

Katedra štatistiky,
Fakulta hospodárskej informatiky,
Ekonomická univerzita v Bratislave



**Pravdepodobnostné modelovanie
inverznými distribučnými funkciami:
Verifikácia kvantilových modelov**

Ľubica SIPKOVÁ

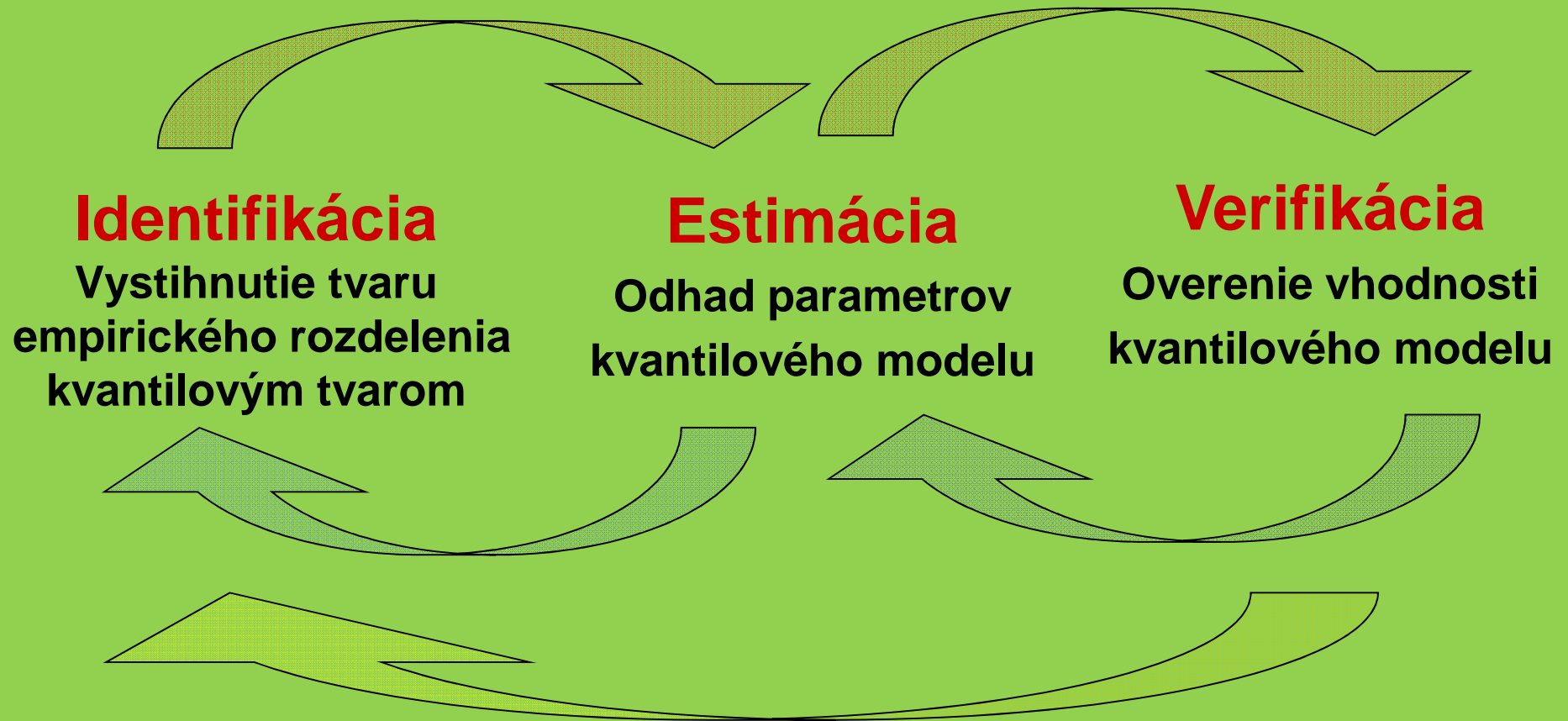
marec 2010

7. z cyklu prezentácií

Záver estimácie – východisko verifikácie

- Výber metódy odhadu v konkrétnej aplikácii závisí aj od požiadaviek praxe, od účelu, na ktorý má model slúžiť, t. j. **ktorú časť je potrebné najtesnejšie modelovať**
- na dosiahnutie čo najlepšej zhody je vhodné použiť v konkrétnom prípade viaceré metódy, **výsledky vzájomne porovnať**, prípadne výsledky jednej metódy použiť ako vstupné hodnoty optimalizačných procedúr ďalších metód
- metódy, sú teoreticky, technicky a hlavne časovo náročné a ich výsledky možno **vylepšovať opakovaním procedúr**
- nemožno voliť najvhodnejší odhad kvantilovej funkcie podľa vlastností odhadov pri jednotlivých metódach, **dôraz na dôslednú analýzu kvality a vhodnosti získaných kvantilových tvarov**

Fázy kvantilového modelovania



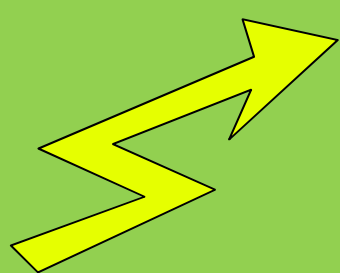
Proces interaktívny, revidujúci predchádzajúce výsledky novými zisteniami s vrátením sa s novými informáciami k predchádzajúcim fázam, postupom a technikám.

Verifikácia – pojem a zameranie

Process rozhodovania o tom, či identifikovaný a odhadnutý model je primeraný a platný pre empirické dáta a uvažovanú situáciu

- **fáza hodnotenia kvality** (kvantilových) modelov, ktorá zahŕňa:
 - porovnania s empirickými údajmi v detailoch
 - zváženie primeranosti funkčného tvaru pre nadväzujúce analýzy
 - porovnanie s predchádzajúcimi výsledkami a zisteniami (sa známymi a v praxi overenými informáciami)

Všeobecné požiadavky na QM



- **„dobrá zhoda“** - kvalitne modelovať konkrétne empirické dáta (testy dobrej zhody rozdelení, grafické, numerické porovnania)
- **jednoduchosť** – môže byť päť a viacparametrický
- **vhodnosť pre vyjadrenie charakteristík** rozdelenia
možnosť porovnať krivky s rovnakým analytickým vyjadrením, s rovnako definovanými charakteristikami v časových alebo prierezových analýzach
- **viacnásobný prístup** v analýzach - smerujúci k systému viacparametrických funkcií (modelovanie v podsúboroch - prierezové analýzy z rôznych hľadísk)
- **jednoduchá aktualizácia** v budúcnosti aj možnosť transformácií prehlbujúcich elasticitu
- aplikovateľnosť **simulačných metód**

Stratégia Q-modelovania

Odhadnúť viaceré modely a porovnávaním
vybrať najvhodnejší:

- je výhodné navrhnuť **viac analytických tvarov** rozdelení, ktoré spĺňajú všeobecné požiadavky a možno ich považovať za vhodné podľa predbežných porovnaní
- odhadnúť ich parametre aplikáciou **viacerých estimačných metód**, prípadne **aj ich kombinovaním** v minimalizačných procedúrach
- pri výbere najlepšieho analytického tvaru spomedzi nich je žiaduce použiť **viaceré techniky hodnotenia kvality** a vhodnosti štatistických modelov



Stratégia 2

- Prvý krok:
odhadnúť „odskúšaný“ komplexný elastický tvar (použiť známe **zovšeobecnené tvary**)
- 2. Až po neúspechu:
kombinovať jednoduché kvantilové tvary pomocou matematického aparátu a dať im príslušnú váhu v častiach rozdelení (skladať **kvantilový tvar „na mieru“**)

Existujú výnimky: v rozsiahlych analýzach, keď modelujeme v podsúboroch s významnou históriou v danej oblasti (ak už známe možné tvary - je vhodné priamo skladať odskúšané rozdelenia)

Problémy pri verifikácii

- **nedostatočná zhoda** teoretického modelu s empirickým rozdelením náhodného výberu (takmer určite zhoda s ktorýmkoľvek ďalším náhodným výberom by bola ešte horšia) - model nespĺňa podmienku dostatočnej reprezentatívnosti. (Testy Chí-kv. a KS test pri veľkých výberoch takmer vždy znamenajú neúspech)

- pri modelovaní rozdelenia s dlhým koncom príčinou významného skreslenia výsledkov



- **korelácia hodnôt** (posúdenie intenzity korelácie vstupných dát a jej vplyvu na odhad parametrov modelu)
- **heteroskedasticita**, t. j. väčšia variabilita hodnôt z koncov rozdelenia (testy známe z teórie ČR)

Ako nakladať z autokoreláciou v grafoch?

