Pravděpodobnost a matematická statistika

Doc. RNDr. Gejza Dohnal, CSc.

dohnal@nipax.cz

Pravděpodobnost a matematická statistika
2010

1.týden (20.09.-24.09.) Data, typy dat, variabilita, frekvenční analýza (histogramy, četnosti absolutní, relativní, prosté, kumulativní), základní statistické charakteristiky (průměr, výběr.rozptyl, minimum, maximum, medián, kvartily, boxplot), sešikmenná rozdělení (vzájemná poloha mediánu a střední hodnoty), chvosty, kvantily

2. týden (27.09.-01.10.) Princip statistické indukce, výběr, vlastnosti výběru, experiment. Náhodná veličina, rozdělení pravděpodobnosti a jeho souvislost s histogramem. Pravděpodobnost, pravidla pro počítání s pravděpodobností, podmíněná pravděpodobnost, závislost náhodných veličin.

3.týden (04.10.-08.10.) Pohádka o Zbohatlíkově

4.týden (11.10.-15.10.) Rozdělení chyb měření - normální rozdělení a počítání s ním. Odhady parametrů normálního rozdělení. Intervaly spolehlivosti pro normální data. Jednovýběrové testy o střední hodnotě

5.týden (18.10.-24.10.) Výběrový poměr jako odhad pravděpodobnosti sledovaného jevu. Alternativní rozdělení, binomické rozdělení. Intervalový odhad výběrového poměru. Výběry s vracením a bez vracení (binomické a hypergeometrické rozdělení)

6.týden (25.10.-29.10.) odpadá

7.týden (01.11.-05.11.) Poruchy v čase (Poissonův proces). Poissonovo rozdělení, exponenciální rozdělení, jeho výhody a nevýhody, modelování doby do poruchy pomocí Weibullova rozdělení, lognormálního rozdělení, případně useknuté normální rozdělení.

8.týden (08.11.-12.11.) Testy dobré shody, Q-Q graf (pouze vysvětlení), testy normality. Některé neparametrické testy

9.týden (15.11.-19.11.) Dvě náhodné veličiny - srovnání dvou výběrů (dvouvýběrové testy)

10. týden (22.11.-26.11.) Dvě náhodné veličiny. Dvourozměrné četnosti jako odhad dvourozměrného rozdělení, frekvenční tabulka. Marginální rozdělení (vše pouze diskrétně s tabulkou)

11. týden (29.11.-03.12.) Závislost náhodných veličin, míry závislosti (kovariance, korelace), test významnosti korelačního koeficientu

12. týden (06.12.-10.12.) Regrese, lineární regresní model (přímková, kvadratická, polynomická regrese), analýza reziduí, pásy spolehlivosti

13. týden (13.12.-17.12.) Více výběrů, jednoduché třídění, ANOVA.

14. týden (20.12.-22.12.) Rezerva, opakování, testy normality (náhrada za 28.10.)

Pravděpodobnost a matematická statistika

Semestrální práce:

1. Popis experimentu
   a. Co chci zjistit?
   b. Jak to zjistím?
   c. Jak pořídím data?

Formulace v jazyce experta daného oboru (technika)

Formulace v jazyce statistika (popis náhodných veličin, stratistických metod

2. Analýza dat
   a) Základní charakteristiky
   b) Grafické zobrazení dat

5-Tukey, průměr, výběr. rozptyl, šikmost, špičatost

histogram, kruhový graf, x-y graf, krabicový graf, empirická d.f.

3. Vyhodnocení experimentu
   1. Popis metody
      a. Výpočet
      b. Závěry (interpretace výsledku)

(t-test, chí-kvagrát test, regrese, ANOVA)

Formulace v jazyce statistika

4. Celkové shrnutí.

Formulace v jazyce experta

Pravděpodobnost a matematická statistika

Semestrální práce:

1. Popis experimentu
   a. Co chci zjistit?
   b. Jak to zjistím?
   c. Jak pořídím data?

Formulace v jazyce experta daného oboru (technika)

Formulace v jazyce statistika (popis náhodných veličin, stratistických metod

2. Analýza dat
   a) Základní charakteristiky
   b) Grafické zobrazení dat

5-Tukey, průměr, výběr. rozptyl, šikmost, špičatost

histogram, kruhový graf, x-y graf, krabicový graf, empirická d.f.

