Katedra štatistiky,
Fakulta hospodárskej informatiky,
Ekonomická univerzita v Bratislave

Pravdepodobnostné modelovanie inverznými distribučnými funkciami : Kvantilová deskriptívna analýza ako východisko ku kvantilovému modelovaniu

Ľubica SIPKOVÁ

Marec 2009

2. z cyklu prezentácií

Východiská kvantilového modelovania

- teória kvantilového pravdepodobnostného modelovania

  (**identifikácie**, estimácie, verifikácie)

- teória „**Order statistics**“ – poriadkových (usporiadaných) štatistík

- vstupné empirické dáta – výberové hodnoty *NP*:
  - **vzostupne usporiadaný empirický súbor údajov spojitej kvantitatívnej náhodnej premennej**
  - **výberové podiely *p***
  - charakteristické črty empirického rozdelenia *NP* odhalené metódami deskriptívnej štatistiky na rôznych základoch

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

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

Východiská grafickej analýzy empirického rozdelenia:

- vstupné empirické dáta – výberové hodnoty *NP*:

  (vzostupne usporiadaný empirický súbor $x_{i:n}$, $i = 1,2,...r,...n$)

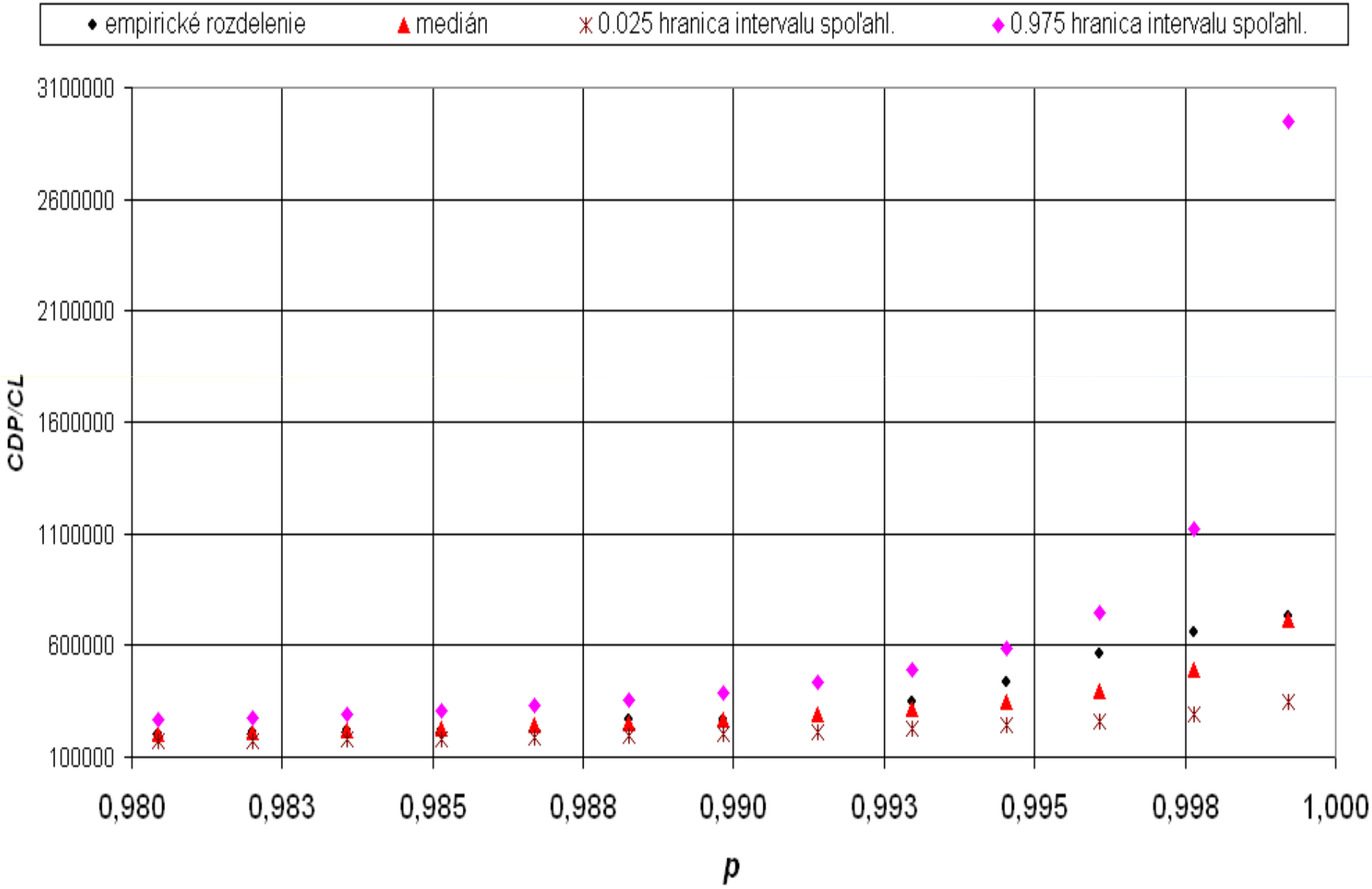
- teória „Order statistics“ – poriadkových (usporiadaných)