V knihe Gilchrista z roku 2000 str.227 časť s názvom:

„Unit exponential spacing control chart“

Aplikácia transformácie cez exponenciálne rozdelenie:

$$y_r = -\ln\left[1 - F\left(x_{(r)}\right)\right]$$

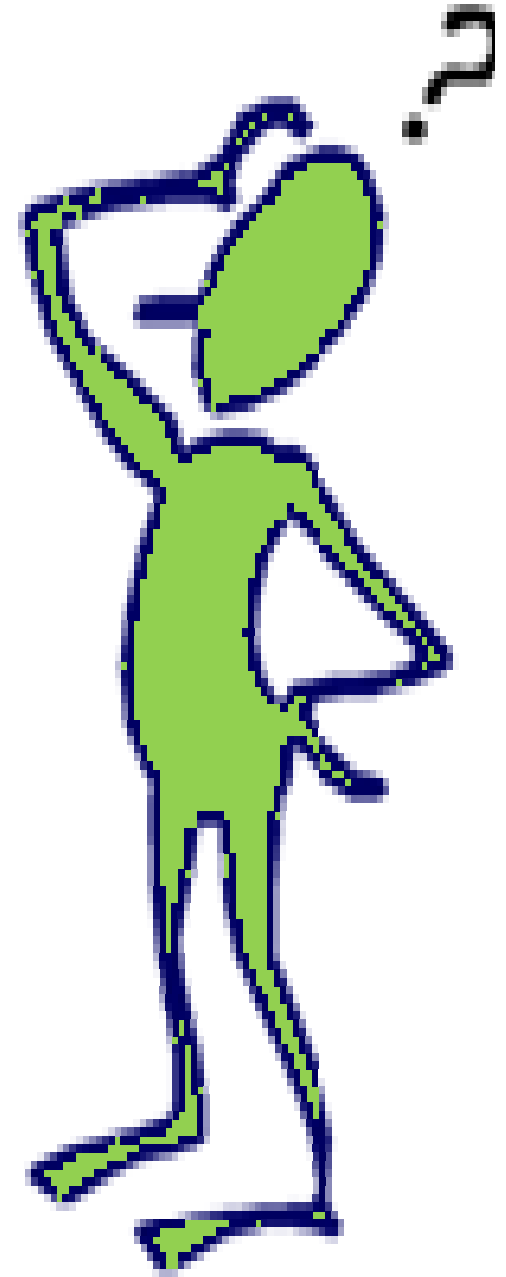
- vypočítaj „medzere“ podľa:

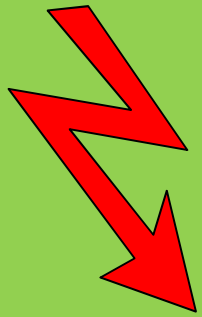
$$v_{(r)} = (n + 1 - r)(y_{(r)} - y_{(r-1)}), \quad y_{(0)} = 0$$

- nezávislé „medzere“ jednotkového exponenciálneho rozdelenia možno znázorniť oproti poradiu ***r***

Ako postupovať pri verifikácii?

- odporúčajú sa testy modelu **na viacerých výberových súboroch** dát (pri veľkých súboroch deliť súbor na časť pre estimáciu a časť pre verifikáciu)
- **grafické a kvantitatívne porovnanie** odlišností medzi teoretickým a empirickým rozdelením
 - charakteristík rozdelení
 - ich grafických znázornení





Nepoužívať simulované náhodné hodnoty jednotlivých modelov

Podľa odporučení Gilchrista používať - **rady odhadov mediánového rankitu rozdelenia populácie:**

- vypočítať presné hodnoty mediánových rankitov odhadnutých tvarov, t. j. odhady mediánov rozdelení n -usporiadaných štatistík podľa jednotlivých odhadnutých kvantilových funkcií

Metódy porovnávania rozdelenia empirických hodnôt s rozdelením mediánového rankitu modelu by mali byť uprednostnené pred metódami porovnávania empirických hodnôt so simulovanými hodnotami odhadnutých modelov.

Prečo mediánový rankit?

- predstavujú rady prostredných hodnôt teoretických rozdelení kvantilov pre každý odhadnutý kvantilový tvar
- predstavujú rady odhadov mediánového rankitu rozdelenia populácie
- rozsah zodpovedajúci rozsahu výberového súboru
- simulované náhodné hodnoty jednotlivých modelov majú príliš veľkú variabilitu – v koncoch najväčší problém, rad stredných hodnôt – mediánov má výrazne menšiu variabilitu – v koncoch stred rozdelení

Ako vypočítať n -teoretických hodnôt



- Mediány rozdelení kvantilov sa dajú vyjadriť a vypočítať **presne pre každý tvar kvantilovej funkcie**, zatiaľ čo stredné hodnoty rozdelení kvantilov často len pomocou aproximácií
- Vstupnými **p-hodnotami** do vzťahov pre tvary teoretických rozdelení by teda mali byť výberové p-hodnoty mediánového rankitu hľadaného rozdelenia, vyjadrené cez **BETA IN_V** funkcie odhadnutých kvantilových tvarov

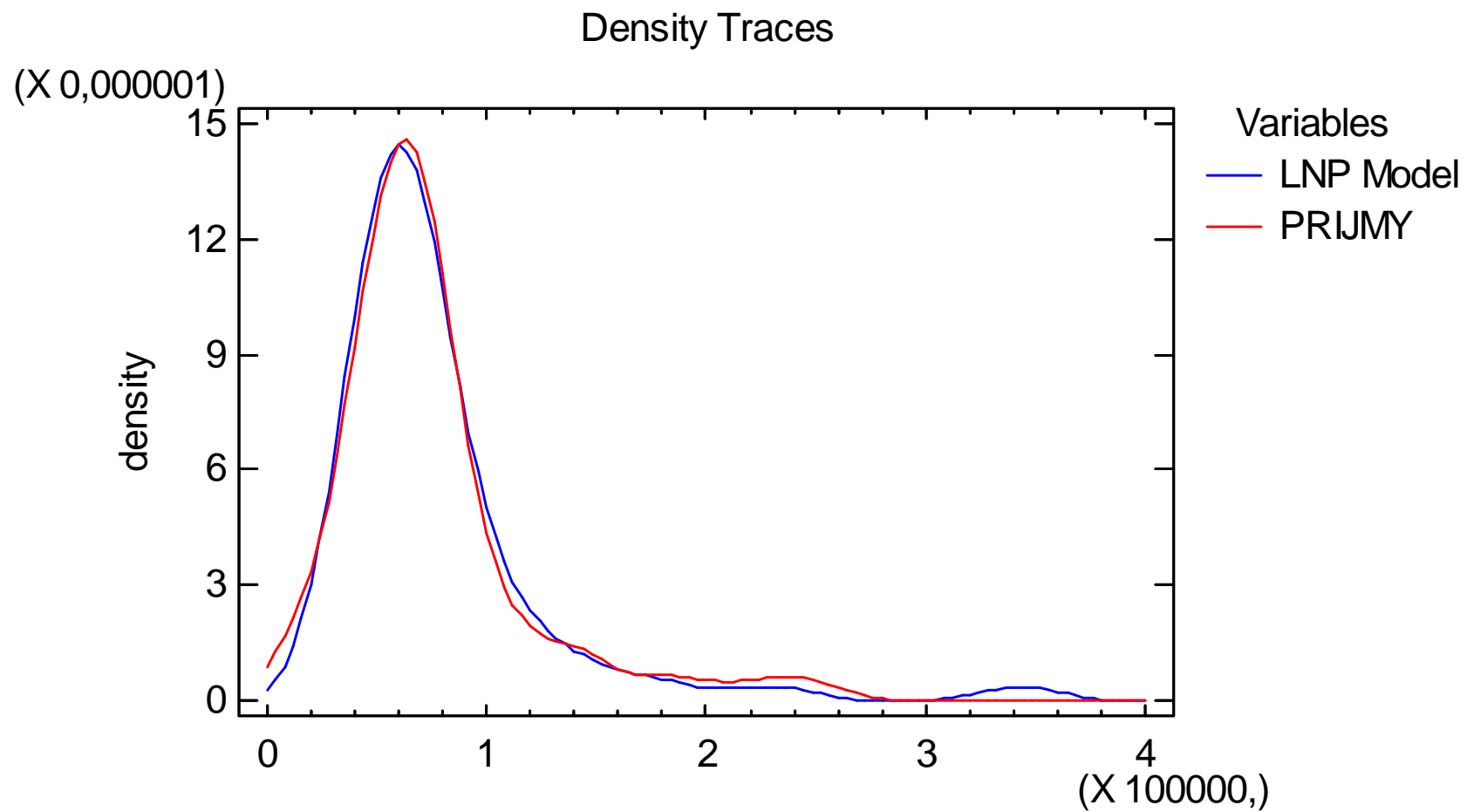
Grafické (vizuálne) analýzy

Vhodné robiť pomocou ponúkaných procedúr programových štatistických, prípadne iných grafických balíkov:

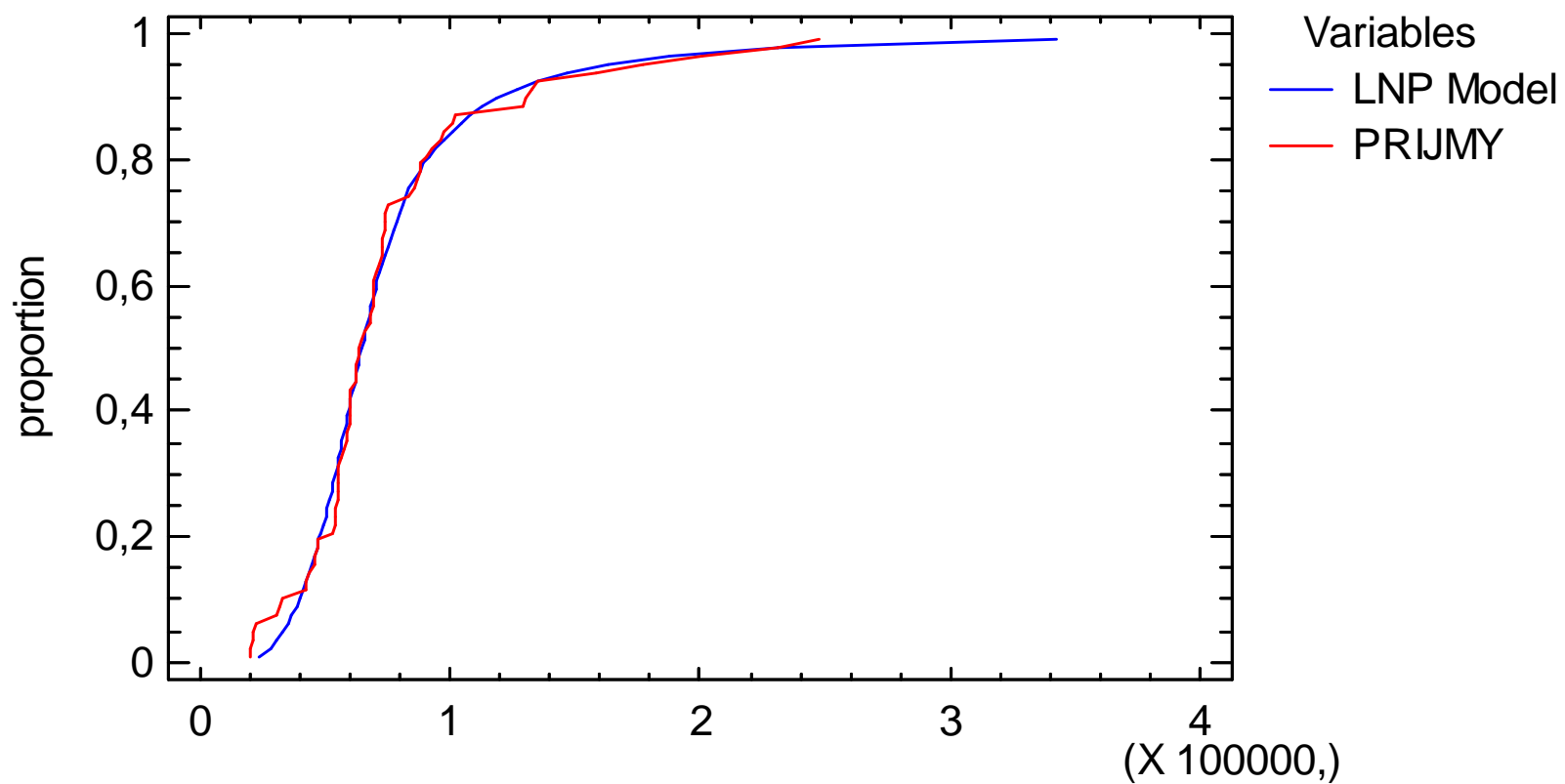
- skladaných grafov funkcií hustôt
- Q-Q grafov
- škatuľkovitých grafov
- grafov rezíduí
- štandardizovaných rezíduí
- grafov kvantilových charakteristík
- porovnanie ich stredových umiestnení pomocou výberových mediánov / priemerov a ich 95 %-ných intervalov spoľahlivosti

a ich porovnaním pre jednotlivé tvary

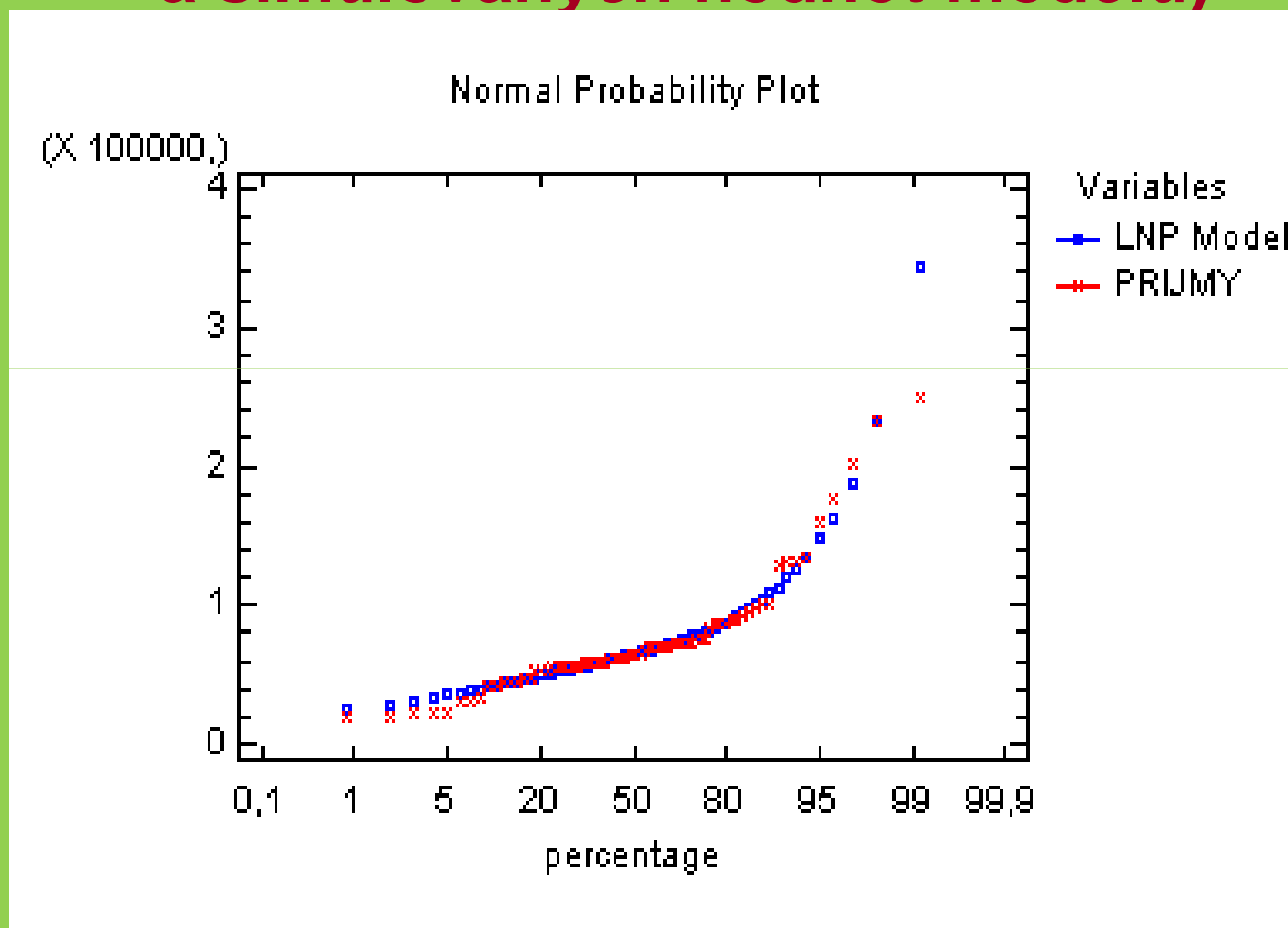
Porovnanie funkcií hustôt



Porovnanie distribučných funkcií

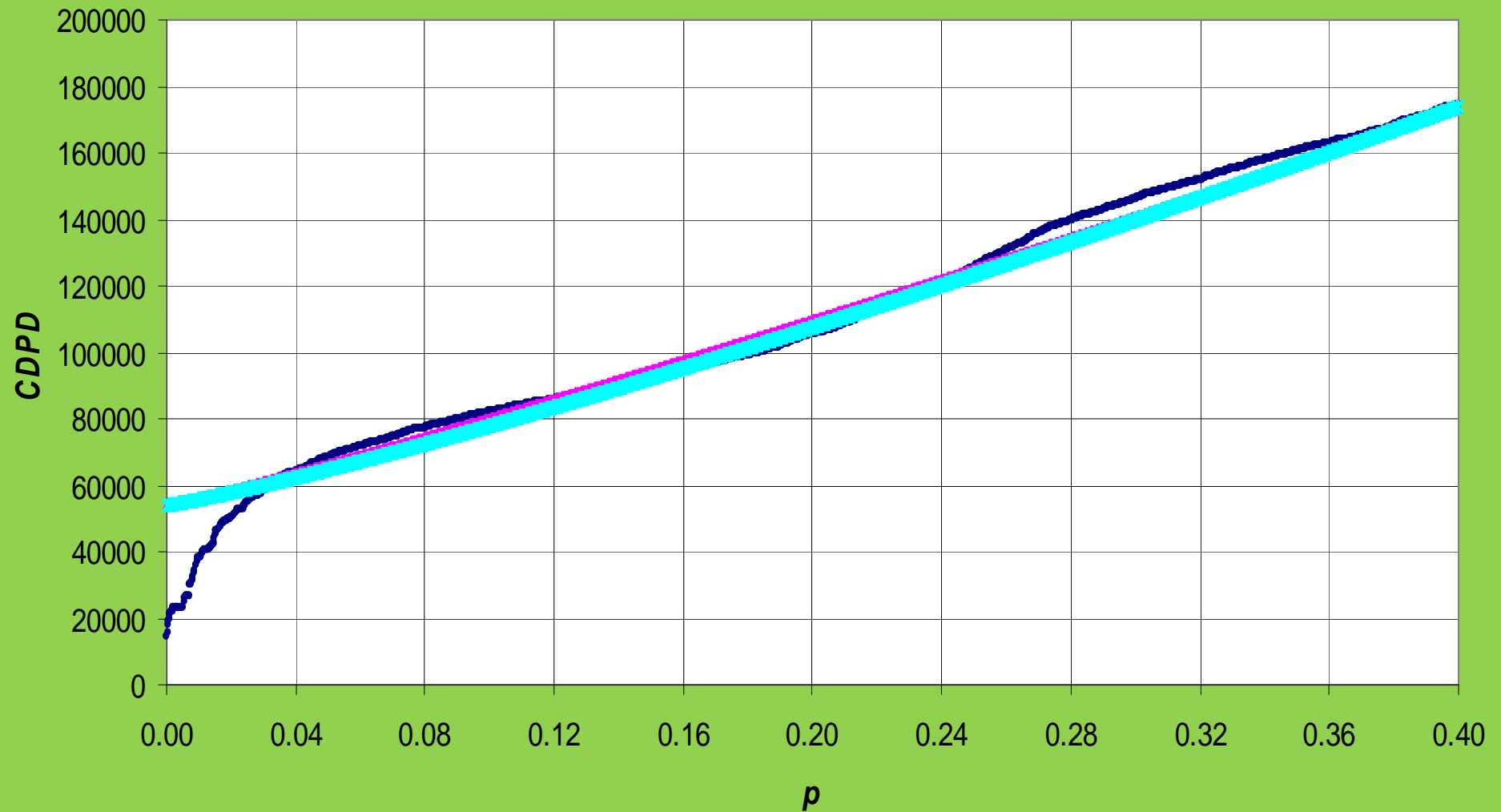


Porovnanie kvantilových funkcií (empirickej a simulovaných hodnôt modelu)