3. Vyhodnocení experimentu
   1. Popis metody
      a. Výpočet
      b. Závěry (interpretace výsledku)

(t-test, chí-kvagrát test, regrese, ANOVA)

Formulace v jazyce statistika

4. Celkové shrnutí.

Formulace v jazyce experta

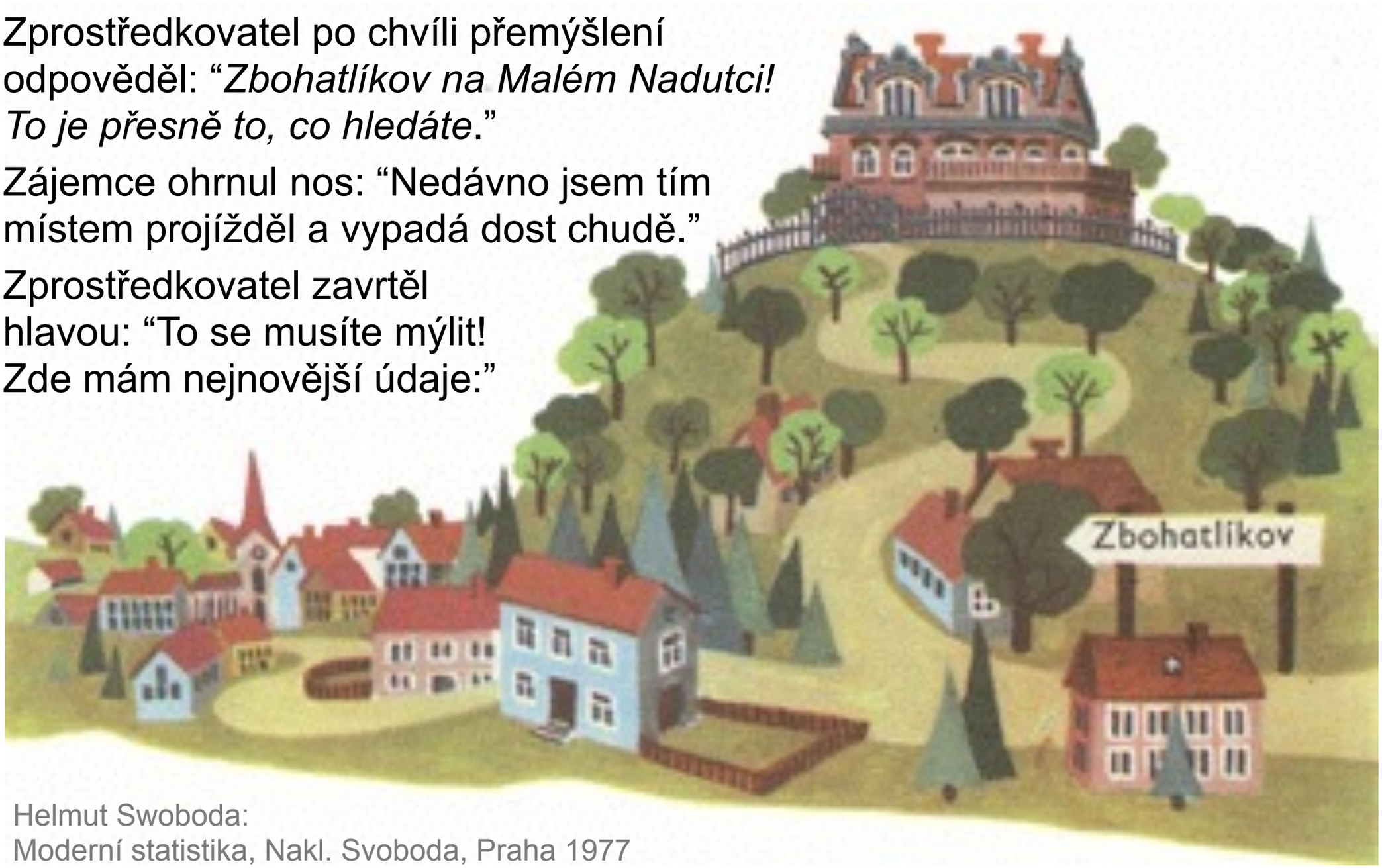
Pohádka o Zbohatlíkově

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, průměrný 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		
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