  štatistík pre určenie výberových podielov p_r

p_r *r*-tá *p*-hodnota

zodpovedá *r*-tému poradiu zistenej hodnoty x_r

Ako získať pravdepodobnosti p_r ?



Rozdelenie *r*-tej poriadkovej štatistiky

n poriadkových štatistík $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$

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

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

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

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

v jednoduchšom tvare zápisu bez parametrov Q- funkcie:

$$X_{r:n} \sim Q_r\,(p_r\,) = Q\,(\,\mathrm{BETAINV}\,(\,p_r\,,\,r\,,\,n\text{-}r\text{+}1)\,)$$

Dvojaký prístup k určeniu p_r

$$X_{r:n} \sim Q_r(p_r)$$

- **Kvantilový**
  - pomocou mediánového rankitu

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

$p_{r:n}^{*}$ - mediánová p_r hodnota

- **Momentový**
  - pomocou aproximácie rankitu cez **rovnomerné** rozdelenie

$$\mu_{r:n} \approx Q\,[\,E\,(U_{r:n})\,;\,\Theta\,]$$

Mediánový rankit

$X \sim Q[p; \Theta]$ budeme písať v tvare $X \sim Q(p)$

- medián r-tej usporiadanej štatistiky:

$M_{r:n} = Q[p_{r:n}^*; \Theta]$ v tvare $M_{r:n} = Q(p_{r:n}^*)$

kde:

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

$p_{r:n}^*$ - mediánová p_r hodnota

Rankit

- stredné hodnoty usporiadaných štatistík:

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

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

$X \sim Q_U(p)$ je r-tá hodnota rankitu definovaná:

$$\mathrm{E}(U_{r:n}) = \mu_{U;r:n} = r/(n+1)$$

jeho apriximácia: $\mu_{U;r:n} \approx (r-0{,}5)/n$

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

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

$$\mu_{r:n} \approx Q\,[\,(r-0{,}5)/n\,]$$

Usporiadané dvojice hodnôt (x_r, p_r)

- hodnota x je výberovým p-kvantilom

výpočet p momentovou metódou

$\mu_{U;r:n} \approx (r\text{-}0{,}5)/n$

v EXCEL-i napr.:

r	$CP = x$	$p = (r\text{-}0{,}5)/n$
1	54,325.00	0.000319
2	58,625.00	0.000958
3	60,997.00	0.001596
4	61,458.00	0.002235
5	61,798.00	0.002874
6	61,965.00	0.003512
7	61,998.00	0.004151
8	62,078.00	0.004789
9	62,895.00	0.005428
10	63,630.00	0.006066
11	64,387.00	0.006705

Usporiadané dvojice hodnôt (x_r, p_r)

- hodnota x je výberovým p-kvantilom

výpočet p kvantilovou metódou

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

v EXCEL-i napr.:

	A	B	C
1	r	n-r+1	BETAINV(0.5,r,n-r+1)
2	1,00	5 103,00	0,00013582
3	2,00	5 102,00	0,00032887
4	3,00	5 101,00	0,00052398
5	4,00	5 100,00	0,00071954
6	5,00	5 099,00	0,00091527
7	6,00	5 098,00	0,00111107
8	7,00	5 097,00	0,00130692
9	8,00	5 096,00	0,00150279
10	9,00	5 095,00	0,00169868
11	10,00	5 094,00	0,00189459
12	11,00	5 093,00	0,00209050
13	12,00	5 092,00	0,00228642
14	13,00	5 091,00	0,00248234

Štatistické prístupy k pravdepodobnostnému modelovaniu

- **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.*

Druhy kvantilových deskriptívnych grafov

Usporiadané dvojice hodnôt (x_r, p_r)

možno znázorniť graficky (v súlade s definovaním funkcií):

1. p **v závislosti od** x **(DF)**
2. x **v závislosti od** p **(QF)**
3. Dp/Dx **vzhľadom na stred príslušnej diferencie** Dx **(**f**)**
4. Dx/Dp **vzhľadom na stred** Dp **(**q**)**
5. Dp/Dx **nie vzhľadom na** x**, ale vzhľadom na** p

