Základy pravděpodobnosti a matematické statistiky

7. Úvod do matematické statistiky

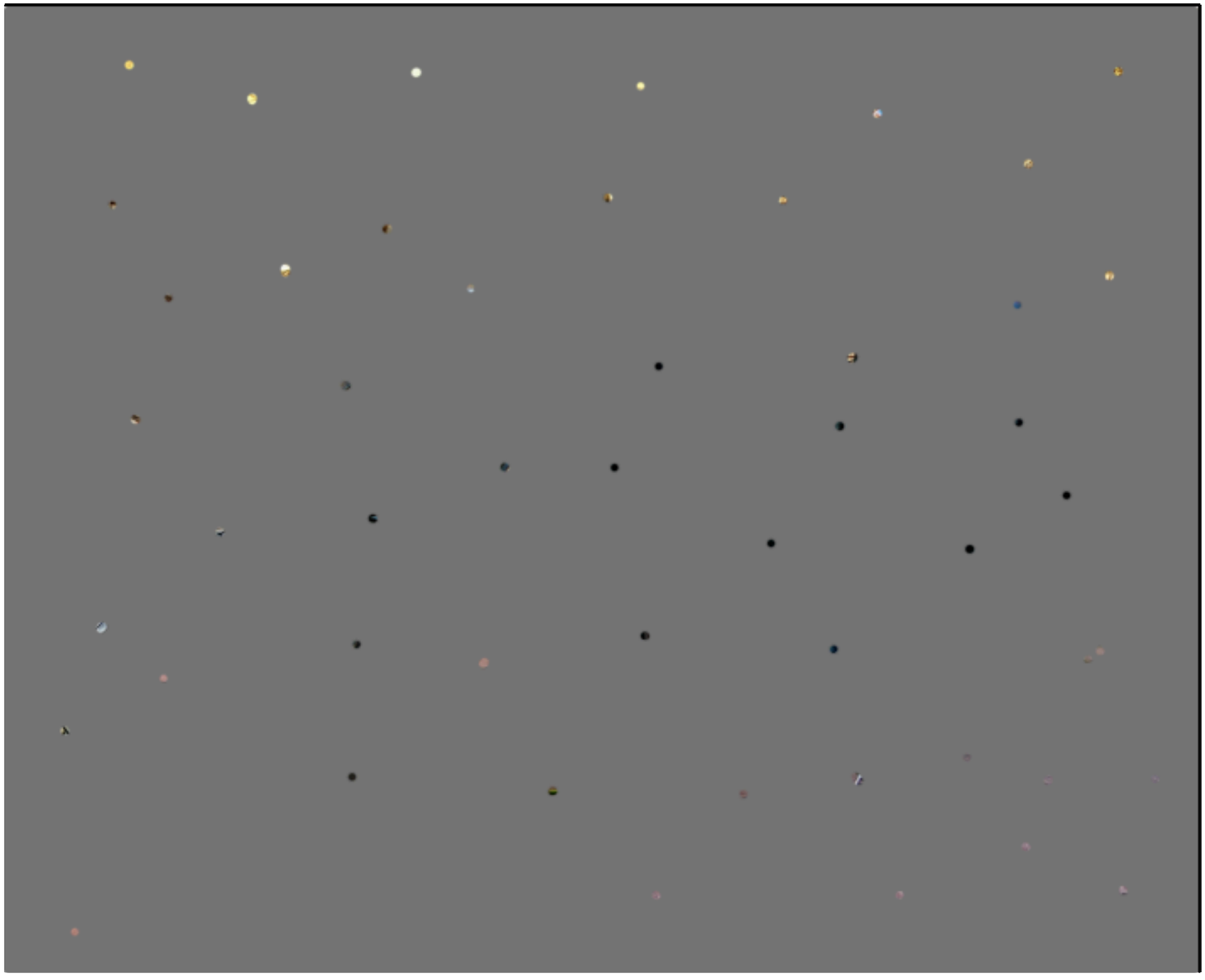
FAKULTA STROJNÍ ČVUT V PRAZE

prof. RNDr. Gejza Dohnal, CSc.

ak. rok 2023/2024

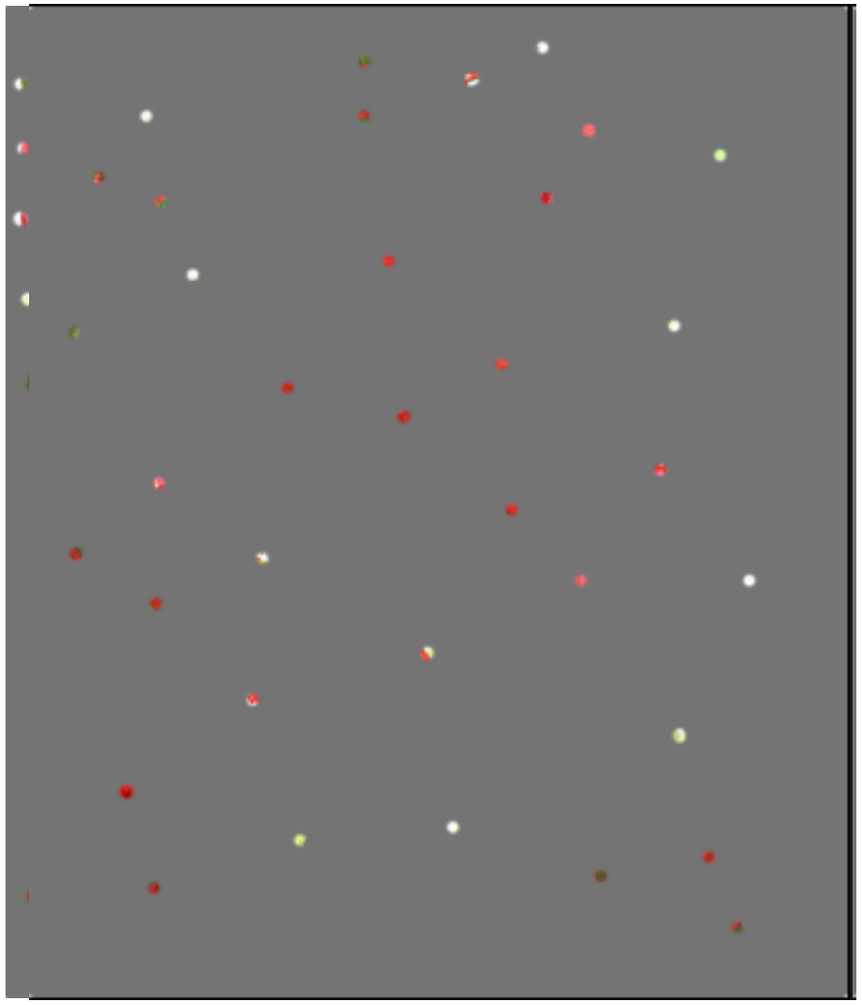
7. Úvod do matematické statistiky

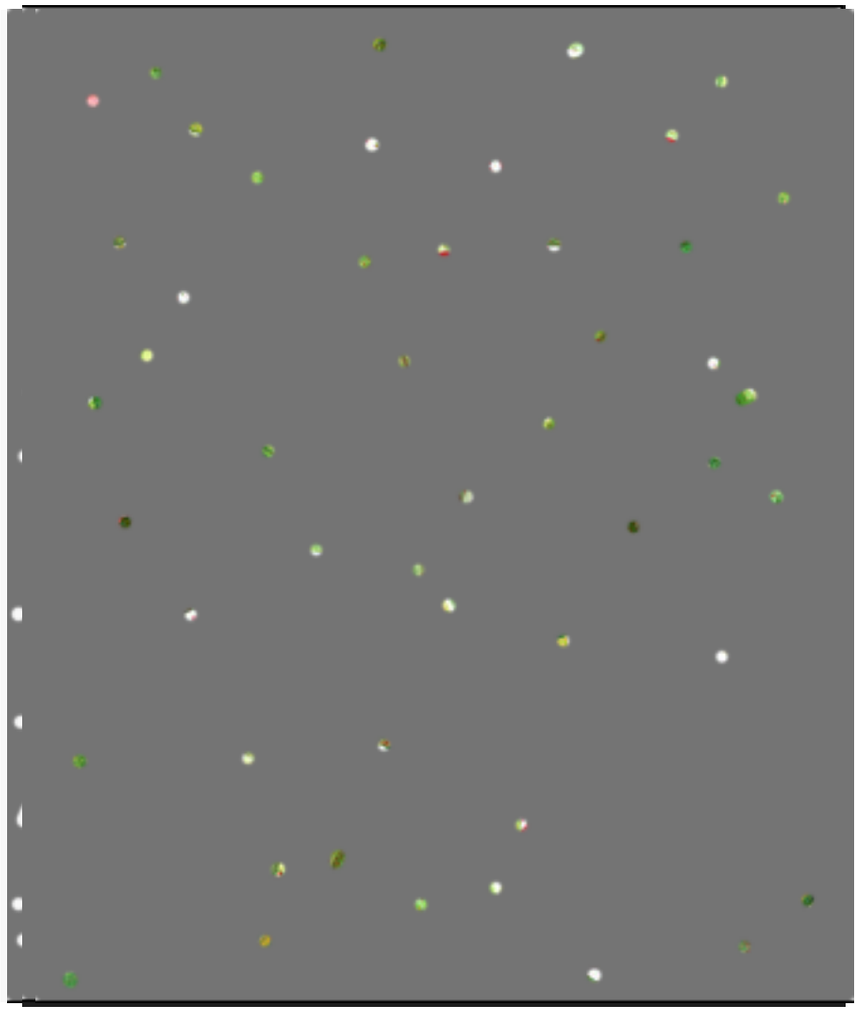
Úloha statistické indukce



Kawasaki
EN 500

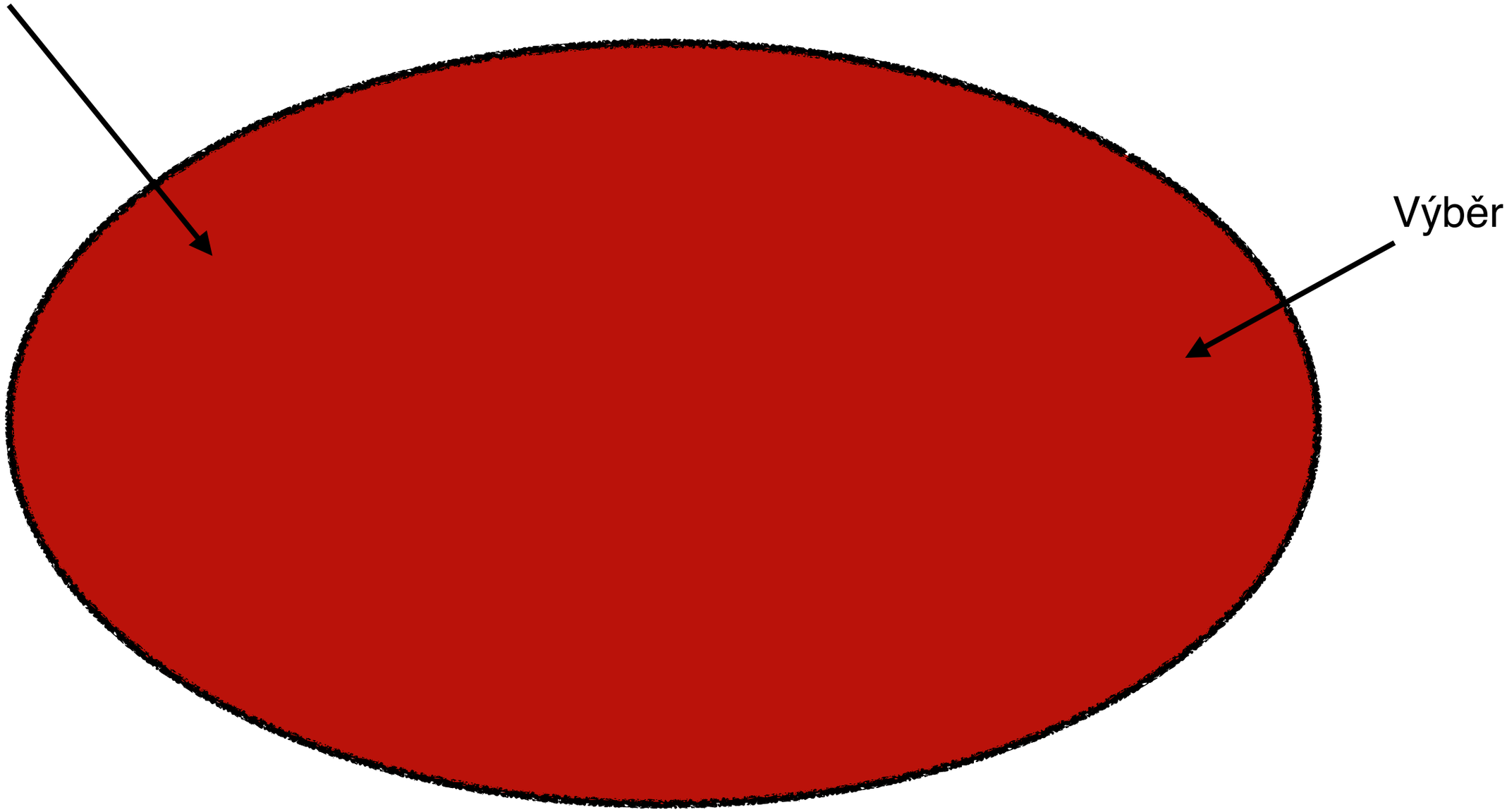
Úloha statistické indukce





Úloha statistické indukce

Základní soubor - nositel sledovaného znaku (veličiny)

