KMA/MAT1 Přednáška č. 3,

Úvod do pravděpodobnosti a statistiky

3. října 2016

1 Pravděpodobnost [Otipka, Šmajstrla]

1.1 Náhodný pokus, náhodný jev

Teorie pravděpodobnosti vychází ze studia náhodných pokusů.

Náhodný pokus

- je proces, který při opakování dává za stejných podmínek rozdílné výsledky.
- Výsledek pokusu není předem znám (výsledek není jednoznačně určen jeho podmínkami), ale je předem dána množina možných výsledků.

Náhodný jev

— Každý možný výsledek náhodného pokusu nazýváme **elementárním náhodným jevem**,

značíme E_1, E_2, ..., E_n.

— Všechny elementární jevy tvoří tzv. **základní prostor elementárních jevů**;

značí se Ω.

— Každá podmnožina základního prostoru Ω se nazývá **náhodný jev**,

značíme A, B, ...,

— přičemž prázdná podmnožina se nazývá **jev nemožný**,

označujeme $\emptyset$

— a celý základní prostor **jev jistý**,

označujeme I.

Klasickým příkladem náhodného pokusu je hod hrací kostkou

Příklad 1.1 (Náhodný pokus ... hod hrací kostkou)**.**

- *Elementární jevy:*
  - *„padne 1“* ... E_1,
  - *„padne 2“* ... E_2,
  - *„padne 3“* ... E_3,
  - *„padne 4“* ... E_4,
  - *„padne 5“* ... E_5,
  - *„padne 6“* ... E_6.
- *Jevy* E_1, E_2, ..., E_6 *vymezují základní prostor* Ω.
- *V tomto základním prostoru mohou být například následující jevy:*
  - *náhodný jev* A ... *„padne liché číslo“* ... $A = E_1 + E_3 + E_5$,
  - *náhodný jev* B ... *„padne číslo* ≥ 4*“* ... $A = E_4 + E_5 + E_6$
  - *jev nemožný ... „padne číslo* > 6*“*,
  - *jev jistý ... „padne číslo* < 7*“*,
  - *neslučitelné jevy ... „padne sudé číslo“, „padne liché číslo“*

Operace s jevy

- **Součet jevů** A, B
  - jev, který nastane právě tehdy, když nastane alespoň jeden z jevů A, B. Zavádíme označení $A + B$ nebo množinově $A \cup B$.
- **Součin jevů** A, B
  - jev, který nastane právě tehdy, když nastanou oba jevy současně. Zavádíme označení $A \cdot B$ nebo množinově $A \cap B$.
- **Rozdíl jevů** A, B
  - jev, který nastane právě tehdy, když nastane jev A a nenastane jev B. Zavádíme označení $A \setminus B$.

Speciální jevy

- Jev A' nazýváme **jevem opačným** k jevu A, je-li

$$A' = \Omega \setminus A.$$

- Náhodné jevy se nazývají **neslučitelné** (disjunktní), jestliže platí

$$A \cap B = \emptyset.$$

- Jevy $A_1, A_2, \ldots, A_n$ tvoří **systém neslučitelných jevů**, je-li

$$A_i \cap A_j = \emptyset \qquad \text{pro všechna} \quad i \neq j.$$

- Tento systém se nazývá **úplný**, je-li

$$A_1 + A_2 + \cdots + A_n = I = \Omega.$$

Klasická definice pravděpodobnosti

Definice 1.2. *Nechť je dáno n elementárních jevů $E_1, E_2, \ldots, E_n$, které tvoří úplný systém neslučitelných jevů a jsou stejně možné. Rozkládá-li se jev A na m ($m \leq n$) elementárních jevů z tohoto systému, pak pravděpodobnost jevu A je reálné číslo*

$$P(A) = \frac{m}{n}.$$

Klasická definice pravděpodobnosti se užívá, je-li:

- **konečný** počet elementárních jevů,
- **stejná míra** výskytu elementárních jevů.

Všechny elementární jevy se obvykle označují jako **všechny možné případy**. Všechny elementární jevy, na které se rozkládá jev A, se nazývají **všechny příznivé případy**. Pak daný vztah přejde na známý tvar:

$$P(A) = \frac{\text{počet příznivých výsledků jevu}}{\text{počet všech možných výsledků}}.$$

- $0 \leq P(A) \leq 1$;
- $P(\emptyset) = 0$ — past nemožného jevu;
- $P(I) = 1$ — past jistého jevu;

- $P(A') = 1 - P(A)$ — vztah pro past opačného jevu.

