Katedra štatistiky,
Fakulta hospodárskej informatiky,
Ekonomická univerzita v Bratislave

Pravdepodobnostné modelovanie inverznými distribučnými funkciami: Úvod do kvantilového modelovania

Ľubica SIPKOVÁ

november 2008

1. z cyklu prezentácií

Všeobecné požiadavky na pravdepodobnostný model

- **„dobrá zhoda“**
- **jednoduchosť –** môže byť viacparametrický
- **vhodnosť pre vyjadrenie charakteristík rozdelenia**
- **viacnásobný prístup v analýzach -** smerujúci k **systému** viacparametrických funkcií (modelovanie v podsúboroch)
- **možnosť aktualizácie -** umožňujúci transformácie prehlbujúce elasticitu
- **aplikovateľnosť simulačných metód**

Štatistické prístupy k pravdepodobnostnému modelovaniu – DF, f

- **Klasický prístup (aplikuje sa ako prvá alternatíva):**
  - náhodný výber $X = (X_1, X_2, \ldots X_n)'$
  - spojité náhoné premenné $X_1, X_2, \ldots X_n$

X_i štatisticky nezávislé a s identickým rozdelením

  - náhodné pozorovania, realizácia náhodného výberu

$$x_1, \; x_2, \ldots x_n$$

  - definovanie rozdelenia:
    - **distribučnou funkciou**
    - **funkciou hustoty pravdepodobnosti**

Štatistické prístupy k pravdepodobnostnému modelovaniu – QF, q

- **Kvantilový prístup (aplikuje sa len ak prvá alternatíva zlyhá – komplikovanejší):**
  - *n* usporiadaných štatistík $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$
    (vzostupné usporiadanie *n* spojitých náhodných premenných $X_1, X_2, \ldots X_n$)

$X_{i:n}$ štatisticky závislé a s neidentickým rozdelením

  - hodnota *i*-tej usporiadanej štatistiky vo výbere o rozsahu *n*

$$X_{i:n},\ i = 1,2,\ldots r,\ldots n$$

  - *n* usporiadaných pozorovaní $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$
  - definovanie rozdelenia:
    - **kvantilovou funkciou – inverznou DF**
    - **kvantilovou funkciou hustoty**

Kvantilový prístup

- **Parametrické metódy** pravdepodobnostného modelovania **spojitej náhodnej premennej** na kvantilovom základe vo všetkých fázach štatistického modelovania (identifikácie, estimácie a verifikácie)
- Metódy indukcie z výberových údajov na populáciu vychádzajú z **teórie poriadkových štatistík**
- Východiskom teórie poriadkových štatistík je náhodný výber :
  n –rozmerný vektor vzostupne usporiadaných, teda **vzájomne závislých náhodných premenných**..
- Každá náhodná premenná je kvantilom rozdelenia náhodnej premennej X - je poriadkovou štatistikou a má rozdelenie prevdepodobností závislé od rozdelenia náhodnej premennej X.
  - **Rozdelenia poriadkových štatistík nie sú identické**.
- Rad n vzostupne usporiadaných pozorovaní náhodnej premennej X predstavuje **jednu realizáciu n poriadkových štatistík**, čiže **výberové hodnoty poriadkových štatististík** náhodného výberu **o rozsahu n**

Definovanie tvaru podľa dvoch prístupov

- **Klasický**
  - **distribučnou funkciou**

$$F(x) = P(X \leq x) = p$$

  - **funkciou hustoty pravdepodobnosti**

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

- **Kvantilový**
  - **kvantilovou funkciou**

$$Q(p) = F^{-1}(p) = x, \qquad 0 \leq p \leq 1$$

  - **kvantilovou funkciou hustoty**

$$q(p) = \frac{\mathrm{d}Q(p)}{\mathrm{d}p}, \quad 0 \leq p \leq 1$$

Rozdelenie akéhokoľvek druhu, vyjadrené vo forme kvantilovej funkcie QF alebo kvantilovej funkcie hustoty, sa nazýva skrátene *kvantilovým rozdelením*, presne *kvantilovým pravdepodobnostným rozdelením kvantitatívnej spojitej náhodnej premennej.*

Kvantilová funkcia

$$Q(p) = F^{-1}(p) = x, \qquad 0 \leq p \leq 1$$

- **V literatúre je nazývaná aj ako**

percentilová (pravdepodobnostná) funkcia

- **Je inverziou k distribučnej funkcii pravdepodobnostného rozdelenia (DF: vstupom hodnoty a výstupom pravdepodobnosti)**

- **Vstupom do nej sú pravdepodobnosti a výstupom hodnoty náhodnej premennej**

Napr.: Štyri spôsoby definovania exponenciálneho rozdelenia

- **Klasický**
  - **distribučnou funkciou**

$$F_{EXP}(x) = 1 - e^{-\gamma x}, \qquad 0 \le x \le \infty, \quad \gamma \ge 0$$

  - **funkciou hustoty pravdepodobnosti**

$$f_{EXP}(x) = \begin{cases} \gamma e^{-\gamma x}, & 0 \le x < \infty \\ 0, & -\infty < x < 0 \end{cases}$$

- **Kvantilový**
  - **kvantilovou funkciou**

$$Q_{EXP}(p) = -\eta \ln(1-p), \qquad \eta = \frac{1}{\gamma}, 0 \le p < 1$$

  - **kvantilovou funkciou hustoty**

$$q_{EXP}(p) = \eta / (1-p), \qquad \eta \ge 0, 0 \le p < 1$$

$X \sim Q[p; \Theta]$

Definovanie kvantilového modelu

$$Q(p) = F^{-1}(p), \qquad 0 \leq p \leq 1$$

Jednoduchý kvantilový model v tvare :

$$Q(p) = \lambda + \eta\, S(p) \qquad 0 \leq p \leq 1$$

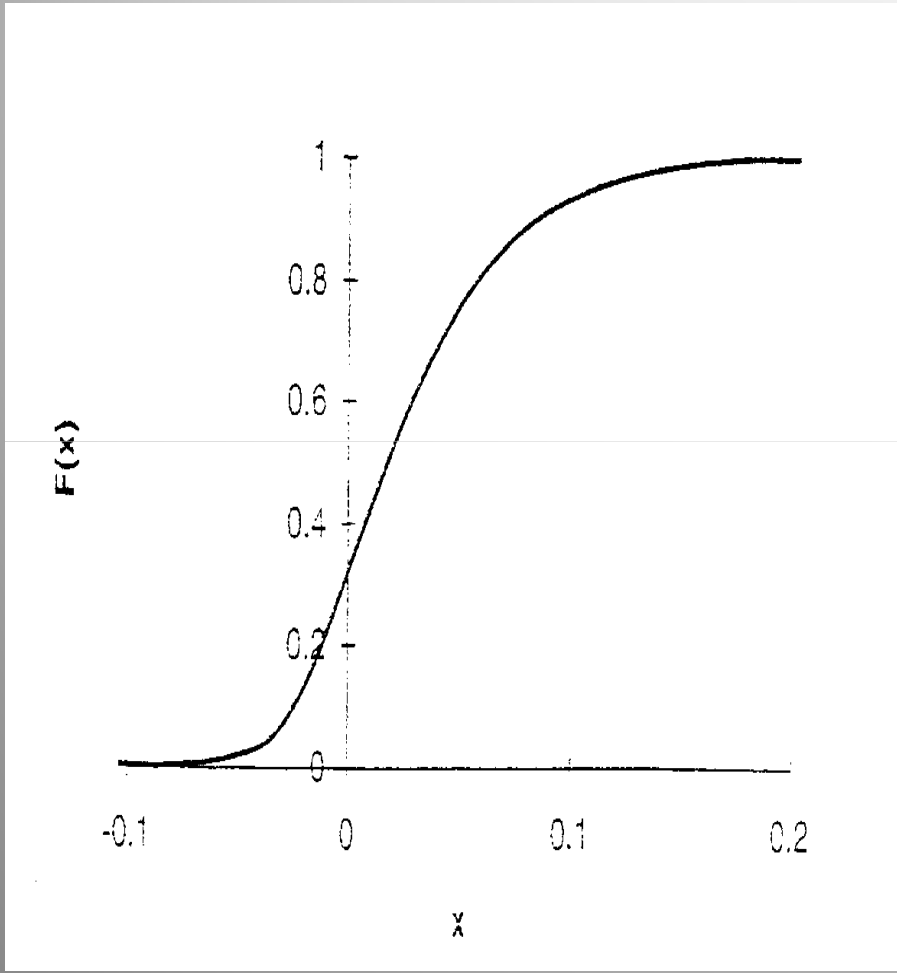
môže byť lineárnym, alebo semilineárnym
dvoj a viacparametrickým
kvantilovým distribučným modelom

v závislosti od jeho **základného kvantilového tvaru** :

