Transformace dat

Transformace dat: vede ke stabilizaci rozptylu,
zesymetričtění rozdělení,
k normalitě,

1. ***Stabilizace rozptylu*:** rozptyl $\sigma^2(y)$ lze určit

$$\sigma^2(y) \approx \left[\frac{dg(x)}{dx}\right]^2 f_1(x) = C$$

kde C je konstanta. Hledaná transformace g(x) je řešením diferenciální rovnice

$$g(x) \approx C \int \frac{dx}{\sqrt{f_1(x)}}$$

Instrumentální metody, přístroje: konstantní relativní chyba $\delta(x)$, $\sigma^2(x)$ je dán funkcí $f_1(x) = \delta^2(x)\, x^2 = \text{konst}\, x^2$ a dosazením do $g(x) = \ln x$

2. ***Zesymetričtění rozdělení*** výběru je možné provést jednoduchou mocninnou transformací..

TRANSFORMACE DAT

ORIGINÁLNÍ DATA $x_1, x_2, \dots, x_n$

① MOCNINNÁ
② Box–Cox

TRANSFORMOVANÁ DATA $y_1, y_2, \dots, y_n$

~~PŘÍMO~~

ODHAD POLOHY: $\bar{x}$, $I_D \leq \mu \leq I_H$

ODHAD ROZPTYLENÍ: s^2

RE-TRANSFORMACE

(1) TRANSFORMACE PRO KONSTANTNÍ ROZPTYL: $\sigma^2(x) \neq$ konstanta

STABILIZACE ROZPTYLU $\sigma^2(y) =$ konstanta

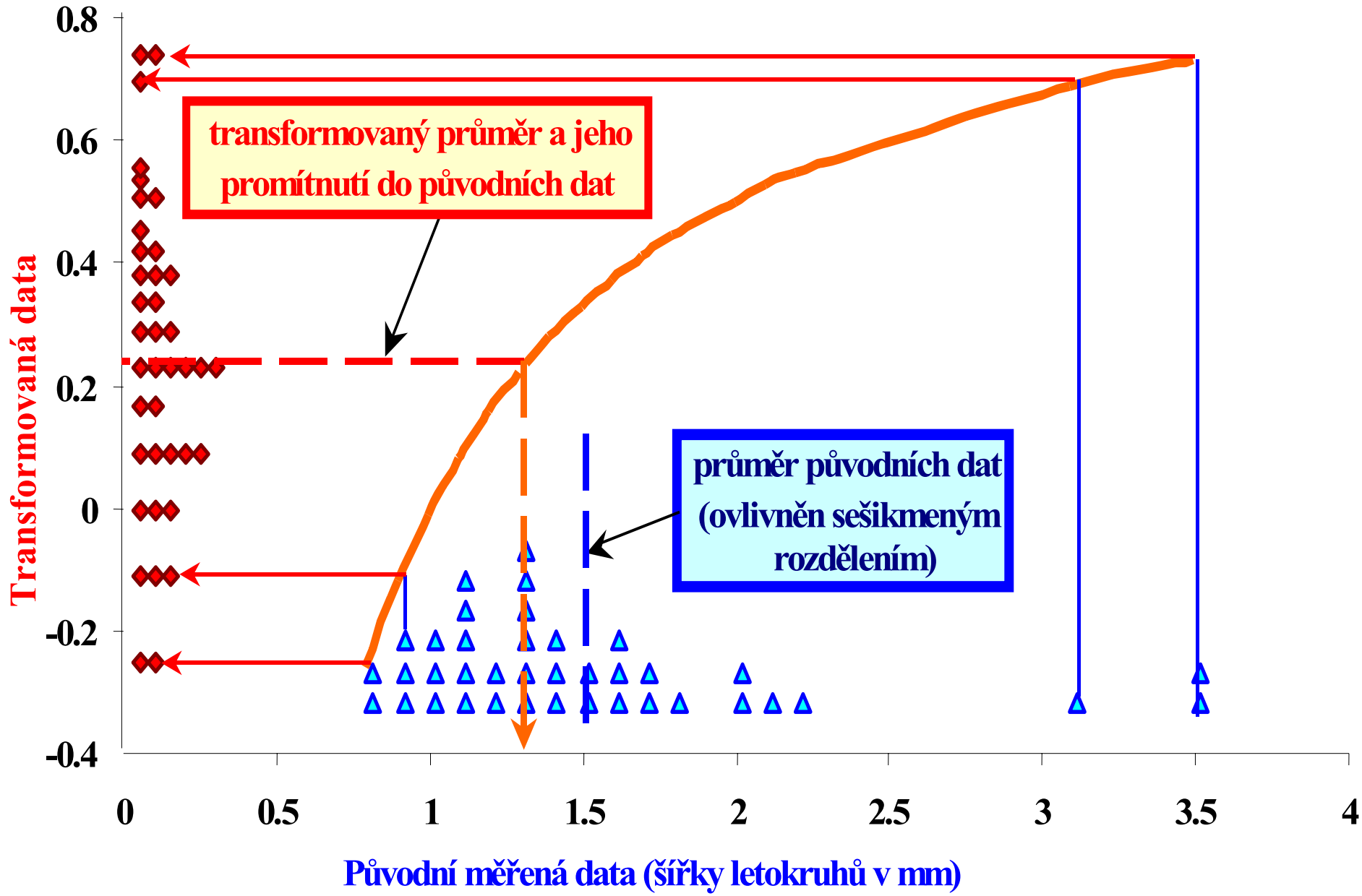
$$\sigma^2(y) \approx \left(\frac{\partial g(x)}{\partial x}\right)^2 \cdot f_1(x) = C$$

(2) TRANSFORMACE PRO SYMETRII:

(a) ZESYMETRIČTĚNÍ ROZDĚLENÍ: mocninné transformace

$$y = g(x) = \begin{cases} x^{\lambda}, & \text{pro } \lambda > 0 \\ \ln x, & \lambda = 0 \\ -x^{\lambda}, & \lambda < 0 \end{cases}$$

TRANSFORMACE DAT - ukázka



Symetrizující mocninná tranformace

zesymetričtění rozdělení výběru provede

$$y = g(x) = \begin{cases} x^{\lambda} & \lambda > 0 \\ \ln x & \text{pro} \quad \lambda = 0 \\ -x^{-\lambda} & \lambda < 0 \end{cases}$$

1. Nezachovává měřítko,
2. Není k λ všude spojitá,
3. Zachovává pořadí dat ve výběru,
4. Hodí se pouze pro kladná data.

Normalizační Box-Coxova transformace

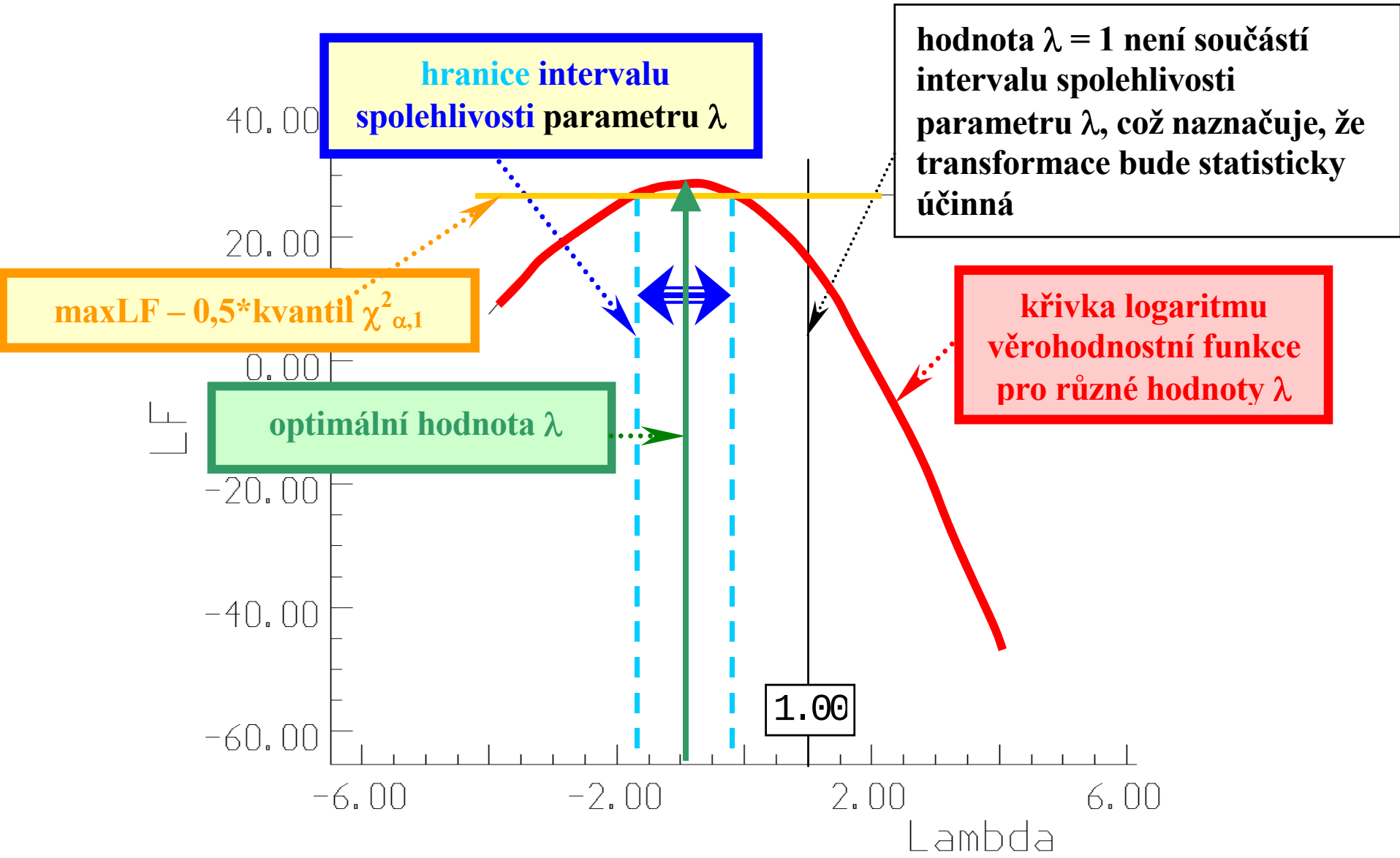
přiblížení rozdělení výběru k normálnímu

$$y = g(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \ln x & (\lambda = 0) \end{cases}$$

Transformace

1. Je vzhledem k λ spojitá,
2. Je použitelná pouze pro kladná data, jinak $(x - x_0)$.

TRANSFORMACE DAT – odhad optimálního λ



Zpětná transformace

$$\bar{y},\ s^2(y),\ \bar{y} \pm t_{1-\alpha/2}(n-1).\ s(y)/\sqrt{n}$$

1. **Nekorektní přístup**:

$$\bar{x}_R = \bar{x}_\lambda = \left[\frac{\sum_{i=1}^{n} x_i^{\lambda}}{n}\right]^{1/\lambda}$$

$\bar{x}_R = \bar{x}_{-1}$ představuje *harmonický průměr*,
$\bar{x}_R = \bar{x}_0$ představuje *geometrický průměr*,
$\bar{x}_R = \bar{x}_1$ představuje *aritmetický průměr*,
$\bar{x}_R = \bar{x}_2$ představuje *kvadratický průměr*.

2. **Korektní přístup:** z Taylorova rozvoje funkce $y = g(x)$ v okolí $\bar{y}$,

$$\bar{x}_R \approx g^{-1}\left[\bar{y} - \frac{1}{2}\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)\right]$$

Pro rozptyl vyjde $s^2(x_R) \approx \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$.

(a) Pro speciální případ $\lambda = 0$, tzn. logaritmickou transformaci typu $g(x) = \ln x$, bude

$$\bar{x}_R \approx \exp\left[\bar{y} + 0.5\, s^2(y)\right]$$

a rozptyl $s^2(x_R) \approx \bar{x}_R^2\, s^2(y)$.

(b) Pro případ $\lambda \neq 0$:

$$\bar{x}_{R,1,2} = \left[0.5(1 + \lambda \bar{y}) \pm 0.5\sqrt{1 + 2\lambda(\bar{y} + s^2(y)) + \lambda^2(\bar{y}^2 - 2\, s^2(y))}\right]^{1/\lambda}$$

a rozptyl $s^2(x) = \bar{x}_R^{-2\lambda+2}\, s^2(y)$.

Rozptyl se vyčíslí

$$s^2(x_R) \approx \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$$

kde derivace jsou vyčísleny v bodě $x = \bar{x}_R$.

100(1 - α)%ní interval spolehlivosti se vyčíslí dle

$$\bar{x}_R - I_D \leq \mu \leq \bar{x}_R + I_H$$

kde

$$I_D = g^{-1}\left(\bar{y} + G - t_{1-\alpha/2}(n-1)\,\frac{s(y)}{\sqrt{n}}\right)$$

$$I_H = g^{-1}\left(\bar{y} + G + t_{1-\alpha/2}(n-1)\,\frac{s(y)}{\sqrt{n}}\right)$$

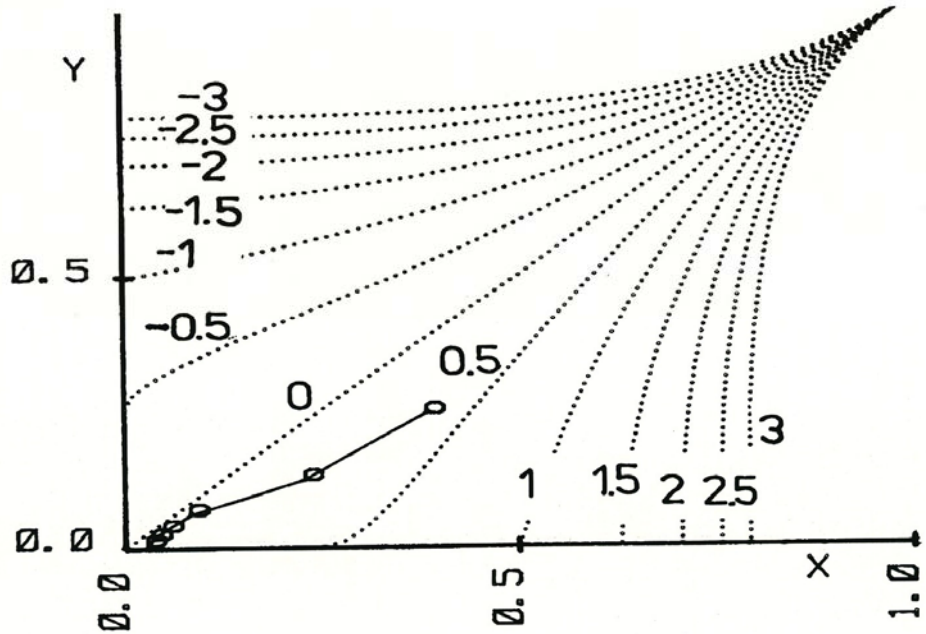
$$G = -0.5\,\frac{d^2g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$$

kde $t_{1-\alpha/2}(n - 1)$ je $100(1-\alpha/2)\%$ní kvantil Studentova rozdělení s (n-1) stupni volnosti.

Hines - Hinesův selekční graf

Osa x: $\tilde{x}_{0.5}/x_{1-P_i}$,

Osa y: $\tilde{x}_{P_i}/\tilde{x}_{0.5}$,



Selekční graf pro výběr z lognormálního rozdělení

Požadavek symetrie kvantilů kolem mediánu

$$\left(\frac{\tilde{x}_{P_i}}{\tilde{x}_{0.5}}\right)^{\lambda} + \left(\frac{\tilde{x}_{0.5}}{\tilde{x}_{1-P_i}}\right)^{-\lambda} = 2$$

kde pro pořadové pravděpodobnosti jsou voleny písmenové hodnoty, $P_i = 2^{-i}$, $i = 2, 3$.

Porovnání: zakreslují se řešení rovnice

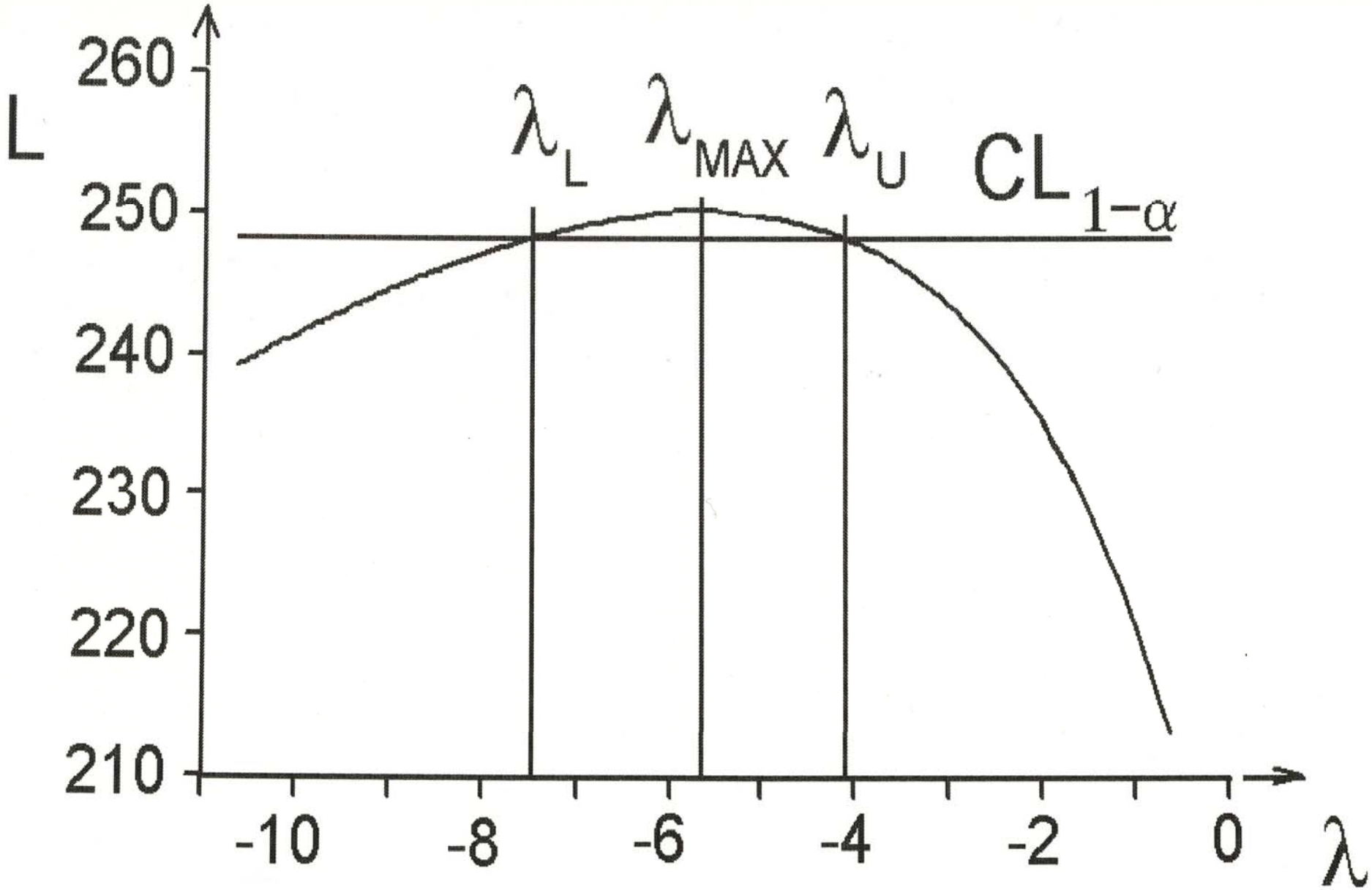
$$y^{\lambda} + x^{-\lambda} = 2 \text{ pro } 0 \leq x \leq 1 \text{ a } 0 \leq y \leq 1$$

(a) pro $\lambda = 0$ je řešením přímka $y = x$,

(b) pro $\lambda < 0$ je řešením vztah $y = (2 - x^{-\lambda})^{1/\lambda}$,

(c) pro $\lambda > 0$ je řešením vztah $x = (2 - y^{\lambda})^{-1/\lambda}$.

