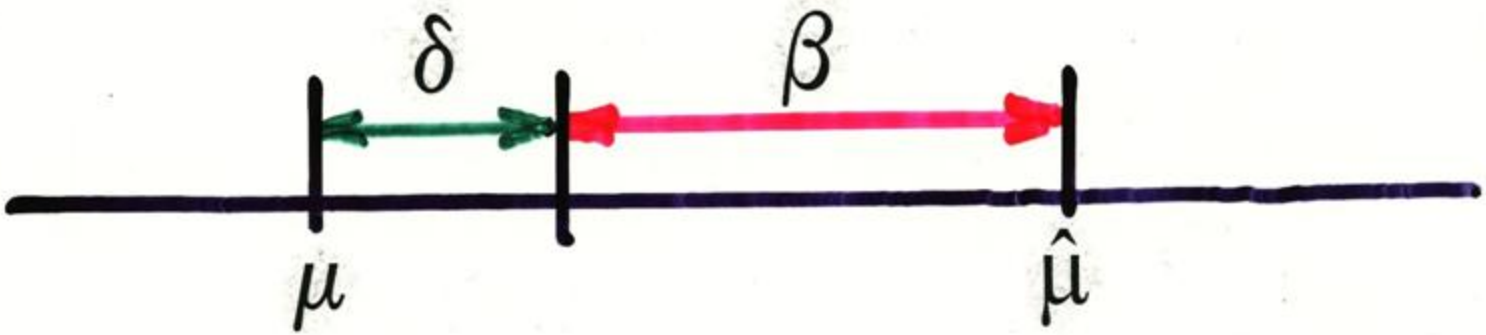
Správnost

Systematické chyby:

β ***vychýlení*** (bias) špatným vyhodnocením, matematickou metodou, nevhodným software

δ ***odchylka*** špatným experimentem, nezvládnutou strategií



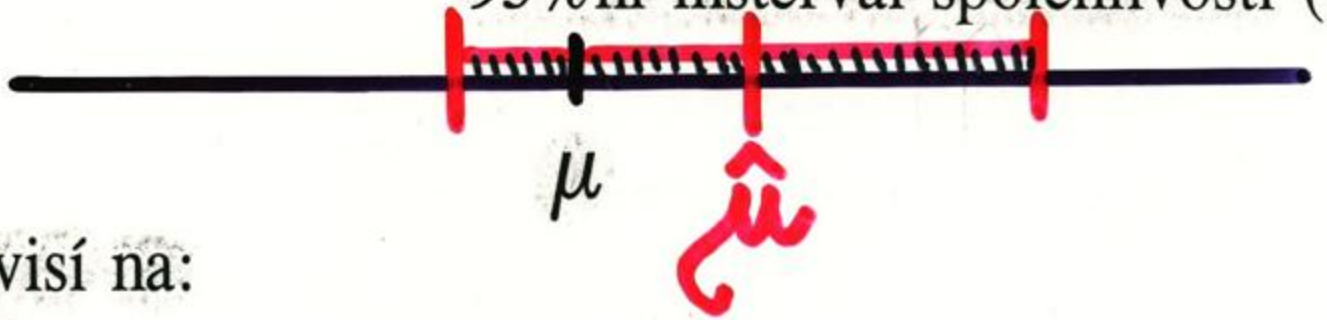
$(\bar{x}, \bar{x}_{kv}, \bar{x}_g, \bar{x}_R, \hat{x}_M, \hat{x}_P, \tilde{x}_{0.5}, P_L, \ldots..)$

Přesnost

Náhodné chyby:

ε ***odchylka*** šum, neodstranitelná kolísání (exp., instrum.)

95%ní insterval spolehlivosti ($\pm$ 2s)



Šíře i. s. závisí na:

1. Četnosti n.
2. Velikost s. (Směrodatná odchylka, nejistota, zákon propagace).
3. Statistické jistotě 90%, 95%, 99%, …

Postup analýzy dat

1. Průzkumová analýza dat:

Diagnostické grafy: *stupeň symetrie rozdělení*
lokální koncentrace dat
vybočující data

2. Ověření předpokladů výběru dat:

Diagnosticky, testy: *ověření normality*
ověření nezávislosti
ověření homogenity
určení minimální četnosti

3. Transformace dat:

Analýza dat: *originální data*
data po mocninné transformaci
data po Box-Coxově transformaci

4. Parametry polohy, rozptýlení a tvaru:

Analýza 1 výběru:

klasické odhady
- *průměr*
- *rozptyl*

robustní odhady
- *medián*
- *uřezané průměry*
- *winsorizovaný rozptyl*
- *interkvantilové rozpětí*

adaptivní odhady

5. Testování dvou výběrů:

a) Testy polohy

b) Testy rozptýlení

Postup analýzy dat

1. Průzkumová analýza dat:

Diagnostické grafy: *stupeň symetrie rozdělení*
lokální koncentrace dat
vybočující data

2. Ověření předpokladů výběru dat:

Diagnosticky, testy: *ověření normality*
ověření nezávislosti
ověření homogenity
určení minimální četnosti

3. Transformace dat:

Analýza dat: *originální data*
data po mocninné transformaci
data po Box-Coxově transformaci

4. Parametry polohy, rozptýlení a tvaru:

Analýza 1 výběru: *klasické odhady* - *průměr*
- *rozptyl*

robustní odhady - *medián*
- *uřezané průměry*
- *winsorizovaný rozptyl*
- *interkvantilové rozpětí*

adaptivní odhady

PRŮZKUMOVÁ ANALÝZA DAT (EDA)

Cílem **průzkumové analýzy dat je nalezení zvláštností statistického chování dat a ověření jejich předpokladů pro následné zpracování „klasickými“ statistickými metodami.**

EDA – Exploratory Data Analysis (Tuckey, Chambers)

Statistické zvláštnosti dat	Základní předpoklady o výběru
Nesymetrie (levostranné nebo pravostranné)	Homogenita rozdělení
	Shoda s teoretickým rozdělením (obvykle normálním)
Lokální koncentrace dat (špičatost nebo plochost)	Potřebná velikost výběru
Extrémní data, odlehlé hodnoty	Nezávislost dat

Průzkumová analýza dat

Účel: (a) Odhalit zvláštnosti v datech (lokální koncentrace dat, symetrie rozdělení výběru),

(b) ověřit předpoklady výběru.

Jednorozměrný výběr: sledován pouze jeden znak.

Počet prvků: (1) může být ***omezený*** a soubor je konečný

(2) může být ***neomezený*** a soubor je nekonečný.

Náhodný výběr, $\{x_i\}$, $i = 1, ..., n$, jehož prvky, tj. naměřené hodnoty, jsou chápány jako realizace jisté náhodné veličiny.

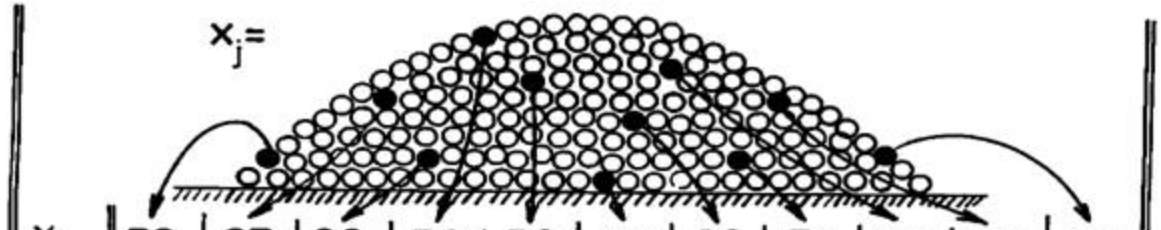
Reprezentativní náhodny výběr je charakterizován předpoklady:

1. Jednotlivé prvky výběru x_i jsou vzájemně **nezávislé**.

2. Výběr je **homogenní**, tj. všechny prvky x_i pocházejí ze stejného rozdělení s konstantním rozptylem, nejsou žádné vybočující prvky.

3. Předpokládá se **normální** rozdělení pravděpodobnosti.

4. Všechny prvky souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru.



x_j	52	37	90	56	58	45	66	75	36	100	68
j	1	2	3	4	5	6	7	8	9	10	11

$x_{(i)}$	36	37	45	52	56	58	66	68	75	90	100
$R_{P_i}=i$	1	2	3	4	5	6	7	8	9	10	11
K_{P_i}	11	10	9	8	7	6	5	4	3	2	1
H_i	1	2	3	4	5	6	5	4	3	2	1

Ze základního souboru je vybrán náhodný výběr x_j, $j = 1, \ldots, n$, o velikosti $n = 11$ a jsou určeny pořádkové statistiky $x_{(i)}$.

Technika pořadí a hloubek: každá pořádková statistika $x_{(i)}$ má své

rostoucí pořadí $R_{P_i} = i$,

klesající pořadí $K_{P_i} = n + 1 - i$,

hloubku H_i , což je menší číslo z obou uvedených pořadí

$$H_i = \min(R_{P_i}, K_{P_i}).$$

Pořádkové statistiky: vzestupně setříděné prvky výběru $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Pořadová pravděpodobnost je dána vztahem

$$P_i = \frac{i}{n+1} \quad \text{nebo v praxi} \quad P_i = \frac{1 - \frac{3}{8}}{n + \frac{1}{4}}$$

$100P_i$ procentní výběrový kvantil je hodnota, pod kterou leží $100P_i$ % prvků výběru.

Rozptyl kvantilu $\tilde{x}_\alpha$: vypočte se dle

$$D(\tilde{x}_\alpha) = \frac{\alpha\,(1 - \alpha)}{n\,[f(\tilde{x}_\alpha)]^2}$$

kde $f(\tilde{x}_\alpha)$ je výběrová hustota pravděpodobnosti v bodě $\tilde{x}_\alpha$.

Písmenové hodnoty: kvantily pro pořadové pravděpodobnosti

$$P_i = 2^{-i},\ i = 1, 2, \ldots ,$$

i	i-tý kvantil	Pořadová pravděpodobnost P_i	Symbol písmenové hodnoty L
1	medián	$2^{-1} = 1/2$	M
2	kvartily	$2^{-2} = 1/4$	F
3	oktily	$2^{-3} = 1/8$	E
4	sedecily	$2^{-4} = 1/16$	D

Dolní a horní písmenové hodnoty:

dolní písmenová hodnota je pro $P_i = 2^{-i}$,

horní písmenová hodnota je pro $P_i = 1 - 2^{-i}$

$i = 2$,	$L_D = F_D$,	dolní kvartil,
	$L_H = F_H$,	horní kvartil,
$i = 3$,	$L_D = E_D$,	dolní oktil,
	$L_H = E_H$,	horní oktil,
$i = 4$,	$L_D = D_D$,	dolní sedecil,
	$L_H = D_H$,	horní sedecil, ... atd.

Hloubka dolních písmenných hodnot: vypočte se dle

$$H_L = \frac{1 + \text{int}\,(H_{L-1})}{2}$$

kde L jsou indexy F, E, D a int(x) značí celočíselnou část čísla x.

Pokud je L = F, bere se L - 1 = M.

H_L celé číslo: dolní kvantil $L_D = x_{(H_L)}$

horní kvantil $L_H = x_{(n+1-H_L)}$

H_L číslo necelé: lineární interpolace podle vztahů

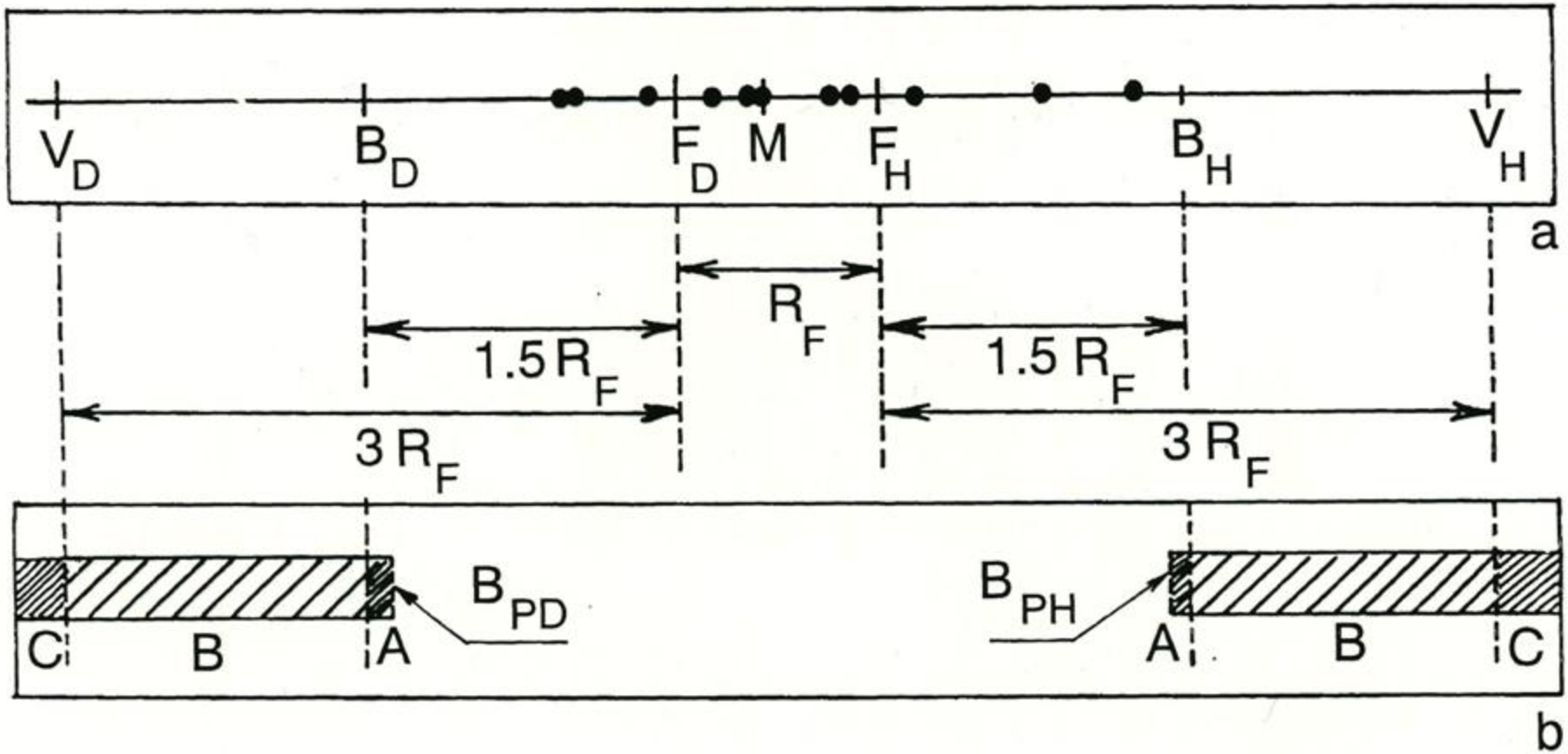
$$L_D = \frac{x_{int(H_L)} + x_{int(H_L)+1}}{2}$$

$$L_H = \frac{x_{n+1-int(H_L)} + x_{n+2-int(H_L)}}{2}$$

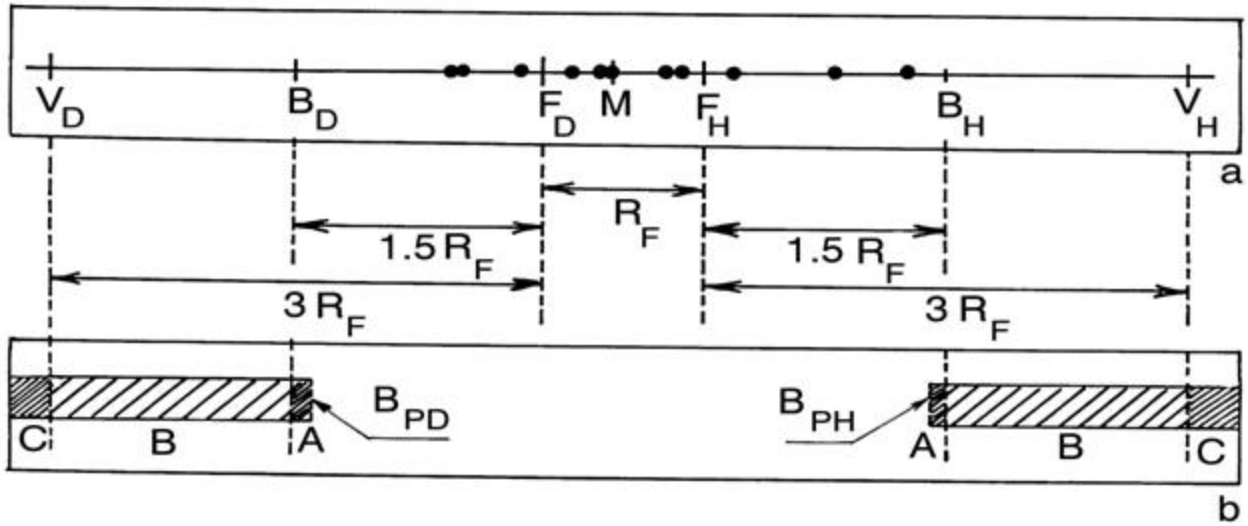
Počet písmenových hodnot n_L: včetně mediánu se určí dle

$$n_L \approx 1.44 \ln(n + 1)$$

Užívání písmenových hodnot



Kvartilové rozpětí: $R_F = F_H - F_D$



Přilehlé hodnoty:

$$B_H = F_H + 1.5\ R_F$$ Horní přilehlá hodnota

$$B_D = F_D - 1.5\ R_F$$ Dolní přilehlá hodnota

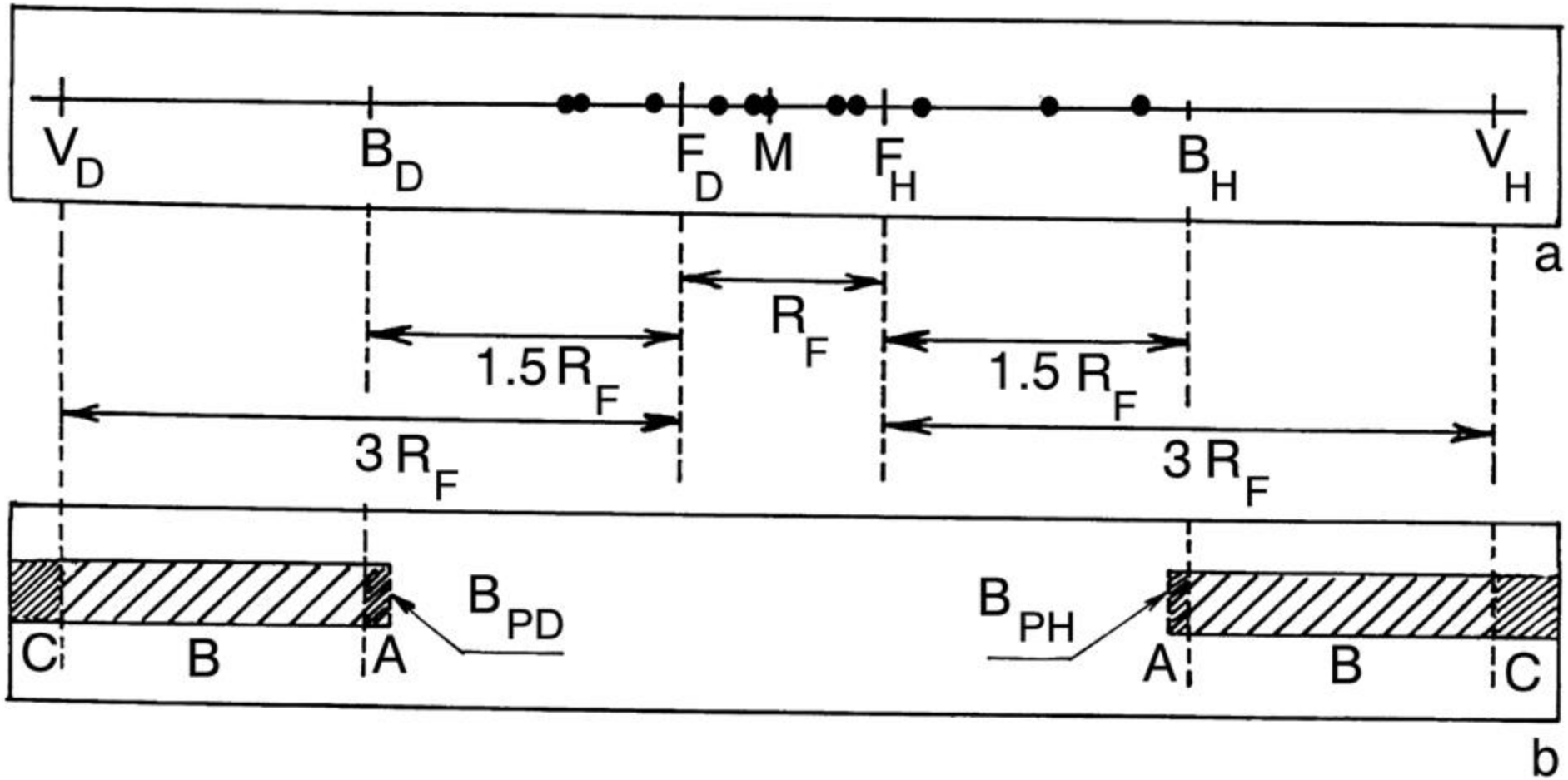
Vnitřní hradby:

$$V_H = F_H + 3\ R_F$$ Horní vnitřní hradba

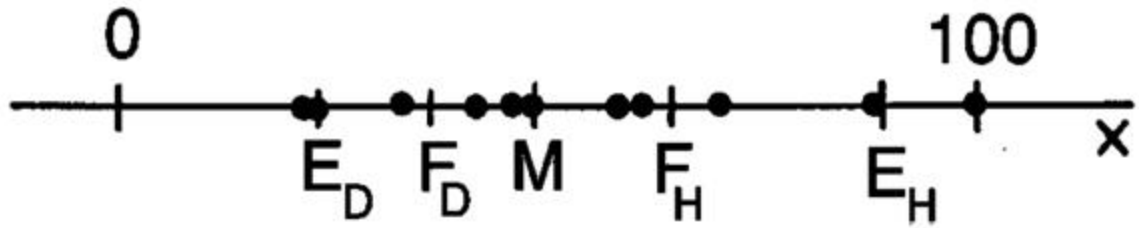
$$V_D = F_D - 3\ R_F$$ Dolní vnitřní hradba

"Podezřelá měření" (=vybočující hodnoty):

Body mimo interval $\{V_D, V_H\}$ vnitřních hradeb



Konstrukce barierově-číslicového schématu indikujícího vybočující hodnoty: a) diagram rozptýlení s mediánem M, kvartily F_D (dolní) a F_H (horní) vnitřními hradbami B_D (dolní) a B_H (horní), vnějšími hradbami V_D (dolní) a V_H (horní); b) oblast vybočujících hodnot - A označuje přilehlé body (B_{PD} je blízké B_D a B_{PH} je blízké B_H), B označuje oblast vnějších a C oblast vzdálených bodů.



i	1	2	3	4	5	6	7	8	9	10	11
$x_{(i)}$	36	37	45	52	56	58	66	68	75	90	100

(a)

L	H_L	L_D	(P_L)	L_H	R_L
		n=...			
M	H_M		M		
F	H_F	F_D	(P_F)	F_H	R_F
E	H_E	E_D	(P_E)	E_H	R_E
D	H_D	D_D	(P_D)	D_H	R_D
...	...	...	...	...	...
...	1	$x_{(1)}$	(P_c)	$x_{(n)}$	R_c

(b) n=11

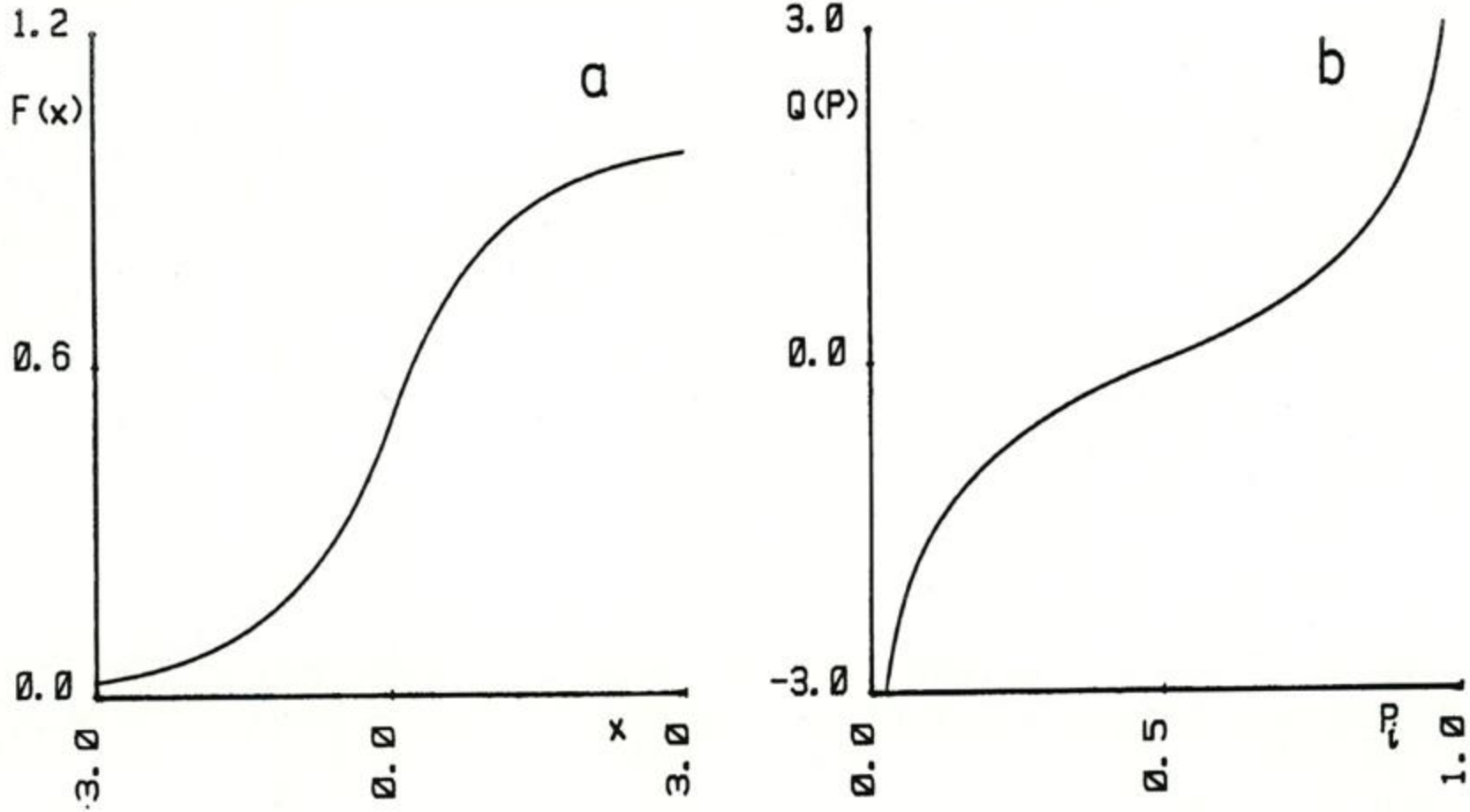
M			58		
F	3.5	48.5	(60)	71.5	23
E	2	37	(63.5)	90	53
	1	36	(68)	100	64

Graficko-tabelární schéma sumarizace dat z předešlé části obrázku:

a) obecné schéma písmenově-číslicového zápisu výběru,

b) sedmipísmenový zápis výběru.