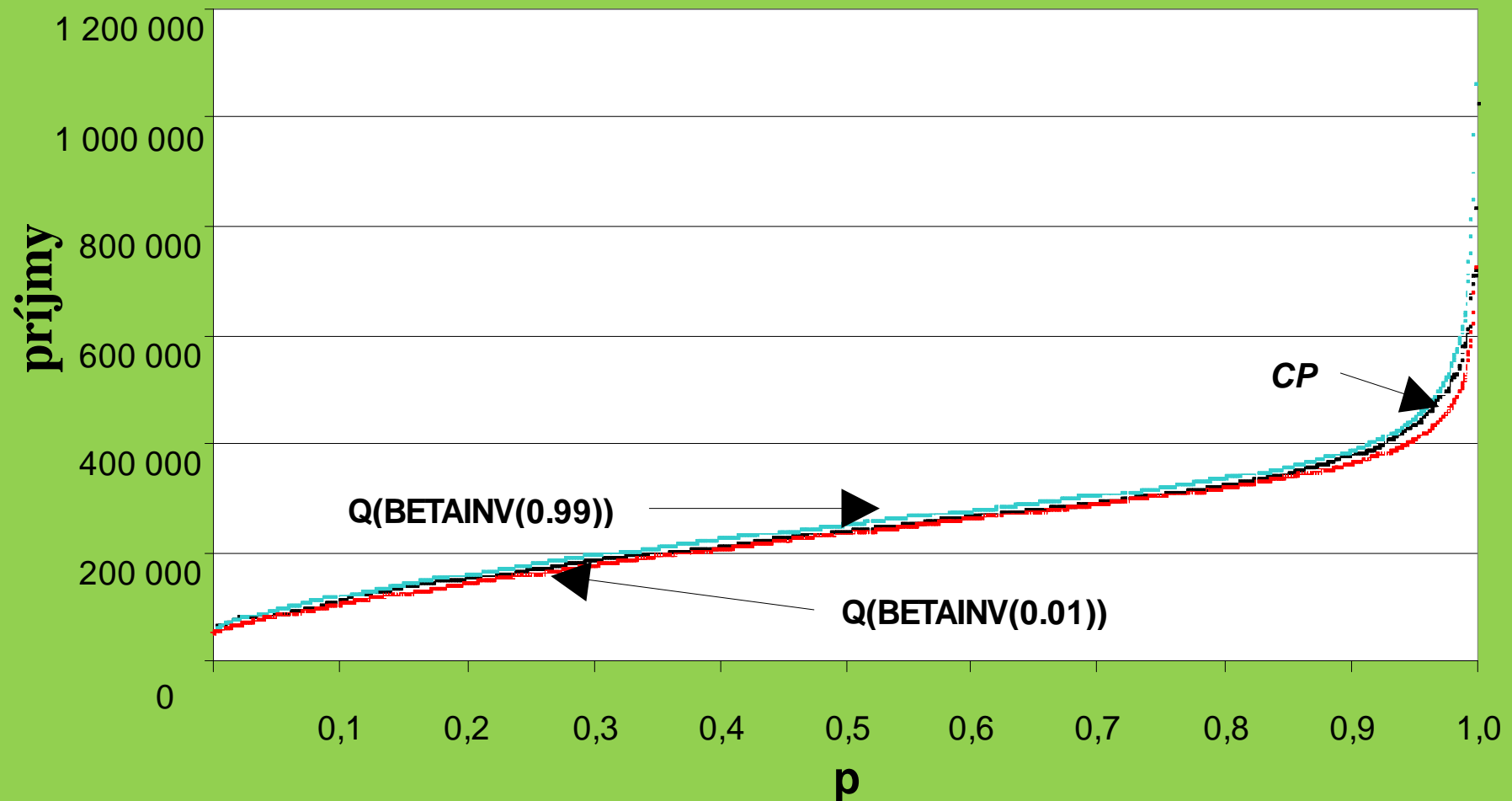
Porovnanie kvantilových funkcií modelu a empirického rozdelenia

• empirické rozdelenie ▲ met Abs. distrib. rezíduí ✕ met Štvorca distrib. rezíduí

