ZPRACOVÁNÍ DAT ZÁKLADNÍMI STATISTICKÝMI METODAMI

Jiří Fišer

18. března 2024

Obsah

3.1 Úvod do statistiky

Pravděpodobnost vs statistika

- **Pravděpodobnost: matematický model reality**
  - idealizovaný, abstraktní model
  - pracuje s jednou nebo více náhodnými veličinami, jejichž rozdělení je známo
  - z podstaty věci nepozorovatelný, jde jen o naši představu

- **Statistika: pozorování (měření) hodnot nějaké veličiny**
  - zkoumá jevy rozsáhlém souboru případů a činí o nich závěry pomocí statistické indukce
  - zobecňuje výsledky na rozsáhlejší soubor než je ten, ze kterého byly skutečně odvozeny
  - žádný konečný náhodný výběr nemůže poskytnout úplnou informaci o rozdělení pravděpodobností náhodné veličiny - máme jen **odhady**
  - příklady:
    - ★ Mají děti/mladiství, jejichž otec nebo matka kouří, horší fungování plic než jejich vrstevníci, jejichž rodiče kteří nekouří?
    - ★ Mají kuřáci větší riziko na onemocnění rakovinou než nekuřáci?
    - ★ Jak dlouho obvykle bezporuchově funguje počítač daného typu?

Data

- pozorování, která činíme kvůli zodpovězení položené otázky
- matematicky: data = realizace náhodné veličiny
- datové tabulky
  - řádky: pozorování týkající se nezávislých subjektů náhodného výběru (osob, experimentů,. . .)
  - sloupce: jednotlivé měřené veličiny
- software: např. databázové systémy, Excel,
- statistický software: SAS, Statistica, SPSS, R, Python, . . .

3.2 Popisná statistika

- pojmový aparát statistiky
- základní nástroj analýzy dat
- prostředky pro prezentaci dat a výsledků

Základní pojmy

- **Statistická jednotka:** objekt, který chceme zkoumat
  např. osoby, domácnosti, firmy, organismy, obce, kraje

- **Statistický soubor:**
  - **základní:** množina všech statistických jednotek, jejichž vlastnosti chceme poznat
  - **výběrový:** množina skutečně vyšetřovaných statistických jednotek (tzv. náhodný výběr)

- **Statistický znak:** vlastnost zjišťovaná na každé statistické jednotce (tj. náhodná veličina)
  např. pohlaví, výška, hmotnost, počet dětí, barva očí, dopravní prostředek, počet úrazů, jméno, věk

- **Rozsah souboru:** počet zkoumaných statistických jednotek

Typy statistických znaků

- **kvalitativní**: slovní, kategoriální
  např. pohlaví, barva očí, dopravní prostředek
- **kvantitativní**: číselné, numerické
  - spojité: např. výška, hmotnost, věk
  - diskrétní: počet dětí, počet úrazů
- **alternativní**: 2 hodnoty
- **množné**: 3 a více hodnot

Jednorozměrný statistický soubor

Označení:

- $\{\varepsilon_1, \ldots, \varepsilon_n\}$ výběrový soubor
- X statistický znak
- x_i hodnota znaku X na objektu ε_i, $i = 1, \ldots, n$
- $(x_1, \ldots, x_n)$ datový soubor
- $(x_{(1)}, \ldots, x_{(n)})$ uspořádaný datový soubor, tj.

$$x_{(1)} \leq \ldots \leq x_{(n)}$$

- $(x_{[1]}, \ldots, x_{[r]})$ vektor variant znaku X, tj. $x_{[i]} \neq x_{[j]}$

Rozložení četností

- slouží ke zpřehlednění datového souboru

- **Bodové:**
  - diskrétní znak s malým počtem variant
  - četnost přiřazujeme jednotlivým variantám

- **Intervalové:**
  - diskrétní znak s velkým počtem variant
  - spojitý znak
  - četnost přiřazujeme třídícím intervalům

Bodové rozložení četností

- **(Absolutní) četnost** varianty $x_{[j]}$: n_j (počet výskytů hodnoty $x_{[j]}$)
- **Relativní četnost** varianty $x_{[j]}$: $p_j = \frac{n_j}{n}$ (empirická pravděpodobnostní funkce)
- **(Absolutní) kumulativní četnost** prvních j variant:

$$N_j = n_1 + \cdots + n_j$$

- **Relativní kumulativní četnost** prvních j variant:

$$F_j = \frac{N_j}{n} = p_1 + \cdots + p_j$$

- **Empirická distribuční funkce:**

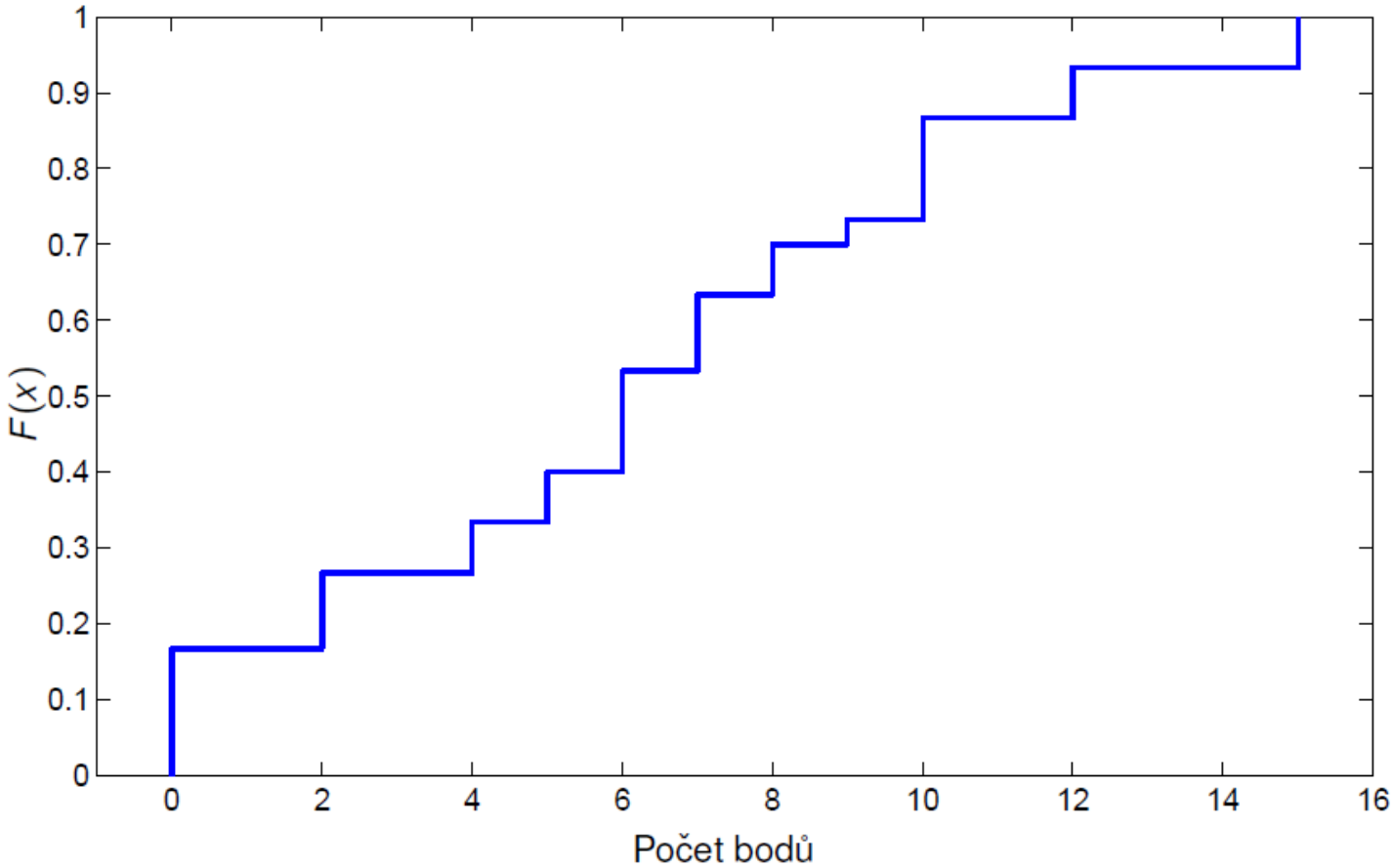
$$F(x) = \begin{cases} 0 & x < x_{[1]} \\ F_j & x_{[j]} \leq x < x_{[j+1]},\ j = 1, \ldots, r-1 \\ 1 & x \geq x_{[r]} \end{cases}$$

Příklad (Bodové rozložení četností)

Při zápočtu ze statistiky se studenti podrobili testu, ve kterém mohli získat 0 až 15 bodů. Výsledky jsou následující: 5,10,6,7,0,2,2,4,8,10,12,15,0,0,4,2,7,10,15,0,6,5,6,9,8,7,10,12,6,0.