Kvantilová funkce Q(P): $x_{(i)}$ proti P_i, i = 1, ..., n, je inverzní k funkci distribuční $F(x_i)$ čili $Q(P_i) = F^{-1}(P_i)$



(a) Distribuční funkce $F(x_i)$ a (b) kvantilová funkce $Q(P_i)$ Laplaceova rozdělení s nulovou střední hodnotou a rozptylem rovným 2

Střední hodnota i-té pořádkové statistiky: rovna $100P_i$ %nímu kvantilu rozdělení výběru $E(x_{(i)}) = F^{-1}(P_i) = Q(P_i)$, kde $F(x)$ je distribuční funkce a $Q(P_i)$ kvantilová funkce výběru.

100α %ní kvantil $\tilde{x}_\alpha$: pro α z intervalu <0, 1> se vyčíslí dle

$$\tilde{x}_\alpha = (n+1)\left(\alpha - \frac{i}{n+1}\right)(x_{(i+1)} - x_{(i)}) + x_{(i)}$$

Pro index i musí být splněna nerovnost

$$\frac{i}{n+1} \leq \alpha \leq \frac{i+1}{n+1}$$

Identifikace statistických zvláštností výběru dat

(1) Stupeň symetrie rozdělení výběru

(2) Stupeň špičatosti rozdělení výběru

(3) Lokální koncentrace dat

(4) Přítomnost vybočujících hodnot (měření)

Pomůcky identifikace statistických zvláštností dat v EDA

Grafické diagnostiky:

Spojité rozdělení:

- G1 Kvantilový graf
- G2 Diagram rozptýlení
- G3 Rozmítnutý diagram rozptýlení
- G4 Krabicový graf
- G5 Vrubový krabicový graf
- G6 Graf polosum
- G7 Graf symetrie
- G8 Graf špičatosti
- G9 Diferenční kvantilový graf
- G10 Graf rozptýlení s kvantily
- G11 Odhad hustoty pravděpodobn.
- G12 Histogram (polygon)
- G13 Kvantil-kvantilový Q-Q graf
- G14 Rankitový graf
- G15 Podmíněný rankitový graf
- G16 Pravděpodobnostní P-P graf
- G17 Kruhový graf

Diskrétní rozdělení:

- G18 Graf poměrů frekvencí
- G19 Poissonův graf
- G20 Modifikovaný Poissonův graf

Spojité rozdělení (transformace):

- G21 Hinesové-Hinesův selekční graf
- G22 Graf logaritmu věrohodnost. funkce

Testy:

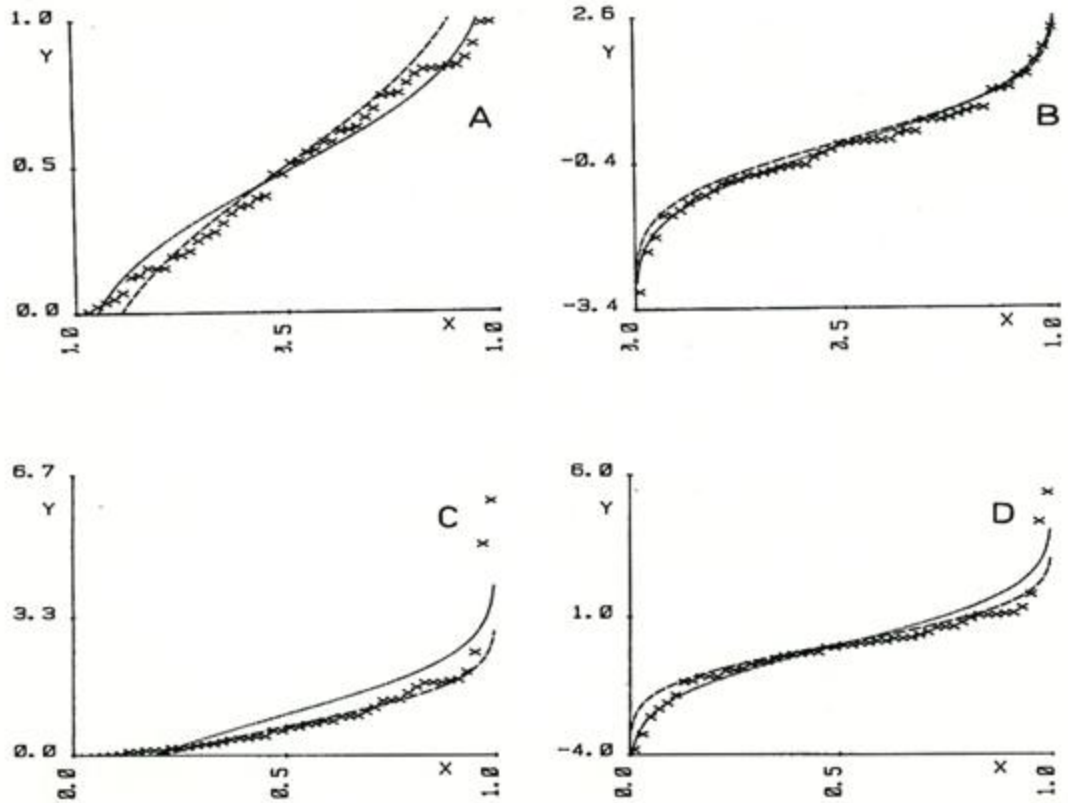
- Testy normality rozdělení dat
- Testy homogenity dat
- Test nezávislosti dat
- Výpočet minimální velikosti výběru dat

Grafické diagnostiky EDA

Kvantilový graf (G1)

Osa x: pořadová pravděpodobnost P_i,

Osa y: pořádková statistika $x_{(i)}$,



Kvantilové grafy (robustní --- a klasické ...) pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova

(G1)

Diagnoza:

- přehledně znázornit data, lokální koncentraci dat,
- rozlišit tvar rozdělení (symetrický, sešikmený)
- identifikovat vybočující data, atd.,
- zakreslují se i kvantilové funkce normálního rozdělení

$N_{P_i} = \hat{\mu} + \hat{\sigma} . u_{P_i}$ pro $0 \leq P_i \leq 1$:

(. . . . .) Klasických odhadů $\hat{\mu} = \bar{x}$ a $\hat{\sigma} = s$,

(- - - - - -) Robustních odhadů $\hat{\mu} = \tilde{x}_{0.5}$ a $\hat{\sigma} = R_F / 1.349$.

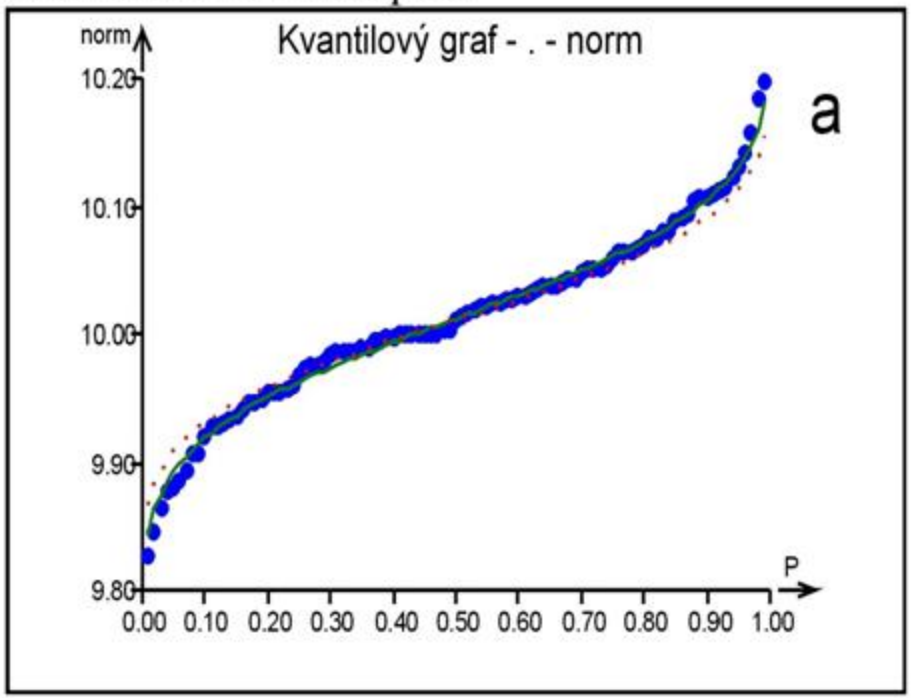
(G1)

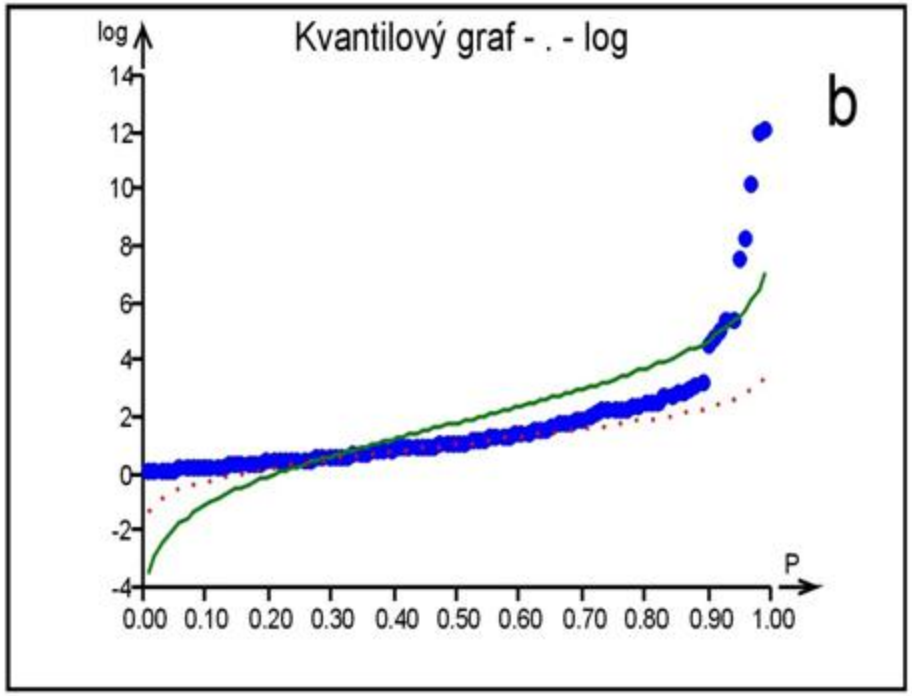
Kvantilový graf (osa x: pořadová pravděpodobnost P_i, osa y: pořádková statistika $x_{(i)}$).

Zakreslují se kvantilové funkce normálního rozdělení, $N_{P_i} = \hat{\mu} + \hat{\sigma}\ u_{P_i}$, pro $0 \le P_i \le 1$:

(1) *klasických odhadů* parametrů polohy a rozptýlení, tj. aritmetického průměru a směrodatné odchylky $\mu = \bar{x}$ a $\hat{\sigma} = s$, a dále

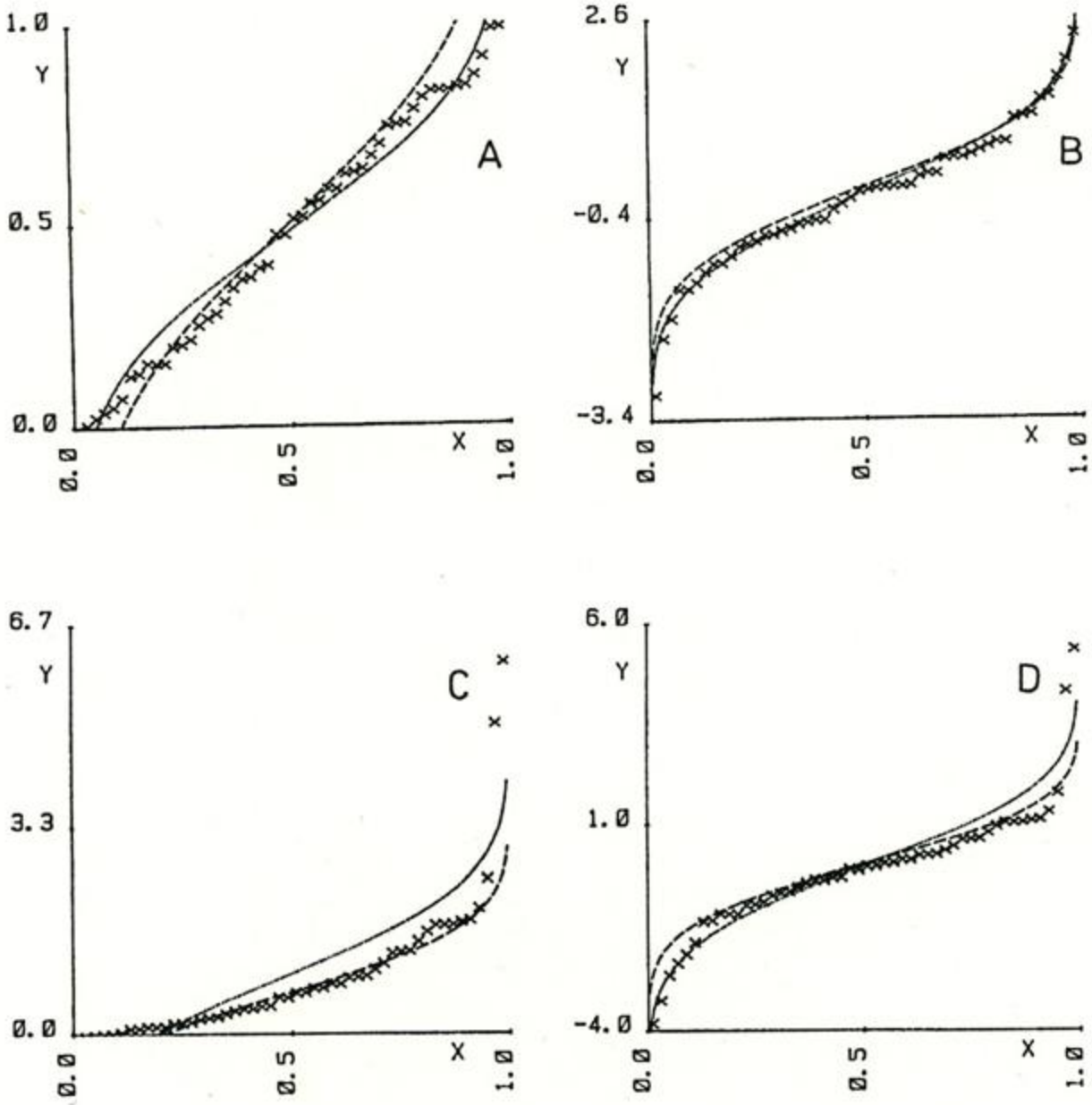
(2) *robustních odhadů*, tj. mediánu M, $\hat{\mu} = M$ a $\hat{\sigma} = R_F/1.349$, kde $R_F = F_H - F_D$ je interkvartilové rozpětí.





Obr. 2.6 Kvantilový graf (robustní --- a klasický ...) pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

(G1)

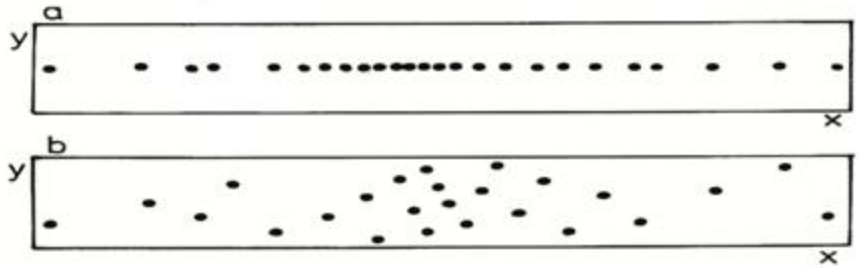


Obr. 2.5 Kvantilové grafy (robustní --- a klasické ...) pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova

Diagram rozptýlení (G2)

Osa x: hodnoty x,
Osa y: libovolná úroveň, obyčejně y = 0

Obsah: jednorozměrná projekce kvantilového grafu do osy x.



Konstrukce a) diagramu rozptýlení a b) rozmítnutého diagramu rozptýlení

Diagnoza:

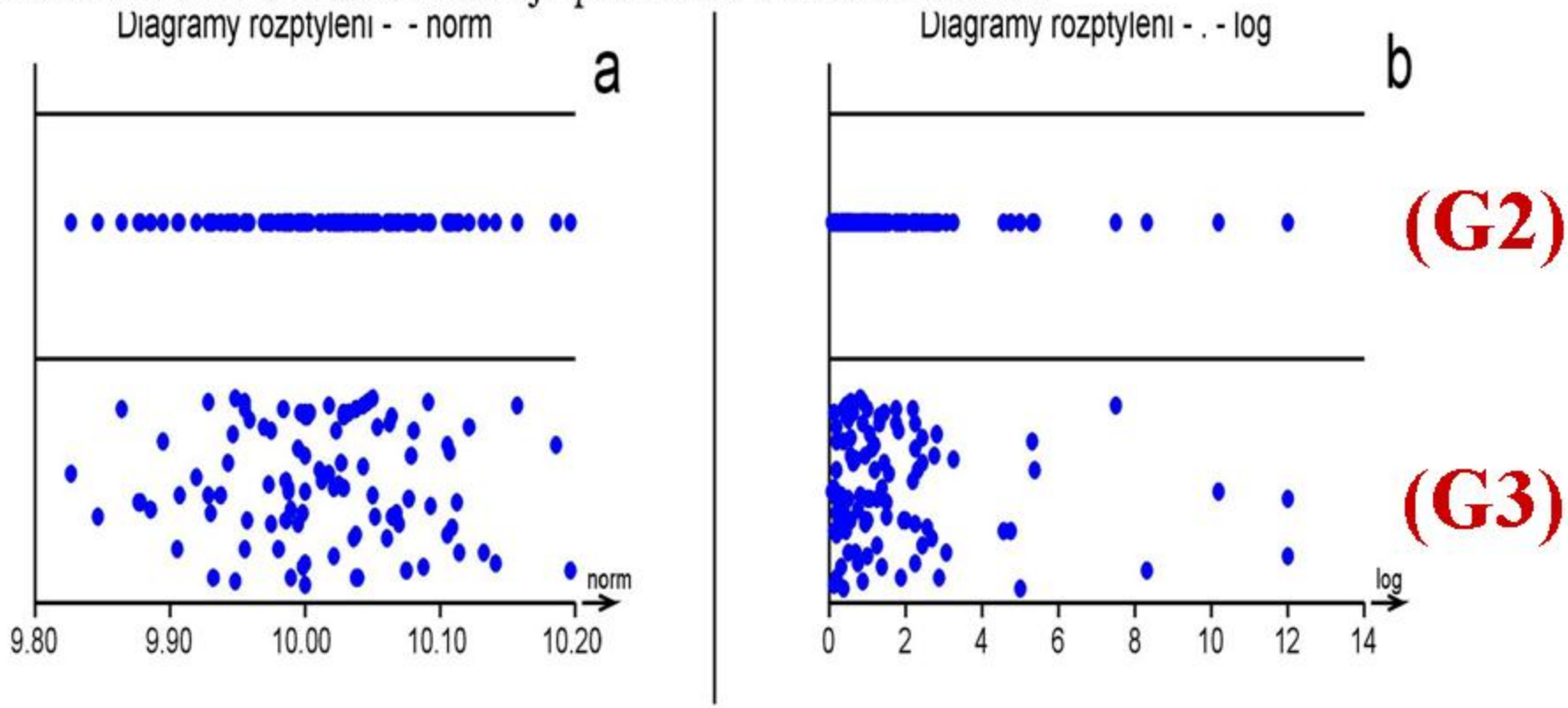
(a) přehledně znázornit data, lokální koncentraci dat,
(b) indikuje i podezřelá a vybočující měření.

Rozmítnutý diagram rozptýlení (G3)

Osa x: hodnoty x,
Osa y: libovolný interval náhodných čísel

Obsah: představuje rozmítnutou projekci kvantilového grafu.
Diagnoza: přehledně znázornit data, lokální koncentraci dat, indikuje i podezřelá a vybočující měření.

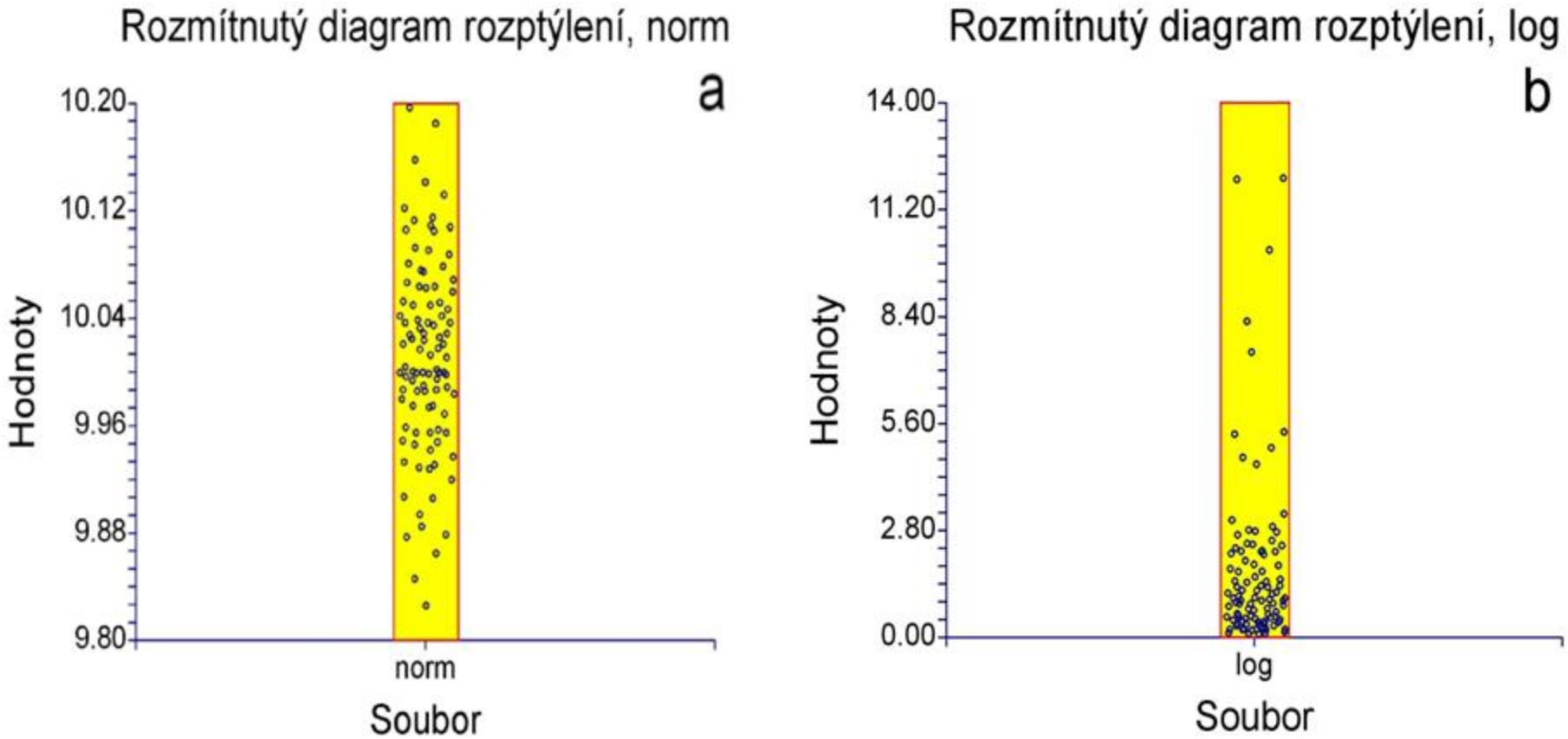
Diagram p rozptýlení (osa x: hodnoty x_i, osa y: libovolná úroveň, např. $y = 0$). Představuje jednorozměrnou rojekci kvantilového grafu do osy x. I při své jednoduchosti ukazuje na lokální koncentrace dat a indikuje podezřelá a odlehlá měření.



Obr. 2.2 Diagram rozptýlení a rozmítnutý diagram rozptýlení pro výběry (shora dolu): (a) *norm*, symetrického (normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

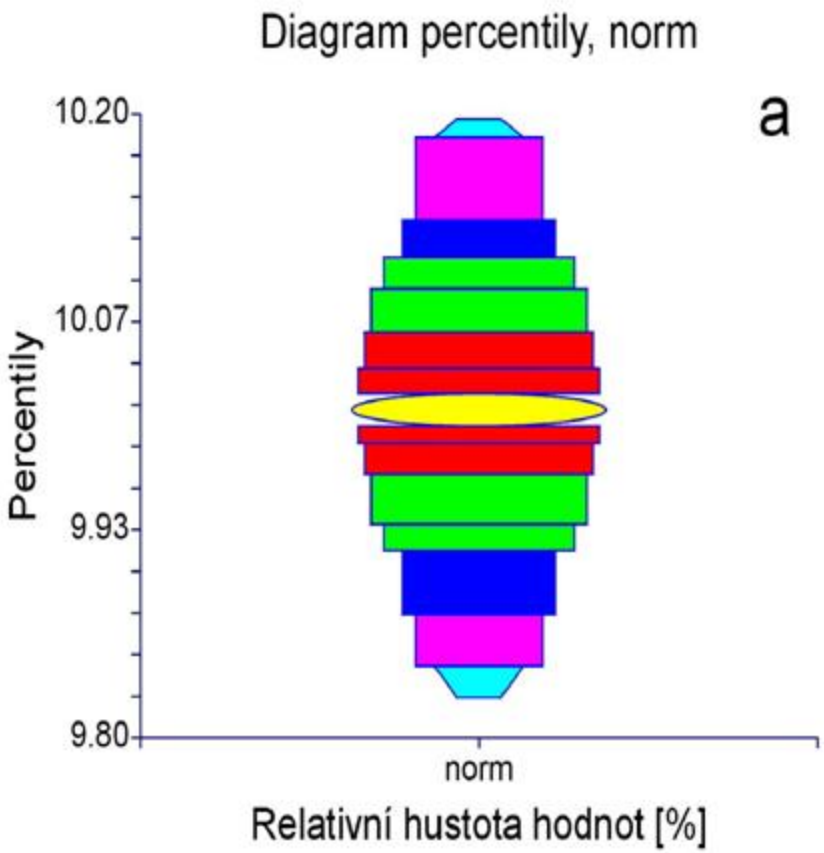
(G3)

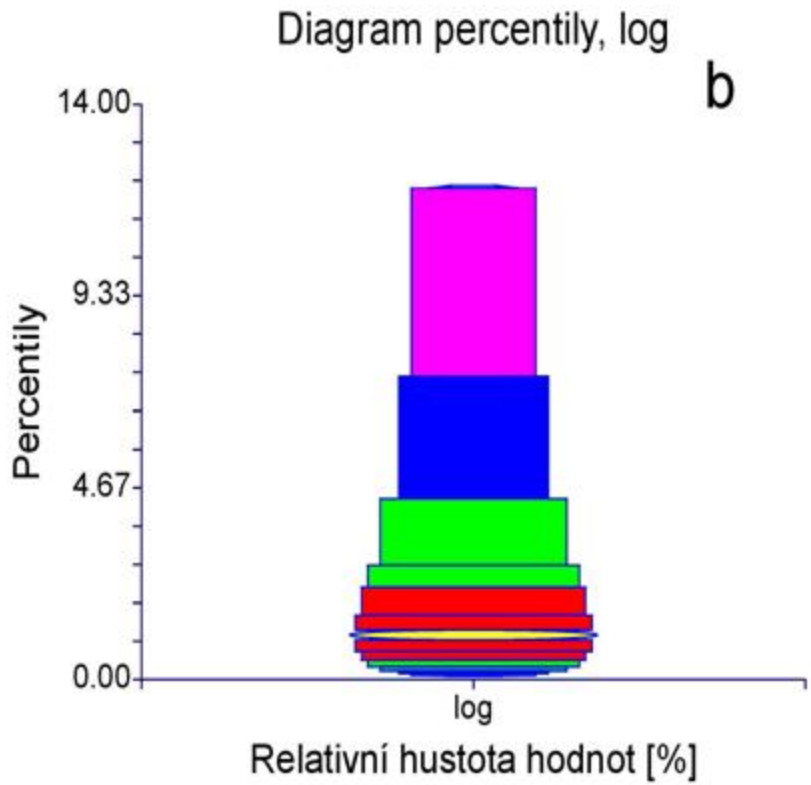
Rozmítnutý diagram rozptýlení (osa *x*: hodnoty *x*, osa *y*: interval náhodných čísel). Diagram představuje rovněž projekci kvantilového grafu, body jsou však pro lepší přehlednost vhodně rozmítnuté.