$$S(p)$$

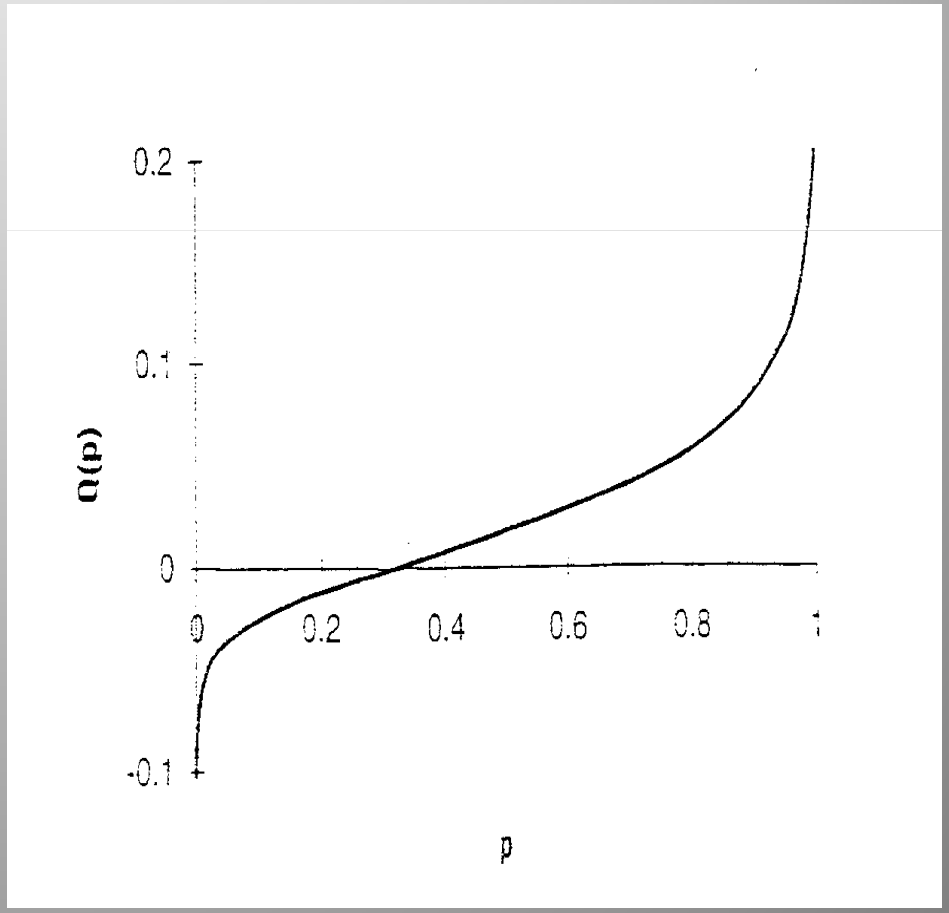
Porovnanie grafov DF a QF

zošikmeného logistického rozdelenia



DF - Kumulatívna distribučná funkcia

QF - Kvantilová (distribučná) funkcia



Východiská kvantilového modelovania

- teória kvantilového pravdepodobnostného modelovania (fázy identifikácie, estimácie, verifikácie)
- teória štatistickej indukcie „Order statistics“ – poriadkové (usporiadané) štatistiky
- vstupné empirické dáta – výberové hodnoty *NP*:
  - vzostupne usporiadaný empirický súbor
  - výberové podiely *p*
  - charakteristické črty empirického rozdelenia *NP* odhalené metódami deskriptívnej štatistiky na rôznych základoch

Nevýhody kvantilového modelovania

- matematicky náročná rozpracovanosť metód
- nedostupnosť jednoducho spracovanej komplexnej literatúry (ani zahraničnej)
- zriedkavejšia aplikácia (s rôznym označením a pojmovým aparátom, zameranie na riešenie konkrétnych oblastí problémov z praxe)
- chýbajúci softvér
- Potrebná hlbšia znalosť matematicko-štatistických metód (náročnejší prístup z hľadiska teoretického aj technického)

Výhody kvantilového modelovania

- v prípadoch, keď klasické prístupy neprinášajú uspokojivé výsledky umožňuje nájsť vhodný tvar modelu
- aktualizácia modelu v budúcnosti novou estimáciou parametrov – veľká flexibilita tvarov
- možnosť spoločného modelovania deterministickej a náhodnej zložky v štatistických modeloch
- pri estimácii parametrov toho istého tvaru modelu možnosť modelovať v podsúboroch, v štruktúrach
- nový pohľad na riešený problém, niekedy jednoduchší, jasnejší – prostredníctvom kvantilových mier
- smerovanie k modelovaniu zmesí
- jednoduché simulácie

História kvantilového prístupu v štatistickej analýze

- článok **Francisa Galtona** uverejnený v roku **1875,** potom sa na kvantilový prístup takmer na 100 rokov zabudlo a rozvíjal sa Pearsonov prístup na báza momentov

- empirická kvantilová funkcia v súvislosti s normálnym rozdelením nazvaná **ogivou** (nie empirická funkcia hustoty - ako v súčasnosti)

- medián a kvartil boli prvýkrát popísané v roku **1882**

- pojem **kvantil** bol zavedený M. G. **Kendalom** až v roku **1940**

História kvantilového modelovania

- Použitie kvantilovej funkcie pre definovanie tvaru pravdepodobnostného rozdelenia –

  Emanuel Parzen 1979
- Handbook of Statistics, časť 16 a 17 editovaná Balakrishnan a Rao (**Arnold, Balakrishnan, Balanda, David, MacGilivray, Moors, Nagaraja a hlavne Parzen)**
- **A First Course in Order statistics - Arnold, Balakrishnan, Nagaraja**
- r. **2000** Modelovanie kvantilovými funkciami (**Statistical modelling with quantile function) Gilchrist, W.G.**

Rozdelenie *r*-tej poriadkovej štatistiky

N poriadkových štatistík $X_{1:n}$, $X_{2:n}$,... , $X_{n:n}$

hodnota *i*-tej poriadkovej štatistiky vo výbere o rozsahu *n*

$$X_{i:n},\ i = 1,2,...r,...n$$

Rozdelenie *r*-tej poriadkovej štatistiky $X_{r:n}$ možno vyjadriť kvantilovou funkciou:

$$X_{r:n} \sim Q(r)[p(r);\Theta(r)] = Q[\mathrm{BETAINV}(p(r), r, n\text{-}r\text{+}1); \Theta]$$

Z nej možno vyjadriť rozdelenie **prvej**, či *n*-tej poriadkovej štatistiky

Rankit

$X \sim Q[p;\Theta]$

- **stredné hodnoty poriadkových štatistík:**

$$E(X_{i:n}) = \mu_{i:n}\,,\; i = 1,2,\ldots r,\ldots n$$

$$E\left(X_{r:n}^{k}\right) = \{n!/[(r-1)!(n-r)!]\}\int_0^1 Q^k(p)\,p^{(r-1)}[1-p]^{(n-r)}\,dp$$

Pre **rovnomerné rozdelenie** (Uniform distribution – ozn. **U**):

$$X \sim Q_{ROVN}(p) \quad \text{potom} \quad \mu_{r:n} = r/(n+1)$$

Odhady $\mu_{r:n}$ **ľubovolného rozdelenia** pomocou **U**-transf.pravidla:

$$\mu_{r:n} = E(Q(U_{r:n})) \approx Q(E(U_{r:n})) = Q(r/(n+1))$$

$$\approx Q((r-0{,}5)/n)$$

Mediánový rankit

$X \sim Q[p;\Theta]$

- mediány poriadkových štatistík:

$M_{r:n} = Q[p_{r:n}^{*} ; \Theta]$, kde:

$$p_{r:n}^{*} = \text{BETAINV}(0{,}5 ; r, n-r+1)$$

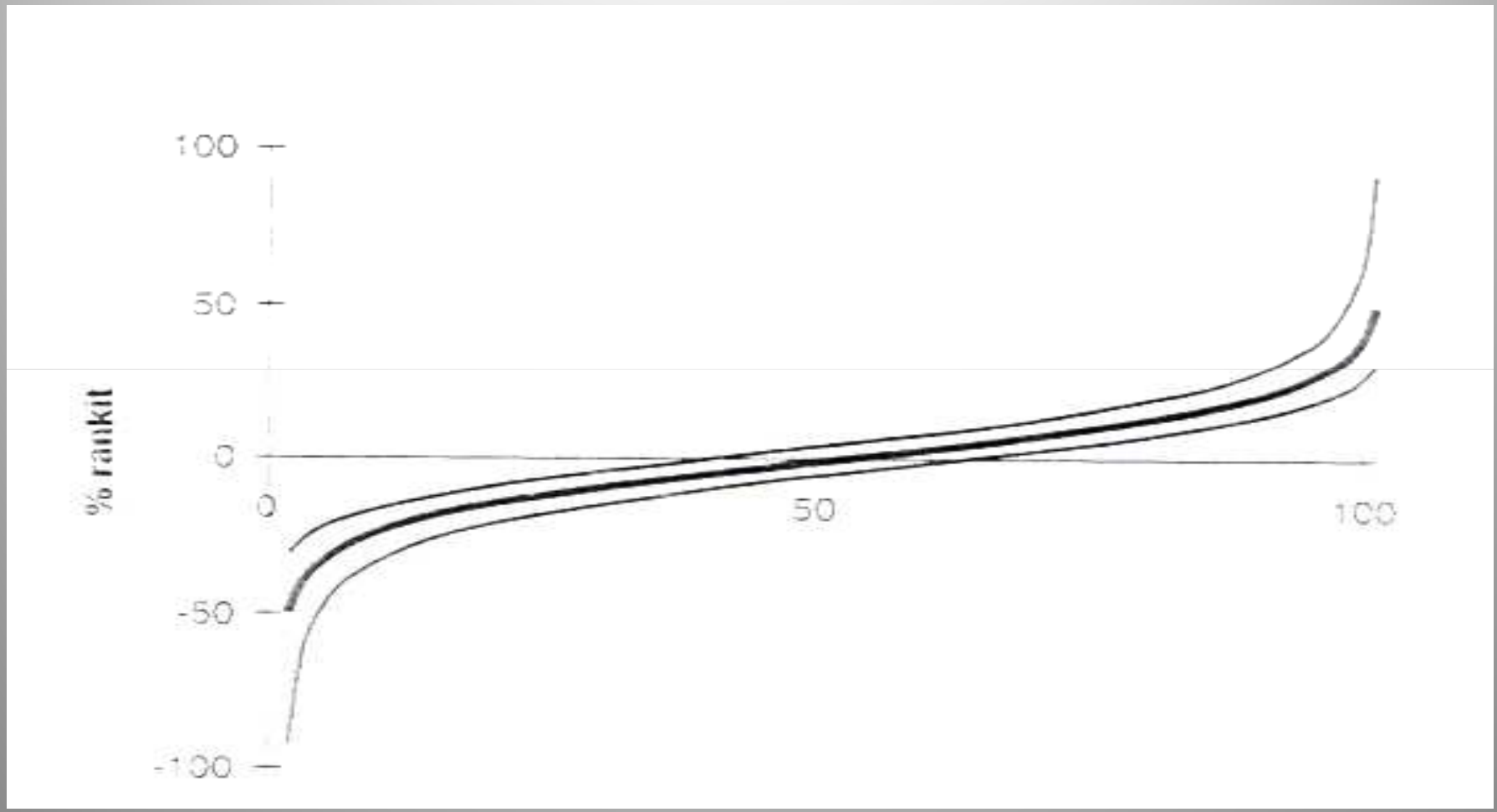
$p_{r:n}^{*}$ **- mediánová** p_r **hodnota**

Rozdelenia $X_{r:n}$ charakterizované **kvantilovými štatistikami**:

dolný percentil $X_{r:n\,(0{,}01)} = Q[\text{BETAINV}(0{,}01 ; r, n-r+1); \Theta]$

horný percentil $X_{r:n\,(0{,}99)} = Q[\text{BETAINV}(0{,}99 ; r, n-r+1); \Theta]$