Body	n_j	p_j (%)	F_j (%)	
0	5	16,7	16,7	
2	3	10,0	26,7	
4	2	6,7	33,4	
5	2	6,7	40,1	
6	4	13,3	53,4	
7	3	10,0	63,4	
8	2	6,7	70,1	
9	1	3,2	73,3	korekce hodnoty 3,3
10	4	13,3	86,6	
12	2	6,7	93,3	
15	2	6,7	100,0	
Celkem	30	100,0	-	

Empirická distribuční funkce



Intervalové rozložení četností

Třídící intervaly: obor hodnot znaku X rozdělíme na disjunktní intervaly

$$(-\infty, u_1\rangle,\ (u_1, u_2\rangle, \ldots, (u_{k-1}, \infty)$$

Stanovení třídících intervalů: subjektivní

- Počet třídích intervalů: různá pravidla
  - k blízké $\sqrt{n}$
  - **Sturgesovo pravidlo:** $k = 1 + 3.3 \log n$
- Zpravidla volíme intervaly stejné délky
- U nesymetrických rozdělení volíme krajní intervaly širší, aby zahrnovaly extrémní hodnoty
- Názvy četností podobné jako u bodového rozložení četností
- všechny body z j-tého intervalu $(u_{j-1}, u_j\rangle$ ztotožníme se středem

$$a_j = \frac{u_j + u_{j-1}}{2}$$

  - $(-\infty, u_1\rangle: \quad a_1 = u_1 - \frac{u_2 - u_1}{2}$
  - $(u_{k-1}, \infty): \quad a_k = u_{k-1} + \frac{u_{k-1} - u_{k-2}}{2}$

Příklad (Intervalové rozložení četností)

U 70 žen byl změřen hemoglobin s přesností 0.1 g/100 ml:
10,2, 13,7, 10,4, 14,9, 11,5, 12,0, 11,0, 13,3, 12,9, 12,1, 9,4, 13,2, 10,8, 11,7, 10,6, 10,5, 13,7, 11,8, 14,1, 10,3, 13,6, 12,1, 12,9, 11,4, 12,7, 10,6, 11,4, 11,9, 9,3, 13,5, 14,6, 11,2, 11,7, 10,9, 10,4, 12,0, 12,9, 11,1, 8,8, 10,2, 11,6, 12,5, 13,4, 12,1, 10,9, 11,3, 14,7, 10,8, 13,3, 11,9, 11,4, 12,5, 13,0, 11,6, 13,1, 9,7, 11,2, 15,0, 10,7, 12,9, 13,4, 12,3, 11,0, 14,6, 11,1, 13,5, 10,9, 13,1, 11,8, 12,2

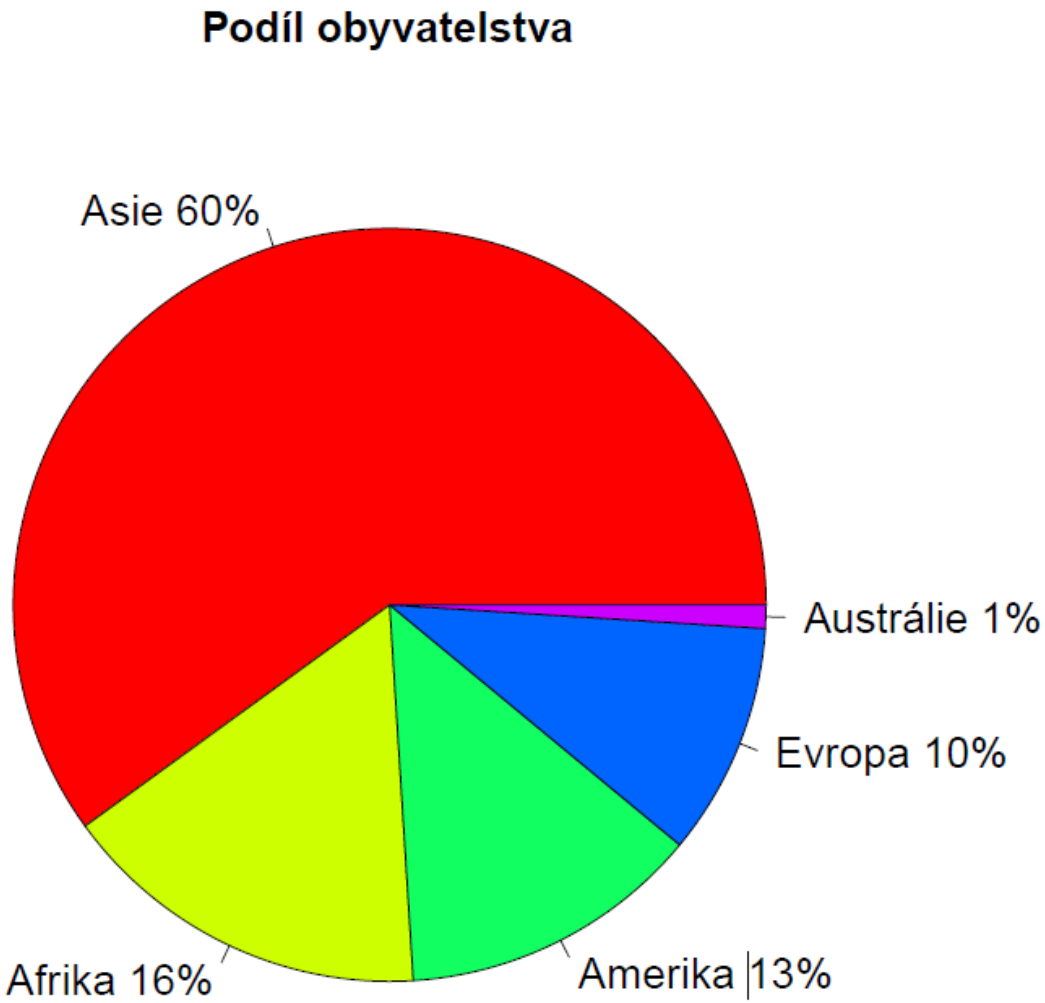
Hladina hemoglobinu v g/100 ml	n_j	p_j (%)	F_j (%)
8,0-8,9	1	1,4	1,4
9,0-9,9	3	4,3	5,7
10,0-10,9	14	20,0	25,7
11,0-11,9	19	27,1	52,9
12,0-12,9	14	20,0	72,9
13,0-13,9	13	18,6	91,4
14,0-14,9	5	7,1	98,6
15,0-15,9	1	1,4	100,0
Celkem	70	100,0	-