Obr. 2.3 Rozmítnutý diagram rozptýlení pro výběry: (a) *norm*, symetrického (normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000.***

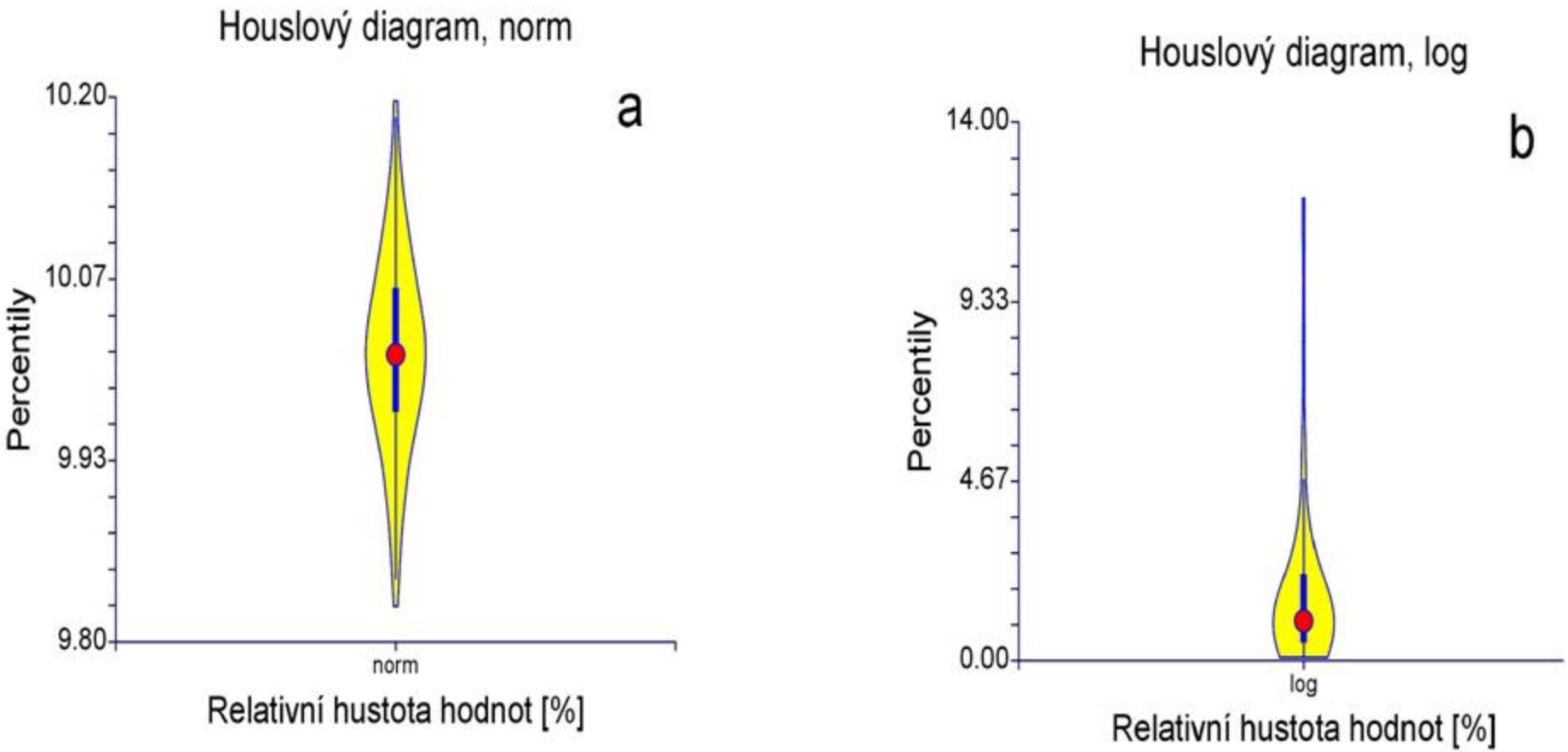
Diagram percentilů (osa *x:* proměnná, osa *y:* percentily).
Diagram zobrazuje vybrané percentily. Jsou to obvykle intervaly 0-2, 2-5, 5-10, 10-15, 15-25, 25-35, 35-45, 45-55, 55-65, 65-75, 75-85, 85-90, 90-95, 95-99, 99-100. Z výsledného obrazce lze usoudit na symetrii rozdělení nebo na jeho tvar.





Obr. 2.4 Diagram některých percentilů pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000.***

Houslový diagram (osa *x*: název výběru proměnné, osa *y*: percentily, hodnoty proměnné). Medián je zobrazen černým kolečkem a začátek a konec úsečky zobrazuje dolní a horní kvantil. Normální rozdělení se projeví v symetrickém tvaru houslí, zatímco logaritmicko-normální v silně asymetrickém tvaru.

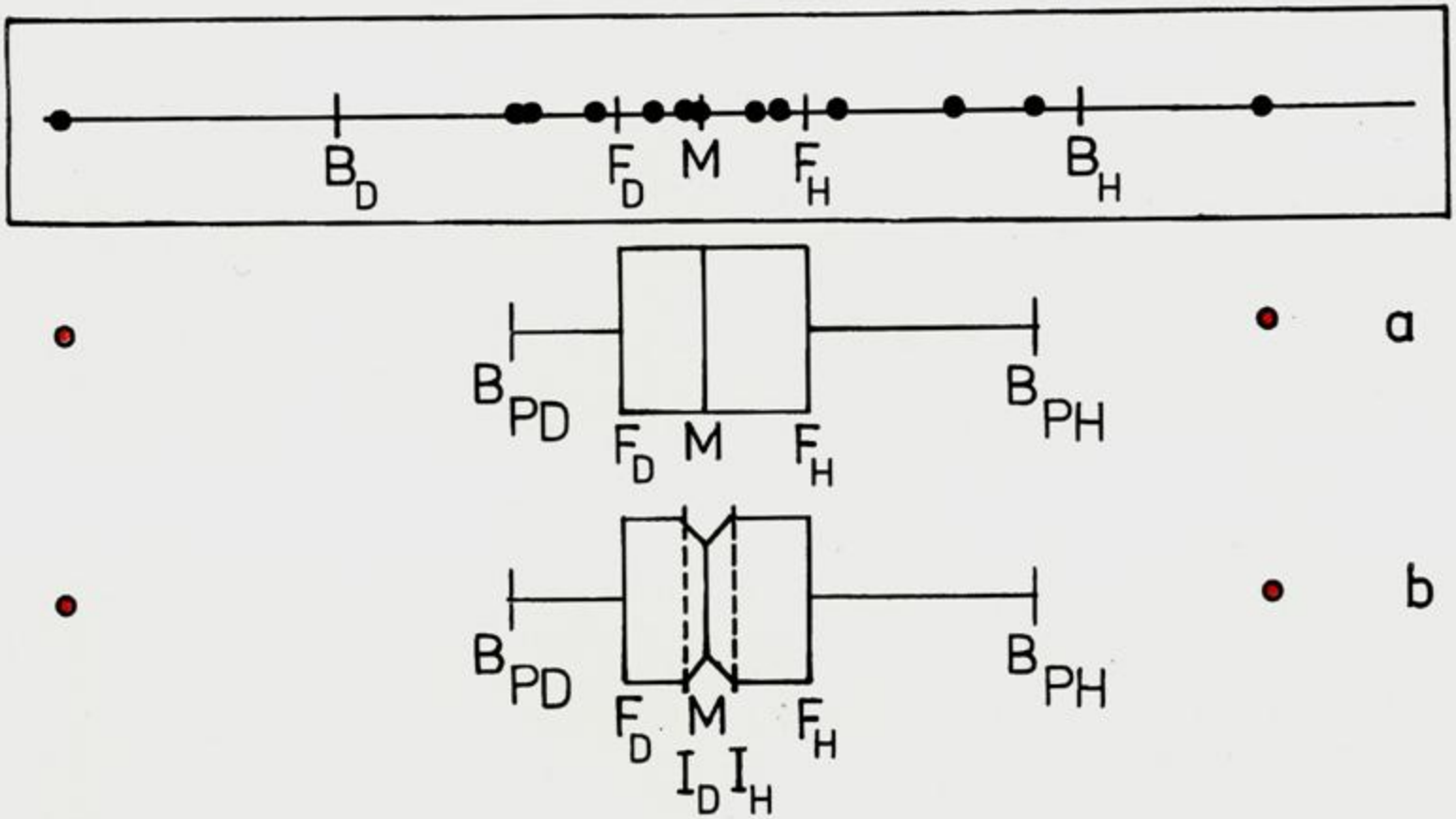


Obr. 2.5 Houslový diagram pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000.***

Krabicový graf (G4)

Osa x: úměrná hodnotám x,

Osa y: libovolný interval,



Konstrukce: a) krabicového grafu,
b) vrubového krabicového grafu

z dat diagramu rozptýlení. Prázdná kolečka indikují vybočující hodnoty.

(G4)

Diagnoza:

a) znázornění robustního odhadu polohy, mediánu M,

b) posouzení symetrie v okolí kvantilů (kvartilů F_H, F_D)

c) interkvantilové rozpětí $R_F = F_H - F_D$,

d) posouzení symetrie u konců rozdělení, u přilehlých hodnot B_{PH} a B_{PD},

d) identifikaci odlehlých dat mimo vnitřní hradby B_H a B_D

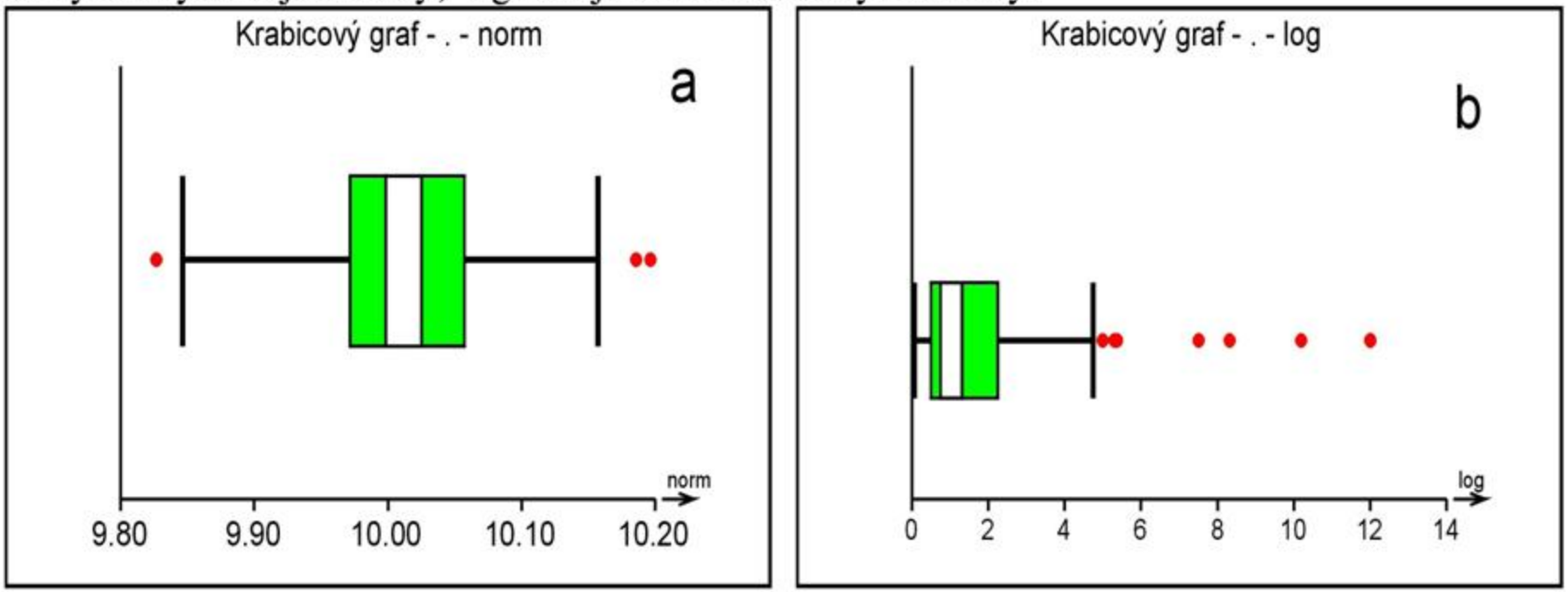
$$B_H = F_H + 1.5\ R_F$$

$$B_D = F_D - 1.5\ R_F$$

Krabicový graf (osa x: úměrná hodnotám x, osa y: interval úměrný hodnotě $\sqrt{n}$). **(G4)**
V místě mediánu M je vertikální čára. Od obou protilehlých stran tohoto obdélníku pokračují úsečky. Ty jsou ukončeny *vnitřními hradbami* B_H, B_D, pro které platí

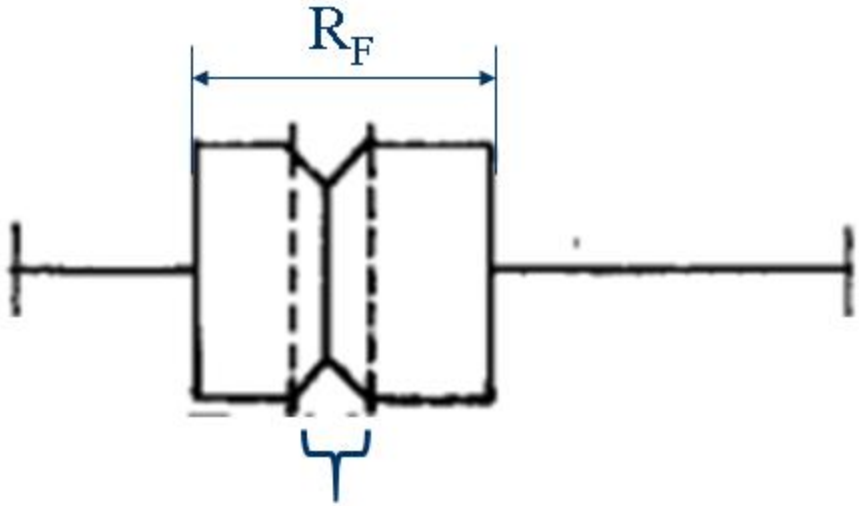
$$B_H = F_H + 1.5\, R_F \,, \quad B_D = F_D - 1.5\, R_F \,.$$

Prvky výběru, ležící mimo interval vnitřních hradeb $[B_H, B_D,]$ jsou považovány za podezřelé, obvykle vybočující body; v grafu jsou znázorněny kroužky.



Obr. 2.7 Vrubový krabicový graf pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Vrubový krabicový graf (G5)



intervalový odhad mediánu

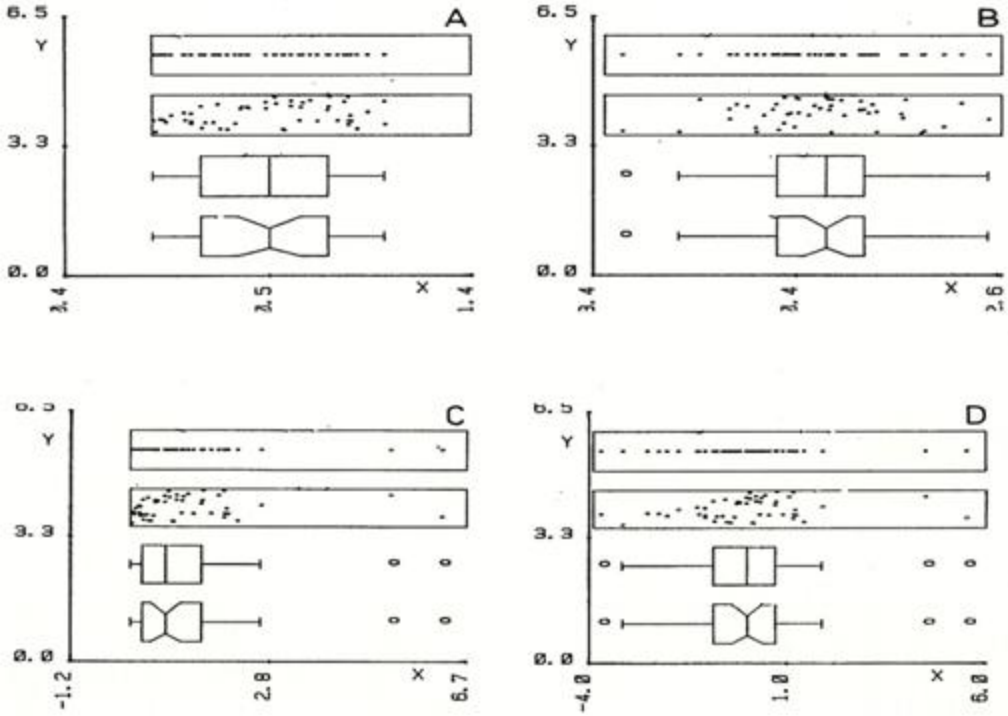
$$I_{D,H} = M \pm \frac{1{,}57 \cdot R_F}{\sqrt{n}}$$

Vrubový krabicový graf

(G5)

Osa x: úměrná hodnotám x,

Osa y: libovolný interval



Diagramy rozptýlení a krabicové grafy výběrů z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova

Diagnoza:

Robustní interval spolehlivosti $I_D \leq M \leq I_H$ s mezemi

$$I_D = M - \frac{1.57\ R_F}{\sqrt{n}}$$

$$I_H = M + \frac{1.57\ R_F}{\sqrt{n}}$$

Krabicový graf: (G4)

obsahuje medián, kvartily, přilehlé hodnoty a vnitřní hradby

Vrubový krabicový graf: (G5)

obsahuje navíc i interval spolehlivosti mediánu

Diagnoza:

(1) Robustní odhad polohy (medián)

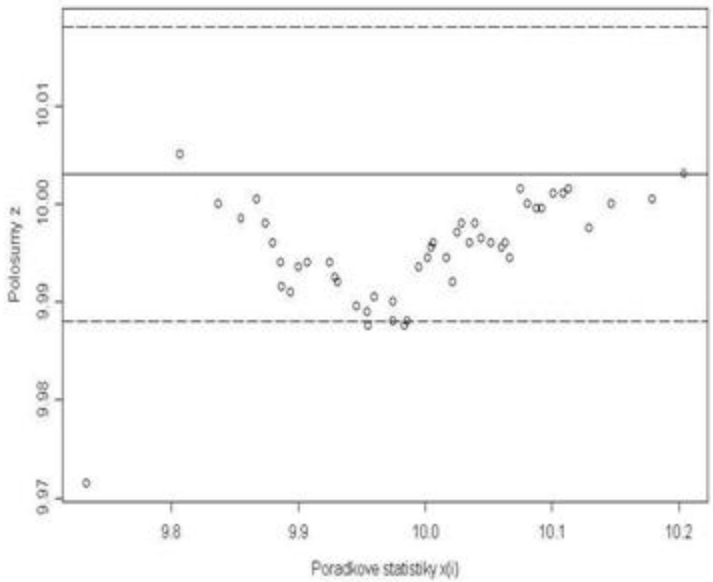
(2) Interval spolehlivosti mediánu

(3) Posouzení symetrie kvantilů

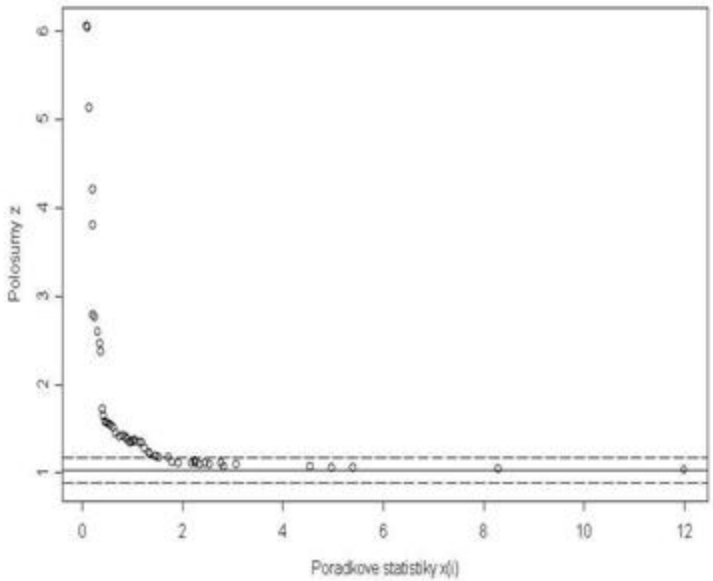
(4) Posouzení symetrie v okolí konců rozdělení

(5) Identifikace odlehlých bodů, podezřelých bodů

Graf polosum



Gra polosum Gaussova normálního rozdělení ***norm*** N(10, 0.1)

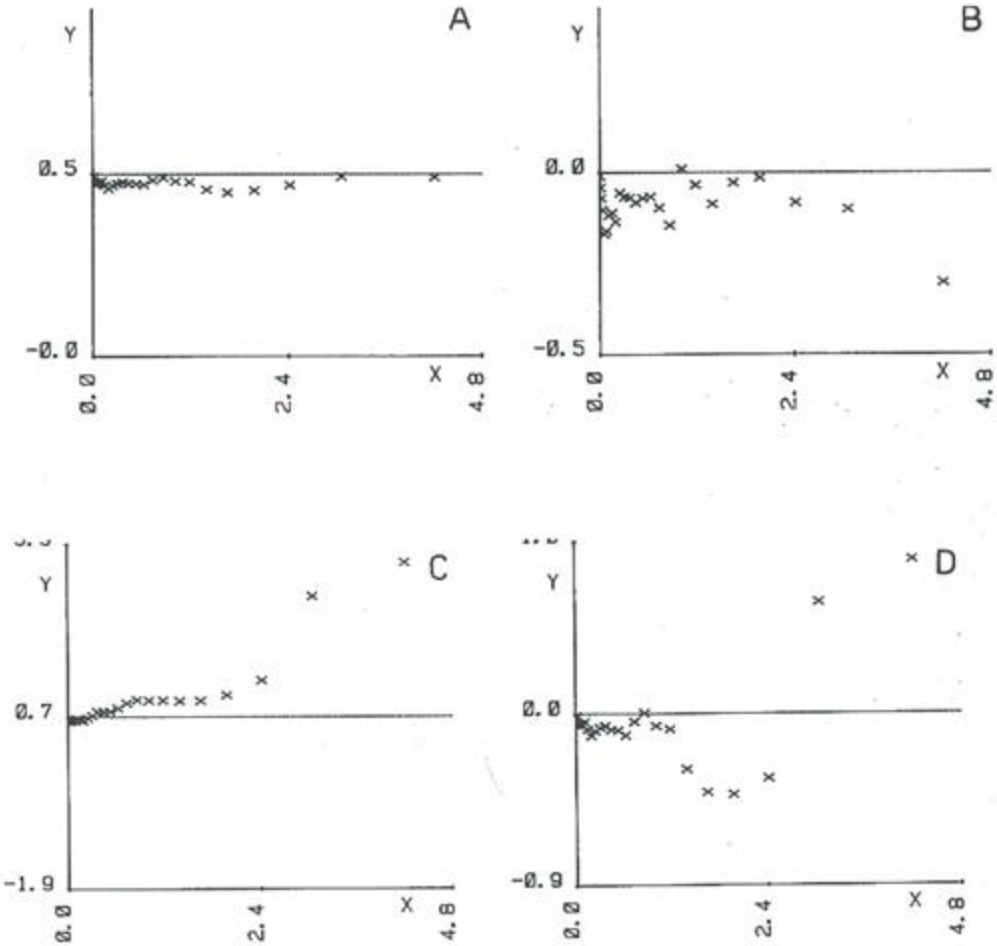


Graf polosum logaritmicko-normálního rozdělení ***log*** LN(5, 2)

Graf polosum (G6)

Osa x: pořádkové statistiky $x_{(i)}$,

Osa y: $Z_i = 0.5\,(x_{(n+1-i)} + x_{(i)})$



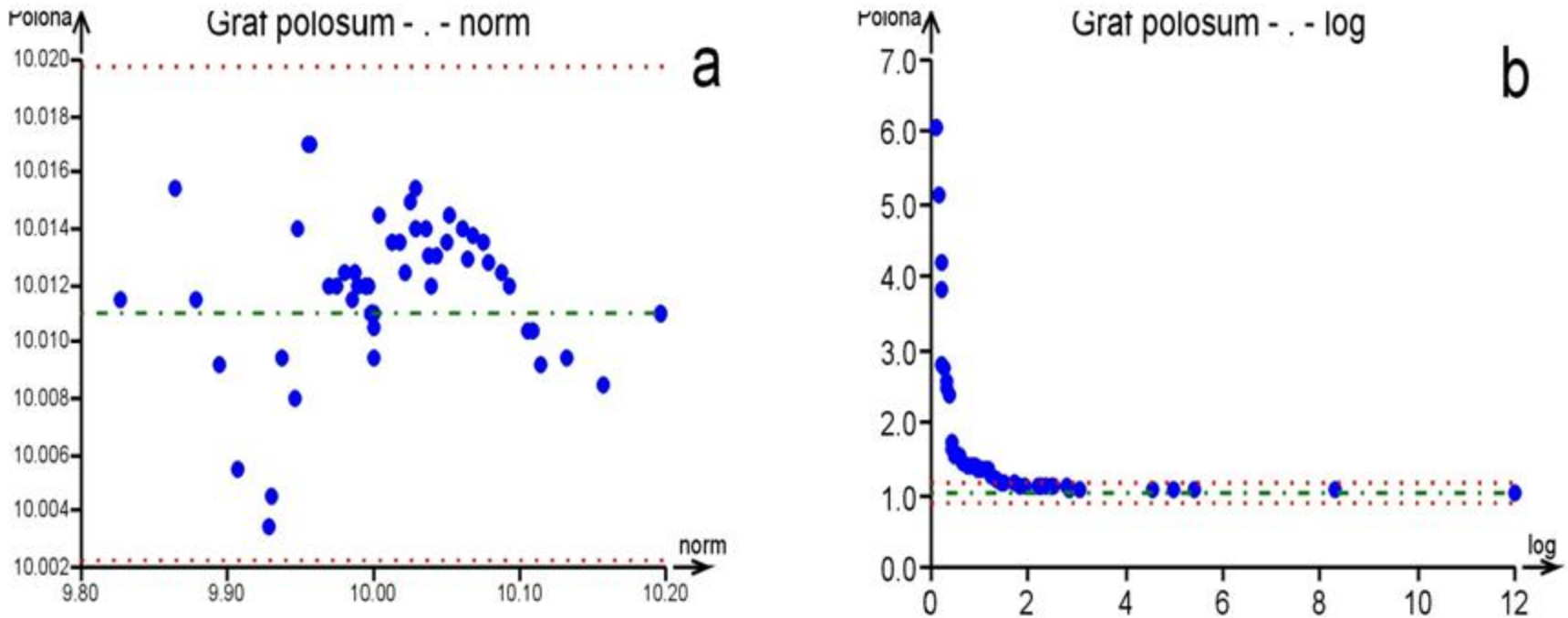
Grafy symetrie pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova, (---- značí symetrii)

Diagnoza:

Symetrická rozdělení: grafem horizontální přímka $\tilde{x}_{0.5} = M$.