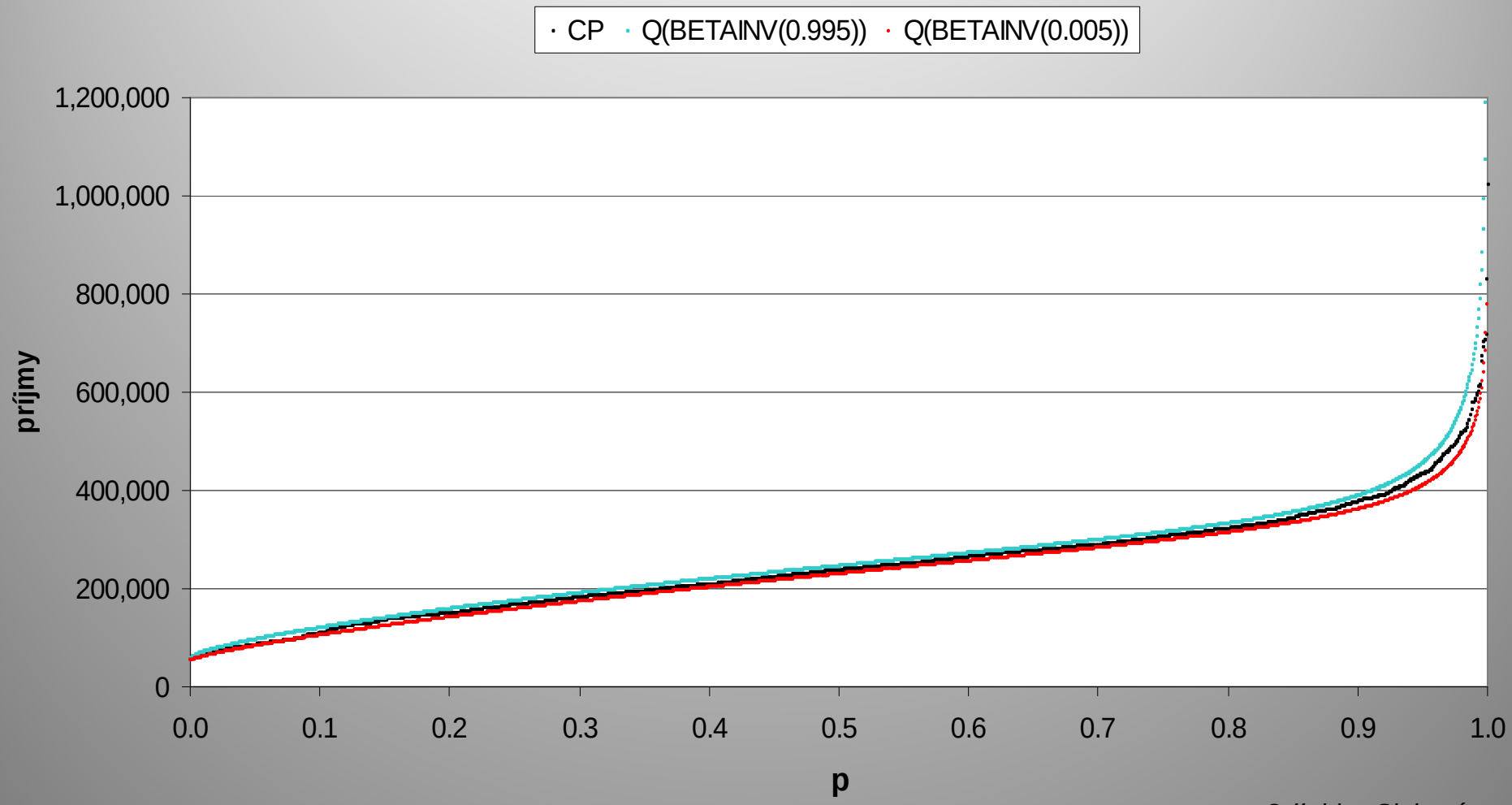
Graf mediánového rankitu a kvantilov poriadkových štatistík



Kvantily (dolné percentily, mediánový rankit a horné percentily) poriadkových štatistík logistického rozdelenia

Weibullov-Paretov tvar Q-modelu príjmov
(Odhad metódou minimalizácie absolútnych distribučných rezíduí)

Graf kvantilov CP s 99% hranicami teoretického rozdelenia



Vystihnutie tvaru empirického rozdelenia - identifikácia

- **Grafická analýza empirického rozdelenia príjmov**
- **Kvantitatívna analýza empirického rozdelenia príjmov:**
  - na momentovom základe
  - na kvantilovom základe

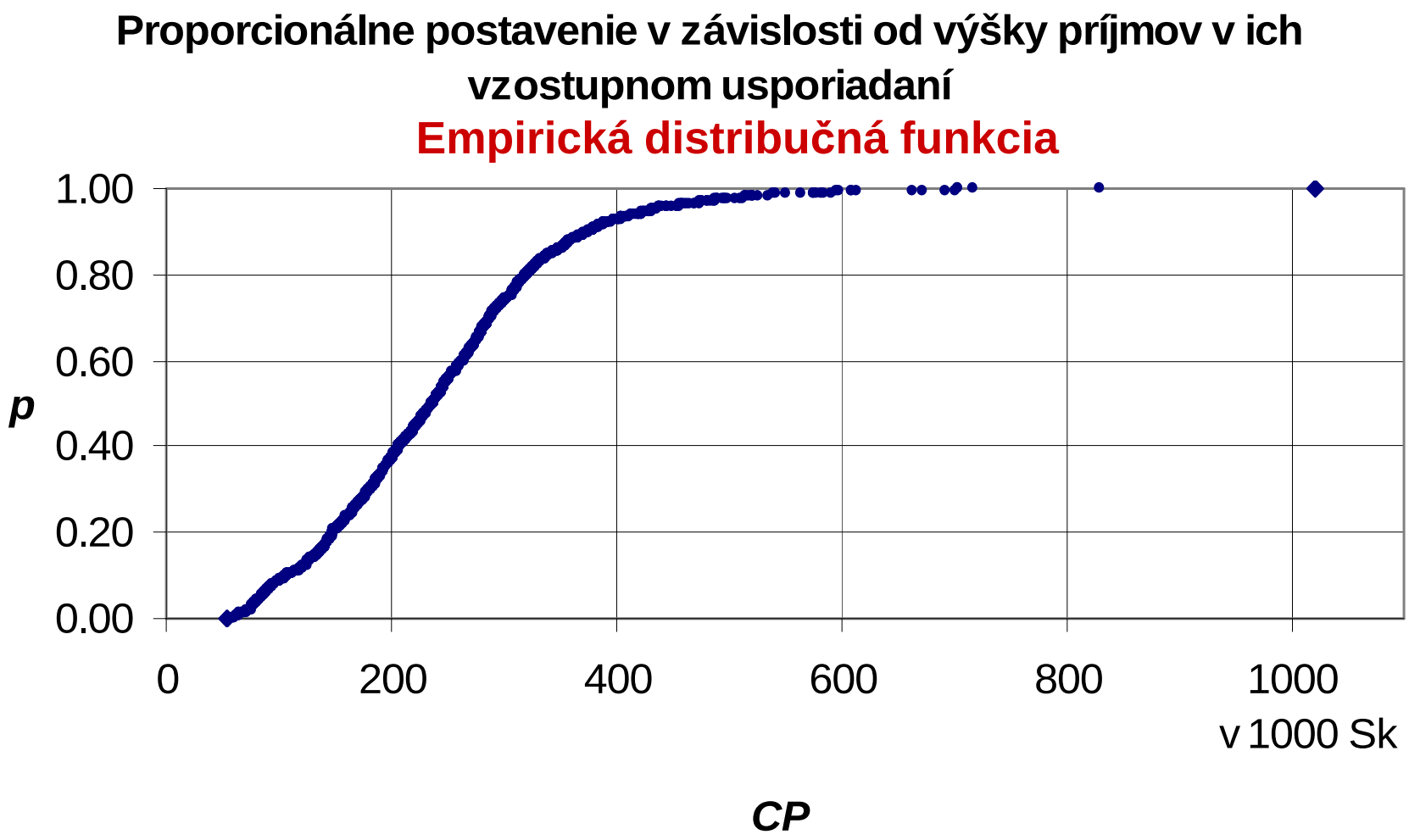
- **Identifikácia rozdelenia jednoduchými tvarmi**
  - Celého rozdelenia
  - Zvlášť pre dolný a horný koniec , prípadne iné časti rozdelenia

- **Voľba tvaru váh pre konce rozdelenia**
  - Ako funkcia *p*
  - Ako parametre modelu