Grafické znázornění setříděných dat

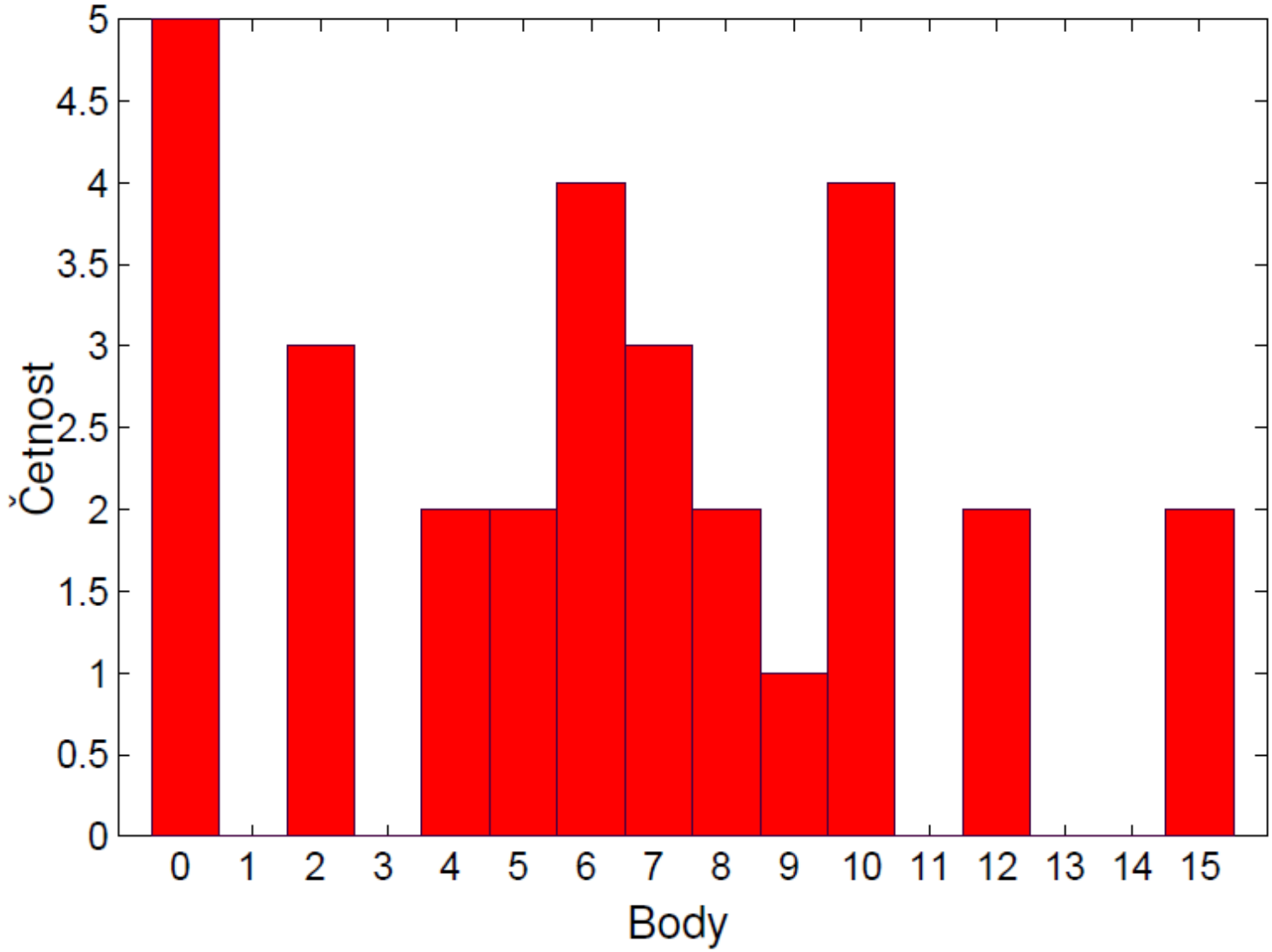
Znázorňujeme relativní a absolutní četnosti nebo relativní a absolutní kumulativní četnosti.

- **Koláčový graf:** pouze pro zobrazení relativních četností
- **Histogram:** sloupcový graf
  - bodové rozložení četností: bodu $x_{[j]}$ přiřadíme obdélník, jehož výška je úměrná zjištěné četnosti
  - intervalové rozložení četností:
    - ★ šířka sloupku rovna délce intervalu
    - ★ bodu a_j přiřadíme obdélník, jehož plocha odpovídá relativní četnosti
    - ★ někdy jsou konstruovány i pro další typy četností, tj. výška je úměrná zjištěné četnosti

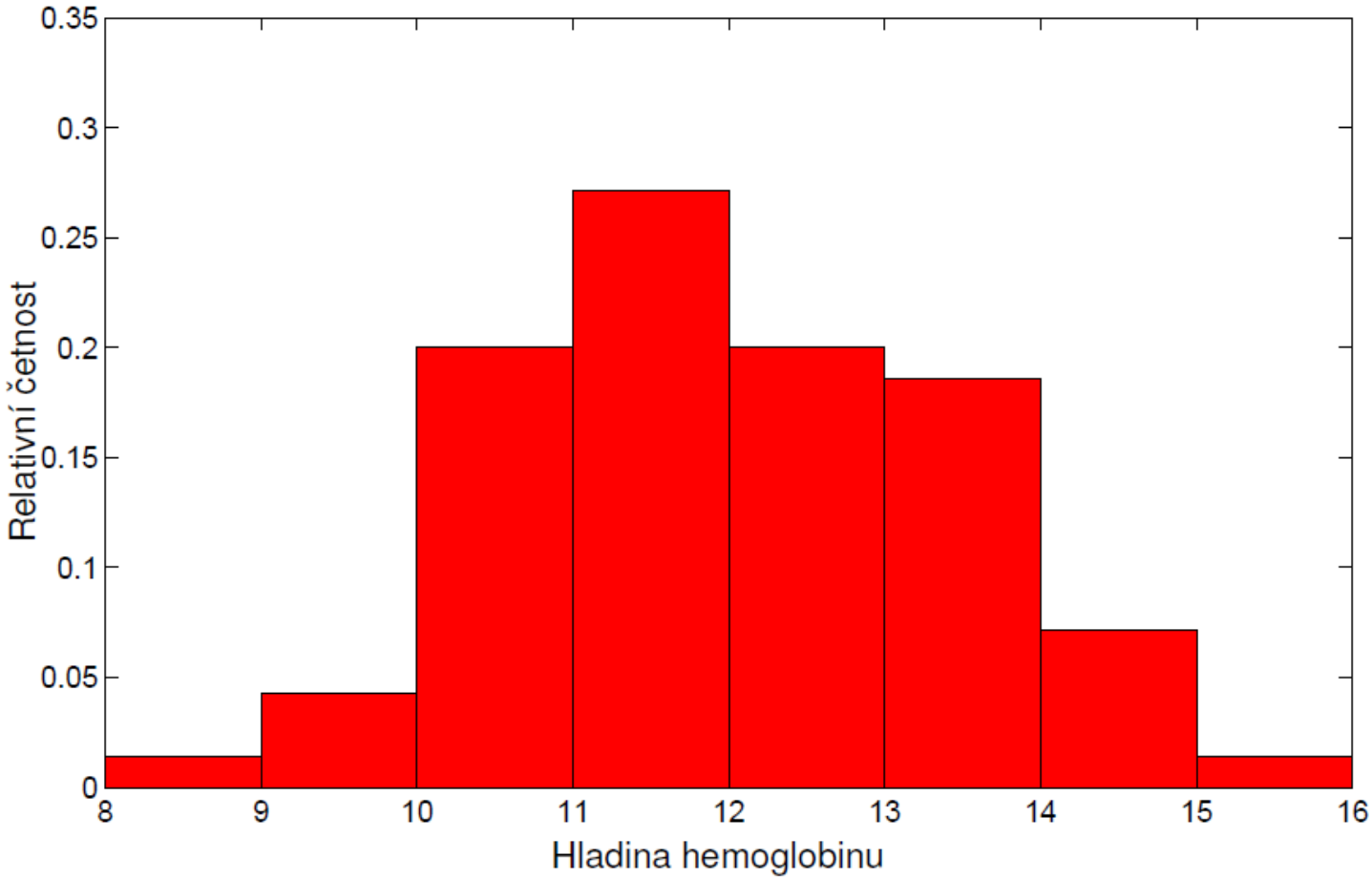
Koláčový graf rozložení obyvatelstva na kontinentech



Histogram absolutních četností počtu získaných bodů v testu ze statistiky



Histogram


0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Relativní četnost
8
9
10
11
12
13
14
15
16
Hladina hemoglobinu

Míry polohy

Aritmetický průměr

- pozorování $x_1, \ldots, x_n$
- nesetříděný soubor:

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- setříděný soubor:

$$\overline{x} = \frac{x_{[1]} n_1 + \cdots + x_{[r]} n_r}{n_1 + \cdots + n_r} = \frac{1}{n} \sum_{i=1}^{k} x_{[i]} n_i$$

- **vážený aritmetický průměr:** soubor rozdělen do s dílčích souborů se známými průměry $\bar{x}_i$ a rozsahy n_i, $i = 1, \ldots, s$

$$\overline{x} = \frac{\sum_{i=1}^{s} \overline{x}_i n_i}{n_1 + \cdots + n_s}$$

Vlastnosti aritmetického průměru

- zvětšíme-li všechna pozorování o konstantu c, zvětší se průměr též o c
- násobíme-li všechna pozorování nějakou konstantou b, pak nový průměr bude roven průměru původních dat krát konstanta b
- **Citlivý** na odlehlá pozorování (outliery)
  - 5,4,5,6,7,4,5,5,4,5: $\overline{x} = 5$
  - 5,4,5,6,7,4,500,5,4,5: $\overline{x} = 54{,}5$, což je hodnota nic neříkající
- odhadem **střední hodnoty** $\mathrm{E}(X)$ náhodné veličiny X
  - př. X značí počet ok při hodu férovou kostkou

$$\mathrm{P}(X = 1) = \mathrm{P}(X = 2) = \cdots = \mathrm{P}(X = 6) = \frac{1}{6}$$

$$\mathrm{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{7}{2} = 3{,}5$$

Třikrát jsme hodili kostkou s těmito výsledky: 3,3,5

Výběrový průměr je $\overline{x} = \frac{3+3+5}{3} = \frac{11}{3} = 3{,}\overline{6}$