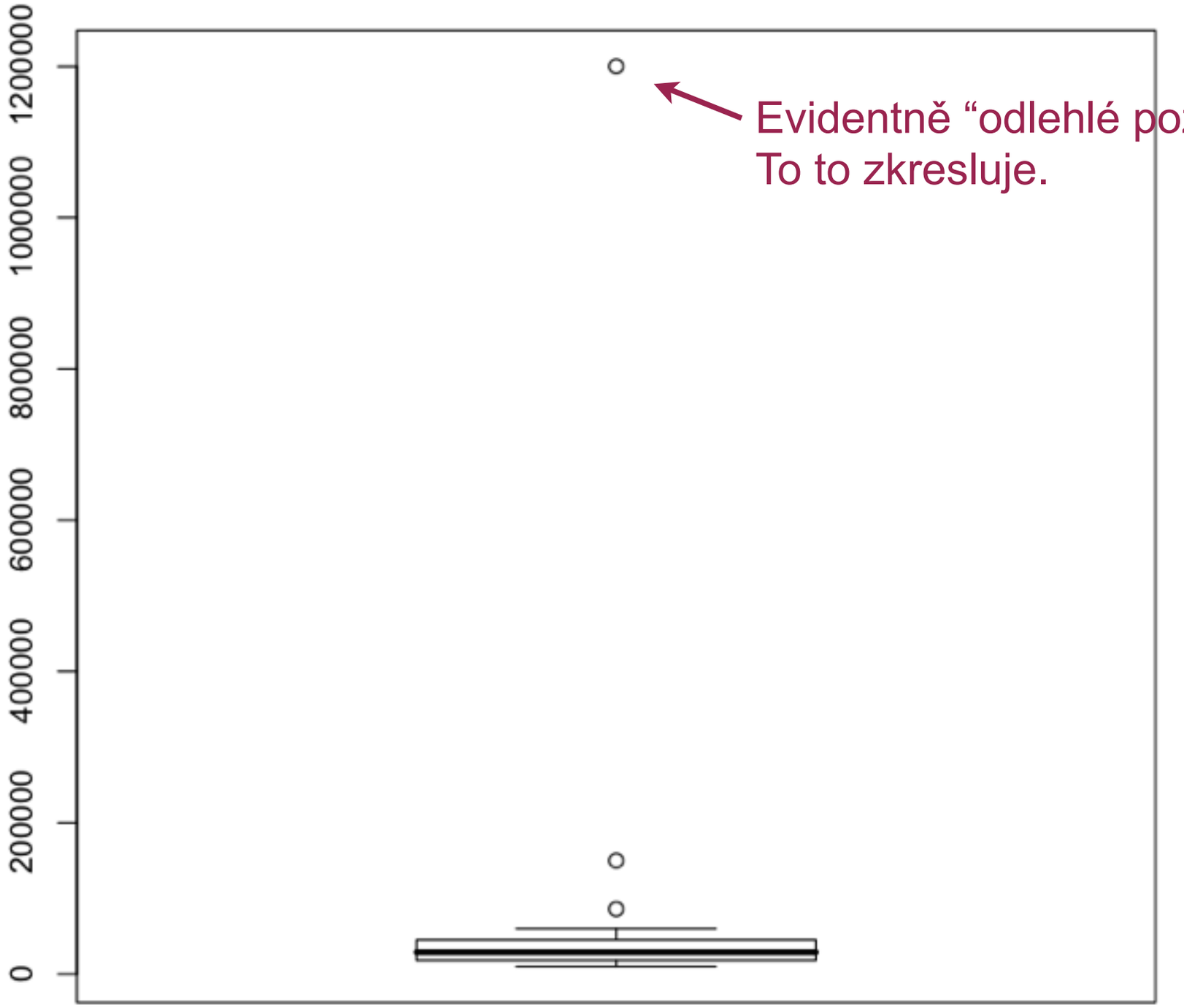
Zkusíme "stem&leaf" diagram:

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

Nic moc!