Grafická analýza empirického rozdelenia

Príjmy domácností SR v roku 2003

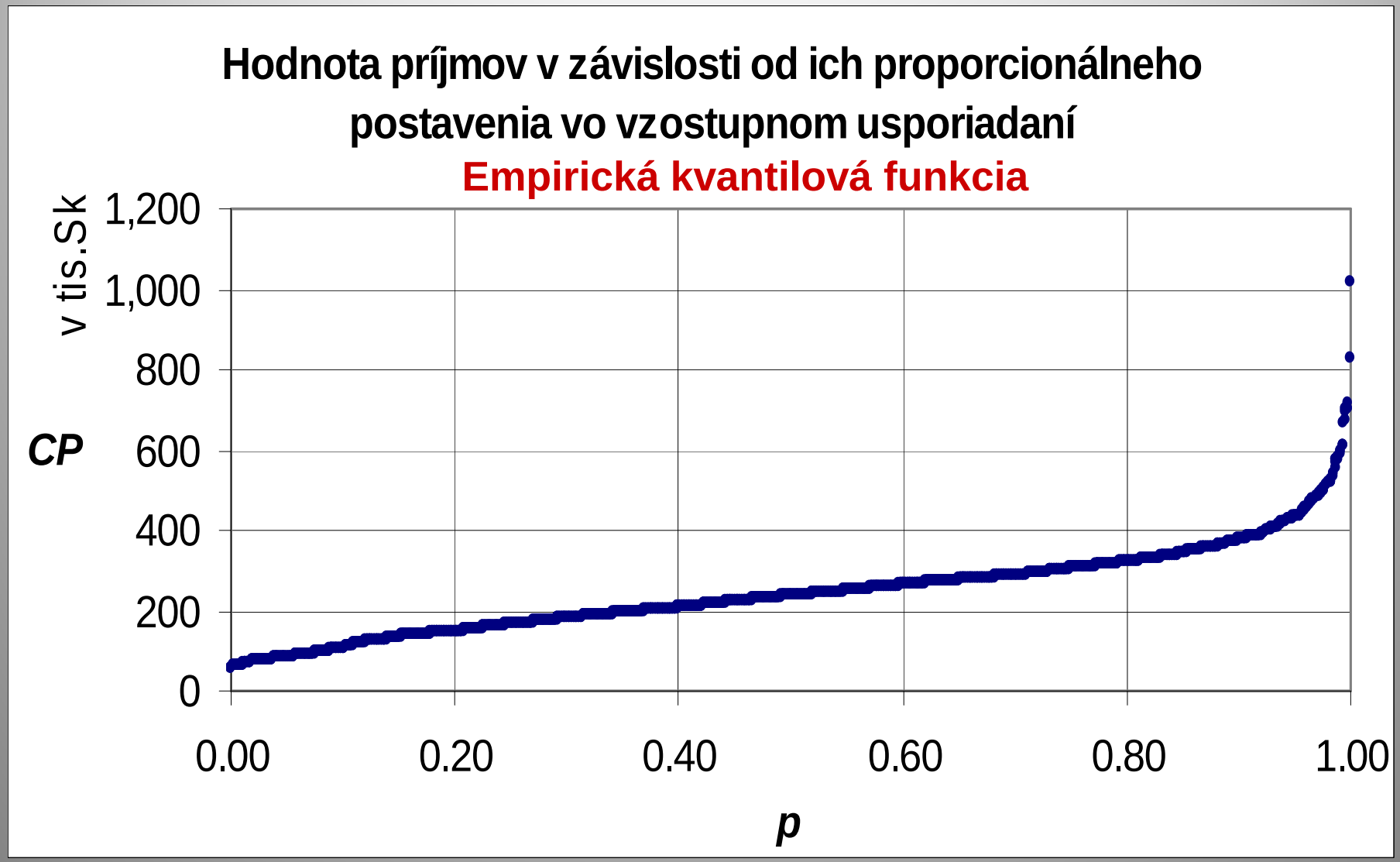
$(F_{emp}(CP) = p)$



Grafická analýza empirického rozdelenia

Príjmy domácností SR v roku 2003

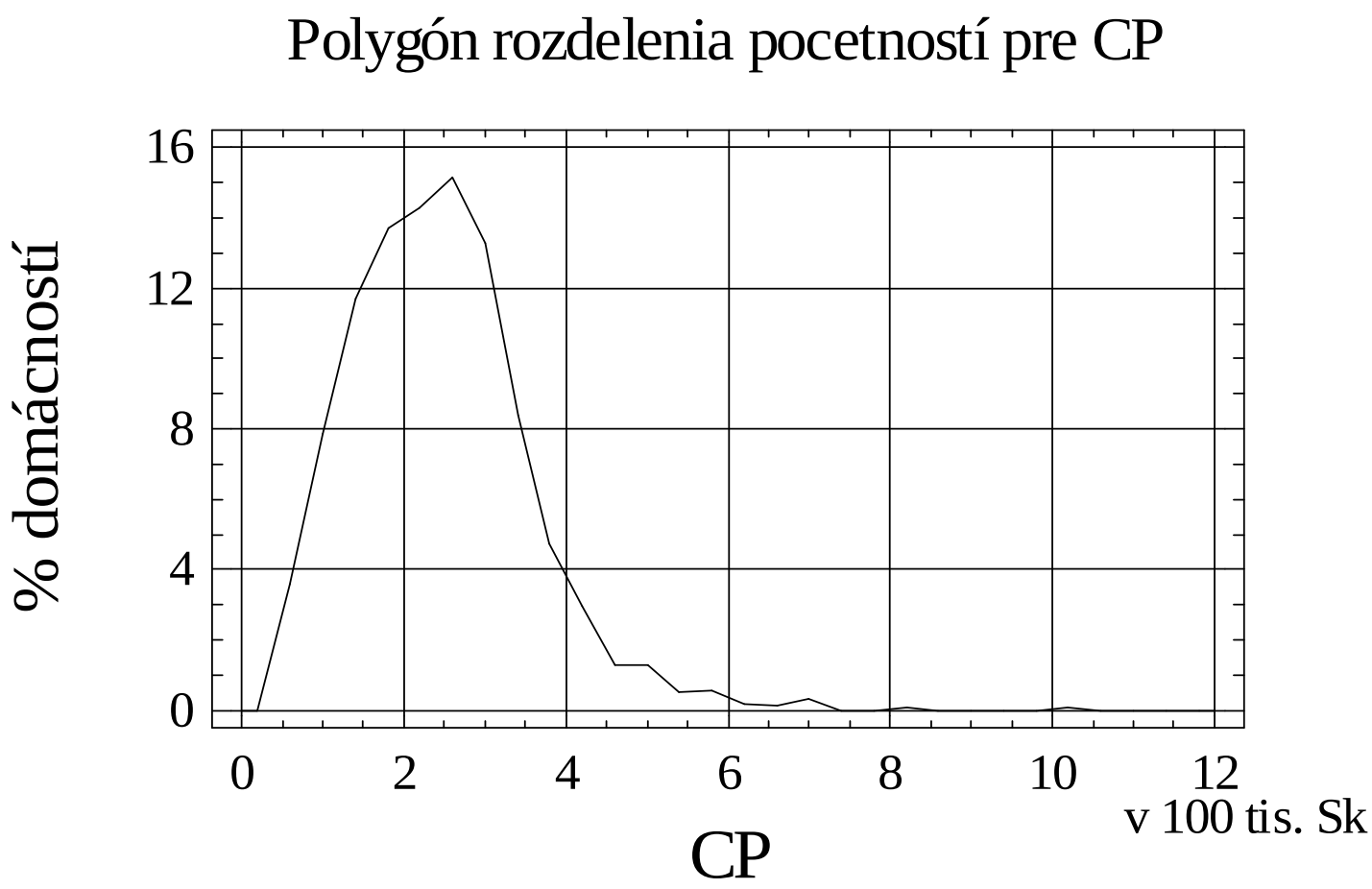
$(Q_{emp}(p)=CP)$



Grafická analýza empirického rozdelenia

Príjmy domácností SR v 2003

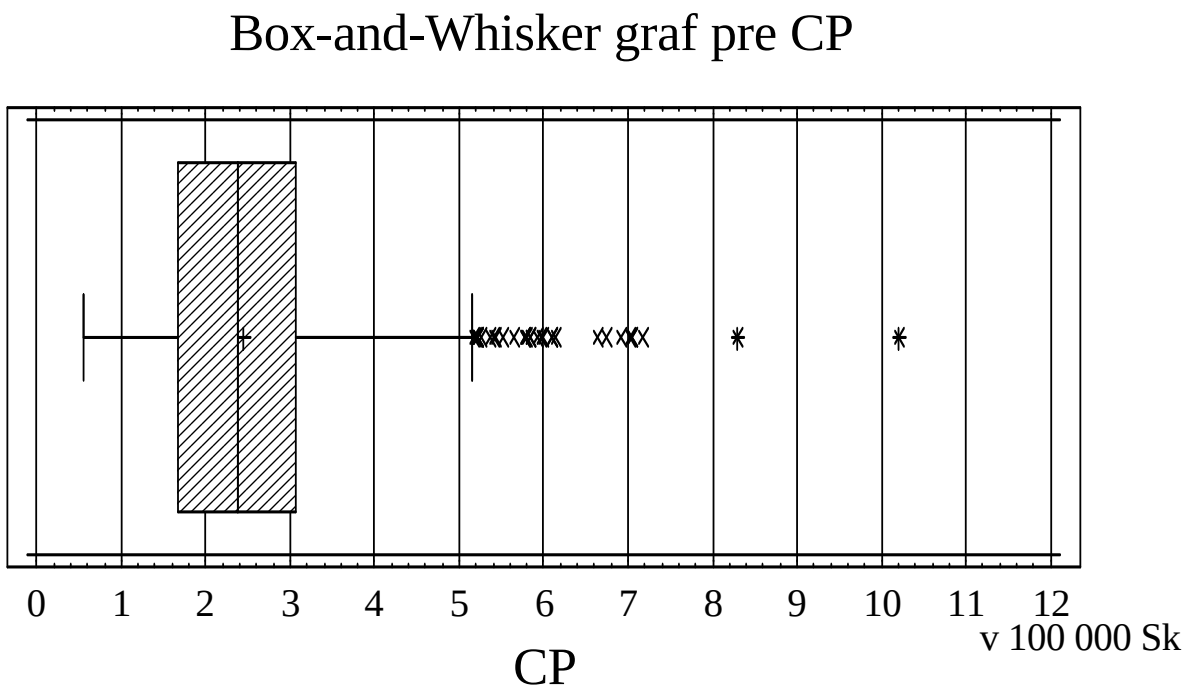
Polygón



Grafická analýza empirického rozdelenia

Príjmy domácností SR v 2003

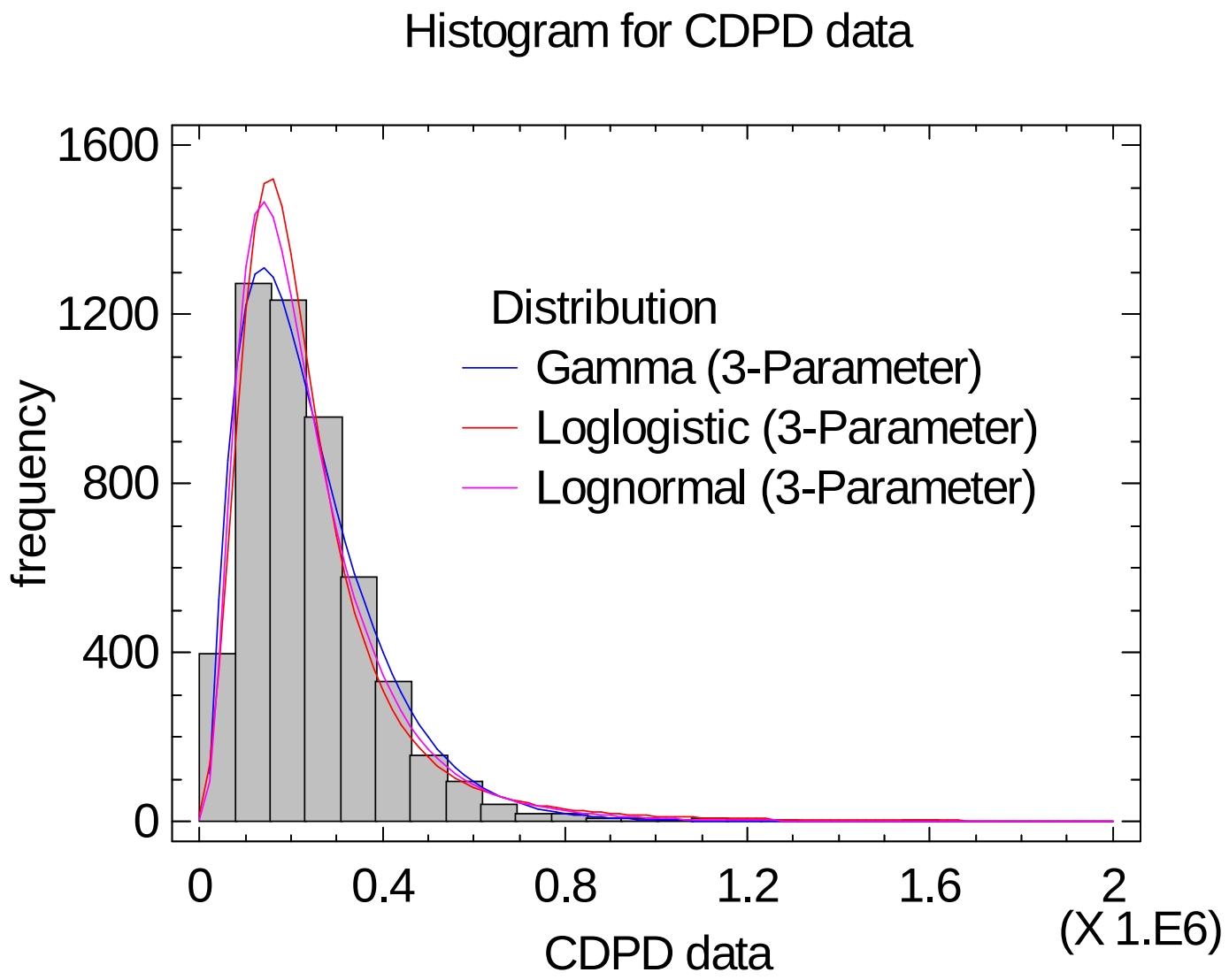
(Box-Plot)



Comparison of Alternative Distributions

procedúra na porovnanie známych tvarov v STATGRAPHICS CENTURION

Distribution	*Est. Parameters*	*Log Likelihood*	*KS D*
Lognormal (3-Parameter)	3	-67677.7	0.0362935
Lognormal	2	-67704.9	0.0397763
Loglogistic	2	-67730.3	0.0496562
Gamma (3-Parameter)	3	-67743.3	0.0244112
Birnbaum-Saunders	2	-67763.8	0.0479773
Gamma	2	-67775.9	0.0292224
Inverse Gaussian	2	-67789.4	0.0550491
Largest Extreme Value	2	-67873.8	0.0483855