Graf maximální věrohodnosti



Metoda maximální věrohodnosti:

pro $\lambda = \hat{\lambda}$ je rozdělení y normální, $N(\mu_y, \sigma^2(y))$

$$\ln L(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i$$

kde $s^2(y)$ je výběrový rozptyl transformovaných dat y.

Pro asymptotický 100(1 - α)%ní interval spolehlivosti parametru λ platí

$$2\left[\ln L(\hat{\lambda}) - \ln L(\lambda)\right] \leq \chi^2_{1-\alpha}(1)$$

kde $\chi^2_{1-\alpha}$ (1) je kvantil χ^2-rozdělení s jedním stupněm volnosti. V intervalu spolehlivosti leží všechna λ, pro která platí

$$\ln L(\lambda) \geq \ln L(\hat{\lambda}) - 0.5\chi^2_{1-\alpha}(1)$$

Čím je konfidenční interval širší, tím je transformace méně výhodná.

Pokud obsahuje interval i hodnotu $\lambda = 1$, není transformace přínosem.

Příklad 2.26 *Atestace stopového Cu v kaolinu*

Ve vzorku kaolinu byl stanoven obsah Cu (ppm) a hodnoty ve vzestupném pořadí jsou 4, 5, 7, 7, 7, 8, 8.3, 8.4, 9.4, 9.5, 10.0, 10.5, 12.0, 12.8, 13.0, 22.0, 23.0.

Stanovte míry polohy a rozptýlení uvedeného výběru.

Řešení:

1. Výpočet na kalkulačce

Aritmetický průměr	$\bar{x}$	= 10.406
Medián	$\tilde{x}_{0.5}$	= 9.400

2. *Interaktivní statistická analýza na PC*

1. Průzkumová analýza dat (EDA):

Diagnostické grafy: stupeň symetrie rozdělení
lokální koncentrace
vybočující data

2. Ověření předpokladů výběru dat:

Diagnostiky: ověření normality,
oveření nezávislosti,
ověření homogenity,
oveření minimální četnosti.

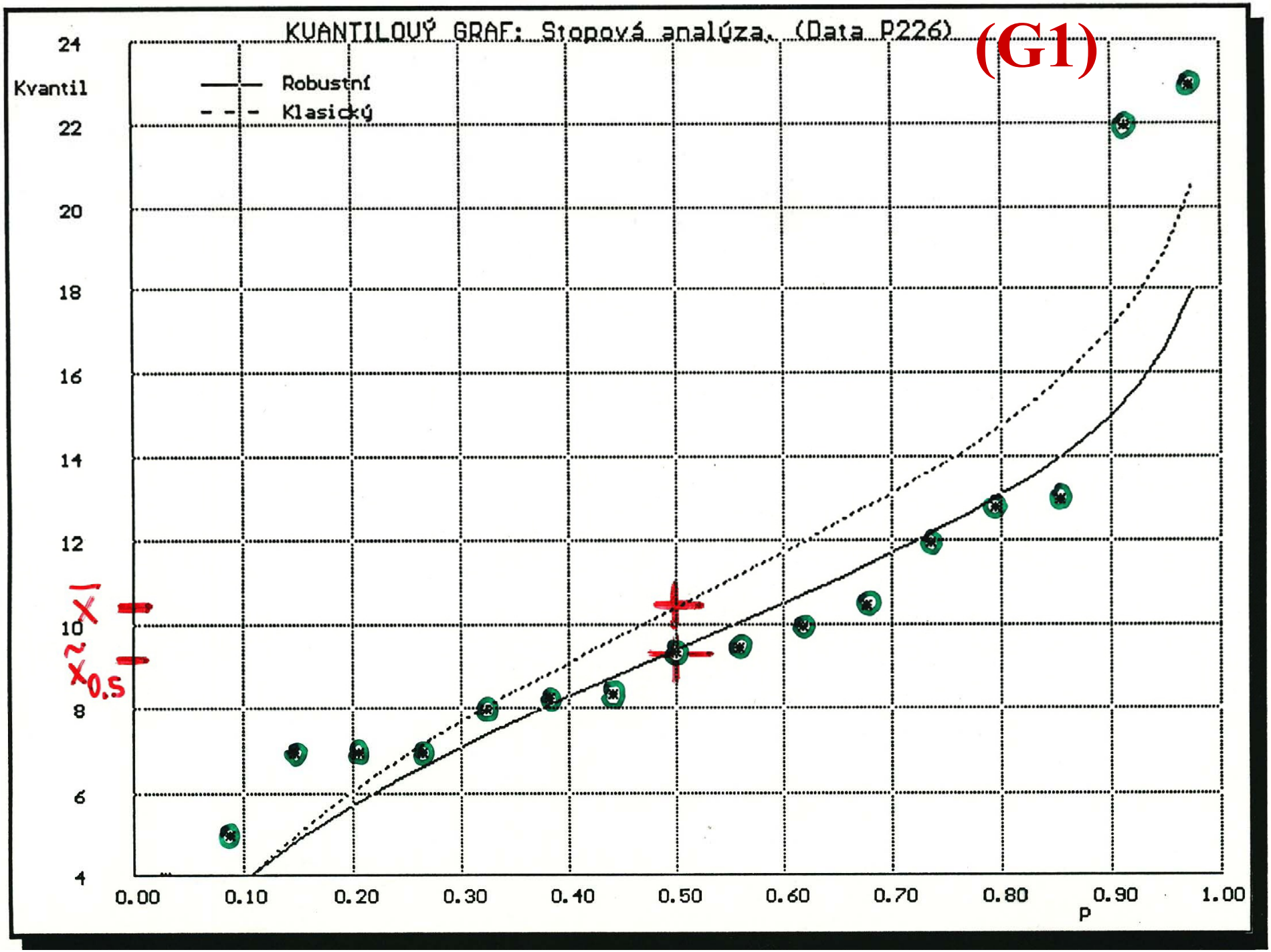
3. Transformace dat: pro symetrii a pro heteroskedasticitu výběru

Box-Coxova transformace,
Mocninná transformace.

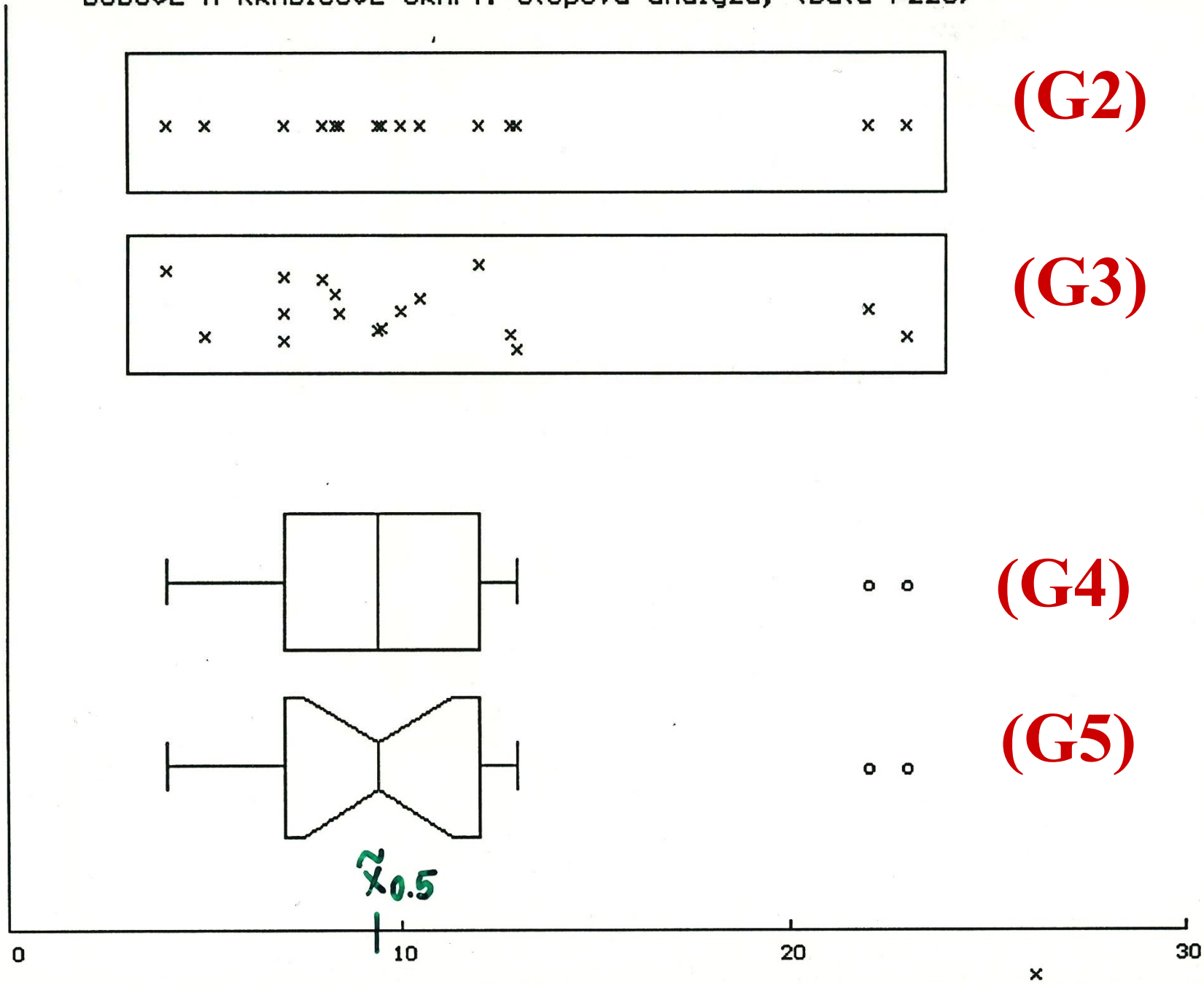
4. Odhady parametrů polohy, rozptýlení a tvaru:

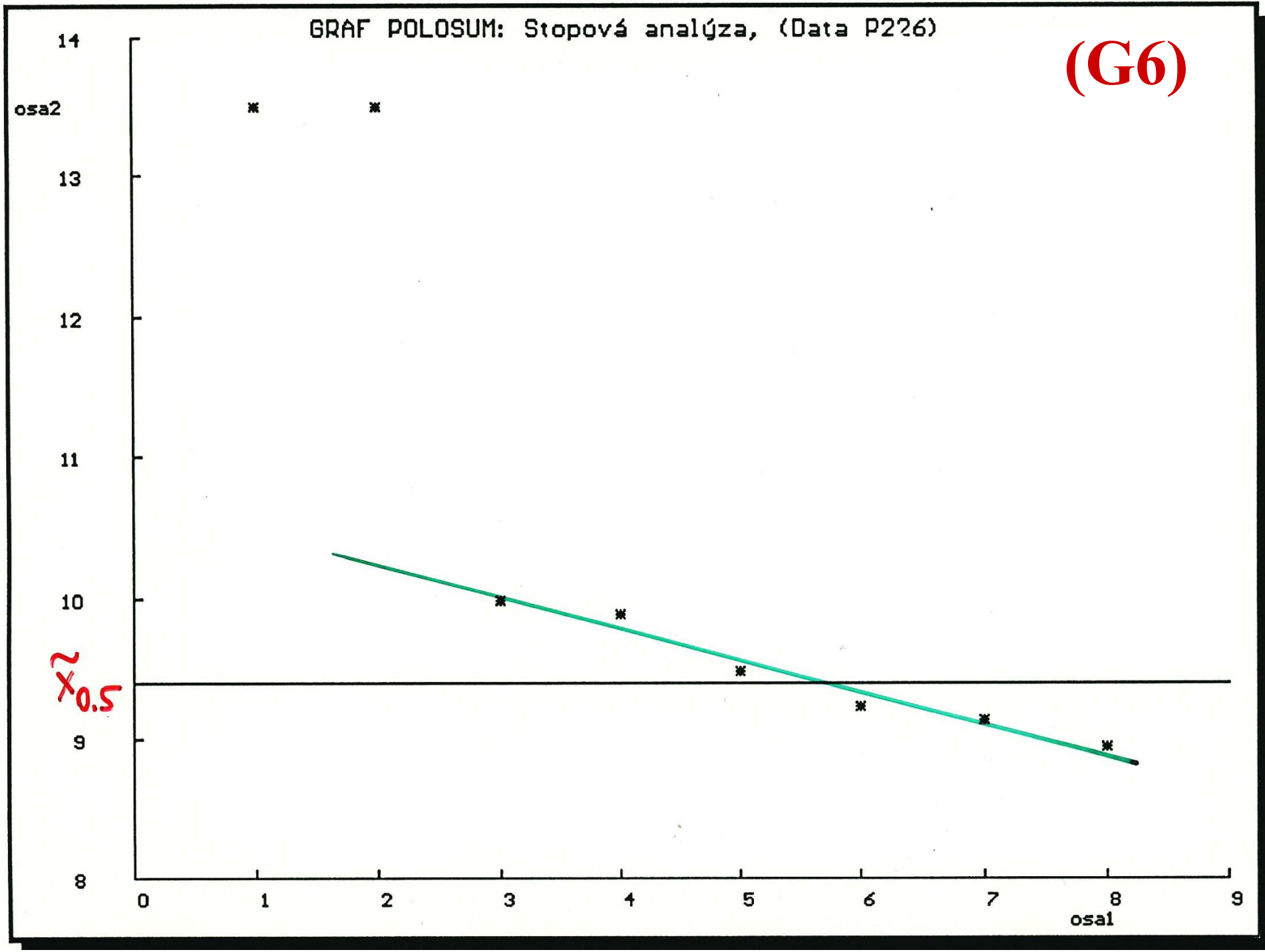
Klasické odhady: aritmetický průměr, rozptyl
Robustní odhady: medián, uřezané průměry,
neparametrické odhady rozptylu.

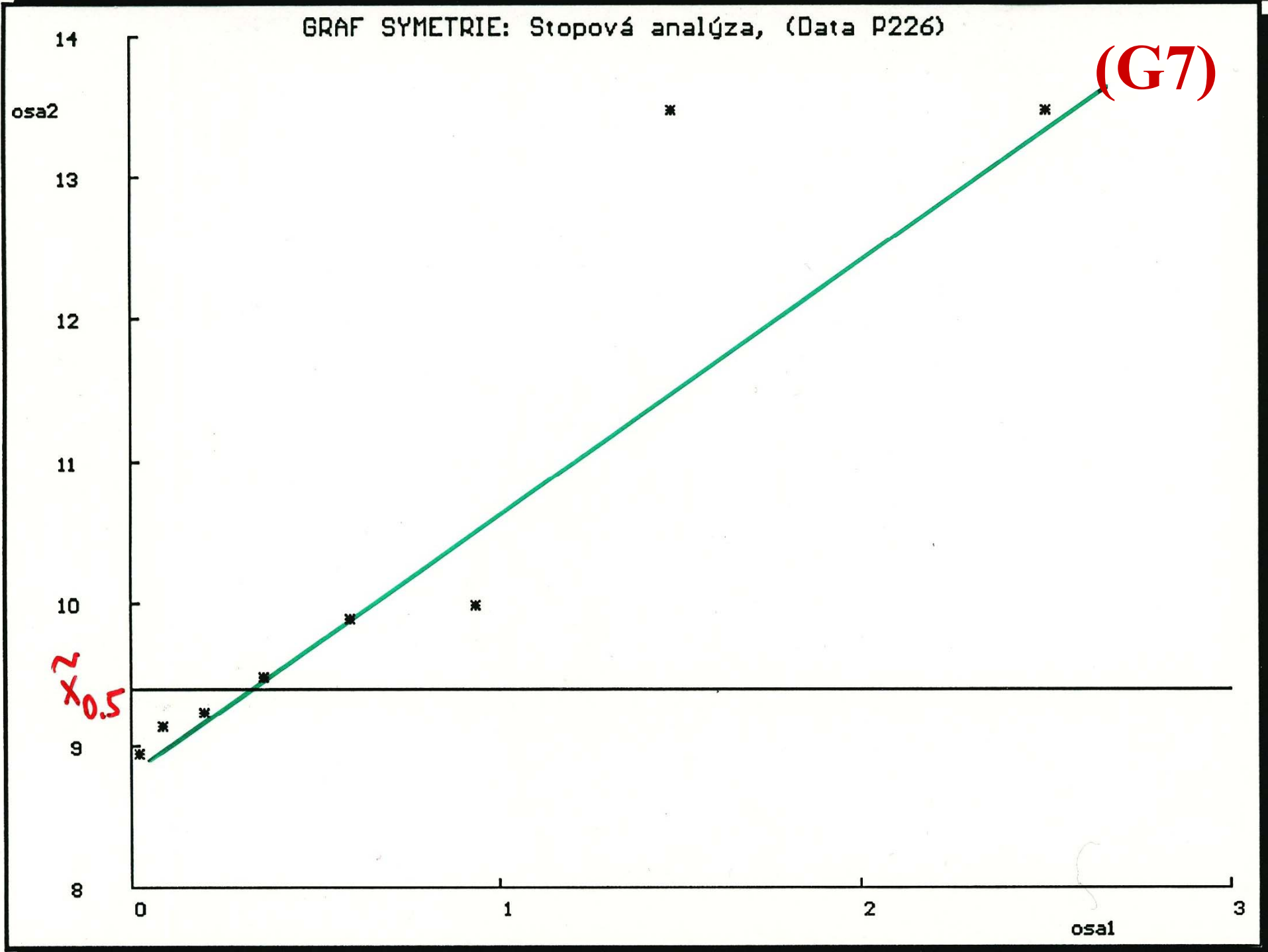
Adaptivní odhady polohy a rozptýlení.