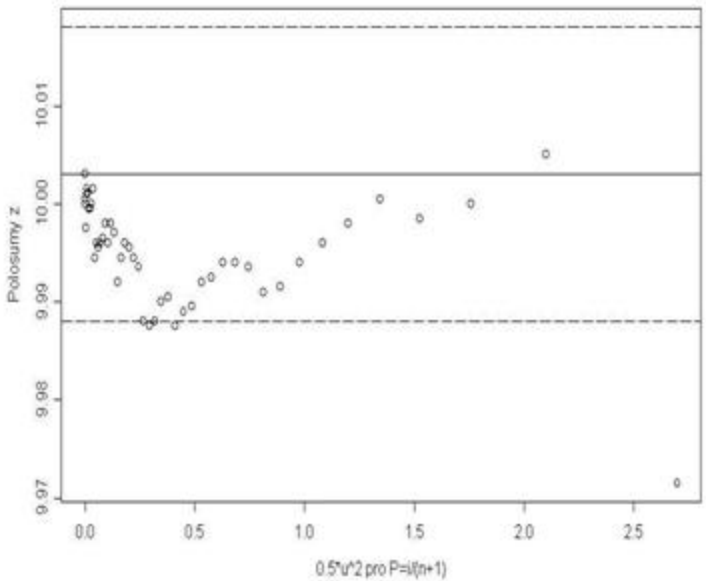
(G6)

Graf polosum (osa x: pořádkové statistiky $x_{(i)}$, osa y: $Z_i = 0.5\,(x_{(n+1-i)} + x_{(i)})$). Pro symetrické rozdělení je grafem polosum horizontální přímka, určená rovnicí $y = M$. Asymetrické rozdělení vykazuje nenáhodný trend a body pak neoscilují okolo horizontální přímky a měřítko osy y není detailní.

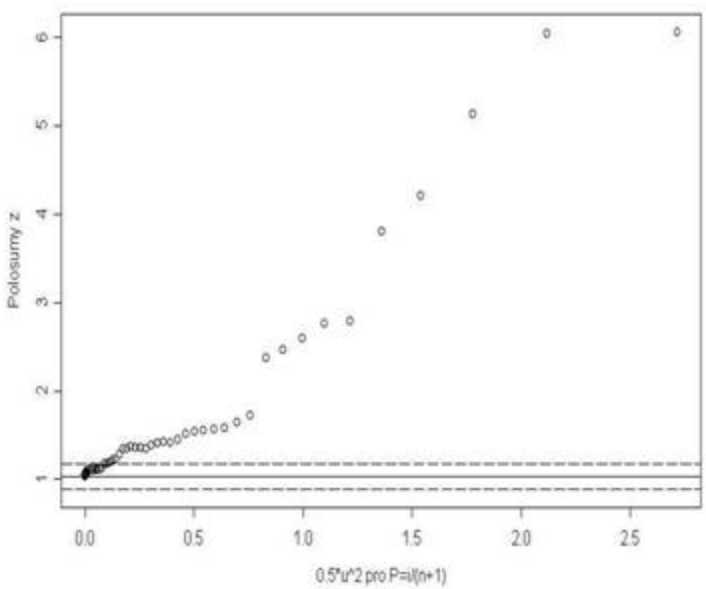


Obr. 2.8 Graf polosum pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Graf symetrie



Graf symetrie Gaussova normálního rozdělení ***norm*** N(10, 0.1)

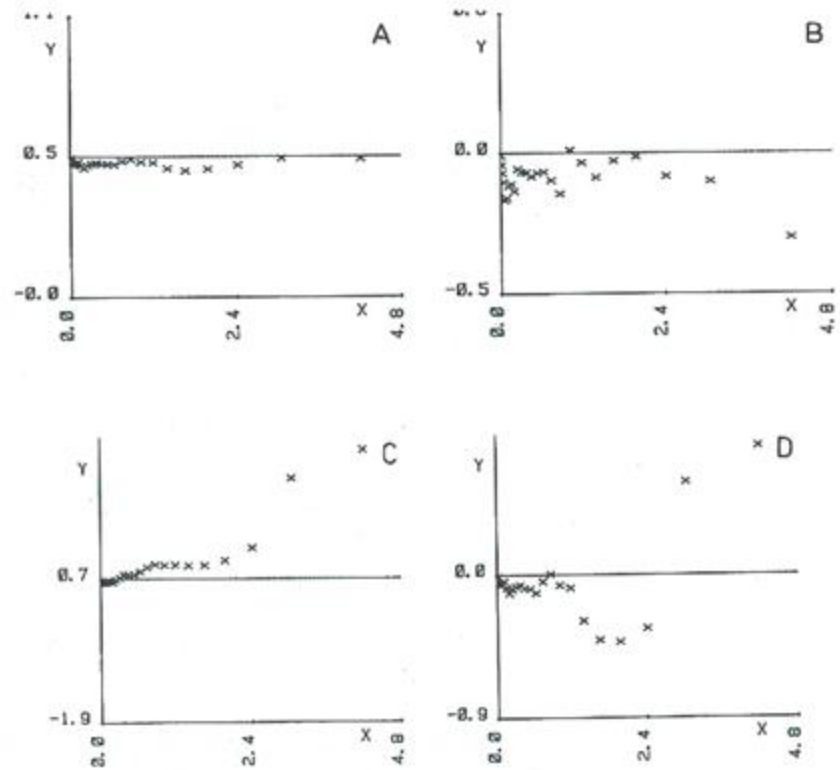


Graf symetrie logaritmicko-normálního rozdělení ***log*** LN(5, 2)

Graf symetrie (G7)

Osa x: $\frac{u_{P_i}^2}{2}$ pro $P_i = \frac{i}{n+1}$

Osa y: $Z_i = 0.5\,(x_{(n+1-i)} + x_{(i)})$



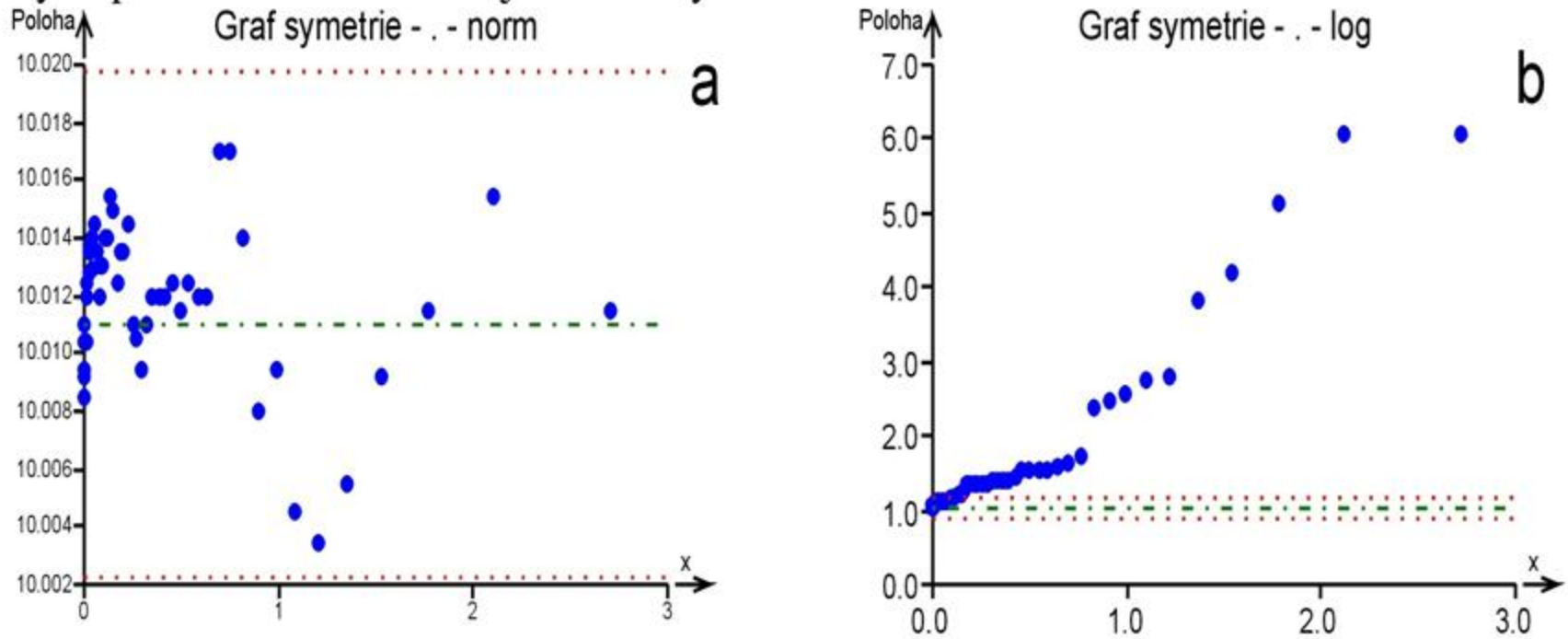
Grafy symetrie pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova, (---- značí symetrii)

Diagnoza:

Symetrická rozdělení jsou charakterizována horizontální přímkou $y = \tilde{x}_{0.5} = M$. Směrnice je odhadem šikmosti.

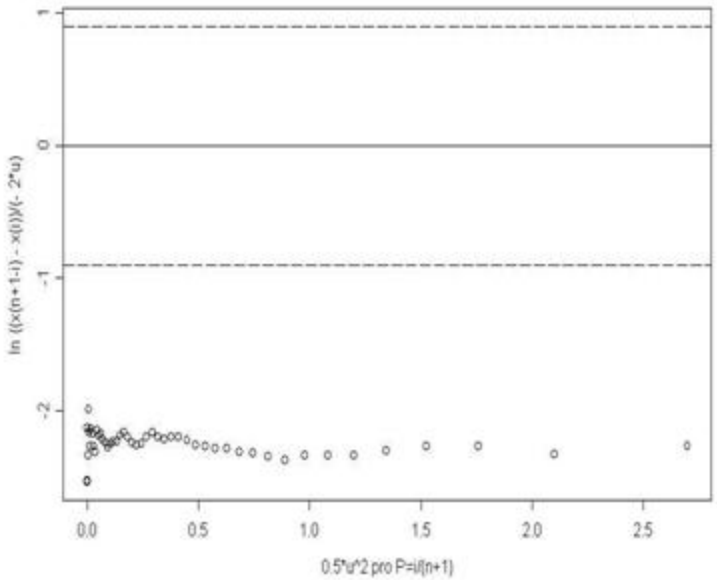
(G7)

Graf symetrie (osa x: M - $x_{(i)}$, osa y: $x_{(n+1-i)}$ - M). Symetrická rozdělení jsou charakterizována přímkou $y = M$. Pro asymetrické rozdělení tato přímka nemá nulovou směrnici a v tomto grafu je směrnice odhadem parametru šikmosti. Asymetrické rozdělení vykazuje body uspořádané v trendu nějaké křivky.

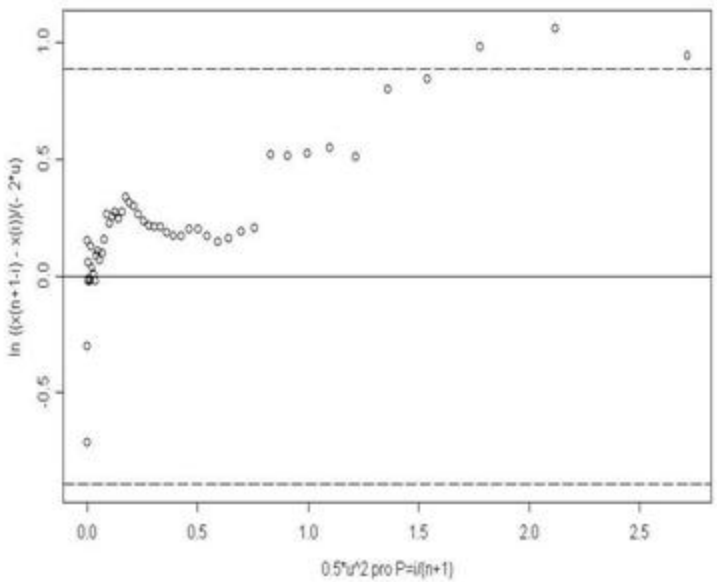


Obr. 2.9 Graf symetrie pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Graf špičatosti



Graf špičatosti Gaussova normálního rozdělení ***norm*** N(10, 0.1)

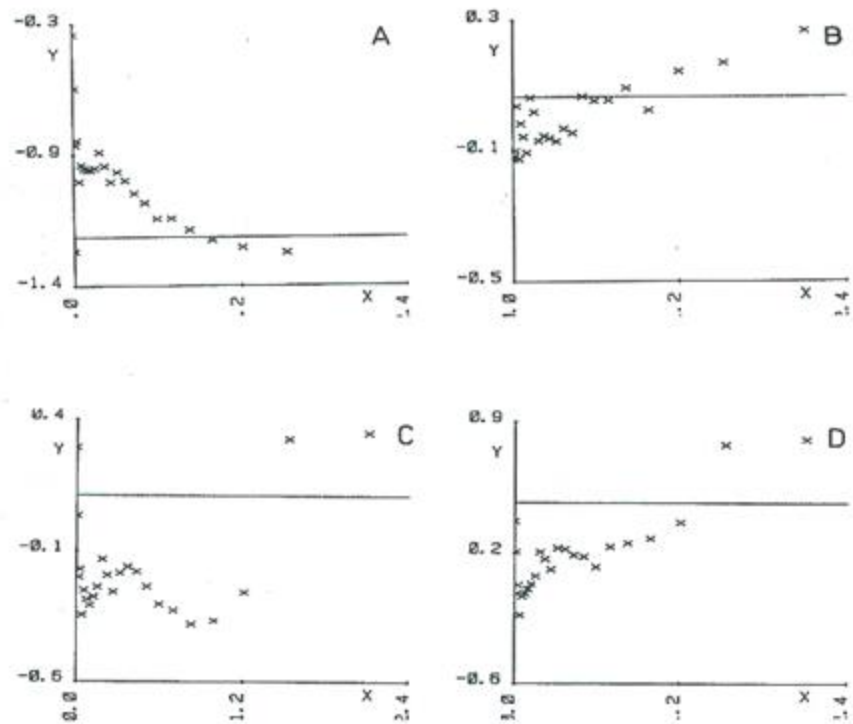


Graf špičatosti logaritmicko-normálního rozdělení ***log*** LN(5, 2)

Graf špičatosti (G8)

$$\text{Osa x}: \frac{u_{P_i}^2}{2} \quad \text{pro } P_i \approx \frac{i}{n+1},$$

$$\text{Osa y}: \ln \frac{x_{(n+1-i)} - x_{(i)}}{-2\,u_{P_i}}$$



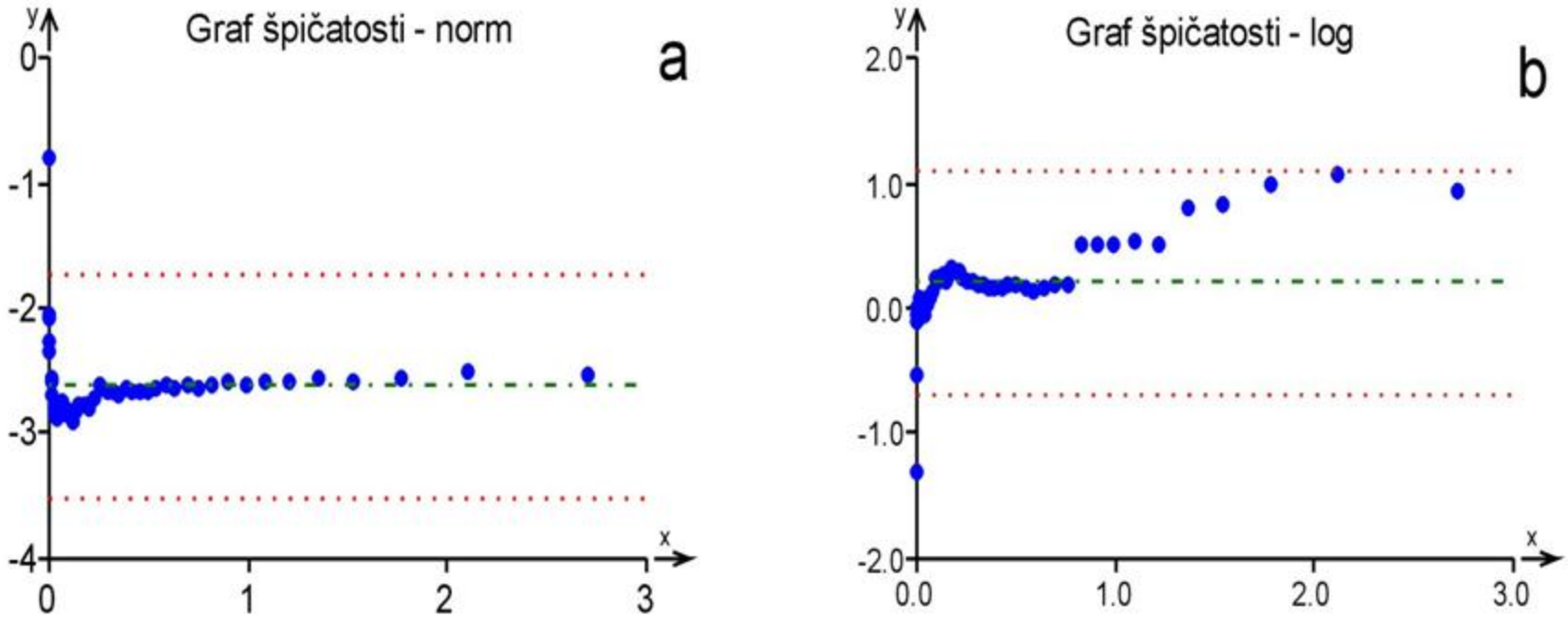
Grafy špičatosti pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova, (_____ značí normalitu)

Diagnoza:

Symetrická normální rozdělení mají grafem horizontální přímku. Směrnice odhadem parametru špičatosti.

Graf špičatosti (osa x: $u_{P_i}^2/2$ pro $P_i = i/(n+1)$, osa y: $\ln(x_{(n+1-i)}/-2u_{P_i})$. **(G8)**

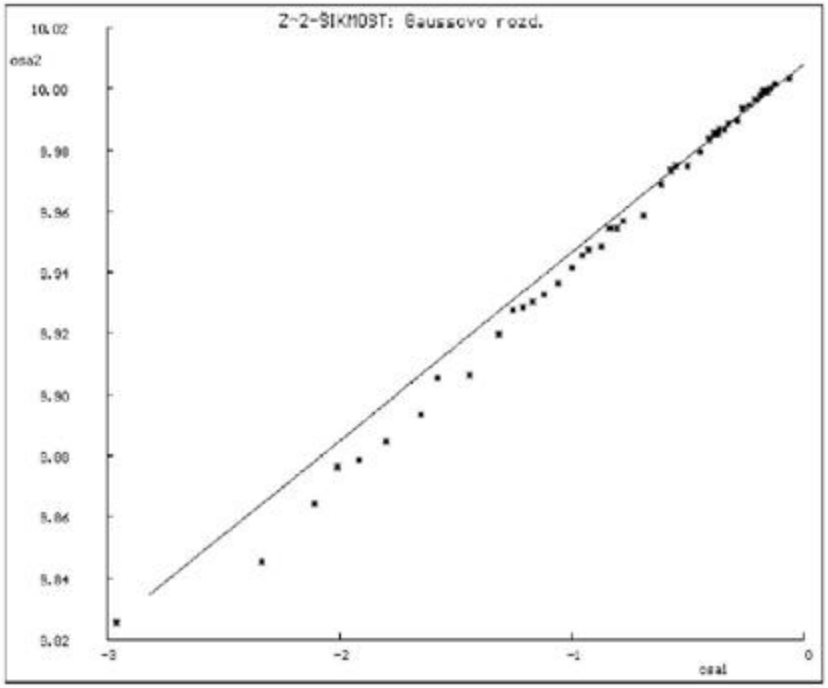
Pro normální rozdělení je grafem horizontální přímka a body leží převážně na této přímce. Pokud body tvoří nenáhodný trend, odpovídá hodnota směrnice této aktuální přímky parametru špičatosti.

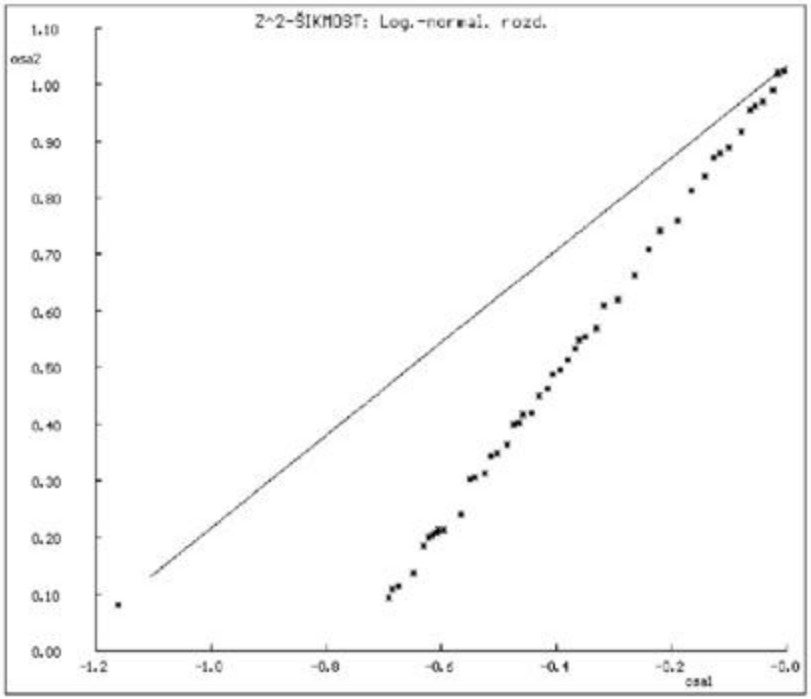


Obr. 2.11 Graf špičatosti pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Graf šikmosti (osa x: $u_{P_i}^2/2$ pro $P_i = i/(n+1)$, osa y: $Z_i = 0.5\ (x_{(n+1-i)} + x_{(i)})$).

Pro případ symetrického rozdělení rezultuje u grafu šikmosti přímková závislost s nulovým úsekem a jednotkovou směrnicí. Body leží těsně na této přímce. U asymetrického rozdělení body neleží na této přímce a vykazují jinou směrnici.



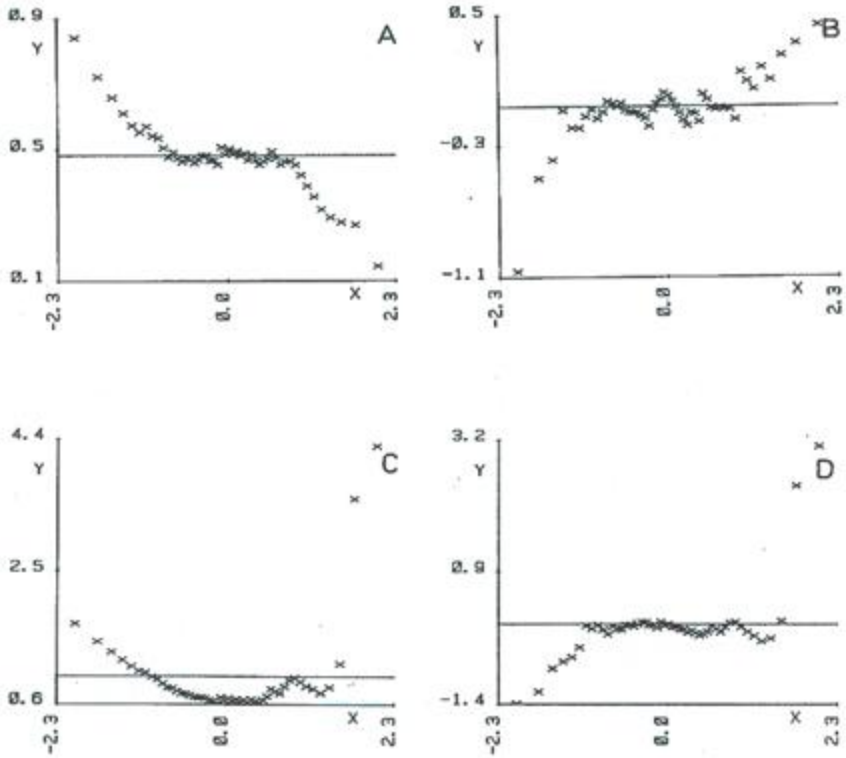


Obr. 2.10 Graf šikmosti pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Diferenční kvantilový graf (G9)

Osa x: kvantil u_{P_i},

Osa y: $d_{(i)} = x_{(i)} - \tilde{s}\ u_{P_i}$



Diferenční kvantilové grafy pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova, (____ značí normalitu)

Diagnoza:

(a) posouzení rozdělení se špičatostí normálního rozdělení,

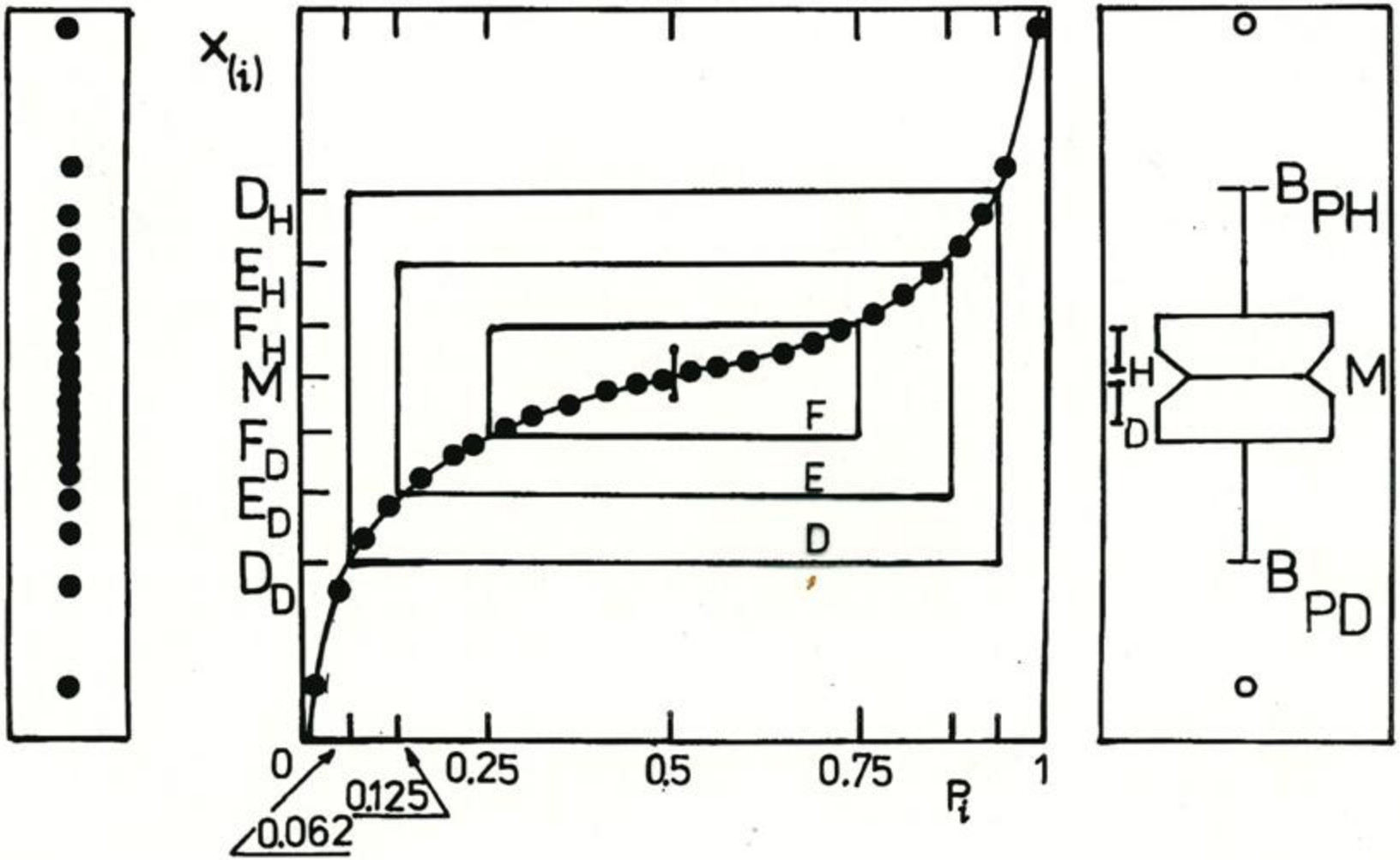
(b) $\tilde{s}$ představuje robustní odhad směrodatné odchylky,

(c) horizontální přímka indikuje symetrické rozdělení.

