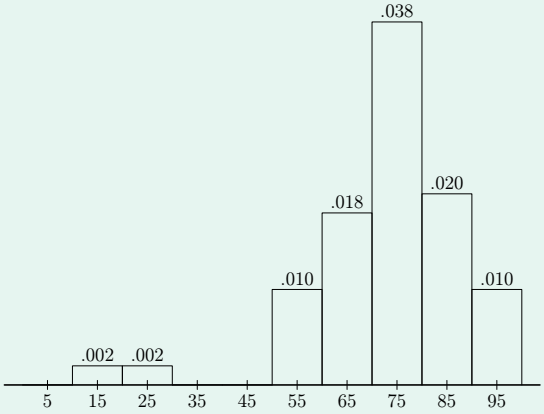
$$\begin{array}{cccccccccccccccccccc} 75 & 70 & 86 & 76 & 61 & 76 & 88 & 74 & 67 & 96 & 93 & 73 & 65 & 53 & 98 & 86 & 75 & 83 & 64 & 15 \\ 76 & 50 & 88 & 75 & 75 & 21 & 86 & 60 & 57 & 76 & 63 & 79 & 71 & 80 & 67 & 87 & 76 & 90 & 70 & 59 \end{array}$$

| interval | meze | $f_i$ | $h_i$ | střed |
|---|---|---|---|---|
| $[10, 20)$ | $(15, 15)$ | $1$ | $0.002$ | $15.0$ |
| $[20, 30)$ | $(21, 21)$ | $1$ | $0.002$ | $25.0$ |
| $[50, 60)$ | $(50, 59)$ | $4$ | $0.010$ | $55.0$ |
| $[60, 70)$ | $(60, 67)$ | $7$ | $0.018$ | $65.0$ |
| $[70, 80)$ | $(70, 79)$ | $15$ | $0.038$ | $75.0$ |
| $[80, 90)$ | $(80, 88)$ | $8$ | $0.020$ | $85.0$ |
| $[90, 100)$ | $(90, 98)$ | $4$ | $0.010$ | $95.0$ |

## Příklad (Histogram relativních četností a číslicový dendrogram)

Výběr obsahující skóre $40$ studentů na zkoušce:

$$\begin{array}{cccccccccccccccccccc} 75 & 70 & 86 & 76 & 61 & 76 & 88 & 74 & 67 & 96 & 93 & 73 & 65 & 53 & 98 & 86 & 75 & 83 & 64 & 15 \\ 76 & 50 & 88 & 75 & 75 & 21 & 86 & 60 & 57 & 76 & 63 & 79 & 71 & 80 & 67 & 87 & 76 & 90 & 70 & 59 \end{array}$$




| stonek | listy | $f_i$ |
|---|---|---|
| 1 | 5 | 1 |
| 2 | 1 | 1 |
| 5 | 3079 | 4 |
| 6 | 1754037 | 7 |
| 7 | 506643565569160 | 15 |
| 8 | 68638607 | 8 |
| 9 | 6380 | 4 |

# Příklad (Modifikace číslicového dendrogramu)

| stonek | listy | $f_i$ |
|---|---|---|
| 1 | 5 | 1 |
| 2 | 1 | 1 |
| 5 | 3079 | 4 |
| 6 | 1754037 | 7 |
| 7 | 506643565569160 | 15 |
| 8 | 68638607 | 8 |
| 9 | 6380 | 4 |

$\Longmapsto$

| stonek | listy | $f_i$ |
|---|---|---|
| $1^\bullet$ | 5 | 1 |
| $2^*$ | 1 | 1 |
| $5^*$ | 30 | 2 |
| $5^\bullet$ | 79 | 2 |
| $6^*$ | 1403 | 4 |
| $6^\bullet$ | 757 | 3 |
| $7^*$ | 04310 | 5 |
| $7^\bullet$ | 5665655696 | 10 |
| $8^*$ | 30 | 2 |
| $8^\bullet$ | 686867 | 6 |
| $9^*$ | 30 | 2 |
| $9^\bullet$ | 68 | 2 |

# Uspořádané výběry a číslicové dendrogramy

## Definice (Uspořádaný výběr)

Hodnoty výběru $x_1, \ldots, x_n$ uspořádané vzestupně podle jejich velikosti se nazývají **uspořádaný výběr**, angl.: *sample order statistics*.

## Příklad (Uspořádaný číslicový dendrogram)

| stonek | listy | $f_i$ |
|---|---|---|
| 1 | 5 | 1 |
| 2 | 1 | 1 |
| 5 | 3079 | 4 |
| 6 | 1754037 | 7 |
| 7 | 506643565569160 | 15 |
| 8 | 68638607 | 8 |
| 9 | 6380 | 4 |

$\Longmapsto$

| stonek | listy | $f_i$ |
|---|---|---|
| 1 | 5 | 1 |
| 2 | 1 | 1 |
| 5 | 0379 | 4 |
| 6 | 0134577 | 7 |
| 7 | 001345555666669 | 15 |
| 8 | 03666788 | 8 |
| 9 | 0368 | 4 |

# Výběrové percentily a kvantily

**Číselné charakteristiky rozložení dat odvozené z uspořádaných výběrů**

## Definice (Výběrový percentil, angl.: *sample percentile*)

Mějme reálné číslo $0 \leq p \leq 1$. Pak **$(100p)\%$ výběrový percentil** výběru $x_1, \ldots, x_n$ je číslo $\widetilde{\pi}_p \in \mathbb{R}$, které rozděluje výběr na dvě části tak, že (přibližně) $(n-1) \cdot p$ hodnot z výběru je menších než $\widetilde{\pi}_p$ a $(n-1) \cdot (1-p)$ hodnot je větších než $\widetilde{\pi}_p$.

- **Výběrový kvantil s hladinou $p = (100p)\%$** výběrový percentil.
- **Terminologie:** výběrový precentil (kvantil) / empirický percentil (kvantil)

## Příklad

Pro $p = 0.25$ je $25\%$ výběrový percentil (to jest výběrový kvantil s hladinou $0.25$) hodnota $\widetilde{\pi}_{0.25}$, pro kterou je (přibližně) $\frac{1}{4}$ ostatních hodnot z výběru menších než $\widetilde{\pi}_{0.25}$ a $\frac{3}{4}$ ostatních hodnot z výběru větších než $\widetilde{\pi}_{0.25}$.

# Stanovení výběrových percentilů

## Algoritmus (stanovení $\widetilde{\pi}_p$)

Pro uspořádaný výběr $x_0, \ldots, x_{n-1}$ (indexovaný od 0) a číslo $0 \leq p \leq 1$ stanovíme $(100p)\%$ výběrový percentil $\widetilde{\pi}_p$ podle jednoho z následujících bodů:

- Pokud $(n-1)p$ *je celé číslo*, pak stanovíme hodnotu $\widetilde{\pi}_p$ jako prvek z uspořádaného výběru, který se nachází na pozici $(n-1)p$, to jest $\widetilde{\pi}_p = x_{(n-1)p}$.
- Pokud $(n-1)p$ *není celé číslo*, to jest pokud $(n-1)p = r + s$, kde $s \in (0,1)$ a $r$ je nezáporné celé číslo, pro které $r < n-1$, pak

$$\widetilde{\pi}_p \;=\; x_r + s \cdot (x_{r+1} - x_r) \;=\; (1-s) \cdot x_r + s \cdot x_{r+1} \,,$$

kde $x_r$ a $x_{r+1}$ označují čísla z uspořádaného výběru na pozicích $r$ a $r+1$.