Úloha 1.3. *Vypočtěte pravděpodobnost uhádnutí všech šesti čísel při tažení šesti čísel ze čtyřiceti devíti.*

Řešení.

$$P(A) = \frac{1}{C_6(49)} = \frac{1}{\frac{49!}{6!43!}} = \frac{1}{13\,983\,816} \doteq 7{,}1 \cdot 10^{-8} = 0{,}000\,000\,071.$$

□

2 Statistika

Statistika zkoumá určitý **statistický soubor** (např. skupinu osob, věcí, událostí, časových období, ...) na základě určitého **znak**u, či několika znaků.

Znaky dělíme na

- *kvantitativní* — hodnoty takového znaku se liší číselnou velikostí: např. hmotnost nebo výška osoby;
- *kvalitativní* — hodnoty znaku se liší „kvalitou“: např. obor studia, barva očí, ...

Statistický soubor se skládá ze *statistických jednotek*, jejich počet označujeme n.

Každý znak nabývá jen určitý počet (opakujících se) hodnot. Označme si:

- $x_i, i = 1, \ldots, n$ hodnoty znaku x (x_5 značí hodnou znaku x u statistické jednotky číslo 5);
- $x_j^*, j = 1, \ldots, k$ jedinečné hodnoty znaku x;
- $n_j, j = 1, \ldots, k$ četnost hodnoty x_j^*, udává, u kolika jednotek v souboru byla hodnota x_j^* zjištěna;
- $p_j, j = 1, \ldots, k$ relativní četnost hodnoty x_j^* je četnost dělená počtem jednotek v souboru: $p_j = \dfrac{n_j}{n}$ (relativní četnost lze také vyjádřit v %).

Jistě platí:

$$\sum_{j=1}^{k} n_j = n \qquad \text{a} \qquad \sum_{j=1}^{k} p_j = 1.$$

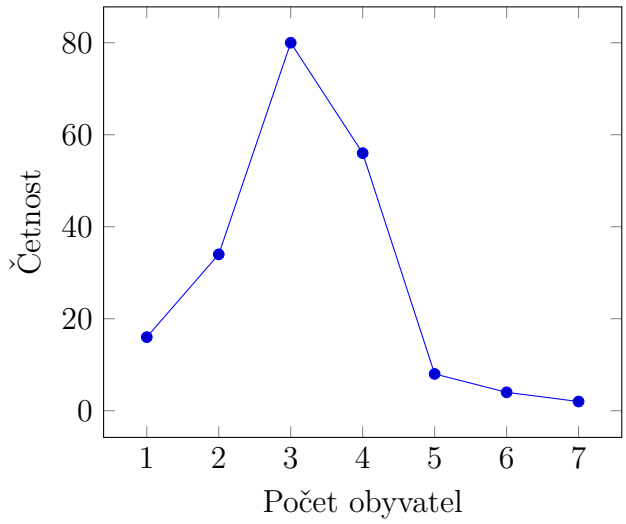
2.1 Grafické znázornění četností

Sloupcové grafy (diagramy)

Úloha 2.1. *Mějme seznam 200 bytů s údaji o lidech, kteří v nich bydlí. Můžeme získat následující rozdělení četností a k nim dopočítat příslušné relativní četnosti.*

počet obyvatel v 1 bytu	1	2	3	4	5	6	7	Σ
četnost	16	34	80	56	8	4	2	200
relativní četnost	$16/200 = 0{,}08$	0,17	0,40	0,28	0,04	0,02	0,01	1,00

Četnosti z tabulky můžeme znázornit i graficky. Na obrázku 1 je spojnicový graf a na obrázcích 2 a 3 jsou grafy sloupcové.



Obrázek 1: Spojnicový diagram.