Weibull	2	-68026.0	0.0646936
Logistic	2	-68477.7	0.0910885
Laplace	2	-68514.9	0.108414
Exponential	1	-68795.9	0.199527
Normal	2	-69321.9	0.120725


Histogram for CDPD data
1600
1200
800
400
0
frequency
Distribution
Gamma (3-Parameter)
Loglogistic (3-Parameter)
Lognormal (3-Parameter)
0
0.4
0.8
1.2
1.6
2
(X 1.E6)
CDPD data

KS a CHi-kvadrát testy potvrdili nevhodnosť aplikácie jednoduchých tvarov modelu CP

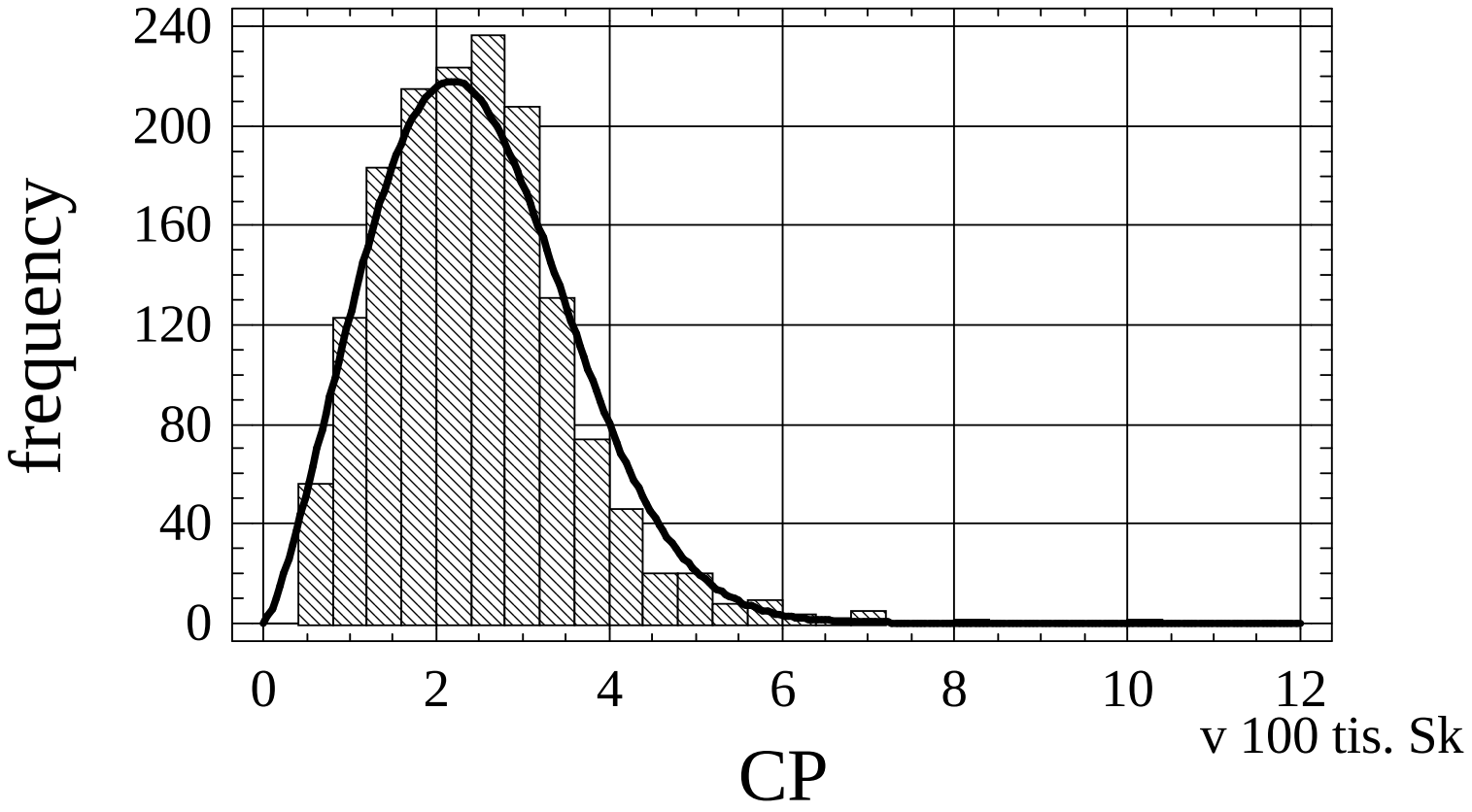
KS Test	Erlang	Gamma (3-Parameter)	Loglogistic (3-Parameter)	Weibull (3-Parameter)	Lognormal (3-Parameter)
DPLUS	**0.0**	**0.0244112**	**0.0494693**	**0.0399911**	**0.0362935**
DMINUS	**1.0**	**0.0227598**	**0.0365427**	**0.0527996**	**0.0303411**
DN	**1.0**	**0.0244112**	**0.0494693**	**0.0527996**	**0.0362935**
P-Value	**0.0**	**0.00437615**	**0.0**	**0.0**	**0.00000264**

Chi kv. Test	Gamma	Gamma (3-Parameter)	Lognormal	Lognormal (3-Parameter)
Chi-Squared	399.179	379.992	394.717	354.507
D.f.	97	96	97	96
P-Value	0.0	0.0	0.0	0.0