**Význam:** $\widetilde{\pi}_p$ je vážený průměr $x_r$ a $x_{r+1}$ s vahami $1-s$ a $s$ (lineární interpolace).

**Složitost:** $O(n \log n)$ (obecný výběr je potřeba uspořádat).

# Implementace

## Algoritmus (stanovení $\widetilde{\pi}_p$)

```
(defun percentile (sample p)
  "Return (100P)-th SAMPLE percentile."
  (let* ((sample (sort (copy-seq sample) #'<=))
         (n (length sample))
         (r (* (1- n) p)))
    (if (integerp r)
        (elt sample r)
        (multiple-value-bind (r s)
            (truncate r)
          (+ (* (- 1 s) (elt sample r))
             (* s (elt sample (1+ r))))))))
```

## Příklad (Výběrové kvantily)

Uvažujme následující uspořádané výběry:

1. $x_0 = 10$, $x_1 = 20$, $x_2 = 30$, $x_3 = 40$ a $x_4 = 50$
   - Pro $p = 0.25$ máme $(n-1) \cdot 0.25 = 4 \cdot 0.25 = 1$, to jest $\widetilde{\pi}_{0.25} = x_1 = 20$.
   - Pro $p = 0.5$ máme $(n-1) \cdot 0.5 = 4 \cdot 0.5 = 2$, to jest $\widetilde{\pi}_{0.50} = x_2 = 30$.
2. $x_0 = 10$, $x_1 = 20$, $x_2 = 30$, $x_3 = 40$, $x_4 = 50$ a $x_5 = 60$
   - Pro $p = 0.25$ máme $(n-1) \cdot 0.25 = 5 \cdot 0.25 = 1.25$, to jest $\widetilde{\pi}_{0.25} = x_1 + 0.25 \cdot (x_2 - x_1) = 20 + 0.25 \cdot 10 = 22.5$.
   - Pro $p = 0.5$ máme $(n-1) \cdot 0.5 = 5 \cdot 0.5 = 2.5$, to jest $\widetilde{\pi}_{0.50} = x_2 + 0.5 \cdot (x_3 - x_2) = 30 + 0.5 \cdot 10 = 35$.

# Vlastnosti výběrových percentilů

## Věta (o výběrových percentilech)

*Uvažujme uspořádaný výběr $x_0, \ldots, x_{n-1}$. Pak pro každé $k = 0, \ldots, n-1$ platí, že $x_k$ je výběrový kvantil s hladinou $\frac{k}{n-1}$. To jest, $x_k$ je $\left(100\frac{k}{n-1}\right)\%$ výběrový percentil.*

## Důkaz.

Triviálně plyne z faktů, že

$$(n-1)\cdot\frac{k}{n-1} = k, \quad (n-1)\cdot\left(1-\frac{k}{n-1}\right) = (n-1)\cdot\frac{n-1-k}{n-1} = n-1-k.$$

Jinými slovy, právě $k$ prvků z výběru je ostře menších než $x_k$ a právě $n-1-k$ prvků z výběru je ostře větších než $x_k$. Odtud $\widetilde{\pi}_{\frac{k}{n-1}} = x_k$ dle definice. □

**Důsledek:** každý prvek výběru je percentil pro nějaké $0 \leq p \leq 1$.

# Výběrové percentily se speciálními názvy

## Výběrový medián $\widetilde{m}$

Definujeme $\widetilde{m} = \widetilde{\pi}_{0.50}$, to jest:

- prostřední hodnota ve výběru,
- rozděluje hodnoty uspořádaného výběru „na dvě poloviny“.

## Výběrové kvartily $\widetilde{q}_n$

**dolní kvartil:** $\widetilde{q}_1 = \widetilde{\pi}_{0.25}$

**prostřední kvartil:** $\widetilde{q}_2 = \widetilde{\pi}_{0.50} = \widetilde{m}$

**horní kvartil:** $\widetilde{q}_3 = \widetilde{\pi}_{0.75}$

## Výběrové decily, . . .

**$n$tý decil:** $10n\%$ percentil, to jest:

**první decil:** $10\%$ percentil $= \widetilde{\pi}_{0.10}$

**druhý decil:** $20\%$ percentil $= \widetilde{\pi}_{0.20}$

$\vdots$

**devátý decil:** $90\%$ percentil $= \widetilde{\pi}_{0.90}$

## Příklad

Uspořádaný výběr obsahující skóre $40$ studentů na zkoušce:

$$\begin{array}{cccccccccccccccccccc} 15 & 21 & 50 & 53 & 57 & 59 & 60 & 61 & 63 & 64 & 65 & 67 & 67 & 70 & 70 & 71 & 73 & 74 & 75 & 75 \\ 75 & 75 & 76 & 76 & 76 & 76 & 76 & 79 & 80 & 83 & 86 & 86 & 86 & 87 & 88 & 88 & 90 & 93 & 96 & 98 \end{array}$$

**Vybrané percentily:**

$$\widetilde{\pi}_0 = 15.00, \quad \widetilde{\pi}_{0.5} = 48.55, \quad \widetilde{\pi}_{0.1} = 56.60, \quad \widetilde{\pi}_{0.15} = 59.85, \quad \widetilde{\pi}_{0.2} = 62.60,$$
$$\widetilde{\pi}_{0.25} = 64.75, \quad \widetilde{\pi}_{0.3} = 67.00, \quad \widetilde{\pi}_{0.35} = 70.00, \quad \widetilde{\pi}_{0.4} = 72.20, \quad \widetilde{\pi}_{0.45} = 74.55,$$
$$\widetilde{\pi}_{0.5} = 75.00, \quad \widetilde{\pi}_{0.55} = 75.45, \quad \widetilde{\pi}_{0.6} = 76.00, \quad \widetilde{\pi}_{0.65} = 76.00, \quad \widetilde{\pi}_{0.7} = 79.30,$$
$$\widetilde{\pi}_{0.75} = 83.75, \quad \widetilde{\pi}_{0.8} = 86.00, \quad \widetilde{\pi}_{0.85} = 87.15, \quad \widetilde{\pi}_{0.9} = 88.20, \quad \widetilde{\pi}_{0.95} = 93.15,$$

$$\widetilde{\pi}_{0.9525} = 93.44, \quad \widetilde{\pi}_{0.955} = 93.74, \quad \widetilde{\pi}_{0.9575} = 94.03, \quad \widetilde{\pi}_{0.96} = 94.32, \quad \widetilde{\pi}_{0.9625} = 94.61,$$
$$\widetilde{\pi}_{0.965} = 94.90, \quad \widetilde{\pi}_{0.9675} = 95.20, \quad \widetilde{\pi}_{0.97} = 95.49, \quad \widetilde{\pi}_{0.9725} = 95.78, \quad \widetilde{\pi}_{0.975} = 96.05,$$
$$\widetilde{\pi}_{0.9775} = 96.25, \quad \widetilde{\pi}_{0.98} = 96.44, \quad \widetilde{\pi}_{0.9825} = 96.64, \quad \widetilde{\pi}_{0.985} = 96.83, \quad \widetilde{\pi}_{0.9875} = 97.03,$$
$$\widetilde{\pi}_{0.99} = 97.22, \quad \widetilde{\pi}_{0.9925} = 97.42, \quad \widetilde{\pi}_{0.995} = 97.61, \quad \widetilde{\pi}_{0.9975} = 97.81, \quad \widetilde{\pi}_{1} = 98.00.$$

# Další hodnoty odvozené z uspořádaných výběrů

## Definice

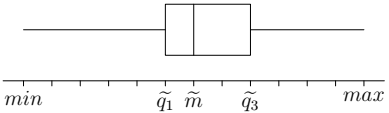
Pro uspořádaný výběr $x_1, \ldots, x_n$ zavádíme následující pojmy.

- **Průměr extrémů:** $\dfrac{x_n - x_1}{2}$, angl.: *midrange*.
- **Variační rozpětí:** $x_n - x_1$, angl.: *range* (Přednáška 1).
- **Interkvartilové rozpětí (IQR):** $\widetilde{q}_3 - \widetilde{q}_1$, angl.: *interquartile range*.
- **Percentilový průměr:** $\dfrac{\widetilde{q}_1 + 2 \cdot \widetilde{m} + \widetilde{q}_3}{4}$, angl.: *trimean*.
- **Pětihodnotový souhrn:** $x_1,\ \widetilde{q}_1,\ \widetilde{m},\ \widetilde{q}_3,\ x_n$, angl.: *five-number summary*.

# Krabicové grafy s anténami

**Grafická reprezentace pětihodnotového souhrnu**

angl.: *box-and-whisker diagram, boxplot*




## Postup při kreslení diagramu

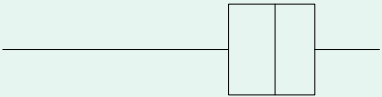
Pro uspořádaný výběr $x_1, \ldots, x_n$ postupujeme následovně:

1. Nakreslíme horizontální osu pro hodnoty $x_i$ (ve vhodném měřítku).
2. Nad osou zakreslíme obdélník (tzv. **krabice**) jehož krajní strany (obě kolmé k ose) jsou zakresleny na pozicích daných kvartily $\widetilde{q}_1$ a $\widetilde{q}_3$. Obdélník je dále předělen vertikální čárou na pozici dané mediánem $\widetilde{m} = \widetilde{q}_2$.
3. **Levá anténa** je nakreslena jako čára rovnoběžná s osou jdoucí od minima výběru k levé straně obdélníku.
4. Analogicky **pravá anténa** jde od pravé strany obdélníku do maxima výběru.