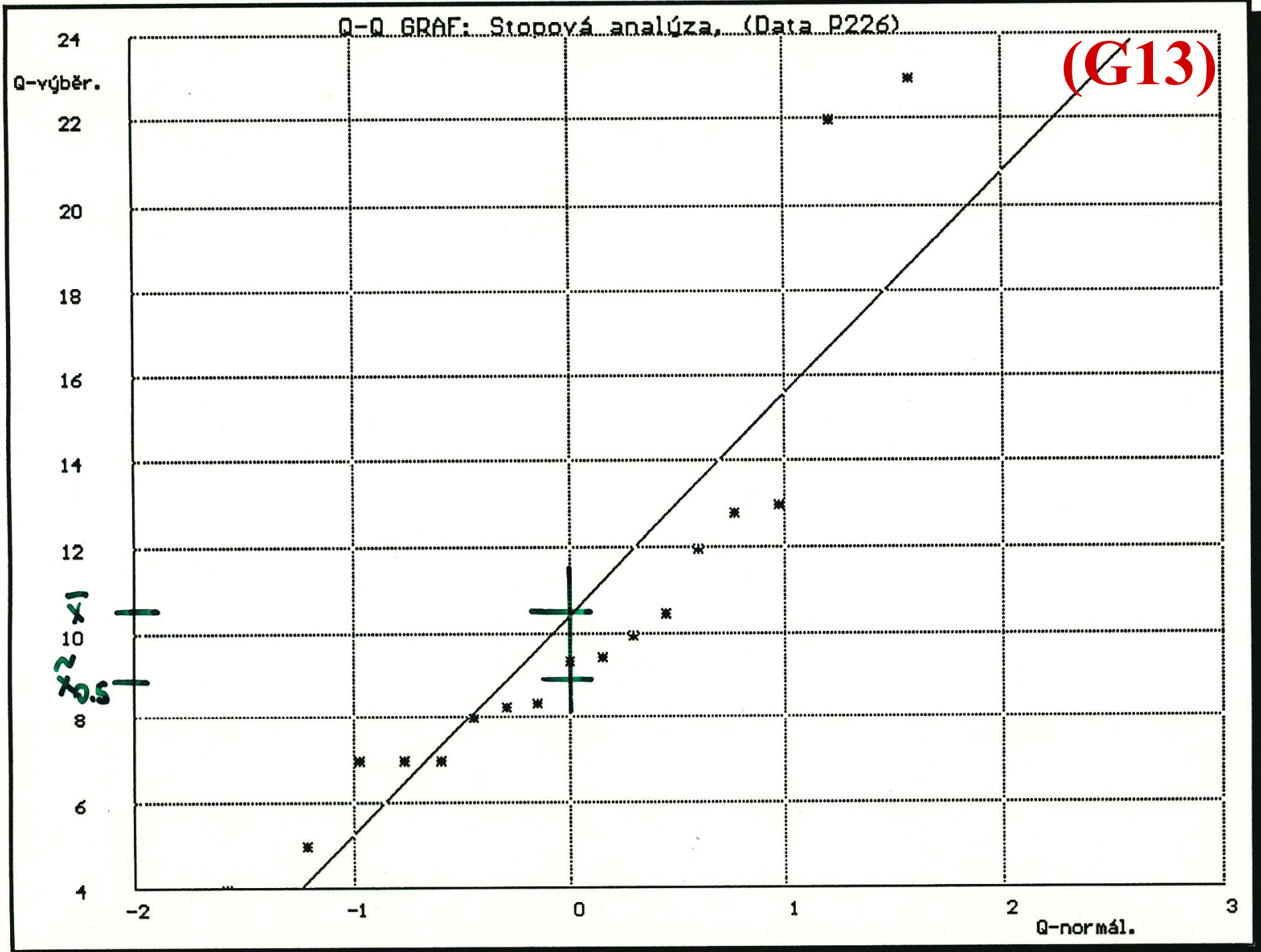

KVANTILOVÝ GRAF: Stopová analýza, (Data P226)
(G1)
Kvantil
Robustní
Klasický
24
22
20
18
16
14
12
10
8
6
4
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
P
$\tilde{x}_{0.5}$
$\bar{x}$

BODOVÉ A KRABICOVÉ GRAFY: Stopová analýza, (Data P226)

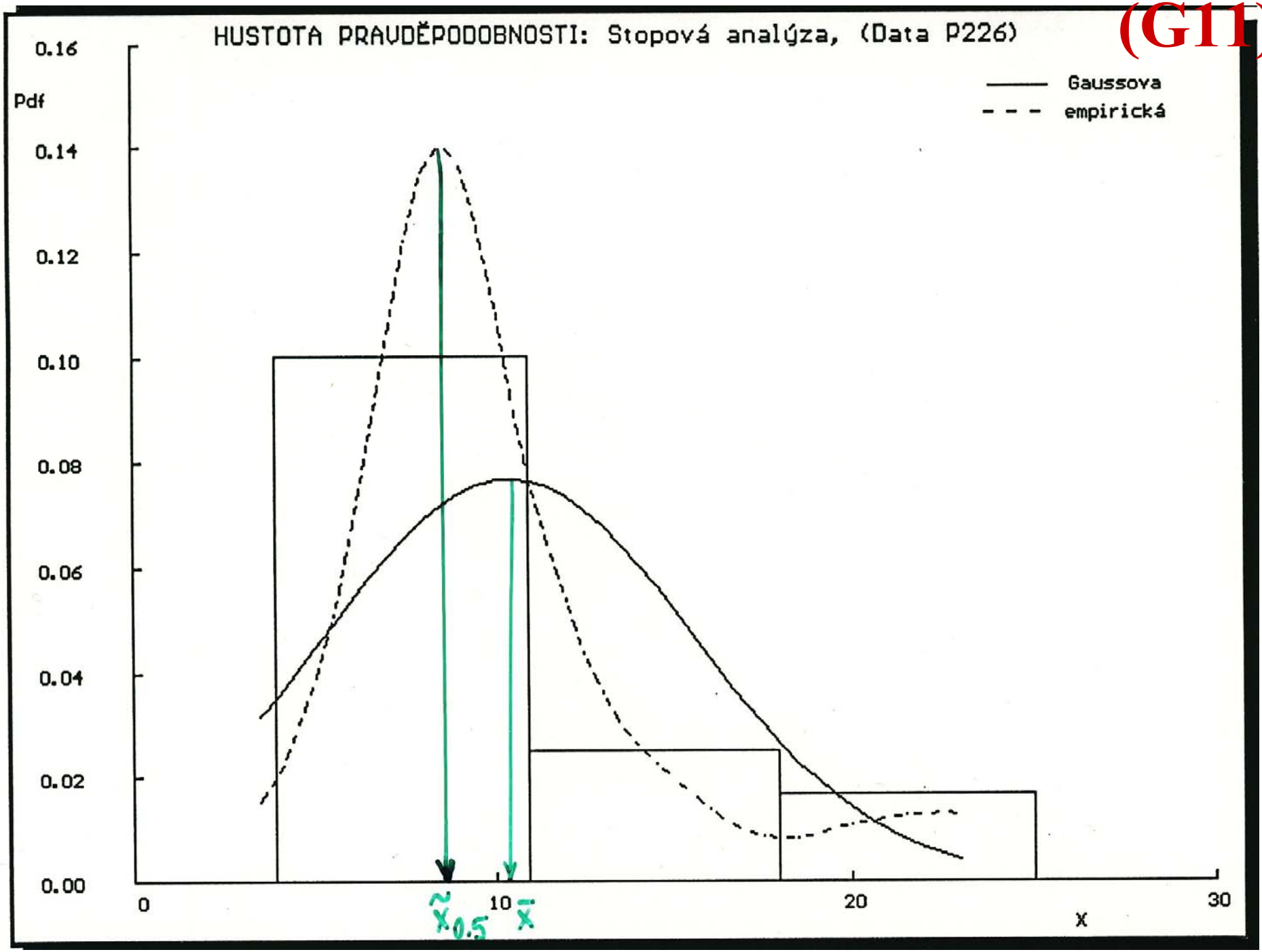


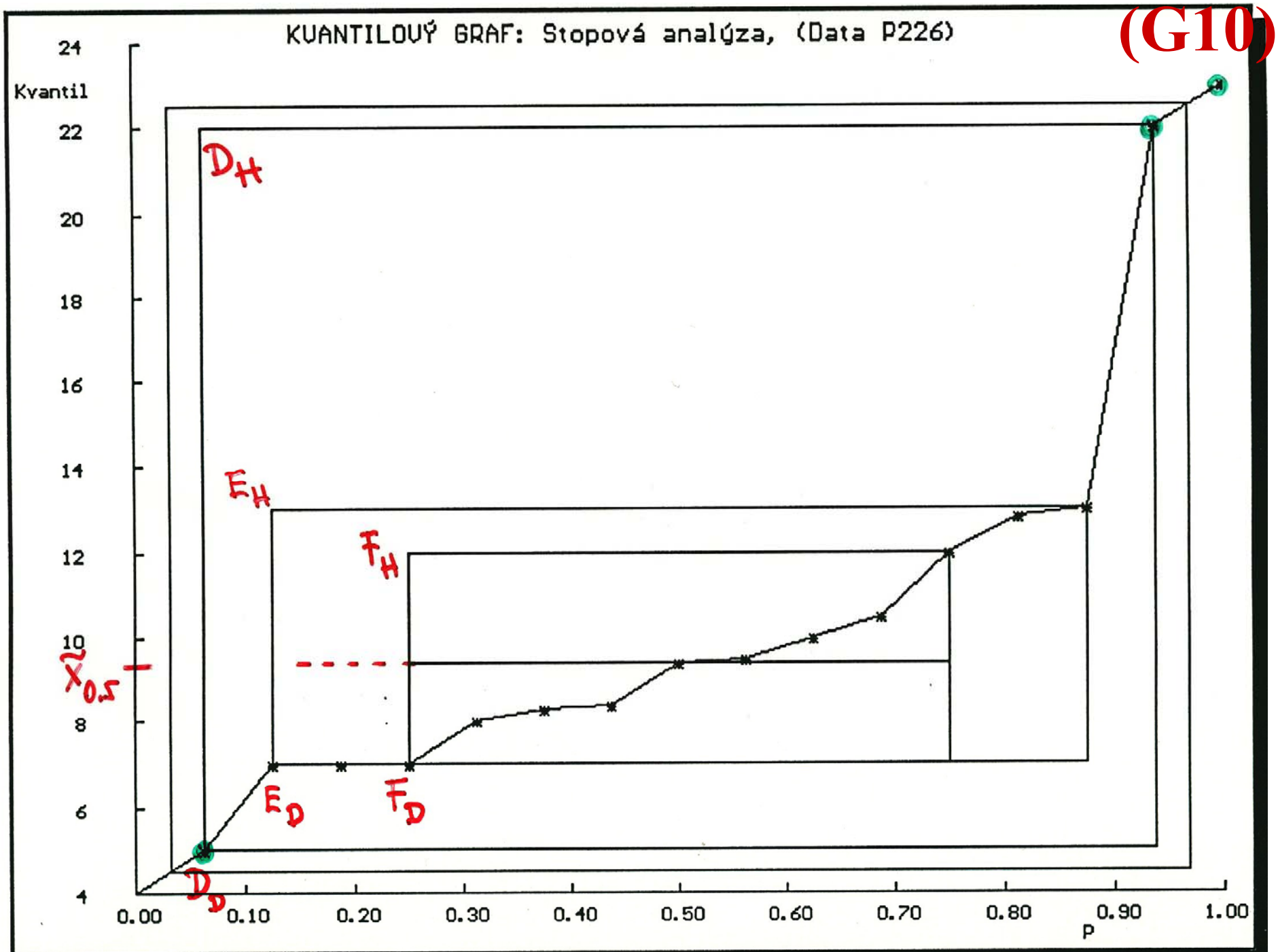

GRAF POLOSUM: Stopová analýza, (Data P226)
(G6)
osa2
14
13
12
11
10
9
8
$\tilde{x}_{0.5}$
0
1
2
3
4
5
6
7
8
9
osa1

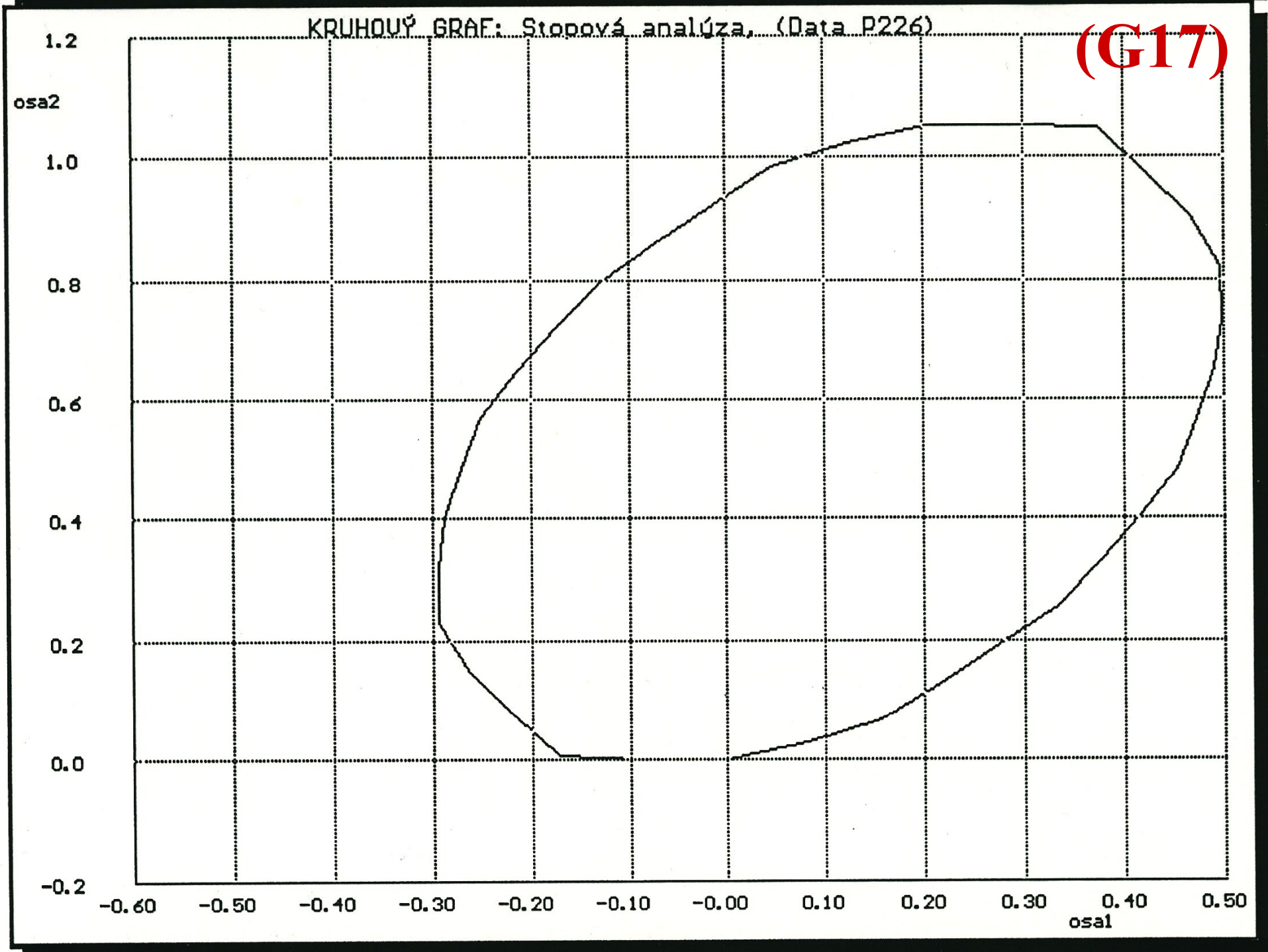

GRAF SYMETRIE: Stopová analýza, (Data P226)
(G7)
osa2
14
13
12
11
10
9
8
$\chi^2_{0.5}$
0
1
2
3
osa1

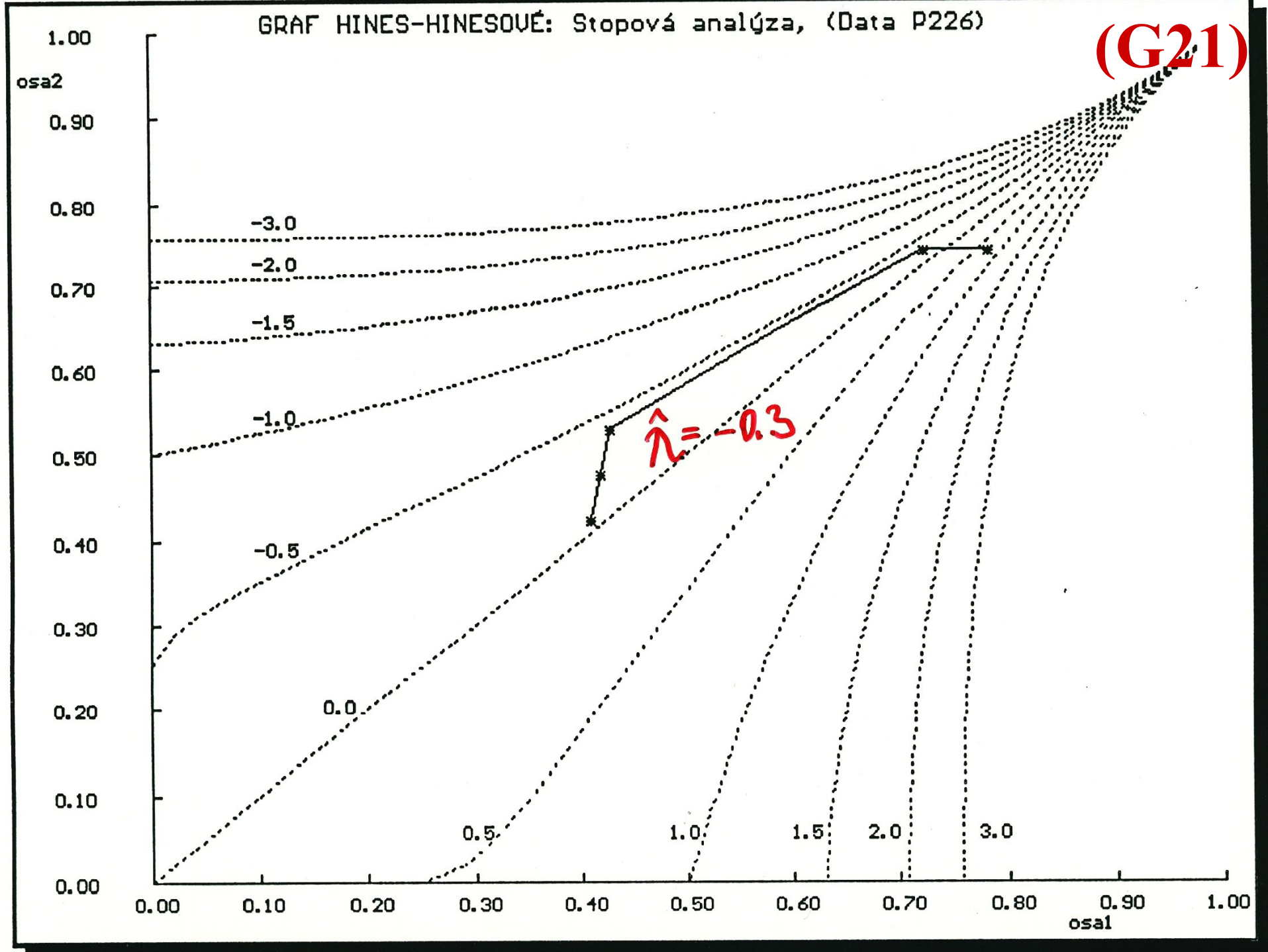

Q-Q GRAF: Stopová analýza, (Data P226)
(G13)
Q-výběr.
Q-normál.
24
22
20
18
16
14
12
10
8
6
4
-2
-1
0
1
2
3
$\tilde{x}_{0.5}$
$\bar{x}$