Pohádka o Zbohatlíkově

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

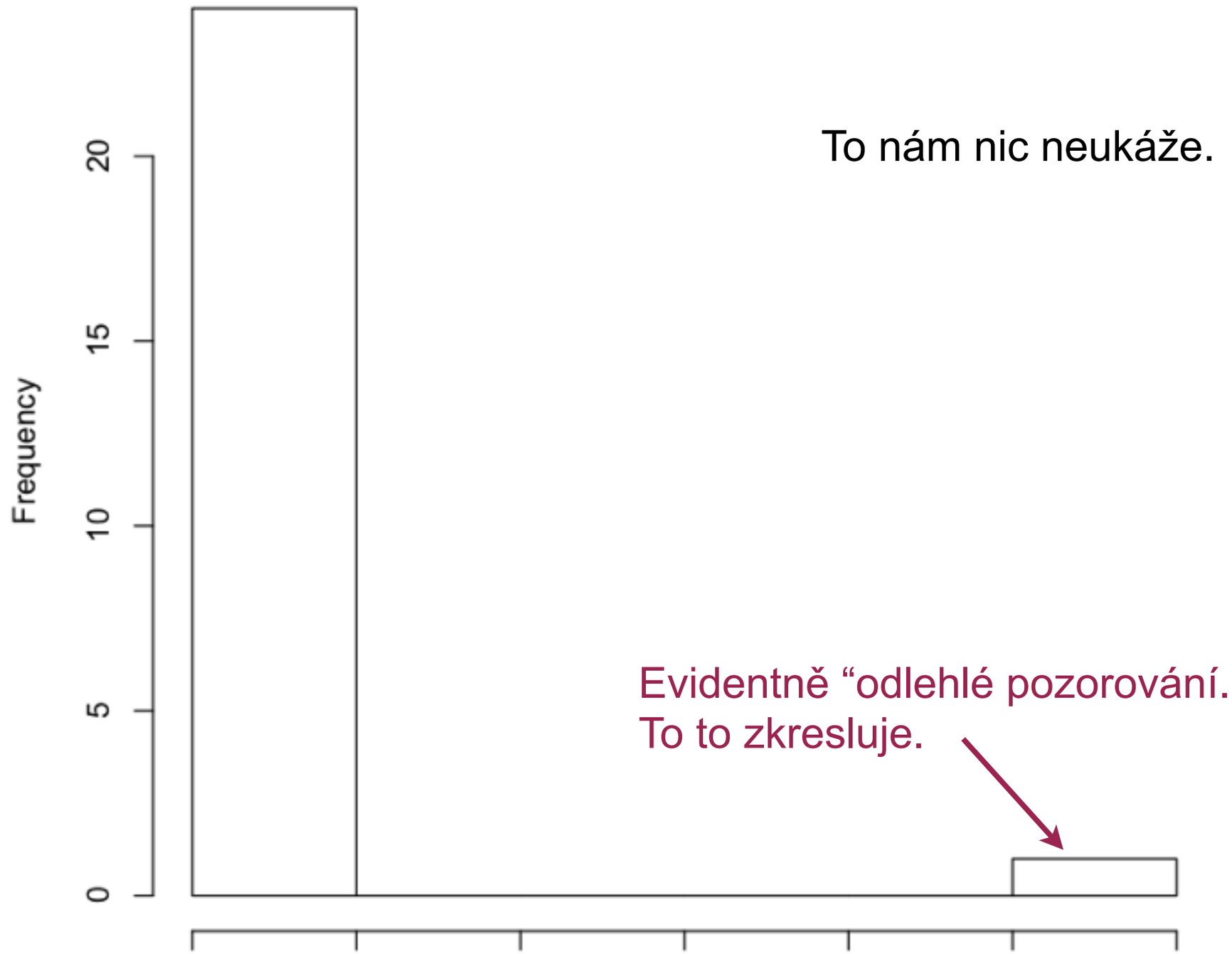


Evidentně "odlehlé pozorování.
To to zkresluje.