Míry polohy - kvantily

Pro $\alpha \in (0,1)$ je **výběrový α-kvantil** definován jako číslo $\widetilde{x}_\alpha$, které rozděluje datový soubor na dvě části tak, že

1. alespoň 100α % všech dat je menších nebo rovných $\widetilde{x}_\alpha$
2. alespoň $100(1-\alpha)$ % všech dat je větších nebo rovných $\widetilde{x}_\alpha$

Data uspořádáme vzestupně $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

Medián

$$\widetilde{x}_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{pro } n \text{ liché,} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{pro } n \text{ sudé.} \end{cases}$$

Míry polohy - kvantily

Data uspořádáme vzestupně $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

α kvantil

- je-li $n\alpha = c$, kde c je přirozené číslo, potom

$$\widetilde{x}_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2}$$

  - $100 \cdot 0{,}21 = 21$, tedy $\widetilde{x}_{0,21} = \frac{x_{(21)} + x_{(22)}}{2}$

- jestliže $n\alpha$ není přirozené číslo, zaokrouhlíme $n\alpha$ nahoru na nejbližší přirozené číslo c a položíme

$$\widetilde{x}_\alpha = x_{(c)}$$

  - $20 \cdot 0{,}21 = 4{,}2$, tedy $\widetilde{x}_{0,21} = x_{(5)}$

Kvantily - příklady

Příklad (n sudé)

Ve výrobě se v posledním půl roce v jednotlivých měsících vyskytl následující počet úrazů: 1, 3, 2, 4, 2, 4. Určete medián, dolní a horní kvartil počtu úrazů za měsíc.

Počty uspořádáme vzestupně

$$1,\ 2,\ 2,\ 3,\ 4,\ 4$$

- medián: $\widetilde{x}_{0,5} = \frac{x_{(3)}+x_{(4)}}{2} = \frac{2+3}{2} = 2{,}5$
- dolní kvartil: $n\alpha = 6 \cdot 0{,}25 = 1{,}5 \ \Rightarrow \ \widetilde{x}_{0,25} = x_{(2)} = 2$
- horní kvartil: $n\alpha = 6 \cdot 0{,}75 = 4{,}5 \ \Rightarrow \ \widetilde{x}_{0,75} = x_{(5)} = 4$

Kvantily - příklady

Příklad (n liché)

*Ve výrobě se v posledním půl roce v jednotlivých měsících vyskytl následující počet úrazů: 1, 3, 2, 4, 2, 4, **1**. Určete medián, dolní a horní kvartil počtu úrazů za měsíc.*

$$1,\ 1,\ 2,\ 2,\ 3,\ 4,\ 4$$

- medián: $(n+1)/2 = 8/2 = 4 \ \Rightarrow \ \widetilde{x}_{0,5} = x_{(4)} = 2$
- dolní kvartil: $n\alpha = 7 \cdot 0{,}25 = 1{,}75 \ \Rightarrow \ \widetilde{x}_{0,25} = x_{(2)} = 1$
- horní kvartil: $n\alpha = 7 \cdot 0{,}75 = 5{,}25 \ \Rightarrow \ \widetilde{x}_{0,75} = x_{(6)} = 4$

Příklad 2: vypočtěte 0,1 kvantil, dolní a horní kvartil

Uvažujme data x daná tabulkou:

hodnota x_i	1	2	3	4	5
počet výskytů n_i	10	12	6	3	0

- Tedy
  - $x_{(1)} = \cdots = x_{(10)} = 1$,
  - $x_{(11)} = \cdots = x_{(22)} = 2$,
  - $x_{(23)} = \cdots = x_{(28)} = 3$,
  - $x_{(29)} = \cdots = x_{(31)} = 4$.
- $n\alpha = 31 \cdot 0.1 = 3.1 \;\Rightarrow\; \widetilde{x}_{0.1} = x_{(4)} = 1$
- $n\alpha = 31 \cdot 0.25 = 7.75 \;\Rightarrow\; \widetilde{x}_{0.25} = x_{(8)} = 1$
- $n\alpha = 31 \cdot 0.75 = 23.25 \;\Rightarrow\; \widetilde{x}_{0.75} = x_{(24)} = 3$

Vlastnosti výběrového mediánu

- **Není citlivý** na odlehlá pozorování (outliery)
  - 5,4,5,6,7,4,5,5,4,5: $4,4,4,5,5,5,5,5,6,7 \Rightarrow \widetilde{x}_{0,5} = 5$
  - 5,4,5,6,7,4,500,5,4,5: $4,4,4,5,5,5,5,6,7,500 \Rightarrow \widetilde{x}_{0,5} = 5$
- Odhadem **mediánu**
- splňuje požadované vlastnosti (posun o konstantu, změna měřítka)
  - $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$
  - $(x + c)_{(1)} \leq (x + c)_{(2)} \leq \cdots \leq (x + c)_{(n)}$
  - $(x \cdot b)_{(1)} \leq (x \cdot b)_{(2)} \leq \cdots \leq (x \cdot b)_{(n)}, \quad b > 0$

Vlastnosti výběrového mediánu

4. Vztah výběrového průměru a výběrového mediánu

hodnota x_i	1	2	3	4	5
počet výskytů n_i	10	12	10	0	0

$$\overline{x} = \widetilde{x}_{0,5} = 2$$

hodnota x_i	1	2	3	4	5
počet výskytů n_i	8	10	8	4	2

$$\overline{x} = 2.4 > \widetilde{x}_{0,5} = 2$$

hodnota x_i	1	2	3	4	5
počet výskytů n_i	2	4	8	10	8

$$\overline{x} = 3.2 < \widetilde{x}_{0,5} = 4$$

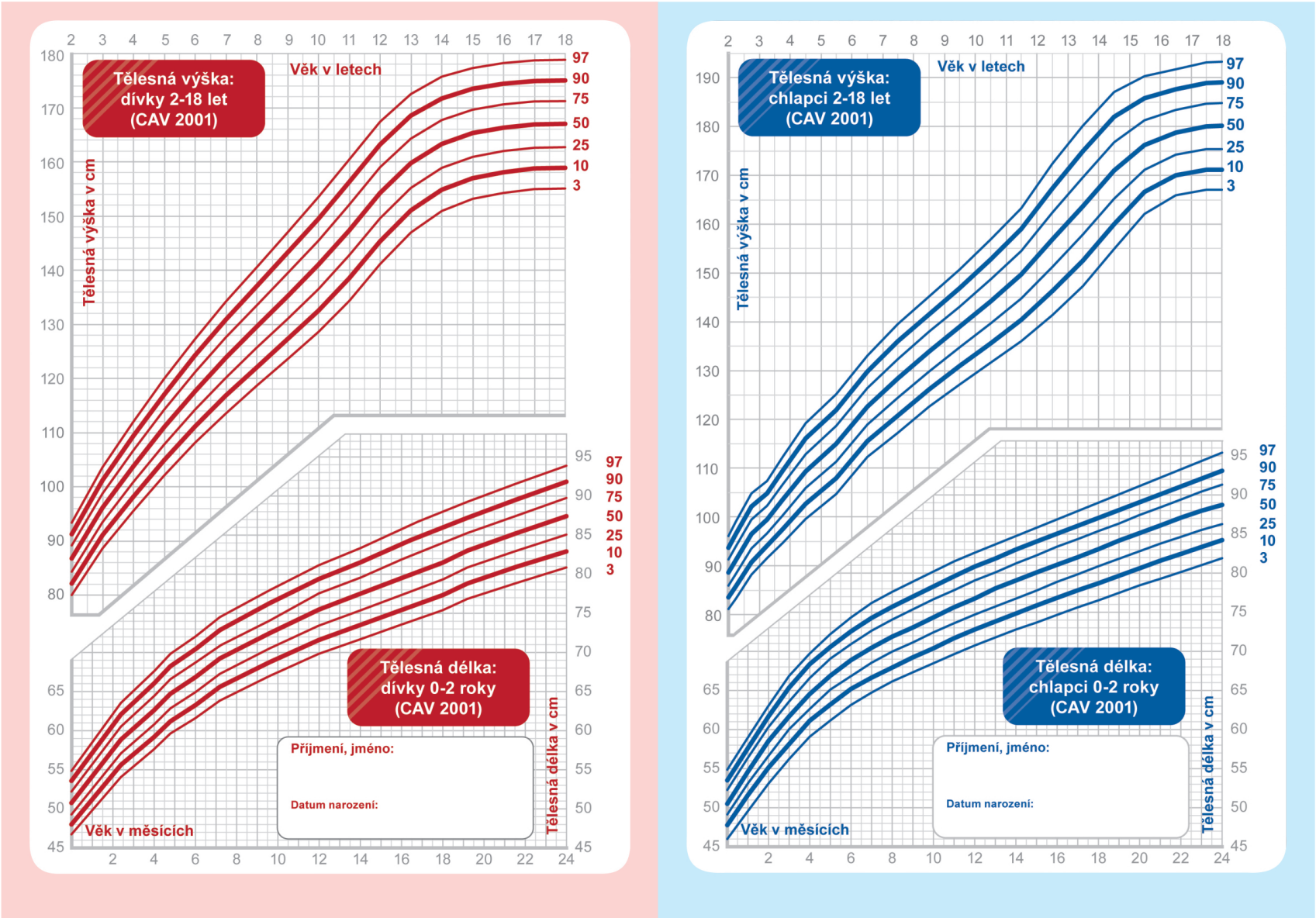
Hrubá měsíční mzda zaměstnanců podle vzdělání ČR, rok 2021