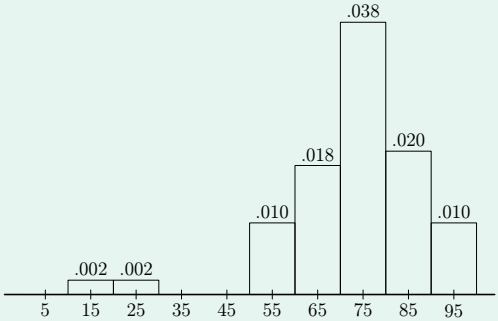
## Příklad (Krabicový graf s anténami)

Výběr obsahující skóre 40 studentů na zkoušce:

75 70 86 76 61 76 88 74 67 96 93 73 65 53 98 86 75 83 64 15
76 50 88 75 75 21 86 60 57 76 63 79 71 80 67 87 76 90 70 59



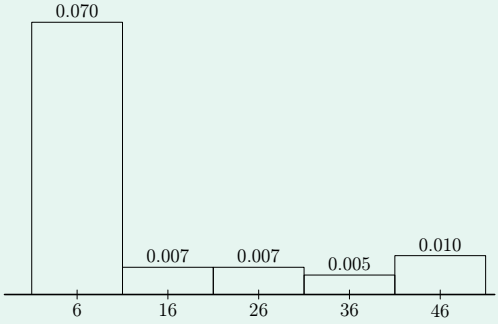
1. kvartil: $\widetilde{\pi}_{0.25} = \widetilde{q}_1 = 64.75$
medián: $\widetilde{\pi}_{0.50} = \widetilde{q}_2 = \widetilde{m} = 75.00$
3. kvartil: $\widetilde{\pi}_{0.75} = \widetilde{q}_3 = 83.75$



## Příklad (Krabicový graf s anténami)

Uspořádaný výběr soubor obsahující počty mrtvých při $40$ živelních pohromách:

$$2\ 2\ 2\ 2\ 2\ 2\ 2\ 3\ \ 3\ \ 3\ \ 3\ \ 3\ \ 4\ \ 4\ \ 4\ \ 4\ \ 4\ \ 5\ \ 5$$
$$5\ 5\ 6\ 6\ 6\ 6\ 8\ 9\ 12\ 16\ 18\ 22\ 25\ 29\ 32\ 39\ 41\ 47\ 48\ 50$$

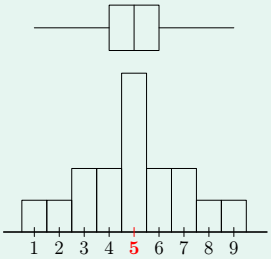



1. kvartil: $\widetilde{\pi}_{0.25} = \widetilde{q}_1 = 3.0$
medián: $\widetilde{\pi}_{0.50} = \widetilde{q}_2 = \widetilde{m} = 5.0$
3. kvartil: $\widetilde{\pi}_{0.75} = \widetilde{q}_3 = 16.5$

# Příklady (Výběrový průměr $\times$ výběrový medián)

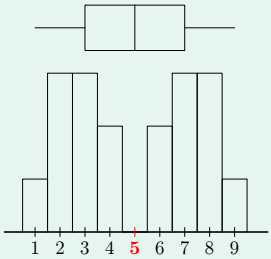
1 2 3 3 4 4 5 5 5 5 5
6 6 7 7 8 9

$\widetilde{q_1} = 4$
$\widetilde{m} = 5$
$\widetilde{q_3} = 6$



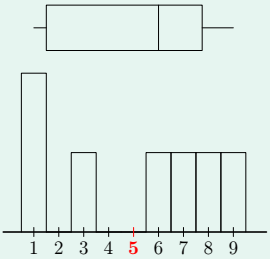

1 2 2 2 3 3 3 4 4 6 6
7 7 7 8 8 8 9

$\widetilde{q_1} = 3$
$\widetilde{m} = 5$
$\widetilde{q_3} = 7$




1 1 1 1 3 3 6 6 7 7 8
8 9 9

$\widetilde{q_1} = 1.5$
$\widetilde{m} = 6$
$\widetilde{q_3} = 7.75$



# Zkoumání přítomnosti odlehlých hodnot

**Odlehlá hodnota:**

- potenciálně podezřelá hodnota ve výběru,
- hodnoty, které jsou „příliš daleko“ od většiny hodnot ve výběru,
- mohou vznikat jako *chyby v datech* (např. chyby při měření).

**Krabicový graf s anténami a bariérami**

- **vnitřní bariéra** – vlevo a vpravo od obdélníka ve vzdálenosti $1.5\,\mathrm{IQR}$;
- **vnější bariéra** – vlevo a vpravo od obdélníka ve vzdálenosti $3\,\mathrm{IQR}$;

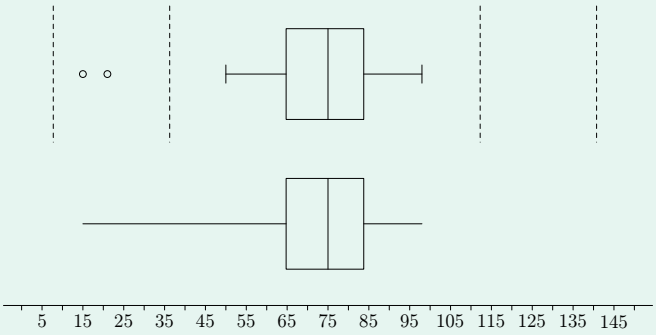
**Typy odlehlých hodnot** (v grafu se zakreslují kolečky)

- **mírně odlehlá hodnota** je hodnota nacházející se mezi vnitřní a vnější bariérou
  angl.: *mild outlier, suspected outlier*
- **silně odlehlá hodnota** je hodnota nacházející se za vnější bariérou
  angl.: *extreme outlier, outlier*

## Příklad

Výběr obsahující skóre 40 studentů na zkoušce:

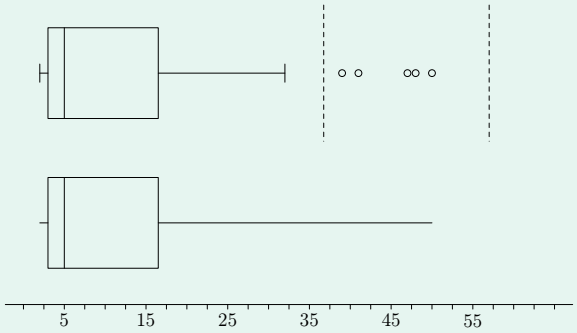
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 75 | 70 | 86 | 76 | 61 | 76 | 88 | 74 | 67 | 96 |
| 93 | 73 | 65 | 53 | 98 | 86 | 75 | 83 | 64 | 15 |
| 76 | 50 | 88 | 75 | 75 | 21 | 86 | 60 | 57 | 76 |
| 63 | 79 | 71 | 80 | 67 | 87 | 76 | 90 | 70 | 59 |



## Příklad

Uspořádaný výběr soubor obsahující počty mrtvých při $40$ živelních pohromách:

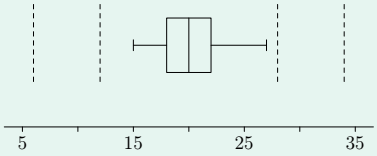
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 5 | 5 |
| 5 | 5 | 6 | 6 | 6 | 6 | 8 | 9 | 12 | 16 |
| 18 | 22 | 25 | 29 | 32 | 39 | 41 | 47 | 48 | 50 |