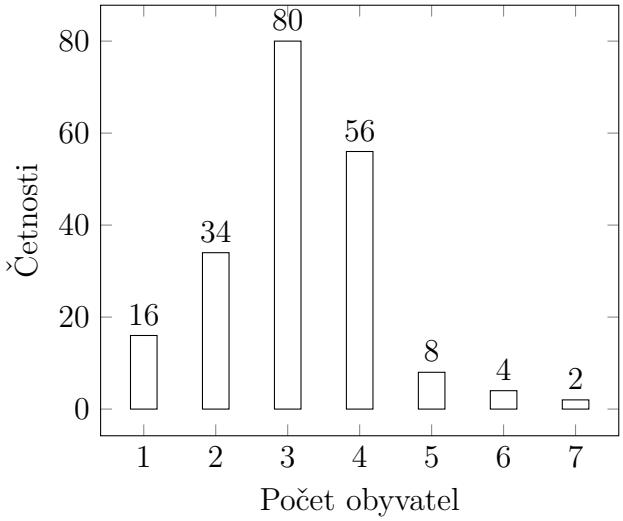
Histogramy

Pokud jsou hodnoty studovaného znaku pro nás příliš podrobné, můžeme je sdružovat do intervalů.

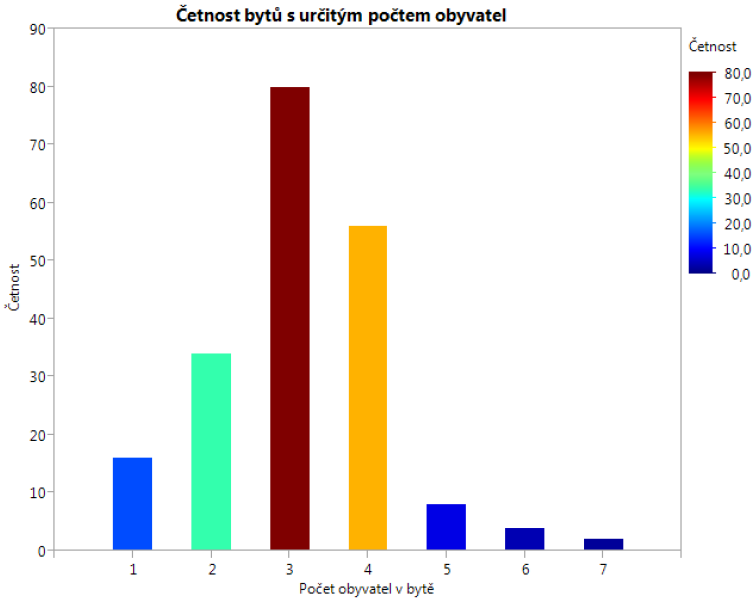
Úloha 2.2. *Podle věku můžeme obyvatele předchozích bytů, kterých je podle tabulky* 626*, neboť*

$$16 \cdot 1 + 34 \cdot 2 + 80 \cdot 3 + 56 \cdot 4 + 5 \cdot 5 + 4 \cdot 6 + 2 \cdot 7 = 626,$$

rozdělit například následovně:



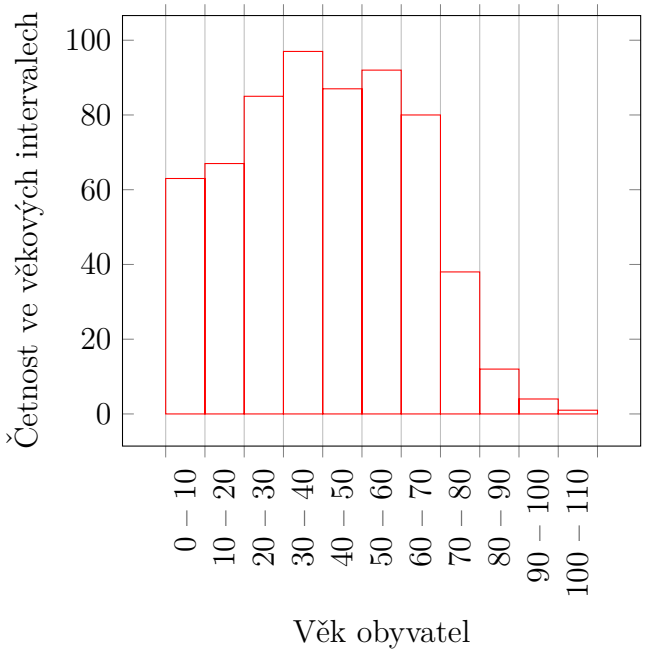
Obrázek 2: Sloupcový diagram jednoduchý s hodnotami.