	průměr	medián
CELKEM	40 777	35 169
základní a nedokončené	28 672	27 023
střední bez maturity	31 111	29 567
střední s maturitou	39 609	36 051
vyšší odborné a bakalářské	47 271	41 662
vysokoškolské	61 334	50 472

Využití výběrových kvantilů

- Jakou hladinu cholesterolu v krvi nepřekročí 90 % zdravé populace ČR? (pro krevní obraz jsou stanoveny referenční hladiny pro jednotlivé ukazatele)
- Jakou délku nepřekročí 95 % lišek? rozmezí 58-90 cm (5% a 95% kvantil)
- Jak definovat pojem stoletá voda, který odpovídá průtoku, jenž je maximálním ročním průtokem překročen jenom v 1 % případů?
- Jakou výši kapitálu musí pojišťovny EU držet, aby snížily riziko platební neschopnosti v průběhu roku? (99,5% kvantil, směrnice Solvency II)
- Percentilové grafy

Percentilové růstové grafy dětí



Míry polohy

Modus

- varianta znaku, která má největší četnost

Míry variability

Míry absolutní variability

- **Variační obor** $\langle x_{(1)}, x_{(n)} \rangle$
- **(Variační) rozpětí:** $R = x_{(n)} - x_{(1)}$
- **Kvartilové rozpětí:** $R_Q = \widetilde{x}_{0.75} - \widetilde{x}_{0.25}$
- **Kvartilová odchylka:** $\frac{R_Q}{2}$
- **Rozptyl:**

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

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{k} (a_i - \bar{x})^2 n_i$$

- **Směrodatná odchylka:** $s_x = \sqrt{s_x^2}$

Poznámka: Někdy se ve vztahu pro rozptyl používá koeficient $1/n$. s_x^2 má lepší vlastnosti pro malá n

Následující tabulka četností udává životnost (v hodinách) určité komponenty. Stanovte průměrnou životnost, směrodatnou odchylku a modus životnosti této komponenty.

živ.	$300 \leq t < 400$	$400 \leq t < 500$	$500 \leq t < 600$	$600 \leq t < 700$	$700 \leq t < 800$
čet.	13	25	66	58	38

- a_i ... střed i-tého intervalu životnosti

živ.	$300 \leq t < 400$	$400 \leq t < 500$	$500 \leq t < 600$	$600 \leq t < 700$	$700 \leq t < 800$
střed	350	450	550	650	750
čet.	13	25	66	58	38

- Modus: 550 h

živ.	$300 \leq t < 400$	$400 \leq t < 500$	$500 \leq t < 600$	$600 \leq t < 700$	$700 \leq t < 800$
střed	350	450	550	650	750
čet.	13	25	66	58	38

- Průměrná životnost komponenty = ??

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{5} a_i \cdot n_i$$

$$= \frac{1}{200}(350 \cdot 13 + 450 \cdot 25 + 550 \cdot 66 + 650 \cdot 58 + 750 \cdot 38) = 591{,}5 \text{ h}$$

živ.	$300 \leq t < 400$	$400 \leq t < 500$	$500 \leq t < 600$	$600 \leq t < 700$	$700 \leq t < 800$
střed	350	450	550	650	750
čet.	13	25	66	58	38

- Směrodatná odchylka životnosti komponenty = ??

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{5}(a_i - \overline{x})^2 \cdot n_i}$$

$$= \sqrt{\frac{(350 - 591,5)^2 \cdot 13 + \cdots + (750 - 591,5)^2 \cdot 38}{199}}$$

$$= 112,66 \text{ h}$$

Míry variability

Míry absolutní variability

Poznámka

Nelze srovnávat variabilitu dvou a více znaků, jestliže se výrazně liší úrovní znaku nebo jsou vyjádřeny v různých jednotkách!
$\rightarrow$ *nutno použít relativní míry variability*

Míry variability

Míry relativní variability

- **Variační koeficient:**

$$V_x = \frac{s_x}{\bar{x}}$$

- **Relativní kvartilová odchylka:**

$$Q_r = \frac{\widetilde{x}_{0.75} - \widetilde{x}_{0.25}}{\widetilde{x}_{0.75} + \widetilde{x}_{0.25}}$$

Příklad

Zjišťováním hmotnosti mužů a žen ve věku 50 let, byly zjištěny následující údaje:

- *průměrná hmotnost mužů: 95 kg*
- *směrodatná odchylka u mužů: 4 kg*

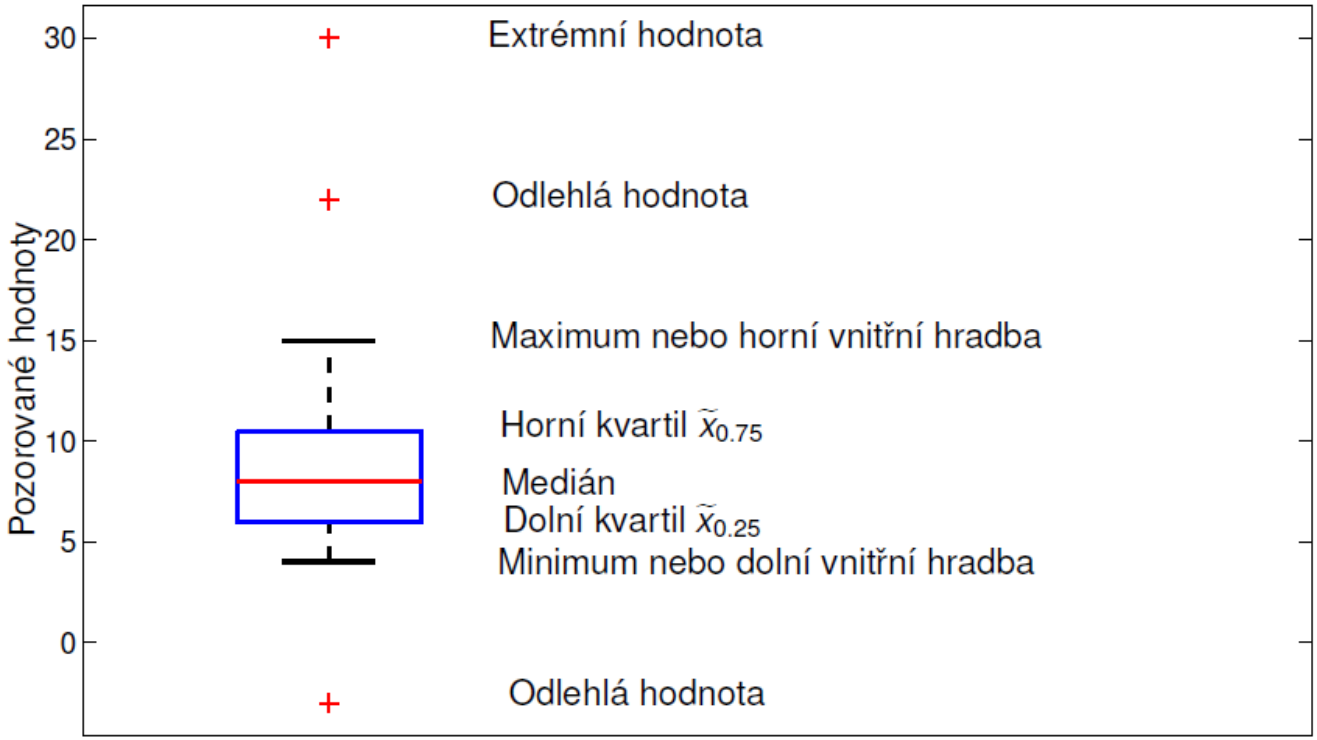
- *průměrná hmotnost žen: 65 kg*
- *směrodatná odchylka u žen: 3,32 kg*

Je správná interpretace: muži jsou v průměru těžší a mají větší výkyvy hmotnosti?