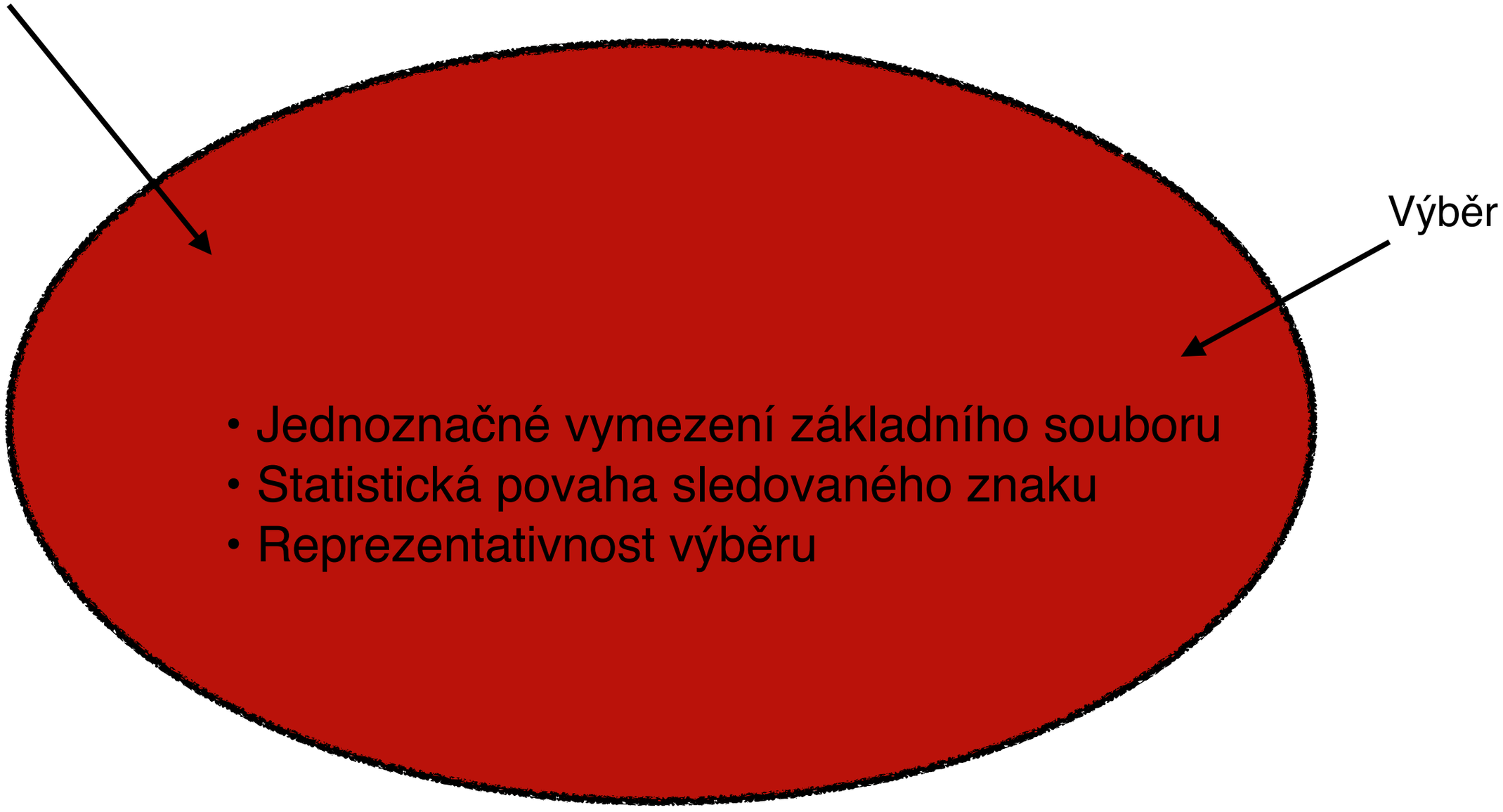


Pozorování výběru (měření sledovaného znaku) => Zjištění vlastností výběru

=> zobecnění na celý základní soubor

Úloha statistické indukce

Základní soubor - nositel sledovaného znaku (veličiny)



Pozorování výběru (měření sledovaného znaku) => Zjištění vlastností výběru

=> zobecnění na celý základní soubor

Statistické charakteristiky

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ výběru $X_1, X_2, \ldots, X_n$.

Pravděpodobnostní charakteristiky	Výběrové charakteristiky
Střední hodnota $E(X) = \int_{-\infty}^{\infty} x f(x) dx$	Výběrový průměr $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$
Momenty $\mu_k(X) = E(X - E(X))^k$	Výběrové momenty $m_k(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^k$
Rozptyl $var(X) = E(X - E(X))^2$	Výběrový rozptyl $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2$
Kvantily $\tilde{x}_{100\alpha}$	Výběrové kvantily $X_{([np]+1)}$

Statistické charakteristiky

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ výběru $X_1, X_2, \ldots, X_n$.

Za předpokladu, že náhodný výběr je nezávislý a je z normálního rozdělení, t.j. $X_1, X_2, \ldots, X_n$ jsou i.i.d. a $X_k \sim N(\mu,\sigma^2)$, $k = 1, 2, \ldots, n$, lze určit rozdělení pravděpodobnosti některých charakteristik:

- Pokud je μ a σ^2 známé, má výběrový průměr $\bar{X}_n$ rozdělení $N(\mu,\sigma^2/n)$
- Pokud μ a σ^2 neznáme, má veličina $T=(X - \bar{X})/s$ tzv. Studentovo neboli t-rozdělení $t(n\text{-}1)$
- Veličina $S^2 = (n\text{-}1).s^2/\sigma^2$ má $\chi^2(n\text{-}1)$ rozdělení (o $n\text{-}1$ stupních volnosti)

Statistické charakteristiky

Další důležité výběrové charakteristiky:

- Výběrová šikmost (skewness): $Skew(X) = \frac{m_3(X)}{m_2^{3/2}(X)}$

pro $X \sim \mathrm{N}(\mu,\sigma^2)$ je

$$E(Skew(X)) = 0 \qquad var(Skew(X)) = \frac{6(n-2)}{(n+1)(n+3)}$$

- Výběrová špičatost (kurtosis): $Kurt(X) = \frac{m_4(X)}{m_2^2(X)} - 3$

pro $X \sim \mathrm{N}(\mu,\sigma^2)$ je

$$E(Kurt(X)) = -\frac{6}{n+1} \qquad var(Kurt(X)) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

Máme-li dostatečný počet pozorování (řádově stovky), mají statistiky

$$T_3 = \frac{S_{kew}^{norm}}{\sqrt{Var(S_{kew}^{norm})}} \qquad T_4 = \frac{K_{urt}^{norm} - E(K_{urt}^{norm})}{\sqrt{Var(K_{urt}^{norm})}}$$

přibližně standardní normální rozdělení pravděpodobnosti.

Statistické charakteristiky

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ výběru $X_1, X_2, \ldots, X_n$.

- *Uspořádaný výběr*: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ vznikne z původního výběru $X_1, X_2, \ldots, X_n$ uspořádáním podle velikosti pozorovaných hodnot $x_1, x_2, \ldots, x_n$.
- *Pořadová statistika*: $X_{(k)}$ je náhodná veličina X_m, která je k-tá v pořadí podle velikosti pozorovaných hodnot $x_1, x_2, \ldots, x_n$. Index k nazýváme *pořadím veličiny* X_m a zapisujeme to $R_m = k$.
- Statistika $X_{(1)}$ se nazývá minimum, $X_{(n)}$ je maximum
- medián $\tilde{x}_{50}$: je-li n liché, je roven $X_{([n/2]+1)}$
  pro n sudé je roven $(X_{(n/2)} + X_{(n/2+1)})/2$
- dolní kvartil $\tilde{x}_{25}$: $X_{([n/4]+1)}$ resp. $(X_{(n/4)} + X_{(n/4+1)})/2$
- horní kvartil $\tilde{x}_{75}$: $X_{([3n/4]+1)}$ resp. $(X_{(3n/4-1)} + X_{(3n/4)})/2$

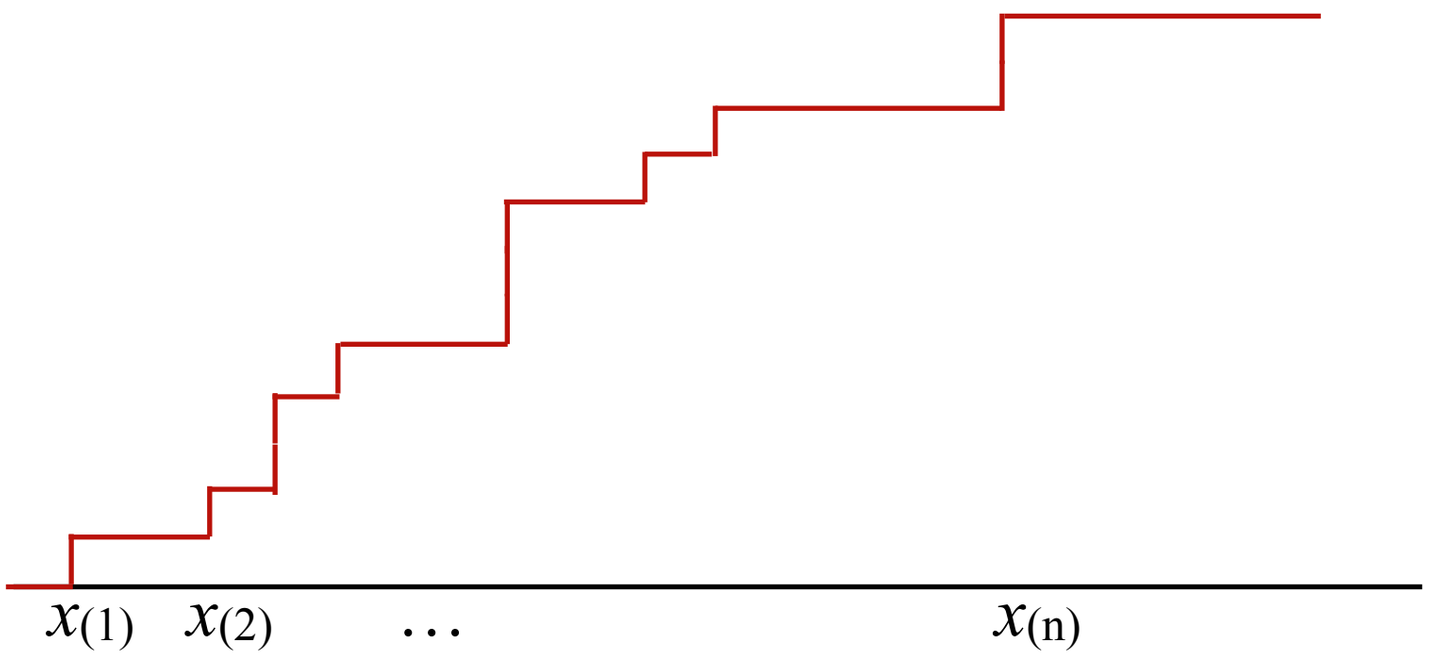
Grafická analýza

Statistické charakteristiky: jsou spočteny na základě pozorování $x_1, x_2, \ldots, x_n$ výběru $X_1, X_2, \ldots, X_n$.

Empirická distribuční funkce:

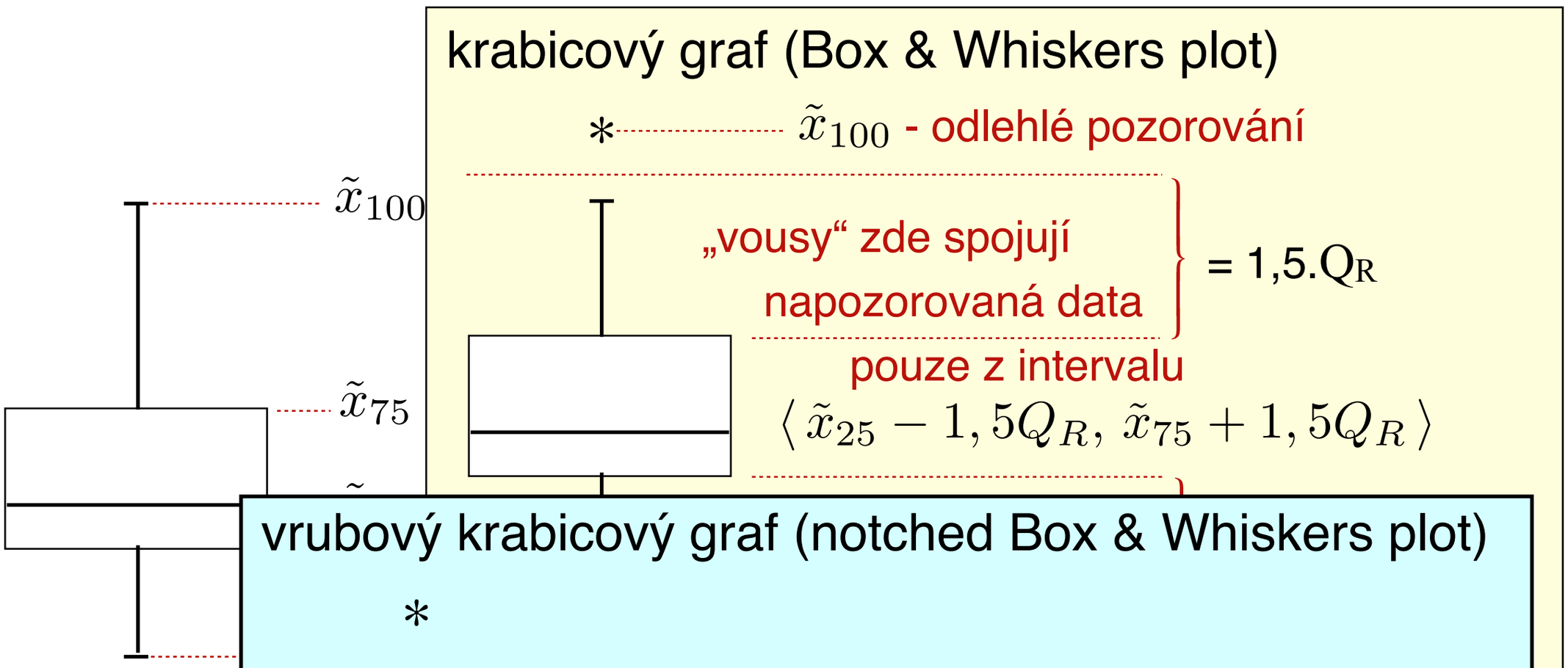
vycházíme z uspořádaného výběru: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. Potom $F_n(x_{(i)}) = \frac{i}{n}$