# Příklady (Citlivost $\overline{x}$ a $s^2$ na odlehlé hodnoty)

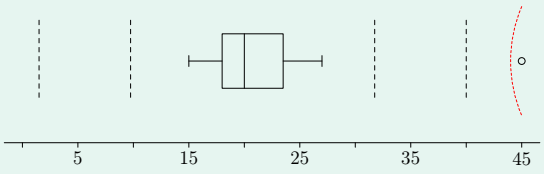
15 17 18 18 20 20 20 21 22 24 27

$$\overline{x} = 20.18$$
$$s^2 = 11.16$$
$$s = 3.34$$
$$\widetilde{q}_1 = 18$$
$$\widetilde{m} = 20$$
$$\widetilde{q}_3 = 21.5$$




15 17 18 18 20 20 20 21 22 24 27 10000

$$\overline{x} = 851.82$$
$$s^2 = 8\,299\,741$$
$$s = 2\,880.93$$
$$\widetilde{q}_1 = 18$$
$$\widetilde{m} = 20$$
$$\widetilde{q}_3 = 22.5$$



# Modus a modální intervaly

**Výběrový modus** = typická hodnota ve výběru (vyskytuje se nejčastěji)

## Definice (Výběrový modus)

Hodnota $y$, která má ve výběru $x_1, \ldots, x_n$ nejvyšší absolutní četnost, se nazývá **výběrový modus**, angl.: *mode*. Pokud ve výběru existuje několik různých hodnot se stejnou maximální četností, všechny hodnoty se považují za mody výběru.

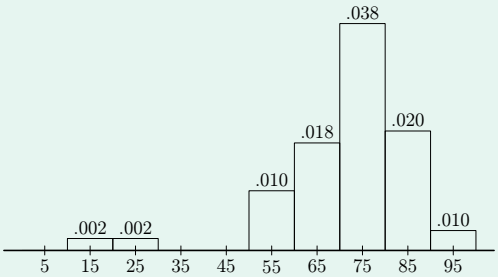
Pro data dělená do intervalů zavádíme následující:

## Definice (Modální interval)

Interval hodnot, který má největší relativní četnost, se nazývá **modální interval**, angl.: *modal class* (obecně jich může být více se stejnou maximální relativní četností). Střed modálního intervalu nazýváme **modus** výběru při zvoleném rozdělení na intervaly.

**Poznámka:** nejvyšší relativní četnost intervalu $= f_i \cdot h_i$ je nejvyšší

## Příklad (Histogram intervalového rozdělení četností)




Modální interval: $[70, 80)$; modus: $75$.

## Příklad (Výběrový modus $\neq$ výběrový průměr $\neq$ výběrový medián)

Pro výběr $1, 1, 1, 2, 3, 4, 5$ máme:

- výběrový průměr: $\overline{x} = \frac{17}{7} \approx 2.43$,
- výběrový medián: $\widetilde{m} = 2$,
- výběrový modus: $1$.

# Analýza závislostí dat

- Dva číselné výběry nebo více výběrů.
- Výběry jsou buď ze stejné populace nebo z jiné.
- Data mohou být stejného nebo různého typu (různé jednotky).
- **Základní otázka:** Jak porovnat data ve výběrech?
  - netriviální problém (obvykle vysoká složitost, pro velká data problém)
  - různé typy závislostí (lineární, nelineární, logické, . . . )
  - korelační analýza, regresní analýza, metody data-mining (MAGISTERSKÁ ETAPA)

**Dva základní přístupy**

- **Grafický:**
  - jednoduše interpretovatelný uživateli (laiky),
  - nedává přesný popis závislostí (posouzení je věcí intuice/psychologie).
- **Numerický:**
  - závislost v datech se vyjádří číselnou charakteristikou.

# Grafické metody založené na porovnávání diagramů

**Slouží k porovnávání dvou výběrů** $x_1, \ldots, x_n$ a $y_1, \ldots, y_n$ se

- stejnými jednotkami a se
- stejnou velikostí.

**Obvyklé znázornění:**

1. **Dva číslicové dendrogramy se společnými stonky**
   - Vypíšeme všechny možné hodnoty stonků z obou výběrů do jednoho sloupce.
   - Listy z prvního výběru se zapisují od stonku nalevo.
   - Listy z druhého výběru se zapisují od stonku napravo.
2. **Dva krabicové grafy s anténami kreslené nad sebe**
   - Nakreslíme společnou horizontální osu pro hodnoty z obou výběrů.
   - Krabicový graf prvního výběru se zakreslí nad osu.
   - Krabicový graf druhého výběru se zakreslí nad první diagram.

## Příklad (Porovnání výběrů pomocí dendrogramů se společnými stonky)

79 87 42 74 73 86 66 65 87 21 58 96 72 77 57 80 64 87 74 59
80 86 86 99 74 22 83 74 72 61 49 63 87 72 78 52 78 76 74 66

80 88 80 28 84 67 95 81 80 78 68 89 89 68 80 70 77 70 77 65
77 48 37 19 71 64 86 94 32 30 86 90 85 90 72 57 59 72 61 70

| $f_i$ | listy | stonek | listy | $f_i$ |
|---|---|---|---|---|
| 0 | | 1 | 9 | 1 |
| 2 | 21 | 2 | 8 | 1 |
| 0 | | 3 | 027 | 3 |
| 2 | 92 | 4 | 8 | 1 |
| 4 | 9872 | 5 | 79 | 2 |
| 6 | 665431 | 6 | 145788 | 6 |
| 14 | 98876444443222 | 7 | 0001227778 | 10 |
| 10 | 7777666300 | 8 | 000014566899 | 12 |
| 2 | 96 | 9 | 0045 | 4 |

# Příklad (Porovnání výběrů pomocí dendrogramů se společnými stonky)

54 56 76 84 70 72 70 71 50 64 72 86 61 69 87 71 72 89 10 59
76 37 62 78 77 93 83 52 30 65 78 69 55 62 72 79 62 94 66 95

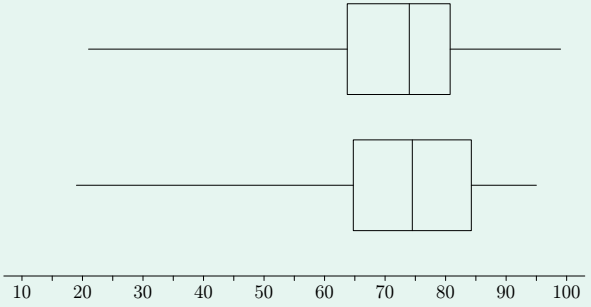
32 89 57 18 46 45 62 40 80 88 21 89 30 61 57 73 12 83 89 29
27 72 21 24 75 70 29 70 60 30 75 30 16 14 69 29 87 37 40 61

| $f_i$ | listy | stonek | listy | $f_i$ |
|---|---|---|---|---|
| 1 | 0 | 1 | 2468 | 4 |
| 0 | | 2 | 1147999 | 7 |
| 2 | 70 | 3 | 00027 | 5 |
| 0 | | 4 | 0056 | 4 |
| 6 | 965420 | 5 | 77 | 2 |
| 9 | 996542221 | 6 | 01129 | 5 |
| 14 | 98876622221100 | 7 | 002355 | 6 |
| 5 | 97643 | 8 | 0378999 | 7 |
| 3 | 543 | 9 | | 0 |

## Příklad (Porovnání výběrů pomocí krabicových diagramů s anténami)

79 87 42 74 73 86 66 65 87 21 58 96 72 77 57 80 64 87 74 59
80 86 86 99 74 22 83 74 72 61 49 63 87 72 78 52 78 76 74 66

80 88 80 28 84 67 95 81 80 78 68 89 89 68 80 70 77 70 77 65
77 48 37 19 71 64 86 94 32 30 86 90 85 90 72 57 59 72 61 70

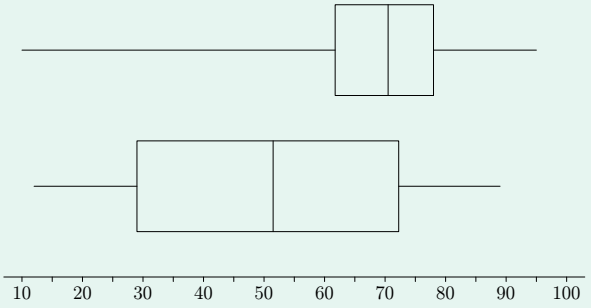



$\overline{x} = 70.90$
$s^2 = 284.19$

$\overline{x} = 70.35$
$s^2 = 362.75$

## Příklad (Porovnání výběrů pomocí krabicových diagramů s anténami)

54 56 76 84 70 72 70 71 50 64 72 86 61 69 87 71 72 89 10 59
76 37 62 78 77 93 83 52 30 65 78 69 55 62 72 79 62 94 66 95
32 89 57 18 46 45 62 40 80 88 21 89 30 61 57 73 12 83 89 29
27 72 21 24 75 70 29 70 60 30 75 30 16 14 69 29 87 37 40 61




$\overline{x} = 68.20$
$s^2 = 291.86$

$\overline{x} = 50.92$
$s^2 = 625.71$

# Kvantilové porovnání: kk-grafy

**Kvantily odpovídajících hladin jsou znázorněny jako body**
Uspořádané výběry $x_0, \ldots, x_{m-1}$ a $y_0, \ldots, y_{n-1}$, kde $m \leq n$.