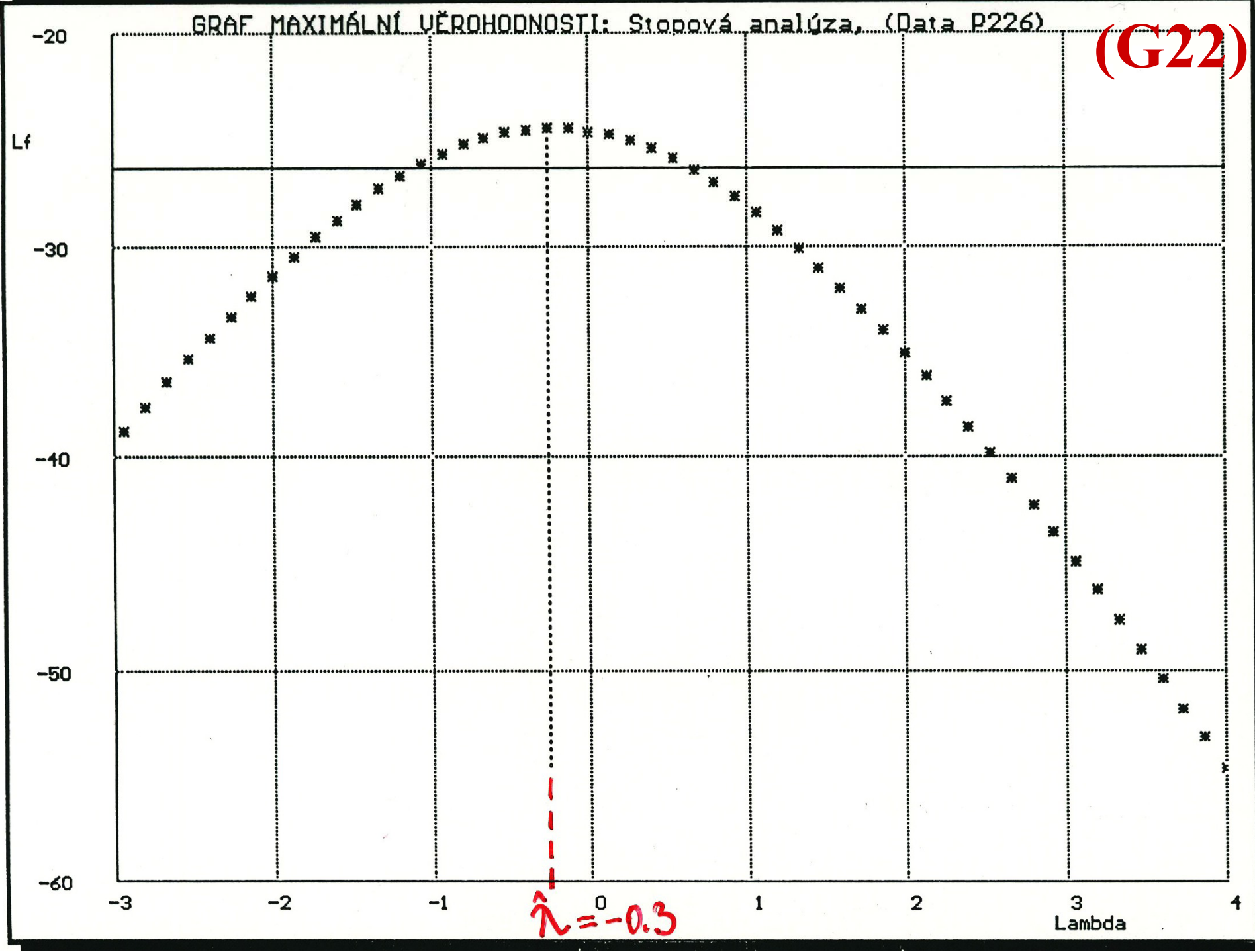
(G11)

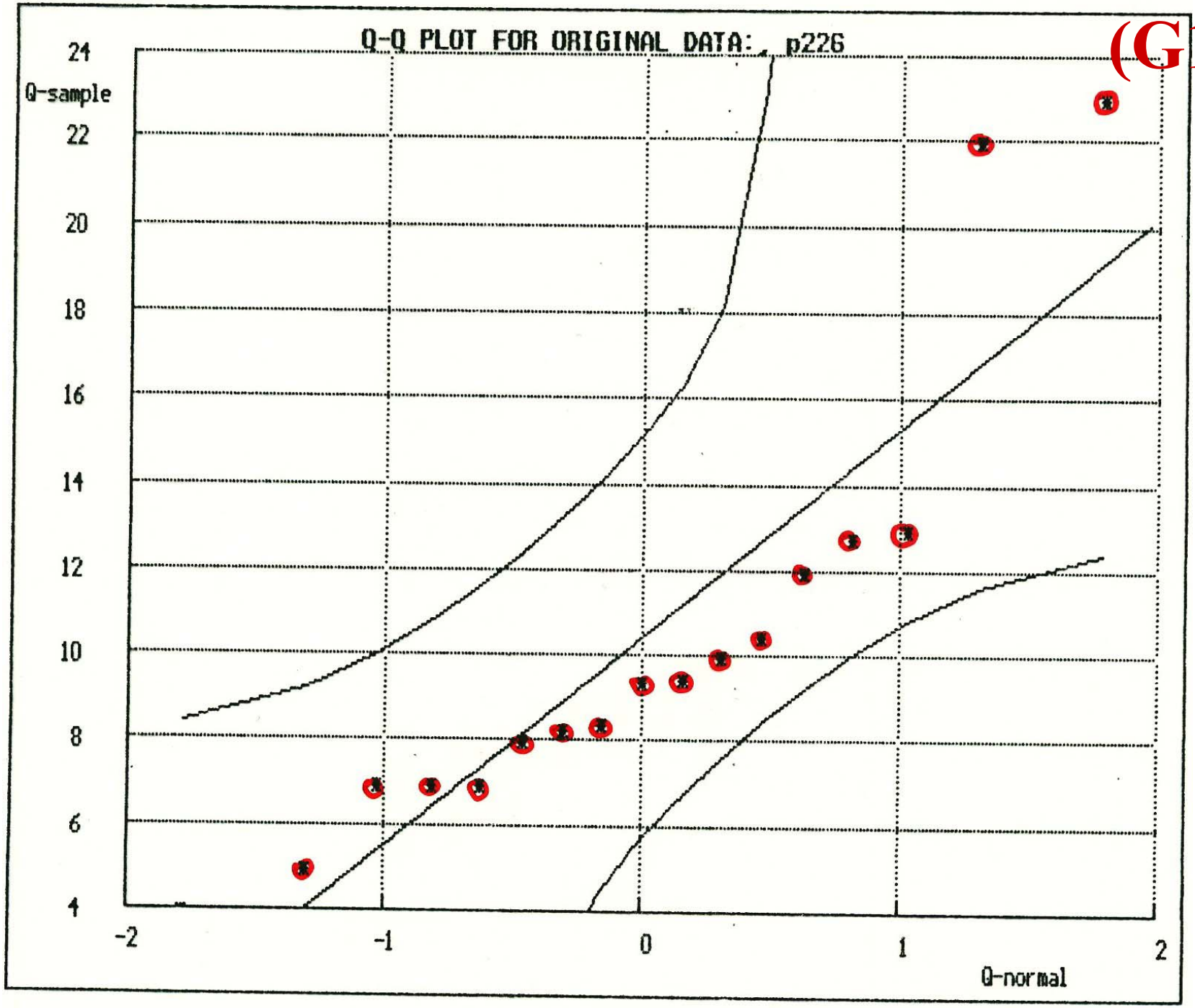

HUSTOTA PRAVDĚPODOBNOSTI: Stopová analýza, (Data P226)
Gaussova
empirická
Pdf
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
x
$\tilde{x}_{0.5}$
$\bar{x}$

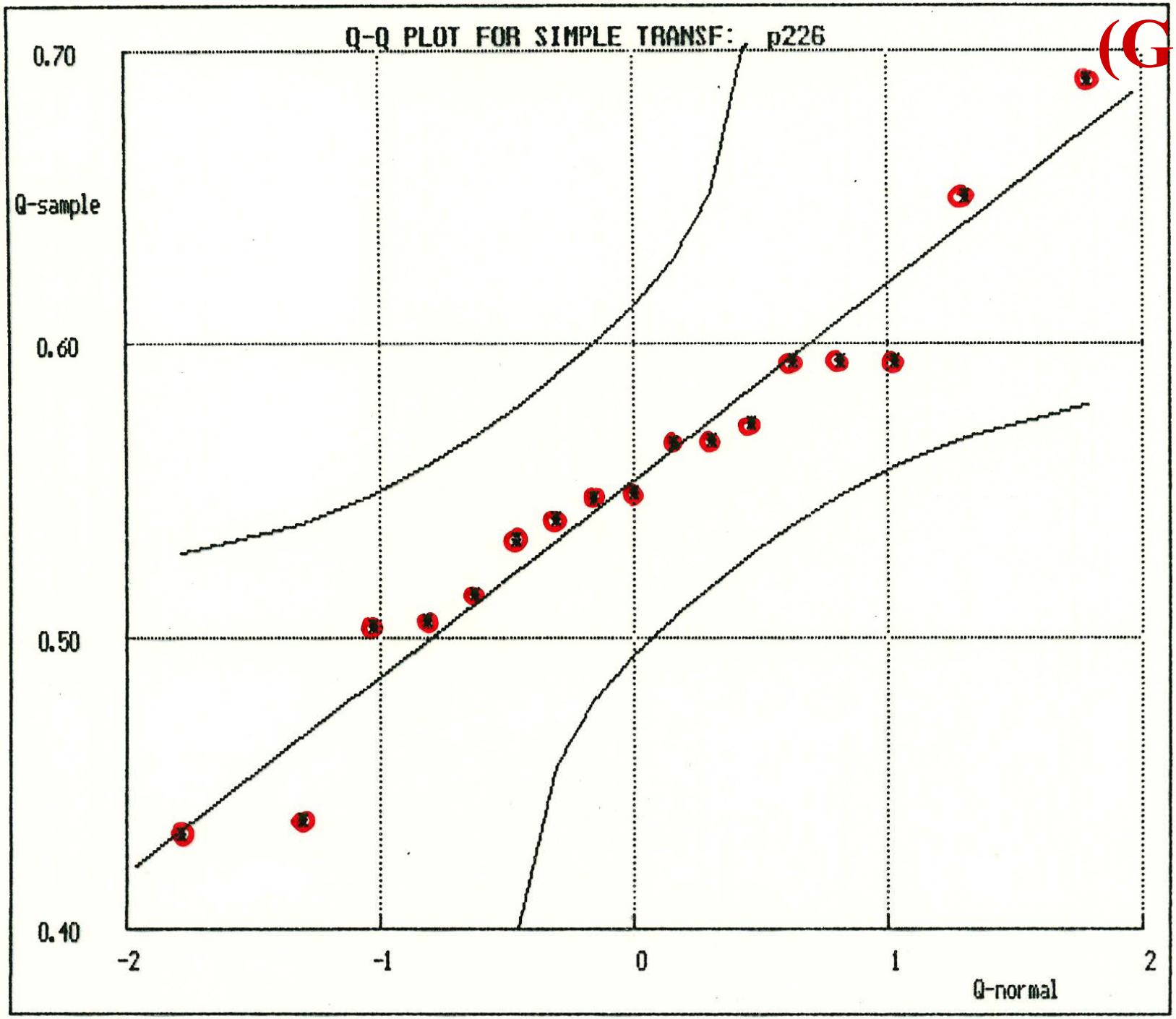

(G10)
KUANTILOVÝ GRAF: Stopová analýza, (Data P226)
Kvantil
24
22
20
18
16
14
12
10
8
6
4
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
p
D_H
E_H
F_H
$\tilde{x}_{0.5}$
E_D
F_D
D_D

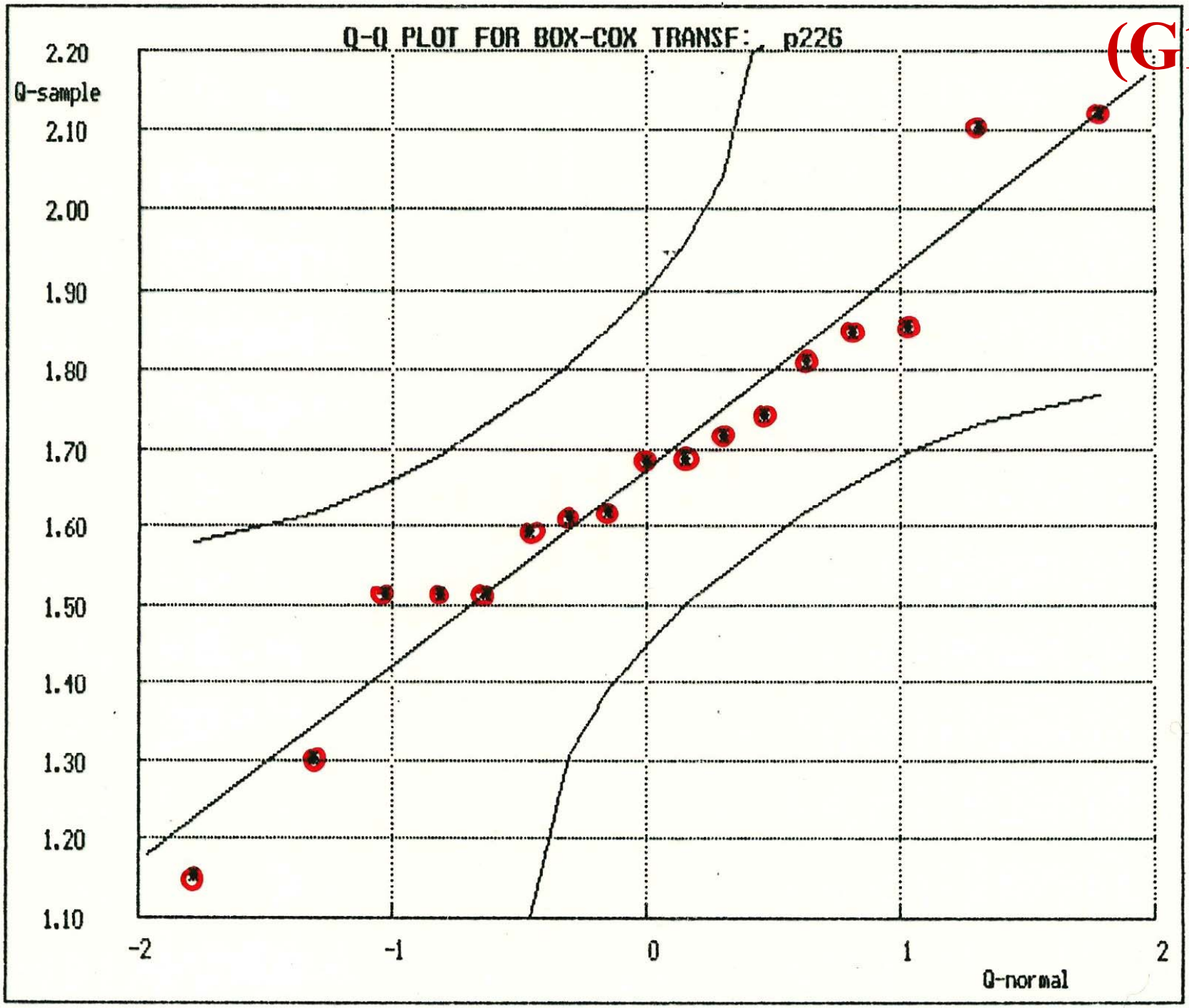

KRUHOVÝ GRAF: Stopová analýza, (Data P226)
(G17)
osa2
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.60
-0.50
-0.40
-0.30
-0.20
-0.10
-0.00
0.10
0.20
0.30
0.40
0.50
osa1

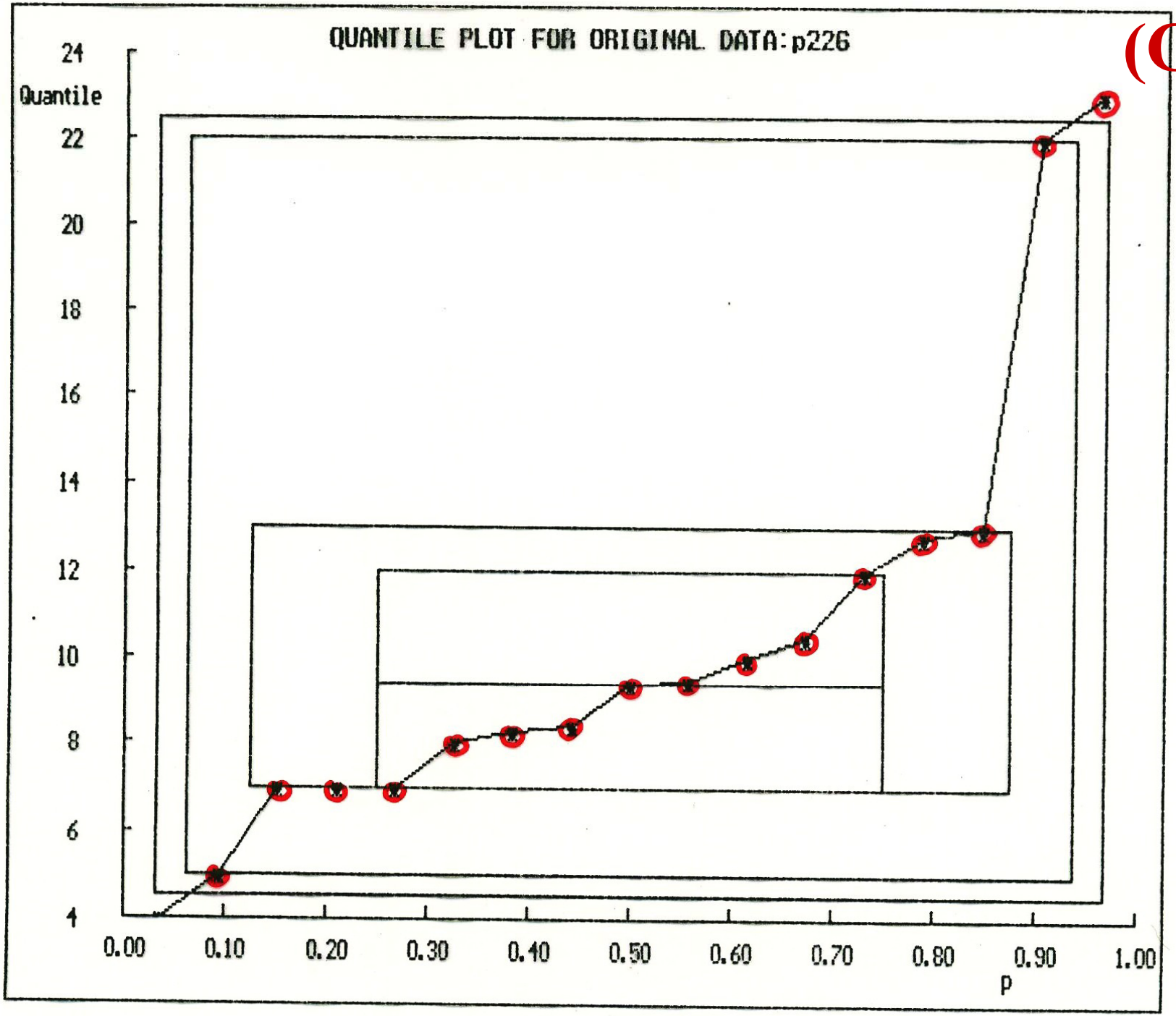

GRAF HINES-HINESOVÉ: Stopová analýza, (Data P226)
(G21)
osa2
osa1
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
3.0
$\hat{\lambda} = -0.3$