Graf rozptýlení s kvantily (G10)

Osa x: P_i,

Osa y: $x_{(i)}$



Diagnoza: **(G10)**

(a) pro symetrická rozdělení má kvantilová funkce sigmoidální tvar,

(b) pro rozdělení sešikmená k vyšším hodn. je konvexně rostoucí,

(c) pro rozdělení sešikmená k nižším hodn. konkávně rostoucí,

(d) zakreslují se tři obdélníky F, E a D:

(1) ***Kvartilový obdélník:*** má na ose y kvartily F_D a F_H, na ose x jsou $P_2 = 2^{-2} = 0.25$ a $1 - 2^{-2} = 0.75$.

(2) ***Oktilový obdélník E***: má na ose y oktily E_D a E_H, na ose x jsou $P_3 = 2^{-3} = 0.125$ a $1 - 2^{-3} = 0.875$.

(3) ***Sedecilový obdélník D***: má na ose y sedecily D_D a D_H, na ose x jsou $P_4 = 2^{-4} = 0.0625$ a $1 - 2^{-4} = 0.9375$.

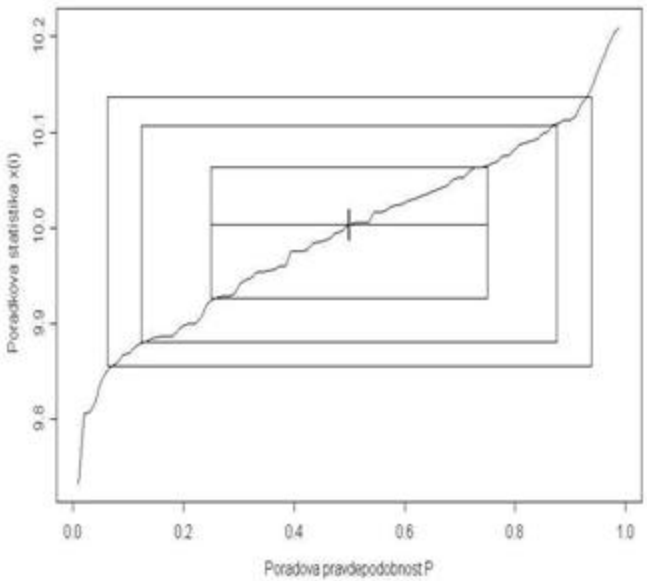
(4) Medián $\tilde{x}_{0.5}$ tvoří úsečku a robustní odhad konfidenčního intervalu mediánu $\tilde{x}_{0.5} \pm 1.57 \; R_F / \sqrt{n}$.

Identifikace řady zvláštností výběru: **(G10)**

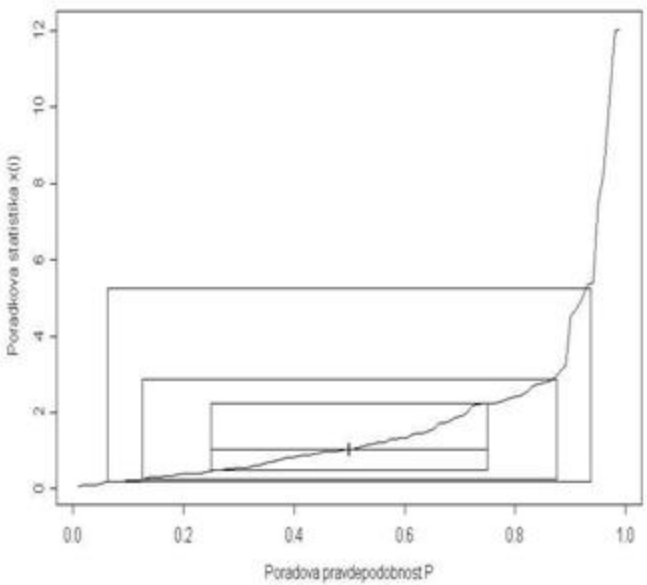
1. ***Symetrické unimodální rozdělení výběru***: obdélníky symetricky uvnitř sebe.
2. ***Nesymetrická rozdělení***: pro sešikmené rozdělení rozdílné vzdálenosti mezi dolními a horními hranami obdélníků.
3. ***Odlehlá pozorování***: náhlý vzrůst na křivce, kdy směrnice roste nade všechny meze. Body vně sedecilového obdélníku.

Graf rozptýlení s kvantily

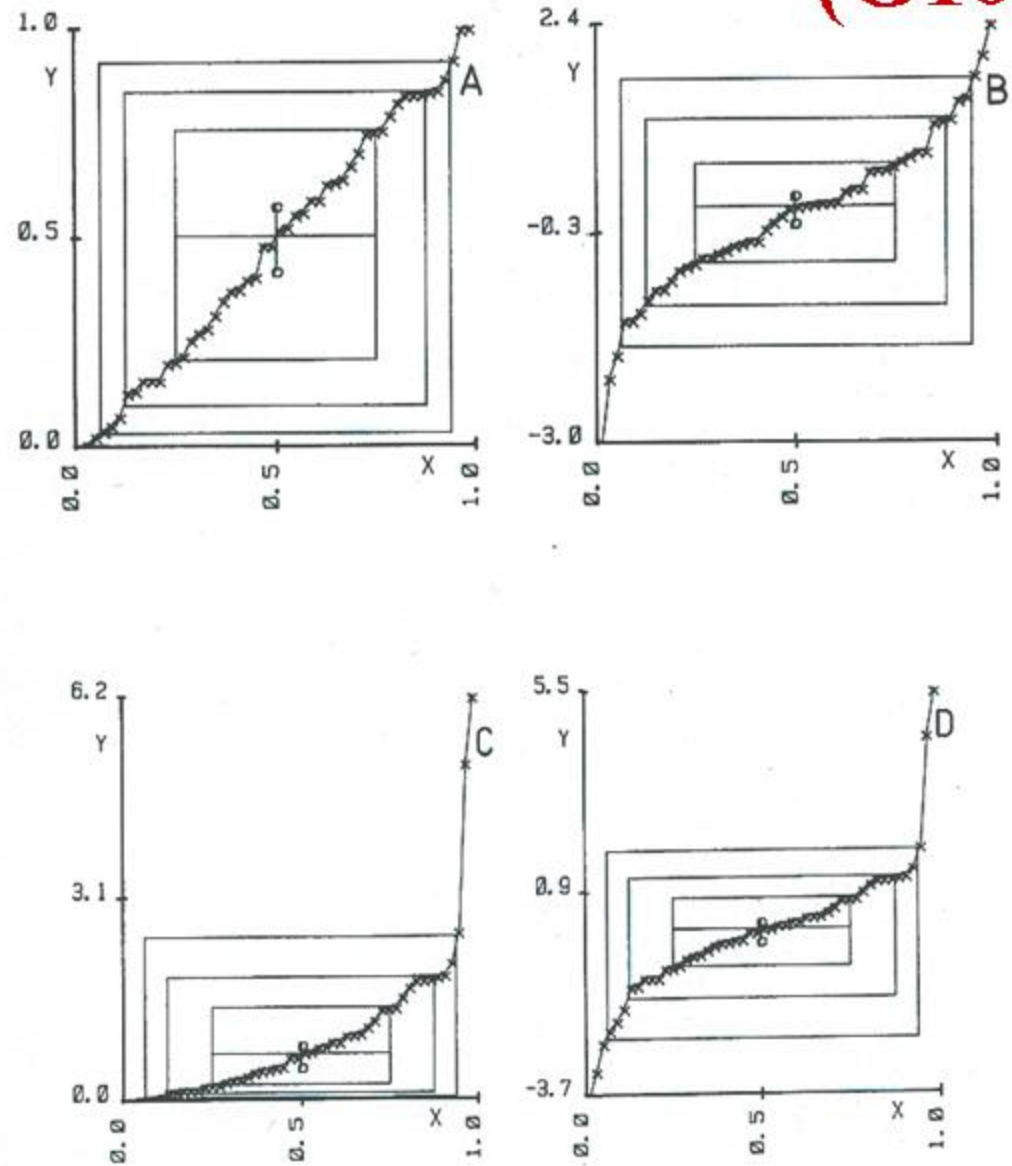
(G10)



Graf rozptýlení s kvantily Gaussova normálního rozdělení ***norm*** N(10, 0.1)



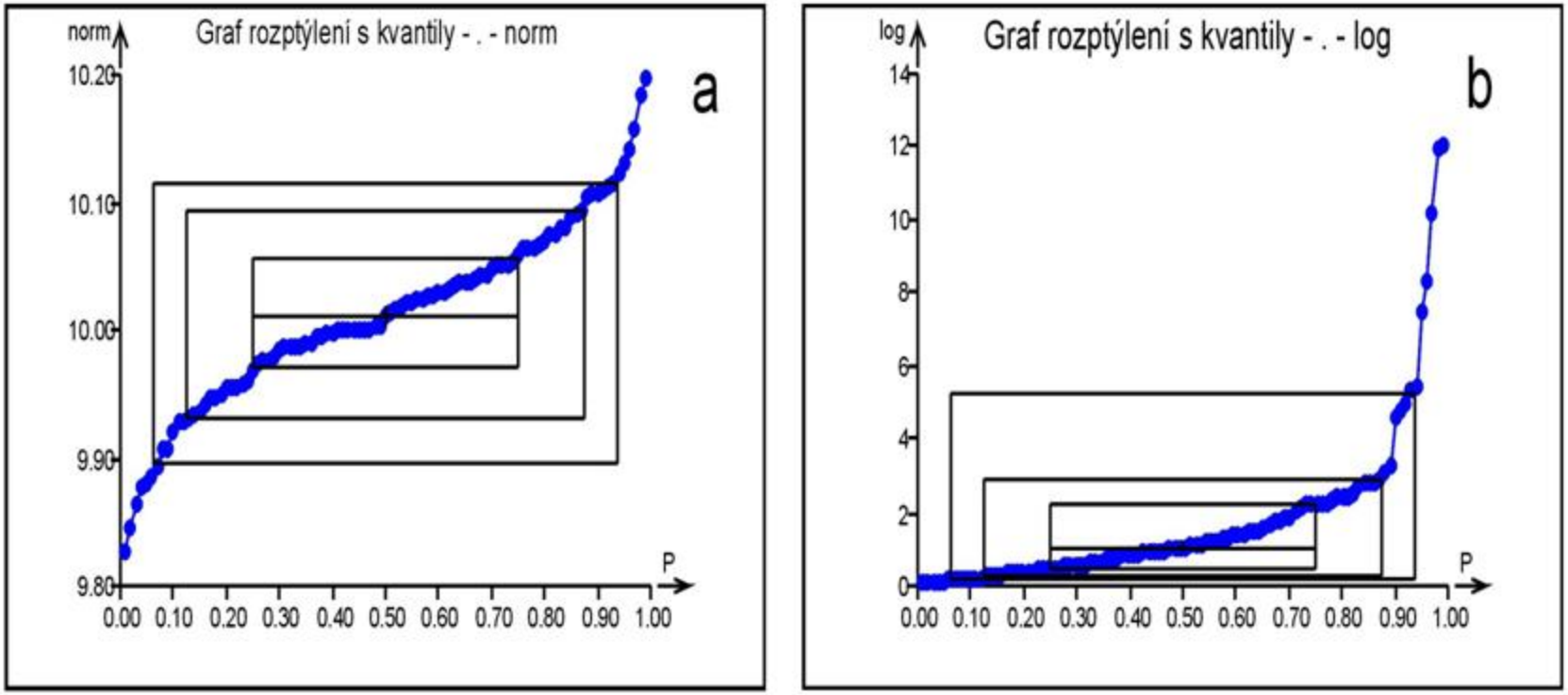
Graf rozptýlení s kvantily logaritmicko-normálního rozdělení ***log*** LN(5, 2)



Grafy rozptýlení s kvantily pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova

(G10)

Graf rozptýlení s kvantily (osa x: P_i, osa y: $x_{(i)}$). Pro symetrická rozdělení má kvantilová funkce sigmoidální tvar. Pro rozdělení zešikmená k vyšším hodnotám je konvexně rostoucí a pro rozdělení zešikmená k nižším hodnotám konkávně rostoucí.



Obr. 2.12 Graf rozptýlení s kvantily pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Charakteristiky šikmosti a špičatosti

Charakteristiky šikmosti a špičatosti

Název	Definice	Charakterizuje	Platí pro L
Polosuma Z_L	$0.5\ (L_D + L_H)$	symetrii při $Z_L = 0$	F, E, D, ...
Rozpětí R_L	$(L_H - L_D)$	rozptýlení	F, E, D, ...
Šikmost S_L	$(M - Z_L) / R_L$	symetrii při $S_L = 0$	F, E, D, ...
Pseudosigma*) G_L	$R_L / (-2\ u_{P_i})$	špičatost (Gaussovo G_L = konst.)	F, E, D, ...
Délky konců T_L	$\ln (R_L / R_F)$	špičatost	E, D

*) u_{P_i} je kvantil standardizovaného normálního rozdělení pro $P_i = 2^{-i}$.

Délka konců (T_E, T_D):

- normální rozdělení (0.534; 0.822),
- rovnoměrné rozdělení (0.405; 0.559)
- Laplaceovo rozdělení (0.693; 1.098).

Šikmost S_L:

- rozdělení sešikmená k vyšším hodnotám S_L záporné,
- rozdělení sešikmená k nižším hodnotám S_L kladné.

Pseudosigma G_L:

- pro rozdělení s delšími konci než normální hodnota G_L roste.

Identifikace rozdělení výběru

Posouzení zvláštností rozdělení výběru:

(1) Jádrový odhad hustoty pravděpodobnosti

(2) Histogram (resp. polygon, rootogram, atd.)

Porovnání rozdělení výběru s teoretickým rozdělením:

(1) Kvantil-kvantilový (Q-Q) graf

(2) Rankitový graf

(3) Podmíněný rankitový graf

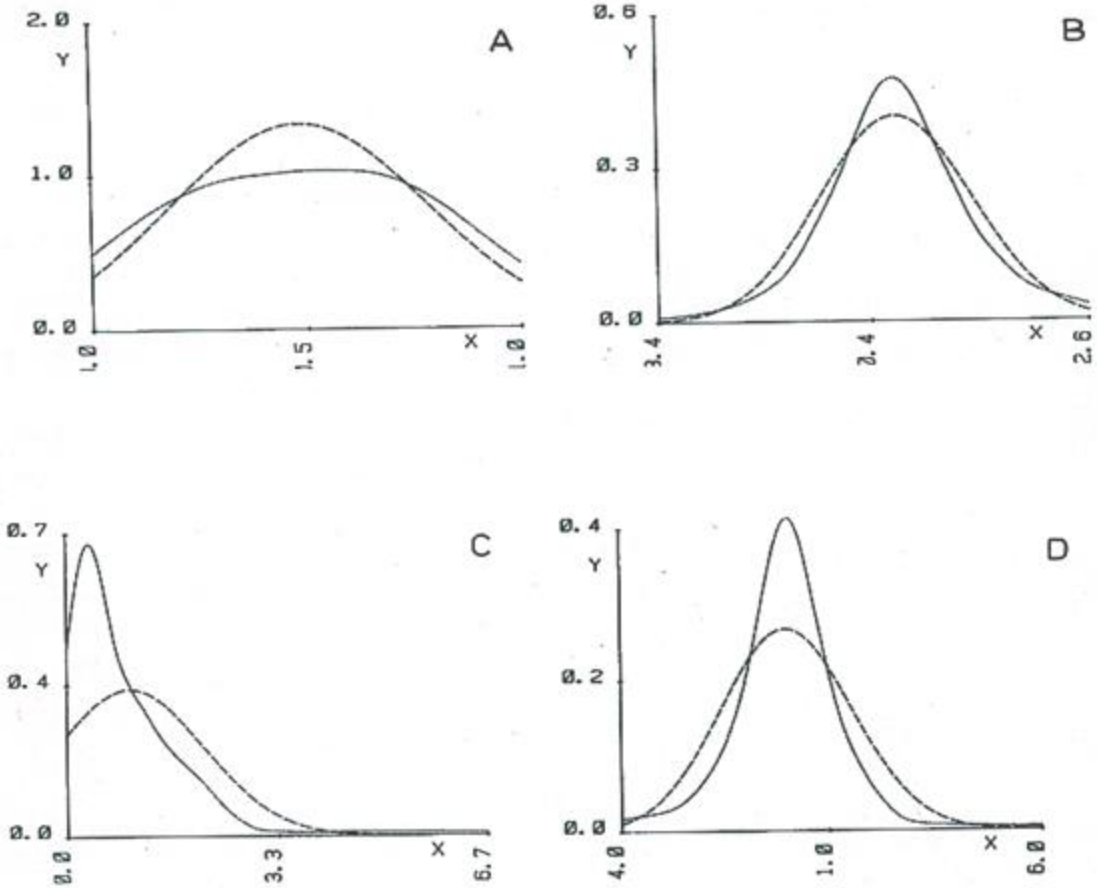
(4) Pravděpodobnostní graf

(5) Kruhový graf

Jádrový odhad hustoty pravděpodobnosti (G11)

Osa x: proměnná x,

Osa y: odhad hustoty pravděpodobnosti $\hat{f}(x)$



Jádrové odhady hustoty pravděpodobnosti pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova.

Čárkovaně je znázorněna hustota Gaussova rozdělení s parametry $\bar{x}$ a s^2 a plnou čarou jádrový odhad hustoty pravděpodobnosti empirického rozdělení výběru

Pro malé výběry se konstruují jádrové odhady hustoty dle **(G11)**

$$\hat{f}(x) = \frac{1}{n\,h} \sum_{i=1}^{n} K\left[\frac{(x - x_j)}{h}\right]$$

kde šířka pásu h určuje stupeň vyhlazení.
Jádrová funkce K(x) je symetrická kolem nuly.

Bikvadratické jádro se vyčíslí dle

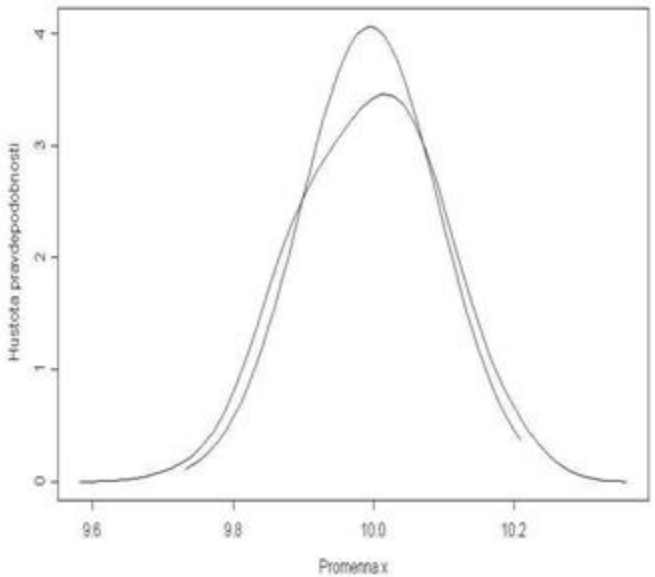
$$K(x) = 0.9375\,(1 - x^2)^2 \qquad \text{pro} \quad -1 \le x \le 1$$

$$K(x) = 0 \qquad \text{pro x mimo interval} \quad -1 \le x \le$$

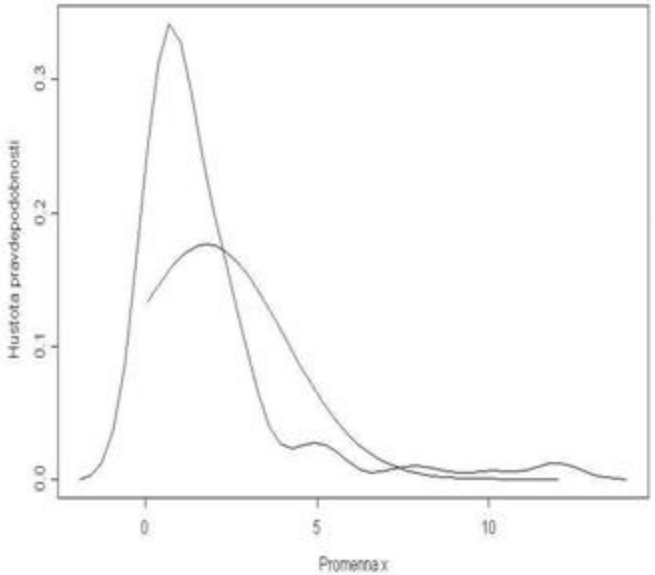
O kvalitě odhadu $\hat{f}(x)$ rozhoduje volba parametru h

$$h_{opt} = 2.34\,\sigma\,n^{-0.2}$$

Graf hustoty pravděpodobnosti



Graf hustoty pravděpodobnosti Gaussova normálního rozdělení ***norm*** N(10, 0.1)



Graf hustoty pravděpodobnosti logaritmicko-normálního rozdělení ***log*** LN(5, 2)

Postup:

(G11)

1. Zkonstruuje se předběžný odhad $\hat{f}_0(x)$ se šířkou pásu

$$h_0 = 0.75 \left(\frac{n}{100}\right)^{-0.2} \min_{(i)} \left[\left(x_{(i + \mathrm{int}(n/2))} - x_{(i)} \right) \right]$$

2. Sestrojí se vlastní odhad hustoty pravděpodobnosti $\hat{f}_K(x)$ s využitím jádrové funkce ale s nekonstantní šíří pásu

$$\hat{f}_K(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left[\frac{x - x_i}{h_i}\right]$$

Pro lokální šíři pásu platí

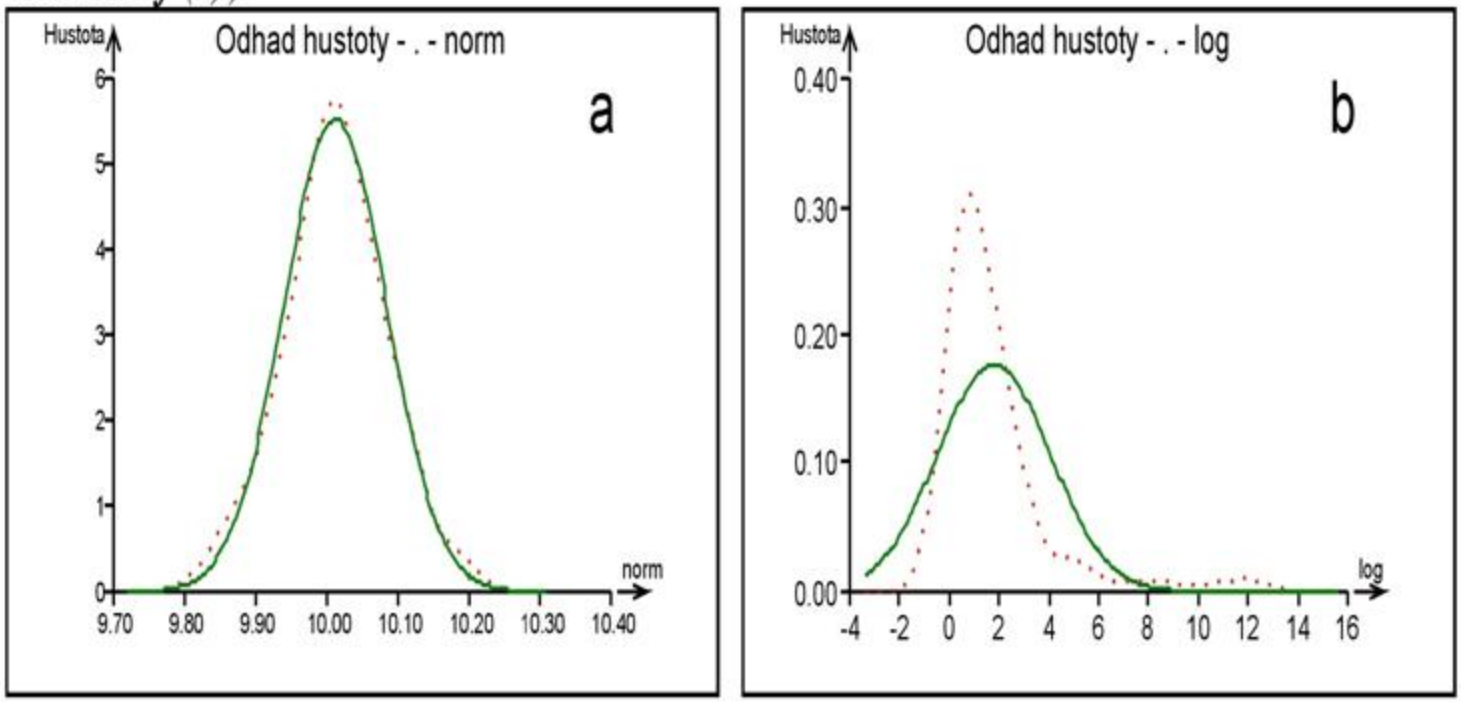
$$h_i = h_0 \left[\frac{\hat{f}_0(x_i)}{\max_{(i)} \hat{f}_0(x_i)} \right]^{-\alpha}$$

Parametr α <0, 1> ovlivňuje hladkost odhadu $\hat{f}(x)$

Pravidlo: čím je α vyšší, tím je $\hat{f}_K(x)$ hladší ale více zkreslené.

(G11)

Jádrový odhad hustoty pravděpodobnosti (osa *x*: proměnná *x*, osa *y*: hustota pravděpodobnosti $\hat{f}(x)$).