## Sestrojení kk-grafu, angl.: *quantile-quantile-plot, qq-plot*

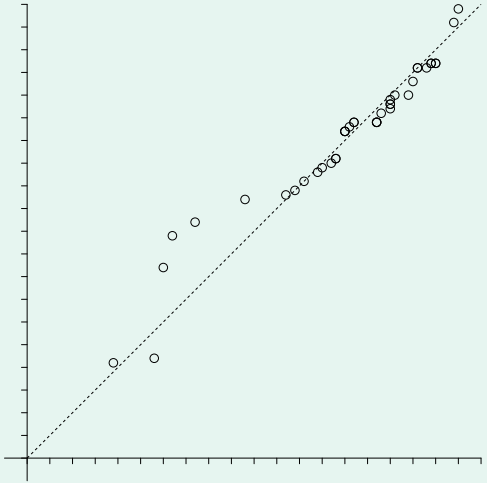
Kvantily dvou výběrů odpovídajících hladin jsou znázorněny jako body ve dvourozměrné souřadné soustavě: pro každou hodnotu $x_k$ z prvního výběru zakreslíme do grafu bod o souřadnicích $\left[x_k, \widetilde{\pi}_p[y]\right]$, kde $p = \frac{k}{m-1}$ a $\widetilde{\pi}_p[y]$ označuje výběrový kvantil s hladinou $p$ z výběru $y_0, \ldots, y_{n-1}$.

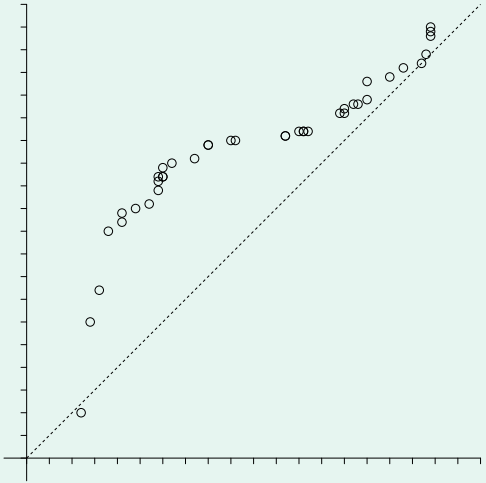
**Speciální případ: dva výběry se stejnou velikostí**

- pokud $m = n$ a $p = \frac{k}{m-1}$, pak $\widetilde{\pi}_p[y] = y_k$ (viz předchozí Větu), to jest
- v grafu se zaznamenají body $[x_0, y_0], [x_1, y_1], \ldots, [x_n, y_n]$.

**Čtení grafu:** body (blízko) na diagonále = (skoro) stejné výběry

# Příklad (Kvantilové porovnání pro dvojice dat z předchozích příkladů)





## Příklad (Kvantilové porovnání výběrů o různých velikostech)

Setříděné výběry $x_0, \ldots, x_{20}$ (velikost 21) a $y_0, \ldots, y_{39}$ (velikost 40):

16 24 33 35 38 40 45 46 51 53 56 59 60 63 66 68 69 70 71 72 83

12 13 20 27 34 37 40 40 41 42 42 42 44 50 50 51 51 52 53 54
54 54 55 55 56 58 59 59 59 60 67 70 70 77 81 82 83 84 86 99

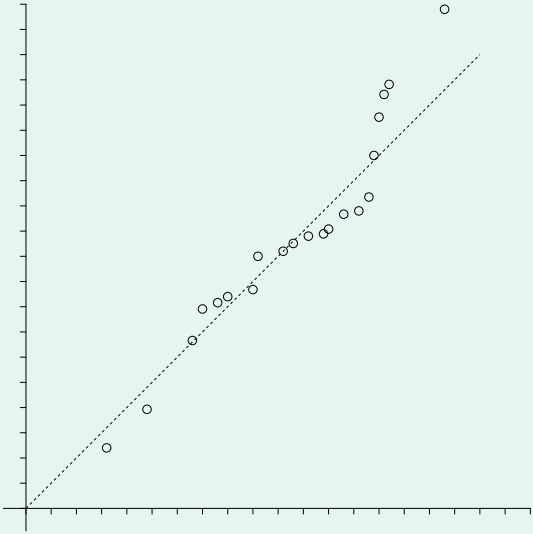
**Tabulky s porovnáním kvantilů:**

| $x_i$ | $k_i$ | $\widetilde{\pi}_{k_i}[y]$ |
|---|---|---|
| 16 | 0.00 | 12.00 |
| 24 | 0.05 | 19.65 |
| 33 | 0.10 | 33.30 |
| 35 | 0.15 | 39.55 |
| 38 | 0.20 | 40.80 |
| 40 | 0.25 | 42.00 |
| 45 | 0.30 | 43.40 |

| $x_i$ | $k_i$ | $\widetilde{\pi}_{k_i}[y]$ |
|---|---|---|
| 46 | 0.35 | 50.00 |
| 51 | 0.40 | 51.00 |
| 53 | 0.45 | 52.55 |
| 56 | 0.50 | 54.00 |
| 59 | 0.55 | 54.45 |
| 60 | 0.60 | 55.40 |
| 63 | 0.65 | 58.35 |

| $x_i$ | $k_i$ | $\widetilde{\pi}_{k_i}[y]$ |
|---|---|---|
| 66 | 0.70 | 59.00 |
| 68 | 0.75 | 61.75 |
| 69 | 0.80 | 70.00 |
| 70 | 0.85 | 77.60 |
| 71 | 0.90 | 82.10 |
| 72 | 0.95 | 84.10 |
| 83 | 1.00 | 99.00 |

# Příklad (Kvantilové porovnání výběrů o různých velikostech: graf)



# Úvod do korelační a regresní analýzy

**Čím se zabývají:**

- **Korelační analýza**
  - zkoumá vztahy hodnot dvou (nebo více) veličin,
  - grafické metody, číselné charakteristiky (korelační koeficienty).
- **Regresní analýza**
  - zkoumá jaké typy vztahů mezi proměnnými existují (lineární, kvadratický, ...),
  - statistická disciplína (teorie odhadů).

**Zajímáme se o vztahy mezi hodnotami ze dvou výběrů** tvořených

1. hodnotami **nezávislé veličiny** (například hmotnost auta, ...) a
2. hodnotami **vysvětlované veličiny** (například spotřeba benzínu ...)

**Aplikace:**

- vytváření zjednodušených modelů (například lineárních modelů);
- predikce hodnot založená na interpolaci/extrapolaci.

## Příklad (Analýza závislostí spotřeby automobilu na jeho hmotnosti)

| typ auta | hmotnost | spotřeba |
|---|---|---|
| AMC Concord | 3.4 | 5.5 |
| Chevrolet Caprice | 3.8 | 5.9 |
| Ford Country Squire Wagon | 4.1 | 6.5 |
| Chevrolet Chevette | 2.2 | 3.3 |
| Toyota Corona | 2.6 | 3.6 |
| Ford Mustang Ghia | 2.9 | 4.6 |
| Mazda GLC | 2.0 | 2.9 |
| AMC Sprint | 2.7 | 3.6 |
| VW Rabbit | 1.9 | 3.1 |

1. **nezávislá veličina:** hmotnost auta (v tisících liber)
2. **vysvětlované veličina:** spotřeba (počet galonů paliva na 100 mil).