Identifikácia rozdelenia jednoduchými tvarmi

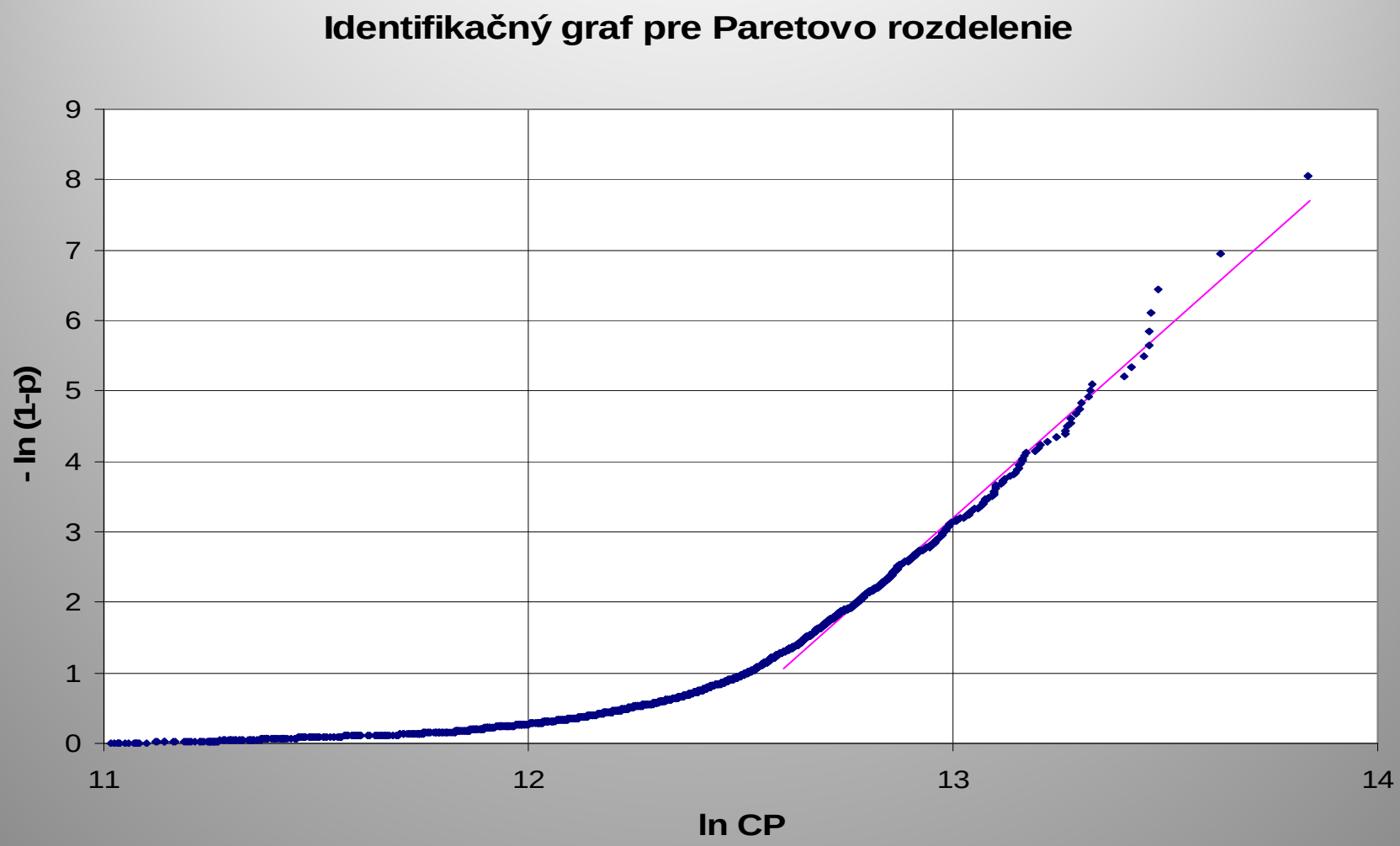
Useknutý Weibullov tvar - len pre dolnú a strednú časť rozdelenia CP

Histogram pre CP s funkciou hustoty Weibullovho rozdelenia



Identifikácia rozdelenia jednoduchými tvarmi

Paretov tvar je vhodný len pre horný koniec rozdelenia



Kvantitatívna analýza empirického rozdelenia

Príjmy domácností SR v 2003

Výberové charakteristiky

➢ na momentovom základe

Priemer	**= 244 284 Sk**	**Koeficient šikmosti**	**= 1.0498**
Rozptyl	**= 1.2012E10**	Štand. šikmosť	= 16.959
Štandardná odchýlka	= 109 598 Sk	**Pearsonova špicatosť**	**= 2.997**
Variačný koeficient	= 44.87 %	Štand. špicatosť	= 24.209

➢ na kvantilovom základe

Medián		**= 237 407 Sk**
Kvartilové rozpätie		**= 139 770 Sk**
Galtonov koeficient šikmosti	g=qd/iqr	**= -0.0123**
Kvartilová diferencia	qd=lq+uq-2	= -1718.50
Moorsova špicatosť	k=[(e7-e5)+(e3-e1)]/iqr	**= 1.174**

Ako može vyzerať kvantilový model ?

Jednoduchý kvantilový model v tvare :

$$Q(p) = \lambda + \eta\, S(p) \qquad 0 \le p \le 1$$

- parameter polohy
- parameter stupnice, variability
- vlastné parametre základného tvaru kvantilovej funkcie

Poskladaný (zložený) kvantilový model napr. v tvare :

$$Q_{CP}(p) = \lambda + \eta\left\{(1-p)\left[-\ln(1-p)\right]^{\beta} + p\left[\frac{1}{(1-p)^{\gamma}}\right]\right\},\ 0 < p < 1,\ \beta > 0,\ \gamma > 0$$

- váhy jednotlivým rozdeleniam
- kvantilový základný tvar rozdelenia pre dolný koniec s jeho parametrami
- kvantilový základný tvar rozdelenia pre horný koniec s jeho parametrami

Možné tvary váh pre dva konce rozdelenia

Ako funkcia *p*

$$(1-p) \quad \text{a} \quad p$$

$$\left[1-p^2(3-2p)\right] \quad \text{a} \quad p^2(3-2p)$$

$$\left[1-p^3(10-15p+6p^2)\right] \quad \text{a} \quad p^3(10-15p+6p^2)$$

$$1-\omega(p) \quad \text{a} \quad \omega(p)$$

Ako parametre modelu

$$(1-\omega) \quad \text{a} \quad \omega$$

$$\frac{(1+\omega)}{2} \quad \text{a} \quad \frac{(1-\omega)}{2}$$

$$\omega_1(1-p) \quad \text{a} \quad \omega_2 p$$

Zošikmené logistické rozdelenie

Logistické rozdelenie (Logistic):

$$S(p) = S_1(p) + S_2(p) \qquad 0 \leq p \leq 1$$

$$S(p) = - \ln (1 - p) + \ln (p) ,$$

$$S(p) = \ln [p/(1 - p)]$$

$$Q_1(p) \leq Q(p) \leq Q_2(p)$$

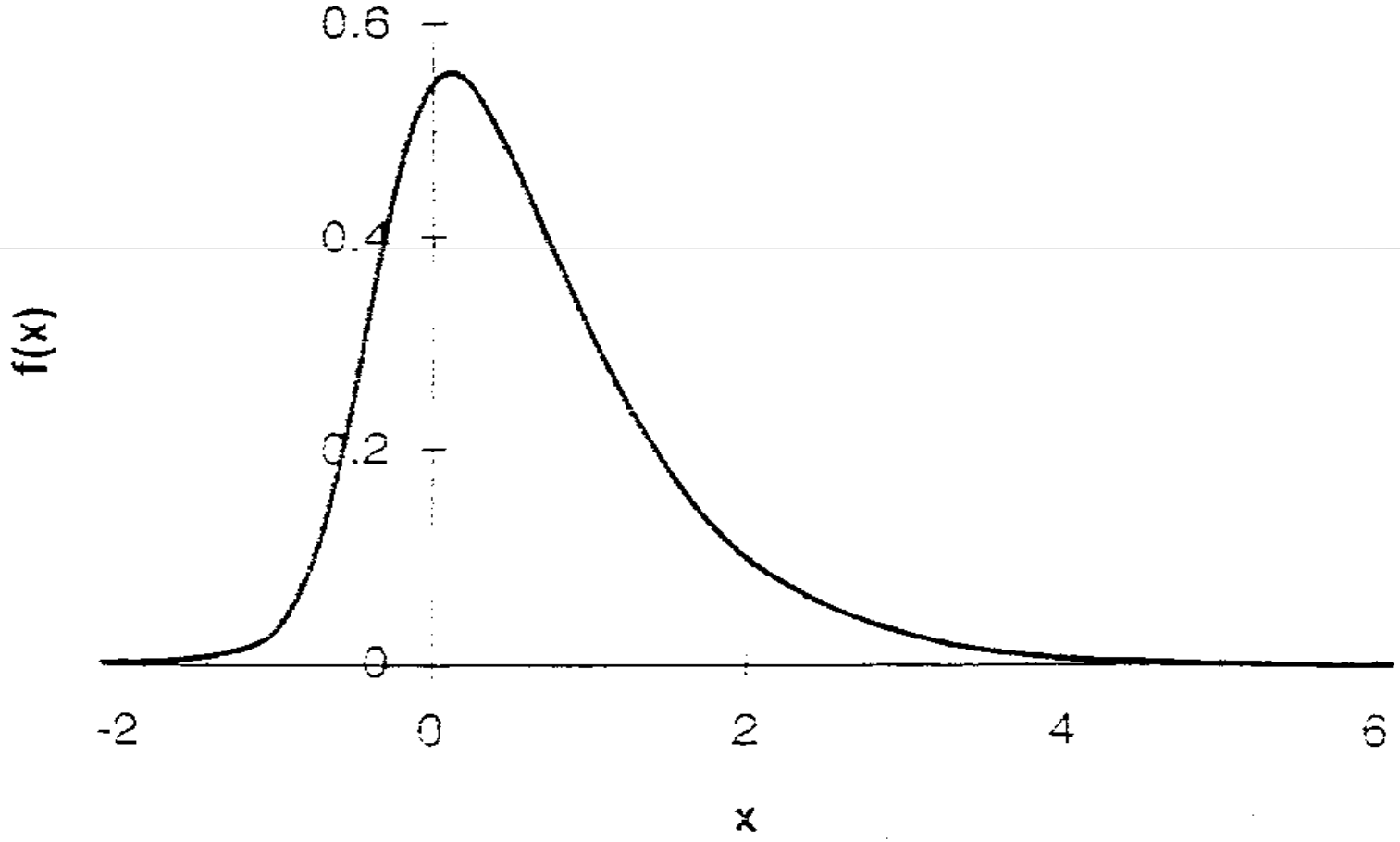
Zošikmené logistické rozdelenie:

$$S(p) = \omega\, S_1(p) + (1-\omega)\, S_2(p), \text{ kde } 0 \leq \omega \leq 1$$

$$S(p) = [(1 + \delta)/2]\, [-\ln (1-p)] + [(1 - \delta)/2]\, [\ln (p)]$$

$$\text{kde } -1 \leq \delta \leq 1, 0 < p < 1$$

Funkcia hustoty
Zošikmeného logistického rozdelenia



Súčin mocninového a Paretovho rozdelenia

Mocninové rozdelenie – Power rozdelenie (Po(α)):

$S_1(p) = p^{\alpha}$, kde $\alpha > 0$, $0 \leq p \leq 1$

Paretovo rozdelenie (Pa(β)):