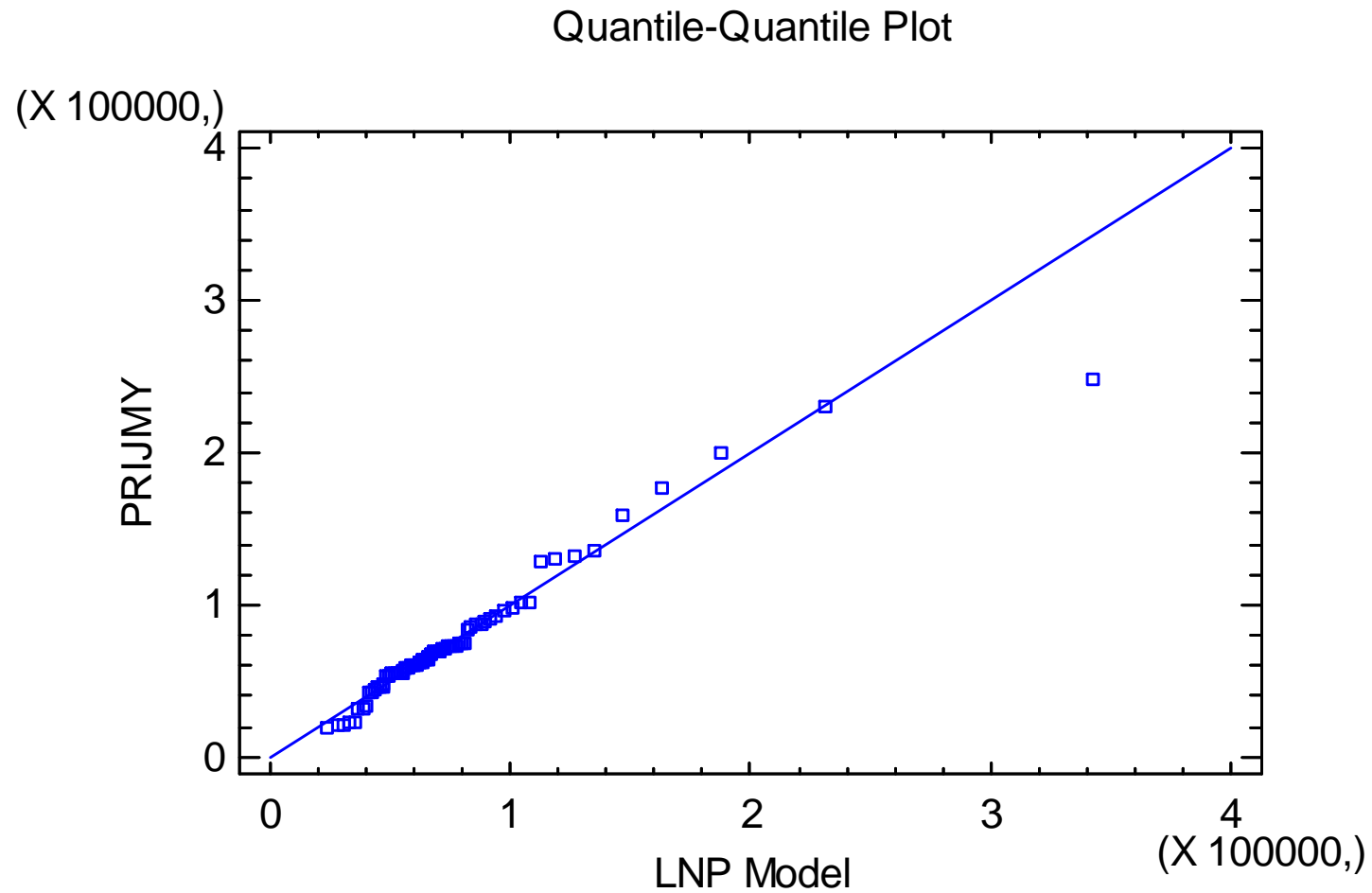


Kvantilové funkcie s ich intervalmi spoľahlivosti

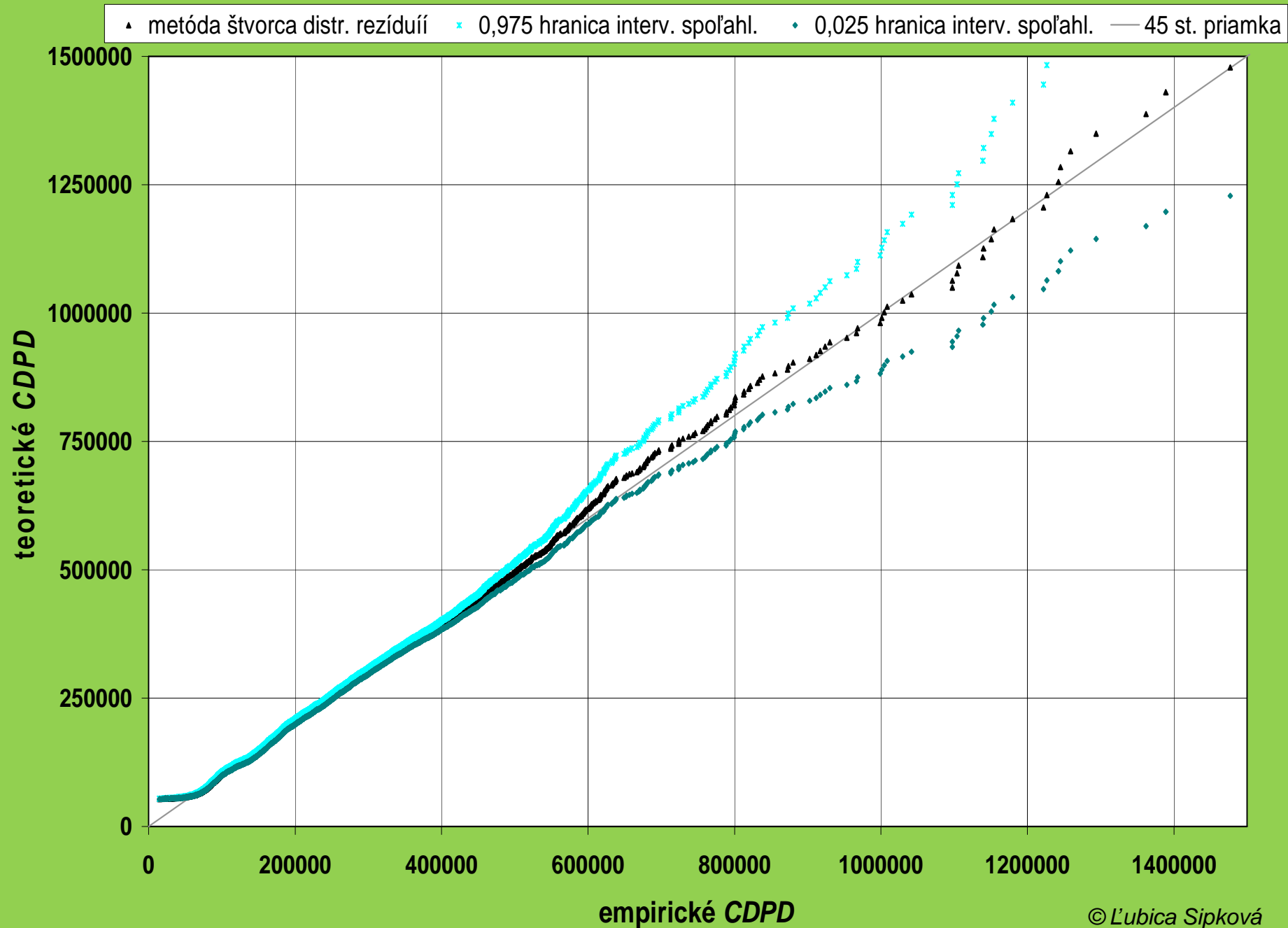
Graf kvantilov CP s 99% hranicami teoretického rozdelenia



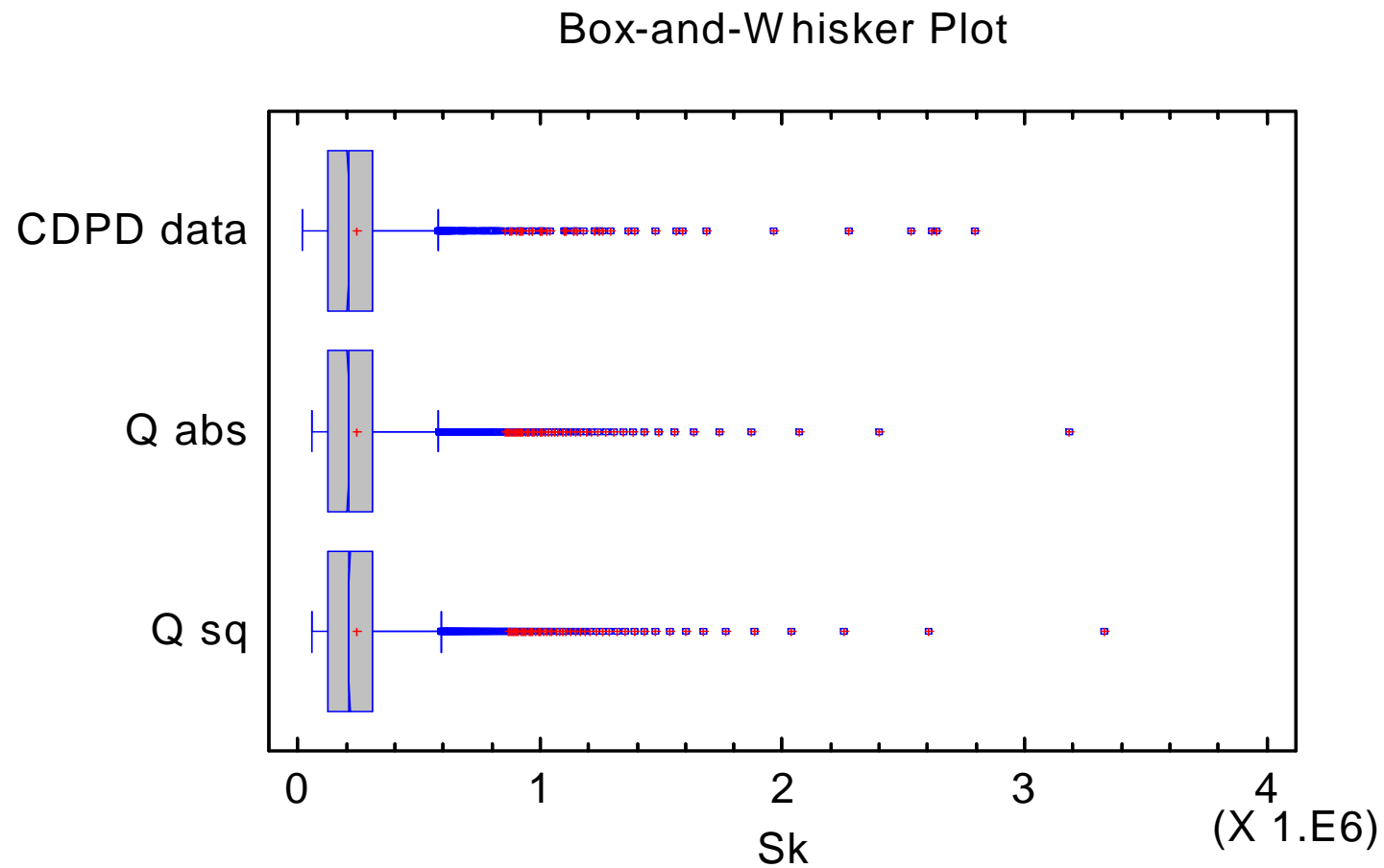
Q-Q graf modelu a empirického rozdelenia