Nicméně, medián
je opravdu 29.000

Pohádka o Zbohatlíkově

stejně dopadne i histogram:



To nám nic neukáže.

Pohádka o Zbohatlíkově

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

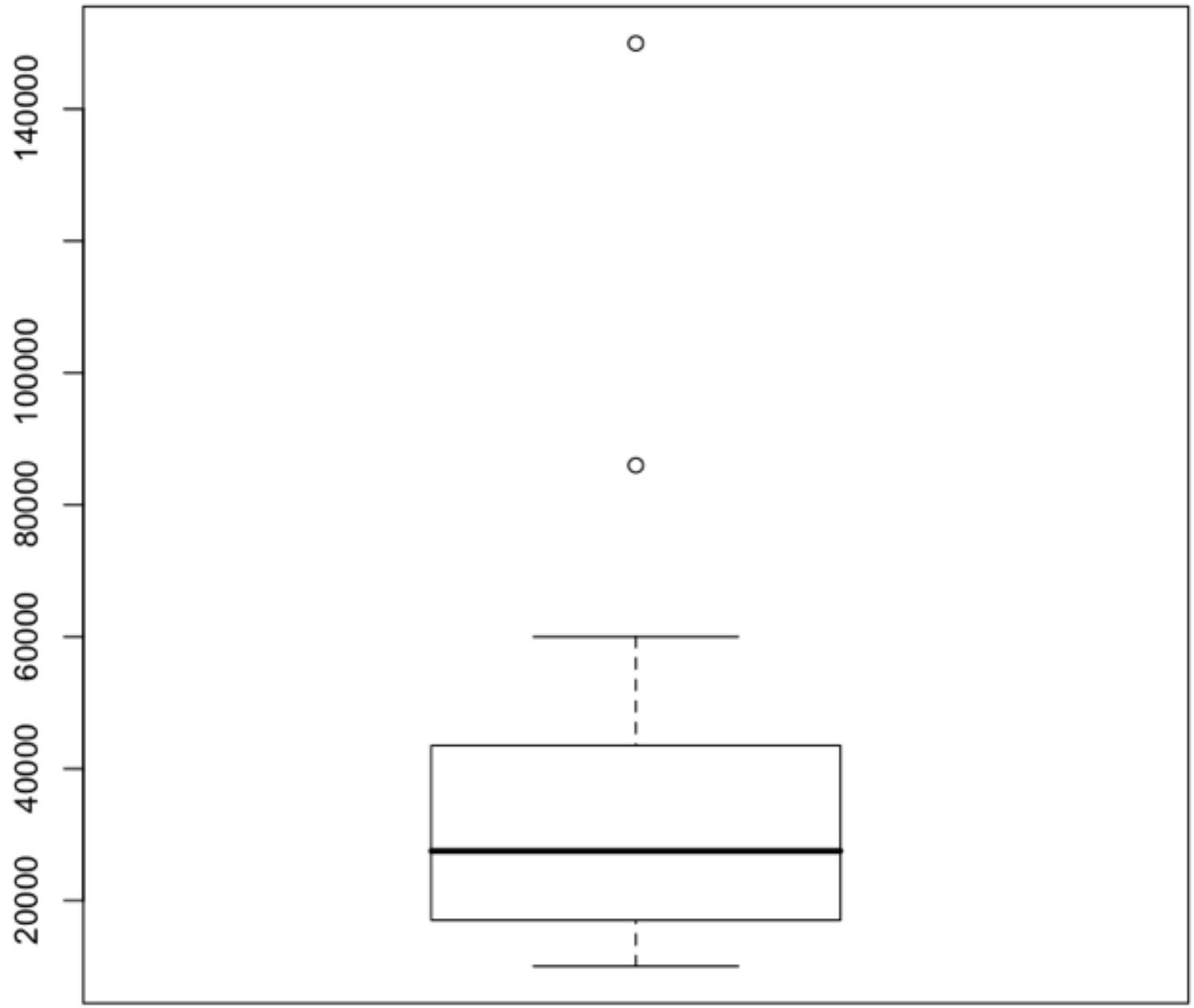
Odstraníme na chvíli extrémní (odlehlou) hodnotu :

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

To už je trochu lepší!