Obrázek 3: Sloupcový diagram barevný a s legendou.

věk	1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100	101–110	Σ
četnost	63	67	85	97	87	92	80	38	12	4	1	626
relativní četnost	$\frac{63}{626}$	$\frac{67}{626}$	$\frac{85}{626}$	$\frac{97}{626}$	$\frac{87}{626}$	$\frac{92}{626}$	$\frac{80}{626}$	$\frac{38}{626}$	$\frac{12}{626}$	$\frac{4}{626}$	$\frac{1}{626}$	1,00

Toto sdružování se dá graficky znázornit pomocí tzv. histogramu (viz obrázek 4). Jde o sloupcový graf, jehož sloupce svou šířkou vyjadřují interval hodnot studovaného znaku a svou výškou četnost hodnot spadajících do příslušného intervalu.



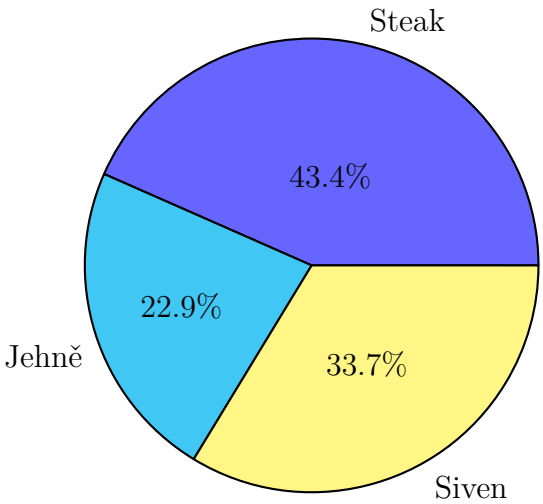
Obrázek 4: Histogram rozložení věku 636 osob (z 200 bytů).

Koláčové grafy

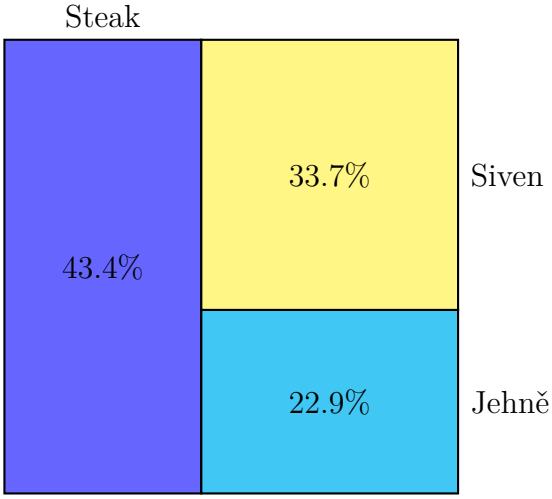
Úloha 2.3. *415 strávníků si oběd vybralo následujícím způsobem:*

Jídlo	***Steak z argentinské roštěné*** *s pečenou kořenovou zeleninou a omáčkou Bordelaise*	***Jehněčí kolínko na majoránce*** *se smetanovým špenátem*	***Filet ze sivena*** *s quinoa salátem*
Četnost	180	95	140
Relativní četnost v %	43,4	22,9	33,7

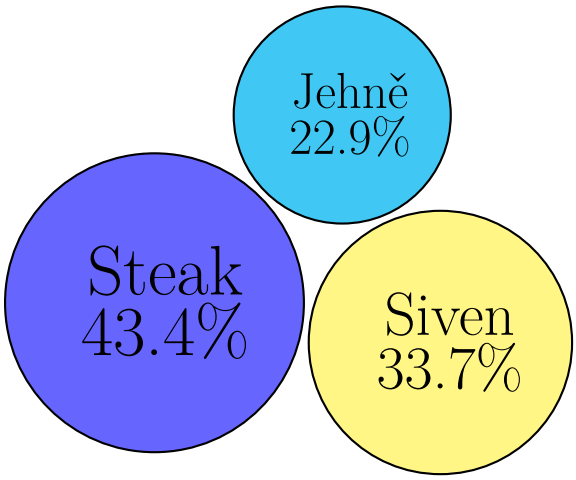
Poměrné rozdělení těchto preferencí se dá graficky znázornit pomocí tzv. koláčových grafů (viz obrázek 5), případně i dalších typů grafů (viz obrázky 6 a 7).



Obrázek 5: Koláčový graf objednávek.



Obrázek 6: Obdélníkový graf objednávek.



Obrázek 7: Obláčkový graf objednávek.

2.2 Charakteristiky kvantitativního znaku

Modus znaku

je hodnota znaku s největší četností (nemusí být dána jednoznačně).