1. *p* v závislosti od *x*

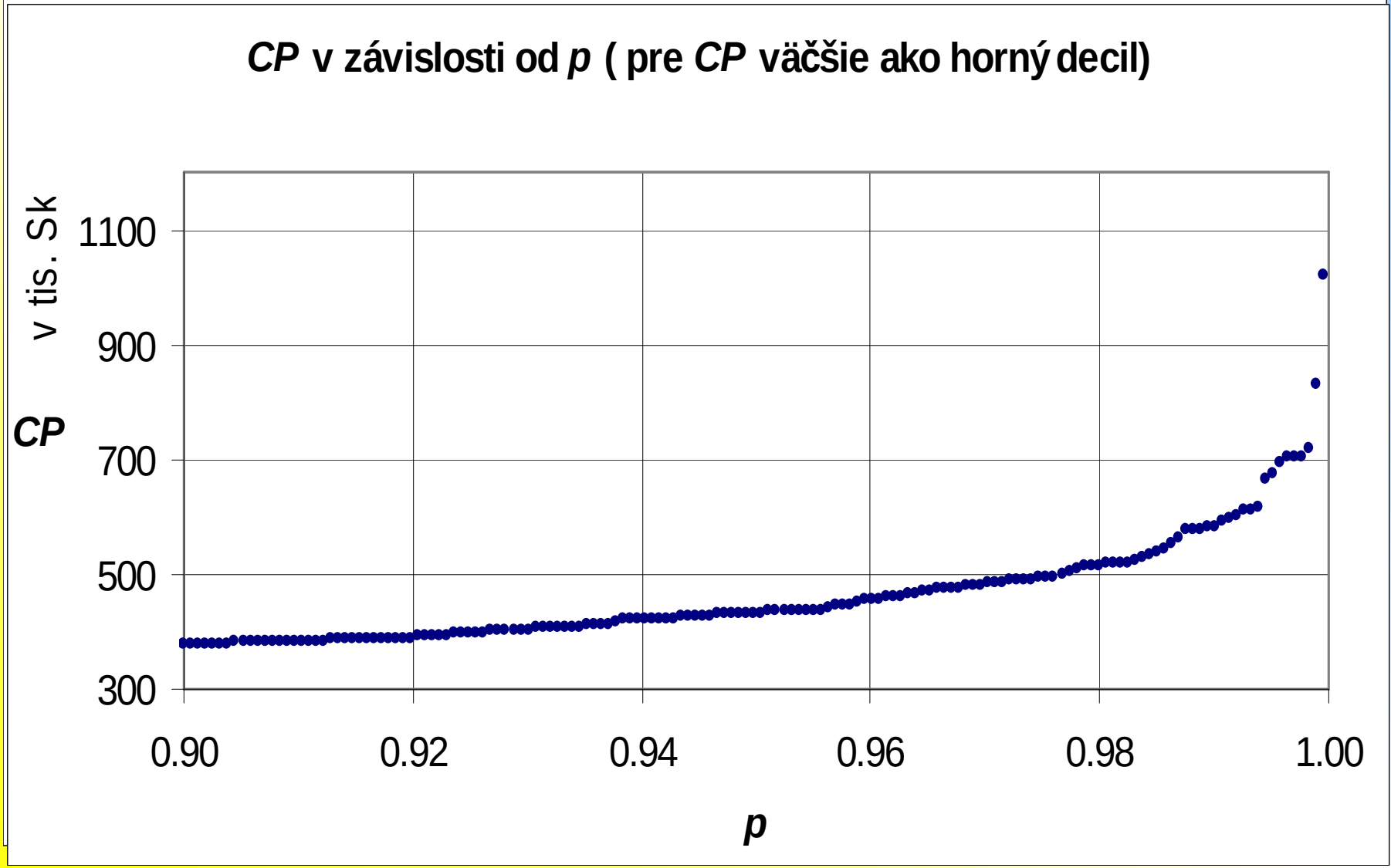
Príjmy domácností SR v roku 2003

$F_{emp}(x) = p$

2. *x* v závislosti od *p*

Príjmy domácností SR v roku 2003

$Q_{emp}(p)=x$



Diferencie *Dx* a *Dp* a z nich vychádzajúce výpočty

Napr. v EXCEL-i:

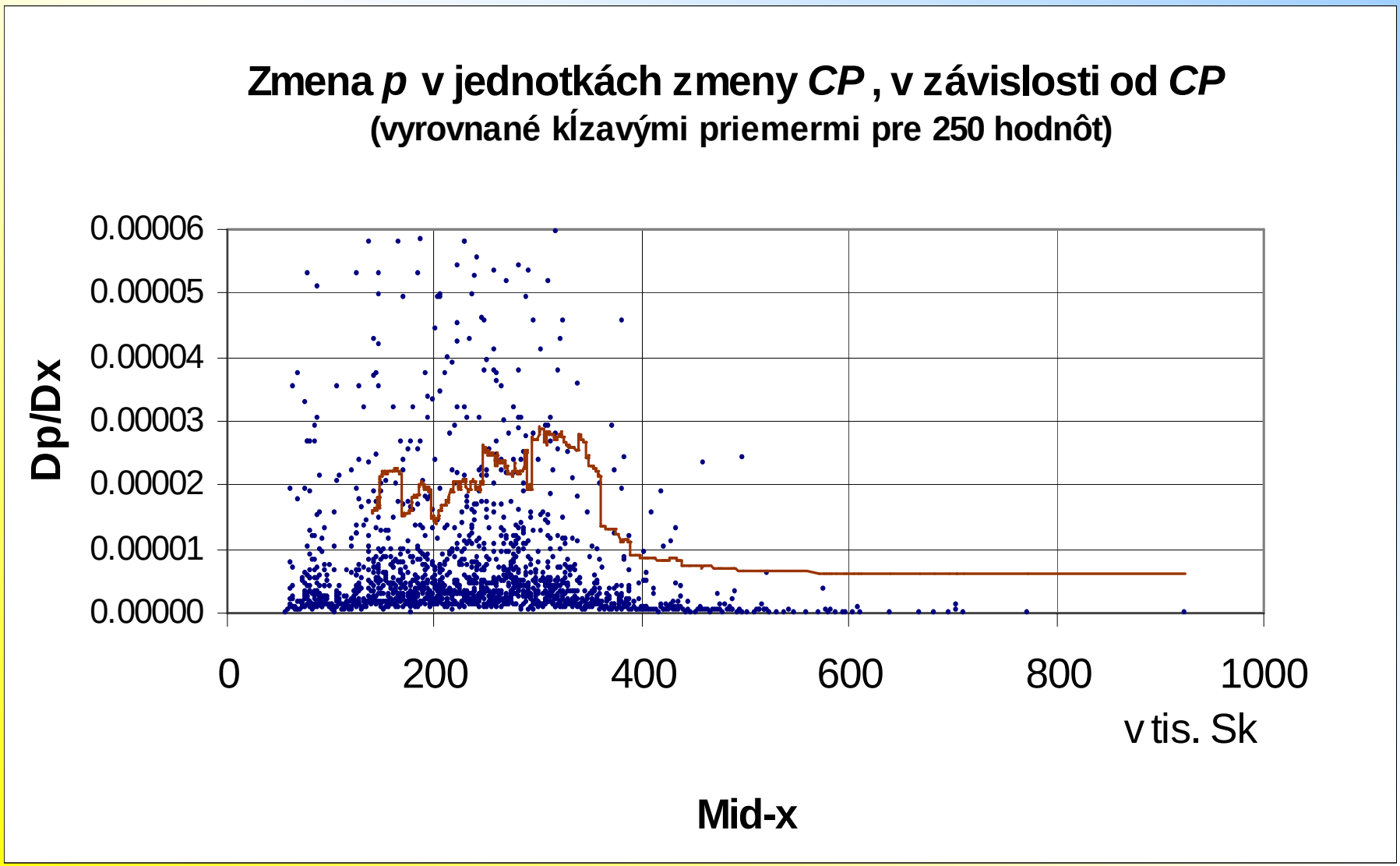
r	*CP = x*	*p = (r-0,5)/n*	*Dx*	*Dp*	*Mid-x*	*Mid-p*	*Dx/Dp*	*Dp/Dx*
1	54,325.00	0.000319	x	x	x	x	x	x
2	58,625.00	0.000958	4,300.00	0.000639	56,475.00	0.000639	6,733,800.00	0.000000148505
3	60,997.00	0.001596	2,372.00	0.000639	59,811.00	0.001277	3,714,552.00	0.000000269211
4	61,458.00	0.002235	461.00	0.000639	61,227.50	0.001916	721,926.00	0.000001385184
5	61,798.00	0.002874	340.00	0.000639	61,628.00	0.002554	532,440.00	0.000001878146
6	61,965.00	0.003512	167.00	0.000639	61,881.50	0.003193	261,522.00	0.000003823770
7	61,998.00	0.004151	33.00	0.000639	61,981.50	0.003831	51,678.00	0.000019350594
8	62,078.00	0.004789	80.00	0.000639	62,038.00	0.004470	125,280.00	0.000007982120
9	62,895.00	0.005428	817.00	0.000639	62,486.50	0.005109	1,279,422.00	0.000000781603
10	63,630.00	0.006066	735.00	0.000639	63,262.50	0.005747	1,151,010.00	0.000000868802
11	64,387.00	0.006705	757.00	0.000639	64,008.50	0.006386	1,185,462.00	0.000000843553
12	64,543.00	0.007344	156.00	0.000639	64,465.00	0.007024	244,296.00	0.000004093395