Pohádka o Zbohatlíkově

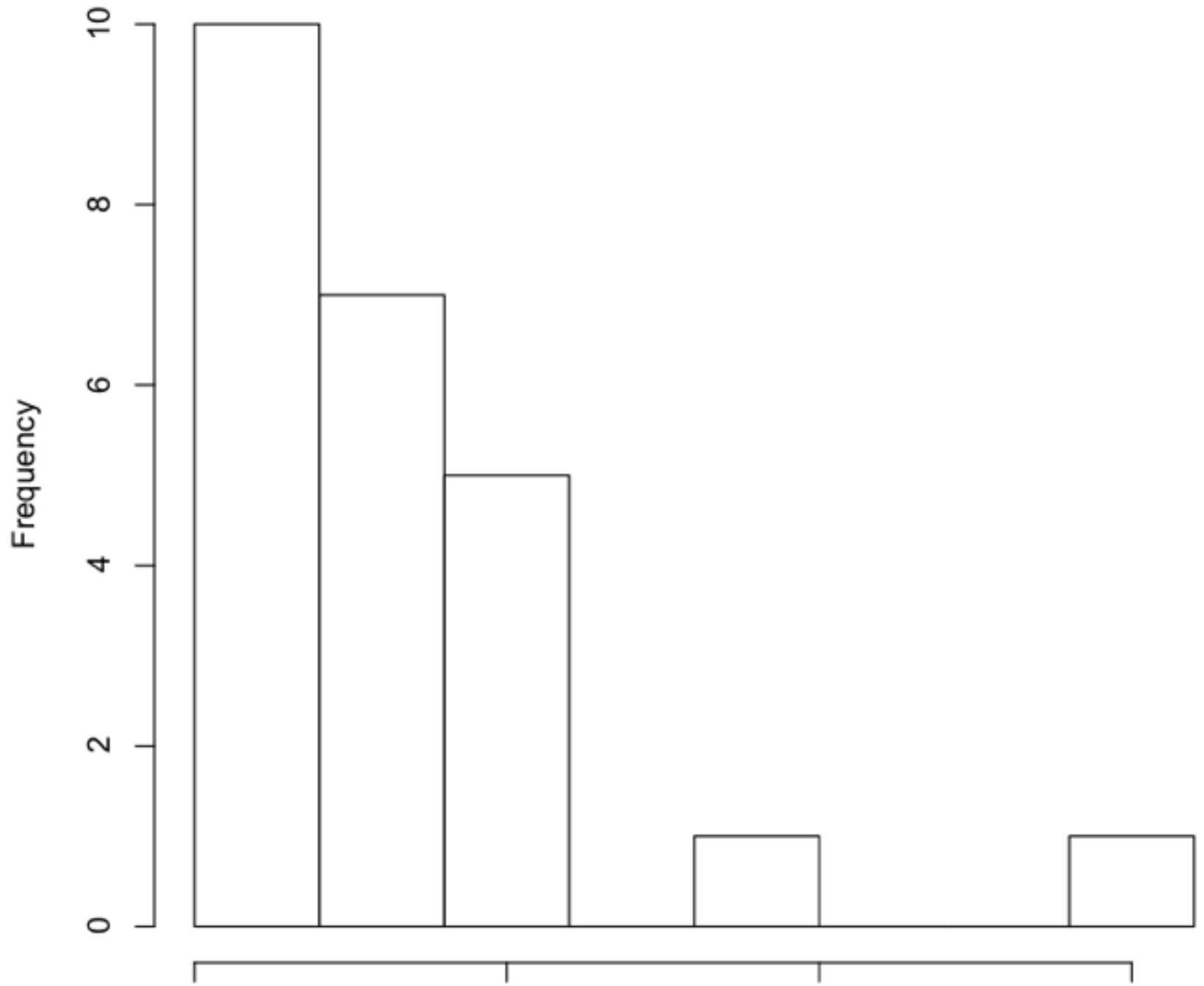
Odstraníme na chvíli extrémní (odlehlou) hodnotu :