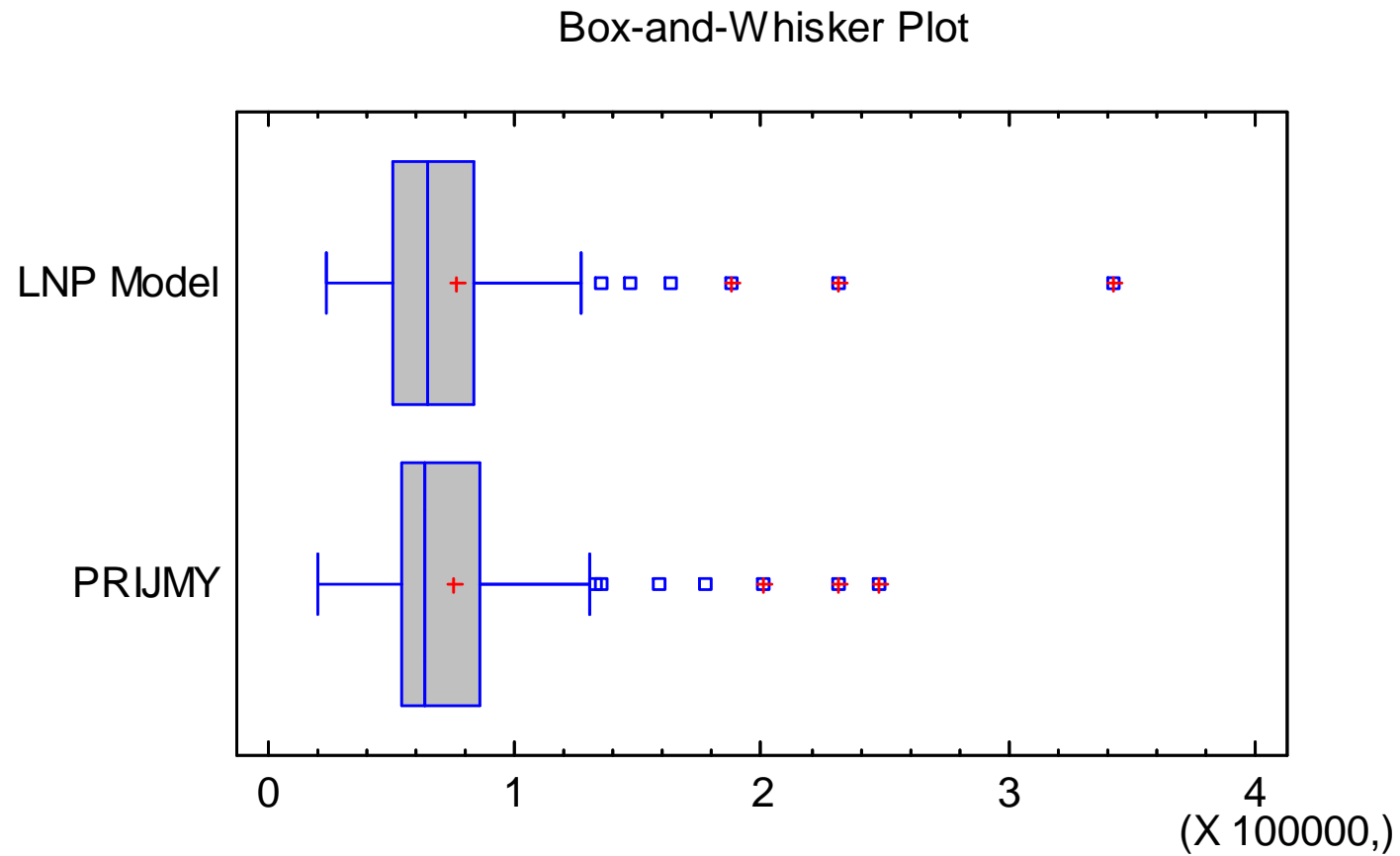
Q-Q graf pre gamma-Paretovo rozdelenie



Porovnanie škatuľkovitých grafov

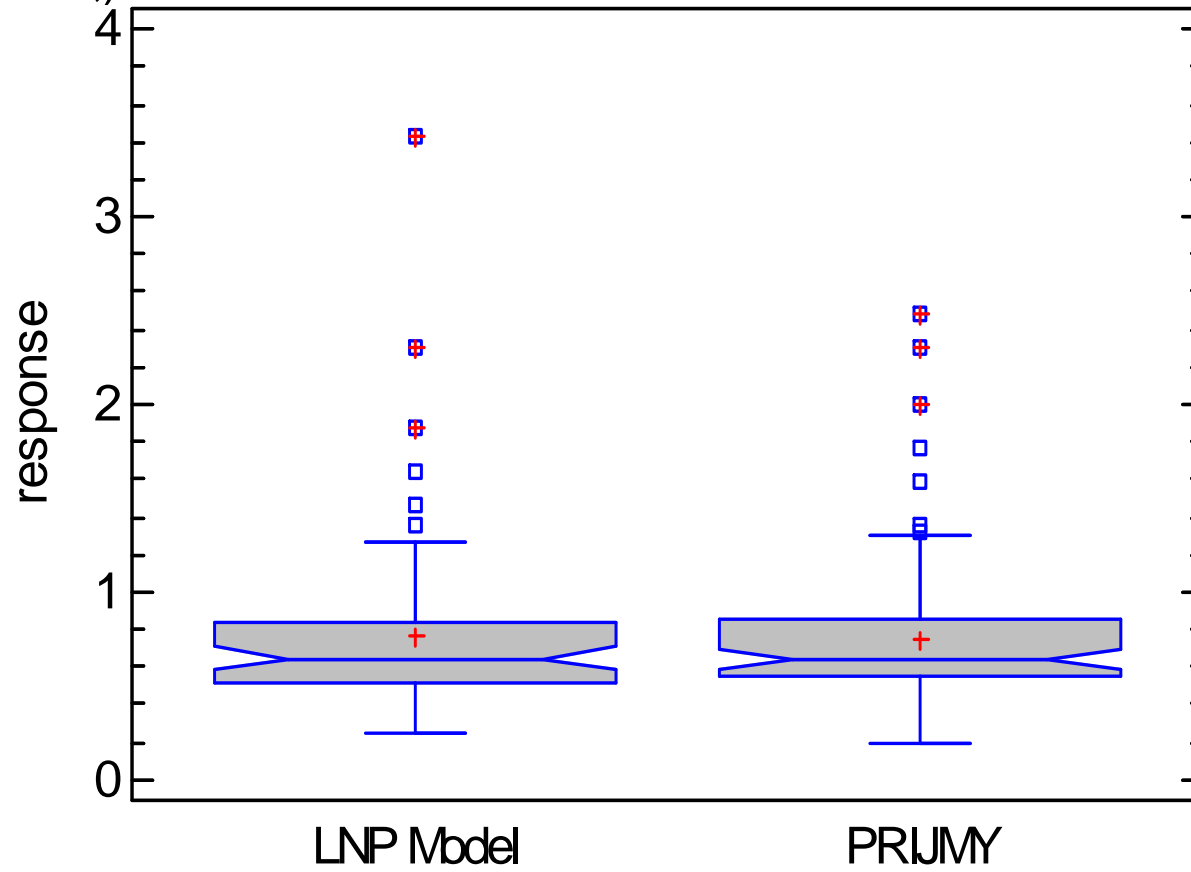


Porovnanie škatuľkovitých grafov



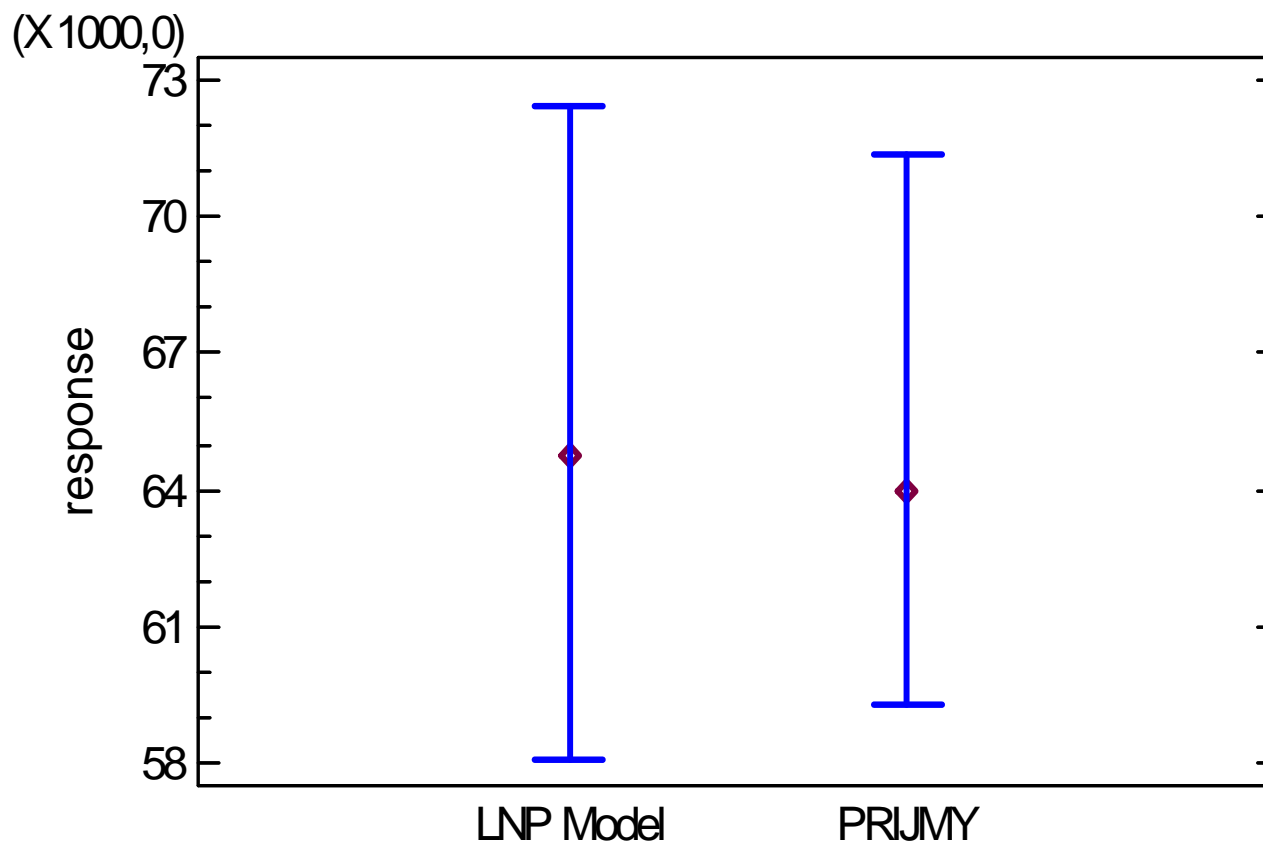
Box-and-Whisker Plot

(X 100000,)