a tedy $$F_n(x) = \frac{\max\{k : X_{(k)} \leq x\}}{n}, \quad x \in \mathbf{R}$$



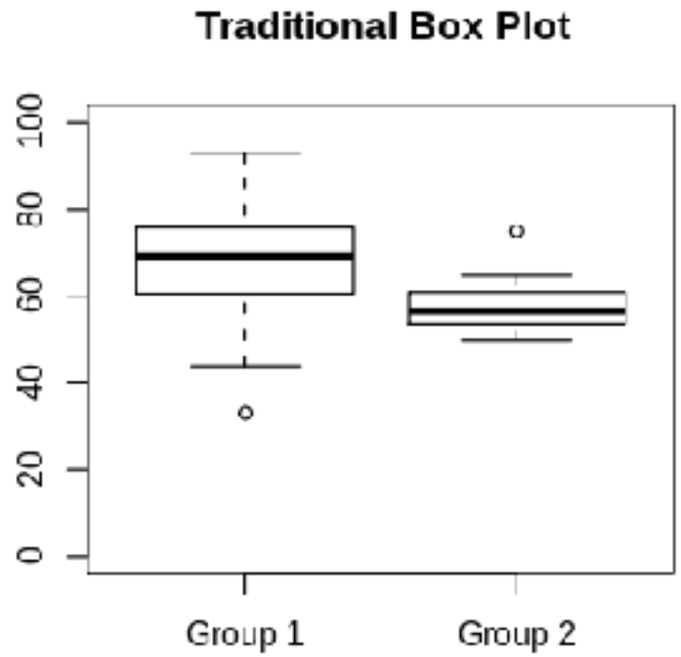
Grafická analýza

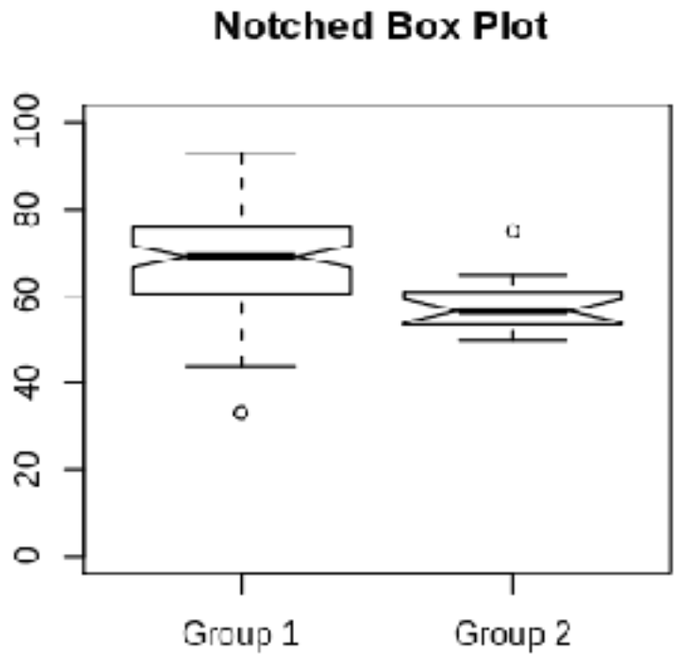
krabicový graf (Box & Whiskers plot)

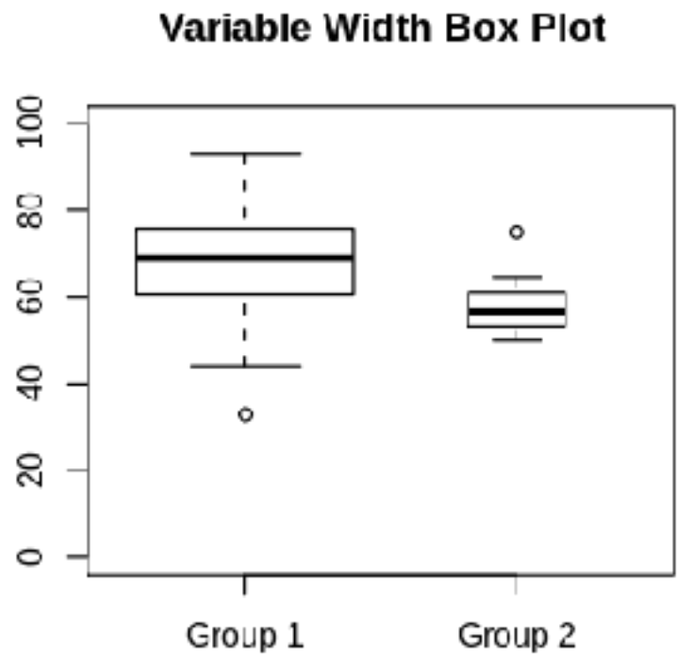


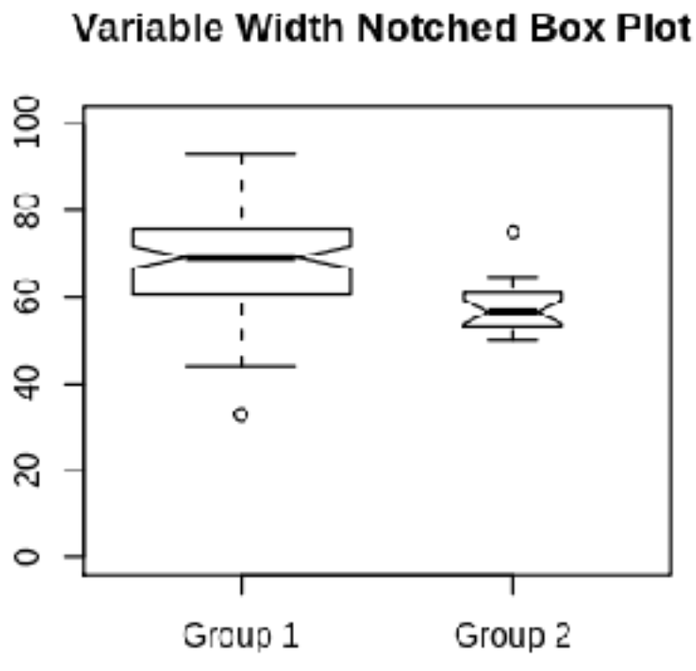
Grafická analýza

krabicový graf (Box & Whiskers plot)



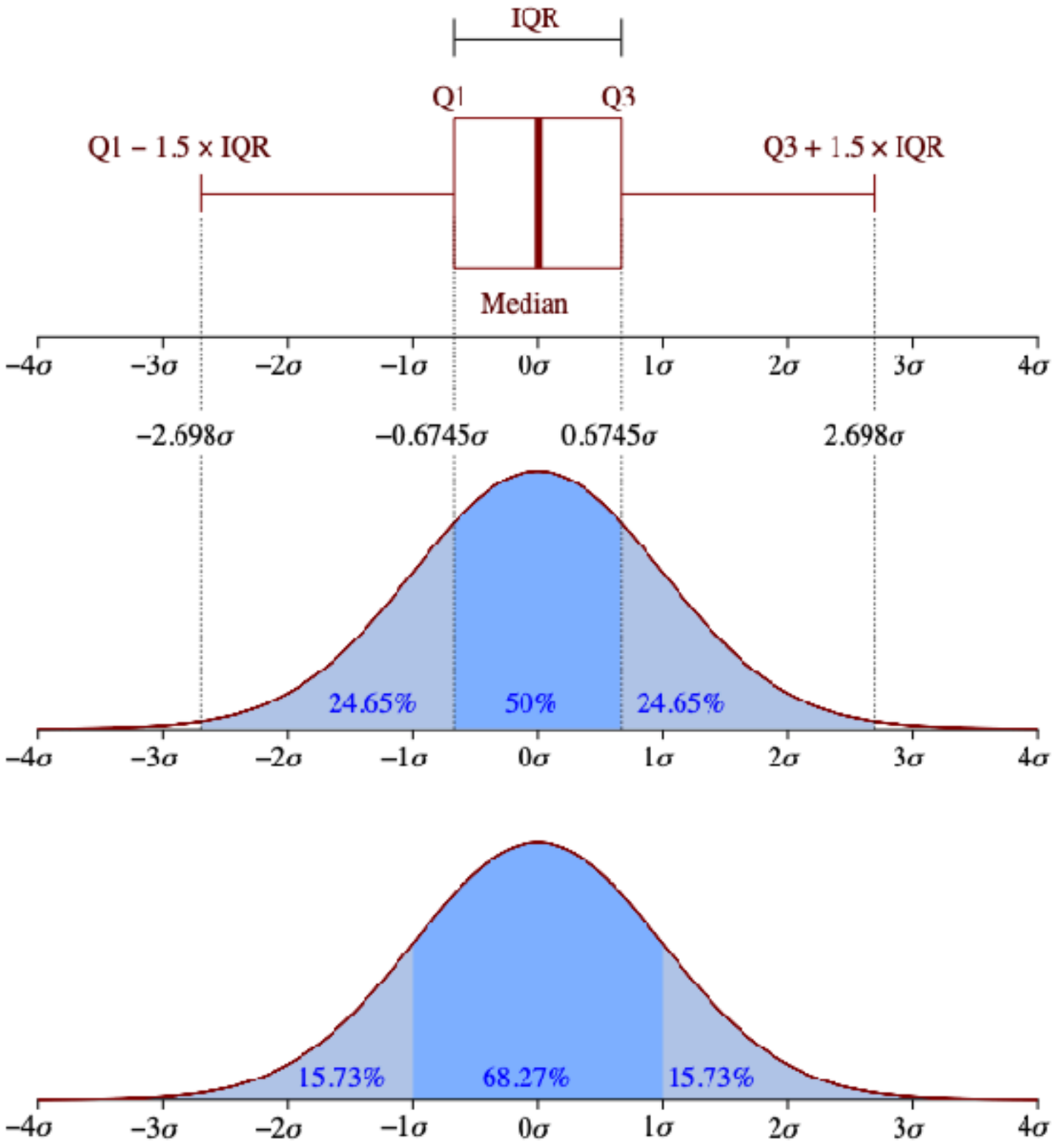






Grafická analýza

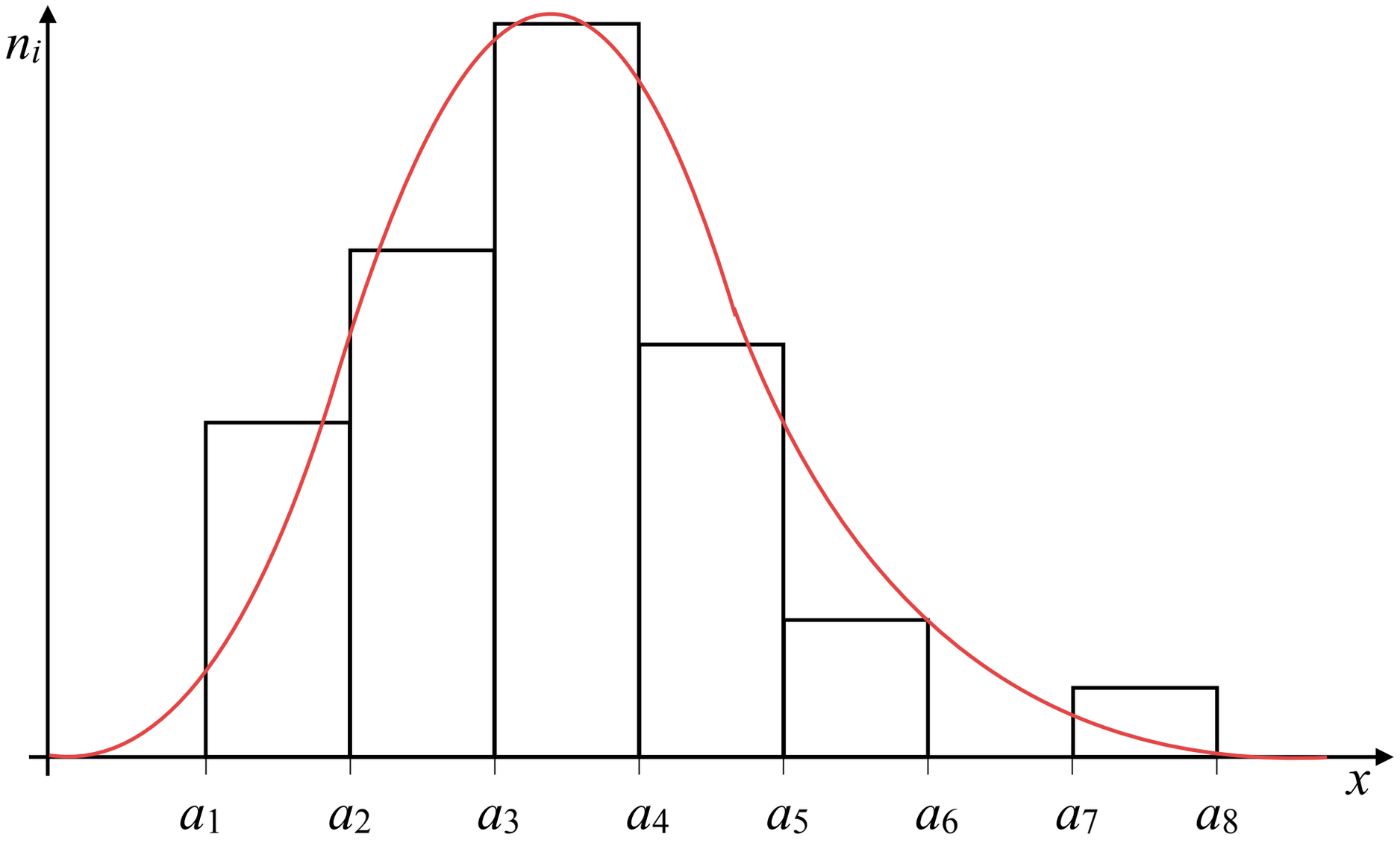
krabicový graf (Box & Whiskers plot)



Frekvenční analýza

Histogram

Máme pozorování $x_1, x_2, \ldots, x_n$ náhodného výběru $X_1, X_2, \ldots, X_n$.