**Otázka:** Lze (přibližně) vyjádřit spotřebu na základě hmotnosti auta?

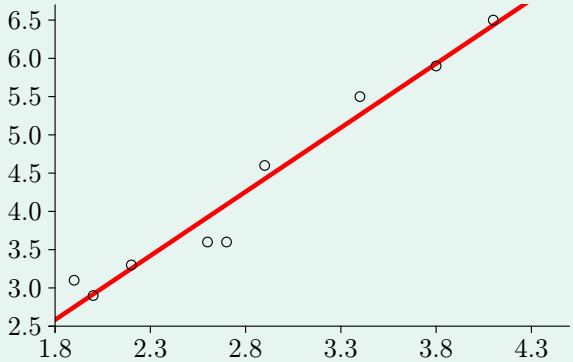
H. V. Henderson, P. F. Velleman: Building Multiple Regression Models Interactively, *Biometrics* **37**(1981), 391–411.

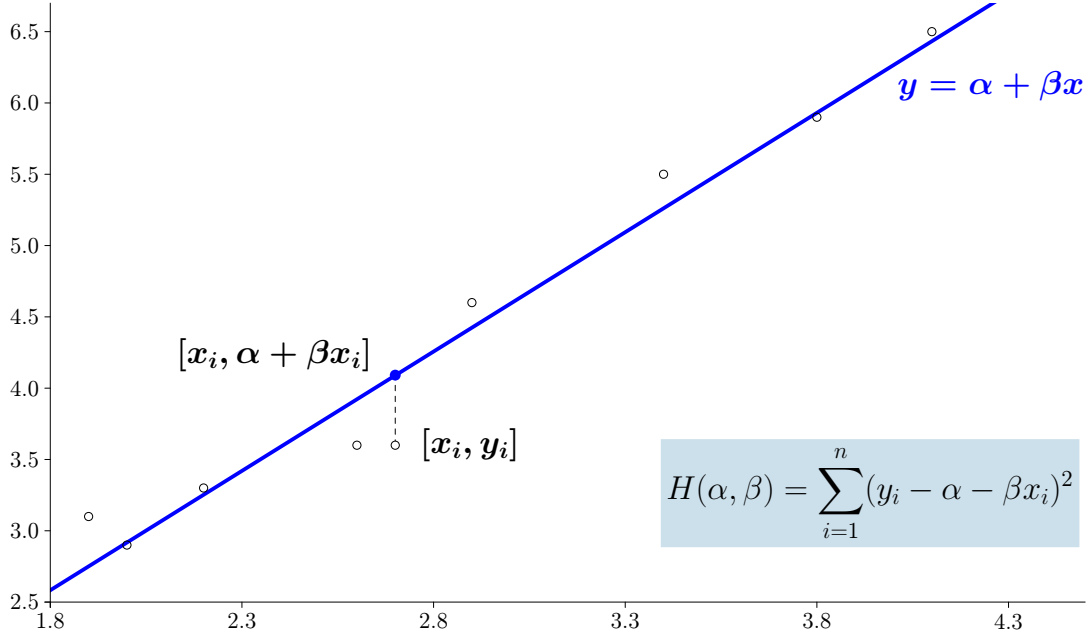
# Příklad (Bodový diagram a regresení přímka)

**Hodnoty vstupních dat (výběrů):**

| $x_i$ | 3.4 | 3.8 | 4.1 | 2.2 | 2.6 | 2.9 | 2.0 | 2.7 | 1.9 |
|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 5.5 | 5.9 | 6.5 | 3.3 | 3.6 | 4.6 | 2.9 | 3.6 | 3.1 |

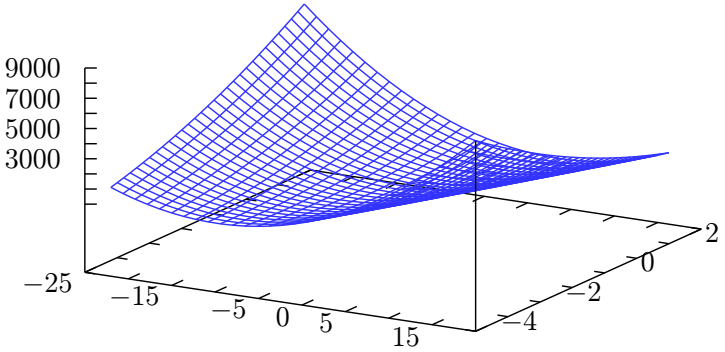
**Bodový diagram:**




6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
1.8
2.3
2.8
3.3
3.8
4.3
$y = \alpha + \beta x$
$[x_i, \alpha + \beta x_i]$
$[x_i, y_i]$
$H(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

# Aplikace metody nejmenších čtverců

Předpis $H(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ definuje funkci dvou proměnných $\alpha$ a $\beta$.




Snaha minimalizovat $H(\alpha, \beta)$ (aplikace vyšetřování průběhu fce. dvou proměnných).

# Stanovení parciálních derivací $H(\alpha, \beta)$

Pro $H(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ máme:

$$\frac{\partial H(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i) \cdot (-1) = \sum_{i=1}^{n} (-2y_i + 2\alpha + 2\beta x_i)$$

$$= -2 \sum_{i=1}^{n} y_i + 2 \sum_{i=1}^{n} \alpha + 2 \sum_{i=1}^{n} \beta x_i = -2 \sum_{i=1}^{n} y_i + 2n\alpha + 2 \left( \sum_{i=1}^{n} x_i \right) \beta,$$

$$\frac{\partial H(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i) \cdot (-x_i) = \sum_{i=1}^{n} (-2x_i y_i + 2x_i \alpha + 2x_i^2 \beta)$$

$$= -2 \sum_{i=1}^{n} x_i y_i + 2 \left( \sum_{i=1}^{n} x_i \right) \alpha + 2 \left( \sum_{i=1}^{n} x_i^2 \right) \beta.$$

# Hledání bodu minimalizujícího $H(\alpha, \beta)$

Položíme $\dfrac{\partial H(\alpha,\beta)}{\partial \alpha} = 0$ a $\dfrac{\partial H(\alpha,\beta)}{\partial \beta} = 0$, to jest:

$$-2\sum_{i=1}^{n} y_i + 2n\alpha + 2\left(\sum_{i=1}^{n} x_i\right)\beta = 0, \quad -2\sum_{i=1}^{n} x_i y_i + 2\left(\sum_{i=1}^{n} x_i\right)\alpha + 2\left(\sum_{i=1}^{n} x_i^2\right)\beta = 0.$$

Zjednodušení:

$$n\alpha + \left(\sum_{i=1}^{n} x_i\right)\beta = \sum_{i=1}^{n} y_i, \qquad \left(\sum_{i=1}^{n} x_i\right)\alpha + \left(\sum_{i=1}^{n} x_i^2\right)\beta = \sum_{i=1}^{n} x_i y_i.$$

Hledaný bod $[\widehat{\alpha}, \widehat{\beta}]$ minimalizující $H(\alpha, \beta)$ je bod, který je řešením této soustavy.

Vynásobením rovnic

$$n\alpha + \left(\sum_{i=1}^{n} x_i\right)\beta = \sum_{i=1}^{n} y_i, \qquad \left(\sum_{i=1}^{n} x_i\right)\alpha + \left(\sum_{i=1}^{n} x_i^2\right)\beta = \sum_{i=1}^{n} x_i y_i$$

konstantami $\sum_{i=1}^{n} x_i$ a $n$ dostaneme:

$$\left(\sum_{i=1}^{n} x_i\right) n\alpha + \left(\sum_{i=1}^{n} x_i\right)^2 \beta = \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right), \tag{1}$$

$$n\left(\sum_{i=1}^{n} x_i\right)\alpha + n\left(\sum_{i=1}^{n} x_i^2\right)\beta = n\sum_{i=1}^{n} x_i y_i \tag{2}$$

Odečtením (1) od (2) eliminujeme $\alpha$ a dostaneme:

$$n\left(\sum_{i=1}^{n} x_i^2\right)\beta - \left(\sum_{i=1}^{n} x_i\right)^2 \beta = n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right).$$

