Statistika

(MD360P03Z, MD360P03U)
ak. rok 2007/2008

Karel Zvára

karel.zvara@mff.cuni.cz
http://www.karlin.mff.cuni.cz/∼zvara

16. října 2007

charakteristiky polohy v geografii/demografii

- často známe jen průměry v dílčích souborech a četnosti: průměry se použijí jako x_j^*, četnosti standardně
- příklad: věk nových profesorů a docentů UK 2002:
  41 profesorů, průměrný věk 51,1 ($n_1 = 41$, $x_1^* = 51{,}1$)
  77 docentů, průměrný věk 47,8 ($n_2 = 77$, $x_2^* = 47{,}8$)
  celkový průměr (**vážený průměr**):

[weighted.mean(c(51.1,47.8),c(41,77))]

$$\frac{41 \cdot 51{,}1 + 77 \cdot 47{,}8}{41 + 77} = 48{,}9$$

nikoliv [mean(c(51.1,47.8))]

$$\frac{51{,}1 + 47{,}8}{2} = 49{,}4$$

charakteristiky polohy v geografii/demografii (2)

- **geografický střed**
  - bod
  - průsečík průměrné zeměpisné šířky a průměrné zeměpisné délky; průměry vážené velikostí sledovaného jevu
- **geografický medián** – obdoba mediánu,
  - čára, která rozděluje geografické objekty do dvou disjunktních skupin
  - hodnocená vlastnost určí váhy objektů
  - uspořádání hodnocení znaků dáno zvolenou geografickou vlastností (např. zeměpisnou délkou)

míry nerovnoměrnosti

- **Giniho index** charakterizuje nerovnoměrnost rozdělení bohatství (příjmů, ...) jediným číslem $\boxed{G = \Delta/(2\bar{x})}$
- průměrný rozdíl v bohatství vztažený k dvojnásobku průměru
- mají-li všichni stejně ($x_{(1)} = \ldots = x_{(n)} > 0$), je nutně $\Delta = 0$ a tedy $G = 0$
- má-li jeden všechno, ostatní nic ($0 = x_{(1)} = \ldots = x_{(n-1)} < x_{(n)} = a$), pak je

$$\bar{x} = \frac{a}{n} \qquad \Delta = \frac{2(n-1)a}{n^2}$$

$$G = \frac{2(n-1)a}{n^2} \cdot \frac{n}{2a} = \frac{n-1}{n}$$

- Lorenzova křivka je jemnějším nástrojem

příklad: tolary (rozdělení příjmů)

jaké procento nejchudších získá **desetinu** celkového bohatství?
četnosti 99 osob (celkový měsíční příjem je 1687)

x_j	10	11	12	13	14	15	16	17	18	19	20
n_j	7	14	16	10	6	3	9	3	1	5	3

x_j	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1

sčítejme příjmy nejchudších, dokud nenasčítáme 10 % z 1687

$$(7 \cdot 10 + 8 \cdot 11)/1687 = 158/1687 = 0{,}0937 = 9{,}37\ \%$$
$$(7 \cdot 10 + 9 \cdot 11)/1687 = 169/1687 = 0{,}1002 = 10{,}02\ \%$$

u jaké části z 99 osob jsme sčítali příjmy?

$$(7 + 8)/99 = 15/99 = 0{,}152 = 15{,}2\ \%$$
$$(7 + 9)/99 = 16/99 = 0{,}162 = 16{,}2\ \%$$

příklad: tolary (rozdělení příjmů)

jaké procento nejchudších získá **polovinu** celkového bohatství?
četnosti (celkový měsíční příjem je 1687)

x_j	10	11	12	13	14	15	16	17	18	19	20
n_j	7	14	16	10	6	3	9	3	1	5	3

x_j	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1

sčítejme příjmy nejchudších, dokud nenasčítáme 50 % z 1687

$$(7 \cdot 10 + \ldots + 9 \cdot 16 + 17)/1687 = 836/1687 = 0{,}4956 = 49{,}56\ \%$$
$$(7 \cdot 10 + \ldots + 9 \cdot 16 + 2 \cdot 17)/1687 = 853/1687 = 0{,}5056 = 50{,}56\ \%$$

u jaké části z 99 osob jsme sčítali příjmy?