Frekvenční analýza

Histogram

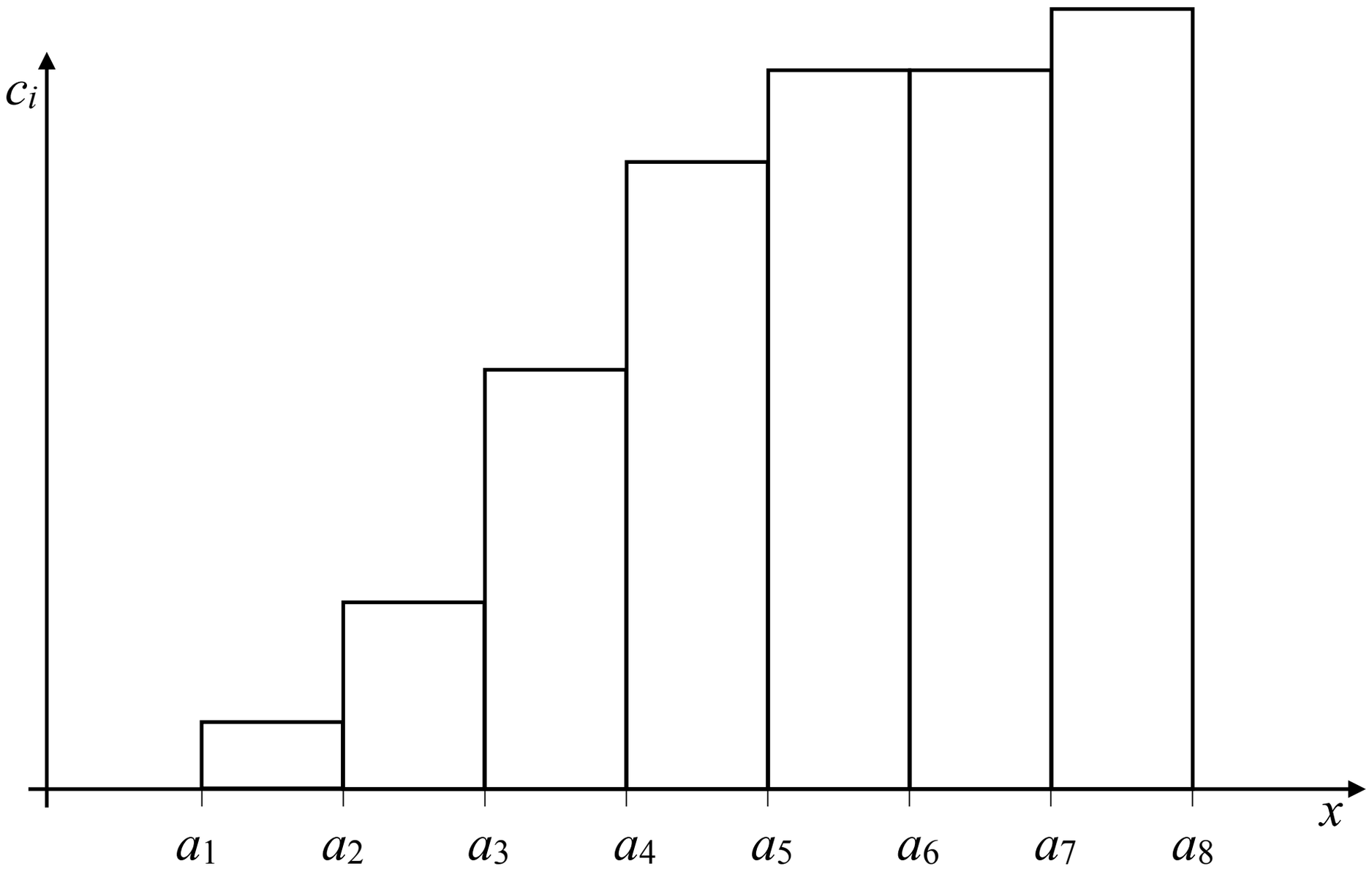
Máme pozorování $x_1, x_2, \ldots, x_n$ náhodného výběru $X_1, X_2, \ldots, X_n$.

pořadí třídy	třídní intervaly	(prosté) absolutní četnosti	(prosté) relativní četnosti	kumulativní četnosti	kumulativní relativní četnosti
1	$a_2 - a_1$	n_1	$f_1=n_1/n$	$c_1=n_1$	$d_1=c_1/n$
2	$a_3 - a_2$	n_2	$f_2=n_2/n$	$c_2=n_1+n_2$	$d_2=c_2/n$
⋮	⋮	⋮	⋮	$c_j = \sum_{i=1}^{j} n_i$	⋮
k	$a_k - a_{k-1}$	n_k	$f_k=n_k/n$	$c_k=n$	$d_k=c_k/n=1$

Frekvenční analýza

Histogram

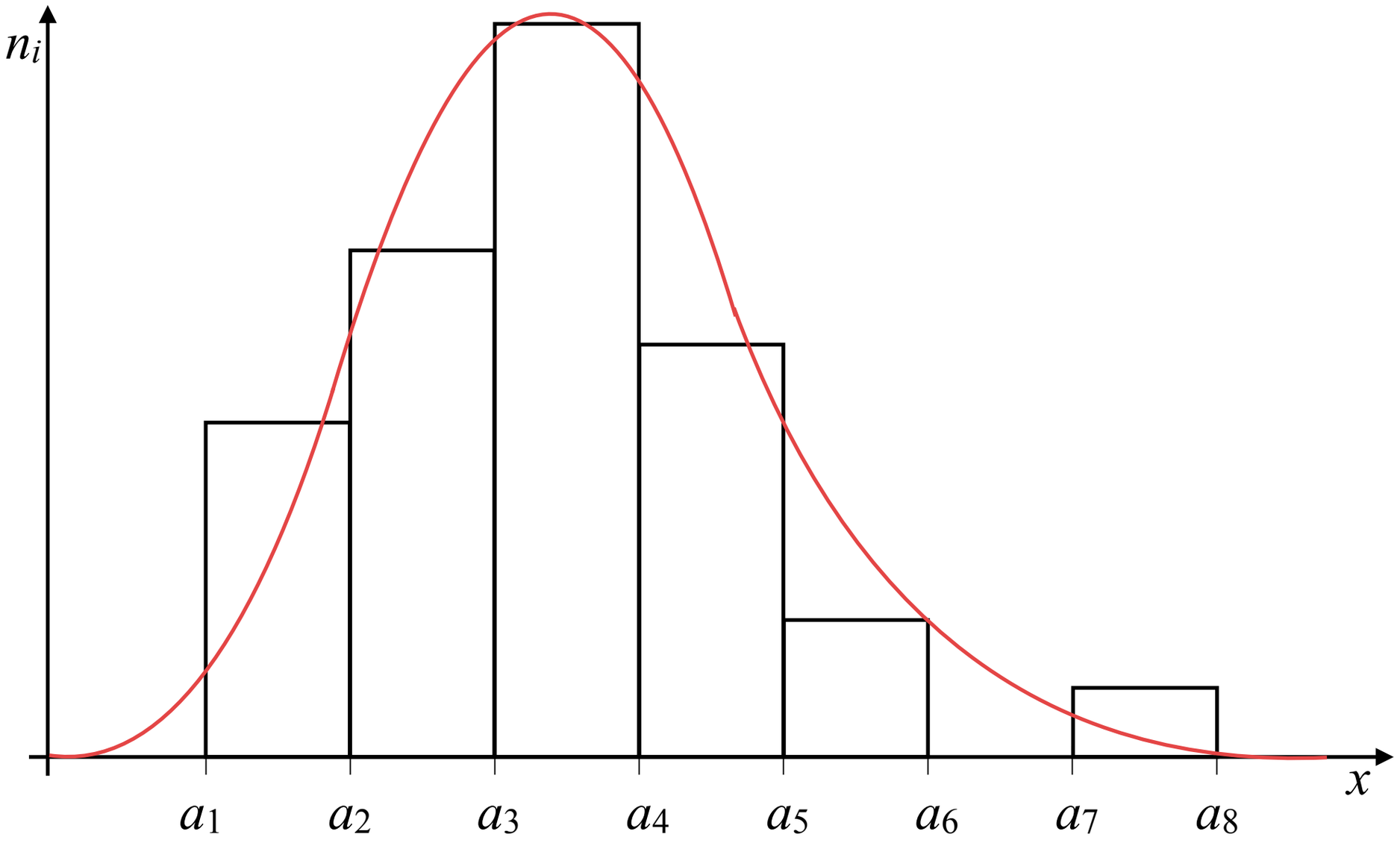
Máme pozorování $x_1, x_2, \ldots, x_n$ náhodného výběru $X_1, X_2, \ldots, X_n$.



Frekvenční analýza

Histogram

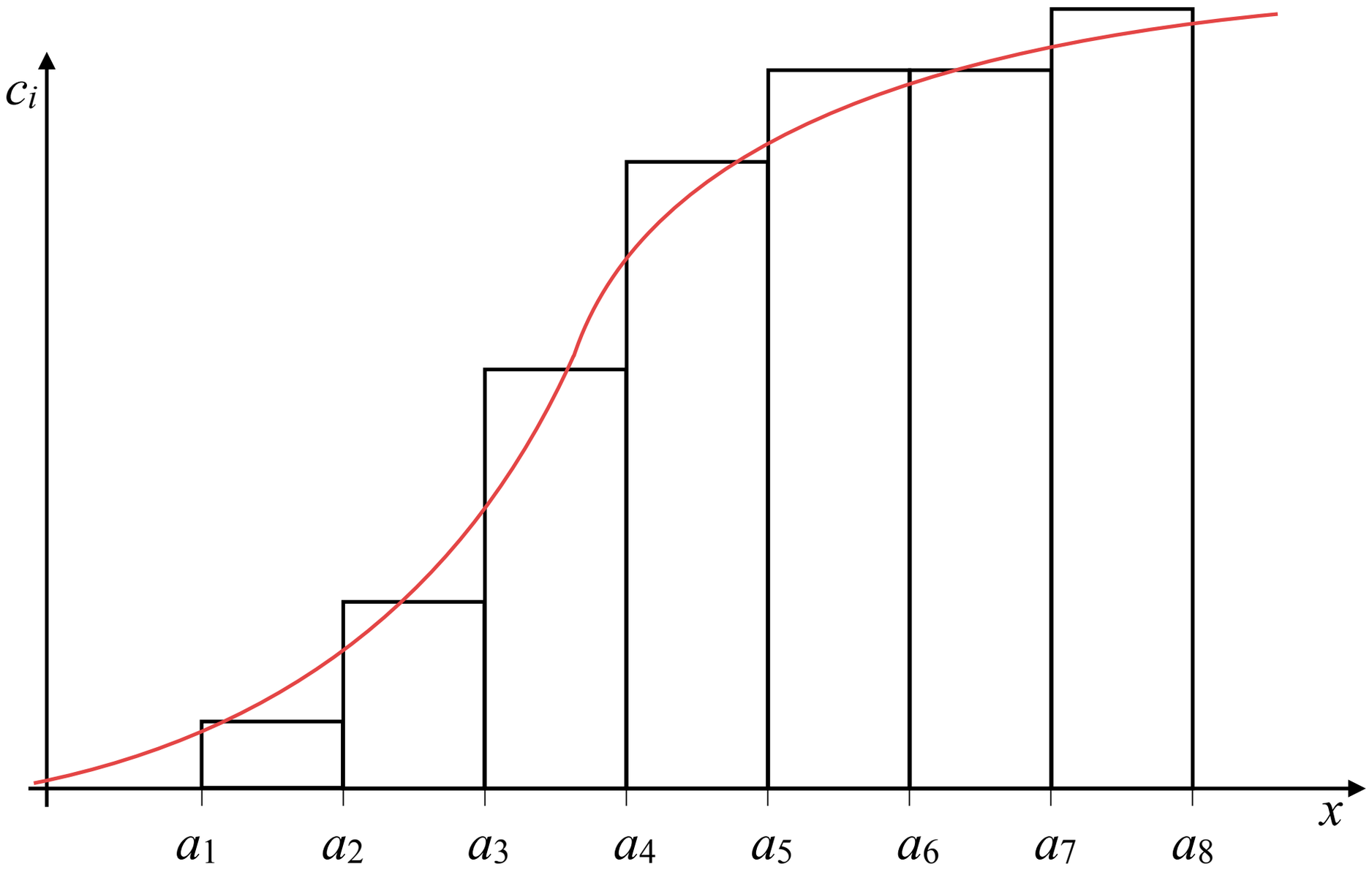
Máme pozorování $x_1, x_2, \ldots, x_n$ náhodného výběru $X_1, X_2, \ldots, X_n$.