- variační koeficient u mužů: 4/95 = 0,0421 (4,21 %)
- variační koeficient u žen: 3,32/65 = 0,0511 (5,11 %)

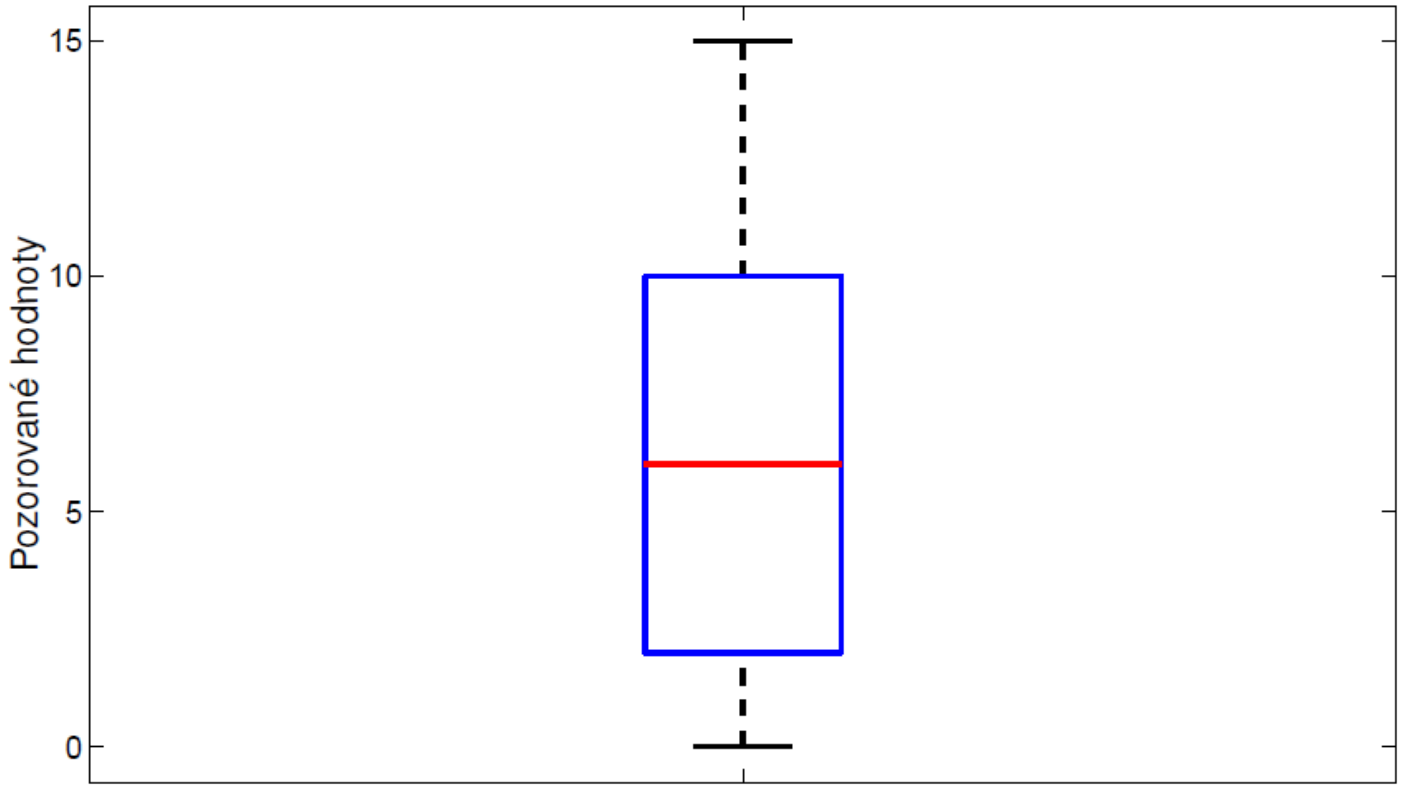
Závěr: Muži jsou v průměru skutečně těžší, ale relativně větší výkyvy hmotnosti mají ženy.

Krabicový diagram (boxplot)

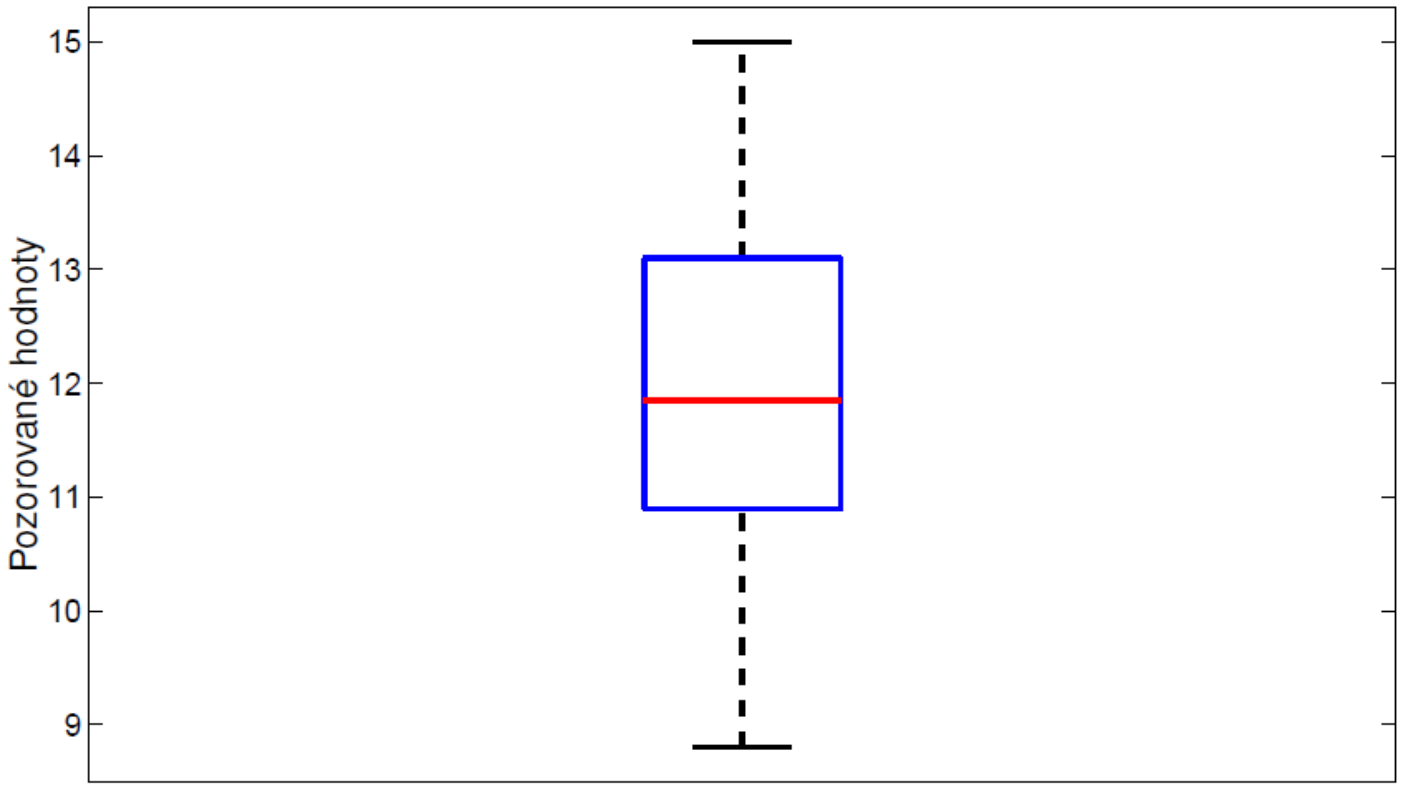


- **Dolní vnitřní hradba:** $\widetilde{x}_{0.25} - 1.5(\widetilde{x}_{0.75} - \widetilde{x}_{0.25})$
- **Horní vnitřní hradba:** $\widetilde{x}_{0.75} + 1.5(\widetilde{x}_{0.75} - \widetilde{x}_{0.25})$
- **Dolní vnější hradba:** $\widetilde{x}_{0.25} - 3(\widetilde{x}_{0.75} - \widetilde{x}_{0.25})$
- **Horní vnější hradba:** $\widetilde{x}_{0.75} + 3(\widetilde{x}_{0.75} - \widetilde{x}_{0.25})$
- **Odlehlá hodnota** leží mezi hradbami
- **Extrémní hodnota** leží za vnějšími hradbami

Krabicový diagram - výsledky testů ze statistiky



Krabicový diagram - hladina hemoglobinu



Dvourozměrný statistický soubor

Na každé statistické jednotce vyšetřujeme dva znaky X, Y.