$$(7 + \ldots + 9 + 1)/99 = 66/99 = 0{,}6667 = 66{,}67\ \%$$
$$(7 + \ldots + 9 + 2)/99 = 67/99 = 0{,}6768 = 67{,}68\ \%$$

příklad: tolary (rozdělení příjmů)

jaké procento získají čtyři (tj. asi 4 %) nejbohatší resp. nejchudší?
četnosti (celkový měsíční příjem je 1687)

x_j	10	11	12	13	14	15	16	17	18	19	20
n_j	7	14	16	10	6	3	9	3	1	5	3

x_j	21	22	24	26	27	28	32	35	36	40	43	45	47
n_j	4	3	3	1	2	1	1	1	2	1	1	1	1

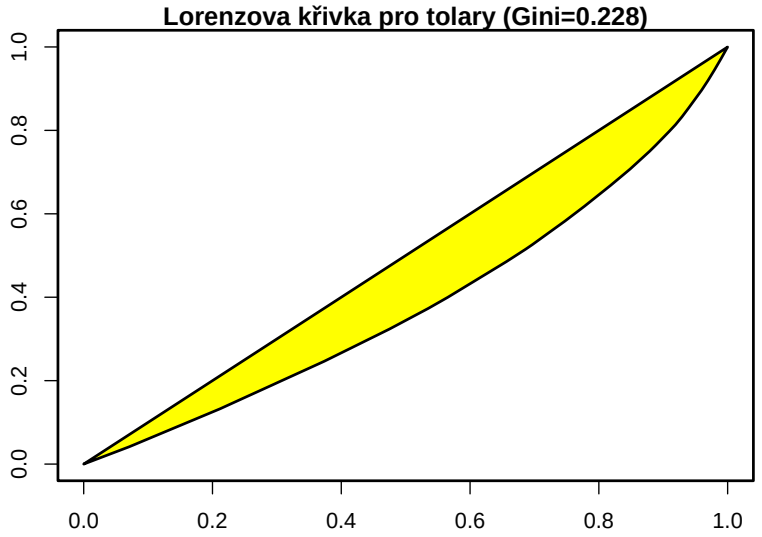
sečteme příjmy oněch čtyř nejbohatších

$$(47 + 45 + 43 + 40)/1687 = 175/1687 = 0{,}1037 = 10{,}37\ \%$$

čtyři nejbohatší tedy dostanou přes 10 % bohatství,
kdežto čtyři nejchudší dostanou

$$(4 \cdot 10)/1687 = 40/1687 = 0{,}0237 = 2{,}37\ \%$$

Lorenzova křivka (Tolary)



Lorenzova křivka

- variační řada: $0 < x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ [sort(x)]
- kumulativní součty pro $j = 0, 1, \ldots, n$ [cumsum(sort(x))]
  (kolik patří celkem j nejchudším)

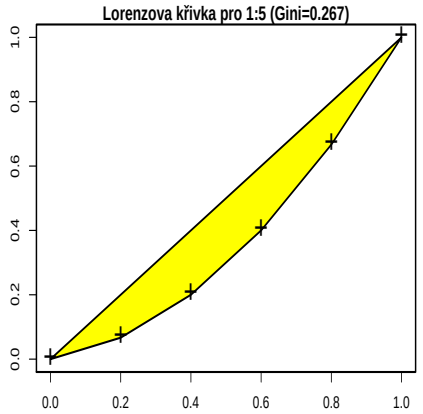
$$X_0 = 0 \qquad X_j = x_{(1)} + x_{(2)} + \ldots + x_{(j)} = \sum_{i=1}^{j} x_{(i)}$$