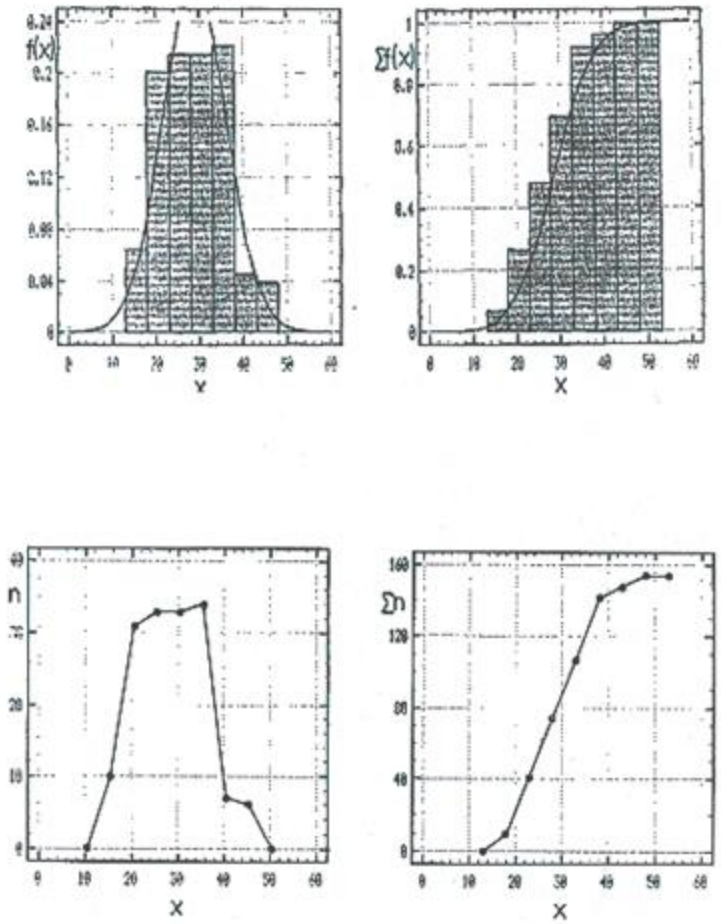


Obr. 2.14 Jádrový odhad hustoty pravděpodobnosti pro výběry. Empirická křivka rozdělení (čárkovaně) a aproximační křivka Gaussova rozdělení (plná čára): (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

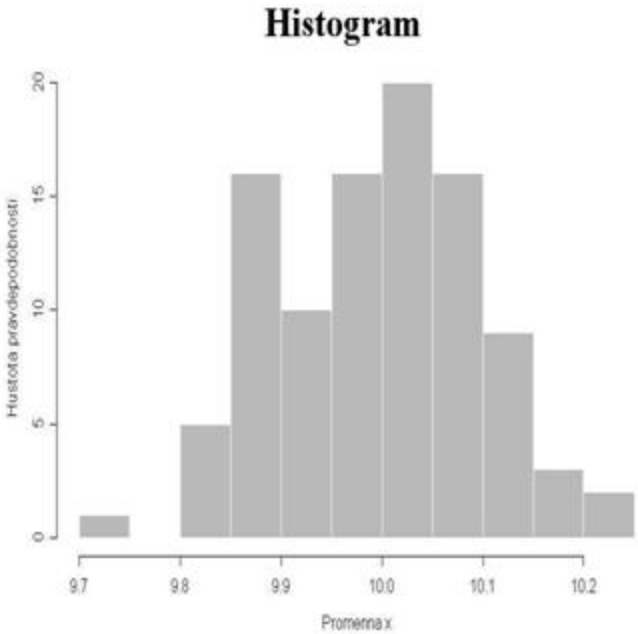
Histogram (G12)

Osa x: proměnná x,

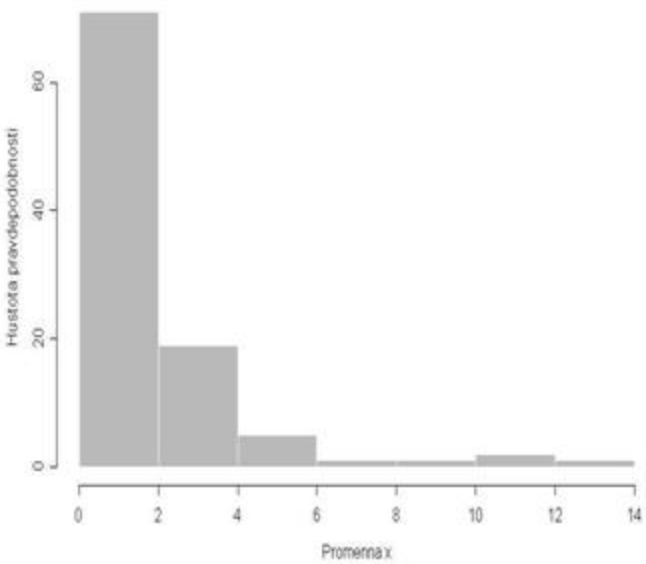
Osa y: hustota pravděpodobnosti



(A) Histogram s grafem hustoty pravděpodobnosti (čárkovaně), (B) kumulativní histogram četností, (C) polygon četností, a (D) polygon kumulativních četností



Histogram Gaussova normálního rozdělení ***norm*** N(10, 0.1)



Histogram logaritmicko-normálního rozdělení ***log*** LN(5, 2)

(G12)

Délka třídního intervalu:

$$\Delta x_j = x^+_{j+1} - x^+_j$$

Volba počtu tříd L:

$$L = \mathrm{int}(2\sqrt{n})$$

$$L = \mathrm{int}(2.46\,(n-1)^{0.4})$$

Optimální délka třídních intervalů:

$$\Delta x_{opt} = 3.49\,\frac{s}{n^{1/3}}$$

kde s je směrodatná odchylka.

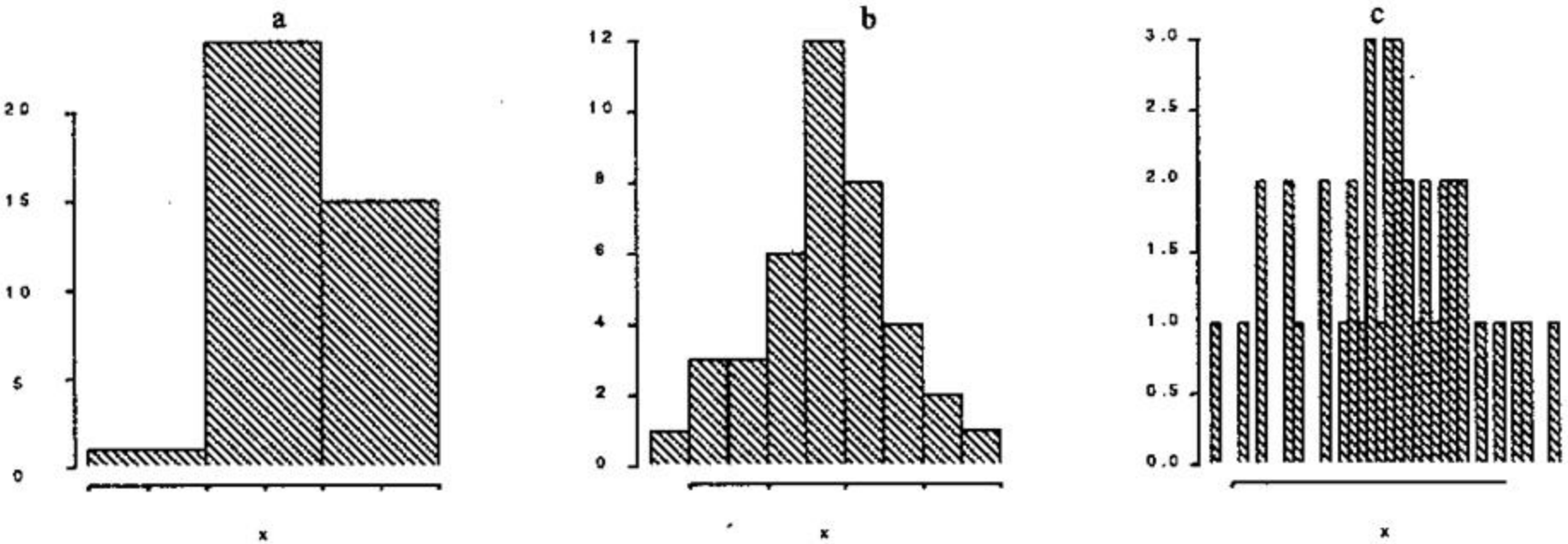
Robustní odhad délky třídních intervalů

$$\Delta x_{rob} = 2\,\frac{F_H - F_D}{n^{1/3}}$$

kde F_D, F_H jsou výběrové kvartily.

HISTOGRAM – diagram četností (G12)

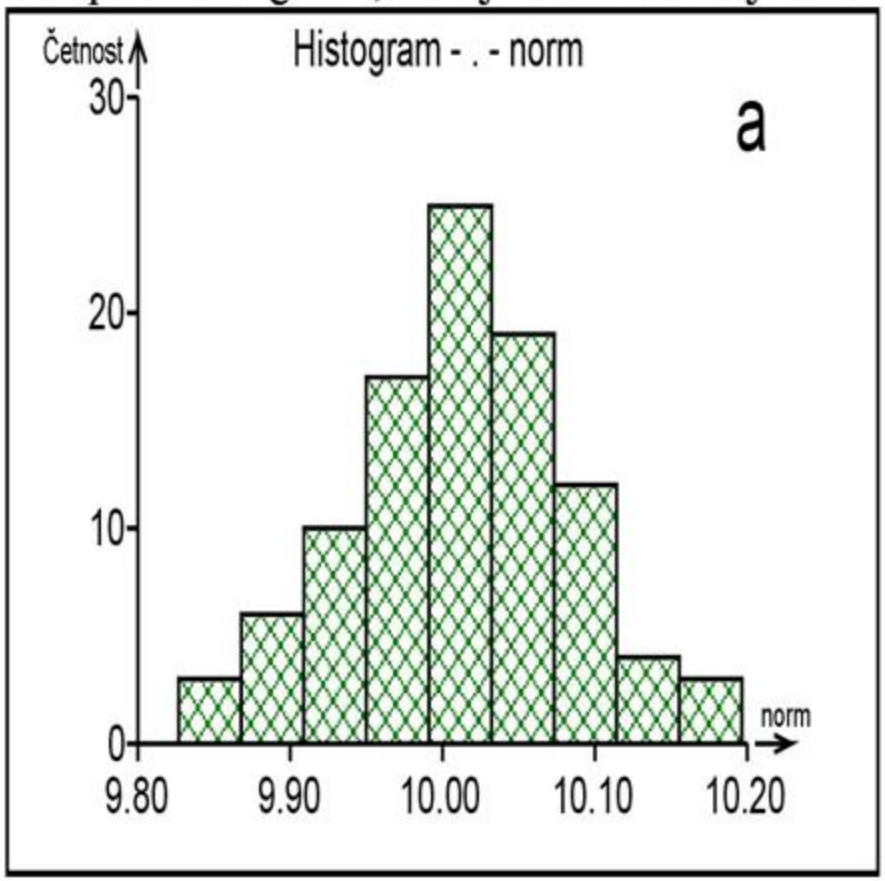
důležitá je správná volba šířky třídy:

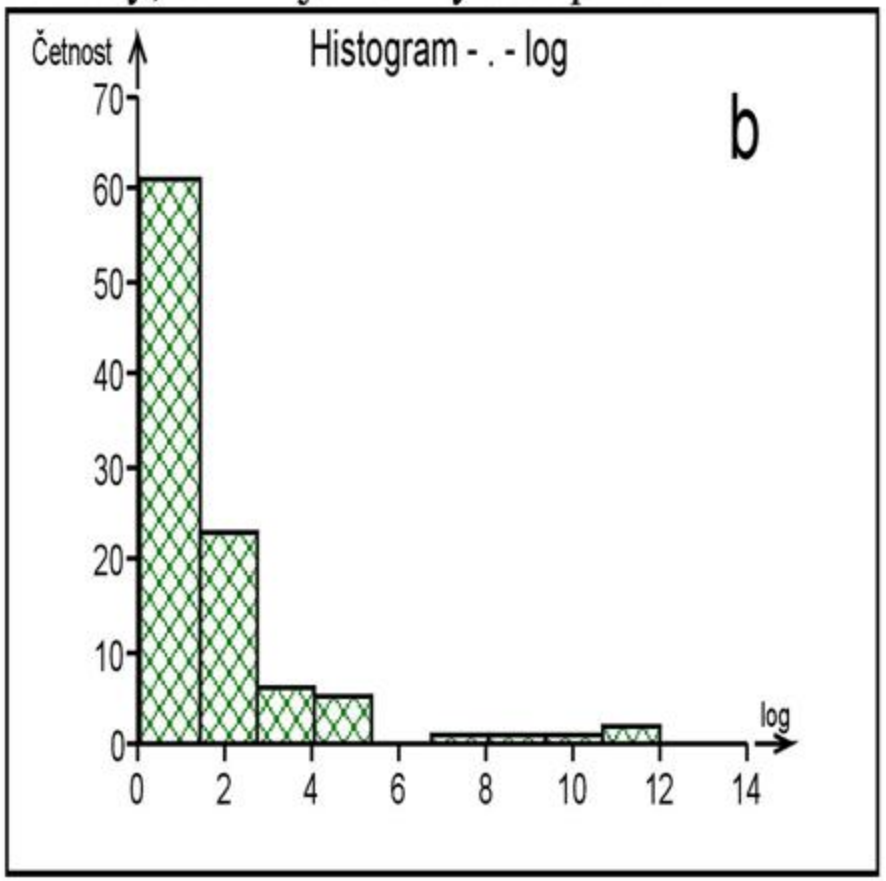


$$L = \mathrm{int}\left[2{,}46 \cdot (n-1)^{0{,}4}\right]$$

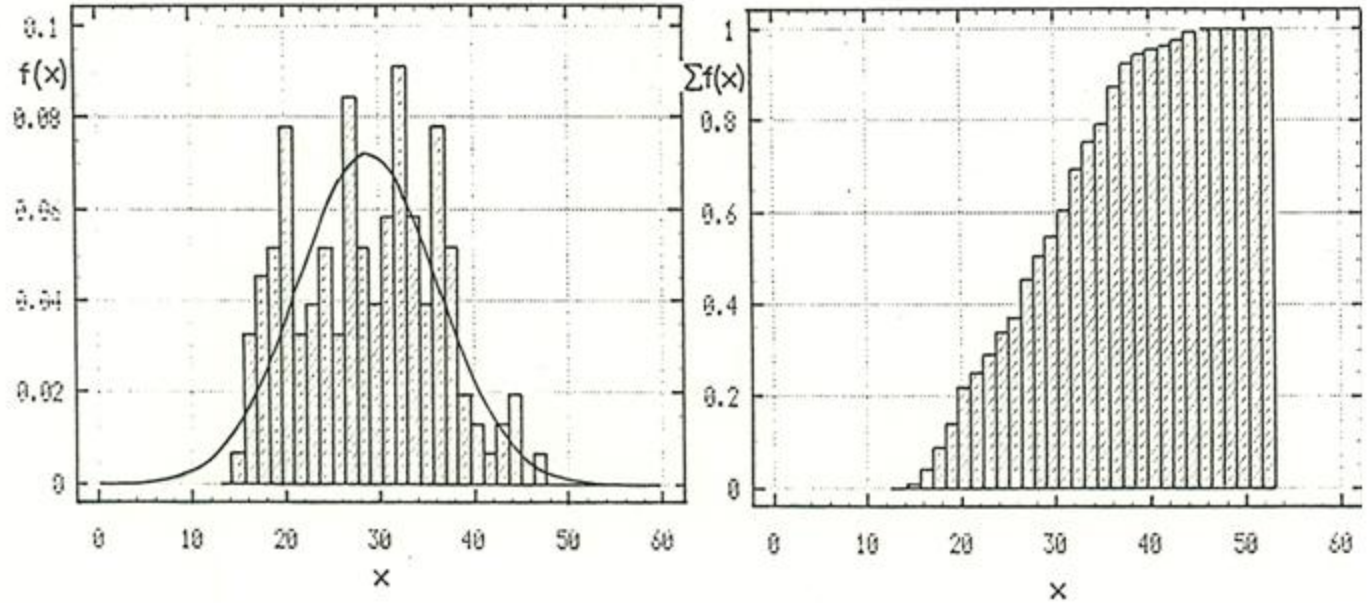
$$\boldsymbol{L} = \mathrm{int}\left(2 \cdot \sqrt{\boldsymbol{n}}\right)$$

Histogram (osa *x*: proměnná *x*, osa *y*: úměrná hustotě pravděpodobnosti). Jde o obrys sloupcového grafu, kde jsou na ose *x* jednotlivé třídy, definující šířky sloupců.



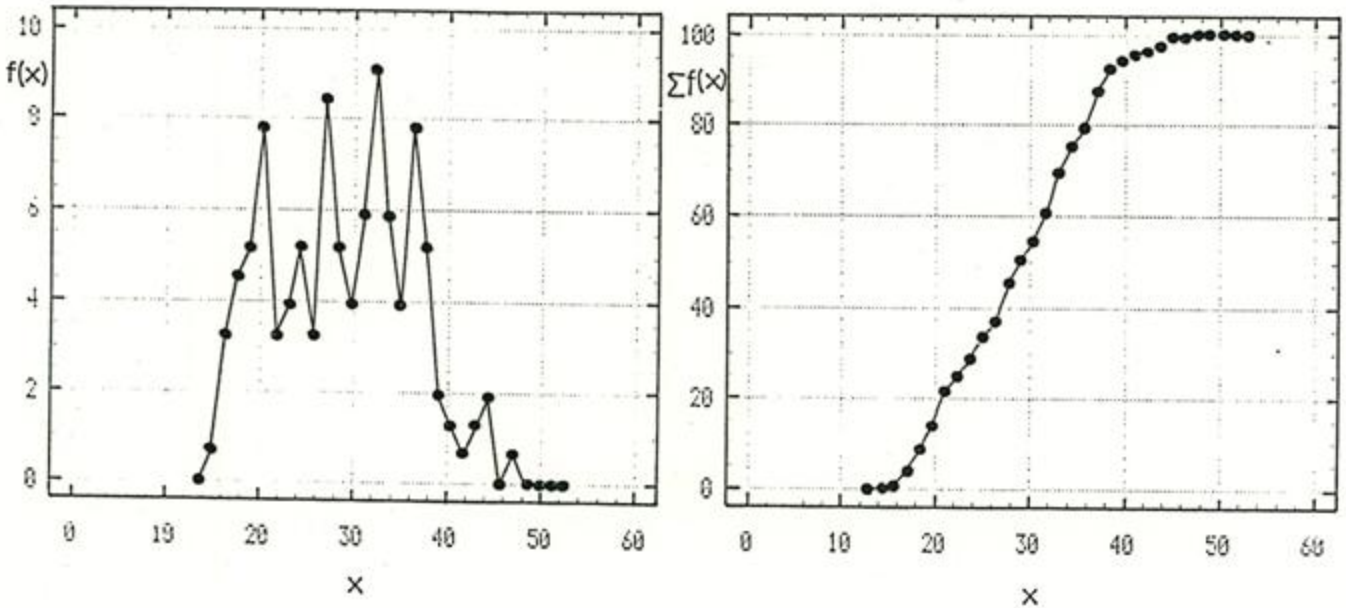


(G12)



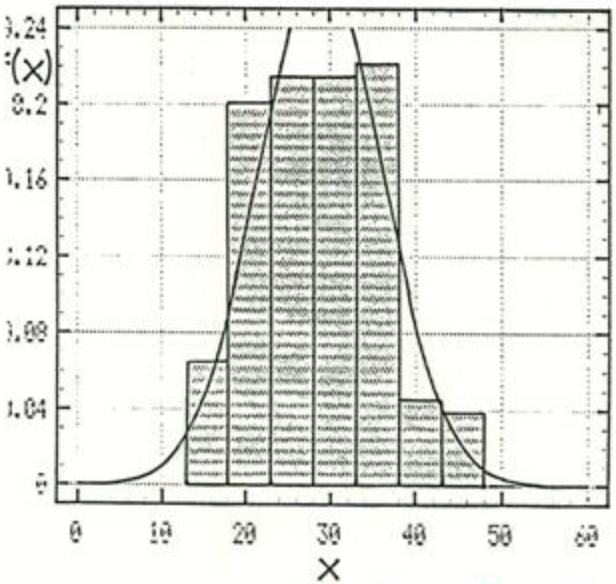
Chybně zvolená šířka tříd histogramu

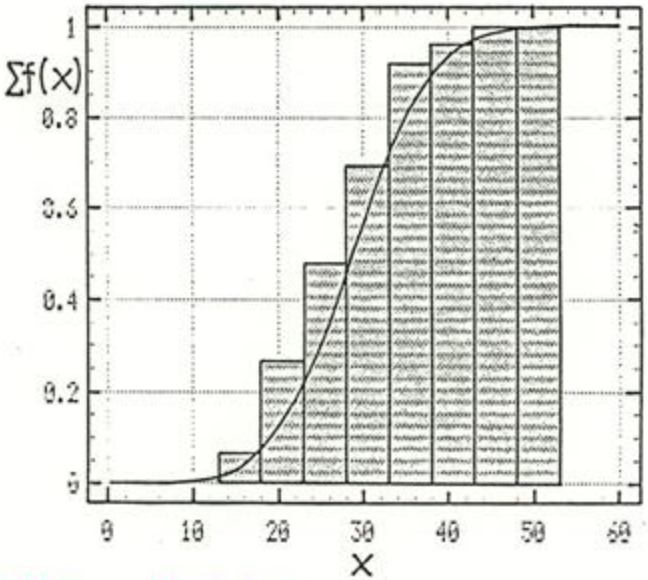
FREQUENCY POLYGON



HISTOGRAM

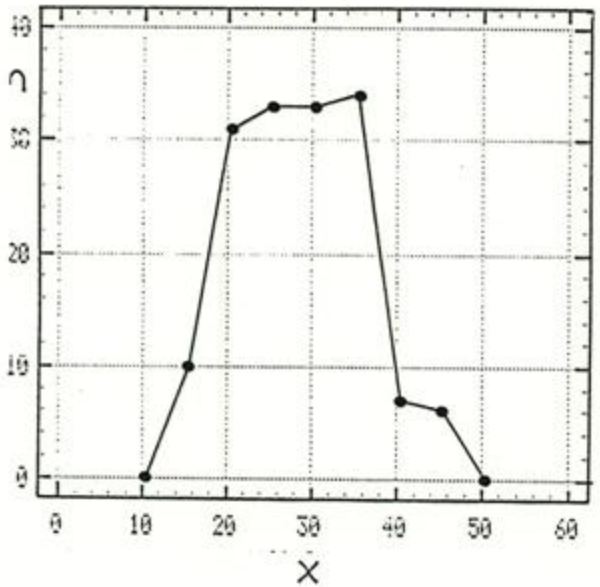
(G12)

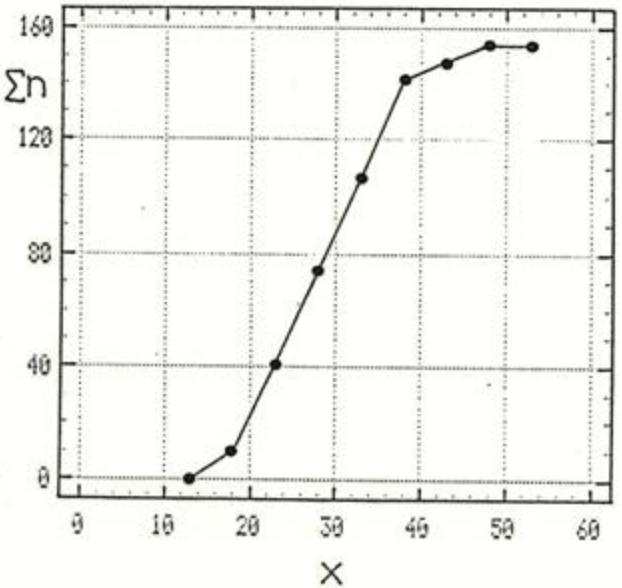




Správně zvolená šířka tříd histogramu

FREQUENCY POLYGON



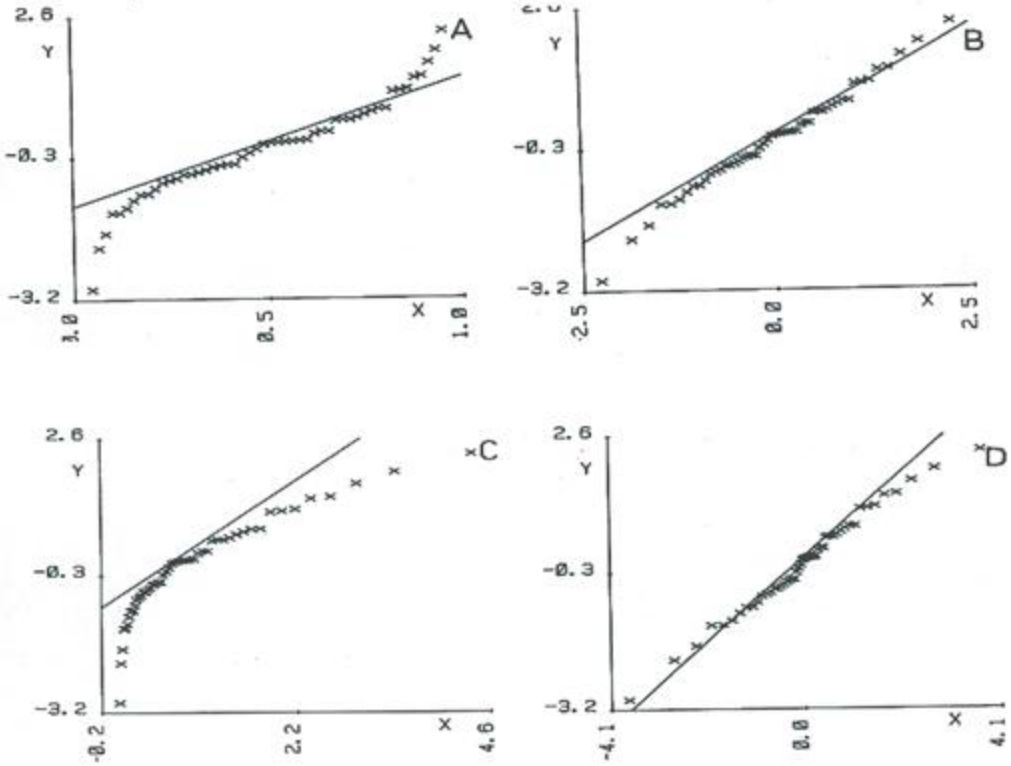


Kvantil-kvantilový graf (Graf Q-Q)

(G13)

Osa x: $Q_S(P_i)$,

Osa y: $x_{(i)}$



Grafy Q-Q pro porovnání rozdělení výběru normálního rozdělení s teoretickým rovnoměrným (A), normálním (B), exponenciálním (C) a Laplaceovým (D) rozdělením. Čarou je vyznačen teoretický průběh

$$x_{(i)} \approx Q_T(P_i)$$

kde P_i je pořadová pravděpodobnost.

Při shodě rozdělení s teoretickým, je $x_{(i)}$ na $Q_T(P_i)$ lineární funkcí.

Obvykle lze normovat použitím proměnné

(G13)

$$s = \frac{x - Q}{R}$$

kde Q je ***parametr polohy***,
R je ***parametr rozptýlení*** mající význam ***měřítka***.

Při shodě rozdělení s výběrovým, bude grafem $x_{(i)}$ vs. $Q_s(P_i)$ přímka

$$x_{(i)} = Q + R \,.\, Q_s(P_i)$$

Standardizované hustoty a distribuční funkce vybraných rozdělení a odpovídající souřadnice grafů Q-Q

Rozdělení	$F_T(s)$	$f_T(s)$	y	x
rovnoměrné	s	1	$x_{(i)}$	P_i
exponenciální				
$x \geq Q$	$1 - \exp(-s)$	$\exp(-a)$	$x_{(i)}$	$-\ln(1 - P_i)$
normální*)	$\Phi(s)$	$\frac{\exp(-0.5\, s^2)}{\sqrt{2\pi}}$	$x_{(i)}$	$\Phi^{-1}(P_i)$
Laplaceovo				
$x \leq Q$	$0.5 \exp(s)$	$0.5 \exp(s)$	$x_{(i)}$	$\ln(2\, P_i)$ pro $P_i \leq 0.5$
$x > Q$	$0.5\,(2 - \exp(-s))$	$0.5 \exp(-s)$	$x_{(i)}$	$-\ln(2\,(1 - P_i))$ pro $P_i > 0.5$
lognormální	$\Phi[\ln(s)]$	$\frac{\exp[-0.5 \ln(s)^2]}{\sqrt{2\pi}}$	$x_{(i)}$	$\exp[\Phi^{-1}(P_i)]$

(G13)

Posouzení linearity: metodou lineární regrese,

Diagnoza: konfidenční pásy, ve kterých bude se zvolenou pravděpodobností ležet skutečná kvantilová funkce

$$Q_T(P_i - C_{1-\alpha}) \leq x_{(i)} \leq Q_T(P_i + C_{1-\alpha})$$

kde $C_{1-\alpha}$ je parametr závislý na velikosti výběru a zvolené hladině významnosti, (pro 95% platí)

$$C_{0.95} = 1.385 \left[N + \frac{\sqrt{N + 4}}{3.5} \right]^{-1/2}$$

Pro $U = P_i$ jde přímo o kvantily u_{P_i}.

