

# Přednáška č. 1

## Popisná statistika

16. 2. 2022

P. Pecherková

# Organizace předmětu

- Přednášející: P. Pecherková, E. Uglickich
- Cvičící: P. Pecherková, E. Uglickich, P. Provinský, R. Likhonina
- Pavla Pecherková:
  - Konzultace – domluva přes chat v MS Teams (preferováno), e-mail
  - Mail: [nemcopav@fd.cvut.cz](mailto:nemcopav@fd.cvut.cz) (preferováno), [pavla.pecherkova@cvut.cz](mailto:pavla.pecherkova@cvut.cz)
  - Webové stránky předmětu: <https://zolotarev.fd.cvut.cz/sis/>
  - TEAMS „stránky“ předmětu - přednášky: ??

- Výuka:
  - V každém týdnu se bere obsahově totéž ve všech hodinách (přednášky/cvičení).
  - Cvičení se nemusí navštěvovat, ale testy se musí psát u kmenového cvičícího.
- Zápočet
  - 2 testy
  - více informací – na cvičeních
- Zkouška
  - Zkoušky probíhají pouze v **letním zkouškovém období**
  - Zkouška je ústní
  - Znamky A-E – zkouší všichni
    - 5 otázek: 1 příklad (slovní úloha) + 4 teoretické otázky (otázky jsou včas umístěny na webových stránkách)

# Soubor, výběr, data

- Soubor: množina všech možných odpovědí
- Výběr: podmnožina získaných informací
  - pro správně definovaný výběr musíme znát další statistické údaje, uděláme později
- Data: (náš předpoklad) výběr, který je dostatečně velký a dává relativně shodné výsledky se souborem

- Soubor: všichni lidé na Zemi



- Výběr:
  - ~~všichni lidé v ČR~~
  - 200 000 lidí náhodně vybraných po celé planetě
- Data:
  - 200 000 lidí náhodně vybraných po celé planetě



## DATA:

- Původní (raw)
  - data, která nám někdo dá z měření, ještě neupravená, bez jakéhokoliv zásahu. Většinou to bývá změřených nepřehledných údajů
  - $\{x_1; x_2; x_3; \dots; x_n\}$
- Uspořádaná
  - data, která seřadíme podle velikosti. Při malých počtech dat je zde možné sledovat nějaký trend a základní povahu souboru.
  - $\{x_2; x_{15}; x_{45}; \dots; x_1\}$  kde  $x_2 < x_{15} < x_{45} < \dots < x_1$
- Tříděná data
  - data, která ukazují unikátní hodnoty v souboru a ke každé hodnotě přiřadí jejich četnost (frekvenci)

|                    |       |       |         |       |
|--------------------|-------|-------|---------|-------|
| Hodnoty ( $x_i$ )  | $x_1$ | $x_2$ | $\dots$ | $x_k$ |
| Četnosti ( $n_i$ ) | $n_1$ | $n_2$ | $\dots$ | $n_k$ |

# Charakteristiky

- V případě, že chceme data jednodušeji popsat, ale i pochopit – použijeme charakteristiky
- Použitím charakteristik získáme zjednodušení (neúplný) popis dat
  - Proč neúplný?
    - Nevíme přesně jak se chová celý soubor, zjistíme o něm jen jednu věc
    - Příklad: Průměr – pro různá data může být stejný průměr. Minimum – známe minimální hodnotu, ale nic jiného
  - Proč ho přesto používáme?
    - Protože je to jednoduché a srozumitelné
    - Pro popis **musíme** využít víc charakteristik, tím zajistíme, že popis bude správně pochopen

Průměrná mzda v ČR

Zisky firmy KupTo stoupli průměrně o 100%, zisky firmy ProdejTo stoupli v průměru jen o 1%.

Lidé vykoupili sněžnice v cenové relaci od 1000 do 5000 Kč.