- úsečkami spojit body $[j/n; X_j/X_n], \quad 0 \leq j \leq n$
- zajímá nás plocha nad touto lomenou čarou a pod úhlopříčkou jednotkového čtverce
- plocha měří nerovnoměrnost rozdělení nějakého zdroje
- kdyby dostal každý stejně, bude velikost plochy nulová
- Giniho koeficient koncentrace je dvojnásobkem této plochy

umělý příklad

$x_1, \ldots, x_5$: 1, 2, 3, 4, 5

j	j/n	$x_{(j)}$	X_j	X_j/X_n
0	0,0		0	0,000
1	0,2	1	1	0,067
2	0,4	2	3	0,200
3	0,6	3	6	0,400
4	0,8	4	10	0,667
5	1,0	5	15	1,000



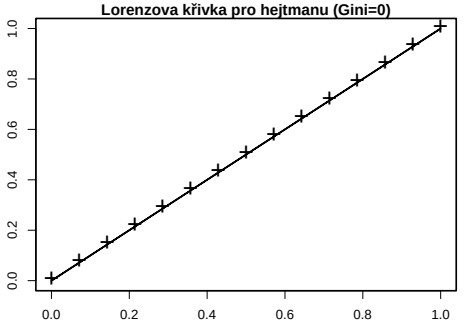
příklad - pokračování

výpočet Giniho koeficientu ($n = 5$)

$$\begin{aligned}
5^2 \cdot \Delta &= |1-1| + |1-2| + |1-3| + |1-4| + |1-5| \\
&\quad + |2-1| + |2-2| + |2-3| + |2-4| + |2-5| \\
&\quad + |3-1| + |3-2| + |3-3| + |3-4| + |3-5| \\
&\quad + |4-1| + |4-2| + |4-3| + |4-4| + |4-5| \\
&\quad + |5-1| + |5-2| + |5-3| + |5-4| + |5-5| \\
&= 10 + 7 + 6 + 7 + 10 \\
\Delta &= 40/25 = 1{,}6 \\
\bar{x} &= 3 \\
G &= \frac{1{,}6}{2 \cdot 3} = \frac{1{,}6}{6} = 0{,}267
\end{aligned}$$

Lorenzova křivka počet hejtmanů v krajích ČR

- v každém kraji je stejně hejtmanů, proto postupné součty rovnoměrně rostou, totéž platí pro X_j/X_n
- lomená čára Lorenzovy křivky přejde v úsečku a plocha zmizí
- průměrná diference je nulová (všechny rozdíly $|x_i - x_j|$ u počtu hejtmanů jsou nulové)