GRAF MAXIMÁLNÍ VĚROHODNOSTI: Stopová analýza, (Data P226)
(G22)
Lf
-20
-30
-40
-50
-60
-3
-2
-1
0
1
2
3
4
Lambda
$\hat{\lambda} = -0.3$


Q-Q PLOT FOR ORIGINAL DATA: , p226
(G13)
Q-sample
24
22
20
18
16
14
12
10
8
6
4
-2
-1
0
1
2
Q-normal

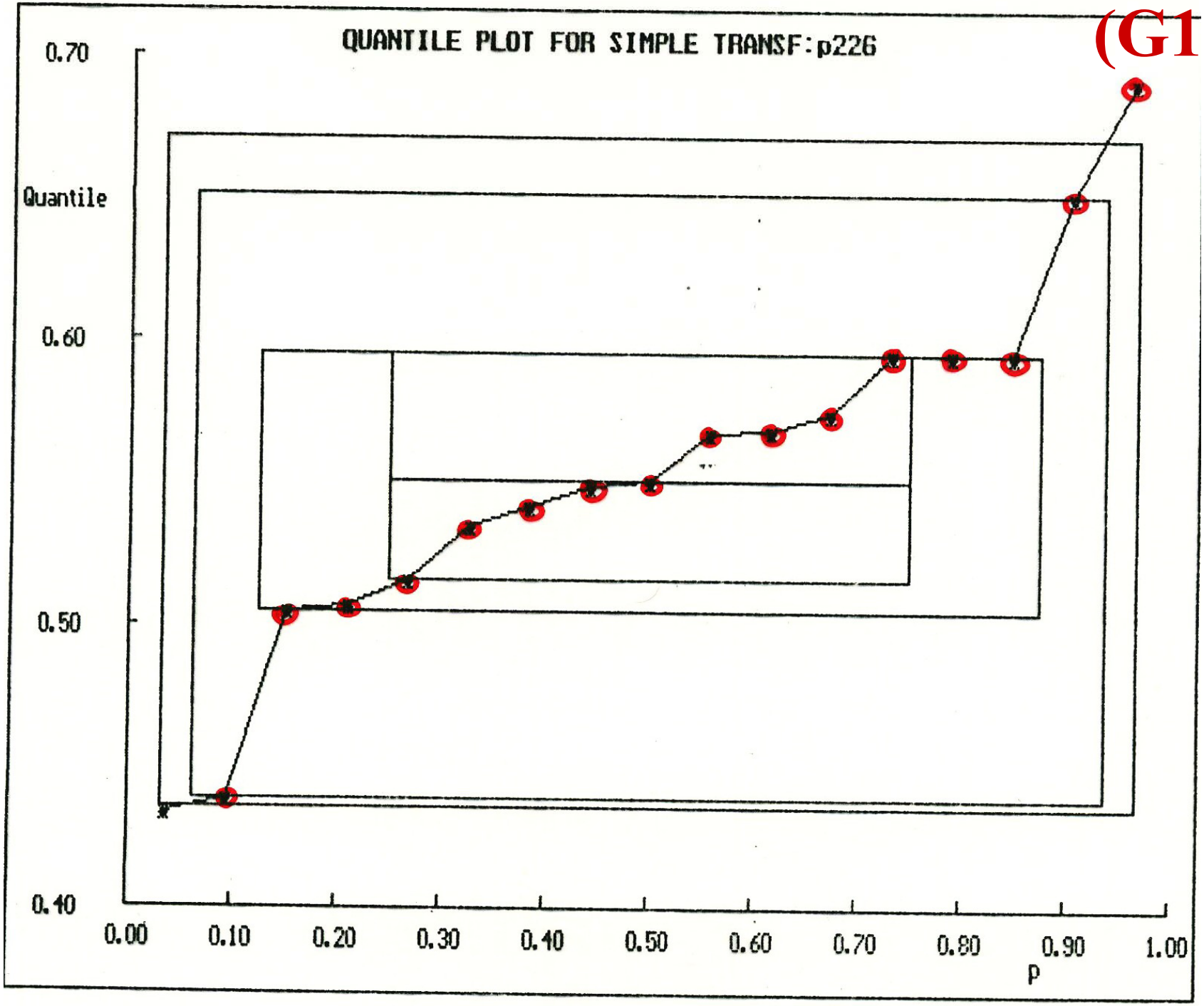

Q-Q PLOT FOR SIMPLE TRANSF: p226
(G13)
0.70
0.60
0.50
0.40
Q-sample
-2
-1
0
1
2
Q-normal

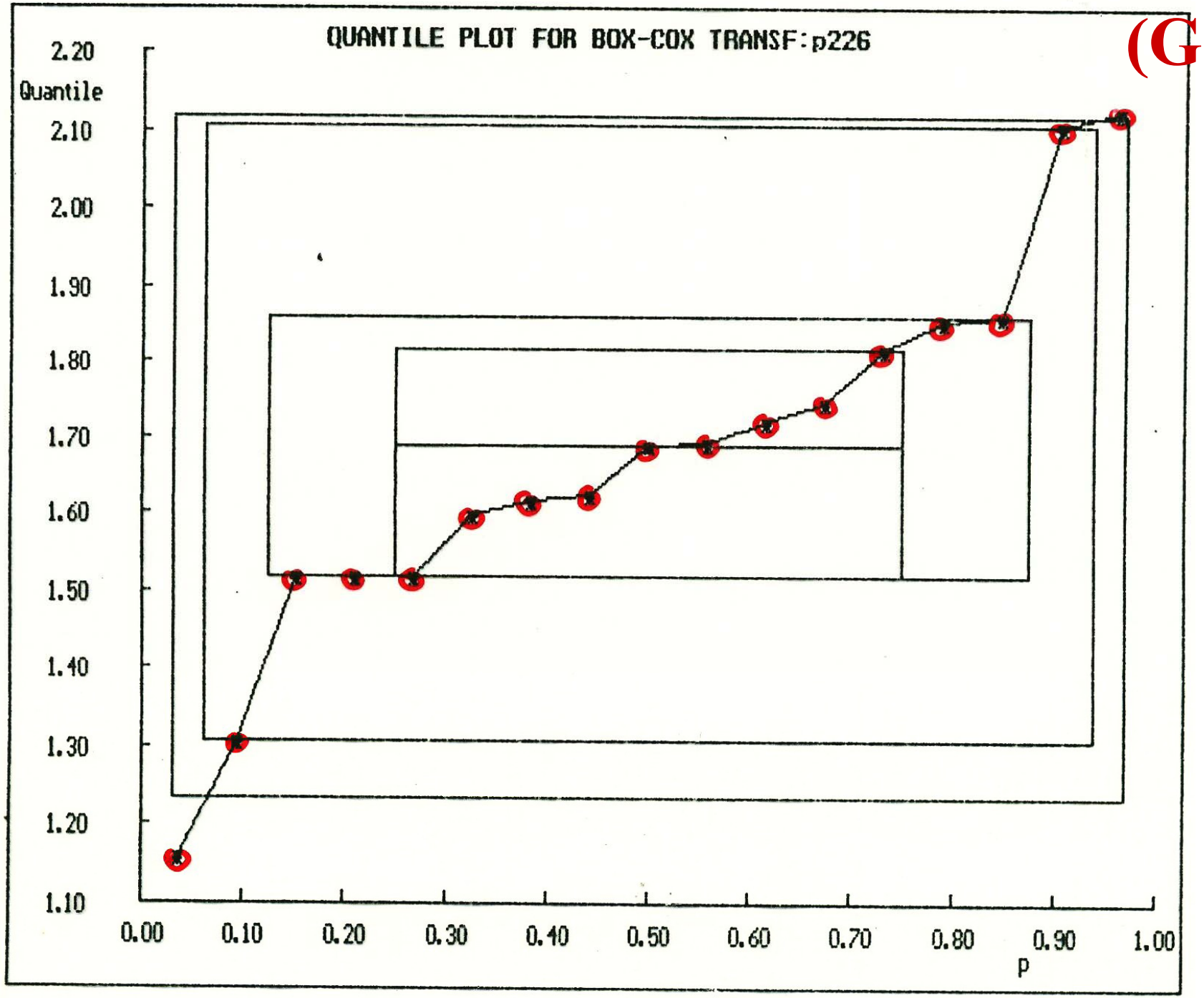

Q-Q PLOT FOR BOX-COX TRANSF: p226
(G13)
Q-sample
2.20
2.10
2.00
1.90
1.80
1.70
1.60
1.50
1.40
1.30
1.20
1.10
-2
-1
0
1
2
Q-normal


QUANTILE PLOT FOR ORIGINAL DATA:p226
Quantile
24
22
20
18
16
14
12
10
8
6
4
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
p

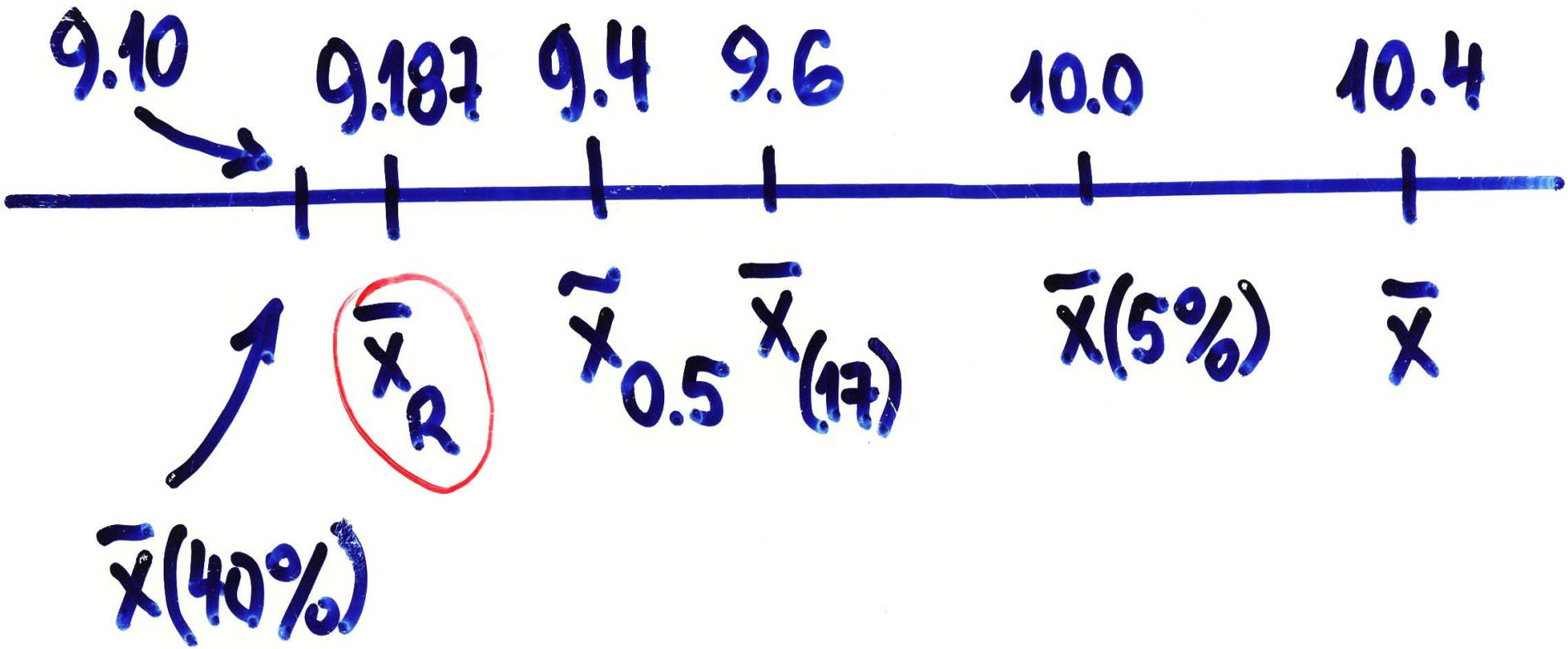
(G10)

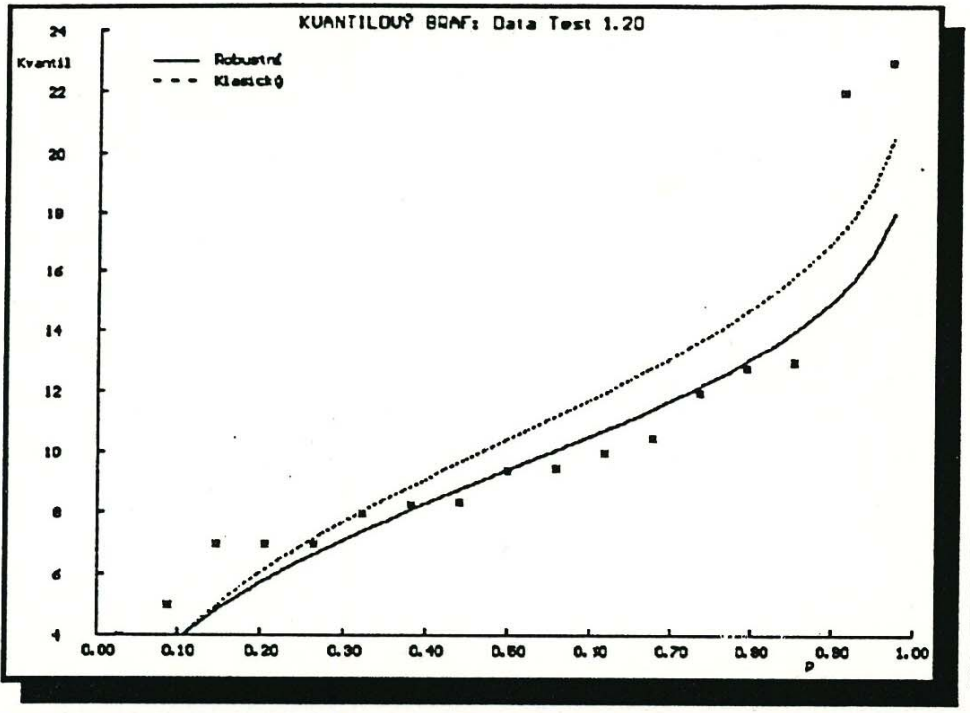
(G10)

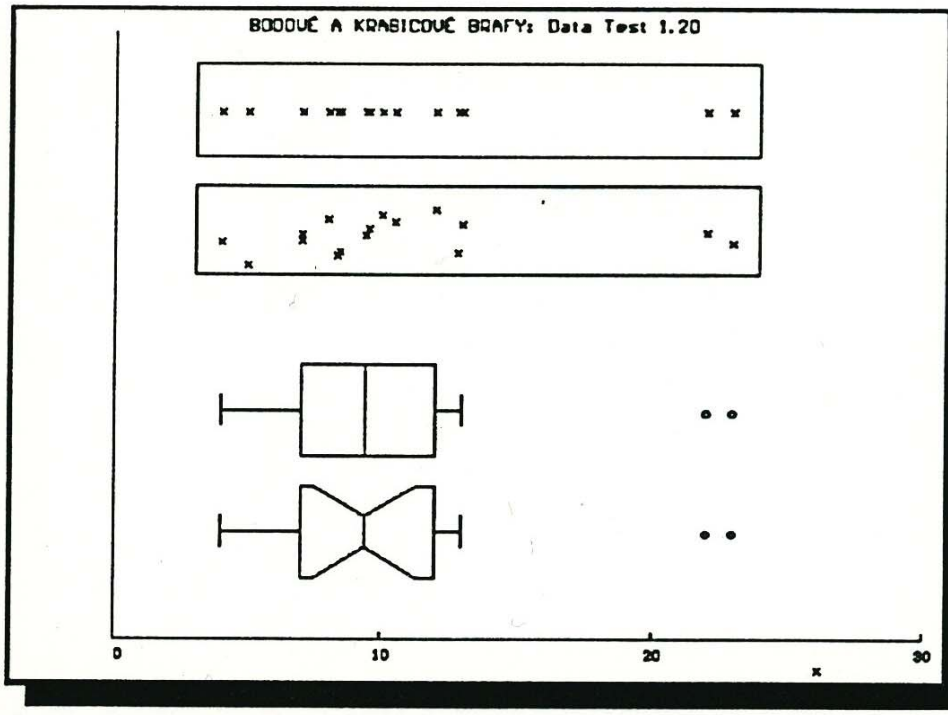

QUANTILE PLOT FOR SIMPLE TRANSF:p226
Quantile
0.70
0.60
0.50
0.40
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
p

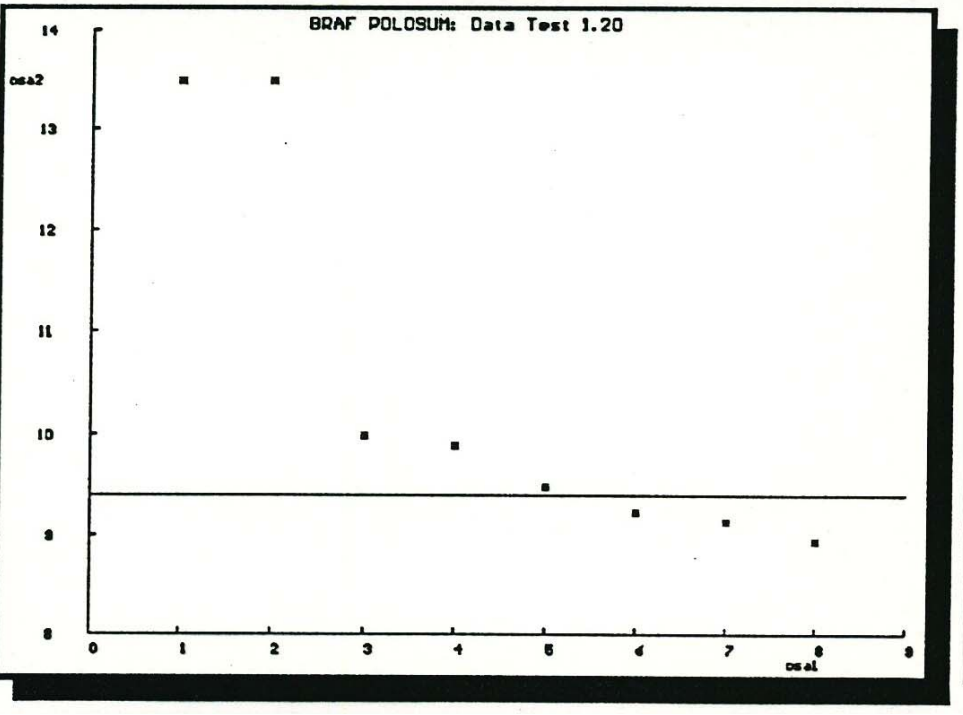

QUANTILE PLOT FOR BOX-COX TRANSF:p226
(G10)
Quantile
2.20
2.10
2.00
1.90
1.80
1.70
1.60
1.50
1.40
1.30
1.20
1.10
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
p