Frekvenční analýza

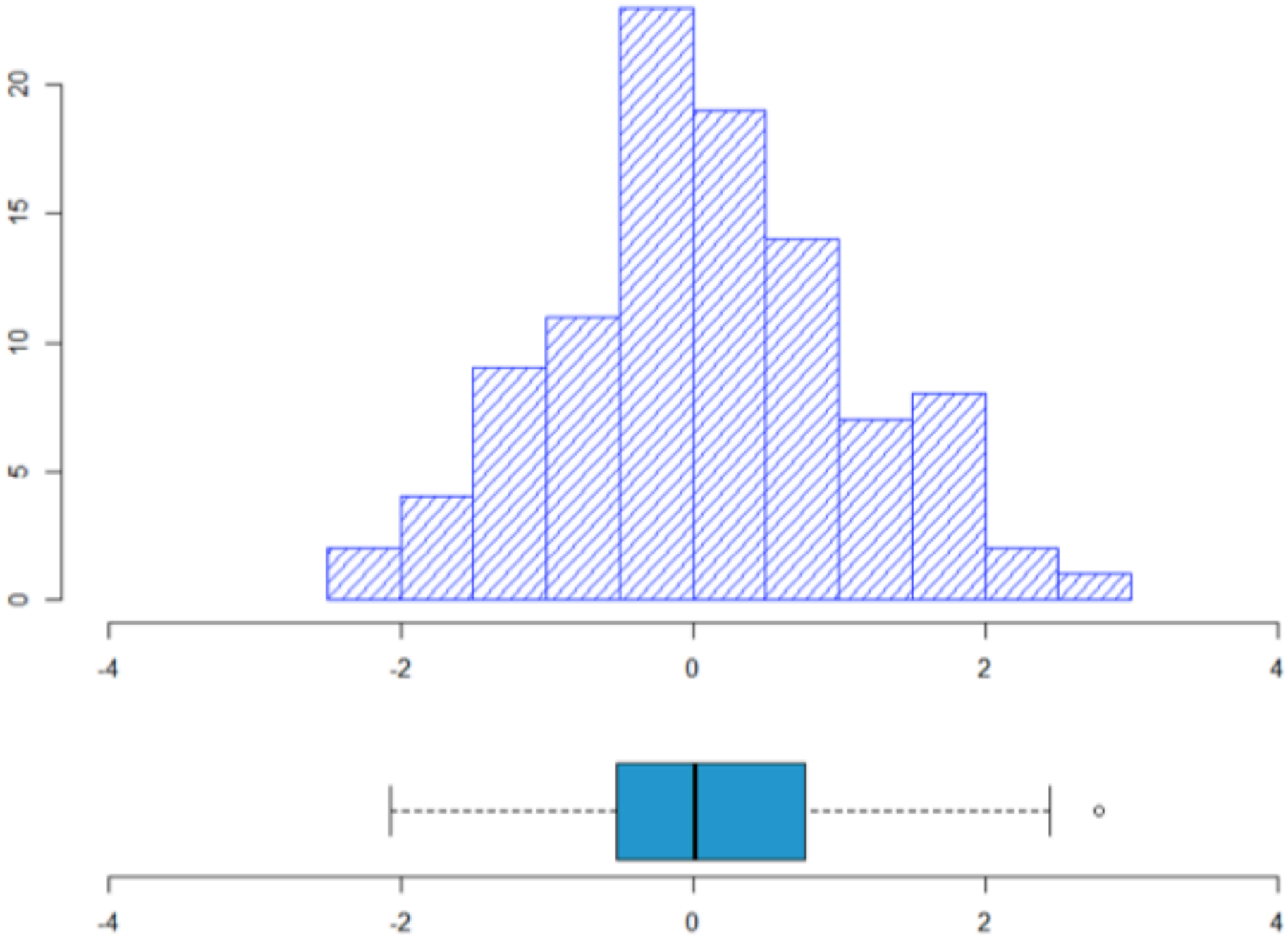
Histogram

Máme pozorování $x_1, x_2, \ldots, x_n$ náhodného výběru $X_1, X_2, \ldots, X_n$.



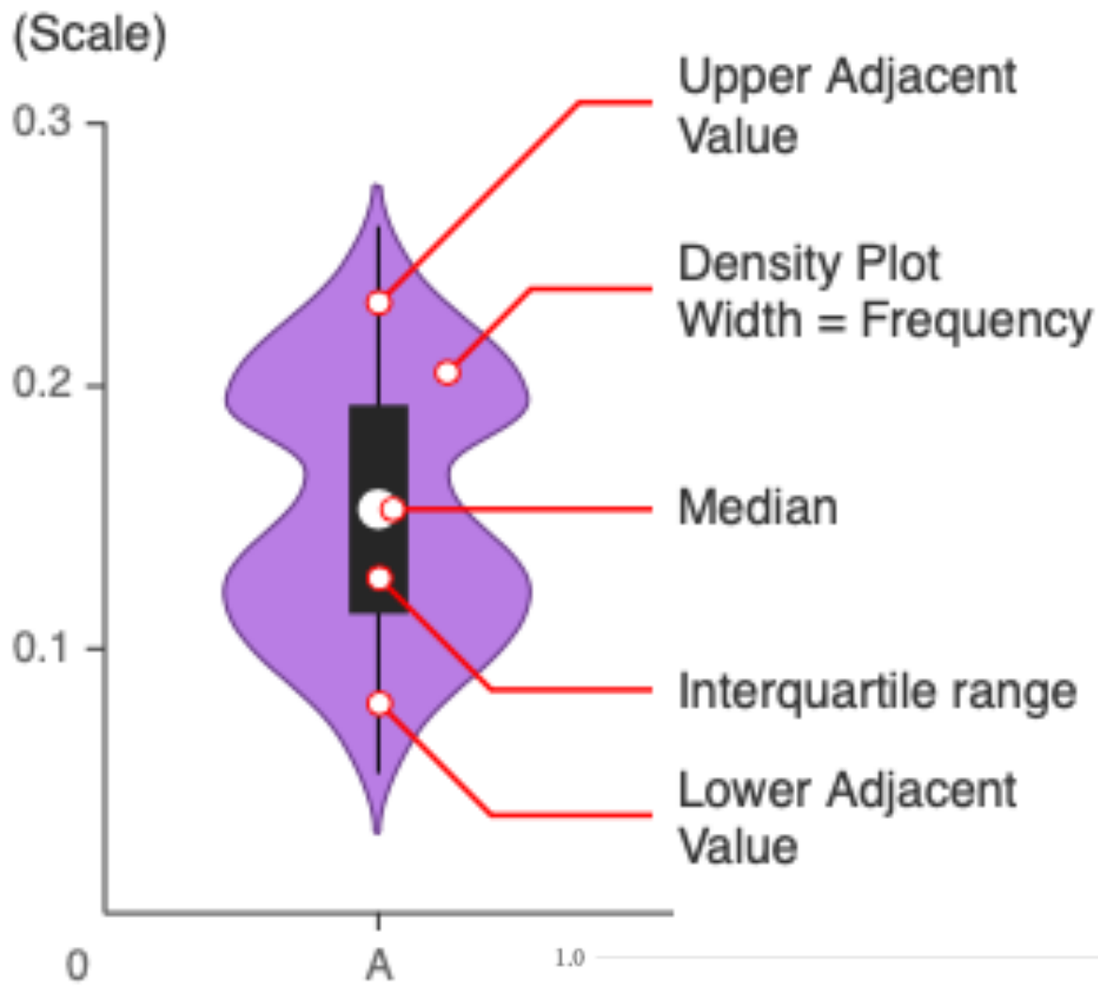
Frekvenční analýza

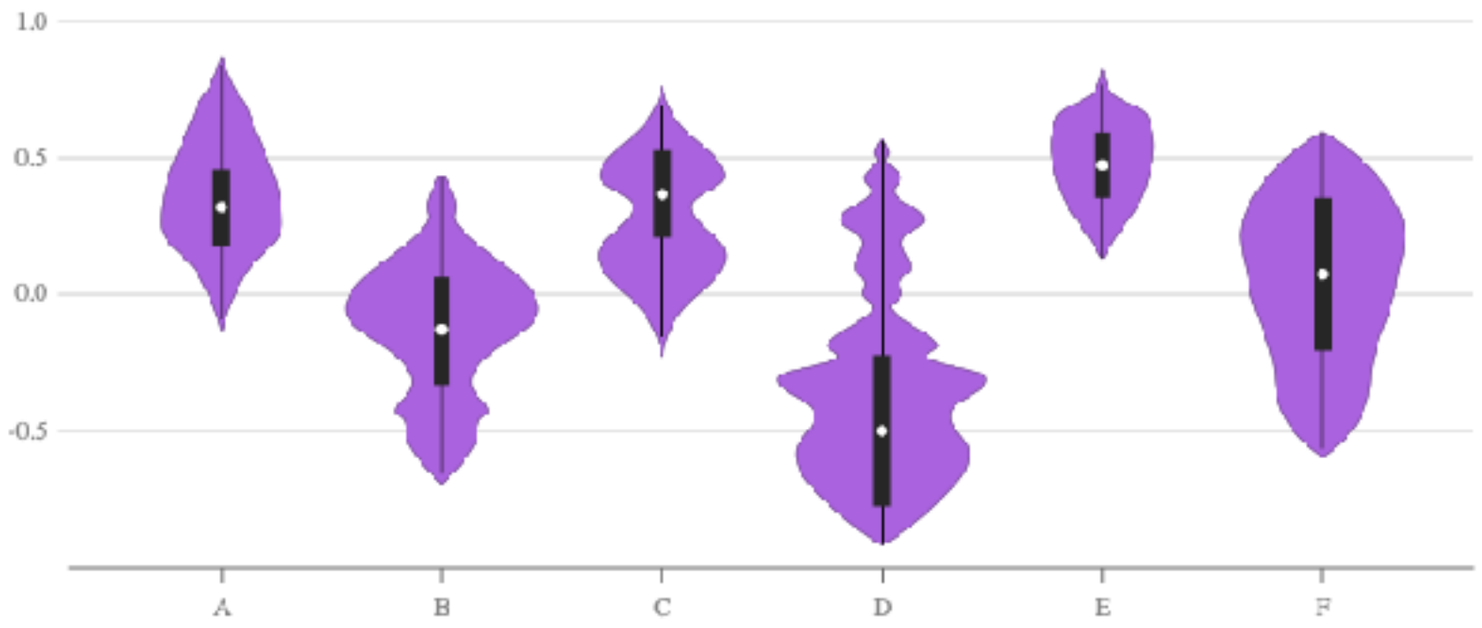
Histogram x Box plot



Frekvenční analýza

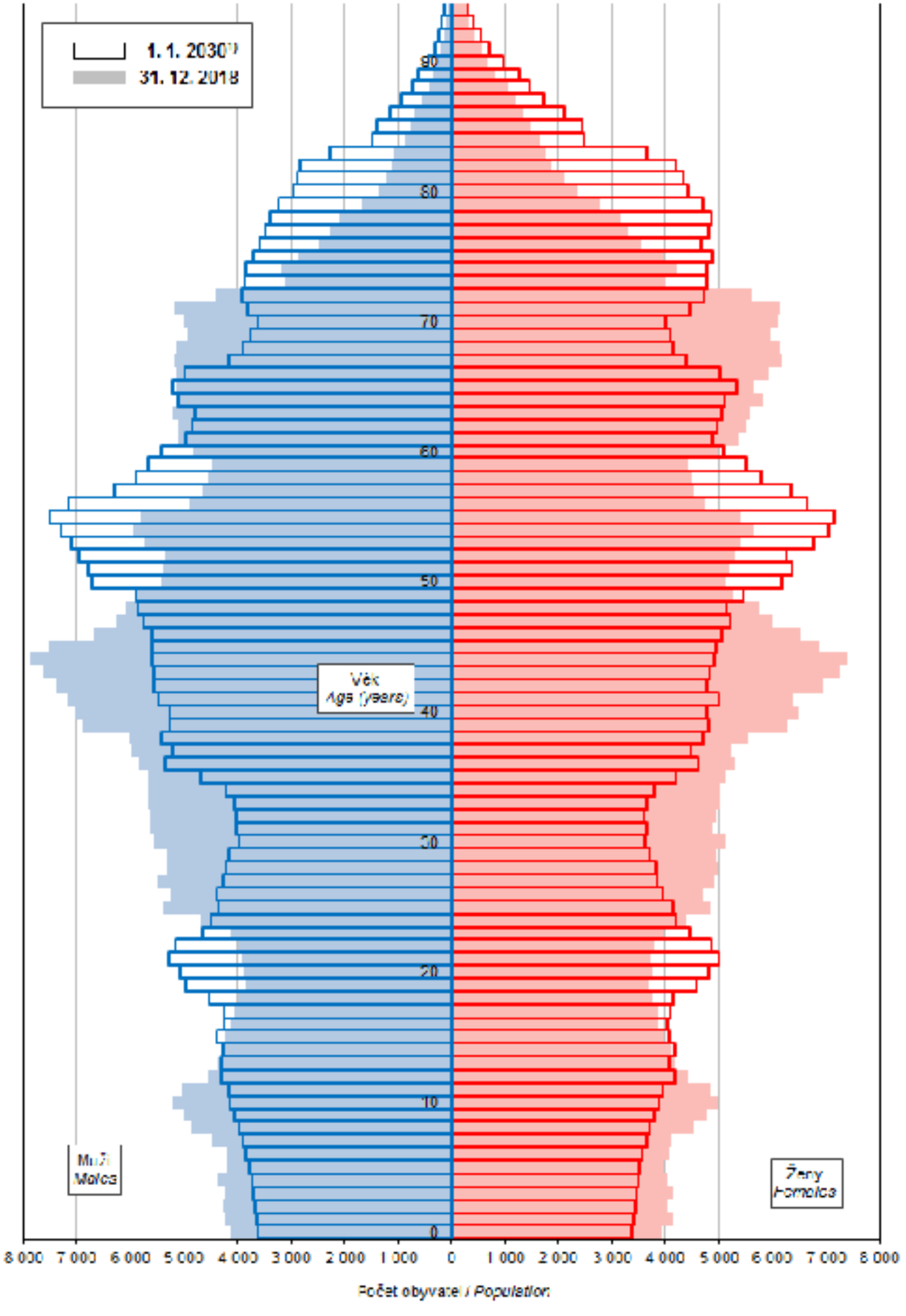
Houslový graf (Violin plot)





Věkové složení obyvatelstva Ústeckého kraje k 31. 12. 2018 a k 1. 1. 2030

Age distribution of the population in the Ústecký Region as at 31 December 2018 and 1 January 2030