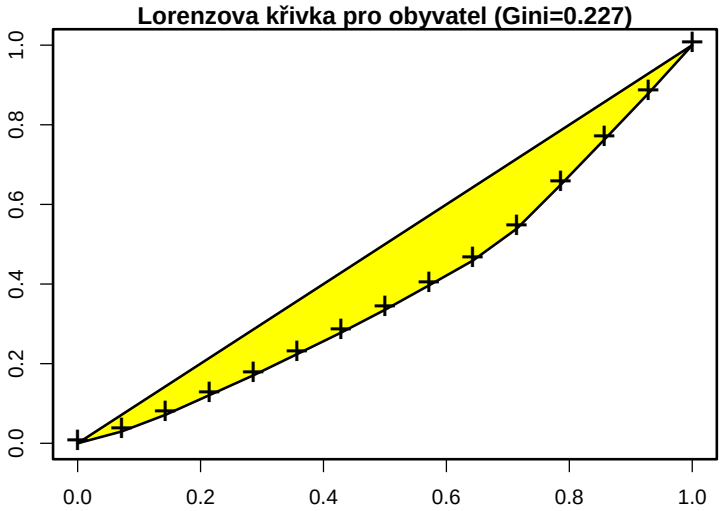


příklad: kraje ČR ke konci roku 2006

kraj i	obyvatel y_i	rozloha[km^2] n_i	hustota na km^2 x_i
Hlavní město Praha	1 188 126	496,1	2 395,0
Středočeský kraj	1 175 254	11 014,7	106,7
Jihočeský kraj	630 006	10 056,9	62,6
Plzeňský kraj	554 537	7 561,1	73,3
Karlovarský kraj	304 602	3 314,6	91,9
Ústecký kraj	823 265	5 334,5	154,3
Liberecký kraj	430 774	3 163,0	136,2
Královéhradecký kraj	549 643	4 758,4	115,5
Pardubický kraj	507 751	4 518,6	112,4
Vysočina	511 645	6 795,6	75,3
Jihomoravský kraj	1 132 563	7 196,3	157,4
Olomoucký kraj	639 894	5 266,8	121,5
Zlínský kraj	589 839	3 963,5	148,8
Moravskoslezský kraj	1 249 290	5 427,0	230,2
celkem	10 287 189	78 867,0	130,4

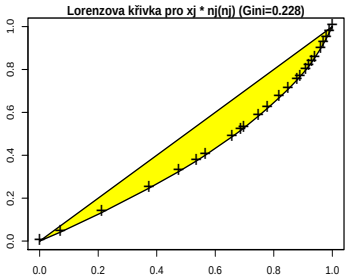
Jdi zpět | Jdi zpět ke grafu | Jdi zpět k teorii

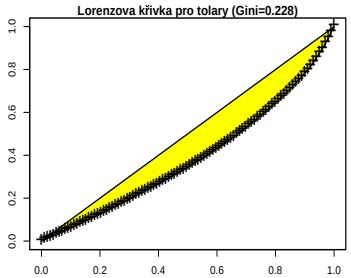
Lorenzova křivka (obyvatelé – kraje)



Lorenzova křivka pro tolary ještě jinak

- spousta hodnot proměnné tolary se opakuje, mohli jsme použít četnosti
- hodnota $x_{(j)}$ se vyskytuje $n_j$krát
  - o 10*7=70 tolarů se rozdělilo 7 „nejchudších“ osob
  - o 11*14=154 tolarů se rozdělilo 14 druhých „nejchudších“
  - ...
  - posledních 47 tolarů připadlo jedinému nejbohatšímu





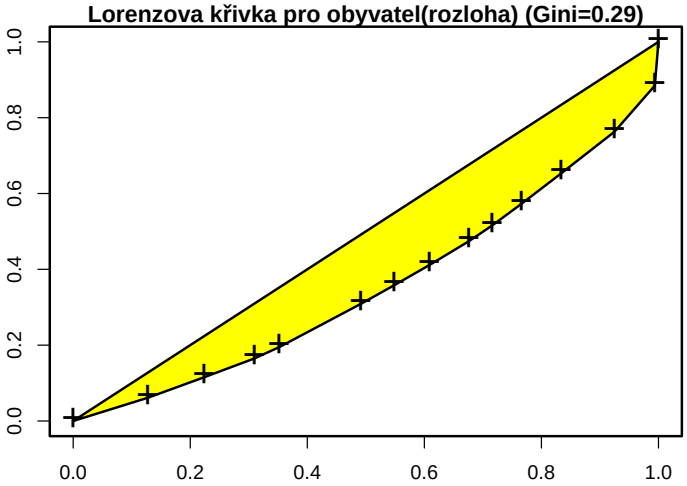
případ s vahami - příklad

- nerovnoměrnost rozmístění obyvatel v republice, ale údaje jen podle krajů
- potřebovali bychom pro každý jednotlivý km^2 znát počet obyvatel zde žijících
- známe jen počty obyvatel y_i v krajích a rozlohu krajů n_i
- předpokládáme rovnoměrné rozmístění uvnitř kraje, tedy $x_i = y_i/n_i$ obyvatel na každý km^2 v i-tém kraji
- každou takovou hustotu x_i musíme započítat $n_i$krát
- celková plocha $n = n_1 + \ldots + n_{14} (= N_{14})$
- průměrný počet obyvatel na km^2

$$\bar{x} = \frac{\sum_i n_i x_i}{\sum_i n_i} = \frac{\sum_i n_i (y_i/n_i)}{n} = \frac{\sum_i y_i}{n} = \bar{y}$$