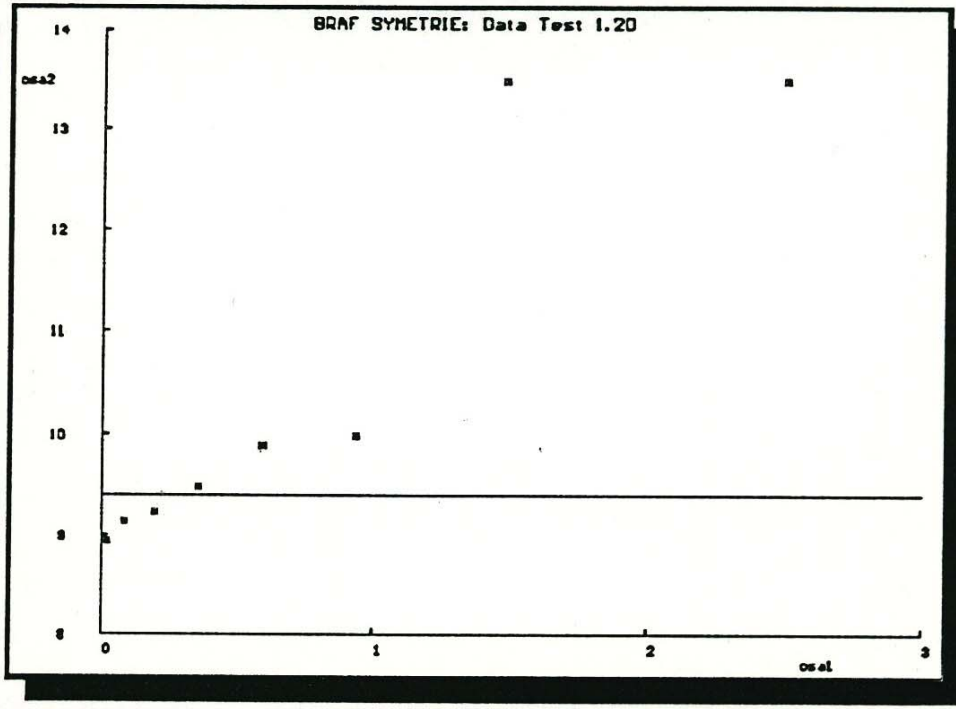
STŘEDNÍ HODNOTA

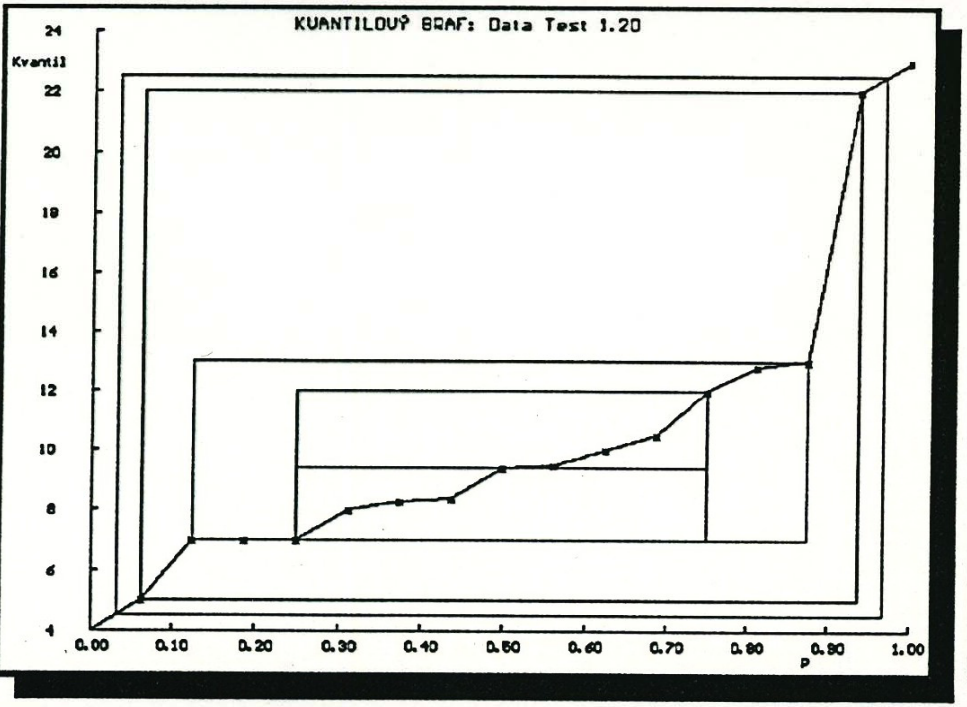


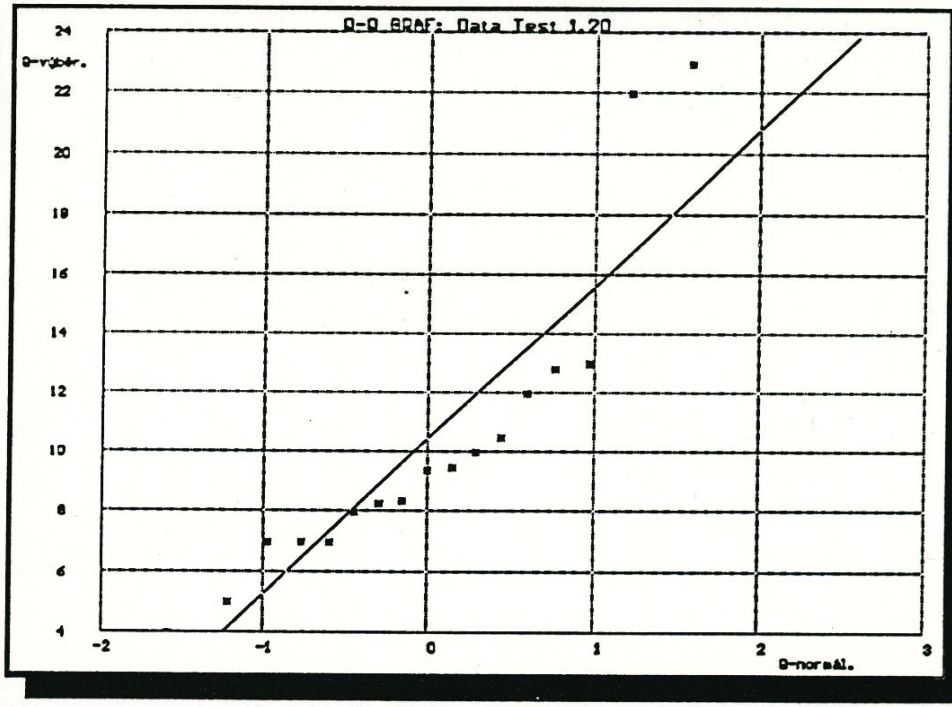

KVANTILOVÝ GRAF: Data Test 1.20
Kvantil
Robustní
Klasický
p

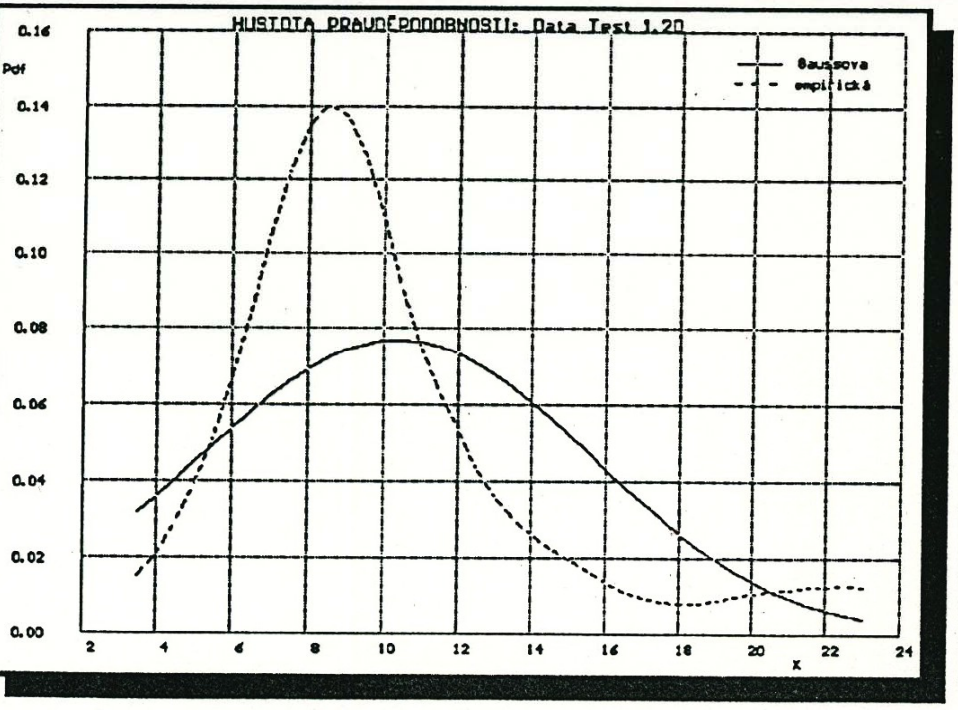

BODOVÉ A KRABICOVÉ GRAFY: Data Test 1.20
x

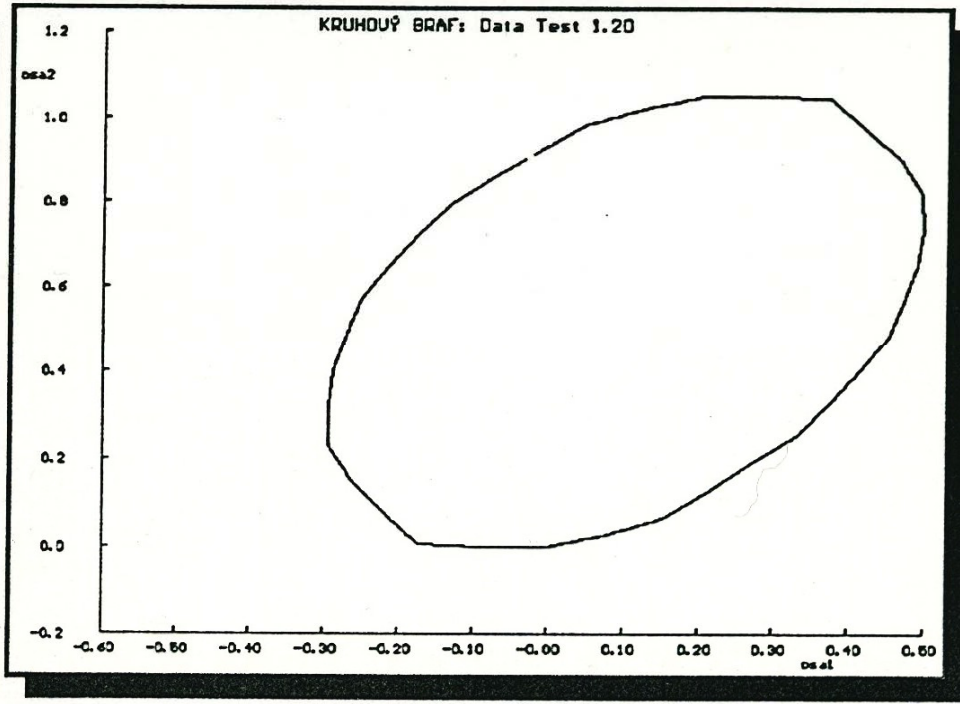

BRAF POLOSUM: Data Test 1.20
osa2
osa1


BRAF SYMETRIE: Data Test 1.20
osa2
osa1


KVANTILOVÝ GRAF: Data Test 1.20
Kvantil
24
22
20
18
16
14
12
10
8
6
4
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
p


Q-Q GRAF: Data Test 1.20
Q-výběr.
24
22
20
18
16
14
12
10
8
6
4
-2
-1
0
1
2
3
Q-normál.


HUSTOTA PRAVDĚPODOBNOSTI: Data Test 1.20
Pdf
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Gaussova
empirická
2
4
6
8
10
12
14
16
18
20
22
24
x


KRUHOVÝ GRAF: Data Test 1.20
osa2
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.60
-0.50
-0.40
-0.30
-0.20
-0.10
-0.00
0.10
0.20
0.30
0.40
0.50
osa1

Basic Assumptions

Title: p226

I N P U T

(1) CONDITIONS:

Sample size, n : 17
Significance level, alpha: 0.050
Output file name : P226A.TXT

(2) INPUT OF ORIGINAL DATA:

(1)	4.0000E+00	(6)	8.0000E+00	(11)	1.0000E+01	(16)	2.2000E+01
(2)	7.0000E+00	(7)	8.3000E+00	(12)	1.0500E+01	(17)	2.3000E+01
(3)	7.0000E+00	(8)	8.4000E+00	(13)	1.2000E+01		
(4)	7.0000E+00	(9)	9.4000E+00	(14)	1.2800E+01		
(5)	5.0000E+00	(10)	9.5000E+00	(15)	1.3000E+01		

O U T P U T

(1) CLASSICAL ESTIMATORS OF LOCATION AND SHAPE:

Mean : 1.0406E+01 Variance : 2.6834E+01
Standard deviation : 5.1802E+00 Skewness : 1.3991E+00
Kurtosis : 4.2721E+00

(2) TEST FOR NORMALITY:

Quantile, Chi^2(1-alpha,2) : 5.9915E+00
Chi^2 statistics : 7.9084E+00
Conclusion: Normality rejected
Computed significance level : 1.9174E-02

(3) TEST FOR INDEPENDENCE:

Quantile, t(1-alpha/2,n+1) : 2.1009E+00
t statistics : 1.0381E+00
Conclusion: Independence accepted
Computed significance level : 1.5649E-01

Assumption of sample homogenity:

Arithmetic mean : 1.0406E+01
Variance : 2.6834E+01
Standard deviation : 5.1802E+00

Internal bounds:
Lower bound: -3.1909E+00
Upper bound: 2.2191E+01

(4) MINIMAL SAMPLE SIZES:

for 25% relative error of standard deviation: n = 14
for 10% relative error of standard deviation: n = 83
for 5% relative error of standard deviation: n = 328

(5) OUTLIERS DETECTION:

point number 17 (high) : 2.3000E+01
Number of outliers : 1

Measures with excluded outliers

Mean	9.6188E+00	Variance	1.7388E+01
Standard deviation	4.1699E+00	Skewness	1.6101E+00
Kurtosis	6.3399E+00		

ADSTAT (c) Trilobyte

UNIVARIATE DATA ANALYSIS

One Sample Analysis

Title: p226

I N P U T

(1) CONDITIONS:

Sample size, n : 17
Significance level, alpha: 0.050
Output file name : P226.TXT

(2) INPUT OF ORIGINAL DATA:

(1)	4.0000E+00	(6)	8.0000E+00	(11)	1.0000E+01	(16)	2.2000E+01
(2)	7.0000E+00	(7)	8.3000E+00	(12)	1.0500E+01	(17)	2.3000E+01
(3)	7.0000E+00	(8)	8.4000E+00	(13)	1.2000E+01		
(4)	7.0000E+00	(9)	9.4000E+00	(14)	1.2800E+01		
(5)	5.0000E+00	(10)	9.5000E+00	(15)	1.3000E+01		

O U T P U T

(1) ESTIMATORS OF DISTRIBUTION SHAPE:

Skewness : 1.3991E+00
Kurtosis : 4.2721E+00

(2) CLASSICAL ESTIMATORS OF LOCATION AND SPREAD (Normality assumed):

Mean : 1.0406E+01
Standard deviation : 5.1802E+00
Variance : 2.6834E+01
95.0 % confidence interval:
Lower limit: 7.7425E+00 Upper limit: 1.3069E+01

(3) OTHER ESTIMATORS OF LOCATION:

Mode estimate : 7.5000E+00
Mid-Sum estimate : 1.3500E+01

(4) ROBUST STATISTICS OF LOCATION AND SPREAD (Normality not assumed):

Median measures:
Median : 9.4000E+00
Standard deviation : 4.4596E+00
Variance : 1.9888E+01
Standard deviation : 2.0494E+00
Variance (nonpar.) : 1.4316E+00
Standard deviation : 1.0003E+00
Variance (Marritz) : 1.0001E+00
95.0 % confidence interval:
Lower limit: 7.2798E+00 Upper limit: 1.1520E+01

Trimming 5% (for P=0.05):

Mean	:	1.0062E+01
Standard deviation	:	5.5644E+00
Variance	:	3.0963E+01
Winsorized mean	:	1.0406E+01
Winsorized Standard deviation	:	5.1209E+00
Winsorized variance	:	2.6223E+01

95.0 % confidence interval:

Lower limit: 7.1450E+00 Upper limit: 1.2979E+01

Trimming 10% (for P=0.10):

Mean	:	9.6324E+00
Standard deviation	:	3.8390E+00
Variance	:	1.4738E+01
Winsorized mean	:	1.0406E+01
Winsorized Standard deviation	:	5.1209E+00
Winsorized variance	:	2.6223E+01

95.0 % confidence interval:

Lower limit: 6.7152E+00 Upper limit: 1.2549E+01

Trimming 40% (for P=0.40):

Mean	:	9.1059E+00
Standard deviation	:	3.5294E+00
Variance	:	1.2457E+01
Winsorized mean	:	9.1412E+00
Winsorized Standard deviation	:	1.4739E+00
Winsorized variance	:	2.1722E+00

95.0 % confidence interval:

Lower limit: 7.2116E+00 Upper limit: 1.1000E+01

Biweight:

Mean	:	8.8034E+00
Standard deviation	:	2.5962E+00
Variance	:	6.7403E+00
Weight sqrt(w)	:	3.7120E+00

95.0 % confidence interval:

Lower limit: 7.3207E+00 Upper limit: 1.0286E+01

(5) ADAPTIVE ESTIMATORS OF LOCATION AND SPREAD:

Hoog's estimates:

Relative tail length	:	2.7797E+00
Mean	:	9.3353E+00
Standard deviation	:	3.7223E+00
Variance	:	1.3855E+01

95.0 % confidence interval:

Lower limit: 7.4215E+00 Upper limit: 1.1249E+01

Úloha E219. *Odhad střední hodnoty obsahu kadmia v bramborách*

Je třeba určit spolehlivý odhad střední hodnoty obsahu kadmia v bramborách v oblastech jižní Moravy a jižních Čech.

Data: z úlohy E219: obsah kadmia 100.x v bramborách [mg/l]

8.0	1.8	1.5	5.2	8.0	2.0	3.6	1.3	3.2	6.0	7.0	1.9	1.5	5.5
1.5	1.0	1.0	3.3	3.2	1.1	11.5	3.3	6.0	2.0	0.9	0.5	2.7	5.5

(1) Přehled popisných statistik: odhady je třeba vynásobit 10^{-2} [mg/l].