# Charakteristiky

## Charakteristiky

### polohy

### variability

Střední hodnota

Ostatní

Pracuje se  
střední hodnotou

Ostatní

Průměr

Medián

Modus

Min/max

Kvantil

Kritická  
hodnota

Rozptyl

Směrodatná  
odchylka

Rozpětí

Mezikvartilové  
rozpětí

# Charakteristiky polohy

## 1. Průměr

- Pro data netříděná:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$
- Pro data tříděná:  $\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i$

## 2. Medián: je prostřední hodnota uspořádaného souboru

- Pro lichý počet dat:  $\tilde{x}_{0.5} = \left(\frac{n+1}{2}\right)^{\text{tý prvek}}$  (prostřední prvek)
- Pro sudý počet dat:  $\tilde{x}_{0.5} = \frac{\left(\frac{n}{2}\right)^{\text{tý prvek}} + \left(\frac{n}{2}+1\right)^{\text{tý prvek}}}{2}$  (průměr z prostředních dvou prvků)

### Příklad:

Stav silnic jsme v dotazníku známkovali jako ve škole.

Odpovědělo nám 10 řidičů:

$x = \{2\ 2\ 3\ 5\ 1\ 1\ 5\ 5\ 4\ 2\}$

|       |   |   |   |   |   |
|-------|---|---|---|---|---|
| $x_i$ | 1 | 2 | 3 | 4 | 5 |
| $n_i$ | 2 | 3 | 1 | 1 | 3 |



3. Minimum / maximum: minimální a maximální hodnota prvku

4. Modus: prvek s největší četností

- Nejčtenější hodnota je 1 = data s jedním módem
- Nejčtenější hodnoty jsou 2 = data s dvěma mody
- Nejčtenějších hodnot je víc = data jsou bez modu

5. Kvantil: je hodnota prvku, která rozděluje data tak, že je  $\alpha \cdot 100\%$  rovno nebo menších a  $(1 - \alpha) \cdot 100\%$  je rovno nebo větších než daná hodnota.

- Dolní kvantil (Q1) – rozdělí data na 25 a 75 %
- Horní kvantil (Q3) – rozdělí data na 75 a 25 %

6. Kritická hodnota: je hodnota prvku, která rozděluje data tak, že je  $(1 - \alpha) \cdot 100\%$  rovno nebo menších a  $\alpha \cdot 100\%$  je rovno nebo větších než daná hodnota.

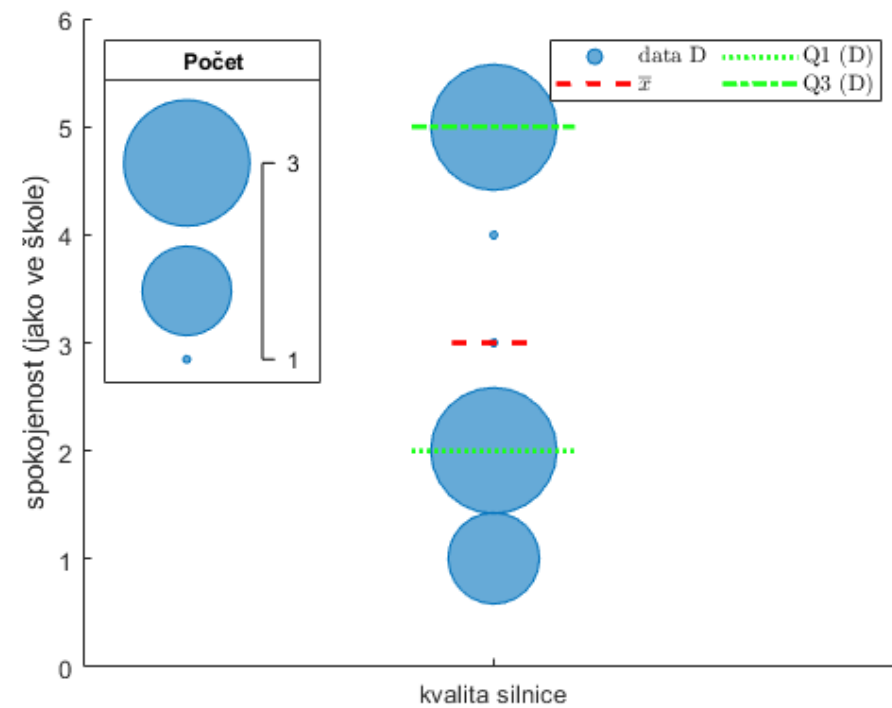
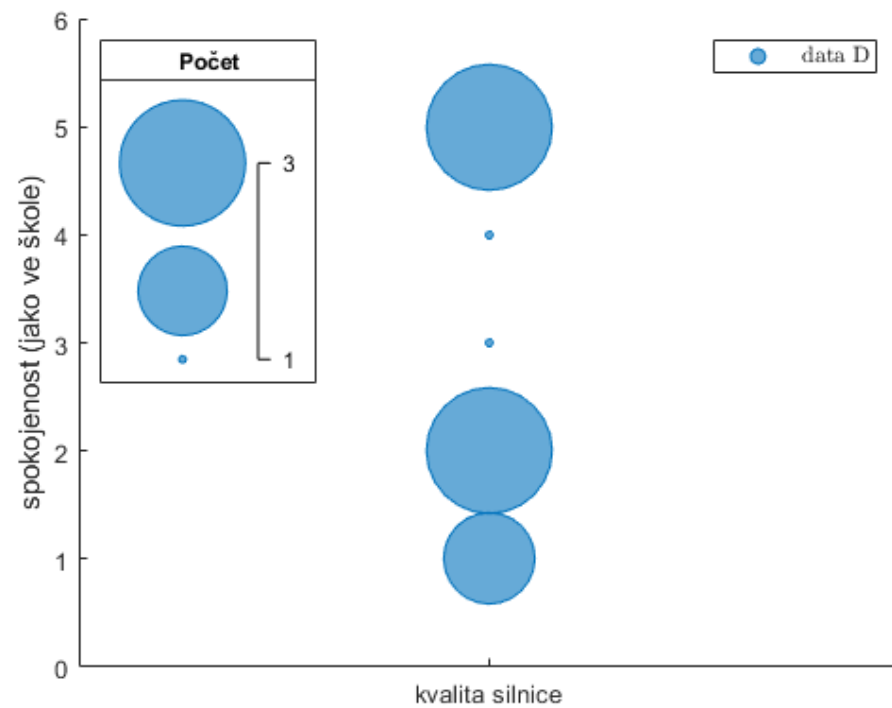
### Příklad:

Stav silnic jsme v dotazníku známkovali jako ve škole.

Odpovědělo nám 10 řidičů:

$x = \{2\ 2\ 3\ 5\ 1\ 1\ 5\ 5\ 4\ 2\}$

|       |   |   |   |   |   |
|-------|---|---|---|---|---|
| $x_i$ | 1 | 2 | 3 | 4 | 5 |
| $n_i$ | 2 | 3 | 1 | 1 | 3 |



# Charakteristiky variability

## 1. Rozptyl: míra rozptylu dat okolo aritmetického průměru

- Netříděná data  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  (výběr, data)
- Tříděná data  $s^2 = \frac{1}{n-1} \sum_{i=1}^k n_i (x_i - \bar{x})^2$  (výběr, data)

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \text{ (soubor)}$$

## 2. Směrodatná odchylka:

- Netříděná data  $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$  (výběr, data)
- $\sigma = \sqrt{\sigma^2}$  (soubor)

### Příklad:

Stav silnic jsme v dotazníku známkovali jako ve škole. Odpovědělo nám 10 řidičů:

$x = \{2 \ 2 \ 3 \ 5 \ 1 \ 1 \ 5 \ 5 \ 4 \ 2\}$

$\bar{x} = 3$

### 3. Rozpětí: rozdíl mezi maximem i minimem

- $R = \max(x) - \min(x)$

### 4. Mezikvartilové rozpětí: rozdíl mezi horním a dolním kvantilem

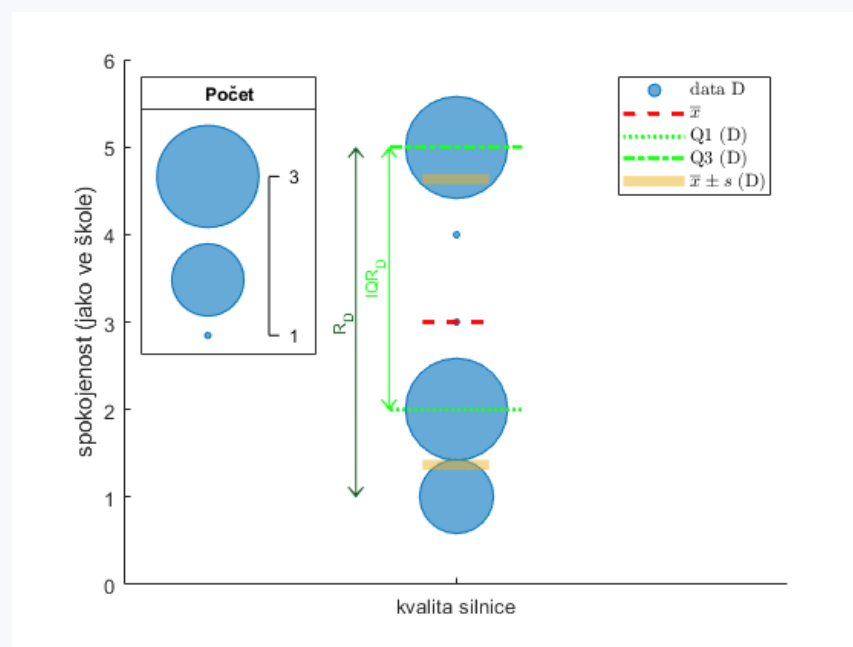
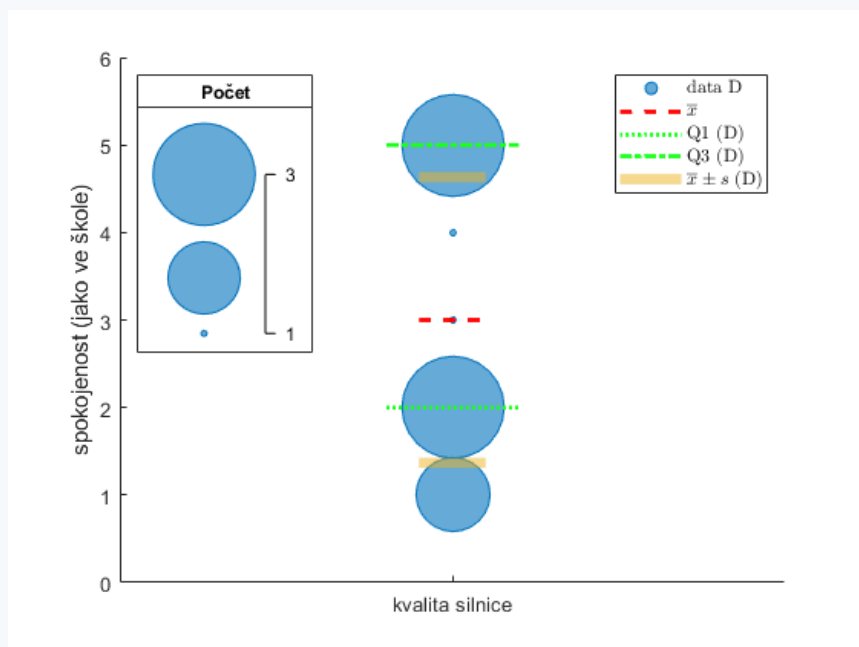
- $IQR = Q3 - Q1$

#### Příklad:

Stav silnic jsme v dotazníku známkovali jako ve škole. Odpovědělo nám 10 řidičů:

$$x = \{2 \ 2 \ 3 \ 5 \ 1 \ 1 \ 5 \ 5 \ 4 \ 2\}$$

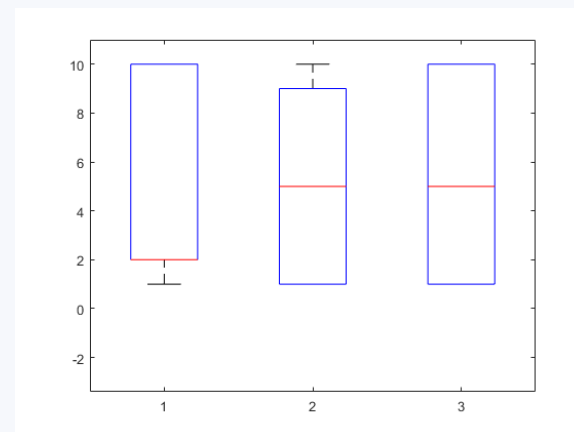
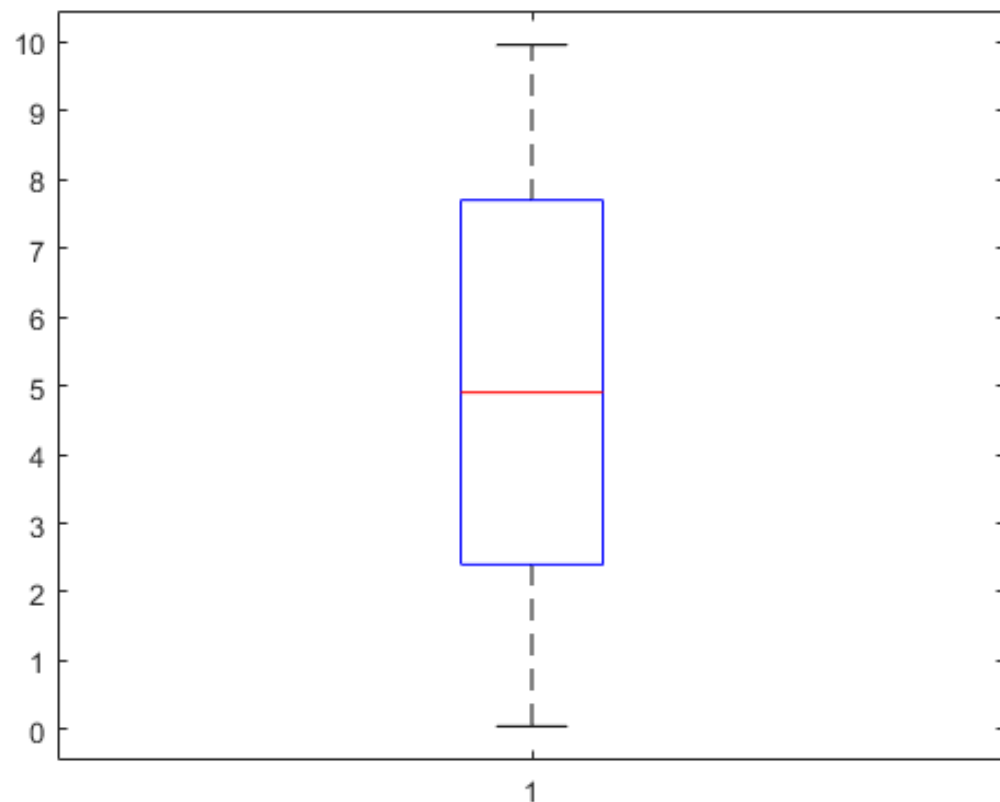
$$\bar{x} = 3$$



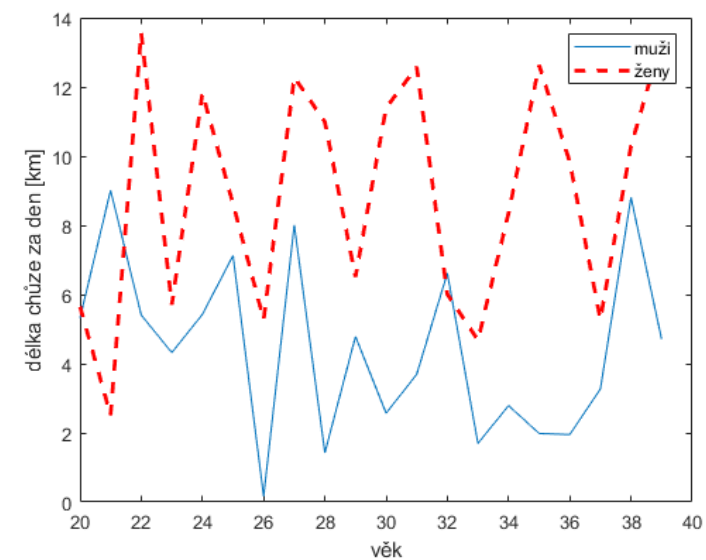
# Grafy

- Existuje velké množství grafů, pro každý problém se hodí jiný graf
- My budeme pracovat nejčastěji s následujícími grafy:
  1. Krabicový diagram (boxplot)
  2. Spojitý graf (časový)
  3. Sloupcový graf
  4. Histogram