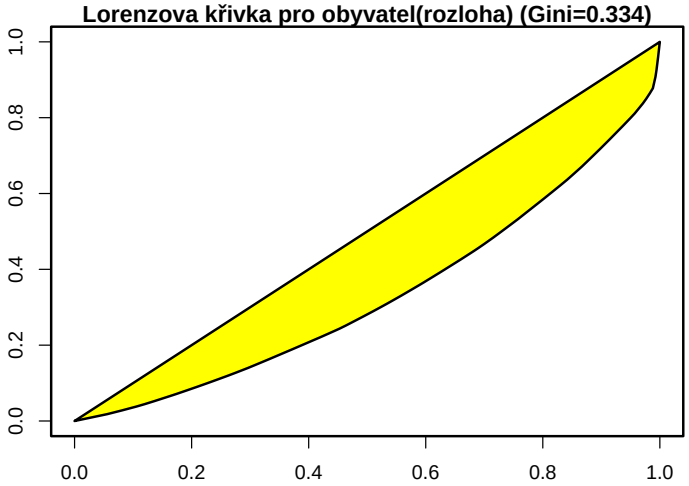
Jdi zpět k tabulce

Lorenzova křivka: obyvatelé krajů, vztaženo k rozloze



Jdi ke grafu okresů Jdi zpět k tabulce

Lorenzova křivka: obyvatelé okresů, vztaženo k rozloze



Jdi zpět ke grafu krajů

poznámky

- hrubší hodnocení (kraje, nikoliv okresy) znamená **menší** hodnotu Giniho indexu!
- nezáleží na zvolených jednotkách
- na vodorovné ose jde o umístění v řadě od nejchudších k nejbohatším
- označme kumulativní součty $N_i = \sum_{j=1}^{k} n_j$
- na svislé ose jde o podíl na bohatství
- označme kumulativní součty od nejchudších $Y_i = \sum_{j=1}^{i} y_j$
- pro zajímavost: $N_k = n$, rozděluje se bohatství Y_k
- ve všech případech je **pořadí** sčítanců dáno pořadím „hustot“ $x_i = \frac{y_i}{n_i}$ (např. obyvatel/rozloha)

výpočet v případě vah

- kumulativní součty $N_i = \sum_{j=1}^{k} n_j, \qquad Y_i = \sum_{j=1}^{k} Y_j$
- střední diference průměrných počtů obyvatel na km^2 (hustot)

$$\Delta = \frac{1}{n^2} \sum_{i=1}^{k} \sum_{j=1}^{k} n_i n_j \left| x_i - x_j \right| = \frac{1}{(\sum n_t)^2} \sum_{i=1}^{k} \sum_{j=1}^{k} n_i n_j \left| \frac{y_i}{n_i} - \frac{y_j}{n_j} \right|$$

$$= \frac{1}{n^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \left| n_j y_i - n_i y_j \right| = \frac{2}{n^2} \sum_{i=1}^{k-1} \left(N_i Y_{i+1} - N_{i+1} Y_i \right)$$

$$G = \frac{\Delta}{2\bar{y}} = \sum_{i=1}^{k-1} \left(\frac{N_i}{N_k} \frac{Y_{i+1}}{Y_k} - \frac{N_{i+1}}{N_k} \frac{Y_i}{Y_k} \right)$$

- Lorenzova křivka spojuje body $\left[\frac{N_i}{N_k}; \frac{Y_i}{Y_k} \right]$

Jdi zpět k tabulce dat

příklad Pavlík, Kühnl: str. 114 (okresy středočeského kraje)

Okres	plocha [km^2]	obyvatel	hustota na km^2
i	n_i	y_i	x_i
BN	1443	88288	61,2
RA	930	56489	60,7
PB	1629	106266	65,2
KH	937	81890	87,4
MB	1067	109766	102,9
NB	881	94377	107,1
BE	662	79764	120,5
KO	819	99408	121,4
PZ	634	77940	122,9
ME	713	96104	134,8
PH	597	94328	158,0
KL	692	154445	223,2
AB	496	1175522	2370,0
celkem	11500	2314587	201,3

příklad Pavlík, Kühnl: str. 114