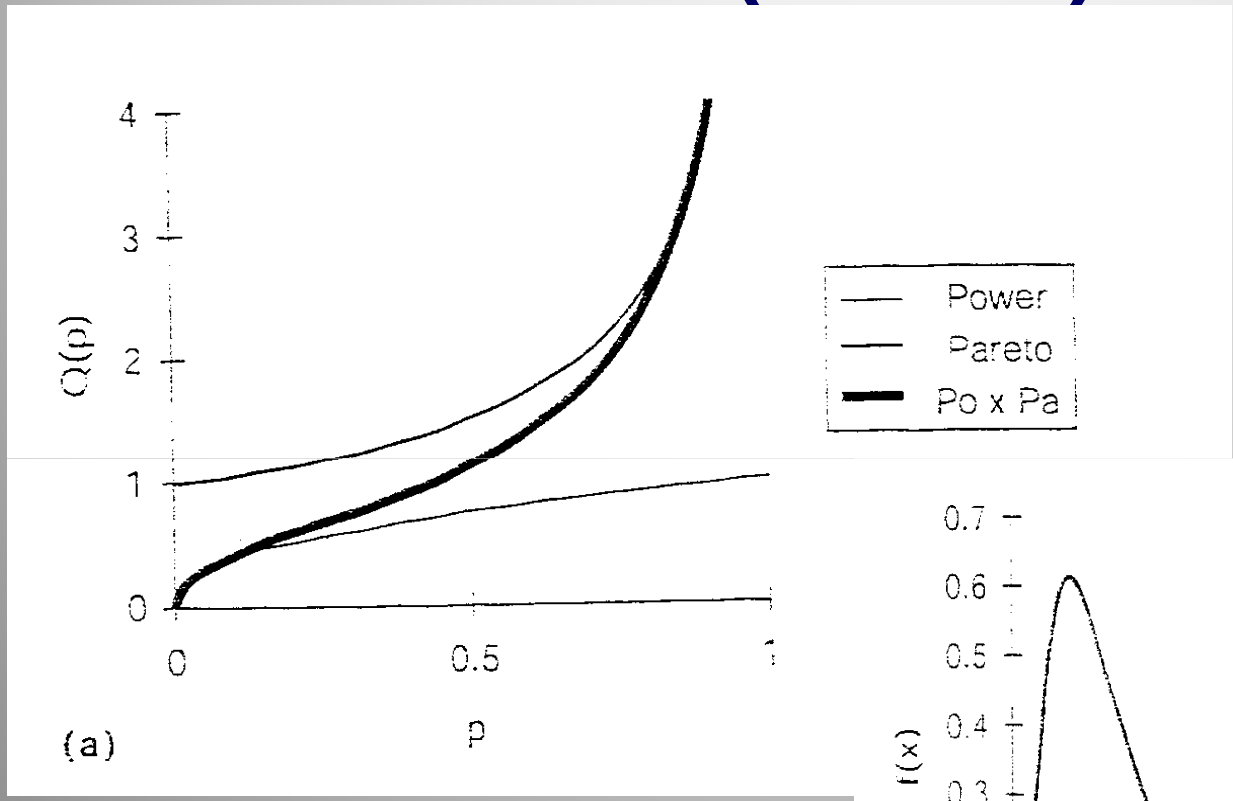
$S_2(p) = 1/(1 - p)^{\beta}$, kde $\beta > 0$,

Mocninovo-Paretovo rozdelenie (Po(α)Pa(β)):

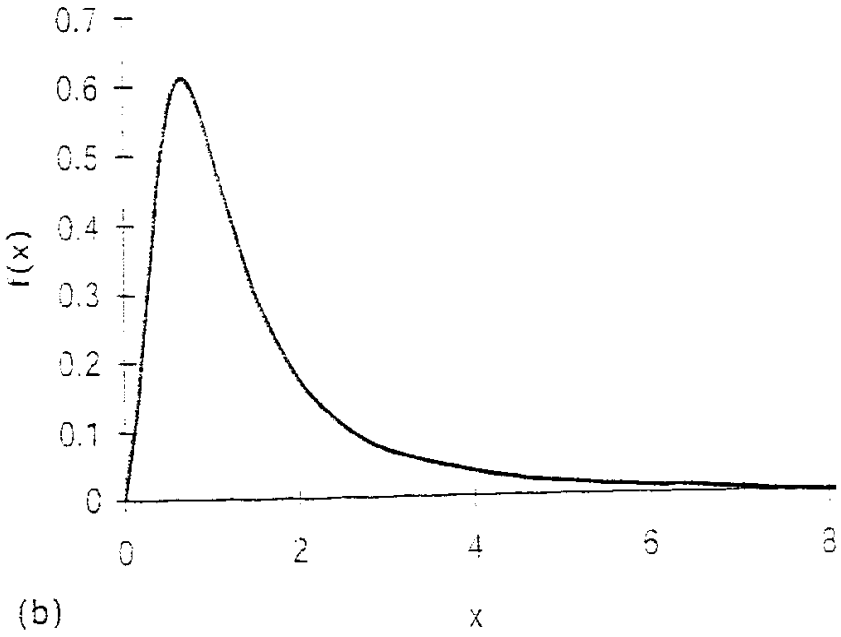
$S(p) = S_1(p) \times S_2(p)$

$S(p) = p^{\alpha}/(1 - p)^{\beta}$, kde $\beta > 0$, $\alpha > 0$,

Graf mocninovo-Paretovho rozdelenia (Po x Pa)



Funkcia hustoty Mocninovo-Paretovho rozdelenia (Po x Pa)



Kvantilová funkcia Mocninového (Power) a Paretovho (Pareto) rozdelenia

Zovšeobecnené lambda rozdelenie

Komplexný známy elastický tvar – len odhadnúť jeho parametre

Ramberg-Schmeiserov tvar:

$$RSGLD(\lambda_1,\lambda_2,\lambda_3,\lambda_4)=\lambda_1+\frac{p^{\lambda_3}-(1-p)^{\lambda_4}}{\lambda_2},\quad 0\le p\le 1$$

Dôvody výberu RS GLD:

- **historický aspekt**
- **teoretický aspekt**
- **komplexnosť, elasticita a univerzálnosť jeho tvaru**
- **konvergencia k Paretovmu tvaru**
- **vhodnosť pre simulačné štúdiá**

Semilineárny päťparametrický Gamma-Paretov tvar

(skladanie použitím pravidiel modifikácie kvantilových funkcií)

$$S\ (p) = p$$

$$R[S(p)] = \frac{1^{\downarrow}}{S(1-p)}$$

pravidlo o reciprocite

$$P(S(p)) = [S(p)]^{\beta}$$

pravidlo o Q-transformácii

$$\downarrow$$

$$S_{PAR}(p) = \frac{1}{(1-p)^{\gamma}}$$

PARETOINV

GAMMAINV

pravidlo o súčte dvoch kvantilových funkcií

$$Q(p) = \alpha + \left(\omega(1-p)\,\mathrm{GAMAINV}(p;\beta.\gamma) + \kappa p \Big/ (1-p)^{\delta} \right), \quad 0 < p < 1|$$

Metódy estimácie kvantilových modelov

- Metódou maximálnej vierohodnosti
- Metódou momentov
- Metódou kvantilov
- Öztürk a Dale aproximatívnou metódou
- Metódou minimalizácie absolútnych distribučných rezíduí
- Špeciálne metódy u zovšeobecnených tvarov (Starship)

Odhad parametrov QF

Zovšeobecnenou Öztürk a Dale aproximatívnou metódou

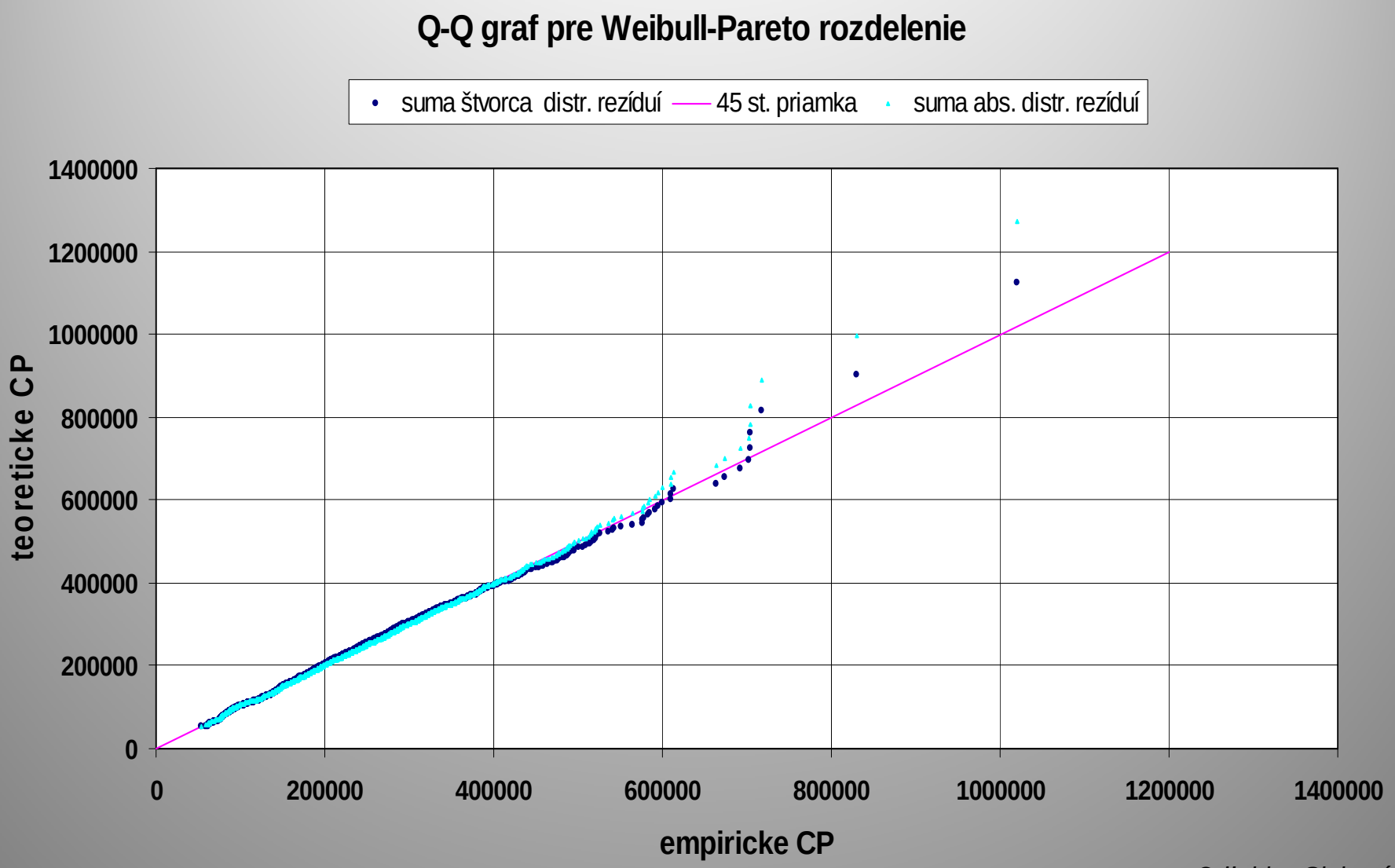
$$S(\Theta)=\sum_{r=1}^{n}\left[x_r - Q\left(\frac{r}{(n+1)},\Theta\right)\right]2$$