# Vyjádření $\widehat{\alpha}$ a $\widehat{\beta}$ minimalizujících $H(\alpha, \beta)$

Vyjádřením $\beta$ z rovnice

$$n\left(\sum_{i=1}^{n} x_i^2\right)\beta - \left(\sum_{i=1}^{n} x_i\right)^2\beta = n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

dostáváme řešení $\widehat{\beta}$:

$$\widehat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}.$$

# Vyjádření $\widehat{\alpha}$ a $\widehat{\beta}$ minimalizujících $H(\alpha, \beta)$

Použitím předchozí rovnice $n\alpha + \left(\sum_{i=1}^{n} x_i\right)\beta = \sum_{i=1}^{n} y_i$ vyjádříme $\widehat{\alpha}$:

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} x_i\right)\widehat{\beta}}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\widehat{\beta} = \overline{y} - \overline{x}\widehat{\beta}.$$

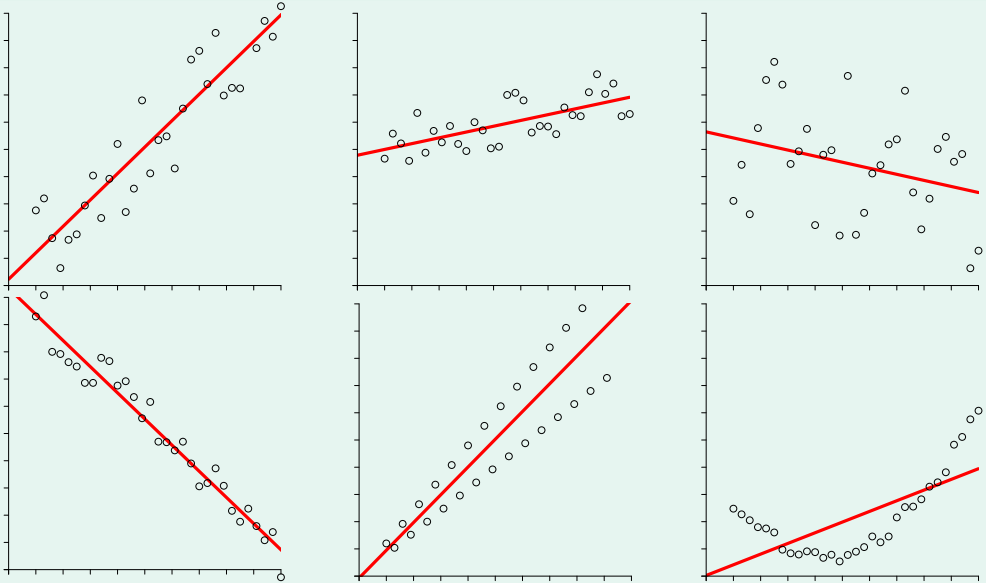
## Důsledek (tvar rovnice regresní přímky)

$$\widehat{y} \;=\; \widehat{\alpha} + \widehat{\beta}x \;=\; \overline{y} - \overline{x}\widehat{\beta} + \widehat{\beta}x \;=\; \overline{y} + \widehat{\beta}(x - \overline{x})$$

**Legenda:**

- $x$ je hodnota nezávislé veličiny, $\overline{x}$, $\overline{y}$ jsou výběrové průměry,
- $\widehat{\alpha}$, $\widehat{\beta}$ jsou vypočtené koeficienty, $\widehat{y}$ je (odhadovaná) hodnota vysvětlované veličiny.

# Příklady (Regresní přímky)



# Zjednodušení čitatele koeficientu $\widehat{\beta}$

Čitatele ve výrazu pro $\widehat{\beta}$ lze zjednodušit následovně:

$$\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n} x_i y_i - \overline{x}\left(\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \overline{x} y_i$$

$$= \sum_{i=1}^{n} (x_i y_i - \overline{x} y_i) = \sum_{i=1}^{n} (x_i y_i - \overline{x} y_i) - \overline{y} \cdot 0 = \sum_{i=1}^{n} (x_i y_i - \overline{x} y_i) - \overline{y} \cdot \sum_{i=1}^{n} (x_i - \overline{x})$$

$$= \sum_{i=1}^{n} (x_i y_i - \overline{x} y_i) - \sum_{i=1}^{n} (\overline{y} x_i - \overline{y}\,\overline{x}) = \sum_{i=1}^{n} (x_i y_i - \overline{x} y_i - \overline{y} x_i + \overline{y}\,\overline{x})$$

$$= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

# Zjednodušení koeficientu $\widehat{\beta}$

Použitím předchozího pozorování:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

Použitím výpočtového tvaru $s^2$ (Přednáška 1) lze zjednodušit:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

# Kovariance a korelační koeficient

## Definice (Kovariance, korelační koeficient)

Mějme $x_1, \ldots, x_n$ a $y_1, \ldots, y_n$ reprezentující hodnoty nezávislé a vysvětlované veličiny. **Výběrová kovariance** (angl.: *covariance*) je číslo

$$c_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \,.$$

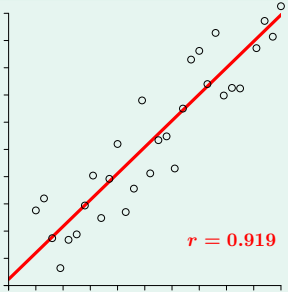
Číslo $r$ definované vztahem

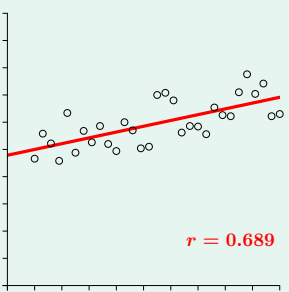
$$r = \frac{c_{xy}}{s_x \cdot s_y},$$

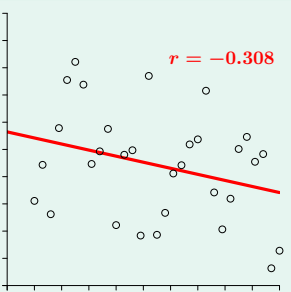
kde $s_x$ je směrodatná odchylka $x_1, \ldots, x_n$ a $s_y$ je směrodatná odchylka $y_1, \ldots, y_n$, se nazývá **výběrový korelační koeficient** (angl.: *correlation coefficient*).

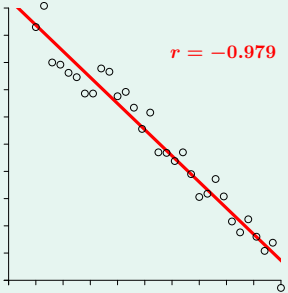
**Terminologie:** někdy uveden jako „Pearsonův korelační koeficient“.

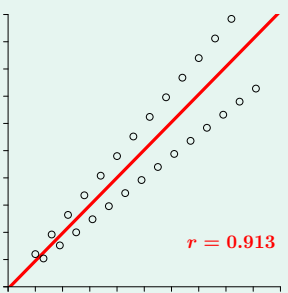
# Příklady (Regresní přímky a korelační koeficienty)

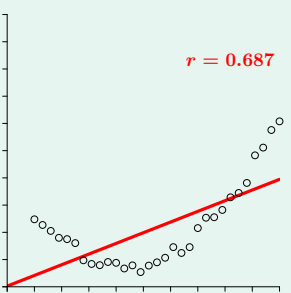


# Vlastnosti kovariance a korelačního koeficientu

## Věta

*Pro hodnoty $\widehat{\beta}$, $r$ a $\widehat{y}$ platí následující:*

$$\widehat{\beta} = \frac{c_{xy}}{s_x^2} = r \cdot \frac{s_y}{s_x},$$

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \overline{x}}{s_x} \right) \left( \frac{y_i - \overline{y}}{s_y} \right),$$

$$\widehat{y} = \overline{y} + \frac{c_{xy}}{s_x^2} \cdot (x - \overline{x}) = \overline{y} + r \cdot \frac{s_y}{s_x} \cdot (x - \overline{x}).$$

## Důkaz.

První rovnost plyne z faktů:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{c_{xy}}{s_x^2}, \qquad r \cdot s_x \cdot s_y = c_{xy} = \widehat{\beta} \cdot s_x^2,$$

odtud $\widehat{\beta} = \dfrac{c_{xy}}{s_x^2} = \dfrac{r \cdot s_x \cdot s_y}{s_x^2} = r \cdot \dfrac{s_y}{s_x}$. Druhá rovnost je důsledkem

$$r = \frac{c_{xy}}{s_x \cdot s_y} = \frac{1}{s_x \cdot s_y} \cdot \frac{1}{n-1} \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \overline{x}}{s_x} \right) \left( \frac{y_i - \overline{y}}{s_y} \right).$$