(G14)

Pořádkové statistiky $U_{(i)}$

$$U_i = \Phi\left[\frac{x_i - \hat{\mu}_R}{\hat{\sigma}_R}\right]$$

kde $\Phi(x)$ značí distribuční funkci stand. normálního rozdělení,

$$\hat{\mu}_R = \tilde{x}_{0.5}$$

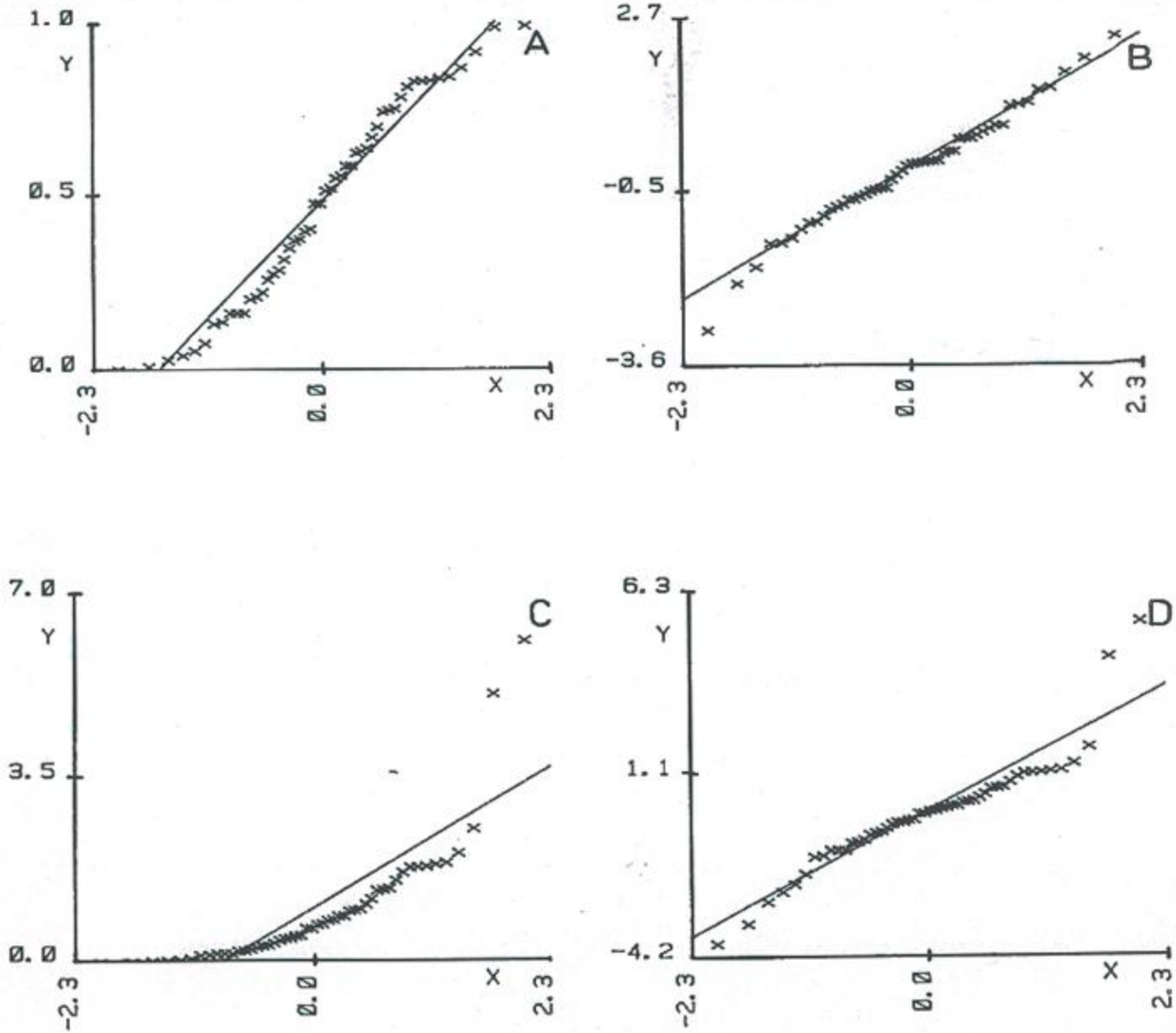
$$\hat{\sigma}_R = 0.75\ (\tilde{x}_{0.75} - \tilde{x}_{0.25})$$

$$U_{(0)} = 0 \text{ a } U_{(n+1)} = 1.$$

Diagnoza:

- lineární závislost dokazuje normalitu rozdělení výběru

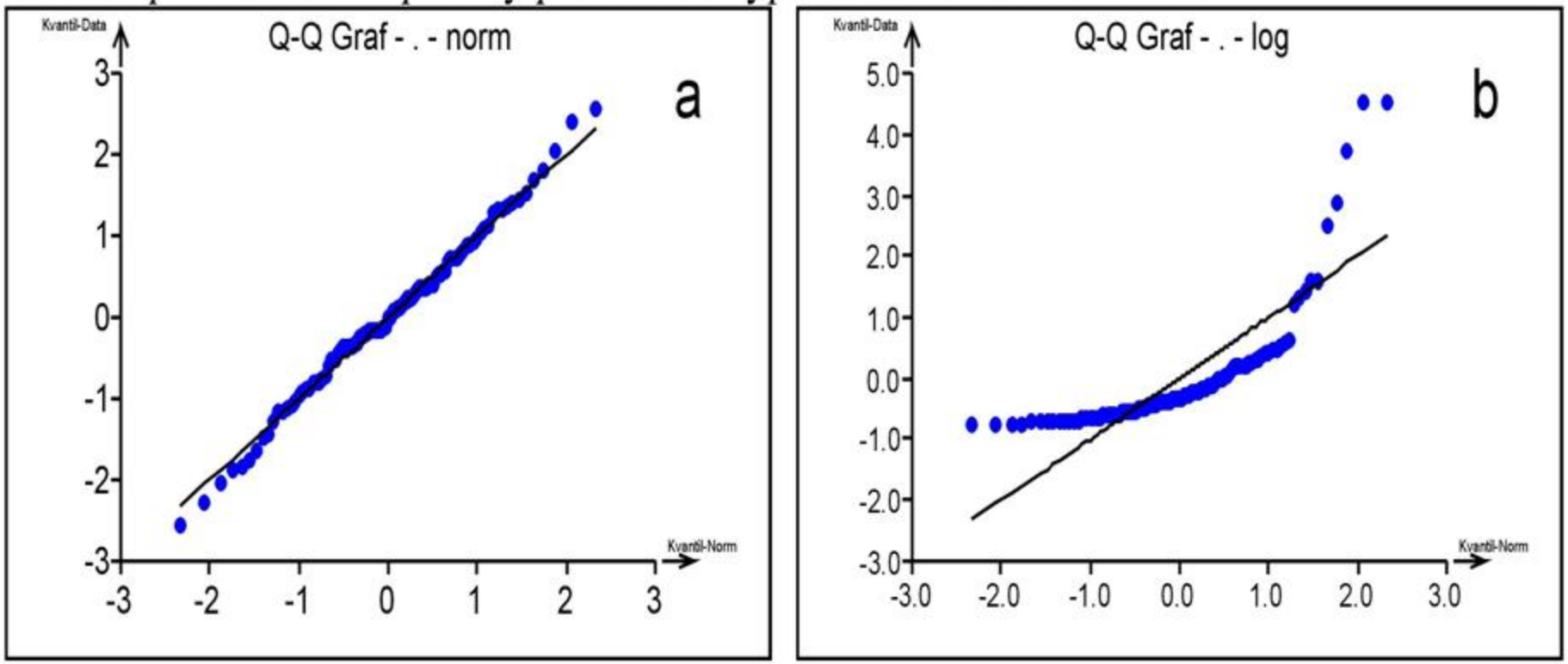
(G14)



Rankitové grafy pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova

(G13)

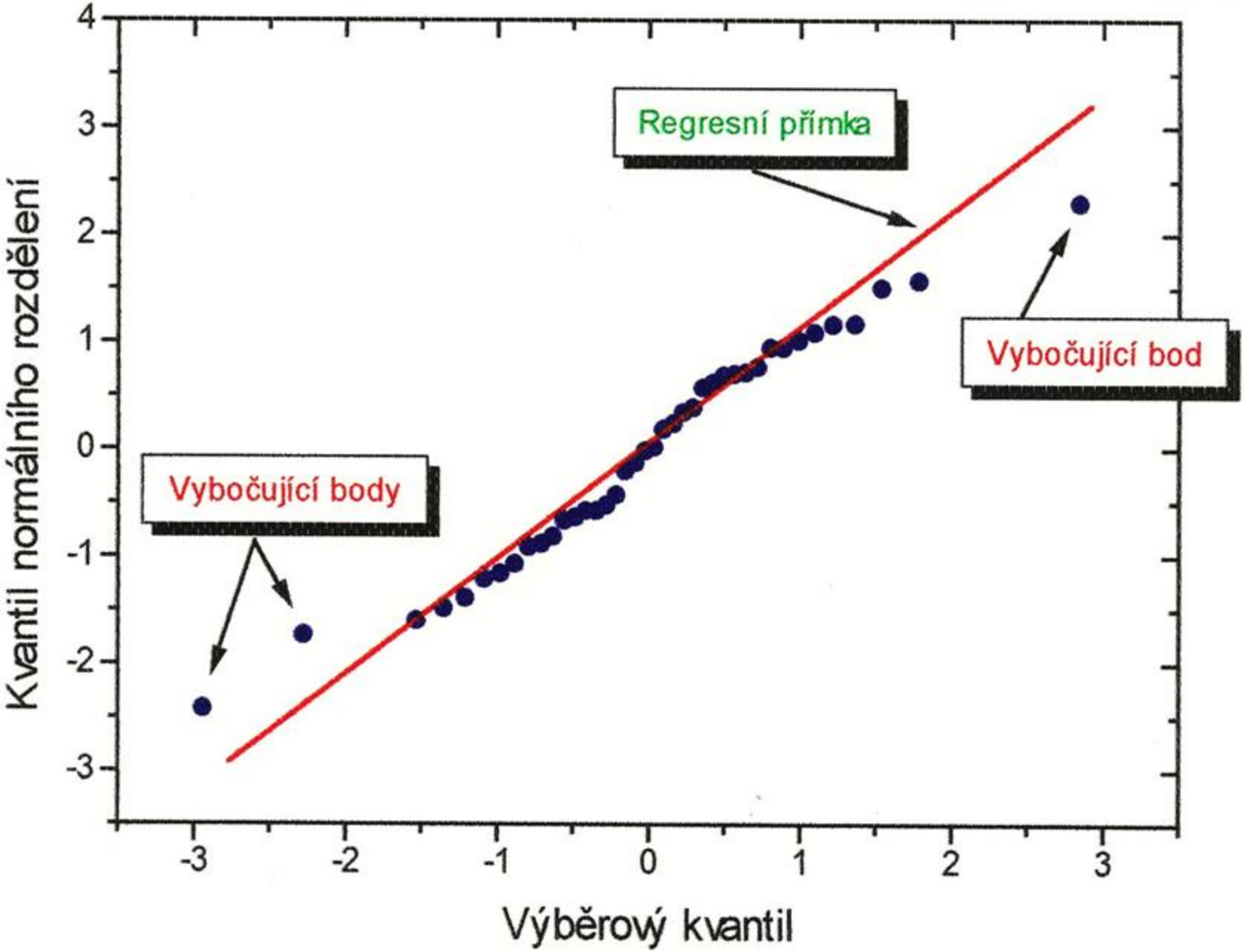
Kvantilově-kvantilový graf (graf *Q-Q*) (osa x: $Q_T(P_i)$, osa y: $x_{(i)}$). Umožňuje posoudit shodu výběrového rozdělení, jež je charakterizováno kvantilovou funkcí $Q_E(P)$ s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$. Korelační koeficient r_{xy} je pak kritériem těsnosti proložení této přímky při hledání typu neznámého rozdělení.



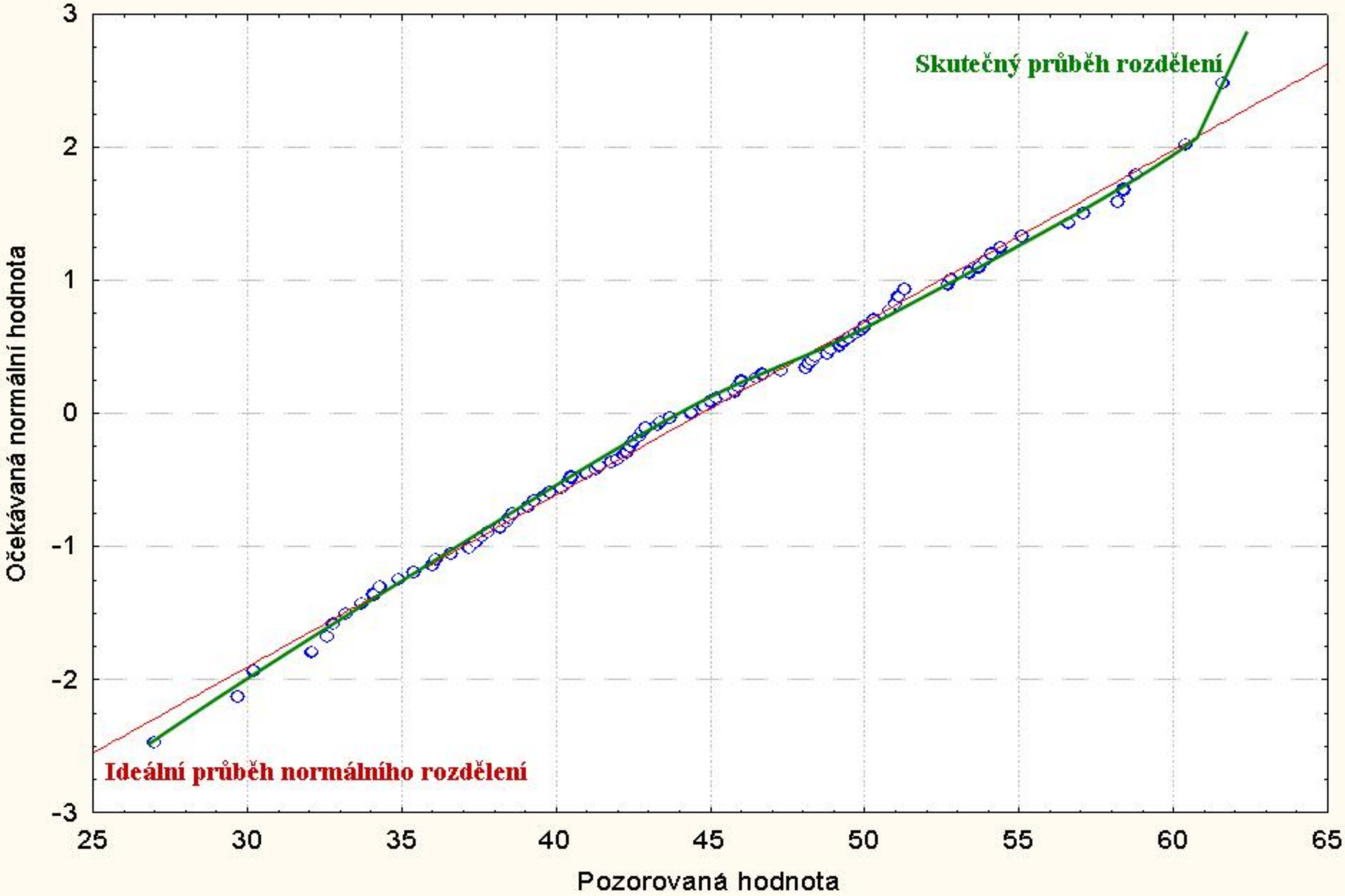
Obr. 2.15 Rankitový čili kvantil-kvantilový graf (*Q-Q* graf) pro ověření shody s teoretickým normálním rozdělením: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Q-Q graf

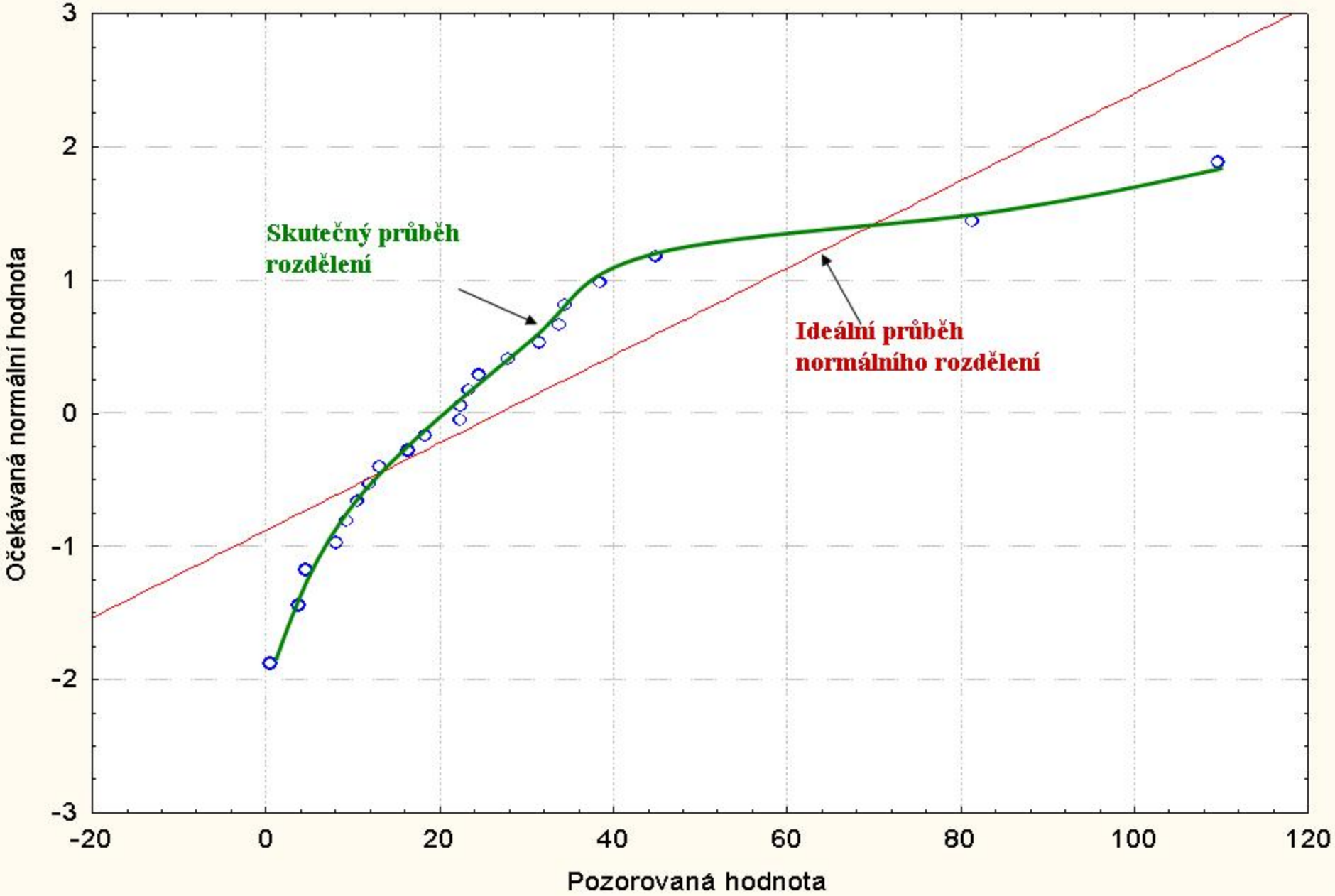
(G14)


Regresní přímka
Vybočující bod
Vybočující body
Kvantil normálního rozdělení
Výběrový kvantil
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

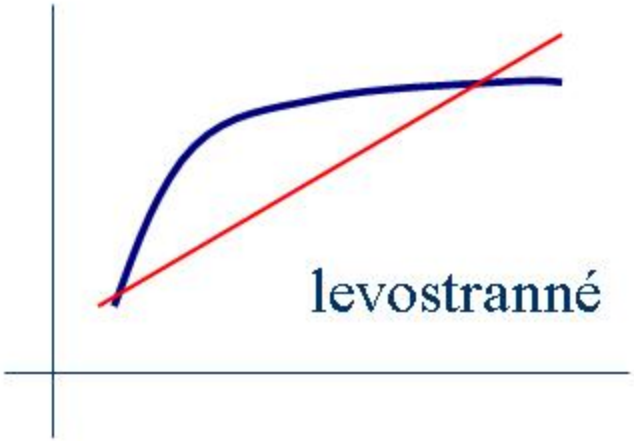
Příklad: **Porovnání rozdělení výběru u dobré shody s ideálním průběhem pro předvolené normální rozdělení** **(G14)**

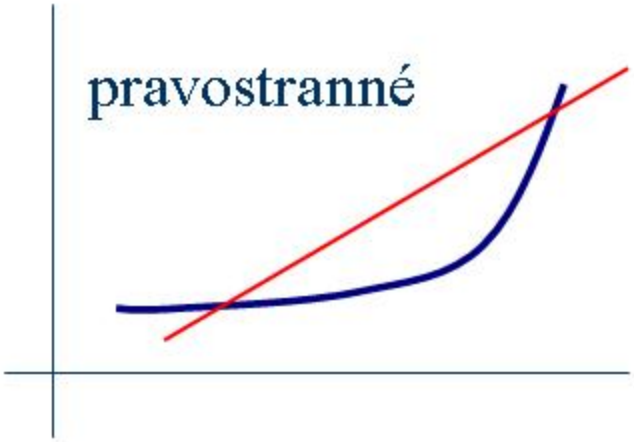


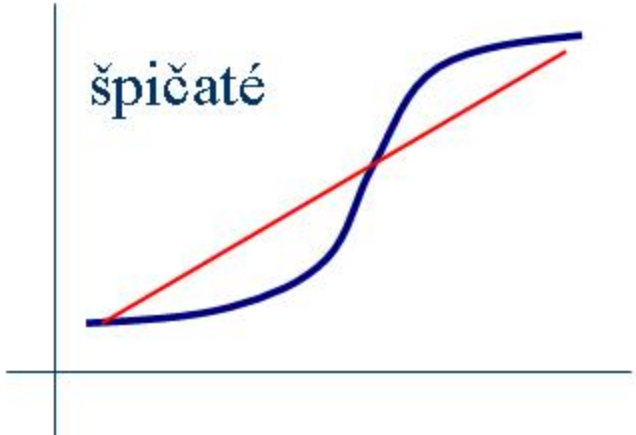
Příklad: **Porovnání rozdělení výběru u špatné shody skutečného průběhu rozdělení s ideálním průběhem pro normální rozdělení** **(G14)**

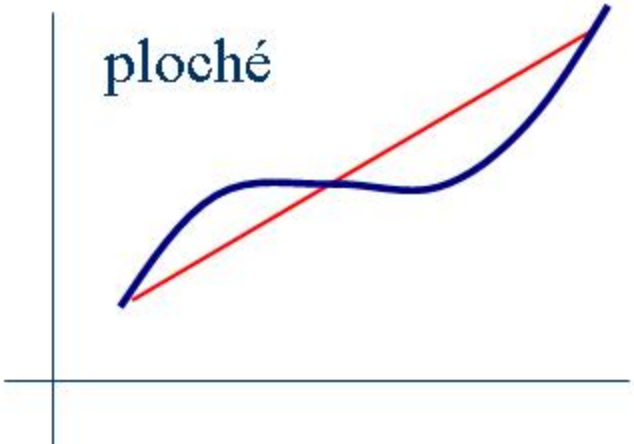


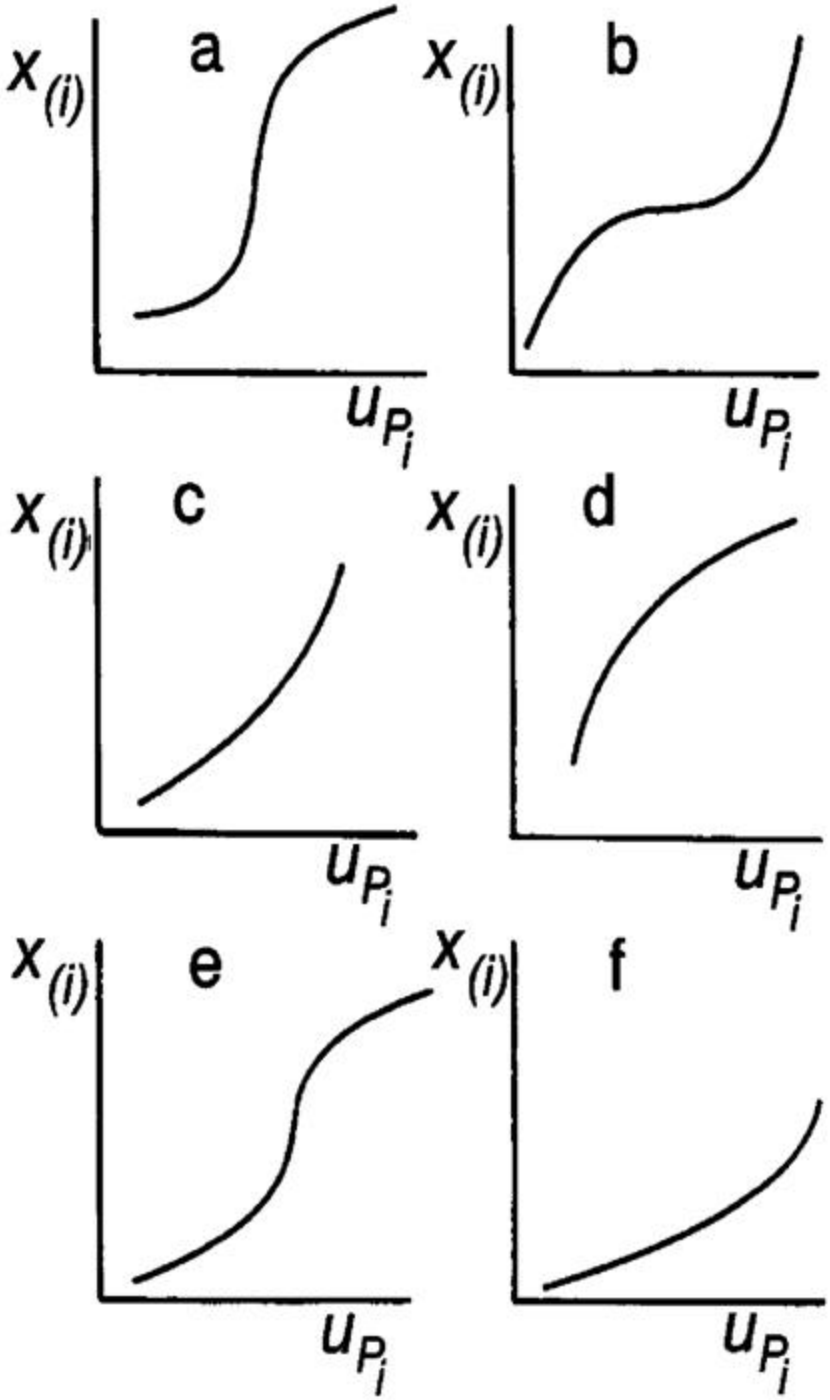
Diagnostikování rozdělení kvantil-kvantilovým grafem