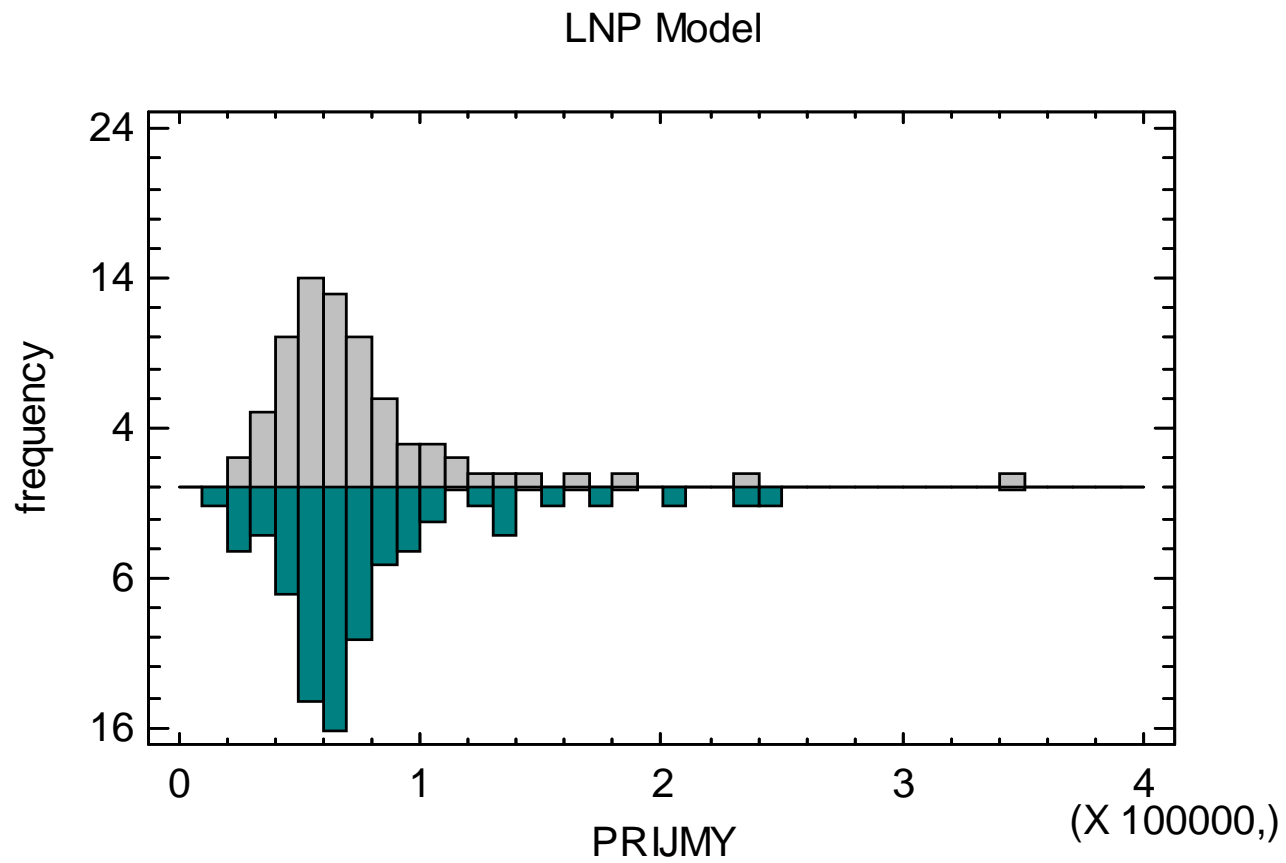


Porovnanie mediánov a ich intervalov spoľahlivosti

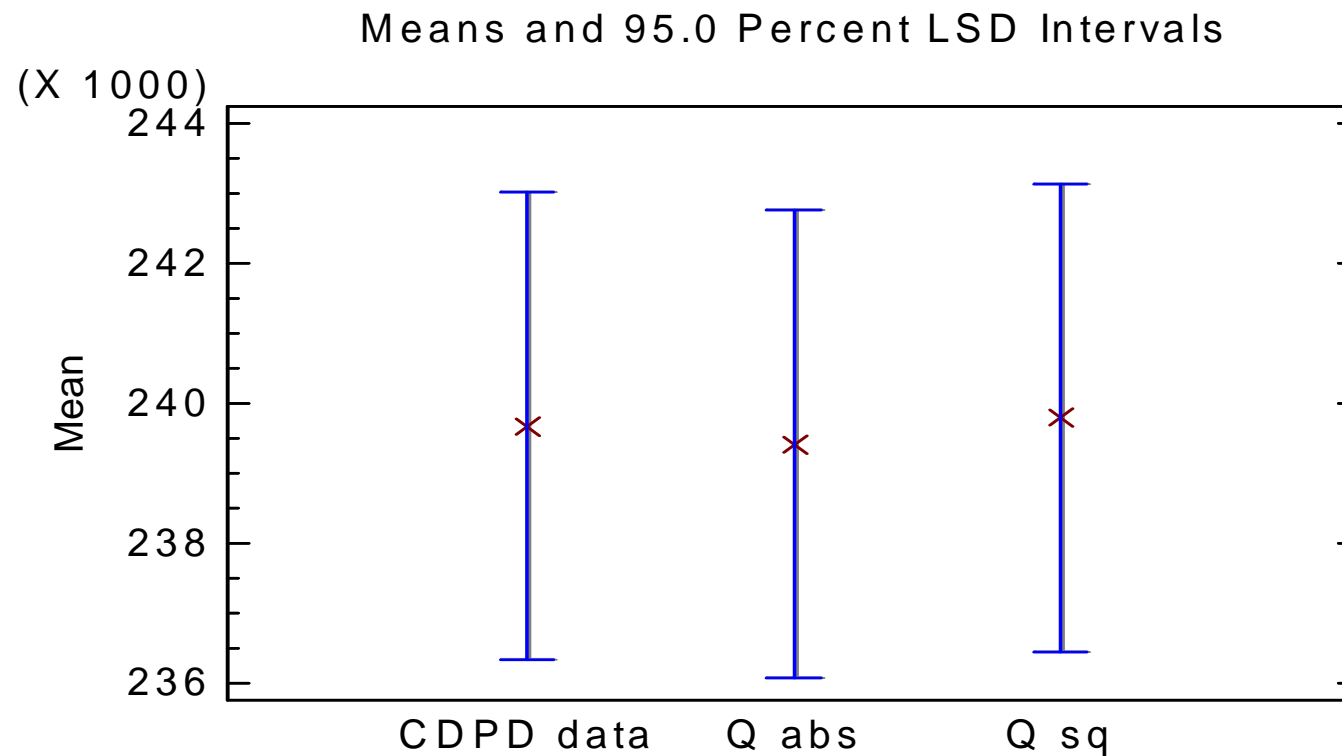
Median Plot with 95,0% Confidence Intervals



Porovnanie histogramov

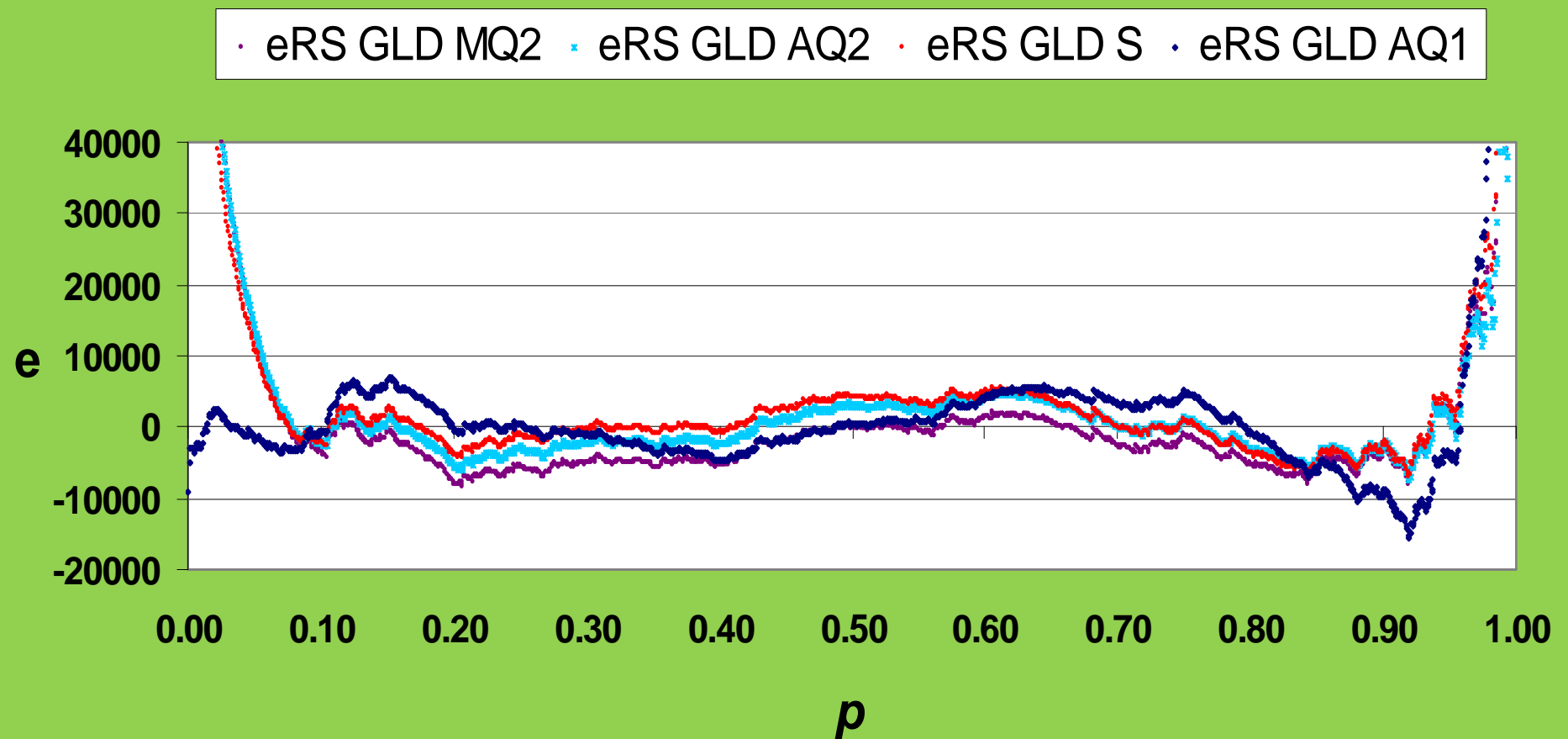


Grafické posúdenie kvality Gamma-Paretových tvarov modelov rozdelenia príjmov domácností SR



Grafy rezíduí modelov a empirického tvaru

Graf rezíduí



Kvantitatívne analýzy

- aplikácia **klasických metód** overenia vhodnosti štatistických modelov pomocou štatistických programových balíkov
- porovnania **základných charakteristík**
- analýzu rozdelenia **rezíduí**
- porovnanie **koeficientov korelácií** na posúdenie miery lineárnej závislosti hodnôt teoretického rozdelenia od empirických hodnôt znázornenej v Q-Q grafoch.