Odhady parametru polohy a rozptýlení (NCSS2000 a ADSTAT):

Střední hodnota parametru polohy	Bodový odhad	Dolní m.	Horní m.	n
Aritmetický průměr	3.571	2.512	4.631	28
Geometrický průměr	2.677	---	---	28
Harmonický průměr	1.983	---	---	28
Medián	2.95	1.433	4.467	28
Módus	1.20	---	---	28
5%ní uřezaný průměr	3.351	2.344	4 357	25
(jeho winsorizovaný odhad)	3.461	---	---	25
10%ní uřezaný průměr	3.210	2.117	4.303	22
(jeho winsorizovaný odhad)	3.468	---	---	22
40%ní uřezaný průměr	2.739	1.247	4.232	6
(jeho winsorizovaný odhad)	2.668	---	---	6
M-odhad	3.264	2.21	4.317	28
Hoggův *M*-odhad	3.571	2.512	4.631	28
Střední hodnota parametru rozptýlení				
Směrodatná odchylka	2.733	2.161	3.72	28
Rozpětí	11.0	---	---	28
Interkvartilové rozpětí	4.0	---	---	28

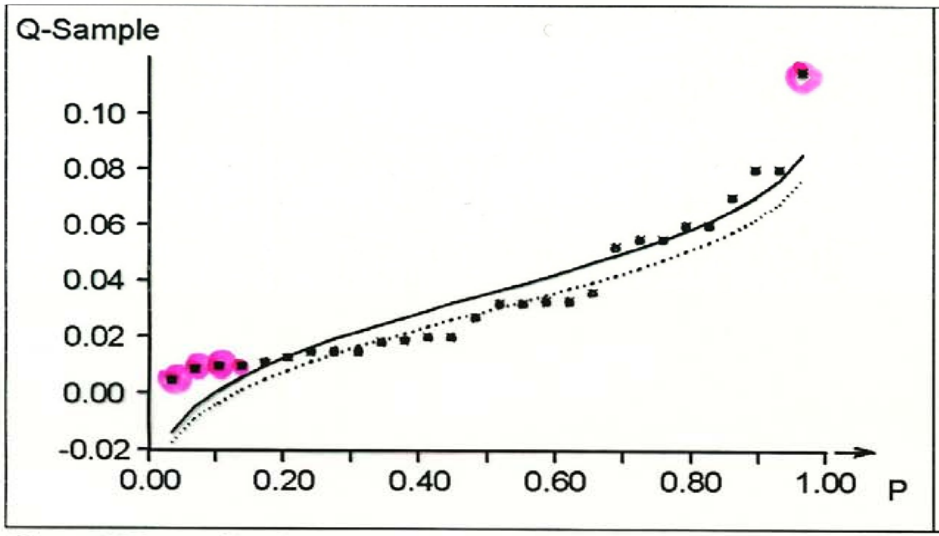
Odhady parametru tvaru:

šikmost $\hat{g}_1 = 1.12$ (odchylka od 0 je statisticky významná, nenormální rozdělení),

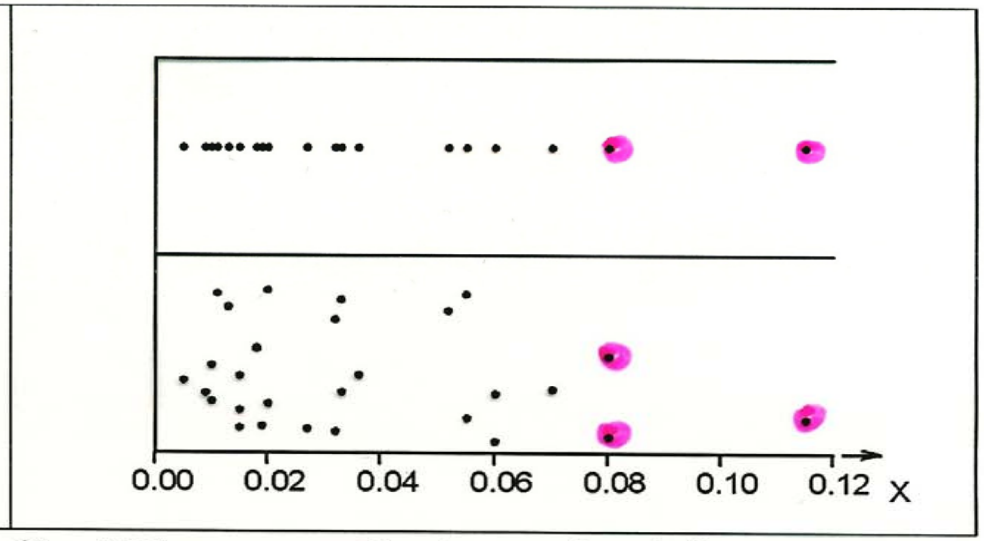
špičatost $\hat{g}_2 = 3.67$ (odchylka od 3 je statisticky nevýznamná, normální rozdělení).

(2) Základní diagnostické grafy EDA:

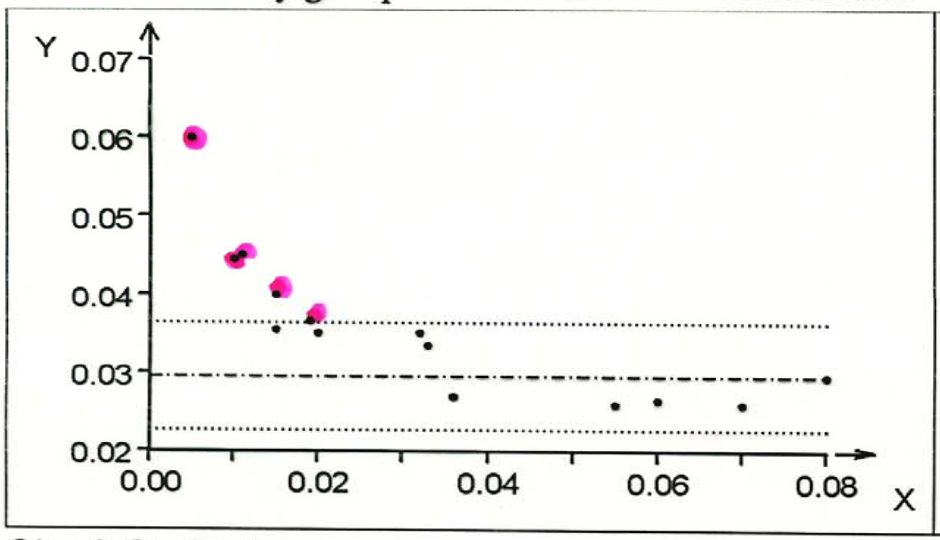
(obsah kadmia v bramborách)



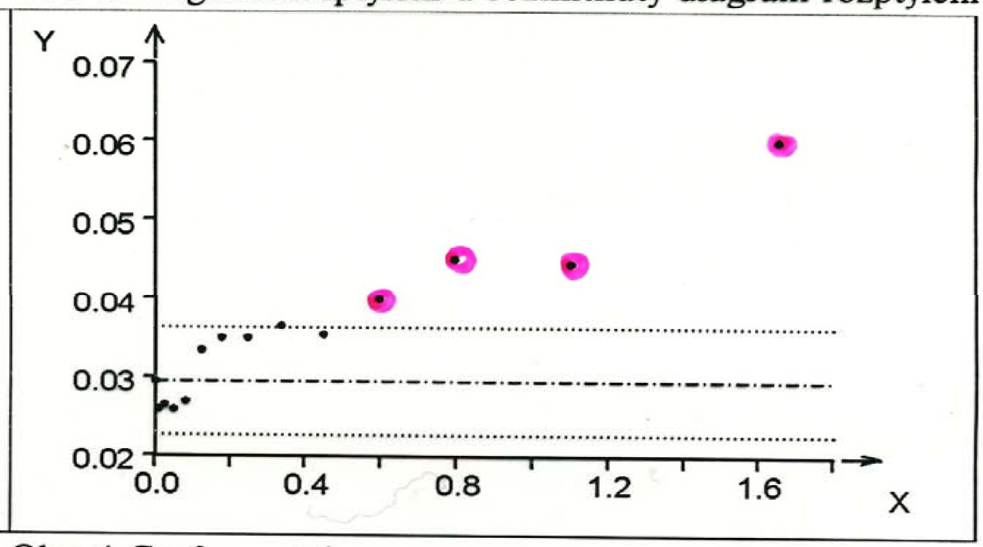
Obr. 1 Kvantilový graf pro obsah kadmia v bramborách



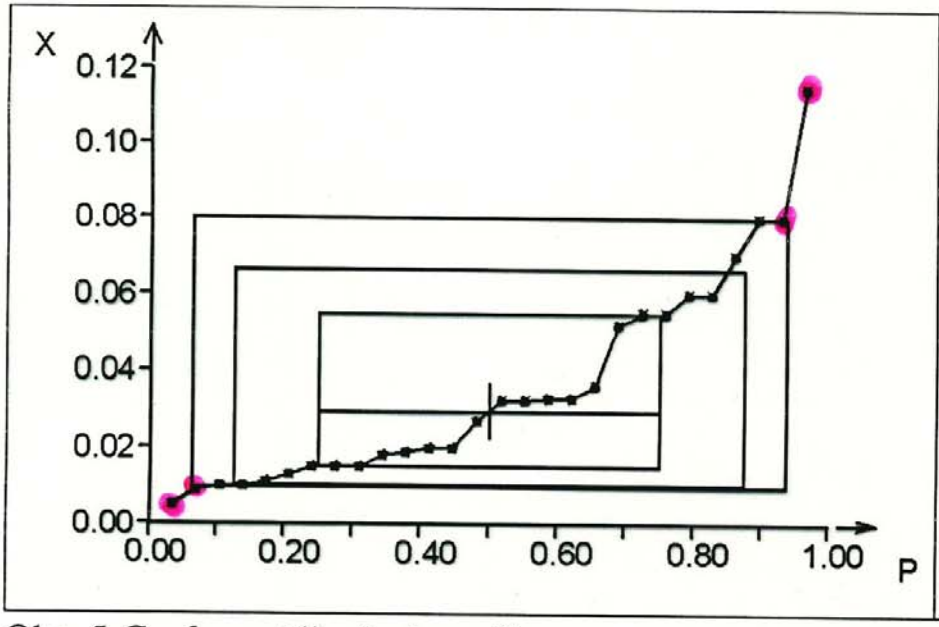
Obr. 2 Diagram rozptýlení a rozmítnutý diagram rozptýlení



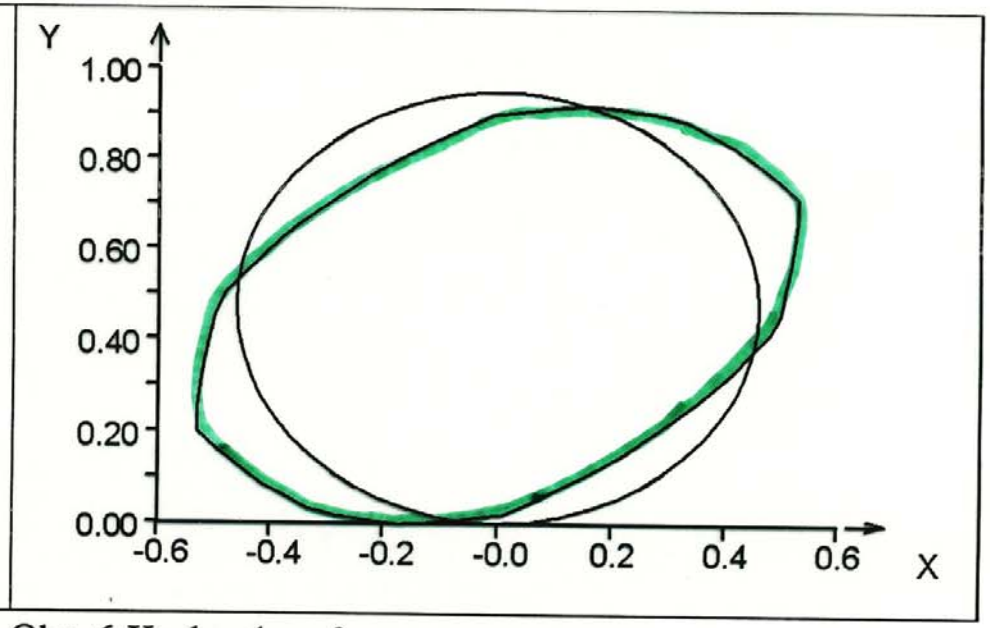
Obr. 3 Graf polosum



Obr. 4 Graf symetrie



Obr. 5 Graf rozptýlení s kvantily

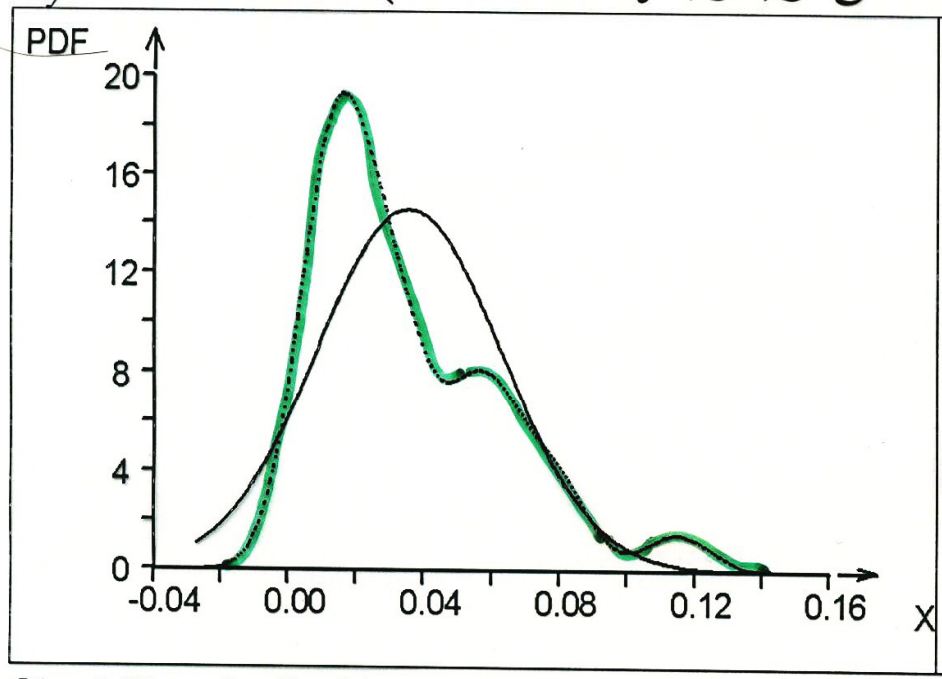


Obr. 6 Kruhový graf

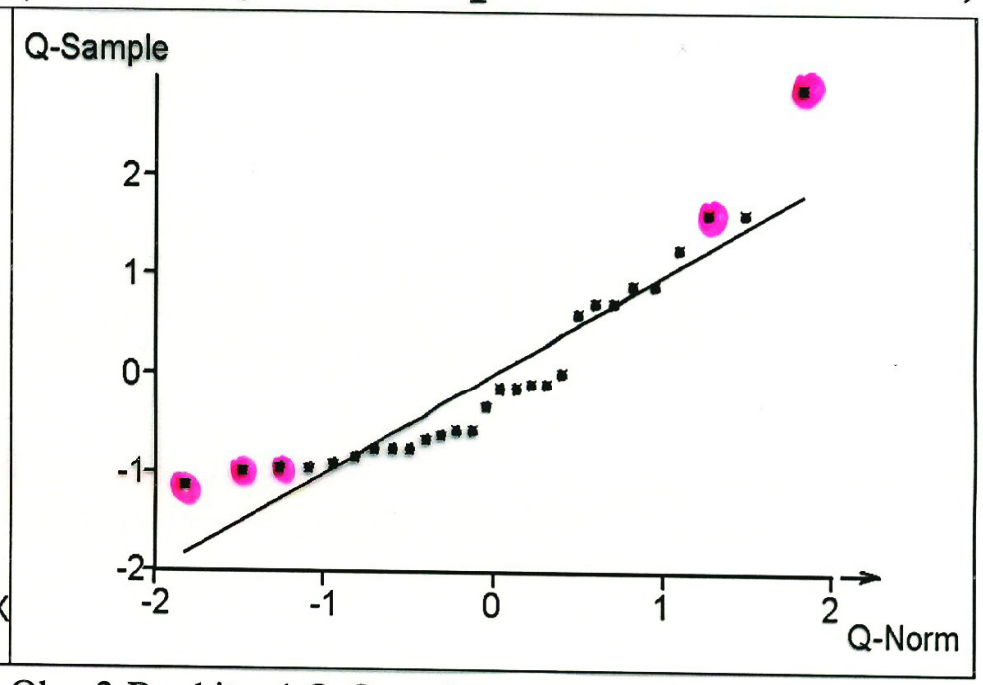
(3) Určení výběrového rozdělení EDA:

(obsah kadmia v bramborách)

r_{xy} = 0.9879 (rankitový *Q-Q* graf) ukazuje na exponen. rozdělení,



Obr. 1 Kernel odhad hustoty pravděpodobnosti



Obr. 2 Rankitový Q-Q graf

Kvantilové míry polohy, rozptýlení a tvaru pro obsah kadmia v bramborách [mg/l]

Kvantil	P	Dolní kvantil Q_D	Horní kvantil Q_H	Rozsah R_Q	Polosuma Z_Q	Šikmost S_Q	Délka konců T_Q
Median	0.5	2.95	2.95	-			
Kvartil	0.25	1.50	5.50	4.00	3.50	-0.004	0.000
Oktil	0.125	1.04	6.63	5.59	3.83	0.093	**0.334**
Sedecil	0.0625	0.97	8.00	7.03	4.48	0.093	**0.564**

Tabulovaná délka oktilových konců pro norm. rozdělení T_E = 0.534, (nalezeno 0.334),
Tabulovaná délka sedecilových konců pro norm. rozdělení T_D = 0.822, (nalezeno 0.564),

Bodový odhad šikmosti $\hat{g}_1$ = 1.12,

Bodový odhad špičatosti $\hat{g}_2$ = 3.67.

Závěr: výběrové rozdělení je sešikmené a nedá se aproximovat normálním.

(4) Ověření základních předpokladů o výběru:

(obsah kadmia v bramborách)

(a) ***Vyšetření nezávislosti prvků výběru*:**

von Neumannův test nezávislosti prvků $t_{17} = 0.083 < t_{0.975}(28+1) = 2.045$, a proto je nezávislost přijata.

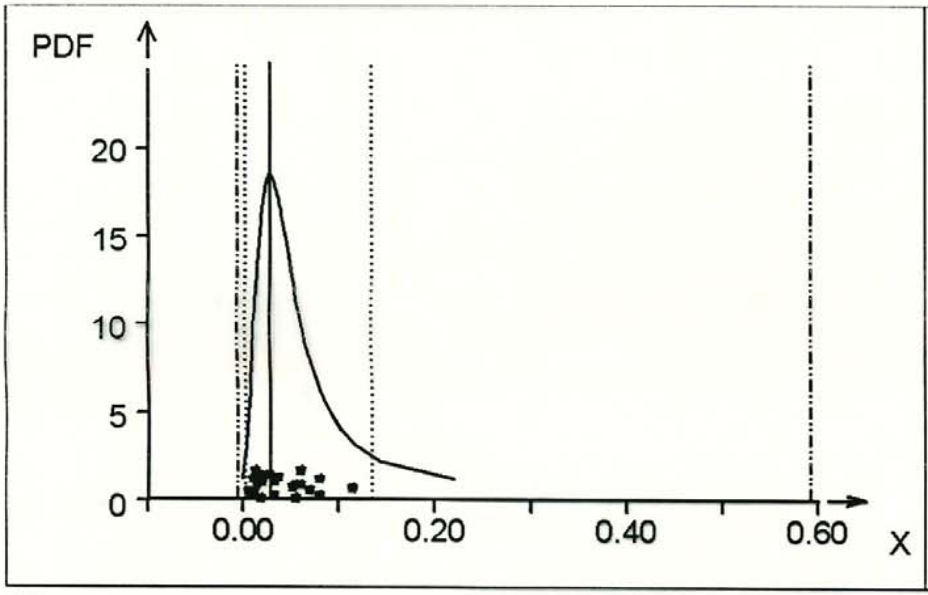
(b) ***Vyšetření normality výběrového rozdělení:***

Jarque-Berrův test šikmosti a špičatosti $C_1 = 8.737 > \chi^2(0.95, 2) = 5.992$ dokazuje, že předpoklad normality je zamítnut.