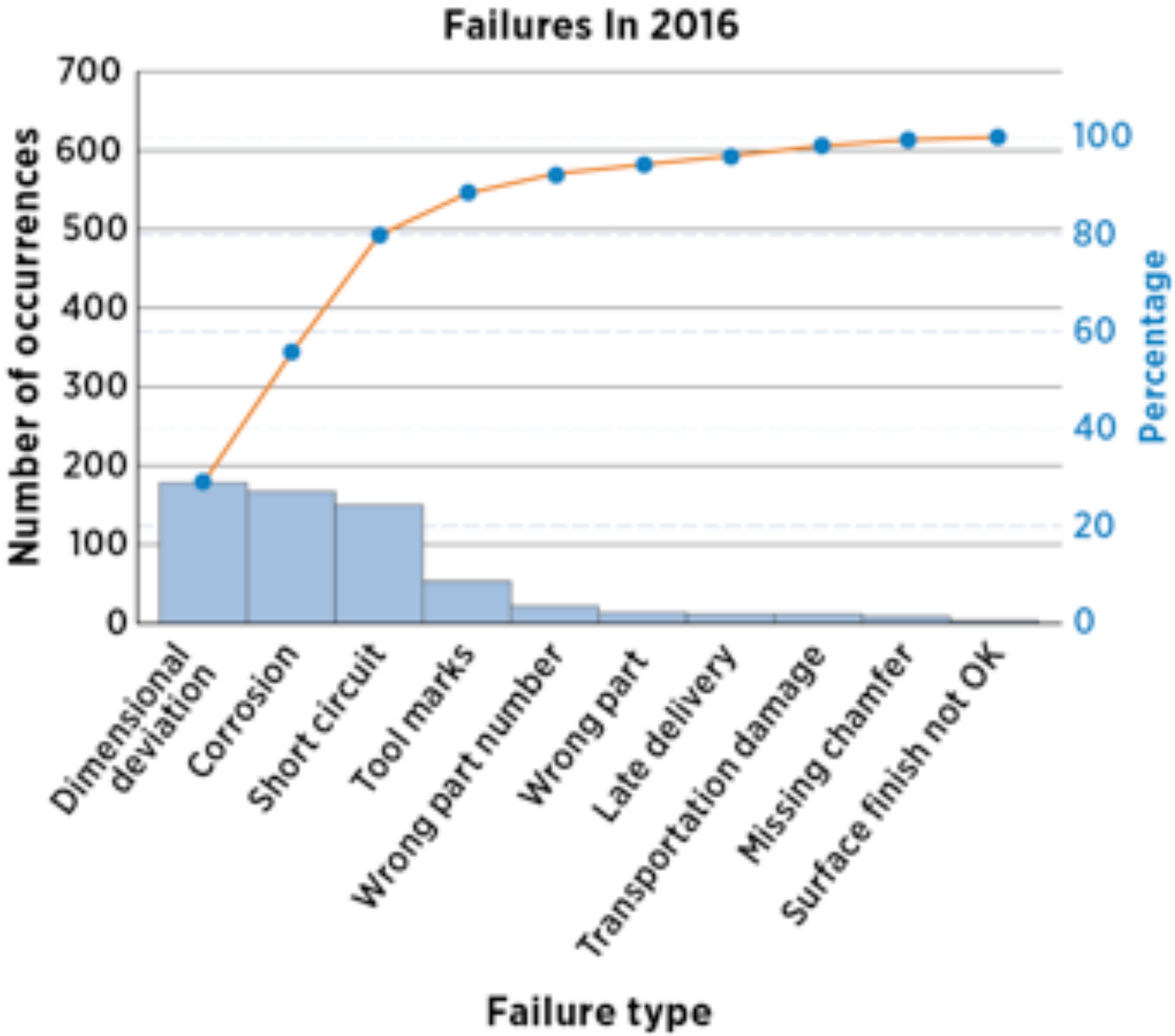


[1] Zdroj: Projekce obyvatelstva v krajích ČR do roku 2070

[1] Source: CZSO publication "Projekce obyvatelstva v krajích ČR do roku 2070" (Czech only)

Frekvenční analýza

Paretův graf



Malý slovníček statistických pojmů, aneb

co je třeba znát a porozumět tomu:

- **Základní soubor (populace, universum)**: množina objektů, na nichž provádíme statistické zkoumání; musí být přesně specifikována
- **Výběr (ze základního souboru)**: n-tice náhodných veličin $X_1, X_2, \ldots, X_n$ odpovídající nezávislým pozorováním vybraných objektů základního souboru na nichž pozorujeme nějakou veličinu X reprezentující určitou měřitelnou (a přesně danou) vlastnost všech objektů základního souboru.
- **Rozsah výběru** je počet objektů n zahrnutých do výběru.
- **Reprezentativnost výběru** je vlastnost výběru, zaručující rovnoměrné zastoupení charakteristických vlastností objektů základního souboru.
- **Náhodný výběr** vznikne tehdy, když každý objekt základního souboru má stejnou pravděpodobnost být zahrnut do výběru.
- **Realizace výběru**: je množina naměřených (napozorovaných) číselných hodnot $x_1, x_2, \ldots, x_n$ veličin z výběru.

Malý slovníček statistických pojmů, aneb

co je třeba znát a porozumět tomu:

- **Uspořádaný výběr**: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ vznikne z původního výběru $X_1, X_2, \ldots, X_n$ upořádáním podle velikosti pozorovaných hodnot $x_1, x_2, \ldots, x_n$.
- **Pořadová statistika**: $X_{(k)}$ je náhodná veličina X_m, která je k-tá v pořadí podle velikosti pozorovaných hodnot $x_1, x_2, \ldots, x_n$.
- **Pořadí** m-tého pozorování veličiny X_m ve výběru: pokud $X_m = X_{(k)} \quad => \quad R_m = k.$
- $X_{(1)}$ se nazývá (**výběrové) minimum**, $X_{(n)}$ je (**výběrové) maximum**
- **medián** je prostřední hodnota ve výběru: je-li n liché, je roven $X_{([n/2]+1)}$ pro n sudé je roven $(X_{(n/2)} + X_{(n/2+1)})/2$
- **dolní kvartil**: $X_{([n/4]+1)}$ resp. $(X_{(n/4)} + X_{(n/4+1)})/2$
- **horní kvartil**: $X_{([3n/4]+1)}$ resp. $(X_{(3n/4-1)} + X_{(3n/4)})/2$
- **Výběrový modus** je nejčastější hodnota, která se vyskytuje v realizaci výběru. Tato hodnota nemusí existovat.

Malý slovníček statistických pojmů, aneb

co je třeba znát a porozumět tomu:

- **Výběrový průměr** nahrazuje neznámou střední hodnotu veličiny X:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Výběrový rozptyl** je charakteristika odpovídající rozptylu náhodné veličiny X

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- **Výběrová směrodatná odchylka** s je druhou odmocninou z výběrového roptylu
- **Výběrový index šikmosti** je výběrovou variantou indexu šikmosti a je mírou symetrie pozorované veličiny X

$$Skew(X) = \frac{m_3(X)}{m_2^{3/2}(X)}$$

- **Výběrový index špičatosti** je výběrovou variantou indexu špičatosti a je mírou soustředění hodnot pozorované veličiny X kolem průměru.

$$Kurt(X) = \frac{m_4(X)}{m_2^2(X)} - 3$$

Malý slovníček statistických pojmů, aneb

co je třeba znát a porozumět tomu:

- **Třídní intervaly** rozdělují maximální rozsah pozorovaných hodnot náhodné veličiny (od minima do maxima) na k stejných dílů.
- **(prostá absolutní) četnost *i*-té třídy** je počet pozorování náhodné veličiny X v i-té třídě, $i = 1, \ldots, k$.
- **(prostá) relativní četnost *i*-té třídy** je poměr počtu pozorování náhodné veličiny X v i-té třídě ku rozsahu výběru n, $i = 1, \ldots, k$.
- **kumulativní (absolutní) četnost *i*-té třídy** je počet pozorování náhodné veličiny X od minima až do i-té třídy včetně, $i = 1, \ldots, k$.
- **kumulativní relativní četnost *i*-té třídy** je součet relativních četností pozorování náhodné veličiny X až do i-té třídy včetně, $i = 1, \ldots, k$.
- **Histogram četností** je grafické zobrazení četností ve formě sloupkového grafu. Relativní četnosti lze zobrazovat i ve formě kruhového (koláčového) grafu. Existuje celá řada variant.

Malý slovníček statistických pojmů, aneb

co je třeba znát a porozumět tomu:

- **Krabicový diagram** je grafické zobrazení rozdělení pozorovaných hodnot pomocí pěti (Tukey's) charakteristik: minima, dolního kvartilu, mediánu, horního kvartilu a maxima.
- **Empirická distribuční funkce** je grafické zobrazení realizace výběru formou grafu po částech konstantní funkce

vycházíme z uspořádaného výběru: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. Potom

$$F_n(x_{(i)}) = \frac{i}{n} \quad \text{a tedy} \quad F_n(x) = \frac{\max\{k : X_{(k)} \leq x\}}{n}, \quad x \in \mathbf{R}$$

Pohádka o Zbohatlíkově

V jedné malé rozvinuté zemi, na kraji Evropské unie, přišel mladý podnikatel do realitní kanceláře a řekl: “*Chtěl bych pozemek na venkově, s lesem, loukami, ne příliš daleko od města, v pěkné krajině, za kterou by se člověk nemusel stydět. Samozřejmě že cenově výhodný*. ”

Zprostředkovatel po chvíli přemýšlení odpověděl: “*Zbohatlíkov na Malém Nadutci! To je přesně to, co hledáte*.”

Zájemce ohrnul nos: “Nedávno jsem tím místem projížděl a vypadá dost chudě.”

Zprostředkovatel zavrtěl hlavou: “To se musíte mýlit! Zde mám nejnovější údaje:”

Helmut Swoboda:
Moderní statistika, Nakl. Svoboda, Praha 1977

Pohádka o Zbohatlíkově

- Zprostředkovatel tvrdí:

  ***Průměrný roční příjem ve Zbohatlíkově činí* 82.320 *tolarů*.**

- Kupec zašel za známým ředitelem banky:

  roční příjem více než poloviny obyvatel je **29.000** *tolarů a více*.

- To je podivné! Co řekne okresní úřad?:

  Dosti chudé místo, prostřední příjem je kolem **29.000** *tolarů.*

- Vrátil se k řediteli banky pro nové informace:

  Nejsilněji zastoupená příjmová kategorie je od **12.000** *do* **24.000** *tolarů*

  Nejčetnější příjem je poměrně přesně **18.000** *tolarů.*

- Rozhněvaný kupec jede za učitelem Počtářem, kam ho poslali. Ten tvrdí, že situace je neutěšená:

  Dvě třetiny rodin mají méně než **30.000** *tolarů.*

  Příjem na hlavu není u většiny lidí ani **7.500** *tolarů ročně.*

  80% obyvatel má ročně méně než **25.000** *tolarů*

Kdo z nich lže?

Pohádka o Zbohatlíkově

Údaje o ročním příjmu 25 rodin ze Zbohatlíkova, n je počet členů domácnosti:

roční příjem	n	roční příjem	n	roční příjem	n	roční příjem	n
1,200.000	3	60.000	1	45.000	2	29.000	3
150.000	5	51.000	3	42.000	2	26.000	4
86.000	4	49.000	4	38.000	4	24.000	4
37.000	3	20.000	7	14.000	1	18.000	4
35.000	5	18.000	3	13.000	4	16.000	3
32.000	3	18.000	8	11.000	1	16.000	2
						10.000	2

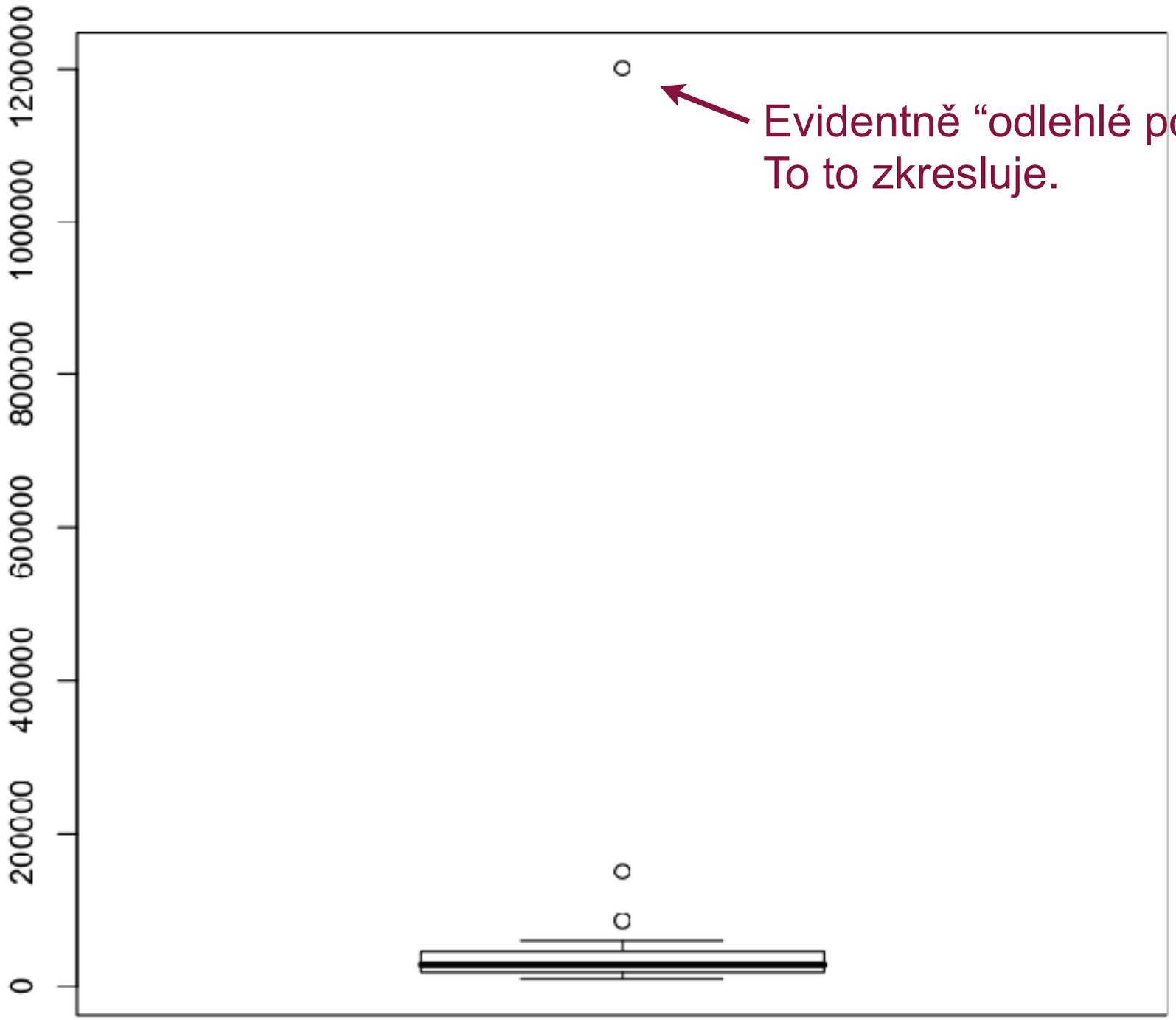
Zkusíme "stem&leaf" diagram:

```
0 | 00000000000000000001112
1 | 2
```

Nic moc!