Třetí rovnost je důsledkem předcházejících. □

# Vlastnosti korelačního koeficientu a regresní přímky

## Věta

*Pro hodnoty $r$, $\overline{x}$ a $\overline{y}$ platí následující:*

1. *bod $[\overline{x}, \overline{y}]$ leží na regresní přímce,*
2. *rovnice $\sum_{i=1}^{n} \big((y_i - \overline{y}) - t(x_i - \overline{x})\big)^2 = 0$ má nejvýš jeden reálný kořen,*
3. *pro korelační koeficient $r$ máme $-1 \leq r \leq 1$.*

## Důkaz (začátek).

První tvrzení plyne přímo z toho, že $\widehat{y} = \overline{y} + \widehat{\beta} \cdot (x - \overline{x})$.
Druhé tvrzení plyne z toho, že $\big((y_i - \overline{y}) - t(x_i - \overline{x})\big)^2 \geq 0$ pro každé reálné číslo $t$.
To jest, suma je nulová, právě když jsou všechny $\big((y_i - \overline{y}) - t(x_i - \overline{x})\big)^2 = 0$. To nastane, právě když $t = \dfrac{y_i - \overline{y}}{x_i - \overline{x}}$ $(1 \leq i \leq n)$; to jest, $t$ je jednoznačně dané.

## Důkaz (dokončení).

Rovnici $\sum_{i=1}^{n}\big((y_i-\overline{y})-t(x_i-\overline{x})\big)^2=0$ můžeme zapsat jako $at^2+bt+c=0$:

$$\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)t^2-2\left(\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right)t+\left(\sum_{i=1}^{n}(y_i-\overline{y})^2\right)=0. \qquad (3)$$

Dle předchozího bodu má (3) nejvýš jeden reálný kořen, to jest pro diskriminant je

$$D=4\left(\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right)^2-4\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\overline{y})^2\right)\leq 0.$$

Odtud dostáváme

$$\left(\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right)^2\Bigg/\left[\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\overline{y})^2\right)\right]=r^2\leq 1.$$

□

# Poznámky o významu korelačního koeficientu

Ze tvaru regresní přímky $\widehat{y} = \overline{y} + r \cdot \frac{s_y}{s_x} \cdot (x - \overline{x})$ lze odvodit následující:

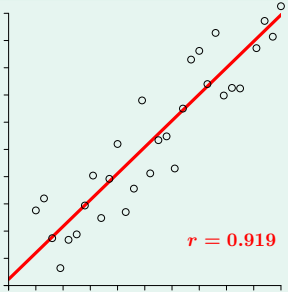
## Poznámky: Kladná a záporná asociace hodnot

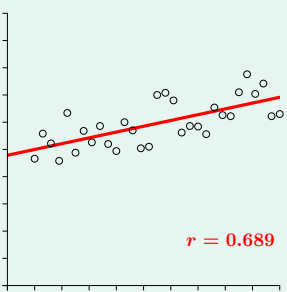
Korelační koeficient je *míra lineární asociace* mezi hodnotami.

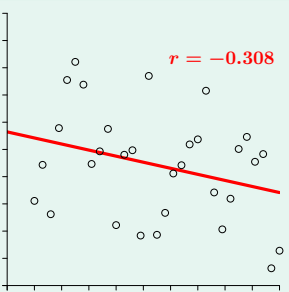
- Pokud $r = 1$, pak všechny body $[x_i, y_i]$ leží na přímce se směrnicí $\frac{s_y}{s_x}$;
- pokud $r = -1$, pak všechny body $[x_i, y_i]$ leží na přímce se směrnicí $-\frac{s_y}{s_x}$;
- **kladná asociace hodnot**: $r > 0$ (čím bližší $1$, tím silnější);
- **záporná asociace hodnot**: $r < 0$ (čím bližší $-1$, tím silnější);
- **slabá asociace hodnot**: $r = 0$.

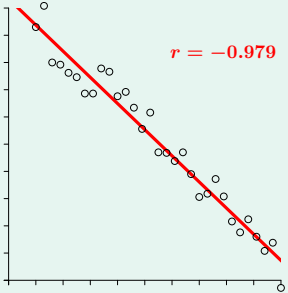
Korelační koeficient = numerická charakteristika síly asociace.

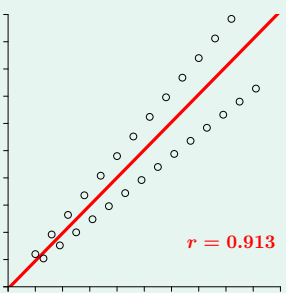
# Příklady (Regresní přímky a korelační koeficienty)

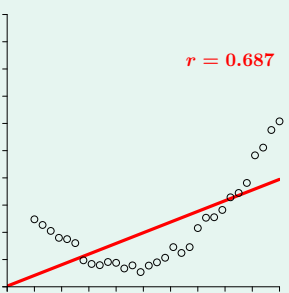


# Nelineární regrese

**Obecný problém:**

- místo funkce $\alpha + \beta x$ vezmeme obecně (nelineární) funkci $H(\alpha, \beta, \dots)$,
- nalezení analogické lineárnímu případu (technické komplikace).

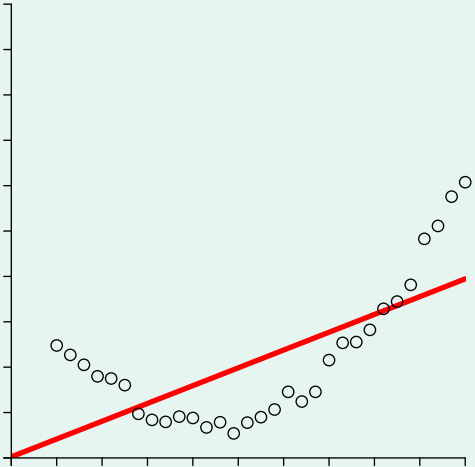
## Příklad (Kvadratická regrese: uvažovaná funkce $\alpha + \beta x + \gamma x^2$)

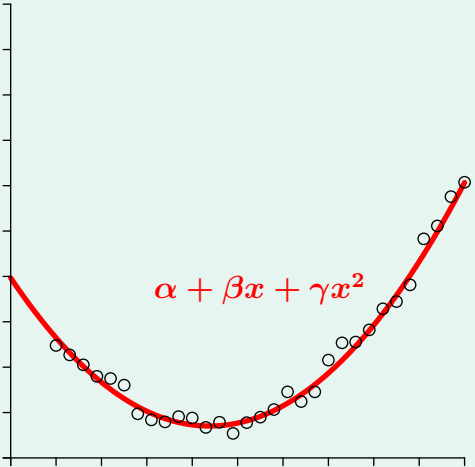
Hledáme bod minimalizující hodnotu $H(\alpha, \beta, \gamma) = \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i - \gamma x_i^2\right)^2$.

Vede na nalezení řešení soustavy:

$$\begin{aligned}
\alpha n + \beta \sum_{i=1}^{n} x_i + \gamma \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i, \\
\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 + \gamma \sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} x_i y_i, \\
\alpha \sum_{i=1}^{n} x_i^2 + \beta \sum_{i=1}^{n} x_i^3 + \gamma \sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} x_i^2 y_i.
\end{aligned}$$

## Příklad (Příklady lineární a kvadratické regrese)





# Přednáška 2: Závěr

**Pojmy:**

- uspořádaný výběr, percentil, kvartil, medián, kvantil
- číslicový dendrogram, krabicový graf, kk-graf,
- nezávislá veličina, vysvětlovaná veličina, regresní přímka
- výběrová kovariance, výběrový korelační koeficient

**Použité zdroje:**



Bílková D., Budínský P., Vohánka V.: *Pravděpodobnost a statistika*
Vydavatelství a nakl. Aleš Čeněk, s.r.o. 2009, ISBN 978–80–7380–224–0.


Hogg R. V., Tanis E. A.: *Probability and Statistical Inference*
Prentice Hall; 7. vydání 2005, ISBN 978–0–13–146413–1.


Tukey J. W.: *Exploratory Data Analysis*
Addison Wesley; 1. vydání 1977, ISBN 978–0–201–07616–5.