V příkladu 2.1 je $\mathrm{Mod}(x) = 3$.

Medián znaku

je „prostřední“ hodnota znaku. Jsou-li hodnoty znaku x uspořádány podle velikosti

$$x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n,$$

potom

$$\mathrm{Med}(x) = \begin{cases} x_{\frac{n+1}{2}}, & \text{je-li } n \text{ liché,} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), & \text{pro } n \text{ sudé.} \end{cases}$$

V příkladu 2.1 je $n = 200$, tedy sudé. Použijeme tedy druhý vzorec:

$$\mathrm{Med}(x) = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) = \frac{1}{2}\left(x_{\frac{200}{2}} + x_{\frac{200}{2}+1}\right) = \frac{1}{2}\left(x_{100} + x_{101}\right) = \frac{1}{2}(3+3) = 3,$$

neboť

$$x_1, x_2, \ldots, x_{200} = \underbrace{1, 1, \ldots, 1}_{16\times}, \underbrace{2, 2, \ldots 2}_{34\times}, \underbrace{3, 3, \ldots, 3}_{80\times}, 4, \ldots, 7.$$

Platí tedy, že na 100. a 101. pozici jsou trojky:

$$x_1 = \cdots = x_{16} = 1, x_{17} = \cdots = x_{50} = 2,\ x_{51} = \cdots = x_{100} = x_{101} = \cdots = x_{130} = 3, \ldots.$$

Aritmetický průměr znaku

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{j=1}^{r} x_j^* \cdot n_j.$$

V příkladu 2.1 je

$$\bar{x} = \frac{1}{200}(1 \cdot 16 + 2 \cdot 34 + 3 \cdot 80 + 4 \cdot 56 + 5 \cdot 8 + 6 \cdot 4 + 7 \cdot 2) = \frac{626}{200} = 3{,}13.$$

Rozptyl znaku

je definován jako aritmetický průměr druhých mocnin odchylek hodnot znaku od jeho aritmetického průměru:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{r}(x_j^* - \bar{x})^2 \cdot n_j.$$

V příkladu 2.1 je

$$\begin{aligned} s_x^2 &= \frac{1}{200}\Big[(1-3{,}13)^2 \cdot 16 + (2-3{,}13)^2 \cdot 34 + (3-3{,}13)^2 \cdot 80 + (4-3{,}13)^2 \cdot 56 \\ &\quad +(5-3{,}13)^2 \cdot 8 + (6-3{,}13)^2 \cdot 4 + (7-3{,}13)^2 \cdot 2\Big] \doteq 1{,}25. \end{aligned}$$

Směrodatná odchylka znaku

je druhou odmocninou z rozptylu:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum_{j=1}^{r}(x_j^* - \bar{x})^2 n_j}.$$

V příkladu 2.1 je

$$s_x \doteq \sqrt{1{,}25} \doteq 1{,}12.$$

2.3 Korelační koeficient

Nyní již budeme zkoumat dvojici statistických znaků x a y (dosud jsme studovali vždy jen jeden).

Zajímá nás míra jejich (statistické) závislosti. Jako charakteristiku míry této závislosti znaků x a y budeme používat jejich tzv. korelační koeficient:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x \cdot s_y},$$

kde x_i, y_i, $i = 1, \ldots, n$, jsou hodnoty znaků x a y, $\overline{x}$ a $\overline{y}$ jsou jejich aritmetické průměry a $s_x \neq 0$ a $s_y \neq 0$ jsou jejich směrodatné odchylky.

Pro hodnoty korelačního koeficientu z definice platí, že

$$-1 \leq r_{xy} \leq 1.$$

Korelační koeficient popisuje pouze tzv. lineární závislost. Takto můžeme popsat tři základní případy:

- s rostoucími hodnotami znaku x rostou i hodnoty znaku y (r_{xy} je „v blízkosti“ 1),
- s rostoucími hodnotami znaku x klesají hodnoty znaku y (r_{xy} je „v blízkosti“ -1),
- s rostoucími hodnotami znaku x si hodnoty znaku y dělají „co chtějí“ (r_{xy} je „v blízkosti“ 0).