Pohádka o Zbohatlíkově

... a co "Box&Whiskers" diagram?

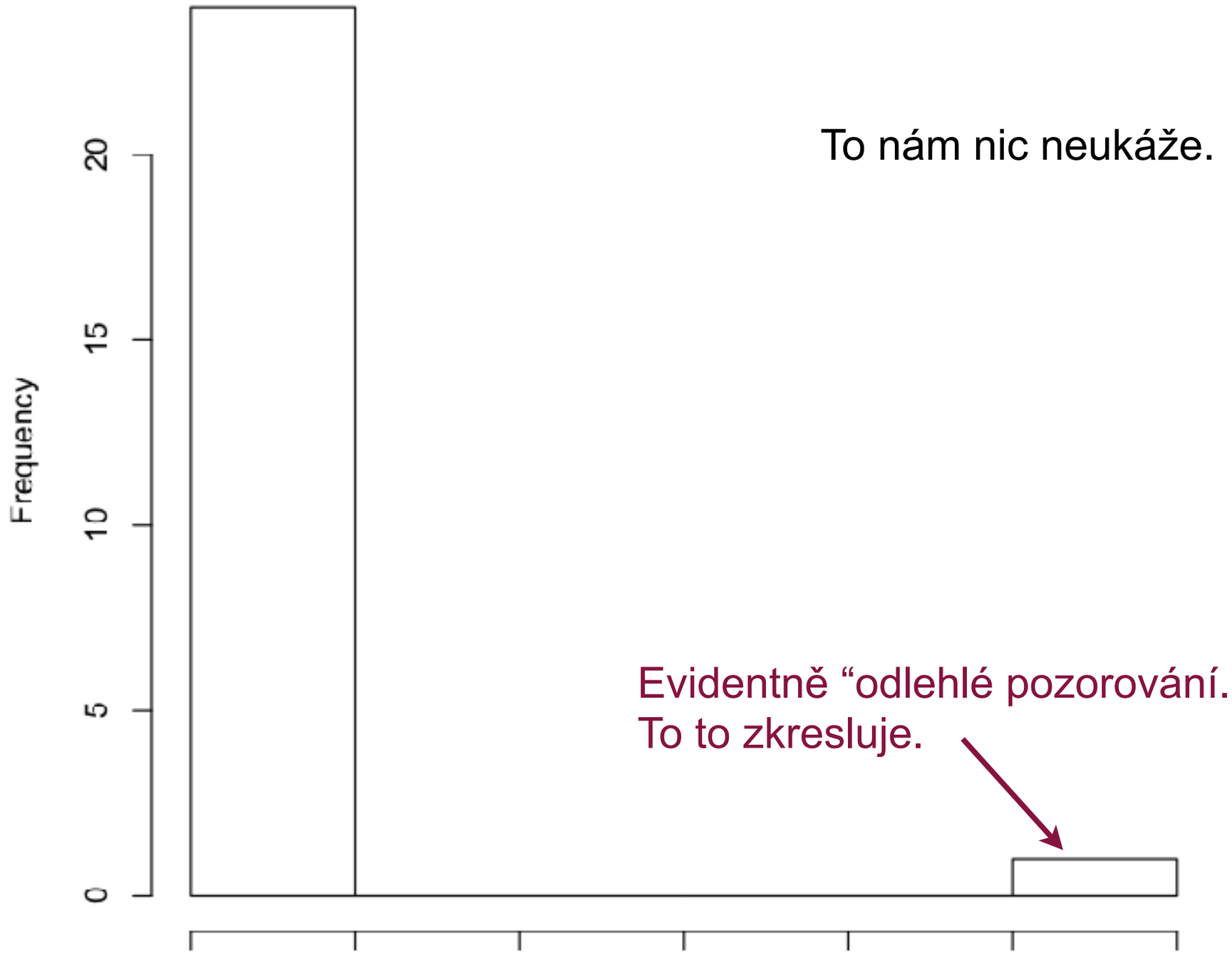


Evidentně "odlehlé pozorování.
To to zkresluje.

Nicméně, medián
je opravdu 29.000

Pohádka o Zbohatlíkově

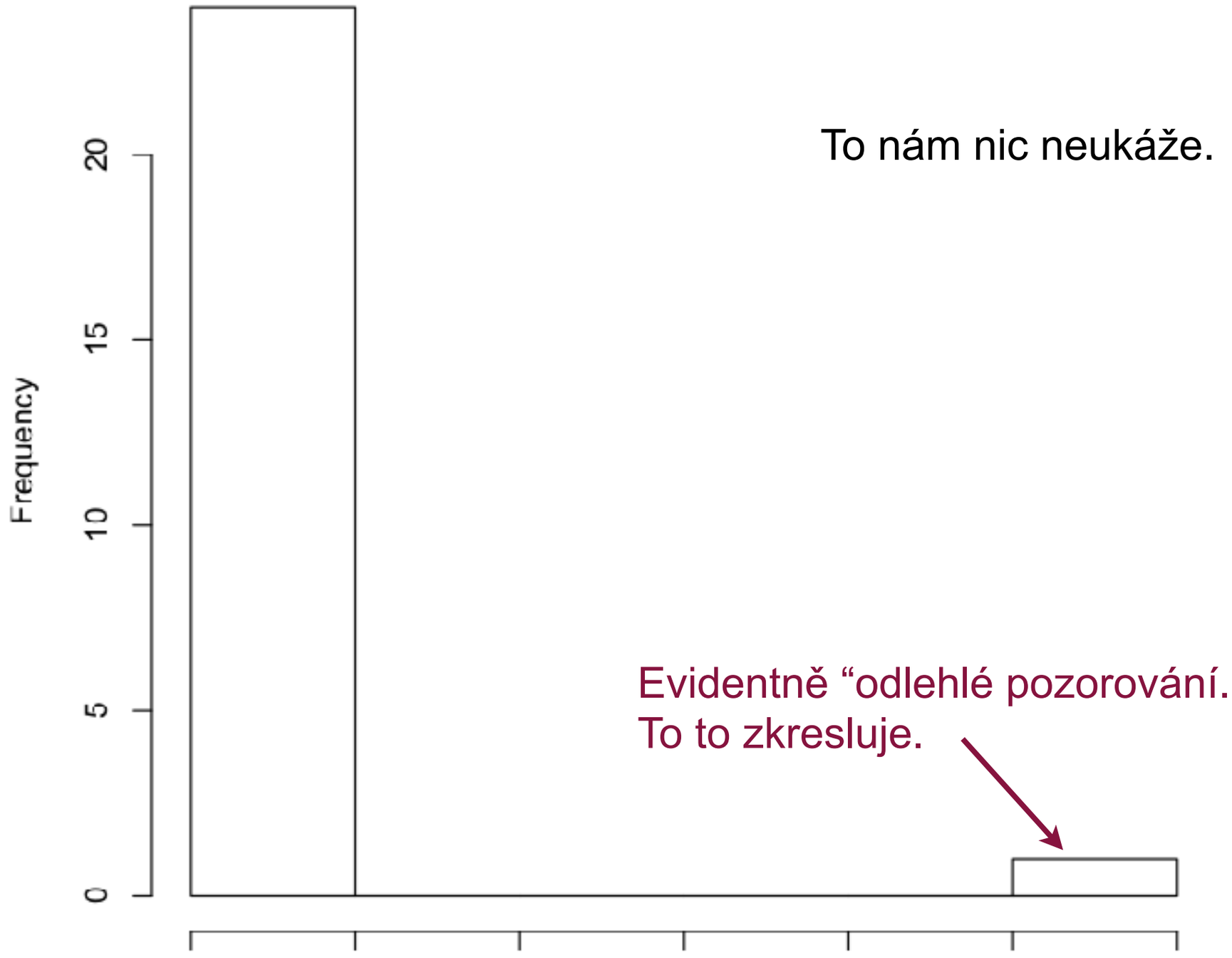
stejně dopadne i histogram:



To nám nic neukáže.

Pohádka o Zbohatlíkově

stejně dopadne i histogram:



To nám nic neukáže.

Pohádka o Zbohatlíkově

Údaje o ročním příjmu 25 rodin ze Zbohatlíkova, n je počet členů domácnosti:

roční příjem	n	roční příjem	n	roční příjem	n		
1,200.000	3	60.000	1	45.000	2	29.000	3
150.000	5	51.000	3	42.000	2	26.000	4
86.000	4	49.000	4	38.000	4	24.000	4
37.000	3	20.000	7	14.000	1	18.000	4
35.000	5	18.000	3	13.000	4	16.000	3
32.000	3	18.000	8	11.000	1	16.000	2
						10.000	2

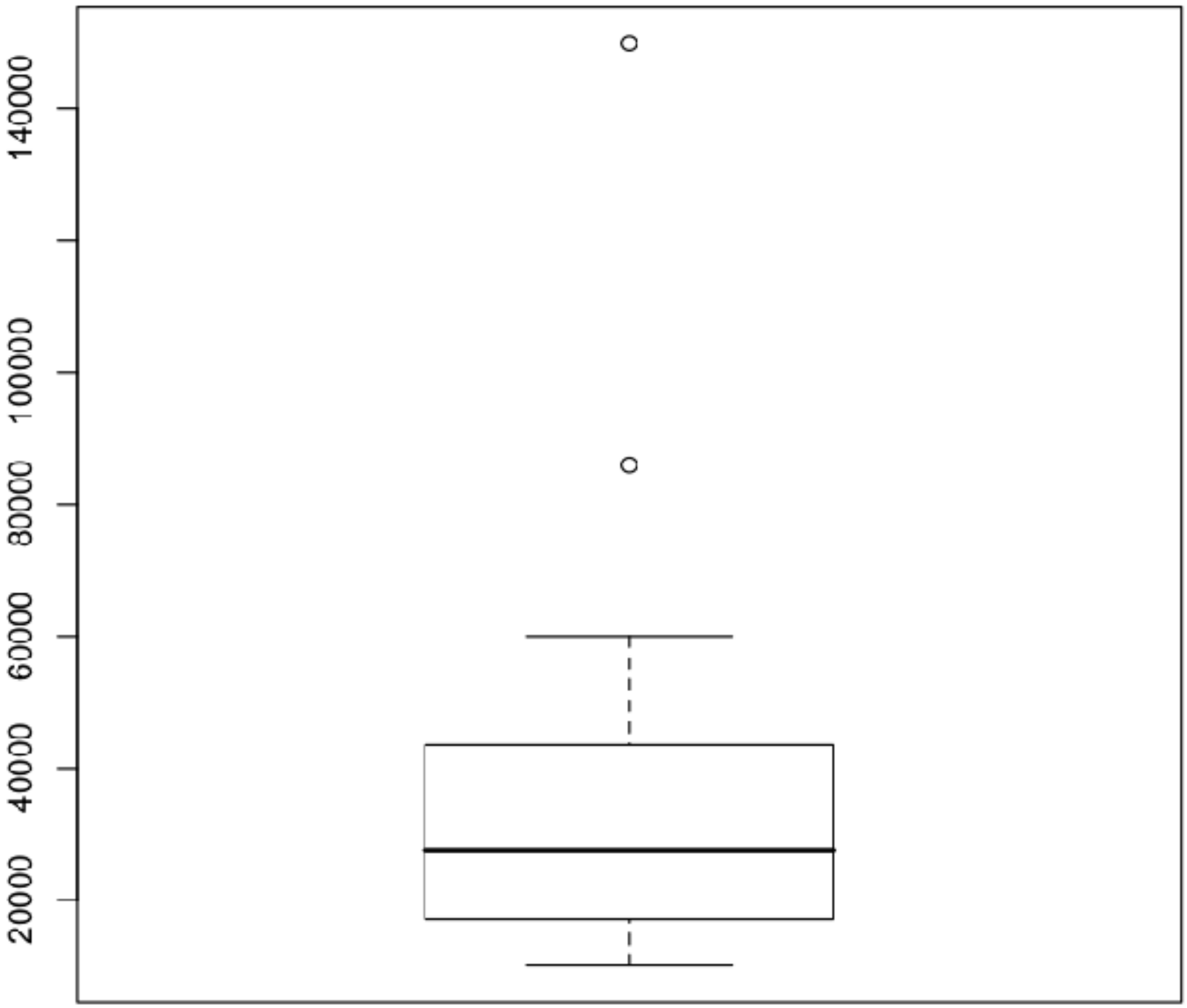
Odstraníme na chvíli extrémní (odlehlou) hodnotu :

```
0 | 11112222223334444
0 | 55569
1 |
1 | 5
```

To už je trochu lepší!