. . .

3. *Dp/Dx* vzhľadom na stred príslušnej diferencie *Dx*

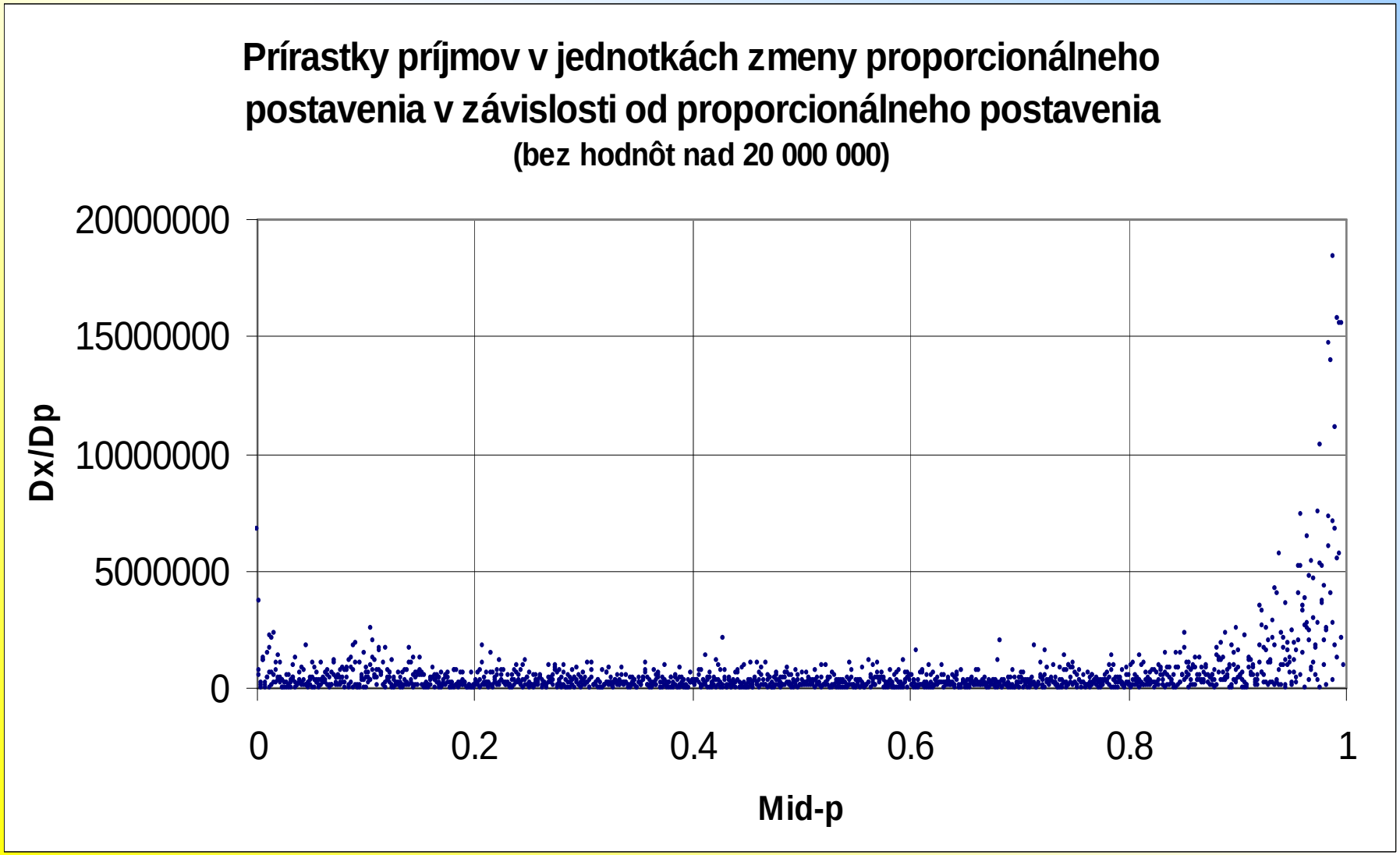
Príjmy domácností SR v roku 2003

$f_{emp}(x)$



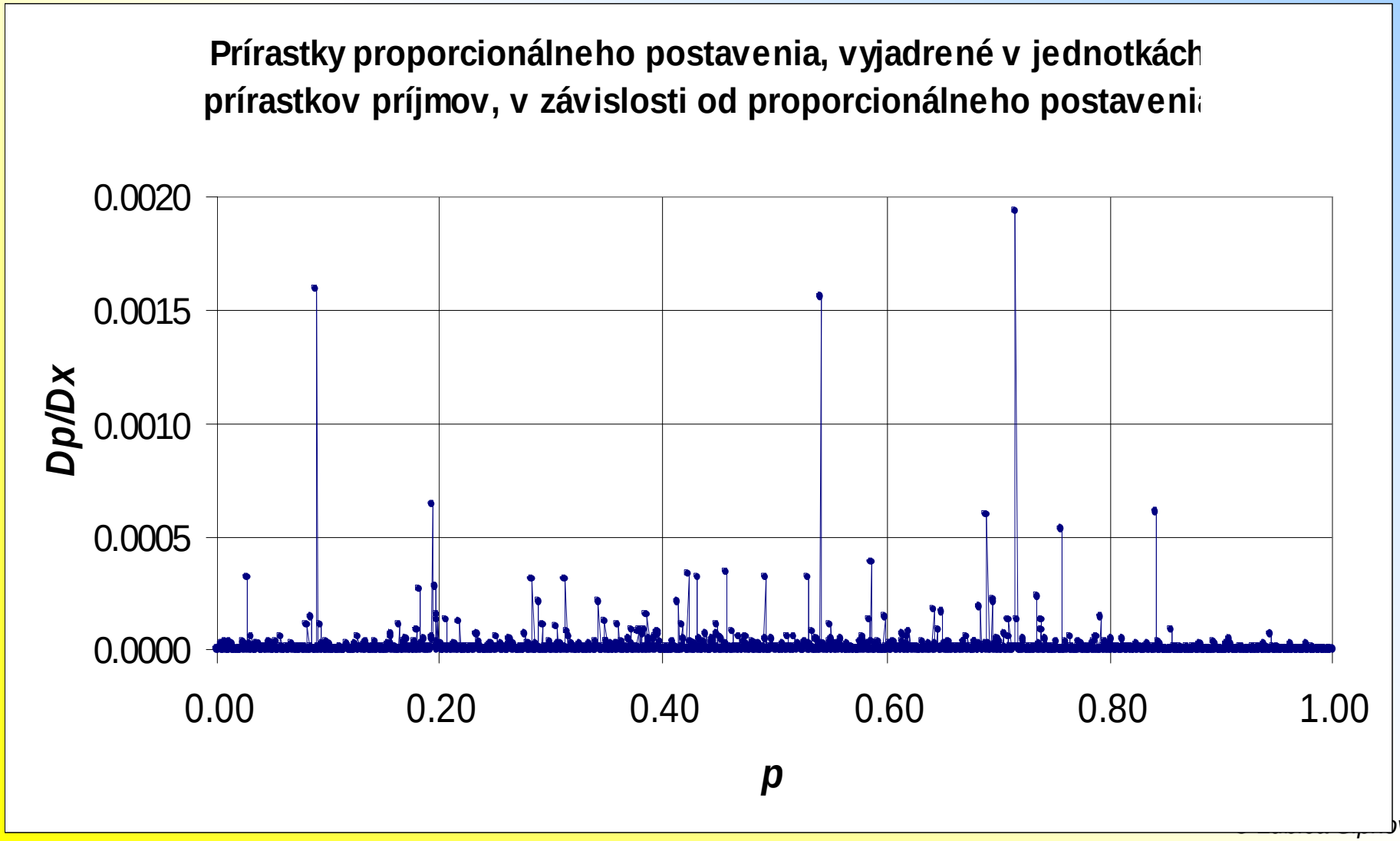
4. *Dx/Dp* vzhľadom na stred príslušnej diferencie *Dp*

Príjmy domácností SR v roku 2003 $q_{emp}(p)$



5. *Dp/Dx* nie vzhľadom na *x*, ale vzhľadom na *p*

$f_{p;emp}(p)$



Vystihnutie tvaru empirického rozdelenia - identifikácia

- Grafická analýza empirického rozdelenia príjmov
- **Kvantitatívna analýza** empirického rozdelenia
  - na momentovom základe
  - **na kvantilovom základe**
- Identifikácia rozdelenia jednoduchými tvarmi
  - Celého rozdelenia
  - Zvlášť pre dolný a horný koniec
- Voľba tvaru váh pre konce rozdelenia
  - Ako funkcia *p*
  - Ako parametre modelu

Najznámejšie kvantilové charakteristiky výberového súboru

- **polohy** (kvartily – dolný lq, horný uq, decily, percentily, t.j. aj medián - m)
- **variability** (variačné rozpätie - vr, kvartilové rozpätie - iqr, kvart. odchýlka – qo)
- **šikmosti** (kvartil. diferencia – qd, kvartilová miera šikmosti – Galtonov koef. šikmosti – $g=qd/iqr$)
- **špicatosti** (? nie je bežne používaná)

Päť výberových charakteristík polohy:

- **minimálna hodnota (*s*),**
- **dolný kvartil (*lq*),**
- **medián (*m*),**
- **horný kvartil (*uq*),**
- **maximálna hodnota (*l*)**

výberové *p*-kvantily pre $p = 0; 0{,}25; 0{,}5; 0{,}75; 1$

sú východiskom výberových charakteristík variability

a Box-Plot-u

Výberové charakteristiky variability:

- **horná kvartilová diferencia** $uqd = uq - m$
- **dolná kvartilová diferencia** $lqd = m - lq$
- **dolná *p*-diferencia** $ld(p) = m - \tilde{Q}(p)$,

  kde $\tilde{Q}(p)$ je p-ty výberový kvantil a $0 \le p \le 0{,}5$
- **horná *p*-diferencia** $ud(p) = \tilde{Q}(1-p) - m$,

  kde $\tilde{Q}(1-p)$ je $(1-p)$ -ty kvantil a $0 \le p \le 0{,}5$
- **kvantilové rozpätie** $ipr(p) = ud(p) + ld(p) = \tilde{Q}(1-p) - \tilde{Q}(p)$
- **kvantilová odchýlka** $qo(p) = ipr(p)/2$
- **mediánová absolútna odchýlka** $\tilde{d} = \text{med}\left| x_i - \bar{x} \right|$

  alebo $\tilde{d} = \text{med}\left| x_i - m \right|$

Výberové charakteristiky šikmosti:

- **kvartilová diferencia** $qd = lq + uq - 2m$
- **Galtonov koeficient šikmosti** $g = qd/iqr$
- **kvantilová diferencia, *p*-diferencia** $pd(p) = ud(p) - ld(p)$
- **Galtonova šikmosť** $g(p) = pd(p)/iqr$

t. j. normovaná kvantilová diferencia pomocou kvartilového rozpätia

- **Galtonov *p*-index šikmosti** $g^*(p) = pd(p)/ipr(p)$

t. j. štandardizovaná kvantilová diferencia pomocou kvantilového rozpätia

- **kvantilový podiel, kvantilová proporcia** $sr(p) = ud(p)/ld(p)$

Výberové charakteristiky špicatosti:

- **index špicatosti** $t(p) = ipr(p)/iqr$, pre $0 \le p \le 0{,}5$
- **horný a dolný index špicatosti** $ut(p) = ud(p)/uqd$

  $lt(p) = ld(p)/lqd$
- **Moorsova špicatosť** k

  vychádza z oktilov $e1, e2, \ldots, e7$

  t. j. kvantilov deliacich súbor na osem častí

  $$k = [(e7 - e5) + (e3 - e1)]/iqr$$
- **horná a dolná špicatosť**

  $$uk(t) = [\tilde{Q}(1 - t) + \tilde{Q}(0{,}5 + t) - 2\tilde{Q}(0{,}75)]/[\tilde{Q}(1 - t) - \tilde{Q}(t)]$$

  kde $0 < t < 0{,}25$

  $$lk(t) = [\tilde{Q}(0{,}5 - t) + \tilde{Q}(t) - 2\tilde{Q}(0{,}25)]/[\tilde{Q}(1 - t) - \tilde{Q}(t)]$$

  kde $0 < t < 0{,}25$

Moorsova sumarizácia o tvare rozdelenia

Štyri miery (m, iqr, g, k) :

1. **medián**
2. **kvartilové rozpätie**
3. **Galtonov koeficient šikmosti**
4. **Moorsova špicatosť**

poskytujú podľa **Moorsa** jednoduchú sumarizáciu o tvare rozdelenia na kvantilovom základe

Zodpovedá **Pearsonovej** štvorčíselnej sumarizácii – priemer, rozptyl, koeficient šikmosti a Pearsonova špicatosť na báze momentov.

Kvantilové deskriptívne štatistiky a grafické znázornenie kvantilov

Ich aplikácia vo všetkých fázach modelovania:

identifikácie

estimácie

verifikácie

Štatistické modelovanie je proces interaktívny, takže postup a výber metód v určitej fáze vychádza zo záverov predchádzajúcej.

Je to nevyhnutné aj preto, že výber metód, napr. estimácie, priamo závisí od predchádzajúcich výsledkov, t.j. vybraného tvaru vo fáze identifikácie.

Pravdepodobnostné modelovanie inverznými distribučnými funkciami: Kvantilová deskriptívna analýza ako východisko ku kvantilovému modelovaniu

Spracovanie druhej 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 2009