Rankitový graf (G14)

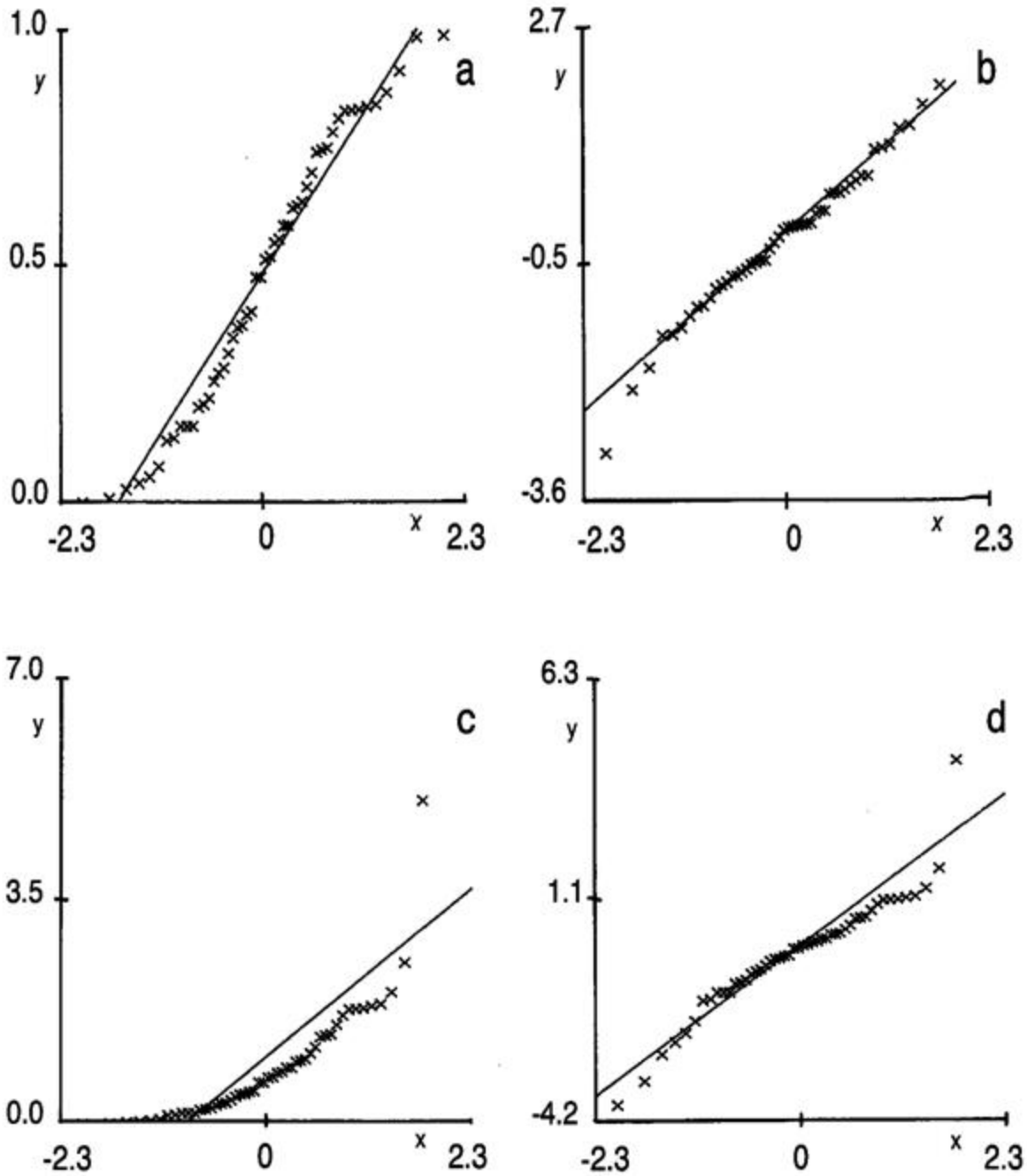
K porovnání rozdělení výběru s rozdělením normálním se Q-Q graf nazývá ***rankitový graf.***

Umožňuje zařazení výběrového rozdělení do skupin podle šikmosti, špičatosti a délky konců.

Konvexní, popř. konkávní, průběh Q-Q grafu indikuje zešikmené rozdělení výběru, zatímco esovitý průběh ukazuje na rozdílnost v délce konců ve srovnání s normálním rozdělením.

Je možné indikovat i směs normálních rozdělení nebo přítomnost vybočujících bodů.

(G14)



Rankitové grafy pro výběry z rozdělení:

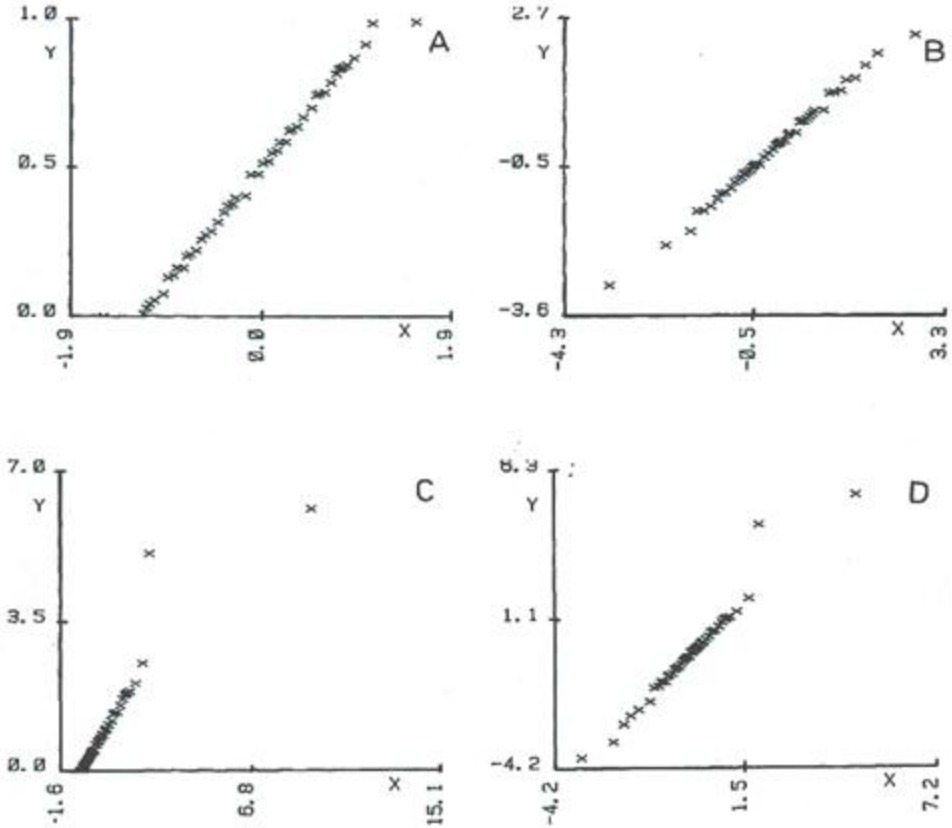
a) rovnoměrného, b) normálního, c) exponenciálního a d) Laplaceova.

Plnou čarou je vyznačen teoretický průběh.

Podmíněný rankitový graf (G15)

Osa x: $\Phi^{-1}\,[0.5\,(U_{(i-1)} + U_{(i+1)})]$,

Osa y: $x_{(i)}$

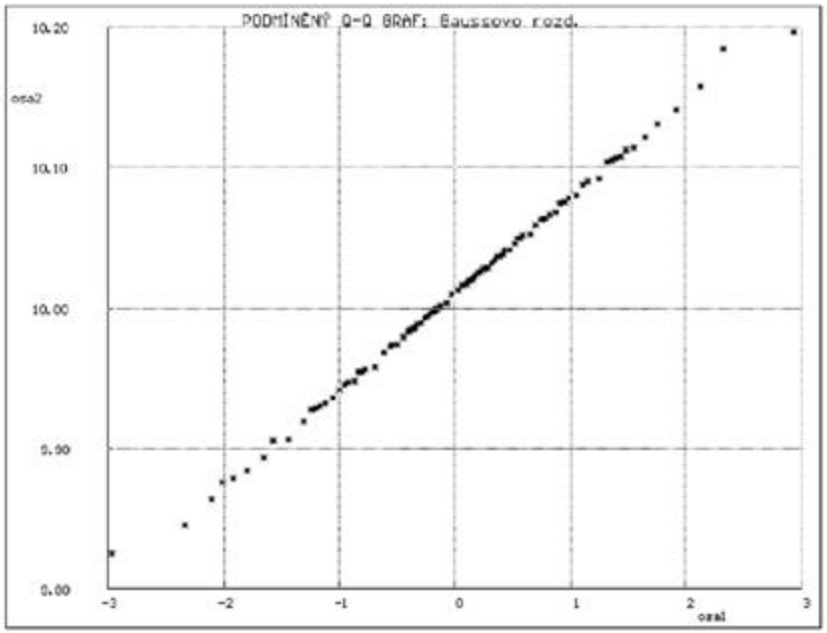


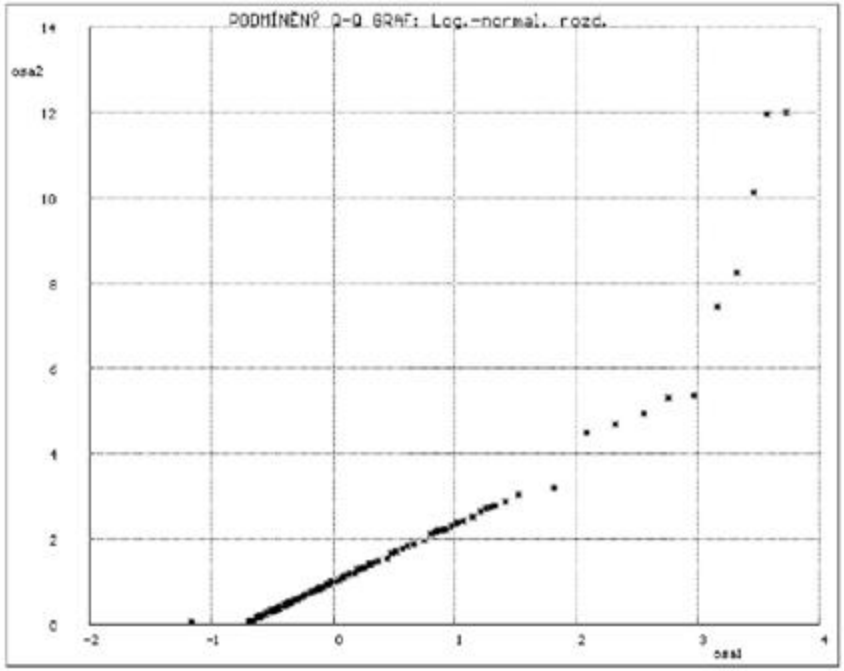
Podmíněné rankitové grafy pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova

Symbol $\Phi^{-1}(U)$ značí standardizovanou kvantilovou funkci normálního rozdělení.

(G15)

Podmíněný rankitový graf (osa x: Φ^{-1} [0.5 $(U_{(i-1)} + U_{(i+1)})$] , osa y: $x_{(i)}$). Přibližná lineární závislost je v podmíněném rankitovém grafu důkazem normality testovaného rozdělení výběru. Z grafu normálního rozdělení je patrná výrazně menší lokální variabilita ve srovnání s rankitovými grafy.



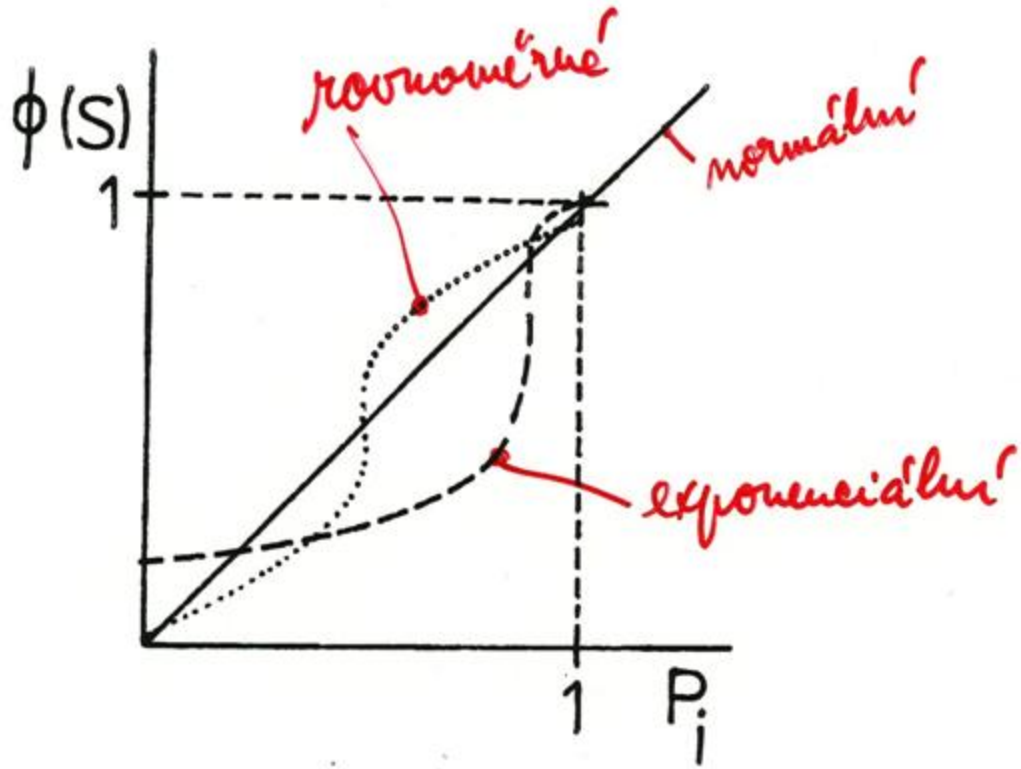


Obr. 2.16 Podmíněný rankitový graf pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Pravděpodobnostní graf (P-P graf) (G16)

Osa x: P_i,

Osa y: $F_T(S_{(i)})$



P-P graf pro ověření normality se zakreslenými křivkami pro normální (_), exponenciální (---) a rovnoměrné (...) rozdělení.

(G16)

K porovnání distribuční funkce výběru se standardizovanou distribuční funkcí teoretického rozdělení.

Standardizovaná proměnná je

$$S_{(i)} = \frac{x_{(i)} - Q}{R}$$

kde Q je parametr polohy a R je parametr rozptýlení

(G16)

Diagnoza:

(a) P-P grafy jsou citlivé na odchylky od teoretického rozdělení v okolí módu,

(b) Q-Q grafy jsou citlivé na odchylky od teoretického rozdělení v oblasti konců.

Postup:

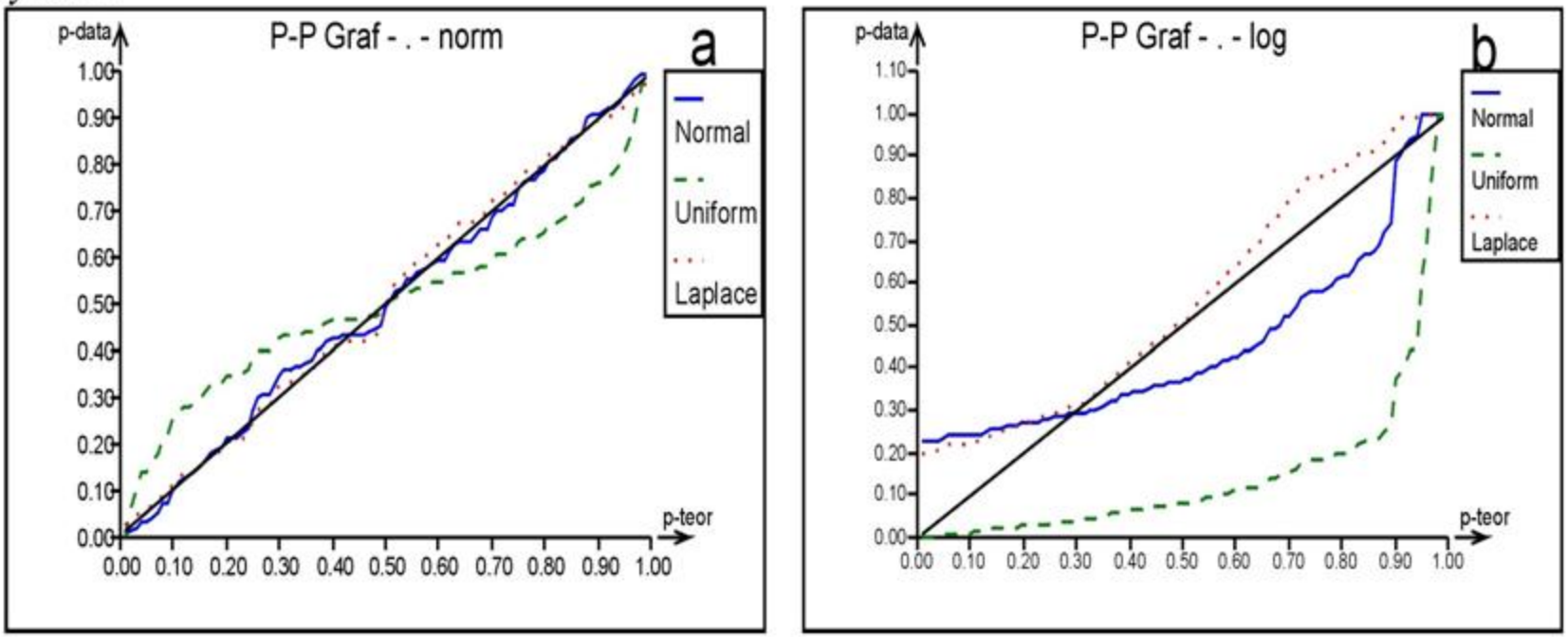
1. Osa x: $P_i = i/(n + 1)$, osa y: $\Phi[(x_{(i)} - \bar{x}) / s]$,
2. Výběr k hodnot (cca k = 200) a výpočet $P_i = i/(k + 1)$.
3. Generace standardizovaných kvantilů $Q_{si} = G^{-1}(P_i)$ a Q, R pro rozdělení $F_T(.)$. Pro normální rozdělení

$$\hat{Q} = \frac{1}{k}\sum_{i=1}^{k} Q_{si} \quad \text{a} \quad \hat{R} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(Q_{si} - Q)^2}$$

4. Kreslení závislosti $F_T\,[(Q_{si} - \hat{Q}) / \hat{R}]$ vs. P_i.

(G16)

Pravděpodobnostní graf (***P-P* graf**), (osa x: P_i, osa y: $F_T(S_{(i)})$. Slouží k porovnání distribuční funkce výběru, vyjádřené přes pořadovou pravděpodobnost, se standardizovanou distribuční funkcí zvoleného teoretického rozdělení. Standardizovaná proměnná je zde definována vztahem $S_{(i)} = (x_{(i)} - Q)/R$, kde Q je *parametr polohy* a R je *parametr rozptýlení*.



Obr. 2.17 Pravděpodobnostní graf (*P-P* graf) pro výběry: (a) *norm*, symetrického (Gaussova, normálního), a (b) *log*, asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Kruhový graf (G17)

Diagnoza: slouží k ověření, že výběr pochází z rozdělení $F_e(x)$, Transformací

$$Z_{(i)} = F_e(x_{(i)})$$

vyjdou náhodné veličiny $Z_{(i)}$ rozdělené přibližně rovnoměrně na intervalu [0, 1].

Konstrukce: soustava vektorů $\bar{V}_i$ o stejné délce

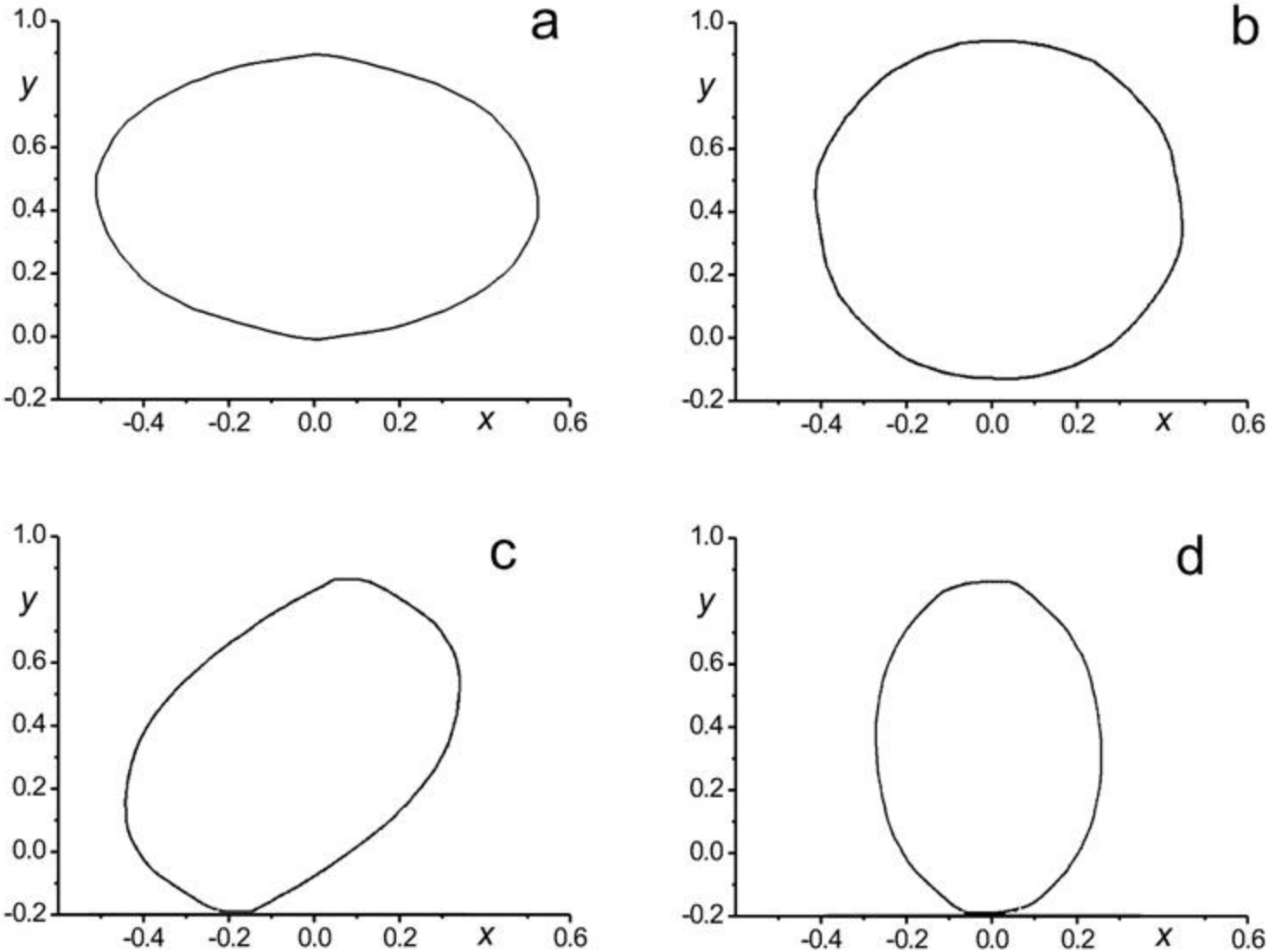
$$l_0 = \left[\frac{N\,(N-1)}{2}\right]^{-1/2}$$

a směru $\pi\, Z_{(i)}$.
x-ovou a y-ovou složka vektoru $\bar{V}_i$:

$$V_{x_i} = l_0 \cos(\pi\, Z_{(i)}) \qquad V_{y_i} = l_0 \sin(\pi\, Z_{(i)})$$

Úhly se uvažují v radiánech.

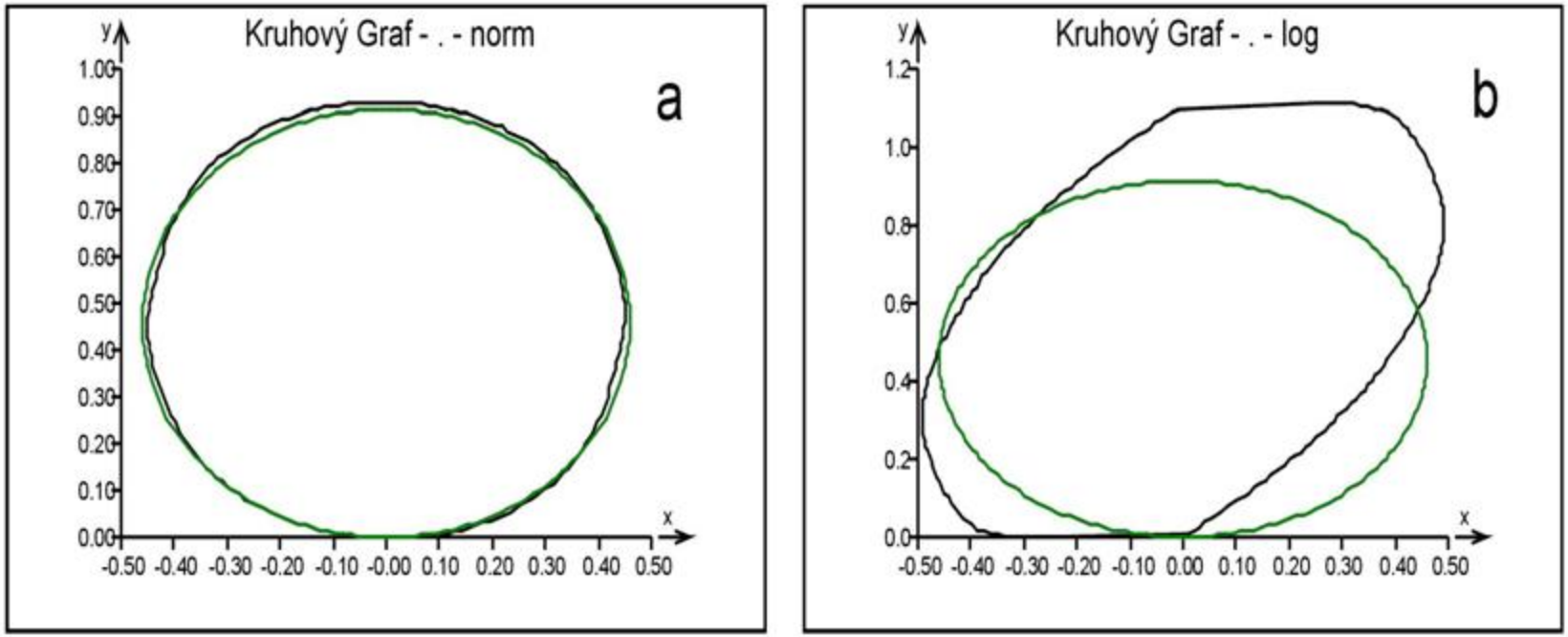
(G17)



Kruhový graf pro výběry z rozdělení a) rovnoměrného, b) normálního, c) exponenciálního a d) Laplaceova.

(G17)

Kruhový graf slouží k vizuálnímu ověření hypotézy, že výběr pochází ze symetrického (nejčastěji Gaussova) rozdělení. Odchylky od kružnice ukazují na jiné než symetrické rozdělení výběru: (a) protáhlý elipsovitý tvar, s hlavní osou umístěnou úhlopříčně, ukazuje na asymetrické rozdělení, (b) elipsovitý tvar podél osy *x* ukazuje na rovnoměrné rozdělení.



Obr. 2.18 Kruhový graf pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Analýza cholesterolu u 8 tisíc pacientů