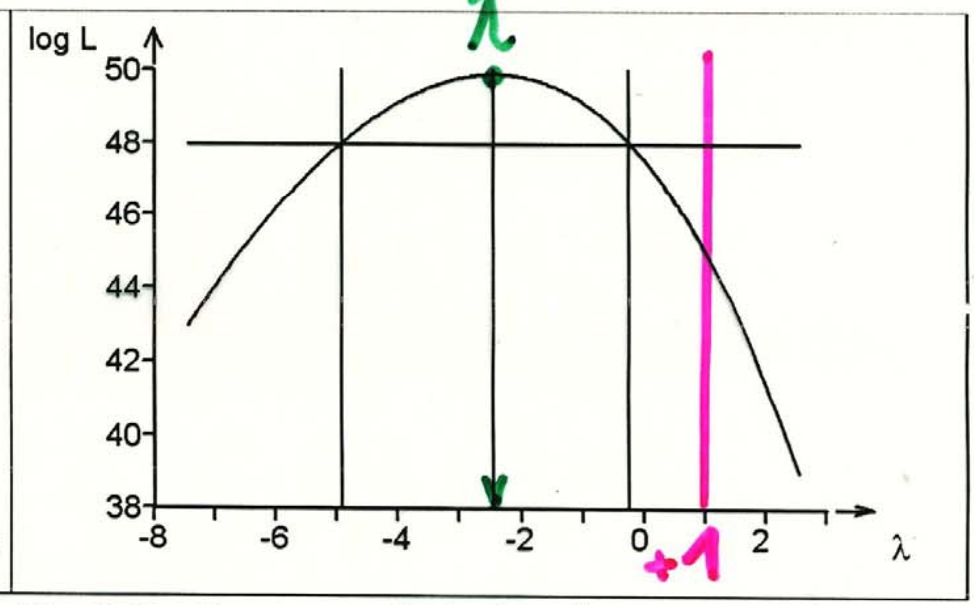
(c) ***Vyšetření homogenity výběru:***

vně intervalu Hoaglinových mezí $[B^*_L = -6.98; B^*_U = 13.99]$ nejsou žádné odlehlé hodnoty.

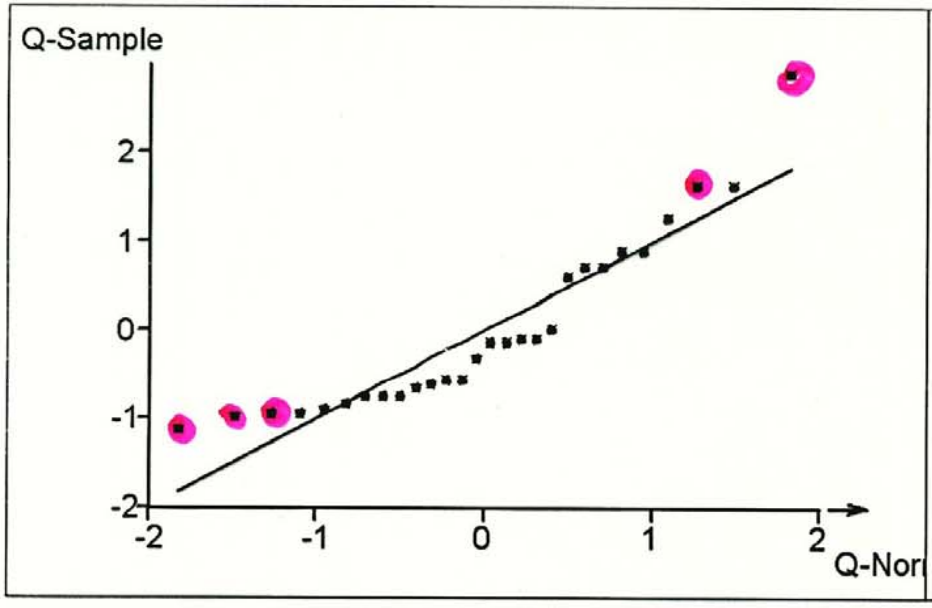
(5) Transformace dat:



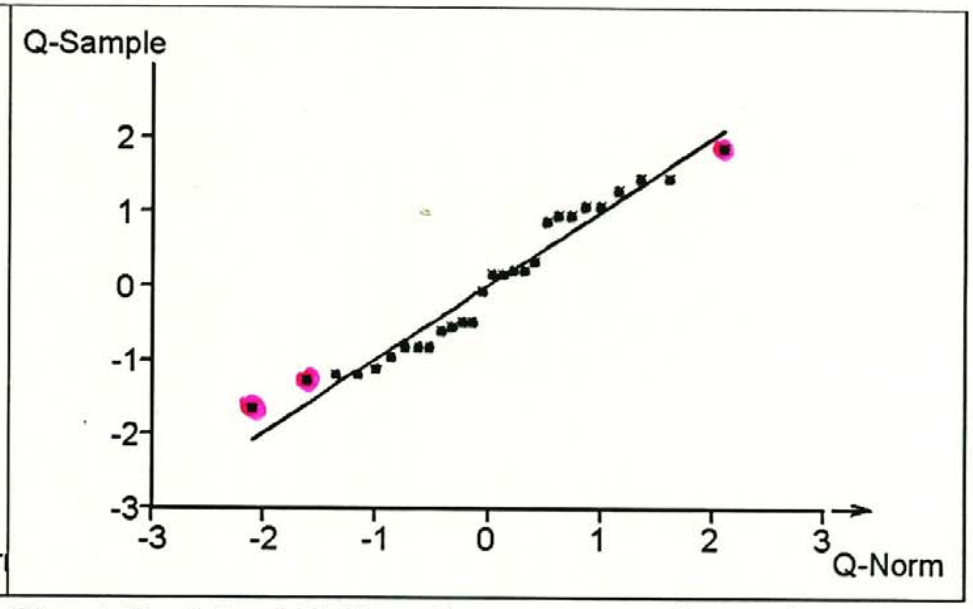
Obr. 1 Kernel odhad hustoty pravděpodobnosti pro obsah kadmia v bramborách



Obr. 2 Graf log. max. věrohodnosti po exp. transformaci



Obr. 3 Rankitový Q-Q graf po Box-Cox transformaci



Obr. 4 Rankitový Q-Q graf po exp. transformaci

Schema mocninné a Box-Coxovy transformace

x	y (mocninná)	y (Box-Cox)
0.5	0.9117	-0.6621
0.9	0.9861	-1.0462
1.0	1.0000	0.0000
1.0	1.0000	0.0000
1.1	1.0128	0.0959
1.3	1.0356	0.2670
1.5	1.0556	0.4166
1.5	1.0556	0.4166
1.5	1.0556	0.4166
1.8	1.0815	0.6114
1.9	1.0893	0.6701
2.0	1.0968	0.7262
2.0	1.0968	0.7262
2.7	1.1416	1.0620
3.2	1.1678	1.2582
3.2	1.1678	1.2582
3.3	1.1726	1.2942
3.3	1.1726	1.2942
3.6	1.1862	1.3968
5.2	1.2459	1.8439
5.5	1.2552	1.9140
5.5	1.2552	1.9140
6.0	1.2699	2.0239
6.0	1.2699	2.0239
7.0	1.2962	2.2217
8.0	1.3195	2.3963
8.0	1.3195	2.3963
11.5	1.3849	2.8869

Průměr	$\bar{x} = 3.5714$	$\bar{y} = 1.1455$	$\bar{y} = 1.0987$
Směrodatná odchylka	$s = 2.7333$	$s = 0.1214$	$s = 0.9107$
Retransform. průměr (mocnina)	$\bar{x}_R = 2.788$		
	$s = 2.2147$		
Retransform. průměr (Box-Cox)	$\bar{x}_R = 2.788$		
	$s = 2.2147$		

Schema mocninné a Box-Coxovy transformace

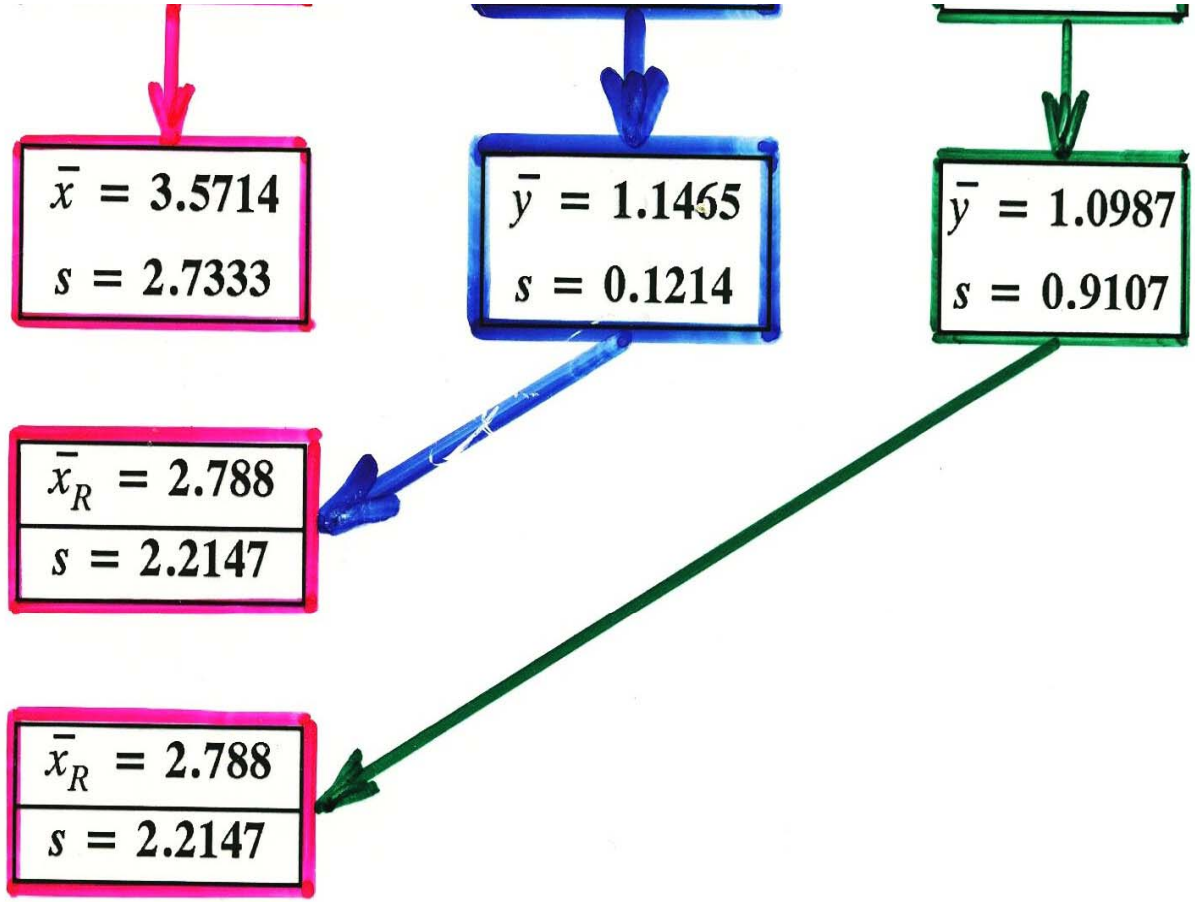
x	y (mocninná)	y (Box-Cox)
0.5	0.9117	-0.6621
0.9	0.9861	-1.0462
1.0	1.0000	0.0000
1.0	1.0000	0.0000
1.1	1.0128	0.0959
1.3	1.0356	0.2670
1.5	1.0556	0.4166
1.5	1.0556	0.4166
1.5	1.0556	0.4166
1.8	1.0815	0.6114
1.9	1.0893	0.6701
2.0	1.0968	0.7262
2.0	1.0968	0.7262
2.7	1.1416	1.0620
3.2	1.1678	1.2582
3.2	1.1678	1.2582
3.3	1.1726	1.2942
3.3	1.1726	1.2942
3.6	1.1862	1.3968
5.2	1.2459	1.8439
5.5	1.2552	1.9140
5.5	1.2552	1.9140
6.0	1.2699	2.0239
6.0	1.2699	2.0239
7.0	1.2962	2.2217
8.0	1.3195	2.3963
8.0	1.3195	2.3963
11.5	1.3849	2.8869

Průměr

Směrodatná odchylka

Retransform. průměr (mocnina)

Retransform. průměr (Box-Cox)



(6) Závěr:

(určení obsahu kadmia v bramborách, odhady nutno vynásobit rozměrem 10^{-2} [mg/kg])

Klasické odhady: (aritmetický průměr a směrodatná odchylka)

$$\bar{x} = 3.57,\ s = 2.73 \text{ a } L_D = 2.51 \text{ a } L_H = 4.63$$

Re-transformované odhady:

1. Symetrizující mocninná transformace:

$\hat{\lambda} = 0.13$ dává $\bar{x}_R = 2.79,\ s_R = 2.21$ a $L_D = 2.04$ a $L_H = 3.77$

2. Normalizující Box-Coxova transformace:

$\hat{\lambda} = 0.13$ dává $\bar{x}_R = 2.79,\ s_R = 2.21$ a $L_D = 2.04$ a $L_H = 3.77$

3. Exponenciální transformace

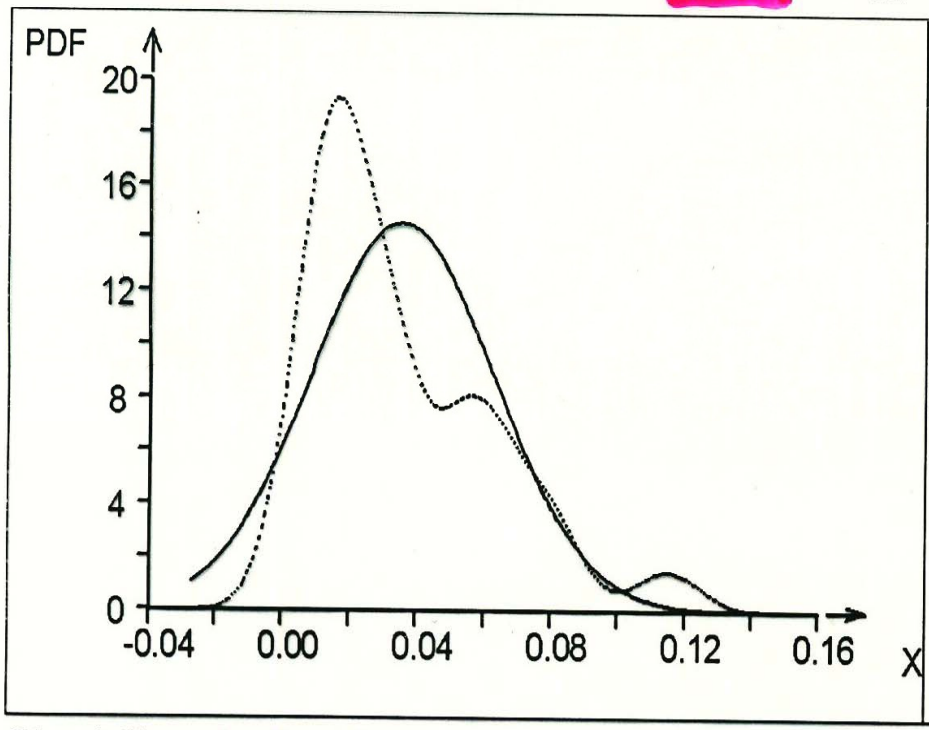
$\hat{\lambda} = 1.07$ dává $\bar{x}_R = 2.67,\ s_R = *.**$ a $L_D = 2.03$ a $L_H = 3.53$

(*.** značí nevyčísleno)

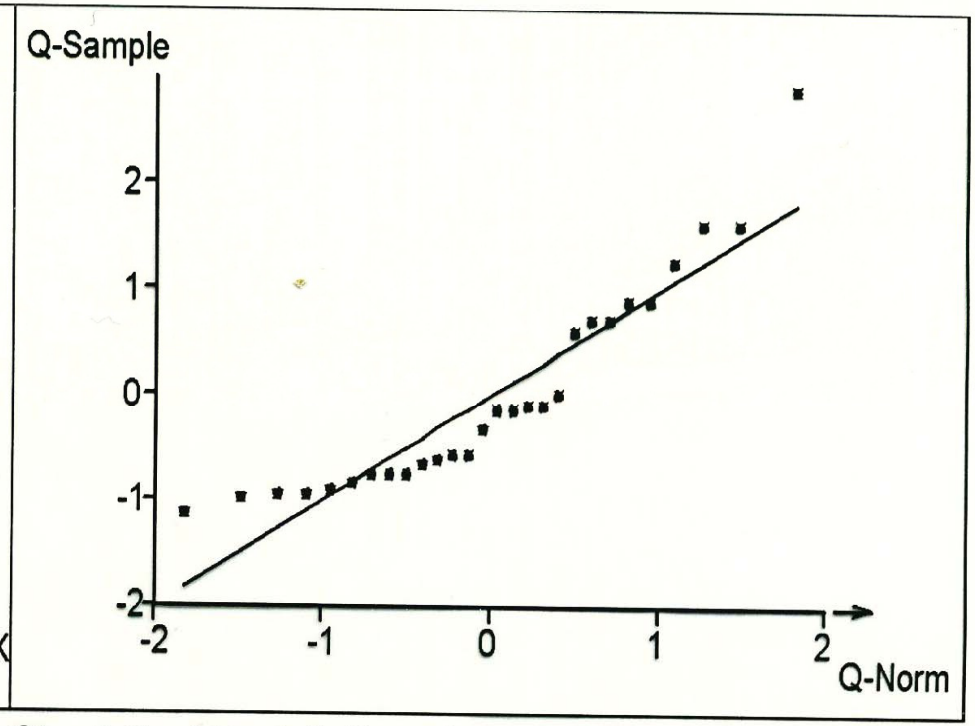
Adaptivní odhady (robustní): selhávájí u asym. rozdělení

M-odhad $\hat{\mu}_M = 3.26$, $\hat{\sigma}_M = 2.60$ a $L_L = 2.21$ a $L_U = 4.32$

M-odhad (Hoog $\hat{\mu}_M = 3.27$, $\hat{\sigma}_M = 2.73$ a $L_L = 2.51$ a $L_U = 4.63$



Obr. 1 Kernel odhad hustoty pravděpodobnosti



Obr. 2 Rankitový Q-Q graf