Statistický soubor: uspořádané dvojice (x_i, y_i), $i = 1, \ldots, n$

Rozložení četností:

- bodové
- intervalové

Kontingenční tabulka

	nekouří	kouří
ženy	12 (48 %)	13 (52 %)
muži	21 (66 %)	11 (34 %)

Číselné charakteristiky

Nesetříděný soubor

- **Aritmetické průměry a rozptyly:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- **Výběrová kovariance:**

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

  - kvantifikace vztahu mezi dvěma **kvantitativními** proměnnými
  - $s_{xy} \in \mathbb{R}$
  - závisí na jednotkách, ve kterých jsou znaky X a Y zaznamenány

$\rightarrow$ nevýhoda pro porovnávání

Pearsonův korelační koeficient

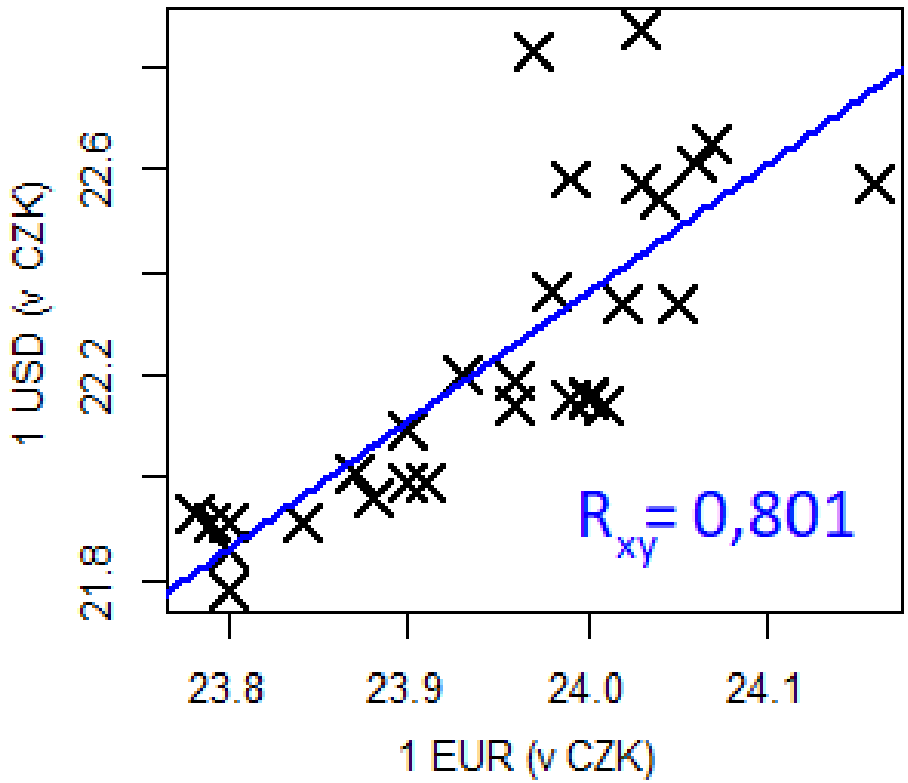
- **míra lineární závislosti** mezi dvěma kvantitativními proměnnými

$$R_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right]\left[n\sum y_i^2 - (\sum y_i)^2\right]}}$$

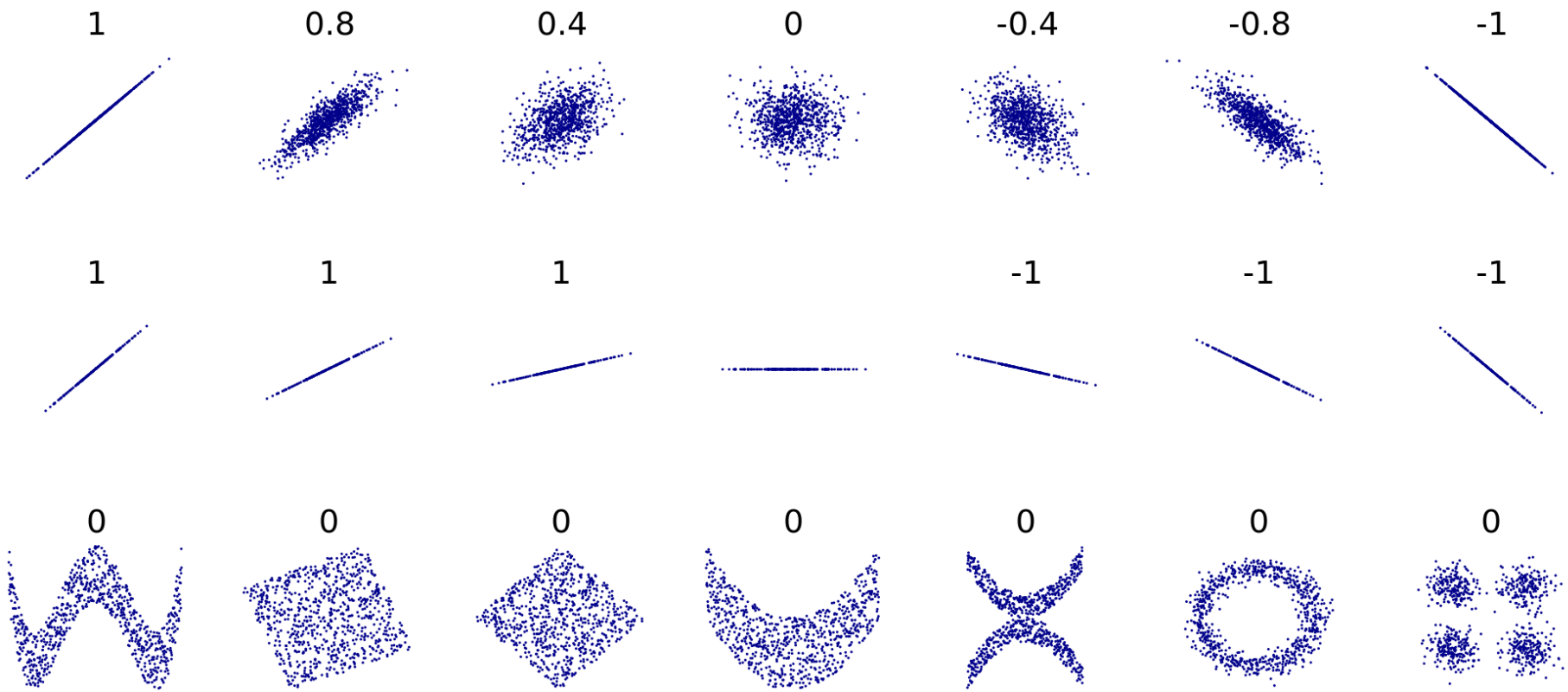
- normovaná podoba kovariance: hodnotu kovariance vztáhneme k jednotlivým směrodatným odchylkám
- $R_{xy} \in \langle -1; 1 \rangle$
- R_{xy} blízké 1: (silná) kladná lineární závislost
  čím vyšší X, tím větší Y
- R_{xy} blízké -1: (silná) záporná lineární závislost
  čím vyšší X, tím menší Y

Korelace - příklad

- Sledujeme kurzy české koruny k americkému dolaru a české koruny k euru během 30 dnů.
- Každý bod odpovídá kombinaci kurzů za daný den.
- Modrá přímka ukazuje lineární regresní vztah mezi oběma znaky.