Metódou minimalizácie absolútnych distribučných rezíduí

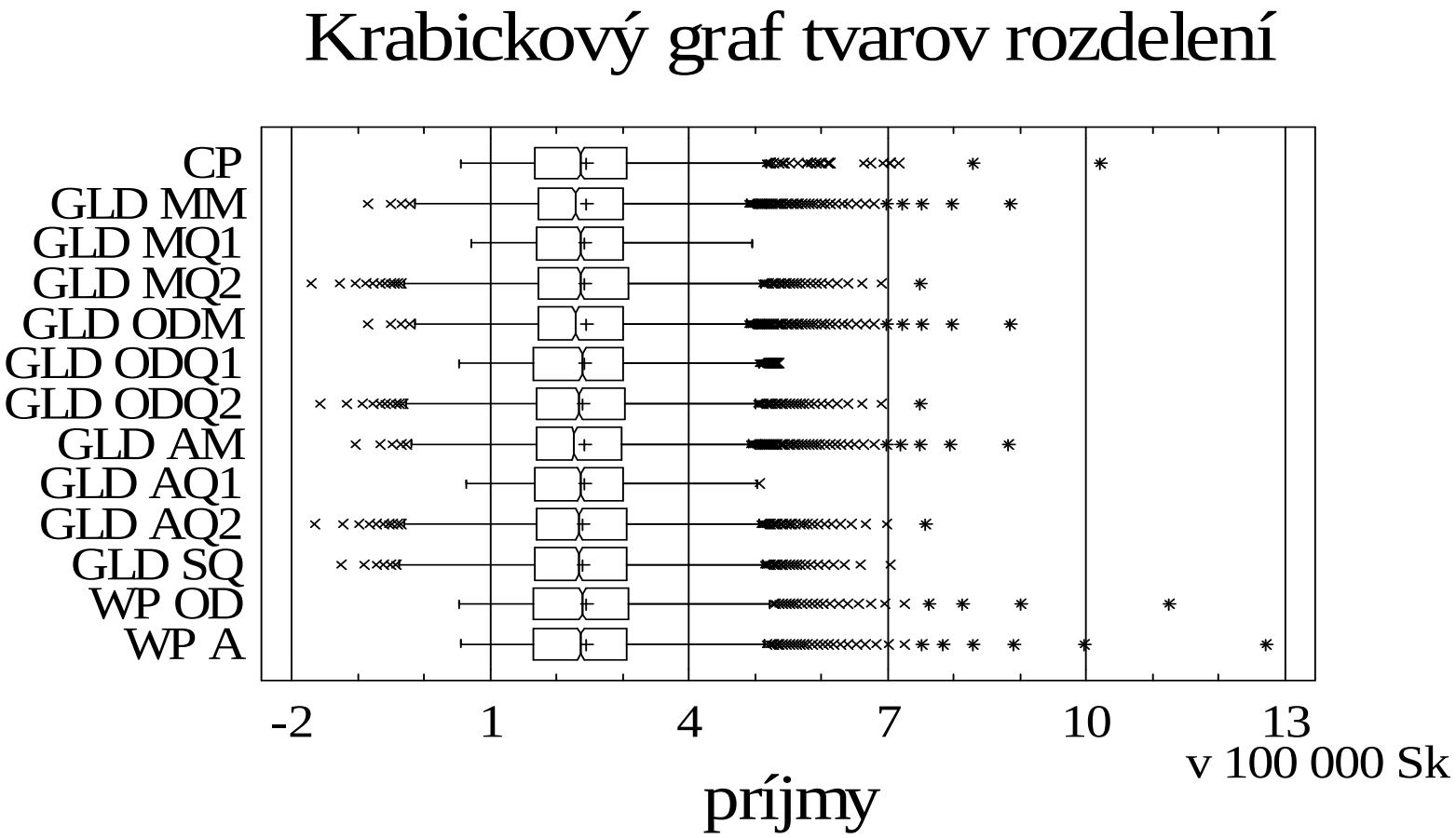
$$S(\Theta)=\sum_{r=1}^{n}\left|x_r - Q\left(p_r^*,\Theta\right)\right|$$

kde $p_r^* = \text{BETAINV}(0{,}5;\ r,\ n\text{-}r\text{+}1)$

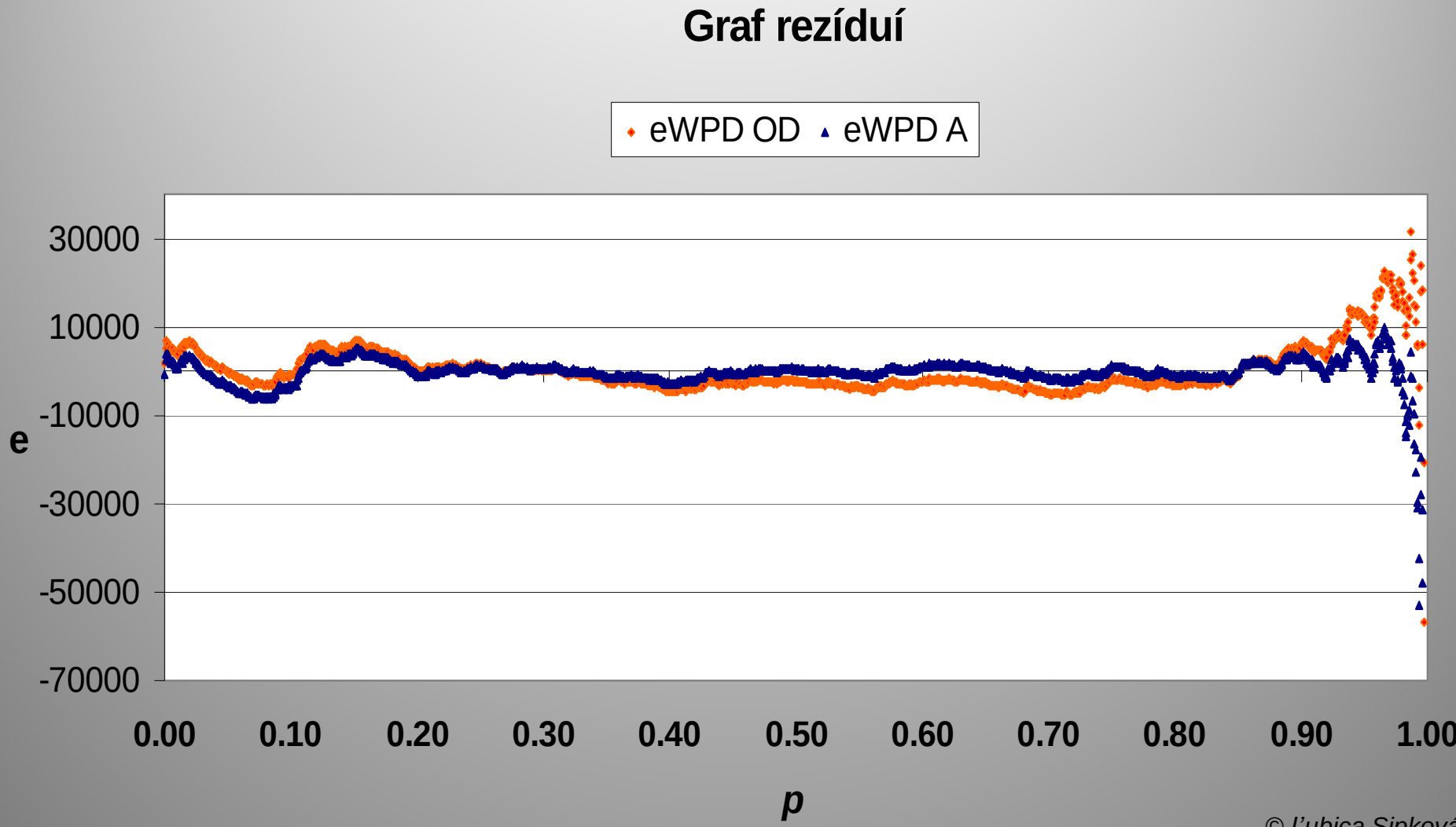
Odhadnutý Weibullov – Paretov tvar



Grafické posúdenie kvality kvantilových modelov rozdelenia príjmov domácností SR - verifikácia



Grafické posúdenie kvality Weibullových-Paretových tvarov modelov rozdelenia CP v SR



Kvantitatívne posúdenie kvality kvantilových modelov rozdelenia CP - verifikácia

Vybrané tvary	WPD_{OD}	WPD_{A}	$RSGLD_{AQ1}$	$RSGLD_{AQ2}$
$\chi^2(19)$	28,64	28,99	35,40	114,08
$\chi^2(19)$ do 0.1	0,07	4,03	0,55	86,28
$\chi^2(19)$ nad 0.95	2,76	0,28	0,36	0,04
Koef.korelácie	0,9982	0,9963	0,9795	0,9867
Dolný percentil	62 tis.Sk	65 tis.Sk	68 tis.Sk	(-12 tis.Sk)
Medián	240 tis.Sk	237 tis.Sk	237 tis.Sk	234 tis.Sk
Horný percentil	570 tis.Sk	601 tis.Sk	490 tis.Sk	546 tis.Sk

výb.dol.perc. CP = **65** tis.Sk

výb.medián CP = **237** tis.Sk

výb.hor.perc. CP = **584** tis.Sk

$\chi^2_{0,95}(19) =$ **30,14**

$\chi^2_{0,99}(19) =$ **36,19**

Pravdepodobnostné modelovanie inverznými distribučnými funkciami: Úvod do kvantilového modelovania

Spracovanie prvej z cyklu prezentácií o kvantilovom modelovaní.

Podrobnejšie možno nájsť v monografii:

Sipková, Ľ; Sodomová, E.: Modelovanie kvantilovými funkciami, Vydavateľstvo EKONÓM, Bratislava, 2007; 175 s.

ISBN 978-80-225-2346-2

Ľubica SIPKOVÁ

november 2008