Úloha 2.4. *Ukažte, že v případě analytické závislosti znaků x a y vyjádřené vztahem*

$$y_i = a \cdot x_i, \qquad i = 1, \ldots, n, \quad a \neq 0,$$

vyjde

$$r_{xy} = \frac{a}{|a|} = \operatorname{sgn} a = \begin{cases} 1, & \textit{pro } a > 0, \\ -1, & \textit{pro } a < 0. \end{cases}$$

Úloha 2.5. *Patnáct chlapců uvedlo výšku svoji a svého otce. Údaje jsou zaznamenány v tabulce 1.*

Data z tabulky 1 jsou znázorněna na obrázcích 8 a 9.

Pro zadaná data vypočteme:

$$\overline{x} \doteq 174{,}667, \qquad \overline{y} \doteq 174{,}533, \qquad s_x \doteq 6{,}831, \qquad s_y \doteq 6{,}323.$$

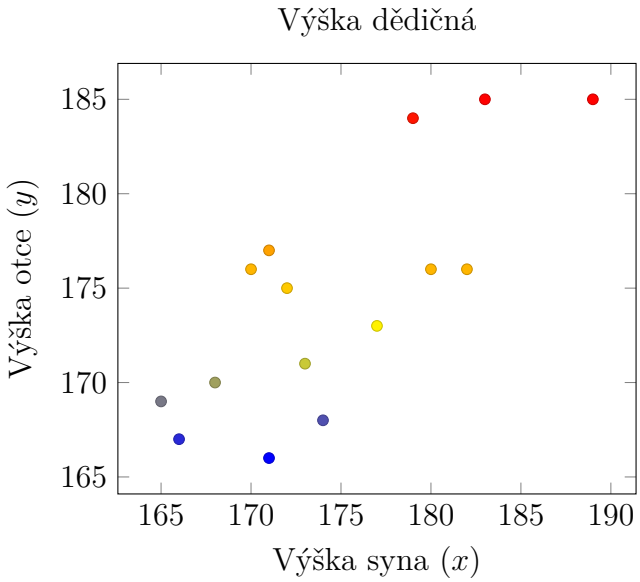
x	172	168	183	182	174	166	173	170	180	171	165	171	179	189	177
y	175	170	185	176	168	167	171	176	176	166	169	177	184	185	173

Tabulka 1: Výška dětí (x) a jejich otců (y).

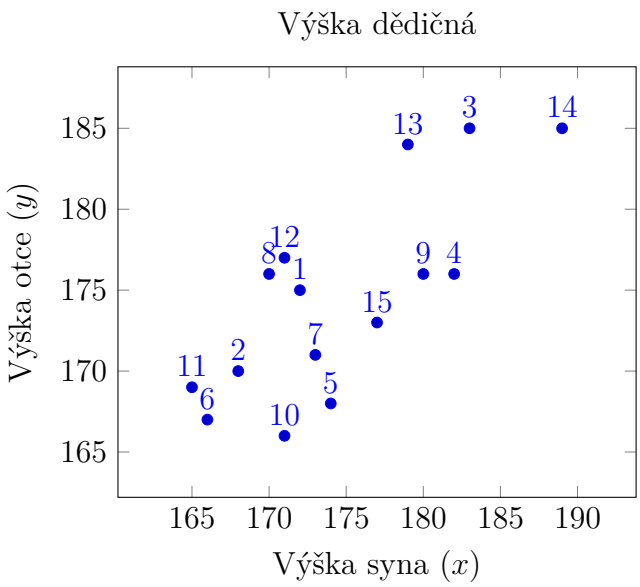
Můžeme tedy přikročit k výpočtu korelačního koeficientu

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x \cdot s_y} \doteq \frac{\frac{1}{15}\sum_{i=1}^{15}(x_i - 174{,}667)(y_i - 174{,}533)}{6{,}831 \cdot 6{,}323} \doteq 0{,}772.$$

Můžeme tedy nahlédnout, že je tu přítomna ne úplně přesvědčivá pozitivní lineární závislost „čím vyšší syn, tím vyšší otec“, resp. „čím vyšší otec, tím vyšší syn“. Ostatně něco takového můžeme odezřít i z obrázků 8 a 9.



Obrázek 8: Základní grafické znázornění dat z tabulky 1. Každé jedno kolečko odpovídá jedné dvojici syn–otec tak, že má souřadnice výšku syna (x) a výšku otce (y).



Obrázek 9: Základní grafické znázornění dat z tabulky 1. Oproti obrázku 8 je u každého kolečka ještě číslo dvojice.