Vizualizace (ne)lineární závislosti

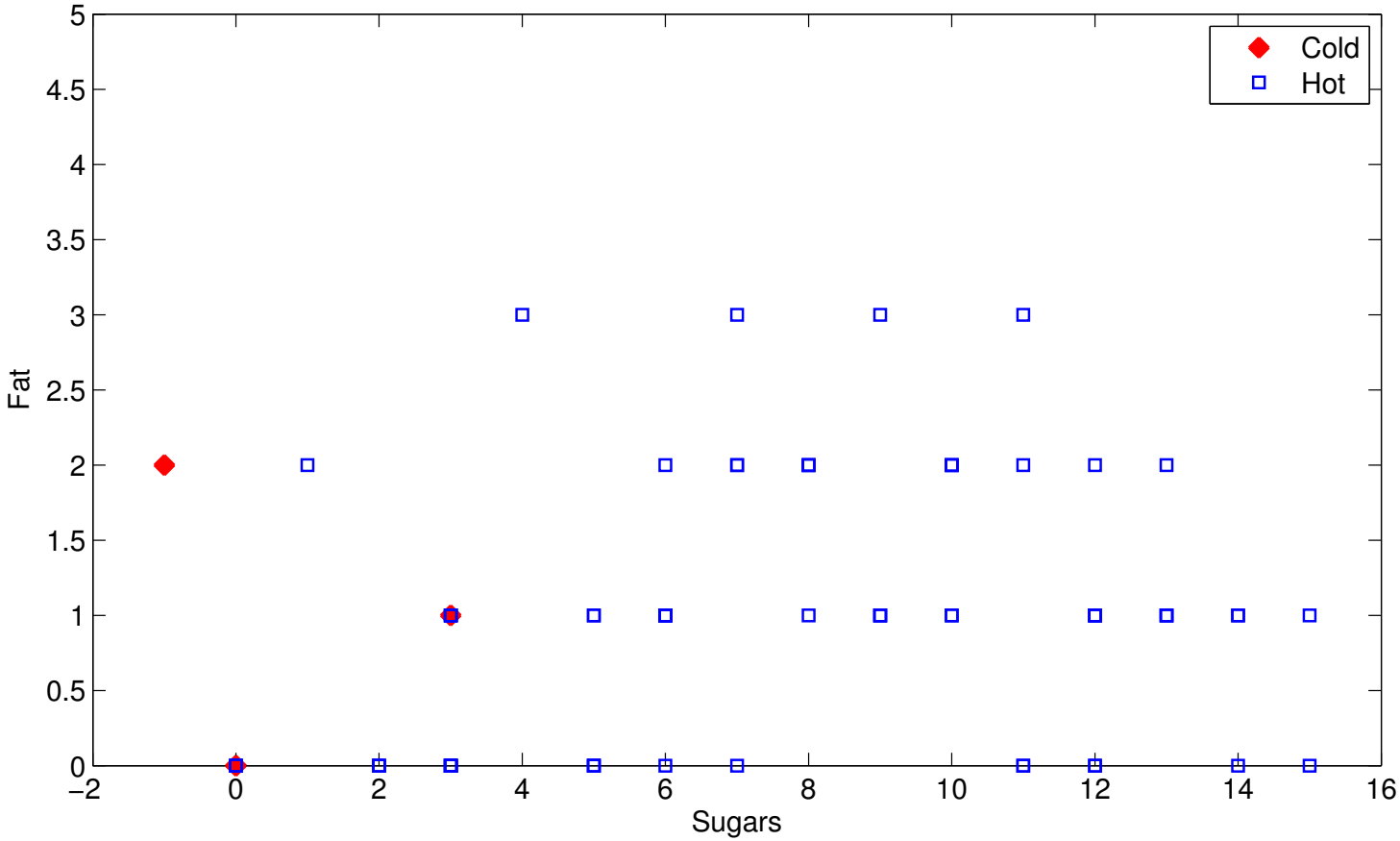


https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/1920px-Correlation_examples2.svg.png

- Interpretace hodnot Pearsonova korelačního koeficientu
- 0-0,19: mezi znaky X a Y není lineární vztah
- 0,20-0,39: mezi X a Y je slabý pozitivní lineární vztah
- 0,40-0,59: mezi X a Y je středně silný pozitivní lineární vztah
- 0,60-0,79: mezi X a Y je silný pozitivní lineární vztah
- 0,80-1: mezi X a Y je velmi silný pozitivní lineární vztah
- analogicky pro záporné hodnoty

Vizualizace vícerozměrných statistických souborů

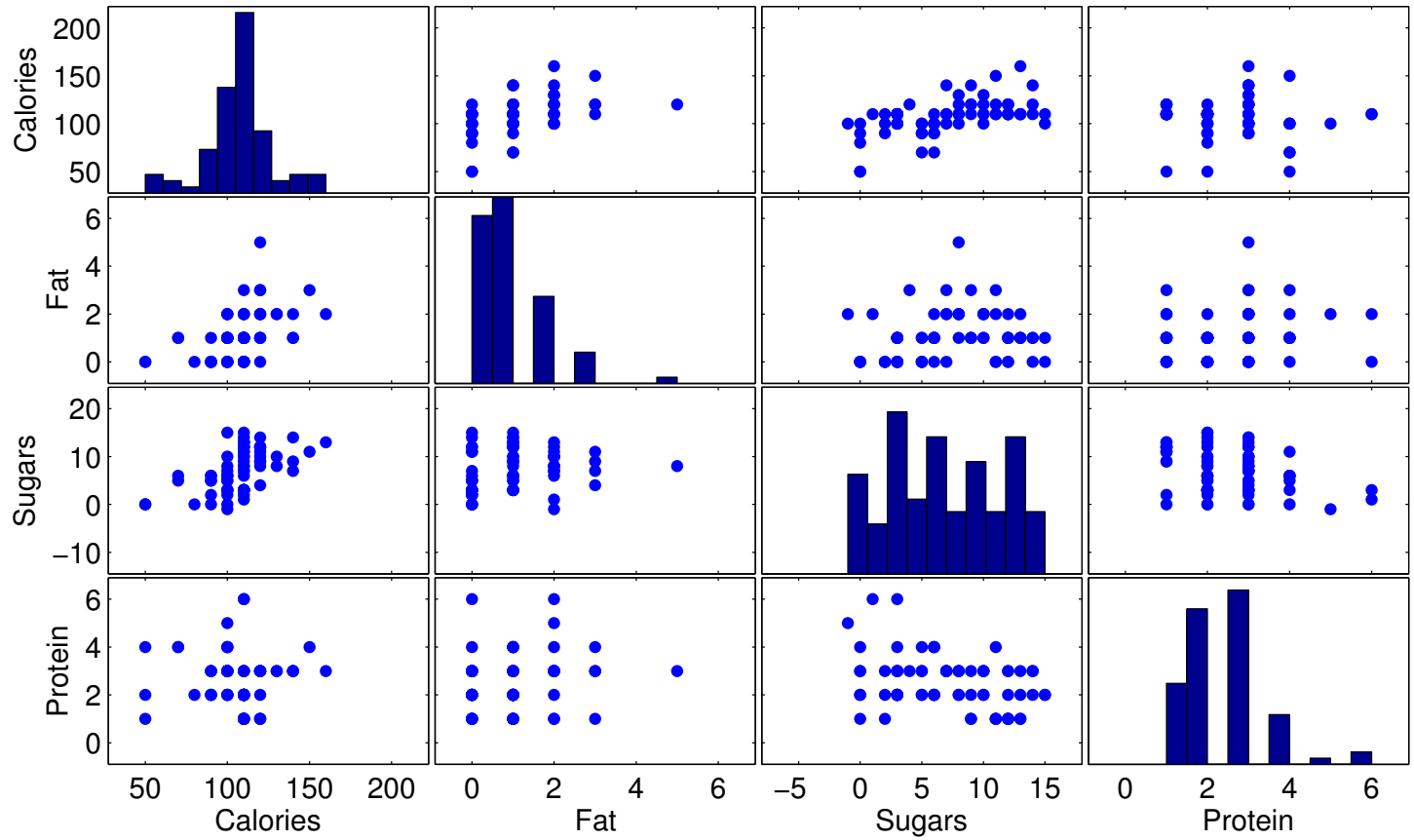
Bodový graf



- chybná hodnota množství cukrů (-1g)
- $r = 0.27 \implies$ slabý lineární vztah mezi množstvím cukrů a tuků

Vizualizace vícerozměrných statistických souborů

Matice bodových grafů



- středně silná lineární závislost mezi počtem kalorií a množstvím tuku $r = 0.50$, resp. množstvím cukrů $r = 0.56$