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

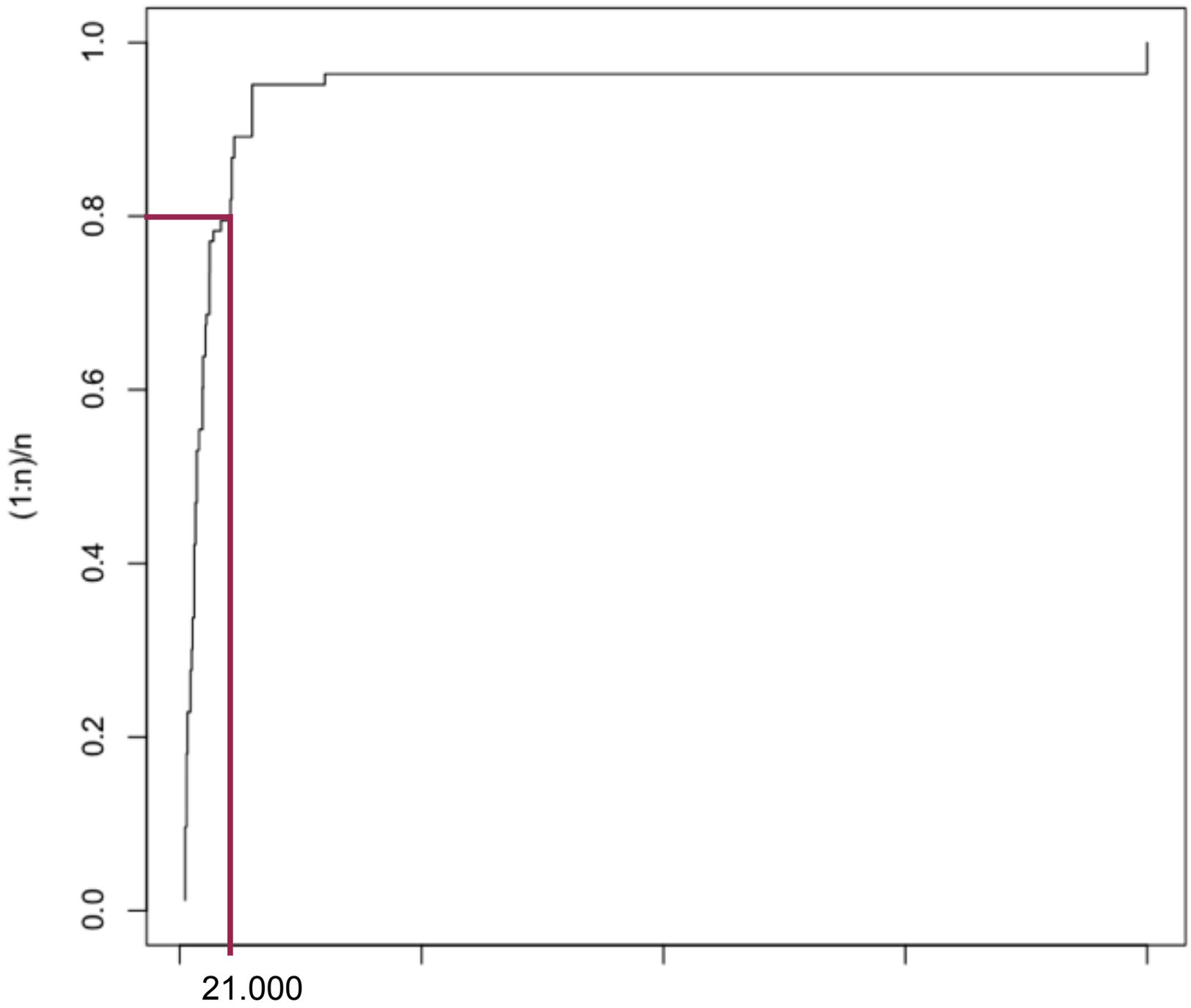
Uspořádané příjmy na hlavu: 80% je 66,4 lidí

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)