Charakteristiky teoretického a empirického rozdelenia a ich zhody

názov štatistiky	Empirické r.	WPD OD	WPD A
min. hodnota	54325	52532	54938
dolný percentil	65388	61722	64733
dolný decil	107956	108792	111364
dolný kvartil	166663	165241	165621
median	237407	239865	237112
horný kvartil	358324	308089	305196
horný decil	377394	371291	373608
horný percentil	584179	569548	600799
max. hodnota	1020600	1124060	1271774
SSE	x	6.9029E+10	1.6290E+11
rozptyl rezidui		44,080,072	104,020,307
standard.odch.rez		6,639	10,199
SAE	x	6,026,843	3,611,875
priem SAE		3,849	2,306
Výber. Koef. Korel.	x	0.9982	0.9963

Summary Statistics

	<i>Count</i>	<i>Average</i>	<i>Median</i>	<i>5% Winsorized mean</i>	<i>Variance</i>
LNP Model	75	76186,4	64728,2	72794,8	2,22434E9
PRIJMY	75	74836,0	64000,0	72932,4	1,85885E9
Total	150	75511,2	64165,6	72749,0	2,02835E9

□

	<i>Standard deviation</i>	<i>Coeff. of variation</i>	<i>5% Winsorized sigma</i>	<i>MAD</i>
LNP Model	47162,9	61,9046%	33775,2	15278,1
PRIJMY	43114,4	57,6118%	38032,5	10454,6
Total	45037,2	59,6431%	34548,5	14006,6

	<i>Minimum</i>	<i>Maximum</i>	<i>Range</i>	<i>Lower quartile</i>	<i>Upper quartile</i>
LNP Model	24086,6	342632,	318546,	51113,5	84036,9
PRIJMY	19928,0	247992,	228064,	54712,0	85483,9
Total	19928,0	342632,	322704,	53545,4	84036,9

	<i>Interquartile range</i>	<i>1/6 sextile</i>	<i>5/6 sextile</i>	<i>Intersextile range</i>	<i>Skewness</i>
LNP Model	32923,4	45771,0	97313,6	51542,6	3,24576
PRIJMY	30771,9	46134,0	96111,2	49977,2	2,1047
Total	30491,5	45952,5	96712,4	50759,9	2,73489

	<i>Std. skewness</i>	<i>Kurtosis</i>	<i>Std. kurtosis</i>
LNP Model	11,4755	14,4053	25,4653
PRIJMY	7,44124	5,38266	9,51529
Total	13,6744	10,4553	26,1382

Table of Means with Standard error intervals

			<i>Std. error</i>		
	<i>Count</i>	<i>Mean</i>	<i>(individual)</i>	<i>Lower limit</i>	<i>Upper limit</i>
LNP Model	75	76186,4	5445,9	70740,5	81632,3
PRJMY	75	74836,0	4978,42	69857,5	79814,4
Total	150	75511,2			

Variance Check

	<i>Test</i>	<i>P-Value</i>
Bartlett's	1,00403	0,441919

Since the the P-value is greater than or equal to 0,05, there is not a statistically significant difference amongst the standard deviations at the 95,0% confidence level.

ANOVA Table

<i>Source</i>	<i>Sum of Squares</i>	<i>Df</i>	<i>Mean Square</i>	<i>F-Ratio</i>	<i>P-Value</i>
Between groups	6,83853E7	1	6,83853E7	0,03	0,8550
Within groups	3,02156E11	148	2,04159E9		
Total (Corr.)	3,02224E11	149			

Since the P-value of the F-test is greater than or equal to 0,05, there is not a statistically significant difference between the means of the 2 variables at the 95,0% confidence level.

Kruskal-Wallis Test

	<i>Sample Size</i>	<i>Average Rank</i>
LNP Model	75	75,5733
PRJMY	75	75,4267

Test statistic = 0,000427373 P-Value = 0,983507

The Kruskal-Wallis test tests the null hypothesis that the medians within each of the 2 columns is the same. The data from all the columns is first combined and ranked from smallest to largest. The average rank is then computed for the data in each column. Since the P-value is greater than or equal to 0,05, there is not a statistically significant difference amongst the medians at the 95,0% confidence level.

Mood's Median Test

Total n = 150

Grand median = 64165,6

<i>Sample</i>	<i>Sample Size</i>	<i>n<=</i>	<i>n></i>	<i>Median</i>	<i>95,0% lower CL</i>	<i>95,0% upper CL</i>
LNP Model	75	37	38	64728,2	58114,9	72412,5
PRJMY	75	38	37	64000,0	59315,5	71344,2

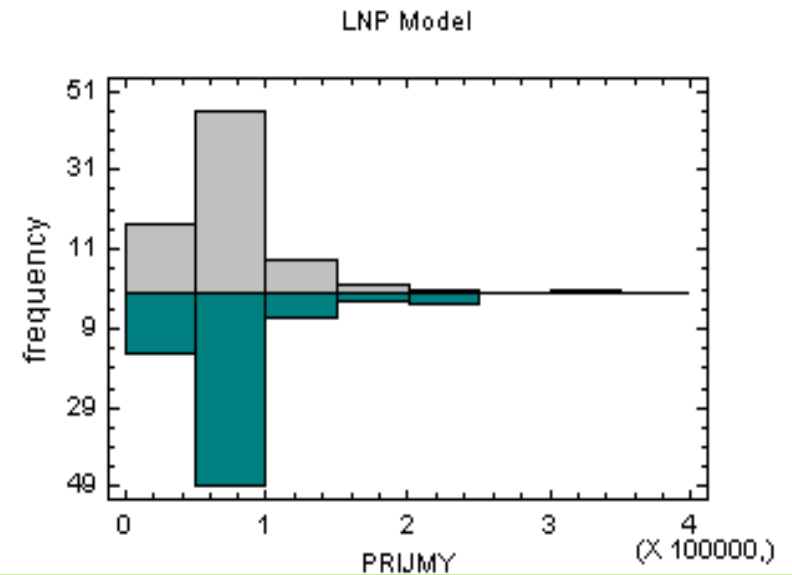
Test statistic = 0,0266667 P-Value = 0,870283

Mood's median test tests the hypothesis that the medians of all 2 samples are equal. It does so by counting the number of observations in each sample on either side of the grand median, which equals 64165,6. Since the P-value for the chi-squared test is greater than or equal to 0,05, the medians of the samples are not significantly different at the 95,0% confidence level.

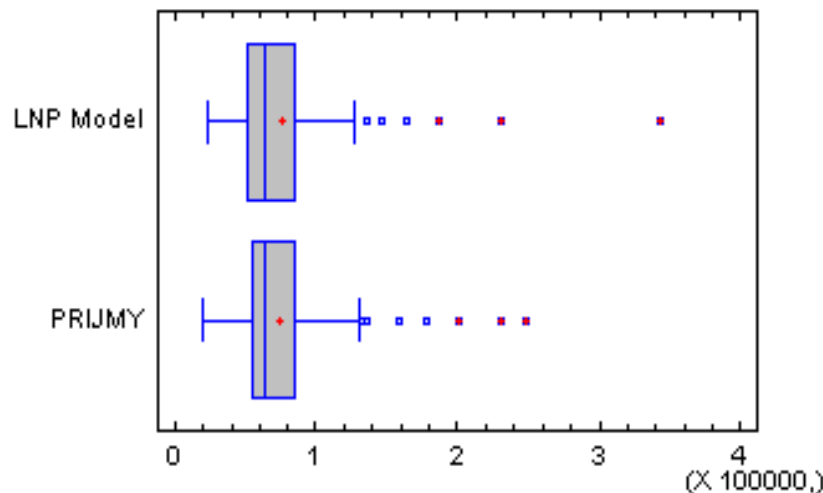
SnapStat: Two Sample Comparison - LNP Model & PRIJMY

SnapStat: Two Sample Comparison

	LNP Model	PRIJMY
Count	75	75
Average	76186,4	74836
Standard deviation	47162,9	43114,4
Coeff. of variation	61,9046%	57,6118%
Minimum	24086,6	19928
Maximum	342632	247992
Range	318546	228064
Std. skewness	11,4755	7,44124
Std. kurtosis	25,4653	9,51529



Box-and-Whisker Plot



95% confidence intervals

Diff. of means: 1350,41 +/- 14580,9 [-13230,5; 15931,3]

Ratio of variances: [0,756003; 1,89404]

Comparison of Means

Null hypothesis: difference = 0

t statistic = 0,183019 Two-sided P-value = 0,8550

Comparison of Sigmas

Null hypothesis: ratio = 1

F statistic = 1,19662 Two-sided P-value = 0,4419

Diagnostics

Shapiro-Wilks P-values = 0,0000 and 0,0000

Lag 1 autocorrelation = 0,7285 +/- 0,2263, 0,8663 +/- 0,2263

**Pravdepodobnostné modelovanie
inverznými distribučnými funkciami:
Modifikačné pravidlá pre kvantilové funkcie**

Spracovanie **siedmej** 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
marec 2010