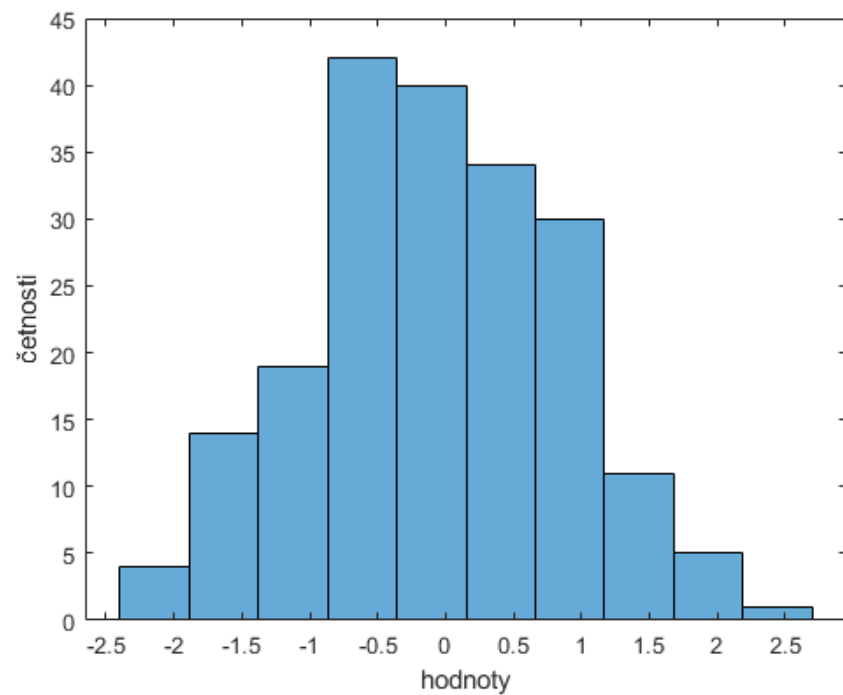
## Krabicový diagram



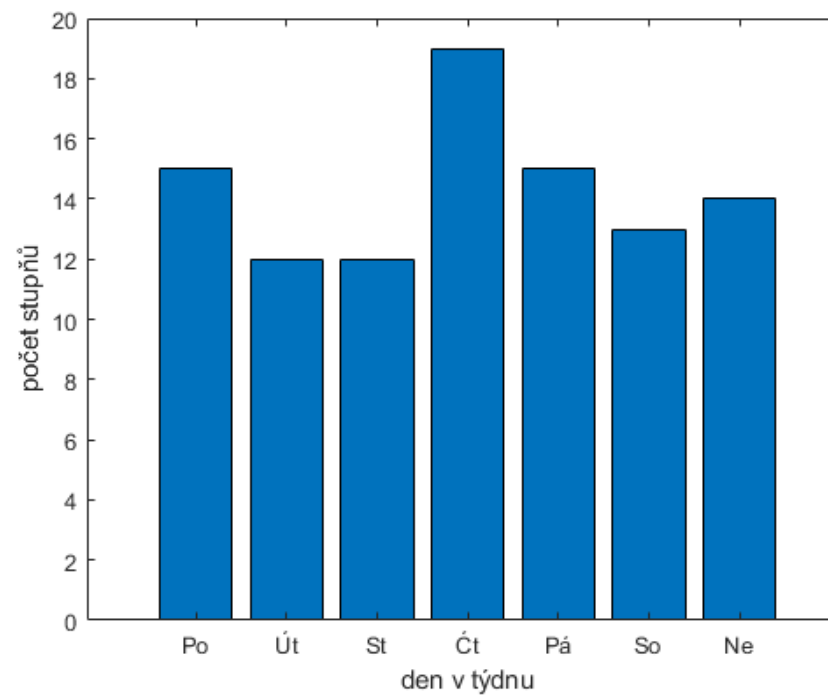
## Spojité graf (časový)



Histogram (graf četností)



Sloupkový graf



# Definice pravděpodobnosti

- Pravděpodobnost – „šance“ s jakou dosáhneme úspěchu  $P = \frac{\text{počet příznivých výsledků}}{\text{počet všech výsledků}} = \frac{p_0}{p}$
- Základní vlastnosti statistiky:
  - $P = 0$  : nemožný výsledek – Jaká je pravděpodobnost, že tato přednáška nikdy neskončí?
    - Nic méně než nemožné řešení není, proto  $P \geq 0 \rightarrow$  pravděpodobnost je nezáporná
  - $P = 1$  : jistý výsledek – Jaká je pravděpodobnost, že nebude-li pršet nezmoknem?
    - Nic víc než něco jistého být nemůže, proto  $P \leq 1 \rightarrow$  pravděpodobnost normovaná
  - Sečtu-li všechny možné možnosti, které ale nemohou zároveň nastat (nemohu zároveň hodit rub a líc), musí být pravděpodobnost  $\sum p_i = 1 \rightarrow$  pravděpodobnost je aditivní