Pohádka o Zbohatlíkově

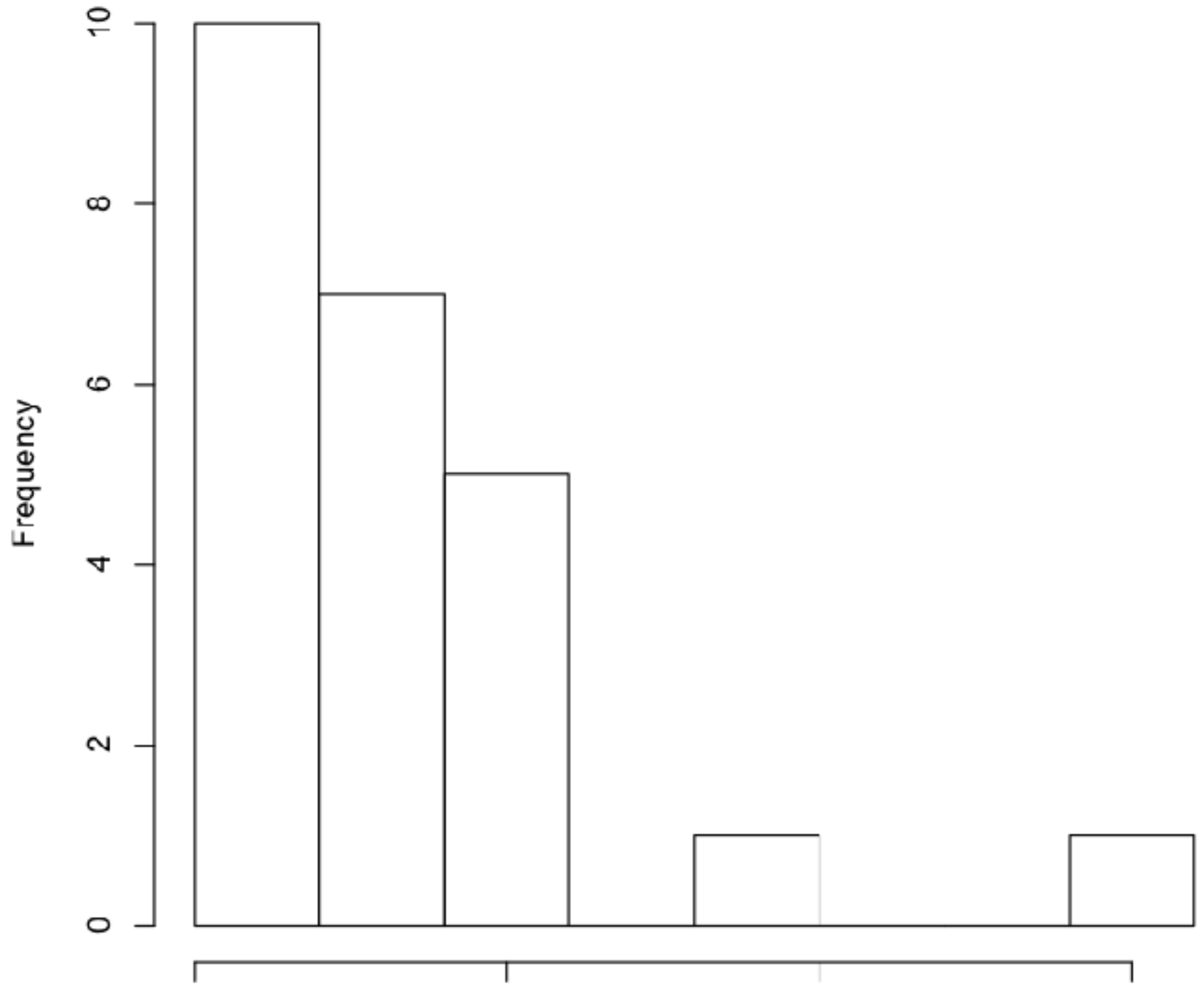
Odstraníme na chvíli extrémní (odlehlou) hodnotu :



X_{med} = 29.000

Pohádka o Zbohatlíkově

Odstraníme na chvíli extrémní (odlehlou) hodnotu :



Pohádka o Zbohatlíkově

Údaje o ročním příjmu 25 rodin ze Zbohatlíkova, n je počet členů domácnosti:

roční příjem	n	roční příjem	n	roční příjem	n		
1,200.000	3	60.000	1	45.000	2	29.000	3
150.000	5	51.000	3	42.000	2	26.000	4
86.000	4	49.000	4	38.000	4	24.000	4
37.000	3	20.000	7	14.000	1	18.000	4
35.000	5	18.000	3	13.000	4	16.000	3
32.000	3	18.000	8	11.000	1	16.000	2
						10.000	2

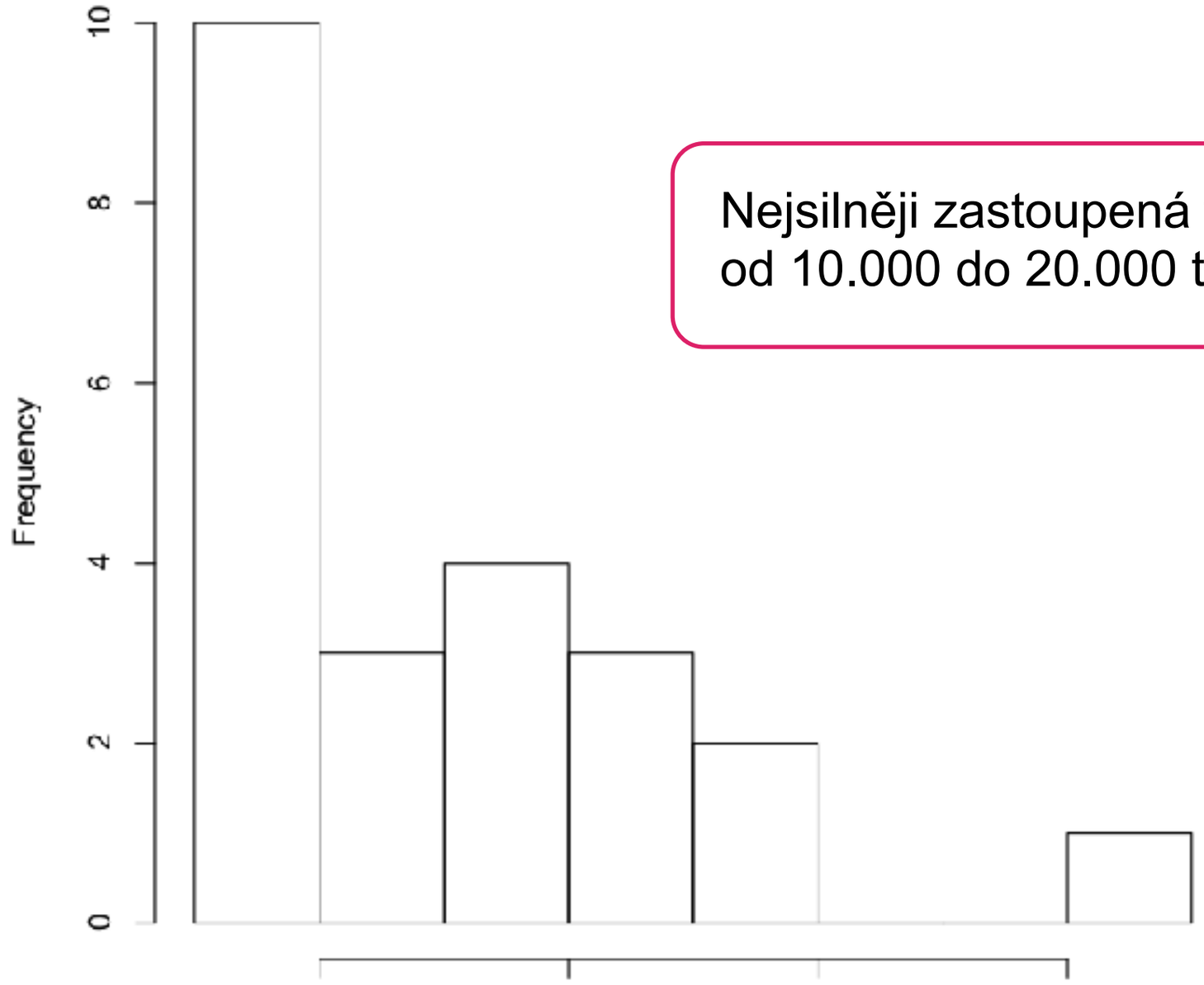
Odstraníme na chvíli dvě extrémní (odlehlé) hodnoty :

```
0 | 013466888
2 | 04692578
4 | 2591
6 | 0
8 | 6
```

Nejčetnější hodnota
je 18.000 tolarů

Pohádka o Zbohatlíkově

Odstraníme na chvíli dvě extrémní (odlehlé) hodnoty:



Nejsilněji zastoupená třída je od 10.000 do 20.000 tolarů

Pohádka o Zbohatlíkově

Příjmy na hlavu (83):

400.000, 400.000, 400.000, 30.000, 30.000, 30.000, 30.000, 30.000, 21.500, 1.500,
21.500, 21.500, 12.333, 12.333, 12.333, 7.000, 7.000, 7.000, 7.000, 7.000,
10.666, 10.666, 10.666, 9.666, 9.666, 9.666, 6.500, 6.500, 6.500, 6.500,
6.000, 6.000, 6.000, 6.000, 60.000, 17.000, 12.250, 12.250, 12.250,12.250,
2.857, 2.857, 2.857, 2.857, 2.857, 2.857, 2.857, 6.000, 6.000, 6.000,
2.250, 2.250, 2.250, 2.250, 2.250, 2.250, 2.250, 2.250, 4.500, 4.500,
4.500, 4.500, 5.333, 5.333, 5.333, 8.000, 8.000, 22.500, 22.500,21.000,
21.000, 9.500, 9.500, 9.500, 9.500, 14.000, 3.250, 3.250, 3.250, 3.250,
11.000, 5.000, 5.000

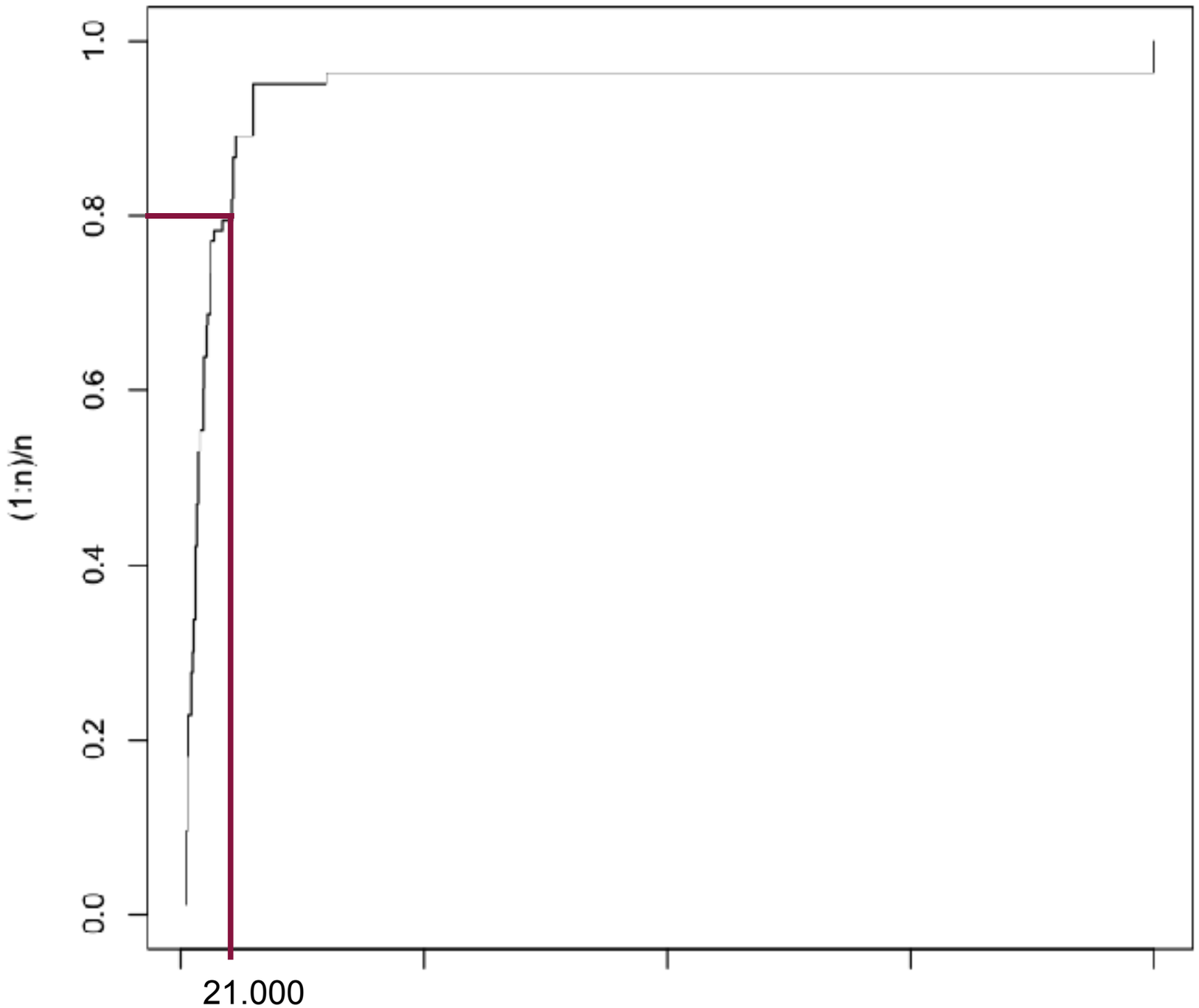
Uspořádané příjmy na hlavu: 80% je 66,4 lidí

2.250	2.250	2.250	2.250	2.250	2.250	2.250	2.250	2.857	2.857
2.857	2.857	2.857	2.857	2.857	3.250	3.250	3.250	3.250	4.500
4.500	4.500	4.500	5.000	5.000	5.333	5.333	5.333	6.000	6.000
6.000	6.000	6.000	6.000	6.000	6.500	6.500	6.500	6.500	7.000
7.000	7.000	7.000	7.000	8.000	8.000	9.500	9.500	9.500	9.500
9.666	9.666	9.666	10.666	10.666	10.666	11.000	12.250	12.250	12.250
12.250	12.333	12.333	12.333	14.000	17.000	21.000	21.000	21.500	21.500
21.500	21.500	22.500	22.500	30.000	30.000	30.000	30.000	30.000	60.000
400.000	400.000	400.000							

80% lidí má menší roční příjem než 20.000 tolarů

Pohádka o Zbohatlíkově

Empirická distribuční funkce:



Pohádka o Zbohatlíkově

Lhal tedy zprostředkovatel?

Nelhal, neboť:

Aritmetický průměr $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{25}\sum_{i=1}^{25} X_i = 82.320$$

V případě příjmů je však lépe použít:

Geometrický průměr $$\hat{X} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}} = 32.730$$

(Neboť mzdy mají zpravidla silně sešikmené rozdělení)